Proteomics and Protein Mass Spectrometry 2006

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

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**
- **Evaluations will be made from in-class presentations of assigned papers plus 2 projects/exams**
- **Where possible, materials from each class will be placed on the proteomics website (go to http://www.uab.edu/proteomics - click on Class)**

Recommended 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)**
- **Both available at Amazon.com**

Suggested course reading material

- **Kenyon G, et al. *Defining the mandate of proteomics in the post-genomics era: workshop report*. Mol Cell Proteomics, 1: 763-780 (2002)**
- **Hood L. *Systems biology: integrating technology, biology and computation.* Mechan Aging Develop, 124: 9-16 (2003)**
- **Kim H et al. *Proteomics and mass spectrometry in nutrition research*. Nutrition, 20: 155-165 (2004)**
- **Yates JR 3rd et al. *Proteomics of organelles and large cellular structures.* Nature Rev Mol Cell Biol, 6: 702-714 (2005)**
- **Righetti P. et al. *Prefractionation techniques in proteome analysis: the mining tools of the third millennium*. Electrophoresis, 26: 297-319 (2005)**
- **Graham DR et al. *Broad-based proteomic strategies: a practical guide to proteomics and functional screening.* J Physiol, 563:1-9, (2005)**

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

BMG 744 Course content

Jan 6, Fri	Barnes/Kim	The world of proteins – beyond genomics. The proteome, proteomics and where to start
Jan 10, Tu	M Niemann	Techniques of protein separation
Jan 13, Fri	H Kim	Protein separation by electrophoresis and other 2D-methods
Jan 17, Tu		***Student presentations***
Jan 20, Fri	S Barnes M. Renfrow	Mass spectrometry of proteins and peptides: principles and principal methods
Jan 24, Tu	S Barnes	MALDI of proteins and peptide mass fingerprinting; LC analysis and peptide sequencing
Jan 27, Fri	M Sabripour S Barnes	Connecting proteomics into bioinformatics; MUDPIT and SEQUEST
Jan 31, Tu	Staff	***Class demo of methods I***
		Mid-term exam – take-home
Feb 3, Fri	K Resing	The challenge of proteomics and mass spectrometry – required seminar attendance – Dr. Resing is an expert on proteomics/statistics and integration with bioinformatics
Feb 7, Tu	S Meleth	Statistical issues in proteomics and mass spectrometry
Feb 10, Fri	S Barnes	Qualitative and quantitative burrowing of the proteome
Feb 14, Tu		***Student presentations***
Feb 17, Fri	P Prevelige	Liquid phase protein structure and protein-protein surface mapping
Feb 21, Tu	S Barnes	Enzymology, proteomics and mass spectrometry
Feb 24, Fri	CY Chen	Analysis of protein-protein interactions by affinity purification
Feb 28, Tu	CC Wang J Mobley (Vanderbilt)	Tissue and fluid proteomics – imaging
Mar 3, Fri	M Renfrow	***Class demo of methods II***
Mar 7, Tu		***Student presentations***
Mar 10, Fri	J Novak M Renfrow	Application of proteomics to IgA nephropathy
Mar 14, Tu	J Felton (LLNL)	Protein adduct analysis by accelerator mass spectrometry
Mar 17, Fri	H Kim	Application of proteomics to the brain proteome
Mar 20	***Final exam due***	

What you should get from this course

- What is proteomics?
- Why do proteomics when we can already do genomics?
- Concepts of systems biology
- The elusive proteome
- Cells and organelles
- Separating proteins - 2DE, LC and arrays
- Mass spectrometry - principal tool of proteomics
- The informatics and statistics of proteomics
- Applications to biological systems

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**
- **Human Proteome Organization (HUPO) had meetings in 2002 in Versailles, France, 2003 in Montreal, Canada, 2004 in Beijing, China, 2004 and 2005 in Munich, Germany.**

What proteomics is not

"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*"

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)

Exploring information space - the *Systems Biology* approach

- Systems biology means measuring everything about a system at the same time
- For a long time 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

