Application of Biomolecular Computing to Medical Science: A Biomolecular Database System for Storage, Processing & Retrieval of Genetic Information & Material

John H Reif[1], Michael Hauser[2], Michael Pirrung[3], and Thomas LaBean[4]

Summary. A key problem in medical science and genomics is that of the efficient storage, processing and retrieval of genetic information and material. This paper presents an architecture for a Biomolecular Database system which provide a unique capability in genomics. It completely bypasses the usual transformation from biological material (genomic DNA and transcribed RNA) to digital media, as done in conventional bio-informatics. Instead, biotechnology techniques provide the needed capability of a Biomolecular Database system, without ever transferring the biological information into a digital media.

The inputs to the system are DNA obtained from tissues: either genomic DNA, or reverse transcript cDNA. The input DNA is then tagged with artificially synthesized DNA strands. These "information tags" encode essential information (e.g., identification of the individual from which the DNA was obtained, as well as the date of the sample, gender, date of birth, etc.) about the individual or cell type that the DNA was obtained from. The resulting Biomolecular Database is capable of containing a vast store of genomic DNA obtained from many individuals (e.g., multiple divisions of an army, etc.). For example the DNA of a million individuals requires about 6 pedabits ($6x10^{15}$ bits); but due to the compactness of DNA, a volume the size of a conventional test tube with a few milliliters of solution contains that entire Biomolecular Database. Known procedures for amplification and reproduction of the resulting Biomolecular Database are discussed.

The Biomolecular Database system has the capability of retrieval of subsets of the stored genetic material, which are specified by associative queries on the tags and/or the attached genomic DNA strands, as well as logical selection queries on the tags of the database. We describe how these queries can be executed by applying recombinant DNA operations on this Biomolecular Database, which have the effect of selection of subsets of the database as specified by the queries. In particular, we describe how to execute these queries on

[1] Dept. of Computer Science, Duke University, Durham, NC 27708. Email: reif@cs.duke.edu

[2] Dept. of Ophthalmology, Duke University Medical Center, Durham, NC 27708.

[3] Dept. of Chemistry, Duke University, Durham, NC 27708.

this Biomolecular Database by the use of Biomolecular computing (also known as DNA computing) techniques, including the execution of parallel associative search queries on DNA databases, and the execution of logical operations using recombinant DNA operations. We also utilize recent biotechnology developments (recombinant DNA technology, DNA hybridization arrays, DNA tagging methods, etc.) that are quickly being enhanced in scale (e.g., output is via DNA hybridization array technology).

The paper also discusses applications of such a Biomolecular Database System to various medical science and genomic processing capabilities, including: (a) rapid identification of subpopulations possessing a specific known genotype, (b) large-scale gene expression profiling using DNA databases, and (c) streamlining identification of susceptibility genes: high throughput screening of candidate genes to optimize genetic association analysis for complex diseases. Such a Biomolecular Database system may provide a revolutionary change in the way that these genomic problems are solved.

1 Introduction

1.1 Motivation: The Need for a Compact Database System for Storage, Processing & Retrieval of Genetic Information & Material. The recent advances in biotechnology (recombinant DNA techniques such as rapid DNA sequencing, cDNA hybridization arrays, cell sorters, etc.) have resulted in many benefits in the health fields. However, these advances in biotechnology have also brought risks and considerable further challenges. The *risks* include the use of biotechnology for weaponry: e.g., diseases (or environmental stresses) engineered to attack and disable military personnel. The *challenges* include the difficulties associated with the acquisition, storage, processing and retrieval of individual genetic information. In particular, it is apparent that the sequencing of the human genome is not sufficient for many medical therapies, and instead one may require information about the specific DNA of the diseased individual, as well as information concerning the expression of genes in various tissue and cell-types. In the scenario of biological warfare, such individual specific information can be essential for therapies or risk-mitigation (e.g., identification of individuals likely to be susceptible to a particular biological attack). To do this, there must be a capability to store this biological information, and also a capability to execute queries that identify

[4] Dept. of Computer Science, Duke University, Durham, NC 27708.

individuals containing certain selected subsequences in their DNA (or transcribed RNA). Hence, what is needed is essentially a database system capable of storing and retrieving biological material and information.

This biological information is quite data-intensive; the DNA of a single human contains about 6 gigabits of information, and the number of genes that potentially may be expressed may total approximately 30,000 (up to 15,000 genes may be expressed in each particular cell-type, and there are thousands of cell-types). The DNA of a single individual contains about 3 x 10^9 bases which (with 4 bases) is 6 x 10^9 bits. The DNA of a million individuals (for example, a large military force) therefore requires 6 pedabits (a pedabit is 10^{15} bits). The expression information for a few dozen cell-types in each of a million individuals may also require multiple pedabits. Although the acquisition of such a vast DNA databank may be feasible via standard biotechnology, the rapid transfer of the DNA of such a large number of individuals into a digital media seems infeasible, due to the tedious and time-consuming nature of DNA sequencing. Even if this large amount of information could be transferred into digital media, it certainly would not be compact: current storage technologies require considerable volume (at least a few dozen cubic meters) to store a pedabit. Furthermore, even simple database operations on such a large amount of data require vast computational processing power (if executed in a few minutes).

1.2 Overview of the Biomolecular Database System

This paper presents architecture for a Biomolecular Database system for the efficient storage, processing and retrieval of genetic information and material. It completely bypasses the usual transformation from biological material (genomic DNA and transcribed RNA) to digital media, as done in conventional bio-informatics. Instead, biotechnology techniques provide the needed capability of a Biomolecular Database system, without ever transferring the biological information into a digital media. It may provide a potentially unique and revolutionary capability in genomics.

DNA: An Ultra-Compact Storage Media. The storage media of this database system are strands of DNA that are (in comparison to RNA) relatively stable and non-reactive: they can be stored for a number of years without significant degradation. In particular, the genetic information can be stored in the form of DNA strands containing fragments of genomic DNA as well as appended strands of synthesized DNA ("information tags") encoding information relevant to the genomic DNA. This Biomolecular Database is capable of containing a vast store of genomic DNA obtained from many individuals (e.g., multiple divisions of an army, etc.). We can provide the store with a redundancy (i.e., number of copies of each DNA in the database) that range from a few hundred or thousand to downwards to perhaps 10, as the stringency of the methods increase. As mentioned above, the DNA of 1,000,000 individuals contains 6 pedabit, but due to the compactness of DNA, a volume the size of a conventional test tube can contain the entire Biomolecular Database. A pedabit of information can be stored (with 10-fold redundancy) in less than a few milligrams of dehydrated DNA, or when hydrated may be stored within a test tube containing a few milliliters of solution.

Construction of the Biomolecular Database System. The inputs to the system are DNA obtained from tissues: either genomic DNA, or reverse transcript cDNA obtained from mRNA expressed from the DNA of a particular cell type. The Biomolecular Database is constructed as follows:

(a) The input DNA strands are first fragmented, e.g., they may be partially digested into moderate length sequences by the use of restriction enzymes. We describe a variety of methods for fragmentation protocols, and compare them by their distribution of strand lengths, and the predictability of the end sequences of the fragmented DNA.

(b): The DNA are then tagged with artificially synthesized DNA strands. These "information tags" encode essential information (e.g., identification of the individual from which the DNA was obtained, as well as the date of the sample, gender, date of birth, etc.) about the individual or cell type that the DNA was obtained from. These "information tags" are represented by a sequence of distinct DNA words, each encoding variables over a small domain. We describe and test tagging protocols based on primer extension and utilizing the predictability of the end sequences of the fragmented DNA.

Processing Queries in the Biomolecular Database System. The paper then discusses how to execute queries on this resulting Biomolecular Database. The system makes the use of Biomolecular computing (also known as DNA computing) methods to execute these queries, including the execution of parallel associative search queries on DNA databases, and the execution of logical operations using recombinant DNA operations. We also describe the use of conventional biotechnology (recombinant DNA technology, DNA hybridization arrays, DNA tagging methods, etc.), e.g., output is via DNA hybridization array technology. These queries include retrieval of subsets of the stored genetic material, which are specified by associative queries on the tags and/or the attached genomic DNA strands, as well as logical selection queries on the tags of the database. These queries are executed by applying recombinant DNA operations on this Biomolecular Database, which have the effect of selection of subsets of the database as specified by the queries. We describe two distinct methods for processing logical queries: a surface-based primer-extension method, as well as a solution-based PCR method. The query processing is executed with vast molecular-level parallelism by a sequence of biochemical reactions requiring time that remains nearly invariant of the size of the database up to extremely large database sizes (e.g., up to 10^{15}). This is because the key limitation is the time for DNA hybridization, which is done in parallel on all the DNA. The output of the queries would be via DNA hybridization array technology.

Computer Simulations and Software. We describe computer simulations and software that can be used for the analysis and optimization of the experimental protocols. In particular, we describe the use of computer simulations for the design of hybridization targets for the readout of information tags and SAGE tags by microarray analysis. We also discuss the scalability of these methods to do logical query processing within Biomolecular Databases of various sizes.

Applications. The paper also discusses applications of a Biomolecular Database System to provide various genomic processing capabilities, including: (a) rapid identification of subpopulations possessing a specific known genotype, (b) large-scale gene expression profiling using Biomolecular Databases, and (c) streamlining identification of susceptibility genes: high throughput screening of candidate genes to optimize genetic association analysis for complex diseases. Such a Biomolecular Database system provides a revolutionary change in the way that these genomic problems can be solved, with the following advantages: (i) the avoidance of sequencing for conversion from genomic DNA to digital media, (ii) the extreme compactness and portability of the storage media, (iii) the use of vast molecular parallelism to execute the operations, and (iv) scalability of the technology, requiring volume that scales linearly with the size of the database, and query time that is nearly invariant of that size. These unique advantages may potentially provide a number of opportunities for a variety of applications beyond medicine, since they also impact defense and intelligence in the biological domain. Applications discussed include reasonable scenarios in (a) medical applications (e.g., oncology: rapid screening, among a selected set of individuals, for expressed genes characteristic of specific cancers), (b) biological warfare (e.g., biological threat analysis: rapid screening of a large selected set of personnel for possible susceptibility to natural or artificial diseases or environmental stresses, via their expressed genes), and (c) intelligence (e.g., identification of an individual, out of a large selected subpopulation, from small portions of highly fragmented DNA).

