BIOLOGICAL SCIENCES AND TECHNOLOGY

Director's R&D Fund

Elucidating Eukaryotic Gene Regulatory Networks

B. H. Jones,[1] J. R. Snoddy,[1] S. Kirov,[1] E. J. Michaud,[1] D. Schmoyer,[2] M. Leuze,[2]
L. C. Pouchard,[2] and D. K. Johnson[1]
[1]*Life Sciences Division*
[2]*Computer Science and Mathematics Division*

The availability of fully sequenced genomes from an increasing number of species creates an unprecedented opportunity to integrate experimental and computational approaches for elucidating gene regulatory networks (GRNs). For this project, we exploit ORNL's ability to analyze differential gene expression on a semigenome-wide scale using cDNA microarrays and to use large-scale computing to perform comparative genome sequence analysis among multiple genomes and implement algorithms to recognize functional elements in DNA. Our goal is to establish an ongoing ORNL program in GRN discovery.

Introduction

Coordinate regulation of gene expression provides the basis for activation and repression of specific cellular pathways and the adaptive response to changing cellular conditions. Although the coding regions of each gene involved in any given pathway are now/will soon be known as mouse and human genome sequence are annotated, the molecular mechanisms that mediate shared regulation and determine which subsets of genes interact in specific pathways are poorly understood. These molecular mechanisms, consisting of transcription factor proteins and the regions of DNA to which they bind, form the architecture of gene regulatory networks, which in turn underlie basic physiology and adaptive responses within an organism.

While GRNs are important in the cellular physiology of both unicellular and multicellular life, the GRNs in multicellular animals are likely to differ from and be more complex than those in unicellular microbes. In the Metazoa, the changes created by cascades of developmental GRNs during the life cycle of an organism help build up the complex body structure and behaviors seen within that multicellular organism.[1] The diverse and complex processes that are found in a human, a fruit fly, a nematode, or a mouse could be seen as largely the result of cascading changes of GRNs. Ultimately these networks work to transform one-dimensional genomic information into complex, multi-dimensional biological phenotypes. Because of their core importance to several aspects of biology, GRN discovery and characterization are a major goal of the DOE Genomes to Life Program. Other agencies, as well, are interested in investing more in specific areas of research that will arise from this project.

The conventional approach to identifying genes involved in specific cellular pathways and the regulatory elements that control expression of those genes has been largely experimental. Although eventually effective, the methods employed are laborious, tedious, and limited by a need for a priori knowledge of the genes and proteins involved in the pathway of interest. However, genomic sequencing has been paralleled by the development of high-throughput techniques such as microarrays that yield expression information for thousands of genes simultaneously. These techniques provide the identities of many, often hundreds, of genes that behave similarly under a specific condition(s), behavior that is due in part to the presence of common transcription factor binding sites in the noncoding regions of each gene. This "guilt-by-association" approach to identify sets of co-regulated genes continues to gain popularity as an entry point for computational studies into molecular mechanisms of co-regulation.[2] This wealth of both expression and DNA sequence data, coupled with the complexity arising from the fact that ~95% of the mammalian genome is non-coding, highlights the need for computational approaches to interpret emerging data and to make predictions that can be extended to the entire genome.

Technical Approach

Optimally exploiting the wealth of sequence information for this application requires development of new tools and integration of the diverse areas of computing and experimental biology. Accordingly, this proposal integrates expertise in functional genomics, computer sciences and computational biology resident at the laboratory. Our general scheme is as follows:

1. Identify genes with shared expression patterns from microarray experiments.
2. Develop and implement computational tools to analyze the regions of the genes likely to contain regulatory elements, using a local database.
3. Identify potential regulatory elements for subsets of the genes.
4. Use bioinformatics to collect functional and pathway information for each subset of genes.

Details of the experimental and computational approaches are outlined separately below:

Experimental Approach

Microarray data can be used to identify genes differentially expressed in a specific condition or mutant mouse model, to group genes according to shared regulatory patterns across multiple experimental conditions, and to classify unknown samples based on their expression profiles. We used microarrays as a high-throughput way to identify sets of genes with shared expression patterns under specific conditions. We chose the following two experimental models: primary mouse keratinocytes during the time course of differentiation and a unique ORNL mouse model of hair loss, near naked hairless (Hr^n). These mice carry a regulatory mutation in the hairless (*hr*) gene that manifests as a sparse ($Hr^n/+$) or nonexistent (Hr^n/Hr^n) coat of hair shortly after birth.[3] The *hr* gene encodes a putative transcription factor, making the phenotype of *hr* mutants an excellent starting point for the studies included herein. Collectively these two models allowed us to examine gene expression related to aspects of skin biology, an area of long-standing interest in the ORNL mouse program.

We printed our microarrays in-house using a collection of cDNAs enriched for genes involved in hair and skin development and physiology, cell cycle control, DNA repair, apoptosis, and oncogenesis. We used cDNA clones that were originally obtained from Research Genetics and were amplified and sequence verified at ORNL in an earlier project. Microarrays were fabricated from these clones according to standard protocols. Each array represented ~1000 unique mouse genes, and each gene was spotted in triplicate.

Bioinformatics Approach

The work described herein requires vast amounts of DNA sequence accession and manipulation. To facilitate this effort as well as to make it possible to easily access multiple types of information about genes, we decided to create a local version of the National Center for Biotechnology Information (NCBI) database and to expand our database to include other information about genes, such as gene ontology (GO) assignment. Such a database would allow querying and processing sets of genes in the high-throughput manner necessary for this project.

Computational Approach

A common core task in the elucidation of gene regulatory networks is to identify the transcription factor (TF) binding sites that are shared by co-regulated genes. Identification of TF binding sites is complicated by the fact that these motifs are often small (8–15 bp), and that a single TF protein can bind to variations in a consensus sequence for that protein.[4] In addition, many true binding sites exist in a single gene, making it difficult to visually compare all of the sites from collections of co-regulated genes. Furthermore, the task is compounded in eukaryotes, in which collections of binding sites referred to as modules or composite elements, rather than individual sites, act synergistically to affect enhancement or suppression of transcription. Therefore the task of identifying the sets of DNA binding sites that mediate gene response to even a single biochemical factor is daunting.

Two main strategies were employed for computational identification of TF binding sites. The first approach involves searching for known TF binding sites in DNA sequence of interest. This tactic is based on consensus binding-site sequences archived in a library of binding sites, such as TRANSFAC (http://transfac.gbf.de/TRANSFAC/). While known binding sites can sometimes be identified with a high degree of confidence, this approach is limited to sites present in the database, and those that have been very well annotated experimentally. An alternative approach, and the one we employed herein, is to develop algorithms that search for de novo binding sites. This method allows the identification of novel sites not yet described, as well as known sites.

Results and Accomplishments

Our experimental effort has led to the identification of sets of genes that we believe to be important in the etiology of hair loss in the ORNL mouse model Hr^n. We first used microarrays to identify a set of genes that were differentially expressed in skin of adult Hr^n mutant mice compared to their wild-type littermates. Further characterization of this list of genes identified many known to be important in the development and maintenance of the hair follicle. We followed these results by repeating the analyses in young animals (7 days of age) in which the lack of hair phenotype was just beginning to manifest, with the goal identifying the initial genes that respond to the hairless protein and are primary rather than secondary responders to the mutation. Many of the genes identified in adult skin were also differentially expressed in skin of young mutant animals, underscoring their potential role in development of the phenotype. Of particular interest is

that many of the genes we identified as differentially expressed reside in gene clusters. These clusters have been proposed to be regulated as a group with specific timing of expression for each gene related to its position within the cluster, much like has been demonstrated for the Hox gene clusters. This raises the possibility that the hairless protein acts as a regulatory switch for the clusters of genes that are required for normal hair follicle development. This possibility is being pursued and will likely become the focus of an upcoming proposal to NIH. Collectively, this set of genes with altered regulation in response to a mutation in *hr* should contain many genes that share mechanisms of co-regulation, making them an excellent input set for the computation efforts. These genes are currently being parsed through the analysis algorithms, and we expect that the results will be included in an upcoming manuscript.

To facilitate the extraction of DNA sequence and the processing and querying of sets of genes, we created a gene key database, GeneKeyDB, to relate our genes of interest to standard databases such as NCBI's RefSeq and LocusLink (see Fig. 1). The current GeneKeyDB is largely a relational representation of the LocusLink data with some additional tables to identify the genomic locations of the genes and interactions with other databases at ORNL. Unlike the NCBI organization of this data, the current database allows for processing sets of genes by computerized analysis, and not just inferencing by

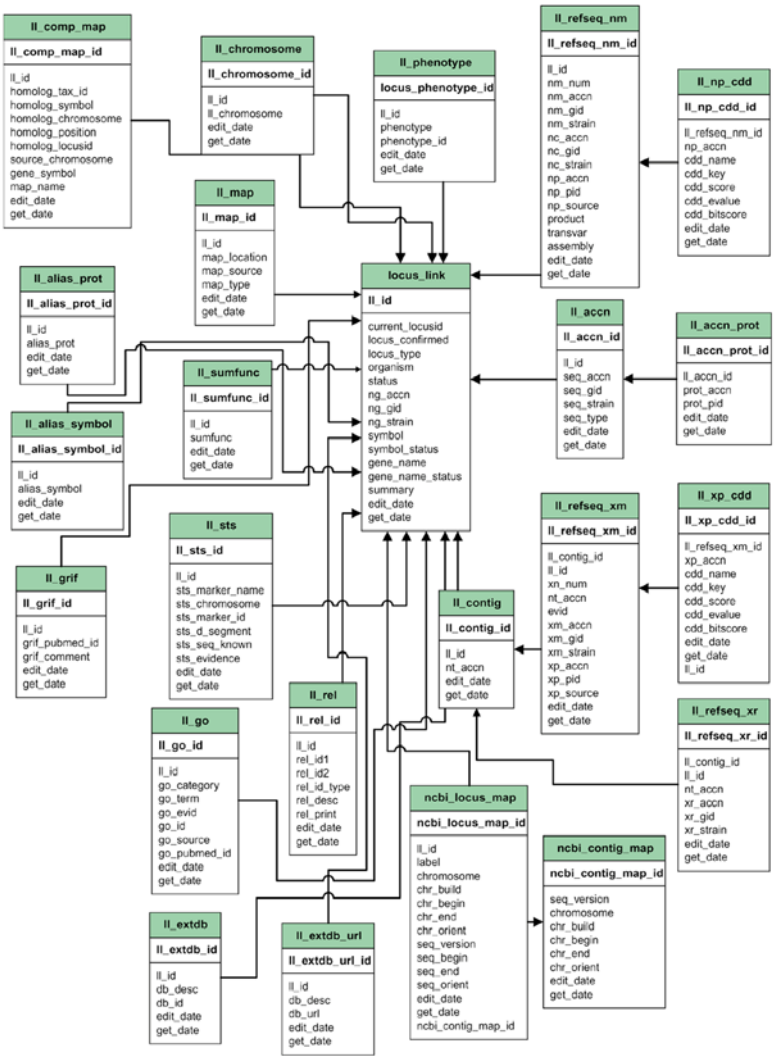


Fig. 1. Schema of the GeneKeyDB database created as a repository for data from NCBI and other sources and as a source for processing sets of genes with high-throughput.

