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

Santa Fe, New Mexico, April 1–2, 2003

GTL Facility for Whole Proteome Analysis

*Organizers: Jean Futrell, Pacific Northwest National Laboratory
George Church, Harvard University
*Facilitator: Karin Rodland, PNNL

This report summarizes the Workshop on the Facility for Whole Proteome Analysis presented by Pacific Northwest National Laboratory (PNNL) on April 1–2, 2003, in Santa Fe, New Mexico. The workshop purpose was to support the U.S. Department of Energy's (DOE) Genomes to Life (GTL) program by eliciting input on global proteomics needs, technologies, and facilities from the scientific community.

More than 30 biologists, microbiologists, technologists, and informatics specialists from industry, academia, and DOE laboratories attended (Appendix A). The agenda (Appendix B) included presentations of research scenarios (Appendix C). These scenarios illustrated how the capabilities of a global proteomics facility would allow researchers to answer questions or apply a systems biology approach to a challenge that has resisted solution by more conventional approaches. Presentations on "toolkits" or existing technologies also were given (Appendix D).

In small breakout sessions, attendees addressed questions relating to the capabilities and potential of a global proteomics facility.

Contents

Genomes to Life Facility Workshop Report 1
Opening Remarks and Workshop Objectives ...2
Genomes to Life Facility Plans 2
Application of Proteomics to Systems Biology .2
Scenario Presentations 3
Technology Toolkit Presentations 8
Breakout Sessions 15
Closing Presentation: GTL Biosystems—Integrating Measures and Models 19
Appendix A: Workshop Attendees 20
Appendix B: Workshop Agenda.....................23
Appendix C: Research Scenarios 24
Appendix D: Toolkits 25
Appendix E: Marvin Frazier's Presentation.....26
Appendix F: Application of Proteomics to Systems Biology...27

Opening Remarks and Workshop Objectives

Jean Futrell, Pacific Northwest National Laboratory

The workshop opened with an introduction and explanation of format by Jean Futrell, PNNL. He referred to a recent *Nature*[1] article by Ruedi Aebersold, Institute for Systems Biology, which stated the need to think of proteomes in terms of complexity and understanding—similar to understanding the Milky Way. This challenge for our time will take a lot of effort by many people.

[1] R. Aebersold. "Constellations in a Cellular Universe," *Nature* 222, 115–116 (2003).

*The organizers and facilitator 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/whole_proteome_facility_workshop040103.pdf

Unlike the universe, the problem of trying to understand what goes on in the cell—even though extremely complex—is bounded and finite. With the appropriate focused effort, we clearly will be able to do it. This complex process requires infrastructure and teams of specialized personnel to work in facilities. Aebersold also makes the point that like the Human Genome Project that preceded it, proteomics research must be done in the public domain. Results must be shared widely in an understandable format.

Workshop objectives were to establish a link between the science drivers and technological capabilities and to generate a more focused set of technologies. DOE white papers dealing with the GTL program have tried to provide an overview of these challenges at a 10,000-foot level. We need to bring them down to 7000 feet.

Futrell gave the following instructions to workshop participants:

- Think of a 5-year time frame—where we are now compared to where proteomics will be in 5 years.
- Understand the science drivers—the rationale for what is done.
- Discuss the toolmakers and technologies that will be shared—what's available now and what will be built in the future?

Genomes to Life Facility Plans

Marvin Frazier, DOE

Marvin Frazier, Office of Biological and Environmental (OBER) GTL program manager, gave a presentation about GTL's Facility for Whole Proteome Analysis (Appendix E). He summarized BER's plans for the GTL program as part of the R&D research program and infrastructure and facilities.

Frazier noted that DOE wants a global proteomics facility to be a bridge between small and big laboratories and that the best way to build that bridge is through good computational capabilities. DOE wants the entrepreneurial spirit of individuals to be available to the larger community.

He also noted that the facility design process will be circuitous and will be done in stages. The conceptual design will be examined thoroughly by scientists in workshops and by R&D. As with the creation of the Joint Genome Institute (JGI), how the facility starts out and what it actually becomes are very different. DOE expects big changes in 2 years because this is not a facility that will be put in place, have the lights turned on, and then run for 10 years in the same mode. It will be very dynamic for the first 5 years, if not longer.

Application of Proteomics to Systems Biology

Lee Hood, Institute for Systems Biology

Hood discussed his views of systems biology, with an emphasis on proteomics (see Appendix F).

Scenario Presentations

Three microbiologists gave presentations on different organisms of interest to the Genomes to Life program:

- Himadri Pakrasi, Washington University—*Synechocystis*
- Tim Donohue, University of Wisconsin-Madison—*Rhodobacter sphaeroides*
- Jim Fredrickson, Pacific Northwest National Laboratory—*Shewanella oneidensis* MR-1

Their presentations are included in Appendix C. The following key points were made:

- Reproducibility is needed. What is the minimum number of experiments to achieve this?
- Controlled cultivation, either controlled batch or continuous in fermenters, is crucial.
- Statistics are needed.
- Global proteomics is a snapshot. Life is kinetics and fluxes. Tie to metabolomics.
- Five years out: Single-cell proteomics? Microbial communities?

- Simulation and modeling will be used as an approach to connect data.
- Data quality is important.

Once the baseline proteome of an organism has been determined, biological insight can come from comparative approaches (i.e., comparative proteomics, functional proteomics).

Global Analysis of Cyanobacterial Proteomes: A User's Perspective

Himadri Pakrasi, Washington University

The National Science Foundation, DOE Office of Basic Energy Sciences, United States Department of Agriculture, and National Institutes of Health fund this work. In regard to GTL aims and DOE missions, the work relates most closely to carbon sequestration, about which much new information is emerging. The field is at an exciting stage. The take-home message is that the carbon fixation process is an interplay between photosynthetic redox reactions and carbon acquisition.

The following cyanobacteria, all of which have very high-quality genome sequences available, are being studied.

- *Synechocystis* 6803
- *Synechococcus* WH8102
- *Anabaena* 7120
- *Prochlorococcus*

Subcellular fractions. These are a critical issue for cyanobacteria. The bacterial cells have intracellular compartments important for the carbon sequestration process. In particular, the carboxysome is being studied in great detail. The peptidoglycan layer is another subcellular region of interest.

Synechocystis has a relatively complex cellular structure. An issue confronting scientists 5 years ago was the relationship among the outer membrane, plasma membrane, and peptidoglycan layer. Investigators developed a procedure to purify the thylakoid and plasma membranes using a two-phase partitioning system to obtain a relatively pure preparation of the plasma and outer membranes. They also separated the thylakoid membrane, but it still contained impurities. The problem was that the majority of thylakoid membranes migrate exactly like the majority of plasma membranes in a sucrose gradient, resulting in one-dimensional fractionation. This area is ripe for technology development.

Study results

- Two-phase partitioning followed by sucrose-gradient centrifugation yielded pure thylakoid and plasma membrane vesicles from *Synechocystis* 6803.
- Photosystem (PS) I and PS II pigment protein complexes function in thylakoid membranes.
- Several proteins of PS I and PS II are found in the plasma membrane.
- The core centers of PS I and PS II are integrated and assembled in the plasma membrane.

These discoveries lead to the following questions: How are the PS components transported to the thylakoid membranes? Via thylakoid-plasma membrane attachment sites? Via membrane vesicle migration between membranes? The two classes of membranes come close but never appear to touch. Through electron microscopy, small vesicles are in evidence. Again, this is an area that needs technology development and imaging.

Discussion. Pakrasi's laboratory found that comparing data from different laboratories is very difficult. If all data are controlled and created in a centralized manner, experiments are designed accordingly.

Rhodobacter sphaeroides Proteomics Perspective

Timothy Donohue, University of Wisconsin-Madison

This work is part of the GTL consortium, "Molecular Basis for Metabolic and Energetic Diversity," which is focusing on generation and production of the reducing power of bioenergetic pathways in *Rhodobacter*. As of April 1, 2003, Donohue's group had not done a proteomics experiment. Cells had been sent to Dick Smith at PNNL for accurate mass tag analysis.

Rhodobacter sphaeroides is an alpha-protobacterium. Strain 2.4.1 has been sequenced, assembled, and annotated by JGI, Oak Ridge National Laboratory, and members of the community. It has a 4.5-Mb genome, 2 chromosomes, 5 plasmids, and ~4500 ORFs. *R. sphaeroides* is an energetically versatile organism. It is photosynthetic, makes hydrogen, removes organic toxins, and can synthesize biodegradable plastics.

Donohue illustrated proteomics needs by comparing photosynthetic and aerobic respiratory cycles. The genome sequence revealed many new insights into *Rhodobacter* biology. The genome sequence predicts many different electron carriers and at least five different oxidases. Most of the proteins are membrane bound, so it's a challenge to analyze them. Many cellular components are of variable abundance and need high sensitivity and dynamic range. The heme group in *c*-type cytochromes has a covalently attached polypeptide, so MS methods must be able to account for this common protein modification.

During the time the protein complexes are being assembled, investigators want to be able to assay the time-dependent appearance of proteins in spectral complexes. They want to dissect regulatory basis for differential kinetics of photosynthesis gene expression. They currently do not know what other reactions are going on when photosynthesis is shut down.

Investigators are making the PS membrane from a few sites in the cell. They want to determine the factors that are responsible for assembling the vesicles.

Another point in discussing where we want "omics" technology to be in 5 years is that not all RNAs are mRNA, tRNA, and rRNA. Small RNAs are key regulators of metabolic and genetic networks. We need to be able to analyze these as well as other macromolecules in high-throughput facilities. The facility really needs to identify and characterize carbohydrates as well as proteins and metabolites.

Discussion. One question to be addressed is, What happens when we go from photosynthetic back to aerobic conditions? Chlorophyll does not turn over, yet in four generations those membranes cannot be found. Do they undergo differentiation? The answer requires examining proteins, which no one has done. One theory is that chlorophyll differentiates into oxidative membrane, but there's no data to support that.

When the organism is growing, the energy requirements for maintenance may be very high. Most energy is not going into growth but into maintenance. Biosynthesis is very slow and has to be minimized to maintain complex, diverse pathways. We find that minimizing the sample's complexity gives a better chance of understanding what goes on. A cell in slow growth has less complex systems, and we may have a better chance of interpreting results from high-throughput measurements.

Sam Kaplan, University of Texas Medical School reported that they have just developed data that fly in the face of *Escherichia coli* researchers. These data show varying growth rates over broad ranges using transcriptome data. Messenger RNA levels should change with growth rate but instead remain constant. The level of mRNA does not vary according to growth rate. Protein analysis would give a sense of productive turnover.

The question was asked, if there were a technology that gave perfect quantitation of everything, what would we do with it? The response was, Get metabolic and regulatory maps. The community already has RNA, pools, mutants, and biochemistry to do more functional genetics. If the flow of reducing equivalents is changed, how does that change expression and other parameters?

Response: Manipulate to make more hydrogen and increase the efficiency of the photosynthetic apparatus. Use the stamp-collecting snapshot data to plan the next round of experiments. The global proteomics facility will provide guidance for the next round of experiments.

Investigators now know about a lot of post-transcriptional activity. The mRNAs are produced in overwhelming abundance relative to complexes, and proteins are produced more abundantly, too. Chlorophyll is a critical factor.

Shewanella oneidensis MR-1

James Fredrickson, Pacific Northwest National Laboratory

PNNL is studying *S. oneidensis* for DOE as part of the *Shewanella* Federation (SF). They are interested in *Shewanella* because of its effectiveness in reducing metals. A tomographic image of *Shewanella* incubated with uranium showed crystals of reduced uranium in the cell periplasm and on the outside of the outer membrane. Electron transport systems can be coupled to the reduction of metals.

In short, *S. oneidensis*

- Effectively reduces metals and radionuclides.
- Readily forms aggregates, flocs, and biofilms and likes to attach to surfaces.
- Is a facultatively aerobic Gram-negative gamma-proteobacterium.
- Has been sequenced (MR-1 genome, ~5 Mb).
- Has developed genetic systems.
- Is a respiratory versatile organism of eight decaheme c-type cytochromes, with three outer membrane (OM) lipoproteins.
- Is widely distributed in the environment (soil, sediment, water column, clinical).
- Is a gradient organism, adaptive to changing environment.
  - Some 88 predicted 2-component regulatory proteins.
  - Some pathogenic strains (e.g., to fish).

Phased Microbial Genomics. In the near term, SF is trying to link gene sequence to proteomics data, make metabolic connections, link physiology to genomic information, uncover gene function, and explore metabolic and regulatory networks. The mid-term will focus on ecofunctional genomics such as environmental sensing and response; cell-cell interactions, consortia, and assemblages; and cell function in an environmental context. In the long term, SF will do community genomics such as structure and function, intracellular metabolic and signaling networks, and linking to predictable community ecology.

Shewanella does not live alone. It uses fermentation products for energy and interacts with other microorganisms. Other organisms use products from *Shewanella*.

The federation is using genome sequence, informatics, controlled cultivation, linked measurements, information synthesis and interpretation, imaging, metabolites, proteomics, and gene expression to investigate global response and regulation in *Shewanella*. Controlled cultivation generates sample, but it is an invaluable research tool as well. Currents gaps are in metabolite analysis, quantitative proteomics, and modeling.

SF is considering a phased approach to characterize the community in which *Shewanella* lives. Diversa and others are developing high- throughput cultivation technologies. What if they sequence lots of genomes, put them back together two at a time, build up the numbers, do linked measurements, look at who is expressing what, and measure signaling molecules? The federation is doing this first on pure cultures of MR-1. This type of information would be coupled into community models where we can look at cellular and intracellular regulatory networks. We need to gradually increase the level of complexity to understand interactions.

The proteomics facility wish list for *Shewanella* includes

- Proteomics: Consortia, monocultures, fractions, complexes (including protein DNAs)
  - Comprehensive, quantitative
  - Extent and type of modifications
  - Rapid turnaround, user-friendly data interface
  - Single-cell measurements
  - Cellular location
- Metabolite and small-molecule analyses
  - Comprehensive and quantitative
  - Intracellular and extracellular concentrations
  - Capacity for rapid sample stabilization
  - Isotope labeling and pathway analyses
- Gene expression
  - Global quantitative expression (as opposed to relative levels)
  - Single-cell measurement
- Cultivation
  - High-throughput, difficult-to-culture organisms

  - Culture maintenance and preservation
  - Continuous or semicontinuous monitoring of soluble and gaseous metabolites
  - Controlled experimental systems (planktonic, biofilm, multispecies)
- Computation
  - Data storage, retrieval, integration
  - Data-analysis tools (especially proteomics)
  - Metabolic and regulatory network models
  - Cell-community models and simulations

Discussion. Attendees agreed that quorum sensing is important at all levels, particularly in cell signaling and communication, but even in bioreactors and cell cultures.

Mike Knotek, Consultant: The *Shewanella* group obviously is the most developed. What sort of informatics environment is used? Fredrickson responded that this a real gap in SF. They were formed differently from the rest of GTL as part of the Microbial Cell Project, with no infrastructure for data sharing to facilitate collaboration. They used the collaboratory environment and are trying to adapt what they're doing into that environment. Eugene Kolker is working on data integration. This is a key point for other GTL projects, and the SF group has been working for other GTL projects and adapting their systems.

Darrell Chandler, ANL: To what extent are disparate technologies applied in the *Shewanella* Federation and other groups contributing to the data-integration problem, and how this could be simplified?

Fredrickson: We are open to ideas. It would help to have integrated data-generation platforms. These things need to be developed hand in hand, and there is not a lot of cross feeding. If technology platforms can be simplified and unified, it may help the informatics.

Kaplan: If researchers wanted to ask specific questions of computer databases (e.g., whether they could predict how *Rhodobacter* would work under low light conditions), they could go and do the experiment. These systems must be available to nonexperts as well, so they can ask questions and be able to move seamlessly back and forth among databases. This comparative-biology approach would be very useful for cross- and integrative understanding.

Donohue: This is a critical issue. Even among GTL people, the issues are the same for *Rhodobacter* as *Shewanella* but there are no links in databases for investigators. Creating a platform for people with different organisms and in different fields can enable researchers to know immediately what is available. A large scientific community outside of DOE should know about that.

Carol Giometti, Argonne National Laboratory: ANL is generating thousands of 2D gel patterns and uses an Oracle-relational database platform that is finite but is a start. They want to get protein-expression data rapidly to the scientific community. They currently have a password-protected site for collaborators to look at and download data and a public site for published data. They need input from the research community on what kind of scientific questions to ask so the query structure of the database can be developed further.

Kaplan: This kind of information should be made available to undergraduates and high school students so they can click on the databases and think about how biology works in moving toward the browsing stage. People ask why yeast resources are not available for other organisms. Databases need to be standardized so that anyone coming in from outside the discipline can get to the important information they need. Aebersold talked about this in his *Nature* article.

Yuri Gorby, PNNL: Two obstacles are that

- High-throughput–generating technologies and large data files often have proprietary software and gated distribution. Commitment is needed with companies. Identify a company that can standardize these data sets, or a lot of time must be spent in transforming data sets to browsable platforms.
- These metabolic flux analyses and models are the type of computation links from observational to predictive science. They have to be developed and thoroughly understood. Quality must be high because it's easy to get lost.

George Church, Massachusetts Institute of Technology: To get to a browsing stage requires an investment and trust in companies and databases that may be difficult to achieve. Putting flat files out on the Web is an option; they are intuitive, and no Oracle query is needed. A high school student can read them, and an undergraduate student can line up two organisms. The major computational resource we need for now is tons of disk drives.

Charles Auffray, Genexpress: Regarding data quality and precision, one example given is the curve for sequencing throughput and cost. The transition phase around 1998 occurred after almost 20 years of technology development because of Phred-Phrap tools. We need to think of community quality standards for proteomics and imaging. We must be able to measure quality and precision, but currently we are not at the right stage for precision studies. What platforms are ready to develop such quality standards? Flat files are a good option.

Harvey Bolton, PNNL: After hearing about the three systems discussed today, single-cell analysis and isolation of cell fraction seem to be key. But how key are they in the 5-year outlook? Some are doing fractionation on chemostat cultures.

Donohue: Part of what we've done successfully is fractionation. In an experiment to make membranes de novo, there are only five or six machinations per cell. We don't really know how to isolate them, and we need to image them on a single-cell basis.

Fredrickson: All cell fractionation techniques are imperfect.

Knotek: There is an egalitarian beauty in having the data available. But if the information environment cannot be taken to a wildly higher level such as having huge computational resources and taking a sophisticated approach to data management and the long view, we won't make progress in systems biology.

George Michaels, PNNL: In biology, GenBank is the paradigm. We need a data depository and tools for proteomics. Flat files contain sequencing information, and tools are developed to analyze the information. We need a repository and analysis tools. This is a good opportunity area for DOE. We can't look at petabytes of data—it would take 30 years to look at each technology—so we need metadata. No one tries to do this by hand, and we need to give up the idea that they can. The parallel example is weather modeling, when one could get all the flat files from these people, but we don't want to do it.

Church: As data files get larger, they're not necessarily more complex. So something with a lot of modalities, even with less than a petabyte, would be complex but would not require sophisticated databases. More modalities might require them. When databases are scaled, get to the application. We should not stigmatize flat files, but we do need to think about applications and we won't be browsing through petabytes.

Kaplan: Cell fractionation is a crude thing, and this implies growth and reproducibility. We will not be able to grow everything in every way. Much greater use needs to be made of chemostats. For example, if cells are being grown at 1% dissolved oxygen, local oxygen concentrations will not be at the desired 1% level as cells increase in number. Chemostats are the only answer to that question. Single-cell analysis would be lovely, but unless it's available today another approach must be taken—synchronous cell populations, where the majority of cells are single. Think about growth and how we are doing that.

Donohue: Single-cell technology is primed to help us in cell cycle. It is still an average population. Wouldn't you really like to know what's going on in the cell? This is a clear “go” point.

Auffray: One way to organize things is with layers of information. We need technology core integration and semantic integration. There are many estimates of the number of genes because of the lack of quality standards and the definition of a gene. So then what are the right experiments, and what are the right questions? It's not only data collection and standards but also semantics and vocabularies. The power of these platforms will make them more usable by a broader, more diverse audience.

Eugene Kolker, BIATECH: In regard to what to do with different types of data, *E. coli* has the largest number of databases but little is available to the public. They've exchanged data as Excel flat files, which is not a solution, but it is put on the Web and now they are trying to have comparisons enabled across platforms. The problems are with comparing apples and oranges—cDNA array vs oligonucleotide data—two types of expression data sets that cannot be compared. Even though array analysis of gene expression has been around for several years now, we don't even have standards in this area. In many more areas, we need to establish standards and establish experiment validity in proteomics and systems biology, and this is clearly a daunting challenge.

Technology Toolkit Presentations

Four proteomics technologists presented toolkits of technologies being used in the field and in their laboratories:

- Marvin Vestal, Applied Biosystems Inc.—Proteomic Technologies
- Carol Giometti, Argonne National Laboratory—2D Gels for Proteomics
- Darrell Chandler, Argonne National Laboratory—Microarrays
- Richard Smith, Pacific Northwest National Laboratory—Global Proteomics

Their presentations are included in Appendix D, although summaries of their talks and ensuing discussions are provided here.

Proteomic Technologies

Marvin Vestal, Applied Biosystems Inc.

Vestal reviewed past, current, and future technologies for proteomics.

Components of Proteomic Analyzers

- Sample prep (e.g., separation, concentration)
- 1D and 2D gel interface with MS
- Liquid chromatography (LC) interface to MS
- Chemistry for proteomics with MS
- Sample plates and matrix-assisted laser desorption ionization (MALDI) matrices
- MS and MS/MS
- Applications software
- LIMS and results management
- Bioinformatics

In 1990, MALDI (1%) and electrospray ionization (ESI) (99%) were available and used. Today, hybrid systems are used with both ESI and MALDI interfaces. QQQ is still ESI (50%) and time of flight (TOF) (50%) for proteomics, but that's not a firm ratio.

By 2006, Vestal sees a nearly complete melding of TOF with traps and combinations, as stated in the article by Aebersold and Mann.

Advantages of LC coupled to ESI and MALDI for proteomics:

LC ESI

- Direct coupling of LC to MS
- Fast, lots of MS and MS/MS
- Accepted MS/MS ionization mode.

LC MALDI

- Sample in solid state
- Not time limited for MS/MS
- Analysis can be faster or slower than separation
- More sophisticated workflows
- Fast, lots of MS and MS/MS
- Results-dependent acquisition stop criteria

Need to apply established principles of analytical chemistry to assessing proteomics data quality, including

- Replicate measurements
- Objective statistical evaluation of spectral quality
- Improved scoring algorithms that provide reliable statistical estimation of the probability that a reported hit is correct
- Validation of methods using complex mixtures of known samples covering a broad range of concentrations

Until all of these have been done, we have to take the data with a grain of salt. Sensitivity, speed, and data quality all must be high for routine high-throughput proteomics, and if we get speed by sacrificing data quality, we're going in the wrong direction.

Sensitivity is expressed by

- Detectable concentration (moles/L)
- Sample consumed (moles or grams)
- Sample loaded

- Determinations per second
- Copies per cell (of prime interest here)

Copies per cell is a hard number to get, and we don't know how to do it at this time. It's a good goal to work toward.

Factors determining sensitivity

- Chemical noise
- MS efficiency
- Sampling efficiency
- Dynamic range
- Molecules consumed per pulse
- Pulse rate
- Ions required per measurement
- Measurement time
- Ions required per measurement: ~10 ions at peak minimum
- Total number depends on number of peaks and dynamic range required (100 to 1 million)
- Molecules consumed per shot depend on laser and matrix

MS efficiency

- Ions detected per sample molecule consumed (detection ~0.5, transmission 10^{-4} – 1, ionization 10^{-4} – 1)
- Relative ionization efficiency (sample background)

A strength of ESI and MALDI is that solvent and many common impurities are not ionized. The major difference between instruments is in transmission efficiency.

Sensitivity can be improved by

- Reducing chemical noise
- Better separation and fractionation (fewer peptides per sample)
- Improving ionization efficiency
- Increasing sample utilization: More shots, smaller sample volume, more sample per shot at constant ionization efficiency
- Simplifying spectra
- Increasing resolution of precursor selection
- Improving analyzer transmission efficiency

TOF is becoming increasingly important. Increasing laser rate improves results in many ways (see presentation).

Applications for using MS only

- Precise MW of intact proteins
- MW profiles of pathogens
- MW of noncovalent complexes
- Tissue imaging
- Biomarkers

MALDI TOF-TOF and MALDI Q-TOF operating at 10 kHz are the only practical analyzers for meeting these requirements and specifications. They should be commercially available in 2 years.

The proteomics analyzer of the future will interface with high-throughput separations and will be rugged and fully automated.

Two-Dimensional Electrophoresis

Carol Giometti, Argonne National Laboratory

The gold standard of proteomics is 2DE. At ANL, 2DE methods have been used for two decades for high-volume, high-throughput analyses of complex protein mixtures of interest to DOE, starting with high-volume mouse samples. Wasinger et al. first used the term "proteome" in a 1996 *Electrophoresis* article discussing 2DE methods and results.

The technology can provide a lot of data such as relative abundance (with or without metabolic labeling), pI and MW, post-translational modifications, and identifications. At ANL, flat files are provided, and everything is put into an Oracle database. ANL investigators are integrating with protein databases and have the completed genome sequences on their machines so they can go back and forth. They are downloading Kegg metabolic databases and also want to be able to cross-compare microarray profiles.

Bottlenecks in 2DE include tedious methodologies such as protein separation, detection, and identification; dynamic range limitations; and the inability to determine function.

Commercially available immobilized pH gradient strips and prepoured slab gels improve analysis reproducibility and ease of gel handling. People who are new to the field need this, and the commercial products are expected to get even better.

Automated Protein Separation, Innovations, and Identification. Automated protein separation includes production and use of standardized separation matrices and automation of all sample loading, gel handling, and protein detection protocols. Innovations in protein detection include such multiple detection methods as phosphoproteins, glycoprotein, and total protein (with automatic image capture) on a single 2DE image. Accelerated protein identification is in the conceptual stage with digestion of entire 2DE pattern with specific protease, impregnation with matrix, and MALDI-TOF.

Theoretical 2DE maps of proteins can be computed based on genome sequences, but often the theoretical position of a protein doesn't match the observed. Such theoretical maps, however, could be improved with input of knowledge about post-translational modifications, for example. It's a matter of learning the rules. As more data are collected from 2DE experiments and compared with theoretical patterns, predictions can be improved. Eventually, protein identifications will be done computationally rather than through protein excision from gels and subsequent identifications based on tryptic peptides.

Sample Fractionation. Sample fractionation for improved dynamic range has been done using differential centrifugation, affinity purification, chromatographic enrichment, and sequential extraction (membrane proteins). Automated protocols to minimize effort and increase reproducibility (applicable to all proteome analytical approaches) are needed for high-throughput use of similar protocols.

Characterization of Function. Giometti has been developing a method to separate proteins that are still intact to keep multimeric components intact. Separation by 2DE under nondenaturing conditions provides retention of function by identifying specific enzymatic activity and characterizing components of protein complexes and protein-ligand associations. This is an approach to the description of function for "hypotheticals." She would like to think of the nondenaturing 2D gels as protein chips produced by the microbe itself.

Detection and Characterization of Metalloproteins (X-Ray Fluorescence). Currently there are no methods for global screening to detect all metalloproteins. Ken Kemner at ANL is using the Advanced Photon Source for X-ray fluorescence (XRF) to look at metals outside of cells. Now, in collaboration with Giometti, Kemner is using XRF to detect metalloproteins expressed by cells in one of ANL's LDRD projects. Proteins are separated by electrophoresis and then put into the X-ray beam for detection of specific metals such as Fe, Mn, and Cu, and maybe more.

In a global proteomics facility's future vision, 2DE can play a part through the following:

- Automated sample preparation
- Automated protein separation and detection
- Automated protein identification
- Streamlined image acquisition and data assimilation and integration
- State-of-the art data interrogation and management tools

Discussion. X-ray fluorescence (XRF) enables researchers to see metals associated with proteins. At ANL, XRFs have been done of known metalloproteins in 2DE gel spots cut from both silver and Coomassie blue-stained gels, and the iron has been detected.

In response to a question about sample isolation and storage, Giometti noted that, once proteins are denatured, they can be stored at –80°. Nondenatured proteins would have to be analyzed as quickly as possible. The other aspect of nondenaturing technology is that it picks up on where researchers have started to go with biology—ligands, assuming the interactions are stable enough. If compatible with separation matrices, sensitive spectroscopic techniques could be used to detect ligands by using ligand-specific stains. Different matrices should be tested to obtain larger pore sizes for resolution of large protein complexes. Five years out, someone in the market will develop this.

For traditional 2DE to be done for quantitative analysis, Giometti requests samples in triplicate in a volume sufficient for running four to five 2DE gels.

Microarrays in a Proteomics Facility

Darrell Chandler, Argonne National Laboratory

Philosophy of microarray technology

- Start with the end in mind
- Identity does not equal characterization
- Complex does not equal machine
- Cell is not a community
- Culture is not a natural environment

When investing in or developing technology, how far forward should one look? What is the end state? How far out we look does impact the technology-development path.

A smorgasbord of nucleic acid arrays includes the following:

- Planar arrays—glass substrates, SAMs, coatings
- Flow-through chips
- Coded beads
- Electronic chips
- Gels

An array technology is more than just the substrate. The recognition element and the signal or measurement are included, so the array of technologies becomes complex. Need to ask the biology question, What do we want to do with this technology?

Fabrication methods

- In situ synthesis
- Quill-style pins
- Pin and ring
- Ink-jet piezoelectric
- Positive displacement and capillaries

Measurement Scale: Is the investigator interested in single cells, subcellular components, or communities of different types of cells? Chandler comes at this from an analytical chemistry viewpoint. Variation in the experiment results in image variations. The issue of standards and control needs to be addressed up-front.

Measurement noise defines replication requirements (nine-mer probes, planar array on a glass substrate). They did 24 replicates before making sense of the data. They felt that 60 to 70 replicates were needed just to capture the noise in biology. The greatest source of variability was in printing the array.

QA and QC in production mode

- Garbage in, garbage out: Image analysis and statistics can’t solve everything.
- How do we ensure substrate, probe, and chip quality?
- How does the choice of technology platforms affect the QA-QC pipeline?
- Does each QA-QC system support DOE’s long-term goal of predictive biology?
- Who will be responsible for QA and QC?

The past year Chandler has worked with military customers who want QA, and the science being discussed here is no different.

Computation is part of QA and QC. All those tools and techniques talked about on the back end must be on the front end.

Protein chips and beyond: This has all the challenges of DNA arrays and more.

- Peptides
- Aptamers
- Carbohydrates and lipids
- Antibodies
- Functional proteins and enzymes: Soluble, membrane
- Function under such extreme conditions as anaerobic, thermophiles, halophiles

We may have to fabricate protein chips in a glove box, an additional challenge.

How prepared is the existing technology? We have to consider the following because we think everything is out there ready to interact. But it’s really all a big pile of “stuff.” We don’t understand how all these interact with a piece of glass. And if we can’t understand this, how can we extrapolate into biology? We must consider

- Post-translational modifications
- Attachment chemistries and active sites
- Surfaces and steric effects
- Stability: Content, substrate

- Sensitivity

The ideal is to have antibodies stuck down nicely on glass plates with reactive sites up; that's not the case, however, unless we do more basic research in chemical interactions to really learn how best to develop these technologies.

ANL's trajectory is to leave the surface behind and get back to an environment in which molecules are functioning normally and go from antibody, protein, and enzyme arrays to a synthetic cell. The current GTL call emphasizes tags. Can fluorescent tags be generated for everything? How do optical tags respond to interesting environments? What other signal-transduction methods could or should be incorporated into a microarray format? How does one detect, identify, and characterize?

Visualizing global protein function

- Can fluorescent tags be generated for everything?
- How do optical tags respond to interesting environments?
- What other signal-transduction methods could or should be incorporated into a microarray format?
- How does one detect, identify, and characterize the unknown? What is the end state?

The cost is in the content:

- Probe and protein synthesis and preparation
- Volumetrics and liquid handling and quantification equipment
- Performing the experiment
- Analyzing the experiment

The use of satellite vs central facilities brings up the following questions:

- If a facility produces content, should it also produce the assay?
- Is it necessary or advisable to select one or a few array technologies?
- Are chips an integral part of evaluating content irrespective of the user's scientific goals and experiments?

A production line for custom chips would help Joe Researcher, but companies won't invest in low-volume products, and cost currently keeps many out of arrays.

- Is the customer part of the chip-production process?
- How much use and training is in a user facility?
- Should DNA, protein, and other types of arrays accompany every sequenced genome?
- What are the standards of production and performance?

Summary and perspective

- Predictive biology and natural environments are stated GTL end states.
- Arrays have a place in facilities and GTL science.
- Prediction places a premium on the mundane: QA and QC.
- Environment implies what is unknown.
- Arrays in or for a facility are not necessarily congruent with arrays for scientific inquiry and biology.
- What do we want from a facility?

Discussion

Knotek: DOE is thinking of making production wholesale rather than retail. If people want these things in quantity, they need to use private vendors so they can mass-produce for broader use. This is a better way to separate government and private companies.

Donohue: Five years from now, technologies will avoid the big up-front costs. Companies are positioning themselves to make designer chips. The break-even point occurs when the analysis has been done and it makes sense to build chips in the lab. Donohue can synthesize the chips for DNA arrays cheaper than the commercial vendors.

In situ synthesis of DNA arrays is an issue. How can we do it for peptide, protein, and carbohydrate chips? It's very costly.

Knotek: This may end up being the difference between using Wal-Mart vs a mom-and-pop shop. We may need to certify vendors to use protocols in ways people can trust.

Donohue: The Cystic Fibrosis Association has driven the price of chips down; this could be a model to explore.

Kaplan: How important is the information? Expensive is cheap, depending on how much imperative there is (e.g., bioterrorism). Cost can be irrelevant, and, in any case, demand can bring costs down.

Michaels: What scientific questions are key to national imperatives? What are the main scientific drivers? Saying you want to understand a cell isn't enough.

Marv Stodolsky, DOE/BER: The Human Proteome Organization (HUPO) is setting up a competition like CASP for microarrays (serum). These platforms should be talking to each other. It may be up to us to set competitive standards for both government and commercial facilities and make the results publicly known.

Analyzing Complex Biological Systems: The Roles of Separations and Mass Spectrometry

Richard Smith, Pacific Northwest National Laboratory

Predictions and assumptions: Proteome analyses in the next decade will be based largely on combined separations and MS, and peptide-level analyses will continue to dominate but will be augmented by intact protein-level analyses for reasons of sensitivity.

Given the constraint of a sequenced genome, the combination of high-accuracy mass measurements and separation times (e.g., LC elution) provides unique marker peptides for essentially all proteins.

The two stages are (1) initial generation of accurate mass and time (AMT) tags by "shotgun LC-MS/MS" measurements with conventional instrumentation and validation by LC-FTICR, and (2) application of AMT tags in repeated measurements with the same organism. This avoids the routine need for identifying peptides by MS/MS and is the basis for better quantitation, higher throughput, and proteome coverage. Some of these processes are becoming more and more effective and are evolving rapidly. Tandem MS will not be sufficient, so we will dig down deeper and deeper.

PNNL has done a capillary LC-FTICR 2D display of *Deinococcus radiodurans* and has identified peptides and ORFs. Once this is done, spots can be annotated rapidly. This is a truly global comprehensive coverage of proteins based on peptide tags. Some 2582 (83%) of predicted proteins have been identified and validated. Once AMTs or subsets are available, repeated measurements of a protein can be made.

Automation improves throughput and data quality. Analyses can be replicated and variation seen. When a step in the overall analysis is automated, the data get better. Some variations are still found for unknown causes. Internal calibrations are used for both the mass spectra, and a more complicated statistical procedure, a genetic algorithm, is used for separation.

An ordered list of 1667 *Shewanella* proteins was observed under aerobic conditions. Order was shown by decreasing rate of relative abundance. These analyses now move into the nanoflow mode by using long pack capillaries, which are close to 100% efficient. Below 100 ng of total proteome sample, the electrospray response becomes proportional to sample quantity. In a linear response, the matrix and ionization effects are eliminated. Anyone making proteome measurements needs to migrate to this regime.

Marvin Vestal made the point that numbers can be fudged in many ways. In his lab, they used a 10-µL solution with 5 ng of trypic digest of n14- and n15-labeled *Deinococcus* and also spiked it with albumin that was many times less than other sample components, and it worked fine.

When the sensitivity of a measurement is increased, new sources of noise become apparent, so improved procedures and cleaner solids are needed. Intact protein measurement augments peptide-level analyses, which generally are much less complex and yields more information on protein-modification states. This same approach can be taken to the whole-proteome level. If it is done under nondenaturing conditions, proteome-wide information could be obtained on interacting partners.

Nanoflow LC separations with ESI MS can

- Increase overall specificity and sensitivity
- Decrease or eliminate matrix and ionization suppression
- Provide linear response, better quantitation

Dynamic Range Enhancement Applied to MS (DREAMS) FTICR. This technique expands the dynamic range of measurements and allows use of the full dynamic range of FTICR after removing the most abundant species during a separation. PNNL has analyzed a mixture of ^{14}N and ^{15}N-labeled *D. radiodurans* cells. The combined proteome coverage was 3264 AMT tags (40% of the predicted proteomes in a single analysis).

Technological limits: The ultimate in MS analysis

- Micro- and nanofluidic single-cell manipulations and separations
- 100% efficiency nano-ESI and use
- Multiplexed individual ion analysis

Candidate facility technologies: Peptide-level proteomics

- Automated capillary LC-FTICR
- Capillary LC with various other MS/MS instrumentation for peptide identification (AMT tag development)
- Intact protein-level proteomics: CIEF and capillary LC-FTICR and TOF

Ancillary capabilities and instrumentation

- Stable-isotope labeling
- Protein and peptide fractionation
- Subcellular fractionation

Informatics supporting

- Protein ID and quantitation
- QA and QC

Discussion. Sample preparation is crucial. It must be completely automated because variations impact data analysis. If investigators don't ask the right questions, they won't get the right answers. One size doesn't necessarily fit all.

Some high-throughput work is being done with microfluidics system within the GTL Goal 1 work. Affinity purification is an issue. In Goal 1, Smith and Rodland at PNNL are splitting a postdoctoral position. A big push is to start with cell lysate automation in 6 months to a year from now (e.g., robotics, microfluidics, automated sample capture, and washing).

Chandler: Regarding sample purity, many biological samples of interest are attached to dirt. This goes beyond just getting sample into the detector, and I don't know to what extent we have to think about that. We need the ability to control the culture under a wide range of environmental conditions.

Smith: Protein complexes and quantitation are very important issues. Complexes are almost never clean; they always have fellow travelers from the proteome. Backing out the data is a computational exercise.

Vestal: This becomes the researcher's responsibility.

Kaplan: Here we're describing the gold standard. But my microbiologist colleagues have their own cottage industries. I know they don't do it as I do. If you look at *E. coli*, most of the work is done under anaerobic conditions and is nonreproducible from one lab to the next. So what's the truth?

Smith: A facility also plays another role as a core of expertise where researchers can learn and teach each other. Currently, there are no answers to the question about what needs to be built into a facility to ensure accuracy in preparation, but the same basic tools are useful. The differences are more in the front end.

Breakout Sessions

To enable more in-depth discussion of a global proteomics facility, workshop participants were assigned to one of three breakout groups. The moderated groups included a mix of biologists, technologists, and informaticists who discussed the following questions.

1. What is the science driver for a facility?

For any of the proposed GTL facilities, we need a large, practicing systems biology community that will use the facilities properly. Much of the scientific community doesn't understand what global proteomics is. To justify a multibillion-dollar facility, this community must be fostered by DOE and its laboratories and provided with a vision.

Scientists want to predict how communities of microbes will adapt to changes in environment. If this facility can help them learn how to do this, it will make a major impact on the science.

The bulk of biological science will continue to be done in individual laboratories. If researchers can do it at home they should, and if they can buy it for home, they should buy it. This facility, however, would provide specialized capabilities unavailable at individual facilities or laboratories and would be coupled with available expertise. If a global proteomics capability is made available, researchers will come up with innovative uses for it, but it's hard to sell a facility on that basis.

If a program is having impact on DOE missions, it has priority in the facility. And in turn, if a program can do global proteomics on a problem, the customer base will expand quickly.

2. What would a facility provide that could not be done at home?

Capabilities envisioned at a facility are

- Generic biology studies (not limited to microbes).
- Specialized needs such as culture conditions, sample-preparation procedures, and metabolic labeling.
- Integrated experiments using different technologies and methods to look simultaneously at the transcriptome, proteome, and metabolome. Multiple methods will reduce errors.
- Identification and quantitation of proteins and in what stoichiometries and metabolites.
- Kinetics, fluxes, not just snapshots.
- High-end, very large mass spectrometry (and multidimensional protein and peptide separations).
- Chemostats for some (not all) organisms. Synchronous time-series simple communities (not all).
- Separate R&D component; new technology import.
- Stable isotope analyses.

Conversely, the question of what wouldn't be done in a facility was discussed. The example of the genome centers was given: They became more focused as time went by, and sequencing assembly no longer has to be done there.

One observation is that a global proteomics facility would be a magnet to draw in expertise—both permanent and short-term (i.e., collaborative).

Another possibility is to think of the facility as a development entity that can transfer a capability to other sites and then transfer it back—more of an engineering paradigm. For example, much is being done in the Netherlands for continuous cultivation in metabolomics and measuring offgases. Not everyone would want this kind of capability in their own lab, but they conceivably could come use it at the facility. Those kinds of measurements could be very helpful. Investigators can come in, do controlled experiments, and then go back to their labs.

Some users will have a very sophisticated grasp of their goals, and others will not. Whole projects have been stymied at JGI because a collaborator couldn't provide decent DNA.

Organisms. Opinions on this topic varied. Some felt the facility should serve consortia of scientists for specific organisms and should be neither organism specific nor organism limited. Others suggested considering transformable organisms. Some but not all GTL organisms are amenable to transformation. For most organisms of interest, people develop systems sooner or later. This may not be an issue but may at least impact priority. Specific comments included the following:

- A facility would be ideal for doing in-depth, detailed analyses for an organism of choice—perhaps even organism design.
- We don't want a pilot plant but rather a small facility that can be used to show how to do controlled measurements. Then individuals can grow their particular organisms and have access to the facility by transferring samples.
- Will there be a choice between many organisms at some high level vs a few organisms in

gory detail? A small debate ensued about favorite organisms vs many organisms.

- How many microbes should be done per year—hundreds or tens of thousands? Realistically, the number probably is somewhere between to get enough data for the comparative analyses required for predictive understanding.

How far do we want to look at engineering organisms? It's not all that far out. Take the best parts of different organisms and put in a matrix of one's own design? Or take the farmer's approach and breed to induce mutations? Both of these work, but global approval of created organisms is necessary, and there are big ethical implications.

3. What specific technologies are desired in a facility?

Most participants focused on MS, but other technologies were discussed. Attendees noted that DOE is good at developing new technologies that would need to be incorporated into the facilities.

Most biologists have an interest in proteomics and say, "This is the short list of proteins I'm interested in." They're not thinking in a systems biology paradigm. How much of the facility will be a global systems biology-driven enterprise, and how much will be a very focused, productionist, conventional approach? These categories are not mutually exclusive, but different tools are needed to accomplish each. How do we connect these initially disconnected approaches?

One suggestion was to ask how the facility would enable new people to bring their expertise to the field, including knockouts. Wasn't proposing this to be a knockout microbe community. The yeast is considered to be a model. This is an opportunity, and someone in the community will be familiar with it.

A huge influence on this facility will be instrumentation that is not static; of necessity, it will change immensely. Thus, a mechanism to prevent obsolescence must be built in. Technology assessment and integration are needed to keep the facility and technologies current. Look to integrate new technologies as we go along. Being able to do comparisons would be valuable. For example, look at JGI's technology evaluation component where they evaluate new arrays, matrices for megabases.

Microbe cultivation at the facility should be limited. Each investigator will know best for his particular organism, but the capabilities in a facility should be flexible enough for outside users to do the following:

- High-end MS.
- Chemostats (for QC, for some but not all organisms).
- Synchronous time series for simple communities.
- Offgas.
- Controlled environmental parameters.
- Arrays as multidimensional separation component, not just for "omics."
- Ability to identify, quantify proteins in context of other "omics."
- Quantitative and qualitative capabilities.
- Governance (two-way user plan).
- Up-front plan for data distribution, management, integration, maintenance.
- Sample QC, tracking, and handling.

Biologists will want annotated proteomes and physical property (native mass, association state) and functional information. Calorimetry and surface plasmon resonance might be useful. Proteomics includes more than MS.

The goal of systems biology is to understand the cell as a whole. To do this, we need to know redox and post-translational indications. No one has talked about the role of metabolomics, which has been more or less lumped into a category. If DOE expends all this money and effort, metabolites need to be included, especially for microbial systems. This brings up another realm of processes. Even if a handful of metabolites are being done, investigators have to work from the same sample if they want to coordinate their activities with others. This would be a good consortia goal.

If one has the broad spectrum of possible metabolites for which a signature can be identified, correlative work on an experiment can show cause and effect. A group of standards will require measurements with good quantitation and dynamic range. There also will be very specific questions.