1.3 Organization of the Paper.

In this section we have provided a brief medical science motivation for a Biomolecular Database system, and a brief overview of the system. In the next Section 2 we briefly discuss relevant conventional biotechnologies and we briefly overview the Biomolecular computing (also known as DNA computing) field. In Section 3 we describe in detail our Biomolecular Database system. In that section we make use of various relevant Biomolecular computing methods, including the use of word designs for synthetic DNA tags, execution of parallel associative search queries on DNA databases, and the execution of logical operations using recombinant DNA operations. In Section 4 we discuss a number of genomic processing applications of Biomolecular Database systems. In Section 5 we conclude the paper with a review of potential advantages of Biomolecular Database systems, and acknowledgements.

2. Review of Biotechnologies for Genomics and the Biomolecular Computing Field.

2.1 Conventional Biotechnologies for Genomics. There have been considerable commercial biotechnology developments in the last few decades, and many further increases in scale can reasonably be expected in the next five years. For example, the DNA hybridization array technology developed by Affymetrix, Inc. (capability is currently up to 400,000 output spots, and within 5 years, a projected 1,000,000 outputs) can be adapted for output of queries to conventional optical/electronic media. Other biotechnology firms (e.g., Genzyme Molecular Oncology, Inc.,) also have developed competing biotechnologies.

Genomics. In the research field known as *genomics*, there are a number of main areas of focus, each with somewhat different goals. These include:

(a) DNA Sequencing. *Sequencing* is the determination of a specific base pair sequence making up the DNA. This tells us all the possible genes that a given organism may express, its genetic make-up. In conventional bio-informatics, it is generally assumed that the genes discussed have been previously sequenced and placed in a computer database.

(**b) Gene Expression Analysis.** *Expression analysis* attempts to determine which genes are being expressed in a given tissue or cell-type at a specific moment in time. The objective, to identify all the genes that are being expressed, is challenging because of the great complexity of the mixture of mRNA being analyzed--each cell may express as many as tens of thousands of genes (Carulli et al. 1996). SAGE Tagging and cDNA hybridization arrays, as discussed below, are techniques for determining comprehensive gene expression data for a given cell-type or tissue. The technique of differential expression analysis compares the level of gene expression between two different samples. Variations in the level of expression of individual genes or groups of genes can provide valuable clues to the underlying mechanism of the disease process. A number of methods currently being used to obtain comprehensive gene expression data are described below.

***(i) cDNA* hybridization arrays.** A *cDNA* hybridization array is composed of distinct DNA strands arrayed at spatially distinct locations. A cDNA hybridization array operates by hybridizing the array with fluorescent-tagged probes made from mRNA, which anneal to its DNA strands. This generates a fluorescent image defining expression, which provides a very rapid optical readout of expressed genes. However, cDNA hybridization arrays are generally manufactured for use with a given set of expressed genes, for example those of a given cell type. The design and manufacture of cDNA hybridization arrays for a given expression library of size over 10,000 can be quite costly and lengthy. Affymetrix has recently developed an oligonucleotide array, known as a UniversalChip that is not specialized to any gene library; it consists of 2000 unique probe sequences that exhibit low cross-hybridization and broad sampling of sequence space It can be used with fluorescent-tagged probes made from DNA rather than mRNA. This technology can be used for output in a Biomolecular Database system.

(ii) ***Serial Analysis of Gene Expression (SAGE)*** is a technique for profiling the genes present in a population of mRNA. By the use of various restriction enzymes, SAGE generates, for each mRNA, a 10-base tag that usually uniquely identifies a given gene. In the usual SAGE protocol, the resulting SAGE tags are blunt-end ligated together and the results are sequenced. The sequencing is faster than sequencing the entire expressed genes because the tags are much shorter than the actual mRNA they represent. Once sequencing is complete, the tag sequences can be looked up in a public database to find the corresponding gene. Using the sequence data and the current UniGene clusters, a computer processing stage determines the genes that have been expressed. SAGE can be used on any set of expressed genes and it is not specialized to a particular set. This technology can be adapted for use for additional information tags appended to the DNA in a Biomolecular Database. Genzyme Molecular Oncology, Inc. is the developer of this SAGE technology.

(iii) ***Differential Expression Analysis*** is a technique for finding the difference in gene expression, e.g., between two distinct gene types. Lynx Therapeutics, Inc. has developed a randomized tagging technique for differential expression analysis. The randomized tagging techniques of Lynx Therapeutics, Inc. may be adapted to determine the difference between two Biomolecular Database subsets.

2.2 Relevant Biomolecular Computing Techniques

Biomolecular Computing. In the field known as *Biomolecular Computing* (and also known as *DNA Computing*), computations are executed on data encoded in DNA strands, and computational operations are executed by use of recombinant DNA operations. Surveys of the entire field of DNA-based computation are given in (Reif, 1998, Reif, 2002).

The first experimental demonstration of Biomolecular Computing was done by Adelman (1994) who solved a small instance of a combinatorial search problem known as the Hamiltonian path problem. Considerable effort in the field of Biomolecular Computing methods has been made to solve Boolean satisfiability problems (SAT) problems, which is the problem of finding the Boolean variable assignments that satisfy a Boolean formula. Frutos, Thiel, Condon, Smith, Corn, (1997); Faulhammer, Cukras, Lipton, (2000); Liu, Liman, Frutos, Condon, Corn, (2000), applied surface chemistry methods and Pirrung, et al. (Pirrung, Connors, Montague-Smith, Odenbaugh, Walcott, Tollett, (2000), improved their fidelity. Recently Adelman's group Braich, Chelyapov, Johnson, Rothemund, Adleman, (2002), solved a SAT Problem with 20 Boolean variables using gel- separation methods. While the 20 Boolean variables size problem is impressive, Reif (2002), has pointed out that the use of Biomolecular Computing to solve very large SAT problems is limited to at most approximately 80 variables, so is not greatly scalable in the number of variables.

In contrast, the use of Biomolecular Computing to store and access large databases appears to be a much more scalable application. Baum (1995), first discussed the use of DNA for information storage and associative search and Lipton (1996), Bancroft, Bowler, Bloom, Cleeland (2001), also discussed this application. Reif, LaBean (2000), developed and Reif, LaBean, Pirrung, Rana, Guo, Kingsford, Wickham (2001), experimentally tested the synthesis of very large DNA-encoded databases with the capability of storing vast amount of information in very compact volumes. Reif, et al. (2001), tested the use of DNA hybridization to do fast associative searches within these DNA databases. Reif (1995), also developed theoretical DNA methods for executing more sophisticated database operations on DNA data such as the database join operations and various massively parallel operations on the DNA data. [Gehani, 1998] investigated methods for executing DNA-based computation using micro-fluidics technologies. Also, [Gehani, et al 1999] describes a number of methods for DNA-based cryptography and counter-measures for DNA-based steganography systems as well as discuss various modified DNA steganography systems which appear to have improved security. Kashiwamura, Yamamoto, Kameda, Shiba, Ohuchi (2002), describe the use of nested PCR to do hierarchical memory operations. Suyama, Nishida, Kurata, Omargari (2000), and Sakakibara, Suyama (2000), has developed Biomolecular Computing methods for gene expression analysis. Recently Garzon, Deaton, Neathery, Murphy, Franceschetti, Stevens (1997), analyzed the efficiency and reliability of associative search in DNA databases, and Chen, Deaton, Wang (2003), discuss DNA databases with natural DNA based the prior work of Reif, et al. (2001), and this work.

3 A Biomolecular Database System

3.0 Overview. The inputs to the system are natural DNA obtained from tissues: either genomic DNA, or reverse transcript cDNA obtained from mRNA expressed from the DNA of a particular cell type. A short piece of synthetic DNA is added to each natural DNA strand. This piece of synthetic DNA, called an information tag, is used to code information about that piece of DNA. This information can include the age or gender of the person from whom the DNA came, or the clinical symptoms of individuals suffering from a disease. In a typical application, the Biomolecular Database consists of a mixture of DNA strands from many different people (or other organisms). This Biomolecular Database system is capable of storage, processing and retrieval of genetic information and material. Individual molecules of DNA in the Biomolecular Database can be selected and removed from the mixture on the basis of the information that is encoded in their information tag. This paper describes several innovative biological applications for Biomolecular Databases; in particular, we discuss the application of our Biomolecular Database system to a number of genomic information processing applications.

3.1 Biological Inputs. The inputs to the system are DNA obtained from tissues. Typically this input DNA is either (i) genomic DNA, or (ii) reverse transcript cDNA obtained from mRNA expressed from the DNA of a particular cell type. (To insure stability and non-reactivity, we suggest the database be composed of DNA rather than RNA.)

3.2 Preprocessing the DNA. Biochemical operations can be used to partially digest the DNA by restriction enzymes (insuring the resulting DNA strands are of modest size), and then label the resulting genomic DNA fragments with synthetic DNA information tags.

Fragmentation of the input DNA. The creation of a Biomolecular Database must involve some degree of fragmentation of genomic DNA. While this may at first seem a very simple step, in fact it is critical to later processing that this fragmentation step be done in a highly controllable way. We describe several methods to produce DNA strands of the desired length. In these, one requires that the methods both produce a predictable distribution of lengths, and also that at least one of the resulting ends have a defined sequence.

(a) Mechanical shearing. This is a method that produces a size distribution, however it is not so useful in our context since the resulting ends have undefined sequences.

