Genomes to Life Facility Workshop Report
http://DOEGenomesToLife.org

La Jolla, California, May 12–14, 2003

Bioinformatics in the GTL Facility for Whole Proteome Analysis

*Organizer: George Michaels, Pacific Northwest National Laboratory

*Facilitators: Michael Liebman, University of Pennsylvania; Gary Montry, Southwest Parallel Software; Deborah Gracio, PNNL

Workshop Introduction and Planning

Meeting organizer George Michaels (Pacific Northwest National Laboratory, PNNL) introduced the workshop and defined the approach and expected outcomes. The workshop's purpose was to identify bioinformatics issues related to the U.S. Department of Energy's Genomes to Life (GTL) Facility for Whole Proteome Analysis.

Contents

The workshop focused around the following questions.

If the technology exists to quantitatively and qualitatively determine amount and location of the nearly complete complement of proteins and metabolites for any cell, tissue, or microbial community:

- What quality parameters are needed to make the data useful to a biologist?
- What experiment design issues will need to be addressed?
- How do we capture these new data types?
- How would these data be integrated with other databases and experiments?
- How would these high-quality data drive computational new approaches to modeling and simulation of biological behavior at the molecular machine, cell, tissue, or microbial community level?
- What are the Grand Challenge experiments to focus on?

These questions were the basis for discussion in breakout groups during the workshop. Participants also were asked to help draft the approach to the interaction and structure of core and satellite facilities, data management, modeling, and

*The organizer and facilitators shown above planned and implemented the meeting and prepared this report. Their purpose was to provide a forum for the broad biological research community to discuss scientific and technical issues associated with planned user facilities for the Genomes to Life program. The report does not identify potential sites, leadership teams, final technical details, or funding for the facilities.

Published: September 29, 2003, http://doegenomestolife.org/pubs/bioinform_proteom_analysis_facility_051203.pdf

data fusion. Their task for the workshop was to develop a list of issues relating to PNNL's bioinformatics programs and the whole proteomics facility and to conduct a requirements analysis based on these issues.

The bioinformatics workshop assembled experts in proteomics-related fields with far-ranging experience in industry, quality assessment and control, quantitative experimental methodologies, database information management, and development and implementation of national and international standards (see p. 15, Appendix A). At the beginning of each workshop session, experts made short presentations on how they had addressed issues and challenges in their fields of expertise and examined potential roadblocks, technical challenges, requirements for relationship building, and possible approaches. The workshop agenda is included as Appendix B, p. 17.

Overview

Participants agreed that a significant component of success will be the ability to create a bioinformatics infrastructure and process that permits easy communication among the facility and satellite user facilities. Enduring standards, protocols, and quality-control methodologies will be needed to ensure that data generated by the facility is optimized for analysis and reuse.

Other key points:

- The bioinformatics slate is clean now; participants would be helping to draft the approach to facilities, data management, modeling, and data fusion.

- The proteomics facility is only one of several types planned for GTL. One goal of this workshop was to determine what kind of data will be needed from the other facilities and how the research proposals, processes, and protocols planned for the proteomics facility will impact the others.

- Proteomics technologies are in the very early phases. Technologies at national research laboratories and academic institutions have not been implemented in a robust fashion that accommodates high-throughput production.

- In the new facility, researchers will be able to do new types of global proteomics experiments. What new questions need to be asked in this research?

Overall Needs Assessment

Participants identified basic requirements for a successful bioinformatics program:

- An overarching design to tie facility bioinformatics with bioinformatics and other data-management methodologies occurring across the field.

- A good business flow, including
  - Defining the bioinformatics pathway,
  - Having good teams in place for fault-tolerant experimental design, and
  - Defining, tracking, and addressing quality issues.

- Definition of the optimal set of experiments that would explore the full range of the proteomics domain.

- Flexibility and ability to evolve over time without disrupting the work or losing the capability to access and reanalyze old data.

Issues Within the Workshop Scope

Participants agreed that the proteomics facility represents a fundamental shift in basic biology theory and practice. It will require entirely new and as yet undeveloped technologies; methodologies; and understanding of data gathering, analysis, archiving, retrieval, and interpretation. New, explicit, and highly detailed standards of process and quality control; cross-disciplinary exchange of data; and longevity and reinterpretation of old data will be critical to the success of an effective, long-term proteomics facility. Processes, standards, and infrastructure developed for this facility will need to be translated across all of GTL as the various facilities work together to optimize the quality and availability of experimental output.

Discussion Areas

Participants agreed on the following issues.