Looking at the small RNAs is not easy if they are not abundant. Microarrays could be done inexpensively at the facility on the same samples grown in the chemostat at little cost.

In terms of a global, high-throughput facility, how all these data will be used is unclear. Conceptually, we want to use them similarly to array data—tease out the networks and see what's coexpressed. The facility concept is a much larger picture. Frankly, not many people or groups are engaged in systems biology research, which is still in the early stages. It is analogous to gene sequencing and genomics, however, in that systems biology will become more commonplace as tools and capabilities are developed.

If a facility has high-throughput data available, more complex experiments can be planned. At Monsanto, for example, they are generating lots of genetic mutants. Isolines of organisms differ from 5% of genome. They can look at many, many metabolites and do transcription by environmental conditions. A great deal of confidence in the process is required to even begin looking on that scale.

4. How much has technology changed over the last 5 years? From that, can we extrapolate 5 years out? How do we plan for change?

This area was the least touched-on question at the workshop. The only comments were the following:

- The big MS meeting 5 years ago didn't include much on proteomics. Half to two-thirds of this year's presentations will be on proteomics.
- Dick Smith's system (at PNNL) can be bought now but not the front end HPLC. Some parts are not commercially available.

5. What kind of data will be given to the biologists?

Not being able to guarantee data quality is worse than having no data at all. Use of global proteomic data is one of the biggest questions we need to bring forward. It's in the early stages now, but that will change. The way genomics has come on, so will proteomics. We want GTL and the microbiology community to find out how everyone's data relates.

Proteomics requires several ways to answer more and different questions than sequencing does. The kind of data will depend on the kinds of questions asked in the facility work. If the goal is to model the cell, the model will be comprehensive enough to answer thousands of questions.

One goal should be to provide raw data access in a short period of time. Some people will want all the files. When JGI did *Rhodopseudomonas* and placed contigs on the physical maps, one or two labs wanted all the data. Some data will be raw, and some will have been worked up, but algorithms for developing quality scores will develop. All data must be archived.

Most people want flat files with data, but many want a user-friendly interface to compare proteomics data under a variety of conditions. This should run on a Java platform.

GTL should provide data and tools but not necessarily in this facility. They could be provided by other organizations funded by GTL or industry. Multiple ways are needed for the data to get there, and it needs to be done in dialog with facility staff and investigators. Investigators must be able to get raw data as well as tools to analyze and compare data.

Critical validation aspects will be different for different experiments and must include written and accepted universal validation methods.

Depending on where investigators are in the chain, they may want low-level granularity. Someone interested in a biological problem will want to know what's there and the experiment. The value of information will be not only to an individual researcher, but to the community.

It seems that we will be building a huge database. When considering various conditions, is there a way to think about filling out a global matrix on an organism for which a concerted approach is possible? Then the informatics can be built. What is the critical set of experiments for an organism that would give the biggest bang for a minimum number of samples?

Will data be enhanced as we go along? Say an investigator sees a clipped protein or crosslink that is not in the sequence. These things add up to one's aggregate view of functionality. Enhanced annotation must be self-contained or disseminated.

We need computational tools that will help us see data at a glance. This means designing systems that can do what we can't do right now.

JGI has various QA and QC scores that accompany sequence data. Nobody really works with raw data unless there's a question. They look at the final numbers.

Regarding data availability, the safe side is to put all the raw data on the Website and then eventually move to an official repository with accession numbers. Currently, there is no such thing for proteomics data.

Consider setting a time limit for making data publicly available. Tying data availability to publication does not recognize that it could be years before the data are used in a publication.

Specific questions:

- What is the data-validation process for managing data?
- What minimum level of data processing and integration is expected?
- What is the ideal?
- How will we evaluate the success of a particular assay?
- How many different types of data could we expect to integrate?

6. What are the preferred processes for working with a facility?

This area covered training and education of people accessing the facility. DOE will have preferred research areas, and they need the opportunity to assemble consortia that will pick apart those areas in a variety of detail and approaches and will have access to the facility. Our goal is to create that matrix.

Training. This is an essential facility piece. However, how often a certain user will be involved in facility use is unknown. Thought must be given to just how much training is given or required. Training should begin well in advance of facility use, especially in relation to sample preparation, exposure, and measurement. And it also needs to be part of experiment design. Addressing a lot of design issues up-front will take care of technology issues.

Workshops. Workshops should be held to bring various communities together for discussions and guidance at various levels (e.g., postdocs) and to help organize the communities. They could be held both for preparing users and as a mode of standard operation. Taking this approach would make the facility unique in DOE experience. The facility will dictate a different training model from undergraduates to faculty (e.g., Cold Spring Harbor). Focused, dedicated time is required on technology and facility use.

Core facilities should have dialogue to tell scientists how to prepare samples. The facility will do quality checks of samples as they come in to ensure that analyzing “junk” does not waste resources. Specifications are needed for sample preparation and delivery.

Some people will push the envelope and will need an R&D component to meet their needs. The facility must be dynamic and have the ability to be modified. Some people doing R&D will be in the facility or at other institutions.

7. What should the balance be between consortia and principal investigator use?

A real user facility should level the playing field and encompass and accommodate both kinds of users. Long-term scientific impact and breakthroughs, however, probably will come from larger teams, consortia, and multidisciplinary groups. The facility will be used not only by people who want to send their sample in and get data back but also by others who will want more details, integration, interaction, and student participation. This is a challenge, but the facility needs to accommodate all of them. Mechanisms should be in place to determine access, with demand being one criterion. A review and prioritization committee is critical.

One aspect of the facility is the user, and another is the facility goal of multiple microbes, multiple conditions, and unlimited outcome. In addition to massive global approach, we need a targeted approach for processes of interest to DOE (e.g., how photosynthesis works and is regulated). Another example would be decontamination. Several priority model organisms should be chosen. This kind of model fits in with the DOE Grand Challenge idea. Facility goals are fairly broad and, at this point, we need to focus more on collective biology goals.

In all cases, the facility must coordinate with the other three facilities. In some instances, the overall output is the sum of that process.

The facility possibly could be a "virtual facility" spread across multiple sites, with a common access portal. Samples could come in and be parsed out, which would save on a lot of building and give technology validation up-front. Another thought is to centralize it all. Both models have been used in the past. A large grant site requires that technologies be operational very quickly. JGI is an example of a large site with concentrating technologies. There still is room for diverse approaches.

We need to get feedback from others in terms of what they want this facility to do for their biology. DOE doesn't want to go forward without real demand. Areas are based on current thoughts and needs, but the larger biology community will dictate what the facility will become.

Closing Presentation: GTL Biosystems—Integrating Measures and Models

George Church, Harvard University

In Church's closing presentation of the workshop (see Appendix G), he discussed the need for improving models and measures in proteomics. We know the size of the genome and the size of proteins, but we don't know how big the environment is or how much metabolic data there will be.

Appendix A: Workshop Attendees

Charles Auffray
Genexpress - CNRS FRE 2571
Génomique Fonctionnelle et
Biologie Systémique en Santé
Functional Genomics and
Systemic Biology for Health
7, rue Guy Moquet - BP 8
94801 VILLEJUIF Cedex - FRANCE
charles.auffray@vjf.cnrs.fr

Peter Beernink
BBRP-LLNL
7000 East Ave., L-448
Livermore, CA 94551
Beernink1@llnl.gov

Jim Bixler
Facility Program Manager
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: P7-50
Richland, WA 99352
jim.bixler@pnl.gov

Harvey Bolton, Jr.
Biological Sciences Division Director
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: P7-50
Richland, WA 99352
harvey.bolton@pnl.gov

Andrew Bradbury
Los Alamos National Laboratory
P.O. Box 1663
Bikini Atoll Road, SM 30
Los Alamos, NM 87545
amb@lanl.gov

Darrell Chandler
Biochip Technology Center
Argonne National Laboratory
9700 Cass Avenue
202 Bldg, Room A249
Argonne, IL 60439
dchandler@anl.gov

George Church
Harvard University
Alpert 513B
200 Longwood Ave.
Boston, MA 02115
church@rascal.med.harvard.edu

Timothy Donohue
University of Wisconsin-Madison
Bacteriology Department
1550 Linden Dr.
Madison, WI 53706
tdonohue@bact.wisc.edu

Sharon Doyle
Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
sadoyle@lbl.gov

Marvin Frazier
U.S. Department of Energy
Life Sciences Division
SC-72 - Building: GTN
Germantown, MD 20874
MARVIN.FRAZIER@science.doe.gov

Jim Fredrickson
Laboratory Fellow
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: P7-50
Richland, WA 99352
jim.fredrickson@pnl.gov

Jean Futrell
Battelle Fellow
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: K9-95
Richland, WA 99352
jean.futrell@pnl.gov

Julie Gephart
Scientific and Technical Communications
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: K9-41
Richland, WA 99352
julie.gephart@pnl.gov

Carol Giometti
Argonne National Laboratory
9700 Cass Avenue
202Building - Room B117
Argonne, IL 60439
csgiometti@anl.gov

Mark Gomelsky
Department of Molecular Biology
Ag C Bldg., Rm. 6007
University of Wyoming
Laramie, WY 82071-3944
gomelsky@uwyo.edu

Christopher Hack
Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
cachack@lbl.gov

Bob Hettich
Oak Ridge National Laboratory
P.O. Box 2008 MS6131
Oak Ridge, TN 37831-6131
hettichrl@ornl.gov

Lee Hood
Institute for Systems Biology
4225 Roosevelt Way NE
Suite 200
Seattle, WA 98105
lhood@systemsbiology.org

Greg Hurst
Oak Ridge National Laboratory
P.O. Box 2008 MS6131
Oak Ridge, TN 37831-6131
hurstgb@ornl.gov

Samuel Kaplan
Dept. of Microbiology
And Molecular Genetics
University of Texas Medical School
P.O. Box 20708
Houston, Texas 77225-0708
Samuel.Kaplan@uth.tmc.edu

Mike Knotek
10127 N. Bighorn Butte Dr.
Oro Valley, AZ 85737
m.knotek@verizon.net

Eugene Kolker
BIATECH
19310 N. Creek Parkway
Suite 115
Bothell, WA 98011
ekolker@biatech.org

Frank Larimer
Genome Analysis and Systems Modeling
Life Sciences Division
Oak Ridge National Laboratory
1060 Commerce Park, Rm 211, MS-6480
Oak Ridge, TN 37831
larimerfw@ornl.gov

Reinhold Mann
Deputy Lab Dir, Science & Technology
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: K1-46
Richland, WA 99352
reinhold.mann@pnl.gov

Betty Mansfield
Human Genome Management Information
System
Oak Ridge National Laboratory
1060 Commerce Park – MS6480
Oak Ridge, TN 37831-6480
mansfieldbk@ornl.gov

Vera Matrosova
USUHS, Dept. of Pathology
4301 Jones Bridge Rd.
Bethesda, MD 20814
vmatrosova@usuhs.mil

F. Blaine Metting
Biological & Environmental Sciences
Program Manager
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: K9-76
Richland, WA 99352
blaine.metting@pnl.gov

George Michaels
Bioinformatics Director
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: P7-50
Richland, WA 99352
george.michaels@pnl.gov

Ed Michaud
Life Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6445
michaudejiii@ornl.gov

Michael Murphy
Joint Genome Institute
2800 Mitchell Drive
Walnut Creek, CA 94598
mbmurphy@lbl.gov

Marina Omelchenko
USUHS, Dept of Biology
4301 Jones Bridge Rd.
Bethesda, MD 20814
omelchen@ncbi.nlm.nih.gov

Himadri Pakrasi
Department of Biology, Box 1137
Washington University
St. Louis, MO 63130
PAKRASI@BIOLOGY.WUSTL.EDU

Karin Rodland
Protein Function Group Leader
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: P7-56
Richland, WA 99352
karin.rodland@pnl.gov

R.D. (Dick) Smith
Battelle Fellow
Pacific Northwest National Laboratory
P.O. Box 999, MSIN: K8-98
Richland, WA 99352
dick.smith@pnl.gov

Michael Thelen
Protein Biochemistry and Molecular Biology
Computational and Systems Biology Division
Biology and Biotechnology Research Programs
Lawrence Livermore National Laboratory
Livermore, CA 94550
mthelen@llnl.gov

Marvin Vestal
Applied Biosystems
500 Old Connecticut Path
Framingham, MA 01701
vestalml@appliedbiosystems.com

Julian P. Whitelegge
University of California – Los Angeles
Department of Chemistry
405 Hilgard Avenue
Los Angeles, CA 90095
jpw@chem.ucla.edu

Appendix B: Workshop Agenda

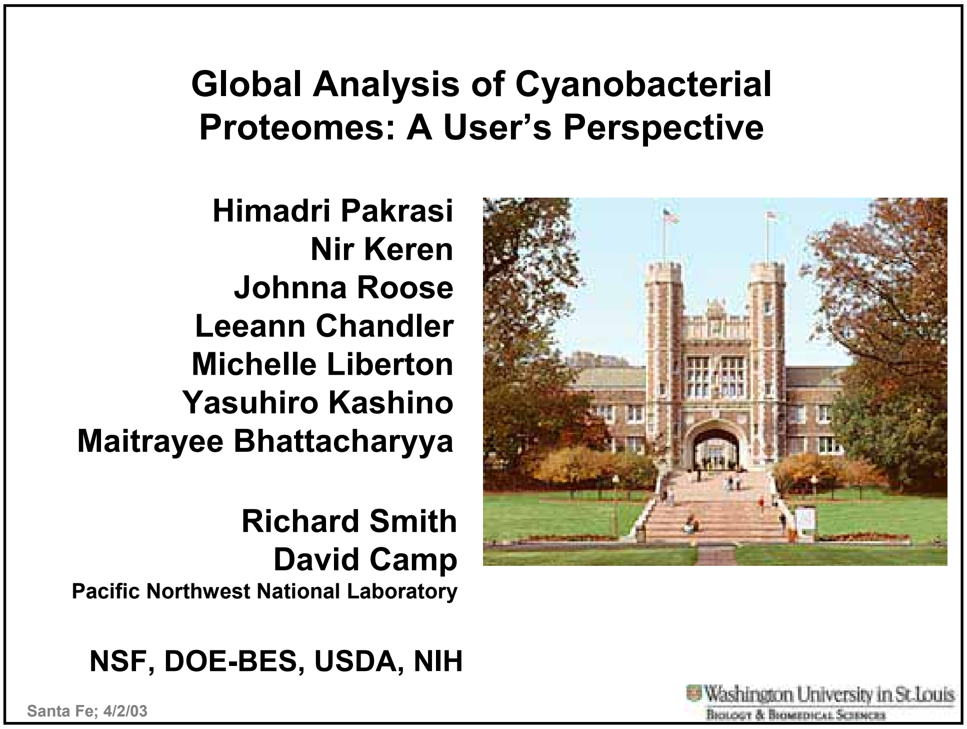
Tuesday, April 1, 2003 – Coronado Room

7:30 p.m. – 7:45 p.m.	Introduction and Explanation of Format – Jean Futrell
7:45 – 8:30	Overview of DOE Facilities Concept – Marv Frazier
8:30 – 9:30	Application of Proteomics to Systems Biology – Lee Hood

Wednesday, April 2, 2003 – New Mexico Room

8:00 a.m. – 8:10 a.m.	Objectives for Breakout Sessions – Karin Rodland
8:10 – 9:45 a.m.	Presentation of Three Scenarios:
8:10 – 8:30	Himadri Pakrasi – *Synechocystis*
8:30 – 8:50	Tim Donohue – *Rhodopseudomonas*
8:50 – 9:10	Jim Fredrickson – *Shewanella*
9:10 – 9:45	Discussion of Scenarios
9:45 – 10:00 a.m.	Break
10:00 – 12:00 n	Tool Kit Presentations:
10:00 – 10:30	Marvin Vestal – *Proteomic Technologies*
10:30 – 11:00	Carol Giometti – *2D Gels for Proteomics*
11:00 – 11:30	Darrell Chandler – *Microarray Technologies for Proteomics*
11:30 – 12:00 n	Dick Smith – *Global Proteomics*
12:00 – 2:30 p.m.	Breakout Sessions (working lunch) – New Mexico, Santa Fe, and Exchange Rooms
2:30 – 2:45	Break
2:45 – 3:00	Reports from Breakout Sessions
3:00 – 4:00	Open Discussion
4:00 – 4:15	Closing Remarks – George Church
4:30 – 7:00	Wrapup Session for Presenters, Breakout Moderators, and Reporters Only

Appendix C: Research Scenarios



Global Analysis of Cyanobacterial Proteomes: A User's Perspective

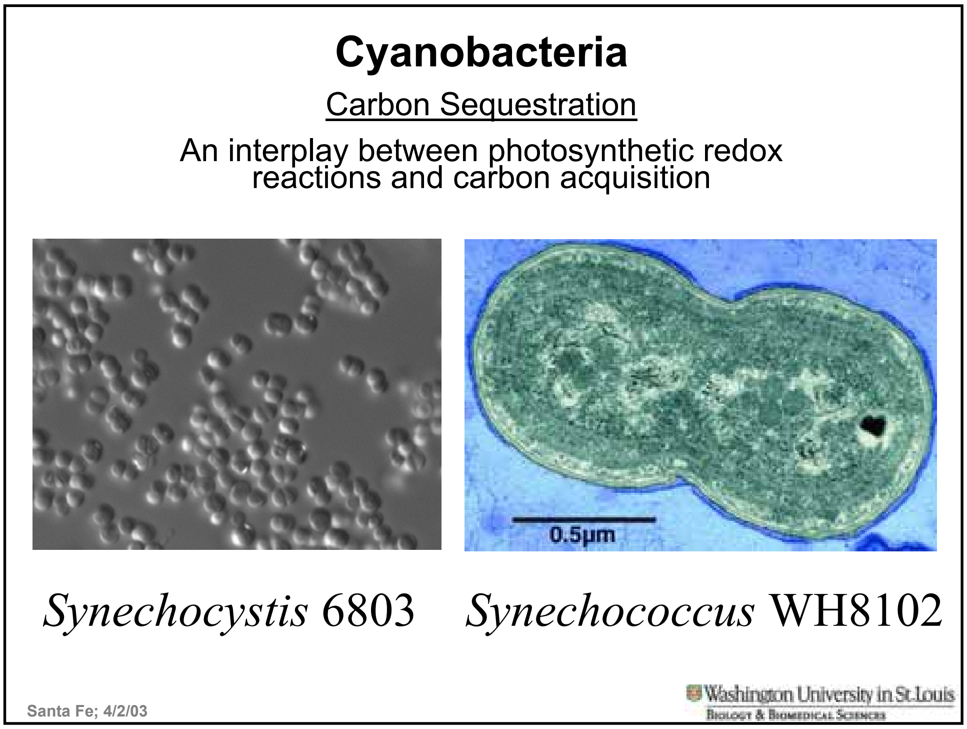
Himadri Pakrasi
Nir Keren
Johnna Roose
Leeann Chandler
Michelle Liberton
Yasuhiro Kashino
Maitrayee Bhattacharyya

Richard Smith
David Camp
Pacific Northwest National Laboratory

NSF, DOE-BES, USDA, NIH

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Cyanobacteria

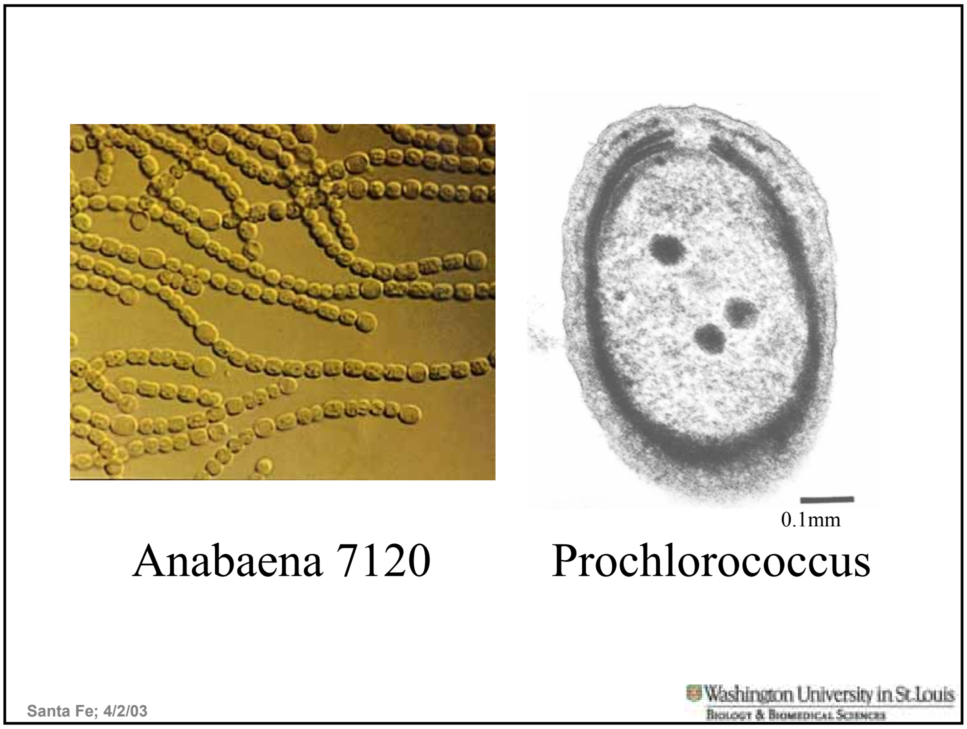
Carbon Sequestration

An interplay between photosynthetic redox reactions and carbon acquisition

Synechocystis 6803 *Synechococcus* WH8102

Santa Fe; 4/2/03

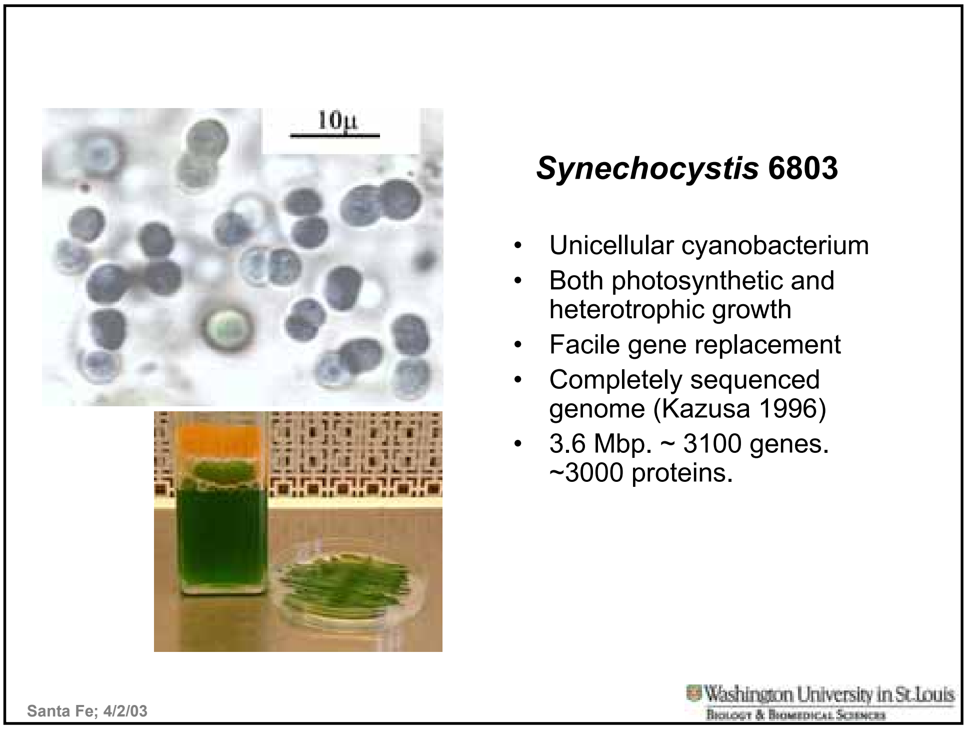
Washington University in St.Louis
Biology & Biomedical Sciences



Anabaena 7120 Prochlorococcus

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences

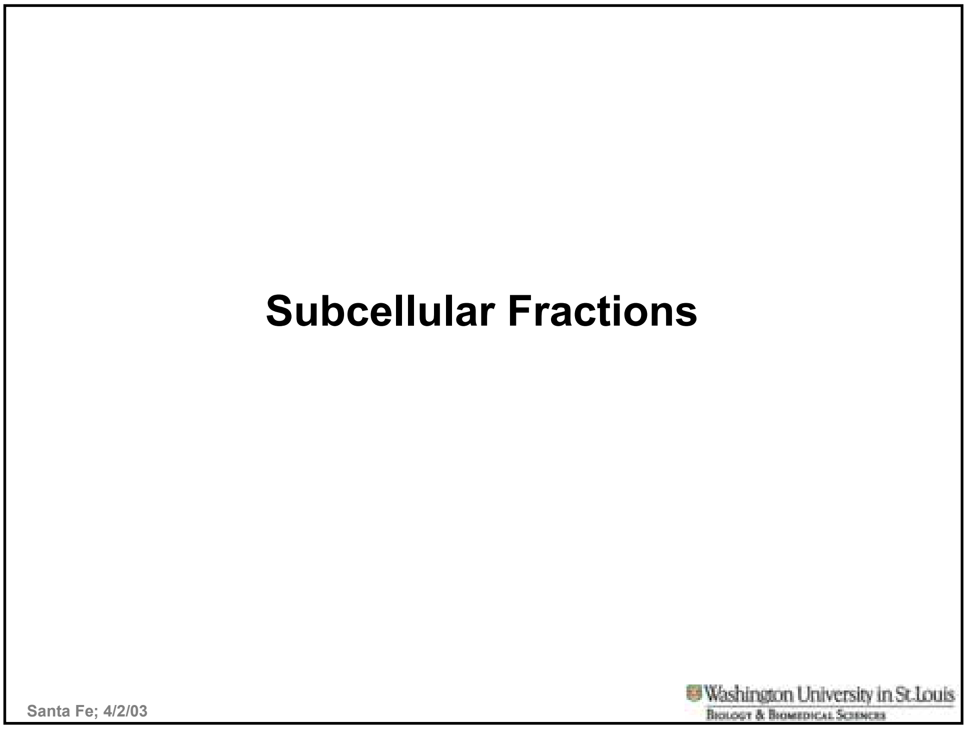


***Synechocystis* 6803**

- Unicellular cyanobacterium
- Both photosynthetic and heterotrophic growth
- Facile gene replacement
- Completely sequenced genome (Kazusa 1996)
- 3.6 Mbp. ~ 3100 genes. ~3000 proteins.

Santa Fe; 4/2/03

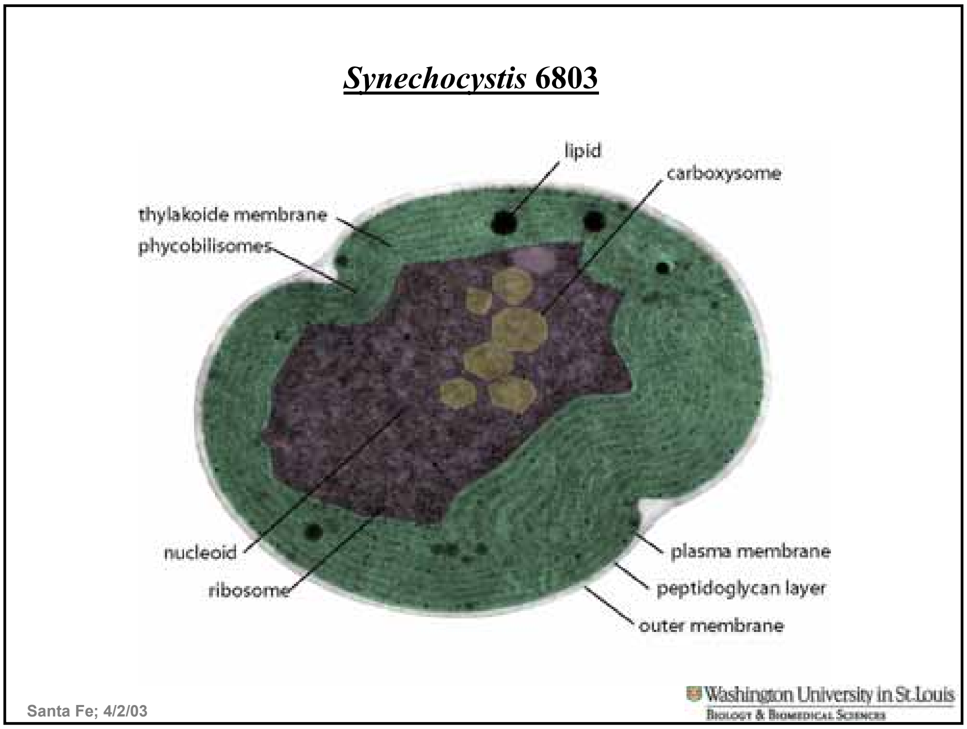
Washington University in St.Louis
Biology & Biomedical Sciences



Subcellular Fractions

Santa Fe; 4/2/03

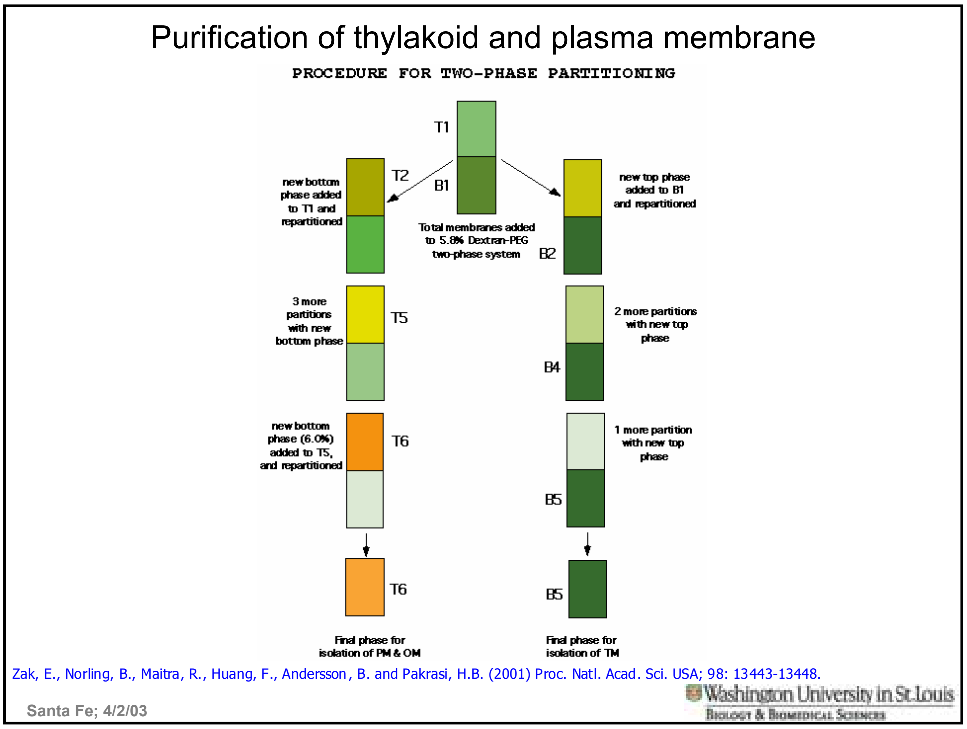
Washington University in St.Louis
Biology & Biomedical Sciences



***Synechocystis* 6803**

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



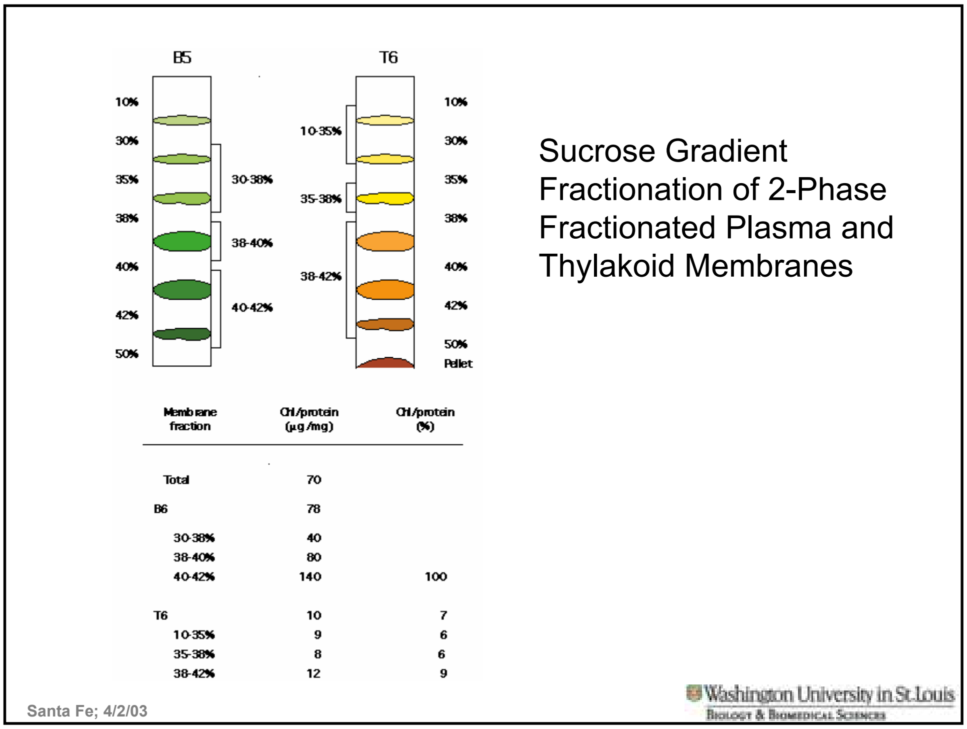
Purification of thylakoid and plasma membrane

PROCEDURE FOR TWO-PHASE PARTITIONING
T1
B1
Total membranes added to 5.8% Dextran-PEG two-phase system
T2
new bottom phase added to T1 and repartitioned
new top phase added to B1 and repartitioned
B2
3 more partitions with new bottom phase
T5
2 more partitions with new top phase
B4
new bottom phase (6.0%) added to T5, and repartitioned
T6
1 more partition with new top phase
B5
T6
B5
Final phase for isolation of PM & OM
Final phase for isolation of TM

Zak, E., Norling, B., Maitra, R., Huang, F., Andersson, B. and Pakrasi, H.B. (2001) Proc. Natl. Acad. Sci. USA; 98: 13443-13448.

Washington University in St.Louis
Biology & Biomedical Sciences

Santa Fe; 4/2/03

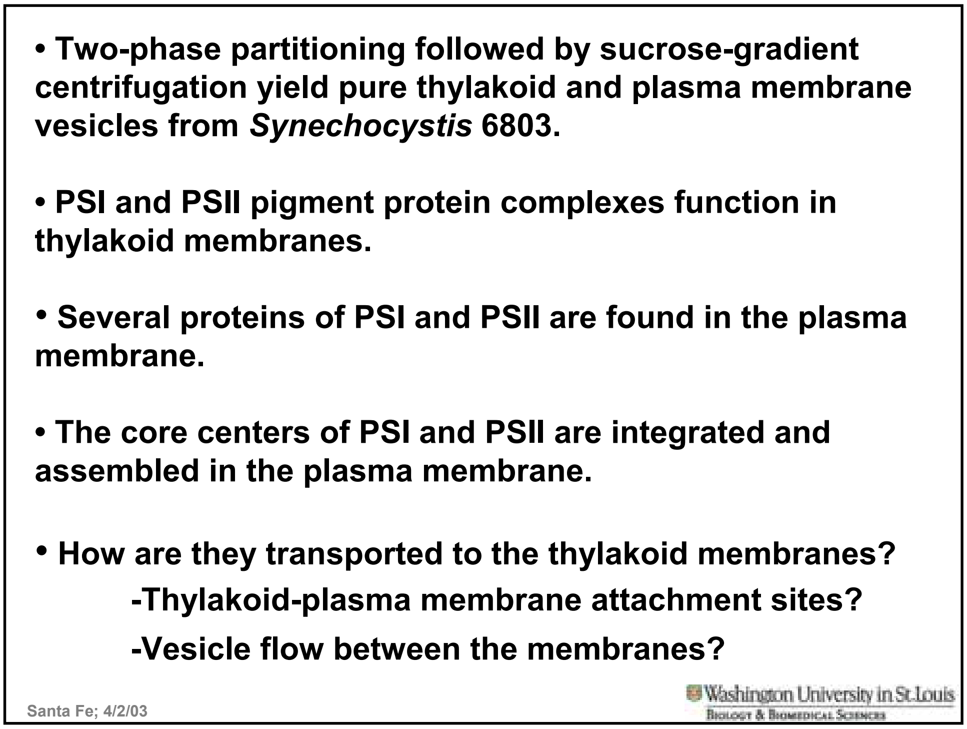


Sucrose Gradient Fractionation of 2-Phase Fractionated Plasma and Thylakoid Membranes

Membrane fraction	Chl/protein (µg/mg)	Chl/protein (%)
Total	70	
B6	78	
30-38%	40	
38-40%	80	
40-42%	140	100
T6	10	7
10-35%	9	6
35-38%	8	6
38-42%	12	9

Washington University in St.Louis
Biology & Biomedical Sciences

Santa Fe; 4/2/03



- Two-phase partitioning followed by sucrose-gradient centrifugation yield pure thylakoid and plasma membrane vesicles from *Synechocystis* 6803.

- PSI and PSII pigment protein complexes function in thylakoid membranes.

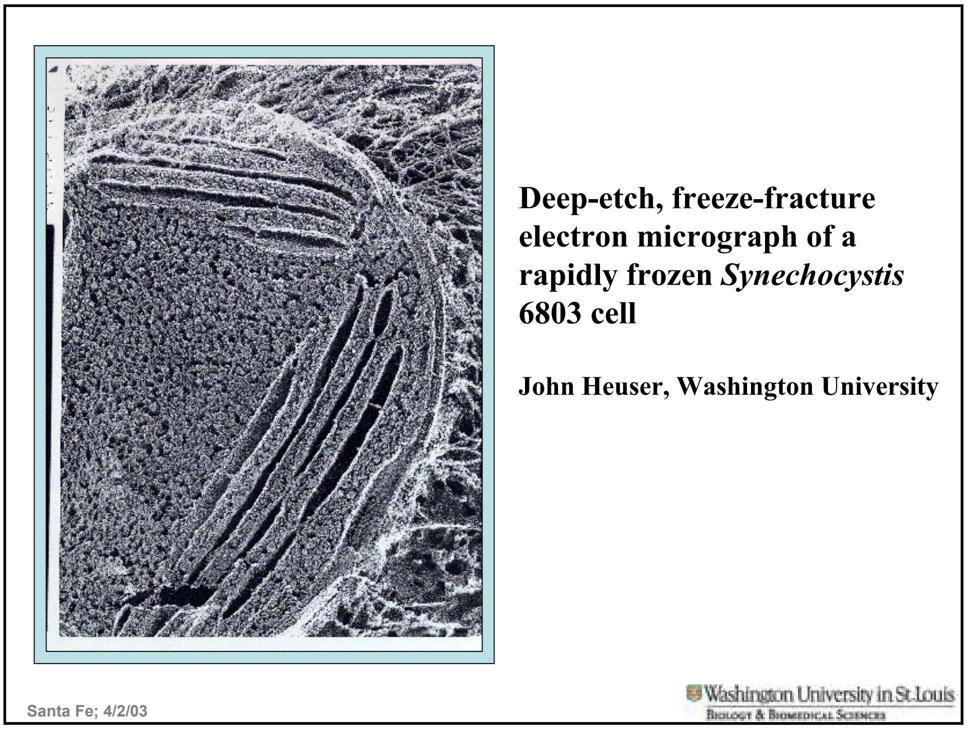
- Several proteins of PSI and PSII are found in the plasma membrane.

- The core centers of PSI and PSII are integrated and assembled in the plasma membrane.

- How are they transported to the thylakoid membranes?
  - -Thylakoid-plasma membrane attachment sites?
  - -Vesicle flow between the membranes?

