# Experimental Construction of Very Large Scale DNA Databases with Associative Search Capability

John H. Reif[1,ϕ], Thomas H. LaBean[1], Michael Pirrung[2],
Vipul S. Rana[2], Bo Guo[1], Carl Kingsford[1], and Gene S. Wickham[1]
Departments of Computer Science[1] and Chemistry[2], Duke University, Durham, NC 27708

**Summary**. DNA has the theoretical capability of storing vast databases in a very compact volume, for example, a gram of DNA can store 4.2 x $10^{21}$ bits of information. Subsequently, encoded data can be retrieved by associative search queries. However, until now no large scale experiments have verified this. We describe the experimental creation of very large databases of artificially synthesized DNA sequences designed for encoding digital data. A database, or library, consists of sequences of single-stranded DNA, each sequence encodes a number which provides the index to the database element. DNA subsequences, also referred to as words, were designed using computer search algorithms to ensure significant Hamming distance between distinct words to allow for annealing discrimination. The largest libraries are constructed in two phases: (1) An initial DNA library is constructed on plastic microbeads by combinatorial, mix-and-split methods. (2) Half the library is cleaved from the beads and concatenated onto the remaining bead-bound strands to generate a new library containing elements of twice the original length and library diversity which is the square of the original. We have completed the first stage in a number of experiments of increasing size, currently the largest of which has $12^7$ microbeads, each carrying approximately $10^7$ strands of DNA. This already constitutes by far the largest number of distinct DNA strands synthesized in a library of this type. Following successful completion of the second construction phase, the resulting DNA library will contain $12^{14}$ or 1.28 x $10^{15}$ distinct data elements.

We describe our on-going experiments for executing associative search queries within the synthesized DNA databases. These queries are executed by hybridization of a target database strand with a complementary query strand probe. In our initial annealing experiments for processing associative search queries, we employed fluorescently labeled query strands and performed separation of fluorescent versus non-fluorescent beads using Fluorescence Activated Cell Sorting (FACS or flow cytometry). We also tested polymerase chain reaction (PCR) as an output method, and developed a PCR technique for search in the pair-wise constructed library that exploits the particular properties of words in that library. We have also implemented computer software that provides a simulation (viewable on the internet) of the experimental search procedures, as well as a simulation of input/output from conventional 2D images.

## 1 Introduction

**1.1 Overview.** All known biological organisms make use of the sequential ordering of monomeric bases in long-chain nucleic acid molecules for storage, processing, and transmission of biological information. Researchers in the field of DNA-based computing are now investigating the possibility of encoding, storing, manipulating, and retrieving non-biological information in DNA sequences. The present study aims to test one such application, specifically, the creation of large databases of DNA sequences and methods for associative search queries within the databases.

The extreme compactness of DNA as a data storage is nothing short of incredible. Since a mole contains 6.02 x $10^{23}$ DNA base monomers, and the mean molecular weight of a monomer is approximately 350 grams/mole, then 1 gram of DNA contains 2.1 x $10^{21}$ DNA bases. Since there are 4 DNA bases, each DNA base can encode 2 bits, and it follows that 1 gram of DNA can store approximately 4.2 x $10^{21}$ bits. In contrast, conventional storage technologies can store at most roughly $10^9$ bits per gram, so DNA has the potential of storing data on the order of $10^{12}$ more compactly than conventional storage technologies.

**1.2 Prior Work**. Eric Baum [B95] 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 data base size.* However, there were further technical issues to be resolved. For example, the query may not be an exact match or even partial match with any data in the database, but DNA annealing affinity methods work best for exact matches. Reif and LaBean [RL00] proposed improved biotechnology methods to do associative search in DNA databases. These methods adapted some information processing techniques (Error-Correction and VQ Coding) to

[ϕ] Address correspondence to reif@cs.duke.edu optimize input and output (I/O) to and from conventional media, and to refine the associative search from partial matches to exact matches. Prior to our project, the use of DNA annealing to achieve parallel associative search had not been experimentally implemented.

The current study follows from significant prior work by Lynx Therapeutics on construction of bead-bound DNA libraries [BWV+00, BJB+00]. The Lynx methods were developed for the purpose of differential expression analysis which is the comparison of the ensemble of mRNA transcribed in different cell types or at different times. The Lynx method begins with the synthesis of a large combinatorial library of oligonucleotides by mix-and-split technique on plastic beads. They are appended with cDNA such that each cDNA is linked with high probability to only a single unique synthetic tag. Differential analysis is accomplished by hybridization of fluorescent labeled probes and sorting by FACS. The prior work made significant strides toward construction of specific purpose DNA-encoded databases. The current study utilizes a similar synthetic technique for construction of immobile libraries, however we have designed modified methods for construction of libraries of much larger size, we are interested in implementation of associative search, and we have allowed for output via not only FACS but by PCR.

There is also considerable prior work on DNA codeword design [MCC00, A96, B96, DMGFS96, M96, GDNMF97] and on word design used for surfaced-based DNA computing [GFB+96, FTCSC97]. [CRF+96] shows that surface morphology may be an important factor for discrimination of mismatched DNA sequences. A three-base design has been described [CFLL98]. Evolutionary search methods for word designs are described in [DMR+97]. Laboratory experiments of word designs have been performed [KCL96] and ligation experiments are described by Jonoska and Karl [JK97]. Wood [W98] considers the use of error correcting codes for word design and to decrease mismatch errors. Hartemink *et al.* [HGK98] described an automated constraint-based procedure for nucleotide sequence selection in word design. We will utilize and improve on these methods for DNA word design, including evolutionary search methods, and error correcting codes.