- Implementing a high-dimensional approach:
  - What would be the set of experiments, each with its own series of parameters, that would explore the full range of a domain to define the larger picture for a particular phenotype?
  - Look at three dimensions to define the microbiology event. This is not the norm for microbiology. The third dimension is the discriminator that allows the definition of causal relationships.
- Defining the business case for the work:
  - What are the success metrics for an experiment?
  - What information has to be captured?
  - How does that information integrate with the rest of the operation to support quality control and the final analysis of an entire experiment? An entire experiment is the complete collection of samples gathered around a particular scientific question. How can investigators be sure enough samples are coming through the process to answer the experimental questions posed at the beginning?
- Quality control: Experimental design, characterization and reusability of data. Many experimentalists have no experience in developing high-throughput experiments; a lot of bad microarray information already is out there.
  - How do we address quality specification issues with regard to proteomics?
  - How do we address data-quality issues for data generated by different experimental questions? A participant noted that data quality is driven by process understanding and control of the chain of operations required to generate the data. In the case of proteomics, that includes all the sample preparation, instrument calibration, and data characterization and processing. The experimental question does not drive data quality. The experimental question, however, does drive the accuracy and repeatability of the data required.
  - How do we set up experiments that incorporate controls from one experiment to the next so data relationships can be held in common? This is actually a process design issue. In mRNA expression profiling processes, this is addressed by using spike-ins of known concentrations and through the linear response range of the sensor being employed.
- In practice, discipline is required to achieve high throughput in a sufficiently consistent fashion so the biological variation by far outstrips the process variation.
  - High-throughput proteomics generates more information about the proteome; how are these data differentiated from all the rest? Need to do a good job of characterizing all sample attributes; characterization is a critical piece of describing the identified proteins in a biological functional context.
  - Protein scans are very ripe for new interpretation by algorithms. Currently, data are analyzed to some arbitrary cutoff; how can “seeds be pulled from the weeds?”
  - From the computational end, the ability to see the path forward makes a big difference in avoiding duplicative work. This requires a sufficient number of experiments and the generation of sufficient data.
- Quality control: Standards, metadata, archiving, and retrieving data. Need to support generation of new, arbitrary metadata types and reference them to specific samples. For example, a whole proteomics facility will be creating a huge data resource; we must capture data in a sufficiently robust and flexible man-

ner that a broad spectrum of the user community can access and use it.

  - Must keep original data including original images of gel electrophoresis methods because of evolving interpretation.
  - Need new standards for data quality and errors, especially for P.I.'s working with data from unfamiliar technologies. For example, the current PNNL proteomics resource will not, by itself, generate these standards. Need a discussion of how to press the issue of standards, which the Interoperable Informatics Infrastructure Consortium (I3C) has been debating (e.g., for mass spectroscopy and electrophoresis).
  - At minimum, we must define standards to be used throughout the facility and via satellite facilities and use established standards where possible.
  - Must also consider normalization of the many standards (some of which have been developed for clinical trials).
  - Because of the overwhelming data volume generated by early whole proteomics efforts such as PNNL's, initial standards will become de facto standards.

- Technology implementation
  - Informatics is driven by technology implementation.
  - How will possible parallel development and codevelopment among labs and within a lab be handled?
  - Systems exist for particular methods of throughput, but the models on which they were built do not necessarily match the model that PNNL is trying to implement. What process and pieces should be in place to implement a proteomics-metabolomics factory?
  - Need a plan to deal with each technology's way of looking at proteins. The facility will be evolving constantly to improve existing technologies and create new ones for high-throughput outcomes.
  - How can variations in instrument experiment ratios (e.g., a specific instrument that works about 50% of the time) be tracked?

- Existing models and lessons learned: Commercial ventures such as Oxford GlycoSystems, Bristol-Myers Squibb, Merck, Pfizer, and Monsanto hold these systems internally and to their competitive advantage. They're all in the big pharmaceutical sector and focused on a particular form of throughput.

- Data access and security: Investigators in many projects may want to perform other types of analysis (e.g., statistical access) on data that they may not have generated. Who will have access to data from research conducted at the facility core? What data protection and distribution agreements must the core and satellites negotiate? Who are the parties to the negotiation? In other words, who controls the data access and security?
  - How can the facility work as a distributed data system?
  - What types of tools should be in place to deal with data-perturbation issues?

