Imagine a group of cars are obliterated in a parking lot explosion. Based on license plates and hood ornaments, it's easy to determine the type and number of cars that were involved. But to go beyond this, to determine which starter corresponds to which vehicle or which car had the fuzzy dice hanging from the rear-view mirror, that's a tougher nut to crack.

In a sense, this is what happens in the proteomics workflow known as “bottom up,” in which protein mixtures are digested into peptides, fractionated, and then sequenced and characterized in a tandem mass spectrometer. From these data researchers can determine which proteins were present in the sample, and whether or not they were phosphorylated.

But suppose one protein actually contained phosphorylation sites at either end of its sequence. This protein can exist in any of four states: completely unphosphorylated, doubly phosphorylated, or phosphorylated at either site, explains Detlev Suckau, head of MALDI Applications and Proteomics Research & Development at Bruker Daltonik in Bremen, Germany.

"A bottom-up approach does not distinguish between the four forms," says Suckau. It can identify them, but not determine how frequently or under what conditions they appear. Nor can it determine whether those two modifications coexist on the same protein at the same time, or are mutually exclusive.

To answer those questions, researchers need the so-called “top-down” approach. Top-down proteomics explores proteins without digesting them—that is, from the “top down.” Because top-down analyses consider intact proteins rather than peptide pools, they can distinguish our four hypothetical isoforms based on their different molecular weights—not to mention closely related family members and splicing isoforms that bottom-up strategies often cannot differentiate.

Researchers have used top-down approaches on individual proteins or protein families, and are pushing toward the proteomics scale, where it can interrogate hundreds or even thousands of proteins in a single sample. The approach even has a dedicated research alliance: the Consortium for Top Down Proteomics was formed in March 2012 with the goal of “promot[ing] innovative research, collaboration and

Using this four-dimensional, liquid-phase separation strategy, Neil Kelleher's team at Northwestern characterized some 3,000 human protein isoforms, the largest top-down proteomics study to date. Courtesy of Neil Kelleher

education accelerating the comprehensive analysis of intact proteins.”

But making that transition won't be easy. The technique challenges mass spectrometrists on every level—from protein separation, to the mass spectrometry itself, to bioinformatics.

The top-down workflow clearly has the upper hand when it comes to proteome analyses: If post-translational modifications are molecular toggles that tweak protein activity, then it makes sense that researchers would want to study the different isoforms that wax and wane as cells develop, respond to stimuli, and die.

Yet because it requires separating and sequencing large protein isoforms differing by just a methyl group or two, the workflow demands the highest of high-end instrumentation, the kind of mass specs that can weigh molecular masses out to two or three decimal places.

Ljiljana Paša-Tolić, an investigator in the Consortium for Top Down Proteomics and lead for mass spectrometry at the Environmental Molecular Sciences Laboratory, uses a Fourier-transform ion cyclotron resonance (FTICR) mass spectrometer to drive her top-down analyses.

Driven by massive, supercooled, superconducting magnets, FTICRs are like the Lamborghinis of mass spectrometry, offering the very highest resolution and mass accuracy. “In FTICR all of the performance parameters scale with the magnetic field linearly or quadratically,” explains Paša-Tolić. “So, the higher the field, the better performance you have.”

Currently, her lab uses a 15-Tesla instrument from Bruker Daltonics, with which she can routinely study LC-separated proteins of 70-80 kDa. But using a 21-Tesla instrument, currently under construction with Department of Energy funding, Paša-Tolić estimates they may be able to push that to 150 kDa. “That would then enable us to assess almost the whole human proteome,” she says.

Consortium member and top-down evangelist Neil Kelleher of Northwestern University, who has long advocated FTICRs for top-down proteomics, recently switched to a Thermo Fisher Scientific Orbitrap Elite, a high-resolution, high-mass accuracy instrument that's in many ways a cross between a traditional ion trap and an FTICR, but with slightly lower resolution.

Although many credit Thermo's Orbitrap line as a "game-changer" for proteomics in general, Kelleher says the Elite is the first mass spectrometer in that

line to have the speed and performance to make top-down practical. “The Elite changed everything about top-down on an Orbitrap. It really is a sea change.”

In part, he explains, that’s because proteins in a mass spectrometer don’t really behave like big peptides. They are harder to fragment and take longer to weigh accurately, for instance. “I’m here to tell you that a 40, 50-plus charge state, a big fat protein, is different than a little 2+ tryptic peptide,” he says.

Using an Orbitrap Elite (and a 12-T FTICR-MS), Kelleher’s team in 2011 performed the most comprehensive top-down proteomics analysis to date: the identification, quantification, and characterization of some 1,043 proteins in 3,000 isoforms from human HeLa cells, along with similar-scale analyses on two other cell lines (1).