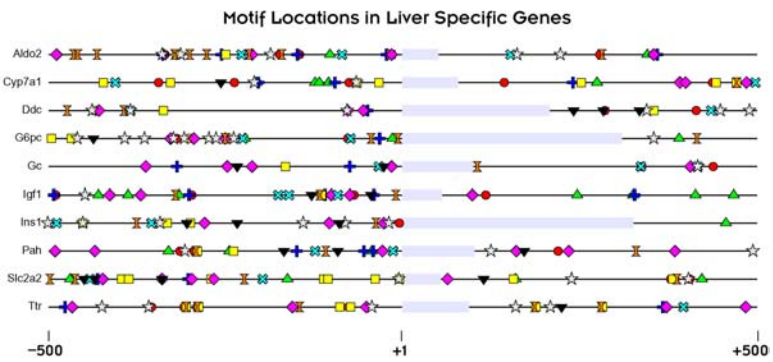


Fig. 2. Example output of the motif discovery toolkit applied to a set of genes co-regulated in liver, as described by Krivian and Wasserman.[5]

analyzing one gene at a time by humans using web pages and web links. The current GeneKeyDB links to our expression array database (GIMS) and the ORNL gene annotation database (GGPIW).

The suite of novel computational tools that we have developed for the motif discovery toolkit constitute a new approach for identifying functional DNA in non-coding regions by sequence comparison. Our approach differs from other motif-finding methods in a number of important ways, which collectively make it unique:

- it is applicable to any collection of related genomic sequences,
- it does not depend on the global alignment of these sequences,
- it does not require a motif to be present in each sequence of the collection, and it is able to discover multiple copies of a motif in a single sequence,
- it can find multiple motifs in a single run, and
- it is a general computational method; it does not require specific biological information about the sequences being examined.

Our motif discovery toolkit contains tools for a large number of tasks, including filtering sequences, finding short inexact matches, combining short matches, constructing graphs to represent relationships between motifs, identifying interesting graph structures, producing position weight matrix and profile hidden Markov models of motifs, scoring sequences relative to a motif model, and searching for inexact inverted and direct palindromes. An example of the toolkit as applied to a set of genes co-regulated in liver[5] is shown in Fig. 2.

A significant feature of our motif discovery toolkit is its extensive use of graph algorithms. The integration into the motif discovery toolkit of a number of innovative graph algorithms, some of which execute three orders of magnitude faster than competitors, will allow consideration of problems much larger than any previously addressed. Additionally, although the motif discovery toolkit uses global search methods, it examines only those patterns actually found in the input sequences, rather than all 4^n possible patterns for an n-base sequence, thus greatly improving efficiency. As a result, the graphs we examine are generally smaller than would be the case if an edge between two patterns were weighted only by the number of matches. Our metric tends to connect only those patterns for which there is a biologically significant relationship. Collectively, these new ideas make the toolkit well suited to the task of motif discovery.

Summary and Conclusions

In summary, we have developed a set of novel computational tools that can be customized for a variety of applications within the general area of identifying DNA regulatory elements. We have also created a local database that allows for rapid retrieval and processing of sets of genes from genomic databases. Experimentally, we have identified an interesting set of genes that we believe to be disregulated as a direct consequence of a mutation in the transcription factor *hr*, and these genes serve as an input dataset for further analysis using our motif discovery toolkit. The results of our work have been submitted in manuscript form to The Eighth Annual Conference on Research in Computational Molecular Biology (RECOMB 2004). A manuscript focusing on the biology and molecular mechanisms of the Hr^n mutants is in preparation. The infrastructure, collaborations, and tools that we have developed during this project provided the basis for two proposals submitted in FY 2003 to DOE Genomes to Life and to the Joint DMS/NIGMS Initiative to Support Research Grants in the Area of Mathematical Biology, NSF 02-125. In addition, the resources developed in this project will serve as the basis for continuing collaborative efforts between the Life Sciences and Computer Science and Mathematics Divisions, as well as with the Department of Computer Science at the University of Tennessee, Knoxville.

References

[1]V. F. Hinman, A. T. Nguyen, R. A. Cameron, and E. H. Davidson, "Development of Gene Regulatory Network Architecture Across 500 Million Years of Echinoderm Evolution," Proceedings of the National Academy of Science, *PNAS* **100**, 13356 (2003).

[2]J. Q. Quackenbush, *Science* **302**, 240 (2003).

[3]K. F. Stelzner, *Journal of Heredity* **74**, 193–6 (1983).

[4]W. W. Wasserman and W. Krivian, *Naturwissenchaften* **90**, 156–66 (2003).

[5]W. Krivian and W. W. Wasserman, *Genome Research* **11**, 1559–66 (2001).

High-Throughput Analysis and Modeling of Protein Complexes

F. W. Larimer,[1] Y. Xu,[1] D. Xu,[1] P. F. Britt,[2] R. L. Hettich,[2] and G. Hurst[2]
[1]*Life Sciences Division*
[2]*Chemical Sciences Division*

We developed methods to accelerate the study protein complexes by covalently crosslinking the protein subunits of a complex, digesting the crosslinked complex enzymatically to produce crosslinked pairs of peptides, and subsequently identifying these crosslinked pairs of peptides using mass spectrometry. These crosslinked peptides thus define a potential interaction face of the subunits making up the complex. We have examined a biotin-labeled, heterobifunctional (lysine-reactive N-hydroxysuccinimide ester and non-selective photoactivatable groups) crosslinking reagent, sulfosuccinimidyl [2-6-(biotinamido)-2-(p-azidobenzamido)-hexanoamido]ethyl-1,3'-dithiopropionate (sulfo-SBED), to understand the chemistry, side reactions, affinity isolation, and mass spectrometric behavior of sulfo-SBED crosslinked peptides.

Introduction

A major goal this project was to develop techniques for studying protein complexes by covalently crosslinking the protein subunits of the complex, digesting the crosslinked complex enzymatically to produce crosslinked pairs of peptides, and to identify these crosslinked pairs of peptides using mass spectrometry. Because of generally low yields from crosslinking reactions, we proposed to use affinity-labeled crosslinking reagents that would allow selective extraction of crosslinked peptide pairs from the many non-crosslinked peptides that would result from the proposed scheme. A commercially available reagent, sulfo-SBED (Fig. 1), was chosen because it contains a biotin affinity group, which has a strong affinity for the protein avidin. Separation media derivatized with avidin are also commercially available.

Technical Approach

To understand the chemistry, side reactions, affinity isolation, and mass spectrometric behavior of sulfo-SBED crosslinked peptides, we reacted sulfo-SBED with a small peptide, neurotensin. Bovine hemoglobin was chosen as a model protein complex for developing crosslinking methodology because it is commercially available in reasonable purity and at low cost. Because of the complications that we encountered in identifying crosslinked peptide pairs from the sulfo-SBED crosslinking of bovine hemoglobin, it was not possible to study crosslinking of mixtures of protein complexes.

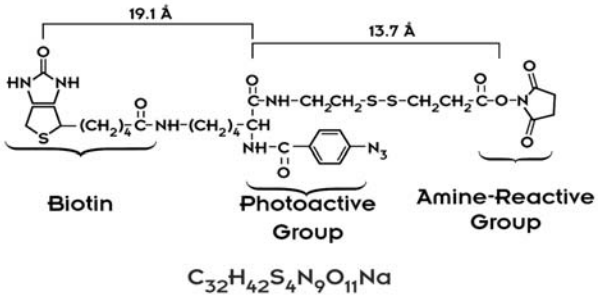


Fig. 1. Sulfo-SBED (Pierce).

Results and Accomplishments

Reaction of sulfo-SBED with a peptide model system. As described in a manuscript submitted to the *Journal of the American Society for Mass Spectrometry*, we studied the reaction of the crosslinking reagent sulfo-SBED with neurotensin, a small peptide.[1] This manuscript describes side reactions such as "one-sided" products between the crosslinking reagent and the peptide, and oxidation of the biotin sulfur in sulfo-SBED. A "double-blocking" protocol is also described. This protocol was found to reduce non-specific binding of non-crosslinked peptides to the avidin affinity chromatography medium used to isolate crosslinked peptides. We demonstrated the detection of neurotensin/sulfo-SBED reaction products in the presence of a background of non-biotinylated peptides.

Reaction of Sulfo-SBED with bovine hemoglobin. Using techniques similar to those described in the sulfo-SBED/neurotensin manuscript, we crosslinked bovine hemoglobin with sulfo-SBED. Hemoglobin is protein complex containing four protein subunits—A, B, C, and D. Subunits A and C are of identical amino acid sequence (the α chain), as are subunits B and D (the β chain). Each subunit contains a heme group, noncovalently bound

between two histidine residues. The spatial relationships among the four subunits in hemoglobin are known, and several three-dimensional structures for this protein are available in the Protein Data Bank.

