Proteomics and Mass Spectrometry 2009

Stephen Barnes, PhD
4-7117, MCLM 452
sbarnes@uab.edu

Helen Kim, PhD
4-3880, MCLM 460A
helenkim@uab.edu

Matt Renfrow, PhD
6-4681, MCLM 570
Renfrow@uab.edu

Jeevan Prasain, PhD
6-2612, MCLM 456
jprasain@uab.edu

Peter E. Prevelige, PhD
5-5327, BBRB 416
prevelig@uab.edu

Course plan

- **Meet Tuesdays/Fridays in MCLM 401 from 9-11 am (Jan 6-Mar 20)**
- **Graduate Students taking this course are required to attend each session (unless advance communication with instructor)**
- **Evaluations will be made from projects/exams**
- **Where possible, class notes will be available on the UAB proteomics website (go to http://www.uab.edu/proteomics - click on Class)**

Recommended general texts

- **Suggested text - "*Introduction to Proteomics*" by Daniel C. Liebler, 2002**
- **Also see "*The Expanding Role of Mass Spectrometry in Biotechnology*" by Gary Siuzdak (a 2003 edition of the 1996 first edition)**
- **"*Mass spectrometry data analysis in proteomics*", (ed., Mathiesson, R) in Methods in Molecular Biology, vol 367.**

Suggested readings

- Kenyon G, et al. *Defining the mandate of proteomics in the post-genomics era: workshop report*. Mol Cell Proteomics, 1: 763-780 (2002)
- Kim H et al. *Proteomics and mass spectrometry in nutrition research*. Nutrition, 20: 155-165 (2004)
- Righetti P. et al. *Prefractionation techniques in proteome analysis: the mining tools of the third millennium*. Electrophoresis, 26: 297-319 (2005)
- Anderson NL. *The roles of multiple proteomic platforms in a pipeline for new diagnostics*. Mol Cell Proteomics (2005)
- Venkatesan et al. *An empirical framework for binary interactome mapping*. Nat Methods. 2008 Dec 7. [Epub ahead of print] PMID: 19060904
- Yan W et al. *Evolution of organelle-associated protein profiling*. J Proteomics. 2008 Dec 7. PMID: 19110081
- Pan S, et al. *Mass Spectrometry Based Targeted Protein Quantification: Methods and Applications*. J Proteome Res. 2008 Dec 23. [Epub ahead of print] PMID: 19105742

S Barnes/H Kim BMG 744 1/06/09

BMG/PHR 744 - section 1

- **Jan 6, Tu** **Barnes/Kim**
  The world of biomolecules. The proteome, proteomics and other –omics and where to start
- **Jan 9, Fri** **M. Renfrow**
  Mass spectrometry – gas phase transfer and instrumentation – including ETD
- **Jan 13, Tu** **S. Barnes**
  Methods for the identification of proteins: MALDI-TOF of proteins and peptide mass fingerprinting; LC analysis and peptide sequencing
- **Jan 16, Fri** **S. Barnes**
  Ion fragmentation in mass spectrometry; application to proteomics
- **Jan 20, Tu** **J. Prasain**
  Ion Fragmentation of small molecules
- **Jan 23, Fri** **S. Barnes/E Shonsey**
  Sample preparation for proteomics and mass spectrometry; MS in Forensics

BMG/PHR 744 - section 2

- **Jan 27, Tu** **A. Smith**
  Connecting proteomics into bioinformatics; MUDPIT and SEQUEST; false discovery rates in complex systems
- **Jan 30, Fri** **C. Crasto**
  The bioinformatics of the proteome
- **Feb 3, Tu** **S. Barnes/H. Kim**
  Enhancing proteomic analysis by reducing sample complexity; approaches to protein separations
- **Feb 6, Fri** **J. Prasain**
  The use of mass spectrometry in metabolomics and lipidomics
- **Feb 10, Tu** **S. Barnes**
  Mass spectrometry in qualitative and quantitative burrowing of the proteome
- **Feb 13, Fri** **J. Prasain**
  Qualitative and quantitative analysis/method validation

