Clinical Proteomics: From Research to Practice

Eric Fung, MD, PhD
Vice President of Medical and Clinical Affairs
efung@ciphergen.com
Outline of presentation

- Proteomics defined
- A summary of proteomics technologies
- Focus on biomarkers to diagnostics
- Practical aspects of a clinical proteomics study
Proteome - defined

- All the proteins expressed by a genome.
- "Functional Proteome" = all the proteins produced by a specific cell in a single time frame.
- Can be defined as the total protein complement of a system (be that an organelle/cell/tissue/organism).
- The genome is for the most part static within a generation, whereas the proteome is more dynamic: cell cycle, development, response to environment, etc.
Why is the proteome important?

- It is the proteins within the cell that:
  - Provide structure
  - Produce energy
  - Allow communication
  - Allow movement
  - Allow reproduction

- Proteins provide the structural and functional framework of cellular life
Proteomics, defined

- The study of the expression, structure and function of proteins, and the interactions between proteins.
  - Where and when are proteins expressed? Abundance?
  - Protein modifications and activities
  - Interactions: Protein-protein, protein-DNA, protein-small molecule, etc
  - Protein structure

- It represents the protein counterpart to the analysis of gene function.

- Initial goal was to rapidly identify all the proteins expressed by a cell or tissue – a goal that has yet to be achieved for any species
Why is proteomics important?

- Identification of proteins changed in disease conditions
- Identification of pathogenic mechanisms
- Promise in novel drug discovery via analysis of clinically relevant molecular events
- Contributes to understanding of gene function
Why Proteomics?

Same genome
Different proteome
Proteomics vs Genomics

- **Proteins actually do the work of the cell**
  - DNA/RNA analysis cannot predict the amount of a gene product made (if and when)
  - RNA quantitation does not always reflect corresponding protein levels

- **Genomics cannot predict post-translational modifications and the effects thereof**
  - Post-translational modification is extensive: so far more than 200 different types of modifications have been reported! How does modification alter protein function?
  - ~30,000 human genes yields 1,200,000+ protein variants?!

- **Multiple proteins can be obtained from each gene (alternative splicing)**
38,016 possible isoforms! How many are produced and functional?
Proteomic Methodologies

- **Analysis of protein expression patterns**
  - 2-D gel electrophoresis
  - Protein separation – mass spectrometry
    - SELDI, MALDI, ICAT, LC(n)-MS(n), ESI

- **Analysis of protein sequence information**

- **Analysis of protein structure/function relationships**

- **Bioinformatics/statistics**

- **Affinity approaches**
  - Protein arrays

- **Genetic methods – tagging every protein**
  - Prototype – Yeast two-hybrid assay
Components of Human Serum

Tirumalai et al., Figure 1
“Unbiased” proteomics approaches

Mocellin et al, TMM, 2004
Studying the proteome

Patient sample → extraction method (i.e. lipid solubilisation) → separation methods (2D gels) → comparison methods to identify changed proteins in gels

protein fragmentation and mass spec. → protein extraction from gels (spot cutting) → matching spectrum to virtual spectrum for protein identification
Two-dimensional gel
ProteinChip® technology

Study Design
- Fractionate
  - Q-HyperD, CM-HyperD, IMAC-HyperCel,
  - HIC, DEAE, MEP, IDM

Profile
- Q10, CM10, IMAC30, H50, RS100

Collect and Preprocess
- Baseline subtract, Normalize
- Calibrate

Characterize and Identify
- Peptide Mass Fingerprinting, MS/MS

Univariate, Multivariate analysis to Prioritize Biomarkers
- CiphergenExpress Hierarchical Clustering, PCA;
  - BPS

Independent Validation
- Cross Comparison
- Candidate Markers

Multivariate Models
- Protein ID
- Independent Validation by Immunoassay

ROC curve, area=0.94327, std = 0.094973, alpha=2.791, beta=0.47154

Discovery 1
- Discovery 2

Collect and Preprocess
- Baseline subtract, Normalize
- Calibrate
ProteinChip® Three dimensions of Resolution

Wash Conditions
- Imidazol
- CHAPS
- Urea
- Detergent
- pH
- Salt
- Organic
- Water

Native Mass in Daltons

Surface Type
- Anion Exchange
- Cation Exchange
- Immobilized Metal Affinity
- Reversed Phased
- Normal Phase

Graph showing mass distribution in daltons.
ProteinChip® System, Series 4000
Protein Microarrays

High-Throughput!

