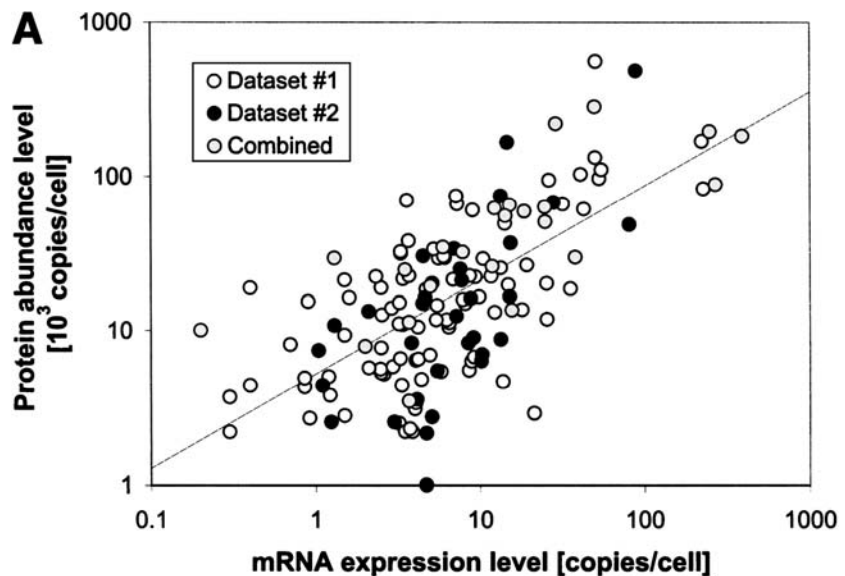


**Overview of Protein Profiling Technologies**  
(9/10/04)

Background: Comparing Protein Abundance and mRNA Expression Levels. The remarkable success of genome level DNA sequencing has placed us at a threshold of knowledge that was unimaginable 25 years ago. To enable this watershed of data to be transformed into knowledge that will be of use in diagnosing, staging, understanding, and treating human diseases will require that we not only know the sequences of the estimated >30,000 human proteins but also that we identify key changes in protein expression which portend the impending onset of disease, accurately classify at the molecular level the disease subtype, and that we understand the functions, interactions, and how to modulate the activities of proteins which are intimately involved in disease processes. One of the most fundamental approaches to understanding protein function is to correlate expression level changes as a function of growth conditions, cell cycle stage, disease state, external stimuli, level of expression of other proteins, or other variable.

While DNA microarray analysis has the ability to interrogate the relative level of mRNA expression of 25,000 or more genes; it is the concentration of proteins, their post-translational modifications (e.g., phosphorylation), and their interactions that are the true causative forces in the cell. Obviously, therefore, it is the corresponding protein concentrations and the changes in these concentrations that we would like to be studying. The difficulty in trying to predict protein from mRNA concentrations is illustrated in Figure 1. Although there is a general trend for protein concentration to rise with mRNA levels, the actual correlation is weak and protein concentrations can vary by more than two orders of magnitude for a given mRNA level. Clearly, Figure 1 provides a powerful impetus to develop improved proteomic biotechnologies.

One approach to this challenge is to find correlations between mRNA and protein expression data that might allow more accurate extrapolation of protein from the more easily obtained mRNA expression data. There have been only a few such studies, most notably in human cancers and yeast cells; and, for the most part, they have reported only minimal and/or limited correlations. In research supported by the Yale/NHLBI Proteomics Center, we found that smaller homogenous subpopulations of yeast genes have significantly higher (and lower) correlations than are found in a global "all against all" comparison. In particular, some sub-cellular localization categories - for example, the nucleolus - have significantly higher correlations ( $r = 0.80$ ) than the global correlation ( $r = 0.66$ ). Other localizations



**Figure 1.** A direct comparison of protein abundance and mRNA expression in yeast (Greenbaum et al, 2001).

present less of a correlation between mRNA and protein data, for example, the mitochondria ( $r = 0.42$ ) - possibly reflecting the heterogeneous nature and function of this organelle. In terms of functional categories, we found that although some categories, such as cell rescue, show a lower than average correlation ( $r = 0.45$ ), other functional categories, such as cell cycle, show a significant increase in correlation ( $r = 0.71$ ). Logically, this increased correlation reflects the co-regulated nature of the proteins in this functional category.

