Proteomics and Protein Mass Spectrometry 2004

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

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

Course plan

- **Meet Tuesdays/Fridays in MCLM 401 from 9-11 am (Jan 6-Mar 19)**
- **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/exams**
- **Where possible, materials from each class will be placed on the proteomics website (go to http://www.uab.edu/proteomics - click on Resources)**

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**

BMG 744 Course content

Jan 6	Barnes/Kim	The world of proteins – beyond genomics
Jan 9	H Kim	The proteome, proteomics and where to start
Jan 13	L Brandon	Isolation of specific cells and subcellular fractions
Jan 16	M Baggott	Techniques of protein separation
Jan 20	H Kim	Protein separation by electrophoresis and other 2D-methods
Jan 23		*Student presentations*
Jan 27	S Barnes	Mass spectrometry of proteins and peptides: principles and principal methods
Jan 30	S Barnes	MALDI and peptide mass fingerprinting
Feb 3	S Barnes	Interpretation of peptide fragmentation spectra – peptide sequencing and posttranslational modifications
Feb 6		*Class demo of methods*
Feb 9		*Mid-term exam*
Feb 13	E Lefkowitz	Connecting proteomics into bioinformatics
Feb 16	S Meleth	Statistical issues in proteomics and mass spectrometry
Feb 20	S Barnes	Qualitative and quantitative burrowing of the proteome
Feb 24	Kim/Townes	Protein-protein networks/Affinity isolation/immunoprecipitation
Feb 27	P Prevelige	Protein structure by H-D exchange mass spectrometry
Mar 2	S Barnes	Enzymology, proteomics and mass spectrometry
Mar 5		*Student presentations*
Mar 9	Barnes/Wang	Tissue and fluid proteomics
Mar 12	H Kim	Application of proteomics to the brain proteome
Mar 16	V Darley-Usmar	The mitochondrial proteome
Mar 19		*Final exam*

Goals of the course

- What is proteomics?
- Why 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

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 has had meetings in November, 2002 in Versailles, France and in October, 2003 in Montreal, Canada

