Technology Development

B45

Comparative Optical Mapping: A New Approach for Microbial Comparative Genomics

Shiguo Zhou[1] (szhou@lmcg.wisc.edu), Thomas S. Anantharaman[2], Erika Kvikstad[1], Andrew Kile[1], Mike Bechner[1], Wen Deng[3], Jun Wei[3], Valerie Burland[3], Frederick R. Blattner[3], Chris Mackenzie[6], Timothy Donohue[4], Samuel Kaplan[6], and **David C. Schwartz**[1,5] (dcschwartz @facstaff.wisc.edu)

[1]Laboratory for Molecular and Computational Genomics, [2]Animal Health and Biomedical Sciences, [3]Laboratory of Genetics, [4]Department of Bacteriology, and [5]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706; and [6]Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, 6431 Fannin St., Houston, TX 77030

The recent plethora of sequenced genomes has just ushered in a new era of genetics-based research. Although an impressive number of species have been, or are planned to be sequenced, the full value of such efforts will be fully accrued when patterns of variation can be discerned and annotated for many strains or isolates within a given species. As such, bacteria are an ideal place to start investigations aimed at the discernment of genome rearrangements, chromosome deletions, and horizontal transfer of foreign DNA, since these events help drive bacterial evolution. For example, genome remodeling events may cause irreversible gene loss, or add novel functionalities to an organism. Unfortunately, current approaches do not adequately identify and characterize such large-scale genomic rearrangements. The Optical Mapping System, developed in our laboratory creates high resolution maps of entire genomes, using DNA directly extracted from cells—this approach obviates the need for libraries, PCR, and probe technologies. The system uses a complex blend of single molecule technologies to enable high throughput and the construction of reliable maps. This capability has been proven by the construction and sequence comparison of 13 bacterial optical maps, and 3 parasites. Recent advances in both throughput and resolution of the Optical Mapping System has enabled genomic comparisons amongst different strains of the same species or closely related species, allowing for the pinpoint discernment of insertions, deletions and rearrangements. Comparisons of optical maps vs. in silico maps, and in silico vs. in silico maps constructed for two strains of *Yersinia pestis* (CO-92 biovar Orientalis and KIM), *E. coli* K12 and *Shigella flexneri* 2a, two strains of *S. flexneri* (2a and Y), and two strains of *Rhodobacter sphaeroides* (2.4.1 and ATCC 17029) have revealed regions of homology, insertion sites and a panoply of rearrangements. These results portend the wide use of Optical Mapping to uniquely provide genome structural details for a large number of strains, isolates or even closely related species, in ways that would complement direct sequence analysis.

B47

Optical Mapping of Multiple Microbial Genomes

Shiguo Zhou[1] (szhou@lmcg.wisc.edu), Michael Bechner[1], Erika Kvikstad[1], Andrew Kile[1], Susan Reslewic[1], Aaron Anderson[1], Rod Runnheim[1], Jessica Severin[1], Dan Forrest[1], Chris Churas[1], Casey Lamers[1], Samuel Kaplan[4], Chris Mackenzie[4], Timothy J. Donohue[2], and **David C. Schwartz**[1, 3] (dcschwartz @facstaff.wisc.edu)

[1]Laboratory for Molecular and Computational Genomics, [2]Department of Bacteriology, and [3]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706; and [4]Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, 6431 Fannin St., Houston, TX 77030

Our laboratory has developed Optical Mapping, which is a proven system for the construction of ordered restriction maps from individual DNA molecules directly extracted from cells. Such maps have utility in large-scale sequencing efforts by providing scaffolds for assembly, and as an independent means for validation, since entire genomes are mapped without the use of clones or PCR amplicons. Given the major increase in the throughput of Optical Mapping, we have used this system to map multiple microbial genomes, which included; *Thalassiosira pseudonanna* (diatom), *Enterococcus faecium*, *Pseudomonas fluorescens*, *Rhodobacter sphaeroides* and *Rhodospirillum rubrum*. These efforts were funded by DOE to help expedite and validate parallel sequencing projects. Our optical mapping data showed that *T. pseudonanna* has a haploid genome size of 33.8 Mb, possessing 22 chromosomes ranging from 658 kb to 3,322 kb, while other genomes had the following features: *E. faecium* (2.8 Mb), *P. fluorescens* (6.9 Mb, which is 1.4 Mb larger than expected size of 5.5 Mb), *R. rubrum* (4.2 Mb instead of the expected 3.4 Mb), and *R. sphaeroides* (4.2 Mb). Overall the map resolution varied from a low resolution map of average restriction fragment size of 50 kb for *R. rubrum* (*Xba*I), to a high resolution map with an average fragment size of 6.0 kb for the *R. sphaeroides* *Hind*III map. Comparison of optical maps of *R. sphaeroides* with the nascent sequence contigs of the genome sequencing project detailed the utility of optical maps in expediting sequencing projects by the identification of misassemblies within nascent sequence contigs, gap characteristics and the orientation of sequence contigs.