Detection of Intact Sulfo-SBED Crosslinked Hemoglobin

To characterize initially the reaction between sulfo-SBED and bovine hemoglobin, we examined the behavior of the intact crosslinked protein prior to digestion by trypsin. Figure 2 shows the affinity isolation of the crosslinked complex, as well as a non-crosslinked control. In this experiment, sulfo-SBED crosslinked hemoglobin was fractionated using monomeric avidin column chromatography.[2] Unreacted hemoglobin is not retained by the column, and elutes in fractions 1–3. Sulfo-SBED crosslinked hemoglobin is eluted from the column in fractions 8–10 via addition of a low-pH buffer.

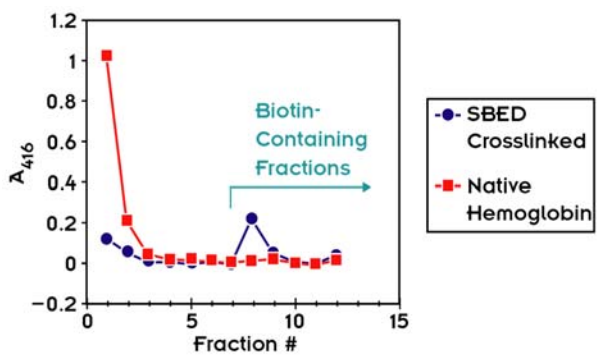


Fig. 2. Avidin fractionation of sulfo-SBED crosslinked hemoglobin.

Various fractions from the avidin separation of sulfo-SBED crosslinked and control hemoglobin were analyzed using matrix-assisted laser desorption time-of-flight mass spectrometry (MALDI-TOF-MS), as shown in Fig. 3. The sulfo-SBED-treated sample shows addition of the crosslinking reagent to both the α and β chains [Fig. 3(d)]. After affinity isolation, however, only the α chain and its sulfo-SBED reaction product appear [Fig. 3(f)]; little or no β chain is observed. These measurements indicate that the separation shown in Fig. 2 indeed yields crosslinked hemoglobin in fraction 8.

Significantly, it appears that the crosslinking reaction allows hemoglobin to retain sufficient tertiary structure to maintain the non-covalently bound heme groups. Evidence for this is that detection of hemoglobin in the fractions eluted from the avidin column was performed using optical absorbance at 416 nm, a characteristic absorption band associated with the heme group. A significant distortion of the hemoglobin structure by the crosslinking reaction could cause loss of the heme group.

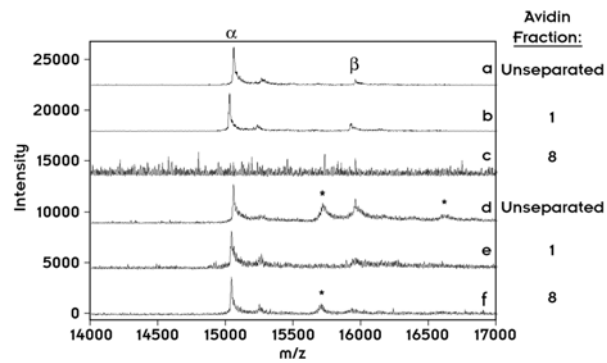


*Fig. 3. MALDI-TOF spectra of hemoglobin. (a,b,c) not crosslinked. (d,e,f) sulfo-SBED crosslinked. * = addition of sulfo-SBED.*

Detection of Crosslinked Peptides in Trypsin-Digested, Crosslinked Hemoglobin

Using conditions under which hemoglobin could be crosslinked using sulfo-SBED, and the procedures described in the sulfo-SBED/neurotensin manuscript, the next step was trypsin digestion of sulfo-SBED crosslinked hemoglobin, followed by avidin isolation of crosslinked peptide pairs, and mass spectrometric identification of these crosslinked peptide pairs. Despite the improvements described in the manuscript, it still proved difficult to detect and identify crosslinked peptide pairs. The probable reason for this is the complexity of the peptide mixture resulting from the digestion, and the low amounts of crosslinked peptides in this mixture. Initial yields of crosslinking reactions are typically 1–10% (or less). Furthermore, digestion results in an *increase* in the number of species present in the mixture, by a factor of roughly 10–50. While the avidin separation should mitigate this problem, there remains a considerable background of non-crosslinked peptides in mass spectra of avidin-purified digests.

Figure 4 shows a narrow range of the MALDI-TOF mass spectra of avidin-purified trypsin digests of a non-crosslinked bovine hemoglobin sample and a sulfo-SBED crosslinked hemoglobin sample. Although non-specific peptides appear in both spectra, one peak at mass-to-charge

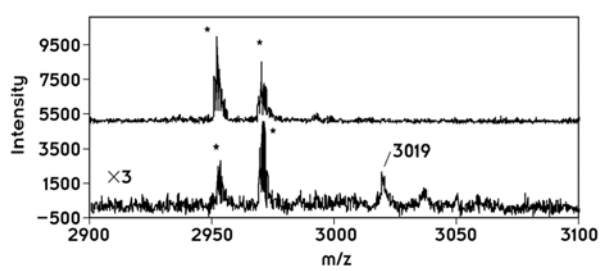


*Fig. 4. MALDI-TOF mass spectra of tryptic digests of (a) non-crosslinked hemoglobin, (b) sulfo-SBED crosslinked hemoglobin. * represents peptides binding non-specifically to avidin affinity medium.*

Table 1. Comparison of putative crosslinks from mass spectrometry with PDB model of hemoglobin 1HDA

Observed (m/z)	Calculated (m/z)	Peptide 1[a] (subunit) RESIDUES positions	Peptide 2[a] (subunit) RESIDUES positions	Distance range[c] (Å)
1924.7	1924.9	(B) MLTAEEK β1–7[b]	(A) GHGAK α57–61	32–37
	"	(D) MLTAEEK β1–7[b]	(A) GHGAK α57–61	41–45
	"	(B) MLTAEEK β1–7[b]	(C) GHGAK α57–61	42–46
	"	(D) MLTAEEK β1–7[b]	(C) GHGAK α57–61	31–36
3019.8	3020.5	(B) AHGKK β61–65	(D) VKAHGKKVLDSFSNGMK β59–75	45–57
"	"	(B) AHGKK β61–65	(D) VKAHGKKVLDSFSNGMK β59–75	54–59
"	"	(B) AHGKK β61–65	(D) VKAHGKKVLDSFSNGMK β59–75	52–57
"	"	(B) AHGKK β61–65	(D) VKAHGKKVLDSFSNGMK β59–75	51–56
"	"	(D) AHGKK β61–65	(B) VKAHGKKVLDSFSNGMK β59–75	45–57
"	"	(D) AHGKK β61–65	(B) VKAHGKKVLDSFSNGMK β59–75	54–59
"	"	(D) AHGKK β61–65	(B) VKAHGKKVLDSFSNGMK β59–75	52–57
"	"	(D) AHGKK β61–65	(B) VKAHGKKVLDSFSNGMK β59–75	51–56
"	3020.5	(B) VKAHGKK β59–65	(D) AHGKKVLDSFSNGMK β61–75	22–32
"	"	(B) VKAHGKK β59–65	(D) AHGKKVLDSFSNGMK β61–75	18–28
"	"	(B) VKAHGKK β59–65	(D) AHGKKVLDSFSNGMK β61–75	25–30
"	"	(B) VKAHGKK β59–65	(D) AHGKKVLDSFSNGMK β61–75	28–33
"	"	(D) VKAHGKK β59–65	(B) AHGKKVLDSFSNGMK β61–75	12–30
"	"	(D) VKAHGKK β59–65	(B) AHGKKVLDSFSNGMK β61–75	12–29
"	"	(D) VKAHGKK β59–65	(B) AHGKKVLDSFSNGMK β61–75	24–28
"	"	(D) VKAHGKK β59–65	(B) AHGKKVLDSFSNGMK β61–75	25–29
3524.0	3523.8	(A) VLSAADKGNVKAAWGK α1–16[b]	(B) LLVVYPWTQR β30–39	14–21
"	"	(A) VLSAADKGNVKAAWGK α1–16	(B) LLVVYPWTQR β30–39	15–26
"	"	(A) VLSAADKGNVKAAWGK α1–16	(B) LLVVYPWTQR β30–39	18–30
"	"	(A) VLSAADKGNVKAAWGK α1–16[b]	(D) LLVVYPWTQR β30–39	26–31
"	"	(A) VLSAADKGNVKAAWGK α1–16	(D) LLVVYPWTQR β30–39	32–38
"	"	(A) VLSAADKGNVKAAWGK α1–16	(D) LLVVYPWTQR β30–39	33–40
"	"	(C) VLSAADKGNVKAAWGK α1–16[b]	(B) LLVVYPWTQR β30–39	26–31
"	"	(C) VLSAADKGNVKAAWGK α1–16	(B) LLVVYPWTQR β30–39	32–38
"	"	(C) VLSAADKGNVKAAWGK α1–16	(B) LLVVYPWTQR β30–39	33–40
"	"	(C) VLSAADKGNVKAAWGK α1–16[b]	(D) LLVVYPWTQR β30–39	15–21
"	"	(C) VLSAADKGNVKAAWGK α1–16	(D) LLVVYPWTQR β30–39	15–26
"	"	(C) VLSAADKGNVKAAWGK α1–16	(D) LLVVYPWTQR β30–39	18–30
"	3523.9	(B) AAVTAFWGKVK β8–18	(B) LLGNVLVVVLARNFGK β104–119	9.5–23
"	"	(B) AAVTAFWGKVK β8–18	(D) LLGNVLVVVLARNFGK β104–119	41–58

Table 1 (continued)