BMG/PHR 744 - section 3

- **Feb 17, Tu** **S. Barnes**
  Enzymology and mass spectrometry
- **Feb 20, Fri** **M. Renfrow**
  Analysis of protein-protein interactions by affinity purification and mass spectrometry
- **Feb 24, Tu** **P. Prevelige**
  Mass Spectrometry as a Tool for Studying Protein Structure
- **Feb 27, Fri** **P. Prevelige**
  Study of macromolecular structures – protein complexes
- **Mar 3, Tu** **C-C. Wang**
  Tissue and body fluid proteomics and mass spectrometry
- **Mar 6, Fri** **K. Schey**
  Applications of MS to tissue imaging

BMG/PHR 744 - section 4

- **Mar 10, Tu** **M. Renfrow**
  Applications of FT-ICR-MS
- **Mar 13, Fri** **J Novak/M Renfrow**
  Mass spectrometry in glycomics research - Application to IgA nephropathy
- **Mar 17, Tu** **S. Barnes**
  Isotopes in mass spectrometry
- **Mar 20, Tu** **H. Kim**
  Use of proteomics and MS methods in the study of the brain proteome and neurodegenerative diseases
- **Mar 20** ***Final report due***

Course learning objectives

- Introduction to the concepts and practice of systems biology
- Sample ionization and mass spectrometers
- Mass spectrometry and its principal methods
  - protein and peptide ID; ion fragmentation; stable isotope labeling; quantification

Course learning objectives

- Informatics, statistics and quality control in mass spectrometry
- Importance of prefractionation in proteomics - 2DE, LC and arrays
- Applying mass spectrometry to protein modifications, function, structure and biological location, and to other biological molecules

Hopes and hazards of biomedical research

It boils down to whether having taken life apart into its distinct pieces, can we reassemble it in new ways? [*strong analogies to what have been the central quests of physics*]

Can we create a form of life that might live in a very hostile extra-terrestial environment and thereby save humanity?

Or will we (as well as our enemies) instead create life forms that can terrorize or even eliminate us? *Will Einstein's and Oppenheimer's moral dilemmas surface in biomedical science?*

History of proteomics

- **Essentially preceded genomics**
- **"Human protein index" conceived in the 1970's by Norman and Leigh Anderson**
- **The term "proteomics" coined by Marc Wilkins in 1994**
- **Human proteomics initiative (HPI) began in 2000 in Switzerland - http://www.hupo.org**
- **Human Proteome Organization (HUPO) had meetings in 2002 in Versailles, France; 2003 in Montreal, Canada; 2004 in Beijing, China; 2005 in Munich, Germany; 2006 in Long Beach, CA; 2007 in Seoul, Korea; 2008 in Amsterdam. The 2009 meeting will be in Toronto and 2010 in Sydney**

What proteomics is, what it isn't

"Proteomics is not just a mass spectrum of a spot on a gel"

George Kenyon,
2002 National Academy of Sciences Symposium

Proteomics is the identities, quantities, structures, and biochemical and cellular functions of all proteins in an organism, organ or organelle, and how these vary in space, time and physiological state.

Collapse of the single target paradigm - the need for systems biology

Old paradigm

Diseases are due to single genes - by knocking out the gene, or designing specific inhibitors to its protein, disease can be cured

But the gene KO mouse didn't notice the loss of the gene

New paradigm

We have to understand gene and protein networks - proteins don't act alone - effective systems have built in redundancy

Research styles

- **Classical NIH R01**
  - **A specific target and meaningful substrates**
  - **Emphasis on mechanism**
  - **Hypothesis-driven**
  - **Linearizes locally multi-dimensional space**
- **Example**
  - **Using an X-ray crystal structure of a protein to determine if a specific compound can fit into a binding pocket - from this "*a disease can be cured*" - this approach ignores whether the compound can get to the necessary biological site, whether it remains chemically intact, and where else it goes**