#### Reasons for lack of apparent correlation between mRNA and protein expression levels.

There are at least two probable reasons for the poor correlation between the level of mRNA and protein. Firstly, mRNAs differ in their rates of translation into protein and secondly, proteins differ in their in vivo half lives.

As one means of examining the first option, we looked at correlations between mRNA and protein abundance for those genes that had varied versus steady levels of mRNA expression over the course of the cell cycle. Broadly speaking, the cell can control the levels of protein at the transcriptional level and/or at the translational level. Logically, we assumed that those genes that show a large degree of variation in their expression are controlled at the transcriptional level. Thus we would expect, and we found, a very high degree of correlation ( $r = 0.89$ ) between the mRNA and protein levels for these particular genes; the cell has already put significant energy into dictating the final level of protein through tightly controlling the mRNA expression, and we assume that there would then be minimal control at the protein level. In contrast, those genes that show minimal variation in their mRNA expression throughout the cell cycle are more likely to have little or no correlation with the final protein level; the cell would be controlling these genes at the translational and/or post-translational level, with the mRNA levels being somewhat independent of the final protein concentration. And indeed, we found strikingly less correlation between protein and mRNA expression for these genes ( $r = 0.2$ ). Furthermore, we found that those genes that have higher than average levels of ribosomal occupancy - that is, a large percentage of their cellular mRNA concentration is associated with ribosomes (being translated) - have well correlated mRNA and protein expression levels ( $r = 0.78$ ). These cases probably represent a situation wherein the cell, having significantly controlled the mRNA expression to produce a specific level of protein, will probably not also employ mechanisms to control the translation. Alternatively, those proteins that have very low occupancy rates have uncorrelated mRNA and protein expression ( $r = 0.30$ ); thus, given that the cell has not tightly controlled the mRNA expression for this gene, it will dictate the resulting protein levels through rigorous controls of its translation (that is, through tight limits on occupancy).

#### Future studies of computational correlation of mRNA and protein expression levels.

Further examination of the protein classes that allow for very high correlations, along with incorporation of data associated with mRNA and protein turnover rates, will allow us to create a more rigorous methodology that should allow more accurate extrapolation of protein from mRNA abundance levels. In addition, we will extend these studies, which are described in Greenbaum et al (2003), by using the same approaches to examine relative changes rather than just the absolute concentrations of mRNA molecules and their corresponding proteins. Because of the paucity of published human protein and mRNA expression data on heart, lung and blood cells, our initial work has been carried out on yeast. One of our goals in both the NHLBI Proteomics and NIDA Neuroproteomics Centers will be to extend a similar type of analysis to the protein expression data that will be obtained during the course of the profiling studies now underway and that will be undertaken in both Centers. One important goal of these studies will be to reach the point where we can more accurately extrapolate mRNA to protein expression data and thereby, for instance, guide the selection of antibodies to be spotted onto microarrays and those proteins that will be targeted by directed MS-based protein browsing and other

technologies to permit the independent measurement of selected proteins of high potential interest.

Currently, no protein profiling technology is available that can approach the ability of microarray analysis to simultaneously profile the relative level of mRNA expression of 25,000 or more genes. Indeed, developing more powerful protein profiling technologies that are applicable and optimized for tissues of interest are one of the major goals of both the NHLBI Proteomics and NIDA Neuroproteomics Centers. As new protein profiling methodologies are implemented in these Centers, those technologies that are likely to be useful for analyzing other systems will be made available to the scientific community through the Keck Laboratory. The following sections provide a brief overview of four protein profiling technologies, MALDI-MS based peptide/protein disease biomarker discovery, differential fluorescence 2D gel electrophoresis (DIGE), isotope-coded affinity tag (ICAT)/MS based protein profiling, and multidimensional LC/MS analysis of tryptic digests of whole cell and partially purified protein extracts (MudPIT) that are currently available through the Keck Laboratory and its closely associated Centers.