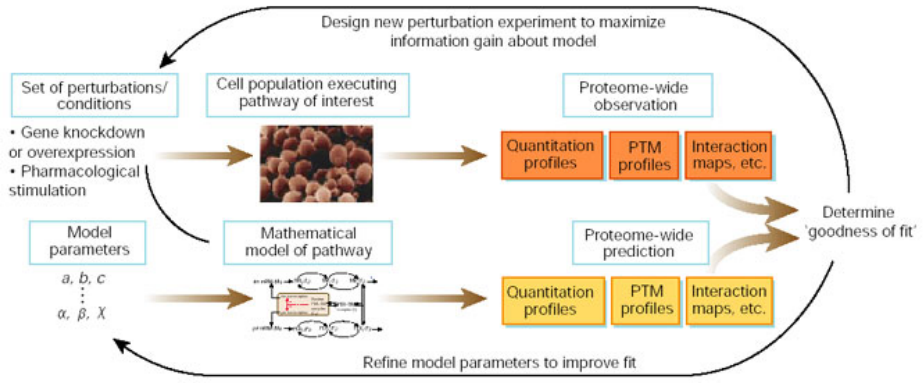
Defining disease from the proteome

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

Techniques in Systems Biology

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

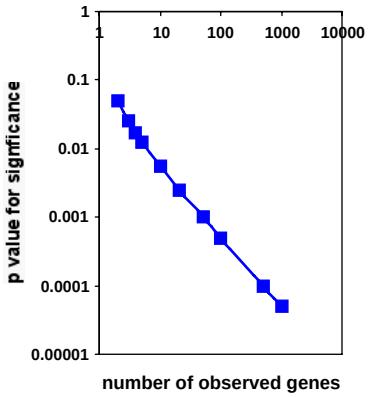
Schematic of systems biology paradigm



Important aspect of systems biology is that the model must undergo continual refinement

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, then catastrophic events will 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

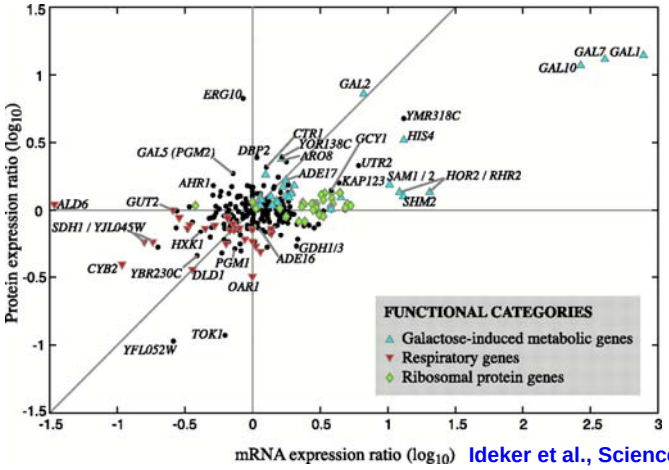
- "Expression patterns are the place where environmental variables and genetic variation come together. Environmental variables will affect gene expression levels."
- "Don't we need to be very careful to understand the environmental inputs that might have an impact on that expression? Perhaps an over-the-counter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

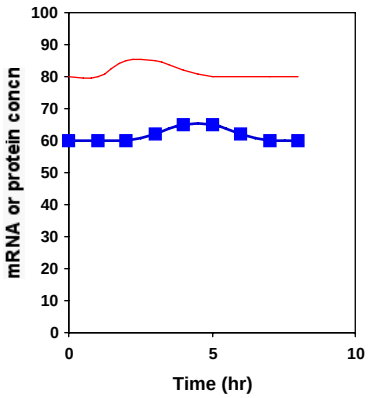


Ideker et al., Science 292: 929 (2001)

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**

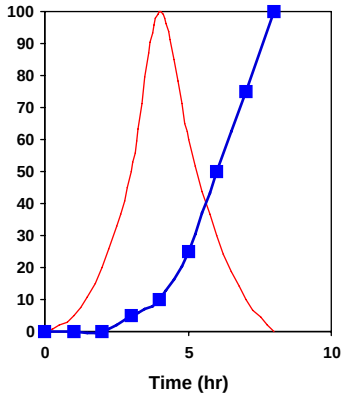


(XXXX)

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

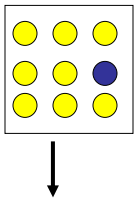
Determining the relationship between mRNA (red) and protein (blue) levels depends totally on when you measure them - for the figure opposite, the ratio at 2.5 hr is 10:1, whereas at 7.5 hr it's 1:100

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



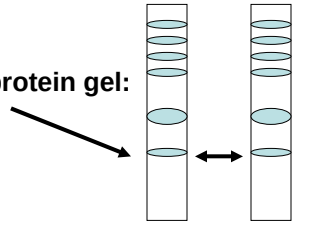
Upregulation of a gene for a crucial protein may have poor correlation at the protein level: example A.

