Laboratory Directed Research and Development Program
Berkeley Lab FY 2000 Coversheet

Project Title: Structural/Functional Genomics and Pathways: *D. radiodurans* and *B. subtilis*

Prop No.

Investigator(s): A. P. Arkin, T. Head-Gordon, S. Holbrook, D. Rokhsar

Budget No.

Division: Physical Biosciences

PAO OFFICE USE ONLY

Funds Requested (FY 2000): $295K

Proposed Project Duration: 3

Out Year Funds Requested: (for multiyear projects only)

New Proposal [x] **Continuation** [ ]

Long-Term Funding (amount, source, likelihood):
DOE (SSI and OBER); NIH

Collaborating Divisions or Institutions:

Summary

Purpose /Goals:
The sequencing of the complete genomes of a variety of microbes, the metazoan *C. elegans*, and the soon to be completed *Drosophila* and human genome projects, are a driving force for understanding biological systems at a new level of complexity. The goal of the computational structural and functional genomics initiative of the future is to link these sequencing efforts to a high-throughput program of annotation and modeling of both molecular structures and functional networks. We propose to build a coherent computational biology program at LBNL by linking research in DNA modeling; protein fold recognition, comparative modeling, and *ab initio* prediction of individual gene products; molecular recognition of protein-protein and protein-nucleic acid complexes; and modeling biochemical and regulatory pathways, using *Deinococcus radiodurans* and *Bacillus subtilis* as our test beds.

Approach/Methods:
To explore the basis for survival of *D. radiodurans* under extreme conditions of DNA damage, and the sporulation/competence switch in *B. subtilis* under nutrient stress, we will construct molecular models of the key components of the DNA repair system including damaged DNA, multigenomic repair intermediates such as Holliday junctions, proteins known or yet to be discovered that are involved in DNA repair, as well as higher order protein-protein or protein-nucleic acid complexes implicated in repair, replication, and transport. A model for sporulation of *B. subtilis* will be developed using Bio/Spice and other modeling tools to identify the control mechanism of the sporulation/competence switches in cell nutrition. An essential element of the proposed program is to develop connections between the analyses of single molecules, complexes, and networks, to establish a coherent program for integrating structural and functional annotation into systems level understanding for any organism.

Relationship to other Berkeley Lab projects sponsored by DOE or other agencies:

Is there human subject data, cells, or tissues and/or animal use on this project? If yes, fill in attached form.
(See instructions)

Yes [ ] **No** [x]

On an attachment (3 pages, maximum), please provide a brief description of the project:
Purpose / Goals; Approach / Methods; potential results or significance *and, if multi-investigator or multi-divisional, proposed organization.*

Introduction and Connections to other projects at LBNL:

All organisms respond to external stresses with defined programs of physiological and morphological change; the description, modeling, and ultimately manipulation of these responses is an central goal of modern biology. In principle, the advent of high throughput methods for genome sequencing, proteomics, and gene-expression assays provides the information needed to understand the relevant molecular and cellular processes. But disentangling the function of these systems from sequence and other data requires an array of computational and theoretical approaches that remain to be developed. Here we propose a concerted effort in the modeling of two response systems -- DNA repair in *D. radiodurans*, and sporulation in *B. subtilis* -- as a testing ground for computational and theoretical approaches from the molecular to the systems levels. Our long-term goal is to establish a robust framework for reliably predicting the three dimensional architecture of all proteins of any genome in order to gain insight into their function, and to integrate this structural and functional annotation into systems level understanding for any genome.

There are multiple experimental "structural and functional genomics" efforts underway at the ALS at Berkeley Lab, including the *Methanococcus jannaschii* project of S. H. Kim, the *pyrobaculum* genome led by T. Alber, and the study of disease of the premature aging syndrome under Mian, Campisi and McDermott, which are core research projects at the Macromolecular Crystallography Facility. We would like to emphasize that the computational approaches outlined here, while focused on the understanding of the DNA repair system of *D. radiodurans* and the sporulation system of *B. subtilis*, are broadly applicable to all structural/functional genomics efforts. By establishing a coherent working group in computational and theoretical biology at LBNL we position the Lab to take advantage of upcoming large-scale DOE and NIH initiatives in this area, including those pertaining to the Strategic Simulation Initiative. To develop connections between analyses at different levels of description (single molecules, complexes, networks) we propose to hire four postdoctoral researchers who would be jointly supervised by the four PI s and whose projects would straddle these three levels. As the project develops, we expect to involve other computational and experimental efforts at the Lab.