Peptide/Protein Disease Biomarker Discovery. A publication by Petricoin et al (2002) is among several recent studies suggesting that naturally occurring peptide disease biomarkers that bind C16 and other supports may be identified by algorithmic analysis of MALDI-MS spectra acquired from comparatively large numbers of disease versus normal serum samples. Since preliminary data obtained in the NHLBI Proteomics Center appears to confirm this approach has merit, a new service has been made available through the Keck Laboratory that utilizes a similar overall approach as that described by Petricoin et al (2002) but which is being carried out on a superior MALDI-MS platform and which utilizes an algorithm written by staff in the NHLBI Proteomics Center. This Keck Laboratory service includes robotic desalting of serum and other biological samples using C-18 ZipTips followed by automated MALDI-MS data acquisition that covers the mass range extending from 700 to 28,000 Da. The latter figure of 28,000 Da is about the upper limit for the alpha-cyano-4-hydroxy cinnamic acid matrix that is used to acquire the MALDI-MS data. Although the mass range is adjustable, meaningful data is not usually acquired below about 700 Da due to interference from the matrix. The resulting spectra are subjected to biomarker discovery analysis using a customized Random Forest-based algorithm written by Dr. Hongyu Zhao's laboratory (Wu et al, 2003). This algorithm is designed to identify a relatively small number (e.g., 20-40) of biomarkers whose relative intensities can be used to best discriminate all disease from normal serum samples in a training set that often is composed initially of approximately 48 sera from disease and 48 sera from normal patients. The validity of the resulting biomarkers are then assessed by using them to classify a testing set which also is composed of approximately equal numbers of sera from disease and normal patients. A straightforward approach to optimize the size of the training set is to gradually increase its size. Following each incremental increase of perhaps 25 samples, the ability of the resulting biomarkers to correctly classify a testing set (of perhaps 50 sera from approximately equal numbers of disease and normal patients) is determined as a function of the size of the training set. Based on our ovarian cancer model system, the optimum training set size is about 150 samples and the success rate at classifying "unknown" samples is about 80%. Although not yet tested, we believe this service may be equally applicable to other biological fluids (e.g., amniotic fluid, urine, and tissue extracts, etc).

A major limitation of the current disease biomarker approach, which is based on analyzing only a single MALDI-MS spectrum/sample, is that its dynamic range of about 100 is eight orders of magnitude less than the approximately 10 orders of magnitude range of protein/peptide concentrations in sera. Over the next several months several approaches will be taken to increase the dynamic range and likely biological importance of the resulting

peptide/protein disease biomarkers. These approaches will include methodologies to remove very abundant albumin and immunoglobulin proteins which may suppress ionization of less abundant peptides/proteins. Additionally, our disease biomarker technology will be further improved by moving it to an LC/MALDI-Tof/Tof mass spectrometer. Two very major advantages of moving this technology to this new platform are the ability to automate the separation of sera into multiple RP-LC fractions by spotting onto MALDI-MS targets and to then have the capability of going back at anytime to any sample spot and acquire the MS/MS fragmentation data that can be used to identify the protein origin of peptide biomarkers. By first separating the sera into multiple RP-LC fractions, we hope to increase the dynamic range of MALDI analysis by removing/separating away more abundant peptides/proteins so that we may detect less abundant peptides/proteins. We hypothesize this will significantly increase the success rate at classifying unknown samples based on the resulting biomarkers.

It is important to note that the disease biomarker discovery technology nicely complements the DIGE technology described below. Since MALDI-MS response decreases with increasing MW (e.g., a  $10^3$ -fold larger amount of a 50,000 Da protein may be needed to generate the same MALDI-MS response as that generated by a 2,500 Da peptide), the disease biomarker service optimally detects the *naturally* occurring forms of small proteins and peptides that span the 800-10,000 Da range while DIGE optimally detects differential protein expression above this range.

2D Fluorescence difference gel electrophoresis (DIGE) utilizes mass- and charge-matched, spectrally resolvable fluorescent dyes (Cy3 and Cy5) to label two different protein samples in vitro prior to 2D electrophoresis. Compared to conventional 2D electrophoresis, DIGE has the major advantage that both the control and experimental sample are run in the same polyacrylamide gel. These samples are then imaged separately but because they were run in the same gel, the images can be perfectly overlaid without "warping". This reduces the number of gels that must be run to make statistically valid comparisons and raises the confidence with which protein changes between samples can be detected and quantified. Hence, changes in relative level of protein expression may be detected that are as little as 1.3-fold above background. Use of a third dye (Cy2) permits an internal standard to be created by pooling an equal aliquot of all biological samples in the experiment. The internal standard is then run on every gel in the experiment. This means that every protein spot from all samples will be represented in the internal standard. This in turn allows more accurate quantification and spot statistics between gels. Based on the literature (Zhou et al, 2002) and our experience, it is possible to profile 1,000 or more protein spots on properly prepared samples that provide well resolved 2D gels. The limit of detection for quantifying protein expression ratios is between 0.25 - 0.95 ng protein which is similar to that for silver staining (Tonge et al, 2001; Gharbi et al, 2002). Because detection is based on fluorescence, the DIGE approach has a large dynamic range of about  $10^4$ , which permits differential expression analysis of relatively low copy number proteins (Tonge et al, 2001). An additional advantage of this system is the ability to detect many protein post-translational modifications, such as phosphorylation, which often play a key role in modulating protein function and which cannot generally be detected by ICAT-MS based protein profiling (see below). The overall DIGE approach involves the following major steps:

