Proteomics and Protein Mass Spectrometry 2003

Stephen Barnes, PhD

4-7117, KHGB 602

sbarnes@uab.edu

Class text - *Introduction to Proteomics,* Daniel C. Liebler*, 2002*

Also see "*Mass Spectrometry for Biotechnology*" by Gary Siuzdak (note a new edition will be out soon)

Course plan

- **Meet Tuesdays/Fridays in KHGB 437 from 9-11 am**
- **Graduate Students taking this course are required to attend each session**
- **Evaluations will be made from in-class presentations of assigned papers plus 1-2 projects**
- **Where possible, materials from each class will be placed on the proteomics website (go to http://www.uab.edu/proteomics - click on resources)**

Course content

Date	Speaker	Topic
1/7	Barnes	What is proteomics?
1/10	Kim	Separating proteins
1/14	Barnes	Introduction to mass spectrometry
1/17	Barnes	Basic mass spec of peptides and proteins
1/21	Kirk/Wilson	Peptide fingerprinting and MSMS
1/24	Wang	Microscale and nanoscale analysis
1/28	Staff	Practical demonstrations of MS and 2DE
1/31		Student presentations
2/4	Barnes/ Brookes	Protein complexes
2/7	Safavy	Chemical modification of proteins
2/11	Orlando	Glycoproteome
2/14	Prevelige	H-D exchange and protein structure
2/18	Barnes	Enzymology and mass spec
2/21		Student presentations
2/25	Barnes	Posttranslational modifications
2/28	Staff	Locating modifications
3/4	Lefkowitz	Integration of bioinformatics in proteomics
3/7	Allison	High dimensional aspects of proteomics
3/11		Student presentations

Goals of the course

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

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 had its first meeting in November, 2002 in Versailles, France

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**
  - **Accent on mechanism**
  - **Hypothesis-driven**
  - **Linearizes locally multi-dimensional space**
- **Example**
  - **Using a 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*"**

Life is just a speck in reality

We have no sense of motion as we live, but

- the earth rotates once a day at 1,000 mph

- it also moves around the Sun at 17,000 mph,

- and around the Milky Way at 486,000 mph

From substrates to targets to systems - a changing paradigm

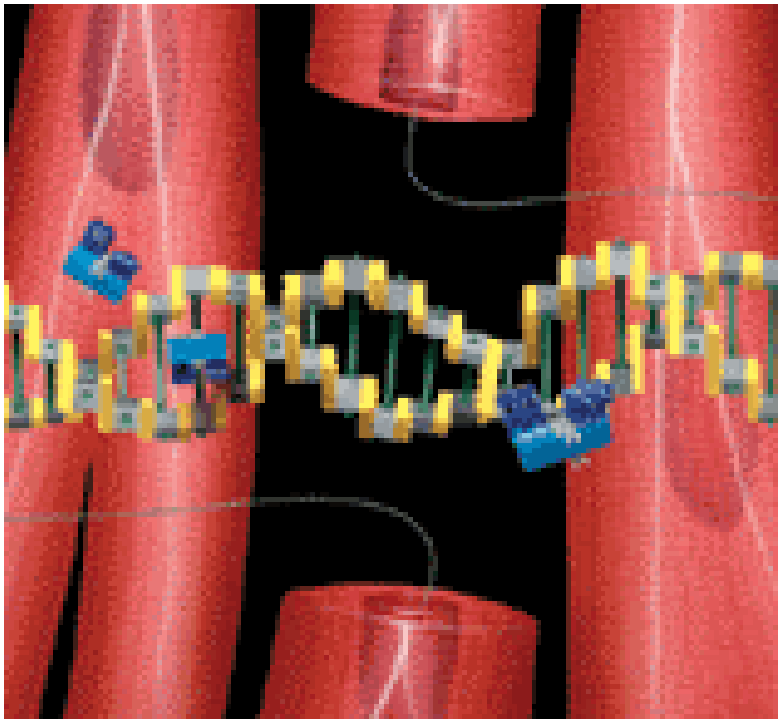
- **Classical approach - one substrate/one target**
- **Mid 1980s - use of a pure reagent to isolate DNAs from *cDNA libraries* (multiple targets)**
- **Early 1990s - use of a *reagent library* (multiple substrates) to perfect interaction with a specific target**
- **2000 - effects of specific reagents using *DNA microarrays* (500+ genes change, not just one)**