Observed (m/z)	Calculated (m/z)	Peptide 1[a] (subunit) RESIDUES positions	Peptide 2[a] (subunit) RESIDUES positions	Distance range[c] (Å)
“	“	(D) AAVTAFWGKVK β8–18	(B) LLGNVLVVVLARNFGK β104–119	41–58
“	“	(D) AAVTAFWGKVK β8–18	(D) LLGNVLVVVLARNFGK β104–119	9.5–23
“	3523.7	(B) VLDSFSNGMKHLDDLK β66–81	(A) MFLSFPTTK α32–40	22–32
“	“	(D) VLDSFSNGMKHLDDLK β66–81	(A) MFLSFPTTK α32–40	28–37
“	“	(B) VLDSFSNGMKHLDDLK β66–81	(C) MFLSFPTTK α32–40	28–37
“	“	(D) VLDSFSNGMKHLDDLK β66–81	(C) MFLSFPTTK α32–40	22–32
3539.1	3539.8	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	43–45
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	43–45
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	43–45
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	45–47
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	22–24
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	26–27
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	26–27
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(A) YR α140–141	36–37
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	22–24
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	26–27
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	26–27
“	“	(B) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	36–37
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	43–44
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	43–45
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	44–45
“	“	(D) VKAHGKKVLDSFSNGMKHLDDLK β59–81	(C) YR α140–141	45–47
“	3538.6[d]	(B) GTFAALSELHCDKLHVDPENFK β82–103	(A) GNVK α8–11	38–43
“	“	(D) GTFAALSELHCDKLHVDPENFK β82–103	(A) GNVK α8–11	67–71
“	“	(B) GTFAALSELHCDKLHVDPENFK β82–103	(C) GNVK α8–11	44–50
“	“	(D) GTFAALSELHCDKLHVDPENFK β82–103	(C) GNVK α8–11	67–73

[a]Residue containing NHS–reactive group is underlined.
[b]NHS reactive group is N–terminal amine of protein.
[c]distance between Cα in NHS–reactive residue and Cα in residues of the other peptide.
[d]Mass match based on oxidized peptide pair.

ratio (m/z) 3019 appears only in the crosslinked sample. Table 1 lists other m/z values for peaks observed similarly in mass spectra of digested crosslinked samples but not observed in the mass spectra of the corresponding non-crosslinked control experiment.

To determine possible identities of the peaks representing potential crosslinked pairs identified by mass spectrometry, a comparison was carried out between observed m/z values and calculated m/z values for all possible complete and incomplete tryptic fragments, inter- and intra-subunit sulfo-SBED crosslinked peptide pairs, and known side products of the sulfo-SBED reaction with bovine hemoglobin.[3] (There are approximately 65,000 possible products.) Table 1 also lists the resulting matches between observed and predicted m/z values; several matches were found within a small m/z error (<1 Da) of each observed peak. These possible matches were then evaluated to determine whether they were consistent with the known tertiary and quaternary structure of bovine hemoglobin, using the Protein Data Bank (PDB) structure 1HDA. The estimated maximum spanning length of sulfo-SBED between alpha carbons of two amino acid residues is approximately 29Å. Therefore, two peptides that have been crosslinked by sulfo-SBED must be separated by 29Å or less in the quaternary structure of hemoglobin. Table 1 lists distances between alpha carbons of relevant amino acid residues in possible crosslinked peptide pairs, as calculated from the PDB 1HDA structure. Because the m/z selection procedure described above cannot distinguish between the two α subunits (A and C) and the two β subunits (B and D) of hemoglobin, all possible combinations of crosslinked peptides are included in Table 1. The two reactive groups of the crosslinking reagent sulfo-SBED are an N-hydroxysuccinimide (NHS) ester, which attaches specifically to the primary amine groups on the side chains of lysine residues and at N-termini of proteins, and a photoactivated arylazide that inserts non-specifically into C-H and N-H bonds. Thus one end of the sulfo-SBED crosslinker is attached to a lysine residue, or to a residue that occupied the N-terminus of a hemoglobin subunit; the other end of the sulfo-SBED crosslinker may be attached to any residue. The distances shown in Table 1 represent a range of distances between a lysine (or terminal) residue on one peptide and the alpha carbons of *all* residues in the other peptide. If the maximum span of sulfo-SBED falls within this range, the pair of peptides could plausibly be crosslinked. If the range of distances separating the peptides is greater than the length of sulfo-SBED, it is unlikely that the pair could be crosslinked without distortion of the hemoglobin structure.

While further proof in addition to the mass spectrometric identification and comparison with the PDB model is required, we have identified the following peptides in bovine hemoglobin that were crosslinked by sulfo-SBED:

β 59–65 of subunit B to β 61–75 of subunit D
β 59–65 of subunit D to β 61–75 of subunit B
α 1–16 of subunit A to β 30–39 of subunit B
α 1–16 of subunit A to β 30–39 of subunit D
α 1–16 of subunit C to β 30–39 of subunit B
α 1–16 of subunit C to β 30–39 of subunit D
β 8–18 of subunit B to β 104–119 of subunit B
β 8–18 of subunit D to β 104–119 of subunit D
β 66–81 of subunit B to α 32–40 of subunit A
β 66–81 of subunit D to α 32–40 of subunit A
β 66–81 of subunit B to α 32–40 of subunit C
β 66–81 of subunit D to α 32–40 of subunit C
β 59–81 of subunit D to α 140–141 of subunit A
β 59–81 of subunit B to α 140–141 of subunit C

Summary and Conclusions

The original proposal outlined a program for identifying and characterizing protein complexes by combining crosslinking, mass spectrometry, and computation. Much of the originally proposed work was not feasible because of the reduced scope of the project necessitated by a budget substantially below that requested. Many of these originally proposed ideas would be of considerable benefit in developing an approach for characterizing protein complexes via an approach involving crosslinking and mass spectrometry. Synthesis of more suitable crosslinking reagents would avoid the problem of the lack of specificity of the photoreactive group in sulfo-SBED. Exploration and development of tandem mass spectrometric approaches, which could potentially give partial amino acid sequence information on each of the crosslinked peptides, would be of great benefit. Finally, more suitable computational methods for analyzing experimental data would be a necessity for carrying out crosslinking/mass spectrometry experiments with a reasonable throughput.

References

[1]G. B. Hurst, T. K. Lankford, and S. J. Kennel, "Mass Spectrometric Detection of Affinity Purified Crosslinked Peptides," *Journal of the American Society for Mass Spectrometry*, accepted for publication (2004).

[2]G. B. Hurst, M. Sega, J. L. Stephenson, Jr., P. K. Lankford, and J. A. Lewis, "Tandem mass spectrometry analysis of crosslinked proteins and protein complexes," Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando, FL, June 2–6, 2002.

[3]M. B. Strader, G. B. Hurst, T. K. Lankford, R. L. Hettich, and S. J. Kennel, "Affinity Purification of Crosslinked Peptides," Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal, Canada, June 9–12, 2003.

Elucidating the Functions of Genes and Pathways That Contribute to Genomic Instability, Cell Death, and Malignancies in Mouse Models with Telomere Dysfunction

Y. Liu
Life Sciences Division

The objective of this project was to build a laboratory capability and explore research tools at ORNL for dissecting complex biological systems, using telomere or telomerase as an initial model. We now have a fully operating lab and have obtained preliminary results. We have initiated genomic screening of chemical induced mutant mouse bank previously generated by the Life Sciences Division for mutations in genes involved in telomerase or telomere regulation. To understand the protein network associated with telomerase or telomere, we isolated putative proteins that may associate with these complexes. Mammalian expression vectors have been used to express tagged telomerase core protein or telomere binding protein in an immortalized cell line. Protein or protein complexes that are copurified with tagged proteins have been isolated and further identified by mass spectrometry. We have analyzed two mutant mouse models that are deficient in telomerase-associated proteins. Our framework allows us to understand the possible function of these proteins in telomere maintenance, which is critical in cell viability and genomic stability. Dysfunction of telomerase or telomere complexes is frequently observed in cancer and premature aging syndromes. Elucidating the molecular mechanism that regulates these complexes would benefit our understanding in these human disorders and their treatment.

Introduction

Telomere is a DNA-protein complex localized on the end of each chromosome, the function of which is to cap and protect chromosomes against degradation or fusion.[1] Telomere integrity in cells thus plays an essential role in the control of genomic stability. Loss of genetic material at chromosome ends ("telomere shortening") is frequently observed in elder populations and premature aging syndromes.[2] Furthermore, telomere dysfunction contributes to genomic instability that leads to cell death or cancer development.[3] Accumulating evidence suggests that telomere integrity depends on the ability to maintain telomere length and/or the ability to mask the telomeres from being recognized as damaged DNA.[4] However, the mechanisms by which short telomeres lose their capacity to suppress DNA damage responses are not well understood. In this proposal, we have planned to search for new mouse models, use the existing mouse models, and identify novel proteins to study the molecules that are important in telomere length maintenance and genomic stability. These molecules may serve as excellent therapeutic targets that may reduce cell viability in cancer cells.

Several key proteins or protein complexes are critical in telomere length maintenance. Telomerase is critical in telomere length maintenance by replenishing telomere loss due to "end replication problem" as the result of DNA replication in almost all eukaryotes. Loss or decrease in telomerase activity would lead to loss in telomeric DNA, which, in turn, triggers genomic instability or cell death.[1] However inadequate expression of this complex is also observed in 90% of human cancers and may be associated with malignant transformation or cancer progression in human.[2] These findings implicate telomerase and telomere homeostasis as key elements in the proliferation of normal and cancerous cells. Elucidating molecular machinery in regulating telomerase activity is critical in our understanding telomere length maintenance and malignant transformation.

Mammalian telomerase is a large ribonucleoprotein complex and contains two core components, telomerase reverse transcriptase and telomerase RNA, the latter serving as an integral template for de novo synthesis of telomeric DNA by telomerase. [4] These molecules are minimal components that are required in reconstituting telomerase activity in vitro.[5–7] Loss function in either component would result in loss of telomerase activity in vivo.[8–10] In addition to the core components, several mammalian telomerase-associated proteins have been identified . The function of these proteins in telomerase regulation or telomere length maintenance is poorly understood and need to be characterized. Elucidating these molecular events will certainly benefit our understanding in telomerase or telomere length regulation and thus cancer development and treatment.

