DNA as a Medium for Storing Digital Signals

Sotirios A. Tsaftaris[1], Vassily Hatzimanikatis[2] and Aggelos K. Katsaggelos[1]

[1] Department of Electrical Engineering and Computer Science
[2] Department of Chemical and Biological Engineering
Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208 USA
s-tsaftaris@northwestern.edu

Abstract

Motivated by the storage capacity and efficiency of the DNA molecule in this paper we propose to utilize DNA molecules to store digital signals. We show that hybridization of DNA molecules can be used as a similarity criterion for retrieving digital signals encoded and stored in a DNA database. Since retrieval is achieved through hybridization of query and data carrying DNA molecules, we present a mathematical model to estimate hybridization efficiency (also known as selectivity annealing). We show that selectivity annealing is inversely proportional to the mean squared error (MSE) of the encoded signal values. In addition, we show that the concentration of the molecules plays the same role as the decision threshold employed in digital signal matching algorithms. Finally, similarly to the digital domain, we define a DNA signal-to-noise ratio (SNR) measure to assess the performance of the DNA-based retrieval scheme. Simulations are presented to validate our arguments.

Introduction

The problem of searching in a database of digital signals can be described as follows. Consider an $M \times l_V$ array (set) V, with rows the $1 \times l_V$ digital signals v_i, each entry of which is a *k*–bit integer. This set V is hereafter termed the *digital database*. Consider also a vector q_d that contains $l_Q < l_V$, *k*–bit integers, hereafter termed the *digital query*. The problem at hand is to find out whether q_d can be found in V. Traditionally a matching criterion must be defined first that describes the similarity between the query and the digital signal at the location under examination [5]. Overall the goal is to find (a) a yes/no answer whether a match has been found (in essence the criterion is minimized and is lower than a user supplied threshold) and (b) the locations and the vector identity of such matches.

There are many criteria that can be used for matching and each of them offers distinct characteristics in terms of performance and computational cost. Examples of these criteria are the mean squared error (MSE), the sum of absolute differences, the sum of squared error, and weighted versions of them. Similarly there are different ways of searching within the database. Traditional correlation and convolution techniques can be used, although the computational cost should be considered. The complexity of the problem at best scales linearly with the size of V, but in most cases the complexity is a polynomial function of V. Although the complexity of the matching operation is low, it is usually the number of such operations dictated by the size of V that renders the problem difficult to implement for practical applications.

DNA offers significant advantages when compared to other media for storing digital signals or data, in general. The DNA molecule, especially in its double stranded form, is very stable, compact, and inexpensive. Polymerase Chain Reaction (PCR) is an economical and efficient way to replicate databases. Querying the database can be implemented with a plethora of techniques. In digital databases the query time increases proportionally to the size of the database. However, in DNA databases when annealing is used as a search mechanism, the querying time is independent of the database size when the target molecules (the molecules representing V in our case) have equal concentrations.

This article's main contribution is the analysis and simulation of a DNA-based database and retrieval mechanism, which mirror the digital world. Specifically, we show that the concentration of DNA molecules plays the same role as the decision threshold used in MSE based matching. Furthermore, similar to the digital domain, we define a signal to noise ratio (SNR) metric to quantify the performance of the DNA retrieval scheme.

We begin our presentation by defining the needed terminology on DNA chemistry and properties and sketch the characteristics of the equivalent DNA database system that can store digital signals in section 3. In section 4 we show how the performance can be estimated in terms of efficiency and sensitivity by modeling hybridization kinetics. In section 5 we offer simulation results to illustrate performance. Finally, in section 6 we conclude this article and show a direct application of our work in biotechnology.

DNA Equivalent

Our research is centered on providing a DNA-based alternative to the problem of querying a digital database. It is the compact nature of the DNA molecule that renders it

an attractive storage medium. Furthermore, the chemical structure of the DNA supplies us with hybridization, an extraordinary tool, which allows for the medium to be part of the computational platform, since data searches can be performed by utilizing it [1], [12].