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

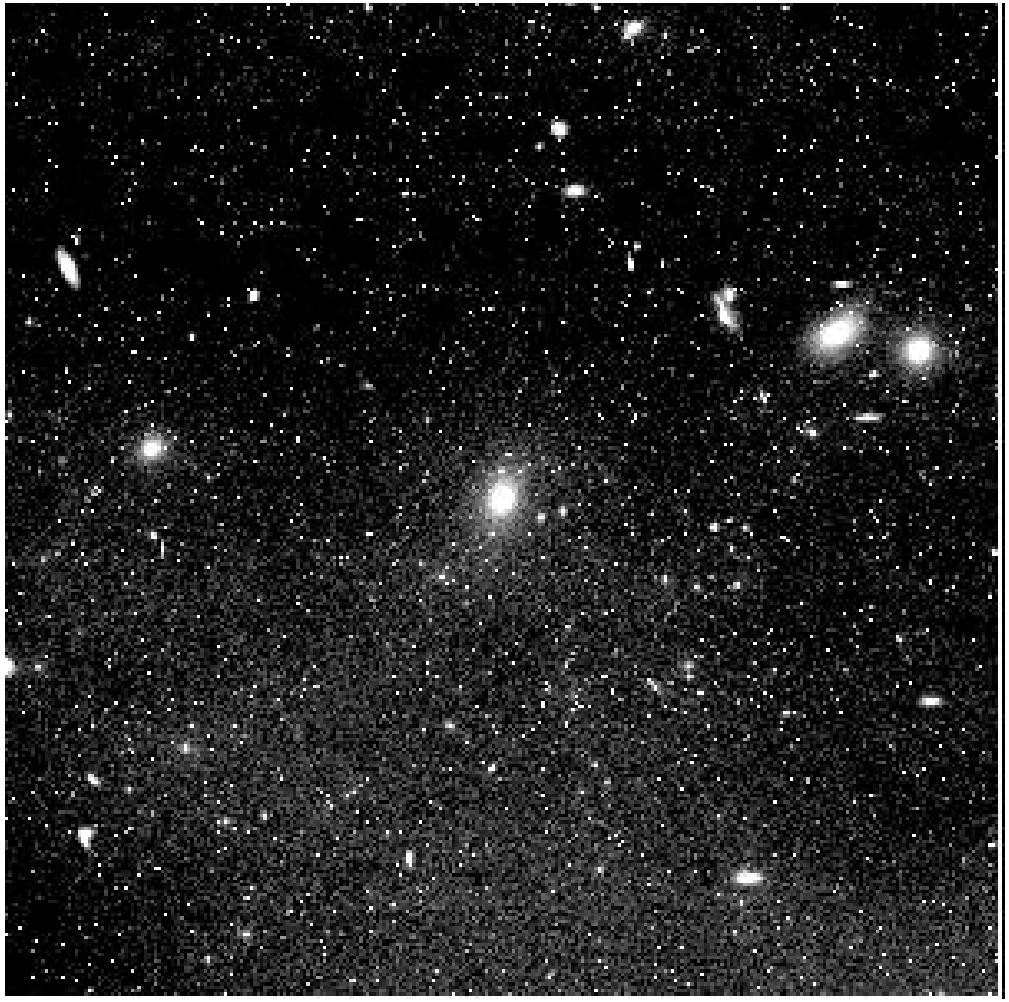
New paradigm

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

- **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

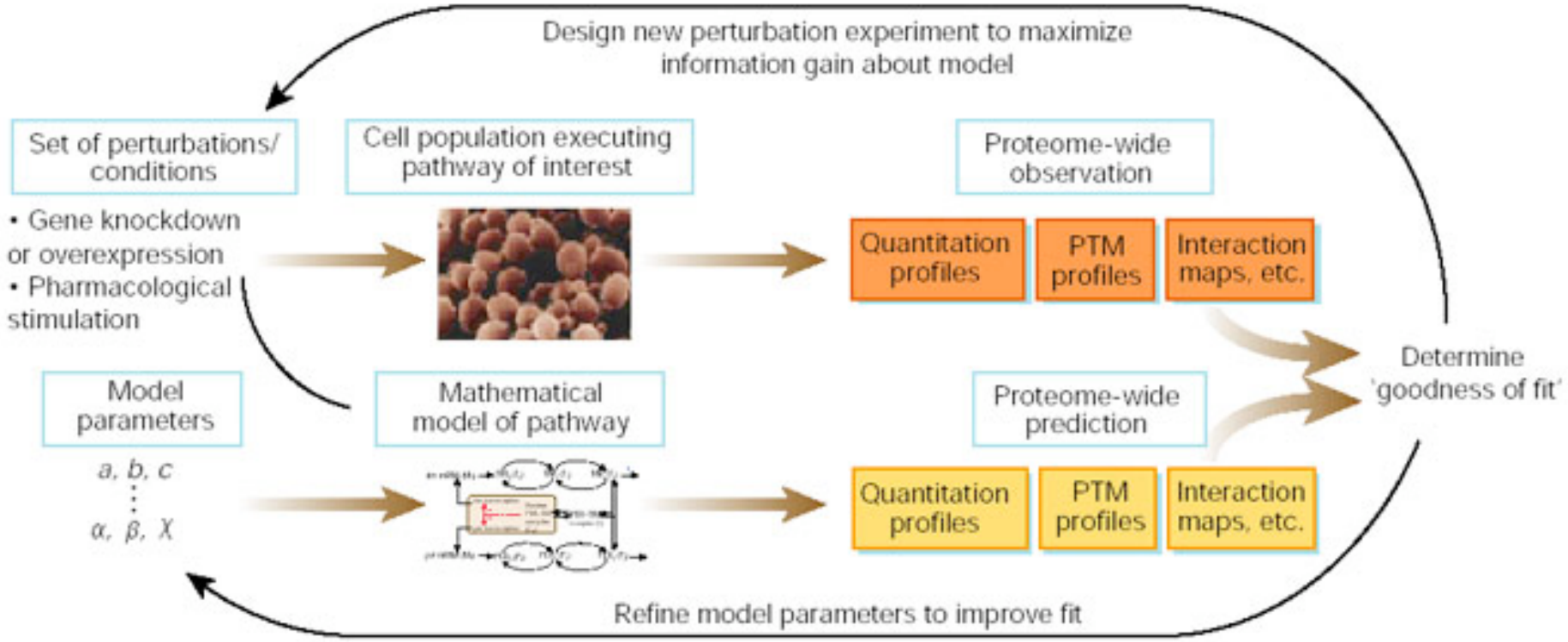
Defining disease from the proteome

- 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**

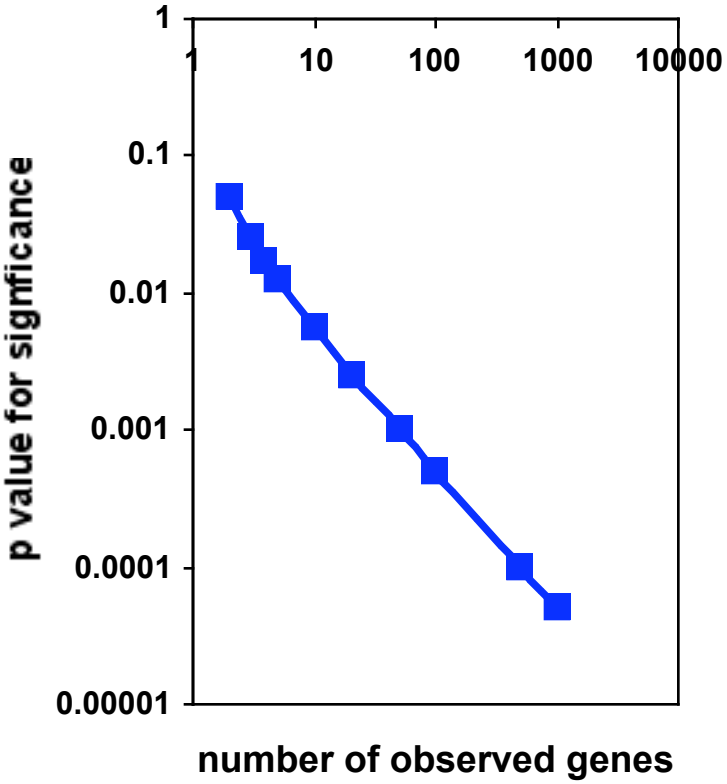
