A. Title

Whole proteome reagents and devices as data collection platforms for biological modeling and design.

B. Table of Contents

C. Statement of objectives *(1 page maximum, what questions are addressed?)*

Long term goals: *Saccharomyces cerevisiae* is a self-replicating, self-fabricating cellular device. This device is the basis for our oldest microbial biotechnologies, fermentation and baking, and, as cultures, collections of these devices have been used by humankind for these purposes at least since the Bronze Age. The long term goal of this project to enable complete physical-chemical computational simulation of this organism during normal growth, stress, growth arrest, meiosis and other changes in internal state caused by interactions of its genome and expressed products with one another and with the environment. The outcomes of this project will include scientific ones, including a better understanding of the contribution of individual genes and proteins to fitness of the organism under selective conditions, and engineering ones, in particular the ability to engineer microbial eukaryotes to perform desired functions such as approximate analog computation.

Introduction: In this project, we view yeast as a rugged, adaptable, highly evolved, highly efficient, and versatile cellular device, in fact the oldest cellular devices used by humankind. We will develop wet lab and computational methods to enable high-throughput identification of low molecular weight compounds and protein reagents that interact with every protein its genome encodes. We will use these reagents to devise methods and instruments to learn how this device functions under diverse conditions. We will use output from the DNA chips to monitor the contribution of individual genes to fitness of these cells under different conditions, in particular collecting information on genes that contribute to fitness but are not essential for it. We will develop protein chips and use them to determine the expressed protein complement under different conditions, and thus the identity and rate of appearance and disappearance of different components under described conditions. We will use small molecules and peptide aptamers as ORF-specific reagents to ablate the activity of different components present in these cells under the described conditions, use the DNA chips to quantify genes in the population that affect these perturbations, and use protein chips to quantify the changes in the protein complement. These experiments will thus give us a first order list of the components of these cellular devices, their centrality to the responses described, and a good deal of insight into the pathways in which these components function. The results of these experiments will contribute to a knowledge base of yeast pathways that will represent the effect of particular reagents on cellular components and pathways. The conjunction of the population genetic information provided by the DNA chips with the protein complement information provided by the protein chips will give us unparalleled information on the components, pathways, and contributions of individual gene products to the function of these devices. The results of these experiments will help us engineer new cells to perform new functions, in particular approximate analog computation.

This project addresses a large number of scientific and engineering questions. Of these, we view the following as most important: What is the contribution of every gene in the genome to fitness under different conditions of selection? Which genes are essential and which are important? What is the contribution of each expressed protein to fitness under different selective conditions? How can we make devices and experimental paradigms (heuristics) to measure the above? What are the changes to the required genetic complement and expressed protein complement when function of individual gene products (device components) is ablated? How can we use that information to construct a dynamic, computational representation of the cell that gives us insight into normal (evolutionarily selected) functions and guide our abilities to engineer new functions?

D. Approach. *(1 page; How will the questions be addressed experimentally, theoretically, and/or computationally in novel and interdisciplinary ways? What will each discipline contribute to the effort?)*

In outline (see below for more detail) the research program we have introduced above breaks into a number of tasks. Here, we list what each discipline contributes to each task, in rough order, and then address general issues of novelty and intrerdisciplinary features

1) We will build the DNA chip reader. M, I, B.
2) We will build microfluidics and supporting informatics to get the small molecules. M, I, B.
3) We will build the procedures and fluid-handling steps to select the aptamers. B, M, I.
4) We will build the protein chip. M, B, I.
5) We will build systematic setups to collect data from these experiments, at the cell level and at the population level. M, I, B.
6) We will use information to construct a dynamic computational representation of the cell and the genetic makeup of the population under these different conditions of selection. I, B, M

Every one of the above tasks requires new technology, new methods, new thinking, and new work. The use of DNA chips to monitor fitness of organisms dates first from the late 1990s. The use of this heuristic to select small molecules dates from 1999. The ability to select aptamers against ORFs is new. The possibility of using these reagents to make a protein chip is new and high risk. The experimental paradigms and liquid handling, and data collection approaches for the gene and protein experiments are new. The ability to incorporate this information into a meaningful computable representation is new and high risk.

What is interdisciplinary about the above? Every one of the tasks is interdisciplinary.

E. Significance (*1 page; Why are the ideas interesting to others? Why are the ideas so important that DARPA, an agency interested only in high-risk, high-impact research, should fund the research? How will this effort revolutionize work among your disciplines or lead to the creation of new disciplines? How will others be able to exploit, reuse and extend the work you propose to do*?)

This project addresses basic scientific issues, including the contributions of genes to fitness during evolution, the means by which networks of regulatory proteins effect cellular decision-making, and the design and construction of new biological devices. The first issue was recognized as important in the early 20th century, the second, in the late 1960s and early 1970s, and the third in the1990s, as engineers began to apprehend the possibilities offered by biological systems. In this project, all four investigators are known for their contribution to understanding of these issues and are in a position within their fields to articulate them well enough to cause them to be interesting to others.

These ideas, from the small molecules to the ORF specific aptamers to the population genetic measurements to the protein chips to the simulations, are all high impact. Their risk varies, from significant (beyond easily fundable by NIH, for example, the gene specific aptamers) to high (for example, the protein chips, or the exact nature and value of the computational representation).

How will work revolutionize work among these disciplines or lead to the creation of new disciplines? This work is enabling to biologists at every turn. For microsystems engineers and bioinformaticians, it presents brand new challenges, from handling the population data to making devices that facilitate the selection of small molecules and peptide aptamers, to making the protein chips, to making the computational representation that helps us think about the gene and protein perturbation experiments. We also expect this project to speed the spread of the selection + contribution to fitness experimental heuristic. Moreover, the physical entities we generate, from small molecules, which will define new classes of lead compounds against products of larger genomes, to ORF-specific aptamers, to protein chips themselves, will lead to novel whole genome and proteome studies. The computational entity, the first- order cell representation, will lead to new kinds of experiments to test its predictions. Although we cannot promise that even wild success would lead to new disciplines per se, but I believe it would speed the emergence of two things, the formation of a strong nucleus of young researchers working at this Bio-Info-Micro interface, and the immigration of students from engineering backgrounds into labs where they can design cellular devices.

How will others extend, exploit, and re-use the results of this work? The labs involved have an unrivalled track record for spreading their influence to academic users and to industrial ventures. Davis' lab is responsible, in no particular order, for developments such as the best lambda derived cloning vectors, including the first versatile phage protein expression vectors (such lambda gt11), and about half the uses of microarray technology. Davis' lab has sent directly vectors to >2000 labs, and via reagent companies, to >10,000. Brent's lab responsible for much yeast and two hybrid technology, including interaction mating. Brent 's lab has sent this technology to >2000 academic labs, and, via reagent companies, at least another 2000. Interaction mating two hybrid technology is an important research component in almost all pharma and biotech companies, it is central to some research programs in genomics companies such Millennium, and it is the primary enabling technology of Myriad and Curagen. Arkin and Harbury have emerged as a forces for change and collaboration in the public scientific community, and they sit on the SAB of several new companies. Among them, the investigators are named as inventors on more than 30 issued US patents.