A double helix of *DNA* (Deoxyribo-Nucleic Acid) is made from two single strands of DNA, each of which is a chain of *nucleotides* (or *bases*) A, G, T, and C. Nucleotides can be joined together in a linear chain to form a *single strand* of DNA. Each base in DNA has its unique *Watson-Crick complement*, which is formed by replacing every A with a T and vice versa, and every G with a C and vice versa. Every strand has a complementary sequence; for example, the complementary sequence of ATG is TAC. If two complementary sequences meet in a solution under appropriate conditions, they will attract each other and form a double stranded helical structure, the *duplex*. This process is called *hybridization* or *annealing*. *Specific hybridization*, refers to cases where the two single strands are perfectly complementary at every position and the double-stranded molecule that is formed is perfect, while *non-specific hybridization*, corresponds to cases with mismatched base pairs.

The first step towards defining a DNA system equivalent to a system implemented digitally is to map the digital information into DNA. The problem is also known as the codeword or word design problem. In our case the problem translates into finding N DNA sequences or words x_i, i=0,…,N–1 ($N = 2^k$), each of length l bases, capable of encoding integer signal values 0,…,N–1.

In most DNA computing applications only specific hybridizations are acceptable. In our case, we design DNA words such that the hybridization strength between them is inversely proportional to the absolute difference of the corresponding encoded integer signal values. To accomplish this, we introduced, the Noise (or inexact match) Tolerance Constraint (NTC) [9], [10]. This constraint and others are needed to ensure that only wanted duplexes will be formed. In other words, we want to minimize the possibility of formation of unwanted duplexes and maximize the possibility of wanted ones. In a laboratory setting this translates to minimizing the concentration of unwanted hybridizations while maximizing the concentration of wanted ones.

For simplicity let us assume that DNA sequences are constructed as in Figure 1, although many other different structures can be found [12]. The left part is the index that identifies the data, which appear on the right. The index part of different elements should be very dissimilar. For clarity we term these structures *database elements* (DE_j). Each DE_j is a single stranded DNA, it has concentration C_j, and is identified by a unique index (not shown). Furthermore it has a data payload of L bases ($L = l \cdot l_V$) hence it is capable of storing l_V signal values. Data are concatenations of DNA words.

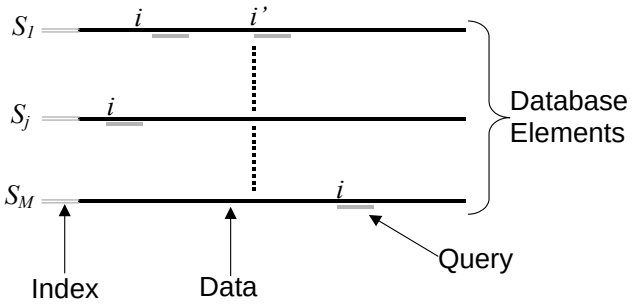

Figure 1: Illustration of hybridizations between query and *DEj*.

The system can be described with the following parameters and inputs:

- M DEs (we have M digital signals), each of concentration C_j and sequence information s_j of length L , j=1…M.
- We therefore need M indices. Each index has a length IN. The discussion on design requirements for the indices and their generation is omitted from this article for brevity.
- A query Q of length L_Q<L, shown in Figure 1 as solid gray line, of concentration $|Q|_o$.
- Temperature T and salt concentration (presently ignored).

The DNA database can be constructed by first mapping the digital signals into DNA sequences, then chemically synthesizing them. Finally, a DNA index is attached to each sequence and the sequences are placed in a test tube in a soluble state [12]. Information capacity is a critical component of a database although in many cases capacity is a function of speed and accuracy. The study of information capacity of DNA databases has to be coupled with the accuracy of the database. Although information theoretic limits and bounds can be found it is of special interest to find the capacity of a database as a function of database size (number and length of elements), volumetric space (the volume of the database), and the error rate. Knowing the volumetric space and database size we can estimate the concentration of the database elements. There is a limit on the concentration for each element imposed by the laboratory methods employed for database management and information extraction. For example if PCR is for information extraction the minimum concentration allowed depends on the length of the amplified product [3], [4]. It is therefore very hard at this point to estimate the capacity of the database since it depends on the length of the DNA words l, the number of DEs M, and the laboratory protocols used parameters that are constantly changing.