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



The mRNA for protein X is increased 6-fold.

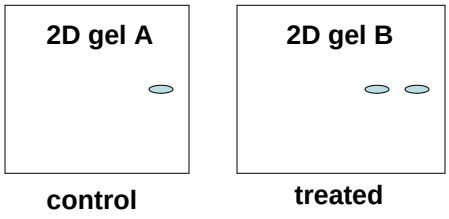
Whereas, on the protein gel:



Control Treated

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

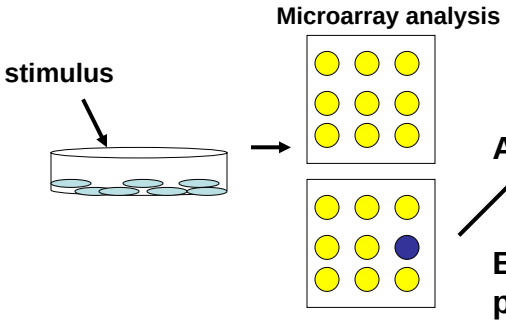
- **May be manifested as a difference in posttranslational modification, as the cell tries to "inactivate" excessive amounts of the protein, by altering it chemically.**



So, the modified protein is not part of the response to the stimulus, but rather, the system telling us that increased amounts of this protein were toxic.

-------How do we know the difference?

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???

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 1/06/06

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)

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

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%.

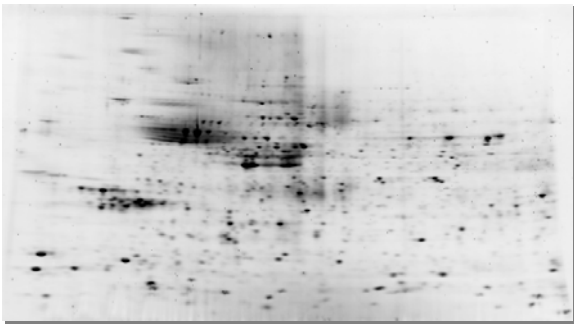
Challenges in proteomics

- So many protein forms!
- 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

The whole proteome contains
proteins, proteins, proteins,….

A typical cell proteome can include more than 60,000+ polypeptide forms;

- A really good 2D gel resolves 3,000 proteins;
- This is only 5% of the total.



–Obviously, any 2D gel of an unfractionated sample is only looking at the "low hanging fruit."

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

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

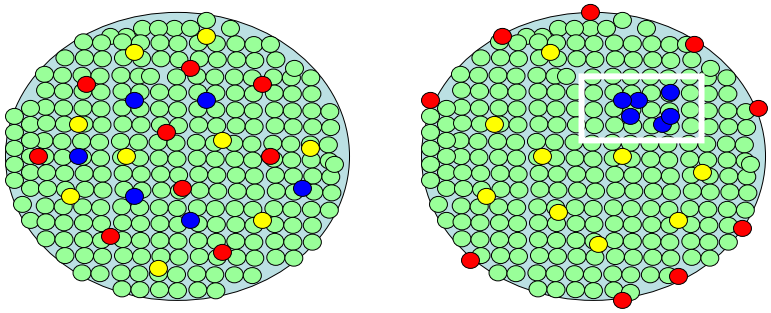
# proteins expressed per cell	need # cells for 100 fmoles
10	6.02×10^{9}
100	6.02×10^{8}
1,000	6.02×10^{7}
10,000	6.02×10^{6}
100,000	6.02×10^{5}

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 3.5 g/100 ml (35 g/L = 0.5 mM) (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 can't analyze it!!!!!**

Proteins are non-randomly distributed within cells.



