Harvard Medical School

DARPA BAA 01-26 (BIO-COMP) **3 May 2001**

Section I. Administrative

Technical Area:	DNA Computing (& Computational Models 2: Experimental Validation)
Proposal Title:	DNA Memory and Input/Output
Type of Business:	Other Educational

Technical Contact: George Church	**Administrative Contact:** Cindy Reyes
Address: Dept. of Genetics 200 Longwood Ave. Boston, MA 02115	**Address:** Dept. of Genetics 200 Longwood Ave. Boston, MA 02115
Phone Number: 617-432-7562	**Phone Number:** 617-432-1278
Fax Number: 617-432-7266	Fax Number: 617-432-7266
E-mail: church@arep.med.harvard.edu	**E-mail:** reyes@arep.med.harvard.edu

Summary of Costs

Total Base:	1,920,154	**Total Option:**	1,404,511
FY1 (8/01-7/02):	620,721	**FY4 (8/04-7/05):**	688,486
FY2 (8/02-7/03):	637,427	**FY5 (8/05-7-06):**	716,025
FY3 (8/03-7/04)	662,006		

http://www.darpa.mil/ito/Solicitations/PIP_01-26.html

Section II: Detailed Proposal Information

A. Innovative Claims

A.1 Bit-per-bp Memory

This proposed research would push the limits of DNA storage devices: density, accuracy, and archival lifetime. We will pursue two novel and compact DNA memory options based on site-specific **recombinases** (SSRs) and error-prone DNA **polymerases** (e.g. eta), with the goal of achieving close to the theoretical density of one bit per base pair (bp) in sub-nm^3 volume. The ability to faithfully duplicate an entire array of recorded bits will be developed.

A.2 I/O: Input

The project will focus on real-world Input/Output issues including **analog-to-digital (A/D) and digital-to-analog (D/A)**. Published DNA computing has focused on computational algorithms rather than I/O or memory. Such emphasis is illogical since the DNA-enzyme clock rates are typically 6 logs slower than the GHz expected of electronic- optical (EO) computing and DNA is not (in the long term) more parallelizable than EO. For system input, we harvest a diverse set of biochemical and biophysical sensors. We also propose a novel fusion of DNA and RNA polymerases to decouple positioning from synthesis.

A.3 I/O: Output

In our Output focus we address the utility of using **Polymerases for positioning** mechanical effectors and hence rapidly synthesize three-dimensionally complex EO computers. This should be compatible with the DNA-bit programming done in the system input (task A.2).

A.4 Optimization: minigenomes

In order to improve the performance of the fabrication and memory tools, we will develop **in vitro replication/translation arrays** for experimental feedback. We will design a 90kbp minigenome capable of replication and protein-synthesis. This minigenome will be 6 times smaller than the smallest living cellular genomes, and display up to 800-fold faster replication, with 1000-fold fewer molecular components.

A.5 Optimization: Computational modeling

These in vitro replicating systems will be **ideal for integrating with detailed computational models,** due to simplicity, knowledge of the 3D structure of nearly all components and extreme experimental accessibility. Also coupling the extremes of modeling (from single base changes to 3D structures to molecular networks to population doubling selection) is likely to be dramatically more transparent and tractable.

A.6 Novel, Useful Applications & technology transfer

We will focus from the start on practical **applications that take advantage of the unique features of DNA** rather than competing head-on with EO. Examples are: (a) proven information archiving and retrieval (up to a billion years as mineralized fossils or living DNA records); (b) interfacing with biochemical, photon, or thermal sensors. (c) A DNA recorder analogous to black-box flight recorder would take early advantage of our ability to record on DNA more easily than reading it. Only rarely would the archived materials be accessed. (d) Polymerases take 0.34 nm steps under control of available dNTPs. Novel methods for separating the positioning from the incorporation of reactive bases will allow nanofabrication.

We will leverage our experience with DNA-tags, single-molecule manipulations, polymerase microarrays, computational genomics, homologous recombination, and transfer of the above to many commercial and academic groups. We believe the tech transfer plan itself is innovative as a focus for the model integration proposed in related DARPA proposals.

B. Proposal Roadmap

Item	Summary Statement	Section
B.1 Main goal of the work	Design & test molecular memory, I/O & (in vitro replication based) nano-fabrication. Develop models & measures of defense & health relevance.	A
B.2 Tangible benefits to end user	A test-bed for community integrated modeling & measurement capable of nano-manufacturing useful proteins & smart-materials.	G
B.3 Technical barriers preventing past achievement	(1) Genome sequence annotation & full set of working in vitro components, (2) comprehensive array readout methods and fluorescent protein encoding, (3) anticipation of stochastic effects with models.	A, H
B.4 Main elements of the proposed approach.	Design & testing of (1) DNA memory at one bit per nm^3 using DNA & RNA polymerases. (2) Real-world A/D Input & (3) D/A Output for the above in vivo and in vitro. Integration of micro- & nano-array based parallelism. (4) A replicator/fabricator system for optimization of manufacturing & readout. Elements of a 90 kbp complete replicator. (5) Optimization & integration modeling.	C.1 C.2 C.3 C.4 C.5
B.5 Specific basis for confidence	Our group track record on design-to-commercialization of relevant & similarly hard projects, e.g. genome sequencing & modeling (GTC), single-molecule DNA conductance (Agilent), in vitro DNA replication (Mosaic), DNA-enzyme arrays (Affymetrix), instruments (IAS), etc.	C.6 G
B.6 Nature of expected results	A shareable rapid prototyping BioSystem integrated with the DARPA/BBN/BioSpice, with ten useful worked examples.	D
B.7 Risk if work is not done	DARPA Bio-modeling & DNA-computing efforts may be generally disconnected from feasible, inspiring, & useful prototype-system.	C.7
B.8 Criteria for evaluating progress and capabilities	Ten working Proof-of-Concept models & experiments with clear engineering/utility specifications.	E, F
B.9 Cost of the proposed effort for each contract year	Summary of costs for the proposed BIO-COMP project . Base Option FY1 (8/01-7/02) $620,721 FY4 (8/04-7/05) $688,486 FY2 (8/02-7/03) $637,427 FY5 (8/05-7/06) $716,025 FY3 (8/03-7/04) $662,006	L, M

C. Technical Goals and Deliverables

(The deliverables at the milestones of section F are indicated by numerals delimited by # below).

C.1 DNA memory

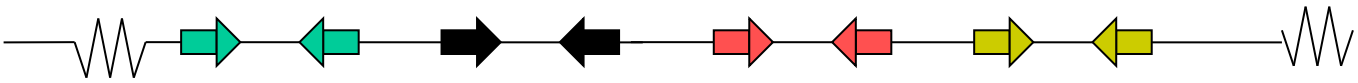
C.1.1 Recombinases

The Site-specific recombinases (SSRs) are a family of enzymes that catalyze specific rearrangements of DNA [Nash 1996]. Broadly speaking, SSRs are capable of catalyzing three such reactions: integration, excision, and inversion. The relative orientation and location of the SSR binding sites specifies the particular rearrangement that the enzyme catalyzes. We describe here a technology-in-development for exploiting the flexibility and specificity of SSRs to "record" molecular events on DNA. Potential areas of application for this idea include high-throughput cell lineage analysis in multicellular organisms, history tracking for monitoring world-wide transport of tagged biomaterials, quantitative in vivo recording of the activation of specific transcriptional programs, and a memory-I/O for DNA computing technologies.

Recently SSRs have proven useful in experiments aimed at cell lineage analysis and/or the study of tissue-specific gene function in multi-cellular organisms. Generally, specific expression of an SSR (generally Flp or Cre, two particularly well-characterized SSRs), catalyzes an excision event of the DNA intervening two SSR binding sites (frt or loxP, respectively) that are oriented in parallel. The central portion of an SSR binding site, its "spacer", is thought to be generally unimportant for SSR activity, but critical for the matching of any two sites between which the SSR will catalyze a recombination event. The enzymatic mechanism of SSR action depends on homology pairing of the spacers of two SSR binding sites. Small differences between the spacer sequences of a given pair of SSR binding sites can drastically reduce the frequency of recombination events between those sites. Thus, directed mutations can result in an SSR binding site that undergoes reactions with the wild-type binding site at a drastically reduced frequency, but the site is still functional, in that it can react efficiently with a binding site that contains an identical set of mutations. Directed changes [Sauer 1996; Schlake 1994] in the SSR binding sites could be used to "insulate" sites from one another, potentially permitting multiple non-interacting genomic manipulations with a single SSR system.

The two possible catalytic activities of an SSR on two binding sites located on a single strand of DNA (excision or inversion) can be abstracted as discreet transitions between two potential states of a single "bit". A pair of sites oriented in parallel can only experience an irreversible transition event (0 -> 1), namely, excision. Sites oriented in anti-parallel, on the other hand, can experience reversible transition events (0 <-> 1), as the intervening DNA flips from one orientation to the other in the presence of an SSR. If more than one of such "bits" (each consisting of a pair of sites), all on the same DNA construct, can be efficiently insulated from reacting with one another (as discussed above), some interesting possibilities arise. Just as with a computer, the number of potential states of such a construct would rise exponentially with the number of two-state bits available. A byte, for example, consisting of 8 bits, is capable of 256 states (2^8). A DNA construct of 30 insulated bits, each consisting of a pair of interacting SSR binding sites, would theoretically be capable of existing in 2^30 (~1 billion) different states. Linear increases in the size of the DNA construct would result in exponential gains in the number of potential states. Read-out of the "bit states" could potentially be accomplished via direct sequencing, directional PCR, or probe hybridization.

Schematic of multi-state construct with reversible bits (SSR binding sites oriented in anti-parallel):



C.1.2 SSR Bit Proof-of-Concept

We are currently pursuing proof-of-principle multi-state constructs in E. coli. We have successfully built a plasmid-based two-bit construct using two pairs of Flp binding sites oriented in anti-parallel. This construct is consequently capable of four states (2^2). State-switching is reversible and essentially random in the presence of the Flp recombinase. Differences in the switching rates of the two bits have not been extensively characterized, but have been observed to differ from one another. This is consistent with observations in the literature that sequence changes in the spacer segment of the binding site can result in variation in the rate of enzyme catalysis. The bits appear to be relatively insulated from one another, as restriction enzyme analysis of plasmids exposed to Flp does not reveal a significant quantity of excision events between elements of the different bits. Our approach still suffers from several limitations, including an insufficient degree of control over expression of the SSR (we are currently using an E. coli strain with a genomic Flp locus under control of a temperature-sensitive promoter), and apparent instability of the constructs over time. The observed instability might arise from inter-plasmid SSR reactions (as we based the construct on a multi-copy plasmid), or the fact that indirect repeats, such as those present on the multi-state construct, can prove unstable in E. coli strains that do not have enough recombination systems knocked out. We are currently planning a similar approach to our previous attempts, with the following changes: (1#1) placing the Flp enzyme under the control of more tightly regulatable promoters (1#2) generating two constructs, each with a greater number of bits (one with reversible bits and the other with irreversible bits) (1#3) cloning the multi-bit construct into a BAC (bacterial-artificial-chromosome) in a BAC-compatible strain, where we expect the construct to be present in low copy number (one or two copies) and wild-type recombination systems to be knocked out, hopefully reducing the instability observed in the previous system.