(b) Reagent-less methods to create breaks. Pirrung, Zhao, Harris (2001), developed a nucleoside analogue whose backbone can be cleaved by long-wavelength UV light, and specific photocleavable T analogs could be used (analogous to the dUTP method). However, again it is not so useful in our context since the resulting ends have undefined sequences.

(c) Controlled digestion of high MW DNA by DNAse I. This is another method that can be used to produce DNA of a specific size range. It relies on careful monitoring of reaction progress and does not produce specific sequences at the ends of the fragments to enable ligation or PCR processes.

(d) Digestion of DNA with restriction endonucleases. This offers the advantage that known sticky ends are generated.

(e) "Rare Cutting" Endonucleases. These can be used to produce DNA fragments of larger sizes. The recognition sequence of such enzymes is as large as 8 bp, meaning that on average, DNA is cut to $1/(0.25^8)$ or 65 kb. In many situations, fragments larger than 65 kb may be desired; for example, complicated loci with many introns might comprise as much as 100 kb, and that is just for one gene.

(f) PCR methods for Fragmentation. One attractive alternative is to use PCR. Random-primed PCR has been used to amplify the whole genome of single sperm (Li, (1990); Arnheim, Li, Cui (1990); Zhang, Cui, Schmitt, Hubert, Navidi, Arnheim (1992). The challenge in using this strategy is to create long amplicons. In principle, amplicon size in random-primed PCR is a function only of the average distance between two inward-facing hybridized primers, which is then a function only of the primer concentration and temperature. Modest flexibility exists in the hybridization temperature in PCR, so a fruitful strategy to make long amplicons is to lower the primer concentration. In order to efficiently amplify with a low primer concentration, the primers should have a high melting temperature (T_m). Increasing length and G/C content increase primer T_m. Random-primed PCR can therefore be examined with novel conditions and primer designs to maximize the amplicon length. Lengthening the random primers by oligonucleotide synthesis is straightforward. Making the primers G/C-rich is challenging, as G/C-rich templates are known to be more difficult to amplify owing to increased secondary structure. Substitution of *nonstandard bases* such as deazaG for G reduces this difficulty.

(g) UV-sensitive nucleoside analogues in Cell Growth for Fragmentation. Another approach is to grow immortalized lymphoblast cell lines in the presence of UV-sensitive nucleoside analogues Pirrung et al. (2001). These analogues can be incorporated into the DNA of the cells, which could subsequently be cleaved by exposure to UV light. The concentration of the analogues determines the frequency of their incorporation, and the size of the resulting fragments.

3.3 Creation of the Tagged Biomolecular Database. These DNA tags are composed of a concatenation of short subsequences, which encode scalar data values. For example, the information tags may contain the individual's unique ID and the cell type (in the case of reverse transcript cDNA obtained from the RNA expressed by a particular cell type) of the genomic DNA and may also encode other useful information (e.g., sex and birth date of the individual). The tagging can be done using known methods, e.g., a primer extension reaction, using the fact that one of the ends of the genomic cDNA can be predicted by the use of the appropriate initial fragmentation process, and further designing the tags with an ends complementary to these sticky ends resulting from the fragmentation process. The resulting database elements have tags on each 5'- and 3'-end. This can be done so that each Biomolecular Database strand bears a universal amplification (primer) sequence at the extreme 5'- and 3'-ends.

3.4 DNA Word Design for the Information Tags. A key problem is the design of a lexicon of short DNA sequences (DNA words) for the information tags in our Biomolecular Database. (Our DNA "information tag" sequences are in general a subset of such a lexicon). Careful word design is crucial for optimizing error control in the queries that are executed the Biomolecular Database. Good word design can be used to minimize unwanted secondary structure, and minimize mismatching, by maximizing binding specificity. There are conflicting requirements on word design: as strand length decreases (which is desirable), the difference between distinct words of information decreases (which is not desirable). Prior work in DNA word design includes a four-base mismatch word design used for surfaced-based DNA computing (Gray, Frutos, Berman, Condon, Lagally, Smith, Corn (1996), and in (Frutos et al. (1997).. [CRFCC+96] (not in references) shows that surface morphology may be an important factor for discrimination of mismatched DNA sequences. A three-base design was used by L. Landweber (1997), and Cukras, Faulhammer, Lipton, and Landweber (1998). Evolutionary search methods for word designs are described in (Deaton, Murphy, Rose, Garzon, Franceschetti, Stevens (1997). Other DNA word designs are described in (A96 not in references, Baum (1996); Deaton, Murphy, Garzon, Franceschetti, Stevens (1996); Mir (1996);, Garzon et al. (1997). Laboratory experiments of word designs are described in Kaplan, Cecchi, Libchaber (1996), and ligation experiments are described by Jonoska and Karl (1997). (not in references) Wood (1998) (not in references) considers the use of error correcting codes for word design and to decrease mismatch errors. One can utilize and improve on these methods for DNA word design, including evolutionary search methods, and error correcting codes. Hartemink, Gifford, Khodor (1998), describes an automated constraint-based procedure for nucleotide sequence selection in word design. In designing the DNA tags used in the database, one needs to determine how many residues should be used for each data block of the tag-sequence (the tag sequence on the database strands binds to the probe sequence on the query strand), and then decide how many words are required at each block position (determined by the number of values available to the variable).

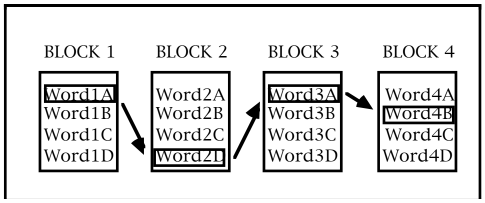

Figure 1

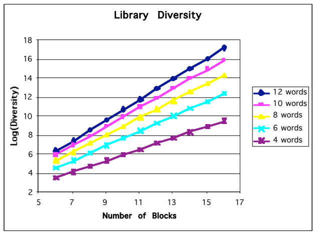

Figure 2

The range of possible sentences entailed by a word-block construction scheme is shown in Figure 1. For each block position in the sequence one word is chosen from the word set and synthesized on the growing DNA strand. Separate reaction vessels are used for each word in the block so that all word choices are utilized but only one is present on any particular strand. For example, the arrows indicate the trace which results in the sentence: word1A-word2D-word3A-word4B. A particular bead is drawn through a particular path in the set of possible word choices, but all possible paths are populated with beads, so all possible DNA sentences are synthesized. Each bead contains multiple copies of a single DNA sequence that can be synthesized by the well known mix-and-split synthesis scheme. Figure 2 shows the scaling of library diversity with increasing sentence length (block count) and increasing number of available words within each block. Diversity is calculated by raising the word count to the exponential power given by block count (i.e. diversity = [word count] $^{\text{block count}}$). As a simple example, to achieve a total diversity greater than one million with sentences containing 6 blocks (for example), one would require a set of 10 word choices per block.

Also, to achieve a total diversity 12^{14} with sentences containing 14 blocks (for example), also requires a set of 12 word choices per block. In designing a DNA-encoded database one must consider several important factors including the following. (i) The overall length of the oligonucleotide sequences used for matching is critical because sequence length directly affects the fidelity and melting temperature of DNA annealing. (ii) Hamming distance (or number of changes required to morph one sequence into another) is another critical consideration. One would like to maximize the Hamming distance between all possible pairs of encodings in the database in order to minimize near neighbor false-positive matching. One strategy for maintaining sequence distance is to assign block structures to the sequences with sets of allowed words (subsequences) defined for each block. (iii) Another important consideration is the choice of the words themselves and the grouping of words into sets for us in the blocks. Sentence length, desired library diversity, and word-pair distance constraints all affect the choices of words in the lexicon. The word design can be made by careful design of the word, lexicon, and database elements, as well as experimental tuning of annealing conditions such as temperature-ramp rate, pH, and buffer and salt concentrations. A useful tool in this task is the computer simulations of DNA hybridization known as BIND developed by Hartemink et al. (1998).

3.5 Additional Tagging Methods for the DNA Strands. Various sophisticated tagging techniques have been developed by the biotechnology industry for expression analysis and differential expression analysis.. These include the SAGE tagging of Genzyme Molecular Oncology, Inc. and the randomized tagging techniques of Lynx Therapeutics, Inc.

(i) ***Serial Analysis of Gene Expression (SAGE)*** is a technique developed by *Genzyme Molecular Oncology, Inc.,* for profiling the genes present in a population of mRNA. By the use of various restriction enzymes, SAGE generates, for each mRNA, a 10-base tag that usually uniquely identifies a given gene. In the usual SAGE protocol, the resulting SAGE tags are blunt-end ligated and the results are sequenced. The sequencing is faster than sequencing the entire expressed genes because the tags are much shorter than the actual mRNA they represent. Once sequencing is complete, one may lookup the tag sequences in a public database to find the corresponding gene. Using the sequence data and the current UniGene clusters, a computer processing stage determines the genes that have been expressed. SAGE can be used on any set of expressed genes and is not specialized to any particular set. This technology can be adapted for use as additional information tags appended to the DNA in our database.

(ii) ***Differential Expression Analysis*** is a technique developed by Lynx Therapeutics, Inc. for finding the difference in gene expression, e.g., between two distinct cell types. The randomized tagging techniques of Lynx Therapeutics, Inc. can be adapted to determine the difference between two DNA database subsets.

(iii) Hybrid Methods. One can modify these methods and extend them to apply to the tagged DNA strands of our database. This requires considerable changes in the protocols, due to unwanted hybridization that may occur due to the combination of synthetic tags with genomic DNA in our database strands. However, these modified methods can provide further powerful capabilities, e.g., the capability for fingerprinting (creating short DNA tags that are nearly unique IDs for longer DNA strands of the database), identification of expressed genes of selected DNA strands, and also the capability for differential expression analysis of distinct selected subsets of the Biomolecular Database.

