Title of proposed idea: Defining the human proteome, digesting, separating and analyzing MSMS derived from it

Nominator: Stephen Barnes

What is the major obstacle/challenge in the field? What is needed to overcome this obstacle/challenge?

Understanding how the genome is expressed at the protein level. We are still working on the premise that "genes" are translated in the linear manner that we were using in 2000 and before, and also that proteins are a result of "the genes" as we currently define them. I'm going to express what may seem naiveté to some – why do we restrain the assembly of mRNAs and hence proteins by joining together exons from the same "gene"? Why not between "genes" And what is an "untranslated" region? The question is, is it (untranslated)?

A limitation in the current standard practice of nanoLC-tandem MS is that analysis is data-dependent. This has to go since it means that lower abundance peptides are ignored. TOF technology, as used in the AB Sciex 5600 TripleTOF, permits a much higher rate of MSMS data acquisition. Technologies must continue to improve. One area that's been ignored is improving chromatographic resolution (above and beyond the existing 2D-technologies such as MuDPIT). The approach will depend on improvements in MSMS acquisition time since peak shapes will become narrower.

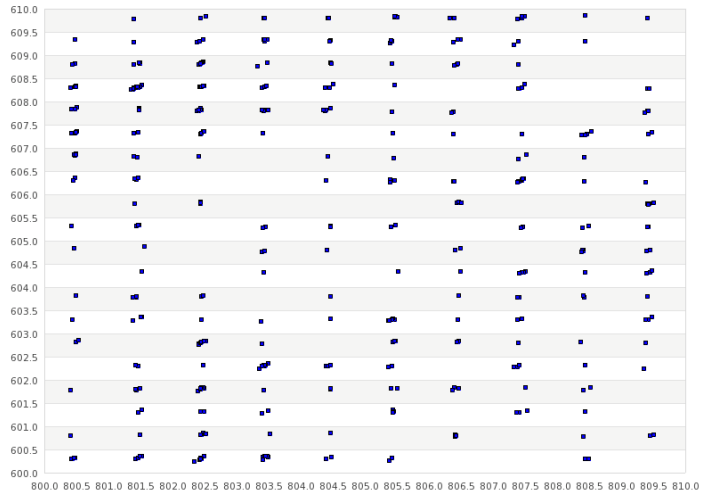
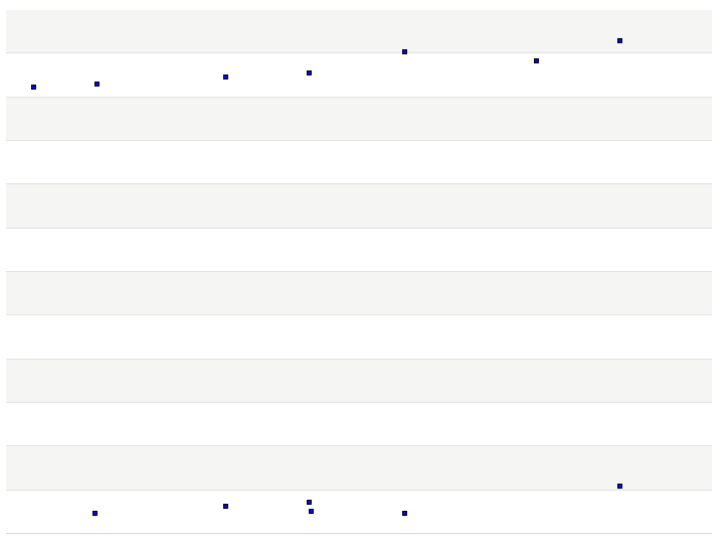
What emerging scientific opportunity is ripe for investment by a Trans-NIH program (e.g. the NIH Common Fund)?

The limiting feature of current proteomics is that it relies on databases for the identification of a particular peptide/protein. For the past 10 years it's been argued that we don't have time to sequence all the peptide MSMS spectra. This story has run its course. There so many peptides that everyone sees in a nanoLC-tandem MS experiment that cannot be identified. There has to be a larger investment in defining the real proteome as well as in *de novo* sequencing. Recent studies using deep DNA sequencing have revealed that while there is a close to faithful copying of DNA from parents to the children at the germ line level, this is not the case in somatic tissues where most of the disease processes. Therefore for most of medicine, there is no such thing as a canonical protein sequence (necessarily) in a particular subject.

So, the scientific opportunity comes from the plethora of information that's being generated by NextGen sequencing. New rules for transcription and translation will emerge. These are so important for the application of proteomics in the human being.

What are the potential Trans-NIH investments that could accelerate scientific progress in this field?

There is a need for investment in translating new genome level information to proteins, novel chromatographic improvements in resolving tryptic peptides and peptides from other proteases, new proteases, and improved de novo sequence interpretation from MSMS data.

If a Trans-NIH program on this topic achieved its objectives, what would be the impact?

What is described above affects every institute and center at NIH that in any way values the importance of proteomics, particularly in clinical applications. At this time the triple quadrupole multiple reaction ion monitoring method for measurement of proteotypic peptides is considered state-of-the-art. However, it has a serious fault, namely the filtering mechanism for the precursor and fragment ions used in this method. The quadrupole detector has a large mass window, typically m/z 0.7, for both parent and fragment ions. Shown below is the output of a tool we have that plots the density of mass space around the m/z values used for a proteotypic peptide from the expected, unmodified, tryptic peptides from the whole human proteome. The range from *m/z* 600-610 (vertical axis) is for the precursor ions and the fragment ions (horizontal axis) are from *m/z* 800-810. What can be seen is that the doubly charged peptide ions and the singly charged fragment ions are grouped approximately *m/z* 0.5 apart rather than continuously. Some of the 0.7 by 0.7 boxes are empty. Other boxes contain many peptides (from many different proteins) that would satisfy the mass window requirements. The second figure is an expanded region from *m/z* 608.3 to 609.0 for the precursor ion and *m/z* 802.15 to *m/z* 802.85 (a 0.7 Da window). Two groups of precursor ions can be seen – the lower one has 7 fragment ions satisfying the window filter criterion and the upper one 8 fragment ions (each coming from a different protein). Given that the complexity of the real human proteome is a result of PTMs including deamidation, C-truncations, known SNPs, differential mRNA splicing, and mutations at the germline and somatic level, and as yet to be defined other gene-protein information transfer, a technology based on even multiple precursor-fragment ion combinations is inadequate. That is why it's essential to use higher resolution mass spectrometry to validate the identity of a peptide. In the above example, a 50 mDa window for the fragment ions would have separated 14/15 out of the two sets of peptides. Also, by collecting the whole MSMS spectrum as occurs using a TOF detector, there are multiple ions to confirm the sequence of the peptide. High chromatographic resolution would also help to ensure that all the characteristic fragment ions for a peptide co-elute at the same time.