LARGE SCALE GENOMIC MONITORING OR PROFILING USING A DNA-BASED MEMORY AND MICROARRAYS

Junghuei Chen*, and Yuzhen Wang
Chemistry & Biochemistry, University of Delaware
Newark, Delaware, 19716

Russell Deaton
Computer Science & Computer Engineering, University of Arkansas
Fayetteville, Arkansas, 72701

ABSTRACT

We describe a new method consisting of enzymatic manipulation of genomic DNA or mRNA with DNA microarrays that is capable of monitoring or profiling any organisms presented in a biological sample without priori knowledge of genomic sequences. The method we have developed seeks to “store” all genomic DNA information in the bacterial communities found in a patch of soil, water or air sample.

The goal is to use genomic information at the population or community scale to monitor and detect the existence of new biota (such as pathogens) in the environment. The scope of organisms with their genomic DNA sequenced is fairly small. Thus, much information at the genomic level is not available with conventional techniques. In addition, many organisms are not amenable to laboratory analysis. However, bio-agents used by terrorist group will be a major threat to our national security. The DNA-based memory potentially provides a way to access information from all organisms in a community (airports, train terminals,…etc.) to assess impact by human and non-human biomaterials, does not require explicit sequence knowledge, and is quick, flexible, and inexpensive to implement. Thus, it could provide a holistic view of the genomic status of the whole environment.

1. INTRODUCTION

We propose a reasoning system based on storage and manipulation of DNA *in vitro* that provides a potentially revolutionary approach to biological information processing, and might be used to screen for human disease. For example, a physiological condition may be indicated by expression levels of many genes and complex relationships among them. The system should be capable of capturing a snapshot of an entire biosystem's state in one memory. At a later time, the memory can be updated, or a new snapshot acquired, and compared to a previous memory to measure change.

In addition, snapshots acquired under different conditions can be compared to reason about common or similar mechanisms or effects, or merged to form a combined representation. The intimate interface of the system to biology might produce a more capable, faster, and efficient system for biological information processing. For example, instead of clustering and analyzing patterns of gene expression from a microarray on a conventional computer, it would be done by the intelligent DNA Memory *in vitro*.

1.1 Background and Significance

We present a DNA-based Memory method that has *in vitro* computational capabilities to learn DNA sequences from the microorganisms to which it is exposed, that can detect ecological changes from the genomic information of all microorganisms, known or unknown, in a sample. The advantages of the DNA memory are a potential capacity in the exabyte (10^{18}) range, and the capability to match patterns and classify data based on content. Thus, it could serve as a large database of heterogeneous information that

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **00 DEC 2004**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **Large Scale Genomic Monitoring Or Profiling Using A Dna-Based Memory And Microarrays**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chemistry & Biochemistry, University of Delaware Newark, Delaware, 19716; Computer Science & Computer Engineering, University of Arkansas Fayetteville, Arkansas, 72701**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2004 in Orlando, Florida., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

could be mined for information based on similar content.

With this technology, intelligence would be incorporated through protocols that separate and cluster sequences into different groups, thus, in a sense, choosing between different classification-categories, *in vitro*. For example, the classification might be harmful pathogen or not. The method here would do the pattern recognition and categorization that is currently done in digital computers for applications such as gene expression studies using DNA micro-arrays. However, unlike digital computers, the DNA-based memory has an intimate connection to the biology and by processing the information *in vitro*, can adapt to and detect new situations without knowing the sequences involved. Thus, the method is an attempt to incorporate intelligence into the contents of test tube.