F. Body of proposal.

1. ***What is the idea****? Emphasize its novelty, importance, interdisciplinary features and potential for overcoming fundamental limits to understanding (2 pages).*

In this project, we are proposing to complete a series of difficult but high payoff tasks. Our overall scientific goals are to understand how individual genes contribute to fitness under selection in organisms and populations, how individual proteins contribute to decision making by cellular regulatory networks, and to use that knowledge to allow predictive simulation of organismic and population behavior. Our parallel engineering goals are generation of gene specific small molecule reagents, gene specific peptide aptamers, chips to detect expressed protein complements, informatics to support the use of the above reagents, new micro and macro instrumentation when appropriate, and overarching computational simulations that use this data to make predictions about organismic behavior and fitness.

In particular, in terms of engineering goals, we will:

1) Build a DNA chip reader.

2) Build fluidic and supporting informatic systems to isolate small molecules against every yeast protein.

3) Develop fluidic and supporting technology to select peptide aptamers against every yeast protein

4) Develop devices based on microarrays of protein affinity reagents (protein chips)

5) Systematize of data collection and analysis of data from these experiments

6) Use the information to build a dynamic computational representation of the cell and the genetic makeup of the population under these different conditions of selection

Novelty, importance, interdisciplinary features and potential for overcoming fundamental limits to understanding.

In previous sections of the proposal, we have covered some of these issues, particularly the novelty and the interdisciplinary nature of the work. Here we list, nonexclusively, some of the ways this project, if successful, would overcome existing (perhaps not fundamental) limitations to biological understanding, engineering, and/or computer science.

1) The selection-population survey heuristic for studying the contribution of individual genes to fitness in a population is new, highly powerful, and arguably under-appreciated. Although the first stab in this direction was the work of Smith et al. (1996), its widespread utility and application depended on the construction of oligonucleotide microarrays and their conjunction

with the genome scale deletion projects in the Davis lab (Winzeler, et al., 1999). Recently, this approach has been extended to the use of populations of diploid organisms haploid at a given locus, for example to identify proteins targeted by small molecule inhibitors (Giaever, et al., 1999)

a) If this work is successful, we will have developed instrumentation and data structures to enable us to identify and quantitate the contribution of every gene in a genome to fitness under given conditions. This includes genes that contribute to but are not essential for a phenotype, genes not detected by the classical approach articulated by Morgan, Sturdevant, and Müller, genes that, outside of agricultural applications, have been ignored in most systems in post 1920's genetics.

b) If this work succeeds, we will have selected of small molecule drugs and peptide aptamers that target every gene in a genome. Use of these reagents will enable a kind of genetics at the protein level, by enabling us and other investigators to conditionally inactivate gene products, then quantify the contribution of other genes that may function in the same process to fitness as above.

2) Small molecule drugs against genes in a yeast genome will suggest classes of lead compounds against genes in the human genome-- they will guide combinatorial medicinal chemistry drug development efforts in pharma companies.

3) The ability to isolate peptide aptamers against ORFs will enable more rapid generation of affinity reagents against the products of larger genomes (human) and newer genomes (pathogens). It may also speed the demise of standard immunological methods to generate antibody reagents.

4) The ability to make protein chips will drop the difficulty and cost of analysis of expressed protein complements by a factor of perhaps 1000X. This qualifies as revolutionary. The application of whole proteome analysis to growth of this organism under genetic and environmental selective conditions will greatly increase our understanding of the robust adaptive behavior of this cellular device. Applied to human gene products, it will lead to similar analyses, and to new kinds of diagnostics.

5) Robust computational representations of this information will be useful for teaching and design-based engineering of new unicellular organisms. Their descendants, incorporating information on higher cells, will be useful in teaching, early phase research, and as guides for therapy.

2. Introduction and Background Information *(5 pages). Help the reviewers understand a) why the proposed research problem is important; b) how the ideas relate to existing knowledge that has been generated by you and others in your fields; and c) why an interdisciplinary approach to the problem involving [BIO: INFO: MICRO], as this term is used in the RA, is appropriate. Identify current trends, fundamental limits that need to be overcome and alternative approaches. Preliminary data, manuscripts, and reprints that are relevant to this section can be submitted as an appendix to the proposal (Section K).*

a) Why the proposed research problem is important.

The problems addressed in this research project are fundamental to an understanding of how robust, adaptive, self-replicating systems function normally and respond and adapt to different conditions. *S. cerevisiae* embodies one basic design principle of living systems: the genetic information, its design program, is separate functionally from the organism, the organismic behavior the program gives rise to. Knowledge of the genetic complement, the DNA sequence, can be compared to knowledge of the source code for an inordinately complex program whose language and function is poorly understood. The research described here will speed the understanding of this information.

If this project were to succeed, its importance to 21st century biology would include a better understanding and quantitation of the genes of a genome to fitness under different conditions, with the genes here including genes that contribute to this fitness as well as those that are essential to it. This understanding would be complemented by an understanding of the proteins expressed under these different conditions. The exhaustive list of these gene and protein components is a first step in a long process of unraveling these adaptive functions that others of our experiments will address. The small-molecular and protein reagents we generate will allow selective ablation of function of the genes of the organism under these different conditions, and these reagents will be important in their own right; the small molecule reagents to guide the choice of lead compounds for other whole genome medicinal chemistry projects (in particular, for humans), the protein reagents as affinity reagents for the proteomic studies above and the quick agents to ablate function in whole organisms and test the effects of those pharmacological interventions before small molecule reagents become available, and as substitutes for antibodies, a late Victorian, 19th century technology that most typically still uses mice and rabbits. Protein chips will find use outside of science; chips that detected proteins present in human blood, saliva and urine would find use as diagnostic tools. The computational representation of gene and protein pathway function will find use as a teaching tool, as a tool to guide research, and as a road map for the construction of analogous but more ambitious representations of human genomic and proteomic functional data.

The project would also have impact on device construction. There is significant new engineering needed for the liquid and compound handling involved in the selection of small molecules against gene products, and here the investigators already feel under great pressure to miniaturize, parallelize, and automate these assays to increase the number that can be done per unit time and to minimize the amounts of the potential inhibitory compounds that are needed. Beyond this effort, the protein chip will break new ground. Here, to the investigators' knowledge, a systematic effort to choose the best physical phenomenon we can use to detect binding of underivatized ligands to affinity reagents has not been attempted, and we expect the eventual approach we adopt to become prototypic, and to have positive consequences throughout the "sensor community". Similarly, although we expect to be able to borrow certain steps used in sample prep from the successful efforts of Alan Northup and his coworkers at Cepheid on DNA-based sensors, the issues in liquid handling, charging, and rinsing the sensor are real and their solution will be generic to any protein sensor.

This project will also have important effects on informatics. Here, we can envision three impacts. First and most pedestrian, as we have stated, we expect that the new devices and new data and new wet lab heuristics will all be prototypic, and they will all bring new data handling, data structure, and data analysis needs. Second, and more important, the construction of the computational representation of the knowledge of gene, protein and pathway function will likely break new ground in that it will require representation of qualitative knowledge stored in the existing biological literature. Construction of this representation will help set the path for construction of defined representations of biological knowledge, and we expect that this path will be followed immediately by more ambitious efforts to codify knowledge of human biology. Third, and most speculative, the investigators believe that knowledge gained from an understanding of the mechanisms that these adaptive self-replicating, self-fabricating devices use to perform analog computation to maximize their fitness under changing environmental conditions will speed time in which these concepts are incorporated into crash proof, approximate, asynchronous, highly parallel computational approaches.

b) How these ideas relate to existing knowledge

The above exposition addresses this issue. Here, however, the investigators would like to offer more perspective. For biology, most of these ideas are at or beyond state of the art. Some of these issues, in the context of 20th century genetics, molecular, and cell biology, are discussed by one of the investigators in a recent review (Brent, 2000). For microsystems, design of systems that would help select the small molecule inhibitors would define state of the art, and successful construction of protein chips would be at or beyond state of the art, where the major current method for probing protein complements is 2-D gel electrophoresis followed by protein mass spectrometry. For informatics, the need to use this information to make computational representations with heuristic and predictive value is similarly at or beyond state of the art. These issues are discussed in Brent (2000) and in Brenner (2000) (see appendix K), and any cellular engineering that this work enables will define a new art.