B49

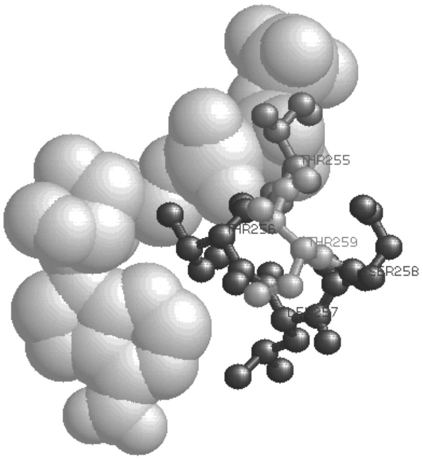
Identification of ATP Binding Proteins within Sequenced Bacterial Genomes Utilizing Phage Display Technology

Suneeta Mandava, Lee Makowski, and **Diane J. Rodi** (drodi@anl.gov)

Combinatorial Biology Unit, Biosciences Division, Argonne National Laboratory, Argonne, IL 60439

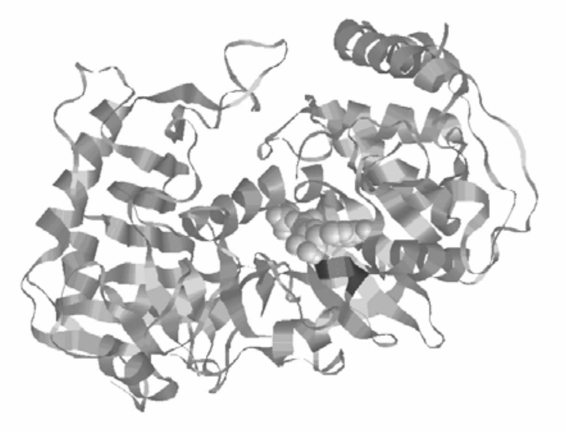
This project is applying a novel approach to genome-wide identification of small molecule binding proteins. Our ability to rapidly sequence prokaryotic genomes is generating an unprecedented amount of DNA sequence data. In spite of the large number of functional genomics tools currently available, typically about 40% of predicted ORFs remain unidentified in terms of function. Our results demonstrate that the similarity between the sequence of a protein and the sequences of phage-displayed peptides affinity-selected against small molecules can be predictive for that protein binding to the small molecule. In this project, we utilize tagged derivatives of the common metabolite ATP fixed to a solid surface as bait to capture peptide-bearing phage particles from solution. Population analysis of hundreds of these captured peptides demonstrates that subpopulations represent portions of known ATP-binding motifs.

To test our ability to predict ATP binding site locations within a protein the similarity between the sequences of our affinity selected ATP peptides and the sequences of known ATP binding proteins from the PDB were calculated. Shown here is an example of this technique applied to the ATP-binding protein phosphoenolpyruvate carboxykinase: (Renderings were carried out using RASMOL in which the segments of maximum similarity are designated in red, and blue the lowest similarity.)



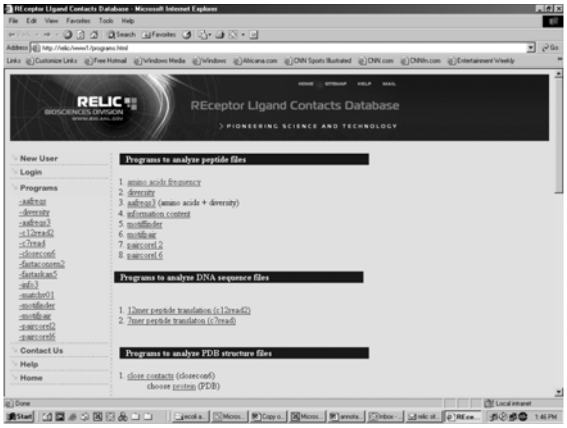
To determine whether or not this method can be used as a global approach to predicting which proteins bind ATP within a sequenced genome, as well as to identify the position of those ATP binding sites, we have developed software which compares the ATP-binding peptide pool to entire genome protein sequences. Comparison of the ATP-binding peptide populations for affinity to two sets of data, ATP-binding proteins in the Protein Data Bank and the entire *E. coli* K12 proteome confirmed that annotation of open reading frames for small molecule binding is possible using this method. Successful identification of residues within 7 Å of the ATP ligand within PDB structures is accomplished in 75% of proteins tested. Alignment of all the proteins in the *E. coli* proteome by peptide-similarity score segregates ATP binders to the top of the list when compared to control scores obtained with alternate ligand-selected pools of peptides.