C.1.3 Applications of SSR systems

(1) The fact that a relatively low number of bits (e.g. 50) can encode a very high number of states (10^15), coupled with a mechanism of stochastic flipping at a low rate, raises the possibility of achieving high-throughput multiplexed cell lineage analysis of multicellular organisms (or monitoring world-wide transport of tagged biomaterials in general). As a transgenic organism (carrying the multi-bit construct and expressing the Flp enzyme) developed, the bit-state of any given cell lineage would change. As cell lineages diverged from one another, so would the bit-states. The bit-state profile of any given cell would be informative of the degree to which the SSR was expressed in its ancestors, as well as of its phylogenetic relationship to other cells with similar profiles. The same sorts of mathematical methods developed for the determination of evolutionary phylogeny on the basis of sequence data would be applicable to these data.

(2) Since the Flp enzyme can essentially be placed under the control of any promoter, the possibility arises of "recording" the activation of specific transcriptional programs in investigations where direct observation of a reporter gene is not a tractable option (e.g. Are transcriptional programs involving anaerobic growth activated during the in vivo infection by Pseudomonas aeruginosa?). The use of a multi-state construct using irreversible bits that undergo transitions with varying efficiencies which depend on the cellular concentration of the FLP enzyme would provide a quantitative record of the extent of activation achieved under a given transcriptional program could be recorded.

(3) Technology development aimed at using DNA to perform computations or store information would benefit from a read-out method: specific changes in DNA that will be faithfully replicated and can be assayed at any subsequent time point. The use of SSR binding

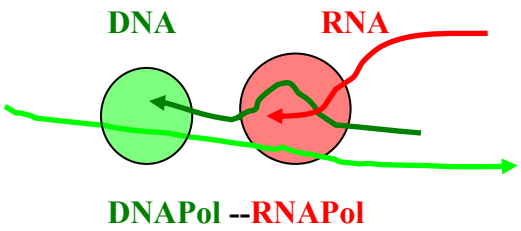
site-based “bits”, using natural or engineered SSRs, could serve as a recording method for DNA and/or cellular computing technologies. In order to have multiple independent sensors running simultaneously, we could either harvest the hundreds of different recombinase site-specificities in various microbial species, to make combinations[Shaikh 2000], or select for variants [*Bulyk 2001].

Comparison with previous DNA storage scenarios. Repressor-based toggle-switch and ring oscillator [Gardner 2000, Elowitz 2000]. Both depended on 3 fairly robust repressors (lacI, tetR, and lambda-CI). Scaling this to a circuit even remotely similar to the achievements in electronic circuits will be very hard. Since the bio"circuit" is homogeneous, the danger of cross-talk is significant. If a two-repressor toggle were considered a "bit", fairly optimistic versions would still probably max-out at 200 bits (and require 400 kbp to do so). An alternative would be to encode the information in the smallest, most accurately replicated bits in nature, the base pairs themselves.

C.1.4 DNA memory using DNA polymerases.

The ultimate goal for memory density would be to toggle sequential base pairs in compacted DNA. Ideally the system would also maintain an integrated I/O system for biochemical concentration time courses. This would consist of an in vitro system containing an error-prone polymerase. The eta-polymerase [Matsuda et al. 2000] has a remarkably high mutation rate (~10%) and high tolerance for 3' mismatch in the presence of all 4 dNTPs and probably even higher if limited to one or two dNTPs. Hence the eta-polymerase could be a "de novo" DNA synthesizer similar to terminal deoxynucleotidyl transferase (TdT) but with the advantage that the product is continuously double-stranded. This has advantages of more predictable and less sticky secondary structure than single stranded DNA, as well as a good support for other polymerases and/or DNA-binding-protein tethers. Assessment of the mutation rate with single dNTPs is covered in task 3#2, below.

To accomplish the recording the levels of one ore more of the dNTPs must be regulated by the directly or indirectly by the environmental components to be recorded. Various degradative, biosynthetic, or salvage nucleotide pathways can be controlled in vitro via a subset of the sensors described below (in section C.2.1). Examples for each dNTP degradation have been characterized -- dCTPases (e.g. phage T4), dATPases (e.g. helicases), dUTPases (e.g. Dut), dGTPases (e.g. E. coli MutT), and TTPases. [Shechter 2000; Frick, et al. 1994; Baldo 1999; Gary et al. 1998; Schultes et al. 1992]. Alternatively (or combined with) photoactivated dNTPs (C.2.2), or chemically-activated primers (C.2.1) can be used. Examples of useful recordings would be for light (see section C.2.2), stress, toxins, or glucose monitoring. The standard (initial) readout would be either conventional electrophoretic sequencing or "in situ" DNA sequencing (see C.3.1). The initial applications would use variable length (unclocked) nucleotide runs (1#4), but increasingly accurately phased syntheses may be possible by appropriate combinations of polymerases and dNTP concentrations. For example, an RNA polymerase fused to the eta polymerase could act as a stepper-positioner moving a precise number of bp based on a sequence of rNTPs. This would be milestone (1#5). Also partial overwriting of previous messages could have interesting applications.



In the above figure the etaDNAPol and RNAPol can only advance (effectively) if the next dNTP and rNTP in line are both present.

C.2 I/O: Real-time sensor input from the cell surface or in vitro

C.2.1 Small ligand sensors

We propose to use small ligands to input data into our engineered cellular and in vitro systems. Currently, there are 58 DNA-binding proteins adapted to E. coli with known ligand/inducers and generally non-cross-reacting DNA-specificity [Robison 1998]. Another set of sensors depends on termination control (His, Trp, Ile, Val, Thr, Phe, etc.). These small ligands allow programming from the outside of the cellular systems, which we engineer. Milestone 1#2 will be expanded to allow multiple regulatory molecules in vivo (2#1). Some of the above will be too challenging initially for the in vitro system (e.g. tetracycline, amino acids, nucleotides). For the in vitro system we have made progress toward establishing a complete empirical code for Zn-fingers and DNA-binding domains in general using phage display and ds-DNA microarrays [Bulyk 2001]. This could provide a general method for long-term fine-tuning of individual promoters by creating concentration gradients of the appropriate DNA binding proteins. This is considerably more cost effective than synthesis of a series of DNA constructs and allows determination of additional network parameters usable in the modeling goals of section C.5. This is milestone 2#2. Allosteric repressors have been used to regulate T7 RNA polymerase [Dubendorff 1991]. Another option would be to use small molecules to trigger allosteric ribozyme switches [Seetharaman. 2001]. The first milestone would be to insert the cAMP-triggered ribozyme in frame at the 5' end of GFP such that upon cAMP-induced cleavage the GFP mRNA becomes more (or less) translatable (2#3). Another application of these ribozymes in in section 3.1.

C.2.2 Photo- & chemo- sensor arrays

A powerful means to record complicated patterns in long DNA molecules would be to use light. Currently photocleavable phosphoramidites are spatially patterned by light projected from megapixel micro-mirror arrays from Texas Instruments [Singh-Gasson 1999] for organic synthesis. Photocleavable nucleotide triphosphates (nitro-benzyl "caged" NTPs) have been used for decades in cell physiology [McCray 1980]. We propose to combine these two approaches with the polymerase-recording concept above. The first milestone will be the incorporation of a fluorescent nucleotide triphosphate by T7 RNA polymerase under control of a caged ATP or GTP (2#4). The second milestone would be to use these to regulate the incorporation of dNTPs in the format of an RNAPol-etaDNAPol fusion. (2#5)

C.3 I/O: Output

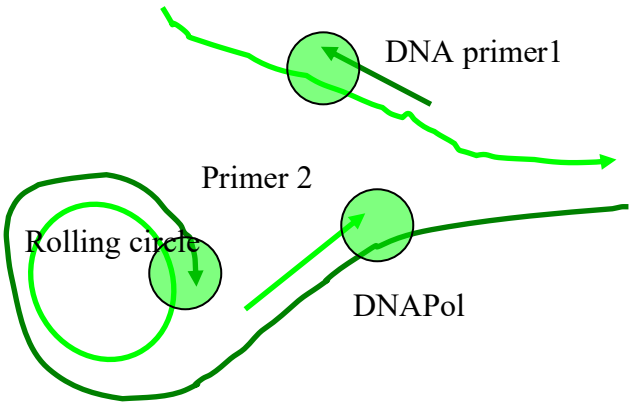
C.3.1 Polony fluorescent in situ sequencing readout

The cellular system we envision will process its small ligand inputs, and then record its computations on one or more DNA molecules. Because we envision this system to be capable of generating large amounts of data (billions of bits), we require a high throughput way of sequencing

these DNA molecules. Currently we are developing a novel technology to address this need [Mitra 1999 and section C.6]. This method is currently working adequately enough for PCR amplicons less than 100 bp [Mitra 1999 and section C.6] that we could use it to assess the patterns produced in task C.2 for spatial and nucleotide fidelity (3#1). Note that using an error-prone polymerase simply means that it will incorporate the bases that we want when and where we want them. So eta polymerase with 10% error in the presence of all 4 dNTPs could be close to 99% "error" if given only one non-template-matching dNTP. In terms of our goals this would really mean 1% error in the specified goal. We have measured the fidelity of incorporation for a variety of high-fidelity polymerases when presented with only one dNTP. We will repeat these quantitations using the eta polymerase (3#2). Chemosensor recording "black-boxes" will be tested in which arrays of allosteric self-cleavage induced by specific small ligands, e.g. cAMP [Seetharaman. 2001] will produce active RNA primers for initiating DNA synthesis (3#3).

C.3.2 Rolling circle amplification

Rolling Circle Amplification (RCA) [*Zhong 2001] represents an alternative to polony amplification (C.6) since it is continuous replication not requiring thermal cycling. With only one primer (or nick), it grows one long tail (figure below) from the original circle at a rate linear with time.



When a second primer from the opposite strand is also included highly branched structures with mass growing initially exponentially with respect to time ($m=k*\exp(t)$, or at least $m=kt^2$).

C.3.3 Interfacing sub-nm to cm scale via chemi-optics combinations.

Our modeling of the RCA process above indicates a novel way to build up layers in a 3D array as a function of time, chemicals and optical patterns outlined below. If replication begins in a uniform layer on the flat surface of a glass slide (or other surface), then the polymerization reaction can only occur in the next nm thick layer up. The strand-displacing activities of polymerases (such as etaPol or an eta-like BstPol) in RCA requires either nicks or primers to initiate strand-displacing DNA synthesis (see the figure above). If some of the RCA primers are immobilized then the hyperbranched-DNA products will be quite stable in space and time. A coarse (micron-scale, 5 Hz) pattern can be set by the megapixel micro-mirror optics, while the finer detail (nm, 250 Hz) is provided by either a free running or RNAPol-etaDNAPol-fusion stepper. It is important to note that the nano-scale recording is not necessarily "redundant" nor limited by the micron scale light patterns, since it contains time components. The thickness and therefore the recording capacity would be a trade-off with the spatiotemporal precision of specific NTP and/or

dNTP pulses used for positioning and recording respectively.