S Barnes/H Kim BMG 744 1/06/09

From substrates to targets to systems - a changing paradigm

- **Classical approach - one substrate/one target**
- **Mid 1980s - use of a pure reagent to isolate DNAs from *cDNA libraries* (multiple targets)**
- **Early 1990s - use of a *reagent library* (multiple ligands) to perfect interaction with a specific target**
- **2000+ - effects of specific reagents on cell systems using *DNA microarrays* (500+ genes change, not just one)**
- **2008 - integration of transcriptomics, proteomics, peptidomics, metabolomics (everything changes, just like in ecology)**

S Barnes/H Kim BMG 744 1/06/09

Exploring information space - the *Systems Biology* approach

- Systems biology means measuring everything about a system at the same time
- For a long time, it was deemed as too complex for useful or purposeful investigation
- But are the tools available today?

Systems Biology

"To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism."

"*Properties of systems, such as robustness, emerge as central issues, and understanding these properties may have an impact on the future of medicine."*

"However, many breakthroughs in experimental devices, advanced software, and analytical methods are required before the achievements of systems biology can live up to their much-touted potential."

Kitano, 2002

The Biological Data of the Future

Current nature of data	Elias Zerhouni, FASEB 2004
• Destructive • Qualitative • Uni-dimensional • Low temporal resolution • Low data density • Variable standards • Non cumulative	• Non-destructive • Quantitative • Multi-dimensional and spatially resolved • High Temporal resolution • High data density • Stricter standards • Cumulative

S Barnes/H Kim BMG 744 1/06/09

Techniques in Systems Biology

- DNA microarrays to describe and *quantify* the transcriptosome
- Large scale and small scale proteomics
- Protein arrays
- Protein structure
- Metabolomics
- Integrated computational models

S Barnes/H Kim BMG 744 1/06/09

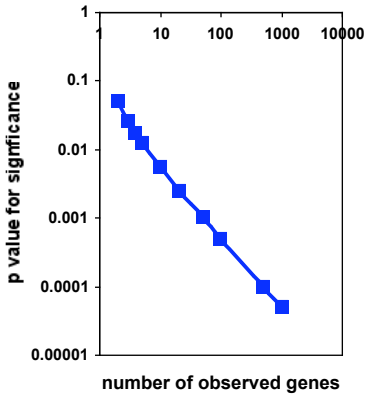
Papers on systems biology

http://www.nature.com/focus/systemsbiology userguide/index.html

This is a cluster of papers in Nature Cell Biology, Nature Reviews in Molecular and Cell Biology, and Molecular Systems Biology that was published in October, 2006

High dimensionality of microarray or proteomics data means you must understand statistics

While reproducible data can be obtained, the large numbers of parameters (individual genes or proteins) require large changes in expression before a change can be regarded as significant



Use of the Bonferroni correction:
A conservative correction

Properties of a system and fold-change

- The primary assumption of most users of DNA microarrays (and proteomics) has been that the cut-off for assessing change is two-fold
- This is a very naïve view of properties of a system
  - Barnes' law "Fold-change is inversely related to biological importance"

Properties of a system and fold-change

- For a system, items that are important are the least likely to change
  - when they do, catastrophic events may occur
  - Proliferation vs apoptosis (PTEN < 50% change)
- Items unimportant to the system can vary a lot (not a core value)
- How can we perceive "importance"?
  - Re-weight the data by dividing by the variance
  - Need to have enough information about each item to calculate its variance (n > 5)

Vulnerability of a system

- To really understand biological systems, you have to appreciate their dynamic state
  - Read about control theory
  - Realize that systems are subject to rhythms
  - Subject them to fourier transform analysis to detect their resonance (requires far more data than we can currently collect)
- A small signal at the right frequency can disrupt the system
  - Analogies "the small boy in the bath" and "the screech of chalk on a chalk board"