The best characterized of the conserved sequence motifs that bind ATP or GTP is a glycine-rich region called a P-loop or Walker A box, which



typically forms a flexible loop between a beta-strand and an alpha-helix. In general, this loop has been shown to interact with one of the phosphate groups of the nucleotide within crystal structures. However, in two recently published P-loop-containing crystal structures, the P-loop motif is located far from the ATP-binding pocket of the protein. In these two cases maximum similarity with our ATP-binding peptide population resides within the actual ATP-binding site as opposed to the P-loop site. This discrepancy may be the result of crystal contacts that alter ATP binding, or a confirmation of the binding funnel theory of Nussinov, which describes the process of ligand binding as a series of short-lived conformational ensembles along a decreasing energy landscape. This interpretation of these observations predicts that although a P-loop motif is predictive of ATP binding, and is the initial site of an early-lived molecular handshake with ATP, the ATP ligand may not remain in the proximity of the P-loop in the final global free energy minimum conformation. This scenario similarly predicts that there may exist alternate primary sequence motifs predictive for ligand binding within the same protein sequence, such as the motif identified with our ATP-binding peptide pool.

Distribution of the software, analysis methods and data generated from this ongoing project is being accomplished by the construction of an ORACLE web-based database named RELIC (for REceptor LIgand Contacts). The architecture of the database is currently in place, and is scheduled to be made accessible to the public within the next three months, subsequent to optimization of software GUIs and a web-based users manual.


REceptor LIgand Contacts Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search Favorites
Address http://relic/www1/programs.html
Go
Links Customize Links Free Hotmail Windows Media Windows Africana.com CNN Sports Illustrated CNN.com CNNfn.com Entertainment Weekly
HOME SITEMAP HELP MAIL
RELIC
BIOSCIENCES DIVISION
WWW.BIO.ANL.GOV
REceptor LIgand Contacts Database
PIONEERING SCIENCE AND TECHNOLOGY
New User
Login
Programs
-aafreqs
-diversity
-aafreqs3
-c12read2
-c7read
-closecon6
-fastaconsen2
-fastaskan5
-info3
-matchv01
-motiffinder
-motifpair
-paircorel2
-paircorel6
Contact Us
Help
Home
Done
Local intranet
Start
1:46 PM

Programs to analyze peptide files

1. amino acids frequency
2. diversity
3. aafreqs3 (amino acids + diversity)
4. information content
5. motiffinder
6. motifpair
7. paircorel 2
8. paircorel 6

Programs to analyze DNA sequence files

1. 12mer peptide translation (c12read2)
2. 7mer peptide translaton (c7read)

Programs to analyze PDB structure files

1. close contacts (closecon6)
   choose protein (PDB)

B51

Development of Vectors for Detecting Protein-Protein Interactions in Bacteria

Peter Agron and **Gary Andersen**
(andersen2@llnl.gov)

Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551

We are interested in developing better approaches for mapping bacterial protein-protein interactions, particularly in *Caulobacter crescentus*, a model system for studying cellular differentiation and the cell cycle. Because of the advantages for high-throughput screening, our focus has been on using *Escherichia coli* as a host for two-hybrid assays. Initially, the BacterioMatch system from Stratagene was tested with 11 pairs of *Caulobacter* genes known to encode interacting proteins. Interactions were detected with several pairs, but the results were not found to be easily reproducible. Also, no interaction was observed with FtsZ, which is known to dimerize with high affinity. Therefore, we have constructed new vectors based on protein-fragment complementation with mouse dihydrofolate reductase (DHFR). In this assay, fusions to two portions of DHFR will reconstitute enzyme activity if tethered by protein-protein interactions, thus conferring trimethoprim resistance to the host as trimethoprim specifically targets the endogenous DHFR. The vectors have compatible replicons with different markers, thus allowing effective screening in *E. coli*. We are initially testing this system with bacteriophage λ cI and Caulobacter *ftsZ*, two genes encoding proteins that dimerize with high affinity. Using this system, we are also placing the interacting domains of additional *C. crescentus* gene pairs in either the 5' or 3' orientation to each of the two portions of DHFR. A library of randomly sheared *C. crescentus* DNA fragments is being placed in either orientation to the carboxy-terminal DHFR protein fragments to screen for known interactions. Based on these results we will test up to 200 additional genes of interest for protein-protein interactions with the random-fragment *C. crescentus* library.

B53

Development and Use of Microarray-Based Integrated Genomic Technologies for Functional Analysis of Environmentally Important Microorganisms

Jizhong Zhou[1] (zhouj@ornl.gov), Liyou Wu[1], Xiudan Liu[1], Tingfen Yan[1], Yongqing Liu[1], Steve Brown[1], Matthew W. Fields[1], Dorothea K. Thompson[1], Dong Xu[1], Joel Klappenbach[2], James M. Tiedje[2], Caroline Harwood[3], Daniel Arp[4], and Michael Daly[5]

[1]Oak Ridge National Laboratory; [2]Michigan State University; [3]University of Iowa; [4]Oregon State University; and [5]Uniformed Services University of the Health Sciences

Microarrays constitute a powerful genomics technology for assessing whole-genome expression levels and defining regulatory networks. Under the support of the DOE Microbial Genome Program, whole genome microarrays containing individual open reading frames were constructed for *Shewanella oneidensis* MR-1 (~4.9 Mb), *Deinococcus radiodurans* (3.2 Mb), *Rhodopseudomonas palustris* (4.8 Mb), and *Nitrosomonas europaea* (2.7 Mb) at Oak Ridge National Laboratory. DNA fragments having less than 75% similarity to other sequences in the genome were selected as specific probes using our automatic primer design program, PRIMERGEN. The majority of the probes have the size of less than 1 kb. To obtain sufficient PCR products for array fabrication, genes were amplified 8 or 16 times in a total reaction volume of 100 ul. The amplified products were then pooled and purified using automated procedures. In total, approximately 4700, 3046, 4508, and 2354 genes were amplified from the *S. oneidensis*, *D. radiodurans*, *R. palustris*, and *N. europaea* genomes, respectively. Additional sets of primers were designed for genes that did not give expected amplification products or that gave low-quality amplification. A 50-mer specific oligonucleotide was designed and synthesized for genes that did not yield desired PCR products after two attempts with PCR primers. The genome coverage for all four bacteria ranged from 95 to 99%. Evaluation of microarray quality by direct scanning, PicoGreen staining and microarray hybridization indicated that the microarray printing quality was very good in terms of spot morphology, intensity and uniformity, and the constructed microarrays have been sent to many collaborators at different institutions.

To evaluate the performance of 50-mer oligonucleotide arrays for gene expression, 96 genes from *S. oneidensis* MR-1 were randomly selected. Microarrays containing various lengths of oligonucleotides (30-70 nt) and PCR products from the same set of genes were constructed. Preliminary hybridization results indicated that the sensitivity of oligonucleotide probes is significantly lower than that of PCR products. Further in-depth comparisons are ongoing.

Genetic mutants are important to functional genomics analysis. However, mutant generation is very time-consuming and labor-intensive. Thus, a simple, high-throughput single-strand oligonucleotide mutagenesis approach has been evaluated in *S. oneidensis*. A new genetic vector containing appropriate sets of genes was constructed. Our preliminary results indicated that the developed genetic vector can replicate well in *S. oneidensis* and confer desired properties. Further work includes the introduction of single-strand oligonucleotides for targeted gene deletion. Also, a protocol for efficient transformation of MR-1 cells by electroporation was developed, which is a critical step for developing a high-throughput single-stranded oligonucleotide-based genetic system.