With one exception, the relationship we have claimed for the above ideas to current practice can be easily verified by reference to the open literature. That exception is the selection of small molecule inhibitors of protein function. That task is of course the major activity of the conventional pharma industry, and for that reason its current state is not so widely accessible to researchers unfamiliar with it. Here, we therefore offer a brief perspective. By far the bulk of conventional drug discovery is directed against individual protein targets, more rarely, individual protein interactions, and still more rarely, pathways. In pharma, the normal high throughput screen for compounds thus abstracts one or two molecular players from the cellular milieu, and searches for binding or inhibition of activity. Targets are assayed serially, with as many as hundreds of thousands of potential inhibitors tested against each one in screens lasting months. Because of past success, the choice of potential target and the libraries of potential inhibitory compounds are heavily biased in favor of a few classes of proteins (e.g. G-protein coupled receptors, proteases), and whole classes of potentially important drug targets (most information processing and gene regulatory proteins) are excluded from these searches. By contrast, the work we are proposing to do here will search in parallel for drugs against all the genes of a genome, without being biased toward particular gene products or pathways by our current incomplete understanding, and with something like 10,000X less effort per gene product. We expect at least two gains from this experience: one is the knowledge of how to do this for the products of the human genome. The other, which we cannot stress enough, is the ability to use the chemical structure of inhibitors against gene products for which there are now no inhibitor compounds as a starting point for design of compound libraries that inhibit these now "undruggable" proteins. The gain to pharma and to medicine from opening now intractable targets such as regulatory proteins to small molecule inhibition is likely to be immense; if this effort succeeds, we expect it to enable the discovery of whole new classes of drugs.

c) Why an interdisciplinary approach is important

We believe that the above exposition amply makes the point that this project will stand or fall on the ability of the investigators to integrate biology with device building and with informatics. Of investigators, two began as biologists, one as a biochemist, and one as a physical chemist with an engineering bias. From the biological perspective, none of what we want to do will work without parallel progress on microsystems and computer science. Framed informatically or from the device standpoint, accomplishing these tasks work points the computer scientists and engineers at highly worthwhile problems that only they can solve. As an example of how this influence might be positive, consider that, for the protein chips, our focus on detection of small amounts of large numbers of underivatized proteins is correct, and compares favorably with much of the current work in "biosensors", where good engineering may have been lavished on badly defined and badly constrained problems. What good is construction of sensor that detects 10 exp 3 anthrax bacteria / ml of blood, if that blood concentration of the organism would only be found after death of an infected person?

Identify current trends, fundamental limits that need to be overcome and alternative approaches. Preliminary data, manuscripts, and reprints that are relevant to this section can be submitted as an appendix to the proposal (Section K).

Again, many of these questions have been at least partially addressed in the sections above. We would therefore like to list the important aspects of the proposed work that are at least somewhat conceptually new, for which our proposed courses of action involve significant technical challenges, and for which existing alternative methods and heuristics are not adequate.

Identification of genes involved in processes, and of small molecule inhibitors of their products. As we have discussed, 20th century transmission genetics can in many cases identify genes central to processes, but largely fails to identify contributory genes and does not quantitate their contribution. Identification and quantitation will be necessary for understanding of traits and fitness, and thus of how living systems evolve. As we have discussed, current methods for identification of small molecule inhibitors rests on startlingly expensive, serial, high throughput screens against a narrow range of targets. Should we succeed in perfecting it, the approach here will decrease the effort involved in finding lead molecules by on the order of 10 ^{4}. Here, the problems to be overcome involve compound handling, cell handling, and analysis of the population information.

Selection of anti-ORF aptamers. As discussed, the generation of antibodies to different proteins is a late 19th century technology that still normally relies on the co-injection of the antigen together with an adjuvant into live mice or rabbits. The process is not normally parallel. Immunogenicity of different proteins varies widely, in part depending on the resemblance of the injected antigen to endogenous proteins to which the injected animal may already be tolerant. Chemically pure antibodies can e obtained by making monoclonals, in a procedure that typically involves killing an immunized animal, grinding up its spleen, fusing cells from its spleen with particular cell lines, selection of individual clones of fused cells that produce the desired antibody. The purified antibody is a complex chemical entity that is oxidized, variably glycosylated, and contains multiple polypeptide chains. Antibodies against different antigens isolated as above differ greatly from one another in their affinities and chemical properties, making the chemistry needed for construction of devices based on arrays of them more difficult. Although a great deal of effort has been expended to overcome the above problems, including for example work to enable single chain antibodies, bacterially produced antibodies, antibodies selected not in animals but from combinatorial libraries, the production system described above remains the norm. The conjunction of the chemically homogeneous peptide aptamers and the in vitro selection modality we have developed promises a drastic change for the better, and the in vivo selection modality we describe below may provide an even faster route to affinity reagents. Here, the problems to be overcome include scaling up our existing in vitro methodology, and adapting experience with the in vivo small molecule selection above to selection of inhibitory peptide aptamers.

The ability to develop and use peptide aptamers and small molecules to probe function of individual gene products offers a useful complement to conventional transmission genetics. As the investigators have articulated since the middle 1990s, these approaches promise to extend the reach of genetics to systems and processes not now genetically tractable. The ability to perform genome wide isolation of gene-product specific small molecules would have even potentially greater effects, as it would afford agents to instantly validate targets for drug therapy in mammalian systems, as well as offering good starting leads for medicinal chemistry efforts against protein targets now thought to be intractable to drug therapy. Here, there are hard problems to be overcome, problems we have discussed above. The substitution of protein chips to determine the protein complement of mixtures for the existing 2-D gel / protein mass spectrometric approaches would greatly increase the sensitivity, and the cost and level of effort required to perform these analysis by a factor of around 1000X. That increase in capability would enable their use in genomic studies and in diagnosis and therapy. We imagine a chip, for example, that can analyze a small amount of tissue from a cancer biopsy or blood sample from a patient infected with a pathogen, determine the presence, amount, and phosphorylation state of proteins that regulate cell death, with that information used immediately to guide the course of treatment. Here, there are problems to be overcome with sample preparation, coupling the affinity reagents to the detection surface, liquid handling, and informatics to work the device and understand the data, but the hardest problem may be choosing or devising a sensitive, scalable, cost effective physical method for detecting binding of underivatized proteins to the affinity surface. This we discuss below.

Computational representation of the knowledge gained, and its incorporation into appropriate simulations of cellular organismic and population behavior. Here, there are decent computational methods for simulating cellular processes when all the players are known, their reactions are known, and their kinetic properties are known, and one of the investigators is a known leader in the further development of simulations based on these ideas. There is by contrast no well-established means to perform this simulation work where some of the knowledge is qualitative, and there is no current theoretical framework to connect quantitative information about the contributions of genes to fitness to the predicted behavior of an organism containing those genes. For this work to succeed, we need to tackle the first hard problem; for it to succeed ultimately, for us to understand how the adaptive behavior of cells and populations arises from random variation and selection, we will need to address the second. Note that the first hard problem is one that we think we can address by taking intelligent advantage of existing computer science and pushing it hard, while the second is not an engineering problem at all, it is rather a biological/ mathematical/ theoretical problem, whose solution may occur in this generation, later, or not at all.