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Deep-etch, freeze-fracture electron micrograph of a rapidly frozen *Synechocystis* 6803 cell

John Heuser, Washington University

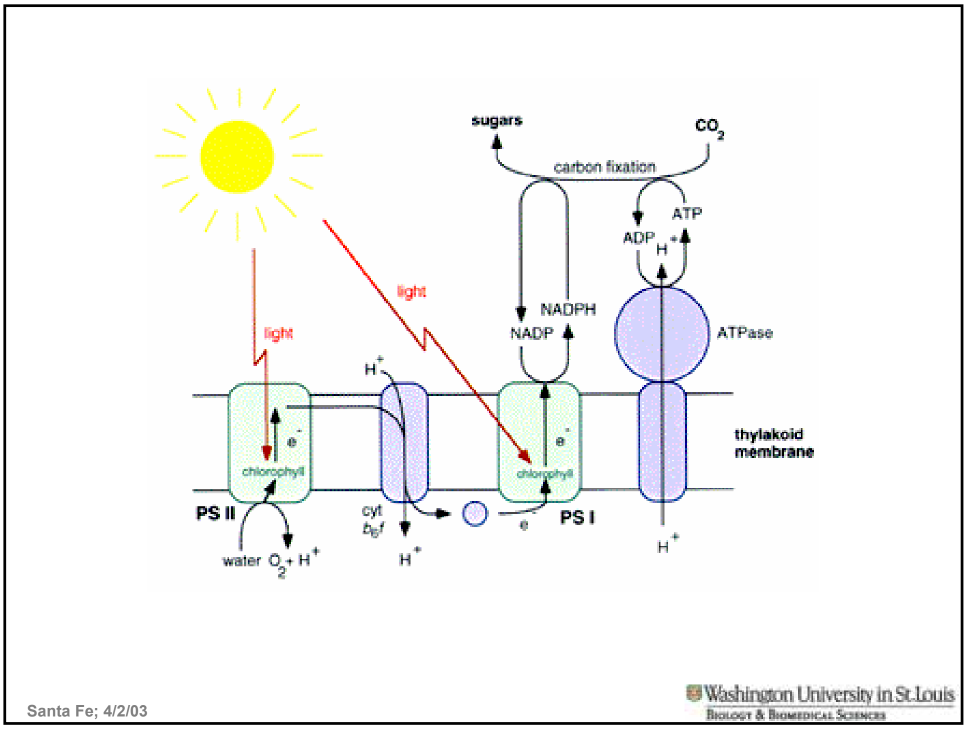
Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences

Isolation of tagged protein complex

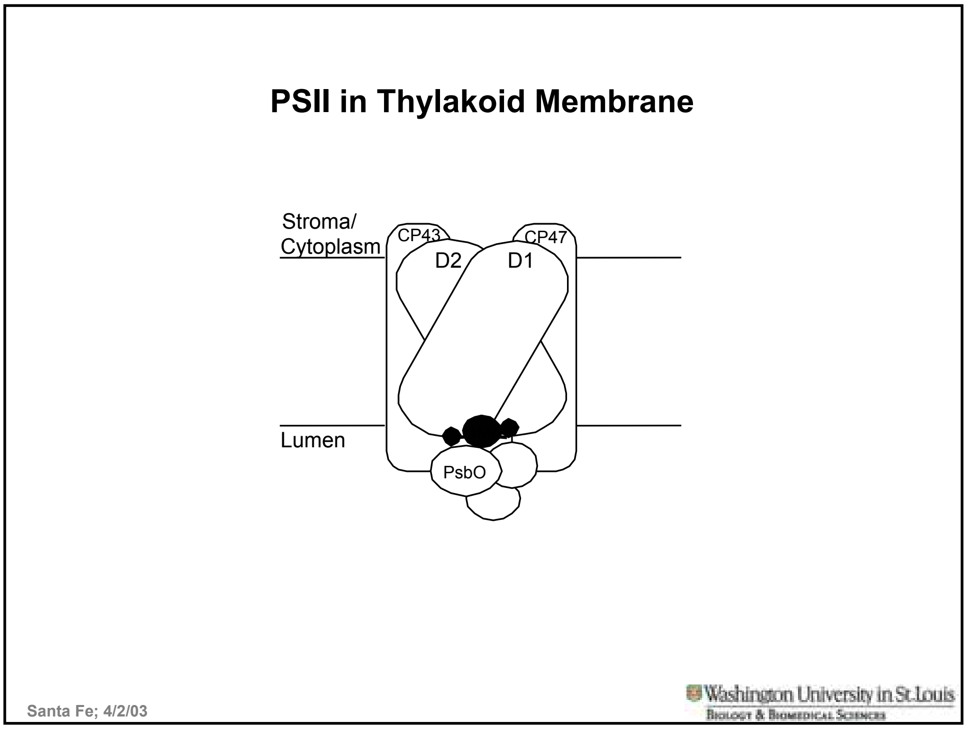
Santa Fe; 4/2/03

Washington University in St. Louis
Biology & Biomedical Sciences



Santa Fe; 4/2/03

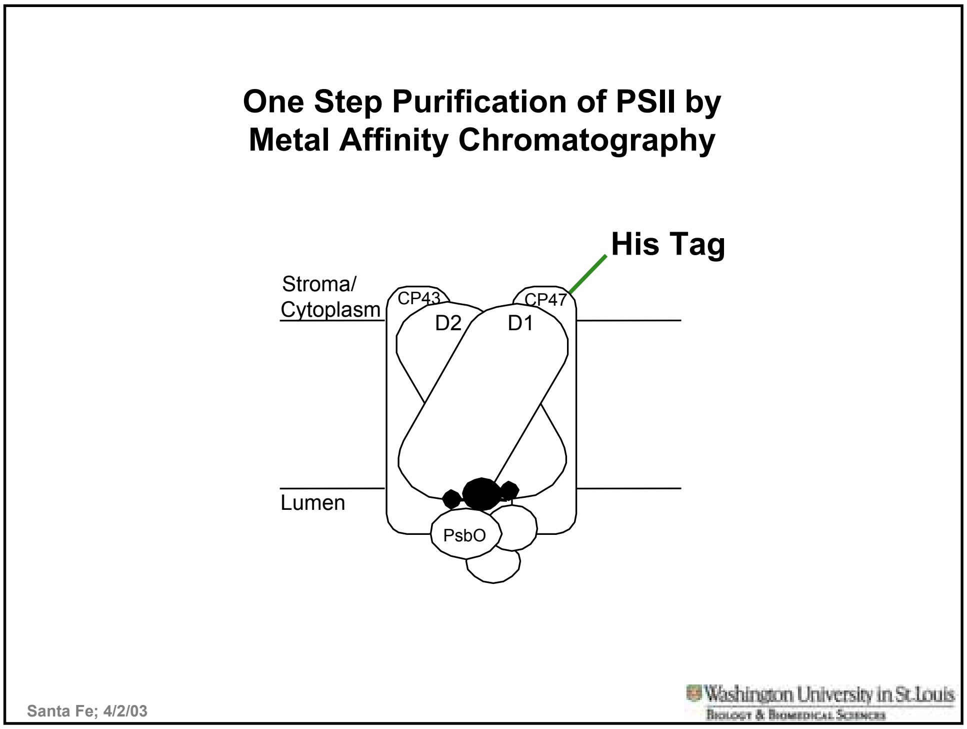
Washington University in St. Louis
Biology & Biomedical Sciences



PSII in Thylakoid Membrane

Santa Fe; 4/2/03

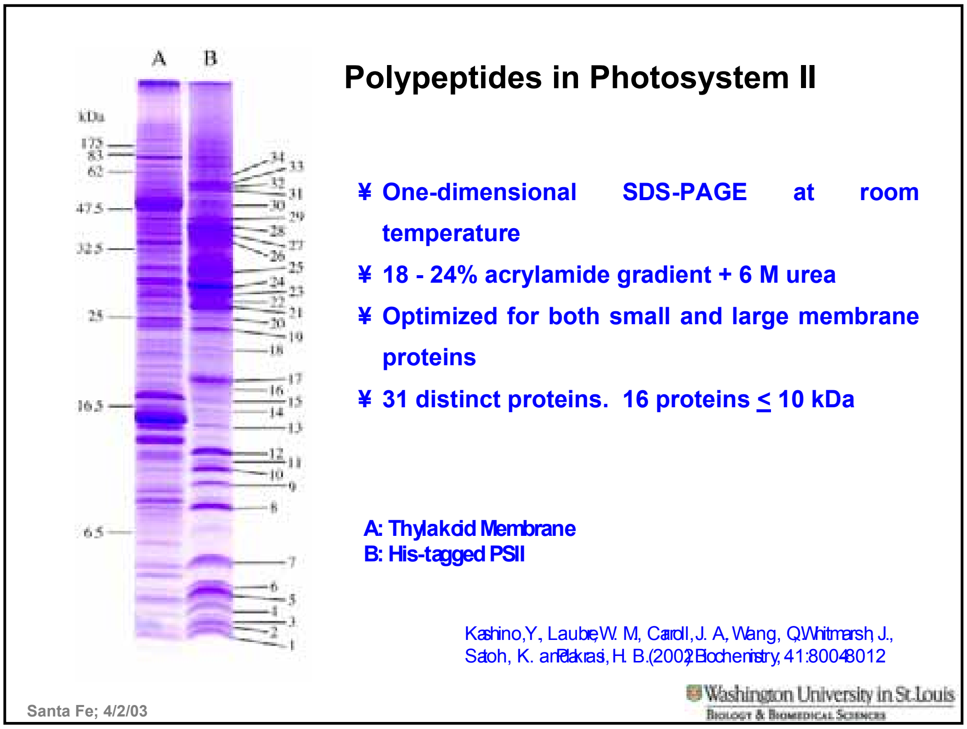
Washington University in St.Louis
Biology & Biomedical Sciences



One Step Purification of PSII by Metal Affinity Chromatography

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Polypeptides in Photosystem II

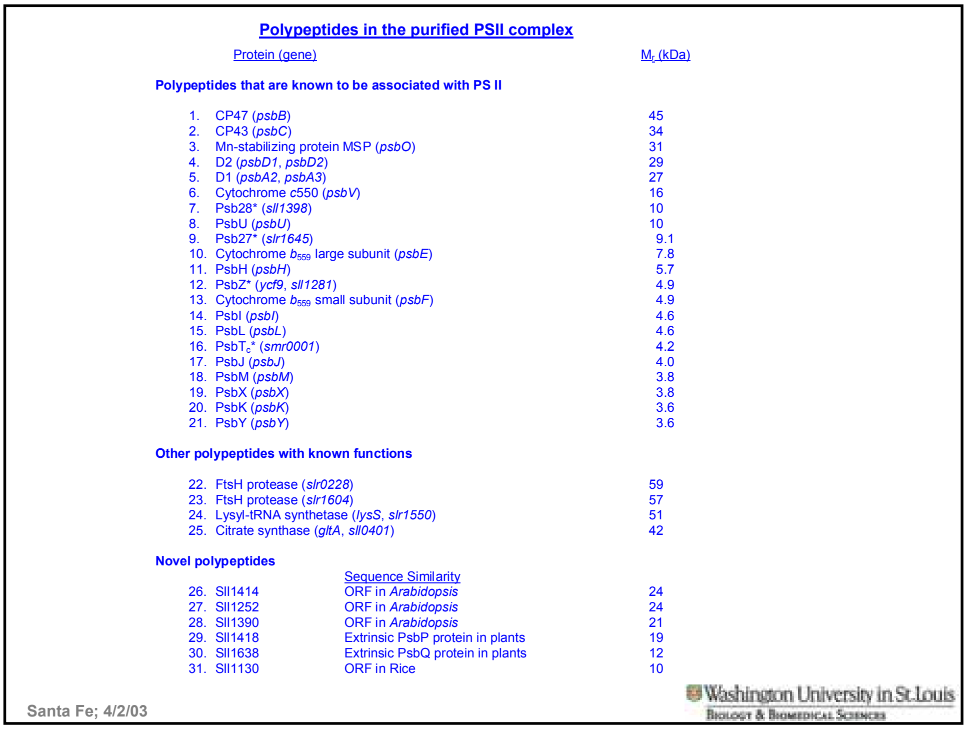
- One-dimensional SDS-PAGE at room temperature
- 18 - 24% acrylamide gradient + 6 M urea
- Optimized for both small and large membrane proteins
- 31 distinct proteins. 16 proteins ≤ 10 kDa

A: Thylakoid Membrane
B: His-tagged PSII

Kashino, Y., Lauber, W. M., Carroll, J. A., Wang, Q., Whitmarsh, J., Satoh, K. and Pakrasi, H. B. (2002) Biochemistry, 41:8004-8012

Santa Fe; 4/2/03

Washington University in St. Louis
Biology & Biomedical Sciences



Polypeptides in the purified PSII complex

Polypeptides that are known to be associated with PS II

Protein (gene)	M_r (kDa)
1. CP47 (*psbB*)	45
2. CP43 (*psbC*)	34
3. Mn-stabilizing protein MSP (*psbO*)	31
4. D2 (*psbD1*, *psbD2*)	29
5. D1 (*psbA2*, *psbA3*)	27
6. Cytochrome c550 (*psbV*)	16
7. Psb28* (*sll1398*)	10
8. PsbU (*psbU*)	10
9. Psb27* (*slr1645*)	9.1
10. Cytochrome b_{559} large subunit (*psbE*)	7.8
11. PsbH (*psbH*)	5.7
12. PsbZ* (*ycf9*, *sll1281*)	4.9
13. Cytochrome b_{559} small subunit (*psbF*)	4.9
14. PsbI (*psbI*)	4.6
15. PsbL (*psbL*)	4.6
16. PsbT$_c$* (*smr0001*)	4.2
17. PsbJ (*psbJ*)	4.0
18. PsbM (*psbM*)	3.8
19. PsbX (*psbX*)	3.8
20. PsbK (*psbK*)	3.6
21. PsbY (*psbY*)	3.6

Other polypeptides with known functions

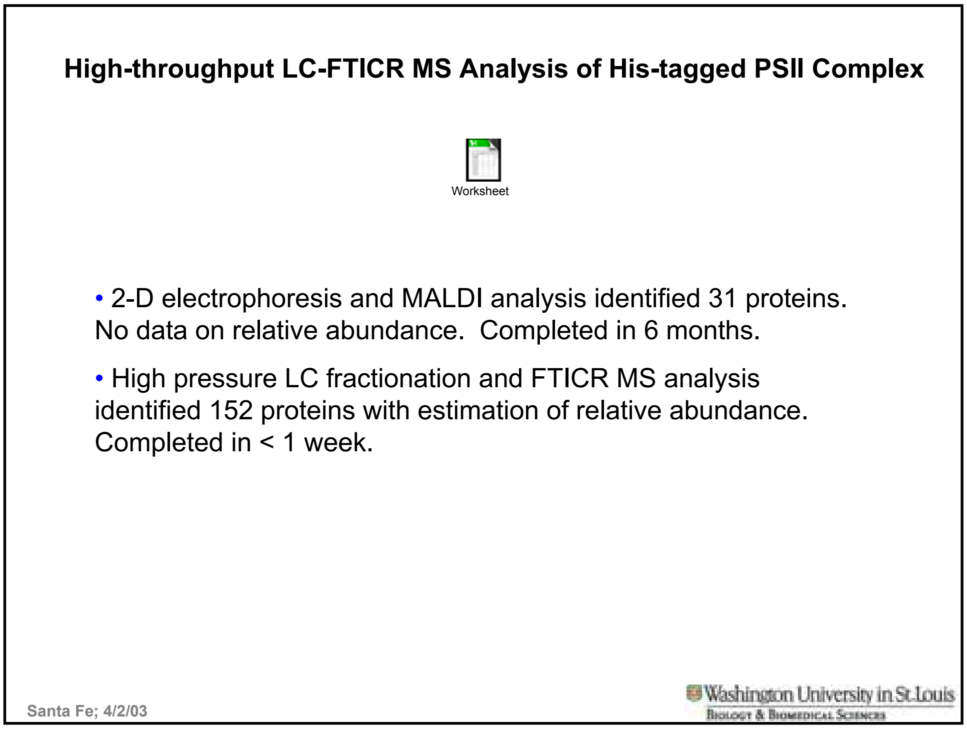
Protein (gene)	M_r (kDa)
22. FtsH protease (*slr0228*)	59
23. FtsH protease (*slr1604*)	57
24. Lysyl-tRNA synthetase (*lysS*, *slr1550*)	51
25. Citrate synthase (*gltA*, *sll0401*)	42

Novel polypeptides

Protein (gene)	Sequence Similarity	M_r (kDa)
26. Sll1414	ORF in *Arabidopsis*	24
27. Sll1252	ORF in *Arabidopsis*	24
28. Sll1390	ORF in *Arabidopsis*	21
29. Sll1418	Extrinsic PsbP protein in plants	19
30. Sll1638	Extrinsic PsbQ protein in plants	12
31. Sll1130	ORF in Rice	10

Santa Fe; 4/2/03

Washington University in St. Louis
Biology & Biomedical Sciences



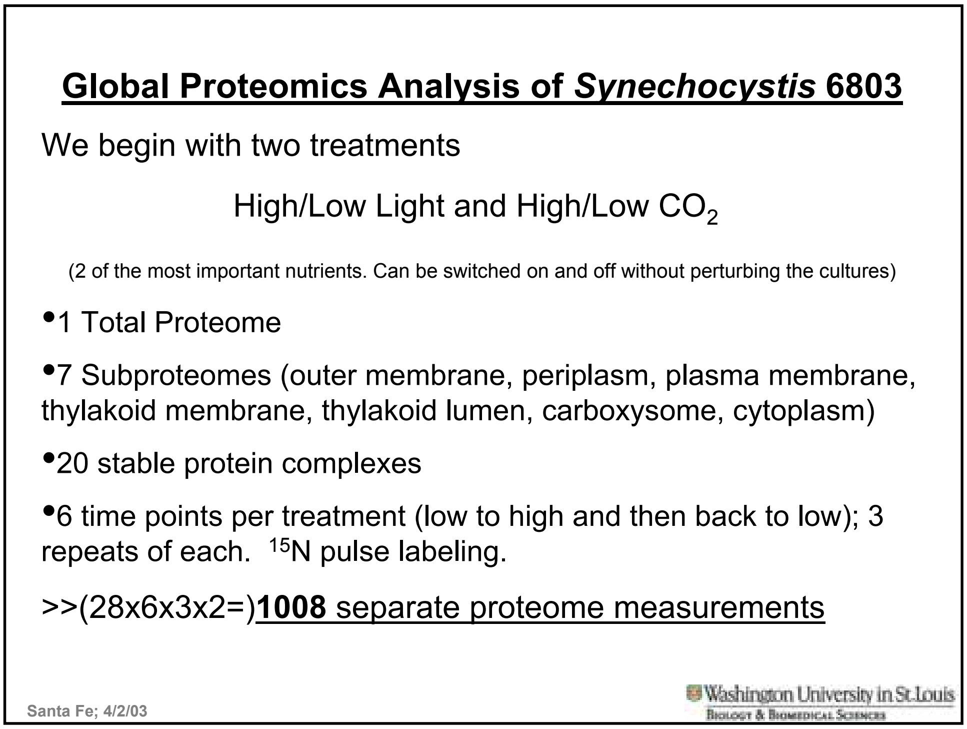
High-throughput LC-FTICR MS Analysis of His-tagged PSII Complex

Worksheet

- 2-D electrophoresis and MALDI analysis identified 31 proteins. No data on relative abundance. Completed in 6 months.
- High pressure LC fractionation and FTICR MS analysis identified 152 proteins with estimation of relative abundance. Completed in < 1 week.

Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Global Proteomics Analysis of *Synechocystis* 6803

We begin with two treatments

High/Low Light and High/Low CO_2

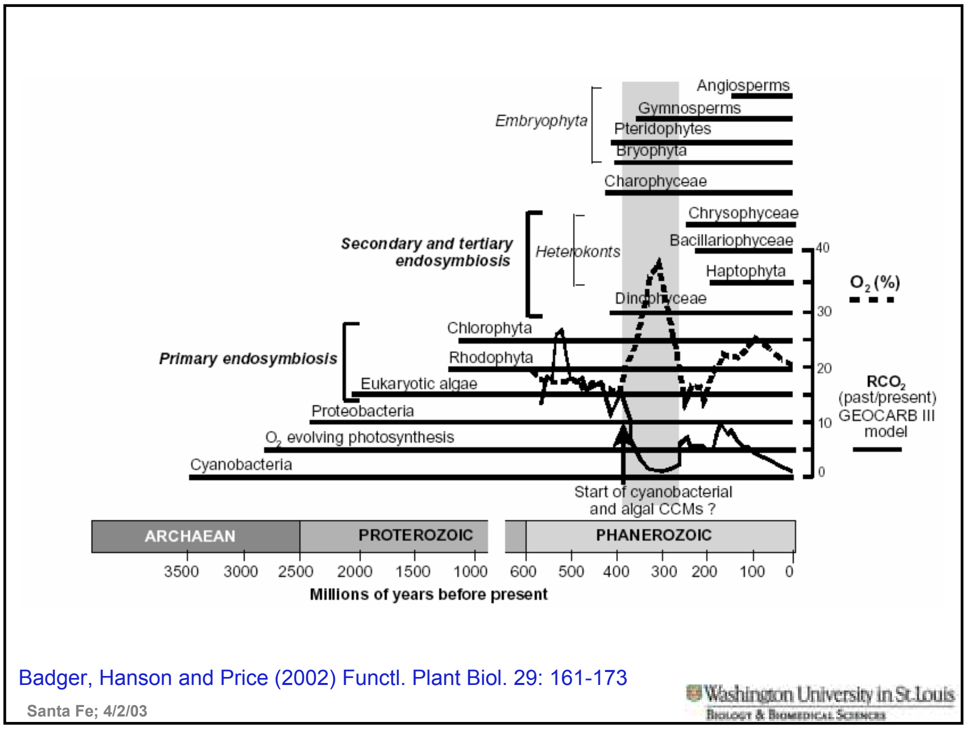
(2 of the most important nutrients. Can be switched on and off without perturbing the cultures)

- 1 Total Proteome
- 7 Subproteomes (outer membrane, periplasm, plasma membrane, thylakoid membrane, thylakoid lumen, carboxysome, cytoplasm)
- 20 stable protein complexes
- 6 time points per treatment (low to high and then back to low); 3 repeats of each. ^{15}N pulse labeling.

>>(28x6x3x2=)**1008** separate proteome measurements

Santa Fe; 4/2/03

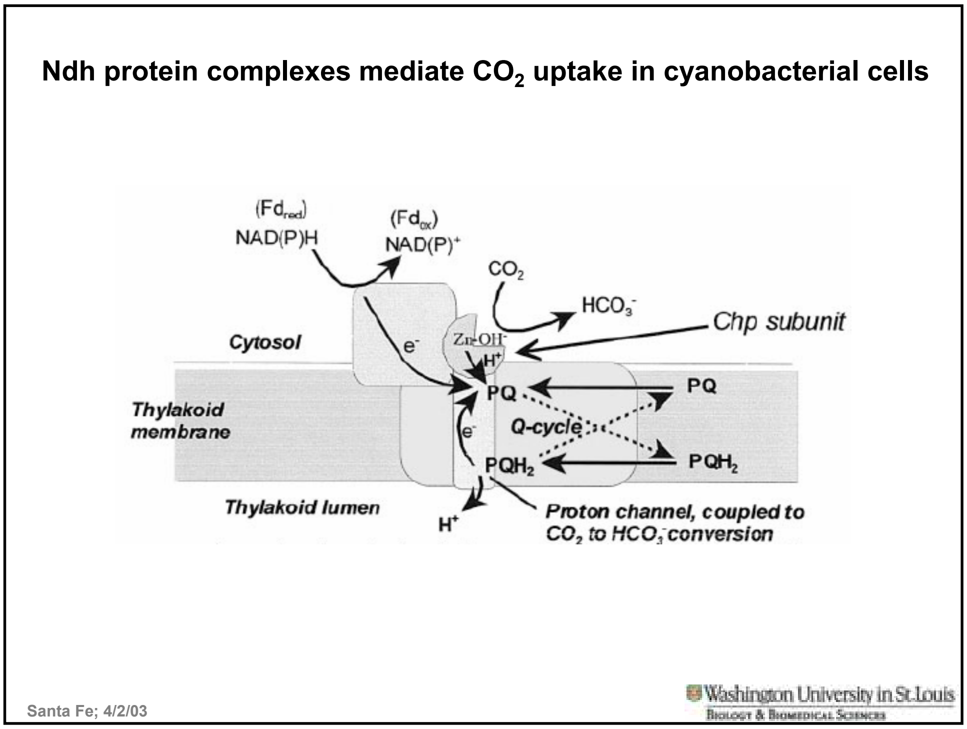
Washington University in St.Louis
Biology & Biomedical Sciences



Badger, Hanson and Price (2002) Functl. Plant Biol. 29: 161-173

Santa Fe; 4/2/03

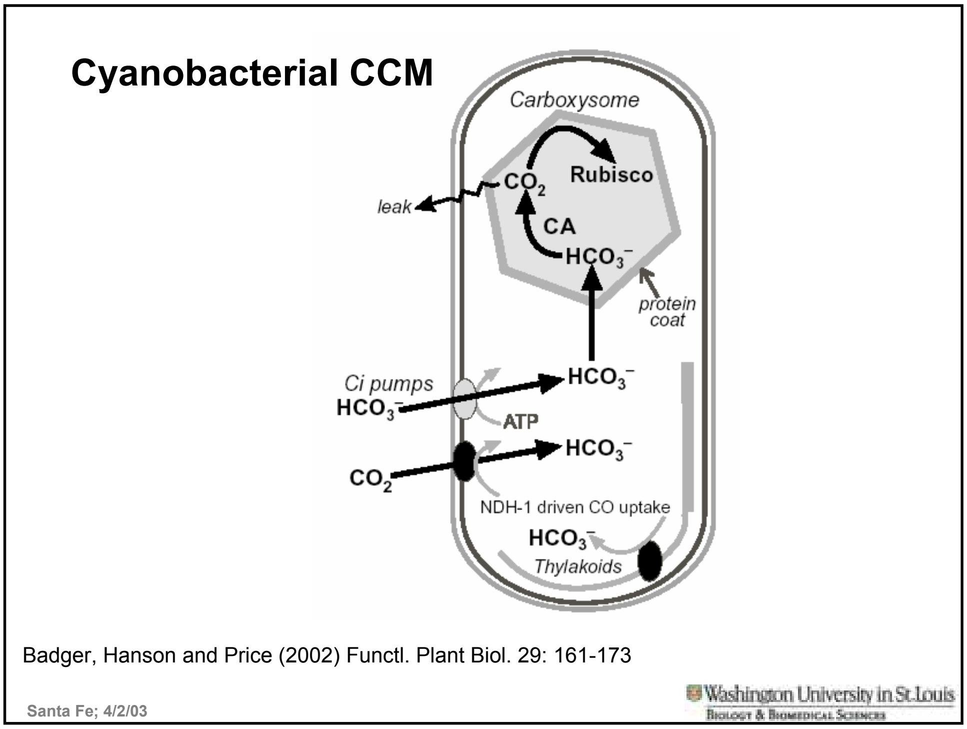
Washington University in St. Louis
Biology & Biomedical Sciences



Ndh protein complexes mediate CO_2 uptake in cyanobacterial cells

Santa Fe; 4/2/03

Washington University in St. Louis
Biology & Biomedical Sciences

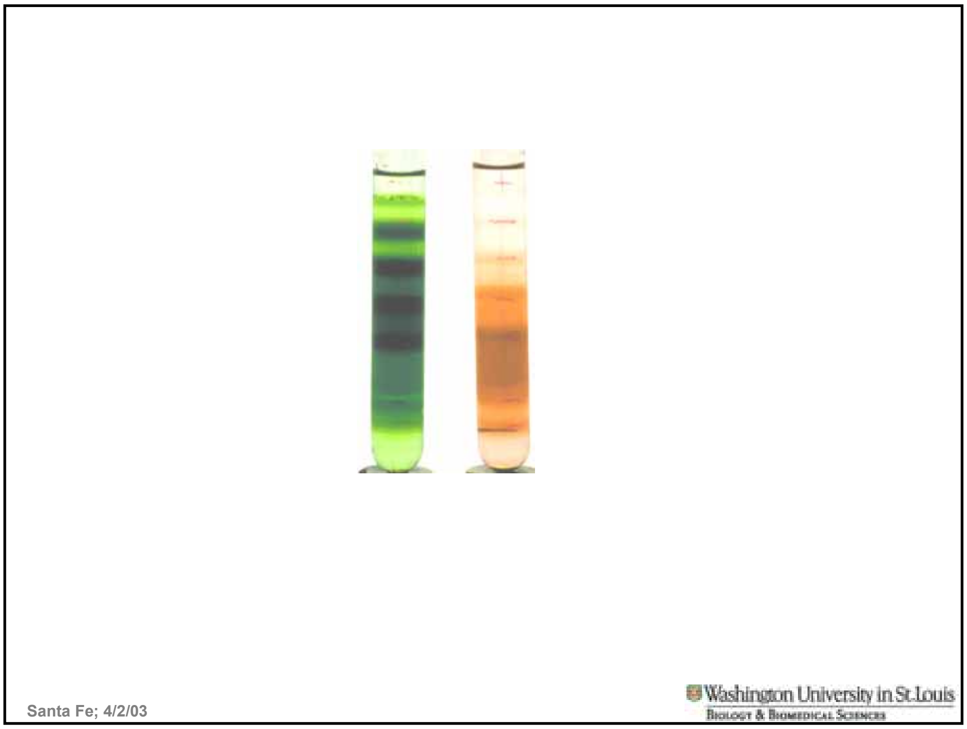


Cyanobacterial CCM

Badger, Hanson and Price (2002) Functl. Plant Biol. 29: 161-173

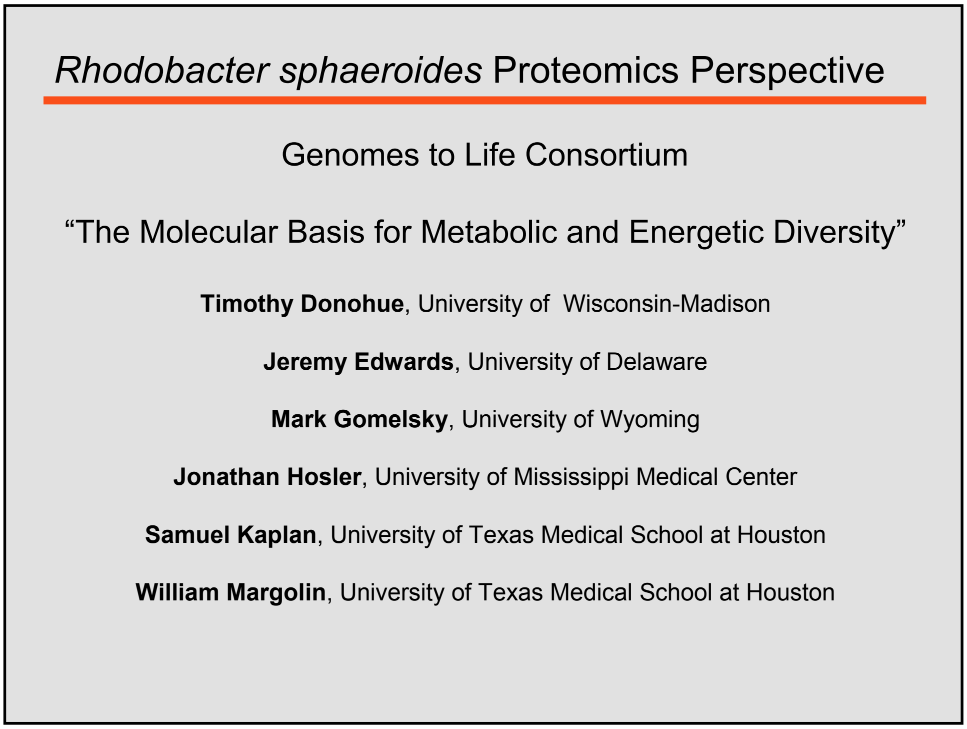
Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Santa Fe; 4/2/03

Washington University in St.Louis
Biology & Biomedical Sciences



Rhodobacter sphaeroides Proteomics Perspective

Genomes to Life Consortium

"The Molecular Basis for Metabolic and Energetic Diversity"

Timothy Donohue, University of Wisconsin-Madison

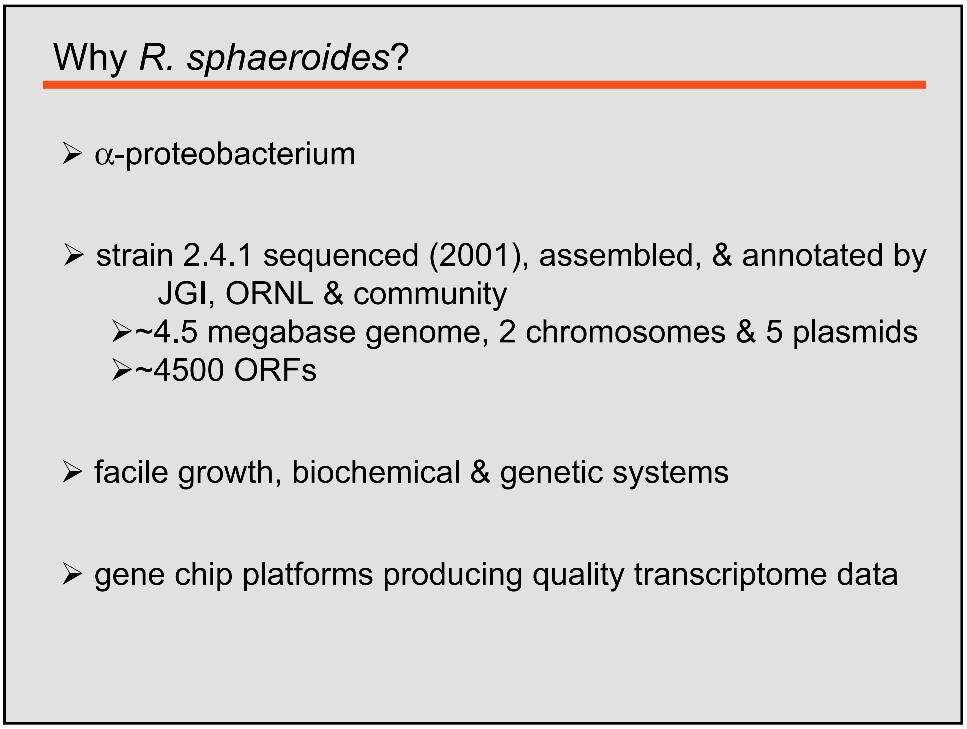
Jeremy Edwards, University of Delaware

Mark Gomelsky, University of Wyoming

Jonathan Hosler, University of Mississippi Medical Center

Samuel Kaplan, University of Texas Medical School at Houston

William Margolin, University of Texas Medical School at Houston



Why *R. sphaeroides*?

- α-proteobacterium
- strain 2.4.1 sequenced (2001), assembled, & annotated by JGI, ORNL & community
  - ~4.5 megabase genome, 2 chromosomes & 5 plasmids
  - ~4500 ORFs
- facile growth, biochemical & genetic systems
- gene chip platforms producing quality transcriptome data

Why *R. sphaeroides*?

Energetic schemes include:

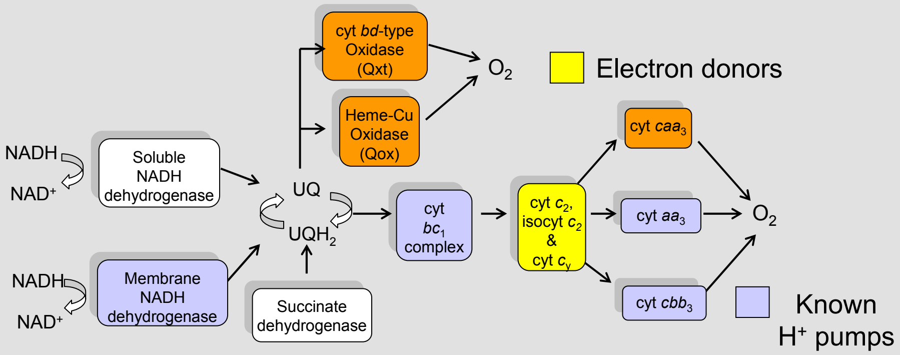
- Photosynthesis
  - Light reactions: Solar energy utilization
  - Dark reactions: CO_2 sequestration
- Respiration (O_2 plus other electron acceptors)
- H_2 production
- Oxidation of organic toxins
- Reduction of metal oxyanions

Synthesis of biodegradable plastics

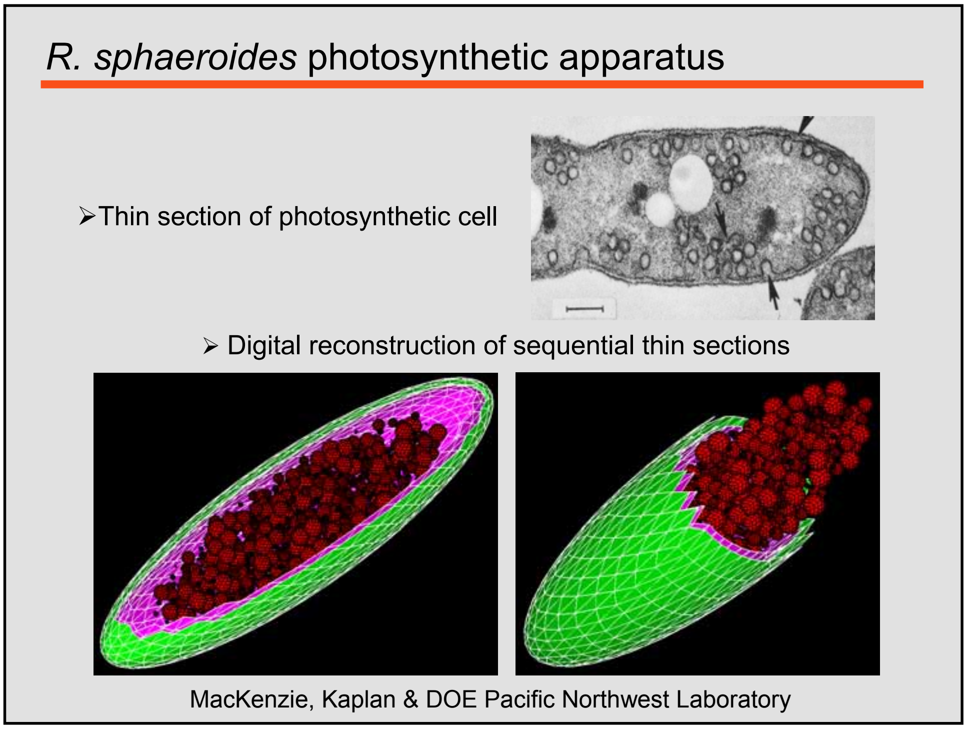
Common link: generation/production of reducing power by bioenergetic pathways

Illustrate proteomics needs by comparing photosynthetic & aerobic respiratory lifestyles

Aerobic respiratory chain



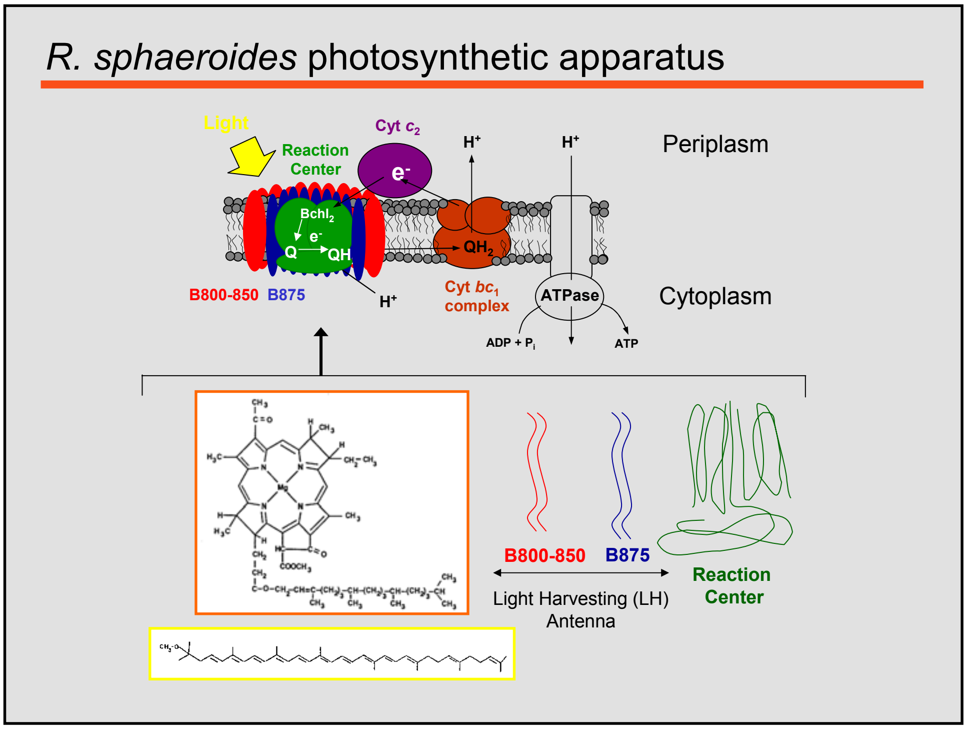
- Many *membrane bound* enzymes
- Variable abundance (*sensitivity*)
- Heme covalenty attached to *c*-type (*post-translational*)



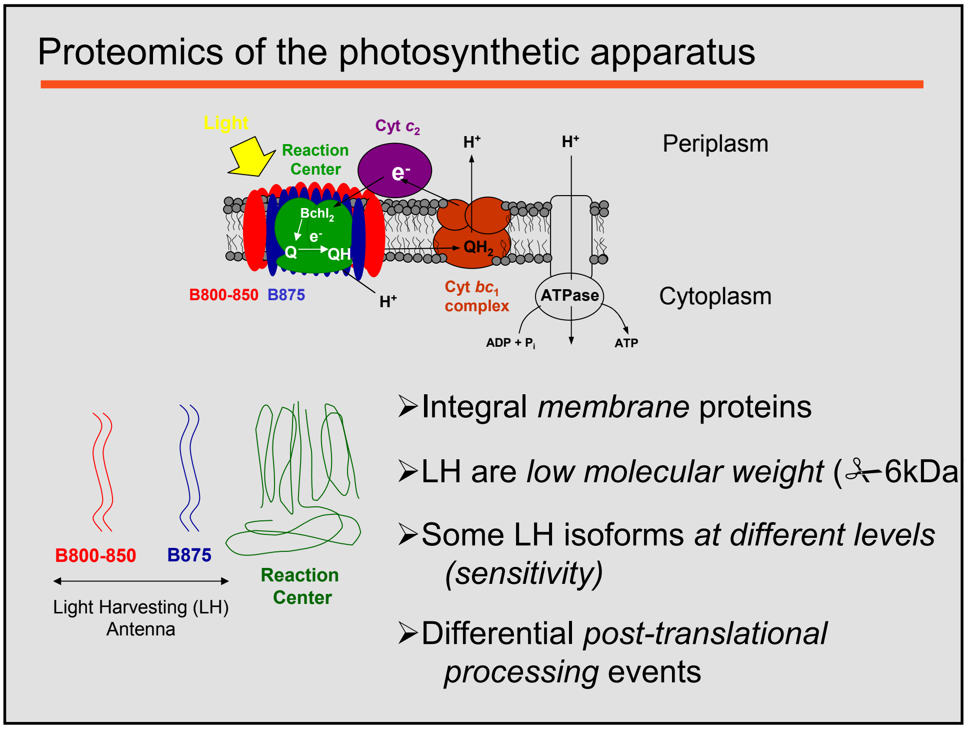
R. sphaeroides photosynthetic apparatus

- Thin section of photosynthetic cell
- Digital reconstruction of sequential thin sections

MacKenzie, Kaplan & DOE Pacific Northwest Laboratory

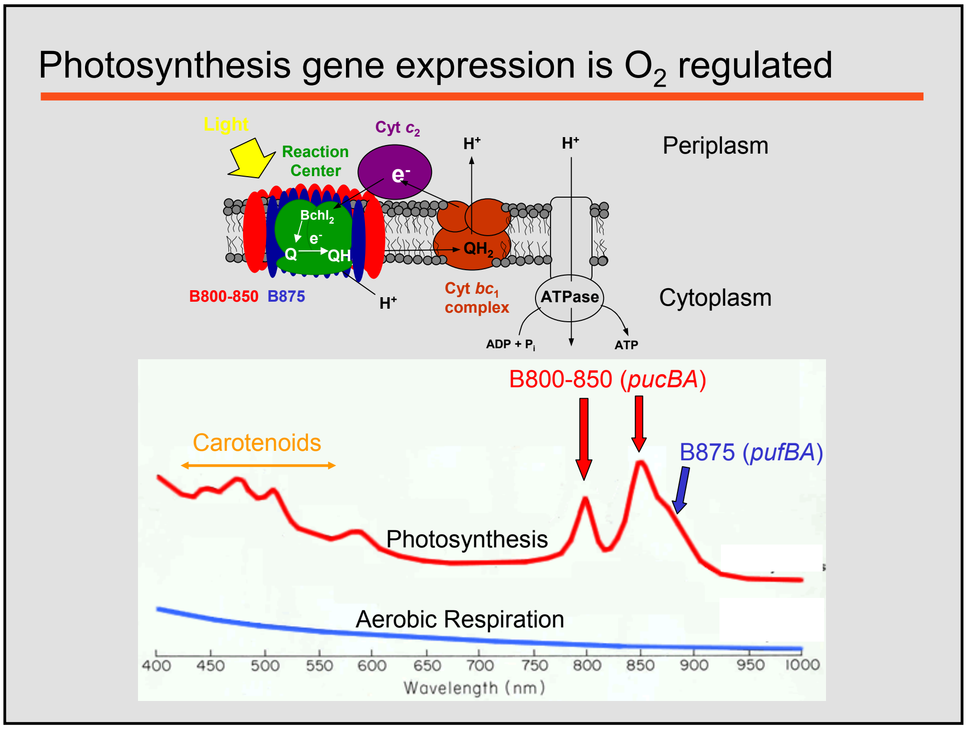


R. sphaeroides photosynthetic apparatus

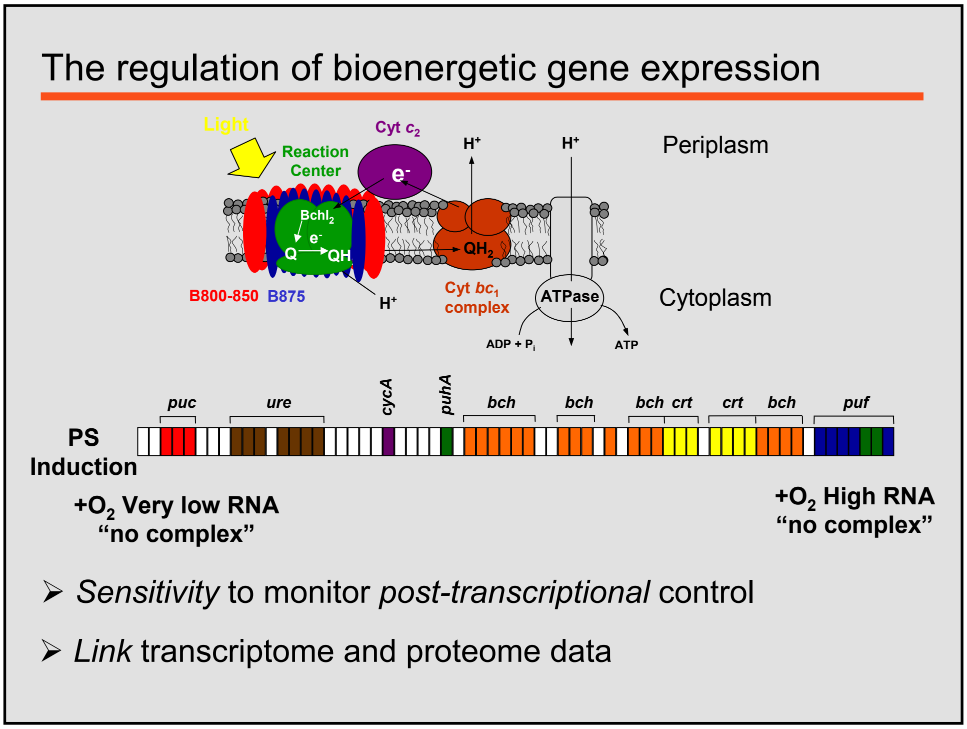


Proteomics of the photosynthetic apparatus

- Integral *membrane* proteins
- LH are *low molecular weight* (✂6kDa
- Some LH isoforms *at different levels (sensitivity)*
- Differential *post-translational processing* events



Photosynthesis gene expression is O_2 regulated

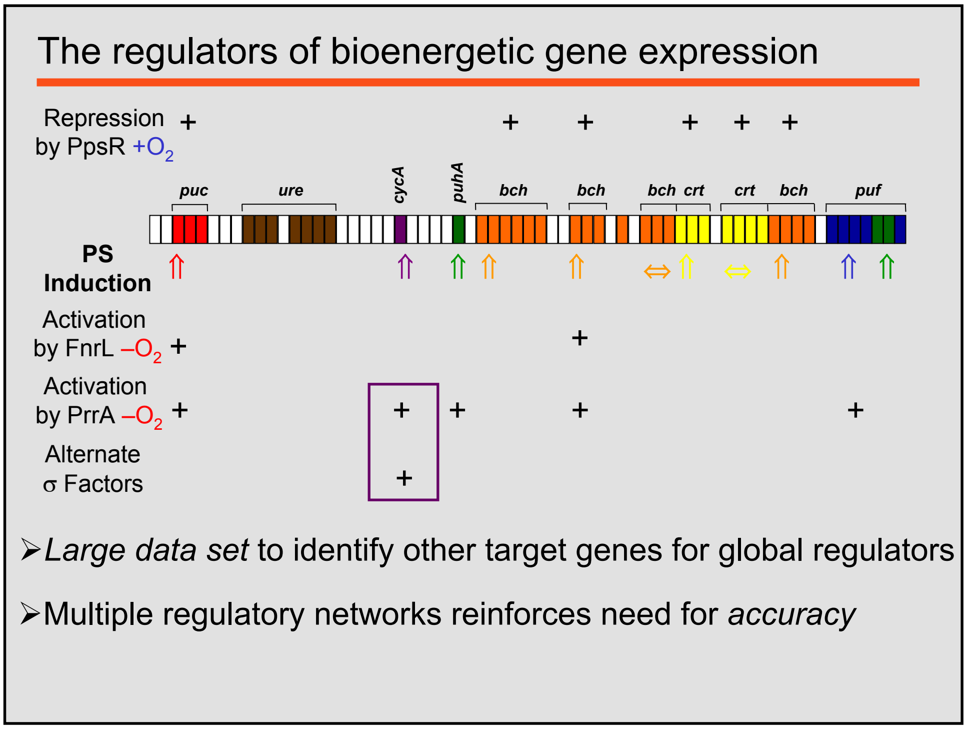


The regulation of bioenergetic gene expression

+O_2 Very low RNA "no complex"