Nature Reviews | Genetics | October 2000

Protein Microarrays
Quantitative Protein Microarrays

- requires previous knowledge about proteins to be studied and that appropriate antibodies exist
- protein complexes can obscure results

MacBeath 2003
Protein Function Microarrays

- Proteins are arrayed on slide and assayed for...
  - Protein-Antibody interactions
  - Protein-protein interactions
  - Protein-small molecule interactions and complexes
  - Activity assays (e.g. phosphorylation, protease activity etc)
Protein Function Microarrays

Results

Calmodulin: identified 6 known and 33 new calmodulin binding proteins

Phosphoinositide binding: 35% of identified proteins are uncharacterized

Zhu et al, 2001
NIH Laser Capture Microdissection

Invented: Dr. Lance A. Liotta
Science’ 96, 97, 98
Case study: Prostate normal epithelium (human)
Liotta et al. (2003) *Cancer Cell*
Use of Novel Protein Array Technology: Signal Pathway Profiling in Human Breast Cancer Biopsy Specimens

Coupling Laser Capture Microdissection With True Signal Pathway Profiling

Normal/Normal (reduction mammoplasty)

2° ALONE  PY-ERK  TOTAL ERK

Ongoing work: cluster analysis with 135 phospho-specific endpoints, all normalized to the self protein for true signal pathway profiling

NCI
The next revolution: Molecular Profiling: Individualized Therapy

Patient

Biopsy

Microdissection

Gene Microarray

EGFR signature, 452 genes in 74 cases

Signal Network Profile

Choose combination therapy tailored to pathogenic defect

Monitor success of therapy

Rationale basis for revising therapy following recurrence

EGFR Pathway

Protein Microarray

Phosphorylated states of signal pathway proteins
## The Need & The Opportunity

Leading Cancer cases and deaths in the U.S. alone, 2004

<table>
<thead>
<tr>
<th>Type</th>
<th>New Cases</th>
<th>Deaths</th>
<th>Lab Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>173,770</td>
<td>160,440</td>
<td>NONE</td>
</tr>
<tr>
<td>Colon</td>
<td>106,370</td>
<td>56,730</td>
<td>CEA, Pre-gen 26</td>
</tr>
<tr>
<td>Breast</td>
<td>217,440</td>
<td>40,580</td>
<td>CA15-3, CA 27-29</td>
</tr>
<tr>
<td>Pancreas</td>
<td>31,860</td>
<td>31,270</td>
<td>CA-19-9</td>
</tr>
<tr>
<td>Prostate</td>
<td>230,110</td>
<td>29,900</td>
<td>PSA (tot, free,cmpx)</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
<td>54,370</td>
<td>19,410</td>
<td>NONE</td>
</tr>
<tr>
<td>Ovary</td>
<td>25,580</td>
<td>16,090</td>
<td>CA-125</td>
</tr>
<tr>
<td>Liver</td>
<td>18,920</td>
<td>14,270</td>
<td>AFP</td>
</tr>
<tr>
<td>Esophagus</td>
<td>14,250</td>
<td>13,330</td>
<td>NONE</td>
</tr>
<tr>
<td>Urinary</td>
<td>60,240</td>
<td>12,710</td>
<td>NONE</td>
</tr>
<tr>
<td>Kidney</td>
<td>35,710</td>
<td>12,690</td>
<td>NONE</td>
</tr>
<tr>
<td>Brain</td>
<td>18,400</td>
<td>12,480</td>
<td>NONE</td>
</tr>
</tbody>
</table>
## Current Tests

### Sensitivity and Specificity

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Indication</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSA</td>
<td>65%</td>
<td>35%</td>
<td>Prostate Cancer Screening (4ng/ml cutoff)</td>
</tr>
<tr>
<td>PAP Smear (ThinPrep)</td>
<td>77-87%</td>
<td>94-98%</td>
<td>Cervical Cancer Screening</td>
</tr>
<tr>
<td>Mammogram</td>
<td>75%</td>
<td>90-95%*</td>
<td>Breast Cancer Screening</td>
</tr>
<tr>
<td>CA-15-3</td>
<td>62%</td>
<td>97%</td>
<td>Breast Cancer Tumor Marker</td>
</tr>
<tr>
<td>HPV DNA</td>
<td>95%</td>
<td>85%</td>
<td>Risk assessment for cervical cancer</td>
</tr>
<tr>
<td>Troponin T- 10 hour</td>
<td>86%</td>
<td>96%</td>
<td>Varies with onset of MI</td>
</tr>
<tr>
<td>CA-125</td>
<td>35-55%</td>
<td>97%</td>
<td>Monitoring Ovarian Cancer</td>
</tr>
<tr>
<td>CA-19-9</td>
<td>Not established</td>
<td>Not established</td>
<td>Pancreatic Cancer</td>
</tr>
</tbody>
</table>