3.6 Amplification and Reproduction of Biomolecular Databases. Once a Biomolecular Database is created, it is be important to be able to accurately replicate it, as they may be consumed in the course of their interrogation. Prudence suggests maintaining each database in an archive, and querying only daughter databases prepared from the archival forms. Since each database member is designed to bear a universal amplification (primer) sequence at the extreme 5'- and 3'-ends, database replication can be performed using PCR. Because the length of the DNA strands in the database might be quite substantial, including both biological DNA information and many flanking tag sequences, the ability to produce full-length amplicons with long templates is crucial to maintaining the fidelity of the databases. "Long accurate" PCR techniques (Taylor, Logan (1995); Taylor, Robinson (1998), using novel thermostable proofreading polymerase enzymes such as *Pfu*, are currently capable of amplifying loci of up to ~40 kb. While powerful, the database design should not be limited to this length by the method for database replication, and it may be easier to enable PCR to produce amplicons of somewhat longer length. One simply needs to enhance by a moderate multiple the (amplicon) length that can be reliably amplified.

Optimized choice of amplicons may be achieved by exploiting two principles: experimental design (Box (1978); Box (1987); Deming (1987) and combinatorial chemistry (Pirrung (1995; Pirrung (1997). Continuous variables that affect PCR reactions include temperatures of the initiation (hot start), annealing, extension, and dissociation steps and concentrations of buffer components, additives, nucleotides, primers, and template. These variables compose a multi-dimensional space. A pervasive challenge in science and technology is identifying specific values for each parameter affecting multivariable processes that result in globally optimum performance and avoid local maxima. Commercial software enables the design of experiments that much more reliably and quickly lead to the global optimum. Non-continuous variables that affect PCR reactions include the identity of the template, primers, and polymerase. An optimum combination of these molecules can be found only by systematic screening of each. For tractable numbers of combinations, all can be examined explicitly. When the diversity space expands beyond that domain, "indexing" techniques are available that permit the optimum performers to be identified even when in a mixture with lower performers (Pirrung (1996). A selection of variable length primers can be examined, including those incorporating modified bases (deazapurines, 2'-OMe RNA) that suppress primer consumption by dimerization. A selection of commercial polymerase enzyme systems can be examined, including MasterAmp™ Taq, ThermalAce,™ Advantage-Tth,™ AdvanTaq™ and KlenTaq™/Pfu. A selection of templates should be examined, including whole viral genomes, bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and the smallest yeast chromosome (225 kb). The analysis of the products of these reactions is challenging due to shearing of large DNA molecules by conventional sieving matrices. Pulsed-field gel electrophoresis (Cantor, Smith, Mathew (1988); Olson (1989), can therefore be used with amplicons of this size.

3.7 Associative Search in Biomolecular Databases

DNA-based associative search. Eric Baum (1995), first proposed the idea of using DNA annealing to do parallel associative search in large databases encoded as DNA strands. The idea is very appealing since it represents a natural way to execute a computational task in massively parallel fashion. Moreover, the required volume scales only linearly with the database size. However, there were further technical issues to be resolved. For example, the query may not be an exact match with any data in the database, but DNA annealing affinity methods work best for exact matches. (Reif and LaBean, (2000), described improved biotechnology methods to do associative search in DNA databases. These methods adapted some information processing techniques (Error-Correction and VQ Coding) to optimize input and output (I/O) to and from conventional media, and to refine the associative search from partial matches to exact matches.

(Reif, et al. 2000) developed and then experimentally tested (Reif et al. (2001), a method for executing associative searches in DNA databases of encoded images, and this method was tested using an artificially synthesized DNA database. Prior to that project, the idea of using DNA annealing to do parallel associative search in synthetic DNA databases had never been experimentally implemented. Reif et al. (2001), details a study involving the design, construction, and testing of large databases for the storage and retrieval of information within the nucleotide base sequences of artificial DNA molecules. The databases consisted of a large collection of single-stranded DNA molecules, which was immobilized on polymer beads. Each database strand carried a particular DNA sequence, which consisted of a number of sequence words drawn from a predetermined lexicon. They made a number of experimental databases of artificially synthesized DNA sequences designed for encoding digital data, scaled in increasing sizes. Each DNA strand of the database is single stranded, and encodes a number that provides the index to the database element. They used an extensive computer search for the design of our DNA word libraries, to insure significant Hamming distance between distinct words and allowing for annealing discrimination. They constructed their largest synthetic databases in two phases. In the first phase they constructed an initial DNA database by combinatorial, mix-and-split methods on plastic microbeads. This constituted by far the largest artificially constructed synthetic databases of this sort. The next phase was the development of a construction method for much larger synthetic databases by combining pairs of the synthetic database strands so as to square the size of the database to approximately 10^{15} distinct data elements (each represented redundantly by over 10 identical stands of DNA). Even with this with over 10-fold redundancy, the DNA database using this construction method is extremely compact, and requires only 10 milligrams of DNA.

Associative Search via PCR. PCR methods can be used for associative search queries in Biomolecular Databases (in particular, on the words of the tagged portions of the Biomolecular Database strands), using known and modified PCR techniques previously developed in [RLP+01]. That paper describes experiments for executing associative search queries within the above described synthetic DNA databases. Associative search queries were executed by hybridization of a database DNA strand with a complementary query strand. Discrimination in annealing experiments is enhanced by the library design, which guarantees a minimum Hamming distance between distinct sequences. In their initial annealing experiments for processing associative search queries, Reif et al. (2001), employed fluorescently labeled query strands and then performed separation of fluorescent versus non-fluorescent beads by Fluorescence Activated Cell Sorting (FACS). Reif et al. (2001), also experimentally tested variants of conventional PCR techniques for executing associative search queries: Reif et al. (2001), developed a PCR technique for associate search in the pair-wise constructed DNA database.

Analysis of Associative Search. Similar error analysis and experimental testing methods can be employed in our proposed generalizations of this prior work (Reif et al. (2001), to tagged genomic DNA. It would be informative to measure rates of various search errors including: false positives from near-neighbor mismatches, partial matches, and non-specific binding as well as false negatives from limit-of-detection problems. It is desirable to directly measure the limits of detection, and to measure the ability to retrieve rare sequences within databases of high strand diversity.

3.8 Logical Query Processing in Biomolecular Databases. Biochemical operations can be used to execute query operations on this Biomolecular Database, so as to retrieve subsets of the Biomolecular Database. Each of the information strands of the database encodes a sequence of data values $v_1,v_2,\ldots v_k$, where the i^th^ value v_i

ranges over a small finite domain D_i (e.g., D_i typically would range over 10 or less possible values, each encoded by a distinct fixed length DNA sequence) The retrieval can be specified by logical queries on the tags of the database as well as associative queries on the attached genomic DNA strands. The associative searches can be executed by recombinant DNA operations, e.g., variants of PCR combined with surface chemistry methods and/or solution based methods. The logical queries include the following: (i) SELECTION: select DNA strands of a given ID or cell type, and (ii) Logical SELECTION: execute logical queries that select those genomic DNA strands whose information strands satisfy a specified logical query formula, whose logical conjunctives include AND as well as OR. These logical conjunctives are applied to selective predicates of the form "Tag(i) = v", where Tag(i) is the ith portion of the information tag of a DNA strand of the database, and v is a fixed value over the domain D_i. (The Boolean NOT of a selective predicate of the form "Tag(i) = v" is not applied directly (since PCR and similar methods do not allow this) by instead applying the OR of selective predicates of the form "Tag(i) = u" for all possible u in D_i. that are not equal to v). These selection operations can be executed by the use of recombinant DNA operations, applying and improving on logical processing methods developed in the field of DNA computing. Furthermore, one can provide the additional operation of selective amplification of the DNA populations. If these amplification operations are also executed, the logical selection and amplification operations results in a test tube whose selected DNA is vastly amplified. After the amplification process is completed, the output strands should vastly predominate all other strands of the Biomolecular Database. Other database operations that can be to implemented by biochemical operations include database unions and limited joins (Reif, (1995).

Scalability of Our Query Processing. These operations can be executed in a scalable way. The required volume never grows significantly; the volume is a fixed linear function of the number of elements of the database. (the constant multiple here is the degree of redundancy that DNA strands are used to store database elements; we expect that one can allow between a few hundred and possibly as few as 10 DNA strands to encode a given database element) The number of required DNA hybridization steps grows only linearly with the size of the query formula. So the time for executing a query grows just linearly with the length of the query formula, which in practice is of very modest size (as compared to the size of the database, which can be enormous), say 20 or so variables. Hence the key time limitation is the time for DNA hybridization. But DNA hybridization time is nearly invariant of the size of the database even if the hybridization is execution on a an enormous number of DNA (up to extremely large database sizes, say 10^{15}). However, there are considerable technical challenges in the design of the protocols; for example biological data strands may be originally dsDNA while search protocols would function best with ssDNA (hence the protocols need to either form ssDNA or be modified appropriately). A key additional technical challenge in scaling the technology is the scale and number of the resulting molecular biology reactions, requiring many tedious laboratory steps, particularly in the case of extremely large database sizes. This can be addressed by subsequent automation. We discuss two distinct methods for logical query processing: the first uses primer-extension techniques on solid support previously developed (Pirrung et al. (2000), to solve SAT problems, and the other uses solution-based PCR amplification techniques. The second has more potential for scalability due to the fact that it is solution-based(so the chemistry operates in 3D) rather than constrained to a surface. (In both cases, one can apply DNA hybridization array technology for output of query results.)