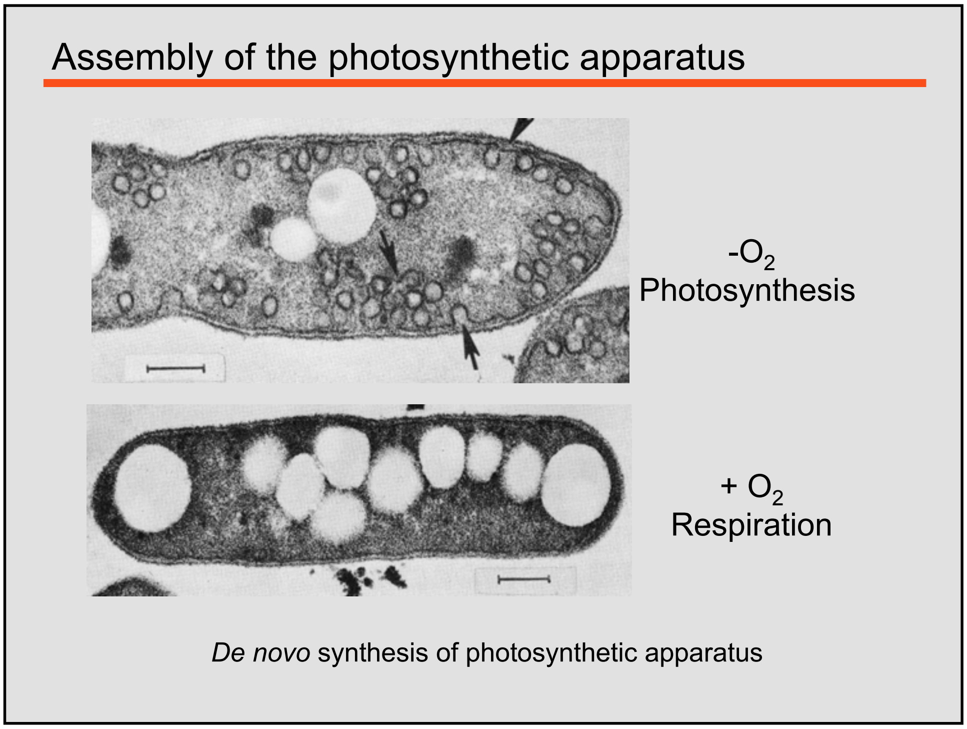
+O_2 High RNA "no complex"

- *Sensitivity* to monitor *post-transcriptional* control
- *Link* transcriptome and proteome data



The regulators of bioenergetic gene expression

- *Large data set* to identify other target genes for global regulators
- Multiple regulatory networks reinforces need for *accuracy*

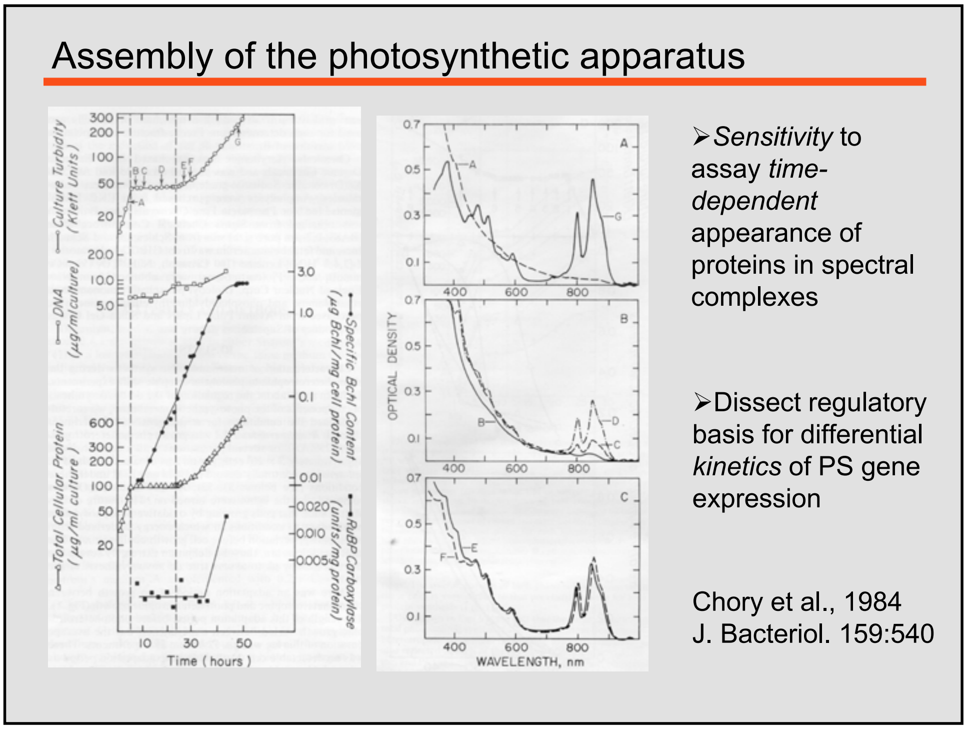


Assembly of the photosynthetic apparatus

$-O_2$
Photosynthesis

$+ O_2$
Respiration

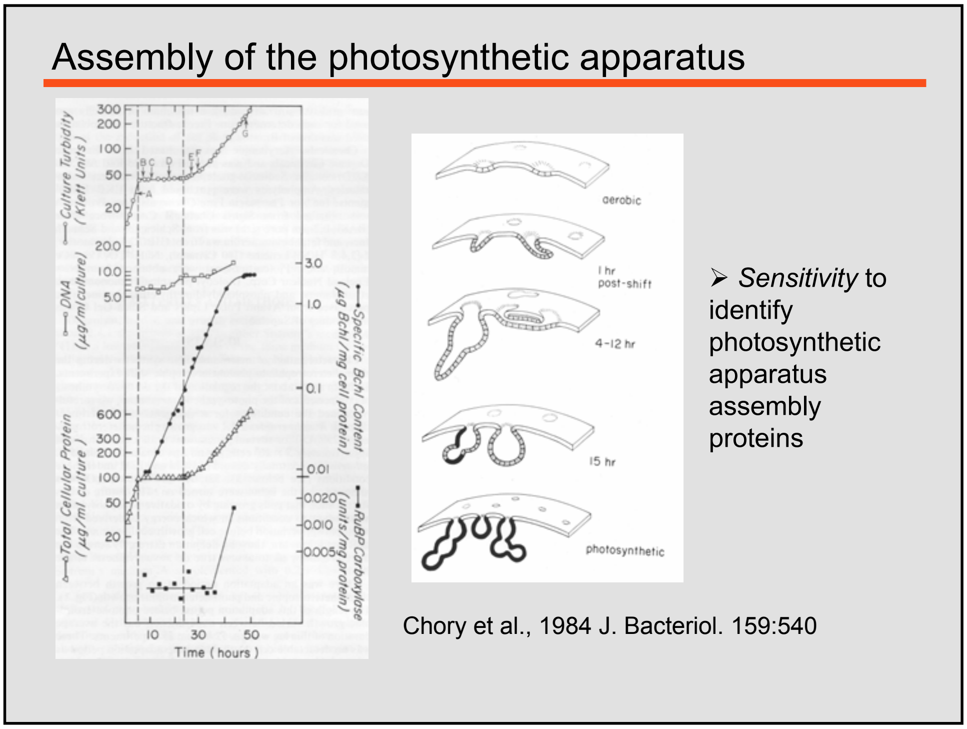
De novo synthesis of photosynthetic apparatus



Assembly of the photosynthetic apparatus

- *Sensitivity* to assay *time-dependent* appearance of proteins in spectral complexes
- Dissect regulatory basis for differential *kinetics* of PS gene expression

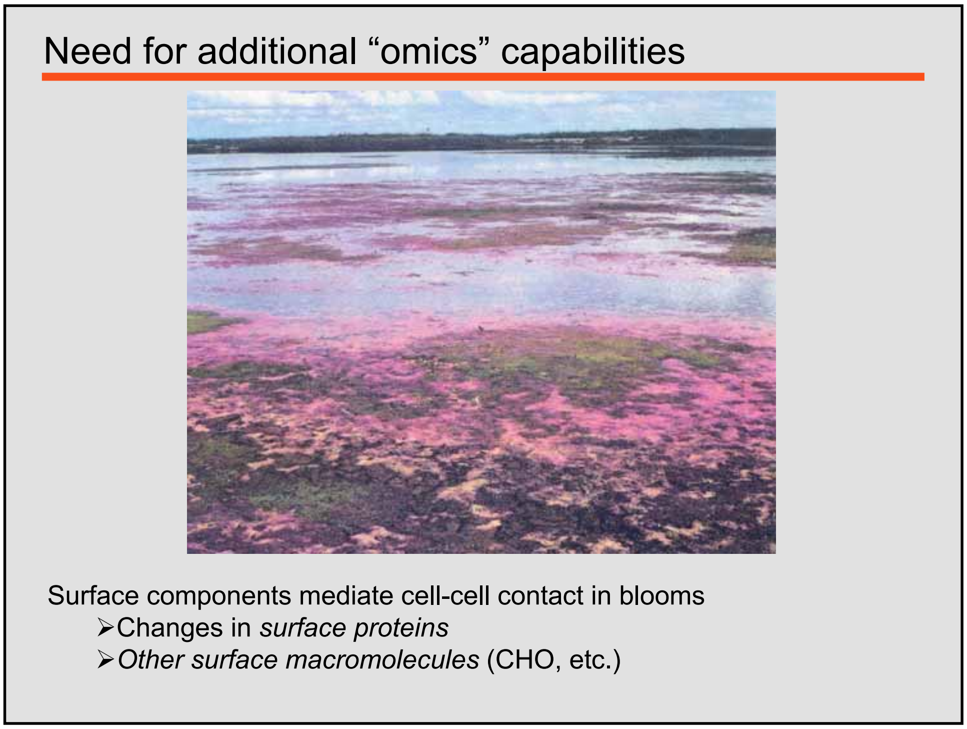
Chory et al., 1984
J. Bacteriol. 159:540



Assembly of the photosynthetic apparatus

- *Sensitivity* to identify photosynthetic apparatus assembly proteins

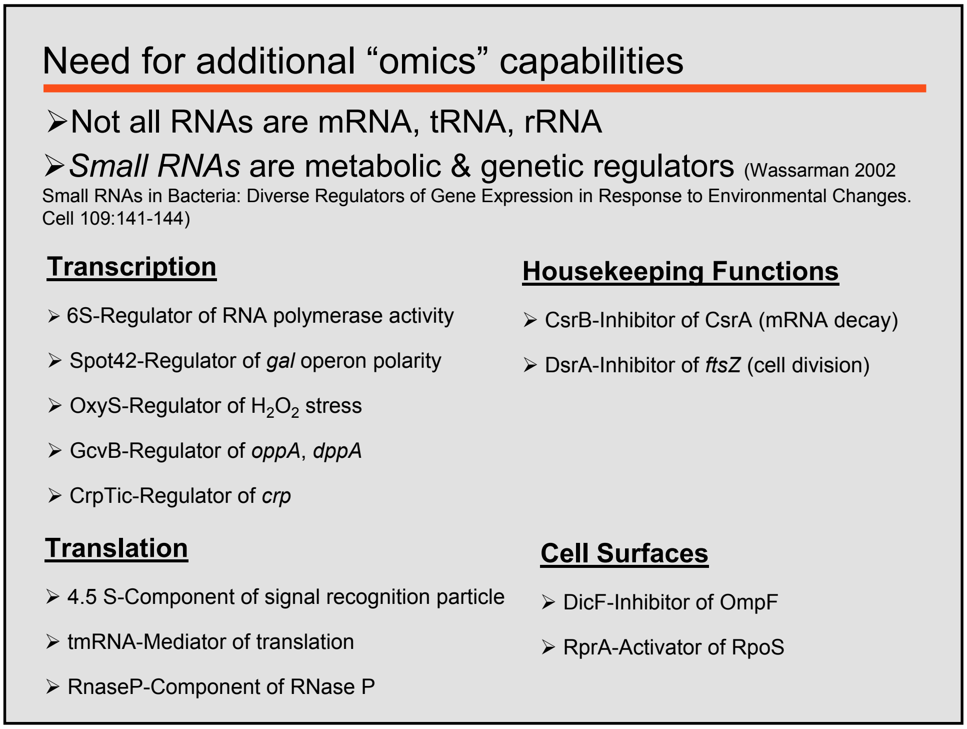
Chory et al., 1984 J. Bacteriol. 159:540



Need for additional “omics” capabilities

Surface components mediate cell-cell contact in blooms

- Changes in *surface proteins*
- *Other surface macromolecules* (CHO, etc.)



Need for additional “omics” capabilities

➢Not all RNAs are mRNA, tRNA, rRNA

➢*Small RNAs* are metabolic & genetic regulators (Wassarman 2002 Small RNAs in Bacteria: Diverse Regulators of Gene Expression in Response to Environmental Changes. Cell 109:141-144)

Transcription

➢ 6S-Regulator of RNA polymerase activity

➢ Spot42-Regulator of *gal* operon polarity

➢ OxyS-Regulator of H_2O_2 stress

➢ GcvB-Regulator of *oppA*, *dppA*

➢ CrpTic-Regulator of *crp*

Housekeeping Functions

➢ CsrB-Inhibitor of CsrA (mRNA decay)

➢ DsrA-Inhibitor of *ftsZ* (cell division)

Translation

➢ 4.5 S-Component of signal recognition particle

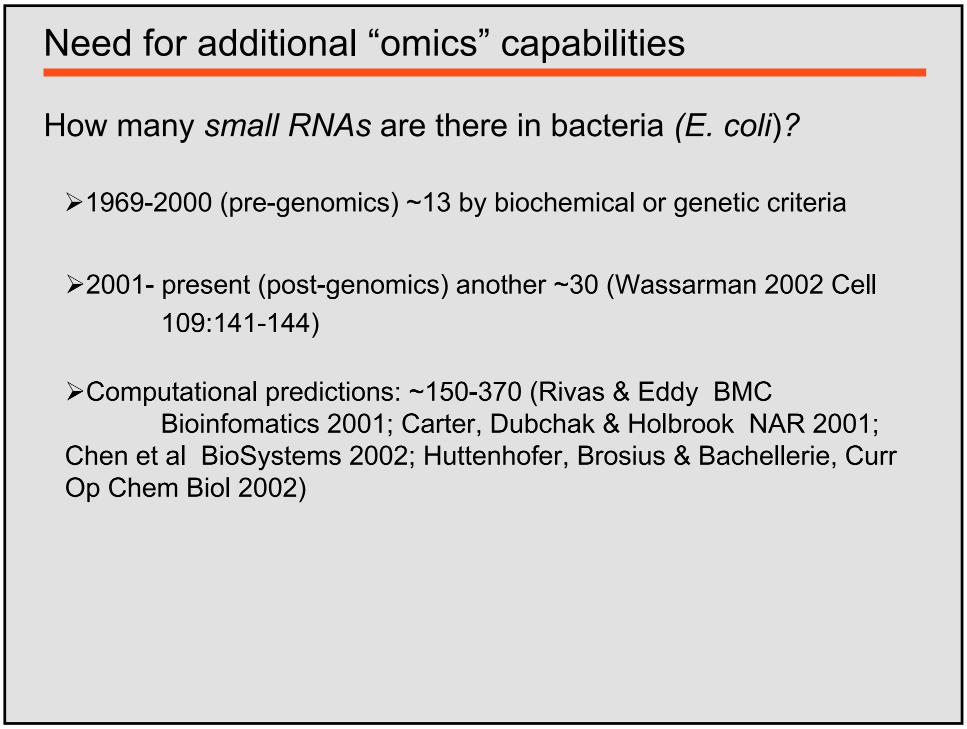
➢ tmRNA-Mediator of translation

➢ RnaseP-Component of RNase P

Cell Surfaces

➢ DicF-Inhibitor of OmpF

➢ RprA-Activator of RpoS



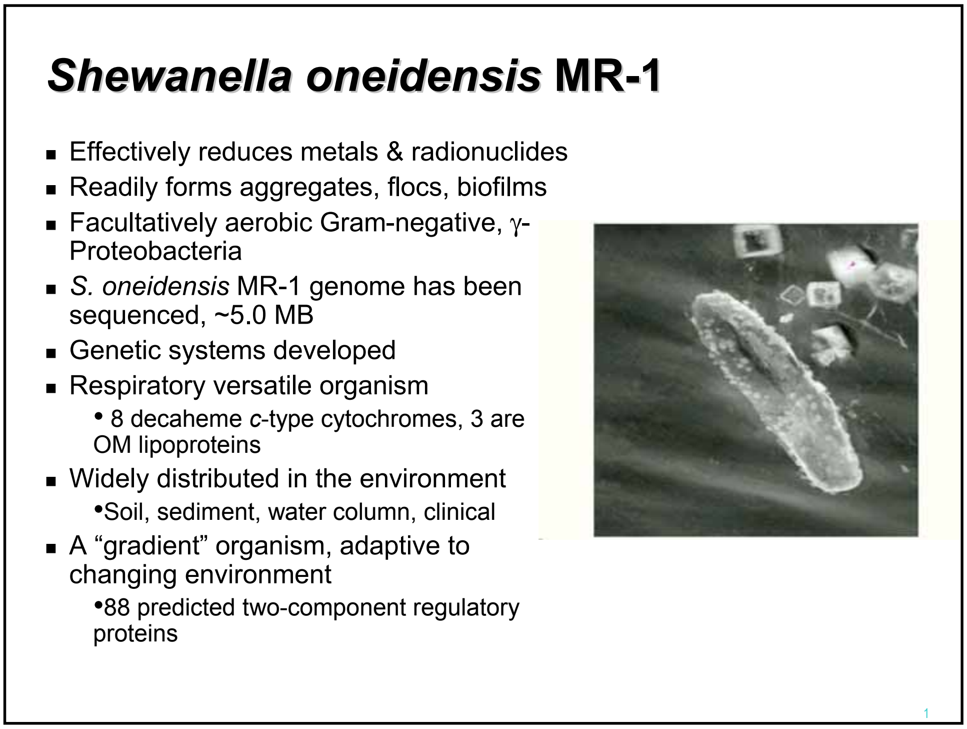
Need for additional “omics” capabilities

How many *small RNAs* are there in bacteria *(E. coli)?*

➢1969-2000 (pre-genomics) ~13 by biochemical or genetic criteria

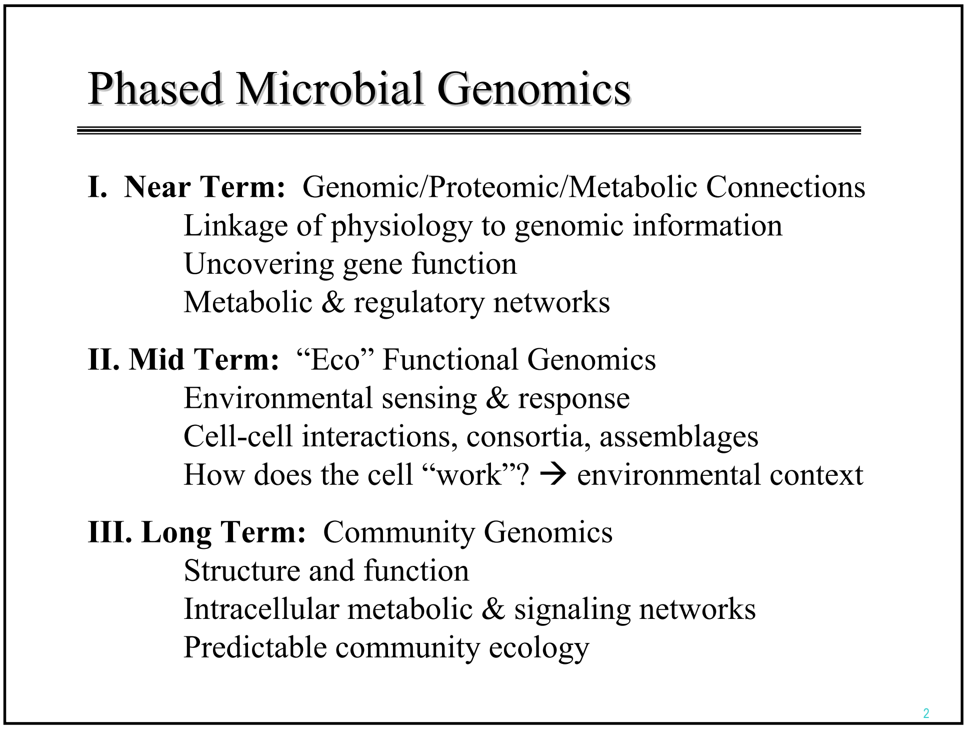
➢2001- present (post-genomics) another ~30 (Wassarman 2002 Cell 109:141-144)

➢Computational predictions: ~150-370 (Rivas & Eddy BMC Bioinfomatics 2001; Carter, Dubchak & Holbrook NAR 2001; Chen et al BioSystems 2002; Huttenhofer, Brosius & Bachellerie, Curr Op Chem Biol 2002)



Shewanella oneidensis MR-1

- Effectively reduces metals & radionuclides
- Readily forms aggregates, flocs, biofilms
- Facultatively aerobic Gram-negative, γ-Proteobacteria
- *S. oneidensis* MR-1 genome has been sequenced, ~5.0 MB
- Genetic systems developed
- Respiratory versatile organism
  - 8 decaheme *c*-type cytochromes, 3 are OM lipoproteins
- Widely distributed in the environment
  - Soil, sediment, water column, clinical
- A "gradient" organism, adaptive to changing environment
  - 88 predicted two-component regulatory proteins



Phased Microbial Genomics

I. Near Term: Genomic/Proteomic/Metabolic Connections

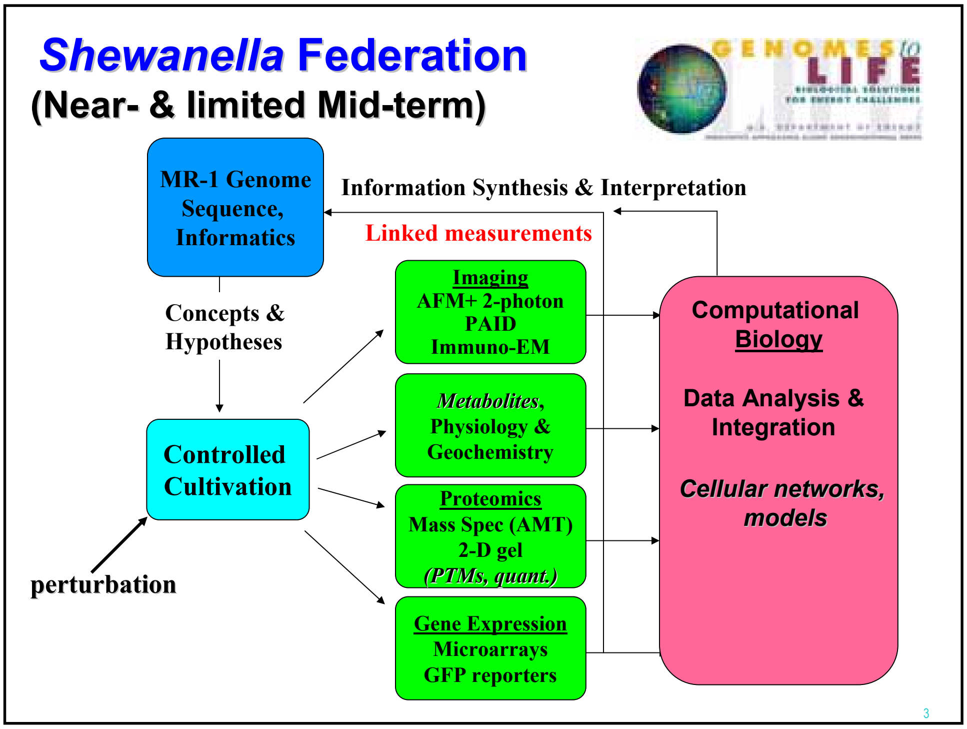
Linkage of physiology to genomic information
Uncovering gene function
Metabolic & regulatory networks

II. Mid Term: "Eco" Functional Genomics

Environmental sensing & response
Cell-cell interactions, consortia, assemblages
How does the cell "work"? → environmental context

III. Long Term: Community Genomics

Structure and function
Intracellular metabolic & signaling networks
Predictable community ecology



Shewanella Federation
(Near- & limited Mid-term)

MR-1 Genome Sequence, Informatics

Information Synthesis & Interpretation

Linked measurements

Concepts & Hypotheses

Controlled Cultivation

perturbation

Imaging
AFM+ 2-photon
PAID
Immuno-EM

Metabolites, Physiology & Geochemistry

Proteomics
Mass Spec (AMT)
2-D gel
(PTMs, quant.)

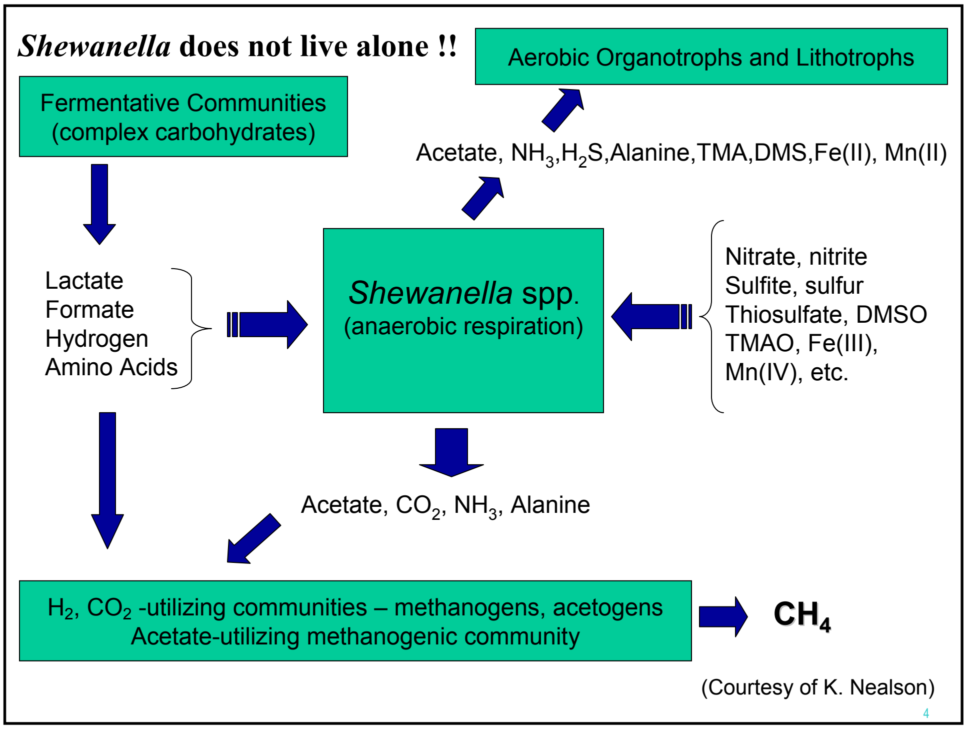
Gene Expression
Microarrays
GFP reporters

Computational **Biology**

Data Analysis & Integration

Cellular networks, models

3



***Shewanella* does not live alone !!**

Aerobic Organotrophs and Lithotrophs

Fermentative Communities (complex carbohydrates)

Acetate, NH_3,H_2S,Alanine,TMA,DMS,Fe(II), Mn(II)

Lactate
Formate
Hydrogen
Amino Acids

Shewanella spp.
(anaerobic respiration)

Nitrate, nitrite
Sulfite, sulfur
Thiosulfate, DMSO
TMAO, Fe(III),
Mn(IV), etc.

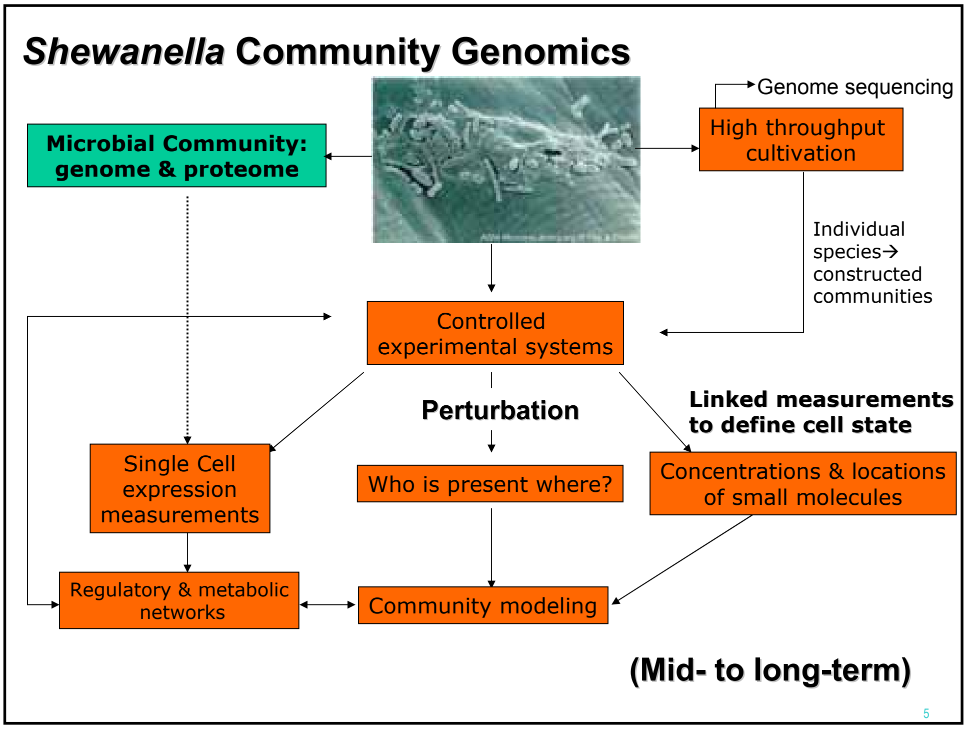
Acetate, CO_2, NH_3, Alanine

H_2, CO_2 -utilizing communities – methanogens, acetogens
Acetate-utilizing methanogenic community

CH_4

(Courtesy of K. Nealson)

4

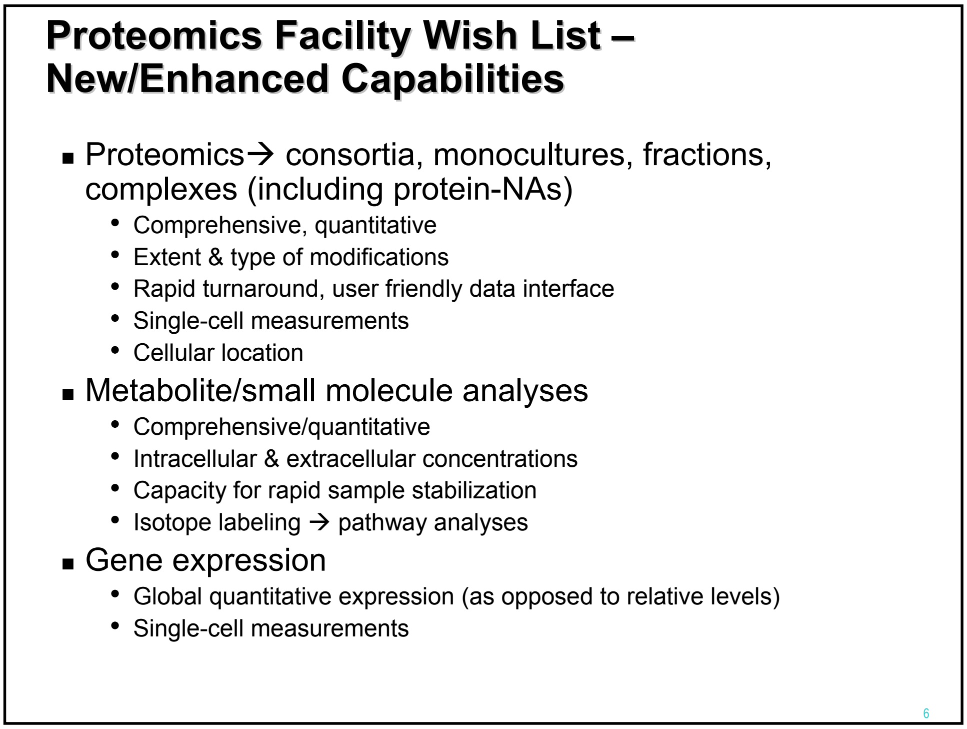


Shewanella Community Genomics

Genome sequencing
High throughput cultivation
Microbial Community: genome & proteome
Individual species→ constructed communities
Controlled experimental systems
Perturbation
Linked measurements to define cell state
Single Cell expression measurements
Who is present where?
Concentrations & locations of small molecules
Regulatory & metabolic networks
Community modeling

(Mid- to long-term)

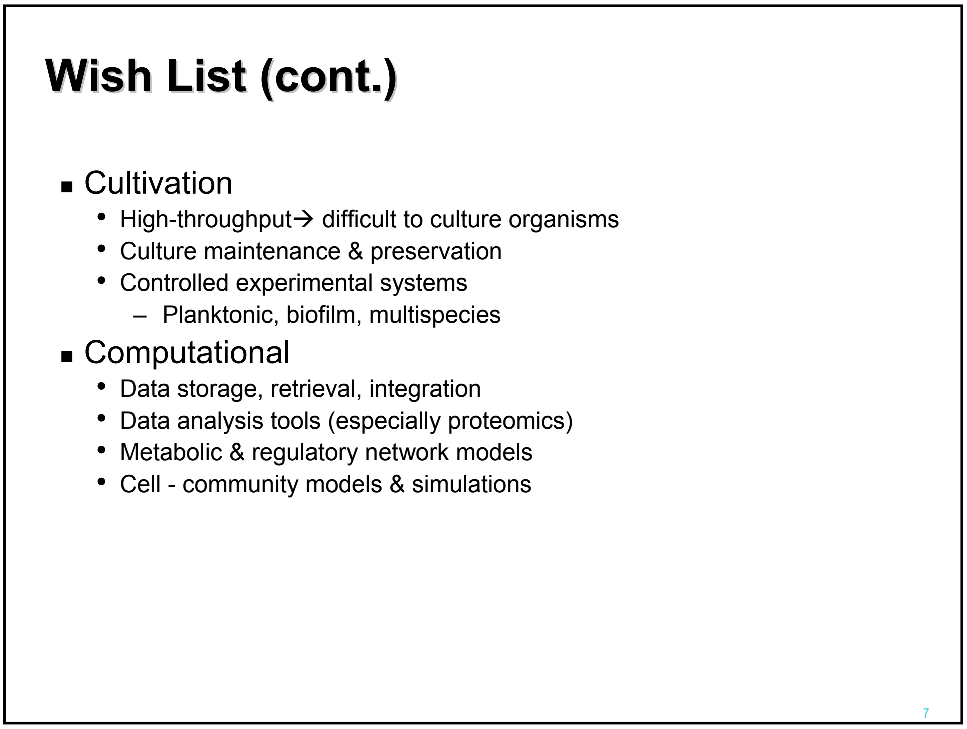
5



Proteomics Facility Wish List – New/Enhanced Capabilities

- Proteomics→ consortia, monocultures, fractions, complexes (including protein-NAs)
  - Comprehensive, quantitative
  - Extent & type of modifications
  - Rapid turnaround, user friendly data interface
  - Single-cell measurements
  - Cellular location
- Metabolite/small molecule analyses
  - Comprehensive/quantitative
  - Intracellular & extracellular concentrations
  - Capacity for rapid sample stabilization
  - Isotope labeling → pathway analyses
- Gene expression
  - Global quantitative expression (as opposed to relative levels)
  - Single-cell measurements

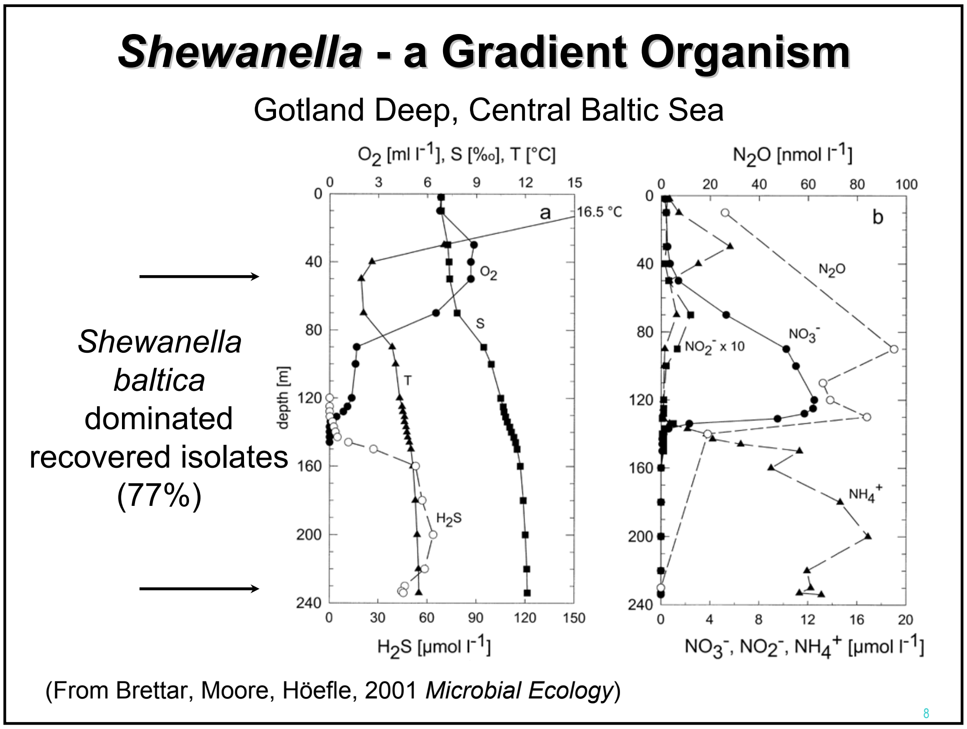
6



Wish List (cont.)

- Cultivation
  - High-throughput→ difficult to culture organisms
  - Culture maintenance & preservation
  - Controlled experimental systems
    - Planktonic, biofilm, multispecies
- Computational
  - Data storage, retrieval, integration
  - Data analysis tools (especially proteomics)
  - Metabolic & regulatory network models
  - Cell - community models & simulations

7



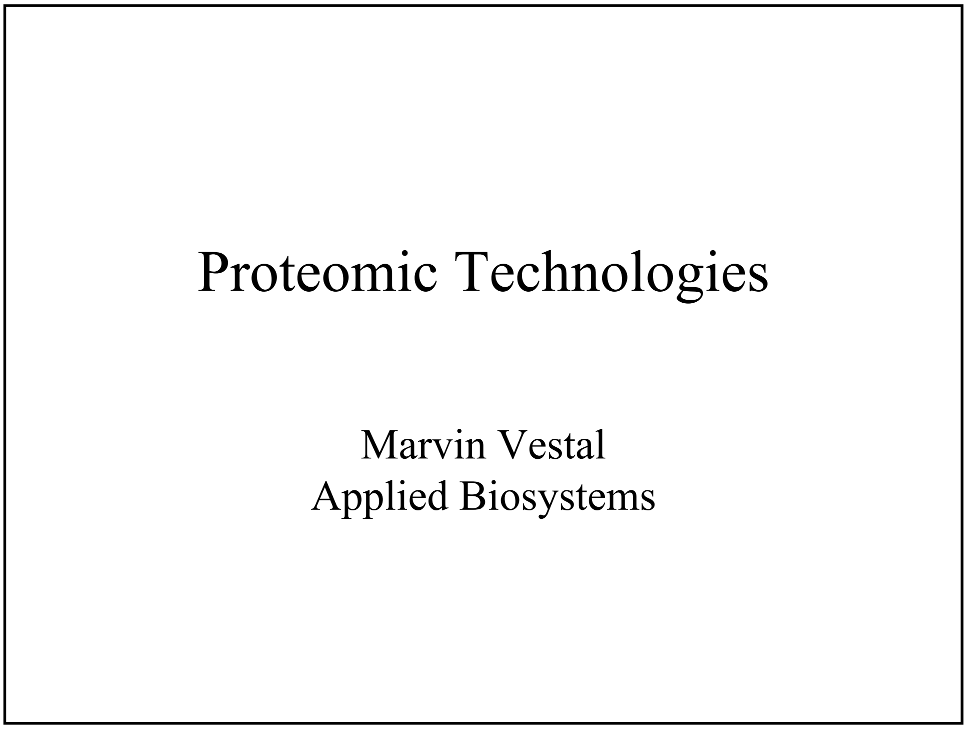
Shewanella - a Gradient Organism

Gotland Deep, Central Baltic Sea

Shewanella baltica dominated recovered isolates (77%)

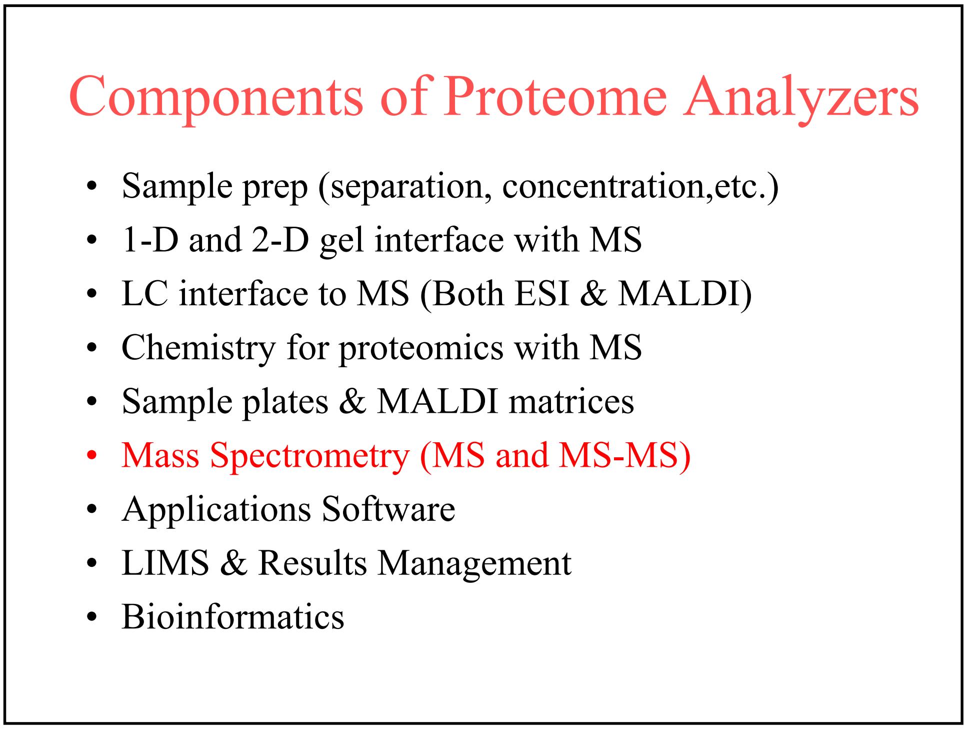
(From Brettar, Moore, Höefle, 2001 *Microbial Ecology*)

8



Proteomic Technologies

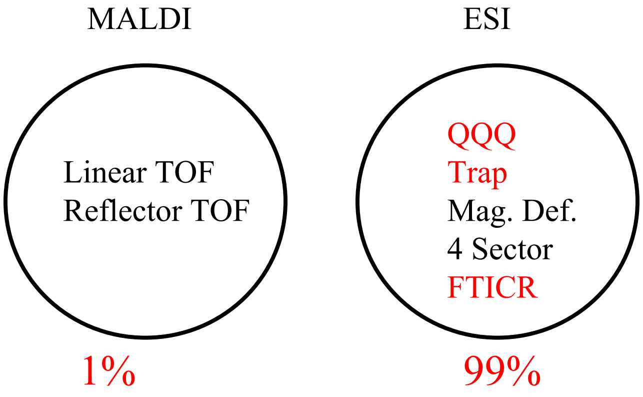
Marvin Vestal
Applied Biosystems



Components of Proteome Analyzers

- Sample prep (separation, concentration,etc.)
- 1-D and 2-D gel interface with MS
- LC interface to MS (Both ESI & MALDI)
- Chemistry for proteomics with MS
- Sample plates & MALDI matrices
- Mass Spectrometry (MS and MS-MS)
- Applications Software
- LIMS & Results Management
- Bioinformatics

In the beginning (ca.1990)



MALDI

Linear TOF
Reflector TOF

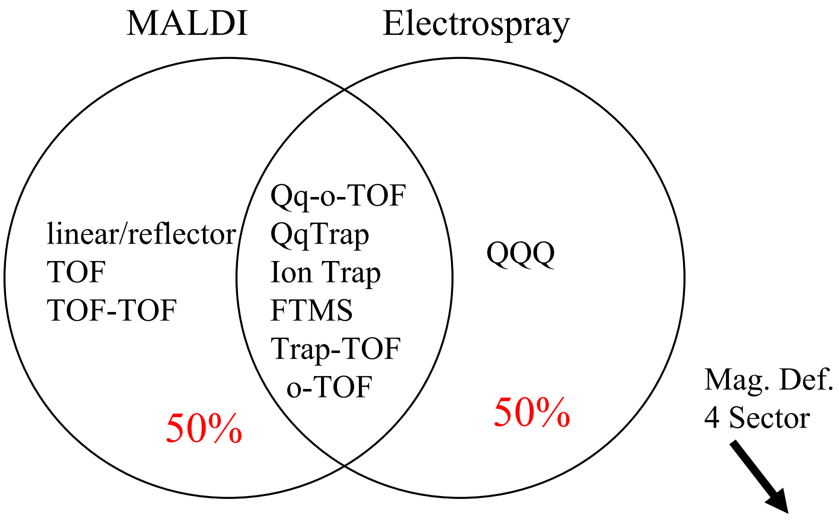
1%

ESI

QQQ
Trap
Mag. Def.
4 Sector
FTICR

99%

Now (2003)



MALDI

linear/reflector
TOF
TOF-TOF

50%

Qq-o-TOF
QqTrap
Ion Trap
FTMS
Trap-TOF
o-TOF

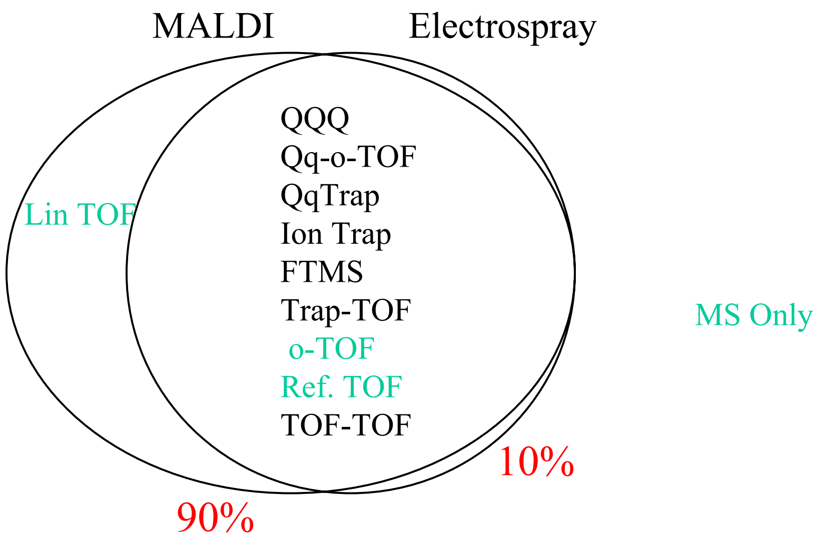
Electrospray

QQQ

50%

Mag. Def.
4 Sector

Will Be (2006?)