**1.3 Current Work.** This paper 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 database consists of a large collection of single-stranded DNA molecules, either free in solution or immobilized on polymer beads, glass slides, or chips. A database strand carries a particular DNA sequence consisting of a number of sequence words drawn from a predetermined set, or lexicon, of possible words. We use at least10 times more DNA than the theoretical minimum of one DNA strand per data item, to provide at least 10-fold redundancy, so each database element is represented by approximately10 identical strands of DNA. This level of redundancy is probably on the lowest end of the range of detectability . These libraries can be used to emulate smaller libraries with much greater redundancy by essentially ignoring the values recorded at some internal positions. We would experience a $12^k$ fold increase in redundancy if values of k blocks are ignored. This strategy will be explained further below.

The experiments described here involve the hybridization of query strands to database strands such that database strands of interest will become marked and separated from the bulk database following query strand binding. A query strand contains the complement of a portion of a database strand, such that the query strand will specifically hybridize (anneal) and co-localize with its complementary database strand. Although the experimental libraries we constructed essentially only hold only a singe bit of information per database element (the data bit is 1 if and only if the element is contained in the library), we can in principle append to each library element a further DNA sequence providing data values.

One goal of the study is 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. False positives should be minimized 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. It is desirable to directly measure the limits of detection. It is also useful to construct a database containing words of known, low probability and some of known, high probability; then word strings of known probability can be queried to gauge the ability to retrieve rare sequences within databases of high strand diversity.

**1.4 Small Test DNA Library Experiments.** For test purposes, we first synthesized by combinatorial, mix-and-split methods a small test library of size $4^6$ on plastic microbeads. We used a biased synthesis technique that made certain sequences have very low probability. For testing of annealing stringency on this library, we used fluorescently labeled query strands and then performed separation of fluorescent versus non-fluorescent beads by fluorescence activating cell sorting (FACS). We have completed the construction of this test library and implemented readout by FACS.

**1.5 Larger DNA Library Experiments.** Then we increased the scale of our experiments and synthesized a larger initial DNA library, again by combinatorial, mix-and-split methods on plastic microbeads. This resulted in an initial library of size $12^7$ = 35,831,808. Each element in the initial database encodes a sequence of 7 base 12 numbers, using a sequence of 7 consecutive 5 base DNA words.

**1.6 Extremely Large DNA Library Experiments.** We are now constructing a large, diverse library of $12^{15}$ DNA sequences. Each DNA strand of the library is single-stranded and encodes a number which provides the index to the database element. To encode the base 12 digit in the $i^{th}$ position, we use a distinct 12 element set $S_i$ of DNA subsequences, each of 5 DNA bases in length. The encoding of each k place base 12 digit integer is thus done using a sequence of k consecutive 5 base DNA sequences. In addition each DNA strand of the library has certain flanking subsequences that are used in the synthesis of the library.

Our strategy entails a two phase synthesis of this DNA library:

(1) First we utilize our initial DNA library of size $12^7$ = 35,831,808 which we have already constructed by combinatorial, mix-and-split methods on plastic microbeads.

(2) Then we square the size of the library by combining pairs of the initially synthesized library strands using DNA hybridization and ligation. The resulting DNA data base elements consist of a concatenation of two of the previously constructed strands. The second phase will result in libraries of size $12^{14} = 1.28 \times 10^{15}$.

Although the constructed library essentially holds only a single bit of information per database element (the data bit is 1 if and only if the element is present in the library), we can easily append to each library element a further sequence providing additional data values).

**1.7 Associative Search in this Extremely Large DNA Library.** For testing of readout of a specific data element in the very large resulting library constructed in the second phase, we will use a two stage process. First, we use FACS to separate out beads whose DNA strands contain a selected suffix. Then we use PCR amplification, exploiting the particular properties of the concatenated words to ensure high fidelity selection of the desired data element. Suppose the database element searched for is indexed by a number written base 12 whose first 7 "digits" base 12 are U and last 7 "digits" base 12" are V. Then that database element is encoded by a DNA word where U is encoded by a 40 base subsequence in the prefix portion of the DNA word, and where V is encoded by a 35 base subsequence in the suffix portion of the DNA word. The PCR amplification can be done by repeated stages, where each stage involves an annealing of a primer U, or V or their complement. The success of the PCR amplification depends critically on annealing stringency of the primers, which is enhanced by the fact that (a) the primers are only 35 bases and (b) that each distinct pair of word sequences within a block differ by at least 3 bases. This PCR amplification can in principle be followed by readout via sequencing and the use of a DNA annealing array (we did not experimentally execute this final readout phase but performed a computer simulation instead.

## 2 Design and Synthesis of DNA Databases

**2.1 Design of DNA Databases.** In designing a DNA-encoded database one must consider several important factors including the following. The overall length of the oligonucleotide sequences used for matching is critical because sequence length directly effects the fidelity and melting temperature of DNA annealing. 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 (see Figure 2). This strategy also lends itself well to chemical synthesis of the database and will be further described below. Another important consideration is the choice of the words themselves and the grouping of words into sets for use in the blocks. Sentence length, desired library diversity, and word-pair distance constraints all effect the choices of words in the lexicon.