Background:

Deinococcus Radiodurans (DR: strange berry that withstands radiation) is a bacteria that was isolated in 1956 from tins of spoiled meat that was thought to be given sterilizing doses of γ radiation. It is a gram positive bacterium and has on the order of $3x10^6$ base pairs, or ~3000 protein products, and is fully sequenced by TIGR under DOE/OBER sponsorship (http://www.ncgr.org/microbe/deinococcustxt.html). DR exhibits unusual tolerance to high radiation doses, surviving with no loss in population to 5,000Gy of γ radiation, and D_{37} doses (37% population survival) up to 15,000Gy, as well as high tolerance to UV and cross-linking agents. By comparison, *E. Coli* shows an exponential decline in population with exposure to γ radiation of 10-100Gy. This resistance to radiation is thought to be a by-product of the adaptative pressure to evolve well-honed DNA repair strategies under dessication conditions, since under dehydration conditions accumulation of DNA damage is known to occur. There seems to be at least three components to DR's successful DNA repair strategy: specifics of the DNA repair mechanism, the fact that it is multi-genomic, and the coupling between repair, replication, and the export of damaged DNA from the intracellular medium. The ability of DR to survive high levels of DNA damaging agents such as: ionizing radiation, ultraviolet light and cross-linking chemicals may be due to the levels and activities of its DNA repair enzymes, protective enzymes such as superoxide dismutase and catalase, the multigenomic state of its DNA, specific repetitive sequences in its DNA, or other factors of which we are currently unaware.

Bacillus subtilis is the best-characterized member of the Gram-positive bacteria and is part of a close clad of bacteria that include *B. cereus*, a food poisoning agent that is also used in plant biocontrol, and *B. anthracis*, a deadly pathogen. Its fully sequenced genome (http://pbil.univ-lyon1.fr/nrsub/nrsub.html) is 4.2 Mb in length and comprises 4,100 protein-coding genes. As a response to nutrient exhaustion much (but not all) of a population of *B. subtilis* enters a program of irreversible differentiation that results in the formation of a dormant life form termed the endospore or spore. This complex process lasts some 8 hours and uses approximately 125 developmental genes. The robust genetics available for *B. subtilis* makes sporulation (and other responses to nutrient stress, including competence) an excellent model system for the study of the basic mechanisms of regulating developmental gene expression. The initiation of a proteome project (http://microbio2.biologie.uni-greifswald.de:8880/) and a wide array of genetic and biochemical studies has yielded an extraordinary amount of direct and suggestive information of how these chemical processes are accomplished in the cell.

DNA Modeling of Damage and Multi-genomic issues:

The fact that DR is multi-genomic may or may not be a component of its DNA repair mechanism. There are examples in which DNA repair efficiency is greater in organisms with higher chromosome multiplicity, although there are examples of radiation-sensitive organisms that are multigenomic. Since DNA double strand breaks from radiation occur randomly, it is thought that each chromosome would show a different distribution of double strand breaks (dsb), and would therefore allow the piecing together of a complete genome from its copies. However, at such high radiation doses, the number of dsb is on the order of 100 (typically 10 can be lethal in eukaryotes), so there is a complementary strand search problem that is more severe for multi-genomic organisms. Experimental attempts to correlate DNA repair efficiency with copy number of DR were inconclusive: attempts to reduce DR to below tetraploid were unsuccessful, and when higher copy numbers were introduced (4-10), DNA repair efficiency showed no obvious correlation with greater multiplicity. One strategy is to avoid the search problem by aligning chromosome copies in such a way that complementary strands are easily compared for damage, and pieced together. Alignment possibilities include Holliday junctions or presence of DNA binding proteins (none have been found yet), but is speculative.

