Meeting Summary

Proteomic Technologies Informatics Workshop

National Cancer Institute

National Institutes of Health

U.S. Department of Health and Human Services

February 8-9, 2005

The Fairmont Olympic Hotel
Seattle, WA

Table of Contents

Welcome and Introductions

Leland Hartwell, Ph.D., President and Director, Fred Hutchinson Cancer Research Center
Gregory J. Downing, D.O., Ph.D.,Director, Office of Technology and Industrial Relations, National Cancer Institute

Dr. Downing welcomed attendees and thanked them on behalf of Dr. Ann Barker, Deputy Director of Advanced Technologies and Strategic Partnerships, NCI and Dr. Andrew von Eschenbach, Director, NCI. He noted that this meeting will look toward developing new strategies to use proteomics to support biomarker discovery and related platforms. The NCI is interested in enabling partnerships that will capitalize on recent developments in genomics and imaging technologies to streamline transformational biologic applications of proteomics. Dr. Downing noted that participants at this meeting bring a wealth of ideas, experiences, and strategies that will help to frame the concepts and challenges of applying proteomics to cancer. He noted that the ultimate focus of these and other discussions is to impact clinical medicine and to improve the outlook and quality of life for cancer patients. Noting that cancer is currently the number one killer in the United States of persons under age 80, Dr. Downing charged participants with discussing information management strategies that will make proteomic data the most useful for developing platforms for the early detection, monitoring, and therapy of cancer. He noted the following goals and objectives for this meeting:

- Prioritize the development agenda of a mouse model serum and tissue proteomics data repository
- Identify the needs of a general clinical proteomics data repository that are not accommodated within the mouse proteomic technology consortia
- Outline a roadmap for developing a general clinical proteomics data repository
- Discuss proteomic data management approaches and develop strategies to frame this information to accelerate discovery and educate the public
- Discuss ways to establish standards in proteomics to minimize experimental variability, share data and information, and facilitate partnerships.

Dr. Downing noted that the NCI seeks to support data cross-validation efforts by developing consortia that are similar in scope to the Mouse Models of Human Cancer Consortium (MMHCC). The Institute wishes to develop a workplan for this project and for NCI integration with other communities to develop standards and resources for protein analysis. With an ultimate goal of facilitating clinical exploration in this area, the Institute will use its cancer Biomedical Informatics Grid (caBIG) to help develop common resources. However, issues of data validation, data mining, data reporting, and standards currently must be addressed.

Dr. Hartwell commented that the application of proteomics to cancer patients has been both encouraging and frustrating. Proteomics has great potential as a tool for the early detection of cancer, but a lack of validated biomarkers has kept the field from achieving its maximum impact in the clinical setting. Markers are currently needed for risk, detection, progression, clinical trial and therapeutic development, and individualized treatment. Dr. Hartwell noted that the combination of markers with imaging technology portends a new era of improved cancer outcomes and for medicine in general. Noting the success of Gleevec in the treatment of chronic myeloid leukemia, he observed that successful cancer drugs can result from the identification of an appropriate marker.

Dr. Hartwell also observed that the proteomics community is currently at a stage similar to that present when the Human Genome Project (HGP) was launched. Current technology enables the sampling of only a small portion of the proteome at differing levels of quality. When the HGP was at a similar juncture, the combination of teamwork, informatics, and standards for quality control and assessment propelled the project to success. Heterogeneity in human cancer provides added challenge when assigning identified markers to a specific meaning. Nonetheless, Dr. Hartwell noted his optimism about the potential for proteomics, especially in terms of outcomes for patients.

Overview of Mouse Proteomic Technology Consortia and Informatics Plans

Samir M. Hanash, M.D., Ph.D., Program Head, Molecular Diagnostics, Fred Hutchinson Cancer Research Center
Martin W. McIntosh, Ph.D.,Proteomics Computational Laboratory,Fred Hutchinson Cancer Research Center

In this session, principal investigators of the mouse proteomic technology consortia outlined their current experimental plans and introduced the data that they will generate and place into a public informatics data repository. In addition, the informatics platform currently being developed by the consortia was outlined to enable attendees to understand the scope (and limits) of the consortia informatics plans.

The "Eastern" Consortium (Samir Hanash, PI):

Dr. Hanash noted that the "Eastern" Consortium (comprised of the Fred Hutchinson Cancer Research Center, the Harvard Partners Center for Genetics & Genomics, Massachusetts Institute of Technology, the Dana Farber Cancer Institute, the Van Andel Research Institute, and Memorial Sloan-Kettering Cancer Center) emphasizes two fundamental questions: 1) Are proteomic technologies suitable for cancer marker discovery in sera from mouse models? and 2) Are mouse models suitable for discovering cancer markers that are applicable to humans? He stated that the consortium leverages the expertise and existing resources to meet program objectives without duplicating work already done or in progress. Leveraged consortium resources include engineered mouse models of different types of adencocarcinoma with genomic and transcriptomic data, extensive studies of corresponding human adenocarcinomas (*e.g.,* the Early Detection Research Network (EDRN) and genomic, transcriptomic, and proteomic data), and multi-investigator studies of human serum and plasma from the Human Proteome Organization (HUPO)'s Plasma Proteome Project (PPP).

Mouse models used at the consortium span a range of adenocarcinomas, including colon/GI, pancreas, lung, and ovarian cancers. Consortium members' experience with these human adenocarcinoma tumors will allow comparisons and transitions from mouse to human studies. A continuum of technologies will be tested, spanning the range from "shotgun" proteomics to extensive fractionation of intact proteins. Also, antibody microarray-based technologies will be applied for discovery and validation. One specific technology is the Whole Proteome Scan using an Intact Protein Analysis System (IPAS). IPAS employs dyes to label proteins, detecting and measuring low-abundance proteins using prostate-specific antigen (PSA) as a reference. This

strategy has been tested in a mouse xenograft lung cancer model to search for human proteins in the mouse plasma following tumor development from implanted human cancer cells.

Validation strategies used by the consortium include antibody microarrays and cross-validation with human tumors. The goal is to determine the relevancy of specific proteins for human cancer. The strategy will generate a large volume of data, thus highlighting the need to discern which information should be captured and placed into a repository. Using IPAS, processed samples are combined into a single set that is subject to fractionation. These fractions are fractionated a second time, and molecular weight information is determined by gel analysis. Identified proteins are then digested, and resultant peptides are analyzed using mass spectrometry (MS). An annotation database will combine all annotations in an extensible markup language (XML) file that can be annotated and deposited in the database. The data system is currently in development and lacks annotations for storage and meaningful query of these data. Dr. Hanash concluded by noting that the consortium wishes to make data publicly-available, although the challenge is to determine which data need to be made available.