In designing the initial small test DNA-encoded database, first it was decided that 24 residues was a good length for the data-encoding, match region of the sequences (the tag sequence to which the probe sequence on the query strand would bind). Next, a word/block synthesis scheme was decided upon. Figures 1 & 2 show high level views of several important aspects of the database design.

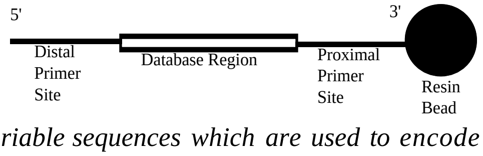

***Figure 1. Schematic Plan for Database Strands.*** *DNA is shown extending toward the left from the spherical resin bead. The 5' and 3' constant regions (identical on each and every bead in the library) contain the proximal and distal sites for ssDNA primer binding for second-strand synthesis, PCR amplification, and sequencing. The database region (rectangular box) contains the variable sequences which are used to encode information. Details of the specific variable regions synthesized are given below.*

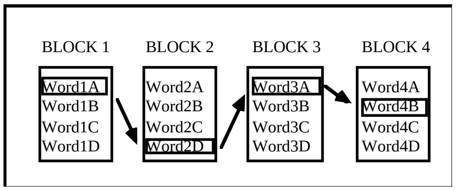

***Figure 2. The Range of Possible Sentences entailed by a Word-Block Construction Scheme.*** *For each block position in the sequence, one word is chosen from the word set and synthesized on the growing DNA strand. Separate reaction vessels are used for each word in the block so that all the word choices 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 will be populated with beads therefore all possible DNA sentences will be synthesized. Each bead contains multiple copies of a single DNA sequence (see below for details of the mix-and-split synthesis scheme).*

If the set of possible sentences greatly exceeds the number of beads used in the synthesis, then probabilistically, each bead will display a unique sequence (one-bead-one-sequence). In the present case, the number of beads in the synthesis exceeds the number of possible sentences, thus multiple beads containing any particular sentence will be present. On all beads the 5' and 3' constant regions contain invariant sequences regardless of the central variable sequence. These constant regions are used for PCR amplification of all database strands present in a test tube. The size of a particular database ( its potential information content) is given by the diversity of possible sequences within the specific library design.

The terms 'library' and 'database' are used somewhat synonymously but 'library' refers more to the physical collection of DNA strands, while 'database' refers to the interpretation of that collection of molecules as a means of storing and manipulating information. Diversity refers to the total size of a library or database, that is, the number of unique sequences consistent with the design and synthesis criteria. Library diversity scales with the number of blocks and with the size of the set of words allowed in each block as shown in Figure 3.

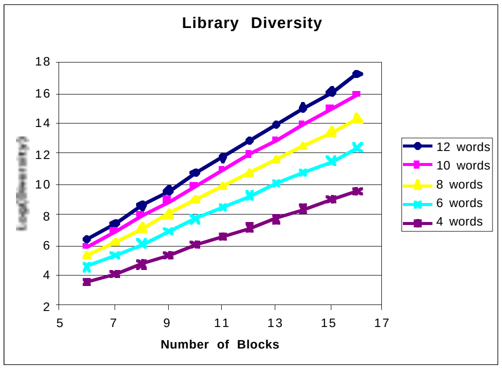

***Figure 3. Library Diversity versus Sentence Length.*** *The figure shows the scaling of library diversity with increasing sentence length (block count) and increasing number of available words within each block. The relationship between words and blocks was described above in Figure 2. Diversity is calculated by raising the word count to the exponential power given by block count (i.e. diversity = [word count]$^{block\ count}$ ). As a simple example, to achieve a total diversity greater than one million with sentences containing 6 blocks, one would require a set of 10 word choices per block. To achieve our goal of a total diversity of $12^{14}$ with sentences containing 14 blocks (for example) requires a set of 12 word choices per block.*

By varying the probabilities for different words within a block, the database provided the capability for testing query searches for rare sequences down to one in approximately half a million. From each set of four words in any given block, word probabilities (concentrations) were set at 2x, x, x, and 1/2x by dividing the resin into four, uneven batches and synthesizing a specific word onto each. Thus the most common database entry or “sentence” was present at $(4/9)^6$ = 1/130 randomly selected database strands and the average sentence was be found with probability $(2/9)^6$ =1/8,304. However, the most rare strand had probability $(1/9)^6$= 1/531,441. This provided us the opportunity to simulate a library of size 531,441 where individual strands have probability 1/531,441. This range of word probabilities provided a very wide range of sentence probabilities while only requiring a 4-fold differential during the division of resin in the oligo synthesis steps.

## 2.2 Word and Block Set Design

The Hamming distance between two DNA words is the total number of base mismatches, for the best possible alignment of the two sequences. A set of programs was written in C++ to assist in balancing the following constraints of good DNA code words: 1) minimize the melting temperature difference ($T_m$) between words so that hybridization of multiple words proceeds simultaneously; 2) maximize Hamming distances between word pairs and between words and complements of word; 3) avoid frame shift binding errors by minimizing overlap between desired words and spurious words straddling boundaries between adjacent blocks.