To understand DNA damage and multigenomic issues, our goal is to use these molecular models to further our understanding of DNA repair in DR and to suggest experimental structural studies in the future. The structures of many types of damaged DNA, such as thymine dimers have been modeled previously. We will model other types of DNA damage and sequences found in DR, such as several types of damaged DNA containing base modifications (i.e. alkylated , apurinic, oxidized, i.e. 8-oxoguanine) as well as DNA with strand breaks, and their complexes to the DR protein models described below. Also molecular models will be constructed of Holliday junctions

that may be involved in intergenomic exchange with special consideration of repeating sequences or protein binding sites which may be used in alignment. In this regard, the crystal structure of the RuvA protein from *E. coli* complexed to a Holliday junction DNA is known. Although the RuvA protein has not yet been identified in DR, the RuvB homologue has been found. We will also model complexes of proteins and damaged DNA using DR molecular models. Software tools for detailed molecular modeling of nucleic acids have not been as extensively developed as the analogous protein tools. Recently several powerful software packages for nucleic acid molecular modeling have appeared which when coupled with the rapidly growing database of DNA and RNA structures allows more complex and improved models to be constructed. This software includes the secondary structure prediction program MFOLD and the Vienna RNA package, the tertiary structure modeling programs MC-SYM and YAMMP, and the conformational analysis tool AMIGOS, We are developing software for modeling RNA double helices and improving secondary structure prediction of RNA using empirical structural information. Adapting these tools for DNA is straightforward. Perhaps even more important than the modeling software is the database of nucleic acid structure stored in the Nucleic Acid Database. These structures include examples of damaged DNA alone (i.e. abasic sites) and in complex with protein (i.e. the uracil-DNA glycosylase-DNA complex).

Toward Complete Structural Annotation of DR, and specific DNA Repair proteins

Our second goal is to structurally annotate the roughly 3000 proteins of DR into either low-resolution fold structure and/or higher resolution structures determined from comparative modeling or ab initio prediction. Although the high G+C content of its genome is reminiscent of thermophiles, DR is a mesophile. Thus, its proteins can be expected to resemble those of other bacterial mesophiles, such as *E. coli*, rather than those of other thermophiles. This makes fold recognition and comparative modeling much more feasible since *E. coli* is so well studied at the structural level. Since DR is an extremophile to dessication conditions, comparisons between the DR genome and to *E. coli* and a thermophile like *Methanococcus* will be of interest, and the Werner repair deficiency syndrome of human and these bacteria might also share proteins involved in DNA repair. One overarching question is whether the DNA repair machinery parts are the same between DR and *E. coli*, but the mechanism is different and/or enhanced, or that there are still missing parts in the DNA repair inventory of DR that do not exist in *E. coli*.

We would also want to search for DNA binding proteins that maybe very sequence specific in order to connect to issues of DR being multi-genomic, and at what level of structure is necessary, lower resolution fold or higher resolution structure, to be useful will be explored. Other known DR repair proteins include uvrA (60% homology to UvrA of *E. coli*), uracil DNA glycosylase, thymine glycol glcosylase, AP endonuclease, deoxy-ribophosphodiesterase, RecA (56% identity with RecA of *E. coli*), and DNA polymerase I (51% identity with DNA polymerase I of *E. coli*). The crystal structures of several repair proteins from *E. coli* are known and can be used for comparative modeling. These include proteins such as uracil DNA glycosylase, which are identified in DR, but also proteins such as photolyase which have not yet been located in the DR genome. Differences in the structures between the two bacteria and in particular in their catalytic regions will be examined. UvrA, uracil DNA glycosylase and RecA will be the first proteins modeled. RecA from DR is unusual in that it shares 56% amino acid identity with RecA from *E. coli*, but when RecA from DR is expressed in *E. coli* it is lethal. RecA is detectable in undamaged *E. coli* cells, but is not detectable in undamaged DR cells, only detectable in post-irradiated DR. Structural details of the active sites, or the multi-functional nature of recA, make this an interesting higher resolution structure to know about. A third case is proteins known to be important for maintaining the DR genome, for which homologues are unknown in other bacteria. An example is the IrrI protein of DR, which is thought to inhibit DNA degradation occurring on irradiation. We will use structure based techniques to search for homologues of both *E. coli* and DR proteins which are not found by sequence based searches.

