Harvard Medical School

DARPA BAA 01-26 (BIO-COMP) Section I. Administrative

22 Aug 2001

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:	750,000	Total Option:	1,548,000
FY1 (8/01-7/02):	350,000	FY3 (8/03-7/04):	500,000
FY2 (8/02-7/03):	400,000	FY4 (8/04-7/05):	516,000
		FY5 (8/05-7-06):	532,000

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

NOTICE: Use and Disclosure of Data

This proposal abstract includes data that shall not be disclosed outside the Government and shall not be duplicated, used, or disclosed— in whole or in part—for any purpose other than to evaluate this proposal. If, however, a contract is awarded to this offeror as a result of—or in connection with—the submission of these data, the Government shall have the right to duplicate, use or disclose these data to the extent provided in the resulting contract. This restriction does not limit the Government's right to use information contained in these data if they are obtained from another source without restriction. The data subject to this restriction are contained in the entire proposal abstract.

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-nm³ 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 threedimensionally 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 proteinsynthesis. 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 blackbox 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.**

Б. С.

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 26 deliverables for 5 sub-tasks are:

(In order to conserve on limited funding and recognizing the serendipity of discovery, we anticipate that 60% of the deliverables for year one will merit immediate follow-up. This subset is planned to expand in year two (with options in later years).

1 DNA memory:

1#1: We will establish the scale-up costs of the SSR system from two bits to the theoretical upper-limit. Oligonucleotides required for reversible, irreversible, and interwoven systems will be designed. (Mr. Shendure & Dr. Church)

1#2: We will do a lab test of an alternative to Flp involving Lambda-Red homologous recombination. (Mr. Reppas)

1#3: We will establish a plan including all necessary existing strains and required constructs for testing a RAG recombination variant usable in mammalian whole organism cell-lineage monitoring. (Mr. Shendure)

1#4: We will test an error-prone polymerase in year two with DNA fluorescent-base extension as the output. (Dr. Mitra)

1#5: A goal of 2 kilobits of storage at one bit-perbase in a year 3-4 option.

2 A/D Input: in vivo allostery, in vitro light

2#1: The tetR or other sensor inputs will be studied and integrated into a recombinational system. (Mr. Reppas)

2#2: We will determine the effects of DNA length on Zinc finger binding specificity and test 5 or more Zn-finger domains harvested from a yeast genome. (Dr. Bulyk).

2#3: A third year (option) 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#4: An early milestone in the highest speed & resolution method proposed here will be the incorporation of a fluorescent nucleotide triphosphate by T7 RNA polymerase under control of a caged ATP or GTP. (Dr. Mitra)

2#5: use these to regulate the incorporation of dNTPs in the format of an RNAPol-etaDNAPol fusion. (option)

3 D/A Output: polonies & micromirrors

3#1:The construction of three-dimensional arrays of DNAs and polymerases will initially build on our "polony" technology [*Mitra 1999]. A method for fluorecent in situ sequencing in such 3D polyacrylamide arrays will be developed using chemically cleavable dNTPs. (Dr. Mitra)

3#2: Photocleavable versions of the above will be tested. (Dr. Mitra)

3#3: A microfluidic synthesis of a microarray of over 3000 oligomers (30-mers) will be synthesized using a micromirror array (Mr. Wright collaboratively with Dr. Gao's lab). 3#4-5 (option): We will bridge a one-cm pattern with nm precision using a chemi-optical combination. Controlled polymerase stepperpositioners will be used to incorporate branching oligonucleotides at nm scale. Incorporation of electronic or optical computing components will be attempted.

In vitro minigenome:

4#1: For the minigenome project, important demonstrations will be the replication dependent in vitro translation of GFP or luciferase using crude E.coli lysates. (Dr. Tian)

4#2: As above but showing dependence on added EF-Tu of measures of replication rate. (Dr. Tian in collaboration with Dr. Blacklow's lab).

4#3: A highly purified translation system will be used to extend the results of 4#1 or 4#2. (Dr. Tian in collaboration with Dr. Blacklow's lab).

4#4: A SSR or homologous recombination method will be tested for the transfer of genes for expression in minigenomes. (Mr. Reppas in collaboration with Harvard Inst. Proteomics)

4#5: We will develop a rolling circle primer immobilization method. (Drs. Tian and Huang)

4#6: We will assess the doubling rates varying number of primers, or amount of nicking.

Computational:

5#1: The computational deliverables will be coordinated with other DARPA BIO-COMP participants. (Drs. Church & Aach)

5#2: The known 3D models will be adapted to include any sequence changes initially EF-Tu mutations. (Dr. Church in collaboration with Dr. Blacklow).