MALDI

Electrospray

QQQ
Qq-o-TOF
QqTrap
Ion Trap
FTMS
Trap-TOF
o-TOF
Ref. TOF
TOF-TOF

Lin TOF

MS Only

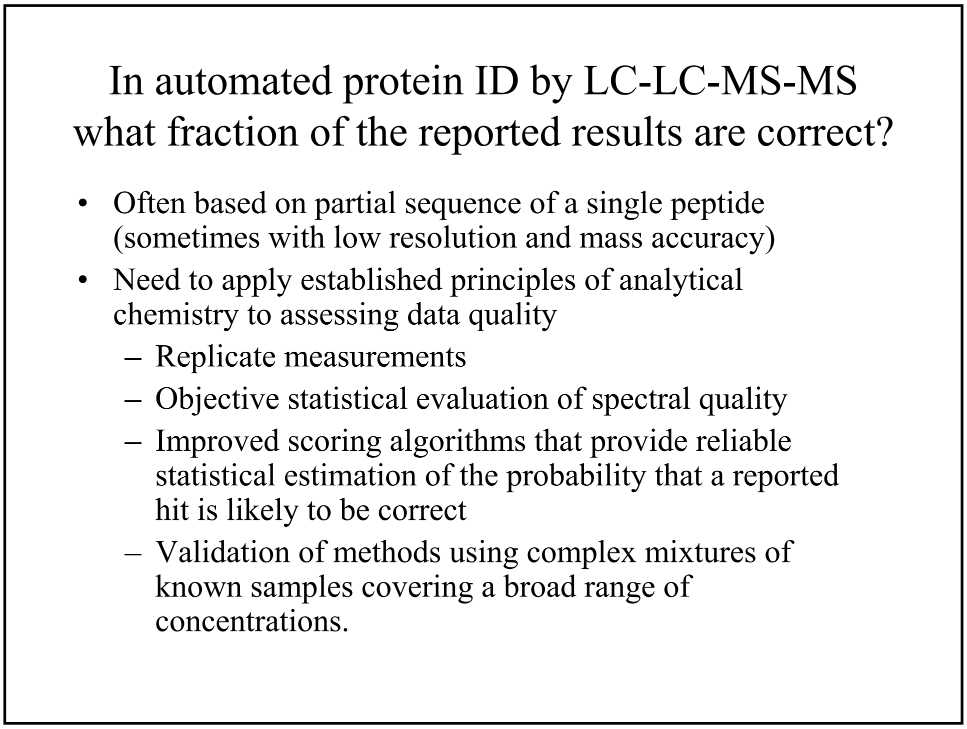
10%

90%

Advantages of LC Coupled to ESI & MALDI for Proteomics:

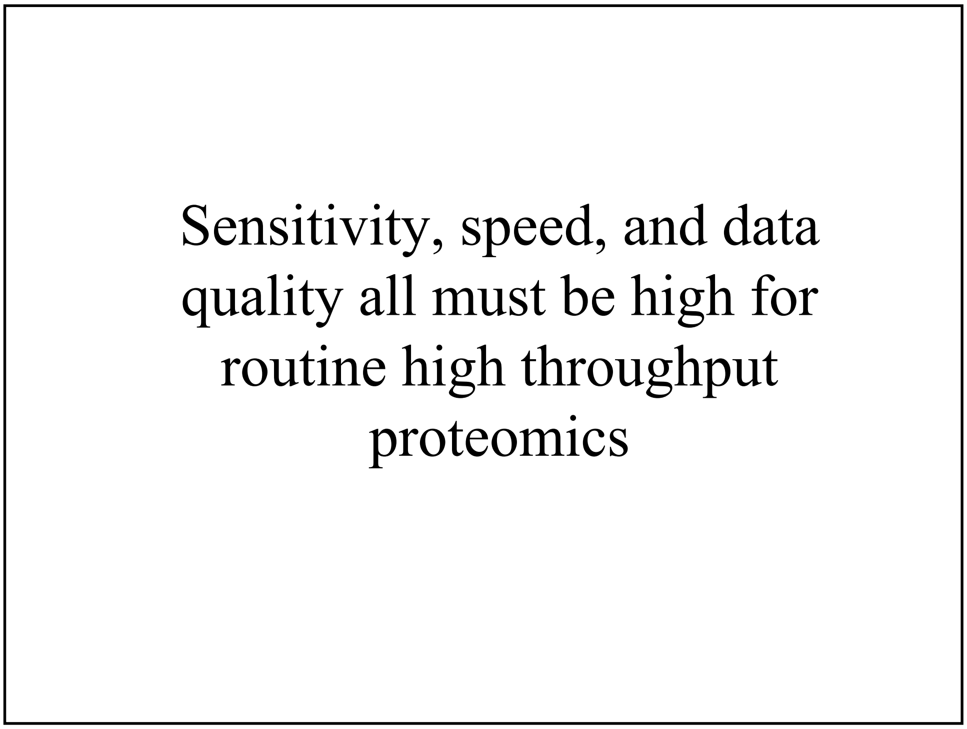
- LC ESI
  - Direct coupling of LC to MS
  - Fast – lots of MS and MS/MS
  - Accepted MS/MS ionization mode

- LC MALDI
  - Sample in solid state
  - Not time-limited for MS/MS
  - Analysis can be faster or slower than separation
  - More sophisticated workflows
  - Fast – lots of MS and MS/MS
  - Results dependent acquisition stop criteria

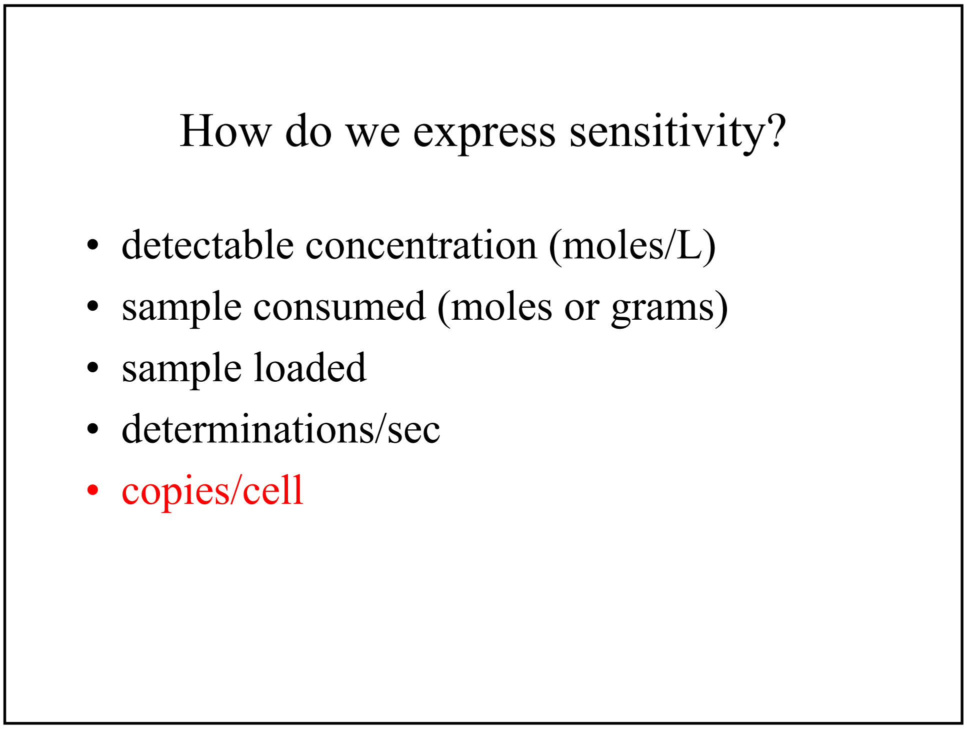


In automated protein ID by LC-LC-MS-MS what fraction of the reported results are correct?

- Often based on partial sequence of a single peptide (sometimes with low resolution and mass accuracy)
- Need to apply established principles of analytical chemistry to assessing data quality
  - Replicate measurements
  - Objective statistical evaluation of spectral quality
  - Improved scoring algorithms that provide reliable statistical estimation of the probability that a reported hit is likely to be correct
  - Validation of methods using complex mixtures of known samples covering a broad range of concentrations.

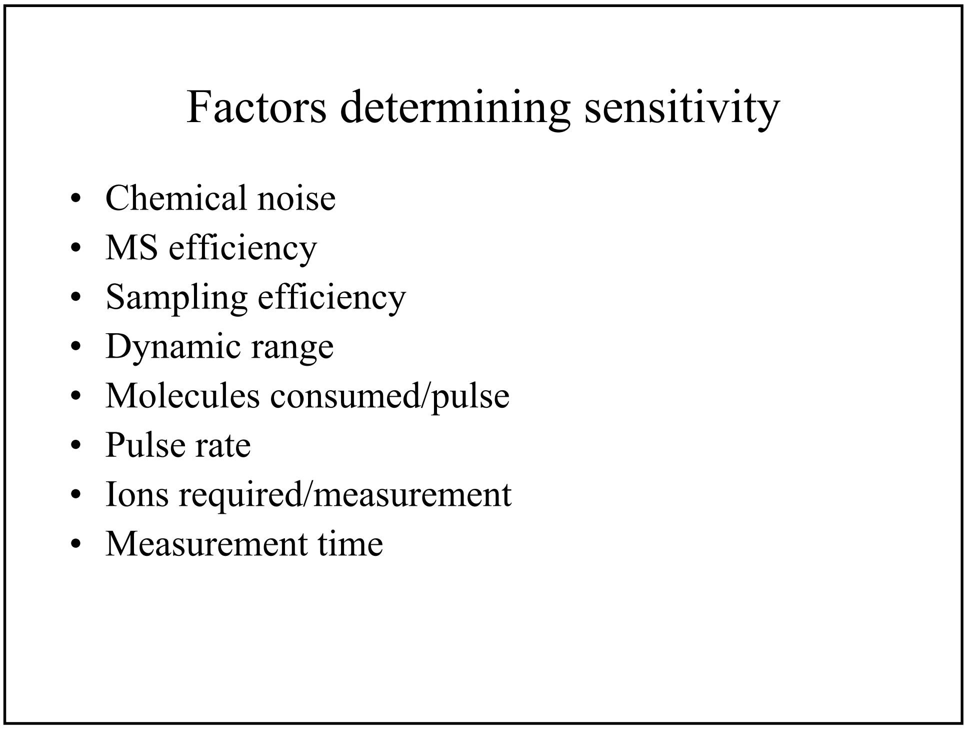


Sensitivity, speed, and data quality all must be high for routine high throughput proteomics



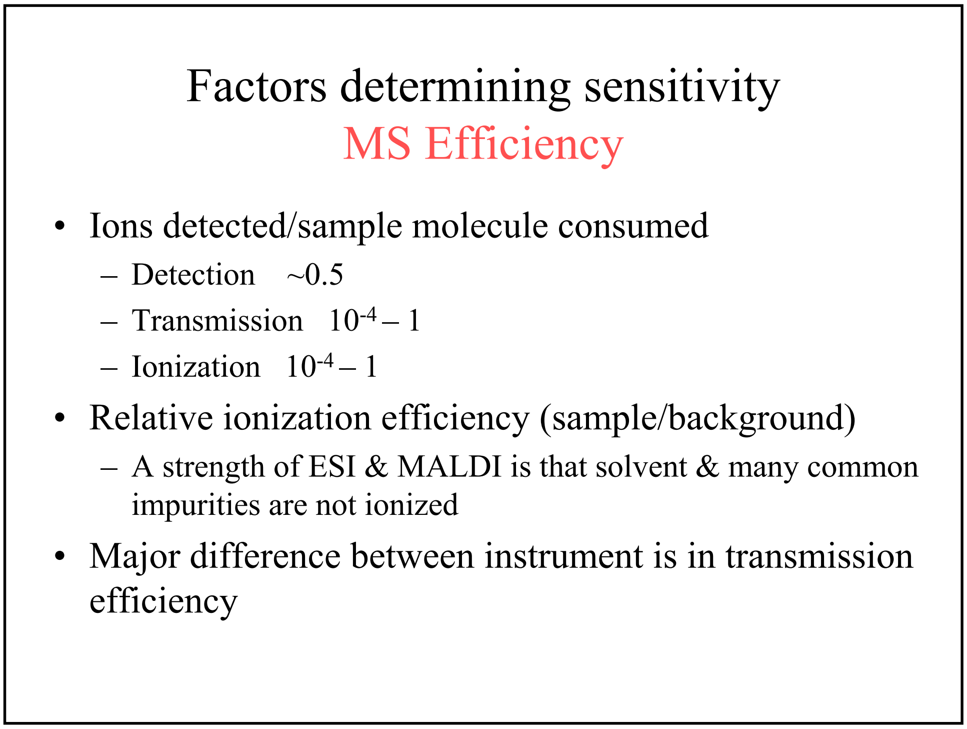
How do we express sensitivity?

- detectable concentration (moles/L)
- sample consumed (moles or grams)
- sample loaded
- determinations/sec
- copies/cell



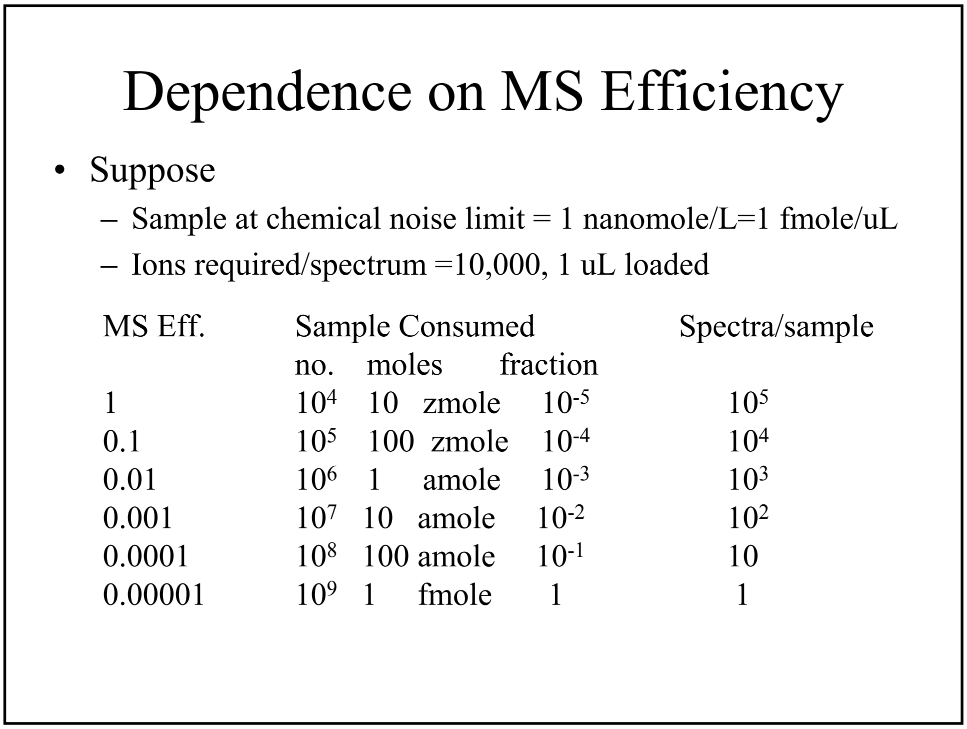
Factors determining sensitivity

- Chemical noise
- MS efficiency
- Sampling efficiency
- Dynamic range
- Molecules consumed/pulse
- Pulse rate
- Ions required/measurement
- Measurement time



Factors determining sensitivity
MS Efficiency

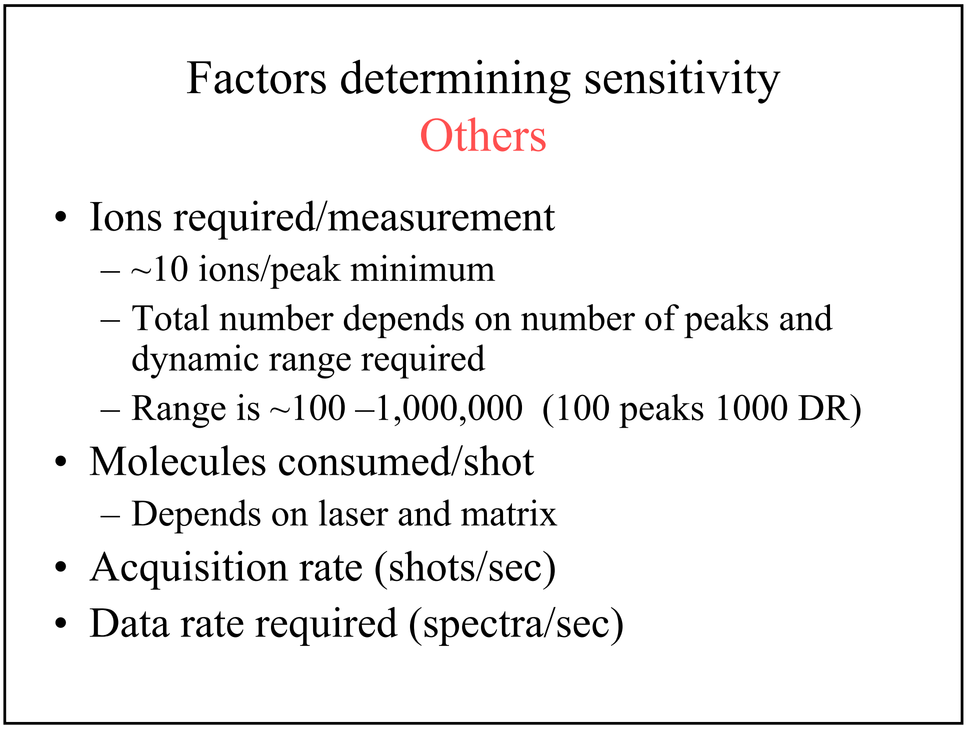
- Ions detected/sample molecule consumed
  - Detection ~0.5
  - Transmission $10^{-4} – 1$
  - Ionization $10^{-4} – 1$
- Relative ionization efficiency (sample/background)
  - A strength of ESI & MALDI is that solvent & many common impurities are not ionized
- Major difference between instrument is in transmission efficiency



Dependence on MS Efficiency

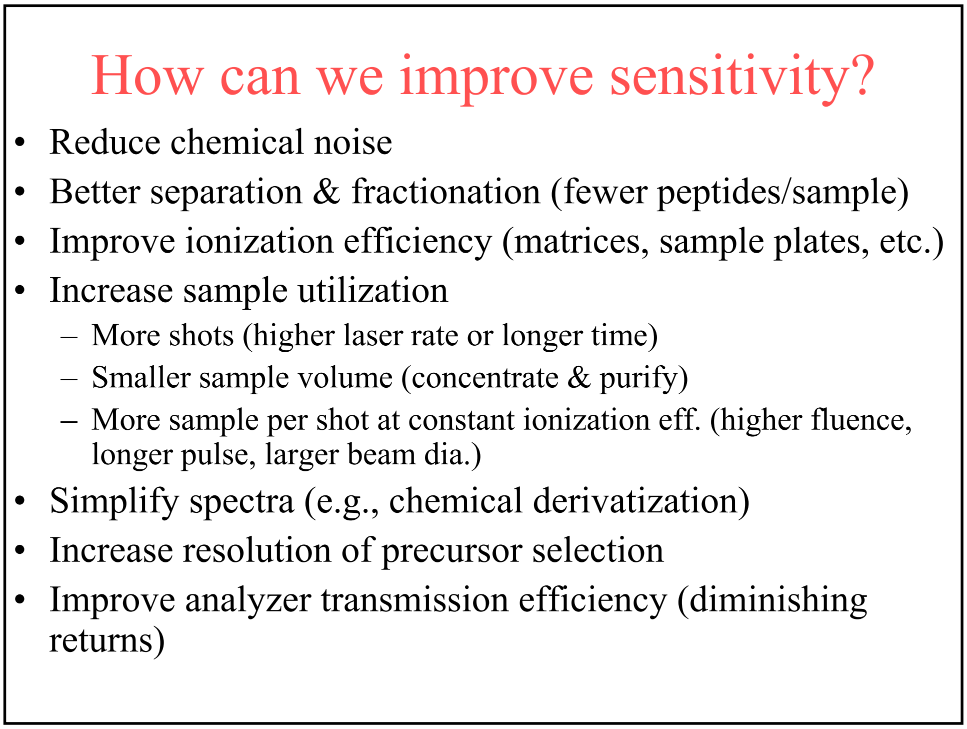
- Suppose
  - Sample at chemical noise limit = 1 nanomole/L=1 fmole/uL
  - Ions required/spectrum =10,000, 1 uL loaded

MS Eff.	Sample Consumed			Spectra/sample
	no.	moles	fraction	
1	10^{4}	10 zmole	10^{-5}	10^{5}
0.1	10^{5}	100 zmole	10^{-4}	10^{4}
0.01	10^{6}	1 amole	10^{-3}	10^{3}
0.001	10^{7}	10 amole	10^{-2}	10^{2}
0.0001	10^{8}	100 amole	10^{-1}	10
0.00001	10^{9}	1 fmole	1	1



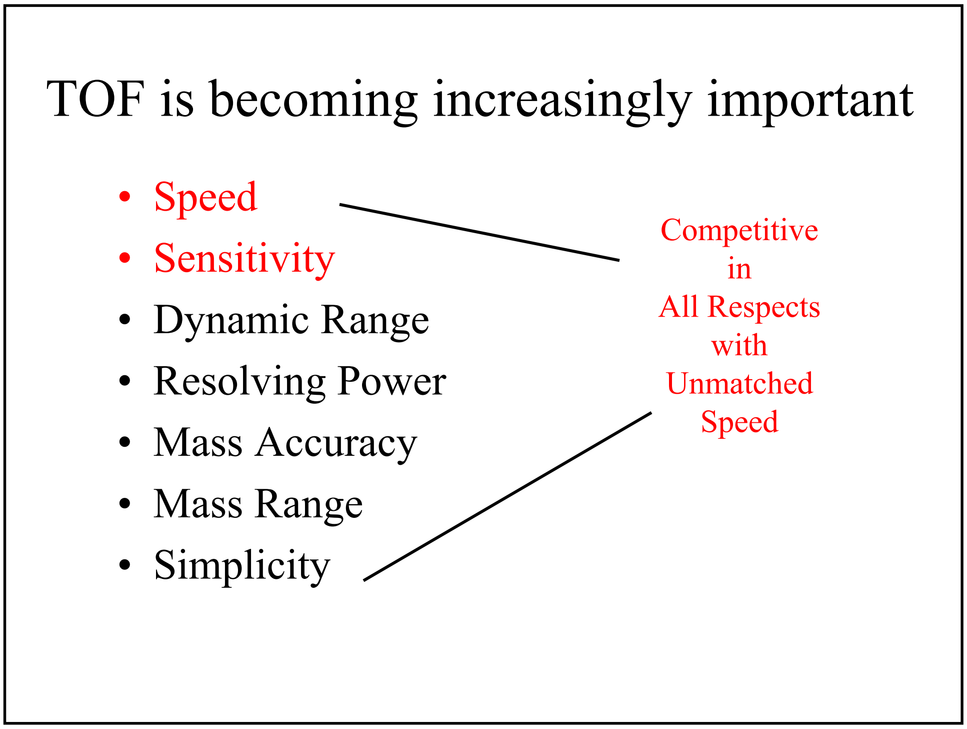
Factors determining sensitivity
Others

- Ions required/measurement
  - ~10 ions/peak minimum
  - Total number depends on number of peaks and dynamic range required
  - Range is ~100 –1,000,000 (100 peaks 1000 DR)
- Molecules consumed/shot
  - Depends on laser and matrix
- Acquisition rate (shots/sec)
- Data rate required (spectra/sec)



How can we improve sensitivity?

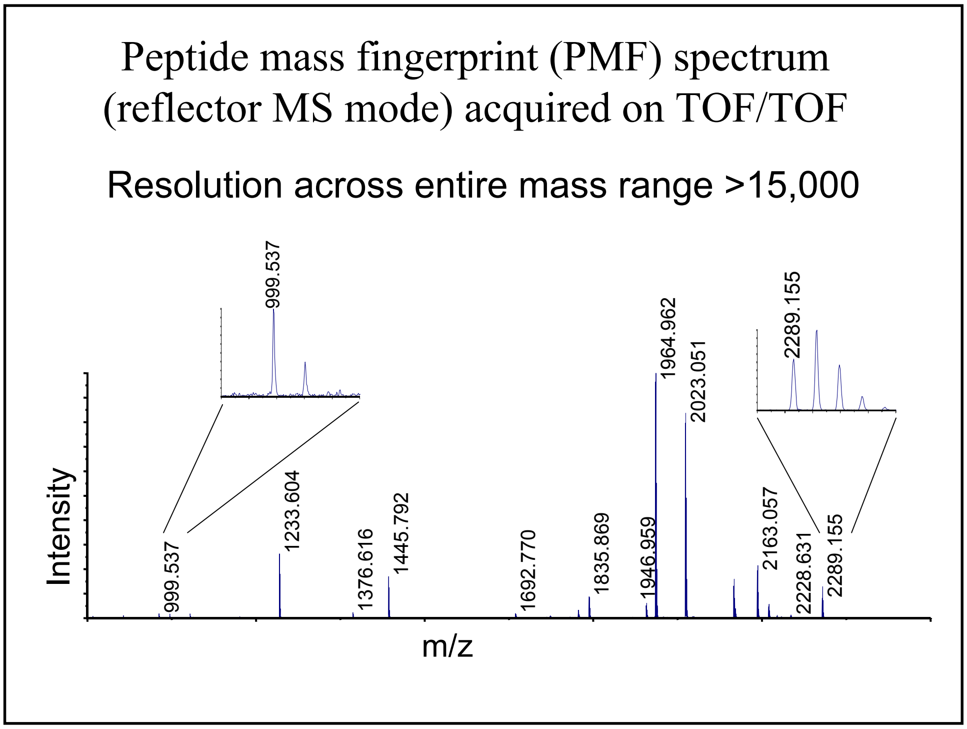
- Reduce chemical noise
- Better separation & fractionation (fewer peptides/sample)
- Improve ionization efficiency (matrices, sample plates, etc.)
- Increase sample utilization
  - More shots (higher laser rate or longer time)
  - Smaller sample volume (concentrate & purify)
  - More sample per shot at constant ionization eff. (higher fluence, longer pulse, larger beam dia.)
- Simplify spectra (e.g., chemical derivatization)
- Increase resolution of precursor selection
- Improve analyzer transmission efficiency (diminishing returns)



TOF is becoming increasingly important

- Speed
- Sensitivity
- Dynamic Range
- Resolving Power
- Mass Accuracy
- Mass Range
- Simplicity

Competitive
in
All Respects
with
Unmatched
Speed



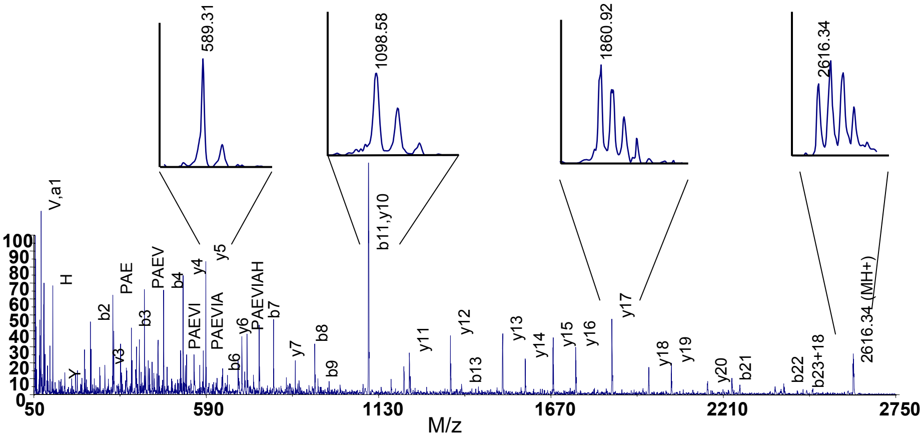
Peptide mass fingerprint (PMF) spectrum (reflector MS mode) acquired on TOF/TOF

Resolution across entire mass range >15,000

MS/MS spectrum
precursor mass 2616.3

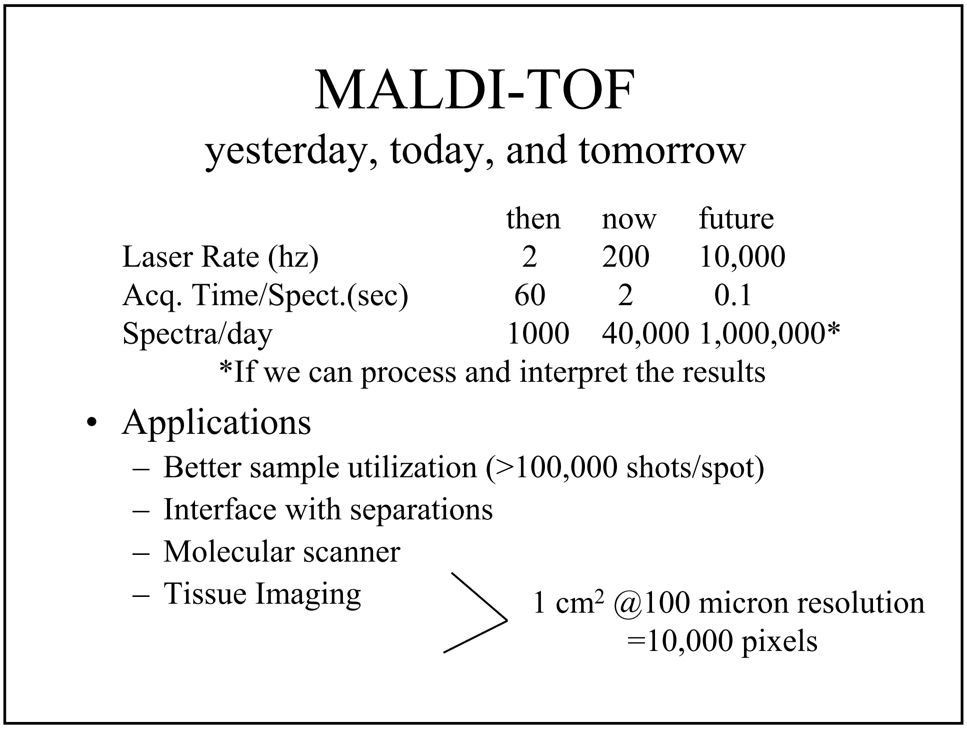
well D3

VQQTIADIASAYEQPAEVIAHYAK



Increasing laser rate improves results in many ways

- Higher quality spectra and more spectra/sample *better use of sample*
  - S/N, dynamic range, mass accuracy
  - Improved sensitivity for low abundance peptides
- Makes applications of other features practical
  - Surface imaging
  - Precursor scanning
  - Interface to LC & Molecular Scanner, etc.
- Higher throughput *more samples*



MALDI-TOF

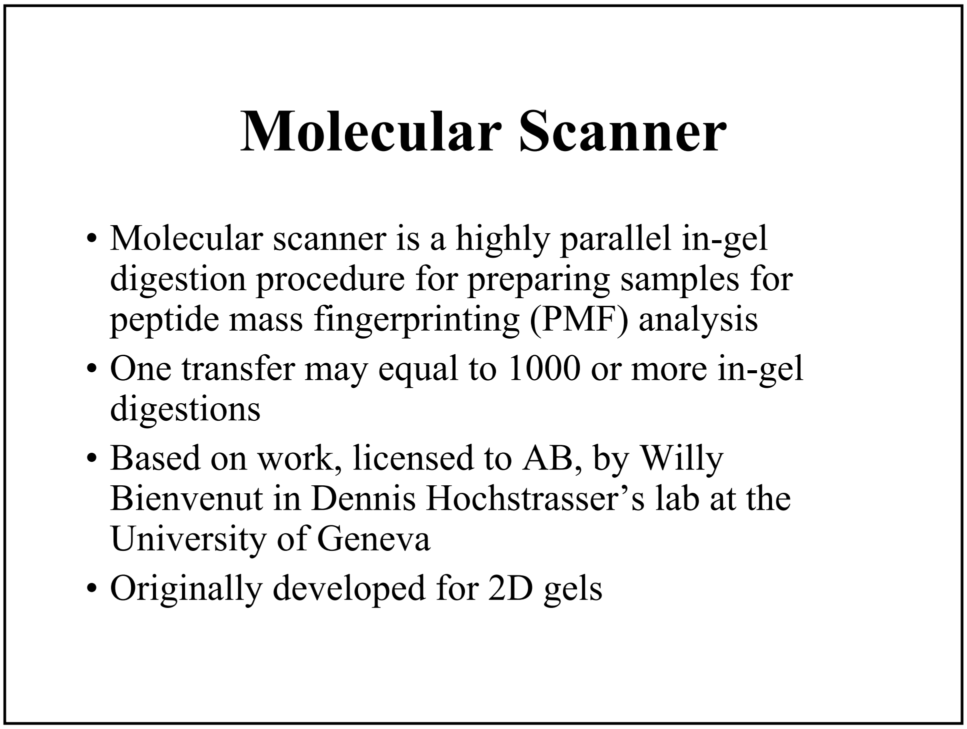
yesterday, today, and tomorrow

	then	now	future
Laser Rate (hz)	2	200	10,000
Acq. Time/Spect.(sec)	60	2	0.1
Spectra/day	1000	40,000	1,000,000*

*If we can process and interpret the results

- Applications
  - Better sample utilization (>100,000 shots/spot)
  - Interface with separations
  - Molecular scanner
  - Tissue Imaging

1 cm^2 @100 micron resolution
=10,000 pixels

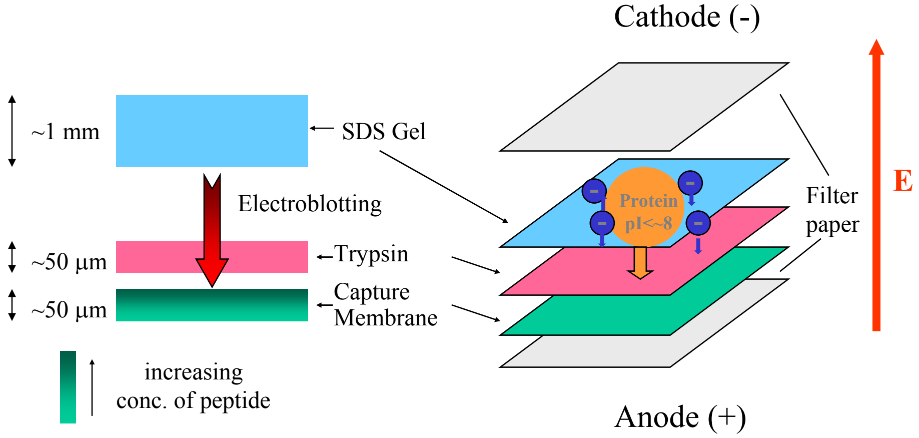


Molecular Scanner

- Molecular scanner is a highly parallel in-gel digestion procedure for preparing samples for peptide mass fingerprinting (PMF) analysis
- One transfer may equal to 1000 or more in-gel digestions
- Based on work, licensed to AB, by Willy Bienvenut in Dennis Hochstrasser’s lab at the University of Geneva
- Originally developed for 2D gels

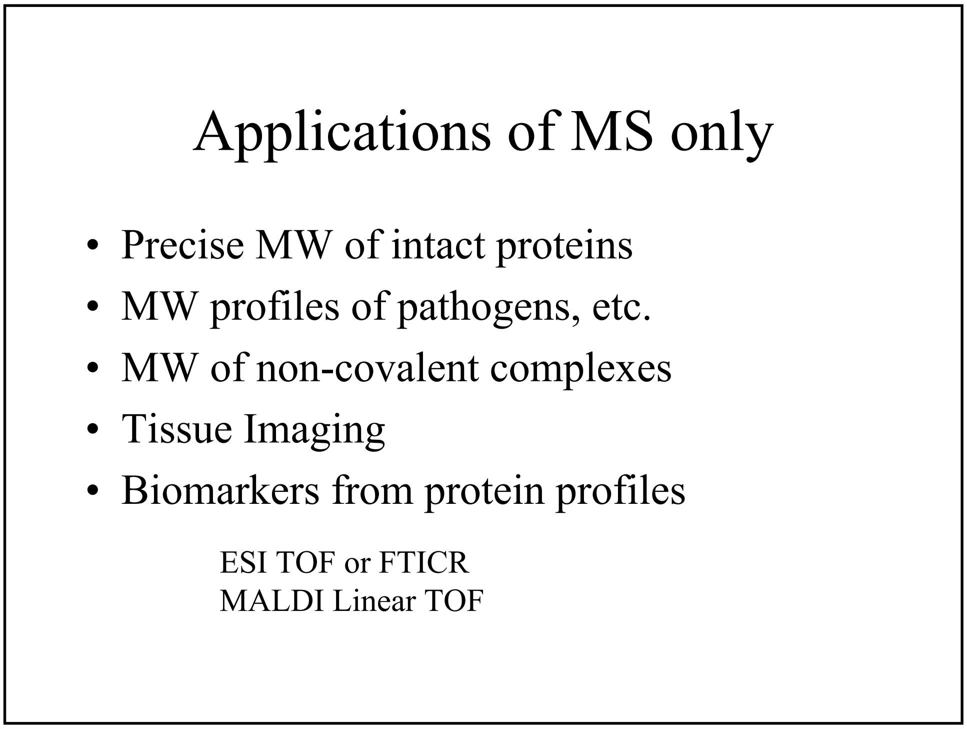
Digestion with Electroblotting

Slide from T. Nadler



Determinations needed

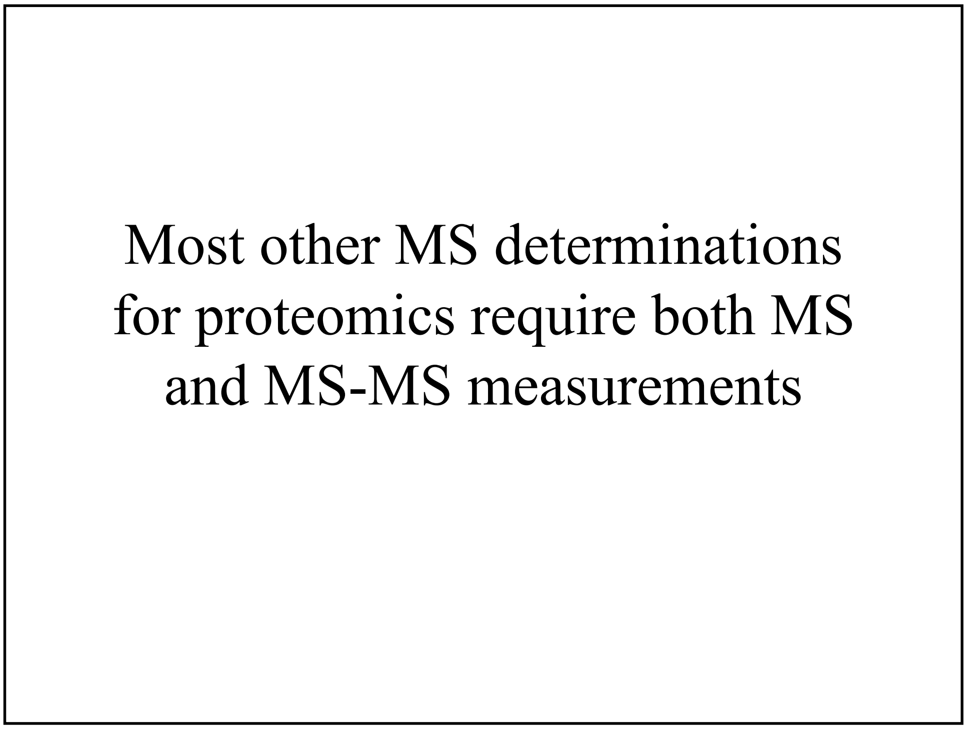
- Identification - correlation with gene product and databases of knowns
- Quantification- absolute or relative, all or selected set
- Differential expression
- Modification- splicing, processing, phosphorylation, glycosylation, etc.
- Association- non-covalent interactions
- Sequence - how does it differ from expected?