Executing Queries into the Biomolecular Database via Primer-Extension Techniques on Solid Support. A number of DNA computing researchers have previously developed microarray methods for DNA-based computing, exploiting the high fidelity of the primer extension reaction to detect complementarity between primer libraries of all solutions to SAT problems and logical queries as templates (e.g., the work of Faulhammer et al. (2000); Liu et al. (2000), and also that of Pirrung et al. (2000), which improved the fidelity). Primer extension is a two-step process, involving first annealing of a template molecule to the primer, the efficiency of which is directly related to sequence complementarity throughout the primer/template complex. Second, a polymerase enzyme binds to the primer-template complex and adds a nucleotide or nucleotides complementary to the base X (see below), the first unpaired base at the 5'-end of the template, only when there is a perfect match in the last portion of base pairs of the primer-template complex. It is important to emphasize that while primer extension was in this case performed on a DNA microarray, the elementary step of a polymerase chain reaction (PCR) is also a primer extension process and thus is subject to the same stringent sequence requirements. The variables (primers) in the SAT computation of [Pir00] were composed of two portions, which can be considered the message (the last few bases "m" at the 3'-end of the oligonucleotide) and the address (a sequence of bases "a" at the 5'-end). With the base-4 encoding of DNA, a message sequence is capable of encoding 10 Boolean variables. For the experiment of [Pir00], all addresses were the same, as only one SAT problem was being addressed. However, this need not be the case. Using similar designs, it is possible to design up to approximately 20 blocks of distinct address sequences, that concatenated form the tag. Each of these blocks of distinct address sequences should exhibit no cross-hybridization under stringent conditions (Hamming distance at least 5), thereby enabling independent encoding and primer extension and therefore interrogation of up to 20 distinct attributes, each with up to 10 scalar values.

PRIMER 5'-aaaaaaaaaaaaaaaaaammmmm-3'

Figure 3. TEMPLATE ←3'-aaaaaaaaaaaaaaaaaammmmmX-5'→

Use of the primer extension method for Logical Queries into the Biomolecular Database is most efficient if performed in solution rather than on a microarray. This creates a challenge in product detection and identification. The following method enables both to be accomplished. An example is presented for the fate of one molecule, though it is appreciated that all molecules in the library are subjected to the same process in parallel. The database member is a DNA molecule that has been created by the methods described earlier, with a biological DNA sequence flanked by one or more created tag sequences, which are to be the templates in a primer extension reaction. The "bottom" strand is interrogated in this example. Primers with the following structure (shown in expanded form below the database element) are created to interrogate each tag. Complements to the address sequence in the tag/template are the same in each primer. Also common to each primer is a "barbed tail" in the form of a 5'-psoralen group. The irradiation of psoralens with long-wavelength UV is widely used to cross crosslink duplex DNA (Helene, Thuong (1991); Pieles (1989); Wellinger, Lucchini, Dammann, Sogo (1999). The message sequence must be unique to each variable value, meaning that up to 10 primers are prepared per variable. The primer is also designed to address a unique X base in the tag/template to be interrogated. The primer extension reaction is performed using a dideoxynucleotide terminator complementary to X and bearing a fluorescent dye with a unique and readily imaged emission spectrum. The dye color is specific to the variable, with the same dye/terminator being used for all interrogations of that variable. Multiple tags can be interrogated simultaneously because their dyes are different. The challenge at this stage is to read out the tags (based on the color(s) of the incorporated fluorescent dye(s)) in the context of the biological DNA. While the primer is still bound to the template, the psoralen is photochemically cross-linked to the bottom strand of the library member, preserving the color of the dye. The bottom strand can then be obtained in single-strand form, which is hybridized to a cDNA microarray. The color(s) of the array element complementary to the biological DNA identify the outcome of the queries of the tag sequences connected with it.

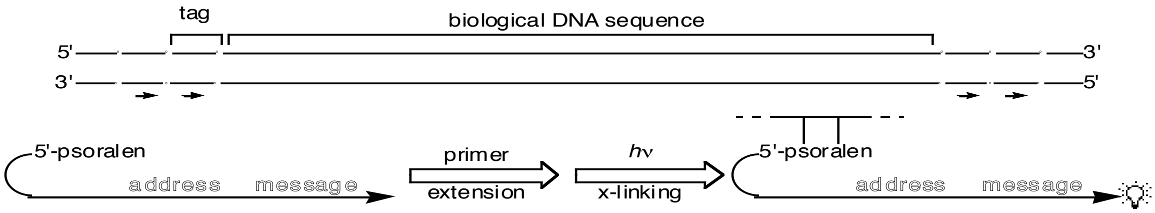

Figure 4.

This concept could be applied in a similar way to a sequential (nested) PCR process by omitting the terminators and psoralen and providing one primer for the top strand and one for the bottom strand in each PCR. The eventual production of a full-length amplicon is dependent on the complementarity of *each* of the primers (logical AND) with its cognate tag sequence. This approach lends well to the use of DNA hybridization array technology for output of query results, providing distinct special locations for distinct outputs.

Another approach for executing Boolean queries on a Biomolecular Database is to use the gel separation-based for SAT of Braich et al. (2002) which have succeeded in solving a 20 variable Boolean satisfiability problem. Although the queries would be executed on the tag portions of the DNA strands of the database, it is not clear how the efficiency of these separation methods would be affected by the genomic portion of the DNA strands in the databases.

Executing Queries in the Biomolecular Database via PCR Amplification Techniques: Another approach for Logical Query Processing is to use a variant of PCR amplification. The goal of this query processing is to selectively amplify only those DNA sequences (the *output strands*) whose information tags satisfy a given logical query. After the amplification process is completed, the output strands would vastly predominate all other strands of the Biomolecular Database.

Initialization Before Logical Query Processing: First, operations are executed that generate, from each DNA strand in the database, a new strand containing a concatenation of multiple copies of the Watson-Crick complement of the original strand. This can be done by a known sequence of routine recombinant DNA operations known as rolling circle replication (Lizardi (1998). This begins by a circularizion of each strand on the databse, and then a primer extension reaction on the circularized strand that repeatedly replicates the complement of the DNA strand to form a repeated sequence, followed by a denature and separation of the result. The length of the resulting DNA strands is predictable (via the time duration and various parameters including temperature) only to a degree, but it is predictable enough to allow us to construct strands that *expect* to have at least a given repeat length (as required by the below protocol). Recall that we have assumed that each database DNA strand is also redundantly represented by a number (ranging from up to a few thousand down to as few as 10) of identical DNA strands. This redundancy aids us here; since this initialization procedure results in Biomolecular Database where most of the redundant strands are lengthened by at least the given multiplying factor.

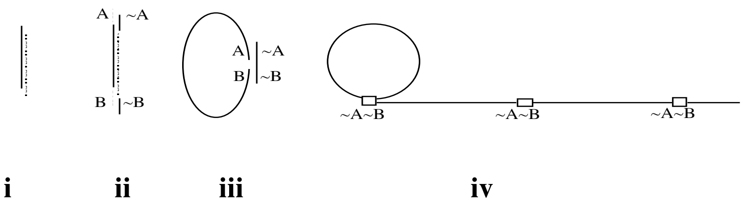

Figure 5

Figure 5 illustrates a scheme for processing genomic DNA into this database format which might include the following steps: (i). Cleave dsDNA into manageable pieces; (ii) Append prefixes to both ends of both strands. Heat denature dsDNA. Anneal to circularizing oligo. (iii) Ligate ssDNA circles. (iv). DNA polymerase reaction with circular templates to produce linear ssDNA containing multiple concatenated database entries. **Note:** The process of converting the DNA into database format may have unintended effects on the representation of entries in the database due to uneven amplification. Artificial bias may take the form of variations in the number of copies present on the average strand (distribution of strand lengths) or differences in the number of strands present for a given database entry. These protocols need to be optimized to take into account these possible affects.

Multiple copies of the database entry are required on a single ssDNA strand so that when Boolean variables recorded in the prefixes (A & B in figure above) are queried by primer binding and PCR, information recorded farther out toward the ends of the strand is not lost by failure to be copied (PCR only amplifies sequence physically between the primer binding sites). The goal is to keep at least a few copies of the prefix information internally within the database strands so that information is not lost to subsequent rounds of query.

Logical Query Processing Using Repeated PCR Operations: We assume the logical query is presented as the logical AND of a list of K logical clauses (each clause needs to be satisfied), where each clause consists of the logical OR of a list of literals (the literals can be Boolean variables or their negation), one of which needs to be satisfied. Each clause C in the formula is processed in turn, selectively amplifying only those DNA strands whose Boolean variables satisfy at least one literal of that given clause. To do, one adds PCR primers encoding the literals of that clause C and their Watson-Crick complements. Then a series of primer-extension reactions are executed that replicate only those DNA strands (or their Watson-Crick complements) that have subsequences that encode one of the literals of clause C. This process, applied as a series of PCR cycles, thus amplifies only those DNA strands whose Boolean variables satisfy at least one literal of that given clause, so that they vastly predominate all other strands of the Biomolecular Database. (Technical Note: On each cycle, the amplified stands have loss of the material prefixing the primer's location, but the initial step of concatenating to each DNA strand in the database multiple copies of the strand insure that that is not a problem.) After the process is completed for each of the clauses in turn, the output strands that satisfy all the clauses would vastly predominate all other strands of the Biomolecular Database. This method for processing a logical query in the database is exquisitely sensitive: to get a result, one requires that the initial database have no more than 10 identical strands of DNA that satisfy the query. Again, DNA hybridization array technology can be used for output of query results, providing distinct special locations for distinct outputs.

Scalability: As discussed above, our query processing is executed with vast molecular-level parallelism by a sequence of biochemical reactions requiring time that remains nearly invariant of the size of the database up to extremely large database sizes (e..g, up to 10^{15}). This is because the key limitation is the time for DNA hybridization, which is done in parallel for all the DNA.