The "Western" Consortium (Martin McIntosh, PI):

Dr. McIntosh discussed the "Western" Consortium, comprised of the FHCRC (laboratory integration and informatics development), the Institute for Systems Biology (ISB; informatics tools and fractionation schemes), the Pacific Northwest National Laboratory (PNNL; fractionation and mass tags table), and the Plasma Proteome Institute (PPI; antibody enrichment, fractionation, and target database). The primary consortium goal is to develop public resources for mining the mouse model serum proteome. Other goals include proof-of-principle of biomarker discovery using high resolution MS as a platform and analyzing the normal variability in serum protein concentrations among and between healthy mice. Deliverables include high quality data, a public database and query tools, and an open-source pipeline for serum proteomics.

For mouse models, genetic variability is minimized through closely-controlled breeding strategies. A two-stage sampling strategy of mouse model plasma includes a discovery cohort of samples collected just prior to sacrifice for cases and controls and a validation cohort of samples collected four weeks apart up to sacrifice for cases and controls. This strategy reflects a plan that attempts to mimic human studies. The mammary adenocarcinoma model will be profiled comprehensively, and validation samples will be banked for models of prostate, epithelial ovarian, GI adenoma, skin papilloma, lung, lung adenoma, and mammary carcinoma. Other consortium resources include an accurate mass tag (AMT) table generated by high-resolution sequencing.

MS platforms include a Micromass LCT Premier electrospray/time-of-flight (ESI/TOF) instrument and tandem instruments, including a Thermo Finnigan LTQ and a Fourier transform ion cyclotron resonance (FTICR) instrument for the AMT database. All algorithms are built into the open-source MS pipeline and allow the identification of the mono-isotopic mass and hydrophobicity of discriminatory peptides. Biomarker discovery will be conducted via generation of a peptide array following image and peptide alignment and normalization.

Consortium efforts during the first year of the funding period will emphasize platform establishment, and efforts during the following year will generate data for the mouse model database. Numerous fractionation schemes and quantitative approaches (*e.g.,* isotope-coded affinity tags (ICAT), ^{18}O reference standards, N-terminal labeling) will be evaluated for optimization. Evaluation criteria for fractionation and quantitation include the number of unique peptides and reproducibility of signal intensity and the ways to best allocate resources in the second year of funding.

The consortium will generate a complete system of open-source tools. Data from both consortia will be stored at a single site and presented using a common analytic strategy. Limitations of this plan include a central focus on MS, use of a single organism and only a few well-defined protocols and platforms, and focus on a subset of specific research questions. The consortium welcomes input from meeting participants regarding strategies to make data and informatics resources of most use to the scientific community, including identifying other uses of consortium data, data elements critical to those use cases, and strategies to make this platform most applicable to clinical proteomics.

Discussion:

One attendee inquired about the interchange of activities between the consortia, and consortia leaders noted that a central data repository will be created for disseminating results to the public. While the consortia have differing philosophies, the groups are currently identifying common principles regarding data entry into public databases. It was also noted that HUPO will discuss the challenges of developing common standards and supporting open-access models at its Proteomics Standards Initiative Spring Workshop, held in Siena, Italy on April 17-20, 2005. Such issues represent components of a larger issue, and the consortium can be viewed as a model for other projects occurring globally. Another participant asked whether the consortia plan to cross-validate data. While time constraints are limiting, the consortia will cross-validate to the extent possible.

One workshop participant asked whether consortia members use a common database and how such a database is updated. It was noted that significant but irreproducible observations are common in proteomic databases. Another challenge has been the changes in gene models that result in different proteins being dropped from the International Protein Index. One participant suggested that reading frames that have been removed could be stored in a searchable archive that features a way to correlate archived and newer entries.

One participant asked whether there has been any discussion in the consortia with respect to informatics for protein separation/fractionation and microarray approaches. It was noted that the consortia wish to adapt tools for microarrays to the extent possible, although the consortia will process data to a point where it can be integrated into standard available microarray approaches. Another participant inquired whether the two consortia have agreed to exchange samples, and Dr. McIntosh commented that the consortia are in negotiation. Ultimately, tissues and samples from both consortia will be banked.

Another attendee commented that two of the issues currently encountered by the consortia, evaluation criteria and the indexing of various genes, were also faced by the Human Genome

Project. It was suggested that the clinical proteomics community explore ways to avoid repeating the ever-shifting mapping issues that consumed resources during the evolution of the HGP. To this end, guidelines from the community will be useful regarding how soon users wish to see proteomic data following its generation.

Discussion Sessions

Each of the following sessions featured brief presentations, with the remainder of time allocated to group discussion and input.

Session 1: Use-Cases for a Proteomics Data Repository

Discussion Leaders:
John J.M. Bergeron, D.Phil., McGill University
Raju Kucherlapati, Ph.D., Harvard Medical School-Partners HealthCare System, Inc.
Philip Jones, M.Sc., European Bioinformatics Institute

Leaders discussed the capabilities of the specific mouse proteomic technology repositories necessary to enable their use to the proteomics community. Speakers were asked to consider two sources of users: consortia and members of the public. Workshop members discussed several anticipated uses of the consortia data, informatics tools development, and the data demands (*i.e.*, raw versus processed) anticipated of the user community.

Dr. Bergeron:

The challenge is to make locally-developed approaches useful to the larger community. At McGill University, knockout models for proteins involved in the damage/regeneration of liver disease are studied. Using enrichment by clathrin-coated vesicles, tandem MS of subunits has been shown to be consistent with stoichiometric abundance. While all of the spectra can be assigned to peptides in various databases, the rat genome is constantly shifting. The vast majority of peptide clusters are assigned to five organelles. Examination of peptide clusters offers a visual tool to sift through large volumes of data.

The CellMapBase application, which is based on primary sequence rather than on protein name, is the backbone of the bioinformatics pipeline. CellMapBase consists of a protocol library plus repositories for files, images, and archive/backup/export. The annotation pipeline, moving from the CellMap database to the annotation database so that proteins are identified correctly, has proven challenging.

Discussion:

One participant inquired about use cases. Dr. Bergeron noted that users may submit tandem mass spectra, and the McGill group can determine with confidence if a spectrum can be assigned to a peptide or protein. To support the consortia, there are in-house methodologies to gather mass spectrometric data. McGill can work with the consortia to obtain raw tandem MS data or to evaluate whether a particular method is best for our database. The data are reprocessed on a regular basis to accommodate changes in reference databases. Consortium members may contact the McGill facility online to determine if specific methods are applicable.

