Oak Ridge National Laboratory and Pacific Northwest National Laboratory

8

Center for Molecular and Cellular Systems: High-Throughput Identification and Characterization of Protein Complexes

Michelle Buchanan[1], Frank Larimer[1], Steven Wiley[2], Steven Kennel[1], Dale Pelletier[1], Brian Hooker[2], Gregory Hurst[1], Robert Hettich[1], Hayes McDonald*[1] (mcdonaldwh@ornl.gov), Vladimir Kery[2], Mitchel Doktycz[1], Jenny Morrell[1], Bob Foote[1], Denise Schmoyer[1],Manesh Shah[1], and Bill Cannon[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

The Center for Molecular and Cellular Systems (CMCS) focuses on Goal 1 of the Genomics:GTL program, which aims to identify and characterize the complete set of protein complexes within a cell to provide a mechanistic basis of biochemical functions. Over the past two years, the CMCS has emphasized developing technologies that can be incorporated into a high throughput "pipeline" for the robust analysis of protein complexes. Several approaches for the isolation and identification of protein complexes from microbial cells were evaluated. Our experience has demonstrated that no single approach will be sufficient to handle the diverse types of complexes present in a cell. Thus, an integrated pipeline has been developed that uses two affinity-based approaches to isolate protein complexes in which tagged proteins are either expressed endogenously or exogenously. The individual technologies have been refined, validated and assembled into a semi-automated pipeline has been in operation for over 30 continuous weeks. A comprehensive laboratory information management system has been developed for sample tracking, process management, and data control. Experiments using over 200 tagged proteins have been conducted using *Rhodopseudomonas palustris* and *Shewanella oneidensis* cultures grown under different states. Data from the two types of pull-down approaches have been compared and have been found to provide complementary information. This suggests that both approaches are needed for comprehensive identification of protein complexes.

During the past year research tasks have been designed to improve the analysis pipeline. "Top-down" mass spectrometry has been used to identify modifications of constituent protein in the complexes. Several types of imaging tools have been employed to observe the complexes in live cells, including co-localization assays and fluorescence resonance energy transfer (FRET)-based assays. Additional effort has been placed on identifying new approaches for minimizing sample handling, such as microfluidic devices and automation. All of these research efforts have focused on development and validation of approaches to provide improved confidence of complex identification, increased sample throughput, and enhanced complex characterization.

* Presenting author

9

High-Throughput Analysis of Protein Complexes in the Center for Molecular and Cellular Systems

Vladimir Kery*[2] (vladimir.kery@pnl.gov), Dale A. Pelletier[1], Joshua N. Adkins[2], Deanna L. Auberry[2], Frank R. Collart[3], Linda J. Foote[1], Brian S. Hooker[2], Peter Hoyt[1], Gregory B. Hurst[1], Stephen J. Kennel[1], Trish K. Lankford[1], Chiann-Tso Lin[2], Eric A. Livesay[2], Tse-Yuan S. Lu[1], Cathy K. McKeown[1], Priscilla A. Moore[2], Ronald J. Moore[2], and Kristin D. Victry[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Pacific Northwest National Laboratory, Richland, WA; and [3]Argonne National Laboratory, Argonne, IL

