Efficiency and Reliability of DNA-Based Memories

Max H. Garzon, Andrew Neel, and Hui Chen

Computer Science, University of Memphis
373 Dunn Hall, Memphis, TN 38152-3240
{mgarzon, aneel, hchen2}@memphis.edu

Abstract. Associative memories based on DNA-affinity have been proposed [2]. Here, the performance, efficiency, reliability of DNA-based memories is quantified through simulations *in silico*. Retrievals occur reliably (98%) within very short times (milliseconds) despite the randomness of the reactions and regardless of the number of queries. The capacity of these memories is also explored in practice and compared with previous theoretical estimates. Advantages of implementations of the same type of memory in special purpose chips *in silico* is proposed and discussed.

1 Introduction

DNA oligonucleotides have demonstrated to be a feasible and useful medium for computing applications since Adleman's original work [1], which created a field now known as biomolecular computing (BMC). Potential applications range from increasing speed through massively parallel computations [13], to new manufacturing techniques in nanotechnology [18], and to the creation of memories that can store very large amounts of data and fit into minuscule spaces [2], [15]. The apparent enormous capacity of DNA (over million fold compared to conventional electronic media) and the enormous advances in recombinant biotechnology to manipulate DNA *in vitro* in the last 20 years make this approach potentially attractive and promising. Despite much work in the field, however, difficulties still abound in bringing these applications to fruition due to inherent difficulties in orchestrating a large number of individual molecules to perform a variety of functions in the environment of virtual test tubes, where the complex machinery of the living cell is no longer present to organize and control the numerous errors pulling computations by molecular populations away from their intended targets.

In this paper, we initiate a quantitative study of the potential, limitations, and actual capacity of memories based or inspired by DNA. The idea of using DNA to create large associative memories goes back to Baum [2], where he proposed to use DNA recombination as the basic mechanism for content-addressable storage of information so that retrieval could be accomplished using the basic mechanism of DNA hybridization affinity. Content is to be encoded in single stranded molecules in solution (or their complements.) Queries can be obtained by dropping in the tube a DNA primer

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2723, pp. 379–389, 2003.

Watson-Crick complement of the (partial) information known about a particular record using the same coding scheme as in the original memory, appropriately marked (e.g., using magnetic beads, or fluorescent tags). Retrieval is completed by extension and/or retrieval (e.g., by sequencing) of any resulting double strands after appropriate reaction times have been allowed for hybridization to take effect. As pointed out by Baum [2], and later Reif & LaBean [15], many questions need to be addressed before an associative memory based on this idea can be regarded as feasible, let alone actually built.

Further methods were proposed in [15] for input/output from/to databases represented in wet DNA (such as genomic information obtained from DNA-chip optical readouts, or synthesis of strands based on such output) and suggested methods to improve the capabilities and performance of the queries of such DNA-based memories. The proposed hybrid methods, however, require major pre-processing of the entire database contents (through clustering and vector quantization) and post-processing to complete the retrieval by the DNA memory (based on the identification of the clusters centers.) This is a limitation when the presumed database approaches the expected sizes to be an interesting challenge to conventional databases, or when the data already exists in wet DNA, because of the prohibitive (and sometimes even impossible) cost of the transduction process to and from electronics. Inherent issues in the retrieval *per se*, such as the reliability of the retrieval *in-vitro* and the appropriate concentrations for optimal retrieval times and error rates remain unclear.

We present an assessment of the efficiency and reliability of queries in DNA-based memories in Section 3, after a description of the experimental design and the data collected for this purpose in Section 2. In Section 3, we also present very preliminary estimates of their capacity. Finally, section 4 summarizes the results and discusses the possibility of building analogous memories *in silico* inspired by the original ideas *in vitro,* as suggested by the experiments reported here. A preliminary analysis of some of these results has been presented in [7], but here we present further results and a more complete analysis.