We’re looking for a few good columns

Kelleher’s analysis required more than just a good mass spec though—the key was a comprehensive liquid-phase separation strategy that could ease the protein load on the mass spectrometer.

High-resolution instruments are relatively slow. The simpler the sample, the more time the instrument can devote to each protein. Liquid chromatography is the easiest way to do this, but unlike peptides, which are relatively simple to fractionate, intact proteins don’t separate easily. Rather than resolving as sharp chromatographic lines, intact proteins tend to slowly extrude from columns, producing broad, complex peaks.

When it comes to top-down proteomics, says Catherine Fenselau, professor of chemistry and biochemistry at the University of Maryland, “the big hang-up” is chromatography. Steven Patrie, an assistant professor of pathology at the University of Texas Southwestern Medical Center, is pursuing one alternative. His team recently combined “superficially porous” reversed-phase resins with capillary columns for LC/MS (SPLC/MS). SP resins have a thin porous shell but a non-porous center. Such a configuration maintains protein surface interactions at the outer most layers of the resin, thus providing the speed of non-porous materials while maintaining the high loading capacity of porous ones.

Using one such column—a 75-micron capillary column packed with 5-micron C18 beads—Patrie and his team showed that hundreds of intact HeLa cell proteins from 10,000 to 50,000 cells could be characterized in as little as five minutes, with low attomole sensitivity (2). That’s the

Jennifer Brodbelt, chemist at the University of Texas, Austin, is developing an alternate fragmentation approach that uses laser power to shatter proteins. Courtesy of Jennifer Brodbelt.

kind of performance researchers get from bottom-up approaches, notes Patrie. “If peptides are being detected at an attomole level, and you can also detect a protein at an attomole level, why bother doing the digestion?” he says rhetorically.

Paša-Tolić’s team devised what they call “a novel online two-dimensional liquid chromatography–tandem mass spectrometry platform” for histone analysis that couples reverse-phase HPLC with weak cation exchange hydrophilic interaction chromatography to separate histone populations first into families (eg, H2A, H2B, H3, and H4) and then by post-translational modification.

“[With] H4 it looks really pretty,” says Paša-Tolić. “You see zero to five different acetylation states. And then within each acetylation state, you have phosphorylations, methylations, and these are not particularly well separated, but you can still get MS/MS data on them.” From 7.5 micrograms of purified histones, her team identified some 700 unique post-translational isoforms.

Kelleher’s analysis used an even more complicated, four-dimensional strategy, separating first by isoelectric point, then size, then hydrophobicity, and finally in the mass spectrometer itself.

The resulting data are beautifully resolved. In one figure, the team documented the shifting post-translational modifications in the HMGA1a protein during senescence. Phosphorylated forms slowly disappear while methylated forms multiply, and a dimethylated form of the protein arose during senescence that was not present in control cells.

Still, the overall study exhibited a strong bias against proteins greater than 50-kDa. “That’s where top-down needs to go,” says Kelleher, “is to really do better at high mass.”

Fragmentation

When Kelleher talks about proteomics data, he distinguishes between protein identification and characterization. “If you completely characterize the protein, that means you know all its splice variants...as well as post-translational modifications, as well as any polymorphisms or mutations that occur in the protein.”

To obtain that kind of information, top-down proteomicists need two pieces of data: the mass of the intact protein and its fragmentation products, as well as its sequence. The latter information arises from tandem MS/MS, in which an intact protein is then fragmented into smaller pieces.

At present, there exists an alphabet soup of fragmentation strategies for mass spectrometers, including collisionally induced dissociation (CID) and high-energy C-trap dissociation (HCD). One popular strategy for top-down work is electron transfer dissociation (ETD), which is basically a chemical reaction inside the fragmentation chamber that causes peptide backbones to break while leaving post-translational modifications intact.

But ETD, says Fenselau, is very hit-or-miss, working “really well for some proteins but not well at all for others.” Instead, Fenselau says she is “very keen to get [her] hands on” a new approach, from Jennifer Brodbelt, a chemist at the University of Texas, Austin.

Brodbelt uses laser power to shatter proteins. In earlier work Brodbelt introduced an infrared laser into an ion trap mass spectrometer, an approach called “infrared multiphoton dissociation” (3). Now she is using an ultraviolet laser and an Orbitrap mass spectrometer to do the same thing with greater efficiency and at higher-resolution.

High-energy UV photons, Brodbelt explains, can fragment peptides in ways that other technologies cannot, perhaps shedding light on tightly folded regions that are refractory to these other approaches. “Pieces of the protein that might have not cleaved well, now might cleave better using this photon-based method,” she says.

The Killer App

Fenselau uses top-down proteomics to study the histones carried by extracellular vesicles called exosome. She also uses it to identify and speciate unsequenced bacteria. To do that, her team applies a combination of MALDI-TOF and liquid-phase Orbitrap mass spectra to first identify, and then sequence, protein biomarkers in crude bacterial lysates.