- Computing system: GTL will be a tremendous data generator, and the bioinformatics program will need new algorithms. This creates new modeling opportunities, so what approach will be taken and what computing resource allocation is needed?
  - How will data reprocessing be handled and the P.I.'s notified of change (e.g., if the algorithms change)? Significant issues must be addressed about archival time, processing time, and communicating new information to users.
  - A future requirement will be the mixing and matching of data from a number of models that are up and running, then seeing the effects at a simulation and predictive level. This has not yet been done in proteomics because current models are incomplete. It is being done now with mouse models.
  - Can a pull system be used to do prediction and then the experiment to validate it? This

has not yet been done in biology because such a system is a huge computer hog. Predictions are a mathematical and computational problem, but they can save a lot of experimental time.

  - Pull system – requires a new kind of biological calculus.
  - Database experts do not understand the needs, approach, and taxonomy of biosystems experts. In biology, researchers may not know the questions they want to ask, whereas there is a better sense of parameters in the physical sciences. Programs such as Atmospheric Research Measurement (ARM), however, can give guidance on mistakes and lessons learned.
  - Microarray is a good place for lessons learned. Part of the problem has been the difficulty in communicating how much effort must be put into quality control. The commercial pharmaceutical model is good for demonstrating the importance and difficulty of communicating experimental design (and what it is!).
  - Once data are collected, they must be moved around and made available to other satellite facilities. At what level is the level of data abstraction?
  - Hardware and software design issues will be huge and difficult. Need to have estimates for software issues and processes. Develop these numbers if possible.
  - Is the intention to use open source software?

- If need is not immediate, investigators can start building or buy temporarily, knowing that they will need to create their own system. This can be expensive but necessary.
  - DOE has said philosophically that it wants everything built to be open source. Investigators can buy to meet immediate needs, but anything added should be open source.
  - Need to define how proprietary interfaces can be examined and tried out before they are purchased.
  - Need to determine how to create the most open software architecture that users can get to.

- What bottlenecks need to be addressed?
  - Generating data.
  - Getting more people involved.
  - Analytical bottlenecks exist where we have data and do not know what to do with it because it does not fit anything done before. One participant noted that if the GTL proteomics facility is to be run on a cost-effective business model, then data should not be generated by the facility unless it is the result of a well-considered, budgeted experimental plan. This statement does make sense if it refers to existing, archived data that scientists have not been able to use in the past. One of the issues with archival data, however, is that a potential user often does not have any information about data accuracy. This makes it difficult at best to use the data effectively.

Presentations

Day One

Exploring the Frontier Between Computing and Biology

John Wooley, University of California, San Diego

Summary: Biology is becoming an information science, and we are entering an era of "mesoscale" biology (e.g., somewhere between a CERN effort and cottage-scale science). The purpose of DOE facilities is to democratize access. The biggest issue facing the GTL facilities is sociology—convincing researchers of the importance of sharing information. The clusters and bridge services model allow "cottage industry" biologists to be able to access and use software and hardware.

DOE needs to emphasize projects that are novel enough to generate funding and be of collaboratory interest. Every government agency needs to have its own portfolio; however, funding crossover should be flexible because biology researchers need funds from all sorts of sources. A single agency does not have enough funds to support the needed research. Therefore, programs need to be developed with overlaps.

The major GTL challenge is to maintain a balance in the triangle of theory, computation, and experiment.

Bioinformatics and Proteomics: Lessons Learned from Argonne National Laboratory's 2D Gel Experience

Gyorgy Babnigg, Argonne National Laboratory

Summary: Database development, integration, renewal, and maintenance, as well as a consistent taxonomy, consumed far more resources than originally envisioned. The database is critical to the success of the program. Ultimately, however, the bottleneck was the availability of time on the mass spectrometer, not actual data management and analysis. Of a total of 60,000 total gels, 5000 are fully annotated and in the database. The Oracle database is currently ~0.25 TB, with 160 processors. For success, it is critical to pay attention to

- authentication methods,
- secure communications,
- flexible user accounting, and the variety of "roles" within the system.

Quality Control for High-Throughput Processes

Robert "Steve" Erb, Gene Logic, Inc.

Summary: For effective quality control, understanding sources of variation in a process is critical. This is similar to manufacturing, in which variation is characterized so the biology can be "unmasked." Variations that can be controlled must be minimized, and the impact of uncontrollable variations on the overall process must be characterized. Key points are as follows:

- How can experiments and processes be designed to assign variation and improve confidence in the measurement? This process is not static and must constantly be assessed and the results normalized by continuous quality control checks.
- "Variation" means something very different to a biologist, a statistician, and a manufacturer. This is an important factor when discussing quality control and assessment in biological experimentation and data gathering and processing.
- From a proteomics standpoint, every protein has its own chemical properties. Therefore, in proteomics, an inherent process variation is not yet known.
- Consistent labeling and terminology, as is done in physics, is critically needed.