3. Technical Approach *(10 pages). Explain the proposed experimental, theoretical, and computational approaches to the problem and highlight the interdisciplinary components of each approach. Provide detailed accounts of how each investigator will work with investigators from other disciplines.*

Task 1) Development of a high throughput DNA chip and analysis software

Task 2) Development of microfluidic and supporting informatic technology to enable identification of small molecule compounds against every ORF product in the yeast genome.

Task 3) Development of fluidic and supporting technology to select peptide aptamers against every ORF product in the yeast genome

Task 4) Development of devices based on microarrays of protein affinity reagents (protein chips)

Task 5) Systematization of data collection and analysis of data from these experiments

Task 6) Use of the information to build a dynamic computational representation of the cell and the genetic makeup of the population under these different conditions of selection

Task 1) High-throughput screening of small molecules

A parallel genome-wide method will be used to identify low molecular weight compounds that inhibit gene function in yeast. This method takes advantage of a complete collection of heterozygous yeast strains, each molecularly tagged with two unique 20bp barcode sequences and precisely deleted for a single gene. Compounds will be obtained from the National Cancer Institute, pharmaceutical companies, and a custom combinational synthesis (developed by one of us, Harbury et al., unpublished). In the experimental protocol, the deletion strains are pooled, grown competitively in the presence of compound, and analyzed for relative growth rate using high-density oligonucleotide arrays. This metric is a measure of the fitness or general health of each strain, allowing the identification of the strains most sensitive to a given compound. These sensitive strains in turn identify the gene product(s) that are directly or indirectly inhibited by compound. These experiments will allow: (1) the identification of novel essential pharmacological targets for yeast and, by inference, human and (2) the identification of novel compounds that act to inhibit these targets.

The ability to perform these screens with large numbers of small molecules is dependent upon a high-throughput system. Figure 1 illustrates the individual steps in the high-throughput screen. The 96-well plates containing growth media are inoculated with the frozen stocks, the small molecules are added into the individual wells, the strains are grown and samples are taken at several time points to measure the growth rate of the individual strains. The chromosomal DNA is extracted from each of these samples, and then the barcodes are amplified and labeled using PCR. The growth rate of all 6000 strains is tracked using Affymetrix DNA microarrays that detect the relative levels of the strains that are tagged with individual barcodes. In this task, Stanford will develop the system of robotics and other equipment, such as the storage and incubation devices, that will fully automate this sample collection process. The data collection and analysis will be the subject of Task 2.

Figur. Flow chart of steps in high-throughput screening.

T ask 2) High throughput DNA chip and analysis technology development

The current DNA microarray technology first developed at Affymetrix in collaboration with the Stanford STDC is not conducive to high-throughput analysis. The 1 cm square chips are enclosed in a bulky package that takes up a large amount of space relative to the size of the actual data collection site. Furthermore, the 1 cm square chips are fabricated on 8 inch wafers, 100 arrays at a time. Performing the hybridization reactions directly on these wafers would reduce the amount of space required for each reaction. Also, these arrays of arrays, combined with the proper fluidics, robotics and optics, will increase the throughput of these experiments by at least a factor of 100. In this task, Stanford will greatly extend the microarray technology by developing the necessary fluidics, robotics, and optics to use the arrays of arrays.

This high-throughput is necessary for identifying the low molecular weight compounds that interact with each protein in yeast, the screen described in Task 1. Since one array is needed for every time point in each experiment involving a pool of compounds or single compounds, many arrays will be needed. In the current configuration with the Affymetrix arrays, the maximum throughput is about 5 per day. With the arrays of arrays, combined with the necessary fluidics, robotics, and optics, the throughput will be over 500 per day, which will greatly facilitate the screening of low molecular weight compounds and their interactions with the various deletion strains.

The steps in this process are illustrated in Figure 1. The arrays of arrays on the wafer are taken directly from the fabrication (supplied by Affymetrix at beginning of project). The samples from the experiments in Task 1 are deposited onto the arrays. This will need to be done with robotic pipettors under environmentally controlled conditions, such as high humidity to reduce the amount of evaporation that may occur with small volumes of liquid deposited directly on these wafers. The use of microfluidics at this stage will be explored as an alternative delivery system for the samples. The arrays will be placed into an oven for incubation, and then removed for the washing steps. Then the data will be obtained from a custom fabricated confocal high-resolution scanner, and collected using a computer with the appropriate image recognition software and database capabilities. The information about what strains are present in what amount at each time point will be extracted from the imaging data.

As in Task 1, this is an interdisciplinary problem that requires the engineering of appropriate components for the high-throughput screen. The engineers involved will learn about the biological applications of the microarrays and the robotics, fluidics, and optics required to support the use of microarrays. The biologists involved in the project will be learning about the robotics, fluidics, and optics that are needed to perform the experiments with the microarrays. The computer scientists involved will be learning about how to deal with the large amounts of data that result from these high-throughput experiments. None of these components can be done in a scientific vacuum, and thus this entire effort requires interdisciplinary collaboration.

Task 3) Development of peptide aptamers against yeast ORF products

Peptide-based aptamers against all yeast proteins will be selected in two ways. First, we will perform affinity selections in vitro. In this modality, a peptide aptamer-encoding plasmid library will be transformed into *E. coli* that will display these peptide aptamers on their main flagellar protein (FliC) (made at TMSI, Colman-Lerner and Brent, unpublished). These *E. coli* will be selected by affinity binding to a set of Gst-fusion proteins that represent the entire *cerevisiae* genome (Martzen, et al., 1999). In our current procedure, we purify the Gst-fused proteins from the lysed Physicki strains by one step precipitation using Glutathione agarose beads, wash the beads thoroughly, mix the beads with bacterial cultures, then spin them out and grow up the E. coli that contain the anti Gst-fusion aptamer. Many of these steps we perform in plates, but not the final selection step, so our process is at this point essentially serial.

To scale this up to whole genome, parallel, isolation, we will elute the proteins from the beads with urea, spot the Gst-fusion protein elutes onto nitrocellulose membranes, wash the membranes in NaCl to renature their epitopes, block the membranes in BSA followed by powdered milk, and then place the membranes in 80 X 40 cm Pyrex baking dishes that also contain 500ml of the FliC library-containing culture at 10^{9} cells / ml. We will agitate the plates slowly (about 0.1-0.2 Hz) for two hours, remove the membranes from the bacterial culture, wash, gently layer the membranes onto solid LB agar medium poured into 80 X 40 cm dishes, incubate at 37o for 12 hours, and collect the colonies formed by *E. coli* that bound the spots. Using 384 well dishes, we can collect Gst-lysates from 384 strains at a time. Depending on the density with which we chose to spot the bead elute onto the membrane, we may be able to fit the entire yeast genome's worth of Gst-ORF proteins onto a single 75 X 35cm membrane.