It was previously noted that a library strand with A, C, G, and T has much greater chance of significant secondary structure than a library strand composed of just A, C, and T [Brenner *et al.*, 2000]. We use only A, C, and T in our DNA code word design. In this document, we represent a DNA code word as a string over the alphabet {A, C, T} and assume that the leftmost end of the string corresponds to the 5' end of the associated DNA code word. The number of C residues present per word profoundly affects the $T_m$ value. From the alphabet {A, C, T} our first program generates all the words containing the desired number of C's. For example, if we specify the minimum number of Cs to be 1, maximum number to be 1, the word length to be four, The following words are generated: AAAC, AACA, AACT, AATC, ACAA, ACAT, ACTA, ACTT, ATAC, ATCA, ATCT, ATTC, CAAA, CAAT, CATA, CATT, CTAA, CTAT, CTTA, CTTT, TAAC, TACA, TACT, TATC, TCAA, TCAT, TCTA, TCTT, TTAC, TTCA, TTCT, TTTC. Beginning with this set of words, the second program generates all subsets which satisfy minimum Hamming distance and minimum set size requirements. The third program generates the satisfied library. It addresses the issue of spurious word repeats due to "frame shifts" at boundaries between neighboring blocks. This program takes the following parameters: If we have block A and its adjacent block B. Take one word from A and another from B, we have a total of 16 word pairs that form an 8 base sequence. The shifting distance is the minimum Hamming distance between every word( in A and B) and every pair (the 8 base sequence) by shifting. We also defined a score function that reflects the probability of generating word repeats. This score was minimized in order to minimize duplicated words.

Due to the high number of sets generated from the second program, it was not possible to test all possible libraries. This program uses instead a greedy algorithm. It gives us a result that satisfies the specified parameters. We ran the whole program multiple times and each time received distinct results. To choose between these viable results, we used two constrains: i) minimize duplicated words; ii) minimize shifting score( as mentioned above ).
To decide which one to use from among the generated sets, we simulated hybridization tests for every sequence in the library as follows: (i). We used a program to generate all the possible sentences; (ii). We concatenated the resulting sentences together to form a huge DNA sequence. (iii). The first sentence, the last sentence and the sentence with the most duplicated words (from i) were used as probes. The software BIND [HG97] was used to calculate free energies of association and melting temperatures for each potential probe binding sites; results were used to compare exact matches with spurious or partial match sites. The $T_m$ value between the best binding site and the next best binding site is compared. The chosen set had larger overall $T_m$ difference and fewer duplicated words.

## 2.3 Synthesis of Database

Desired manipulations of the DNA databases require the strands to remain bound to the resin beads upon which they are synthesized. The resin chosen as solid support was TentaGel M $NH_2$ (Rapp, Inc; see http://rapp-polymere.com/). It is a monodisperse resin consisting of polystyrene microspheres functionalized with amide groups ($NH_2$). The uniformity of bead size enables application of automatic sorting techniques. Release of the DNA can be made by treatment with acid. For the 20 μm bead size, the $NH_2$ capacity is 1.0 pmoles/bead (pmole = $10^{-12}$ mole), which provides attachment of 6.02 x $10^{11}$ strands of DNA per bead. One gram of the 20 μm bead resin contains approximately 2.4 x $10^8$ beads. For the 10 μm bead size, the $NH_2$ capacity is 0.13 pmoles/bead, which provides attachment of 7.83 x $10^{10}$ strands of DNA per bead. A gram of the 10 μm bead resin contains approximately 1.95 x $10^9$ beads.

Figure 4 provides a general outline for the mix-and-split procedure used for the synthesis of combinatorial database libraries. The process can produce vast libraries in which multiple, identical copies of a DNA sequence will be created on any given bead. This one-bead/one-sequence design produces libraries suitable for associative search query experiments because fluorescent probe will localize to and label specific beads carrying copies of target sequence and fail to label beads carrying unrelated sequence.

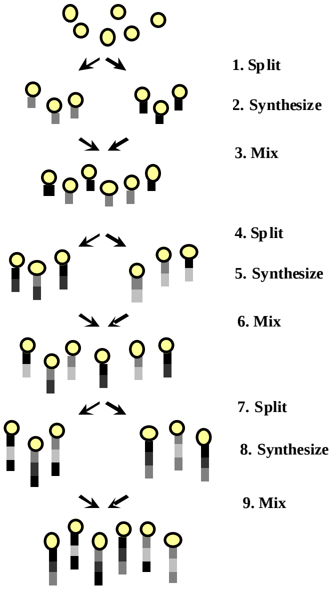

***Figure 4. Flow chart of mix-and-split synthesis scheme.*** *At the beginning of the process (top) bare resin beads are prepared for library construction; this can include synthesis of the 3' constant region in a single reaction vessel. Step 1: resin is split into separate reactions (the figure shows division into two parts). Step 2: a single, specific sequence word is synthesized on all beads in each of the vessels (one vessel - one word). Step 3: All resin is recombined and mixed so that in the next splitting (Step 4) each vessel will receive beads carrying each of the preceding words (both black and white words from Step 2). The next word is added in Step 5; the entire ensemble is remixed in Step 6 and the process can continue with one splitting step for each block in the design. The final result of the process (bottom of figure) is a library of beads where each bead contains multiple copies of a single sequence (for clarity only a single copy is shown per bead). Note that in this summary figure the word syntheses are shown as single steps while the actual chemistry requires the addition of nucleotide monomers one at a time, so each synthesis step in the figure corresponds to a series of chemical cycles – one cycle for each base in a word.*

The database strands were synthesized using an ABI automatic synthesizer and conventional phosphoramidite chemistry. Fresh, dry reagents and solvents were used with coupling times and deprotection conditions designed to optimize synthesis yield. Yields in the high ninety percents were obtained per synthesis cycles. Figure 5 presents detailed sequence information for synthesis of the initial, small library. Figure 6 shows four database sequences which were chosen as targets for experimental queries.