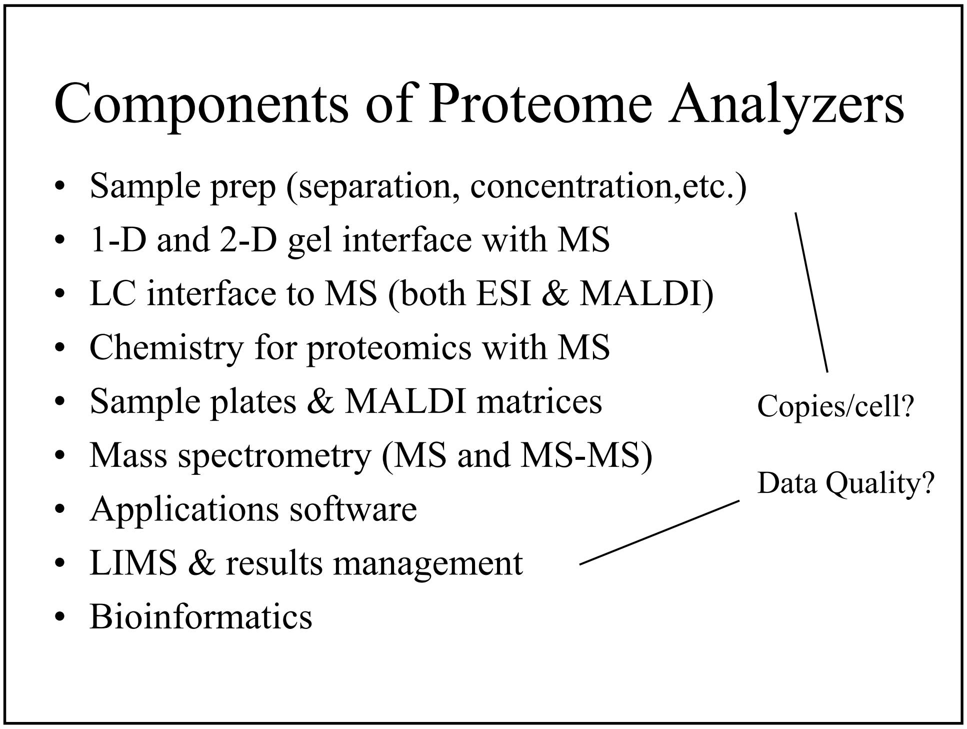
Applications of MS only

- Precise MW of intact proteins
- MW profiles of pathogens, etc.
- MW of non-covalent complexes
- Tissue Imaging
- Biomarkers from protein profiles

ESI TOF or FTICR
MALDI Linear TOF



Most other MS determinations for proteomics require both MS and MS-MS measurements

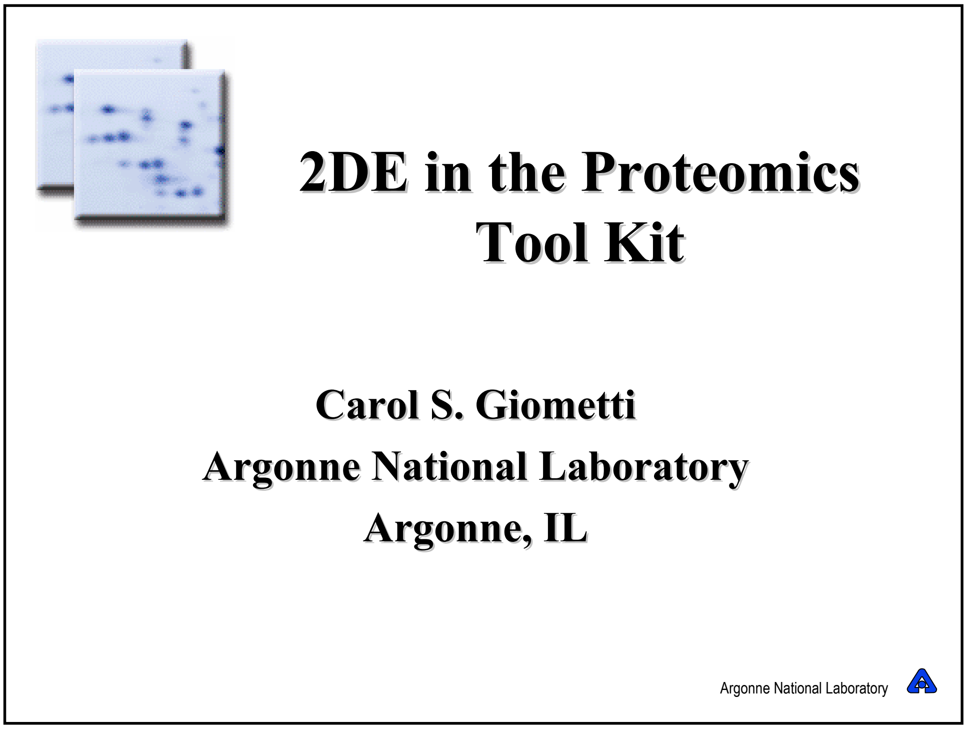


Components of Proteome Analyzers

- Sample prep (separation, concentration,etc.)
- 1-D and 2-D gel interface with MS
- LC interface to MS (both ESI & MALDI)
- Chemistry for proteomics with MS
- Sample plates & MALDI matrices
- Mass spectrometry (MS and MS-MS)
- Applications software
- LIMS & results management
- Bioinformatics

Copies/cell?

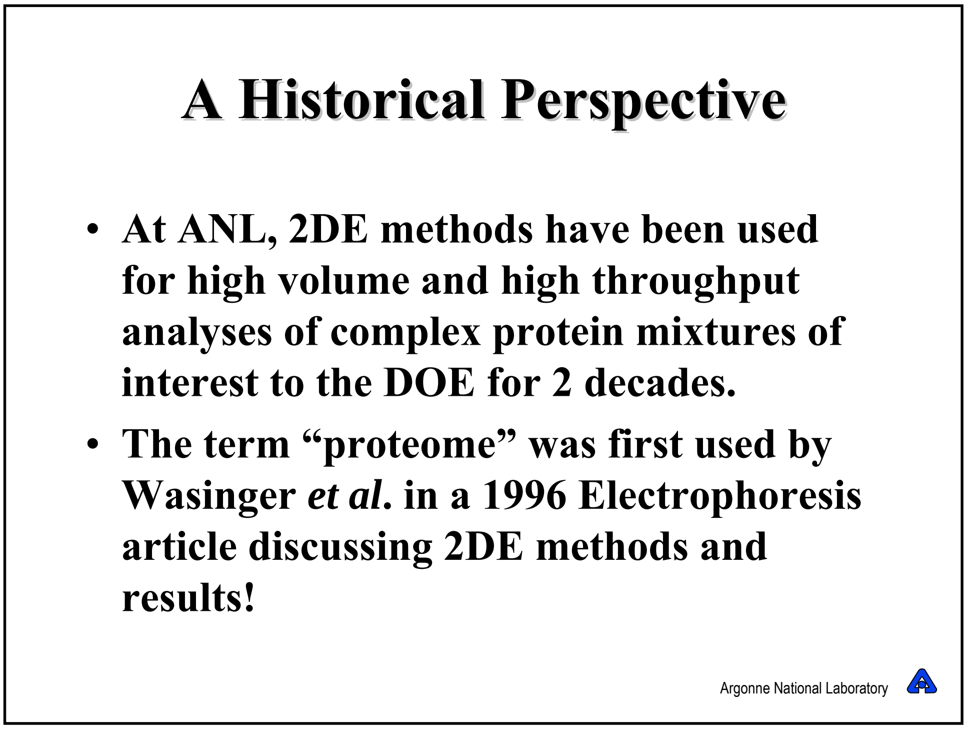
Data Quality?



2DE in the Proteomics Tool Kit

Carol S. Giometti
Argonne National Laboratory
Argonne, IL

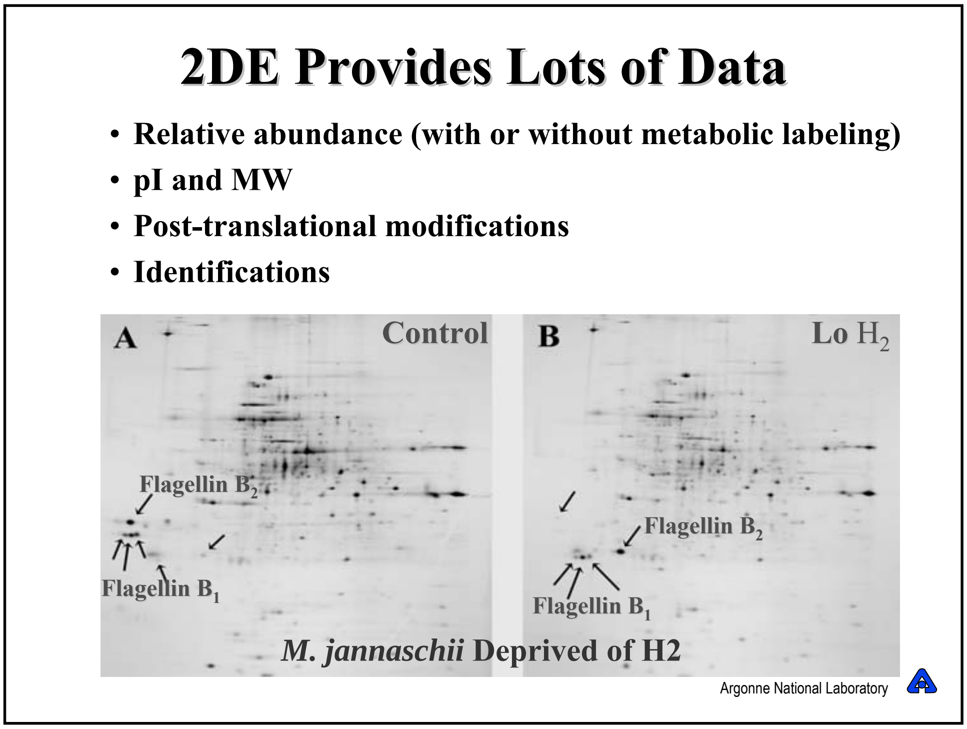
Argonne National Laboratory



A Historical Perspective

- At ANL, 2DE methods have been used for high volume and high throughput analyses of complex protein mixtures of interest to the DOE for 2 decades.
- The term “proteome” was first used by Wasinger *et al*. in a 1996 Electrophoresis article discussing 2DE methods and results!

Argonne National Laboratory

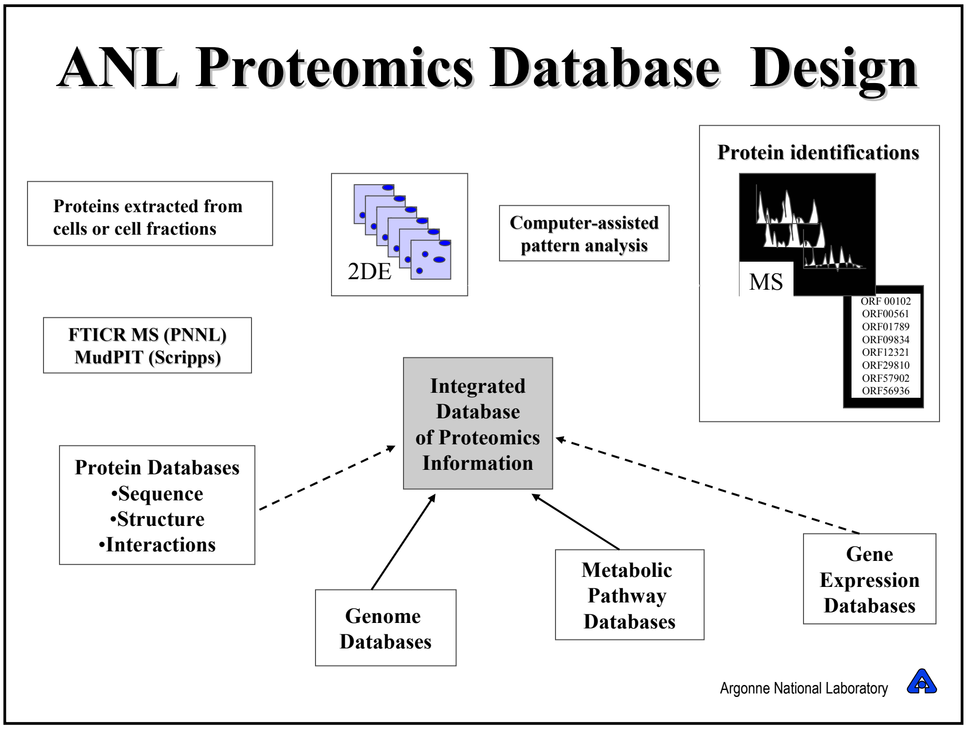


2DE Provides Lots of Data

- Relative abundance (with or without metabolic labeling)
- pI and MW
- Post-translational modifications
- Identifications

M. jannaschii Deprived of H2

Argonne National Laboratory



ANL Proteomics Database Design

Proteins extracted from cells or cell fractions

2DE

Computer-assisted pattern analysis

Protein identifications

MS

ORF 00102
ORF00561
ORF01789
ORF09834
ORF12321
ORF29810
ORF57902
ORF56936

FTICR MS (PNNL)
MudPIT (Scripps)

Integrated Database of Proteomics Information

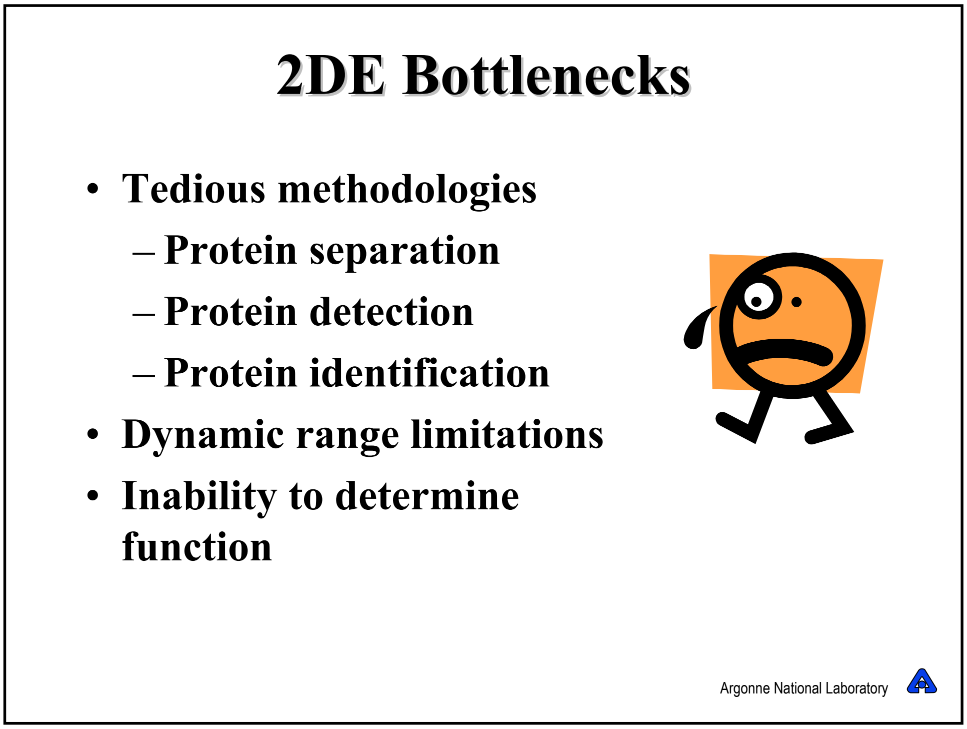
Protein Databases
- •Sequence
- •Structure
- •Interactions

Genome Databases

Metabolic Pathway Databases

Gene Expression Databases

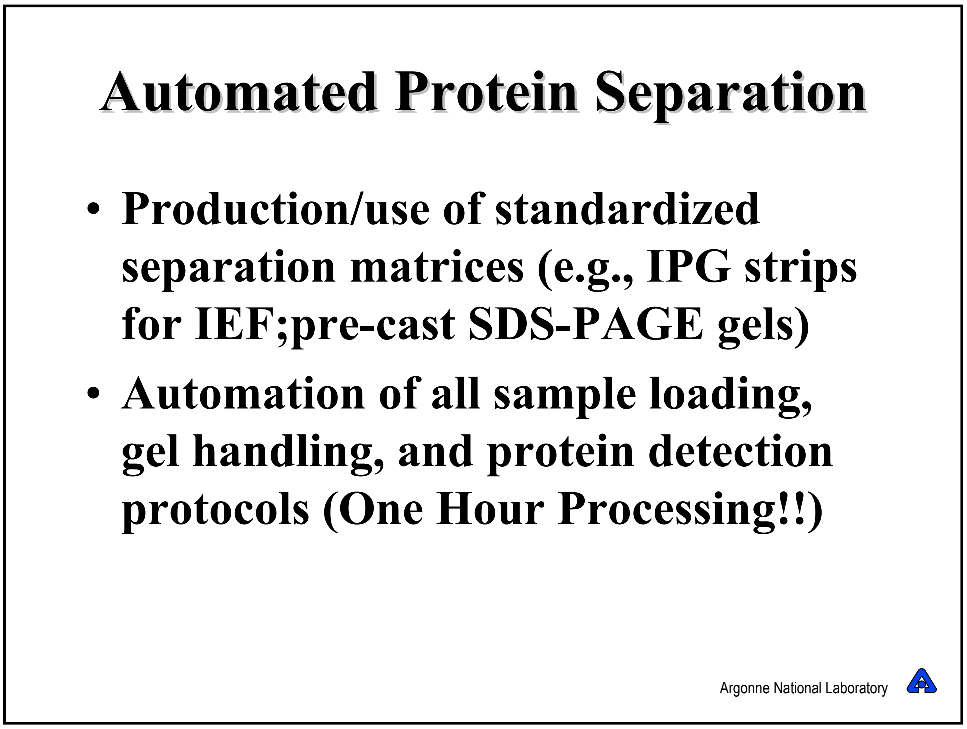
Argonne National Laboratory



2DE Bottlenecks

- Tedious methodologies
  - Protein separation
  - Protein detection
  - Protein identification
- Dynamic range limitations
- Inability to determine function

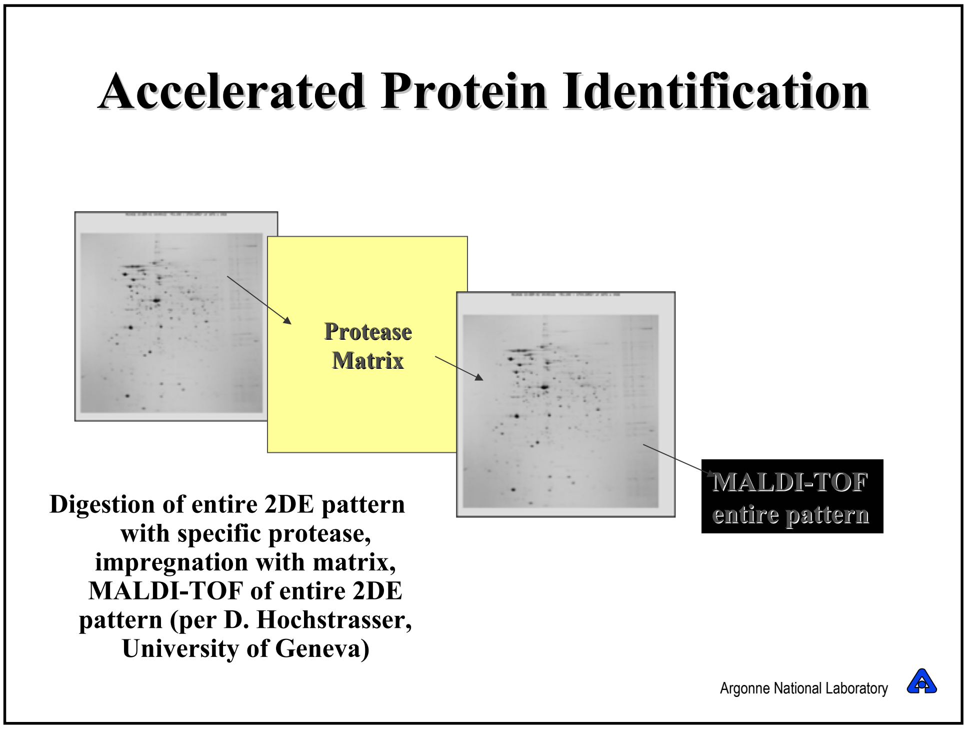
Argonne National Laboratory



Automated Protein Separation

- Production/use of standardized separation matrices (e.g., IPG strips for IEF;pre-cast SDS-PAGE gels)
- Automation of all sample loading, gel handling, and protein detection protocols (One Hour Processing!!)

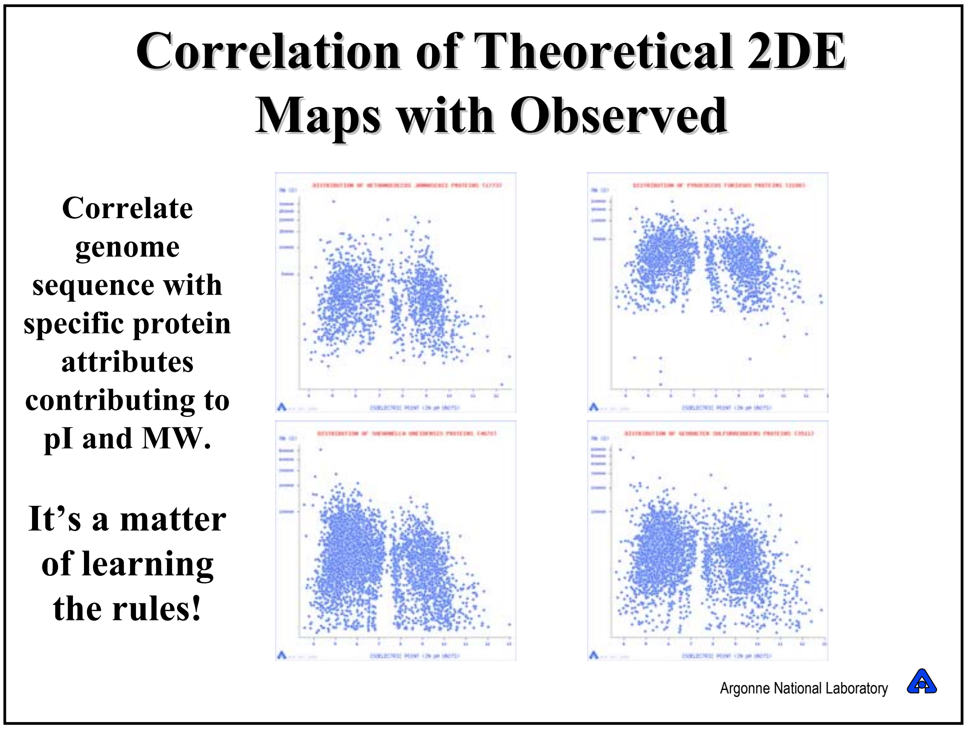
Argonne National Laboratory



Accelerated Protein Identification

Digestion of entire 2DE pattern with specific protease, impregnation with matrix, MALDI-TOF of entire 2DE pattern (per D. Hochstrasser, University of Geneva)

Argonne National Laboratory

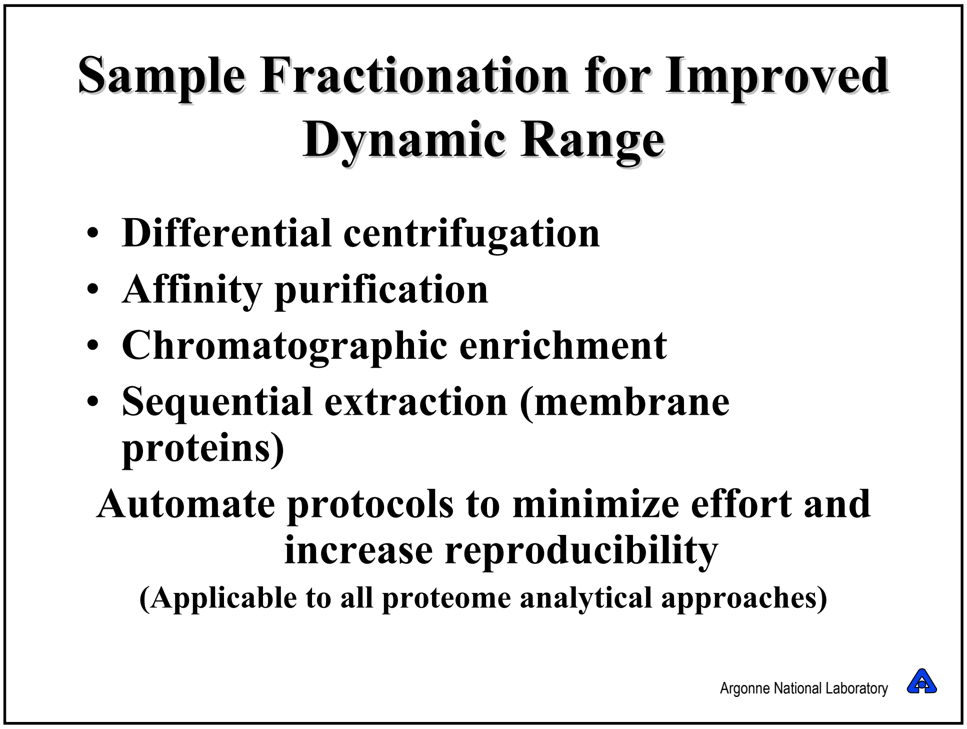


Correlation of Theoretical 2DE Maps with Observed

Correlate genome sequence with specific protein attributes contributing to pI and MW.

It's a matter of learning the rules!

Argonne National Laboratory



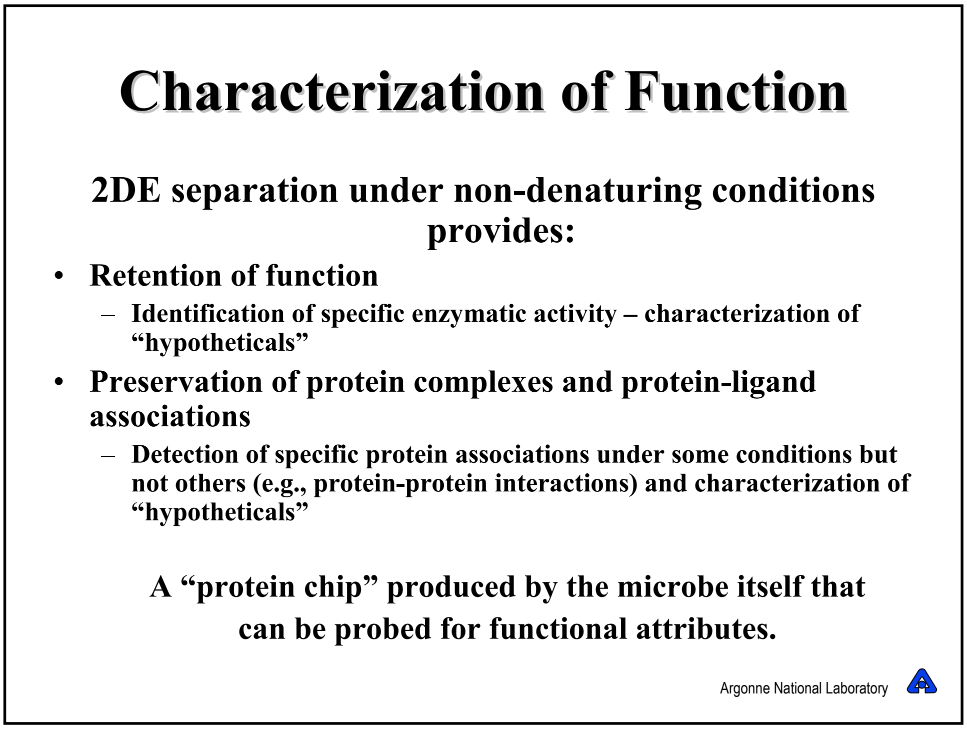
Sample Fractionation for Improved Dynamic Range

- Differential centrifugation
- Affinity purification
- Chromatographic enrichment
- Sequential extraction (membrane proteins)

Automate protocols to minimize effort and increase reproducibility

(Applicable to all proteome analytical approaches)

Argonne National Laboratory



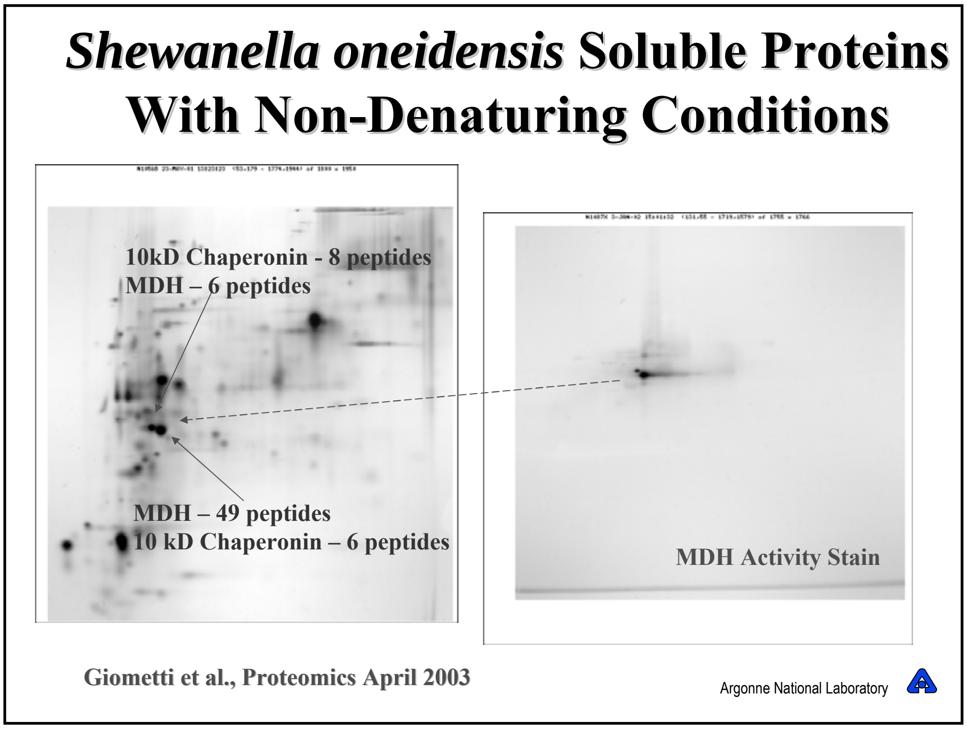
Characterization of Function

2DE separation under non-denaturing conditions provides:

- Retention of function
  - Identification of specific enzymatic activity – characterization of “hypotheticals”
- Preservation of protein complexes and protein-ligand associations
  - Detection of specific protein associations under some conditions but not others (e.g., protein-protein interactions) and characterization of “hypotheticals”

A “protein chip” produced by the microbe itself that can be probed for functional attributes.

Argonne National Laboratory



Shewanella oneidensis Soluble Proteins With Non-Denaturing Conditions

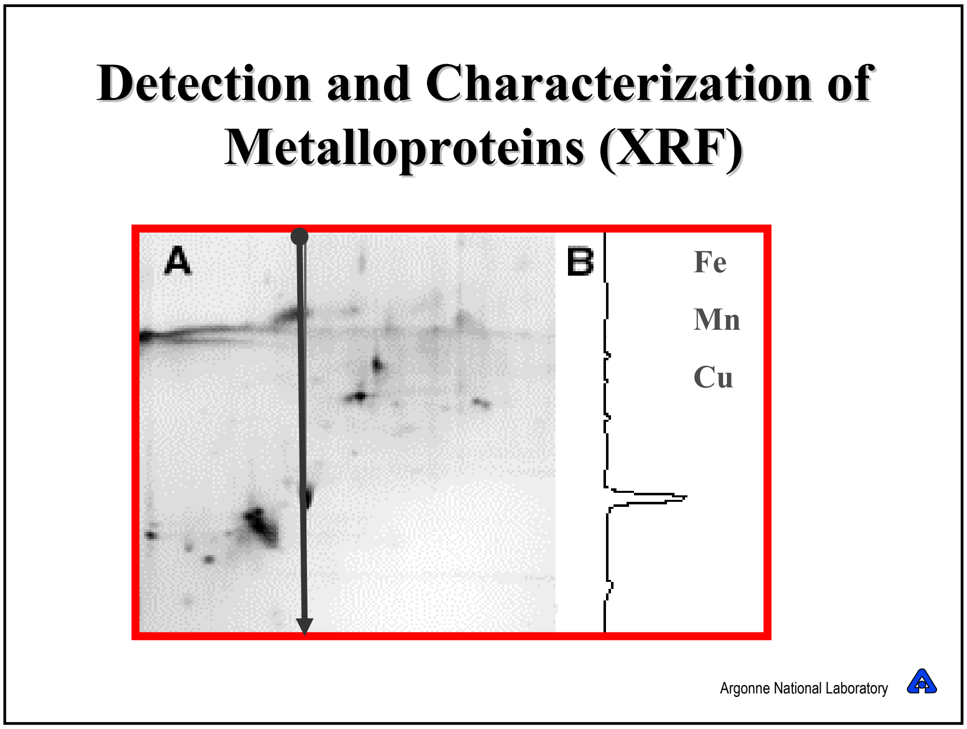
10kD Chaperonin - 8 peptides
MDH – 6 peptides

MDH – 49 peptides
10 kD Chaperonin – 6 peptides

MDH Activity Stain

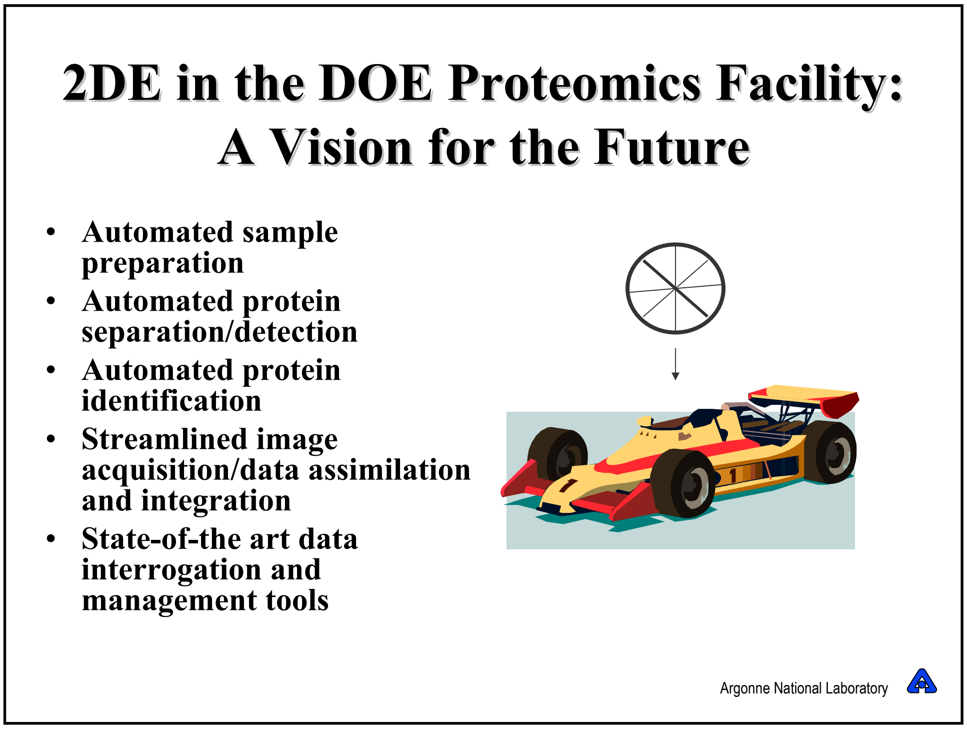
Giometti et al., Proteomics April 2003

Argonne National Laboratory



Detection and Characterization of Metalloproteins (XRF)

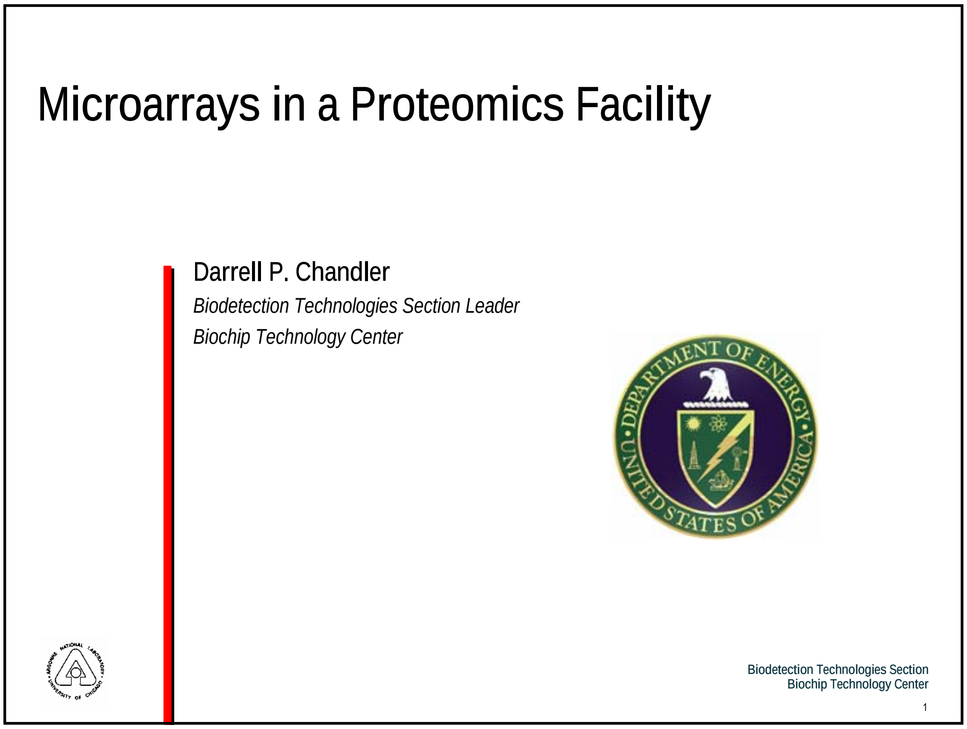
Argonne National Laboratory



2DE in the DOE Proteomics Facility: A Vision for the Future

- Automated sample preparation
- Automated protein separation/detection
- Automated protein identification
- Streamlined image acquisition/data assimilation and integration
- State-of-the art data interrogation and management tools

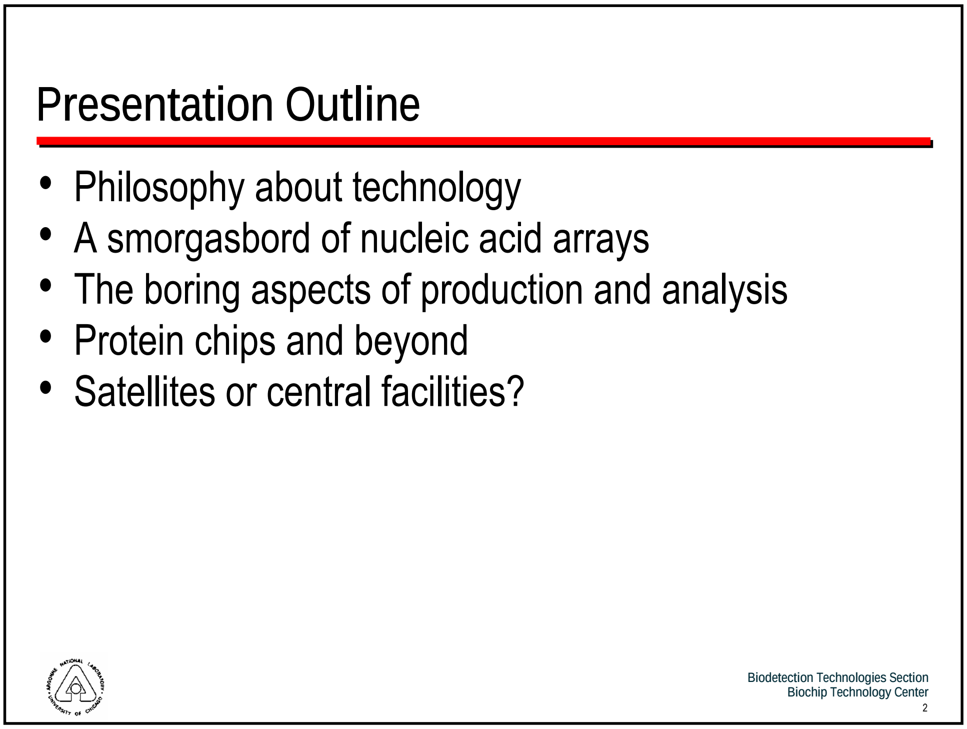
Argonne National Laboratory



Microarrays in a Proteomics Facility

Darrell P. Chandler
Biodetection Technologies Section Leader
Biochip Technology Center

Biodetection Technologies Section
Biochip Technology Center
1



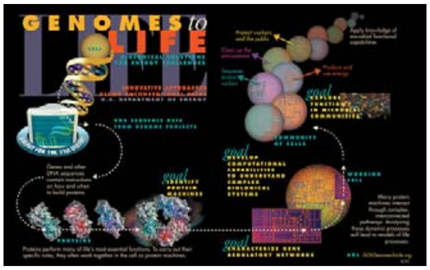
Presentation Outline

- Philosophy about technology
- A smorgasbord of nucleic acid arrays
- The boring aspects of production and analysis
- Protein chips and beyond
- Satellites or central facilities?

Biodetection Technologies Section
Biochip Technology Center
2

Start with the End in Mind (?)

- Identify ≠ characterize
- Complex ≠ machine
- Cell ≠ community
- Culture ≠ natural environment

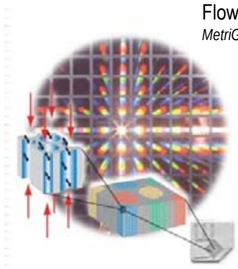


- When investing in or developing technology, how far forward should one look?
- What is your end state?

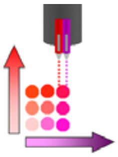
Biodetection Technologies Section
Biochip Technology Center

A Smorgasbord of DNA Microarrays

Planar arrays = glass substrates, SAMs, coatings.

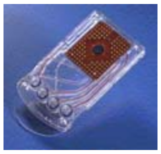


Flow-through chips
MetriGenix



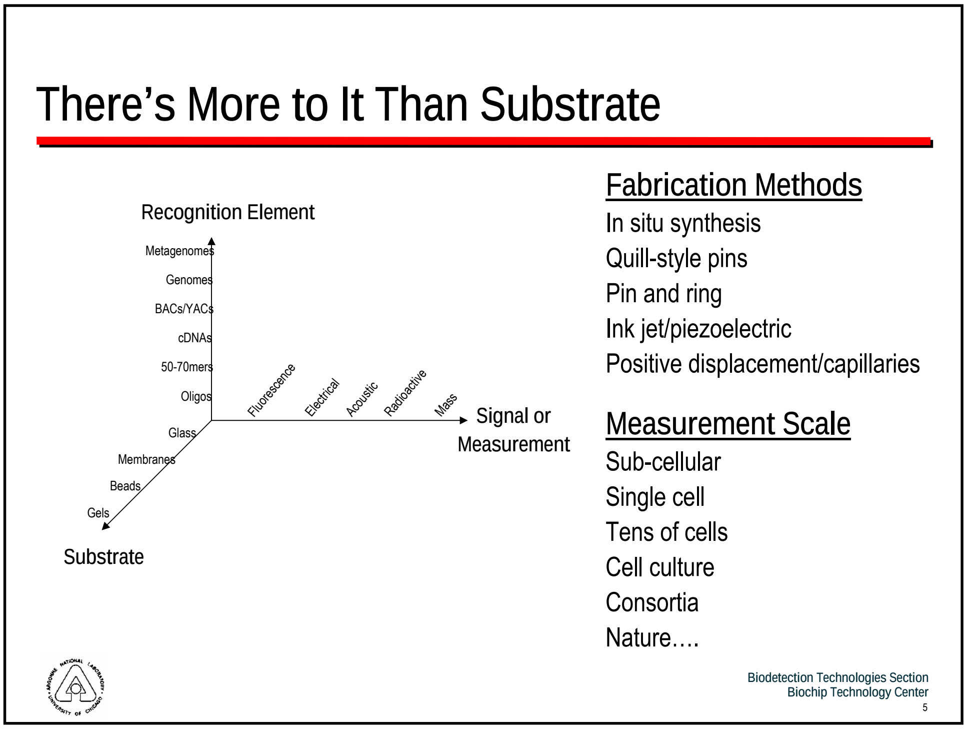
Coded beads = fluidized or suspension arrays.

Gels



Electronic chips
Nanogen

Biodetection Technologies Section
Biochip Technology Center



There’s More to It Than Substrate

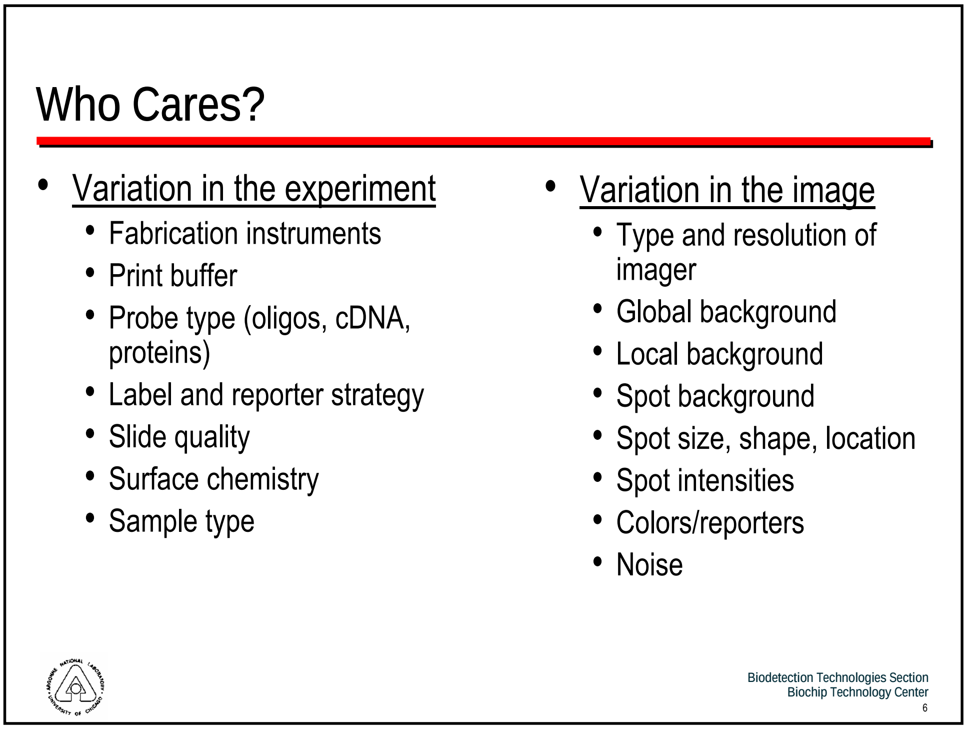
Fabrication Methods

In situ synthesis
Quill-style pins
Pin and ring
Ink jet/piezoelectric
Positive displacement/capillaries

Measurement Scale

Sub-cellular
Single cell
Tens of cells
Cell culture
Consortia
Nature….

Biodetection Technologies Section
Biochip Technology Center
5



Who Cares?