At the DSTC, we will explore a second approach. We will pool all >6000 Gst fused protein elutes. This pool is a normalized reference set of yeast ORFs in which all genome-encoded proteins have roughly the same abundance. Using the membrane screening approach above, we will isolate a subset of the FliC pool that contains aptamers that interact with these yeast proteins, and deplete that subpool for organisms that contain aptamers that interact with native Gst. We will subclone the TrxA-aptamers into our standard yeast vectors. We will now have a collection of ~100,000 yeast aptamers that bind some yeast protein or proteins. We will use the haploid-dropout approach described above to find small molecules, to find targets for the aptamers. Here, we are interested in finding aptamers that interact with specific proteins, but even more interested in finding aptamers that cross react with numerous proteins in distinct ways: these will be useful in building combinatorial protein sensor suites in task 4.

Task 4) Development of devices based on microarrays of protein affinity reagents (protein chips)

The functional peptide aptamers will be used to construct arrays that can detect the identity and amount of protein molecules in small samples from cells. Two approaches will be employed for building such arrays. First, we will construct arrays that include the full set of 6000+ aptamers specific for each yeast protein. No matter the physical basis of the detection phenomenon we use, these arrays, at least at first, will perforce be large and require larger volumes of liquid analyte. Second, we will construct arrays that contain a sufficient number and variety of cross-reactive peptide aptamers to uniquely recognize each yeast ORF, conjoined with appropriate software to decode the array signal and uniquely identify gene products based on the pattern of aptamer reactivity. In both cases, we are planning to spot the aptamer solutions onto the array surface using a pen-spotter we have built from plans developed by Brown and DeRisi for DNA microarrays (Lashkari, et al., 1996).

We will explore a number of ways to couple the aptamers to the array surface. In initial experiments, we have had good luck coupling TrxA aptamers with His-6 at their amino termini to gold layers using the alkane-thiol linkages pioneered by Whitesides and his coworkers, and Nickel complexed NTA at the opposite end of the molecule (invented by Cynthia Bamdad and George Whitesides, see Barberis, et al., 1995). These form the now-canonical self-assembled monolayers, and the His-tagged aptamers bind the Nickel with their variable regions exposed to the analyte (Roger Brent and Cynthia Bamdad, 1995, unpublished). Although adequate for preliminary experiments, for more rugged devices we will probably need to covalently link the aptamer to the detector. At the moment, our working idea is to endow the TrxA aptamers with a 21 residue amino terminal extension comprised of 5 lysine residues followed by 8 glycine-serine dipeptides, and couple the amino terminal tails of the aptamers to the detection surface with lysine chemistry, the amino acid chemistry we are most comfortable with.

Although there are likely to be hurdles in getting this to work, we do not intend to allow problems in the covalent coupling chemistry/ materials science aspects of this portion of the project to slow it down. Accordingly, we have entered into first order discussions with Steffen Nock and his coworkers at Zyomyx, Inc. in Hayward California, which has developed number of proprietary coupling methods, to gain access to their proprietary coupling methods. One of the investigators (RWD) serves on this company's advisory board.

We will also need to decide on the physical phenomenon we will use to detect binding of proteins in analyte to the chip surface. Existing methods in the "sensor community" seem inadequate to this task, and, at the moment, we are considering the further development of a number of means, all based either on the change in mass or dielectric constraint caused by binding of a protein to the detection surface, into devices that manifest the threshold performance we seek. Here is more detail on the effects we are considering.

Planar evanescent wave approaches. Surface Plasmon Resonance (SPR) is an excitation of electrical waves in a metal when light of a particular wavelength, such as a laser, strikes the metal at a particular angle. This is measured as a decrease in the reflected laser light at that angle. Changes in dielectric constant at the surface of the metal can alter the both the angle and the wavelength at which SPR occurs (Wahling, et al., 1979). This is the basis of the BiaCore, a widely used commercial instrument. In fact, the way for us to make a detector would be for us work out the coupling chemistry and actual detection on the BiaCore, but move eventually to a larger gold-coated glass surface on which the aptamers are immobilized. For our purposes, we can easily imagine a gold-coated glass surface that we can interrogate with a laser from underneath. The single laser could raster scan the underneath of the surface. Alternatively, Zyomyx (Steffen Nock, personal communication) and Hitachi (Yuji Miyahara, Hitachi Instruments, Ltd.) have both developed camera based whole surface interrogation and imaging of array surfaces to detect evanescent wave changes, and we may enter into discussions to attempt to access this technology.

Other evanescent wave effects. At the moment, we prefer to take an alternative and less costly approach to the planar SPR techniques. Figure illustrates a technique developed at TMSI whereby individual 60 nm silver Plasmon Resonant Particles (PRPs, Schultz et al., 2000) are used as detectors of approximately 100 protein binding events. Like the planar SPR effect, scattering from the PRPs changes when analyte molecules are bound to the surface. In this case, rather than observing the change in reflected light at a particular angle, we measure the spectral shift in scattered light. The wavelength of the light scattered from PRPs is a strong function of particle diameter. The scattering peak of 60 nm particles is 500 nm. A single PRP is bright enough to see by eye on a dark field microscope (Carlson,unpublished). We anticipate being able to use a photomultiplier tube and a CCD camera to measure an scattering from an array of such beads on a slide **(see figure**).

Detection based on mass changes. We are also considering detecting protein binding based on changes in mass. One established method here (albeit to date for chemicals such as organophosphates and aromatic compounds rather than protein analytes) is the surface acoustic wave. These devices are based on the ability to fabricate piezoelectric material. A voltage applied to a piezoelectric produces a change in thickness; likewise, a change in geometry brought about by squeezing or shaking a piezoelectric can be measured by a change in voltage. In a Surface Acoustic Wave (SAW) device, a bridge material such as quartz is used between two piezoelectric regions as a mechanical filter. Acoustic waves broadcast from one piezoelectric element to the other are filtered by the frequency dependent mechanical response of the bridge (see for example Hierlemann et al., 1999). SAW's can be used to measure protein binding if the bridge region is coated with a protein affinity reagent (Greg Smith, personal communication): in such devices, binding events increase the mass of the bridge, changing its mechanical properties, thus changing the acoustic spectrum transmitted across the bridge.

Piezoelectric devices can also serve as simpler mass sensors. Microfabricated cantilevers oscillated by in-built piezoelectric areas have mechanical resonances measurable by integrated piezoelectric sensors. Proteins bound to the surface of the cantilever causes mass changes which are detectable by changes in the mechanical response of the cantilever. For both kinds of mass sensors, we are working with Greg Smith, now a consultant, who will come to MSI to help with design and fabrication if we adopt this route. Smith is a gifted designer and engineer, who acquired significant experience with cantilever micromechanical devices to study protein binding during years spent on related projects at Affymax, Inc. For the actual fabrication, we are in first-order discussions with Zyomyx, Inc., and with Fujitsu Ltd. (Ryusuke Hoshikawa, President, Advanced Microdevices) for SAW and other MEMS devices.

Detection based on electronic phenomena. In the long term, we would like to get beyond optical and mechanical/acoustical phenomena, and use an electronic effect to detect protein binding. Such devices would likely be much cheaper and much easier to fabricate in large production runs. Our hope that such phenomena can be found and exploited stems from recent explorations of device physics in the condensed matter physics community, which have resulted in measurement techniques capable of detecting the movement of small numbers of electrons. Using these techniques it is already possible to measure differences in the microwave resonance spectra of different proteins, thereby directly identifying proteins by their physical characteristics (Hefti, et al., 1999). Presumably, there is a predictable relationship between sequence and/or structure of a protein and its resonance spectrum. However, whether prediction is possible or not, the spectrum should be consistent, and binding of a known protein analyte to the resonance affinity surface should change it consistently.