One attendee suggested creating semantic meta-data registration, *e.g.,* registering the meaning of all data fields, so that users know immediately whether their fields map to those specified at McGill.

Dr. Kucherlapati:

Dr. Kucherlapati discussed the "information lifecycle" that spans the analytical chemistry lab, collaborative efforts, and repositories of publicly-available data. The analytical chemistry lab creates a protein identification algorithm and laboratory information management system (LIMS) that will enable multitasking, collect required annotations, store instrument files, and facilitate proteomics processes and communication efforts. A Collaboration Data Management System is then needed to integrate data produced at different sites into a unified scheme that potentially enforces minimum annotation sets for collaborative analysis and to provide an environment for analysis across all collaboration data sets. Publicly-available data are then stored in experiment repositories (*e.g.,* PRIDE; see Jones presentation for details) or reference data repositories (*e.g.,* Blind, Swiss-Prot, or Protein Data Bank). The Harvard Partners Center for Genetics & Genomics (HPCGG) leverages its custom-built Gateway for Integrated Genomics-Proteomics Applications and Data System to provide a LIMS environment. Sequest is used for protein identification. The HPCGG is currently planning to leverage a customized version of the NCI's cancer LIMS (caLIMS) for the collaboration data management system.

However, there are several "chokepoints" in the information flow under the present design. First, high-throughput versions of protein identification algorithms rely on incomplete sequence databases. Moreover, proteins that are not adequately represented in the sequence databases may never flow across the link from the LIMS to the collaboration data management system.

Dr. Kucherlapati also noted that, given deficiencies in current sequence databases, polymorphic changes within proteins and post-translational modifications may increase the false-positive rate or incorrect assignments. While it is possible to add specific instances of these items into the database, it is essential to know what one is looking for upfront. However, more robust sequence databases will become available that will be dynamic and consistently improving. Moreover, protein identification algorithms are continuing to evolve in terms of sophistication and utility. However, information loss within the current information flow and problems caused by the transport and storage of large instrument files remain challenging.

Dr. Kucherlapati offered two general directions for potential solutions: facilitating movement of instrument files and facilitating movement of algorithms to data. To enable the former strategy, means to ensure that instrument data files can be transported to researchers who wish to analyze them algorithmically must be created. For the latter, remote reanalysis must be enabled for raw instrument files that are physically dispersed among their sites of creation. Data grid technologies may be useful for such a strategy.

Discussion:

Attendees discussed the key properties and questions that users would require of proteomics data and informatics systems. It was noted that intellectual property management will be critical; once

published, supporting data must be made available. Pre-release of data depends on the nature of the data, although data and annotation must be comparable with that used in academic publication. Also, it will be essential for investigators to provide users with the information necessary to reproduce a given experiment. Recently, HUPO and the Plasma Proteome Consortium sent identical samples to 36 participating labs for analysis, yielding a slate of approaches and techniques for processing, analysis, database searching, and reporting. Thus, there is a great need for standardized, certified processes, which could in turn be referenced when an article is published. One participant noted that standardization may stifle innovation, but it was agreed that reporting to the community must be carried out through standardized processes.

Mr. Jones:

Mr. Jones discussed experiences with the PRoteomics IDEntifications Database (PRIDE; http://www.ebi.ac.uk/pride), a data repository and data transfer format for protein and peptide identifications and supporting evidence. He observed that many requirements must be considered, including the nature of likely queries and of user response, the types of proteomic data to include, ways to promote and encouraged data submission, common standards for data exchange, and the level of detail included. A wide range of queries will likely be posited, including literature reference, protein identification, protein family, peptide, sequence, sample processing methods, environmental conditions, and parameters of search engines and instruments used. Addressing such needs requires common controlled vocabularies and ontologies (*e.g.,* species, tissue, disease, genotype, instrument), clear definitions of the products that will be returned to the user, and the formats of such returns. Controlling the volume of data is also essential; the sheer volume of raw data will swell the database to terabytes in magnitude, and peak lists will initially involve gigabytes and will swell to terabytes at later stage.

In addition to allowing data submission, the flexibility to exchange data is crucial. A successful model of a collaborative effort to achieve this goal is the Protein Standards Initiative (PSI) initiative for the exchange of protein interaction data using the PSI Molecular Interaction XML format. The PSI General Proteomics Standards (GPS) Workgroup is developing data formats for submission and inter-repository exchange that include the Minimum Information about a Proteomics Experiment (MIAPE), the PSI object model, the PSI/GPS ontology, and data exchange formats such as mzData (for instrument output and peak lists) and mzIdent (for peptide and protein identifications).

PRIDE has addressed these problems by offering:

- An XML schema for transfer of proteomics protein identification data
- A relational database implementation for the data repository and a central data repository, with the intention of implementing a network of federated databases
- Secure upload of proteomic data in the PRIDE XML schema
- The ability to search the repository and download results in PRIDE XML or HTML formats
- This set of tools, made available and open-source upon release

Discussion:

Participants made several comments and suggestions regarding the efforts at the European Bioinformatics Institute (EBI), including working to devise ways to link general repositories to specific repositories based on a set of common standards for data transfer. It was also noted that the exact nature of some post-translational modifications, such as glycosylation, cannot be mapped to the EBI vocabularies. Thus, EBI should consider biological questions as well when designing annotations used in its systems.

General Discussion:

Dr. Lance Liotta commented on the public response to the raw proteomics datasets that his group provided online as part of projects with the NCI. He noted that the NCI felt that the field would benefit from raw data sets generated as platforms were modified and developed. These data were partially-analyzed. In response, hundreds of people analyzed the data using their own methodologies, and feedback suggested both improved analytic methods and inabilities to reproduce the data. However, in some instances, the data were analyzed and papers were published without discussions with Dr. Liotta's research group. He therefore stressed the need for communication between those who post data and those who analyze it and publish their results. He noted that the concept of smaller groups that share data initially before posting (*e.g.,* PRIDE) is a good idea. Also, he urged data posters to consider protections of confidentiality for human sample databases, as it cannot be assumed that users will communicate how they plan to analyze the data or communicate the results. One attendee commented that this example illustrates the importance of meta-data standards.

Participants then discussed the needs for biologists as users of proteomic data. It was observed that most biologists will not read tables of proteomic data; identified proteins must be shown to correlate with phenotypic relevance. Because conditions such as the CO_2 level and tissue-culture techniques affect the proteome, this presents a major problem for biological use. Another participant noted that transcriptomic and genomic correlation is a key to making effective use of proteomic data.