The occurrence of several representatives for each fold allows extraction of the common features of its members, and these data can be used to suggest whether a new sequence is or is not a member of a particular folding class. There are several approaches to fold recognition that are based on either sequence-based clustering of folds, neural networks trained to recognize sequence-fold correlations, and explicit structure-based threading approaches. These approaches are viewed as complementary since they each have been developed with different biological empirical input. While assignment of fold class is useful, current *de novo* and fold analysis techniques provide structural information that is "low resolution". Higher resolution determination of protein structure will be critical in extending the information that emerges from fold prediction to structures that are relevant for investigation of more detailed biochemical questions. Our approach is a global optimization method that has been quite successful in the prediction of small homopolymers and α-helical proteins. The global optimization method was tested on the prediction of a 70 amino acid protein, uteroglobin, and further testing on 2 additional α-helical proteins of ~70 amino acids have been predicted with reasonable resolution using a parallel algorithm on the T3E. We will continue to use the ab initio predictions, and propose to extend the global optimization tool to refine low-resolution structures derived from fold recognition algorithms. Once a new sequence is correctly aligned and assigned to a target fold, a large number of constraints are imposed by the fold topology itself, including the positions of α-helical and β-sheet structure, packing of the hydrophobic core, and definition of the loop regions. The global optimization method will be focused on loop regions that exhibit the most variation in structure, and which are often directly tied to function.

Modeling Higher Order Complexes: Coupling of DNA-repair, replication, and transport

Experiments indicate that restoration of population levels is always delayed beyond the time required to repair DNA damage, and length of time lag correlates with radiation dose. There is an unknown mechanism(s) that senses DNA damage, completion of DNA repair, and when to turn off and on replication machinery. Exonucleases seem to be under some tighter control of an unknown inhibitory protein that is coupled to the sensors of DNA damage, and DNA repair completion, and exonucleases may actually be part of the DNA repair strategy. In radiation sensitive organisms, the more uninhibited degradation of DNA by exonucleases contribute to the loss of genetic material that is primarily responsible for lethal outcome. DR exports damaged bases from intracellular medium, probably to avoid elevated mutagenesis arising from reincorporation of damaged bases into genome during DNA synthesis. Formation of macromolecular complexes between protein-nucleic acids and/or protein-protein interactions are important components of higher order structure or

organization necessary in DNA repair mechanisms. For example, the BRCT domain found in eukaryotes is found among several proteins that are involved in repair and cell-cycle regulation, and in fact these proteins mediate formation of protein-protein complexes that themselves may be the coupling mechanism between repair and replication processes. Avoiding replication errors by tuning in the functional state of these complexes to turn off replication while repair is proceeding is in fact one of the strategies enlisted by the prokaryote DR, although how analogous it is to eukaryotic systems is not yet known.

Formation of specific complexes and protein-protein/protein-nucleic acid interactions is required for such diverse cellular processes as signaling, regulation of gene expression, cell division, DNA repair, etc. Two-hybrid screens provide a high-throughput experimental technique for identifying potential interacting partners. But even when the participating proteins are known, the specific structure of a multimeric complex needs to be determined to obtain a molecular scale understanding of the function of the complex, suggest new experiments, and begin to design inhibitory or stimulatory molecules to interfere with or improve the stability/function of the complex. Yet while over 5,000 protein structures are known, only a few hundred protein-protein complexes have been determined. The development of algorithms for protein-protein and protein-nucleic acid complex recognition is therefore a central problem in computational biology, and has generated many computational approaches.