The Genomics:GTL Center for Molecular and Cellular Systems has implemented an integrated high-throughput "pipeline" for identifying the components of protein complexes from two bacterial species of interest to the DOE: *Shewanella oneidensis*, and *Rhodopseudomonas palustris*. This integrated pipeline uses two complementary approaches to isolate and identify protein complexes using affinity-tagged proteins—an endogenous approach, and an exogenous approach. In the exogenous approach, the targets of interest are cloned in a high-throughput procedure. Proteins are then expressed in *E. coli* and purified on Ni^{2+} agarose. Dialyzed purified tagged proteins are reattached to fresh Ni^{2+} agarose and exposed to lysate from the host cell of interest, thus forming protein complexes with host target proteins *in vitro*. In the endogenous approach, plasmids expressing the tagged protein of interest are transformed into the native host, and complexes are purified by tandem affinity purification using resins selective for the hexahistidine tag and the V5 epitope. In both approaches, the complexes are eluted from the beads under denaturing conditions, digested with trypsin and identified using automated liquid chromatography/electrospray tandem mass spectrometry in combination with SEQUEST™ analysis of the data. All main liquid handling procedures in protein and protein complex purification as well as MS sample preparation and MS measurement are automated. We are completing automation of data processing and bioinformatics. A laboratory information management system (LIMS) has been implemented for integrating all aspects of sample tracking, analysis and data flow. Between ORNL and PNNL, we have to date attempted expression of nearly 400 different genes as affinity-tagged fusion proteins, completed over 5000 "pulldown" experiments on these genes (including replicates) and identified several hundred different proteins in the pulldown samples. Distinguishing authentic interactors from non-specific interactors in the identified proteins has been an important aspect of this work. Our process was validated on a number of well known bacterial protein complexes; RNA polymerase, RNA degradosome, F1F0-ATP synthase, GroESL, and others. Some interesting findings on composition of other newly identified protein complexes are being further validated and investigated (e. g. peptidoglycan biosynthesis complexes involving genes Mur A, Mur C and Mur E of *S. oneidensis* etc.). While we have initially focused our efforts on *R. palustris* and *S. oneidensis*, the processes that we have developed are universally applicable to any organism of interest. Our aim is to scale up this process to provide a fully automated capability for high-throughput analysis of protein complexes with the goal of increasing throughput that would allow characterization of greater than 5,000 complexes per year.

* Presenting author

10

Investigating Gas Phase Dissociation Pathways of Crosslinked Peptides: Application to Protein Complex Determination

Sara P. Gaucher* (spgauch@sandia.gov), Masood Z. Hadi, and Malin M. Young

Sandia National Laboratories, Livermore, CA

Chemical crosslinking is an important tool for probing protein structure[1] and protein-protein interactions.[2-3] The approach usually involves crosslinking of specific amino acids within a folded protein or protein complex, enzymatic digestion of the crosslinked protein(s), and identification of the resulting crosslinked peptides by liquid chromatography/mass spectrometry (LC/MS). In this manner, distance constraints are obtained for residues that must be in close proximity to one another in the native structure or complex. As the complexity of the system under study increases, for example, a large multi-protein complex, simply measuring the mass of a crosslinked species will not always be sufficient to determine the identity of the crosslinked peptides. In such a case, tandem mass spectrometry (MS/MS) could provide the required information if the data can be properly interpreted. In MS/MS, a species of interest is isolated in the gas phase and allowed to undergo collision induced dissociation (CID). Because the gas-phase dissociation pathways of peptides have been well studied, methods are established for determining peptide sequence by MS/MS. However, although crosslinked peptides dissociate through some of the same pathways as isolated peptides, the additional dissociation pathways available to the former have not been studied in detail. Software such as MS2Assign[4] has been written to assist in the interpretation of MS/MS from crosslinked peptide species, but it would be greatly enhanced by a more thorough understanding of how these species dissociate. We are thus systematically investigating the dissociation pathways open to cross-linked peptide species. A series of polyalanine and polyglycine model peptides have been synthesized containing one or two lysine residues to generate defined inter- and intra-molecular crosslinked species, respectively. Each peptide contains 11 total residues, and one arginine residue is present at the carboxy terminus to mimic species generated by tryptic digestion. The peptides have been allowed to react with a series of commonly used crosslinkers such as DSS, DSG, and DST. The tandem mass spectra acquired for these crosslinked species are being examined as a function of crosslinker identity, site(s) of crosslinking, and precursor charge state. Results from these model studies and observations from actual experimental systems are being incorporated into the MS2Assign software to enhance our ability to effectively use chemical crosslinking in protein complex determination.

This project is funded by the DOE Genomics:GTL Program Project lead by ORNL/PNNL.

References

1. Young, MM; Tang, N; Hempel, JC; Oshiro, CM; Taylor, EW; Kuntz, ID; Gibson, BW; Dollinger, G. "High throughput protein fold identification by using experimental constraints derived from intramolecular crosslinks and mass spectrometry." *PNAS* 2000, *97*, 5802-5806.
2. Lanman, J; Lam, TT; Barnes, S; Sakalian, M; Emmett, MR; Marshall, AG; Prevelige, PE. "Identification of novel interactions in HIV-1 capsid protein assembly by high-resolution mass spectrometry." *J. Mol. Biol.* 2003, *325*, 759-772.
3. Rappsilber, J.; Siniossoglou, S; Hurt, EC; Mann, M. "A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry." *Anal. Chem.* 2000, *72*, 267-275.
4. Schilling, B; Row, RH; Gibson, BW; Guo, X; Young, MM. "MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides." *J. Am. Soc. Mass Spectrom.* 2003, *14*, 834-850.