*Varies with Radiologist Reading*
Historical Failure of Protein Diagnostics

The Rate of Introduction of New FDA-Approved (CLIA) Diagnostic Protein Analytes Has Decreased to ~Zero
Molecular Imaging of Cells and Blood: Individualized Molecular Medicine

**Protein Microarrays**  “AI” Bioinformatic Tools  Mass Spec

Choose the optimal therapy tailored to the Individual Patient
Monitor Success of Therapy
Early Diagnosis of Disease
Early Warning of Toxicity

Courtesy: Gordon Whiteley, SAIC
How can we discover diagnostic proteomic patterns even without knowing the identity of the proteins ahead of time?
Ciphergen’s Proteomic Diagnostics Process

Expression Difference Mapping™ methods

Discovery

Interacting proteins

Modified proteins

Pathway Discovery

Interaction Discovery Mapping™ methods

Protein identification

SELDI Assisted Purification & MS

Clinical development

Assay development

Diagnostic assay

Increasing medical value
Human Acute Phase Proteins

*Host response proteins ID’ed as markers*

- Albumin (mg/ml)
- Apoliprotein family (100 ug/ml)
- Transthyretin (50 ug/ml)
- Haptoglobin
- Vitamin D Binding Protein
- Inter alpha-trypsin inhibitor 4
- Alpha-1 antichymotrypsin
- Serum amyloid A
- Complement factors
- CCL18 (ug/ml)
- Fibrinogen
- Vitronectin V10

- Transport proteins
- Protease inhibitors
- Immune system
- Clotting
Specific Diagnostic Fragments

Pancreatic Cancer
nvhsgstffkyylqgakipkeasfspr

Ovarian Cancer
mnfrpgvlssrqlgpdpvdpdhaayhpfr

Diabetes
srqlgpdpvdpdhaayhpfr

Host Proteins
Disease Processes

Inter Alpha Trypsin Inhibitor 4 (ITIH4)
The diagram illustrates the natural history of a disease over time, comparing conventional systemic therapy with targeted molecular therapy. The burden of disease is measured on the y-axis, with time on the x-axis.

- **Natural history of disease**:
  - A steep increase in burden of disease followed by a plateau.
  - Represented in black.

- **Conventional systemic therapy**:
  - A slower increase in burden of disease, followed by oscillations.
  - Represented in dark gray.
  - The threshold of detection, conventional, is marked by a dotted line.

- **Surgery**:
  - A sharp decrease in burden of disease, followed by oscillations.
  - Represented in light gray.

- **Targeted molecular therapy**:
  - A gradual increase in burden of disease, followed by oscillations.
  - Represented in medium gray.
  - The threshold of detection, proteomic, is marked by a dotted line.

- **Death** is represented by a horizontal dashed line.
Three basic rules

- **Rule #1. Know what you’re looking for.**
  - Broad questions require more samples to have clinical utility
  - Minimize sample variability
  - Control for all relevant conditions

- **Rule #2: Avoid systematic biases.**
  - Pre-analytical biases
  - Analytical biases

- **Rule #3: Don’t misuse statistics.**
  - Feature selection
  - Independent validation
How many samples do I need?

- **Formula based on univariate analysis**
  - \[ N = 2 \left[ \left( \frac{z_\alpha - z_\beta}{\mu_1 - \mu_2} \right) \sigma \right]^2 \]
  - \( z_\alpha \): z score associated with type I error (p value)
  - \( z_\beta \): z score associated with the power of the study
  - \( \sigma \): standard deviation
  - \( \mu_1 \) and \( \mu_2 \): mean of the two groups

- **Also depends on nature of samples**
  - Cell culture models < animal models < human studies
  - More confounding variables (sex, comorbidities, smoking, etc) require more samples.
Prevalence and study design

<table>
<thead>
<tr>
<th>Healthy controls</th>
<th>Benign disease</th>
<th>Cancer A</th>
<th>Cancer B</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

- Prevalence of Cancer A is $50/200 = 25\%$
- Prevalence of disease X in most studies is almost always higher than it is in the population
- Predictive power of a diagnostic test in a study is usually artificially better than it will be in a real world population
The two languages of clinical proteomics