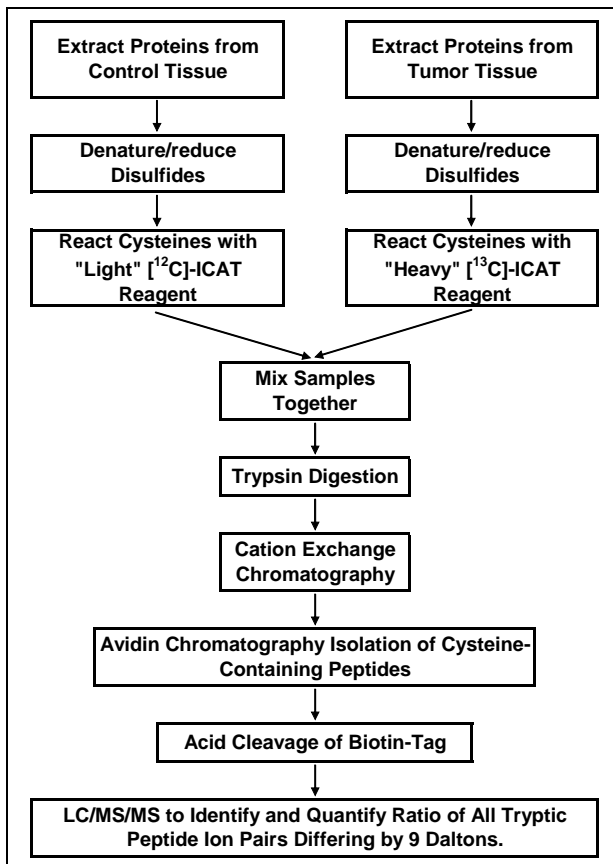
1. In vitro label 50  $\mu$ g of the control and 50  $\mu$ g of the experimental protein extracts with Amersham Biosciences Cy-3 and Cy-5 N-hydroxysuccinimidyl ester dyes. It is recommended that a third dye (Cy-2) be used as an internal (pooled 25  $\mu$ g control + 25  $\mu$ g experimental) standard to permit normalization of multiple gels and for internal normalization.
2. Mix control, experimental, and internal standard samples together (i.e., 150  $\mu$ g total protein) and subject to isoelectric focusing using Immobiline (IPG) Drystrips.

3. Carry out the SDS polyacrylamide gel electrophoresis (second) dimension on a 10 inch wide by 7.5 inch tall by 1 mM thick 12.5% polyacrylamide gel with one glass plate coated with Bind-Silane.
4. Immediately after SDS PAGE, the gel (which is still held between two glass plates) is scanned at all 3 wavelengths simultaneously on an Amersham Typhoon 9400 Imager. After scanning, 16 bit tiff files of each color channel are exported for image analysis using the differential in-gel analysis module of the Amersham DeCyder software package. After spot detection (which includes automatic background correction, spot volume normalization and volume ratio calculation), a user defined "dust filter" may be applied to each gel. This has the effect of automatically removing non-protein spot features from the gel and is followed by recalculation of experimental parameters.
5. The front glass plate is removed and the gel is then fixed and stained with Sypro Ruby, which is the fluorescent stain that will be used as a guide to excise spots of interest from the gel. The reason for using Sypro Ruby, which stains all protein in the gel, is that the Cy-dye labeling is carried out such that the extent of incorporation will be <5% in terms of mole Cy-dye/mole protein. Since the Cy-dye has a MW of about 500 Da when coupled to a protein, low MW proteins (e.g., 10 Kd) labeled with Cy-dyes will not exactly co-migrate in the SDS PAGE dimension with their non-labeled counterparts.
6. Amersham Biosciences DeCyder software is used to quantify the gel image and to identify a "pick list" of differentially expressed protein spots to be excised and subjected to automated trypsin digestion followed by MS-based protein identification. The DeCyder software can analyze any two Cy-dyed gel images, either on the same gel or on different gels, match the spots between the two images, and then identify differentially expressed protein spots. The DeCyder software automatically outputs a listing of statistically significant differences in protein expression including t-test values, using the Cy-2 internal standard. Differentially expressed spots may be identified using a number of criteria including area, volume, 3D peak slope, 3D peak height, and/or statistical variation. Protein spots that show different degrees of intensity between the two samples are highlighted by the software so they can be manually confirmed. The DeCyder software can also analyze Sypro Ruby images, match the spots found with Sypro staining to those identified with the Cy-dye stains, and then choose a 'pick list' from the Sypro stained gel image. DeCyder data can be read by labs without the DeCyder software using an HTML format.
7. The protein spot pick list is transferred to the Amersham Biosciences Ettan Spot Picker instrument which automatically excises the selected protein spots from the gel and transfers them into a 96-well microtiter plate.
8. The excised protein spots are then subjected to automated in-gel trypsin digestion on the Amersham Biosciences Ettan TA Digester.
9. The digests are then subjected to LC/MS/MS analysis on a Micromass Q-ToF instrument with the resulting MS/MS spectra then being subjected to Sequest database searches to identify proteins present in the sample.