One attendee commented on the difficulty of enforcing community standards for analysis and suggested that the field consider the example set by the HGP. When the data are made available, the community will develop the tools necessary to make these data biologically relevant. One attendee asked whether the proteomics informatics community is positioned to influence editors' policies for accepting manuscripts in tandem with the ability to release data. In response, it was noted that the modified data-release policy for most of the proteome data must recognize efforts of the sequencing and bioinformatics groups, with timely release being key.

It was also observed that standards for bioinformatics cannot be divorced from those for methodologies; both must be developed in parallel. Also, the analog data from MS differs from the digital data from the human genome, and the digital format embeds a certain level of objectivity.

Session 2: Specimens, Experimental Annotations, and Data Quality

Discussion Leaders:

Steve Carr, Ph.D.,The Broad Institute

Eric Deutsch, Ph.D., Institute for Systems Biology
David J. States, M.D., Ph.D., School of Medicine, University of Michigan

Leaders discussed the specimen, experimental, and data analysis annotations required of the data repositories to be of use to the scientific community. Attendees discussed practical recommendations for generating informatics systems in the face of rapidly developing standards and responded to the use-case scenarios presented in the previous session.

Dr. Carr: Guidelines for Publication of Peptide and Protein-Identification Data

Speaking on behalf of the *Molecular and Cellular Proteomics* Working Group on Publication Guidelines, Dr. Carr noted that the dramatic increase in the number of large dataset papers being published has led to an inability to determine if results of peptide and protein identification are valid. Published studies often contain insufficient information for the reader to assess methods for data processing or protein identification criteria. Thus, it is likely that many incorrect interpretations are being published. The goals for these publication guidelines include:

- Try to ensure that high-quality, significant data are entering the proteomics literature
- Develop minimal guidelines for publication of peptide and protein identification data in molecular and cellular proteomics
- Focus initially on how identifications are made and validated
- Create guidelines that are neither burdensome nor dictatorial
- Initiate the process for requiring submission of data as a condition of acceptance for manuscripts and the logistics involved in such a process

He noted that finding a peptide match in a database is relatively easy, but knowing whether it is correct is not. It is always possible to match a tandem mass spectrum to a peptide in the database, yet incorrect matches often result from the use of low-quality peptide tandem mass spectrometric data to search the database. Most algorithms use a model based upon an empirical threshold that serves as a "cutoff" value. As such, each algorithm is associated with an unknown and variable false-positive error rate. While statistical methods to validate peptide assignments to tandem mass spectra have shown promising results, none is widely available or accepted at present.

The guidelines proposed by the working group (*Mol Cell Proteomics* 2004;3:531) include:

- Describe the search engine used and how peptide and protein assignments were made using that software, including thresholds and values specific to judging the certainty of identification and description of how applied
- Provide sequence coverage observed for each protein identified
- Increase the stringency of information required to use single peptide identifications for protein assignment
- Describe how the number of unique proteins identified was counted based on the peptides found
- Report the methods used to derive quantitative results from proteomic datasets (under development)

Dr. Carr noted that the proteomics community is data-starved, which inhibits refinement and comparison of new algorithms. Recognizing that integration and collective analysis are likely to yield new knowledge, *Molecular and Cellular Proteomics* strongly encourages submission of all tandem mass spectra mentioned in a paper as supplemental material. The journal is moving toward accepting and serving raw or minimally-processed intact liquid chromatographic/tandem mass spectrometric datasets. However, storage on journal websites is not a viable long-term solution, underscoring the need for creating public repositories.

Recommendations to the mouse models consortia to handle data include:

- Follow the *Molecular and Cellular Proteomics* guidelines
- Use common search algorithms and database to search
- Employ statistical methods to evaluate the false-positive rate
- Plan to integrate data for searching to identify weak associations not evident in single datasets
- Employ common/consistent annotation of results
- Store data in the original instrument vendor format in as minimally-processed a form as possible

Discussion:

One participant inquired if the journal has asked submitters to include a set of standards with their data, and the answer is no. In parallel, however, people are providing sets of highly-curated tandem mass spectrometric data.

Another attendee inquired if this effort reflects wider, community-based efforts and whether the stringency of the guidelines eliminates some biologically-valid identifications. Dr. Carr noted that the guidelines reflect realities by asking submitters to justify their results more stringently. He commented that a consensus view is the ultimate goal; if the community indicates that these guidelines are too stringent, then they will be modified.

Another participant noted that attempts to decrease file space are limited and recommended that the journal request data in a certain format (*e.g.,* mzXML).

Dr. Deutsch:

Dr. Deutsch commented on specimen annotation, noting that the more complex the mechanism for annotating specimens, the richer the query selection can be, the less likely that the annotations will be completed, and the longer time required to develop a good interface. Microarray and mouse community databases are sources for specimen annotation guidelines.

Points to consider include:

- Plan how annotations will map to developing standards (*e.g.,* microarray gene expression object model (MAGE-OM), functional genomics experiment object model (FuGE-OM)). MIAPE and the Minimum Amount of Information about a Microarray Experiment (MIAME) provide good roadmaps, and integration with microarray data will be needed.

- Plan how annotations being captured will integrate with existing repositories from the microarray (*e.g.,* ArrayExpress, the Gene Expression Omnibus) and mouse communities (*e.g.,* eMAGE).
- Require standard characteristics for common queries (*e.g.,* organism, strain, disease state, cell type).
- Use existing ontologies and predefined lists where possible (*e.g.,* Mouse Anatomical Dictionary, Microarray Gene Expression Data Ontology, the Digital Anatomist Foundation Model (FMA), eVOC, Open Biological Ontologies)
- Allow "anything else you've got" annotations (*e.g.,* free text, protocols, arbitrary attached documents). While these may not be searchable, valuable information is retained.
- Consider organismal independence (see the ISB's Peptide Atlas; http://www.peptideatlas.org). Even though the current goal is a repository for mouse model proteomic data, the next requirement will be data from another organism, such as human or rat.
- Hire curators for whom a tidy, complete repository is a passion, to serve as a bridge between programmers and researchers.

Discussion:

One attendee commented that certain peptides in a protein are more likely to be identified, and confidence increases if the number of hits for an entry is high. Does such an observation impact the development of protein identification tools? Dr. Deutsch mentioned the Prototypic Peptide Predictor, a tool currently under development that will show the peptide within the protein and process all possible permutations to predict its likelihood of being identified.

Another participant inquired about transforming datasets to mzXML and mzDATA. The need for generic schema was noted, and one participant commented that MAGE version 2 will have the capacity to describe specimens in any format.