Exploring information space - the *Systems Biology* approach

- **For a long time deemed as too complex for useful or purposeful investigation**
- **But are the tools available today?**
- **Analogies to Einstein's General Theory of Relativity**

Systems Biology is “in”

Science March 1, 2002



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.'***

Kitano, 2002

Systems Biology

"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

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 *quantitate* the transcriptosome**
- **Large scale and small scale proteomics**
- **Protein arrays**
- **Protein structure**
- **Integrated computational models**

The elusive proteome

- The proteome is a subset of the genome (as is the transcriptosome)
- Almost all proteins are modified after translation
  - some are permanent (removal of N-terminal Met, glycosylation, farnesylation)
  - some are transient (phosphorylation, O-glycosylation)
  - others are unintended (nitration, oxidation)

The elusive proteome

- **Even for a given cell type in a tissue, protein expression varies from cell-to-cell (the cell is aware of its *environment*)**
- **Modification of a protein can radically alter its properties**
  - **kinases can increase their activities 100-fold following phosphorylation over a timescale of 10-90 sec and then lose the phosphate group (and hence enzyme activity returns to the basal state)**

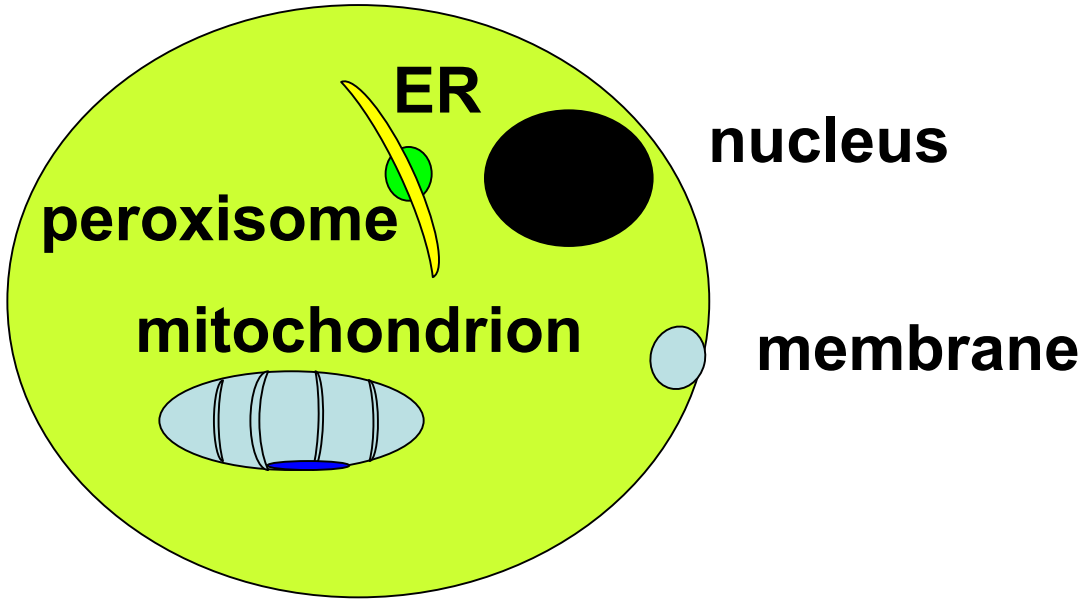
The elusive proteome

- **The cell or cellular site of an expressed protein is crucial for its effectiveness**
- **As in Real Estate, it's all about location, location**