There are several versions of the protein-protein docking problem, in increasing order of generality: (1) atomic scale modeling of protein-protein complexes identified experimentally, (2) identification of potential partners for a given protein when one or both structures are known or inferred, (3) discrimination of two or more homologous partners for a given protein, (4) biophysical characterization of binding energies, affinities, etc., and (5) blind identification of potential partners at the genomic level. Each of these scales of docking is related to the other, in the sense that a target suggested by a higher level prediction is a candidate for further characterization and refinement with input from experiments or computational structure prediction.

For the target and each potential partner, a multifactorial representation of the protein surface will be constructed. Using either known crystal structures, or structures obtained by homology modeling to known structures, fold recognition and refinement, or ab initio methods, we can represent the surface topography, electrostatic field distribution, flexibility, hydrophobicity, etc. of a globular protein by using a spherical coordinate system whose origin is at the center of mass or other reference point. A variety of parameterizations of surface topography properties can be used. The result is a property vector (r,h,V,...) associated with each point on a unit sphere, which represents the available docking surface of the protein. This representation can have problems with overhanging regions in binding pockets; more intricate representations need to be developed for these cases, such as using several origins. The problem of docking rigid protein structures is then reduced to identifying the proper radial distances and orientations of the two representing spheres, and measuring the degree of complementarity of properties at juxtaposed points on the two spheres. This is a five-parameter search (the three coordinates of the center of the second sphere and the orientation of this sphere, with the first one placed at the origin in a fixed orientation). Rather than a "first principles" approach in which binding energies are calculated, we can use neural networks trained on the known sample protein-protein structures to recognize correlated matches in topography, hydrophobicity, etc. as docking partners. For the purposes of identifying potential partners for further study, the network or other discrimination scheme to tolerate a level of "false positives" while minimizing the number of "false negatives." A further important refinement of docking is cooperative changes in structure that occur on both sides of the protein-protein interface upon binding. At the simplest level, these effects are included in our approach by including a B-factor or other measure of local flexibility as a surface property of the protein. An elastic network model, or candidate interacting pairs detected with the sphere-comparison approach can be refined using all-atom methods of molecular dynamics or fast global optimization.

Regulation of sporulation/competence in B. subtilis.

The exact control of the sporulation/competence switches and the role of competence in cell nutrition and DNA repair in *B. subtilis* has not yet been identified. This is partly because: 1) the molecular parts have not all been fully identified, 2) the modes of regulation of these parts, transcriptionally, translationally and post-translationally, have not been completely elucidated, and 3) the sheer number and complexity of the individual chemical interactions preclude a qualitative decription for how this system functions. In order to facilitate the understanding of this model system we propose to begin a quantitative analysis of these pathways. This will require efforts at all levels of biological data analysis. Genome sequence analysis will need to be used in order to identify further members of the operons expressed during these various processes. For example, operons have been identified that coordinately express proteins necessary for both competence and DNA repair thus implying coupling between these processes. Also identification of upstream regulatory sequences and sigma factor binding sites will prove invaluable. Further, the identification of RNA players, an area much neglected in this field, will be undertaken. Recent data indicates that transcript secondary structure in the leader sequences to certain sporulation genes is an important aspect of their regulation. Further, there are implications that small functional RNAs may play a critical role in the sporulation initiation process. Identification of structural motifs is also proving important. For example, it has been shown that post-translational modification of certain proteins is central to signal detection and sporulation. Data on this latter process led to the hypothesis of an sporulation specific ADP-ribosylating protein that has not yet been found. Structural prediction and analysis may point to possible gene that encode such proteins. Comparison with known and predicted operons may provide a clue to its regulation. Other structural problems include prediction of the multimerization state of transcription factors such as Abr (that may be a hexamer) and the basis of sigma factor specificity. Finally, network analysis can be brought to bear to integrate output from these proceeding analyses and data from the literature to produce dynamical models of the sporulation/competence/motility processes in order to understand the factors that lead a particular cell to choose to execute a particular combination of pathways.

Successful completion of this project will not only provide an unprecedented understanding of complex, cross-talking signal transduction pathways but will also provide a research model of how to database heterogeneous biological data and integrate multilevel analytical tools (from sequence to cell function) towards the systems-level understanding that is important to industry, medicine and the military.