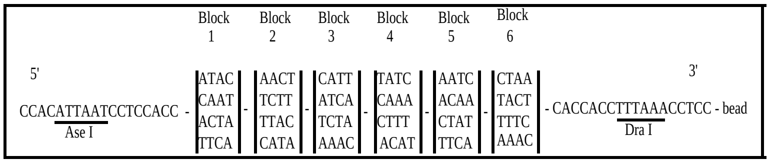

***Figure 5. DNA Sequences for Small Database.*** *The word set design criteria and procedures described above were used to arrive at the word sets listed. The distal and proximal constant regions were designed to contain Ase I and Dra I restriction endonuclease cleavage site (respectively). Restriction sites were included because they may be useful for removing strands from the beads and for possible concatenation or cloning if the necessity should arise. The constant regions each contain 50% C to increase the melting temperature of primer binding.*

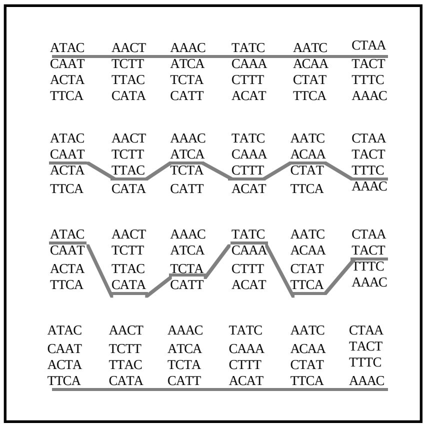
ATAC AACT AAAC TATC AATC CTAA
CAAT TCTT ATCA CAAA ACAA TACT
ACTA TTAC TCTA CTTT CTAT TTTC
TTCA CATA CATT ACAT TTCA AAAC
ATAC AACT AAAC TATC AATC CTAA
CAAT TCTT ATCA CAAA ACAA TACT
ACTA TTAC TCTA CTTT CTAT TTTC
TTCA CATA CATT ACAT TTCA AAAC
ATAC AACT AAAC TATC AATC CTAA
CAAT TCTT ATCA CAAA ACAA TACT
ACTA TTAC TCTA CTTT CTAT TTTC
TTCA CATA CATT ACAT TTCA AAAC
ATAC AACT AAAC TATC AATC CTAA
CAAT TCTT ATCA CAAA ACAA TACT
ACTA TTAC TCTA CTTT CTAT TTTC
TTCA CATA CATT ACAT TTCA AAAC

***Figure 6. Target Sequences shown as Traces through Word Sets.*** *The six blocks of four words are shown four times with lines tracing the patterns of synthesis for selected database sequences. Words are listed with the most common subsequence (2X concentration) at the top, followed by the two moderate (1X) concentration words, and finally the least common word (1/2X) at the bottom of each list. The top trace gives the most common sentence, composed entirely of most common words. The second trace shows a sentence of moderate probability in the database, comprised of moderate concentration words. The third trace gives a sentence of moderate probability, composed of common, rare, and moderate words. The final trace shows the rarest sentence, made up of the least common words.*

The four traces shown in Figure 6 define a set of queries for testing searches over the entire range of difficulty within the small database. Shown below are the target sequences (written 5' to 3') aligned with their complementary probe sequences (written 3' to 5').

*Most common sentence/high probability words (1 copy in 130 sentences)*

```
Target 1:   ATAC AACT AAAC TATC AATC CTAA
Probe 1:    TATG TTGA TTTG ATAG TTAG GATT
```

*Moderate sentence probability/constant moderate word probability (1 copy in 8,304 sentences)*

```
Target 2:   CAAT TTAC ATCA CTTT ACAA TTTC
Probe 2:    GTTA AATG TAGT GAAA TGTT AAAG
```

*Moderate sentence probability/variable word probabilities (1 copy in 8,304 sentences)*

```
Target 3:   ATAC CATA TCTA TATC TTCA TACT
Probe 3:    TATG GTAT AGAT ATAG AAGT ATGA
```

*Least common sentence/low probability words (1 copy in 531,441 sentences)*

```
Target 4:   TTCA CATA CATT ACAT TTCA AAAC
Probe 4:    AAGT GTAT GTAA TGTA AAGT TTTG
```

We first constructed this small initial library for the first phase. This first phase resulted in initial libraries of size $4^6$ where the most rare strand had probability $(1/9)^6$ = 1/531,441 effectively modeling libraries of size up to 531,441.