“We put the whole bacterial lysate into this column, and as the proteins elute we

DON'T OVERLOOK VALUABLE INFORMATION IN YOUR GENE EXPRESSION DATA.

IPA®

Provides biological context for faster, more confident insights.

NEW!

NOW SUPPORTING:
RNA-Seq and
microRNA workflows

www.ingenuity.com

INGENUITY®
SYSTEMS

analyze them, weigh them, record the masses of their fragment ions, and then use the bioinformatics programs that are becoming available to identify the bacteria," she says.

Kelleher calls the approach a "killer app" for the top-down workflow. "It's an awesome use of the scanning power top-down provides across the whole protein." But here's the catch: ProSight PC, the bioinformatics software Fenselau and her colleague Nathan Edwards, a bioinformaticist at Georgetown University Medical Center use, relies on matching detected fragment ion masses against a virtual database of known proteins, splice variants, and modifications.

"If your software identifies proteins from a database, then you'll miss the correct identification if the right answer isn't in the database," says Edwards. That's especially true with unsequenced organisms or novel post-translational modifications.

To circumvent that problem, Edwards and Fenselau configured their searches to match well-conserved proteins, such as ribosomal proteins, based not on the mass of the intact molecule itself, as in bottom-up, but on the mass of its b- and y-ion fragments.

"There are cases where the ribosomal proteins in a related organism are different in only one or two residues, which means that many of its b- and y-ion fragments have the same mass as the b- and y-ion fragments from the true protein."

From that, the team was able to establish sufficient sequence data to compare their protein identifications against known organisms—information that they could use to place the organism in a phylogenetic tree, without first sequencing its genome at the DNA level (4).

But Pavel Pevzner, professor of computer science and director of the NIH Technology Center for Computational Mass Spectrometry at the University of California, San Diego, thinks ProSight PC (which was developed in Kelleher's lab and commercialized by Thermo Fisher Scientific) has a significant flaw.

For Pevzner, ProSight's "Achilles heel" is the use of virtual databases, an approach no standard genomic tool uses, as it is essentially impossible to scale (To wit: Paša-Tolić has calculated there could be 40 trillion theoretical variants of histone H3.1 alone, far too many to populate a virtual database). "As a computer scientist, I cannot agree with the algorithmic design of this tool," says Pevzner.

His alternative, MSAlign+ utilizes "dynamic programming" and spectral alignment to identify proteins (5).

"In every algorithms class students learn that when there is [an] exponential

explosion of variants to consider, you should not try to tabulate these variants but rather come up with an algorithm that makes the problem tractable." Several algorithmic techniques can be used, he explains, including dynamic programming, "the magic trick behind many genomics (and now proteomics) algorithms."

Edwards, who has used MSAlign+, says it is in some ways more powerful than ProSight PC, but also "a bit more research-grade." Still, says Pevzner, MSAlign+ is fast and relatively comprehensive. "It actually discovers practically everything that ProSight PC finds, and more." (Kelleher says that the recently released ProSight PC 3.0, addresses some of these criticisms.)

Such tools could be key to tackling the proteome. Bottom-up experts have identified large swaths of the proteome, but each protein can exist in multiple forms. Indeed, Paša-Tolić has detected some 3,000 unique histone isoforms in one cell type—representing just 30 genes.

And therein lies the strength of top-down, says Kelleher: It's not in the sheer number of identifications a study can produce, but the quality of the resulting data, the unambiguous identification and characterization of discrete protein isoforms.

Biologists, he says, are catching on.

"The more that you understand it, you start to see why top-down has and will continue to develop a following. And it's durable. It's not easy, but it's getting easier all the time."

References

1. Tran, J.C., et al. 2011. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*. 480:254–8.
2. Roth, M.J., et al. 2011. Sensitive and reproducible intact mass analysis of complex protein mixtures with superficially porous capillary reversed-phase liquid chromatography mass spectrometry. *Anal Chem*. 83:9586–92.
3. Madsen, J.A., et al. 2009. Top-down protein fragmentation by infrared multiphoton dissociation in a dual pressure linear ion trap. *Anal Chem*. 81:8677–86.
4. Wynne, C., et al. 2009. Top-down identification of protein biomarkers in bacteria with unsequenced genomes. *Anal Chem*. 81:9633–42.
5. Liu, X., et al. 2012. Protein identification using top-down. *Mol Cell Proteomics*. 11:10.1074/mcp.M111.008524, 1–13.

Written by Jeffrey M. Perkel, Ph.D.

BioTechniques 53:75-78 (August 2012)
doi: 10.2144/000113900

To purchase reprints of this article, contact:
biotechniques@fosterprinting.com