The scalability of the database depends on the available indices. If the length of each index is IN then the

maximum available number of indices is 4^{IN}. In practice this number is much smaller due to certain impose constraints. If the indices used are already reaching the limits for the given index length adding new database elements is rather hard without redesigning the whole database. On the other hand if the available indices are not exhausted the addition of a new database element is rather easy, since it requires only the synthesis of the new database element.

When a search is needed a protocol similar to the following can be employed:

1. Find Q as the DNA mapping of the digital query q_d.
2. Synthesize the complement of Q.
3. Take a sample of the solution.
4. Add the query strand.
5. Cool down to allow hybridization between a database element and the query strand.
6. Detect the hybridization event using one of several spectroscopic techniques, i.e., fluorescent labels attached to the query strand.

In the case where a simple yes or no answer to whether the query can be found in the database was needed, a change in the fluorescent response will indicate success. Q once inserted in the test tube will try to hybridize to the most favorable and stable locations to form complexes [12]. Stability is a function of sequence information, concentrations, and reaction parameters. Therefore the outcome of a search can change dramatically when varying the above parameters. Hence, it is critical to quantify the hybridization behavior.

To simplify our presentation we introduce the notion of fragments. A fragment F_{ip}^j represents the DNA sub-sequence of DE_j at location i of length p with concentration $\left|F_{ip}^j\right|$. It is clear that F_{ip}^j is a subset of s_j. Furthermore, in our case the initial concentration of F_{ip}^j denoted by $\left|F_{ip}^j\right|_o$ is equal to C_j.

Subsequently we can model the query fragment complexes as QF_{ip}^j. Such complexes are illustrated in Figure 1 at various locations. Without loss of generality, to ease our analysis and reduce the complexity we will assume that (i) $p=L_Q$, (ii) complexes will only have internal mismatches (none in the terminal or penultimate positions). Due to (i) and (ii), the total number of complexes is $N_T = M \cdot L / L_Q$. Since p is the same for all cases, we will drop it from our notation. The kinetic analysis will allow us to estimate $\left|QF_i^j\right|$, which can be used to assess whether DNA hybridization can be used as a matching criterion.

Modeling Hybridization Reactions

Equilibrium solution

To estimate $\left|QF_i^j\right|$ we first have to model the hybridization reaction between a query molecule Q and a fragment F_i^j by

$$Q + F_i^j \underset{K_r}{\overset{K_f}{\rightleftharpoons}} QF_i^j, \quad \forall i, j. \tag{1}$$

Eq. (1) can be represented mathematically by time dependent differential equations, the solution of which requires knowledge of the reaction rates K_f and K_r, which can only be estimated through laboratory experiments and in general are not universal. Usually an equilibrium analysis is adopted that renders the above equations time-independent [4]. In equilibrium the following relation between the reaction rates and concentrations holds

$$K_i^j = \frac{K_f}{K_r} = \frac{\left|QF_i^j\right|}{\left|Q\right| \cdot \left|F_i^j\right|} = \exp\left(-\frac{\Delta G_i^j}{R \cdot T}\right), \tag{2}$$

where ΔG is the Gibbs free energy, R is the Boltzman constant, and T is the temperature in Kelvin. The Gibbs free energy for DNA complexes can be estimated using parameters available in the literature [4], [8], which are a function of the sequences Q and F_i^j.

The mass-conservation equation on the query is

$$\left|Q\right|_o = \left|Q\right| + \sum_{i,j}\left|QF_i^j\right|, \tag{3}$$

where $\left|Q\right|$ is the concentration of the free (un-hybridized) query. The sum is over N_T terms.