An attendee observed that writing a standard and convincing a community to use it are distinct challenges. Because most labs do not have the informatics resources to adopt state-of-the-art identification tools, data submission tools are critical.

Dr. States:

Dr. States began by noting that genomics had advantages over proteomics in terms of less tissue variation, one copy of each gene per genome, few sample handling issues, and simpler considerations regarding modification. Large-scale genomics efforts offer many lessons, including developing a framework for error identification, setting standards, and validating lab performance. Noting that applications drive accuracy requirements, Dr. States noted that the error rate falls as sequencing costs increase. He highlighted several quality assurance exercises from the HGP, including using a Cooperative Research and Development Agreement (CRADA) funding mechanism, blind resequencing of test samples, estimation of error rates only after completing a megabase of sequencing, and telescoping the eight sequencing labs into three centers that locked in the major technology choices.

Regarding proteomic analyses, Dr. States commented that abundance is the single most likely predictor of whether a protein will be detected. Identifications may be highly significant even if they are not reproducible. Also, an observation must be reproducible within the original lab. Proteins can be identified at several levels, including member of a gene family, gene product, post-translational modification, transcriptional/splice variant, and complete covalent structure.

Issues in project coordination include multiple permitted formats for data submissions to databases, choice of LIMS, division of responsibilities, data storage, and project coordination. For the Eastern Consortium, the choice of whether to implement a local LIMS and whether to use the NCI's caLIMS (http://calims.nci.nih.gov/developers/) forms and interfaces within the lab is entirely up to the lab. caLIMS offers no explicit support for proteomics or genetics (it was designed for molecular biology), is generic, is integrated with caBIG, and can be adapted to the Mouse Models of Human Cancer Consortium or the Plasma Proteome Project. However, caLIMS data definitions provide a common vocabulary.

Dr. States noted also the danger in imposing too much rigidity in quality control during the early stage of proteomic technology development. Although error processes and accuracy requirements need to be more carefully defined, informatics support in the labs is currently limited. He stressed also the need for project coordination. A division of labor between individual labs and the consortium data center, archiving of data at multiple levels (*e.g.,* raw, processed, analyzed), and the early and inclusive definition of variables will all enhance project progress.

Discussion:

Participants discussed issues related to the publication of proteomic data. It was noted that the literature is ambiguous because results being published are often derived from single experiments. Aggregate data sets across labs will help to make the associations derived from literature analysis much stronger. It was suggested that the number of fractions analyzed and the number of replicate runs per sample be included in publication submissions.

It was also noted that complex mixtures will likely yield divergent and unusual results. Tools such as ProteinProphet were recommended to reduce protein identifications based on single peptides. It was agreed that different labs will continue to display variants in their reporting styles, although this does not preclude concomitant use of a communal standard.

Attendees discussed whether consortia should post raw or processed data. The advantage of making data available online is that the community can view and comment upon the processes of data collection. It was suggested that the consortia make available both minimally-processed and analyzed data, although the mechanism by which the data are posted requires discussion. One participant suggested that data that are processed in multiple steps should be posted in select steps.

The error models associated with processing proteomic data must be understood for the data to be useful. An objective understanding of associated error will enable database users to understand the data without overinterpreting them. To this end, it was suggested to provide a minimal level of filtering to prevent overinterpretation.

One participant responded that a list of candidate peptides or proteins may become a list of biologically-relevant proteins upon validation. Due to fragmentation variances, sample heterogeneity, and the variety of biomarkers associated with one cancer type, panels of biomarkers may become the true indicators of cancer detection. In this case, it will be necessary to determine the number of sera samples from different mice necessary for a marker to be defined as meriting further investigation. Moreover, standard nomenclature should be developed to distinguish between candidate markers (those not yet validated) and "true" biomarkers and enhance the public's understanding of this concept.

Another participant commented that GenBank entries were "owned" by their depositors, and comments added were attributed to the submitter. Thus, the consortia should allow users to add analyses to the consortia database, with conflicting results resolved and the resolution published. It was also suggested to have a separate data warehouse for processed data in addition to a repository for raw data. In summary, the consortia should provide both raw and processed data and an explanation of how conclusions were drawn from these data.

Session 3: Lessons and Challenges of Building Data Repositories

Discussion Leaders:
Kenneth H. Buetow, Ph.D., National Cancer Institute
Ronald Beavis, Ph.D.,Beavis Informatics, Ltd.
Mark Igra,Fred Hutchinson Cancer Research Center

Leaders presented their experiences in developing other data repositories and discussed anticipated challenges for developing proteomics repositories in the current environment of rapidly changing technology and immature standards.

Dr. Buetow:

Dr. Buetow began by discussing ways that resource-development experiences with diverse communities (*e.g.,* the human gene mapping community, caBIG, MMHCC) have contributed to lessons learned, the most basic of which is to understand the scope of the problem that the community is attempting to solve (*e.g.,* goals, needs, users). NCI biomedical informatics initiatives have a goal of creating a virtual web of interconnected data, individuals, and organizations that redefines how research is conducted, care is provided, and patients/participants interact with biomedical research enterprise. caBIG (www.cabig.nci.nih.gov) is an initiative to create a useful tool based on this goal that attempts to cover the watershed of the cancer enterprise. It is being piloted through base agreements in 45 NCI Cancer Centers that have agreed to caBIG principles. caBIG is "open" in many ways, including open source code, open access, data sharing, and "do no harm" licenses. With an understanding that tomorrow's tools will likely be different from those used today, processes are dynamic and evolutionary; an infrastructure must be designed to facilitate rapid exploration of new methods. caBIG is based around smaller, component-based software applications that can "plug-and-play" into new complex structures. Focus areas include boundaries, interfaces, and the metadata infrastructure that joins components, with the shape of boundaries defined by application program interfaces (APIs).

caBIG focuses on standards rather than standardization; data standards are developed to be used as exchange or submission formats. These standards cannot be proprietary and are developed "just in time" as solutions to real, practical problems. The caCORE (cancer Common Ontologic Representation Environment) is comprised of biomedical information objects (to allow extraction of data from their representations in databases and provide conceptual representations so that groups can agree to a common mapping), common data elements (CDEs; structured data reporting elements), and a controlled vocabulary (through the NCI Thesaurus and NCI Meta-Thesaurus).

Standards that support this infrastructure include Enterprise Vocabulary Services (EVS; a toolkit of browsers and APIs), cancer Bioinformatics Infrastructure Objects (caBIO; applications and APIs), the cancer Data Standards Repository (caDSR), and a caCORE software development toolkit. caBIG employs a compatibility matrix that indicates the varying levels of compatibility of a particular system with the grid.