Theoretical analysis and experimental results (see later) indicate the DNA-based memory has a large capacity for separating DNA patterns, and has a fine level of resolution among sample DNA. Although little consensus exists on true species diversity, estimates for the number of living organisms on earth are on the order of 10 ^{7} to 10 ^{8} species, of which most are microbial. Of the known species, only 113 species (8 eukaryotes; 89 prokaryotes; 16 archaea) have their genomes completely sequenced (Embl-ebi, 2002). In the past, genome-enabled studies have focused on individual organisms in specific ecosystems. In addition, standard techniques for studying DNA samples from the environment require some knowledge of the sequence, either for PCR primers or for attachment to a DNA chip. By focusing on single, known organisms, information from other organisms, both known and unknown, is lost. Thus, a challenge is to discover ways to tap the large amounts of information from all organisms in an environment and to use that information to detect deleterious changes to that environment, such as the existence of pathogens.

1.2 Associate DNA Memory

The idea of processing large amounts of information in a test tube, and not on a conventional solid-state computer, presents the possibility of working with genomic DNA on a community or population scale. The DNA computer, in the case described here, is a laboratory protocol that through DNA-to-DNA reactions (hybridizations), matches sequence patterns, thus, stores DNA or RNA sequence information in a DNA-based memory, and then, matches new input to the stored information based upon sequence content.

The storage procedure is called *"learning"* (Valiant, 1984) because it acquires information from examples (the input DNA), and does so without external knowledge of the organisms, or their genomic sequences. In addition, through the "learning" process, the memory DNA acquires information from all organisms in the input, both known and unknown. Moreover, processing the DNA sequences from the enrironment *in vitro* has some additional advantages. The genomic information is processed in one massively parallel step; likewise, matching of stored patterns with new input can be done in parallel, in one step. Similarity is implemented by degree of annealing between new input DNAs and the stored memory DNA sequences, thus providing a technique for recognizing patterns in different environmental samples and detecting change.

As depicted in Figure 1, instead of encoding information directly into individual DNA sequences, the information is stored as combinations of DNA molecules. In other words, instead of encoding a particular piece of information as one DNA sequence, that information is represented as a collection of sequences. For example, unlike a gene chip, in which each spot represents a particular cDNA sequence, here, the input DNA sequence is stored as its constituent subsequences. Moreover, through the DNA memory's learning and recall protocols (see later), the target sequence is decomposed, stored, and recalled *in vitro* through its subsequences.

There are several advantages to this approach. By matching different combinations of subsequences, the memory would be capable of generalizing to new input, at the expense of specificity (Figure 1).

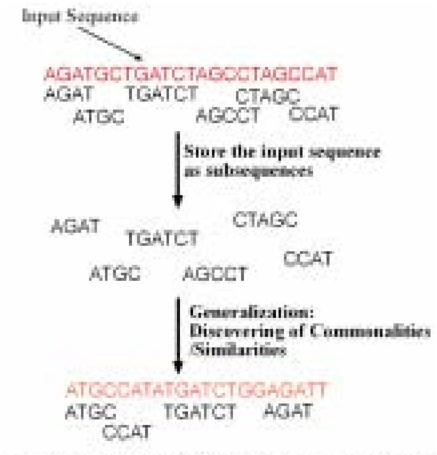

Figure 1: General approach to storing information in memories composed of DNA subsequences.

More conventional techniques would access the genomic information through organisms with known sequences, and then, rely upon a digital computer for processing the data. Pattern recognition and interpretation of large amounts of gene expression or genomic data are difficult problems for conventional computers that the described DNA-memory does *in vitro*.

Finally, DNA's large storage capacity is used to store genomic information from a population or community for subsequent matching. The information is stored in a compact form, and can serve as a database of the status of a specific environment at a given moment in time. Furthermore, when the learned memory DNA of an environment is attached to a DNA microarray, read out can be easily accomplished and interpreted as either a positive or negative match. In addition, the system has the potential for other applications, such as massive parallel detection of pathogens in food sources or the environment, tracking genomic transfer from genetically modified organisms, and storage and retrieval of non-biological information.

The DNA-based memory is an application that has the advantages of DNA computing, mainly the massive parallel ability. For example, the input DNA sequences are learned, stored, and recalled in one massively parallel step. In addition, the proposed memory DNA method uses some of DNA computing's disadvantages, namely the imprecision of matching (the hybridization reaction), to its advantage. Memory recall is implemented by degree of annealing between new input DNAs and the memory sequences, thus providing a technique for recognizing patterns based upon similarities in content. Thus, in the current context of DNA computing, intelligence can be defined as the ability to acquire and apply knowledge to choose among several alternatives on a rational basis.