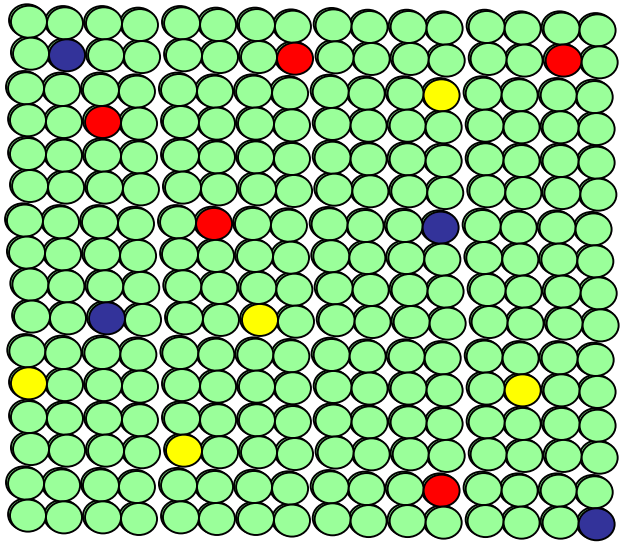
Proteins are different from mRNAs

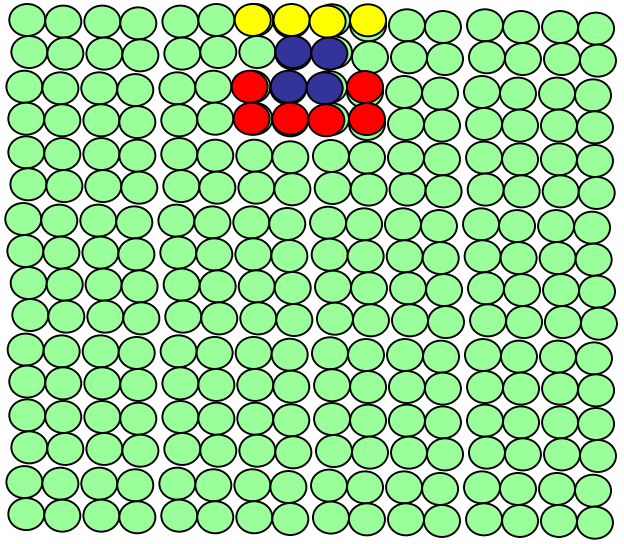
It's all about location, location

For mRNA it doesn't matter.
But where a protein is is everything



Proteins aren't random in cells





So, who's binding to whom?

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

Hazards of interpreting microarray 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

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>**

Separating the proteins of a proteome

- **A typical cell proteome might involve 60,000+ forms**
- **Very wide range in expression**
- **A cell has sets of proteins (~15%) that are essential for all cells to survive and smaller set that characterize the nature of the cell**