3.9 Management of Errors. The logical and associative searches used to select specific molecules and sets of molecules from Biomolecular Databases are not 100% specific or effective. There may be several different kinds of errors: false negatives (appropriate DNA strands are present, but not selected, either because of lack of sensitivity or depletion of the relevant sequences from the database), false positives (inappropriate DNA strands are selected along with desired strands), errors based on degradation of the Biomolecular Database contents, and errors resulting from poorly designed queries, based on incomplete understanding of complex biological parameters. These kinds of errors can affect the results of the applications described here. For example, false negative errors can prevent finding existing individuals with the desired genetic variant. This is less serious than a false positive result, which could lead to sending a non-resistant individual into a contaminated area, under the false belief that he is genetically protected from a biological agent. A similar type or error could arise from database degradation (as for example, from repeated error-prone duplication of the database). This type of error can be easily eliminated by follow-up, confirmatory screening of that single individual's DNA. In general, it would be best to use the Biomolecular Database for very powerful and rapid selections based on genetic information, but then to confirm all results on individual DNA samples. This would require maintenance of individual stocks of DNA for each individual. This is a relatively large task, but well within current technology. A LIMS (laboratory information maintenance software) system and robotic liquid handling capacity are a must for this type of storage. There could also be errors of magnitude. These errors result from preferential amplification of one DNA strand over another. This kind of error is particularly troublesome for it would skew allele frequencies. It may be necessary for us to monitor the frequency and extent of such errors and develop Boolean search strategies that minimize them. The final type of error, based on the incomplete understanding of the human genome, can only be rectified by continued research in other fields. This type of error could result from incomplete knowledge of the way in which genetic variants are distributed among different racial and ethnic groups, For example, the well-described *ccr5* variant that prevents HIV infection has been detected to date only among white males. If one were to select for non-existent African American females expressing this variant, one might well obtain a small number of false positives. This type of error could also arise from mistaken assumptions. A given genetic variant might protect Hispanic females from infection by a given biologic agent, but oriental males carrying the same variant might be fully susceptible to infection because of another independent genetic variant.

3.10 Computer Simulations. Reif et al. (2001), has made computer simulations of their methods for DNA-based associative search. They constructed computer software (viewable on the web) that provide a simulation of the entire experimental process, including the conversion of this attribute database into a DNA database using DNA chips, the PCR method for associative search in this DNA database (using a software simulation of the kinetics of DNA hybridization, and finally the conversion of the result of this query (using extensions of techniques described in Reif,et al. (2000), into conventional media by use of a DNA expression array. Our computer simulation software to the above described query processing provide a basis for future software to simulate and optimize experimental protocols for query processing in Biomolecular Database Systems.

4. Applying our Biomolecular Database System to execute Genomic Processing. There is tremendous potential to apply Biomolecular Databases to the solution of a number of biological problems. The huge amount of data provided by the nearing completion of the sequencing of the human genome has outstripped many conventional methods for DNA analysis.

Genomic Processing Applications. We now discuss applications of such a Biomolecular Database System to provide key genomic processing capabilities, as listed below. Three basic kinds of applications are discussed, which demonstrate different ways in which the massive parallelism of Biomolecular Databases can be used: (1) Rapid identification of individuals either susceptible to or resistant to chemical or biological agents. We describe the selection of a group of DNA molecules based on a common property, then use the information tags to identify the individuals selected. (2) Large-scale gene expression profiling using Biomolecular Databases. Expressed genes from multiple tissues are represented in a Biomolecular Database, from which they can be selected individually or in groups for subsequent expression analysis. (3) High throughput screening of candidate genes to optimize genetic association analysis for complex diseases such as heart disease or Parkinson's disease. Pools of individuals are selected through use of the information tags appended to each DNA molecule in the database. The pools so selected are then subjected to genetic analysis.

Here we describe in detail these three applications that concern genomic information processing, and constitute important genomic processing applications of Biomolecular Database systems for medical science:

4.1 Rapid identification of all individuals possessing a specific known genotype. *For example, a single known genotype can confer properties making the individual either susceptible to or resistant to a particular chemical or biological agent found in the environment.* It is certainly possible with existing biotechnology (e.g., hybridization experiments) to screen individuals for a given genotype. This is done one individual at a time, and is thus a relatively slow process. In addition, the cost of traditional genotyping of an individual ranges from $300 to over $1,000. (At least one genotype databank has executed genotyping of approximately 1,000 individuals with considerable expense and time; however for experimental purposes, that databank provides examples of previous executed individual genotyping at a cost of $0.50-$1.00 per sample.) Clearly an effort to screen a large number of individuals (say 1,000,000) would be slow and very expensive. In contrast, the methodology described here is a *selection* for individuals with a certain genotype rather than a screen. It is correspondingly faster and less expensive. There is currently no available methodology for selection of specific genotypes. Many drugs that are very effective in treating disease are quite toxic to a small portion of the population. Currently, such drugs are removed from the market to avoid these rare but fatal adverse reactions. Such an approach is very costly from the standpoint of untreated disease. The removal of drugs from the marketplace because of rare fatal reactions is very costly in terms of untreated individuals, as well as the money spent on bringing those drugs to the market in the first place. Improved methods for identification of individuals at risk for adverse reactions would eliminate this cost. The capability of screening large numbers of individuals for a given genotype could also avoid a tremendous potential loss of life in the event of the battlefield release of biological weapon or chemical agent.

As another example of a clinical application, one can construct a Biomolecular Database made from blood samples of people suffering from Alzheimer's disease and their families, with the goal of finding genes that may increase people's risk of contracting Alzheimer's disease. The information tags can used to select specific groups of molecules from this database. These molecules, which come from people with similar clinical symptoms, can then be used to test a large number of possible Alzheimer's disease genes. Genes that give promising results can then be tested on the large number of individual samples from which the Biomolecular Database was made. The advantage of this approach is that it allows very efficient use of the limited DNA samples, and it is a good way to look at lots of different combinations of clinical features.

4.2 Large-scale gene expression profiling using Biomolecular Databases. *For example, one may wish to determine the entire set of genes expressed by a particular cell type for a population of individuals who suffered debilitating effects due to a (perhaps unknown) chemical or biological agent. This may allow us to determine if there is a single or small number of genotypes that characterize susceptibility to that agent, over that population.* Gene expression profiling is a labor intensive and slow process. Conventional methods are as follows: *(a) cDNA Hybridization Arrays* (These are 2D arrays of DNA spotted onto a solid support in an addressable way such that the spatial location of a spot identifies to sequence of the DNA bound there . The input cDNA are labeled with a fluorescent dye with a unique and readily imaged emission spectrum. After annealing on this array, the fluorescent cDNA provide a visual readout of the expression.) (b) *SAGE libraries* (These are prepared by extraction from cDNA of very short tag sequences which characterize the expressed gene, followed by concatenation of a number of these tag sequences together for sequencing. Then computer software (using prior information on the relation between these tags and the original expressed cDNA) is used to determine which genes are being expressed.) Gene expression profiling can require the development of new cDNA hybridization arrays, or the construction and s*equencing of SAGE libraries*. The methods for parallel analysis of large numbers of samples described here would streamline this process. In addition, the readout of SAGE data by microarray hybridization would result in significant savings of time and money as compared to the standard method of sequencing SAGE libraries. It would enhance the understanding of acute responses to biologic agents as well as the understanding of complex disease processes.

As another example of a clinical application, one can construct a Biomolecular Database made from a large group of healthy people, with the goal of finding people who are naturally resistant to certain germs, or who respond in certain ways to prescription drugs. One can study the selection of DNA strands from this mixture that have a specific sequence change in a specific gene that is known to change a person's resistance to germs or their response to drugs. Once these strands are isolated, the information tags would be examined to identify the people who have that change in their genes. This would be an extremely useful way to identify people who could have a bad reaction to a drug that is commonly used to treat disease. It could also be very useful in discovering people who are resistant to diseases, either naturally occurring or released during germ warfare.

4.3 Streamlining identification of susceptibility genes: high throughput screening of candidate genes to optimize genetic association analysis for complex diseases. *For example, consider the problem of genomic characterization of those individuals who first were infected by a biological agent, and then died. The death may often have been due to complications involving additional "complex" diseases such as heart disease. Hence mortality resulting from a chemical or biological agent attack may often have been due to* *complications involving a preexisting disease such as heart disease. So mortality can often only be predicted by considering both the individual's susceptibility to that agent, and well as their susceptibility to various preexisting "complex" diseases. For many "complex" diseases, the susceptibility often depends on a number of single nucleotide polymorphisms (SNPs) in the human genome.* Research into the genetic causes of complex disease is currently very expensive, and progress is slow. Complex diseases are quite common, affecting large proportions of the population. Delays in understanding the genetic basis of these diseases slow the development of improved treatments at a significant financial and human cost. In the last 2-3 years, there have been several large-scale efforts to identify single nucleotide polymorphisms in the human genome. The SNP consortium, a non-profit foundation composed of the Wellcome Trust and 11 pharmaceutical and technological companies, has agreed to deposit all SNPs they discover to public databases such as dbSNP, the SNP database maintained by the National Center for Biotechnology Information (NCBI). In the past 3 years, the number of entries in this database has increase from several thousand to over two million. This sudden increase in the number of polymorphic markers has completely overwhelmed current methods for SNP genotyping and high-throughput screening. It has also become apparent that the incidence of single nucleotide polymorphisms varies widely from one region of the genome to another, and large numbers of SNPs must be screened to analyze each candidate gene. Even with unlimited funds and capacity for genotyping, serious challenges to the family based association screening would remain, because the individual screening of a large number of SNPs would quickly exhaust the amount of DNA that can be easily obtained from a single individual. This problem is compounded by the sample cost of preparing pools of DNA from multiple individuals by simple mixing: once samples are mixed, they cannot be separated again, and leftover, pooled DNA is wasted. Indexing of a Biomolecular Database can be of significant assistance in this regard. Large numbers of different groups of individuals can be selected from the Biomolecular Database by logical queries on the information tags. These pools can be used for allelic frequency determinations, and any remaining DNA can be added back to the remaining database**.**

As another example of a clinical application, one can use Biomolecular Databases to help discover what genes are turned on in a specific tissue on the body. Genes that are needed in the brain may not be expressed in the muscles, and genes needed in the muscles may not be needed in the liver. For this reason, measuring what genes are turned on in a specific tissue can help us understand what the possible functions of those genes might be. Biomolecular Databases would provide increased efficiency for these approaches.