Recommended reading: QA and QC definitions in *BioTechniques* 34: 562–3 (March 2003).

Community Databases for Disease-Focused Research

Nathan Goodman, Institute for Systems Biology

Summary: The ultimate goal of the Website is to be useful, manageable, and current. Collaborating with other database and Website owners as fully as possible is critical to reduce the amount of data actually generated and software created, allowing the focus on information that investigators truly want to deliver on their Web sites. If it's not yours, link to it – don't maintain it!

- Set a clear scope for data based on user needs and the specific field. Databases that appear to be similar (e.g., databases for different diseases) may have very different data types and orientations.

- Federation of databases, in terms of quality issues, is technically becoming easier (stable identifiers and other features are more common). Understanding the goals of federation and not overselling what the federation has accomplished are important. Federation may provide a set of databases to work with but does not necessarily produce better science.
  - Conducting analyses across databases that have acknowledged errors (i.e., are error prone) would be extremely helpful.
  - Question why federating is being done. How good is peer-reviewed literature? Is it that much better than bulk data?

- P.I. mindset is important when designing a database: Will it really be used for the intended purpose? (For example, one database designed as the source for publications was never used in that way by researchers.)

- Publication of negative results is needed to avoid duplicative research.

Day Two

Protein Database (PDB)

Philip Bourne, San Diego Super Computer Center at UCSD

Summary: Examining proteins and cells in minute detail is a requirement for successful systems biology. To achieve this, the human-computer interface is critical. P.I.'s who are expert in their biology fields are struggling to get basic information out of their computing systems. This is a major problem across the field of bioinformatics.

- A huge spectrum of computing hardware, software, and capability is distributed throughout the field.

- Conveying information visually is very effective. Literature is almost the worst medium for representing and understanding structure. Journals have been good at using the Web for distribution but very poor at taking advantage of the Web's power to display data effectively (e.g., structure, sequence). A new vision of publishing protein journal articles is online, displaying multiple different views into the information. This makes the article a living document because the database can update information readily.

- Data curation has been critical; original data for PDB was not curated well in terms of consistency, taxonomy, and current scientific understanding and questions (explicit sequence relationships). PDB was built by people with a crystallography mindset, which was not very relevant to current viewpoint. This demonstrates the need for flexibility of data retrieval and interpretation at the most fundamental level (keyword index and appropriate annotation).

- Goal is to enable the user to access information in the greatest detail without downloading every file in PDB and parsing it. This is a big change from the old approach and can be done with a strict API, taxonomy (exchange dictionary), and annotation.

- – Critical elements are visualization, with the ability to do comparisons of structure; query capability; and human-computer interface.
- – Usability testing of the Website has been critical.
- – Two communities of stakeholders are depositors (who want PDB to be a constrained resource with nothing but the purest data) and users.
- – Easily teasing out detailed information about interactions and other processes should be possible from a computer-based resource. The structure of the protein (ligand, chain, residue) should be the interface. This is a critical usability issue because protein experts understand structure, *not* computing.

From Genes to Leads: Expression Profiling in Functional Genomics

Venky Venkatesh, Monsanto

Summary: In the pharmaceutical industry, bioinformatics, statistical analysis, and research design are key factors in deciding whether to initiate and continue experimental work. Key questions and analysis points are as follows:

- Does the experiment make sense in terms of the business?
- Does it make biological sense?
- If a project appears to meet the first two requirements, a biostatistician evaluates and determines what statistical data will be gathered and used to evaluate the experiment as it progresses.
    - – Analysis of samples is in the context of business goals and biology goals established at the beginning of the process.
    - – Statistics must be reexamined as the experiment progresses to determine if it should be terminated.
- Experimental design also estimates sample production and shipping for the receiving labs.
- Good QC, which reduces reliance on process replicates, is a way of controlling cost.
- TxP data that does not pass QC is put into a database and flagged; in general, QC before sample processing eliminates this.

Lessons Learned from Drug-Target Identification for Complex Diseases

Rajeev Aurora, Pfizer Inc.