Besides telomerase, telomere binding proteins and telomere-associated proteins are also key factors in

telomere protection, telomere length maintenance, and chromosome integrity.[11] Several factors are involved in the DNA damage response and also appear to be critical for telomere integrity. For example, cells expressing mutant telomere binding proteins or cells deficient in DNA repair proteins show dysregulated telomere length and chromosome instability in yeast or mammals.[11] Further studies are required to understand at the molecular level how these known proteins or yet unidentified telomere binding or associated-protein proteins protect telomere.

To elucidate telomerase-associated proteins or telomere binding/associated proteins in telomere length regulation, we have proposed to (1) search the established chemical-induced mutant mouse bank for new mouse models that carry point mutations at genes involved in telomerase and/or telomeres regulation, (2) search for novel telomerase or telomere-associated proteins, and (3) elucidate function of telomerase associate protein or DNA repair proteins in telomere length regulation in vivo.

Technical Approach

We have applied the new high-throughput mutation-scanning technique, temperature-gradient capillary electrophoresis (TGCE), to identify point mutations induced by N-ethyl-N-nitrosurea (ENU) in the mouse genome.[12] TGCE detects the presence of heteroduplex molecules formed between a wild-type gene segment and the corresponding homologous segment containing an induced mutation. Partially denatured heteroduplex molecules are resolved from homoduplexes by virtue of their differential mobilities during capillary electrophoresis conducted in a finely controlled temperature gradient. Heteroduplex analysis of three DNA fragments of telomere binding protein 1 and one DNA fragment of telomerase RNA ranging from 500–600 bp in size was applied to identify ENU-induced mutations of a bank of mutant mice.

To identify protein-protein interaction, we used mass spectrometry (MS) coupled with rapid methods for enrichment of protein complexes from total protein lysate. The approach involves expressing tagged proteins (as the bait) in an immortalized cell line, isolating tagged proteins and their binding proteins or protein complexes from cell lysate using anti-Flag immobilized beads, followed by stringent washes. The entire affinity-isolated complex is eluted off the beads through Flag peptide competition and is then digested by trypsin. Mass spectrometry can then separate the resulting array of peptides according to their mass. The combinations of mass and existing trysin cleavage sites provide identifications for the proteins in the complex through bioinformatics database search (Fig. 1).

To study the telomerase activity, we applied an in vitro– based PCR technique called telomerase amplification protocol. This protocol is a highly sensitive

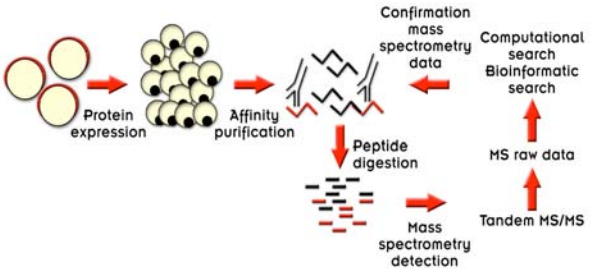


Fig. 1. Schematic presentation of protein tagging and mass spectrometry (MS) detection of protein complex.

assay for detecting telomerase activity in cells. The assay includes two steps. In the first step of the reaction, telomerase adds a number of telomeric repeats onto the oligonucleotide substrate. In the second step, the extended products are amplified by PCR which generates a ladder of products. Finally, the products are separated according to their size through electrophoresis in PAGE gel and the intensity of the products (reflecting the amount of telomerase activity in cells) are visualized through gel photography or imaging.[13]

To measure the average length of telomere repeats at chromosome ends in individual cells, we applied a flow cytometry method combined with fluorescence in situ hybridization (flow FISH) with labeled peptide nucleic acid (PNA) telomeric DNA probes.[14] The signal intensity of telomere fluorescence in each cell reflects the average length of telomeres in that cell. The average of telomere length in the total cell population was calculated from the mean fluorescence obtained from 10^4 cells/sample.

Results and Accomplishments

We initiated, in collaboration with Drs. Edward Michaud and Cymberline Culiat, the genomic screening of the ORNL Cryopreserved Mutant Mouse Bank (CMMB) for mutations in genes involved in regulating telomere, telomerase, or genomic stability. This bank contains DNA from 4000 C57BL/6JRn mice, each carrying a load of ENU-induced mutations inherited from their father. In this bank, every gene in the mouse genome is mutated multiple times, permitting the isolation of an allelic series of mutations for any gene of choice. We use TGCE for identifying single base-pair mutations and then verified new mutations by sequencing. We have intensively searched several databases (NCBI, Ensemble, and UCSC) for genomic sequences of four genes (two telomere-binding proteins and two telomerase core components). Such information is required for the genomic screening of ENU-induced mutations in CMMB. We initially screened for mutations at two genes (one telomerase core component and one telomere-binding protein). By the end of the project, we screened 1421 mutant mice, which total

over 2,131.5 kilobases mouse genomic sequences for mutations at these two genes. Thirty-six putative mutations were identified from the initial screen, which will be verified by sequencing analysis in the future.

To identify telomerase or telomere-associated proteins, we used an immortalized cell line (293T) that expresses a moderate level of Flag-tagged full-length human telomerase catalytic protein or 293T that transiently express a telomere binding protein. These tagged proteins served as the bait protein for isolating their interacting proteins in cells. The tagged proteins and their binding proteins or protein complexes were purified from cell lysate through anti-Flag immobilized beads (Fig. 2) and further identified by MS in collaboration with Dr. Greg Hurst of the Chemical Sciences Division, ORNL. These procedures allowed us to identify the tagged proteins and a number of potential binding proteins. Several previously identified telomerase-associated proteins, for example, Hsp70 and Hsp90, small nuclear ribonucleoprotein,[15, 16] were identified through this procedure, demonstrating that our approach was successful. In addition to these known proteins, we were able to identify 20 proteins that potentially associated with the bait proteins. Further studies are required to confirm the accuracy of MS data and the biological significance of these putative telomerase or telomere associated proteins in telomerase function or telomere length regulation.

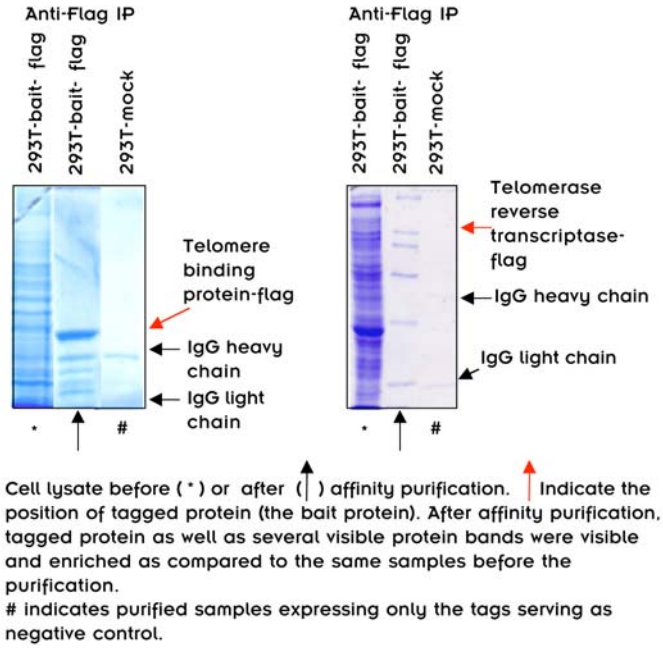


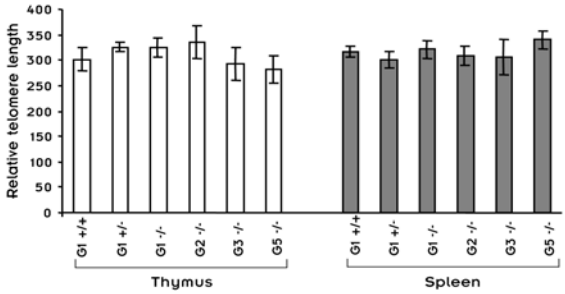
Fig. 2. Affinity purification of telomerase or telomere binding protein complexes using anti-Flag M2 beads. Quality of affinity purification was evident when comparing the same samples before or after affinity purification. After the purification, purified samples were digested by Trypsin and then injected into a mass spectrometer for mass identification.

Previously, Wen Zhou identified a protein component of large cytoplasmic structure, vault, also associated with telomerase activity in vivo (Zhou et al., unpublished data), suggesting that it is a telomerase-associated protein. To elucidate the function of this protein, namely, *V PARP*, in telomerase or telomere length regulation in vivo, we studied the telomerase activity and telomere length in mouse deficient for *V PARP*. This work was carried out in collaboration with Dr. Lea Harrington at the Ontario Cancer Institute, Canada. The knockout mice were viable and fertile for up to several generations. These mice resulted in no apparent change in telomerase activity (data not shown) or telomere length [Fig. 3(a)]. One possibility could be other telomerase associated proteins, such as telomerase-associated protein 1 (*TEP 1*) might share a redundant role with *V PARP* in vivo, as both proteins are commonly shared by two ribonucleoprotein complexes, telomerase and vault. To test this hypothesis, we further generated double knockout mice that are deficient for both *V parp* and *Tep 1*. These mice are also viable and fertile and did not show any detectable change in telomere length at the first generation [Fig. 3(b)]. These data provide direct genetic evidence that neither murine *Tep 1* nor *V parp* is essential for normal mouse development, telomerase catalysis, and/or telomere length maintenance in vivo. A manuscript based on above data is in preparation.