2. RESULTS

2.1 Learning, Recall, and Reasoning Protocols

A schematic of the DNA memory is shown in Figure 2. The initial sequences are a set of tag sequences, to which random sequences are appended during synthesis. In principle, the starting set of random probes contains every possible sequence of a given length. The tag sequences are designed to be independent of each other in that they will not hybridize to each other (Deaton et al., 2003), and can be used for output by hybridizing to their complements on a DNA array, or by biotin-striptavidin bead separation. With simple and common recombinant DNA operations, such as primer extension, exonuclease digestion, and bead extraction, the system learns the DNA sequences to which it is exposed (Figure 2A).

These learned sequences can then be stored as a DNA memory (Memory Strands). Subsequently, the memory strands can be used to recall new input sequences, or sequences that are close under hybridization affinity (Figure 2B). In the learning, the random energy wells present in the hybridization interactions between memory strand and input are deepened by primer extension process (Figure 2C).

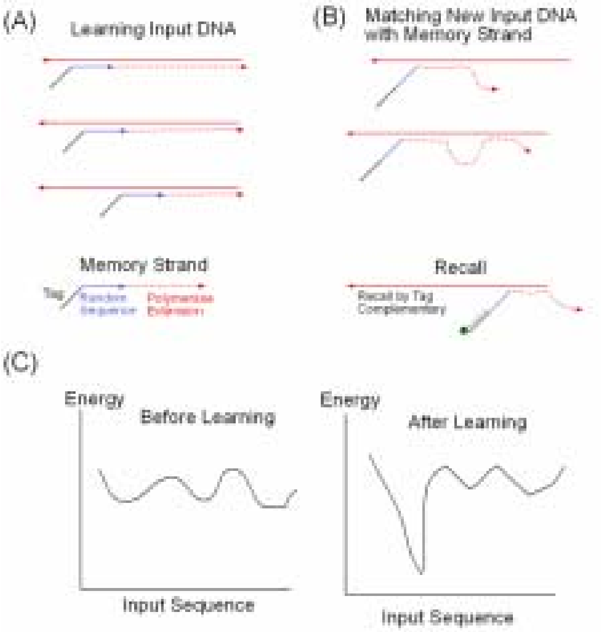

Figure 2: A DNA-based memory with in vitro learning. The memory strands are composed of memory specific sequences and tag sequences that are used for output.

The detailed learning protocol is shown in Figure 3. Initially, the memory strands consist of a tag sequence with biotin attached and short random probe sequences (20-mers). The input DNA, which is learned, is mixed with the initial memory strands (tag plus probes). The probes will hybridize at random locations on the input DNA. After hybridization, a 3' to 5' exonuclease (Exo I) digests probe and input strands from the 3' end until a double-stranded region is encountered (trimming). Then, a 5' to 3' extension by DNA polymerase is done (copying input DNA). The extended memory strands, tag plus extended Watson-Crick complement of input, are separated from the input by the 5' biotin attached to striptavidin beads.

The products of the learning procedure are single-stranded DNAs with a unique tag attached to random length 3' regions that are complementary to the input DNA, and that have undergone some amplification during polymerization. To learn additional inputs, the process is repeated with another initial memory strand with a different tag sequence.

For recall (Figure 4), unknown input is exposed to the different memory strands. The input will hybridize to memory sequences that are close to its Watson-Crick complement. The specific memory that is recalled can be determined from the tag that has the highest concentration of hybridized input. In addition, sequences complementary to the tags, or the memory strands themselves, can be attached to a solid support such as DNA microarray for easy output.