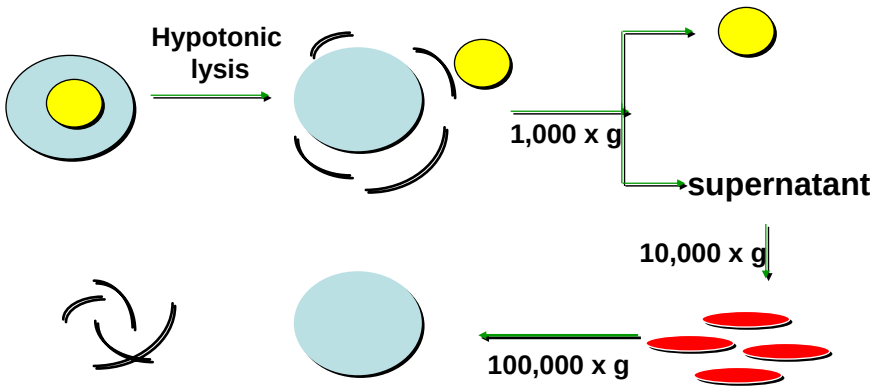
Not this **But this**

This nonrandom distribution of proteins is the basis for subcellular fractionation.

The basis for protein separation:

- **Biological properties:**
  - **Intracellular location**
  - **Protein-protein interactions**
  - **Posttranslational modifications**
- **Intrinsic properties:**
  - **Net charge**
  - **Size**
  - **Extent of tertiary structure**
  - **Hydrophobicity**

Size and density differences allow separation of subcellular compartments.

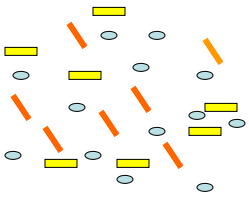


Physical and chemical methods for reduction of the proteome

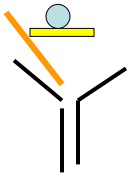
1. Net Charge (+ve or -ve), e.g., DEAE-cellulose
2. Mass, e.g., SDS-PAGE
3. Hydrodynamics, e.g., sucrose density gradient
4. Hydrophobicity, e.g., reverse-phase LC
5. Solubility, e.g., $(NH_4)_2SO_4$
6. Antigenic sites, i.e., immunoprecipitation
7. Metal binding sites

Dr. Niemann will discuss 1-5 in the next lecture.

Biological specificity of antibodies is invaluable in reducing the complexity of the proteome



A cell lysate: 6,739 polypeptides



An immune complex of 1 - 3 polypeptides
(why might there be more than one polypeptide?)

Which sample would you rather resolve on a 2D gel?

What if an antibody to your protein doesn't exist?

Make a protein that has a "tag" for which there *is* an antibody (make molecular biology work for you).

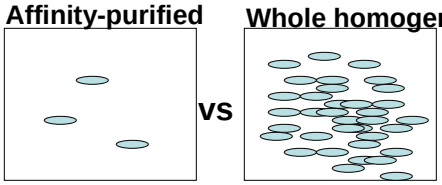
Express one of the following:

- **HA-tagged protein**
- **FLAG-tagged protein**
- **Histidine-tagged protein**

→ **homogenate**

\+ **Antibody to the "tag"**

Affinity-purified vs **Whole homogenate**



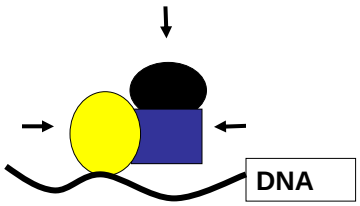
2D gel **2D gel**

← **Enrichment of your protein maybe 10,000-fold**

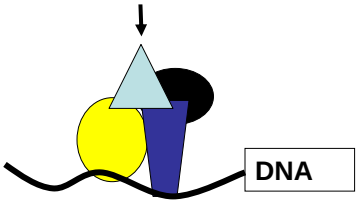
S Barnes/H Kim BMG 744 1/06/06

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.)

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

Types of protein-protein interactions where proteomics can enhance analysis/discovery

- Polymerization to form cytoskeletal structure
- Enzyme:_____
- Transcription factor:_____
- Receptor:_____

Methods to enrich for protein-protein interactions

- Immunoprecipitation
- Recombinant protein with affinity tag, then affinity purifications;
- Adjust buffer conditions to enhance for polymerization of oligomeric structures;
- Solubilize membrane-complexes, but keep “native” interactions intact;

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 in solution structure to be determined

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 proteome special 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