Schematic of systems biology paradigm



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

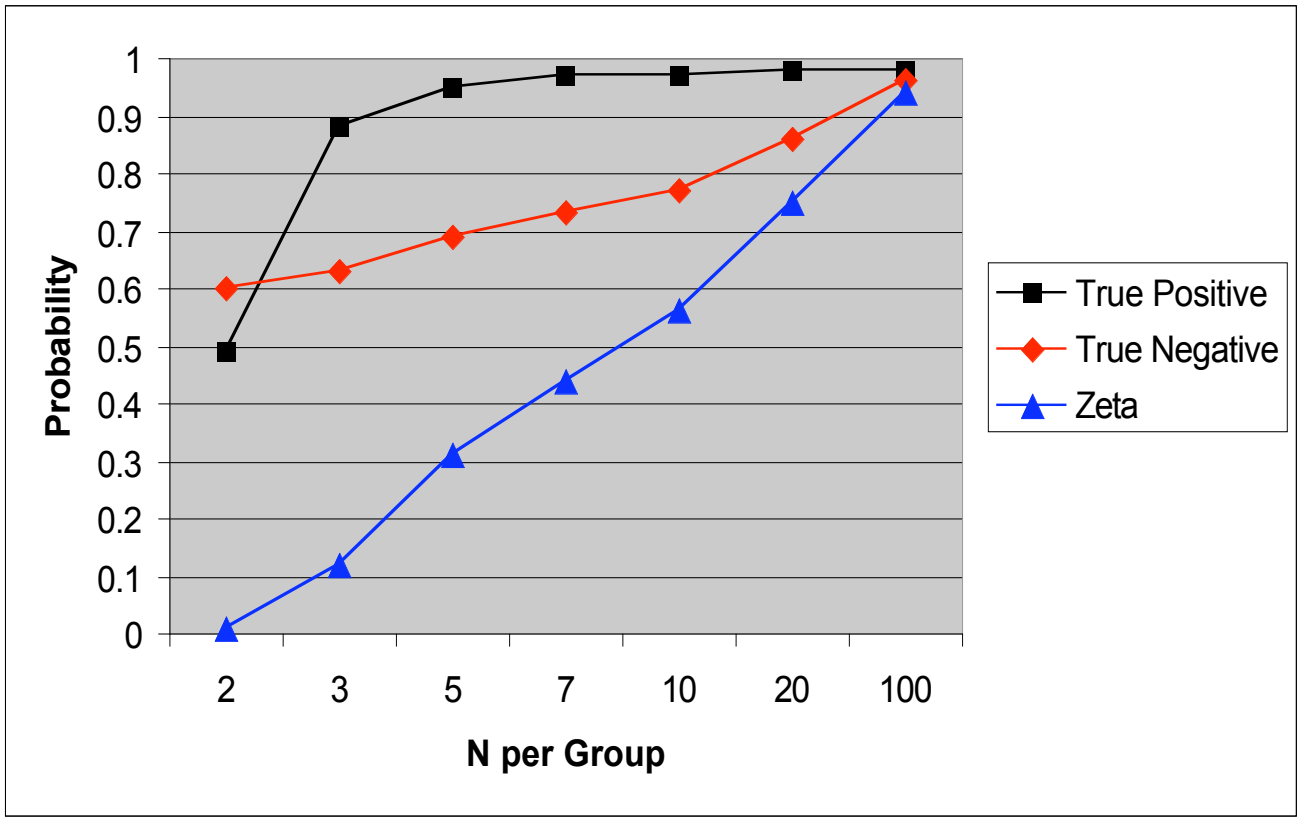
High dimensionality of microarray or proteomics data

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

Statistical realities in systems biology



For n = 3, 90% of the *true positive* changes will be observed and 35% of the *true negatives* will appear to be positive

Properties of a system and fold-change

- The primary assumption of most users of DNA microarrays (and proteomics) is 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”?
  - Reweight 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 data

- **“Expression patterns are the place where environmental variables and genetic variation come together. Environmental variables will affect gene expression levels.”**