2 Experimental Design

The experimental data used in this paper has been obtained by simulations in the virtual test tube of Garzon et al [9]. Recently, driven by efficiency and reliability considerations, the ideas of BMC have been implemented *in silico* by using computational analogs of DNA and RNA molecules [8]. Recent results show that these protocols produce results that closely resemble, and in many cases are indistinguishable from, the protocols they simulate in wet tubes [7]. For example, Adleman's experiment has been experimentally reproduced and scaled in virtual test tubes with random graphs of up to 15 vertices while producing results correct with no probability of a false positive error and a probability of a false negative of at most 0.4%. Virtual test tubes have also matched very well the results obtained *in vitro* by more elaborate and newer protocols, such as the selection protocol for DNA library design of Deaton et Al. [4]. Therefore,

there is good evidence that virtual test tubes provide a reasonable and reliable estimate of the events in wet tubes (see [7] for a more detailed discussion.)

Virtual test tubes thus can serve as a reasonable pre-requisite methodology to estimate the performance and experimental validation prior to construction of such a memory, a validation step that is now standard in the design of conventional solid-state memories. Moreover, as will be seen below in the discussion of the results, virtual test tubes offer a much better insight into the nature of the reaction *kinetics* than corresponding experiments *in vitro*, which, when possible (such as Cot curves to measure the diversity of a DNA pool), incur much larger cost and effort.

2.1 Virtual Test Tubes

Our experimental runs were implemented using the virtual test tube *Edna* of Garzon et al. [7],[8],[9] that simulates BMC protocols *in silico.* Edna provides an environment where DNA analogs can be manipulated much more efficiently, can be programmed and controlled much more easily, at much lower costs, and produce comparable results to those obtained in a real test tube [7]. Users simply need to create object-oriented programming classes (in C++) specifying the objects to be used and their interactions. The basic design of the entities that were put in *Edna* represent each nucleotide within the DNA as a single character and the entire strand of DNA as a string, which may contain single- or double-stranded sections, bulges, and loops or higher secondary structures. An unhybridized strand represents a strand of DNA from the 5'-end to the 3'-end. These strands encode library records in the database, or queries containing partial information that identify the records to be retrieved.

The interactions among objects in *Edna* represent chemical reactions by hybridization and ligation resulting in new objects such as dimers, duplexes, double strands, or more complicated complexes. They can result in one or both entities being destroyed and a new entity possibly being created. In our case, we wanted to allow the entities that matched to hybridize to each other to effect a retrieval, per Baum's design 2]. Edna simulates the reactions in successive *iterations*. One iteration moves the objects randomly in the tube's container (the RAM really) and updates their status according to the specified interactions with neighbor objects, based on proximity parameters that can be varied within the interactions. The hybridization reactions between strands were performed according to the h-measure [8] of hybridization likelihood. Hybridization was allowed if the h-measure was under a given threshold, which is the number of mismatches allowed (including frame-shifts) and so roughly codes for stringency in the reaction conditions. A threshold of zero enforces perfect matches in retrieval, whereas a larger value permits more flexible and associative retrieval. These requirements essentially ensured good enough matches along the sections of the DNA that were relevant for the associative recall.

The efficiency of the test tube protocols (in our case, retrievals) can be measured by counting the number of iterations necessary to complete the reactions or achieve the desired objective; alternatively, one can measure the wall clock time. The *number of iterations* taken until a match is found has the advantage of being indifferent to the

speed of the machine(s) running the experiment. This intrinsic measure was used because one iteration is representative of a unit of real-time for *in vitro* experiments. The relationship between simulation results in simulation and equivalent results *in vitro* has been discussed in [7]. Results of the experiments *in silico* can be used to yield realistic estimates of those *in vitro*. Essentially, one iteration of the test tube corresponds to the reaction time of one hybridization in the wet tube, which is of the order of one millisecond [17]. However, the number of iterations cannot be a complete picture because iterations will last longer as more entities are put in the test tube. For this reason, *processor time* (wall clock) was also measured. The wall clock time depends on the speed and power of the machine(s) running *Edna* and ranged anywhere from seconds to days for the single processors and 16 PC cluster that were used to run the experiments used below.