Hazards of interpreting microarray (proteomic) data

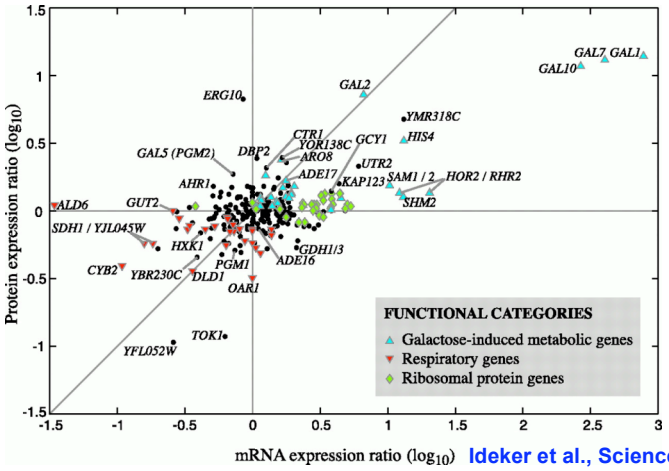
- "Expression patterns are the place where environmental variables and genetic variation come together. Environmental variables will affect gene expression levels."
- "Don't we need to be very careful to understand the environmental inputs that might have an impact on that expression? Perhaps an over-the-counter herbal supplement might cause an expression pattern that looks like that of a very aggressive tumor."

Abridged from Karen Kline, 2002

Why study the proteome when we can do DNA microarrays?

- DNA microarray analysis allows one to examine the mRNA levels of thousands and thousands of genes
- However, the correlation between gene expression and protein levels is poor at best
- Is this a new finding? No, before the age of genetics, it was well known

Apparent poor relationship between gene expression and protein content



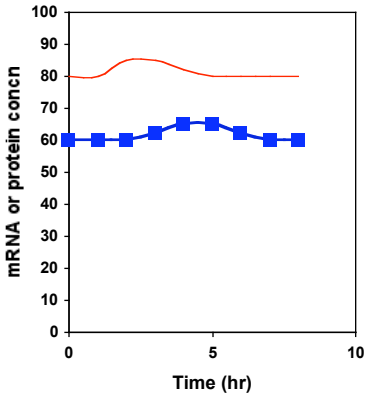
Ideker et al., Science 292: 929 (2001)

S Barnes/H Kim BMG 744 1/06/09

Housekeeping genes and proteins are related

This is the relationship between mRNA (red) and protein (blue) levels expression of a house-keeping gene/protein, i.e., one that has to be expressed at all times

- Even with the small perturbation, the amounts of mRNA and protein are well correlated to each other



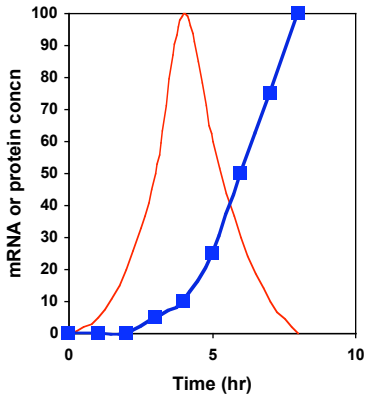
S Barnes/H Kim BMG 744 1/06/09

(Barnes & Allison, 2004)

Sampling time affects interpretation of correlation between mRNA and protein expression for important proteins

Determining the relationship between mRNA (red) and protein (blue) levels depends totally on when you measure them - for the figure opposite, the ratio at 2.5 hr is 10:1, whereas at 7.5 hr it's 1:100

- better to measure the ratio over time and integrate the area under the curve



S Barnes/H Kim BMG 744 1/06/09

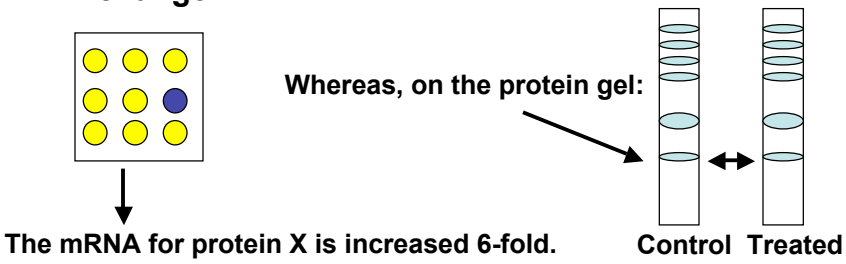
(Barnes & Allison, 2004)