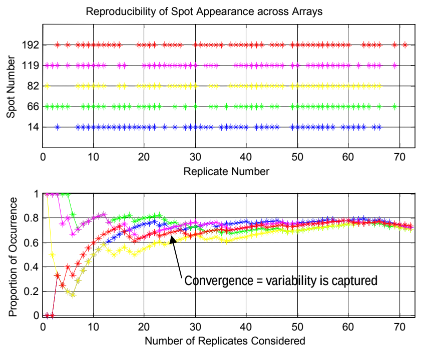
- Variation in the experiment
  - Fabrication instruments
  - Print buffer
  - Probe type (oligos, cDNA, proteins)
  - Label and reporter strategy
  - Slide quality
  - Surface chemistry
  - Sample type
- Variation in the image
  - Type and resolution of imager
  - Global background
  - Local background
  - Spot background
  - Spot size, shape, location
  - Spot intensities
  - Colors/reporters
  - Noise

Biodetection Technologies Section
Biochip Technology Center
6

Measurement Noise Defines Replication Requirements

9-mer probes, planar array

- Every day for 5 days
  - 6 organisms
  - One DNA extraction
  - 3 replicate PCR amplifications
  - 2 hybridizations (to separate chip print lots) per PCR replication
  - 2 arrays per hybridization
  - = 60 replicate arrays per individual
- 24 replicates captures variability in low S/N, informative probe spots



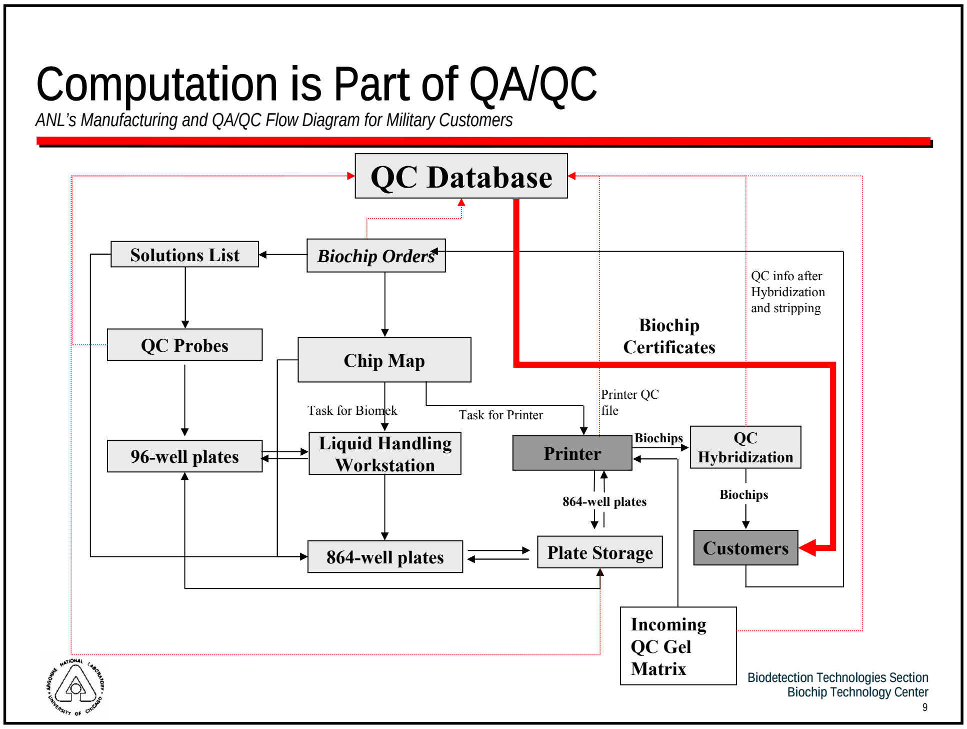
Five probe spots that are ON approximately 70% of the time are considered in this analysis. A minimum of 24 replicate arrays are required to confidently capture the variation in microarray fabrication and hybridization. Similar results are obtained for probe spots that are ON 30, 50 or 90% of the time (not shown).

Biodetection Technologies Section
Biochip Technology Center
7

QA/QC in Production Mode

- Garbage in, Garbage out
  - – Image analysis and statistics can't solve everything
- How does one ensure:
  - – Substrate quality, Probe quality, Chip quality
- How does the choice of technology platform(s) affect the QA/QC pipeline?
- Does your QA/QC system support DOE's long-term goal of predictive biology?
- Who is (going to be) responsible for the QA/QC?

Biodetection Technologies Section
Biochip Technology Center
8



Computation is Part of QA/QC

ANL's Manufacturing and QA/QC Flow Diagram for Military Customers

QC Database

Solutions List

Biochip Orders

QC Probes

Chip Map

QC info after Hybridization and stripping

Biochip Certificates

Task for Biomek

Task for Printer

Printer QC file

96-well plates

Liquid Handling Workstation

Printer

Biochips

QC Hybridization

864-well plates

Biochips

864-well plates

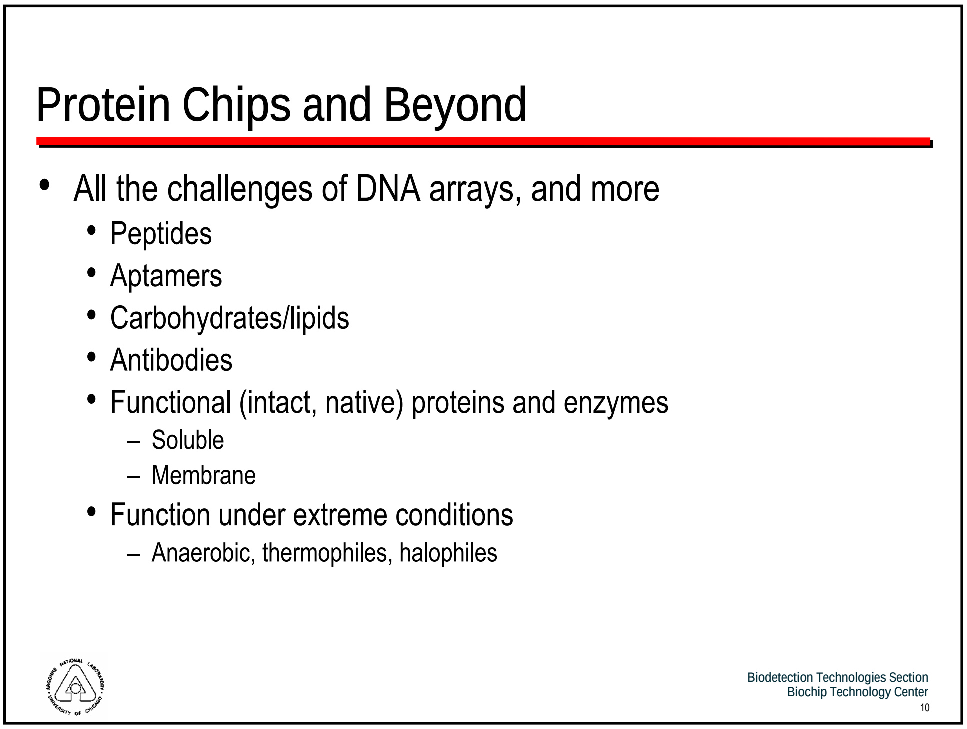
Plate Storage

Customers

Incoming QC Gel Matrix

Biodetection Technologies Section
Biochip Technology Center

9

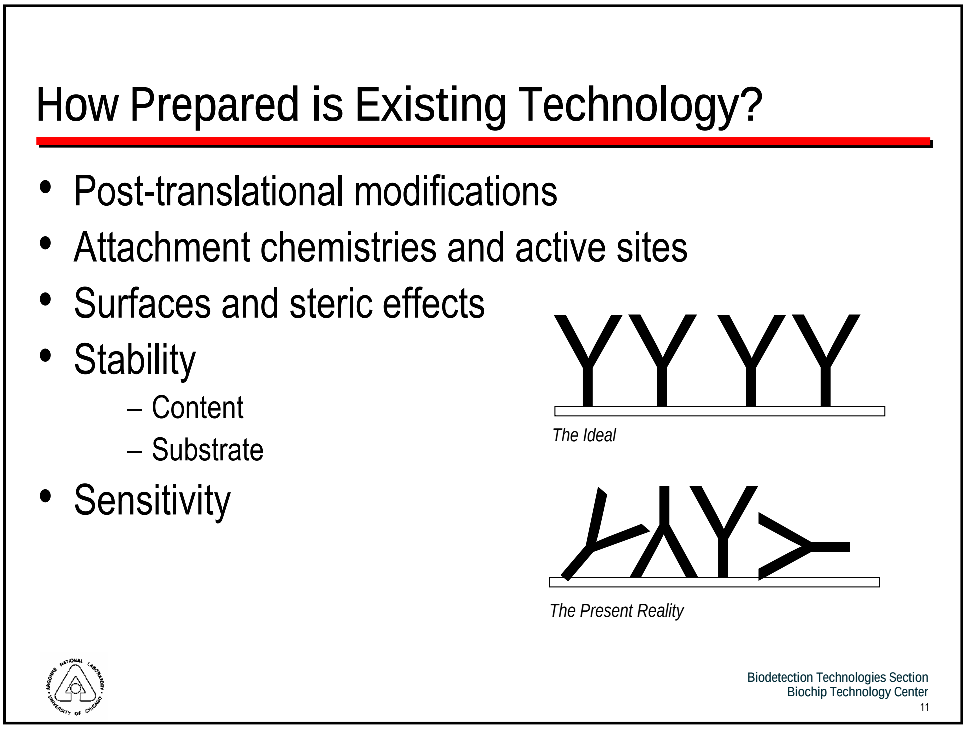


Protein Chips and Beyond

- All the challenges of DNA arrays, and more
  - Peptides
  - Aptamers
  - Carbohydrates/lipids
  - Antibodies
  - Functional (intact, native) proteins and enzymes
    - Soluble
    - Membrane
  - Function under extreme conditions
    - Anaerobic, thermophiles, halophiles

Biodetection Technologies Section
Biochip Technology Center

10



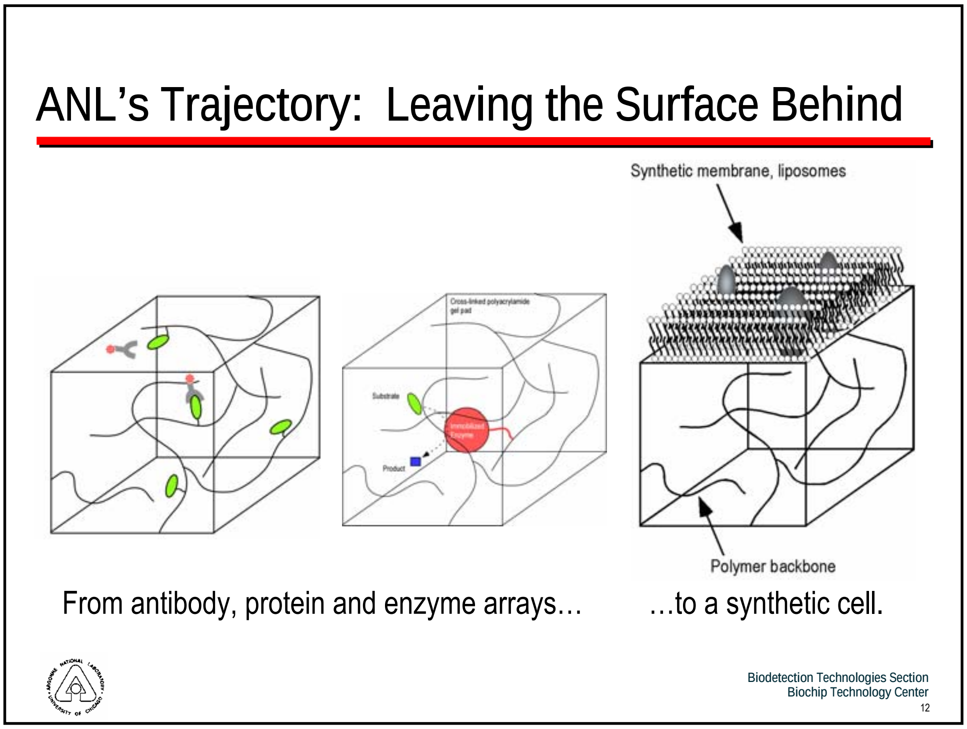
How Prepared is Existing Technology?

- Post-translational modifications
- Attachment chemistries and active sites
- Surfaces and steric effects
- Stability
  - Content
  - Substrate
- Sensitivity

The Ideal

The Present Reality

Biodetection Technologies Section
Biochip Technology Center
11

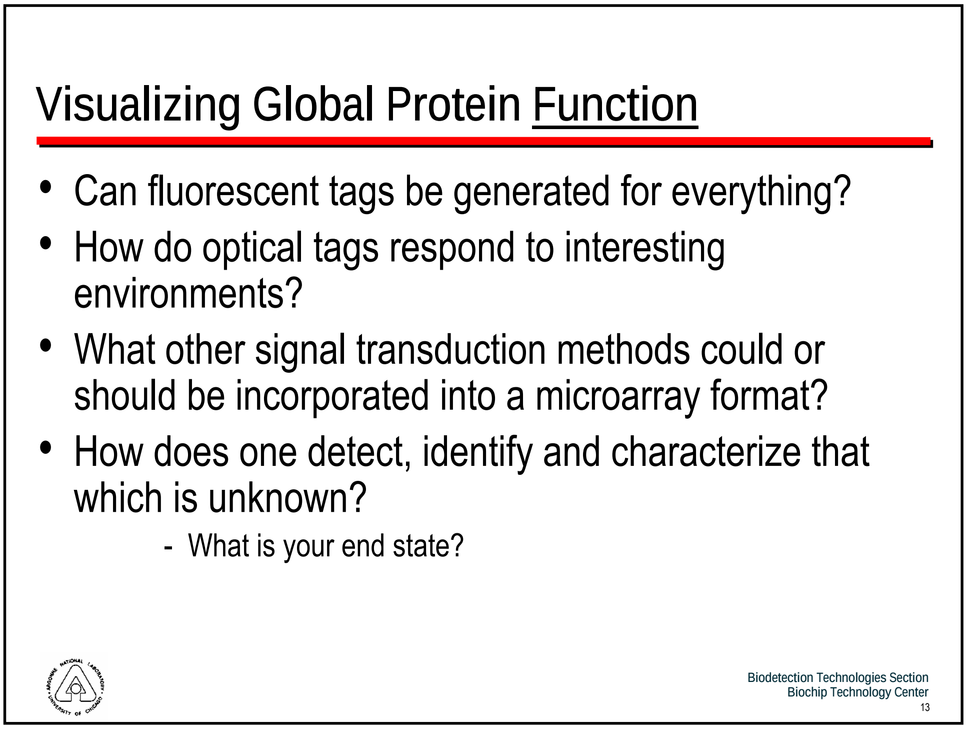


ANL’s Trajectory: Leaving the Surface Behind

Cross-linked polyacrylamide gel pad
Substrate
Immobilized Enzyme
Product
Synthetic membrane, liposomes
Polymer backbone

From antibody, protein and enzyme arrays… …to a synthetic cell.

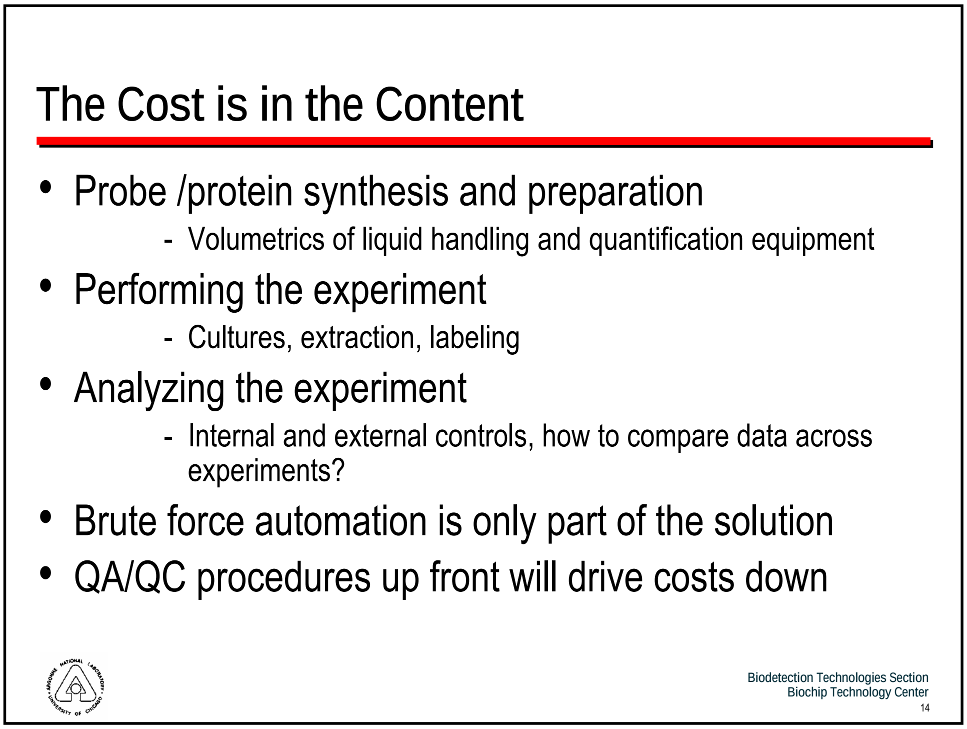
Biodetection Technologies Section
Biochip Technology Center
12



Visualizing Global Protein Function

- Can fluorescent tags be generated for everything?
- How do optical tags respond to interesting environments?
- What other signal transduction methods could or should be incorporated into a microarray format?
- How does one detect, identify and characterize that which is unknown?
  - What is your end state?

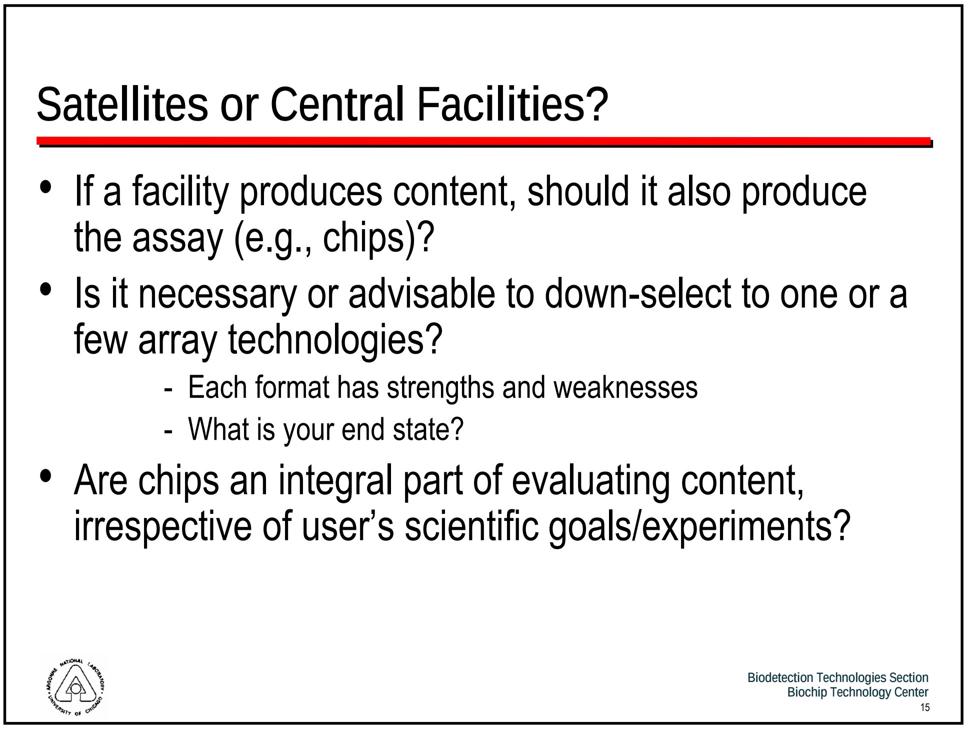
Biodetection Technologies Section
Biochip Technology Center
13



The Cost is in the Content

- Probe /protein synthesis and preparation
  - Volumetrics of liquid handling and quantification equipment
- Performing the experiment
  - Cultures, extraction, labeling
- Analyzing the experiment
  - Internal and external controls, how to compare data across experiments?
- Brute force automation is only part of the solution
- QA/QC procedures up front will drive costs down

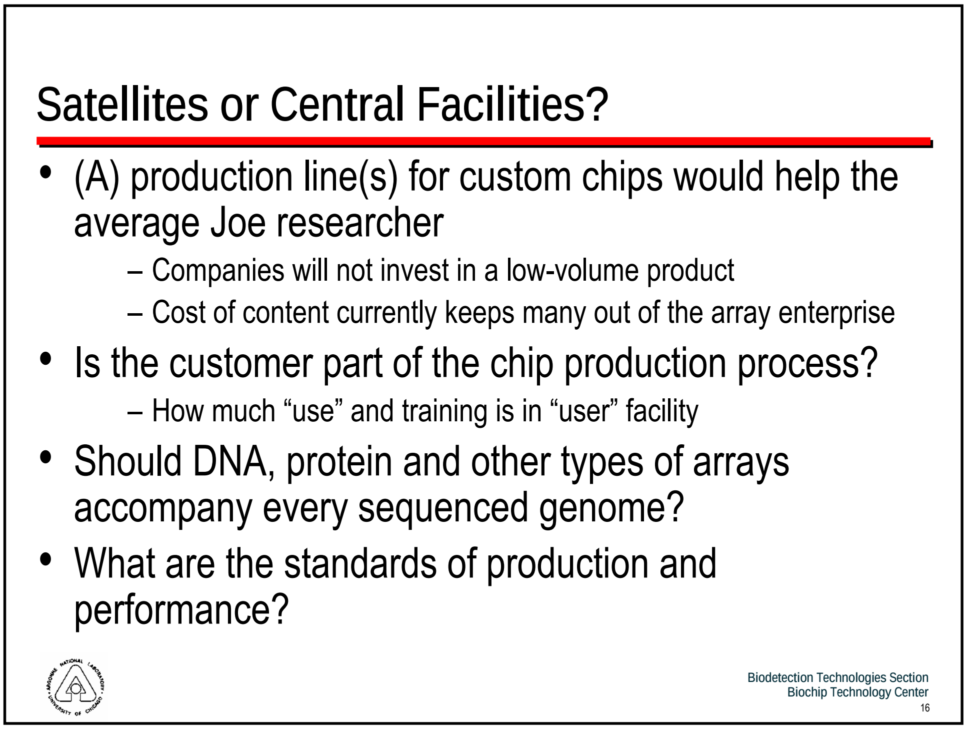
Biodetection Technologies Section
Biochip Technology Center
14



Satellites or Central Facilities?

- If a facility produces content, should it also produce the assay (e.g., chips)?
- Is it necessary or advisable to down-select to one or a few array technologies?
  - Each format has strengths and weaknesses
  - What is your end state?
- Are chips an integral part of evaluating content, irrespective of user’s scientific goals/experiments?

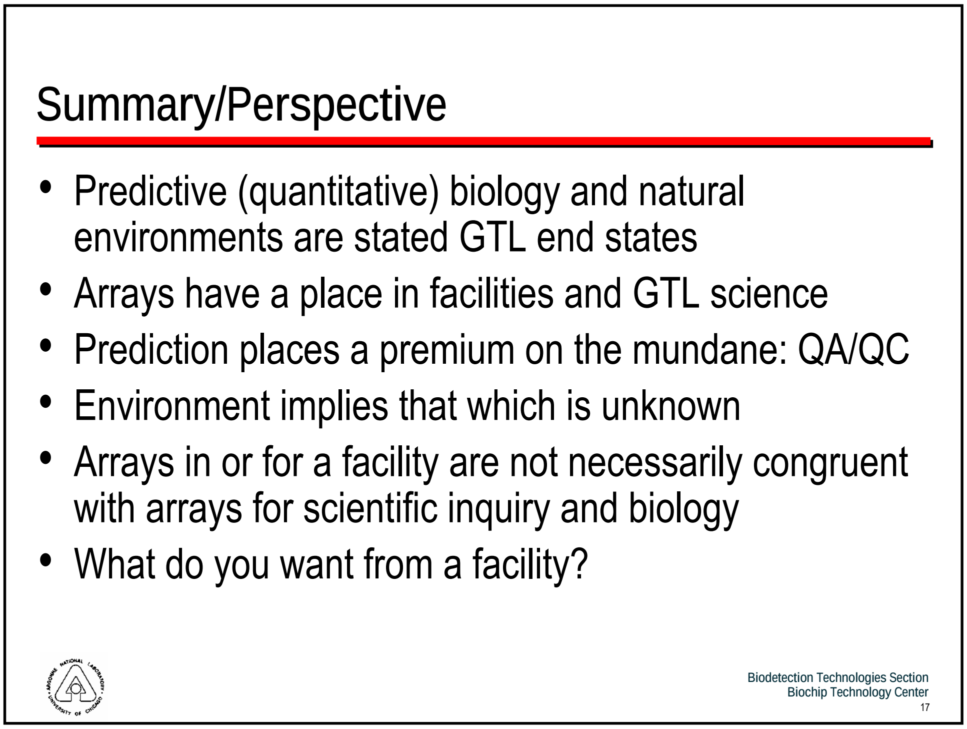
Biodetection Technologies Section
Biochip Technology Center
15



Satellites or Central Facilities?

- (A) production line(s) for custom chips would help the average Joe researcher
  - Companies will not invest in a low-volume product
  - Cost of content currently keeps many out of the array enterprise
- Is the customer part of the chip production process?
  - How much “use” and training is in “user” facility
- Should DNA, protein and other types of arrays accompany every sequenced genome?
- What are the standards of production and performance?

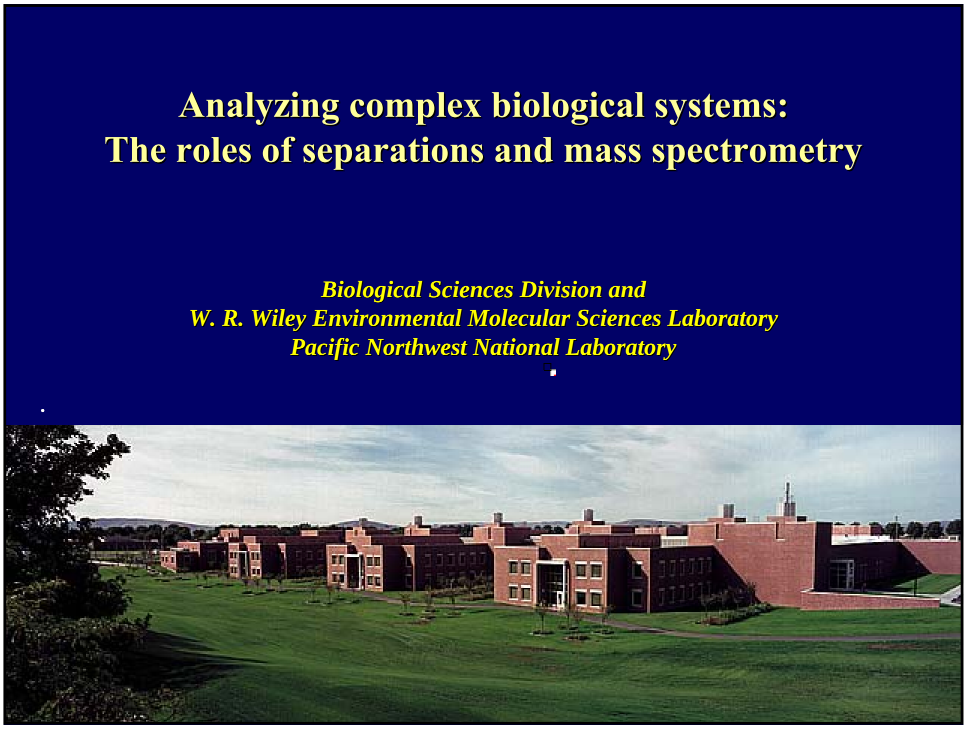
Biodetection Technologies Section
Biochip Technology Center
16



Summary/Perspective

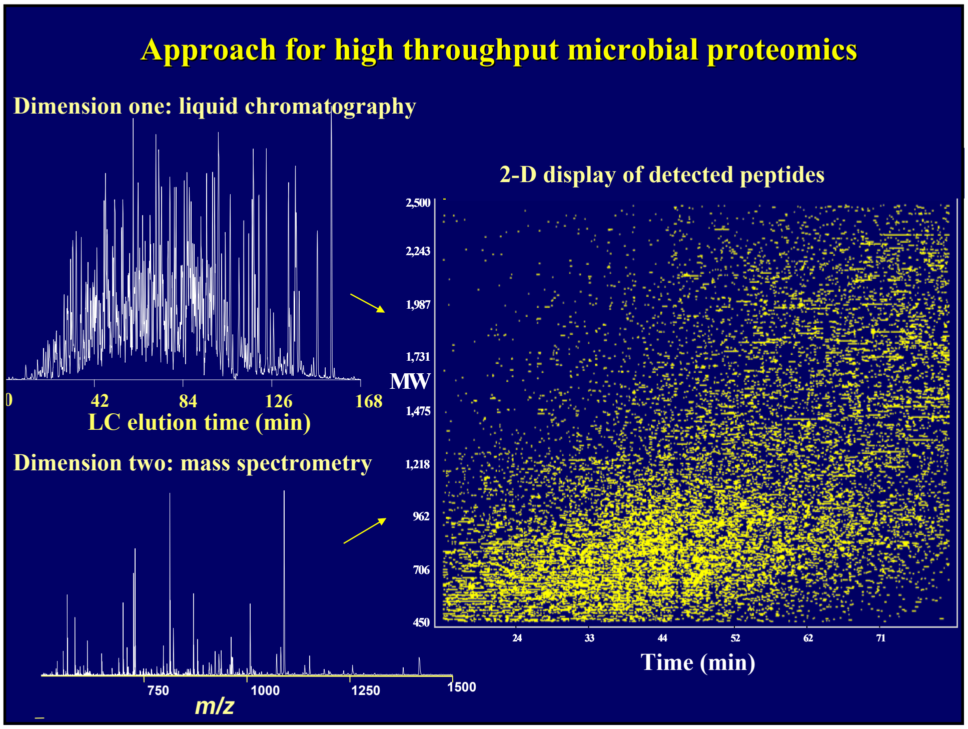
- Predictive (quantitative) biology and natural environments are stated GTL end states
- Arrays have a place in facilities and GTL science
- Prediction places a premium on the mundane: QA/QC
- Environment implies that which is unknown
- Arrays in or for a facility are not necessarily congruent with arrays for scientific inquiry and biology
- What do you want from a facility?

Biodetection Technologies Section
Biochip Technology Center
17



Analyzing complex biological systems: The roles of separations and mass spectrometry

Biological Sciences Division and
W. R. Wiley Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory

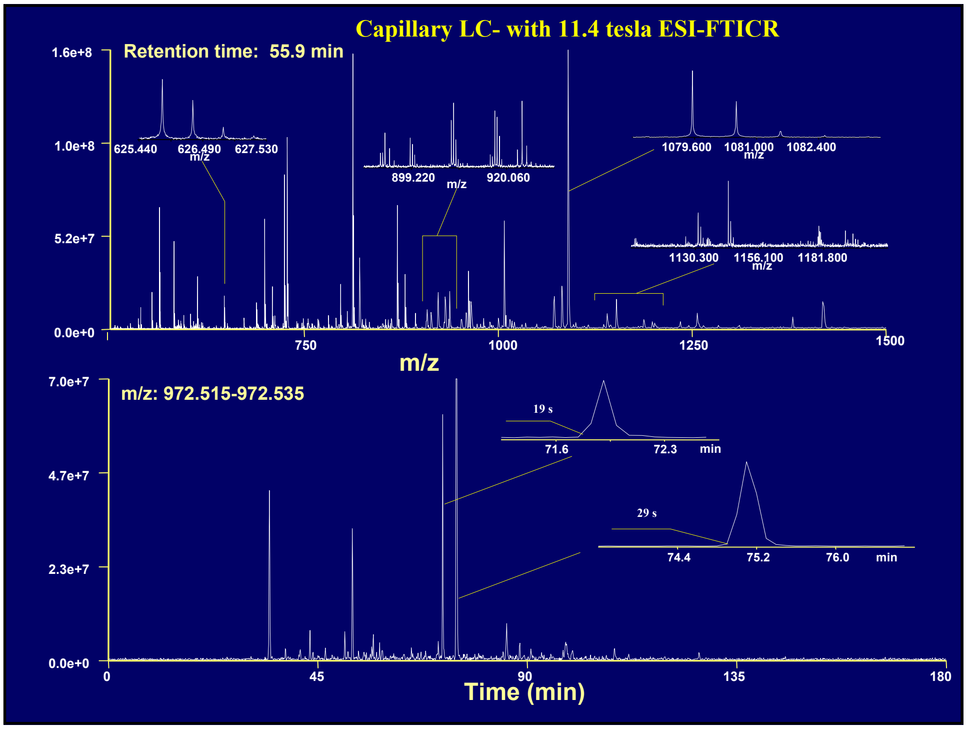


Approach for high throughput microbial proteomics

Dimension one: liquid chromatography

Dimension two: mass spectrometry

2-D display of detected peptides



Capillary LC- with 11.4 tesla ESI-FTICR

Accurate Mass and Time (AMT) Tags

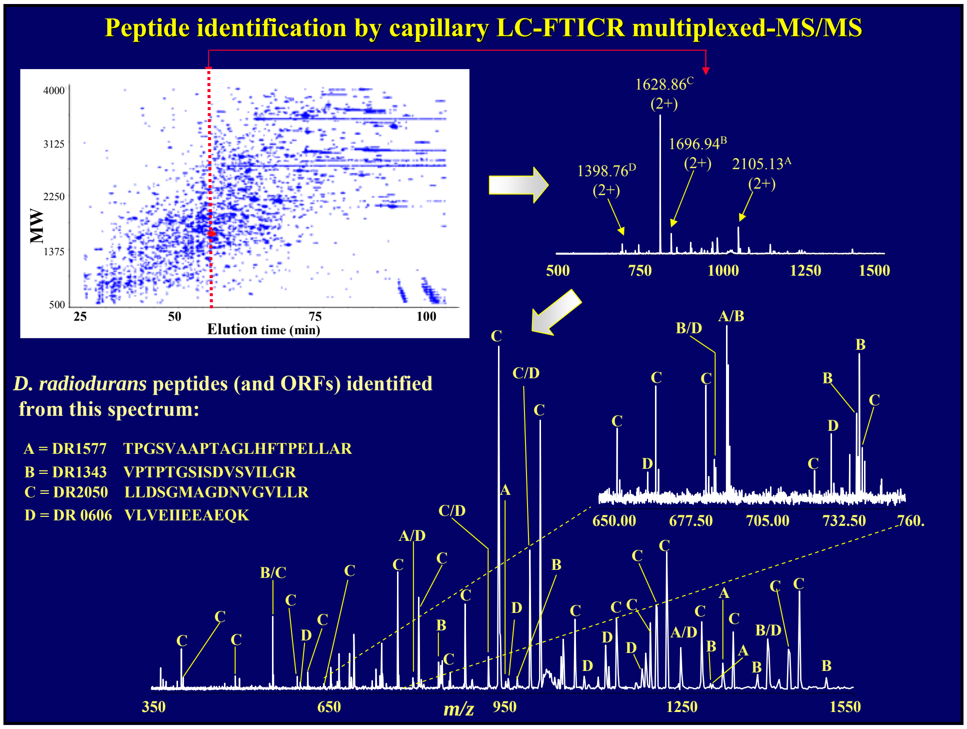
Given the constraint of a sequenced genome, the combination of high accuracy mass measurements and separation (e.g. LC elution) times provides unique marker peptides for essentially all proteins

Two stages:

1. **Initial generation of AMT tags by "shotgun LC-MS/MS" measurements with conventional instrumentation and validation by LC-FTICR**
2. **Application of AMT tags in repeated measurements with the same organism**

- **Avoids routine need for peptide ID by MS/MS**

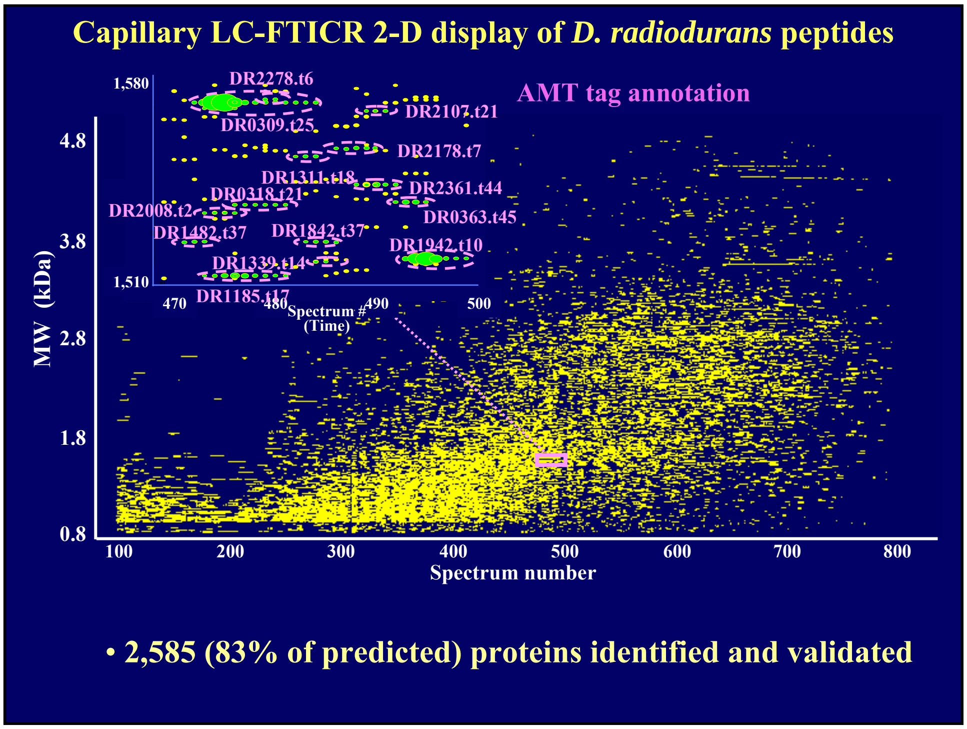
- **Basis for better quantitation, higher throughput and proteome coverage**



Peptide identification by capillary LC-FTICR multiplexed-MS/MS

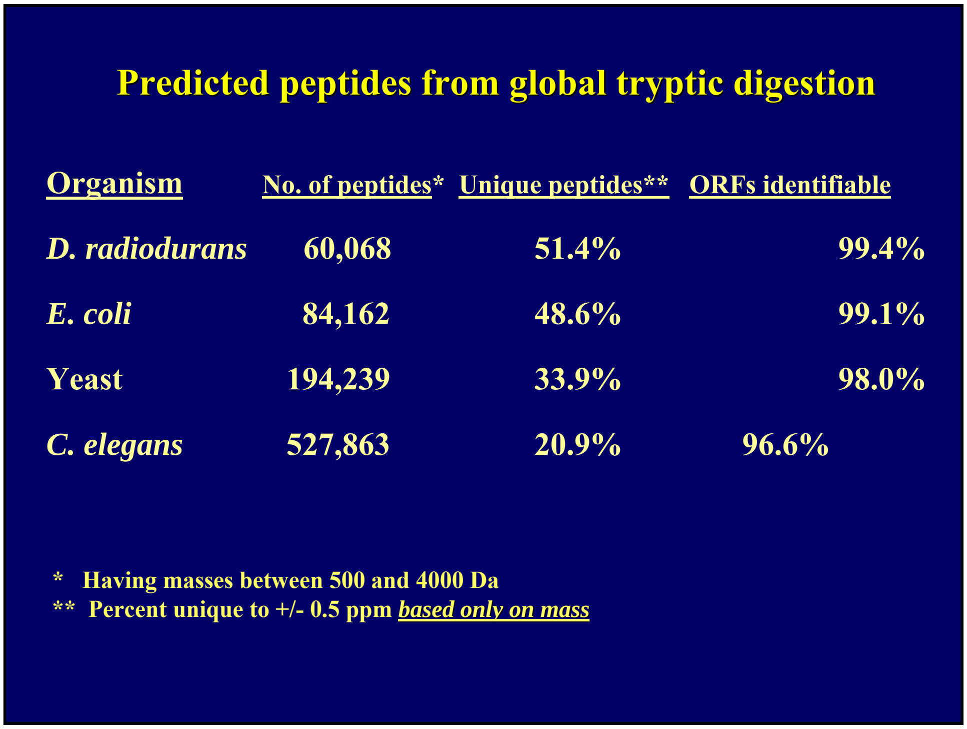
D. radiodurans peptides (and ORFs) identified from this spectrum:

A = DR1577	TPGSVAAPTAGLHFTPELLAR
B = DR1343	VPTPTGSISDVSVILGR
C = DR2050	LLDSGMAGDNVGVLLR
D = DR 0606	VLVEIIEEAEQK



Capillary LC-FTICR 2-D display of *D. radiodurans* peptides

- 2,585 (83% of predicted) proteins identified and validated

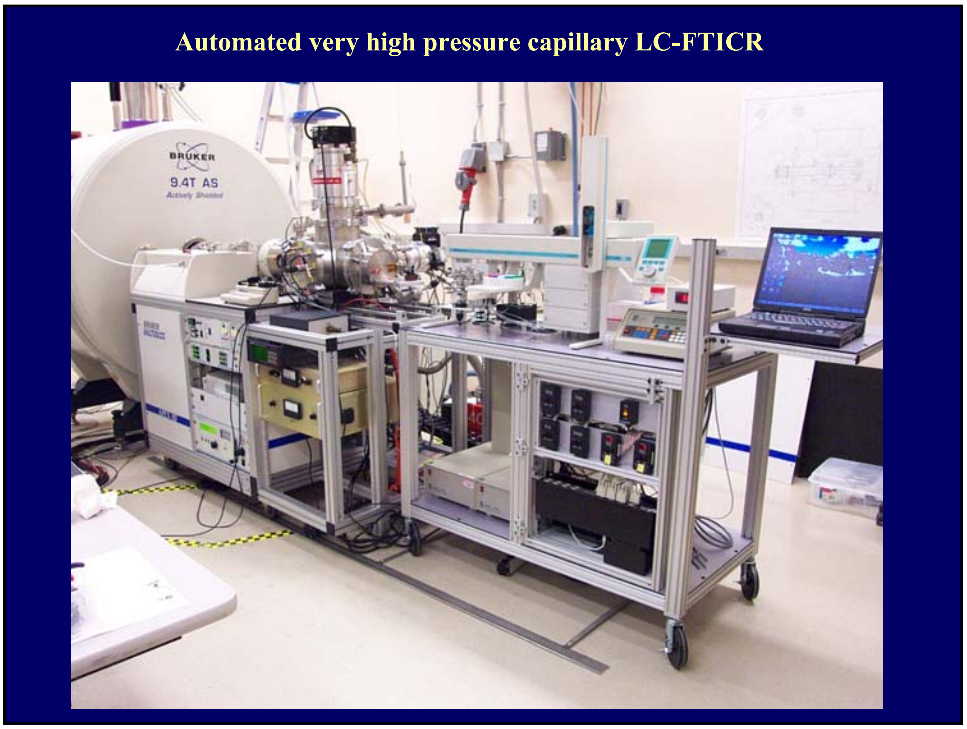


Predicted peptides from global tryptic digestion

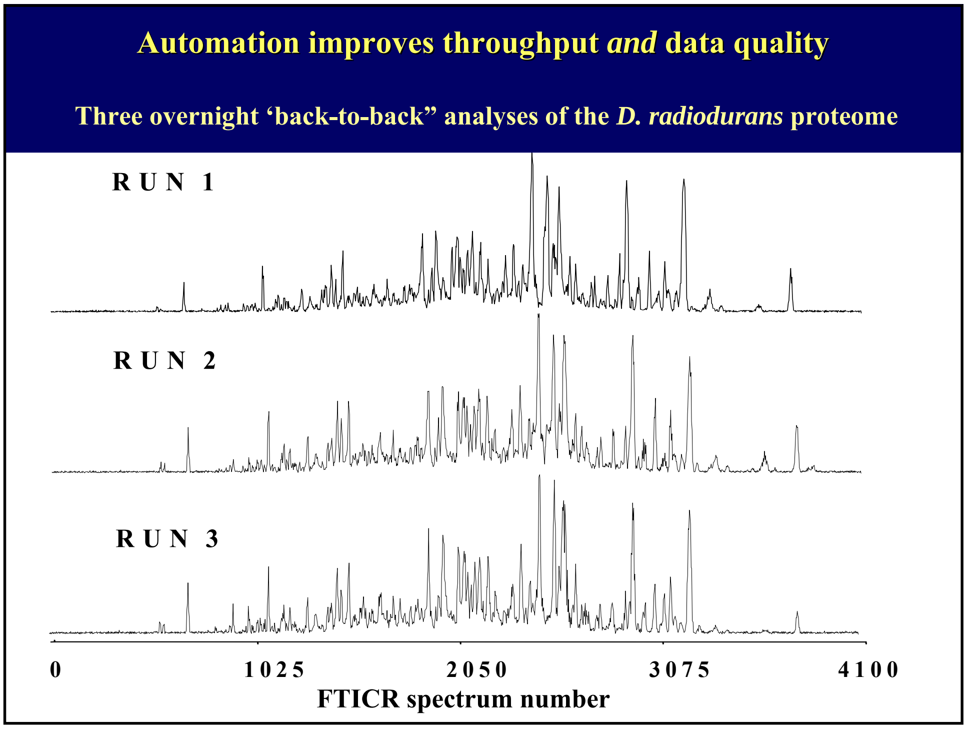
Organism	No. of peptides*	Unique peptides**	ORFs identifiable
D. radiodurans	60,068	51.4%	99.4%
E. coli	84,162	48.6%	99.1%
Yeast	194,239	33.9%	98.0%
C. elegans	527,863	20.9%	96.6%