Isotope coded affinity tag (ICAT)-based protein profiling. While both MALDI-MS based peptide/protein disease biomarker discovery and DIGE analyses comparatively profile the naturally occurring forms of peptides and proteins, ICAT analysis profiles the relative amounts of cysteine-containing peptides derived from tryptic digests of protein extracts. Recognizing that only a single tryptic peptide is needed to quantify the expression of the corresponding parent protein, the ICAT reagent was designed to affinity isolate and quantify via the use of a stable isotope the relative concentrations of cysteine-containing tryptic peptides obtained from digests

of control versus experimental samples. Hence, the newest ICAT reagent from Applied Biosystems has a thiol-specific reactive group adjacent to an alkyl linker which contains either nine  $^{12}\text{C}$  or nine  $^{13}\text{C}$  atoms. This results in a mass difference of 9 daltons between the control and experimental version of the same tryptic peptide. In addition to incorporating a cleavable biotin group (see below), another advantage of this new ICAT reagent is that unlike the previous deuterated ICAT reagent,  $^{12}\text{C}$  and  $^{13}\text{C}$ -ICAT derivatized forms of the same tryptic peptide co-elute on RP-HPLC. This greatly simplifies their relative quantitation by mass spectrometry. A very nice feature of the ICAT approach is that the in vitro incorporation of a stable isotope into one of the two samples being compared obviates the need to analyze by mass spectrometry the control and experimental samples separately. The alkyl linker in the ICAT reagent is connected to a (cleavable) biotin group which allows rapid affinity isolation of cysteine-containing tryptic peptides. While a tryptic digest of a whole cell human protein extract might produce 550,000 peptides, less than 100,000 of these might be expected to contain cysteine. Based on a search of the Swiss Database, <5% of human proteins lack cysteine and would be missed. As depicted in Fig. 2, following derivatization of the control protein extract with the  $^{12}\text{C}$ -ICAT reagent and of the experimental protein extract with the  $^{13}\text{C}$ -ICAT reagent, the pooled samples are subjected to trypsin digestion followed by cation exchange chromatography. Typically, a whole cell or tissue protein extract would be divided into 30 cation exchange fractions with each of them being subjected to avidin chromatography isolation of cysteine-containing tryptic peptides followed by LC/MS/MS analysis to identify ICAT peptide pairs and quantify the relative  $^{12}\text{C}/^{13}\text{C}$  ratios. The resulting ICAT data, which is analogous to that obtained via the use of two different fluorescent dyes in DNA microarray analysis of mRNA or DIGE analysis of protein expression, provides the corresponding ratio for the level of expression of the parent protein in the control versus experimental sample. Currently, the largest number of proteins profiled by this approach from a single sample were 491 proteins contained in microsomal fractions of naïve and in vitro differentiated human myeloid leukemia cells (Han et al, 2001). Based on our experience, we believe it may be possible to confidently identify (i.e., by isolating and quantifying 2 or more peptides/protein) and profile 150 or more proteins/sample that has been separated into 30 cation exchange HPLC fractions prior to LC/MS/MS. Additionally, it is likely that single peptides may be identified from several hundred additional proteins – with some of these also being quantified.