Summary: Large pharmaceutical companies have had to learn how to optimize their research and select molecular targets for drug development that have the most chance of being successful. To date, the industry has experienced an 80% failure rate in clinical trials because the drugs were insufficiently effective, safe, or economically viable. The industry is seeking ways to identify better molecular targets to help improve the success rate, reduce costs, and reduce drug side effects. Lessons learned include the following:

- Choice of the right target will increase the probability of success (low biomass in pathologic state; ideally not expressed in normal tissue; essential to the disease process; and on the cell surface if it is an antibody target).
- Excellent quality control and good metrics are essential to measure success of a target.
- Overall success requires high-throughput methods, including computation to integrate data.
- Experiment design is the key to generating data efficiently. More data means better hypothesis, improved experiment designs, and better conclusions.
- Experiment depth and breadth must be balanced thoughtfully, because no one has the resources to do both completely.
- The new paradigm is a combination of “wet” lab and “dry” (computational modeling)

work. Computational methods of tracking and manipulating data are essential to create and refine models.

Constraint-Based Analysis of Microbial Metabolism

Jeremy Edwards, University of Delaware

Summary: The objective of the chemical engineering group at the University of Delaware was to find an efficient, effective experimental design for inferring information about a biological system. The group used iterative predictive models, validating their models using wet lab experiments. Predictive models were used to guide experiments in terms of determining number of measurements, testing measurement accuracy, and identifying ways in which the system could be or was perturbed.

- Experimental design is a crucial aspect of implementing models successfully. The scope, purpose, and approach must be clearly identified because large amounts of data are generated and separating noise from information is a large task.

- Proposed experiments can be difficult to analyze conceptually.

- Metabolic constraining approaches can generate valuable insight into microbial physiology. In less than one year, the group was able to identify multivariate interactions and propose potential regulators and connectivity that had been missed previously.

- Models will drive technology development and indicate areas of new investigation.

Breakout Session Summaries

Each breakout session was driven by a specific question (see p. 1). Key points from each of the three breakout groups are summarized here.

Day One

Group One

- Investigate existing standards for unique identifiers and make recommendations about the appropriateness of those standards (as with I3C standards). International Union of Pure and Applied Chemistry standards should be considered as well because they represent ICSU (International Congress of Scientific Unions).

- Investigate queueing system used by supercomputing systems (this is for prioritization of entry into the processing and analysis queues.)

- Identify core deliverables to the satellites.
  - Facility deliverable: Interface specification and specific tools with documentation to users for community interaction with the core facility.
  - Manage expectations and communicate clearly with satellite systems.

- Outline workflow processes step by step; circulate these for review within the research community. This involves developing a consensus perspective and commitment among groups producing and analyzing the data.

- Subcommittee to discuss barcoding protocols at satellite facilities and core, unique identifiers. Have identifiers for associated groups, studies, and series. Barcoding, an essential element in tracking sample through the system, transcends the sample range from the microbial (GTL) program to complementary PNNL projects.

- To assist scheduling and QC, consider a forecasting and barcoding system from satellite facility to core. The use of this approach in the complex clinical sample handling of the

Immune Tolerance Network is critical for maintaining coherence in data and analysis streams and in the versioning of analytical tools for updating and qualifying results. ITN is supported by National Institute of Allergies and Infectious Disease, National Institute of Digestive Disorders and Kidney Diseases, and Juvenile Diabetes Foundation.

- Identify common aspects and a common data input-output structure, and have a defined method of labeling and addressing them; define file and output requirements. Adopt and extend existing standards. The compromise may be to provide users with the opportunity to do right (provide them with a template to fill out; identify required fields, and maintain optional fields that will be published back to the user; offer tools to support reuse and analysis of data). Core operations must be governed within a controlled, formatted environment for consistency of data storage, manipulation, versioning, and updating. This may present "limitations" in specific satellite groups for development and implementation of analysis and interpretation methodologies. To facilitate both sets of needs, core-associated standards will govern communication between the core and satellites and reflect a normalized abstraction of data generated at any specific satellite.

- Identify deliverables the core must generate and for whom (e.g., specifications that satellite facilities must follow, processing requirements). This will extend the level of processing provided by the core to each satellite and help to manage the expectations of participants in satellite activities.

- Full-time staff must be involved in investigating and reporting on standards and influencing their development. This key issue must be funded and called out in the proposal. It is essential to guarantee that standards developed or implemented within the core reflect the best among the research community and are capable of evaluating differences between the best and the state of the art.

- Define three scenarios of three groups of people doing the same kind of analysis. What are the associated data issues they might look at? Use this to define templates. Use a focus group approach to identify major categories of satellite users, which may be heterogeneous within certain satellites, to enable the development of appropriate scenarios that will reflect these users and their needs.