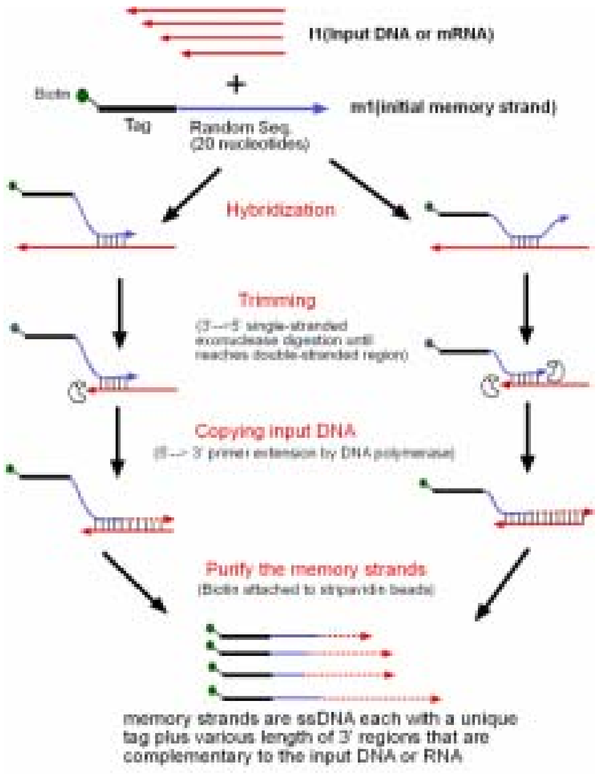
I1(input DNA or mRNA)
+
Biotin
Tag
Random Seq. (20 nucleotides)
m1(initial memory strand)
Hybridization
Trimming
(3'→5' single-stranded exonuclease digestion until reaches double-stranded region)
Copying input DNA
(5'→ 3' primer extension by DNA polymerase)
Purify the memory strands
(Biotin attached to stripavidin beads)
memory strands are ssDNA each with a unique tag plus various length of 3' regions that are complementary to the input DNA or RNA

Figure 3. The learning protocol.

In Figure 4, M1 to M3 are pre-learned DNA memories, and t1 to t3 are their respective tags. The new unknown input DNAs are mixed with memory strands, and hybridize with the memory sequences M1-M3. The hybridization product is then separated by serial ssDNA affinity columns; each of the columns has a specific ssDNA with the sequence complementary to the Tag attached to the cellulose. The DNA that attached to a specific column can be eluted by denaturation. The new input DNAs that are most similar to a specific memory strand can be isolated.

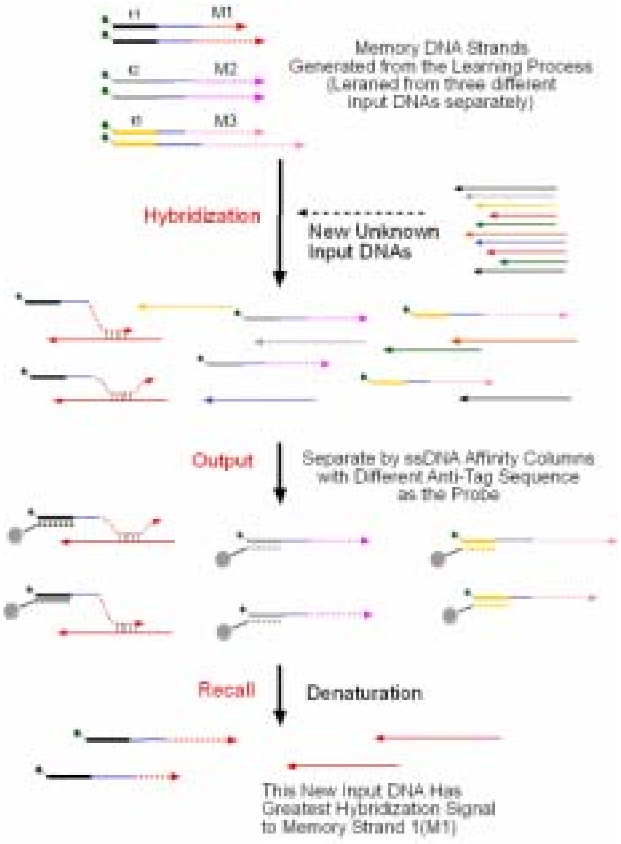

Figure 4. The recall protocol. The lines represent DNA,and arrows represent 3' ends.