Likewise, utilizing the mass-conservation equations for each fragment we have

$$\left|F_i^j\right|_o = \left|F_i^j\right| + \left|QF_i^j\right|, \quad \forall i, j. \tag{4}$$

Our goal is to find $\left|QF_i^j\right|$ from the system of Eqs. (3) and (4). From equations (2) and (4) we have:

$$\left|QF_i^j\right| = \left|F_i^j\right|_o - \left|F_i^j\right| = \ldots = \frac{\left|F_i^j\right|_o \cdot K_i^j \cdot \left|Q\right|}{1 + K_i^j \cdot \left|Q\right|}. \tag{5}$$

In the above equation the only unknown is $\left|Q\right|$. Combining Eqs. (3), (5) and setting $q = \left|Q\right| / \left|Q\right|_o$, after some manipulation we have [13]

$$\overbrace{\sum_{i,j} \frac{\left(\left|F_i^j\right|_o / \left|Q\right|_o\right) \cdot q}{\left(1/\left(\left|Q\right|_o K_i^j\right)\right) + q}}^{h(q)} = \overbrace{1 - q}^{g(q)}. \tag{6}$$

Thus, the problem of determining $\left|QF_i^j\right|$ is equivalent to finding the roots of $f(q)=h(q)-g(q)$, since given q each $\left|QF_i^j\right|$ can be finally estimated by substituting $\left|Q\right| = q \cdot \left|Q\right|_o$

30	13	14	12	5	20	1	30	28	22	22	17	10	22	5	28	16	21	11	18
24	30	29	27	7	9	24	15	17	27	13	23	7	10	23	28	29	27	10	24
6	30	2	1	7	7	15	14	7	1	27	14	7	18	13	19	27	22	11	10

Figure 2: The values of database V.

in Eq. (5). Based on the intermediate value theorem, we have shown that there exists a unique q_s in [0,1] such that $f(q_s)=0$ [13]. Since a solution cannot be found analytically, instead a solution q_B can be found computationally such that $|q_B - q_s| \leq \varepsilon$ using any root-finding method [6].

Estimating Query Selectivity

Selectivity is a term commonly used in analytical chemistry [14]. In our case we are interested in the selectivity of outcomes from annealing reactions. Annealing selectivity is defined as

$$SA_i^j = \frac{\left|QF_i^j\right|}{\sum_{i,j}\left|QF_i^j\right|}. \qquad (7)$$

This dimensionless expression can be seen as the percentage of the complex $\left|QF_i^j\right|$ within all the hybridized complexes or as the probability of such hybridization event. Since in our case we only have QF_i^j type of complexes annealing selectivity can be seen as query selectivity. (Query selectivity can also be found in the literature of designing and optimizing digital databases.) Selectivity can also be used as an indication of matching efficiency, as we will see in the next section.

Another critical component for evaluation is the selectivity per database element (per-DE), which can be defined as

$$SA^{j'} = \sum_i SA_i^{j'}, \qquad (8)$$

essentially indicating the percentage of a particular retrieved database element.

Criteria for Comparison

Similar to [7] we can define the signal to noise ratio as

$$SNR = \frac{\sum_{i,j \in \text{desired}} SA_i^j}{\sum_{i,j \in \text{un-desired}} SA_i^j}. \qquad (9)$$

In our case desired hybridizations QF_i^j are those for which the MSE of their corresponding signal values is less or equal than a threshold T_P, while un-desired hybridizations are all the rest. In addition we can define the error E involved as

$$E = \frac{1}{1+SNR}. \qquad (10)$$

The error is very useful in estimating capacity as mentioned earlier.

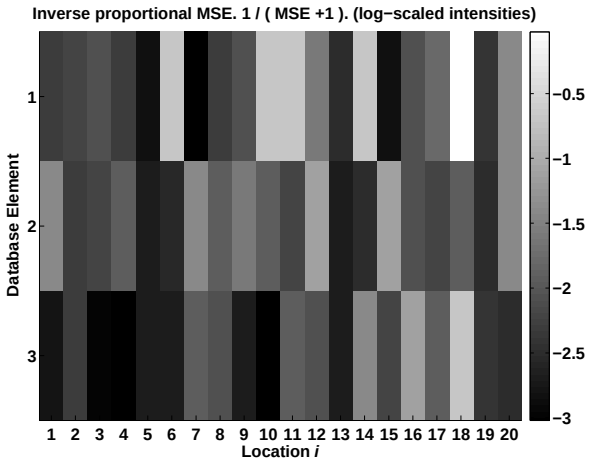

Figure 3: Inverse proportional MSE between V and query {21}. ('White' indicates MSE (min) = 0, while 'black' indicates MSE (max) =400.)