- The core provides the system administration, resources, and tools to develop the software that best serves them; this will then be freely available to everyone and will be kept by the core. The core facilitates research at satellite facilities and supports publishing. The core handles only tools, not output evaluation.
  - Satellite facilities should be given an incentive for building tools on the open source that can be made available via the core.
  - The core will serve as a "clearing house" for satellite-developed software and algorithms that meet the specifications for publishing and other uses established within the core.

- Heads of satellite facilities should form an advisory group to optimize communication among satellites and the core to minimize "surprises" in terms of resource needs and priorities.

Group Two

- The core facility must consider what technology to use now and in the future.

- Data dissemination will be key. Issues include
  - What data should be disseminated,
  - How data gathering and dissemination should be coordinated,
  - Data bottlenecks at the data-integration phase, version control, and how to define queries; and publication of data, community vs collaboration, and focus of the facility (user-centered?). Solutions include having a steering committee, user training, and user focus groups.

- Modeling issues include
  - What are the biological questions, and what resolution of data is necessary?

  - How extensible are the schemas and modeling?
  - Standards for experimental design and a systematic approach are needed.
  - Bottlenecks are human interpretation of data; integration of data; and quality of data. Solutions include frontloading informatics; cross-disciplinary training; and good communications across the two cultures of research and production.

- The facility must support
  - An iterative process with guaranteed consistent quality control,
  - Varying levels of abstraction,
  - Robustness of method to adapt to new paradigm shifts and changes in technology, and
  - Automation of processes.
- Information about operational status and experimental design must be available online. A dashboard approach for operational status is suggested.

Group Three

- Need to understand global regulation and look at multiple parameters all at once. Grand Challenge: Reverse engineer cell and reengineer specific behavior and functions.
- Areas of study to be determined—aerobic vs anaerobic.
  - Set of experiments for proteomics, metabolics, light conditions, specific protein core deregulation, population effect.
  - "Killer application" for research: Engineering of biobased fuel.
- Models
  - Need to understand organism's impact on environment, mechanisms of regulation.
  - Models must be sensitive to the biological hierarchy, have a defined framework based on a formalized hypothesis, and a defined scope. An overall framework needs to be developed to integrate data and models.
  - Models to test: Gene regulation, post-translational modification, localization changes, proteomics, metabolites.
- Experimental design and exact processes must be defined.
- Challenges and needs
  - Computational challenges: Visualization; model that accounts for all of an organism's components; modeling in many timescales; models that address environmental condition and metabolic rates.
  - What will be done with data, how will data be interrogated, and what data sets are necessary?
  - Define minimum variance acceptable for sampling.
  - Homeland security issues.
  - Quality measure: Secondary validation; determine go or no-go based on quality checks during the process. Checks include coefficient variations; organism performance vs prediction; QC cultures; sample validation; random instrumentation and sample checks; protocol checks.

Day Two

Group One

- Models – The proteomics facility is a high-throughput, network model data generator (infer as much information as possible from data to develop the network mode); research on change-based issues. Are experimental observations of proteins consistent with our existing models of interactions at the following levels?
  - – Cell components.
  - – Subnetworks.
  - – Metanetworks.
  - – Cell simulation.
  - – Cell-cell interactions.
- Specific models to test: What complexity should be associated with determining "biological activity" in these organisms, and does this challenge our current perspectives derived from nonmicrobial systems? Network types:
  - – Regulatory (signaling pathways, gene transcription).
  - – Functional (cell migration, chemotaxis; protein interaction).
  - – Metabolic pathway.
  - – Cell cycle (DNA replication).
  - – Multicellular networks.
- Data needs and experiments: These represent some but not all additional bioprocessing involved in establishing a specific gene's biological activity in normal or perturbed behavior.
  - – Post-translational modifications – informatics (predictive); wet lab.
  - – Protein half-life.
  - – Composition.
  - – Quantitation – RNA/transcriptome, proteome, metabolome.
  - – Fractionation – proteins.
  - – Fast response.
- Informatics needs: The breadth of data integrated into overall analysis extends far beyond the volumes of core-generated proteomics data. Moving beyond conventional relational data models will be critical to enabling and enhancing the interactive nature of data analysis and interpretation. Although object-oriented approaches may be most accurate in storing and evolving appropriate relationships, the technology is not in hand. The implementation of an interim object-oriented metalayer may be critical to establish the basis for complex analysis. The metalayer will include
  - – Static and kinetic models.
  - – Post-translational analysis.
  - – Collaborative, interactive, experimental analysis.
  - – Informatics project workbench to infer networks and automated reasoning tools to guide user.
- Experimental design, QC parameters, and go or no-go decision points: These points reflect the next stage in developing needs assessments for QA and QC as outlined by this group on Day One.
  - – Concept and feasibility review by satellite facility members based on data evidence, scientific basis.
  - – QC criteria (defined by the facility as well as the P.I.) —Use software monitor with exit boxes.
    - ◦ Initial conditions QC and experimental flow controller (templates are correctly completed).
    - ◦ Sample status.
    - ◦ Instrumentation status and calibration.
    - ◦ Barcode everything: Time, date, who conducted the work, "chain of custody."