2.2 Libraries and Queries

We assume we have at our disposal a library of non-cross hybridizing (nxh) strands representing the records in the databases. The production of such large libraries has been addressed elsewhere [4], [10]. Well-chosen DNA word designs that will make this perfectly possible in large numbers of DNA strands directly, even in real test tubes, will likely be available within a short time. The exact size of such a library will be discussed below. The nxh property of the library will also ensure that retrievals will be essentially noise-free (no false positives), module the flexibility built into the retrieval parameters (here h-distance). We will also assume that a record may also contain an additional segment (perhaps double-stranded [2]) encoding supplementary information beyond the label or segment actively used for associative recall, although this is immaterial for assumptions and results in this paper. The library is assumed to reside in the test tube, where querying takes place.

Queries are strings objects encoding, and complementary of, the available information to be searched for. The selection operation uses probes to mark strands by hybridizing part of the probe with part of the "probed" strand. The number of unique strands available to be probed is, in principle, the entire library, although we consider below more selective retrieval modes based on temperature gradients. Strictly speaking, the probe consists of two logical sections: the *query* and *tail*. The tail is the portion of the strand that is used with *in vitro* experiments to physically retrieve the marked DNA from the test tube (e.g., biotin-streptavidin-coated beads or fluorescent tags [16]). The query is the portion of the strand that is expected to hybridize with strands from the library to form a double-stranded entity. We will only be concerned with the latter below, as the former becomes important only at the implementation stage, or just be identical to the duplex formed during retrieval.

When a probe comes close enough to a library or probe strand in the tube so that any hybridization between the two strands is possible, an encounter (which triggers a check for hybridization) is said to have occurred. The number of encounters can vary greatly depending directly on the concentration of probes and library strands. It appears that higher concentration reduce retrieval time, but this is only true to a point

since results below show that too much concentration will interfere with the retrieval process. In other words, a large number of encounters may cause unnecessary hybridization attempts that will slow down the simulation. Further, too many neighbor strands may hinder the movement of the probe strands in search of their match. Probing is considered complete when probe copies have formed enough retrieval duplexes with library strands that should be retrieved (perhaps none) according to stringency of the retrieval (here the h-distance threshold.) In single probes with high stringency (perfect matches), probing can be halted when one successful hybridization occurs. Lesser stringency and multiple simultaneous probes require longer times to complete the probe. The question arises how long is long enough to complete the probes with high reliability.

2.3 Test Libraries and Experimental Conditions

The experiments used mostly a library consisting of the full set of 512 non-complementary 5-mer strands, although other libraries obtained through the software package developed based on the thermodynamic model of Deaton et Al. [5] were also tried with consistent results. This is a desirable situation to benchmark retrieval performance since the library is saturated (maximum size) and retrieval times would be worst-case. The probes were chosen to be random probes of 5-mers. The stringency was highest (h-distance 0), so exact matches were required. The experiment began by placing variable concentrations (number of copies) of the library and the probes into the tube of constant size. Once placed in the tube, the simulation begins. It stops when the first hybridization is detected. For the purposes of these experiments, there existed no error margin thus preventing close matches from hybridizing. Introduction of more flexible thresholds does not affect the results of the experiments.

In the first batch of experiments, we collected data to quantify the efficiency of the retrieval process (time, number of encounters, and attempted hybridizations) with single queries between related strands and its variance in hybridization attempts until successful hybridization. Three successive batches of experiments were designed to determine the optimal concentrations with which the retrieval was both successful and efficient, as well as to determine the effect on retrieval times of multiple probes in a single query. The experiments were performed between 5 and 100 times each and the results averaged. The complexity and variety of experiments has limited the quantity of runs possible for each experiment. Over a total of over 2000 experiments were run continuously over the course of many weeks.

3 Analysis of Results

Below are the results of the experiments, with some analysis of the data gathered.