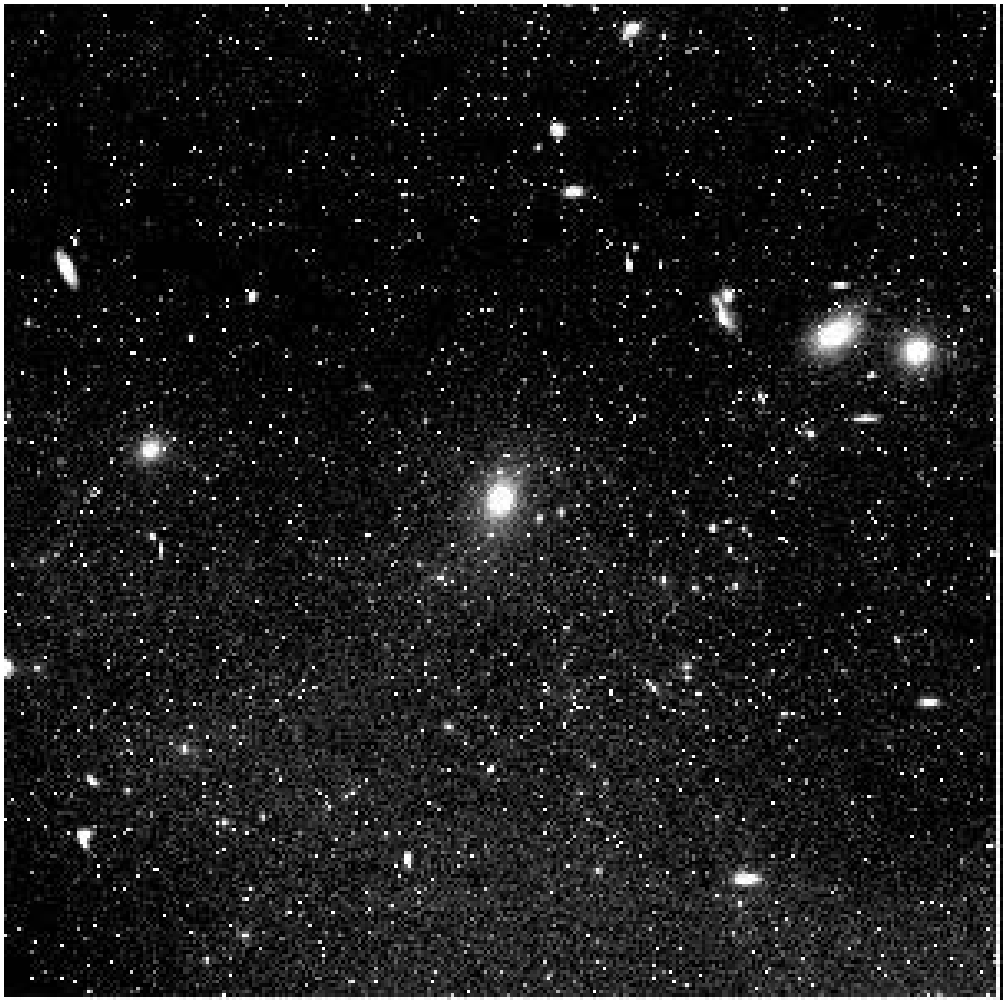
Current methods for proteomic separation

- 2D-isoelectric focusing/SDS-PAGE
  - perhaps 5-10% of an entire proteome
- LC-LC
- Affinity chromatography
  - for certain posttranslational modifications
- Quantitative analysis of peptides with isotopically coded affinity tag (ICAT)

Protein sequence and structure

- The number of possible combinations of the 20 amino acids is mind boggling
- For a 100-mer peptide, the number of distinctly different forms exceeds the number of protons in the universe
- In biology, specific blocks of sequences and their variants are used repeatedly