## 2.4 Experiments in DNA Search and Readout in the Small Library

In the initial experiments, fluorescently labeled query strands were used as probe to anneal to and mark microbeads carrying target database strands. Then, fluorescent and non-fluorescent beads were separated by Fluorescence Activated Cell Sorting (FACS). FACS or flow cytometry was developed for the study and purification of specific cell types from complex mixtures of cells. FACS has also been used for sorting diverse libraries of biomolecules immobilized on micro-scale plastic beads [Brenner *et al.*, 2000]. In FACS, a thin stream of individual particles are passed though one or more laser beams, causing light to scatter and fluorescent dyes to emit light. The light signals are converted to electrical impulses, and data about the observed particles are used to direct particle-containing droplets into appropriate down stream receptacles. The electrostatically charged droplets are deflected into a desired vessel by charged plates. Scattered light indicates a cell or bead is present, and fluorescence signal indicates that a bead has been labeled with a probe strand and thus matches the database query. Besides separation of beads FACS also provides a count of the number of beads sorted into each group.

Search experiments with FACS output (Figure 7) provided data in good agreement with expected results. In control runs, bead counts were within a few percentage points of the designed values. As shown in Figure 7, the location of the F peak shifts down field in the experimental versus the control samples. This shift indicates decreased fluorescence intensity for fluorescein attached to annealed probe compared to the same dye directly attached to bead-bound strand. The expected result for a clean separation is an F "island" fairly well removed from the N peak, as seen in the controls. Further experiments will clarify whether the observed results are due to a difference in fluorescence yield of the dye or due to varying levels of non-specific probe binding. The former explanation currently appears most plausible because increases in annealing and washing stringency decrease the entire F peak and do not preferentially affect the down field portion of the peak. Overall, the current results indicate the viability of the query and output methods.

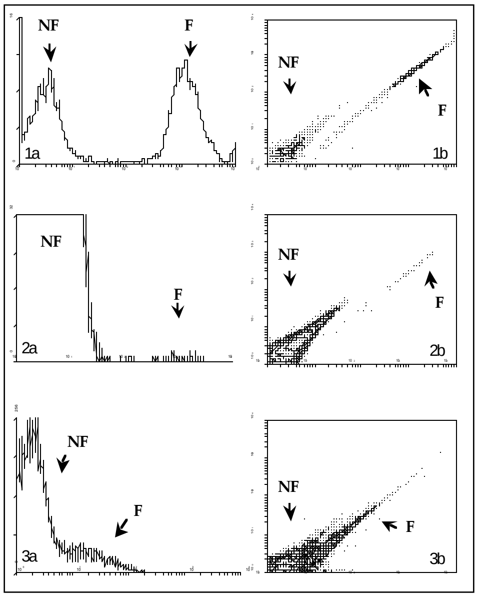

***Figure 7. FACS Data from Query of the Small Library.*** *Data from three experiments is shown in two formats; panels labeled 'a' show histograms of bead count (y-axis) versus fluorescent intensity (x-axis); panels labeled 'b' show fluorescent intensity measures from two separate photomultiplier channels with each dot representing a single bead observation. Groupings of non-fluorescent (NF) and fluorescent (F) beads are indicated. Panels 1a&b show results from a control experiment containing a 50:50 mixture of fluorescently labeled and unlabeled beads. This sample contained 20,000 beads. Control mixtures were prepared using beads carrying a small test oligonucleotide (NF) and the same small oligo with a molecule of fluorescein covalently attached to the 5' end (F). Panels 2a&b show the results of a similar control in which the labeled beads were present at only 1 in 10,000 beads. A sample containing 150,00 beads was sorted. Panels 3a&b show data from a search experiment in which a fluorescently label DNA probe was hybridized with the bead bound library. The sample contained 47,000 beads. The probe sequence was expected to bind specifically to approximately 1 in 130 database beads (Probe 1, described in the previous section). Depending upon where the boundary between the groupings was drawn, the F (labeled) peak contained between 0.6% and 1.0% of the total beads, in reasonable agreement with the expected value.*

**2.5 Increasing the Number of Database Strands.** The next stage of the project is to increase database size. Note that this database size generated by combinatorial, mix-and-split methods on plastic microbeads is limited by a number of factors, including the total number of beads, which is up to 20,000,000,000 for 5 micron beads, but is approximately 1,000,000,000 for the 25 micron beads used here. The maximum library size generated by combinatorial, mix-and-split methods on 25 micron plastic microbeads can thus be at most approximately 100,000,000. We increased database size by two methods:

(i) The number of possible words chosen for each block was increased from 4 to 12 and the length of each word was increased from 4 bases to 5 bases. There is Hamming distance of at least 3 between each pair of distinct words within a block. Seven stages of the combinatorial, mix-and-split synthesis on plastic microbeads then resulted in a library L of size $12^7$ = 35,831,808, where each strand has 35 bases in addition to the fixed flanking sequences at each end.

(ii) The size of the library is then squared by combining pairs of the synthesized library strands.