3.1 Retrieval Efficiency

Figure 1 shows the results of the first experiment at various concentrations averaged over five runs. The most hybridization attempts occurred when the concentration of probes is between 50-60 copies and the concentration of library strands was between 20-30 copies. Figure 2 represents the variability (as measured by the standard deviation) of the experimental data. Although, there exists an abnormally high variance in some deviations in the population, most data points exist with deviations less than 5000. This high variance can be partially explained by the probabilistic chance of any two matching strands encountering each other by following a random walk. Interestingly enough, the range of 50-60 probe copies and 20-30 library copies exhibits minimum deviations.

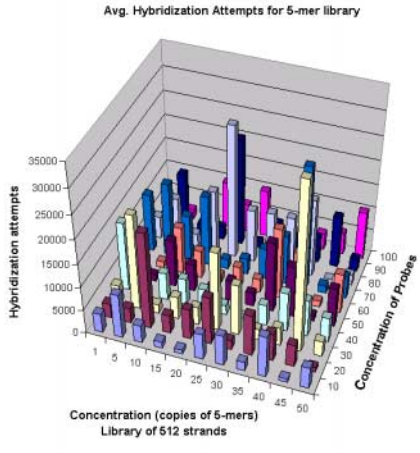

Fig. 1. Retrieval difficulty (hybridization attempts) based on concentration.

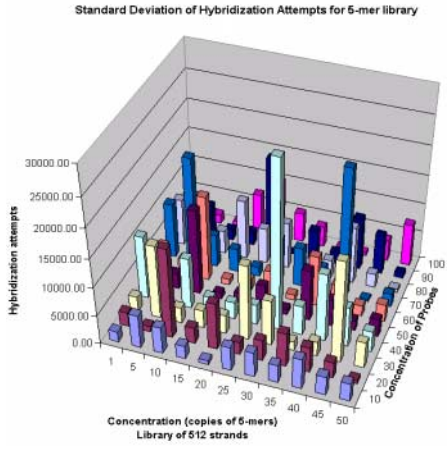

Fig. 2. Variability in retrieval difficulty (hybridization attempts) based on concentration.

3.2 Optimal Concentrations

Figure 3 shows the average retrieval times as measured in tube iterations. The number of iterations decreases as the number of probes and library strands increase, to a point. One might think at first that the highest available probe and library concentration is desirable. However, Fig. 1 indicates a diminishing return in that the number of hybridization attempts increases as the probe and library concentration increase. In order for the experiments *in silico* to be representative of the wet test tube experiments, a compromise must be made. Therefore, if the ranges of concentrations determined from Fig. 1 are used, the number of tube iterations remains under 200. Fig. 4 shows only minimum deviations once the optimal concentration has been achieved. The larger deviations at the lower concentrations can be accounted for by the highly randomized nature of the test tube simulation. These results on optimal concentration are consistent and further supported by comparison with the results in Fig. 1.

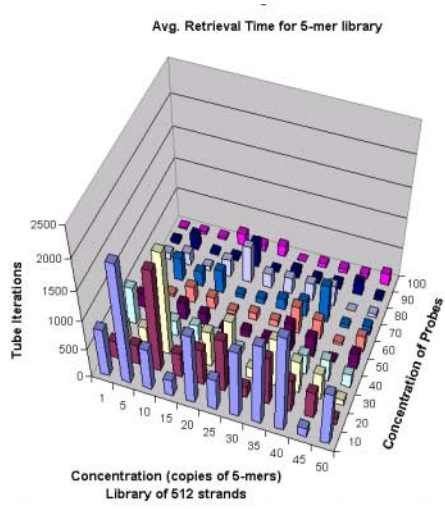

Fig. 3. Retrieval times (number of iterations) based on concentration.