Our ICAT-based, quantitative protein profiling technology is currently being carried out on an Applied Biosystems API QStar XL mass spectrometer. Generally, we follow the Applied Biosystems protocol for reduction and trypsin digestion of 100  $\mu\text{g}$  amounts of extracted protein which then are fractionated into 30 pools by cation exchange HPLC (Han et al, 2001). Avidin chromatography is used to isolate Cys-containing tryptic peptides from each pool which then are individually



**Fig. 3:** Flow chart depicting ICAT/MS-based protein profiling.

subjected to LC-MS/MS on the QStar mass spectrometer at a flow rate of 300 nl/min on a 75 micron x 15 cm Vydac C-18 column equilibrated with 0.5% acetic acid, 5% acetonitrile and eluted with a 60 min acetonitrile gradient.

ICAT derivatized peptide pairs that differ by exactly 9 Da are identified and quantified by the Applied Biosystems ProICAT software. ProICAT can perform modified database searches by extracting and using only data obtained on cysteine-containing peptides, thus significantly reducing search and data analysis times. ProICAT uses a 3-dimensional LC/MS reconstruct algorithm to locate and accurately determine experimental:control (heavy:light) peak ratios in complex proteomic samples. The QStar mass spectrometer is interfaced with a LC Packings Ultimate Capillary/Nanoflow HPLC System which consists of a UltiMate™ Micro Pump and Detection Module for accurate and reproducible micro- and nanoflow delivery, a FAMOS™ Micro Autosampler for automated injections of small volume samples with zero sample loss, and a Switchos™ Micro Column Switching Module that allows for automated sample preparation and multidimensional (e.g., 2D, 3D) LC.

In order to perform high-throughput proteomic profiling using ICAT technology, it is essential to automate as many steps as possible. The standard ICAT procedure from Applied Biosystems requires manual syringe-based purification steps on both cation-exchange and avidin cartridges. These steps are extremely time-consuming (2-3 hours/sample), labor intensive, and are prone to errors. Due to these limitations one person can only process 3-4 ICAT samples in an 8-hour time-period. To address these challenges, we have installed an Applied Biosystems Vision workstation which automates both these steps. By using the Vision workstation we are thus able to automate both the cation exchange HPLC and avidin cartridge chromatography steps. This greatly enhances our ability to process samples, to maximize sample throughput on the QSTAR, and to substantially reduce the possibility of errors associated with manual syringe-based purification. The Vision workstation is a computer controlled biocompatible PEEK™ based HPLC system that enables unattended analysis and fraction collection for protein purification. It is equipped with two positive displacement piston pumps, a robotic sample handling system, eight column switching valve, and UV-Visible, pH, and conductivity monitors. The robotic sample handling system acts as both a fraction collector and autoamplifier, thus allowing automated re-injection of cation-exchange fractions onto the avidin cartridge.

Quantification based on LC-MS peak areas of stable isotope, internal standard analogs of an analyte has been used extensively and thus ICAT technology rests on a very firm foundation in this regard. Protein identification is based on database searches of the resulting MS/MS spectra using Sequest and other algorithms.

Single and multidimensional LC/MS/MS analysis can be utilized to compare tryptic digests of isolated proteins, protein complexes, partially purified and whole cell/tissue extracts. An example where a single dimension LC/MS analysis might be carried out would be to compare the profile of a tryptic digest of the naturally occurring form of a protein isolated from HeLa cells versus the same protein that had been cloned and expressed in E. coli. Since relatively few protein post-translational modifications occur in E. coli, this type of comparative analysis can quickly identify tryptic peptides containing post-translational modifications. These peptides can then be isolated in the mass spectrometer and subjected to MS/MS structural analysis using either collision induced dissociation or other techniques. In this regard, a unique feature of our new FTICR-MS, which we believe will be the platform of choice for this technology, is the capability to also fragment ions using both Infrared Multiphoton Dissociation (IRMPD) and/or Electron Capture Dissociation (ECD) techniques. These three MS/MS approaches often provide complementary structural information which can be particularly helpful when analyzing peptides containing post-translational modifications. Hence, detailed phosphopeptide sequencing has been carried out using ECD FTICR-MS/MS (Stensballe et al,

2000) and combined ECD/IRMPD FTICR-MS/MS has been used to provide detailed structural information regarding glycoproteins (Hakansson et al, 2001).