The layers deposited can include a variety of chemistries attached to (or placed by) the nucleic acids. For example we have developed redox sensitive fluorophore "side-chains" for each of the 4 dNTPs and will develop photosensitive versions (3#4). Metal binding groups and wires [Braun 1998], quantum dots [Michler 2000], quantum-wires [Emiliani 2001], magnetic dots [Cowburn 2000], or refractive dots [Yguerabide 1998], all known to display interesting properties at room temperature (in time scales as brief as femtoseconds) could be assembled by this method. One of the above will be chosen by year five as task 3#5. These explorations are especially important as they may open up ways to dynamically rebuild fast electronic-optical pathways based on signal coincidences and/or traffic levels (analogous to learning and computing in neural circuits). The naturally hyperbranched structures found in RCA seems like an elegant first step in this direction.

C.4 Minigenomes for systems models & nano-engineering tools

Many of the molecular tools for the in vivo and in vitro I/O and memory will be sub-optimal off-the-shelf, especially in a system sense. We have well-established ways to perform site-directed mutagenesis and selection on individual genes in vitro and systems somewhat in vivo. It would be desireable to be able to focus selection on a stripped down version of a living cell that operates "with the hood off" so that any component(s) can be tuned or replaced. It would also be desireable from a computational system engineering viewpoint to have a system composed only of parts whose 3D structure and function are known (which is far from true for any living cell). We propose to make a minigenome (or in vitro cell) which will have exacty such properties.

C.4.1 Minigenome synthesis:

Any effort at biosystem and "genome engineering" [*Link 1997] would benefit greatly from the ability to synthesize arbitrarily long DNA segments accurately and inexpensively. We are testing two systems in which millions of oligonucleotides at attomole scale can be made in hours for a few dollars. One uses maskless photolithography [Singh-Gasson 1999] and improved photo-phosphoramidites [Beier 2000]. The other consists of piezoelectric ink-jet deposition of conventional phosphoramidites [Hughes 2001]. In addition to permanent 3' attachment (for use as immobilized hybridization arrays) both methods can be designed for release, either all at once or in phases to act as self-primers for PCR. Protocols similar to those for DNA-shuffling [Crameri 1996] can be used to make up to 20 kilobase-sized constructs. If some mutagenesis is desired then either the natural synthetic plus PCR errors could be accepted or more purposeful combinatorial steps taken. If more accuracy is sought then automated preparative DHLPC, followed by preparative heteroduplex steps [Jones 1999] could probably bring the error rate down (from 1%) to 0.01% (4#1). Existing biological systems with a two-out-of-three repair mechanism will be sought to reduce error rates even further. These could be assembled by chewing back the 3' ends and extension ligating (a restriction-enzyme independent, ergo more general, strategy). Final constructs in the 50 to 200 kbp range could be size selected and in some cases further combined using homologous recombination established by our group [*Link 1997] or others [Swaminathan 2001]. Applications include converting the codon usage of whole proteins or pathways [Andre 1998] or facilitation of the very projects above that here create some of these basic tools. For example the eta polymerase recorder might

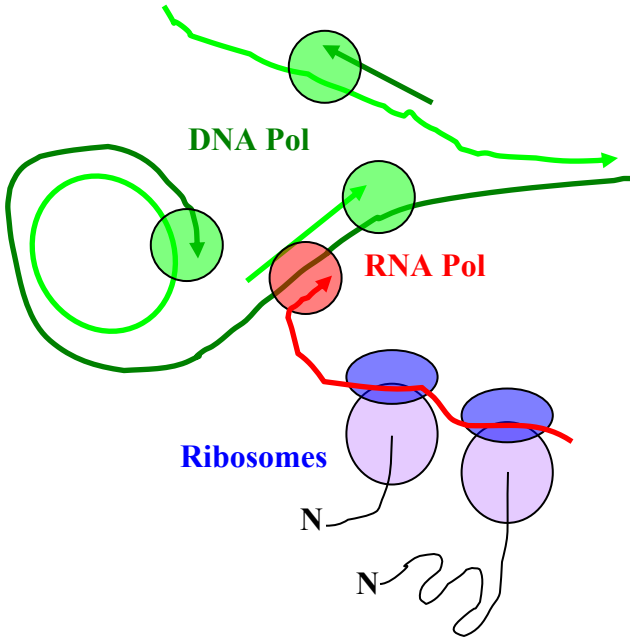
eventually be part of the toolbox. We will also design strategies for accomplishing all of the above tasks in vivo. Our high-throughput sequencing [*Mitra 1999] also would be a component of the toolbox, and facilitated by it.

C.4.2 Modeling the required gene set

The relationship between essential genes defined by evolutionary gene conservation and gene knockout experiments in a minimal genome such as Mycoplasma genitalium has been explored [Tomita 1999; Hutchison 1999;*Robison 1996] . The rules change significantly by removing the requirement for a membrane. This allows full experimental access to components that normally would by hard to supplement from the external environment of the cell.

We base our minigenome on Escherichia (and four other species) rather than on Mycoplasma since E. coli components and pathways are faster and more well characterized in vitro [e.g. http://biochem.roche.com/rts]. We anticipate that only a few of the RNA modification enzymes will be required for sustained growth. Keeping the number down will reduce the chances of unknown interactions at the cost of possibly small changes in efficiency or accuracy. There are 45 modification enzymes for the 31 modified bases in tRNA and rRNA only one of these is thought to be essential for growth (trmA) and even this one does not have a large effect on any of the known in vitro reactions. Indeed even when stripped of all modifications and all 2' hydroxyls a tDNA version of tRNA is very similar in its in vitro enzymatic aminoacylation reaction [Khan 1988] ! Reducing the number of genes further from 144 to 107 by eliminating "dispensable" tRNAs, r-proteins, etc. would probably not be as advisable. The number of tRNAs chosen could range from 20 (the minimum to cover the main 20 amino acids) to 46 (the minimum to cover the main 61 codons) to 85 to cover all tRNAs in E. coli. The advantage of 20 is that it would require less than half the number of tRNA genes and a reduction of 1.7 kbp. However the cost is high in terms of lost flexibility in genome design and initial costs to resynthesize the entire genome due to changing half of the codons in the genome over to the subset chosen. The alternative favored is to harvest by PCR as many adjacent genes as possible without change in codons initially. The cost of de novo synthesis of the smallest genome (74 kb) would be $30K. In contrast, the cost of 80 primer pairs for the alternative would be $1K even for the larger (95 kb) genomes sizes anticipated. This lower unit cost encourages the testing of multiple designs. Once a working basic design is found, the full synthesis becomes more predictable and justifiable. The current best estimate for the starting gene set is given in the table at the end of section III.

There are relatively few in vitro DNA replication/amplification protocols. Those involving thermal denaturation or reverse-transcriptase are likely to result in selection for small products and very disruptive to the function of the overall system. We choose Rolling Circle Amplification (RCA), as robust and familiar to our group [*Zhong 2001]. The overall replication of our minigenome will be set by the polymerase extension rate and the number of initiation sites. The latter are defined either by primer binding sites or by nicking sites.



The above figure adds RNA Polymerase and ribsomes and their polymer products to the branched-rolling circle figure in section C.3.2. The diagram indicates that the transcription will likely be designed to only occur on one strand based on the placement of T7 promoters.

C.4.3 Screening for functions:

Screening will be accomplished by microscopic quantitation of arrays of known variants or by isolation of RCA products from random arrays. Typically the assay would be the level of GFP fluorescence or luciferase luminescence. Physical selection can be accomplished by variations on ribosome display [Mattheakis 1994; Hanes 2000] and puromycin-mRNA display [Hammond 2001; Kurz 2000]. Here the assay would be affinity selection with Nickel columns or antibodies to the epitope fusion proteins. The initial set of screens will simply be for GFP (4#2), then RCA dependent GFP (4#3), then translational production of the polymerase required for RCA & GFP (4#4), then each of the 140 proposed components of the minigenome (typically in naturally adjacent groups of genes from the E. coli genome) to assess whether they aid or inhibit GFP production (4#5). Combinations of regulatory elements and genes will be computationally designed for experimental screening based on the above rounds (4#6).

C.4.4 Replication rate

We will explore both the nicking activity of BstNBI and a subset of the primers used in constructing the minigenome. We will explore design principles to avoid (or exploit) collision of DNA polymerases going 5' to 3' on opposite strands. With one origin per genome and the normal rate of Bst-Polymerase of 250 bp per second, the doubling time of the population of minigenomes would be 7 minutes. If initiating nicks and/or primers occur every 250 bp, then the doubling rate would be one second. These rates go from about 4-fold faster than the fastest rate in bacteria (E. coli) to 800-fold faster. Note that since replication (and competition) is exponential, even 1.01-fold differences on replication rate turn into 2-fold changes in population size after one hundred generations or so, hence changes above are truly enormous. The relevant rate in a Darwinian competition sense will depend on the rate at which the sibling genome complexes typically segregate and express. This will be limited by the protein synthetic rate. The RNA polymerase chosen (from T7) has a rate similar to the DNA polymerase (i.e. 250 nucleotides per minute). The ribosome advances at about 55 nucleotides per minute and the longest gene is 2853 bp, so the expression time is one minute plus folding/assembly time. We will assess the doubling rates varying number of primers, amount of nicking, and protein synthesis (4#7).

The replicating units form groups of related sibling molecules that we call "polonies" for polymerase-colonies. The boundaries of these need to be rigid enough to keep the polonies separate long enough to have a competitive Darwinian advantage (rather than sharing their

"guts" with the others right away), yet flexible enough to allow rapid growth. They also need to pass food and waste. The methods currently in use are (1) isolation by polymer or gels, e.g. acrylamide or agarose [*Mitra 1999] or oil-water emulsion [Tawfik 1998; Ghadessy 2001]. For the porous polymer gels, substrates (amino acids, dNTPs, rNTPs) can be exchanged in and the waste products (dNMPs, rNDPs, PPi, Pi) removed by very rapid dialysis. [biochem.roche.com/rts].