Simulation Results

For our simulations we developed MATLAB routines to estimate $\left|QF_i^j\right|$ as presented above. From a set of experiments we chose: (i) the database V (M=3, l=20, tabulated in Figure 2), (ii) the digital query q_d={21}, and (iii) signal range [0,…,31] (k=5). This set-up will emulate situations where a single word query is used to search inside a database.

We converted the database to a DNA equivalent using the set of 32 words of length 19 presented in [11] and found their equilibrium constants at T=60°C as mentioned in section 3.1. The statistics of their distribution are maximum constant 1.24E17, minimum 2.56, and σ=1.60E16. The words are optimized using the NTC (see section DNA Equivalent) to have large equilibrium constants if their corresponding signal differences are less than T_P=4. In our estimation we ignored cases where a word hybridizes partially with a word and its neighbor. This is driven from the fact that our words are designed to avoid such mishybridizations (see [9], [11] for more details). In all simulations the initial concentration of fragments $\left|F_i^j\right|_o$ was constant $C = 10^{-5}\ {}^{mol}/_{Liter}$. The query concentration $\left|Q\right|_o$ was varying multiples of C.

In Figure 3 we show as a pseudo-gray image $1/(MSE+1)$, between q_d and V. In Figure 5 we show in four pseudo-gray images the selectivity SA_i^j for $\left|Q\right|_o$ equal to, $1/10\,C$ in (a), C in (b), $10C$ in (c) and $100C$ in (d). By comparing Figure 3 with each of the sub-plots in Figure 5 we can see that for low query concentrations the selectivities resemble the inverse MSE of Figure 3,

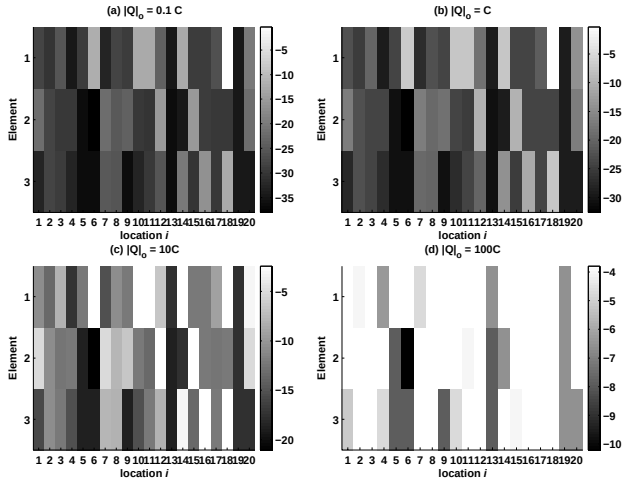

Figure 4: SA_i^j (*j, i* as y- and x-axis respectively) as pseudo images for four query concentrations. ('White' indicates large selectivity, while 'black' indicates small.)

specifically graph (c) is close to Figure 3. As $|Q|_o$ increases the separation between the elements is not adequate. Furthermore, we can see that F_{18}^1 (which corresponds to 21 in *V*) corresponds to a digital value equal to q_d (21 in our case) and hence the *MSE* is equal to zero. SA_{18}^1 is therefore always the highest selectivity. We observe however, that as the query concentration increases the sensitivity of the system decreases and more 'similarity' is allowed, hence the more 'white' in Figure 5(d).