Another lesson learned from previous communal efforts is that quality measures are transforming. Objective measures are critical and should track with both the qualitative and quantitative data. Experimental inputs can be as critical and important as outputs, even though the ultimate use cases may be unclear at present. caBIG has a series of resources and pilot projects that will be online in 2005, including the Tissue Banks and Pathology Tools Workspace (TBPTW) and the Integrated Cancer Research pilot. Community members are encouraged to participate in caBIG activities, submit tools and data infrastructures to caBIG repositories, and work toward making individual applications and solutions caBIG compatible.

Discussion:

One participant asked about proteomic applications for caBIG. Dr. Buetow noted that an interest group is currently working on Proteomics LIMS (estimated deployment: the 3rd quarter of 2005) and also a general-purpose XML system. He noted that caBIG is a federated infrastructure, and anyone may contribute. Another participant noted that community input will help to formulate the shape and capabilities of caBIG.

Another attendee inquired about plans to curate data that are inside repositories that will be integrated with the grid. Dr. Buetow responded that caBIG will attempt to integrate datasets that are identified by the caBIG community as important, as well as new databases when identified.

Dr. Beavis:

Dr. Beavis contextualized his presentation with a quote from Eric Steven Raymond (*The Cathedral and the Bazaar*): "Perfection (in design) is achieved not when there is nothing more to add, but rather when there is nothing more to take away." Based on the design of databases such as MIAPE and RADARS, peaks generated from mass spectra account for the vast majority of the difficulties in use. He thus suggested the following principles of database design:

- Restrict the amount of spectrometric data stored in repositories only to those data necessary to support conclusions

- Accept metadata storage and use a structured data format (*e.g.,* XML) to create a rational, simplified relational database design
- Utilize XML structure to retain object relationships
- Design a relational database for queries that can rapidly access the XML information
- Utilize external resources and do not attempt to create a database that holds all knowledge

Dr. Beavis then discussed the Global Proteome Machine Database (GPMDB) design, which represents the minimum number of tables necessary. XML contains the search parameters, statistics, and other detailed hierarchical structures of ways to put amino acids into domains to identify a protein. Such a design is easier to build and query than are larger, more annotated databases. GPMDB has 5.1 M annotations, and robots troll through the data regularly and highlight outliers. The database includes publicly-available data plus that which is contributed by the public. For a particular protein, a series of mass spectra can be evaluated and compared.

Discussion:

One participant inquired about the minimum data necessary to support conclusions, and Dr. Beavis noted that adding tables into a database is easier than removing them, so the user must decide upfront about desired conclusions. Another participant inquired about the capabilities to analyze differential display data quantitatively and semi-quantitatively. Dr. Beavis noted that, because of the variety in quantitation strategies, it will be best to decide on a method first.

Dr. Igra:

Dr. Igra discussed the repository development strategy at the FHCRC, noting that current capabilities include tracking mice and samples and storage and analysis of tandem mass spectra. Goals include usability, the ability to incorporate experiments and samples from many labs, and helping to establish a widely-used standard. He then contextualized the issues in terms of the development of the World Wide Web and Linux, which were successful due to low barriers to entry, an "evolvable" structure, and widespread utility. The strategy used by the FHCRC was to start with an extensible annotations framework and web ontology language that assigns and reads a unique identifier for any particular item. Tools for annotation are being developed (*e.g.,* sample annotators, experiment annotators, and systems customized for each lab) using an evolutionary model that is both open-source and open-process. The system is being designed for facile community participation and practical use. The object model and other components will be provided to the community.

Discussion:

One participant commented that users may be interested in an attainable "choice" protein as a test case, rather than a common "ocean" protein, as displayed in this presentation. Dr. Igra suggested querying those low-abundance proteins that are of most interest to workshop participants. Bioinformaticians could create a suggested list of proteins to analyze, and the biologists can add value to the quantitative information by contextualizing the relevance.

Another participant noted that establishing close ties between software writers and system users will facilitate the development process and create useful products. Another attendee noted that proteomic technologies develop faster than LIMS systems, making open-source development critical. Also, it is important to put constraints on the system. While input from biologists is critical for interfaces, mass spectrometrists, users, and biologists must communicate to make resources work effectively. Participants agreed that a team approach is necessary; biologists and informatics personnel must collaborate to tie results to relevance. Another attendee highlighted the Molecular Alterations in Breast Cancer initiative, designed to capture all data relative to the specific disease (*e.g.,* heterogeneous data from studies and patients, polymorphisms, epigenetic alterations). While the database design for such an undertaking is relatively trivial, a shared, concrete vision is necessary upfront to harness the data effectively.

Keynote Address: Quality Control for Large, Distributed Data Collection Efforts: Lessons From the Human Genome Project
Maynard V. Olson, Ph.D., Genome Center,University of Washington

A report issued by the National Research Council in 1988, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, D.C., 1988), provided a coherent policy framework for the HGP. The report specifically noted that a special effort should be organized and funded to create the genomic sequence map and that a diversified, sustained effort would be necessary to address technical issues. When this report was published, the total amount of sequence data in GENBANK was approximately 15.5 million base pairs (0.5% of the size of the human genome). The average length of the entries was 1064 base pairs (bp). The late 1990s witnessed an exponential growth in the number of sequences and base pairs downloaded into GENBANK, topping 28 billion bps by 2002.

In contrast to efforts to map the human proteome, the technology base for the HGP proved to be relatively straightforward and stable. By 1990, it was clear that an automated, four-color-fluorescence-based implementation of Sanger dideoxy sequencing would be used. However, many incremental improvements in the technology that occurred in the 1990s were essential to ultimate success. These advances included cycle sequencing (1989), linear polyacrylamide techniques (1994), energy-transfer dyes (1995), mutant DNA polymerases (1995), and, most importantly, quality statistics (*e.g.,* phred, 1998). Dr. Olson noted that by the late 1990s, scale-up of the HGP was imminent, bringing quality control issues to the fore.

The phred/phrap system to evaluate the quality of raw data for a sequencing trace (Ewing B and Green P. *Genome Res* 1998;8:186-194) provided the quality control tool necessary to evaluate data submitted to GENBANK. In a series of inter-center quality control exercises initiated by the National Human Genome Research Institute (NHGRI) in 1997, it quickly became apparent that the phred/phrap system provided an effective, easily adopted approach to quality control. By far the most important activity during these quality control exercises was the exchange of raw data between centers, with subsequent reanalysis by a center other than the data producer. By the time that data production scaled up steeply in 1999, there was a broad consensus that the quality control problem had been solved.