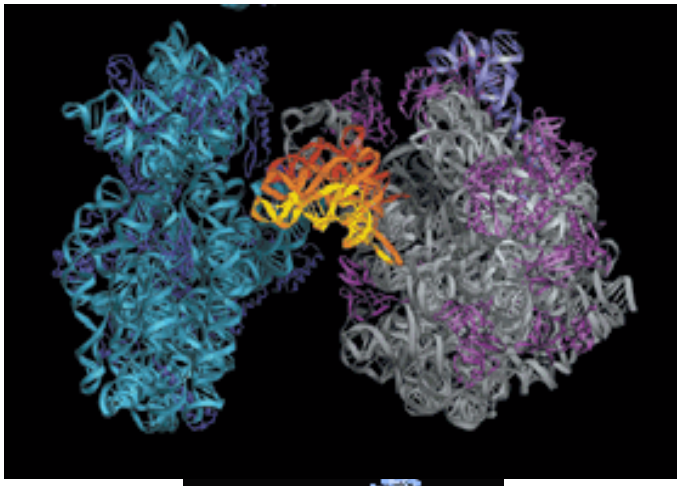
C.5 Computational models for optimization & integration

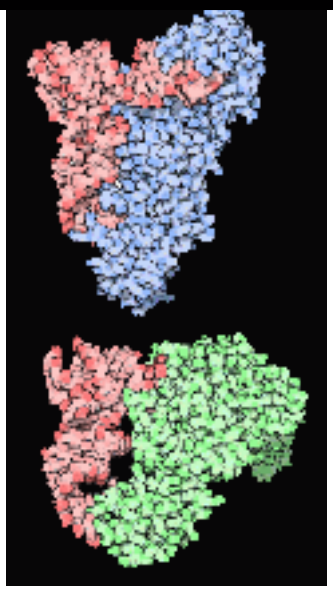
C.5.1 Mutational analyses

We expect to obtain mutations as part of random drift, subtle adaptations to the in vitro system, and specific adaptations to specific challenges (as has recently been done for tRNA-synthetases [Wang 2001] and Taq Polymerase [Ghadessy 2001]. The genome is small enough to allow resequencing. The effects on conserved nucleotides can be computed. The effects on 3D structure can be modeled. Our milestone 5#4 on this will be covered in section C.5.4. One notable advantage of this minigenome is that in contrast to other genomes where only 10% of the proteins are of known 3D structure, the 3D structure of 138 of the most essential 140 components of this proposed replicating system are known through crystallography of closely homologous species.

The three molecular models below represent all but 20 of the 140 gene products of our minigenome (or close homologs). From left top: The 30S ribosomal subunit in blue on the left, 3 tRNAs in yellow, orange, red in the middle, 50S on the right in grey [Wimberly 2000; Yusupov 2001]. The tRNA Synthetases fall into two main classes I and II. (for amino acids CEILMQRVYW, and ADFGHKNPST respectively). Top in pale blue is type I, Gln(Q), and below in pale green is type II, Thr(T).

www.rcsb.org/pdb/molecules/pdb16_2.html





The 70S & 30S ribosome structure from Thermus; [Wimberly 2000; Yusupov 2001] covers most of proteins and RNAs. The known Thermus proteins have about 44% to 53% sequence identity. The

tRNA complement in E. coli consists of 41 anticodon types (86 tRNAs total), all likely to conform to the structure of the first tRNA [Sussman 1978].

C.5.2 Overall 3D & 4D morphology.

The structure of the active replication/translation complex may be considerably less compact than a living cell and hence more accessible to microscopic inspection. A 90-kb circle is likely to have a random coil radius of about 1-micron depending on counterions. Immobilizing a flat surface of sequence-specific DNA binding proteins could force the mini genome into a planar or even linear configuration. In the latter case, the maximal distance between two points on an extended circle would be 15 microns easily visible with conventional optics (5#1). A variety of sub-50nm resolution microscopy methods (e.g. AFM, near field optics) may be applicable and greatly aided computationally by the existence of the component structures. We have a collaboration with Dale Larsen at HMS on this subject (5#2). This system should be a very attractive test-bed for microscopy technology development teams. The 3D & 4D visualization tools will likely help us debug and design by thinking at a molecular scale.

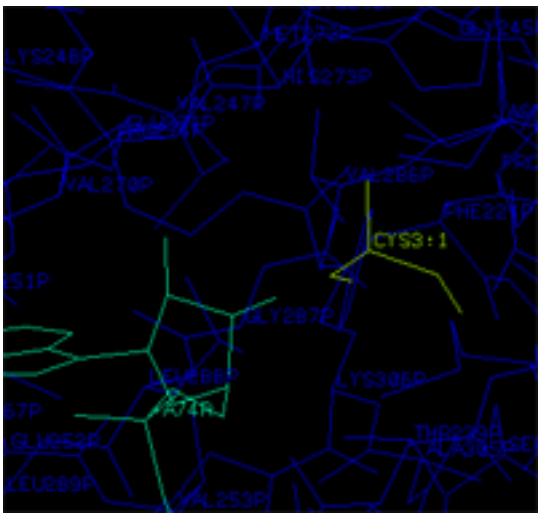
C.5.3 Variance and system modeling

A related computational goal is to apply stochastic modeling [Gibson 2000, 2000b] and/or related petri nets [Goss 1998; Matsuno 2000] to help optimize the reproducibility of the complexes as the number of each molecular species approaches a single copy of the expected stoichiometry (typically one to four per complex). In order to obtain the numerous kinetic reaction probabilities and other parameters for these and other systems models to be "over-determined", we need to exploit the openness of the coupled replication/translation system to collect the needed data. The key issues will be with the broad set of reaction time scales, in particular the fastest time scales. Assays dependent on GFP will be slow due to the lag in the assembly of the fluorophore even in the rapid assembly mutants. The incorporation of labeled nucleotides is on the order of 250 Hz to 0.1 Hz (a bit slower than the diffusion limit). As an example of an interesting way to measure attempts to overcome the stochasticity, we will analyze (sequence) several single molecules created by pulsing single eta-DNAPol and an RNAPol-etaDNAPol fusion to compare the incorporation variance (5#3). Finally population genetic modeling of the selection and drift of sequences in the minigenome replication will be developed to guide system optimization.

C.5.4 Modeling application example: Mirror-image pathways.

Many useful drugs, chemical sensors, and materials require chiral syntheses. Enzymes provide exquisite chiral-specificity as well as low-cost (low temperature/pressure protocols). Billions of years of evolution have fine-tuned these engines of synthesis, which are now being harvested in microbial genome sequencing projects. The enzymes can be mixed by design or combinatorially to create novel substances (e.g. polyketides). However the enzymes are generally all limited to the one chirality for which they originally evolved. If one could engineer in vitro protein synthesis (or a cell) to be capable of flipping from the use of L-aminoacids to D-amino acids, then one could nearly double the number of enzyme reactions available for novel applications (without designing custom selections). In addition, many of the mirror enzymes, RNAs and metabolites would be resistant to common degradative enzymes. Furthermore, the options for programming of combinations of enzymes would be 2^N times greater, where N is the number of enzymatic steps. The major protein that needs to be engineered is elongation factor Tu. Later phases would consist of engineering aminoacyl-tRNA synthases and enzymes involved in amino acid metabolism that are present in the protein synthesis cellular contents. All of these proteins are closely homologous to known 3D-structures, which will aid interpretation and design of

mutagenic strategies. It is evident that we are already fairly close since E. coli EF-Tu and aminoacyl-tRNA synthetases utilize D-amino acids at rates within 25-fold of the efficiencies of normal L-amino acids [Yamane et al 1981] and since alpha,alpha-disubstituted amino acids [Ellman et al 1992] and non-alpha-substituted glycine can be incorporated in proteins efficiently. By my computational and visual inspection, the major positions in EF-Tu that might be involved in D-alpha-aminoacid rejection are Val286 and His273 [Nissen 1999; PDB:1B23]. A variable few codons of insertions and/or deletions at these and nearby positions could be introduced into the gene in a construct allowing selection on the binding properties of the protein [Roberts, et al. 1998; Schumacher et al 1996; Mattheakis 1994; Hanes 2000; Hammond 2001; Kurz 2000]. Selection or screening for binding to a D-aminoacyl-CCA column (and not to an L-aminoacyl-CCA column nor to a non-cognate-aminoacyl-CCA column) would enrich for correct EF-Tu changes. Many of these mutant EF-Tu proteins would then be individually screened for their ability to aid D-aa incorporation into ribosomal peptide synthesis (5#4).



In the figure above, the aminoacyl-nucleotide bond (Cys to Adenosine 74) is shown adjacent to His273 and Val286 (above the Cys).

To create and engineer a full mirror replicating system would be a 4th year milestone aimed at achieving real system optimization through design plus lab selections. It would nicely build on the steps above. Computationally-aided selections similar to the above (for EF-tu) could be designed for the aminoacyl-tRNA synthetases (5#5). The latter are not absolutely necessary for the in vitro version since D-aa charged tRNAs have been made semi-synthetically [Ellman et al. 1992]. The strain providing protein synthesis enzymes can be deleted for the dadA (D-amino acid dehydrogenase) gene, if that activity is found to interfere with stability of D-amino acids and/or their esterification to tRNAs. Additional enzyme knockouts like gdhA, glyA, murL, alr, aspA, tnaA, gcvT, sdaA, deoD, deoA are likely to be required to make a cell entirely dependent on external amino acids and nucleosides and incapable of laundering them to the normal chiral form. To create the genome itself, design of polymerases that accept L-NTPs given poly-D-dNMP templates [Semizarov, et al. 1997]. Alternatively mirror DNA for transforming the cells could be made using L-2'-deoxyribonucleoside 3'-phosphoramidites [Urata, et al. 1992] in the genome synthesis approach described above.

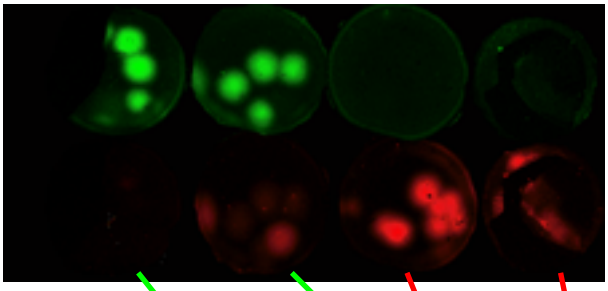
C.6 Demonstrated achievements of our group in this area

C.6.1 Recombination

Our group has developed and tested the first multi-bit SSR based DNA memory [see section C.1]. We coined the term "genome engineering" and have one of the most widely used homology-directed recombination methods [Link 1997]

C.6.2 Polymerase arrays

We have pioneered oligonucleotide synthesis for very high-density genomic microarrays for E. coli [*Selinger 2000] by photolithography as well as piezoelectric deposition of oligonucleotides (relevant to project C.4) [Bulyk 2001]. We have developed a new femtoliter-scale DNA sequencing method and two methods for DNA polymerase action on microarrays [*Bulyk 2000; *Mitra 1999]. These also demonstrate the ability to step along a DNA molecule for many steps as illustrated below (relevant to goals C.3 & C.4)

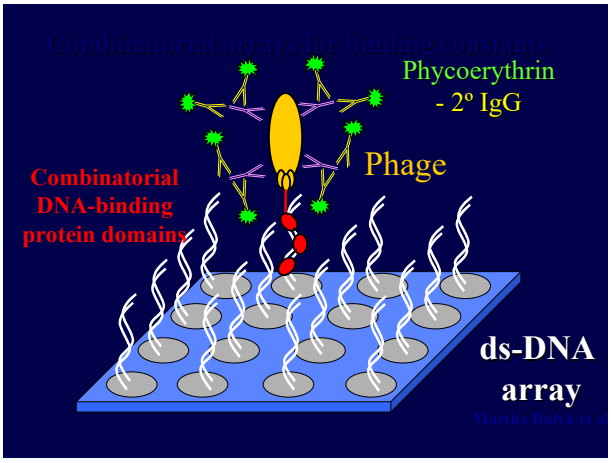


5'...ATAACAATTTCACACAGGAAACAGCTATGACCAT...3'
3'...TATTGTTAAAGTGTGTCCTTTGTCGATACTGGTA...5'

Template strand below, extended primer above Green FITC labeled dCTP and Red Cy5 labeled dTTP were incorporated to check on 4 of the 26 cycles of nucleotide addition, 34 nucleotides in synchrony starting from individual DNA molecules at the center of each circular polony.