Protein space

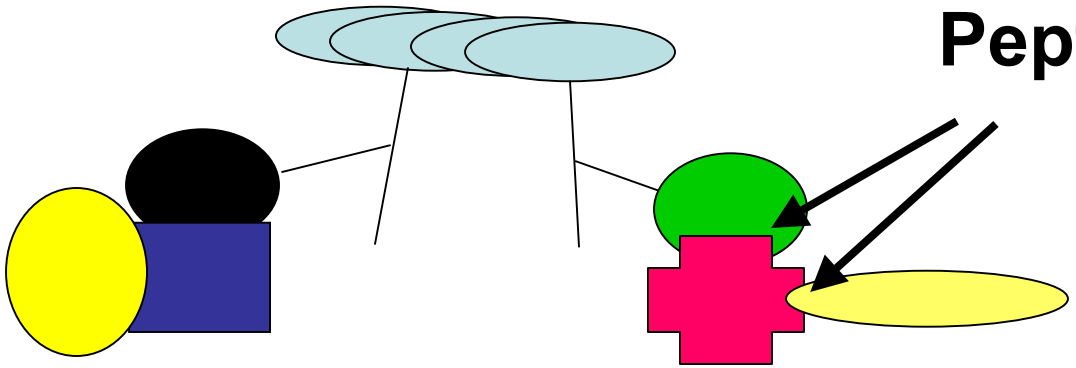


Only a small part of protein space is occupied, rather like the universe

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**

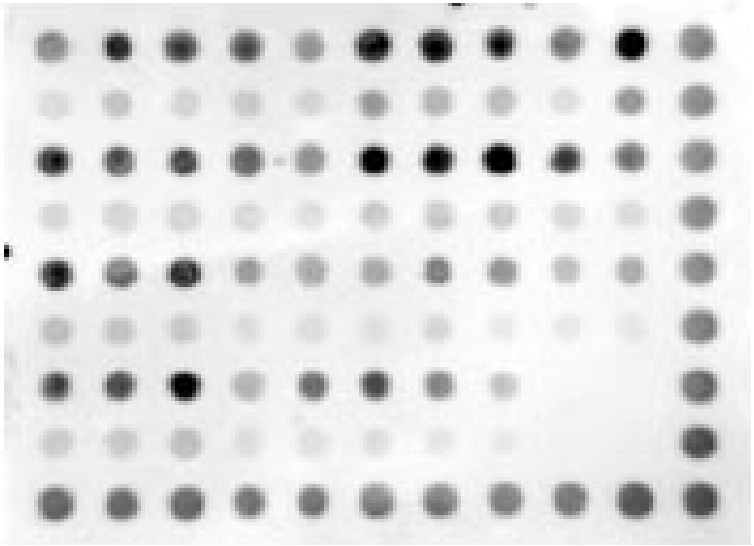
Proteins don’t act alone

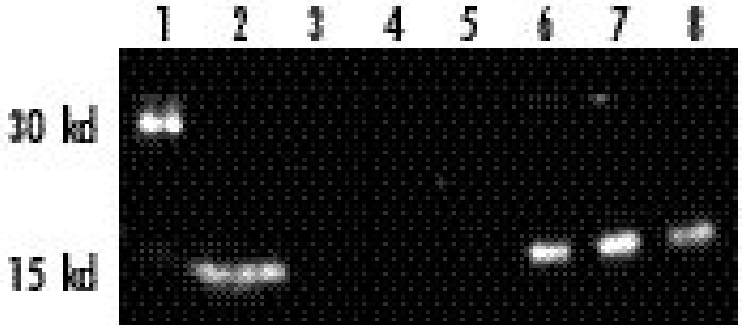


Peptidomimetic targets

Signal transduction complex lying in anticipation

Protein-protein binding





Spotted array of 39 different SH3 domains

Panomics

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 investigators**
- **Urgent need for visualization tools**
- **The importance of new statistical methods for analysis of high dimensional systems**

Suggested course reading material

- *Introduction to Proteomics,* Daniel C. Liebler, Humana Press, 2002
- *Mass Spectrometry for Biotechnology*, Gary Siuzdak, Academic Press, 1996.
- Kenyo, G, et al. *Defining the Mandate of Proteomics in the Post-Genomics Era: Workshop Report*. Mol. Cell Proteomics, 1:763-780 (2002)
- Noble G. *Modeling the heart - from genes to cells to the whole organ.* Science 295:1678-1682 (2002)
- Srinivas PR, et al. *Proteomics for Cancer Biomarker Discovery.* Clin Chem 48:1160-1168 (2002)
- Graves PR and Haystead TAJ. *Molecular Biologist's Guide to Proteomics.* Microbiol Mol Biol Rev 66:39-63 (2002)
- Zhu H and Snyder M. *-Omic approaches for unravelling signaling networks*. Current Opinion Cell. Biol. 14:173-179, (2002)
- Dierick J-F, et al. *Proteomics in Experimental Gerontology*. Exp Gerontol 37:721-732 (2002)