2.2 Validation of the Learning Protocol

An important property to characterize the DNA-based memory is the ability of the learning and recall protocols to distinguish different sets of DNA. In the experiments described here, the goal was to test and verify the basic capabilities of the memory strands, which include learning of input DNA sequences, recall of the learned sequences, differentiation of very different sets of sequences, and generalization to input DNA that is close, but not identical, to the learned sequences. In addition, experiments were done to test the sensitivity of the recall protocol, and to determine the extent to which the random probes cover large DNA input spaces.

To test the learning protocol shown in Figure 3, two plasmids, pBluescript (Input 1; 3 kb) and ϕx 174 (Input 2; 4.9 kb), were selected as the input DNAs. These plasmids were chosen because they have very different sequences, and thus, can be the basis for two unrelated memories. After digestion with DNA nuclease I, the starting sets of input DNAs are between 200 and 500 bases long (data not shown). The starting memory DNAs for the learning protocol were two distinct, non-cross hybridizing tag sequences of length 20 bp that had 20 bases of random DNA appended to them, for a total length of 40 bp. With the learning protocol, Memory 1 strands were trained on pBluescript (input 1), and Memory 2 strands on ϕx 174 (input 2).

A denaturing gel indicates different distributions for the learned sequences (the memory strands) for the two input DNAs, and successful extensions of the initial 40-bp memory strands to between 60 and 100 bases (data not shown). This indicates that the learning protocol is successful at randomly sampling the input space of DNA, creating a Watson-Crick complement of the input DNAs, and polishing the 3' ends.

Next, the capabilities of the recall procedure were tested. In the stained gel shown at the left of Figure 5, the original plasmids, pBluescript and ϕx 174 are in lanes 2 and 4, respectively. In lane 1 is the size ladder, that contains sequences from pBluescript, and in lane 3 is a plasmid that shares an ampicillin resistant gene with pBluescript. As shown at the center of Figure 5, the stained gel shown at the left was blotted, and the Memory 1 DNA, obtained from the learning protocol, was radioactively labeled, and used as a probe for the Southern blotting. As seen, Memory 1 hybridized to the DNA in lanes 1 to 3, thus, recalling the input DNA (pBluescript) on which it was trained (lane 2), and two other

DNAs (lanes 1 and 3) which contain some of the input sequences, but not identical, to the training set. In addition, there was no hybridization, and thus, no recall of the very different set of input DNA (ϕx 174) in lane 4.

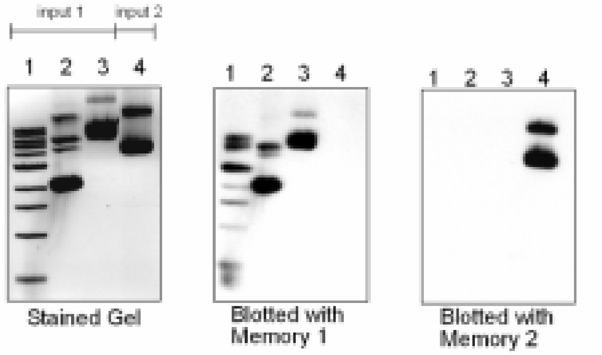

Figure 5. Left is an agarose gel contains 4 DNA samples used as the test set. pBluescript and ϕx 174 are in lanes 2 and 4, respectively. In lane 1 is the size ladder, that contains sequences from pBluescript, and in lane 3 is a plasmid that shares an ampicillin resistant gene with pBluescript. Center is a Southern blot probed with radioactive-labeled Memory 1, which was trained on input 1. Those sets that share some sequences from pBluescript (lanes 1-3) are recalled, but not the dissimilar set (lane 4). Right is a Southern blot with radioactive-labeled Memory 2 as the probe, which was trained on input 2. Input 2 is recalled (lane 4), but not dissimilar DNA (lanes 1-3).