B55

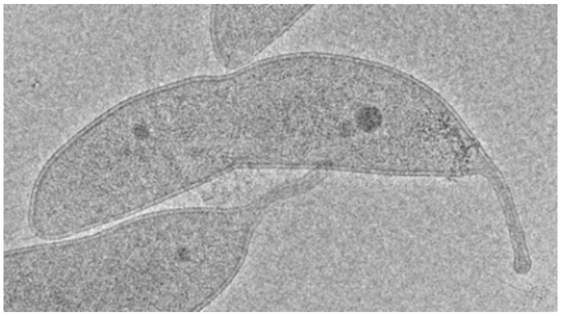
Electron Tomography of Whole Bacterial Cells

Ken Downing (khdowning@lbl.gov)

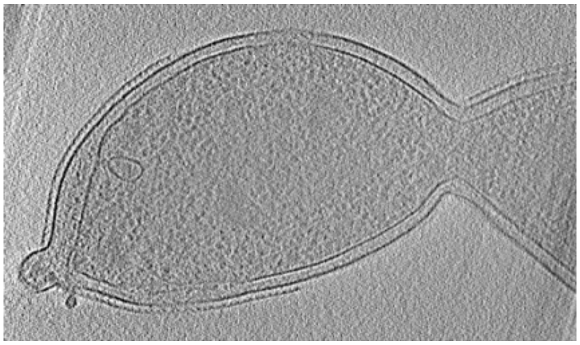
Lawrence Berkeley National Laboratory

In our initial work to develop electron tomography of intact cells and explore its limits of applicability, we have established culture and preparation conditions for a number of small microbes that may be potential targets for this work. These include *Magnetospirillum magnetotacticum*, *Caulobacter crescentis* and *Mesoplasma florum*. We have shown that we can record 2-D projection images by electron microscopy of each of these in frozen-hydrated preparations. We thus retain the native state with no stain or other contrast enhancements, but can see a wealth of internal structures. The mesoplasma is of particular interest since it is among the smallest and simplest living organisms, while these bacteria are sufficiently thin that we can obtain good data with an electron beam energy of 300-400 kV. We have collected several sets of preliminary tomographic data using facilities at the Max Planck Institute in Martinsried, Germany. 3-D reconstructions computed from these data are far more informative than the projection images. As expected, the 3-D maps show textures, representing distribution of proteins and/or nucleic acids, that vary within the organism, as well as some interesting and unexpected internal membrane structures. A number of steps need to be taken before we can begin to relate the densities seen in these reconstructions to individual protein complexes, but the preliminary data does suggest that this will indeed be feasible. Our own electron microscope, which will be especially well suited for this type of tomography, is presently being installed. Once it is in operation we will be able to further optimize our data recording protocols, including implementation of dual-axis tomograms, to improve the resolution and interpretability of the reconstructions.

Downing— Fig. 1. Electron micrograph of frozen-hydrated *Caulobacter crescentis*. The sample was rapidly frozen, with no stain or other contrast enhancement, preserving the native structure of the cell components.



Downing— Fig. 2. Section from a tomographic reconstruction of a cell undergoing division, with flagellum beginning to bud from left end. Patches of the periodic surface layer protein and internal membrane structures are visible in this section.



B57

Single Cell Proteomics—*D. radiodurans*

Norman J. Dovichi (dovichi@chem.washington.edu)

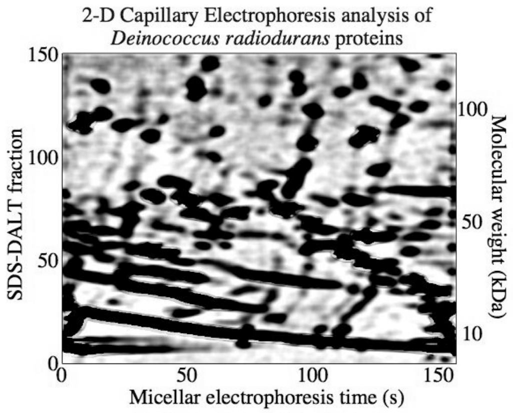
Department of Chemistry, University of Washington, Seattle, WA 98195-1700

We are developing technology to monitor changes protein expression in single tetrads of *D. radiodurans* following exposure to ionizing radiation. We hypothesize that exposure to ionizing radiation will create a distribution in the amount of genomic damage and that protein expression will reflect the extent of radiation damage.