5#3: Software for inclusion of microscopic morphology, dynamics, and stochastics will be developed. (Mr. Wright and/or Mr. Karchenko)) 5#4: As an alternative to the direct allosteric design, we will develop a model for signal transduction cascade, which will be based on two-hybrid and microarray data as (Ms. Petti with Dr. Steffen).

5#5: Metrics for optimality and quality assessment and outlier identification will be developed and refined. (Dr. Segre).

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 goodnessof-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 oligonucleotide design tools and chemical kinetic modeling. He will be assisted by Mr. Peter Karchenko.

Dr. Jingdong Tian will implement an array-based system for coupled replication-transcriptiontranslation. 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]

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.

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. 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 nanofabrication projects.

F. Milestones & Schedule

Our development plan consists of 5 subtasks, 26 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.

Subtasks	2001	2002	2003	2004	2005	2006
0. Major Web updates						
1. DNA-memory: Flp		1#1-3				
eta-Polymerase			1#4	1#5		
2. A/D Input: chemical		2#1		2#2	-	2#3
optical		2#4		2#5		
3. D/A Output: polony		3#1	3#2	3#3		
Mechano-chemical		3#4			3#5	
4. Mini-genome: RCA		4#1		4#3	······	
translation			4#2,4,5	4#6		
5. Computational Optimization: 3D/4D		5#1			5#2	
Stochastic			5#3	5#4		5#5

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).

- 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 "genomeengineering" 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.

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 nanofabrication 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.

H. Comparison with other ongoing research

The approach of using the memory components based on DNA as components of a high-density

I. Key Personnel

Name	Level of Effort to be Expended					Other Major Sources of Support	
	Yr 1	Yr 2	Yr 3	Yr 4	Yr 5	(Current and Proposed)	
Dr. George Church *	6%	6%	6%	6%	6%	Current: DoE, NHLBI-PGA, NSF	
Harvard Med. School						Proposed: Bio-Comp & NIH	
Dr. John Aach *	15%	15%	15%	15%	15%	Current: DoE, NHLBI	
Harvard Med. School							
Mr. Matthew Wright	30%	30%	30%	30%	30%	Current: NSF, NHLBI	
MIT-Chemistry	3070	3070	3070	3070	3070		
Dr. Jingdong Tian	50%	100%	100%	100%	100%	Current: LSRF fellowship	
Harvard Med. School	5070	10070	10070	10070	10070		
Mr. Jay Shendure	30%	30%	30%	30%	30%	Current: NSF	
Harvard Med. School	3070	3070	3070	3070	3070		
Ms. Allegra Petti	100/	100/	100/	4004	400/	Current: NSF	
Harvard Med. School	40%	40%	40%	40%	40%		
Mr. Mik Reppas	100/	100/	100/	4004	400/	Current: NSF	
Harvard Med. School	40%	40%	40%	40%	40%		
Dr. Martha Bulyk						Current: DoE	
Harvard Med. School	50%						
Dr. Rob Mitra			-			Current: DoE	
Harvard Med. School	50%	50%	50%	50%	50%		
Dr. Daniel Segre				0.00/		Current: NSF	
Harvard Med. School	80%	80%	80%	80%	80%		
Graduate students and postdoctoral fellows will be replaced as the move along on their career paths.							

Table I.1 Key Personnel.

J. Description of facilities

K. Experimentation and Integration Plans.

Our results will be integrated with the Biomodeling solutions that DARPA BIO-COMP 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 semiannual 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.

Categories	FY1	FY2	FY3	FY4	FY5	\$K
Equipment	0	0	0	0	0	Lipper Foundation cost sharing
Subcontract	0	0	0	0	0	
Travel	0	0	0	0	0	Lipper Foundation cost sharing
Phone, fax, copying	0	0	0	0	0	Lipper Foundation cost sharing
Personnel	110	126	157	162	167	
Fringe	18	20	26	27	28	
Supplies	75	86	107	111	114	
Indirect costs	146	167	209	215	222	
Totals	350	400	500	516	532	

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

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 possible through our collaboration with the SRI/MIT/BU/HMS BIO-COMP Bio-Spice Integration Project, after year two 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)

August 22, 2001

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

Attention:Dr. Sri KumarSubject:Information Technology Resources, BAA 01-26

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. \$75,000 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

(for now limited to DARPA review and HMS internal use only) : http://arep.med.harvard.edu/darpabiocomp/