Defining disease from the proteome

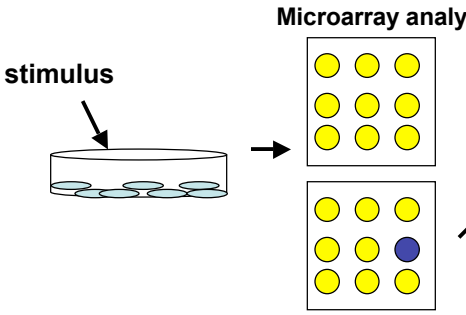
- Numerous examples of a revised picture of disease from analysis of the proteome
  - Aging
  - Cancer
  - Cardiovascular disease
  - Neurodegeneration
- Infectious disease and the microbial proteome

Upregulation of a gene for a crucial protein may have poor correlation at the protein level: example A.

- May be manifested as little or no change in amount, because the cell "damps" out the change.



Gene expression data may predict the <u>nature</u> of protein modifications, *but not which proteins are modified*.



A kinase is upregulated;

But----which substrates are phosphorylated???

S Barnes/H Kim BMG 744 1/06/09

Summary: Why do we need to do broad scale "proteomics"?

- **When you don't have a clue;**
- **When you have only a very small clue;**
  i.e. you've done a microarray experiment, and you have a list of 3,284 genes that are differentially regulated in your system;
- **When you knock out a gene (and hence a protein) that you're convinced is essential for life and health, and the animal pees as usual.**

S Barnes/H Kim BMG 744 1/06/09

Rationales for proteomics approaches in today's research

- Identify a "marker" protein(s);
  - Cancer detection/Monitor response to chemotherapy
  - Identify one pathogen from others;
  - Distinguish a virulent strain of pathogen from nonvirulent.
- Characterize protein differences between disease and normal tissues--
  - For understanding the disease process;
  - To develop drug targets;
- In cancer, there may be novel proteins due to chromosome instability (ETV6-ABL and BCR-ABL), or inappropriate expression may occur (proteins from embryonic or fetal stages of development)

Where there is pathology, but the genetic basis unknown, proteomics can have critical role in identifying proteins to target for therapeutic intervention

Two major disease examples:

- HIV: protease is targeted today;
  - are there other proteins, either viral or host, that could be targeted to better deal with the disease?
- Alzheimer's disease: 3 known mutations (APP, PS1, PS2) and risk factors (ApoE, estrogen loss);
  - 50% of AD patients do not have any of the known genetic abnormalities, yet all become demented, all have amyloid plaques and NFT in their brains.
  - Remember, every AD patient has AD 100%.

Challenges in proteomics

- So many protein forms! How many 100-mers are possible using 20 different amino acids?
- Can protein forms be predicted from genomic sequence data?
  - Posttranslational modifications
  - Differential splicing
  - Unknown transcriptional mechanisms
- The dynamic range (>10^9)
- No equivalent to the PCR reaction
  - Avogadro's number

Avogadro's number comes back to haunt us: a reality check for proteomics

- One gram-mole of anything is 6.02 x 10^{23} molecules
- For mass spec detection, we need at least 100 femtomole (10^{-13} moles) or 6.02 x 10^{10} molecules

protein expression per cell	need # cells for 100 fmoles
10	6.02 x 10^9
100	6.02 x 10^8
1,000	6.02 x 10^7
10,000	6.02 x 10^6
100,000	6.02 x 10^5

Central issue in proteome complexity: dynamic range

Cellular proteomes involve a very wide *dynamic range*: proteins can differ in their amounts by nine (9) orders of magnitude;

- **In blood, albumin is 0.5×10^{-3} M, whereas cytokines are pM (10^{-12} M)**
- **A 2D gel that is overloaded with respect to an abundant protein, may have *barely detectable* amounts of a low abundance protein…….And if you can't see it, you don't know it's changed.**