* Having masses between 500 and 4000 Da

** Percent unique to +/- 0.5 ppm ***based only on mass***

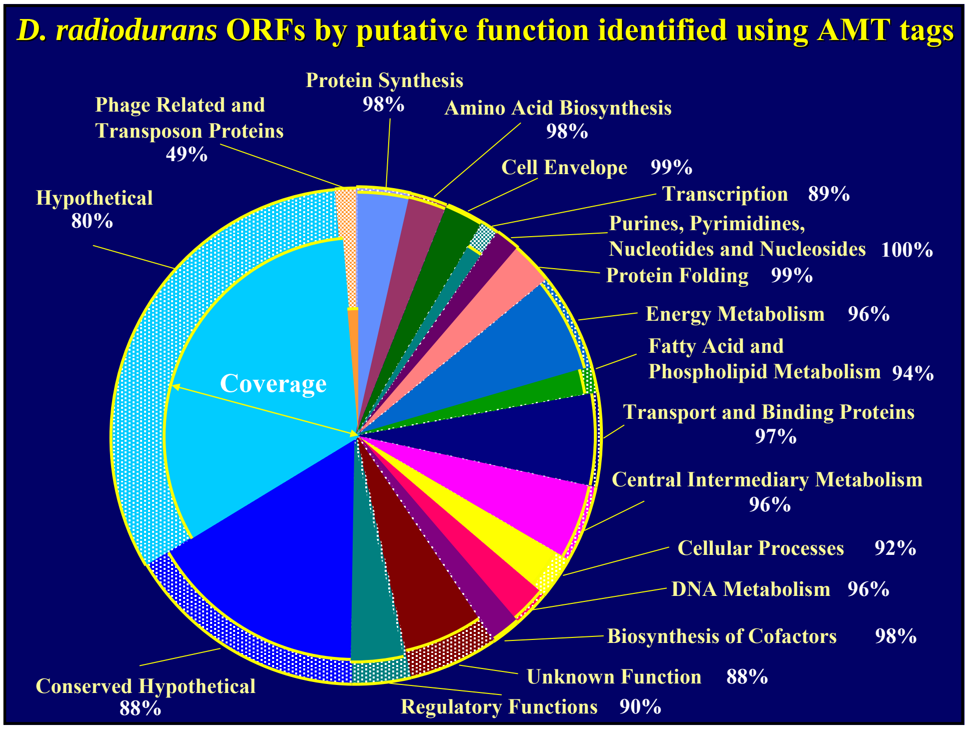


Automated very high pressure capillary LC-FTICR

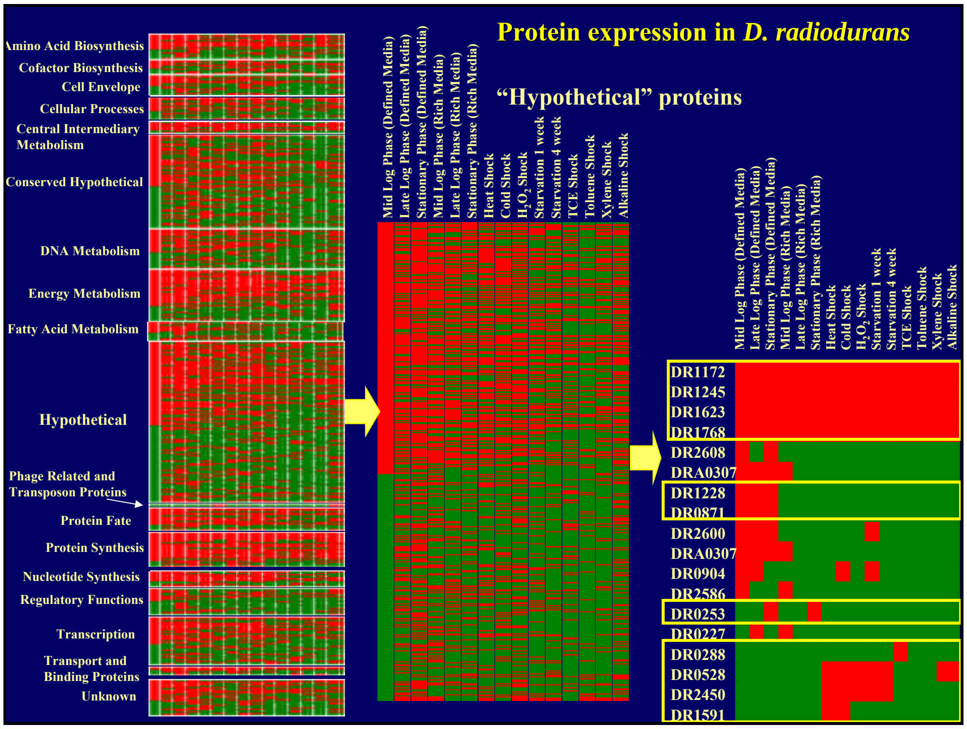


Automation improves throughput *and* data quality

Three overnight 'back-to-back" analyses of the *D. radiodurans* proteome

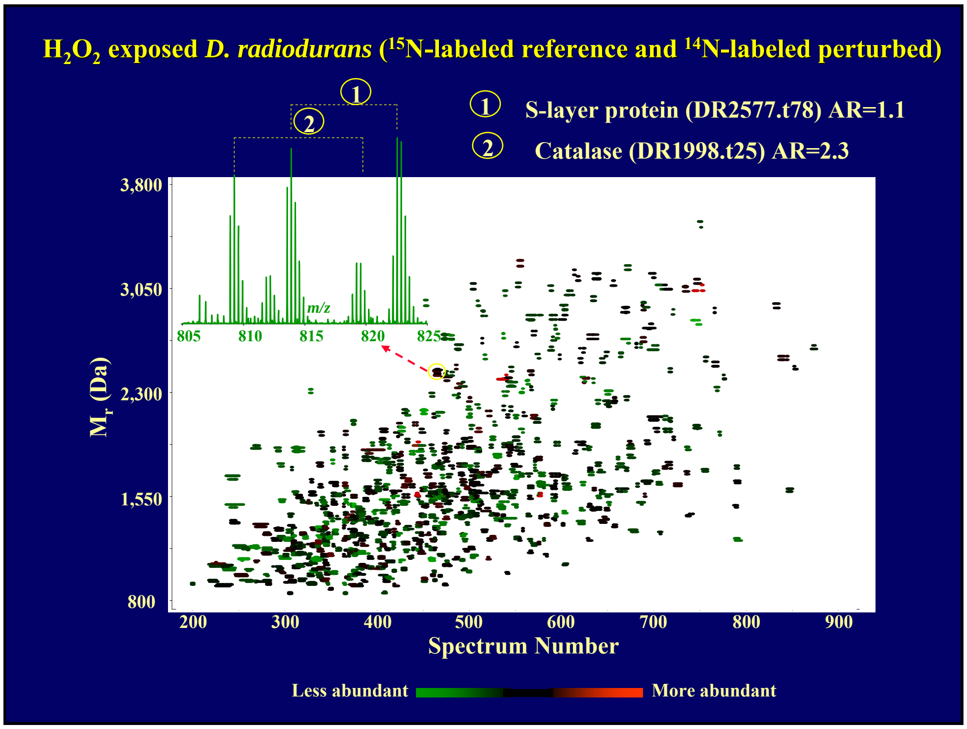


D. radiodurans ORFs by putative function identified using AMT tags



Protein expression in *D. radiodurans*

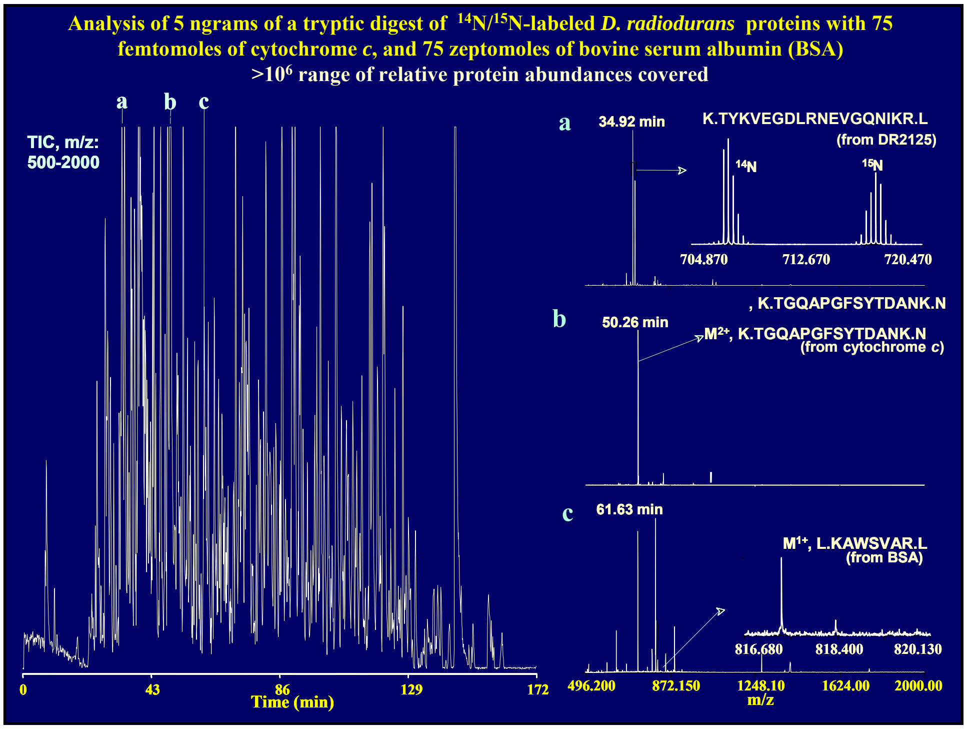
"Hypothetical" proteins



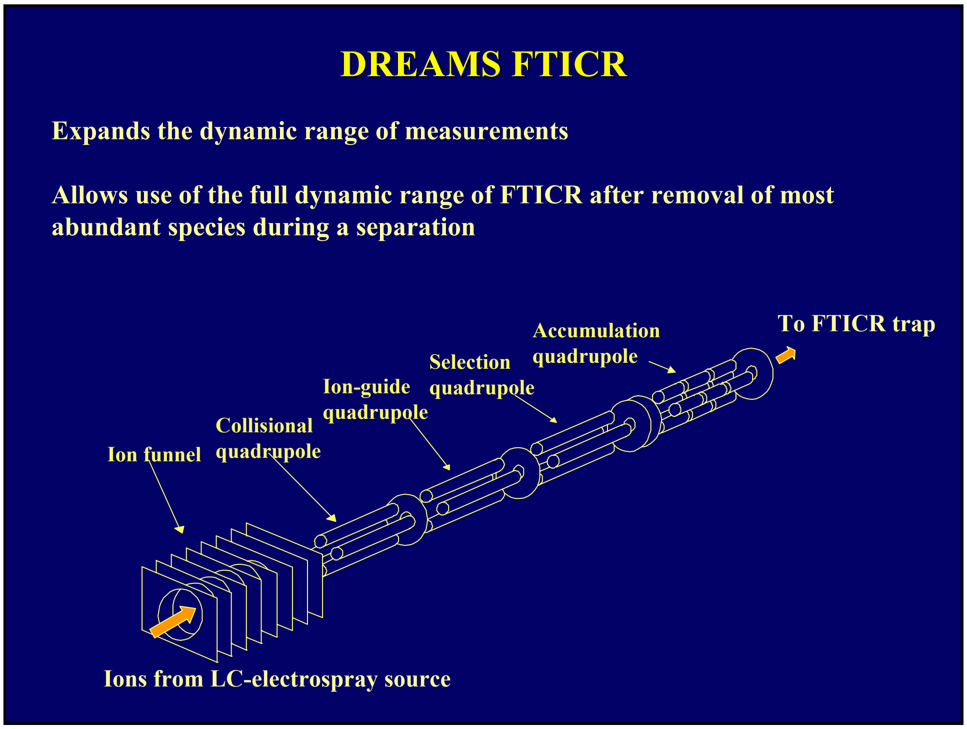
H_2O_2 exposed *D. radiodurans* (^{15}N-labeled reference and ^{14}N-labeled perturbed)

1 S-layer protein (DR2577.t78) AR=1.1

2 Catalase (DR1998.t25) AR=2.3



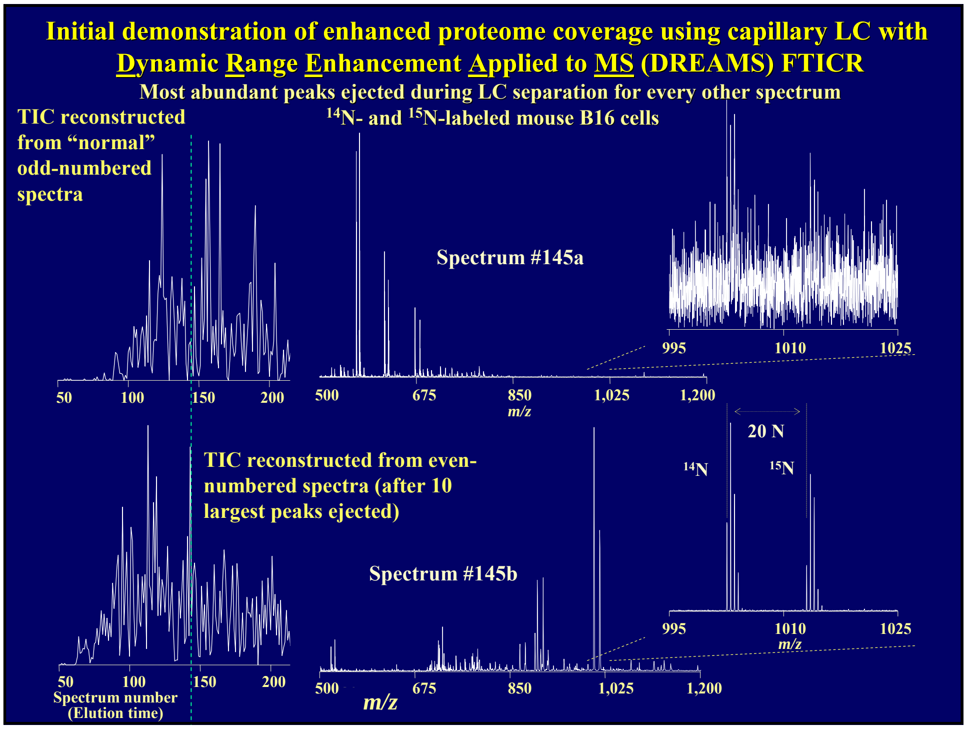
Analysis of 5 ngrams of a tryptic digest of $^{14}N/^{15}N$-labeled *D. radiodurans* proteins with 75 femtomoles of cytochrome *c*, and 75 zeptomoles of bovine serum albumin (BSA)
>10^6 range of relative protein abundances covered



DREAMS FTICR

Expands the dynamic range of measurements

Allows use of the full dynamic range of FTICR after removal of most abundant species during a separation



Initial demonstration of enhanced proteome coverage using capillary LC with Dynamic Range Enhancement Applied to MS (DREAMS) FTICR

Most abundant peaks ejected during LC separation for every other spectrum

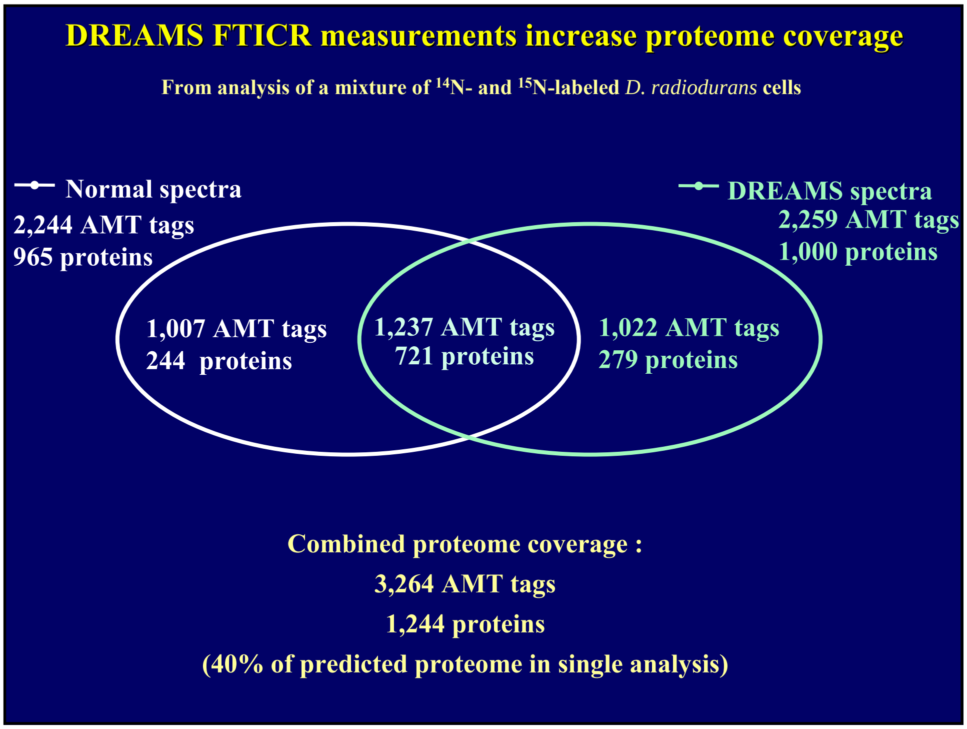
^{14}N- and ^{15}N-labeled mouse B16 cells

TIC reconstructed from “normal” odd-numbered spectra

Spectrum #145a

TIC reconstructed from even-numbered spectra (after 10 largest peaks ejected)

Spectrum #145b



DREAMS FTICR measurements increase proteome coverage

From analysis of a mixture of ^{14}N- and ^{15}N-labeled *D. radiodurans* cells

Normal spectra
2,244 AMT tags
965 proteins

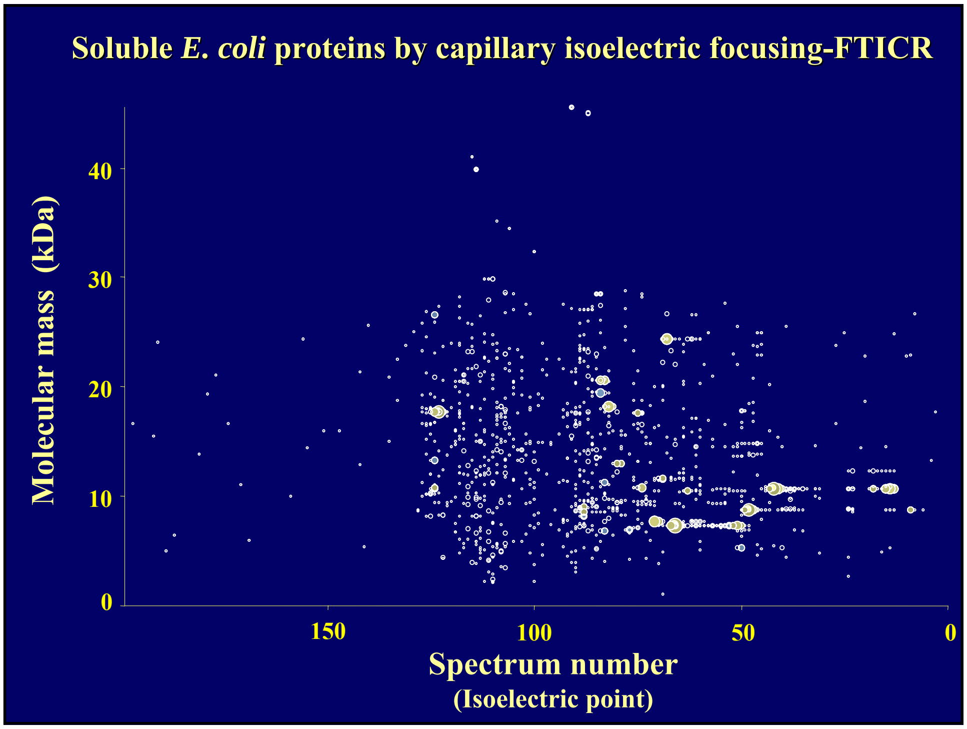
DREAMS spectra
2,259 AMT tags
1,000 proteins

1,007 AMT tags
244 proteins

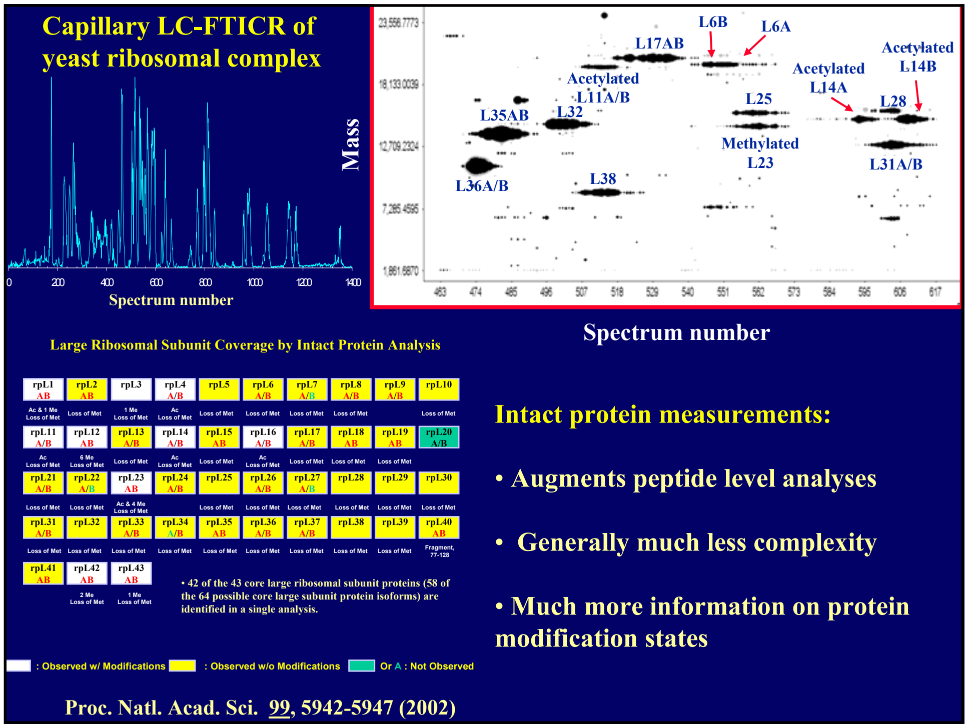
1,237 AMT tags
721 proteins

1,022 AMT tags
279 proteins

Combined proteome coverage :
3,264 AMT tags
1,244 proteins
(40% of predicted proteome in single analysis)



Soluble *E. coli* proteins by capillary isoelectric focusing-FTICR



Capillary LC-FTICR of yeast ribosomal complex

Large Ribosomal Subunit Coverage by Intact Protein Analysis

⬜ rpL1 AB Ac & 1 Me Loss of Met	🟨 rpL2 AB Loss of Met	⬜ rpL3 1 Me Loss of Met	⬜ rpL4 A/B Ac Loss of Met	🟨 rpL5 Loss of Met	🟨 rpL6 A/B Loss of Met	🟨 rpL7 A/B Loss of Met	🟨 rpL8 A/B Loss of Met	🟨 rpL9 A/B	🟨 rpL10 Loss of Met
⬜ rpL11 A/B Ac Loss of Met	⬜ rpL12 AB 6 Me Loss of Met	🟨 rpL13 A/B Loss of Met	⬜ rpL14 A/B Ac Loss of Met	🟨 rpL15 AB Loss of Met	⬜ rpL16 A/B Ac Loss of Met	🟨 rpL17 A/B Loss of Met	🟨 rpL18 AB Loss of Met	🟨 rpL19 AB Loss of Met	🟩 rpL20 A/B
🟨 rpL21 A/B Loss of Met	🟨 rpL22 A/B Loss of Met	⬜ rpL23 AB Ac & 4 Me Loss of Met	🟨 rpL24 A/B	🟨 rpL25 Loss of Met	🟨 rpL26 A/B Loss of Met	🟨 rpL27 A/B Loss of Met	🟨 rpL28 Loss of Met	🟨 rpL29 Loss of Met	🟨 rpL30 Loss of Met
🟨 rpL31 A/B Loss of Met	🟨 rpL32 Loss of Met	🟨 rpL33 A/B Loss of Met	🟨 rpL34 A/B Loss of Met	🟨 rpL35 AB Loss of Met	🟨 rpL36 A/B Loss of Met	🟨 rpL37 A/B Loss of Met	🟨 rpL38 Loss of Met	🟨 rpL39 Loss of Met	🟨 rpL40 AB Fragment, 77-128
🟨 rpL41 AB	⬜ rpL42 AB 2 Me Loss of Met	⬜ rpL43 AB 1 Me Loss of Met							

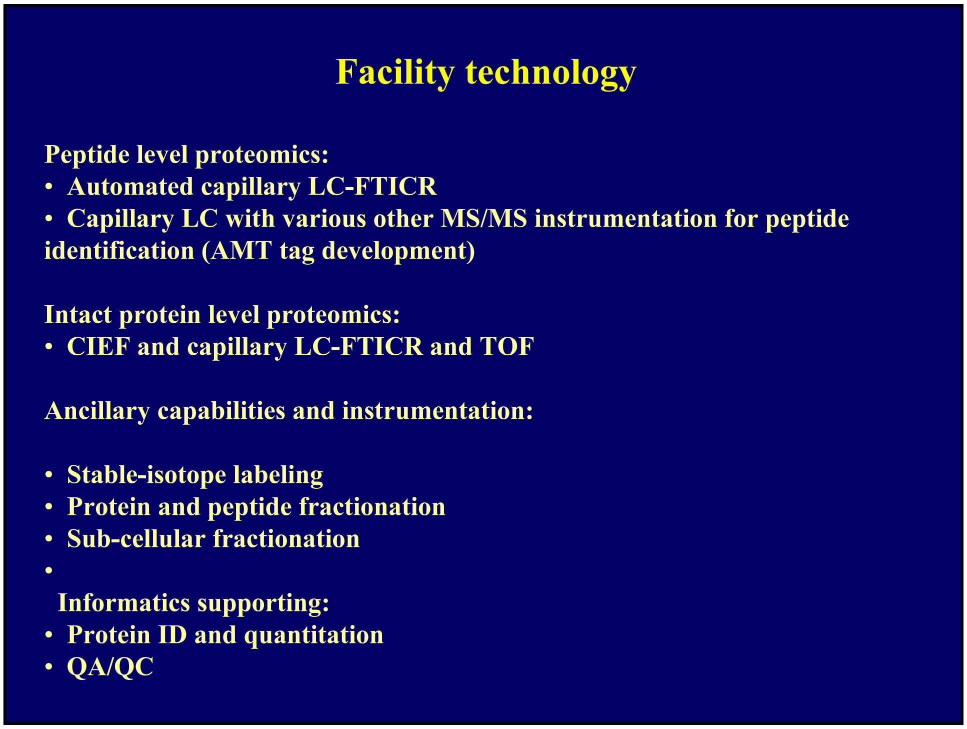
- 42 of the 43 core large ribosomal subunit proteins (58 of the 64 possible core large subunit protein isoforms) are identified in a single analysis.

⬜ : Observed w/ Modifications 🟨 : Observed w/o Modifications 🟩 Or A : Not Observed

Intact protein measurements:

- Augments peptide level analyses
- Generally much less complexity
- Much more information on protein modification states

Proc. Natl. Acad. Sci. 99, 5942-5947 (2002)



Facility technology

Peptide level proteomics:

- Automated capillary LC-FTICR
- Capillary LC with various other MS/MS instrumentation for peptide identification (AMT tag development)

Intact protein level proteomics:

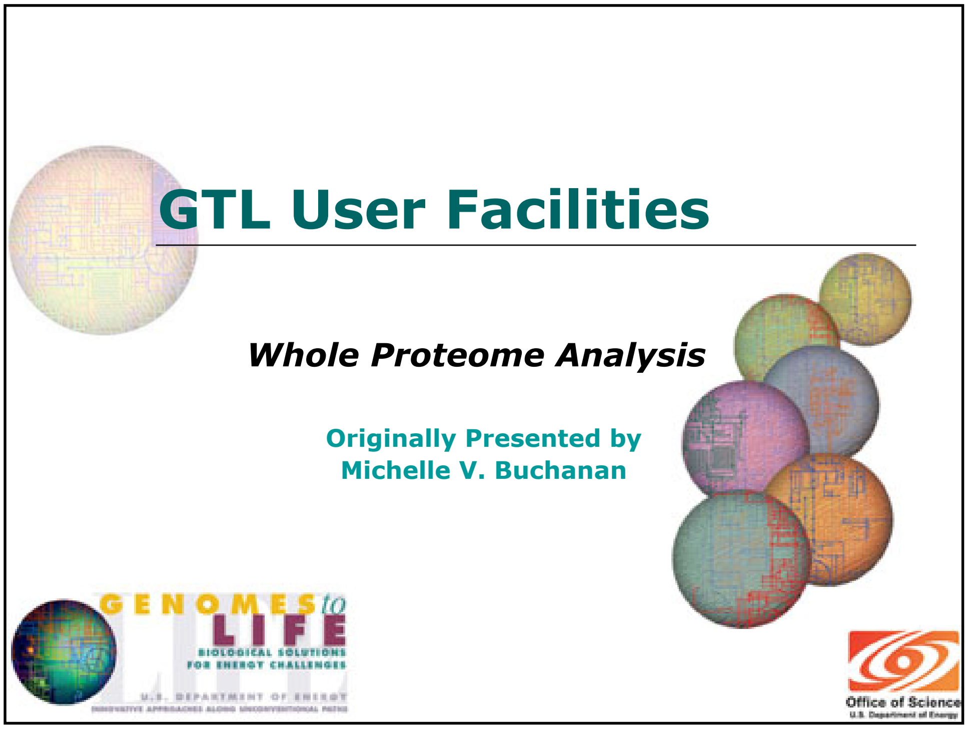
- CIEF and capillary LC-FTICR and TOF

Ancillary capabilities and instrumentation:

- Stable-isotope labeling
- Protein and peptide fractionation
- Sub-cellular fractionation
-

Informatics supporting:

- Protein ID and quantitation
- QA/QC

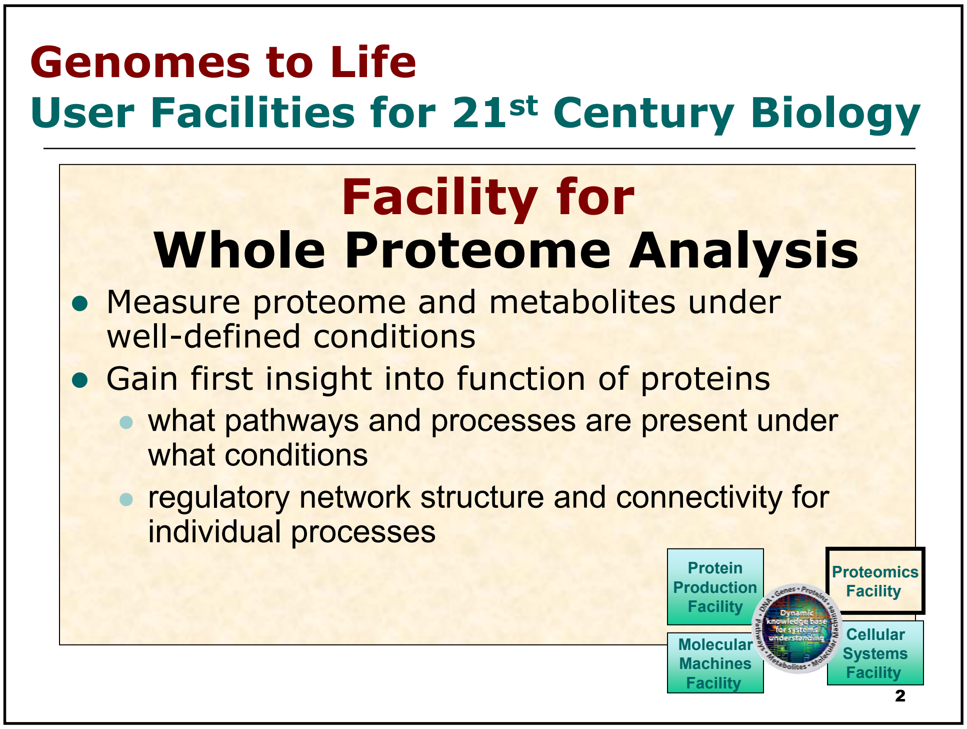


GTL User Facilities

Whole Proteome Analysis

Originally Presented by
Michelle V. Buchanan

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS
Office of Science
U.S. Department of Energy



Genomes to Life
User Facilities for 21st Century Biology

Facility for Whole Proteome Analysis

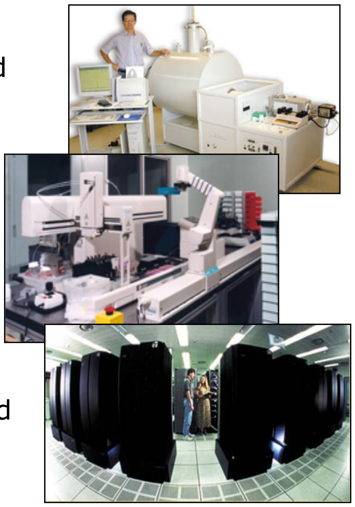
- Measure proteome and metabolites under well-defined conditions
- Gain first insight into function of proteins
  - what pathways and processes are present under what conditions
  - regulatory network structure and connectivity for individual processes

Protein Production Facility
Proteomics Facility
Dynamic knowledge base for systems understanding
DNA • Genes • Proteins • Molecular Machines • Metabolites • Pathways
Molecular Machines Facility
Cellular Systems Facility
2

GTL Facility for Whole Proteome Analysis

Key capabilities will include:

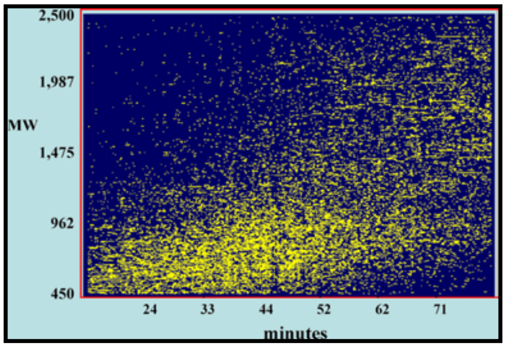
- Growth of organisms under controlled conditions
- High-throughput approaches for sample preparation prior to analysis.
- High throughput techniques for the identification and quantification of proteins, metabolites.
- New computational tools for interpretation and modeling whole proteome data
- Databases and tools for interpreting, archiving, and disseminating data and models to the greater biological community.



Genome-scale, comprehensive determination of microbial proteomes will require high throughput approaches for sample preparation and analysis

- Novel cell cultivation
- On-line analysis of metabolites
- Robotics and automation
- Chip expression assays
- Microsample handling
- Single cell analysis
- Imaging
- Computation and Informatics

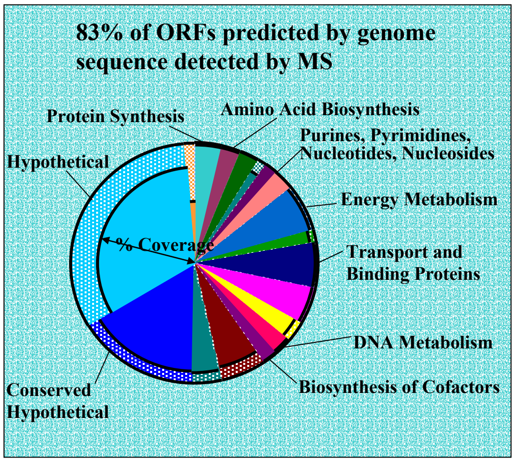
2D Gel of intact proteins



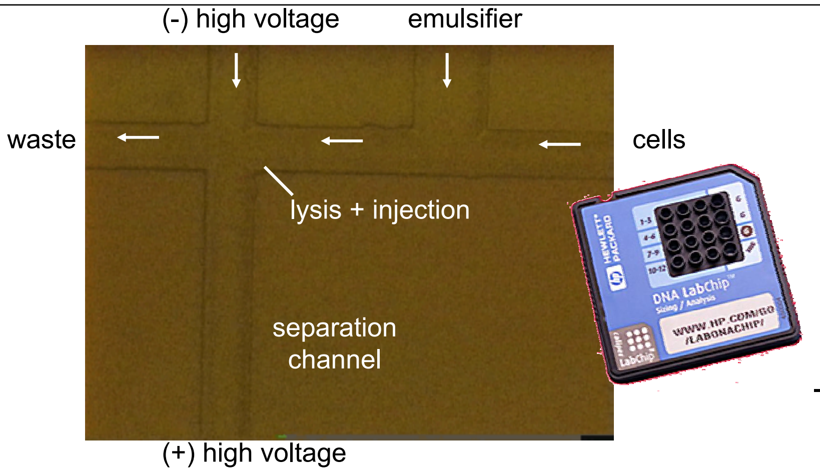
2D representation of LC/MS data from shewanella

Deinococcus radiodurans Proteome

- **BER pilot project developed MS-based approach for proteome analysis using FTICR and accurate mass tags (AMTs)**
  - **Permits proteome to be rapidly identified**
  - **Automated data collection and interpretation**



Protein Analysis of Whole Cells on Microfluidic Devices

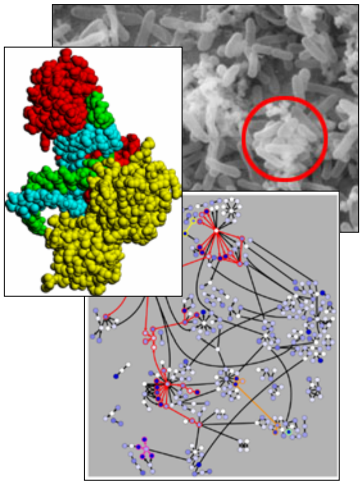


Note: arrows depict direction of flow.

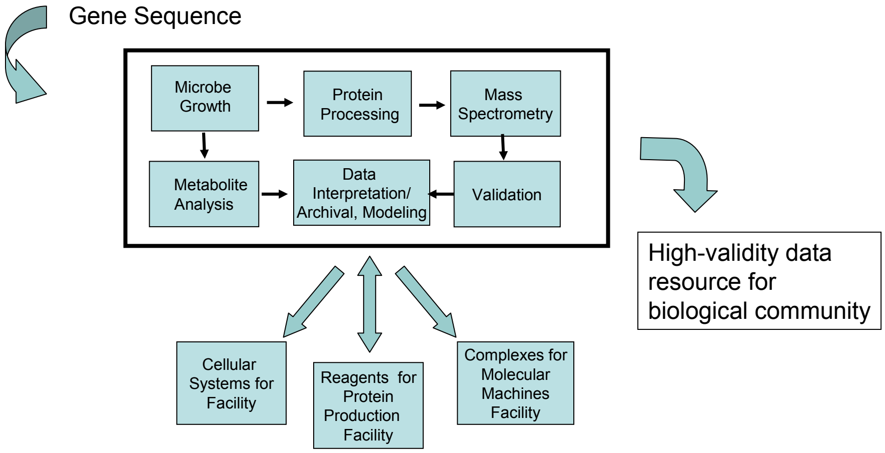
J.M. Ramsey, et al

Impact of Facility

- **High-fidelity data openly accessible to enable scientific studies**
- **Economy of scale**
- **Critical base information for proteomics and cellular systems facilities**
  - Proteome database
  - Models of pathways, processes and regulatory networks
- **Empower the general microbiological and biological communities**



Facility for Whole Proteome Analysis



Gene Sequence

Microbe Growth

Protein Processing

Mass Spectrometry

Metabolite Analysis

Data Interpretation/ Archival, Modeling

Validation

High-validity data resource for biological community

Cellular Systems for Facility

Reagents for Protein Production Facility

Complexes for Molecular Machines Facility

Appendix F: Application of Proteomics to Systems Biology

Lee Hood, Institute for Systems Biology

In this “blue sky” presentation, systems biology pioneer and visionary Leroy (Lee) Hood said that we are at an inflection point in biology and medicine, and we have the opportunity to think of doing things in completely new ways. He said it reminded him of the Human Genome Project beginning, when skepticism was expressed around the country. Similar questions were asked: “Is it really new?” and “Is it just a big fishing expedition?”

Indeed, much about systems biology is hypothesis driven, as opposed to proteomics, for example. So Hood’s view of systems biology is very different, and in explaining it, he put the emphasis on proteomics, acknowledging that we need to understand genomics as well.

Hood said he differs from his colleagues, Ruedi Aebersold in particular, in that he thinks proteomics will be democratized by microfluidics and nanotechnology. The central feature of systems biology is that it is about integrating different kinds of data. We can’t do systems biology with one-dimensional data. A question he gets asked is, “Is it different from the integrative physiology we’ve been doing for 20 years?”

The following summarizes high points from his talk.

Origins of Systems Biology: Why Now?

Systems biology is hypothesis driven, iterative, global, quantitative, and integrative.

- HGP and comparative genomics
- Cross-disciplinary science
- Internet
- Acceptance of biology as information
- High-throughput platforms—global analysis (global implies all the parameters)

All of these are leading us to systems biology approaches.

DNA represents a digital code. It starts with a central core of information—the genome—and that makes biology different from any other discipline. We’ve learned from the genome about two information types in this fundamental digital core: (1) machines and (2) superimposed gene regulatory networks that specify the behavior of genes. We know relatively little about the second area.

Systems biology networks of information also have two types: (1) the network of proteins that go together as biomodules to perform tasks in the cell, and (2) gene superimposed on that is the gene regulatory network (GRN). The linkage of a GRN is the transcription factors. In defining we can pull them apart, but in discussions we need to specify which we are discussing.

Biology and medicine should be the drivers for the technology, which in turn revolutionizes biology and medicine.

Since 1986, sequencing throughput has increased more than 3000-fold. In less than 15 years, Hood believes, there will be another 3000-fold or more increase. He anticipates doing the entire human genome for less than $1000, an accomplishment that will open up the area of predictive medicine.

Promising Technologies

Hood knows of at least seven attempts to do single-molecule sequencing. One exciting possibility is sequence by electrosorting being done by Lyle Mettendorf at LI-COR. They attach a single strand of DNA to a bead, put a primer at the end, then pass it through a microfluidic device. They can label gamma phosphates that will allow cameras to take pictures as the growing chain goes by, and, as the polymerase adds nucleotides to chain, they can read out color-coded nucleotides color by color. Other exciting features:

- Potential to do 20-kb reads.
- Does not require a clone, so it can be done in a single preparation.
- Unclonable DNA (heterochromatin) can be sequenced for the first time.

- Lends itself beautifully to organization and parallelization by microfluidics.
- Potential to do 80 to 100 times the throughput of current instruments.

Hood discussed Ruedi Aebersold's work at some length, specifically protein quantification and identification by the ICAT strategy.

Once one or two peptides from every protein are synthesized, they are quantitated. A typical mammalian cell will have 20,000 expressed proteins and about a million peptides. But Aebersold uses software to instruct the mass spectrometry (MS) to look at only matched pairs. He can go through a one-, two-, or three-dimensional separation with new MS's and run out analyses quickly. The only limitation is the cost of peptide synthesis.

Advantages

- Some 60,000 measurements will be possible once the location of the standard peptide is known.
- Each protein is uniquely identified.
- The absolute quantity of each protein is determined.
- Any subset of proteins can be interrogated.
- Data analysis for quantitative profiling becomes trivial.
- The method is portable and easily standardized.
- The process is substantially cheaper than antibody arrays.

Aebersold also is doing proteomics in serum, which will be useful for looking at diagnostic methods. Albumin makes up more than 50% of proteins in serum, and we can't see low-level proteins because of it. Albumin has no N-linked glycosylation, so Aebersold discovered how to get N-linked peptides by a chemistry in which he can open the ring and attach to a bead, wash it extensively, and release the peptide with N-glycosidase, resulting in an appropriately labeled peptide that can go into the MS. He is now testing this.

Aebersold emphasizes the quantitative proteomics experiment. This involves sample preparation and data collection, analysis, and validation. For proteomics, validation is more than 75% of the time spent, and there are no good tools for doing this yet. MS has a few software packages for sorting out good spectra from bad. Aebersold has some computational biologists looking for statistical criteria operating in context of SEQUEST, which goes from genes to proteins. His group is comparing accurate probabilities from the distribution of database search scores.

Aebersold has put together the critical computational packages for all the steps:

- LC-MS/MS data collection.
- Sequence database searching.
- Quantification.
- Data validation.

All these things lead to a predictive, preventive, and personalized medicine:

- Predictive. Probabilistic health history—DNA sequence; biannual multiparameter blood measurements (integration of RNA + DNA data).
- Preventive. Design of preventive measures via systems approaches.
- Personalized. Using systems biology to identify and manipulate networks on a personalized basis.

Moore's Law has been the driver in the information technology revolution. With sequence information, we are on a sharper curve than Moore's Law, if anything. The key is in knowing how to go from information to knowledge about an organism.

Putting this back into the context of proteomics, DOE has an opportunity to be a pioneer and leader in developing new technologies. Establishing facilities is critical, and partners must be chosen carefully. No one group can do all of this.