Sensitivity of the technique was also investigated. Varying amounts of ϕx 174 were added to a background of pBluescript. The DNA memory strand 2 (trained on ϕx 174) was able to detect target DNA present in a concentration 1% of the background (Figure 6).

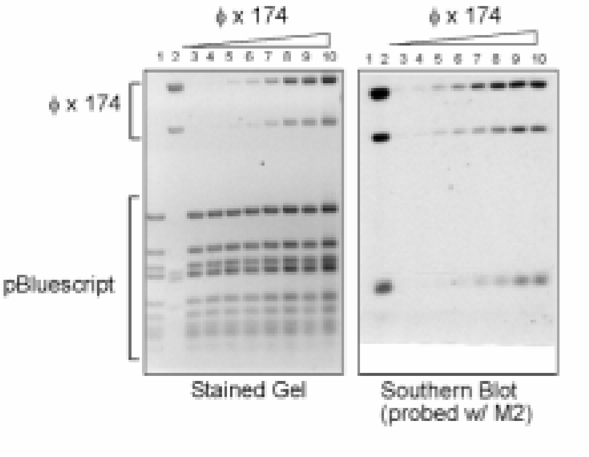

Figure 6. The left panel is an 1 % agarose gel stained with ethidium bromide. Lane 1 contains 14 fragments (from ~100 bp to 600 bp) of pBluescript plasmid (1 ug; ~3 kb) digested with restriction enzyme Hpa II. Lane 2 contains 5 fragments (from ~200 bp to 1.7 and 2kb) of ϕx 174 plasmid (1 μg; ~5 kb) digested with Hpa II. Each of Lanes 3 to 10, contains a fixed amount of pBluescript (1 μg) with increasing amount of ϕx 174 (10 ng, 20 ng, 50 ng, 100 ng, 200 ng, 400 ng, 600 ng, and 800 ng). The right panel is the same gel been blotted to nitrocellulose membrane and probed with Memory strand 2 that has been trained with ϕx 174 as the input DNA. It shows that even with only 1% of ϕx 174 (10 ng) present in pBluescript (1ug), it can still be detected.

Furthermore, we have also measured the ability of the starting random probes to cover a much larger input space. Thus, the genome of *E. coli* (~5 million base pairs (bp)) was learned, and adequately recalled. In addition, the *E. coli* genome was learned with an additional 219 bp fragment of DNA from ϕx 174. The results are shown in Figure 7. After Southern blotting, the DNA memory strand was able to distinguish the 219 bp piece input from among the approximately 5 million bp of the *E. coli* genome (input DNA), showing the capability for an adequate level of resolution.

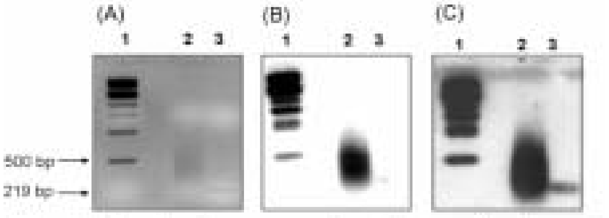

Figure 7. The *E. coli* genome was digested into smaller pieces and learned as memory 1. Then, a 219 bp fragment of ϕX 174 was added to the digested E. coli and learned as memory 2. Recall by blotting showed that memory 1 (*E. coli* alone) was not able to distinguish the ϕx 174 fragment, while learning with the fragment present (memory 2) was. Thus, the learning and recall protocols were able to distinguish a 219 bp sequence from approximately 5 million bp in E. coli. (A) Shown is a stain gel of *E. coli* genomic DNA digested with DNaseI (lane 2), and a 219 bp DNA fragment from ϕx 174 (lane 3). Lane 1 is the markers. (B) Southern blot of gel shown in (A), probed by M1 which has been trained with *E. coli* genomic DNA as input. (C) Southern blot of gel shown in (A), probed by M2 which has been trained with *E. coli* genomic DNA plus the 219 bp ϕx 174 DNA as input.