- ᵒ Staff qualifications and training to do procedures.
- ᵒ Exits for failure of process line or one data stream, signal noise, costs (e.g., additional data gathering), QC failures (need constant QC monitoring and review at each step).

Group Two

- Timescale differences and how to integrate results from multiple models.
- Models need to be validated against real world. Don't model something that can't happen.
- At what level of abstraction should models be created? Within cell, at boundary of cell (cell as a black box).
- How should this multidimensional modeling data be presented to be interpreted (from data to information to understanding to knowledge)?
- Observability of data. A model that needs data that we cannot now observe will drive technology development; model cannot be validated until data are available.
- Technology changes and impacts on sensitivity, ranges, number of organisms sequenced; numbers of perturbation conditions will go up.
- Straw man production process for the facility:
  - – QC must be constructed so that standards are developed and applied to each part of the overall facility production process.
  - – Sample production.
  - – Bioinformatics.
  - – R&D (new instrumentation, processes).
  - – Tech transfer (move R&D to production line).
  - – Advisory steering committee.
- Viewspace is broad on what quality control is; automation implies a highly refined business decision process, but that does not always exist. Continuous QC monitoring needs to occur.

Critical Areas for Future Discussion

Workshop participants identified the following key critical areas that need to be discussed or addressed.

- Need to understand, describe relationship dynamic of all players including core operators, DOE, and users.
- Timely publication in peer-reviewed journals.
- Methods for analyzing data to drive things forward.
- Poster child experimental design is a new process, new way to think about needed experiments, demonstrate success.
- Open system.
- Manage expectation of users about how much data analysis will be done.
- Show reduction of per-sample costs by automation.
- Well-defined protocols for sample prep and core processing.
- Use a variety of instrumentation rather than a farm of the same instrument.
- Easy interfaces, consistent results.
- Papers that can be put together quickly, drawing from the database.
- Educate stakeholders about the importance of protocols. A production-monitoring system will be necessary to assist stakeholders in following existing production protocols and detect variations from those protocols.
- Literature processing such as automated searching, text-data mining.

- Facility has to work as samples come in, high-quality data goes out. After success is achieved, have strong collaborations with good scientists, use facility for well-designed experiments.
- External collaborators and facility people educate each other about what the facility can do and what it should do.
- Initial experiments require new capabilities and are doable.
- Work through details of several scenarios.
- Define appropriate projects for this particular facility to generate good scientific proposals.
- Define the scientific impact of the facility.
- Train users, facility staff.
- Market the facility.
- Communication between prospective users and facility personnel.
- Recognize that the facility will open in about 5 years and be open for at least 20 years.
- What will the facility deliver to its customers? What will justify its cost?
- Facility can drive emerging standards for semantics and ontology (e.g., XML).
- Standards of formats for exchange of information.
- Is the core responsible for creating information?
- Internet protocol issues must be resolved for open system.
- Balance between being a service facility (repeatable, reliable information) and a science facility (research needs). The major educational task will be to properly educate facility users in the difference between (1) production as the use of the existing production facility in a consistent and repeatable manner and (2) research and development of next-generation processes to be introduced into the production facility). Both areas should become a part of the proteomics facility.
- QC ensures that the production process reflects business rules selected by process designers.
- Facilities will need to get information from other instruments worldwide, have it published across facilities.
- Presentation of results back to P.I.'s in proper context; annotation.
- How people other than original P.I.'s can have access to data; security issue; how people can look up results (ontology).
- Issue of culture; support of manager, peers, other scientists.
- Consider stopping production to review the work. This is enabled by continuous QC monitoring of the production process. When results begin to exceed the bounds set to ensure repeatable, consistent data, then production should be stopped until the problem can be identified and corrected. From the business point of view, this makes sense because it minimizes the waste of resources.
- Capture and address real needs of user community.
- Where will data be analyzed? What computing resources will be needed?
- Satellite facilities — How QC will be enforced for data received from satellite facilities is a crucial question that must be properly addressed. The answer is driven by the process and business model adopted for the facility.
- Educate industry.
- Operational and administrative costs need to be identified and the ratio kept to the minimum.
- What technologies will be used?
- New model for publishing results available after the P.I. has published.
- Issues of experimental reproducibility. What is an experiment, and when is it done? Bookkeeping for the facility may hinge on this—10,000 experiments once or 1000 experiments 10 times?