* Presenting author

11

Center for Molecular and Cellular Systems: Statistical Screens for Datasets from High-Throughput Protein Pull-Down Assays

Frank W. Larimer*[1] (larimerfw@ornl.gov), Kenneth K. Anderson[2], Deanna L. Auberry[2], Don S. Daly[2], Vladimir Kery[2], Denise D. Schmoyer[1], Manesh B. Shah[1], and Amanda M. White[2]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

The large-scale analysis pipeline developed at ORNL and PNNL to identify protein complexes uses a high-throughput affinity-tag "pulldown" isolation step followed by denaturing elution, trypsin digestion and combined liquid chromatography tandem mass spectrometry analysis. Each pulldown experiment identifies potential associations between the target proteins and the tagged protein in the isolation step. Any protein complex analysis method may miss some protein:protein interactions and identify artifactual associations.

We are developing informatics-based criteria for assigning significance to protein identifications and associations. Data from blanks and quality assurance standards are used identify analysis problems such as carry-over between samples. Wild-type controls and replicate pull-downs are used to estimate repeatability. Proteins that show up with statistically significant frequency in a large number of experiments are used to establish background profiles. As our Mass Spec dataset of pulldown experiments grows, it will facilitate validation of this approach. With an understanding of the experimental *noise*, a quantitative estimate of the significance of *specific* pulldown results can be estimated.

We are also evaluating published statistical frameworks for interpreting protein association data. In tandem, we have developed a statistical screen for high-throughput pulldown experiments to reduce labeling spurious associations and strengthen identification of true associations. Initial results, though promising, emphasize the difficulties in developing a valid estimator of the probability of association between two proteins.

* Presenting author

12

Center for Molecular and Cellular Systems: Analysis and Visualization of Data from a High-Throughput Protein Complex Identification Pipeline Using Modular and Automated Tools

W. Hayes McDonald[1] (mcdonaldwh@ornl.gov), Joshua N. Adkins[2], Deanna L. Auberry[2], Kenneth J. Auberry[2], Gregory B. Hurst[1], Vladimir Kery[2], Frank W. Larimer[1], Manesh B. Shah[1], Denise D. Schmoyer[1], Eric F. Strittmatter[2], and Dave L. Wabb[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN and [2]Pacific Northwest National Laboratory, Richland, WA

Global or systems level analysis of biological processes is becoming increasingly common and some of the best examples are emerging out the field of proteomics. The Center for Molecular and Cellular systems is focused on high-throughput isolation and characterization of protein complexes. The core experimental pipeline of this effort uses the parallel and complementary approaches of affinity purification of endogenously expressed tagged proteins (endogenous pulldown) and heterologously expressed tagged proteins which are then used to isolate interacting proteins out of a cell lysate (exogenous pulldown). This integrated pipeline is currently being applied to the study of protein complexes from *R. palustris* and *S. oneidensis*. After isolation, constituents of these complexes are identified using high performance liquid chromatography coupled to either tandem mass spectrometry (LC-MS/MS) or high resolution mass spectrometry (LC-MS).

The two affinity isolation and the two mass spectrometry protocols have differing analysis requirements, therefore we have modularized our data analysis and visualization tools. This gives us not only the capabilities to automate and integrate data from these different sources, but also to "plug in" and evaluate new tools readily. Each of the following modules uses one or more software tools to accomplish its task: (a) MS data extraction and preparation - including extraction of data, filtering, and output to necessary format; (b) MS search – MS/MS database searches using SEQUEST or DBDigger or MS searches against an Accurate Mass Tag (AMT) database; (c) Search result filtering and summarization – protein identification and confidence; (d) Experimental filtering – reproducibility and background subtraction built on statistical evaluation and expert analysis; (e) Network visualization – using Cytoscape to view both simple and weighted networks of interactions. Currently, we have major modules automated; future work will require the seamless integration across modules and across PNNL and ORNL. Taken together this tool set gives us not only the ability to automate our data analysis, but also to quickly explore and compare relative strengths and utilities of both the experimental and data analysis pipelines.

* Presenting author