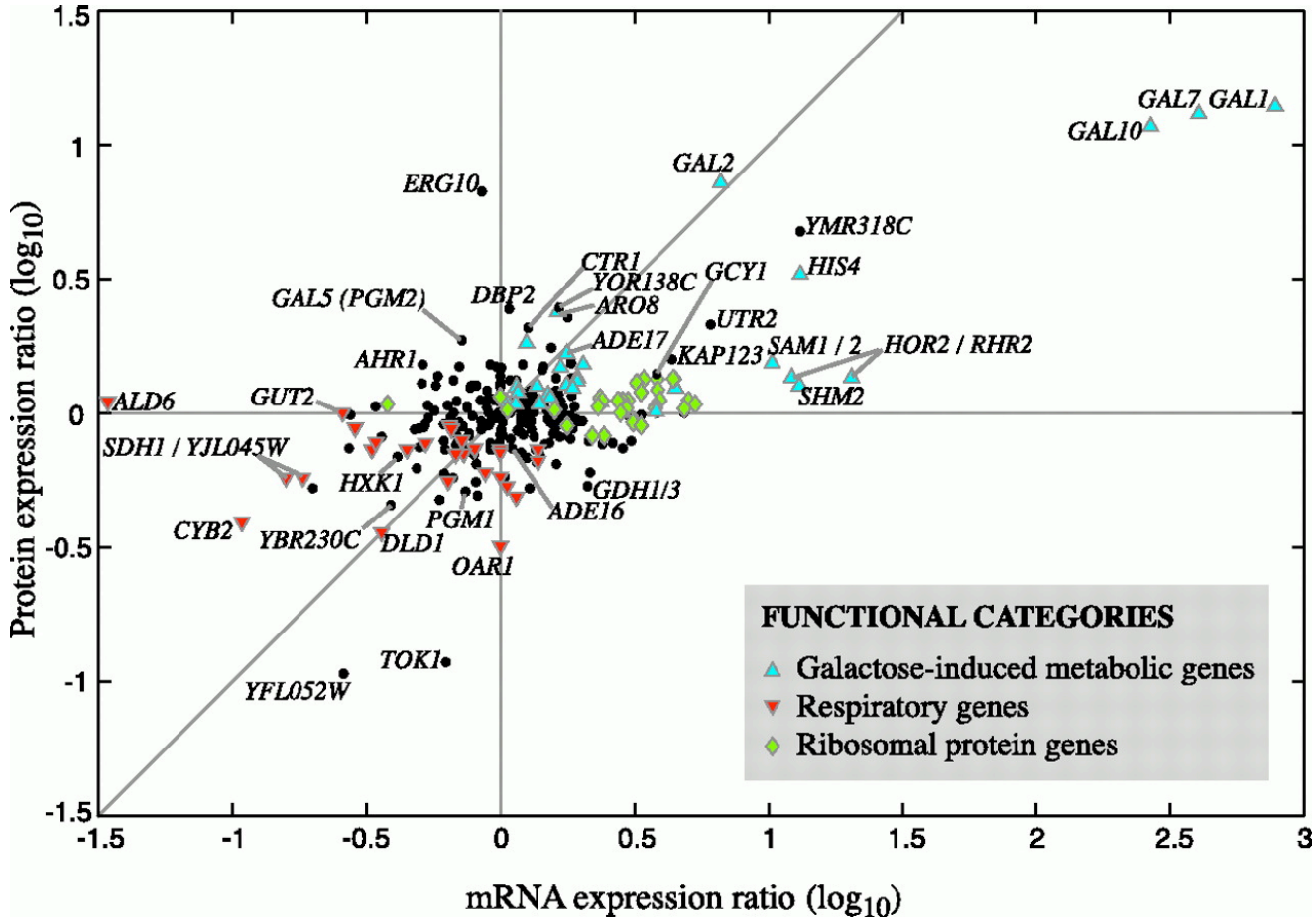
- **“Don’t we need to be very careful to understand the environmental inputs that might have an impact on that expression? Perhaps an over-the-counter herbal supplement might cause an expression pattern that looks like that of a very aggressive tumor.”**