4.4 Further Applications. The applications described above could be of critical value to the US in the case of a terrorist release of a biological (or chemical agent), as in the following brief scenario. A biological agent is released by the terrorist group into a US city or other populated area. The city is evacuated, but it becomes necessary to traverse a potentially contaminated area, or to revisit a known contaminated area. Clearly, any personnel sent into this area, even with protective gear, are at risk for infection. A Biomolecular Database query would be initiated to identify personnel who posses a known genetic variation that prevents or mitigates infection. The personnel sent into the contaminated area could then be selected from the list of genetically resistant personnel.

As an alternative anti-terrorist application, suppose a large population (e.g., of a city) have been exposed to a given biological or chemical agent. It then becomes apparent that a subgroup of individuals requires significantly more aggressive medical therapy to survive, but for logistical reasons, such aggressive therapy cannot be provided to ALL exposed individuals. Stored DNA from resistant and susceptible individuals can be used to determine status of specific groups of genetic markers as described in application C (markers are chosen based on biological and medical inferences). In this way, a series of markers diagnostic for increased susceptibility can be identified. This type of analysis is called class discovery, and has been applied to the treatment of breast cancer, leukemia, and other disorders. However, the use of Biomolecular Databases can greatly streamline this work. Once diagnostic markers have been identified, the techniques worked out in application 4.3 can identify individuals in need of more aggressive care.

5. Discussion and Conclusions

We have described Biomolecular Databases constructed from DNA for rapid genetic analysis of large populations of individuals and complex diseases involving multiple genetic loci. They may improve on conventional methods in size of database and speed of search with the Biomolecular Databases system.

5.1 Comparison with Biomolecular Computing Methods for SAT Problems. As described above, these selection operations can be executed by the use of recombinant DNA operations, using logical processing methods developed in the field of DNA computing. The methods used in DNA computing to solve combinatorial search problems such as the Boolean satisfiability (SAT) problem have the disadvantage that they require a volume that scales exponentially with the size of the problem (number of Boolean variables). This is because the search space of possible Boolean variable assignments scales exponentially. In contrast, the logical queries are executed only on the information tags of the existing database, so the volume therefore only scales linearly with the number of strands of the Biomolecular Database.

5.2 The Key Advantages of Biomolecular Databases appear to be:

(i) Bypassing of Conventional Impasses: In particular, the avoidance of sequencing for conversion from DNA (genomic DNA and transcribed RNA) to digital media.

(ii) Ultra-Compact Storage Media: the extreme compactness and portability of the storage media: A pedabit of information can be stored (with 10-fold redundancy) in less than a few milligrams of dehydrated DNA, or when hydrated may be stored in a few milliliters of solution. A Biomolecular Database is capable of containing the DNA of 1,000,000 individuals (6 pedabits of information) in a volume the size of a conventional test tube.

(iii) Massive Molecular Parallelism: Although a query may require a number of minutes, it is operated on vast numbers of data items (DNA strands), implying a processing power of vast molecular parallelism with at least a few hundred teraflops. The operations can operate in parallel on an entire population of DNA.

(iv) Scalability: The technology requires volume that scales linearly with the size of the database, and query time that remains nearly constant up to extremely large database sizes.

(v) Limitations. The Biomolecular Database technology is limited to applications of a biological nature (where the data is DNA or easily convertible to DNA), and the operations are limited to logical queries in the Biomolecular Database, associative searches, and some essential database operations. It is not intended that the technology compete in any direct way with conventional high performance computers. Instead, the objective is to bypass conventional bioinformatics methodology by processing biological material (genomic DNA and transcribed RNA) in "wet" media, rather than digital media.

5.3 Scalability of Biomolecular Databases Systems. The Key Parameters of Biomolecular Database are: (a) N= number of distinct elements of Biomolecular Database, (b) v= number of variables (each ranging over 10 possible values) used in queries, (c) k=number of individuals in application studies.

For our practical genomic applications of Biomolecular Databases to be fully realized in practice: (i) the database size N should to grow to extremely large values (with a long term goal of approximately 10^{15}), (ii) but for these applications the number of variables v needs only to grow to moderately small constant values (with a long term goal of approximately v=14), since for the genomic applications considered, only a limited number of values need be recorded in the information tag per database element. The relative difficulty of obtaining human genomic material limits the number of individuals k in possible studies to approximately 1000, which is the size of the largest genomic database we are of aware of for which one can legally obtain samples of genomic DNA. However, this figure k=1000 is by no means a limit on the capability of the Biomolecular Database technology, even within the 3 to 5 time period of the proposal funding. In particular, these genomic databases are quickly growing in size, and in may be projected to grow a number of multiples in 5 years. Furthermore, military sources of human genomic DNA may be obtainable, providing alternate routes to obtain the samples of genomic DNA required in large scale studies.

5.4 Acknowledgements. This work has been supported by grants from NSF to J.H.R. (EIA-00-86015, EIA-0218376, EIA-0218359) and DARPA/AFSOR to J.H.R. (F30602-01-2-0561).

REFERENCES

[Ad94] Adleman, L., Molecular Computation of Solution to Combinatorial Problems, Science, 266, 1021--7, (1994).

[AED+00] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC et al (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769), 503-511

[ABN+99] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96(12), 6745-6750

[Arn90] Arnheim, N.; Li, H. H.; Cui, X. F. PCR analysis of DNA sequences in single cells: single sperm gene mapping and genetic disease diagnosis. *Genomics* **8**, 415-9 (1990).

[BCGT96] Bach, E., A. Condon, E. Glaser, and C. Tanguay, Improved Models and Algorithms for DNA Computation, Proc. 11th Annual IEEE Conference on Computational Complexity, J. Computer and System Sciences, (1996).

[BBB+01] Bancroft, C., Bowler, T., Bloom, B., Clelland, C.T.. Long-term Storage of Information in DNA. Science, Vol. 293, No. 5536, pp. 1763-1765 (Sept. 2001).

[B95] Baum, E. B., How to build an associative memory vastly larger than the brain, Science, pp 583-585, April 28, 1995.

[B96] Baum, E. B. DNA Sequences Useful for Computation, 2nd Annual DIMACS Meeting on DNA Based Computers, Princeton University, June 1996.

[Box87a] Box, G. E. P. Empirical model-building and response surfaces. Wiley, 1987.

[Box87b] Box, G. E. P. Statistics for experimenters: an introduction to design, data analysis, and model building. Wiley, 1978.

[BCJ+02] Braich, R. S., Chelyapov, N., Johnson, C.,

Rothemund, P. W. K., Adleman, L., Solution of a 20-Variable 3-SAT Problem on a DNA Computer, Science, Vol. 296, Issue 5567, pp. 499-502, (April 19, 2002).

[CSM88] Cantor, C. R.; Smith, C. L.; Mathew, M. K. Pulsed-field gel electrophoresis of very large DNA molecules. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 287-304 (1988).

[CDW+03] C.J. Chen, R. Deaton, Y. Wang, A DNA-Based Memory with InVitro Learning and Associative Recall, DNA Computing: 9th International Workshop on DNA Based Computers (DNA9), Madison, Wisconsin, June 1-4, 2003, (Edited by Y. Chen and J. H. Reif), Lecture Notes in Computer Science, Springer-Verlag, New York, (to appear 2003).

[CSW+01] Clayton SJ, Scott FM, Walker J, Callaghan K, Haque K, Liloglou T, Xinarianos G. et al, K-ras point mutation detection in lung cancer: comparison of two approaches to somatic mutation detection using ARMS allele-specific amplification. Clinical Chemistry

[CSR+94] Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, Jr., Rimmler JB et al (1994) Protective effect of apolipoprotein e type 2 allele for late onset Alzheimer disease. Nature Genetics 7, 180-184

[CSS+93] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261(5123), 921-923

[DMGFS96] Deaton, R., R.C. Murphy, M. Garzon, D.R. Franceschetti, and S.E. Stevens, Jr., Good encodings for DNA-based solutions to combinatorial problems, Proceedings of the 2nd Annual DIMACS Meeting on DNA Based Computers, June 1996.

[DMRGF+97] Deaton, R., R.C. Murphy, J.A. Rose, M. Garzon, D.R. Franceschetti, and S.E. Stevens, Jr., A DNA Based Implementation of an Evolutionary Search for Good Encodings for DNA Computation, ICEC'97 Special Session on DNA Based Computation,

Indiana, April 1997.

[Dem87] Deming, S. N. Experimental design: a chemometric approach. Elsevier Science, 1987.

[DHM+00] DeRisi J, van den Hazel B, Marc P, Balzi E, Brown P, Jacq C, Goffeau A (2000) Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. Febs Letters 470(2), 156-160

[FCL+00] Faulhammer, D., Cukras, A. R., Lipton, R. J. and L. F. , "Molecular Computation: RNA Solutions to Chess Problems". Proc. Natl. Acad. Sci. USA. 97:1385-1389. (2000).

[FTCSC97] Frutos, A.G., A.J. Thiel, A.E. Condon, L.M. Smith, R.M. Corn, DNA Computing at Surfaces: 4 Base Mismatch Word Design, 3rd DIMACS Meeting on DNA Based Computers}, Univ. of Penns., (June, 1997).

[GDNMF97] Garzon, M., R. Deaton, P. Neathery, R.C. Murphy, D.R. Franceschetti, S.E. Stevens Jr., On the Encoding Problem for DNA Computing, 3rd DIMACS Meeting on DNA Based Computers, Univ. of Penns., (June, 1997).

[GNB+03] M. Garzon, A. Neel, K. Bobba, Efficiency and Reliability of Semantic Retrieval in DNA-Based Memories, DNA Computing: 9th International Workshop on DNA Based Computers (DNA9), Madison, Wisconsin, June 1-4, 2003, (Edited by Y. Chen and J. H. Reif), Lecture Notes in Computer Science, Springer-Verlag, New York, (to appear 2003).