As a comparison, in a second batch of experiments with a smaller (much sparser) library of 64 32-mers obtained by a genetic algorithm [9], the same dependent measures were tested. The results (averaged over 100 runs) are similar, but are displayed in a different form below. In Figure 5, the retrieval times ranged from nearly 0 through 5,000 iterations. For low concentrations, retrieval times were very large and exhibited great variability. As the concentration of probe strands exceeds a threshold of about 10, the retrieval times drop under 100 iterations, assuming a library strand concentration of about 10 strands.

Finally, Figure 6 shows that the retrieval time increases only logarithmically with the number of multiple queries and tends to level off in the range within which probes don't interfere with one another.

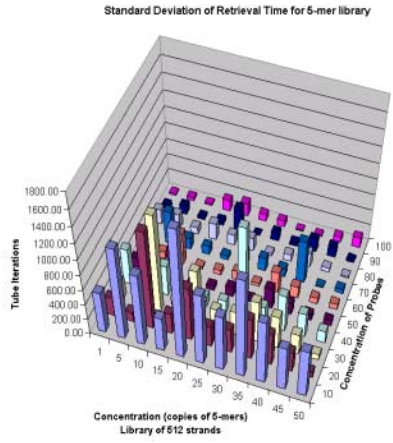

Fig. 4. Variability in retrieval times (number of iterations) based on concentration.

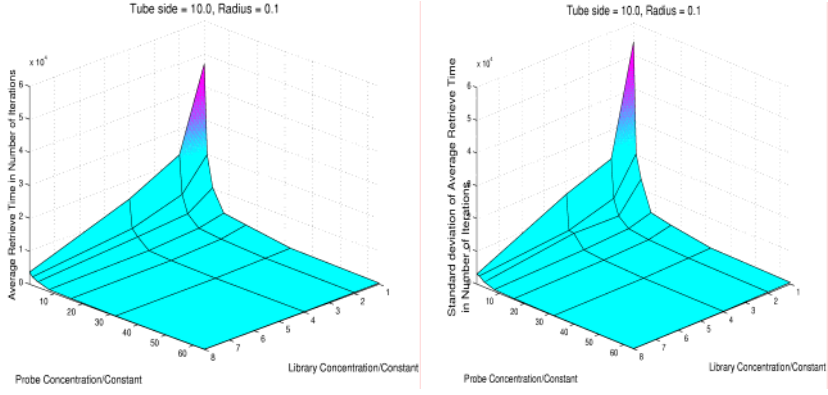

Fig. 5. Retrieval times and optimal concentration on sparser library.

In summary, these results permit a preliminary estimate of optimal and retrieval times for queries in DNA associative memories. For a library of size N, a good concentration of library for optimal retrieval time appears to be in the order of O(logN). Probe strands require the same order, although probably a smaller number will suffice. The variability in the retrieval time also decreases for optimal concentrations. Although not reported here in detail due to space constraints, similar phenomena were observed for multiple probes. We surmise that this hold true up to O(logN) simultaneous probes, past which probes begin to interfere with one another causing a substantial increase in retrieval time. Based on benchmarks obtained by comparing simulations in *Edna* with

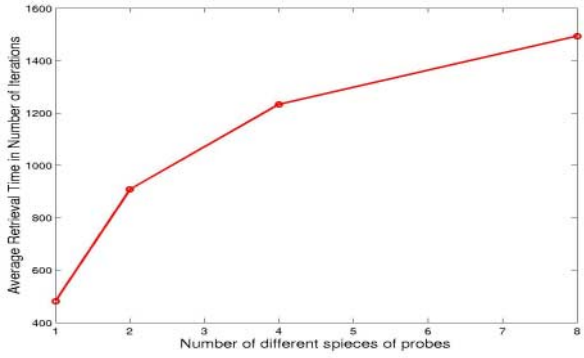

Fig. 6. Retrieval times (number of iterations) based on multiple simultaneous of queries.

wet tube experiments [7], we can estimate the actual retrieval time itself in all these events to be in the order of 1/10 of a second for libraries in the range of 1 to 100 millions strands in a wet tube.