To test these hypotheses, we will develop the following technologies:

- Fluorescent markers for radiation exposure
- DNA/rRNA determination of each cell in a *D. radiodurans* tetrad
- Two-dimensional capillary electrophoresis analysis of the protein content of a single tetrad

- Ultrasensitive laser-induced fluorescence detection of proteins separated by capillary electrophoresis



These technologies will be combined to determine protein expression in single tetrads of *D. radiodurans*, the extent of DNA damage following exposure to Cs-137 radiation, and the amount of chromosomal and rRNA per cell. This technology will be a powerful tool for functional analysis of the microbial proteome and its response to ionizing radiation.

We have generated a number of fully automated two-dimensional capillary electrophoresis separations of proteins extracted from *D. radiodurans*. The figure below presents an example, in which the proteins from *D. radiodurans* are first subjected to capillary SDS-DALT separation, which is the capillary version of SDS-PAGE using replaceable polymers. Like SDS-PAGE, SDS-DALT separates proteins based on their molecular weight, with low molecular weight proteins migrating first from the capillary. Fractions are successively transferred to a second capillary, where proteins are separated in a sub-micellar electrophoresis buffer. Components are detected with an ultrasensitive laser-induced fluorescence detector at the exit of that capillary. Over 150 fractions are successively transferred from the first capillary to the second to generate a comprehensive analysis of the protein content of this bacterium. Data are stored in a computer and manipulated to form the pseudo-silver stain image of figure 1. We estimate that there are between 200 and 300 components resolved in this separation.

B59

Genomes to Proteomes to Life: Application of New Technologies for Comprehensive, Quantitative and High Throughput Microbial Proteomics

Richard D. Smith (rds@pnl.gov), James K. Fredrickson, Mary S. Lipton, David Camp, Gordon A. Anderson, Ljiljana Pasa-Tolic, Ronald J. Moore, Margie F. Romine, Yufeng Shen, Yuri A. Gorby, and Harold R. Udseth

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352

Achieving the Genomes to Life (GtL) Program goals will require obtaining a comprehensive systems-level understanding of the components and functions that give a cell life. At present our understanding of biological processes is substantially incomplete; e.g. we do not know with good confidence all the biomolecular players in even the most studied pathways and networks in microbial systems. It is clear that many important signal transduction proteins will be present only at very low levels (~ hundreds of copies per cell) and will provide extreme challenges for current characterization methods. There is also a growing recognition of the limitations associated with gene expression (e.g. cDNA array) measurements. Increasing evidence indicates that the correlation between gene expression and protein abundances can be low, and that the correlation between gene expression and gene function is even lower. Thus, global protein characterization (proteomic) studies actually complement gene expression measurements.

Successes in genome sequencing efforts have increased interest in proteomics and also provided an informatic foundation for high throughput measurements. As a result, a key capability envisioned for success of the GtL program is the ability to broadly identify large numbers of proteins and their modification states with high confidence, as well as to measure their abundances. The challenges associated with making useful comprehensive proteomic measurements include identifying and quantifying large sets of proteins that have relative abundances spanning more than six orders of magnitude, that vary broadly in chemical and physical properties, that have transient and low levels of modifications, and that are subject to endogenous proteolytic processing. Additionally, proteomic measurements should not be significantly biased against e.g. membrane, large or small proteins. A related need is the ability to rapidly and reliably characterize protein interactions with other biomolecules, particularly their multi-protein complexes. The combined information on protein complexes and the changes observed from global proteome measurements in response to a variety of perturbations is essential for the development of detailed computational models for microbial systems and the eventual capability for predicting their response e.g. to environmental changes and mutations.

We report on development and application of new technologies for global proteome measurements that are orders of magnitude more sensitive and faster than existing technologies and that promise to meet many of the needs of the GtL program. The approaches are based upon the combination of nano-scale ultra-high pressure capillary liquid chromatography separations and high accuracy mass measurements using Fourier transform ion cyclotron resonance (FTICR) mass spectrometry. Combined, these techniques enable the use of highly specific peptide 'accurate mass and time' (AMT) tags. This new approach avoids the throughput limitations associated with other mass spectrometric technologies using tandem mass spectrometry (MS/MS), and thus enables fundamentally greater throughput and sensitivity for proteome measurements. Additional new developments have also significantly extended the dynamic range of measurements to approximately six orders of magnitude and are now providing the capability for proteomic studies from very small cell populations, and even single cells. A significant challenge for these studies is the immense quantities of data that must be managed and effectively processed and analyzed in order to be useful. Thus, a key component of our program involves the development of the informatic tools necessary to make the data more broadly available and for extracting knowledge and new biological insights from complex data sets.