[GLR00] Gehani, A., T. H. LaBean, and J.H. Reif, DNA-based Cryptography, 5th DIMACS Workshop on DNA Based Computers, MIT, June, 1999. DNA Based Computers, V, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, (ed. E. Winfree), American Mathematical Society, 2000.

[GR99] A. Gehani and J.H. Reif, Microflow Bio-Molecular Computation, 4th DIMACS Workshop on DNA Based Computers, University of Pennsylvania, June, 1998. DNA Based Computers, IV, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, (ed. H. Rubin), American Mathematical Society, 1999. Also appeared in a special issue of Biosystems, Journal of Biological and Informational Processing Sciences, Vol. 52, Nos. 1-3, (ed. By L. Kari, H. Rubin, and D. H. Wood), pp 197-216, (1999).

[GFBCL+96] Gray, J. M. , T. G. Frutos, A.M. Berman, A.E. Condon, M.G. Lagally, L.M. Smith, R.M. Corn, Reducing Errors in DNA Computing by Appropriate Word Design, University of Wisconsin, Department of Chemistry, October 9, 1996.

[HGL98] Hartemink, A., David Gifford, J. Khodor,, Automated constraint-based nucleotide sequence selection for DNA computation, 4th Int. Meeting on DNA-Based Computing, Baltimore, Penns., (June, 1998).

[Hel91] Helene, C.; Thuong, N. T. Design of bifunctional oligonucleotide intercalator conjugates as inhibitors of gene expression. *Nucleic Acids Symp. Ser.* 133-7 (1991).

[KYK+02] S. Kashiwamura, M. Yamamoto, A. Kameda, Toshikazu Shiba, Azuma Ohuchi, Hierarchical DNA Memory Based on Nested PCR. DNA Computing: 8th International Workshop on DNA Based Computers (DNA8), Sapporo, Japan, June 10-13, 2002, (Edited by Masami Hagiya and Azuma Ohuchi), Lecture Notes in Computer Science, No. 2568, Springer-Verlag, New York, (2003), pages 112-123.

[KCL96] Kaplan, P., G. Cecchi, and A. Libchaber, DNA based molecular computation: Template-template interactions in PCR, The 2nd Annual Workshop on DNA Based Computers, American Mathematical Society, (1996).

[LCA90] Li, H. H.; Cui, X. F.; Arnheim, N. Analysis of DNA sequences in individual gametes: application to human genetic mapping. *Prog. Clin. Biol. Res.* **340C**, 207-11 (1990).

[Li96] R. J. Lipton, DNA computations can have global memory, DNA Based Computers II (Editors: Laura Landweber and Eric Baum), DIMACS Series in Math. and CS., Vol. 44, AMS, (1996), pp. 259-266

[LFC+00] Liu, Q., Liman W., A. G. Frutos, A. E. Condon, R. M. Corn, and L. M. , "DNA Computing on Surfaces," Nature, **403** 175-179 (2000)

[Liz98] Lizardi, p. et al. Mutant detection and single molecule counting using isothermal rolling circle replication, Nature Genetics, Vol. 19, pp. 225-232 (1998).

[M96] Mir, K.U. A Restricted Genetic Alphabet for DNA Computing, 2nd Annual DIMACS Meeting on DNA Based Computers, Princeton University, June 1996.

[NSK+00] Niculescu AB, Segal DS, Kuczenski R, Barrett T, Hauger RL, Kelsoe JR (2000) Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. Physiol Genomics 4(1), 83-91

[Ols89] Olson, M. V. Separation of large DNA molecules by pulsed-field gel electrophoresis. A review of the basic phenomenology. *J. Chromatogr.* **470**, 377-83 (1989).

[PSE+00] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR et al (2000) Molecular portraits of human breast tumours. Nature 406(6797), 747-752.

[Pie89] Pieles, U.; Englisch, U. Psoralen covalently linked to oligodeoxyribonucleotides: synthesis, sequence specific recognition of DNA and photo-cross-linking to pyrimidine residues of DNA. *Nucleic Acids Res*. **17**, 285-99 (1989).

[Pir95] Pirrung, M. C. Combinatorial libraries. Chemistry meets Darwin, *Chemtracts: Org. Chem*., **8**, 5 (1995).

[Pir96] Pirrung, M. C.; Chau, J. H.-L. Chen, J. Indexed combinatorial libraries: non-oligomeric chemical diversity for the discovery of novel enzyme inhibitors, in Combinatorial Chemistry: A High-Tech Search for New Drug Candidates, Wilson, S. R. and Murphy, R., Eds., John Wiley & Sons, NY, (1996).

[Pir97] Pirrung, M. C. Spatially-addressable combinatorial libraries, *Chem. Rev*. **97**, 473 (1997).

[Pir00] Pirrung, M. C.; Connors, R. V. Montague-Smith, M. P. Odenbaugh, A. L. Walcott, N. G. Tollett, J. J. The arrayed primer extension method for DNA microchip analysis. Molecular computation of satisfaction problems, *J. Am. Chem. Soc.,* **122**, 1873 (2000).

[Pir01] Pirrung, M. C.; Zhao, X. Harris, S. V. A universal, photocleavable, DNA base: nitropiperonyl 2'-deoxyriboside (dP*), *J. Org. Chem*., **66**, 2067 (2001).

[QOR+98] Quillent C, Oberlin E, Braun J, Rousset D, Gonzalez-Canali G, Metais P, Montagnier L et al (1998) HIV-1-resistance phenotype conferred by combination of two separate inherited mutations of CCR5 gene. Lancet 351(9095), 14-18

[R98] Reif, J.H. Paradigms for Biomolecular Computation, First International Conference on Unconventional Models of Computation, Auckland, New Zealand, January 1998. Published in Unconventional Models of Computation, edited by C.S. Calude, J. Casti, and M.J. Dinneen, Springer Publishers, Jan. 1998, pp 72-93.

[R99a] J.H. Reif, Parallel Molecular Computation: Models and Simulations. Proceedings: 7th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA'95) Santa Barbara, CA, July 1995, pp. 213-223. Published in Algorithmica, special issue on Computational Biology, Vol. 25, No. 2, 142-176, 1999a.

[Re02a] J. H. Reif, The Emergence of the Discipline of Biomolecular Computation in the US, invited paper to the special issue on Biomolecular Computing, New Generation Computing, edited by Masami Hagiya, Masayuki Yamamura, and Tom Head, 2002.

[Re02b] Reif, J.H. Perspectives: Successes and Challenges, Science, 296: 478-479, April 19, 2002.

[RL00] John H. Reif and Thomas H. LaBean, Computationally Inspired Biotechnologies: Improved DNA Synthesis and Associative Search Using Error-Correcting Codes and Vector-Quantization, DNA Computing: 6th International Workshop on DNA-Based Computers(DNA6), Leiden, The Netherlands, June 13-17, 2000. Lecture Notes in Computer Science, Springer-Verlag, New York, Vol. 2054, (2001), pp. 145-172.

[RLP+01] John H. Reif, Thomas H. LaBean, Michael Pirrung, Vipul S. Rana, Bo Guo, Carl Kingsford, and Gene S. Wickham, Experimental Construction of Very Large Scale DNA Databases with Associative Search Capability, DNA Computing: 7th Int. Workshop on DNA-Based Computers (DNA7), Tampa, FL, June 10-13, 2001. Lecture Notes in Comp. Sci., (Edited by N. Jonoska and N.C. Seeman), Springer-Verlag, New York, Vol. 2340, (2002), pp. 231-247.

[RM96] Risch N, Merikangas K (1996) The future of genetic studies of complex human disorders. Science 273(5281), 1516-1517

[RS87] B.H. Robinson and N.C. Seeman, The Design of a Biochip: A Self-Assembling Molecular-Scale Memory Device. Prot. Eng. 1, 295-300 (1987).

[RWB+98] Roweis, S., Winfree, E., Burgoyne, R., Chelyapov, N.V., Goodman, M.F., Rothemund, P.W.K., Adleman, L.M., A Sticker-Based Model for DNA Computation, Journal of Computational Biology 5, 615-629 (1998).

[SNK+00] A. Suyama, N. Nishida, K. Kurata, K. Omagari, Gene expression analysis by DNA computing, Currents in Computational Molecular Biology (S. Miyano, R. Shamir, T. Takagi eds.), 12-13 (2000).

[SS00] Y. Sakakibara, A. Suyama, Intelligent DNA chips: Logical operation of gene expression profiles on DNA computers, Genome Informatics 11, 33-42 (2000).

[Sza01] Szatmari, I.; Aradi, J. Telomeric repeat amplification, without shortening or lengthening of the telomerase products: a method to analyze the processivity of telomerase enzyme. *Nucleic Acids Res*. **29**, E3 (2001).

[TL95] Taylor, G. R.; Logan, W. P. The polymerase chain reaction: new variations on an old theme. *Curr. Opin. Biotechnol*. **6**, 24-9 (1995).

[TR98] Taylor, G. R.; Robinson, P. The polymerase chain reaction: from functional genomics to high-school practical classes. *Curr. Opin. Biotechnol* **9**, 35-42 (1998).

[WLD+99] Wellinger, R. E.; Lucchini, R.; Dammann, R.; Sogo, J. M. In vivo mapping of nucleosomes using psoralen-DNA crosslinking and primer extension. *Methods Mol. Biol*. **119**, 161-73 (1999).

[Win98] Winfree, E., Whiplash PCR for O(1) Computing, In Kari, L., Rubin, H., & Wood, D.H., (eds.), Proceedings of the 4th DIMACS Meeting on DNA Based Computers, held at the University of Pennsylvania, June 16-19, 1998.

[ZCS+92] Zhang, L.; Cui, X.; Schmitt, K.; Hubert, R.; Navidi, W.; Arnheim, N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci., USA* **89**, 5847-51 (1992).

[ZGM+00] Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E et al (2000) Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. Genes Dev 14(8), 981-993.