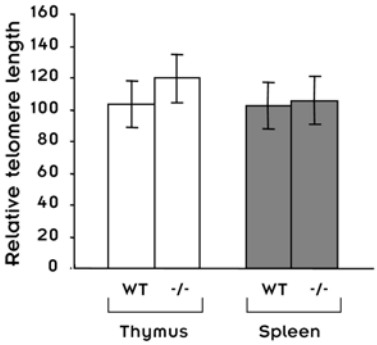
Summary and Conclusions

Our high-throughput screening of CMMB for mutations at the genes regulating telomere length enabled us to identify 36 potential mutations in this mutant mouse bank. Using protein tagging combined with protein affinity purification and MS, we were able to identify several known telomerase-associated proteins and 20 putative telomerase or telomere binding proteins. Using a genetic approach, we demonstrated the possible function of two telomerase-associated proteins in telomerase function or telomere length regulation in vivo. Further studies including verifying mutations at CMMB and characterize putative telomerase or telomere-associated proteins in telomere length regulation would enable us to establish new research tools and understand the mechanism in regulating telomerase or telomere length at the molecular level. A manuscript based on some of our finding is in preparation.

Our projects exploited a leading resource in mouse mutagenesis for generating animal models and in protein



Telomere length was measured by an fluorescent based in situ hybridization method using fluorescent labeled telomere DNA as probe. The intensity of the telomere probe was detected by flow cytometer, which indicate the length of telomere DNA in cells. The average telomere fluorescence in thymocytes and splenocytes derived from wild type (+/+), heterozygous vparp (+/-), or different generations (G_1 and G_5) of V parp deficient mice. In each set, data were pooled from at least five individual mice (error bars represent the standard deviation).



Telomere length was measured by an fluorescent based in situ hybridization method using fluorescent labeled telomere DNA as probe. The intensity of the telomere probe was detected by flow cytometer, which indicate the length of telomere DNA in cells. The average telomere fluorescence in thymocytes and splenocytes derived from wild type (+/+), or knockout mice deficient for both V parp and Tep 1. In each set, data were pooled from at least five individual mice (error bars represent the standard deviation).

Fig. 3. No detectable change in telomere length in mV parp *null or* mV parp /mTep 1 *double null mouse. (a) Telomere length analysis in mouse thymocytes and splenocytes from different generations of* mVparp-*deficient mice by Flow-FISH. The average telomere fluorescence in thymocytes and splenocytes derived from different generations (G_1 and G_5) of* mV parp *knockout mice. (b) Telomere length analysis in thymocytes and splenocytes derived from wild-type and* mV parp / mTep 1 *double-deficient mice by Flow-FISH. In each set, data were pooled from at least five individual mice (error bars represent the standard deviation).*

tagging technique for discovering the nature of the protein-complex "machines" important for telomere maintenance and genomic stability. Such results will form the basis for research directions compatible with DOE's Genomes to Life and Health Effects/Genome programs, particularly in the realms of analysis of protein machines (in this case, telomerase and telomere complexes). Our effort has been expanded to several relevant interesting projects through the internal and international collaboration and formed the basis for establishing new tagging vector system for the GTL program. This project has applications in several programs at Department of Defense and the National Institutes of Health including the National Institute on Aging.

References

[1]C. W. Greider and E. H. Blackburn, "Telomeres, telomerase and cancer," *Sci. Am*. **274**, 92 (1996).

[2]M. W. Djojosubroto et al., "Telomeres and telomerase in aging, regeneration and cancer," *Mol Cells* 15(2), 164–75 (2003).

[3]R. Oulton and L. Harrington, "Telomeres, telomerase, and cancer: life on the edge of genomic stability," *Curr Opin Oncol* 12(1), 74–81 (2000).

[4]E. H. Blackburn, "Switching and signaling at the telomere," *Cell* **106**(6), 661–73 (2001).

[5]T. L. Beattie et al., "Reconstitution of Human Telomerase Activity In Vitro," *Curr Biol* **8**(3), 177–80 (1998).

[6]T. L. Beattie et al., "Functional Multimerization of the Human Telomerase Reverse Transcriptase," *Mol Cell Biol* **21**(18), 6151–60 (2001).

[7]S. L. Weinrich et al., "Reconstitution of Human Telomerase with the Template RNA Component hTR and the Catalytic Protein Subunit hTRT," *Nat. Genet.* **17**(4), 498–502 (1997).

[8]M. A. Blasco et al., "Telomere Shortening and Tumor Formation by Mouse Cells Lacking Telomerase RNA," *Cell* **91**(1), 25–34 (1997).

[9]Y. Liu et al., "The Telomerase Reverse Transcriptase is Limiting and Necessary for Telomerase Function In Vivo," *Curr Biol*. **10**(22), 1459–62 (2000).

[10]X. Yuan et al., "Presence of Telomeric G-Strand Tails in the Telomerase Catalytic Subunit TERT Knockout Mice," *Genes Cells* **4**(10), 563–72 (1999).

[11]T. de Lange, "Protection of Mammalian Telomeres*," Oncogene* **21**(4), 532–40 (2002).

[12]Q. Li et al., "Integrated Platform for Detection of DNA Sequence Variants Using Capillary Array Electrophoresis," *Electrophoresis* **10**, 1499–1511 (2002).

[13]N. W. Kim and F. Wu, "Advances in Quantification and Characterization of Telomerase Activity by the Telomeric Repeat Amplification Protocol (TRAP)*," Nucleic Acids Res*. **25**(13), 2595–7 (1997).

[14]N. Rufer et al., "Telomere Length Dynamics in Human Lymphocyte Subpopulations Measured by Flow Cytometry," *Nat. Biotechnol*. **16**(8), 743–7 (1998).

[15]S. E. Holt et al., "Functional Requirement of p23 and Hsp90 in Telomerase Complexes," *Genes Dev* **13**(7), 817–26 (1999).

[16]A. G. Seto et al., "Saccharomyces Cerevisiae Telomerase is an Sm Small Nuclear Ribonucleoprotein Particle," *Nature* **401**(6749), 177–80 (1999).

Identification and Characterization of Genes and Protein Components in Cell-Cycle Control and Cancer Development

Y. Wang
Life Sciences Division

A "substrate-trapping" strategy to isolate target proteins of cell-division-cycle 14 phosphatase (cdc 14) has been developed, and the substrate-trapping mutants of cdc 14 have been produced. Different epitope-tags of cdc 14 proteins have been engineered, and purified proteins have been made in *E. coli* and mammalian cells. The combination of immunoprecipitation and mass spectrometric analyses has allowed us to identify several potential cdc 14 substrates or interacting proteins. The biological significance of some interacting proteins has been characterized and others are currently under investigation. In collaboration with Dr. Mike Tyers in Samuel Lunenfeld Research Institute, Canada, a systematic functional annotation of a mammalian gene B23 in *S. cerevisiae* gene-deletion set has been performed. We also initiated a collaborative project with Dr. Klas Wiman, Karolinska Institute, Sweden, to study the gene expression profile regulated by the p53 tumor suppressor gene using a high-throughput microarray assay.

Introduction

Tumor cells often acquire damage to genes that directly regulate their cell cycle. A great majority of DNA damage in cancer patients is the result of environmental insults, such as exposure to irradiation and chemical mutagens. Understanding the molecular mechanisms that regulate these events will help us to assess and prevent the risks of cancer associated with the exposure of environmental hazardous materials, which is consistent with DOE's long-term mission. The fundamental task of the cell cycle is to ensure that DNA is faithfully replicated once during the synthesis (S) phase and that identical chromosomal copies are distributed equally to two daughter cells during mitosis. Disruption of this fundamental process could lead to disasters to our life, such as the development of cancer, developmental failure or degenerative diseases.[1]

Numerous genetic interactions and biochemical studies positioned cell division cycle 14 (cdc 14) gene at the bottom of the mitotic exit cascade in budding yeast, where it triggers mitotic exit by inactivating the active cyclin-dependent kinase. Activation of the Cdc 14 is mediated by the release of cdc 14 from nucleolus and tightly controlled through FEAR (Cdc 14 early anaphase release) and MEN (mitotic exit network) pathways. Cdc 14 is a dual-specificity phosphatase conserved from yeast to *C. elegans* to mammals. In mammals, it has been found to play an important role in centrosome splitting at the onset of mitosis, and mutations of the gene have been detected in human breast carcinoma cell lines. The molecular mechanisms underlying the role of cdc 14 in cell cycle regulation and tumor development is largely unknown.[2] Our goal is to isolate and characterize the Cdc 14 phosphatase substrates and interacting proteins in collaboration with the Mass Spectrometry Group at ORNL. This study will enable us to dissect the biological significance of the Cdc 14 protein networks in cell cycle control and cancer development. Our study will not only help us to understand the mechanism of cancer development but also to provide possible protein targets for anti-cancer drug development.

Nucleophosmin/B23 mapped to human chromosome 5q35 is a nucleolar phosphoprotein that is more abundant in tumor cells than in normal resting cells. It has been identified as a substrate of cyclin-dependent kinase 2 (CDK2)/cyclin E in centrosome duplication. B23 is associated with unduplicated centrosomes, and dissociated from centrosomes by CDK2/cyclin E-mediated phosphorylation. B23 has also been found to be involved in at least three distinct forms of hematologic malignancy. The N-terminal region of B23, which contains oligomerization domain, is fused to anaplastic lymphoma kinase in anaplastic lymphoma with t(2;5), retinoic acid receptor, in acute promyelocytic leukemia with t(5;7), and myeloid leukemia factor 1 in myelodyplastic syndrome with t(3;5) translocations.[3] An in-depth annotating the molecular mechanisms of B23 will be of benefit not only to our understanding of the cell cycle regulation but also to the development of cancer.

The p53 tumor suppressor protein is expressed at low levels in normal cells but accumulates in response to DNA damage, leading to cell-cycle arrest and/or apoptosis. The wild-type p53 protein has sequence-specific DNA binding, transcriptional transactivation, and repression activity. Point mutations of p53 found in human tumors abolish its sequence-specific DNA binding activity, as a rule. Recent work demonstrated p53 exerts its effect through a diverse network of transcriptional changes, which makes exploration of the molecular mechanisms that mediate p53-triggered cell-cycle arrest and apoptosis more challenging and difficult.[4] As the first step to solve these puzzles, we have proposed to isolate novel genes that are directly regulated by p53 through microarray approach.