When working with partially purified or whole cell/tissue protein extracts one of the more difficult challenges is dealing effectively with the typically low extent of protein post-translational modifications. For example, only about 30% of the expected 30,000 human proteins are thought to be phosphorylated (Ficarro et al, 2002) and those proteins that are phosphorylated are usually modified at much less than equimolar ratios. Often, less than <5% of a given protein is phosphorylated at a certain position. This typically necessitates enriching the sample prior to MS analysis. Hence, following tryptic digestion of whole cell/tissue extracts, immobilized metal affinity chromatography (IMAC) can be used to enrich for phosphopeptides which can then be analyzed by LC FTICR-MS (Ficarro et al, 2002). Using this type of approach, Ficarro et al (2002) detected more than 1,000 phosphopeptides in a whole cell lysate from *Saccharomyces cerevisiae*.

For very complex samples (e.g., whole cell/tissue protein extracts) the Multidimensional Protein Identification Technology or "MudPit" approach, which utilizes a microcapillary column filled with a strong cation exchange (SCX) packing followed by a reverse phase (RP) column, provides a facile means to enhance both resolution and dynamic range (Wolters et al, 2001). A sample of tryptically digested proteins is loaded onto the columns and a specific sub-set of peptides (related to overall charge) is eluted from the SCX column using a step gradient of increasing salt concentration onto the front of the RP section. Then using an RP gradient, peptides elute from the RP column according to their relative hydrophobicity and enter the mass spectrometer for analysis. After the RP gradient is complete, the next step of the salt gradient releases another sub-set of peptides from the SCX column onto the RP column and the process repeats itself. Comparing the MudPit tandem LC approach to ordinary single stage HPLC, it has been demonstrated that the tandem approach greatly increases the number of peptides identified in a single run and, most importantly, greatly facilitates the identification of peptides from low-abundance proteins. Using this approach on the yeast proteome, Wolters et al (2001) identified 5,540 unique peptides from 1,484 proteins and demonstrated a dynamic range of detection of 10,000. This method may be extended even further to include (LC)<sup>n</sup> separations. In addition to identifying as many proteins as possible in very complex whole proteome samples, the MudPit would be ideal for identifying large numbers of proteins present in sub-cellular organelles and fractions as well as in large protein complexes brought down by immunological and "tag" approaches. It may also be possible to use comparative MudPit analysis to qualitatively identify differentially expressed proteins and their post-translational modifications. As mentioned above, we believe the very high sensitivity, mass resolution, and accuracy of our new Fourier Transform Ion Cyclotron Mass Spectrometer (FTICR-MS) funded by a recent NIH High End Instrumentation Grant (PI: Kenneth Williams, TDC: \$1.4 million) will make it the platform of choice for MudPIT profiling technology.

### Literature Cited

Gharbi, S., Gaffney, P., Yang, A., Zvelebil, M., Cramer, R., Waterfield, M., and Timms, J., (2002) "Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system." *Molecular and Cellular Proteomics* 1, 91-98.

Han, D., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnology* 19, 945-951.

- Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res* 11:1463-1468
- Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4, 117.1-117.8.
- Gygi, S.P., Rochon, Y, Franza, B.P., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19, 1720-1730.
- Hakansson, K., Cooper, H., Emmett, M., Costello, C., Marshall, A., and Nilsson, C. (2001) Electron Capture Dissociation and Infrared Multiphoton Dissociation MS/MS of an N-Glycosylated Tryptic Peptide To Yield Complementary Sequence Information. *J. Anal. Chem.* 73, 4530-4536.
- Petricoin, E.F., Ardekani, A.M., Hitt, B.A, Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simine, C., Fishman, D.A., Kohn, E.C., and Liotta, L.A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 359, 572-77.
- Stensballe, A., Jensen, O., Olsen, J., Haselmann, K., and Zubarev, R. (2000) Electron Capture Dissociation of Singly and Multiply Phosphorylated Peptides. *Rapid Commun. Mass Spectrom.* 14, 1793-1800.
- Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins E., Currie, I., Davison, M. (2001) "Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. " *Proteomics* 1, 377-96.
- Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, in press.
- Zhou, G., Li, H., DeCamp, D., Chen, S., Shu, H., Gong, Y., Flaig, M., Gillespie, J., Hu, N., Taylor, P., Emmert-Buck, M., Liotta, L.A., Petricoin, E.F., Zhao, Y.. (2002) "2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers." *Molecular & Cellular Proteomics.* 1(2), 117-24.