The development of this new technology is proceeding in concert with its applications to a number of microbial systems (initially *Shewanella oneidensis* MR1, *Deinococcus radiodurans* R1 and *Rhodopseudomonas palustris*) in collaboration with leading experts on each organism. This research is providing the first comprehensive information on the nature of expressed proteins by these systems and how they respond to mutations in the organism or perturbations to its environment. Initial studies applying these approaches have demonstrated the capability for automated high-confidence protein identifications, broad and unbiased proteome coverage, and the capability for exploiting stable-isotope (e.g. ^{15}N) labeling methods to obtain high precision relative protein abundance measurements from microbial cultures. These initial efforts have demonstrated the most complete protein coverage yet obtained for a number of microorganisms, and have begun revealing new biological understandings.

Finally, it is projected that the AMT tag approach will be applicable for making much faster and comprehensive 'metabolome' measurements, and can also likely be extended to the characterization of proteomes (and metabolomes) of much more complex microbial communities.

This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute through Contract No. DE-AC06-76RLO 1830.

B61

New Developments in Statistically Based Methods for Peptide Identification via Tandem Mass Spectrometry

Kenneth D. Jarman, Kristin H. Jarman, Alejandro Heredia-Langner, and **William R. Cannon** (William.Cannon@pnl.gov)

Pacific Northwest National Laboratory, Richland, WA 99352

High-throughput proteomic technologies seek to characterize the state of the proteome in a cell population in much the same manner that DNA microarrays seek to characterize the state of gene expression in a cell population. Characterization of the proteins can be done using several different methods, one of which is to digest the proteins first, typically using trypsin, into peptides which are then analyzed using tandem mass spectrometry (MS/MS). A typical procedure may involve extracting cellular proteins followed by tryptic digestion and then separating the peptides with liquid chromatography. The separated peptides are then identified by MS/MS. Ideally, peptides will subsequently be quantitated, post-translational modifications will be determined and the information regarding the peptides will be assembled into a picture of the proteomic state of a cell population.

Just as with DNA microarrays, quality assurance of the high-throughput process is of paramount importance in order for proteomics to be of value to biologists. If peptides are initially identified poorly, then this information and the information on post-translational state and quantitation of protein expression is not of much value. For this reason, there has been much work recently on developing peptide identification methods for MS/MS spectra. This area of research has proceeded on two fronts, the first of which seeks to take advantage of the wide availability of genome sequences. The database search methods try to identify the peptide that resulted in the observed MS/MS spectrum by picking the best candidate from a list of peptides generated from the genome sequence. De novo methods on the other hand, seek to sequence and hence identify a peptide simply from the observed MS/MS spectrum. Regardless of which approach is used, it is essential to have a method for scoring each peptide so that accurate and reliable identifications can be made.

In this work, we present a statistically rigorous scoring algorithm for peptide identification that can be used alone, or incorporated into a database search algorithm or a de novo peptide sequencing algorithm. Our approach is based on a probabilistic model for the occurrence of spectral peaks corresponding to key partial peptide ion types. In particular, the ion frequencies for the most frequently observed ion types are initially estimated from a training dataset of known sequences. These frequencies are then used to construct a fingerprint for any candidate peptide of interest, where the fingerprint consists of a list of spectral peaks and their corresponding probabilities of appearance. A spectrum is then scored against the candidate fingerprints using a likelihood ratio between the hypothesis that the candidate peptide is not present and the hypothesis that the candidate peptide is present. This likelihood ratio can be used for peptide identification. In addition, a probabilistic score that estimates the probability of a candidate peptide being present in the test sample can be constructed from the likelihood ratio. This approach is tested using a large dataset of over 2000 spectra for tryptic peptides of different lengths ranging from 6-mer to 30-mer amino acids. Performance results indicate that this approach is accurate, and consistent across different peptide lengths and experimental conditions. False positive and false negative error rates for sequence length 10-mer and shorter are generally below 5%, while error rates for sequences longer than 10-mers are typically below 3%.

In addition, we present a Genetic Algorithm (GA) for de novo peptide sequencing. Unlike other de novo construction techniques, this methodology does not try to build amino acid chains by piecing together a feasible path through a graph using the spectral information available but starts with complete sequences and attempts to gradually find one that matches the target spectrum optimally. Due to its building approach, the GA is not immediately deterred by incomplete spectra, peaks produced by unusually occurring peptide fragments or background noise.