C.6.3 DNA & protein arrays for binding constants

We have developed and exploited high throughput methods to screen large numbers of sequence-specific DNA binding proteins and all possible combinations of target DNA sequences. This gives a large number of binding constants useful for accurate simulations in objective C.5.



An example of this is shown above, wherein the combinatorial complexity of proteins can be manipulated in an attached phage which is conceptually similar to the mini-replicons described in section C.4. The double-stranded DNA array captures the protein-phage-fusions and after appropriate staining with fluorescent antibody complexes, the image of the complexes can be digitized by 5-micron resolution scanners.

C.6.4 Computational Biology

Dr. Church's first computational efforts beginning in 1977 are still very relevant especially to the 3D modeling in tasks 5.1-2. In particular, the application was to the first high resolution tRNA

structure and the development of tools for rigorously comparing the differences between model predictions and observed data.

We have been involved in various aspects of computational genomics and the modeling of gene regulatory networks too numerous to recite here. (See section G and attached resumes).

We have on our web one of the few dynamic models of a whole cell (the human red blood cell), where the model works and refers to nearly complete data for the relevant reactions.

C.6.5 Other relevant experience

We helped to establish the Human Genome Project and three of the Genome Centers. We have experience with meeting production milestones (even while changing the underlying technology) and with the special problems associated with high-risk/high-payoff technology development projects. We have worked on a similar DARPA project before (supervised by Sonny Maynard). We have a good track record for technology transfer to academic, government and commercial sectors (see section G below).

C.7 Risk if work is not done

DARPA Bio-modeling & DNA-computing efforts need a shareable feasible, inspiring, & useful prototype-system. Otherwise the program will consist of fragmented or impractical models and experimental systems.

Furthermore, the post-genome sequencing world needs arrays to rapidly and accurately read out proteins and metabolites. The in vitro optimization approach will greatly facilitate this goal. The system is naturally resistant to many useful detergents and surfactants, since one of the most labile parts of normal living systems (the lipid membrane) has been deleted. The in vitro mini-replicon opens up the possibility of selecting for variants, which are especially robust in practical environments, while retaining metabolic activity (unlike spores). For example a Taq polymerase resistant to the common anti-coagulant, heparin has been selected [Ghadessy 2001]. Alternative base-pairs [Ohtsuki 2001] and genetic codes [Wang 2001] can be incorporated to expand the options for nanofabrication.

D. Goals & Deliverables

The deliverables for each milestone in Section F are indicated by numerals separated by # in the text of section C. All technical data and computer software will be furnished to the US Government with unlimited rights (DFARS 227). Harvard use of disclaimer and limitations for commercial use would be applied only to the extent that it is consistent with the DARPA policies and the above unlimited rights. We will provide quarterly web updates on our Harvard web site and, as required, mirrored to other DARPA servers.

The 27 deliverables for 5 sub-tasks in outline are: Feasibility of scale-up of the SSR system from two bits to 50 bits as well as an estimate of the theoretical upper-limit will be established (1#1-3). The SSR Flp will also be used in minigenome task 4#3. An eta polymerase construct will be tested in year two with DNA sequencing as the output with an initial goal of 2 kilobits of storage, 1#4-5. The sensor inputs will be studied and integrated into the SSR or polymerase system (2#). The construction of three-dimensional arrays of DNAs and polymerases will initially build on our "polony" technology [*Mitra 1999] (3#1-3). We will bridge of a one-cm pattern with nm precision using a chemi-optical combination. Controlled polymerase stepper-positioners will be used to incorporate branching oligonucleotides at nm

scale. Incorporation of electronic or optical computing components will be attempted (3#4-5). For the minigenome project, important demonstrations will be the RCA dependent expression of GFP, translation itself and measures of replication rate (4#1-7). The computational deliverables will be coordinated with other DARPA BIO-COMP in particular initially the BBN consortium. The known 3D models will be adapted to include any sequence changes and software for inclusion of microscopic morphology, dynamics, and stochastics will be developed (5#1-5). Metrics for quality assessment and outlier identification will be developed and refined.

E. Statement of Work (SOW)

Quarterly updates will be placed on our designated DARPA-BIO-COMP pages. Major updates twice a year. We have a demo web site dedicated to this project as of May 2, 2001. http://arep.med.harvard.edu/darpabiocomp/ (for now limited to DARPA review and HMS internal use only).

Dr. George Church will define and assess statistical tests of the data quality and goodness-of-fit of the model and data, as well as manage the progress toward the milestones detailed in sections D & F. He will assist in high-level debugging of surprises arising in the BioSystems modeling and experimental efforts.

Dr. John Aach will develop the time course and structural genomics computational analyses.

Mr. Matthew Wright will develop genome engineering design tools and chemical kinetic modeling.

Dr. Jingdong Tian will implement an array-based system for coupled replication-transcription-translation. He has extensive experience with in vitro translation and with GFP constructs. He will be assisted by XiaoHua Huang who is codeveloper of rolling-circle amplification methods [*Zhong 2001]

Mr. Jay Shendure will develop higher multiplicity flip-flop memories to allow tests of the D/A input prior to development of the polymerase-based models.

Ms. Allegra Petti will construct and test gene expression system models against the data collected.

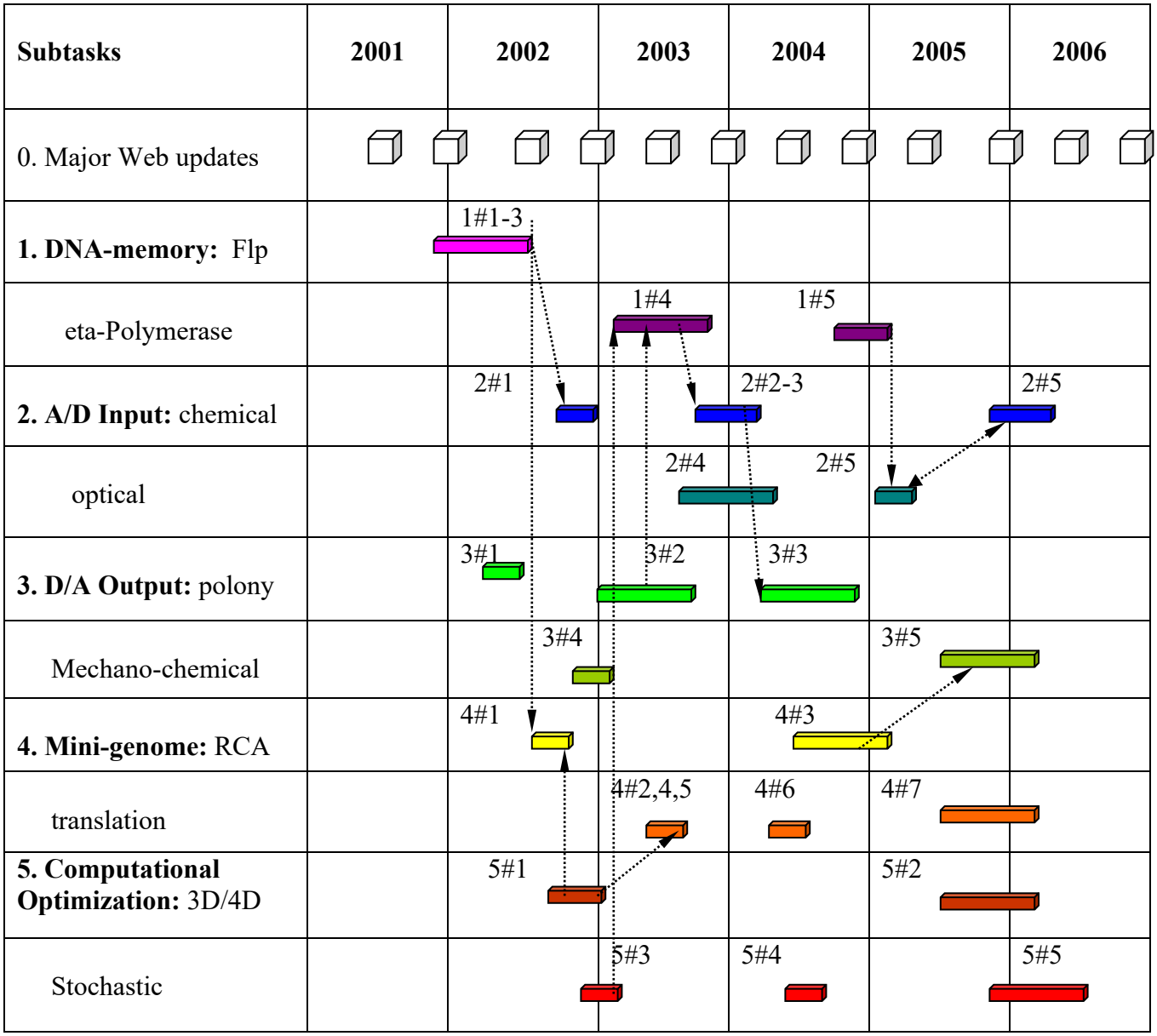
Dr. Rob Mitra will continue to develop advanced replication-array methods.

Dr. Martha Bulyk will determine the binding kinetic constants for a variety of DNA-protein interactions crucial for the D/A input and the mini-genome subprojects.

Dr. Daniel Segre will develop and use variations on our previous flux balance optimization (linear programming) methods to make them applicable to the reactions need for the memory and nano-fabrication projects.

F. Milestones & Schedule

Our development plan consists of 5 subtasks, 27 proof-of-concept demonstrations, phased into 20 quarters and 10 major web updates. The increasingly robust experiments and applicability to the overall program concept are indicated below. This is one possible scenario by which the development will proceed. Clearly, in this rapidly evolving field, other paths will be taken as the new technologies appear (both from our group and others) to make better choices available.



G. Technology Transfer.

We have demonstrated technology transfer of various open-source software from our web/ftp site since 1991 (and by tape since 1977).

(1) Our first software package (for X-ray diffraction model least-squares refinement [*Sussman 1977]) remained in international use for 16 years.
(2) Our more recent AlignACE package (for DNA-motif discovery and searches) is an even more popular software export.
(3) We have transferred a "genome-engineering" system [Link 1997] to over 600 laboratories (about 10% commercial).
(4) Our technology contributed to the first commercial genome sequence (the human pathogen, H. pylori).
(5) Our first patent (on molecular DNA tags & multiplexing) is still actively licensed after 13 years (most recently to Lynx in 2000).
(6) Our patent on single molecule DNA conductance effects is the basis of an Agilent project.
(7) Our single molecule amplification has been licensed to Mosaic Technologies Inc.
(8) Our fluorescent nucleotide tags have been licensed to Pyrosequencing Inc.
(9) The technology developed by this proposed DARPA research will be marketed using standard policy (as were the above) by the Harvard Medical School Office of Technology Licensing (HMS-OTL). At least one company (EnGeneOS, Cambridge, MA) has expressed an interest already.