2.3 Applications of the DNA memory to detect pathogens in environments and medical screening from gene expression.

Two types of DNA microarrays, cDNA type arrays or oligonucleotide arrays, can be used as an output device for the DNA memory. As illustrated in Figure 8, each memory strand learned from the digested genomic DNA of a specific microorganism would be represented as a single spot in the reference micro-array. The references micro-arrays would consist of spots corresponding to memories of many different species of known microorganisms. Differences in the hybridization patterns probed by the new memory strands on the references would indicate changes of the composition of the microorganisms in the ecosystems. The reference chips will be made according to the standard methods of DNA microarray technology (11).

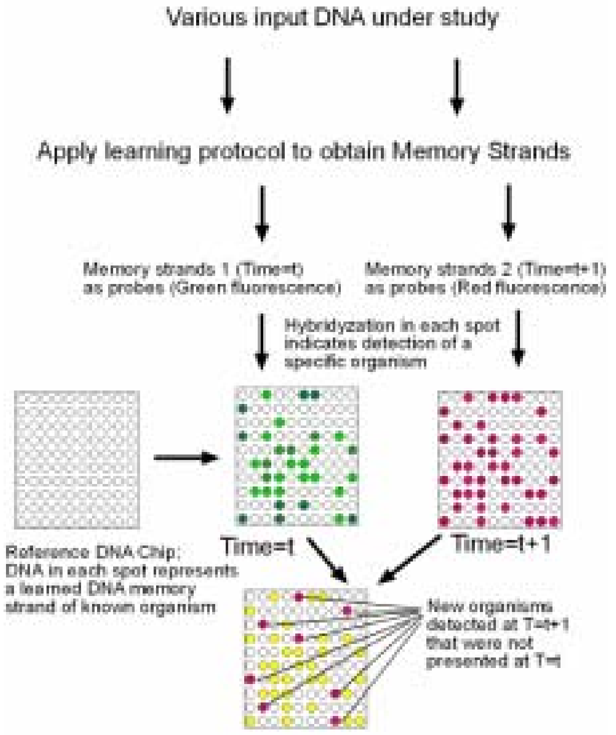
Various input DNA under study
Apply learning protocol to obtain Memory Strands
Memory strands 1 (Time=t) as probes (Green fluorescence)
Memory strands 2 (Time=t+1) as probes (Red fluorescence)
Hybridyzation in each spot indicates detection of a specific organism
Reference DNA Chip: DNA in each spot represents a learned DNA memory strand of known organism
Time=t
Time=t+1
New organisms detected at T=t+1 that were not presented at T=t

Figure 8. Use of DNA Microarrays to Detect Changes.

The basic idea is the whole patterns of spots on the arrays represent a hybridization signature for a particular input DNA sample taken from a specific environment. For these arrays, each spot on the array is whole memory strands, and thus, any particular spot represents the condition under which that memory was formed. Thus, with many spots on the array, the compression of information in the memories enables an array to represent higher-level relationships. For instance, the different patterns, commonalities and similarities among many ecosystems can be obtained by probing each of many reference chips with a specific memory strands learned from a specific ecosystem.

In addition, oligonucleotide arrays (Chee et al., 1996) with specific oligonucleotide DNA synthesized on the chip can also be used as an output device. With this type of chip, instead of DNA memory strands, sets of designed sequences can be synthesized directly onto the arrays (Affymatrix, Santa Clara, CA). These sets will be designed to cover the hybridization space using the tools that we have developed (Deaton et al., 2003a). Synthesized on the oligonucleotide arrays are known sequences that are subsets of all possible sequences of a given length, for example all the possible 10-mers, and will be designed (Deaton et al., 2003 a, b) to have specific hybridization properties. Either memory DNA strands obtained through learning from a specific input DNA or input DNA itself can be hybridized on these arrays, producing a signature identifying the input condition and a spectrum of the sequence composition of the input. Comparison of these results will help characterize the DNA memory with reasoning. Potentially, this might detect minor differences of two closely related input DNAs. For example, gene expression patterns of normal cells *vs* cancer cells.