Even simpler devices may be possible, based on changes in capacitance. Here, the task is to detect the presence of proteins bound to antibodies positioned between electrodes. Instead of quantifying differences in spectra, this technique would simply detect a change in the dielectric constant between two sides of a capacitor and quantitate that change as a function of frequency in a power spectrum.

Task 5) Systematization of data collection and analysis from these experiments

This project will perfect and systematize one powerful data gathering heuristic, the dropout population assay, and bring another, the chip, into being. Both of these have fluid handling and data handling needs, some of which are different, and some of which are common.

For the dropout population assays, we have described the wet parts, the mechanical systematization of the data collection, which constitute significant technical challenges, separately in Task 1. For the protein expression profiling using the chips described in Task 4, our initial approach to systematizing the fluid handling will be quite conservative. We plan to lyse and prepare protein extracts from cultures different of yeast using quite standard, non-automated, procedures in wide use at TMSI and DSTC, and then expose the detection elements of the device to the cell extract in a single chamber. That is, we may have an array that contains more than 6000 individual detection elements, the entire array will be in a single flow chamber or cell, and every element on the array will be exposed to the same liquid. The first order liquid handling for these devices will be by simple pumps and controllers, and we may (Brent Kreider, Phylos, personal communication) be able to adapt the control software and pumps used for fluidics in the BiaCore instrument directly to flowing sample and wash solutions over a protein array.

For the protein chip expression experiments, there are relatively simple (albeit nontrivial) front end informatic needs involved in storing the raw binding information in a way that more sophisticated analysis programs to make use of it. Still, here the "databasology" is simple, in fact the simplest database problem we can anticipate in proteomics so far (compared, for example, to the database related problems inherent in collecting quantitative data from protein mass spectrometry, and identifying proteins from their individual mass peaks). Here, the significant challenges lie downstream.

For the dropout population assays, and for the protein chip expression assays, there are significant common data handling and analysis needs. Development of tools for these efforts will be focused on five areas: 1) Collection of raw fitness and expression data, and error modeling of these data sets 2) "Knowledge/Data-basing", computational representations of yeast biology that we already know, 3) Statistical data modeling, based on information from the first area 4) Analysis of response of the cells to different broad spectrum aptamers and different small molecule inhibitors from task 2, and using that information to identify cellular pathways affected by these compounds 5) Identification of compounds that affect new targets and new pathways based on fitness/expression spectra.

Once a final error model is derived for each kind of data, we will use it together with information from new population and expression experiments to perform quantitative phenotyping of cellular response to challenges. The value of discerning phenotypes from large numbers of quantitative experiments is now becoming widely acknowledged in the gene expression monitoring community (Steve Friend, Rosetta Inc., personal communication), and the analytical tools to extract meaning from this becoming less primitive (Eisen, et al. 1999, Tomayo, et al., 1999). We will build on this ongoing development to develop computational tools that allow us to detect of “significant deviates” from previously observed strain fitness and molecular expression patterns that might indicate collection of a novel bioactive compound. These quantitative phenotypes will include analyses of sensitivity of strain growth and molecular expression patterns of different haplo- and full- mutant strains to different chemical challenges. We will then arrange those phenotypes based together with the knowledge base onto known pathway information to derive molecular level explanations for the observed sensitivities and to predict which genes and systems in the yeast cells are affected by and respond to different types of chemical compounds.

Task 6) Use of the information to build a dynamic computational representation of the cell and the genetic makeup of the population under different conditions of selection

Although this proposal contains a number of highly ambitious tasks, this last is probably the most difficult. Nevertheless, during the project period we hope to make a significant start on the problem, and any gain we can make from doing so can have both long-term utility and can impact favorably in the short term on the success of other aspects of the project.

Conceptually, we divide this task into three subtasks. The subtask that is best developed is still highly challenging. That is the quantitative modeling of intracellular events. For such work, one of us (Arkin) has pioneered the development of computational methods that can effectively simulate network behavior when the players, reactions, and reaction rates are known (McAdams and Arkin, 1997). This work goes beyond the computational numerical solution of systems of differential equations and instead makes use of stochastic reaction models developed by Gillespie (1977). Recently Arkin and Drew Endy at TMSI have further extended these methods by using algorithms developed by Yehoshua Bruck and his coworkers at Caltech (Jehoshua Bruck, personal communication). At TMSI, Brent, Endy, and Colman-Lerner have developed a quantitative representation of one such well understood set of intracellular events in yeast (the response to mating pheromone) and have developed experimental tools to allow the testing and adjustment of the parameters of the representation of this pathway. Extension of these experimental heuristics to other yeast pathways should allow, during the project period, such a representation of the major protein based signal transduction pathways that yeast use to respond to other changes in their environment (heat shock, cold shock, high osmolarity, low nitrogen, loss of "cell integrity"), all of which are nearing the level of understanding that now characterizes the mating response.

We see the second subtask as the construction of computational representations based on less organized, qualitative knowledge of the biology of this organism, and the integration of such representations at need with more quantitative representations where knowledge is sufficient to support quantitative representations. The common term for such representations is "knowledge bases". Although we cannot now state specifically how we are going to construct such representations, we can state that we are casting a wide net. We are exploring a number of approaches for such representations, from those used to represent wiring and plumbing on robotically guided craft (Brent and Nicola Muscettola, NASA Ames) to "qualitative calculus" (Kuipers, 1994, Larry Lok, TMSI, unpublished), to those contemplated by the UT Southwestern Alliance for Cellular Signaling (Arkin). For now, our approach will be to try during the next year to determine which sorts of representation can best be developed to the point that it at least promises the ability to make computable our existing understanding of yeast signaling, DNA replication, meiosis, sporulation, protein transport, and germination pathways during the project period.

The third subtask is to attempt to generate computational frameworks that allow us to understand cellular events using information gained from the population selection heuristic. This is the hardest problem. The beginning of experimental ability to quantitate the contribution of all the genes in an organism to fitness under selective conditions is less than five years old (see Smith et al., 1996) and it was only the development of the bar code paradigm that made it technically approachable (Winzeler et al., 1999). At the moment, there is no analytical framework that connects this experimental data either to general statements about the mechanism by which those genes contribute to fitness (here, even the development of quantitative criteria to classify genes into essential and important and the demonstration that these matched known biology would be useful), much less to the predicted behavior of populations of organisms containing different lesions. As in the second subtask, our approach to this problem is now to cast a wide net for mathematical/ computational/ population biological tools that might be relevant, to decide early in the project period whether any of these seem to justify great effort, and to otherwise confine work under this subtask to mechanism free efforts to classify the importance of genes to phenotype.

Here, however, the use of gene specific small molecule and peptide aptamer reagents will allow conditional ablation of gene function during the selections. The effect of these perturbations on the fitness of individual members of the population will reveal genetically important interactions between the targeted protein and the missing protein in the individual mutant members of the population, and that interaction information will go directly into our knowledge base.

However primitive our first efforts are, computational representations of this information will be useful for teaching yeast biology, and we expect them to be particularly to engineering students who will begin to be working with this organism and attempting to use it as a platform to construct custom microbial small molecule biosynthetic devices and information processing devices; because so much engineering is now design-based, we expect that engineering students will enthusiastically use such simulations to explore the biology of this construction platform. Because both the population selection and protein chip experimental heuristics are particularly applicable to small genome microbial organisms, the facile ability to construct even crude simulations from these data types may facilitate the rapid construction of useful simulations for microbial pathogens (and recall that the population selection methods we describe earlier point directly and proteins to be targeted for anti-microbials, while the small molecule compounds are actually drugs or leads for drugs). Descendants of these simulations incorporating information on higher cells will be useful in teaching, early phase research, and as guides for therapy.