Technical Approach

Transient protein-protein interaction and low-affinity "bait" proteins are the two major blockades in isolation of substoichiometric protein complexes by an immunoprecipitation/mass spectrometric approach. If a target protein is a substrate of an enzyme, it is, in general, extremely difficult to "fish out" the substrates using the enzymes as bait, because of the fugitive enzyme-substrate interactions. Based on the crystal structures of protein tyrosine phosphatase 1B (PTP1B), we have identified a conserved aspartate residue in human Cdc 14 (cell division cycle 14) proteins. Site-directed mutation of this conserved aspartate residue to alanine increases binding affinity to its substrates several-fold while it dramatically decreases their catalytic capacity. These mutants have been used for substrate-trapping followed by mass spectrometric identification. Proteins that are substantially interesting are further evaluated by their subcellular localization through fluorescent imaging, cell cycle distribution, and genetic complementation analyses.

Seventy percent of yeast genes are functionally classified and more than 80% of the genetic pathways in yeast can be found in human. Each of the 5100 nonessential genes in the 6200-gene *S. cerevisiae* genome has been systematically deleted and bar-coded, with the composite deleted genomes assembled into an ordered array of live yeast strains.[5] Overexpression of human genes in wild-type yeast cells either can be lethal or can have no effect. Subsequent crosses with deletion strains, in synthetic complementation or dosage-lethality experiments, respectively, can identify the analogous genetic pathway to which the human gene belongs, thus providing a gene-based method to filter the function of mammalian genes. To systematically annotate the function of the B23 gene in yeast gene-deletion arrays, the B23 gene was first cloned into a yeast expression vector and transformed into wild-type yeast cells. Haploid MATα yeast cells carrying the human B23 were mated with each of MATα deletion mutant by an automated robot. The presence or absence of yeast colonies was chosen as a standard cutoff to map the genetic interacting pathways using an automatic colony analyzing software.

The advanced development of cDNA microarray technology enabled us to study thousands of gene expression profiles instead of one or two genes at a time. To study the expression profiles of p53-regulated genes, mRNAs were isolated and reverse-transcribed into cDNA from experimental and control groups. The cDNAs of the two groups were differentially labeled and then hybridized onto the mouse 22k cDNA chips from Vanderbilt University. Differentially expressed genes were analysed through an array reader using Gene-Pix software, and the presence of consensus p53 binding sequences (two consensus decamers RRRCWWGYYY separated by a gap of 0–13 nucleotides) in the genes of our interest was searched through the NCBI public database.

Results and Accomplishments

We have created point mutations converting the wild-type Cdc 14 into substrate-trapping mutants. The mutants were first tested in complementation assay in yeast and shown to be incapable of rescuing their yeast orthologue defects (Fig. 1), suggesting they could be used as substrate-trapping mutants. These mutants were in-frame cloned into a myc-epitope tagged expression vector and introduced into mammalian cells by stable transfection. Immunoprecipitaion of Cdc $14A^{D/A}$ mutant protein with anti-myc antibody (9E10) from Cdc 14 $A^{D/A}$ expressing cells allowed us to pull down several potential Cdc 14A

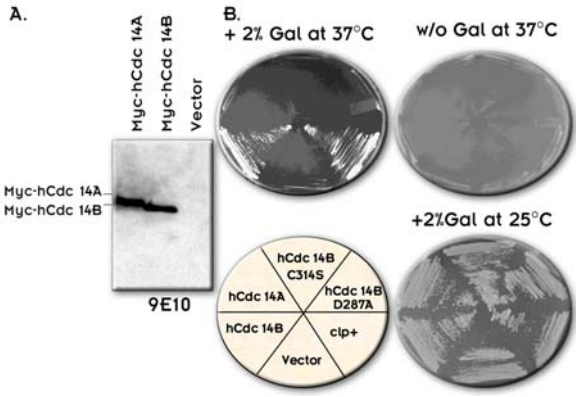


Fig. 1(a) Western blot analysis of Myc-tagged human Cdc 14A and Cdc 14B expression in cdc 14-3 *cells transformed with pRSHisGal-hCdc 14A, pRSHIsGal-hCdc 14B and pRSHisGal vector alone respectively. (b) Expression of human Cdc 4B but not Cdc 14A rescues the lethality of* cdc 14-3 *cells at non-permissive temperature. The temperature-sensitive mutant* cdc 14-3 *cells containing different plasmids as indicated were tested for colony formation on XY plates with and without galactose at 37°C and 25°C. Human Cdc* $14B^{C314S}$ *and Cdc* $14B^{D287A}$ *are catalytic dead mutants.* Clp^{+} *is the* S. pombe *ortholog of* S. cerevisiae *Cdc 14p.*

substrates or interacting proteins. Following SDS-PAGE separation, in-gel peptide digestion with trypsin and mass spectrometric analysis, we were able to narrow down to several potential Cdc 14 substrates (Fig. 2). As a proof of principle, we focused on B23 because of recent findings on its involvement in the control of centrosomal duplication. We pulled out B23 full-length cDNA from a cDNA library and cloned it into a Flag-epitope vector. Co-transfection of both B23-Flag and Myc-Cdc 14A showed that B23 and Cdc 14A could indeed interact with each other in vivo by reciprocal immunoprecipitation (Fig. 3). Other groups have shown unphosphorylated threonine 199 of B23 binds to unduplicated centrosome and inhibits centrosomal duplication. We believe our Cdc14 A is the genuine phosphatase that may dephosphorylate Thr199 of B23, which in turn controls centrosomal duplication in S phase and prevents centrosomal reduplication during mitosis. Our preliminary results indicate they might co-localize on centrosomes in G1 and anaphase. We also constructed GFP-B23 and Red-Cdc 14 vectors in order to study the dynamic or temporal interaction of the two proteins in a living cell. Due to the lack of a live cell imaging system at ORNL, this approach will not be possible for the time being. As an expansion of the project, we also purified centrosomal fractions from Jurkat cells (a T cell line) by sucrose gradient centrifugation. Centrosomal fractions were verified by the expression of γ-tubulin on Western blot. We showed that B23 and Cdc 14 co-migrate in the same centrosomal fractions (Fig. 4). These centrosomal fractions can be used for the isolation of the mysterious unknown centrosomal components using $B23^{T199A}$ mutant, that may be directly regulated by the unphosphorylated B23. A manuscript based on the above studies is in preparation.

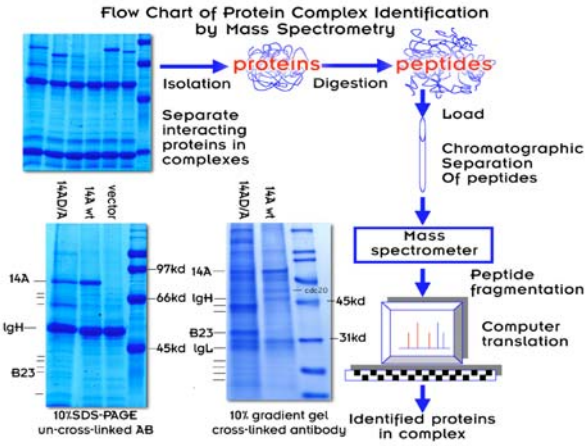


Fig. 2. A flow-chart of immunoprecipitation/mass spectrometric assay to isolate and identify potential associated protein complexes. Substrate-trapping mutant proteins were expressed in cells. After affinity purification, associated protein complexes were separated on SDS-PAGE gel. Individual associated proteins were isolated and digested to generate individual peptides. Peptides were loaded to mass spectrometer for peptide identification. Upon data bank search, individual peptides were assigned to individual proteins. Potential substrates of Cdc 14s were indicated.

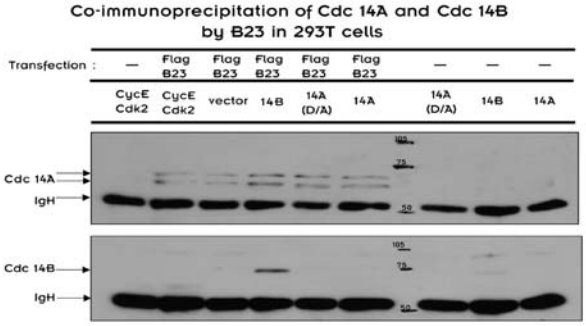


Fig. 3. In vivo interaction of Cdc 14 with B23. Flag-tagged B23 protein complexes were affinity precipitated with anti-Flag antibody. Cdc 14 proteins were detected by anti Cdc 14 antibodies, confirming the mass spectrometric finding.

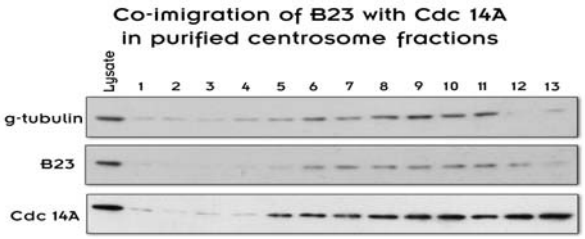


Fig. 4. Co-immigration of B23 with Cdc 14A in purified centrosomal fractions. Centrosomal fractions were purified from T-cells through ultracentrifugation in sucrose cushion. γ-tubulin was used as the centrosome marker. Centrosome fractions that containing B23 also expressed Cdc 14A.