A library was designed with 7 blocks each of 12 possible words using procedures similar to those described above for the smaller library. Each DNA strand of the library is single stranded, and encodes a number which provides the index to the database element. To encode the base 12 digit in the $i^{th}$ position, we used a distinct 12 element set $S_i$ of DNA subsequences, each of 5 bases in length. The encoding of each k place base 10 digit integer is thus done using a sequence of k consecutive 5 base DNA sequences. (In addition each DNA strand of the library has certain flanking subsequences that are used in the synthesis of the library. Designed word sets are:

|  | Block | Block | Block | Block | Block | Block | Block |  |
|---|---|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 |  |
|  | AAACC | AATCC | AACCA | AACCT | AATCC | ACACA | AACCA |  |
|  | ACCAA | ACACT | ACATC | ACCTA | ACAAC | ACCAT | ACACT |  |
|  | ACTCT | ATCAC | ACCAT | ACTAC | ACCTT | ATCTC | ACTTC |  |
|  | ATCTC | CAAAC | ATTCC | ATACC | ATCCA | CACAA | ATCAC |  |
|  | CATAC | CCATA | CACTT | CAAAC | CAACT | CATTC | CATAC |  |
| 5' *Bam HI* | CCTTA | CCTAT | CATAC | CCATT | CCATA | CCATT | CCAAA | *Bgl II* 3' |
| CATCGGATCC | CTACA | CTCTT | CCAAA | CTCAA | CTCAT | CTAAC | CTCTA | AGATCTCACACCCTCCAC |
|  | CTCAT | CTTCA | CTACT | CTTCT | CTTTC | CTTCA | CTTCT |  |
|  | TACCA | TACCT | TAACC | TATCC | TACTC | TAACC | TACTC |  |
|  | TCAAC | TCCAA | TCCTA | TCACA | TCCAA | TCCTA | TCCAT |  |
|  | TCCTT | TCTTC | TCTCT | TCCAT | TCTCT | TCTAC | TCTCA |  |
|  | TTTCC | TTACC | TTCAC | TTCTC | TTACC | TTCCT | TTACC |  |

The second phase of construction of the very large library is shown in Figure 8. At the time of this writing the first phase had been successfully completed, while the second phase had yet to be attempted.

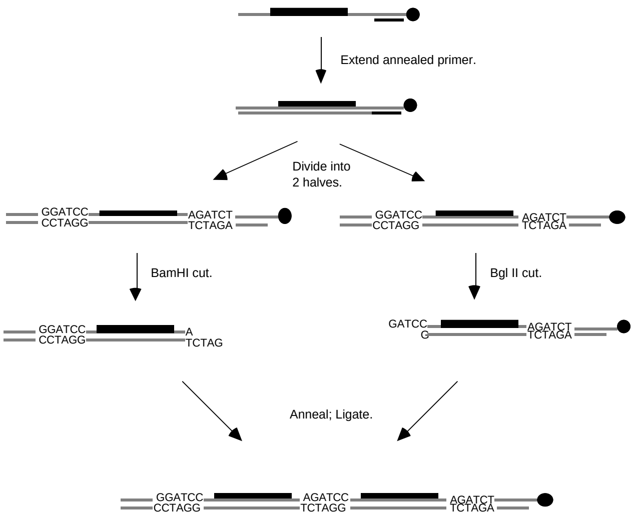

***Figure 8. Procedure for Squaring the Diversity of the Library.*** *The database encoding region of the DNA is shown as a black rectangle, and the resin bead as a black circle. First, a complementary primer is annealed to the bead bound database strand and extended with DNA polymerase to generate the double-stranded library. Next, the library is divided and one half is digested with Bam HI while the other half is digested with Bgl II. Note that any pair of restriction enzymes which generates compatible sticky-ends could be used. Note also that dephosphorylation of the Bgl II cut fraction will prevent tail-to-tail ligation in the subsequent step. Following annealing and ligation of the Bam HI and Bgl II digested fractions, the double length library is obtained. The resulting sentences consist of 14 words, each of 5 bases, plus the remaining invariant sequences. The new library will have half as many total elements as the starting library, but its diversity will be the square of the starting library's diversity. Even with 10-fold coverage in the starting library there is a high probability that all potential sequences will be represented in the final library.*

### 2.6 Experiments in DNA Search and Readout in the Very Large Libraries

Our libraries grew so large that the exclusive use of FACS readout was no longer sufficient to isolate a desired database entry. However, FACS readout can be employed as a first step, isolating sets of DNA strands with a common suffix. Given a search query, we will first separate out (either using magnetic bead separation or a cell sorter) all sequences matching the last 7 blocks of a sentence. This will greatly decrease the amount of DNA to be searched. A simple design for our hybridization experiments will be to use PCR with primers consisting of two sequences, each consisting of 7 blocks. These can be sequenced to provide output.

## 3 Computer Simulations.

We also made computer simulations of our method for DNA-based associative search. We began with the selection of (digital) multispectral images to form a trial image database, and then preprocessed the image data to construct an "attribute database" which was then converted into a DNA database. Using known parameters for the kinetics of DNA hybridization, we simulated the use of PCR to perform an associative search. The result was used to reconstruct the original image using extensions of techniques described in [Reif and LaBean, 2000]. These simulations can be viewed on the internet at the following sites.