- **Clinicalese**
  - Clinical question
  - Clinical trial design
  - Clinical specificity, sensitivity
  - Positive/negative predictive value

- **Analyticalese**
  - Precision
  - Accuracy
  - Dynamic range
  - Analytical specificity, sensitivity
The clinical question

- **Clear-cut clinical question**
  - Unmet clinical need
  - The marker will affect patient management and patient outcome

- **Know the desired results**

- **Controls just as important as disease samples**
Acquire the samples

- Minimize pre-analytical biases
- Multi-institutional roster of collaborators
- Associate with clinical trials
- Retrospective first, prospective when possible
- Controls just as important as disease samples
- Get enough samples to have a statistically meaningful result
Sample types

- **Body fluids**
  - Serum, plasma
  - Cerebrospinal fluid
  - Urine
  - Fine needle aspirate, other cytology

- **Biopsy specimens**
  - Tissue
  - Laser-capture microdissection
Processing the sample

- **Fractionation is key**
  - Anion exchange
  - Subcellular fractionation
  - Other – Imac, phospho, etc.

- **Homogenization conditions**
  - Sonication
  - Dounce
  - Polytron
  - Bead-beater

- **Lysis conditions**
  - Detergent
  - Buffer constituents
Multivariate analysis

- Physicians and biologists employ multivariate analysis routinely

- Advantages over univariate analysis
  - Accommodates biological variation
  - Accommodates systematic variation
  - Eliminates dependence on single analyte

- Many types of multivariate analysis
Types of multivariate analysis

- **Unsupervised learning**
  - No a priori “knowledge” of groups
  - Can discover new groups/subgroups
  - Generally not useful as diagnostic algorithm
  - Examples: PCA, heirarchal clustering, k-means clustering

- **Supervised learning**
  - Class assignments required for input
  - Output is a classification (diagnostic) algorithm
  - Examples: classification trees (BPS), support vector machines, neural networks
T Test Review

\[ t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}} \]
The relationship between CV and t value

<table>
<thead>
<tr>
<th>Fold Change $\overline{X}_T / \overline{X}_C$</th>
<th>CV_C $^*$</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10%</td>
<td>38.73</td>
</tr>
<tr>
<td>2</td>
<td>25%</td>
<td>15.49</td>
</tr>
<tr>
<td>2</td>
<td>50%</td>
<td>7.75</td>
</tr>
<tr>
<td>2</td>
<td>100%</td>
<td>3.87</td>
</tr>
</tbody>
</table>

$^*$: CV of group control; n= 30
The relationship between fold change and t value

<table>
<thead>
<tr>
<th>Fold Change</th>
<th>CV$_C$*</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{\bar{X}_T}{\bar{X}_C}$</td>
<td>20%</td>
<td>19.36</td>
</tr>
<tr>
<td>2</td>
<td>20%</td>
<td>19.36</td>
</tr>
<tr>
<td>1.5</td>
<td>20%</td>
<td>9.68</td>
</tr>
<tr>
<td>1.25</td>
<td>20%</td>
<td>4.84</td>
</tr>
<tr>
<td>1.15</td>
<td>20%</td>
<td>2.90</td>
</tr>
</tbody>
</table>

*: CV of group control; n= 30
Multivariate analysis

- **Multivariate algorithms**
  - Unsupervised learning
  - Supervised learning

- **Feature selection**
  - Univariate analysis (easy but not as powerful)
  - Multivariate analysis (more powerful but requires more expertise)

- Use multivariate techniques to find optimal combination of features

- Independent validation of fixed classification algorithms

- Always report validation results
How to build a model

- Divide discovery data into training and testing sets
- Feature selection: reduce from 1000s to <10 important variables
- Different methods of feature selection can lead to slightly different ranks of features but most important features will be common
- Further reduce number of features by calculating correlation, keeping non-correlated features
- Use selected features to create classification algorithms
- Bootstrapping/cross-validation help describe robustness of algorithms
- Choose the best algorithm and apply to validation data set
From model to diagnostic test

- A fixed laboratory protocol is used to generate the data to train, test, and validate the model

- Diagnostic assay measures discrete, individual components
  - Laboratory report provides values for each component
  - Measuring and reporting values for individual components gives confidence to physician

- Value for individual components applied to classification model
  - Laboratory report provides computed output from classification model
  - A reference range or cutoff value is provided
  - Samples with values outside reference range (or above/below cutoff value) are flagged for further clinical workup