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

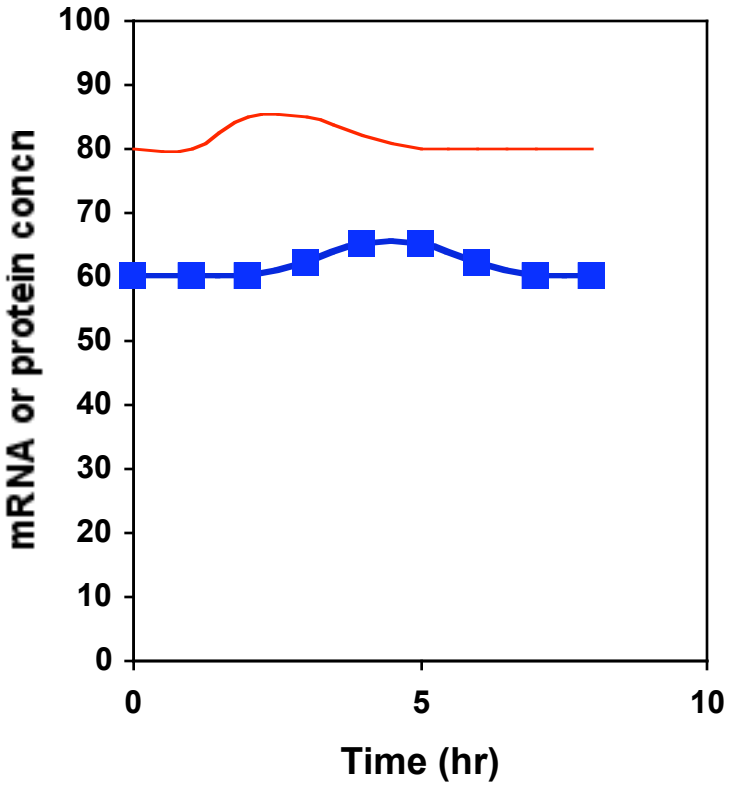


Ideker et al., Science 292: 929 (2001)

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

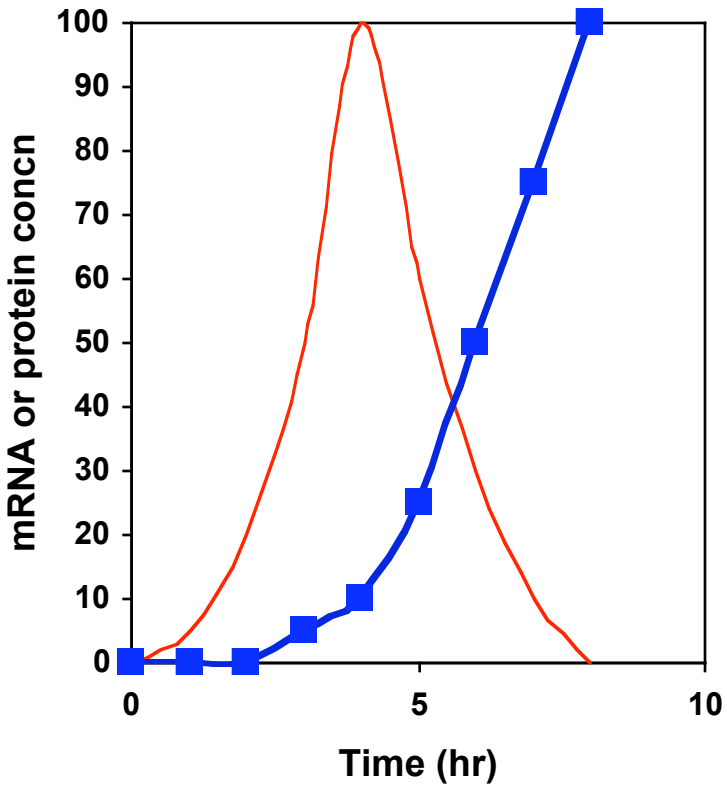
- **Even with the small perturbation, the amounts of mRNA and protein are well correlated to each other**



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**



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 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**