*DNA Database Search Simulation* http://cgi.cs.duke.edu/~clk/dnasrch/dna_search.cgi

| | |
|---|---|
| *DNA Database Selection* | http://cgi.cs.duke.edu/~clk/dnasrch/open_dna_db.cgi |
| *DNA Image Database Construction Page:* | http://cgi.cs.duke.edu/~clk/dnasrch/MakeDNADB.html |

*More information about searching:*

http://cgi.cs.duke.edu/~clk/dnasrch/help/search_info.html#select

## 4 Conclusion and Future Applications

This study explored the use of DNA databases for storage of information and retrieval of that information by associative search techniques. Important procedures in subsequence and sequence design were demonstrated as well as successful application of those designs to combinatorial library synthesis by mix-and-split methods. Query output by bead separation using a fluorescence activated cell sorter was achieved. Test databases of increasing diversity were synthesized, and construction of an extremely large database is now partially complete. Future work includes finishing construction of the large library, further optimization of hybridization conditions for associative searches, sequencing a number of database elements as quality control of the synthesis, evaluating the benefits (increased fidelity?) of increasing the word length from 4 to 5 bases, and examining the limits of detection by searching for very rare database elements.

Bead-bound DNA databases similar to those described here would be useful for applications in which spatial clustering of identical sequences imparts some benefit, for example searches within libraries so huge that a presorting or enrichment procedure would be required as an initial filtering step. It should also be noted that the DNA databases described here can easily be cleaved from the solid support and utilized in soluble form for applications in which solubility would be beneficial. We are currently adapting our database system for construction of large cDNA libraries with synthetic DNA tags on both ends of each element for more complex associative search procedures.

## References

[A96] Amenyo, J.-T., Mesoscopic computer engineering: Automating DNA-based molecular computing via traditional practices of parallel computer architecture design, Proceedings of the 2nd Annual DIMACS Meeting on DNA Based Computers, (June 1996).

[B95] Baum, E. B., How to build an associative memory vastly larger than the brain, *Science*, **268** 583-585 (1995).

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

[BJB+00] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K, Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, *Nature Biotechnology*. **18**(6):630-4, (2000).

[BWV+00] Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J., Luo, S., Kirchner, J.J., Eletr, S., DuBridge, R.B., Burcham, T. & Albrecht G. In Vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. *PNAS*, **97**(4), 1665-1670 (2000).

[CFLL98]. Cukras AR. Faulhammer D. Lipton RJ. Landweber LF. Chess games: a model for RNA based computation, *Biosystems*. **52**(1-3):35-45, (1999).

[CRF+96] Cai, W., E. Rudkevich, Z. Fei, A. Condon, R. Corn, L.M. Smith, M.G. Lagally, Influence of Surface Morphology in Surface-Based DNA Computing, Submitted to the 43rd AVS National Symposium, Abs. No. BI+MM-MoM10, (1996).

[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.

[DMR+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.

[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).
[CRFCC+96]

[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, U. Penn., (June, 1997).

[GFB+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).

[HG97] Hartemink, A. & Gifford, D. (1997) Thermodynamic Simulation of Deoxyoligonucleotide Hybridization for DNA Computation. *Proceedings of the 3rd DIMACS Workshop on DNA Based Computers*, University of Pennsylvania (June 1997).

[HGK98] 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).

[JK97]. Jonoska, N., and S.A. Karl, Ligation Experiments in Computing with DNA. ICEC'97 Special Session on DNA Based Computation, Indiana, (April 1997).

[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).

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

[MCC00] Marathe, A., Condon, A.E. & Corn, R.M. (2000) On Combinatorial DNA Word Design, *DNA Based Computers V*, DIMACS Workshop on DNA Based Computers (5th, MIT, June 14-15, 1999) editors E. Winfree & D.K. Gifford, DIMACS Series in Discreet Mathematics and Theoretical Computer Science, Volume **54**, pp 75-90, (2000).

[RL00] Reif, J.H. & LaBean, T.H. (2000) Computationally Inspired Biotechnologies: Improved DNA Synthesis and Associative Search Using Error-Correcting Codes and Vector-Quantization, Sixth International Meeting on DNA Based Computers (DNA6), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Leiden, The Netherlands, (June, 2000) ed. A. Condon. To be published by Springer-Verlag as a volume in Lecture Notes in Computer Science, (2001).
[PostScript: http://www.cs.duke.edu/~reif/paper/Error-Restore/Error-Restore.ps]
[PDF: http://www.cs.duke.edu/~reif/paper/Error-Restore/Error-Restore.pdf]

[W98] Wood, D. H., Applying error correcting codes to DNA computing, 4th DIMACS International Meeting on DNA-Based Computing, Baltimore, (June, 1998).

(John H. Reif) COMPUTER SCIENCE DEPARTMENT, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* reif@cs.duke.edu

(Thomas H. LaBean) COMPUTER SCIENCE DEPARTMENT, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* thl@cs.duke.edu

(Michael Pirrung) DEPARTMENT OF CHEMISTRY, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* pirrung@chem.duke.edu

(Vipul S. Rana) DEPARTMENT OF CHEMISTRY, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* vipul@chem.duke.edu

(Bo Guo) COMPUTER SCIENCE DEPARTMENT, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* bog@cs.duke.edu

(Carl Kingsford) COMPUTER SCIENCE DEPARTMENT, DUKE UNIVERSITY, DURHAM, NC 27708
Currently at: COMPUTER SCIENCE DEPARTMENT, PRINCETON UNIVERSITY, PRINCETON, NJ 08544
*E-mail address:* carlk@cs.princeton.edu

(Gene S. Wickham) COMPUTER SCIENCE DEPARTMENT, DUKE UNIVERSITY, DURHAM, NC 27708
*E-mail address:* gwickham@acpub.duke.edu