As a result, final data quality in the April 2003 release of the human genome was excellent, demonstrating an error rate of approximately 10^{-5}. Current quality-control issues involve second-

order issues such as misassemblies and gaps in difficult-to-sequence regions and optimizing the tradeoff between quality and utility in sequences that are used primarily for comparison to a small number of gold-standard genomes.

Dr. Olson also noted that the public-private competition to complete the human genome sequence strained basic scientific values. The rapid scientific progress, when combined with exuberant entrepreneurial capitalism, simultaneously created a temporary financial goldmine and misleading advertising about the benefits of solving the sequence as rapidly as possible. He noted that the intense social interest in HGP endeavors helped encourage a series of dynamics that hovers over all large-scale scientific endeavors carried out in the public eye. Dr. Olson commented that an irreducible amount of faith in one's colleagues is essential to maintain a balance for such projects; the ultimate QC issue for the scientific community is how it maintains its basic values in the face of intense social forces.

Discussion:

One audience member, noting that the characterization of the proteome parallels the HGP in terms of the speed at which proteins can be identified, asked about the trajectory of proteomic efforts relative to that observed with the genome. Dr. Olson noted that the key to rapid progress in proteomics will be the exchange of raw, unedited data between labs. Inter-lab cooperation will be the backbone of a successful proteomics initiative.

Another attendee asked for Dr. Olson's thoughts on the role of the private sector in such an endeavor. Dr. Olson replied that this role will change on a case-by-case basis. He noted that the breakdown in the relationship between the public and private sectors in the HGP occurred because parties whose goals did not overlap were encouraged to work together. He noted that candid discourse is essential at an early stage to identify the areas of overlap and shared interests.

Another participant inquired whether the genome results were over-hyped and how to balance the language of such projects to engender public acceptance. Dr. Olson reiterated the central importance of candor, noting that there is a tremendous temptation to oversell the benefits of a public project. He concluded by noting that society does support such efforts, even when it fails to pay much attention to them.

Summary: Workshop Day 1

The consortia informatics groups summarized their development plans in light of the discussions on Day 1. Day 1 discussion leaders offered their reflections on the previous day's sessions, noting a positive trajectory toward identifying action items that will move this activity forward. Reiterating the need to provide users with the tools to extract meaningful information and results from the data, leaders commented that specific action items can be implemented in the next few weeks and months that will set guidelines that will extend beyond the parameters of the mouse model consortia. Dr. Downing also noted that the NCI will launch a website on its clinical proteomics projects and the two consortia on March 7.

Breakout Groups

Two breakout groups were created, divided generally into topics for those who provide data to repositories and those who will be users of the data repositories. Groups discussed specific issues and reported back to the full panel for discussion (see Session 4, below).

Group A: Analytical Tools
Discussion Leaders:
Pierre-Alain Binz, Ph.D.,Proteome Informatics Group,Swiss Institute of Bioinformatics
Michael Ochs, Ph.D., Fox Chase Cancer Center

Group B: Data Standards and Architecture
Discussion Leaders:
Weimin Zhu, M.Eng, European Bioinformatics Institute
Simon Lin, M.D.,Duke University Medical Center

Session 4: Report Out and Discussion
Discussion Group Leaders:
Pierre-Alain Binz, Ph.D.,Proteome Informatics Group,Swiss Institute of Bioinformatics
Michael Ochs,Ph.D., Fox Chase Cancer Center
Philip Jones, M.Sc.,European Bioinformatics Institute
Simon Lin, M.D.,Duke University Medical Center

Group A:

Analytical tools necessary for proteomic analysis include data collection and federation, data processing, data and information validation, data visualization, and data mining. Tools for data collection and federation include tools to exchange information and data between repositories. These require a minimum set of common information in interoperable repositories, complete with technical and biological annotation. Data processing requires the development and assessment of quality metrics. Data and information validation includes the query of experimental design, annotation, experimental data for reference, and tools to allow comparison of results. Data visualization is crucial at each step of the process. Data mining tools include those to normalize between different techniques.

Group B:

This group began by identifying the primary users of the repository and their needs. The ideal repository will serve multiple functions, providing raw data for statisticians, LIMS data for consortia members, analytical and searching tools for community users, links to relational databases for biologists, and data to relate plasma findings to tumor findings for cancer researchers. Data challenges include size, format, meaningful modes of presentation, and the relationship between data and changes and updates in relational datasets. The repository will contain experimental data (*e.g.,* MS, arrays), meta-data (*e.g.,* search parameters, data to search against), and data on sample preparation and animal handling. Interface needs include the ability to download datasets, a browsable interface, tools to support query and analysis, and links to external databases. Major issues identified by group members included the contrast between statistical methods and manual validation of MS spectra, error rates, sample collection issues

(*e.g.,* platelet activation, proteolysis), standardization to evolving reference databases, and differing experimental practices among collaborating investigators.

Group members briefly discussed the relationship between the object model and relational schema, noting that the object model is suitable for a global standard, but relational schema are appropriate for internal implementation. Regarding components that can be standardized, group members listed HUPO PSI standards, such as mzData, mzIdent, and MIAPE.

Discussion:

Participants began by discussing inputs that will contribute to the project design and identified missing pieces. One participant noted that quality measures are necessary for data submitted to a common repository, regardless of their origin. The field needs guidelines from which to calibrate machines and build standards so that data from the consortia and other systems can be compared. Another participant suggested that raw data should be archived, albeit not necessarily in the central repository. Instead, derived conclusions, supporting evidence, and analyzed data should be stored in the central repository, and mechanisms must be created to update data periodically.

Another attendee reflected on the sense of immediacy and urgency to create a public, shared repository, both as a reference tool and a prototype for biomarker studies. caBIG can play an intimate role in this opportunity, and it was observed that this specific group is a major driver of caBIG activity in this particular space. Consortia representatives noted that caBIG is a welcome collaborator in their efforts. caBIG will take the following specific action items to assist with this effort:

1. Act as a broker to share data, tools, and intermediate products that emerge from these consortia and as a vehicle for communication with the broader community.
2. Create two groups to follow up on issues related to Breakouts A and B to create and implement practical standards for public repositories. Dr. Schaefer will coordinate these efforts.
3. Encourage participation in the appropriate caBIG special interest group by members of the assembled attendees.