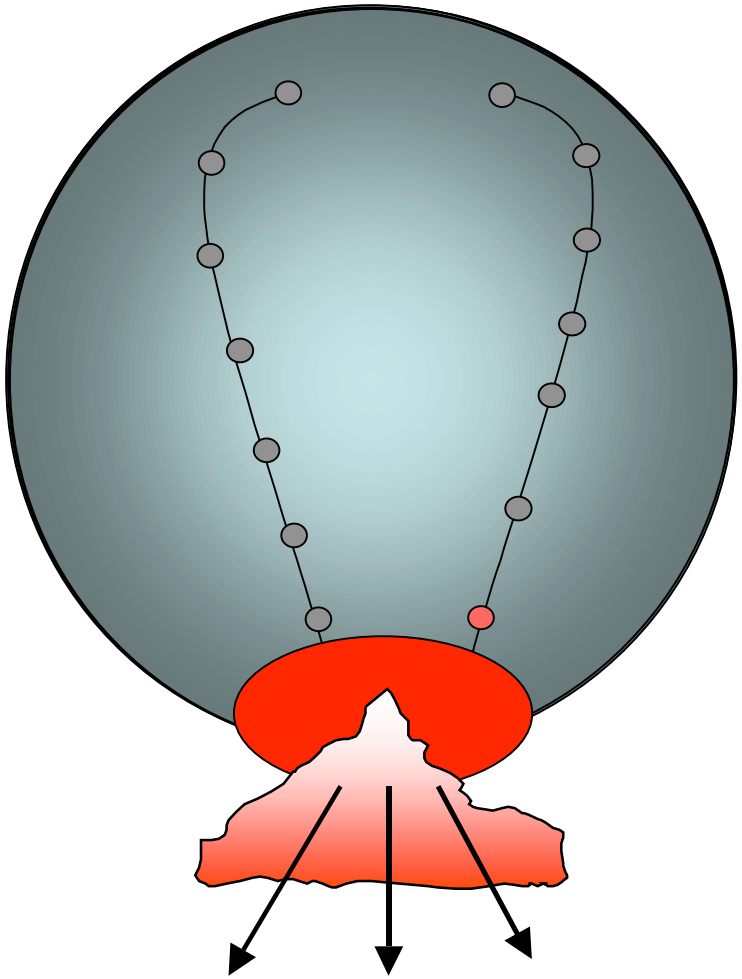
Protein informatics

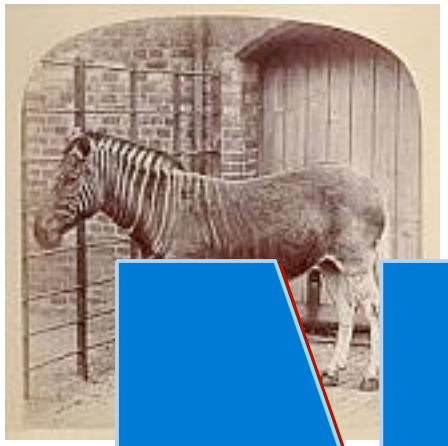
- **The predicted sequences of the proteins encoded by genes in sequenced genomes are available in many publicly available databases (subject to the limitations mentioned earlier)**
- **The mass of the protein is less useful (for now) than the masses of its fragment ions - as we'll see later, the masses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

Visualization at the whole cell level





Chaplin
silent
ie

A g

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)
- Kim, H, Page GP and Barnes S. *Proteomics and mass spectrometry in nutrition research*. Nutrition 20:155-165 (2004).
- Hood L. *Systems biology: integrating technology, biology and computation.* Mechanisms of Aging and Development 124: 9-16 (2003).
- Patterson SD, and Aebersold RH. *Proteomics: the first decade and beyond*. Nature Genetics 33 (suppl):311-323 (2003).
- Aebersold R and Mann M. *Mass spectrometry-based proteomics.* Nature 422:198-207 (2003).
- Noble G. *Modeling the heart - from genes to cells to the whole organ.* Science 295:1678-1682 (2002)
- Graves PR and Haystead TAJ. *Molecular biologist's guide to proteomics.* Microbiol Mol Biol Rev 66:39-63 (2002)
- Ping P. *Identification of novel signaling complexes by functional proteomics*. Circulation Research 93:595-603 (2003)

PROTIG and Videocast

- There is a 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
  - John Fenn, 2002 Nobel Laureate, talks at 2 pm CST on Wednesday, Jan 7th about electrospray ionization in proteomics