H. Comparison with other ongoing research

The approach of using the memory components based on DNA as components of a high-density exhaustive records (analogous to the aircraft black-boxes) and/or for manufacturing other memory and CPU components addresses the key disadvantages of other DNA memory schemes, which are slow speed of DNA enzymes and thermal fragility of network memory. However it creates new challenges of optimizing the nano-fabrication components harvested from microbial genome projects.

The proposed effort requires a deep integrated model of the core components of living systems. This is both an advantage and a disadvantage -- an advantage since it installs a level of accountability for the overarching DARPA BIO-COMP Program; a disadvantage in requiring an disciplined set of very small incremental challenges to allow dissection of the progress and debug interaction failures.

For the mini-genome component, this could be compared with the E-cell project [Tomita 1999], which proposes to simulate a cell with 127 genes. Our system advantages include: (1) working in vitro components rather than a purely “in silico” concepts, (2) derived from rapidly growing and well-understood E. coli rather than M. genitalium, (3) By omitting membranes we eliminate the most uncertain aspects of current E-cell models, i.e. cell division components. [Tomita 1999; Hutchison 1999]. (4) The e-cell does not specify which tRNAs and what impact that would have on codon choices, hence fabrication costs and future flexibility.

Other cell system models represent only a fraction of the full replicating entity and hence prone to major inaccuracies of omitted interactions.

I. Key Personnel

Table I.1 Key Personnel.

Name	Level of Effort to be Expended					Other Major Sources of Support (Current and Proposed)
	Yr 1	Yr 2	Yr 3	Yr 4	Yr 5	
Dr. George Church * Harvard Med. School	30%	30%	30%	30%	30%	Current: DoE, NHLBI-PGA, NSF Proposed: Bio-Comp & NIH
Dr. John Aach * Harvard Med. School	60%	60%	60%	60%	60%	Current: DoE, NHLBI
Mr. Matthew Wright MIT-Chemistry	30%	30%	30%	30%	30%	Current: NSF, NHLBI
Dr. Jingdong Tian Harvard Med. School	100%	100%	100%	100%	100%	Current: LSRF fellowship
Mr. Jay Shendure Harvard Med. School	30%	30%	30%	30%	30%	Current: NSF
Ms. Allegra Petti Harvard Med. School	80%	80%	80%	80%	80%	Current: NSF
Dr. Rob Mitra Harvard Med. School	50%	50%	50%	50%	50%	Current: DoE
Dr. Martha Bulyk Harvard Med. School	50%	50%	50%	50%	50%	Current: DoE
Dr. Daniel Segre Harvard Med. School	80%	80%	80%	80%	80%	Current: NSF
Postdoctoral (TBN)	100%	100%	100%	100%	100%	
* Drs. Church and Aach are also submitting BAA001-26 proposals also in BIO-COMP Topics 3 & 4 to fully support the initiative. The LOEs above are for their involvement in the program as a whole and will be allocated to the effort proposed herein and in their other proposals as appropriate. Graduate students and postdoctoral fellows will be replaced as the move along on their career paths.						

J. Description of facilities

The Harvard/MIT/BU intellectual environment is excellent for multidisciplinary, collaborative and functional genomics research. The Church Laboratory provides some of the glue with students from all three universities and a location in three adjacent buildings at the heart of the HMS campus: 1) the Alpert Building home to the Genetics department, 2) the Seeley-Mudd Building is home to the Harvard Institute of Proteomics (HIP), the Harvard Institute of Chemistry and Cell Biology (ICCB), and the Lipper Center for Computational Genetics, 3) the Thorn Building of the BWH-Hospital Genomics & Bioinformatics Center. Harvard has recently made considerable endowment commitments to

the above and the University-wide Center for Genomics and Research. We work closely with our departmental Biopolymers facility, which has a staff of five, departmental computer facility with a staff of four. We have direct computer network and CAD links to the HMS machine shop, which coordinates with several other university and commercial machining and design facilities. Other resources include:

2 Affymetrix Chip Scanners (HP & MD)
1 Microarrayer prototype (Anorad stages), 150 slide capacity, 16 piezoheads (GeSim)
1 microarray scanner (General Scanning 5000)
1 Automated DNA and protein sequencers, synthesizers and related items (ABI 3700, 377, 373S, 391, 1000S, 394, 380B, 270A, 477A, 430A, 420A, 130A)
1 FPLC and Phast systems (Pharmacia)
1 LCQ HPLC-MSn Ion Trap mass spectrometer (Finnegan)
1 Storm Fluorimager with 29 exposure plates (Molecular Dynamics)
Numerous PCR machines with 96, 384-well, and slide heads (MJR)
1 Microfluidics development platform (Caliper)
5 -20 C freezers and two -80 freezers
7 low-speed centrifuges and ultra-centrifuges (IEC, Sorvall, Beckmann)
2 Oscilloscopes and 2 electrophysiological amplifiers 70 femtoamp rms (Axon)
1 micropipette puller and microforge (Narishige)
5 high voltage (500V to 6000V) power supplies (Biorad and EC)
2 Ultra-thin gel Direct Transfer Electrophoresis (HMS shop, Cykal)
1 96-pin array oligonucleotide synthesizer Primer Station 960 (IAS & HMS)
3 electrotransfer devices (Polytech)
1 pulsed-field CHEF electrophoresis (Genplex)
1 UV crosslinker (HMS shop)
1 Laser-induced fluorescent 4-color capillary electrophoresis (ABI 310)
2 DEC alpha fileservers running Ultrix
2 dual Intel PII, RAID level 1 & 5 based Linux fileservers
15 computers running under WinNT, 10 Linux, 6 Linux&NT, 5 MacOS
1 Silicon Graphics Octane computer
2 Linux Clusters (Beowulf-type) for parallel & associative processing (e.g. 17 Dual-processor Pentium-III 933MHz machines).
1 Terabyte tape jukebox server running Arkeia
1 Confocal Microscope (Biorad)
1 Automatic film processor
1 Bioflo 3000 mammalian and microbial cell culture chemostat (New Brunswick)
1 EPICS ALTRA flow sorter with Autoclone multiwell plates option (Beckman-Coulter)

K. Experimentation and Integration Plans.

Our results will be integrated with the Biomodeling solutions that the BBN DARPA contractors are currently developing. The main point of this proposal is to develop applications of the BioSystems modeling software sufficiently compellingly useful and simple that many of the BIO-COMP teams and related efforts will use our examples as a focus for
Our experience and willingness to work with other contractors in order to develop joint experiments in a common test-bed environment is evident from our web site and many successful past technology transfer examples (Section G above) involving support of over 600 groups worldwide. We expect to participate in teams and workshops to provide specific technical background information to DARPA, attend semi-annual Principal Investigator (PI) meetings, and participate in other coordination meetings via teleconference or Video Teleconference (VTC). As evidence of the latter, our course on Computational Biology has been routinely used for distance learning, integrating streaming internet video and PowerPoint slides, software, and other interactive tools specifically on the topics of this BIO-COMP BAA [www.courses.fas.harvard.edu/~bphys101]. Our budget (Section L below) requests support for these various group experimentation efforts in the form of extra supplies and personnel time for shipping DNA memory and mini-genome prototypes.

L. Cost Breakdown by Category (single task/contract)

Categories	**FY1**	**FY2**	**FY3**	**FY4**	**FY5**	**$K**
Equipment	0	0	0	0	0	Lipper Foundation cost sharing
Subcontract	0	0	0	0	0	Collaboration with BBN
Travel	0	0	0	0	0	Lipper Foundation cost sharing
Phone, fax, copying	0	0	0	0	0	Lipper Foundation cost sharing
Personnel	213	221	229	239	248	Sharing with BBN Integration
Fringe	32	33	34	35	36	
Supplies	116	121	126	131	136	
Indirect costs	259	262	272	283	294	
Totals	620	637	662	688	718	

The Lipper Foundation funding has provided the basic infrastructure for the computational biology subtasks. HHMI and DoE have helped provide instrumentation for the experimental validations. Many of the parts will be upgraded and/or replaced during the course of the proposed research. Additional economies are expected through our collaboration with the BBN/MIT/BU/HMS BIO-COMP Bio-Spice Integration Project, as the latter will provide support for computing environment supportable through a broad community.

M. Contractors requiring the purchase of information technology resources (ITR) as Government Furnished Equipment (GFE)

May 1, 2001

Defense Advance Research Projects Agency
Information Technology Office
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Attention: Dr. Sri Kumar
Subject: Information Technology Resources, BAA 01-26, BBN Proposal P01-BBN-140

Dear Dr. Kumar,

The requirements of the BBA request a letter of justification for Information Technology Resources (ITR) purchased as GFE under our proposal. This letter responds to that requirement.

We expect that the proposed work will require the purchase of ITR to perform the proposed work. It is estimated that the total complement of equipment components will exceed $20,000 per year based on current spending patterns of the individuals proposed for the project. The components will be purchased for and become dedicated resources of this HMS/MIT/DARPA BIO-COMP Project. There will be some cost sharing due to private foundation support for some of our current infrastructure. Nevertheless the software and parts upgrade and replacement (4 year amortization) costs for the proposed project described will be considerable and need to be covered by the budget for this project.

The ITR parts have been priced based on past comparisons of price and quality maximally exploiting Harvard and academic discounts and site-licenses. The ITR items will consist of whatever basic consumer-grade Intel CPUs is cost-effective at the time of replacement, plus disk drives, RAM, operating systems and other software. The rate of change is significant enough that we anticipate significant (>1.5-fold) improvements in price/performance between now and when our first purchases are made. The 900 MHz CPU/RAM/motherboard sets will cost $3000 each; disk drives >16 Gbytes, $300. The remainder of the “Supplies” budget lines (e.g. $116,631 in FY1) consists of lab consumables (oligonucleotides, fluorescent nucleotides, microarrays, polymerases, etc.) also purchased in accordance with the above criteria.

Sincerely,

George M. Church

Section III. Additional Information

References **(* = Church lab contribution)**

Andre S, Seed B, Eberle J, Schraut W, Bultmann A, Haas J. Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. J Virol. 1998 Feb;72(2):1497-503.

Baldo AM, McClure MA. Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. J Virol. 1999 Sep;73(9):7710-21.

Beier M, Hoheisel JD. Versatile derivatisation of solid support media for covalent bonding on DNA-microchips. Nucleic Acids Res. 1999 May 1;27(9):1970-7.

Beier M, Hoheisel JD. Production by quantitative photolithographic synthesis of individually quality checked DNA microarrays. Nucleic Acids Res. 2000 Feb 15;28(4):E11.

Bethke BD, Sauer B. Rapid generation of isogenic mammalian cell lines expressing recombinant transgenes by use of Cre recombinase. Methods Mol Biol. 2000;133:75-84.