It is worth noticing that similar results may be expected for memory updates. Adding a record is straightforward in DNA-based memories (assuming that the new record is noncrosshybridizing with the current memory), one can just drop it in the solution. Deleting a record requires making sure that all copies of the records are retrieved (full stringency for perfect recall) and expunged, which reduces deletion to the problem above. Additional experiments were performed that verified this conclusion. The problem of adding new crosshybridizing records is of a different nature and was not addressed in this project.

3.3 DNA-Based Memory Capacity

An issue of paramount importance is the capacity of the memories considered in this paper. Conventional memories and even memories developed with other technologies have impressive sizes despite apparent shortcomings such as address-based indexing and sequential search retrievals. DNA-based memories need to offer a definitive advantage to make them competitive. Candidates are massive size, associative retrieval, and straightforward implementation by recombinant biotechnology. We address below only the first aspect.

Baum [2] claimed that it seemed DNA-based memories could be made with a capacity larger than the brain, but warned that preventing undesirable crosshybridization may reduce the potential capacity of 4^n strands for a library made of n-mers. Later work on error-prevention has confirmed that the reduction will be orders of magnitude smaller [6]. Based on combinatorial constraints, [14] combinatorially obtained some theoretical lower bounds and upper bounds of the number of equilength DNA strands. However, from the practical point of view, the question still remains of determining the size of the largest memories based on oligonucleotides in effective use (20 to 150-mers).

A preliminary estimation of the runs has been made in several ways. First, a greedy search of small DNA spaces (up to 9-mers) in [10] by exhaustive searches averaged a number of 100 code words or less at a minimum h-distance apart of 4 or more, in a space of at least 4^{10} strands, regardless of the random order in which they the entire spaces were searched. Using the more realistic (but still approximate) thermodynamic model of Deaton et Al. [5], similar greedy searches turned up libraries of about 1,400 10-mers with nonnegative pairwise Gibbs energies (given by the model.) An *in vitro* selection protocol proposed by Deaton et Al. [4] has been tested experimentally and is expected to produce large libraries. The difficulty is that quantifying the size of the libraries obtained by the selection protocol is yet an unresolved problem given the expected size for 20-mers. In a separate experiment simulating this selection protocol, *Edna* has produced libraries of about 100 to 150 *n*-mers (n=10, 11, 12) starting with a full size DNA space of all *n*-mers (crosshybridizying) as the seed populations. Further several simulations of the selection protocol with random seeds of 1024 20-mers as initial population have consistently produced libraries of no more than 150 20-mers. A linear extrapolation to the size of the entire population is too risky because the greedy searches show that sphere packing allows high density in the beginning, but tends to add more strands very sparsely toward the end of the process. The true growth rate of the library size as a function of strand size *n* remains a truly intriguing question.

4 Summary and Conclusions

The reliability and efficiency of DNA-based associative memories has been explored quantitatively through simulation of reactions *in silico* on a virtual test tube. They show that there the region of optimal concentrations for library and probe strands to minimize retrieval time and avoid excessive concentrations (which tend to lengthen retrieval times) is about O(logN), where N is the size of the library. Further the retrieval time is highly dependent on reactions conditions and the probe, but tends to stabilize at optimal concentrations. Furthermore, these results remain essentially unchanged for simultaneous multiple queries if they remain small compared to the library size (within O(log N).) Previous benchmarks of the virtual tube provide a good level of confidence that these results extrapolate well to wet tubes with real DNA. The retrieval times in that case can be estimated in the order of 1/10 of a second. The important question of how the memory capacity grows as a function of strand size is certainly sub-exponential, but remains a truly intriguing open question.

An interesting possibility is suggested by the results presented here. The experiments were run in simulation. It is thus conceivable that conventional memories could be designed in hardware using special-purpose chips of the software simulations. The chips would run according to the parallelism inherent in VLSI circuits. One iteration could be run in nanoseconds with current technology. Therefore, once can obtain the advantages of DNA-based associative recall at varying threshold of stringency *in silico*, while retaining the speed, implementation, and manufacturing facilities of solid-state memories. A further exploration of this idea will be fleshed out elsewhere.