Participants also discussed practical considerations for storing large volumes of data, and two strategies to support the storage of terabytes of data were suggested. First, for groups interested in remaining in a federation, it is necessary to devise a way to make virtual, distributed repositories. Second, for archival repositories, the bottleneck occurs as large amounts of data move through the "pipes" that comprise the public infrastructure. Suggested approaches to solve this problem included the pre-positioning of reference datasets, alternative strategies for pre-packaging and shipping (*e.g.,* overnight shipping of DVDs), and moving the tools to local data sites rather than moving the data to the location of the tools. One attendee noted that the Plasma Proteome Project found that shipping of datasets via DVD worked effectively.

Other suggestions for making such a resource useful for human clinical studies included establishing provisions to protect patient health information while retaining the capability to link samples and data to a specific clinical trial and PI. It was recommended to add these provisions at the front end of the design.

Dr. Downing noted that NCI is willing to meet with instrument makers to discuss common file format downloads for various mass spectrometers. He noted that the Institute would like to have one representative from this workshop participate in the dialog. He then asked attendees for suggestions on ways that NCI can help to leverage its resources for the consortia and other community-based proteomics efforts. Participants offered the following suggestions:

- Serve as an interface between the public and private sectors and help garner the public support necessary for industry to establish a marker for efficacy or treatment
- Assist with IRB issues and translation to clinical studies
- Contribute expertise with human cancer research and the biological variants in human disease.

Session 5: Requirements of a General Clinical Proteomics Informatics Resource

Discussion Group Leaders:
Stephen George Oliver, Ph.D., Faculty of Life Sciences, University of Manchester
Samir M. Hanash, M.D., Ph.D.,Fred Hutchinson Cancer Research Center
Martin W. McIntosh, Ph.D., Fred Hutchinson Cancer Research Center

In this final session, leaders discussed use-case scenarios for a general clinical proteomics resource and emphasized their views of the resource needs required to develop a general clinical proteomics data repository to support biomarker discovery.

Dr. Oliver:

The proteome is central to the functional genomics agenda; proteins are directly linked to the genome. However, identifying the proteome is technically more challenging than the genome and the transcriptome. Dr. Oliver then discussed the Proteome Experimental Data Repository (PEDRo), a database model that was developed in the Consortium for Genomics of Microbial Eukaryotes (COGEME; www.cogeme.man.ac.uk). The PEDRo model was published (Taylor CF, *et.al. Nat Biotechnol* 2003;21:247-254) following feedback from the wider community. At the time of publication, it contained no complete datasets. Recently, however, a database containing PEDRo proteomic data from seven species (Pierre) has been developed that will be online later this month.

PEDRo was designed to provide enough detail to allow analysis and comparison of results from different experiments, allow the sustainability of experiment design and implementation decisions to be assessed, and to allow protein identification to be rerun in the future using new databases or software. The system is not detailed enough to allow experiments to be rerun.

Dr. Oliver also discussed other resources, including the Genome Information Management System (GIMS; download at http://img.cs.man.ac.uk/gims), a Java-based tool that allows close integration of the programming language with the database. Using the object database, FastObjects, GIMS allows rapid access to database data from application programs and allows data to be stored in a way that reflects the underlying mechanisms in the organism. The GIMS user interface allows the user to browse the database, ask canned queries, and store and combine datasets. Results may be saved as txt, html, or XML.

He then highlighted several other resources, including myGrid (www.mygrid.org.uk), an open-source upper-middleware for bioinformatics, and the *In Silico* Proteome Integrated Data Environment Resource (*i*SPIDER; http://www.ispider.ac.uk), an integrated platform of proteomic data resources enabled as grid/web services. Existing infrastructure to support iSPIDER includes myGrid, AutoMed, PSI/Pedro infrastructure and standards, and protein identification tools at the University of Manchester.

General Discussion:

Dr. Hartwell then reflected on the workshop, noting that much of most interesting development activity is uncoordinated and most likely duplicative. He noted that this meeting will help enable these activities to collaborate, through caBIG and other means. Stressing that the activities and discussions should not end today, Dr. Hartwell reiterated Dr. Olson's challenge of having two or more groups analyze each other's raw data as a way to measure achievement of consensus. Also, the community must articulate a grand goal that is currently beyond reach, so that it defines a marker of success. He noted that no current proteomic activities espouse this type of goal, and he suggested using biomarkers for disease as an endpoint that will completely transform medicine.

A participant commented that human cancers arise from numerous mechanisms and are heterogeneous as compared to genetically-induced mouse cancers. Thus, to translate mouse proteomics to clinical studies, proteomic data must be linked to clinical information. Considering that only a subset of cancer patients may respond well to a particular therapy, the link between proteomic and clinical data will inform hypotheses for future clinical trials. Also, patient consent and confidentiality must be built into the system, and the consortia may serve as a model on which to build. Another participant noted that a database that compiles data from a variety of cancers on which biomarkers have passed some sort of empirical process will be a valuable resource.

Dr. Downing noted that the EDRN has developed an architecture for discovery and validation of biomarkers. The NCI would like these consortia to be a pathway that enables discovery in a complementary, yet different, way. How will this new resource facilitate such discovery for the clinic? One attendee suggested defining a specific challenge goal for the consortia. In 2003, attendees at a HUPO/NIH meeting set the goal of reliably identifying and quantitating 5000 proteins in serum, plasma, and tissue in a three-year time frame. While this challenge has not yet been met, posing a similar challenge to detect and quantitate a number of proteins in mouse (or human) serum would represent a goal to be met in time.

Another attendee observed that the heterogeneity of human cancers will necessitate help from the NCI Specialized Programs of Research Excellence (SPOREs) and other members of the clinical community. Although it is possible to post raw data from human specimens, consent forms may prevent publishing of background data, even if deidentified. To this end, it was suggested that the NCI could assist, perhaps by making the data available to a small group, but not the public. Also, the message for the public must be controlled; *e.g.,* a biomarker shall be defined as such only when it has been validated. Dr. Downing noted that the NCI sees these projects as a path forward, noting that the Institute is actively engaged in a pilot project among its prostate cancer SPOREs for the National Biospecimen Network to develop a shared repository for specimens

and data used in an inter-institutional biomarkers study. Several attendees commented on the natural synergy among different data types, noting that the proteomics enterprise is evolving toward a systems biology perspective.

Next Steps and Closing Remarks
Greg Downing, D.O., Ph.D.; National Cancer Institute

Dr. Downing thanked attendees for a productive and spirited meeting. He reminded them that the NCI will launch a website for the consortia and other proteomics initiatives on March 7th. Workshop participants will receive a summary from this meeting in the next ten days. Noting that it is likely that this group will convene again in the near future, Dr. Downing closed the meeting.