Appendix A: Workshop Attendees

External Participants	Institution	E-Mail
Jay Abramovitz	Software Technology Group, Inc.	jay@softwaretechnology.com
Rajeev Aurora	Pfizer Inc.	Rajeev.aurora@pharmacia.com
Gyorgy Babnigg	Argonne National Laboratory	gbabnigg@anl.gov
Hamid Bolouri	Institute for Systems Biology	HBolouri@systemsbiology.org
Philip Bourne	UCSD/San Diego Supercomputer Center	bourne@sdsc.edu
Damon Coffman	Inovise Medical	damonc@softwaretechnology.com
John DePaula	Software Technology Group, Inc.	johnd@softwaretechnology.com
Forbes Dewey	Massachusetts Institute of Technology	cfdewey@mit.edu
Jeremy Edwards	University of Delaware	edwards@che.udel.edu
Robert (Steve) Erb	Gene Logic, Inc.	phdrserb@earthlink.net
Ziding Feng	Fred Hutchinson Cancer Center	zfeng@fhcrc.org
Nathan Goodman	Institute for Systems Biology	nateg@shore.net
Jim Gray	Microsoft	Gray@microsoft.com
Mike Gribskov	UCSD/San Diego Supercomputer Center	gribskov@sdsc.edu
Eugene Kolker	Biatech	ekolker@biatech.org
Michael Liebman	University of Pennsylvania	liebmanm@mail.med.upenn.edu
Krishna Mahadevan	Genomatica, Inc.	rmahadevan@genomatica.com
Natalia Maltsev	Argonne National Laboratory	maltsev@mcs.anl.gov
Gary Montry	Southwest Parallel Software	montry@spsoft.com
Ron Taylor	University of Colorado	ronald.taylor@uchsc.edu
Edward Uberbacher	Oak Ridge National Laboratory	ube@ornl.gov
Venky Venkatesh	Monsanto	tvvenk@monsanto.com
Robert Wildin	Software Technology Group, Inc.	bobw@softwaretechnology.com
John Wooley	University of California, San Diego	jwooley@ucsd.edu

PNNL Participants	E-Mail
Ken Auberry	Kenneth.Auberry@pnl.gov
Jim Bixler	Jim.Bixler@pnl.gov
Harvey Bolton	Harvey.Bolton@pnl.gov
Bill Cannon	William.Cannon@pnl.gov
Deb Gracio	Debbie.Gracio@pnl.gov
Gary Kiebel	grkiebel@pnl.gov
Michaela Mann	Michaela.Mann@pnl.gov
Blaine Metting	Blaine.Metting@pnl.gov
George Michaels	George.Michaels@pnl.gov

PNNL Participants	E-Mail
Matt Monroe	Matthew.Monroe@pnl.gov
Haluk Resat	Haluk.Resat@pnl.gov
Margie Romine	Margie.Romine@pnl.gov
Marla Seguin	Marla.Seguin@pnl.gov
Heidi Sofia	Heidi.Sofia@pnl.gov
Bobbi-Jo Webb	bobbie-jo.webb@pnl.gov
Steve Wiley	Steven.Wiley@pnl.gov

Appendix B: Workshop Agenda

Tuesday, May 13, 2003

Time	Topic	Speaker
8:00 a.m.	Welcome, Introductions, Charge	George Michaels
8:15	Exploring the Frontier Between Computing and Biology	John Wooley
9:00	Bioinformatics and Proteomics	Gyorgy Babnigg
9:50	Discussion	
10:00	Break	All
10:10	Quality Control for High-Throughput Processes	Steve Erb
11:00	Discussion	Nathan Goodman
11:10	Community Databases for Disease-Focused Research	
12:00 noon	Discussion	All
12:15 p.m.	Lunch	
	Breakout Sessions	
1:00– 2:45	Breakout Rooms Announced, Sessions Begin	Deborah Gracio, Michael Liebman
	Posters	Gary Montry
2:45	Break, Prep for Summaries	All
3:15	Breakout Session Summaries	All
5:00	Summaries Continue	All
8:00	Adjourn	