In order to annotate the functions of human B23 gene in yeast, we have performed, in collaboration with Dr. Mike Tyers, Toronto, Canada, two rounds of synthetic dosage lethality screening using both wild-type B23 and $B23^{T119A}$ mutant (a critical phosphorylation site mutant) in ~5100 viable yeast gene-deletion mutant arrays. When a stringent bioinformatic filter was applied, we found that a total of 66 (33 for wild-type B23, 5 for $B23^{T199A}$ mutant and 28 for both) out of 5100 genes were mapped into B23 genetic pathways (Fig. 5). In addition to the verification of known functions of B23 involving ribosome biosynthesis and spindle pole body (centrosome) splitting, we identified several novel roles of B23 such as histone acetylation and actin organization. We are currently investigating the potential relationships between B23 and the mammalian orthologues of the newly discovered B23-related yeast genes in mammalian systems. We will use the data obtained from this study as a proof of concept to conduct a high- throughput annotation of genes on human chromosomes 5, 16 and 19, an ultimate goal of functional genomic program sponsored by DOE.

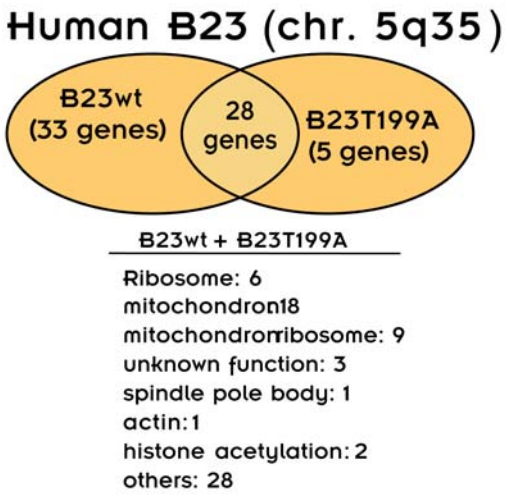


Fig. 5. Summary of B23 genetic pathways based on the yeast gene-deletion array screening. A total of 66 genes were identified. Twenty-eight genes were involved in both wild-type B23 and mutant B23^{T199A}, 33 exclusively interacted with wild-type B23 and 5 only with mutant B23. Actin network assembly and histone acetylation are the new genes that have never been found to associate with B23 before.

Using a temperature-sensitive mouse mutant p53 construct, we have previously shown that wild-type p53 can trigger both G1 arrest and apoptosis in a v-myc retrovirus-induced p53 negative mouse T lymphoma line J3D. We have also found that bcl-2 can prevent p53-induce apoptosis but not G1 arrest, whereas HPV16E7 blocked p53-induced G1 arrest but not apoptosis.[6] These findings let us believe simultaneous expression of both bcl2 and HPV16 E7 would prevent both apoptosis and G1 arrest even in the presence of activated wild-type tsp53. Moreover, this triple-gene expressing cell line could be used to isolate genes that are potentially directly regulated by p53 through cDNA microarray approach, because the downstream signals (the cellular effects of p53) are supposed to be blocked. The experiments were performed on mouse 22k cDNA chips from Vanderbilt University. As a proof of concept, we made a stable line expressing tsp53, bcl2, and HPV16 E7. Western blotting and RT-PCR showed that these cells express all three exogenous proteins. By comparison of the gene expression profiles using cluster analysis at different time points at non-permissive temperature, we have identified several novel genes directly regulated by p53. As a control for the temperature shift, we have also generated stable cell lines expressing E7 and bcl2 from the p53 negative parental J3D cells. Candidate genes that are differentially expressed are currently being confirmed by Northern blot and quantitative Real-PCR analyses. Bioinformatic analysis has enabled us to confirm the presence of p53 consensus sequences (two consensus decamers RRRCWWGYYY separated by a gap of 0–13 nucleotides) in the promoter regions of our newly identified genes. We are currently investigating the ability of those genes in inhibiting or promoting p53-induced G1 arrest and apoptosis in p53$^{-/-}$ and p53$^{+/+}$ cells. These studies will help us to understand how the p53 tumor suppressor protein modulates the cellular response to various genomic insults and to discover novel pathways that mediate the p53-triggered cell cycle arrest and apoptosis.

Summary and Conclusions

The initial LDRD funding has helped me to set up a new molecular biochemical laboratory and obtain preliminary as well publishable data. My lab is now fully functioning and able to perform most of the molecular and biochemical experiments. Using a protein-tagging strategy in combination with mass spectrometric peptide identification, we have identified several Cdc 14 substrates/interacting proteins and characterized their potential involvement in centrosome duplication and cytokinesis. We have identified at least two novel functions of the mammalian B23 gene based on yeast deletion set screening. Using the high-throughput microarray approach, we identified several novel genes that are potentially regulated by p53 tumor suppressor and function as checkpoint controls in response to DNA damage in cells.

This study of the protein complexes involved in the cell cycle control and cancer development is in line with DOE's Genome to Life protein complex efforts and long-standing interest in issues related to human susceptibility to the products of energy production. In fact, as a member directly working on Genome to Life project, I have experienced tremendous mutual benefits between our mammalian protein complex project and the Genome to Life project, though the latter is mainly focused on microbial biology. I believe our interaction and collaboration will continue in the very near future. Since the Cdc 14 regulatory networks, B23, and p53 tumor suppressor pathways are all instrumental in helping us understand cellular response to the exposure of environmental hazardous materials and cancer development, we believe that data generated in these studies will qualify us to apply for follow-up funding from the DOE Genomes to Life and Low Dosage Radiation programs, NASA Low Dosage Radiation program, and the NCI Cancer Genetics and Molecular Biology programs.

References

[1]J. Rouse and S. P. Jackson, "Interfaces between the detection, signaling, and repair of DNA damage," *Science* **297**(5581), 547–51 (2002).

[2]S.Trautmann and D. McCollum, "Cell Cycle: New Functions for Cdc14 Family Phosphatases," *Curr. Biol*. **12**(21), R733–5 (2002).

[3]M. Okuda, "The Role of Nucleophosmin in Centrosome Duplication," *Oncogene* **21**(40), 6170–4 (2002).

[4]A. Gudkov, "Microarray Analysis of p53-Mediated Transcription: Multi-Thousand Piece Puzzle or Invitation to Collective Thinking," *Cancer Biol. Ther.* **2**(4), 444–5 (2003).

[5]A. H. Tong et al., "Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants," *Science* **294**(5550), 2364–8 (2001).

[6]Y. Wang et al., "Abrogation of p53-Induced G1 Arrest by the HPV 16 E7 Protein Does Not Inhibit p53-Induced Apoptosis," *Oncogene* **12**(12), 2731–5 (1996).

Comprehensive Molecular Probing of Live Biological Cells

T. G. Thundat,[1] R. J. Warmack,[2] T. L. Ferrell,[1] M. J. Doktycz,[1] M. D. Barnes,[3] and T. E. McKnight[2]
[1]*Life Sciences Division*
[2]*Engineering Science and Technology Division*
[3]*Chemical Sciences Division*

Imaging, characterizing, and identifying intra- and extra-cellular localization of proteins and protein complexes at nanometer resolution are important goals in nanobiology. The objective of this proposal is to design and develop nanoscale probes and a photonic force microscope (PFM) to image intra- and extra-cellular protein complexes. Such probes reside at the confluence of the smallest of man-made materials and those of natural systems, making them ideal probes for biological cells. Nanoscale probes offer a superlative approach to characterizing the molecular interactions taking place within living systems. The PFM using laser tweezers and nanoprobe optical monitoring will allow probing and imaging outside and inside the cell. This unique microscope will provide information about biomolecular complexes and cellular nanomechanics to complement conventional scanning probe and optical techniques, providing an unprecedented perspective of cellular processes.

Nanoscale imaging of molecules and cells is currently achieved using atomic force microscopy (AFM). In an AFM, a cantilever with a very small spring constant is scanned over the samples to create a force image of the object. Since cantilevers can detect extremely small forces such as van der Waals' forces between the cantilever tip and the sample, a resolution image can be obtained. However, the cantilevers have limited freedom as they are attached to a massive chip and cannot be used for imaging sharp features, for example, the underside of a cell. The PFM is an innovation that will allow extra-cellular probing and, for the first time, nanoscale intracellular probing of cells. The PFM retains the sensitivity of the AFM without the limitations of the cantilever spring and its inherent restriction to surface scanning (two-dimensional). The ability of the PFM to tether a dielectric bead with an optically adjustable force provides full three-dimensional imaging of structures, revealing the nature and magnitude of molecular forces. The extremely small force constant of the optical trap allows detection of very small forces. The ability to detect extremely small forces translates into very high resolution in imaging. Advancing these probes beyond surface analyses will enable extensive characterization of cellular processes in real time. This basic work is expected to promote fundamental discoveries in biomedical imaging and bioengineering programs hoped to impact medical applications.

Progress in the first year included the design, acquisition, assembly, and operation of the PFM, providing state-of-the-art optical imaging of cells plus guided probe analysis of internal and external nano-mechanics. Theoretical studies of light scattering and fluorescence were performed to substantiate work to be performed in FY 2004. We were successful in trapping a bead that executes Brownian motion in the optical trap of the photonic force microscope. The Brownian motion amplitude decreases significantly once the bead is trapped in the optical trap. We were also successful in navigating the trapped bead in a fluid medium by translating the optical trap. Efforts are under way to image a living cell using optically trapped bead. Our next studies will concentrate upon demonstrating the efficacy of these unique nanoprobe methods in conjunction with other advanced imaging studies that will solve issues of interest to DOE and NIH.

The role of high-resolution imaging has applications directly relevant to the Department of Energy's (DOE) Genomes to Life program. It will provide a more complete imaging capability for the exterior and interior of a living cell, bound proteins, receptor sites, etc. Such information will greatly enhance our understanding of living cells and the way cells work. The tools afforded by this work are also quite relevant to Nanoscale Science, Engineering, and Technology Research.

This project also has applications in several NIH programs, especially the National Institute for Biomedical Imaging and Bioengineering. The ability to real-time image cells to external stimuli has significant applications in clinical applications, biomedical engineering, and drug discovery. The combination of unique nanoprobe techniques with other imaging and probing instrumentation at ORNL opens the possibility of becoming an NIH center that would attract and foster very fruitful collaborations with visiting cellular biologists.