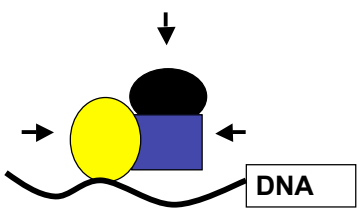
S Barnes/H Kim BMG 744 1/06/09

The need to enrich for subproteomes, and/or isolate the lower abundant proteins

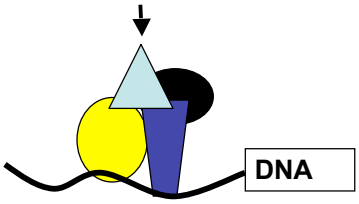
- **See lecture by Steve Barnes and Helen Kim on February 3rd in section 2 of the course**
- **Biological properties:**
  - **Intracellular location; Protein-protein interactions; Posttranslational modifications**
- **Intrinsic properties:**
  - **Net charge; Size; Extent of tertiary structure; Hydrophobicity**

S Barnes/H Kim BMG 744 1/06/09

A valuable role of the proteomics approach:
Discovery & analysis of protein-protein interactions



Inactive transcription complex



Active transcription complex

(The arrows indicate the proteins that could be antigens for immunoaffinity purification of the complex.)

Experimental design and quality control issues

- **How do we carry out a proteomics experiment?**
  - **Randomize sample analysis**
  - **Process samples blinded to identities**
  - **Standardize procedures and vendors of disposable plastics used in experiment-- minimize variation where possible**
  - **Consult with statistician before experiment; ensure enough "power" for the experiment so that statistical analysis yields significant data.**

Take home lessons in analyzing proteins with proteomics methods

- The fewer proteins in the proteome you analyze, the better the chances of detecting the ones that "matter."
- Genomics data can complement proteomics data.
- Understanding the biological properties of the proteins of interest can enhance proteomics analysis.
- Intrinsic properties of proteins form the basis of invaluable prefractionation prior to proteomics analysis.
- Quality control is an issue that becomes increasingly important with large datasets and measurement of small changes

Predicting the proteome

- ***Bioinformatics*** **is the basis of high throughput proteome analysis using mass spectrometry. Protein sequences can be computationally predicted from the genome sequence**

- **However, *bioinformatics* is not able to predict with accuracy the sites or chemistry of posttranslational modifications - these need to be defined chemically (using mass spectrometry)**

Predicting the proteome

- *Predicting the proteome has elements of a circular argument*
  - protein sequences were initially determined chemically and were correlated with the early gene sequences. It then became easier to sequence a protein from its mRNA (captured from a cDNA library). This could be checked (to a degree) by comparison to peptide sequences. Now we have the human genome (actually two of them).
- *So, is it valid to predict the genes (and hence the proteome) from the sequence of the genome?*
  - We're doing this in current research. But as we'll see, the mass spectrometer is the ultimate test of this hypothesis -
  - why? <u>because of its mass accuracy</u>

Protein structure

- Determined by folding - folding rules not yet defined - cannot predict structure *de novo*
- X-ray crystallography has been used to produce elegant structural information
- NMR and H-D exchange combined with mass spec enable the in-solution structure to be determined (see Peter Prevelige's lectures on February 20/24)

Protein informatics

- The predicted sequences of the proteins encoded by genes in sequenced genomes are available in many publicly available databases (subject to the limitations mentioned earlier)
- The mass of the protein is less useful (for now) than the masses of its fragment ions - as we'll see later, the masses of tryptic peptides can be used to identify a protein in a matter of seconds

So, what do we do with all these data?

- Management of the data generated by DNA microarray and proteomics/protein arrays
  - High dimensional analysis
- Beyond the capabilities of individual investigators
- Urgent need for visualization tools
- The importance of new statistical methods for analysis of high dimensional systems

PROTIG and Videocast

- There is an NIH-based proteomics interest group (PROTIG)
  - Sign up at http://proteome.nih.gov
- Proteomics and mass spec talks are available for viewing (using Real Player)
  - Log on at http://videocast.nih.gov
  - Podcasts are also available

S Barnes/H Kim BMG 744 1/06/09