Braun E, Eichen Y, Sivan U, Ben-Yoseph G. DNA-templated assembly and electrode attachment of a conducting silver wire. Nature. 1998 Feb 19;391(6669):775-8.

Buchholz F, Angrand PO, Stewart AF. A simple assay to determine the functionality of Cre or FLP recombination targets in genomic manipulation constructs. Nucleic Acids Res. 1996 Aug 1;24(15):3118-9.

***Bulyk**, M.L., Y. Choo, X. Huang, and G.M. Church. "Exploring DNA binding specificities of zinc fingers with DNA microarrays," *Proc. Nat. Acad. Sci USA 98*: in press. 2001.

***Bulyk,** M.L., Gentalen, E. Lockhart, D.J., Church, G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. Nature Biotechnology 17: 573-7.

Cowburn RP, Welland ME Room temperature magnetic quantum cellular automata Science. 2000 Feb 25;287(5457):1466-8.

Crameri A, Whitehorn EA, Tate E, Stemmer WP. Nat Biotechnol 1996 Mar;14(3):315-9. Improved green fluorescent protein by molecular evolution using DNA shuffling.

Dubendorff JW, Studier FW. J Mol Biol 1991 May 5;219(1):45-59 Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor.

***Edwards** JS, Ibarra RU, Palsson BO. Nat Biotechnol 2001 Feb;19(2):125-30 In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.

Ellman JA, Mendel D, Schultz PG Science 1992 Jan 10;255(5041):197-200 Site-specific incorporation of novel backbone structures into proteins.

Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. Nature. 2000 Jan 20;403(6767):335-8.

Emiliani V, Guenther T, Lienau C, Notzel R, Ploog KH. J Microsc 2001 Apr;202(Pt 1):229-40 Femtosecond near-field spectroscopy: carrier relaxation and transport in single quantum wires.

Frick DN, et al. J Biol Chem 1994 Jan 21;269(3):1794-803. Dual divalent cation requirement of the MutT dGTPase. Kinetic and magnetic resonance studies of the metal and substrate complexes.

Gary TP, Colowick NE, Mosig G. A species barrier between bacteriophages T2 and T4: exclusion, join-copy and join-cut-copy

recombination and mutagenesis in the dCTPase genes. Genetics. 1998 Apr;148(4):1461-73.

Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in Escherichia coli. Nature. 2000 Jan 20;403(6767):339-42.

Ghadessy FJ, Ong JL, Holliger P. Directed evolution of polymerase function by compartmentalized self-replication. Proc Natl Acad Sci U S A. 2001 Apr 10;98(8):4552-7.

Gibson M & Bruck J Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels Michael A. J. Phys.Chem. A 104 (9) 1876-1889 Mar 2000. www.paradise.caltech.edu/~gibson/papers.html

Gibson and Eric Mjolsness, Modeling the Activity of Single Genes, in Computational Methods in Molecular Biology: From Genotype to Phenotype, Bolouri and Bower, eds. 2000b.

Goss PJ, & Peccoud J.Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets.Proc Natl Acad Sci U S A. 1998 Jun 9;95(12):6750-5.

Hammond PW, Alpin J, Rise CE, Wright MC, Kreider BL. In vitro selection and characterization of Bcl-XL binding proteins from a mix of tissue-specific mRNA display libraries. J Biol Chem. 2001 Mar 30

Hanes J, Schaffitzel C, Knappik A, Pluckthun A. Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. Nat Biotechnol. 2000 Dec;18(12):1287-92.

Hughes TR, et al Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol. 2001 Apr;19(4):342-7.

Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO,Venter JC. Science 1999 Dec 10;286(5447):2165-9. Global transposon mutagenesis and a minimal Mycoplasma genome.

Jones AC, et al. Clin Chem 1999 Aug;45(8 Pt 1):1133-40 Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis.

Khan AS, Roe BA. Science 1988 Jul 1;241(4861):74-9 Aminoacylation of synthetic DNAs corresponding to Escherichia coli phenylalanine and lysine tRNAs.

Kurz M, Gu K, Lohse PA. Psoralen photo-crosslinked mRNA-puromycin conjugates: a novel template for the rapid and facile preparation of mRNA-protein fusions. Nucleic Acids Res. 2000 Sep 15;28(18):E83.

***Link** AJ, Phillips D, Church GM. Methods for generating precise deletions and insertions in the genome of wild-type Escherichia coli: application to open reading frame characterization. J Bacteriol. 1997 Oct;179(20):6228-37.

***Link**, A.J., Robison, K. and Church, G.M. (1997b) Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli. Electrophoresis 18:1259-1313

Lowe T & Eddy S, "A Genomic tRNA Database", 2001 http://rna.wustl.edu/GtRDB/Eco/

Matsuno H, Doi A, Nagasaki M, Miyano S. Hybrid Petri net representation of gene regulatory network. Pac Symp Biocomput. 2000;:341-52.

Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. Low fidelity DNA synthesis by human DNA polymerase-eta. Nature. 2000 Apr 27;404(6781):1011-3.

Mattheakis LC, Bhatt RR, Dower WJ. An in vitro polysome display system for identifying ligands from very large peptide libraries. Proc Natl Acad Sci U S A. 1994 Sep 13;91(19):9022-6.

McCray JA, Herbette L, Kihara T, Trentham DR. Proc Natl Acad Sci U S A 1980 Dec;77(12):7237-41 A new approach to time-resolved studies of ATP-requiring biological systems; laser flash photolysis of caged ATP.

Michler P, Imamoglu A, Mason MD, Carson PJ, Strouse GF, Buratto SK. Quantum correlation among photons from a single quantum dot at room temperature Nature. 2000 Aug 31;406(6799):968-70.

***Mitra,** R. and Church, G.M. (1999) In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res. 27(24):e34; pp.1-6.

Nash, HA. Site-specific recombination: integration, excision, resolution, and inversion of defined DNA segments. p. 2363-2376. In F. C. Neidhardt (ed.), Escherichia coli and Salmonella typhimurium, vol. 1. ASM, Washington, DC. (1996).

Nissen P, Thirup S, Kjeldgaard M, Nyborg J Structure Fold Des 1999 Feb 15;7(2):143-56 The crystal structure of Cys-tRNACys-EF-Tu-GDPNP reveals general and specific features in the ternary complex and in tRNA.

Nordstrom T, Ronaghi M, Forsberg L, de Faire U, Morgenstern R, Nyren P. Biotechnol Appl Biochem 2000 Apr;31 (Pt 2):107-12 Direct analysis of single-nucleotide polymorphism on double-stranded DNA by pyrosequencing.

Ohtsuki, T, et al. Unnatural base pairs for specific transcription (2001) PNAS 98: 4922-4925

Roberts RW, Szostak JW Proc Natl Acad Sci U S A 1997 Nov 11;94(23):12297-302 RNA-peptide fusions for the in vitro selection of peptides and proteins.

***Robison,** K., McGuire, A. M., Church, G. M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome. Journal of Molecular Biology (1998) 284, 241-254.

***Robison,** K., Gilbert, W., Church, G.M. (1996) More Haemophilus and Mycoplasma genes. Science 271: 1302-1304.

Sauer B. Multiplex Cre/lox recombination permits selective site-specific DNA targeting to both a natural and an engineered site in the yeast genome. Nucleic Acids Res. 1996 Dec 1;24(23):4608-13.

Schultes BC, Fischbach E, Dahlmann N. Purification and characterization of two different thymidine-5'-triphosphosphate-hydrolysing enzymes in human serum. Biol Chem Hoppe Seyler. 1992 May;373(5):237-47.

Schlake T, Bode J. Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci. Biochemistry. 1994 Nov 1;33(43):12746-51.

***Selinger**, D., Cheung, K., Mei, R., Johanson, E.M., Richmond, C., Blattner, F.R., Lockhart, D., and Church, G.M. (2000) RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nature Biotechnology 18, 1262-7.

Shaikh AC, Sadowski PD. J Mol Biol 2000 Sep 8;302(1):27-48 Chimeras of the Flp and Cre recombinases: tests of the mode of cleavage by Flp and Cre.

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Hutchison, C. Bioinformatics , 15, Number 1, 72-84 (1999) E-CELL: Software environment for whole cell simulation. www.e-cell.org/

Schumacher TN, Mayr LM, Minor DL Jr, Milhollen MA, Burgess MW, Kim PS. Identification of D-peptide ligands through mirror-image phage display. Science. 1996 Mar 29;271(5257):1854-7.

Seetharaman S, Zivarts M, Sudarsan N, Breaker RR. Immobilized RNA switches for the analysis of complex chemical and biological mixtures. Nat Biotechnol. 2001 Apr;19(4):336-41.

Semizarov DG, et al. Stereoisomers of deoxynucleoside 5'-triphosphates as substrates for template-dependent and -independent DNA polymerases. J Biol Chem. 1997 Apr 4;272(14):9556-60.

Shechter DF, Ying CY, Gautier J. J Biol Chem 2000 May 19;275(20):15049-59. The intrinsic DNA helicase activity of Methanobacterium thermoautotrophicum delta H minichromosome maintenance protein. " preferentially uses dATP and DNA-dependent dATPase"

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol. 1999 Oct;17(10):974-8.

***Sussman**, J.L., Holbrook, S.R., Warrant, R.W., Church, G.M., and Kim, S.-H. (1978) Crystal structure of yeast phenylalanine transfer RNA I. Crystallographic refinement. J. Mol. Biol. 123: 607-630. PDB ID: 6TNA

***Sussman,** J.L., Holbrook, S.R., Church, G.M., and Kim, S.-H. (1977) A structure factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. Acta Cryst. A33: 800-804.

Swaminathan S, Ellis HM, Waters LS, Yu D, Lee EC, Court DL, Sharan SK. Rapid engineering of bacterial artificial chromosomes using oligonucleotides. Genesis. 2001 Jan;29(1):14-21.

Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. Nat Biotechnol. 1998 Jul;16(7):652-6.

Umlauf SW, Cox MM. The functional significance of DNA sequence structure in a site-specific genetic recombination reaction. EMBO J. 1988 Jun;7(6):1845-52.

Urata H, et al. Synthesis and properties of mirror-image DNA. Nucleic Acids Res. 1992 Jul 11;20(13):3325-32.

Wang L, Brock A, Herberich B, Schultz PG. Expanding the Genetic Code of Escherichia coli. Science. 2001 Apr 20;292(5516):498-500.

Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. Nature 2000 Sep 21;407(6802):327-39 Structure of the 30S ribosomal subunit.

Wu P, Nossal N, Benkovic SJ Biochemistry 1998 Oct 20;37(42):14748-55. Kinetic characterization of a bacteriophage T4 antimutator DNA polymerase. "Antimutator enzymes exhibit a higher DNA replication fidelity than the wild-type enzyme, at the cost of increased nucleotide turnover."

Yamane T, Miller DL, Hopfield JJ Biochemistry 1981 Dec 8;20(25):7059-64 Discrimination between D- and L-tyrosyl transfer ribonucleic acids in peptide chain elongation.