2.4 Comparison with Current Technology

The essential problem that the proposed techniques try to solve is the production of a signature or fingerprint to identify an organism or a physiological condition from the composition of genomic material, either genomic DNA or mRNA, including those sequences present in low abundance (Lievens et al., 2001). The DNA memory uses techniques similar to many current protocols to implement long-term storage with reasoning capabilities; however, its application is very different from the specific purposes of the current techniques; mainly for identification of individual organisms or differentially expressed genes, and thus, are usually focused on specific answers to one problem at a time. Our goals are not the identification or quantification of individual organisms or differentially expressed genes, but the identification and quantification of populations or physiological conditions. Thus, our DNA memory takes a higher level approach, and as a consequence, provides different information than current conventional techniques.

REFERENCES

M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. A. Fodor, "Accessing genetic information with high density DNA arrays," *Science*, 274, 610–614, 1996.

R. Deaton, J. Chen, H. Bi, and J. A. Rose, "A software tool for generating non-cross hybridizing libraries of DNA oligonucleotides," in DNA Computing: 8th International Workshop on DNA-Based Computers (M. Hagiya and A. Ohuchi, eds.), (Berlin), pp. 252–261, Springer Verlag, 2003b. *Lecture Notes in Computer Science* 2568.

R. Deaton, J. W. Kim, and J. Chen, "Design and test of non-crosshybridizing oligonucleotide building blocks for DNA computers and nanostructures." *Appl. Phys. Lett.,* 82, 1305-1307, 2003 a.

"Embl-ebi" http://www.ebi.ac.uk/genomes/toc.html, 2002.

S. Lievens, S. Goormachtig, and M. Holsters, "A critical evaluation of differential display as a tool to identify genes involved in legume modulation: looking back and looking forward," *Nucleic Acids Research*, 29, 3459–3468, 2001.

L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, 27, 1134–1142, 1984.

CONCLUSION

In sum, the proposed DNA memory seeks to answer whether the population of microbes present in an environment has changed over time and whether the population might contain any dangerous species, not what those specific species are. Likewise, it would answer whether the physiological condition of an organism has change and does it match a set of previously stored conditions, not the specific gene whose expression level has changed.

Therefore, the intelligent DNA memory attempts to capture global information about a population of organisms, or whole genome gene expressions under certain conditions, and recognize patterns of contrast and commonality at that higher level. For example, in gene expression, the learning protocol captures the entire population of genes and their levels of expression, not just those that are differentially expressed. Moreover, the DNA memory incorporates intelligent processing and reasoning capabilities into the test tube. It is not simply a lab technique to gather data for analysis in an electronic computer, but uses the massive scale of storage and parallelism of DNA as the computational tool to draw inferences on the entire *in vitro* knowledge base quickly and efficiently.

This means that the DNA memory can reason and extract knowledge in situations that involve new or unknown information, which conventional lab techniques and electronic computers cannot do. Examples of this include populations of microorganisms with unknown, unsequenced, or mutant organisms and physiological conditions with complicated or unknown patterns of gene expression.

Thus, because of its simplicity of implementation, the proposed work would make large-scale DNA-based associative memories a reality in the near-term, as well as providing a convenient mechanism for applications in biosensing and gene expression. In the future, it is also possible to apply this technique to non-biological data, the advantage of a DNA memory are massive scale and storage density with potentially exabyte (10^{18}) amounts of information in a gram of DNA, the massive parallelism of the search and reasoning protocol, which could supply substantial speed-ups, and the capability to search data based upon context and content, thus providing a semantic component to the memory.