4. *Milestones (1 page).* *Provide an outline of proposed tasks and the research schedule. This material will be used as a statement of work for the purpose of negotiating the grant* **.**

Here again are the separate tasks for this project framed in terms of their engineering goals.

1) Build the DNA chip reader.
2) Build microfluidics and supporting informatics to get the small molecules.
3) Devise the procedures and fluid-handling steps to select ORF-specific aptamers.
4) Build the protein chip.
5) Build systematic setups to collect data and analyze data from these experiments, at the cell level and at the population level.
6) Use information to construct a dynamic computational representation of the cell and the genetic makeup of the population under these different conditions of selection.

Here milestones related to these tasks that we expect to have completed at the end of each designated year.

Year 1. Build chip reader, obtain at least two small molecule libraries from , begin working with Harbury library, obtain compound library from NCI, select in vitro aptamers against ~100 yeast ORF products, determine preferred method to detect binding of underivatized proteins, demonstrate aptamer coupling to detection surface, begin informatic work on data collection and analysis, demonstrate quantitative mating pheromone model, determine data structure for knowledge base.

Year 2. Obtain access to at least one pharma company library of small compounds. Select small molecules against ~100 yeast proteins including against at least 1 against a target the pharma industry would consider "undruggable", obtain aptamers against >6000 yeast proteins and combinatorial aptamers with different cross reactivities against pooled proteins, demonstrate detection of submicromolar amounts of protein to aptamer-linked detector element, demonstrate working data collection and analysis structure for bar code small molecule experiments, demonstrate program specification for collection and analysis from protein chip, demonstrate additional quantitative models of yeast pathways, deploy working first order knowledge base, demonstrate informatic tools and criteria for determining contribution of individual genes to fitness under given selective conditions.

Year 3. Select small molecules against a set of ~1000 yeast proteins including numerous proteins hitherto thought undruggable, demonstrate arrays based on ~1000 ORF specific yeast aptamers and a device based on combinatorial binding to smaller numbers of cross reactive aptamers, demonstrate census by these devices of yeast protein complement under different selective conditions, demonstrate working program for collection and analysis of this data, extend quantitative models for yeast signaling pathways to maximum extend allowed by data, incorporate qualitative knowledge for all major yeast cell biological processes into knowledge base, demonstrate experimental and informatic ability to perturb population of mutants with small molecules and aptamers and measure interacting genetic loci by change in frequency in population, and ability to incorporate that information into knowledge base.

Years 4 and 5. Select small molecules against all yeast proteins. If selection of small molecules against some yeast ORFs proves difficult, design new libraries to find classes of molecules that will. Demonstrate array devices that either contain aptamers against all >6000 yeast proteins or that contain cross-reactive aptamers sufficient in pilot experiments to discriminate among all >6000 ORFs. Continue to fill knowledge base with information on pathways gathered from perturbation experiments. Have worked with math, computation, and population biologists to establish theoretical basis for inference of function from expression monitoring and population selection experiments.

5. *Information and Technology Transition (1 page)*. *Describe the offeror's policies, practices and plans for sharing results with the larger research community and for transitioning information and technologies that emerge from the offeror's efforts.*

Stanford University, the offering institution, as well as the two subcontractors, TMSI and UC Berkeley, have well established technology transfer policies. In the case of Stanford, these policies, continued over more than two generations, arguably have set the pattern for incubation of technology-based industry in the US and the world.

In practice, as mentioned above, these investigators have a good track record for spreading their influence to academic users and to industrial ventures. Davis' lab is responsible for about half the impact of microarray technology in academic and corporate settings, but that is only the most recent large contribution. He and his coworkers have a major share of the credit for the promulgation of phage based recombinant DNA technology in the 1970s and 1980s Davis' lab has sent directly vectors to >2000 labs, and via reagent companies, to >20,000. Brent's lab is responsible for much yeast and two hybrid technology, including interaction mating. Brent lab has sent this technology to >2000 academic labs, and, via reagent companies, at least another 5000. Interaction mating two hybrid technology is an important research component in almost all pharma and biotech companies, is central to some research programs in genomics companies such Millennium, and it is probably the primary enabling technologies for Myriad and Curagen. Arkin shares credit for the promulgation of stochastic simulation methods in biological modeling. Harbury's work with Kim has led to new insights into membrane fusion with consequences for antiviral therapy. The investigators serve as consultants and SAB members for numerous companies and are named as inventors on more than 30 issued US patents.

Perhaps significantly, none of the investigators are well known for starting companies that gain exclusive rights to their own work. All have rather tended to adopt the practice of working with numerous companies. For phage cloning vectors and two hybrid methods, such a mode of operation has arguably vastly speeded commercialization of the methods compared to the rate that might have resulted from more exclusive licensing, albeit with less financial gain for the inventors.

Thus, the plans of the investigators to commercialize technologies resulting from this work are different, depending on the nature of the technology (different routes for protein chips and computer programs), contingent, with a high degree of willingness to decide on the spot as developments dictate, relatively unrestrictive, due to the bias for working with many partners described above, and confident, based on highly successful experience extending over decades.

6. ***Continuation plans (<1 page).*** *Provide an indication of the steps that will be taken by the institution(s) to sustain the research efforts after the DARPA award has ended.*

The most important statement to make here is that the research directions described in this project here are central to the research interests of each investigator. That is to say, each investigator in this project has a strong and ongoing commitment to the research directions described. All three institutions are strongly committed to their investigators and to furthering of this research in its own right: Stanford has launched the Bio-X program, TMSI was established to perform this work, and UC Berkeley has UC Berkeley has allocated funds for this work and related projects through the Health Sciences Initiative that fosters collaboration among biologists, engineers, chemists, physicists and mathematicians..

Here, however, DARPA is in a position to be enabling: for example, although there will be ORF specific aptamers, without this project, there may not be a genome's worth, and there will not be a protein chip; although there will be a knowledge base of protein networks, there may not be an ambitious one that tries to conjoin qualitative information with the information derived from protein expression and population selection experiments.

It is useful to note that the project is larger than can be accomplished with DARPA funding. In this work, however, the issues are so important, and the gain from even incremental progress so great, that the investigators and their institutions anticipate being able to replace DARPA funds with other government funds (for example, from the NIGMS "glue grant" program) and funds raised by their institutions from corporate sponsors or even private philanthropists when the need arises and the risk has decreased. In fact, due to the fact that even early, partial progress will have great impact, this proposed research is a good example of an instance where front-loading any DARPA funding would make great sense.

7. *References to relevant related work*. *This should not be used as a substitute for content (not included in page count).*

Barberis, A, Pearlberg, J., Simkovich, N. Farrel, S., Reinagel, P., Bamdad, C., Sigal, G. and Ptashne, M. (1995) Contact with a component of the polymerase II holoenzyme suffices for gene activation. Cell. 81, 359-368.

Brenner, S. (2000) Theoretical biology in the third millennium. Phil. Trans. R. Soc. London 354, 1963-1965.

Brent, R. (2000) Genomic Biology. Cell. 100, 169-183.

Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J. and Brent, R. (1996) Genetic selection of peptide aptamers that recognize and inhibit Cyclin-dependent kinase 2. Nature. 380, 548-550.

Eisen, M. B, Spellman, , P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14868.

Giaever G., Shoemaker, D. D. , Jones, T. W., Liang, H., Winzeler, E. A., Astromoff, A., and Davis, R. W. (1999). Genomic profiling of drug sensitivities via induced haploinsufficiency. Nat Genet. 21,278-283.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340-2361.

Kuipers, B. (1994). Qualitative reasoning: modeling and simulation with incomplete knowledge. MIT Press. Cambridge, Massachusetts.

Hefti, J., Pan, A., and Kuman, A. (1999). Sensitive detection method of dielectric dispersions in aqueous-based, surface-bound macromolecular structures using microwave spectroscopy. Applied Physics Letters 75, 1802-1804

Hierlemann, A., Ricco, A. J., Bodenhofer, K., and Gopel, W (1999) Effective use of molecular recognition in gas sensing: results from acoustic wave and in situ FT-IR measurements. Anal. Chem. 71, 3022-3035

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. , Davis, R. W. (1996) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA. 24, 13057-13062.

Martzen , M. R., McCraith, S. M., Spinelli, S. L., Torres, F. M., Fields, S., Grayhack, E. J., and Phizicky, E. M. (1999). A biochemical genomics approach for identifying genes by the activity of their products. Science. 286, 1153-1155.

McAdams, H. H. and Arkin, A. (1997) Stochastic mechanisms in gene expression. Proc Natl. Acad. Sci. USA. 94, 814-819.

Schultz, S., Smith, D. R., Mock, J. J., and Schultz, D. A. (2000) Single-target molecule detection with nonbleaching multicolor optical immunolabels. Proc. Natl. Acad. Sci. USA. 97, 996-1001.

Smith V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science. 274,2069-2074.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Sutisak Ketareewan, S. Dimitrovsky, E., Lander, E. S., and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoetic differentiation. Proc. Natl. Acad. Sci. USA. 96, 2907-2912.

Wahling, G. Raether, H. and Mobius, D. (1979) Studies of organic monolayers on thin silver films using the attenuated total reflection method. Thin Solid Films. 58, 391-395.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Davis, R.W., et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science. 285 , 901-906.

G. ***Facilities available at the offeror’s institution or via affiliated institutions, especially unique facilities or capabilities (2 pages).*** *This section should also be used to illustrate the specific steps that the offeror’s institution has taken or will take to demonstrate its commitment to establishing and maintaining an environment that is supportive of interdisciplinary efforts undertaken by its faculty, students, and postdoctoral fellows.*

Unique facilities. Both the DNA Sequencing Technology Center and The Molecular Sciences Institute are dedicated facilities, deliberately set up to enable development of new science and technology in a multidisciplinary environment, and both are outstanding environments to create these new methods and heuristics. At Stanford, the location of the Harbury lab juxtaposes the lab's small molecule libraries with one of the best chemistry departments in the world. Similarly, LBL offers an unparalleled environment for exploration of physical phenomena and their application to biology, and UC Berkeley offers the best computer science departments in the world.

Investigator commitment to establishing interdisciplinary efforts. The investigators involved in this project have a track record of working with investigators outside of their disciplines. Ron Davis is the director of the Stanford DNA Sequencing and Technology Development Center, which consists of a large (75+) interdisciplinary team of engineers, physicists, and molecular biologists working on various aspects of genomic sequencing. Davis has had collaborations with principal investigators from physics: Calvin Quate and Steve Chu; from engineering: Fabian Pease and Bill Reynolds; from chemistry: R. Mathies (UC Berkeley); and others. Davis has also recently entered into a collaborative effort with Harley McAdams, Adam Arkin, Leonard Adleman, and Roger Brent. Roger Brent is the Associate Director of the Molecular Sciences Institute, a 25+ group of biologists, computer scientists, and physicists. In addition to the interdisciplinary team within the Institute, Brent has continuing collaborations with principle investigators from other disciplines, including Ron Davis, Adam Arkin at UC Berkeley, John Clarke at UC Berkeley, Chris Sander at Millennium Predictive Medicine, and Louis Gerstein at Yale. Pehr Harbury has productive collaborations with numerous structural biologists and chemists. Adam Arkin is an assistant professor of Bioengineering and Chemistry at University of California, Berkeley and a Faculty Scientist in the Physical Biosciences Division of Lawrence Berkeley National Laboratory. The nature of his work naturally leads to interdisciplinary collaborations. Currently he has collaborations with Don Naki and Fernando Valle, biologists at Genencor International, Antje Hofmeister, a biologist at UCB, Luke Lee, Dorian Liepmann and Terry Leighton (engineering and biology) UCB, John Little (Biology, University of Arizona), Mark Ettinger (Applied Math, Los Alamos National Laboratory), Susan Graham (CS, UCB) and Phil Coella (Applied Mathematics, LBNL), Ruth VanBogelen at Parke-Davis, Peter Karp (SRI), Milton Saier (UCSD), and Tyrrell Conway (U. Oklahoma) and finally, is part of the Alliance for Cellular Signaling run by Alfred Gilman (UT Southwestern Medical Center). The fact that these investigators are now working together is strong evidence of their commitment to this kind of work.

Institutional commitment to establishing and maintaining environments supportive of multidisciplinary research efforts. Stanford has historically been committed to interdisciplinary research environments. This commitment is most recently evidenced by the Bio-X initiative, which has begun to bring together in one building faculty from biology, chemistry, physics, engineering, and medicine to maximize their impact on one another. The Bio-X building will contain over 50 faculty from the various schools in the university. Similarly, the Molecular Sciences Institute fosters a research environment dedicated to multidisciplinary research, and maintains close ties to both UC Berkeley and UCSF research communities. Finally, the Calvin lab at UC Berkeley now hosts such an environment.

It is worth mentioning that both the DNA Sequencing Technology Center and The Molecular Sciences Institute are off-campus research environments whose predominant personnel are research fellows. Although most of the research activities described here are highly specialized, some are the proper province of graduate students in biology, engineering, and computer science, and would provide outstanding training for young researchers at that phase of their careers. Accordingly, the investigators place great priority on making sure that these two off-campus research sites are outstanding places for graduate students to work, and have explicitly planned (see attached letters, appendix K) to facilitate the flow of qualified students to these project sites.

H. CV's

I. Current and pending support.

1. Project title and summary
2. Source and amount of funding (annual direct costs; identify agencies and provide grant/award numbers for current grants/awards)
3. Percentage of effort devoted to each project
4. Relationship (and degree of overlap) of project to the proposed effort

J. Other agencies to which this proposal was submitted.

K. Appendices. Pertinent preliminary data, manuscripts, reprints, and other supporting materials may be included here. This should NOT be used as a substitute for substantive content in previous sections. (3 copies only).

L. Budget

1. Personnel salaries, wages and fringe benefits (indicate percentage of effort for faculty).
2. Equipment purchases and maintenance (provide purchase justification and basis for estimates, e.g., bids, catalog prices, recent similar purchases).
3. Materials and supplies (itemize major categories).
4. Travel (indicate purpose; identify and justify foreign travel).
5. Other direct costs (itemize major categories).
6. Subcontracts (provide supporting data to justify amount(s)).
7. Indirect costs.
8. Totals (provide annual breakdown and cumulative summary).
9. Supporting cost or pricing information to justify items 1 through 7 above.