References

1. L.M. Adleman: Molecular Computation of Solutions to Combinatorial Problems. *Science* **266** (1994) 1021–1024
2. E. Baum, Building An Associative Memory Vastly Larger Than The Brain. *Science* **268** (1995) 583–585.
3. A. Condon, G. Rozenberg (eds.): DNA Computing (Revised Papers). In: Proc. of the 6th International Workshop on DNA-based Computers, 2000. Springer-Verlag Lecture Notes in Computer Science **2054** (2001)
4. R. Deaton, R., J. Chen, H. Bi, M. Garzon, H. Rubin, D.H. Wood. A PCR-Based Protocol for In-Vitro Selection of Non-Crosshybridizing Oligonucleotides (2002). In [11], 105–114
5. R.J. Deaton, J. Chen, H. Bi, J.A. Rose: A Software Tool for Generating Non-crosshybridizing Libraries of DNA Oligonucleotides. In [11], pp. 211–220.
6. R. Deaton, M. Garzon, R. E. Murphy, J. A. Rose, D. R. Franceschetti, S.E. Stevens, Jr. The Reliability and Efficiency of a DNA Computation. *Phys. Rev. Lett.* **80** (1998) 417–420
7. M. Garzon, D. Blain, K. Bobba, A. Neel, M. West: Self-Assembly of DNA-like structures i*n silico.* Journal of Genetic Programming and Evolvable Machines **4**:2 (2003), in press.
8. M. Garzon: Biomolecular Computation *in silico*. *Bull. of the European Assoc. For Theoretical Computer Science* EATCS (2003), in press.
9. M. Garzon, C. Oehmen: Biomolecular Computation on Virtual Test Tubes. In: N. Jonoska and N. Seeman (eds.): Proc. of the 7th International Workshop on DNA-based Computers, 2001. Springer-Verlag Lecture Notes in Computer Science **2340** (2002) 117–128
10. M. Garzon, R. Deaton, P. Neathery, R.C. Murphy, D.R. Franceschetti, E. Stevens Jr.: On the Encoding Problem for DNA Computing. In: Proc. of the Third DIMACS Workshop on DNA-based Computing, U of Pennsylvania. (1997) 230–237
11. M. Hagiya, A. Ohuchi (eds.): Proceedings of the 8th Int. Meeting on DNA Based Computers, Hokkaido University, 2002, Springer-Verlag Lecture Notes in Computer Science **2568** (2003)
12. J. Lee, S. Shin, S.J. Augh, T.H. Park, B. Zhang: Temperature Gradient-Based DNA Computing for Graph Problems with Weighted Edges. In [11], pp. 41–50.
13. R. Lipton: DNA Solutions of Hard Computational Problems. *Science* **268** (1995) 542–544
14. A. Marathe, A. Condon, R. Corn: On Combinatorial Word Design. In: E. Winfree and D. Gifford (eds.): DNA Based Computers V, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. **54** (1999) 75–89
15. J.H. Reif, T. LaBean. Computationally Inspired Biotechnologies: Improved DNA Synthesis and Associative Search Using Error-Correcting Codes and Vector Quantization In [3], pp. 145–172
16. K.A. Schmidt, C.V. Henkel, G. Rozenberg: DNA computing with single molecule detection. In [3], 336.
17. J.G. Wetmur: Physical Chemistry of Nucleic Acid Hybridization. In: H. Rubin and D.H. Wood (eds.): Proc. DNA-Based Computers III, U. of Pennsylvania, 1997. DIMACS series in Discrete Mathematics and Theoretical Computer Science **48** (1999) 1–23
18. E. Winfree, F. Liu, L.A. Wenzler, N.C. Seeman: Design and self-assembly of two-dimensional DNA crystals. *Nature* **394** (1998) 539–544