Yguerabide J, Yguerabide EE. Light-scattering submicroscopic particles as highly fluorescent analogs and their use as tracer labels in clinical and biological applications. Anal Biochem. 1998 Sep 10;262(2):157-76.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF.Science 2001 Mar 29; Crystal Structure of the Ribosome at 5.5 A Resolution.

***Zhong** Xb, Lizardi PM, Huang Xh, Bray-Ward PL, Ward DC. Visualization of oligonucleotide probes and point mutations in interphase nuclei and DNA fibers using rolling circle DNA amplification. Proc Natl Acad Sci U S A. 2001 Mar 27;98(7):3940-5.

SP (3D % seq.ident.)	Stochi metry	Mge#	Length in bp Big-mini	Tiny-Mini	Access# B#	Gene Name	Left end position	Right end in bp	Orie nta-tion	Sequence 5' end
Total	144	107	89,498	74,310	2853					
16S	1	y	1418	1418	3968	rrsB	4164238	4165779	>	aaattgaag
23S	1	y	2903	2903	3970	rrlB	4166220	4169123	>	ggttaagcg
5S	1		120	120	3971	rrfB	4169216	4169335	>	tgcctggcg
10sb (RNaseP)			375	375	3123	rnpB eg.	3268233	3267857	<	gaagctgac
TRNAs	20-46	y	3136	1364	3939	gltT	4165951	4166026	>	gtccccttc
Cca (no)		?	1236		3056	cca	3199532	3200770	>	gtgaagatt
TrmA (22)		?	1098		3965	trmA	4159749	4160849	<	atgaccccc
BstNBI (no)			1815		AF329098	bstNBI	1	1815	>	atggctaaa
Tri1		?			AP001918	tral	92673	97943	>	atgatgagt
Flp		no	1272		NC_00139 8	flp	5573	523	>	atgccacaa
GFP		no	717		AF302837	gfp	27	743	>	atgagtaaa
Rnpa **(36)**			357	357	3704	rnpA	3882122	3882481	>	gtggttaag
BstPol		Multi	2631	2631	U93028	pol	95	2728	>	atgagattg
Rpol_Bpt7		Multi	2649	2649	NC_00160 4	gp1	3171	5822	>	atgaacacg
EFTu **(50)**		451	1179	1179	3339	tufA	3467782	3468966	<	gtgtctaaa
EFG **(59)**		89	2109	2109	3340	fusA	3469037	3471151	<	atggctcgt
EFTs		433	846	846	170	tsf	190857	191708	>	atggctgaa
EFP (no)		26	561	561	4147	efp	4373277	4373843	>	atggcaacg
IF1		173	213	213	884	infA	925448	925666	<	atggccaaa
IF2 **(25)**		142	2682	2682	3168	infB	3310983	3313655	<	atgacagat
IF3 **(~50)**		196	540	540	1718	infC	1798120	1798662	<	attaaaggc
RF1 (no)		258	1080		1211	prfA	1264235	1265317	>	atgaagcct
RRF		435	555	555	172	frr	192872	193429	>	gtgattagc
RL1 **(~50)**	1	82	699	699	3984	rplA	4176457	4177161	>	atggctaaa
RL2	1	154	816	816	3317	rplB	3448180	3449001	<	atggcagtt
RL3	1	151	627	627	3320	rplC	3449934	3450563	<	atgattggt
RL4	1	152	603	603	3319	rplD	3449318	3449923	<	atggaatta
RL5	1	163	534	534	3308	rplE	3444536	3445075	<	atggcgaaa
RL6	1	166	528	528	3305	rplF	3443244	3443777	<	atgtctcgt
RL7	4	362	360	360	3986	rplL	4178138	4178503	>	atgtctatc
RL9	1	93	447	447	4203	rplI	4423686	4424135	>	atgcaagtt
RL10	1	361	492	492	3985	rplJ	4177574	4178071	>	atggcttta
RL11	1	81	423	423	3983	rplK	4176025	4176453	>	atggctaag
RL13	1	418	426	426	3231	rplM	3375858	3376286	<	atgaaaact
RL14	1	161	369	369	3310	rplN	3445415	3445786	<	atgatccaa
RL15	1	169	432	432	3301	rplO	3441742	3442176	<	atgcgttta
RL16	1	158	408	408	3313	rplP	3446396	3446806	<	atgttacaa
RL17	1	178	381	381	3294	rplQ	3437253	3437636	<	atgcgccat
RL18	1	167	351	351	3304	rplR	3442881	3443234	<	atggataag

RL19	1	444	342	342	2606	rplS	2742203	2742550	<	atgagcaac
RL20 *	1	198	351	351	1716	rplT	1797417	1797773	<	atggctcgc
RL21 *	1	232	309	309	3186	rplU	3330781	3331092	<	atgtacgcg
RL22	1	156	330	330	3315	rplV	3447538	3447870	<	atggaaact
RL23	1	153	300	300	3318	rplW	3449019	3449321	<	atgattcgt
RL24	1	162	309	309	3309	rplX	3445090	3445404	<	atggcagcg
RL25	1	?	282		2185	rplY	2280537	2280821	>	atgtttact
RL27 *	1	234	252	252	3185	rpmA	3330503	3330760	<	atggcacat
RL28 *	1	426	231	231	3637	rpmB	3809065	3809301	<	atgtcccga
RL29	1	159	189	189	3312	rpmC	3446205	3446396	<	atgaaagca
RL30	1	?	174		3302	rpmD	3442180	3442359	<	atggcaaag
RL31 *	1	257	210	210	3936	rpmE	4124593	4124805	>	atgaaaaaa
RL32	1	363	168	168	1089	rpmF	1146590	1146763	>	atggccgta
RL33 *	1	325	162	162	3636	rpmG	3808877	3809044	<	atggctaaa
RL34 *	1	466	138	138	3703	rpmH	3881965	3882105	>	atgaaacgc
RL35 *	1	197	192	192	1717	rpmI	1797826	1798023	<	atgccaaaa
RL36 *	1	174	114	114	3299	rpmJ	3440255	3440371	<	atgaaagtt
RS1	1	?	1671		911	rpsA	961218	962891	>	atgactgaa
RS2	1	70	720	720	169	rpsB	189874	190599	>	atggcaact
RS3	1	157	696	696	3314	rpsC	3446819	3447520	<	atgggtcag
RS4	1	311	615	615	3296	rpsD	3438692	3439312	<	atggcaaga
RS5	1	1168	498	498	3303	rpsE	3442363	3442866	<	atggctcac
RS6	1	90	405	405	4200	rpsF	4422696	4423091	>	atgcgtcat
RS7	1	88	534	534	3341	rpsG	3471179	3471718	<	atgccacgt
RS8	1	165	387	387	3306	rpsH	3443790	3444182	<	atgagcatg
RS9	1	417	387	387	3230	rpsI	3375450	3375842	<	atggctgaa
RS10	1	150	309	309	3321	rpsJ	3450596	3450907	<	atgcagaac
RS11	1	176	384	384	3297	rpsK	3439346	3439735	<	atggcaaag
RS12	1	87	369	369	3342	rpsL	3471815	3472189	<	atggcaaca
RS13	1	175	351	351	3298	rpsM	3439752	3440108	<	gtggcccgt
RS14	1	164	294	294	3307	rpsN	3444216	3444521	<	atggctaag
RS15	1	424	264	264	3165	rpsO	3309056	3309325	<	atgtctcta
RS16	1	446	246	246	2609	rpsP	2743957	2744205	<	atggtaact
RS17	1	160	249	249	3311	rpsQ	3445951	3446205	<	atgaccgat
RS18 *	1	92	222	222	4202	rpsR	4423417	4423644	>	atggcacgt
RS19	1	155	273	273	3316	rpsS	3447885	3448163	<	atgccacgt
RS20 *	1	363.5	258	258	23	rpsT	20815	21078	<	ttggctaat
RS21	1	?	210		3065	rpsU	3208422	3208637	>	atgccggta
RS22	1	?	135		1480	rpsV	1553850	1553987	<	atgaaatcg
Sya	4	292	2628	2628	2697	alaS	2817403	2820033	<	atgagcaag
Syc	1	253	1383	1383	526	cysS	553834	555219	>	atgctaaaa
Syd	2	36	1770	1770	1866	aspS	1946774	1948546	<	atgcgtaca
Sye	1	462	1413	1413	2400	gltX	2517277	2518692	<	atgaaaatc
Syfa	2+2	194	981	981	1714	pheS	1795983	1796966	<	atgtcacat
Syfb	2+2	195	2385	2385	1713	pheT	1793581	1795968	<	atgaaattc
Syga	2+2	251	909	909	3560	glyQ	3722036	3722947	<	atgcaaaag
Sygb	2+2	?	2067		3559	glyS	3719957	3722026	<	atgtctgag
Syh	2	35	**1269**	1269	2514	hisS	2637321	2638595	<	gtggcaaaa

Syi	1	345	2814	2814	26	ileS	22391	25207	>	atgagtgac
Syk1	2	136	1512	1512	2890	lysS	3031677	3033194	<	atgtctgaa
Syl	1	266	2580	2580	642	leuS	671424	674006	<	atgcaagag
Sym	2	365	2028	2028	2114	metG	2192320	2194353	>	atgactcaa
Syn	2	113	1395	1395	930	asnS	986808	988208	<	atgagcgtt
Syp	2	283	1716	1716	194	proS	217057	218775	<	atgcgtact
Syq	1	?	**1659**		680	glnS	705316	706980	>	atgagtgag
Syr	1	378	1731	1731	1876	argS	1958086	1959819	>	gtgaatatt
Sys	2	5	1290	1290	893	serS	938651	939943	>	atgctcgat
Syt	2	375	1926	1926	1719	thrS	1798666	1800594	<	atgcctgtt
Syv	1	334	2853	2853	4258	valS	4478550	4481405	<	atggaaaag
Syw	2	126	1002	1002	3384	trpS	3510272	3511276	<	atgactaag
Syy	2	455	1269	1269	1637	tyrS	1713972	1715246	<	atggcaagc

Table of potential genes for the minigenome. Derived from Escherichia, Bacillus, T7, Saccharomyces, and Aequorea. The first column is the protein or RNA name, with SwissProt prefix where applicable, (in parentheses is the % identity to the closest 3D structure or in bold if more than 85% identical. All rRNA and ribosomal proteins are covered at about 50%). Column two is the stoichiometry if a well-characterized complex is formed (e.g. ribosome or tRNA synthetases). Column 3 is Mycoplasma gene number as an indication of genes conserved in an existing "minimal genome". The next two columns are the size of the gene in bp (including stop codon where appropriate). Column 6 is the accession numbers or the unique B# for the M52 version of the E. coli K12 genome sequence. Column 7 is the Demerec gene name (related to column 1). The next two columns are the bp coordinates for the start and end of each gene in the sequence entry listed in column 7. The last two columns are the orientation of the gene in the database entry and the 5' end of the sequence. This table and related materials for initiating the computational modeling can be found in our web site dedicated to this project (for now limited to DARPA review and HMS internal use only) :
http://arep.med.harvard.edu/darpabiocomp/