The per-DE selectivities found by Eq. (8) are shown in Figure 5(a). We can see that SA^1, is dominant in all cases since it contains F_{18}^1. However we notice that the selectivity SA^3 of DE_3 of, which contains F_{18}^3 (the next smallest *MSE* is equal to 1), is initially small but it increases as $|Q|_o$ increases. Unfortunately this comes at the expense of *SNR* and *E* (since they are defined for all the elements in the database), as we can see in Figure 5(b), that is *SNR* decreases while *E* increases as the query concentration increases. We repeated the previous experiments with a two sample digital query q_{d2}={21 11} and compared its performance with the previous single word scenario. We used the same integer to DNA sequence mapping. We calculated the equilibrium constants between all possible pairs of database-query duplexes at T=60°C as before. The statistics of their distribution are maximum equilibrium constant equal to 2.69E20, minimum equal to 1.49E01, and σ equal to 3.56E19. We shall mention that we only took into consideration duplexes of length 38 bases long equal to the length of the query. It is possible that a query may partially hybridize to only one corresponding word in the database and leave the rest of the query molecule

(a)

	$\|Q\|_o$			
j'	$\frac{1}{10}C$	C	$10C$	$100C$
$SA^{j'}$ 1	9.99E-01	9.99E-01	5.43E-01	3.80E-01
$SA^{j'}$ 2	4.70E-07	1.11E-04	2.19E-01	3.42E-01
$SA^{j'}$ 3	2.72E-06	6.40E-04	2.38E-01	2.78E-01

(b)

SNR	5.29E+09	2.25E+07	1.92E+02	4.98E-01
E	1.89E-10	4.45E-08	5.19E-03	6.68E-01

Figure 5: (a) Eq. (8), (b) *SNR* and *E* for various $|Q|_o$ with query {21}.

overhanging. Such events are possible and are motivated by thermodynamics and molecular kinetics and energy minimization. Recall that the DNA molecules will seek the least energy consuming conformation. We refrained from such analysis for brevity.

In agreement with our previous experiment the initial concentration of fragments was $C = 10^{-5}\ ^{mol}/_{Liter}$. The query concentration $|Q|_o$ was varying as a multiple of *C*.

Similarly to Figure 3 we show as a pseudo-gray image the inverse MSE, between our new q_d and V in Figure 7. As before in Figure 8 we show in four pseudo-gray images the selectivity SA_i^j for various initial concentrations $|Q|_o$. By comparison we see, as in the previous experiment, the resemblance between the two figures, which illustrates that the method can be extended to bigger query sizes. We can see that F_{18}^1 (which corresponds to the consecutive samples {21 11} in *V*) corresponds to a digital value equal to our two-word query q_d and therefore the *MSE* is equal to 0. SA_{18}^1 is therefore always the highest selectivity. Again we can see that the sensitivity of the system can be affected by controlling the query concentrations.

The per-DE selectivity found by Eq. (8) are shown in Figure 6(a). We can see that SA^1, is dominant in all cases since it contains F_{18}^1. The system behaves differently with longer queries. This can be seen by examining the ratio SA^2 / SA^3 from the data in Figure 5(a) and Figure 6(a).

With shorter queries the ratio is much bigger when compared to longer queries. The explanation is that while in the single query case DE_2 contained desired hybridizations in the two-word case it does not. (We shall mention that the threshold utilized here to separate desired from undesired hybridizations is the same since our matching criterion is MSE. The MSE criterion is independent of the query length due to its averaging element.)

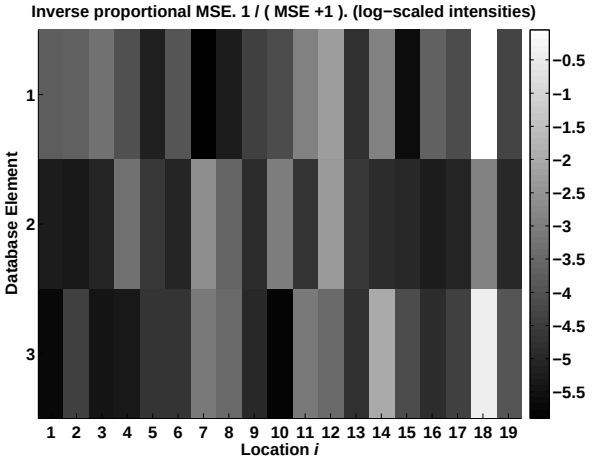

Figure 7: Inverse proportional MSE between V and query {21 11}. ('White' indicates MSE (min) = 0, while 'black' indicates MSE (max) =380.5.)

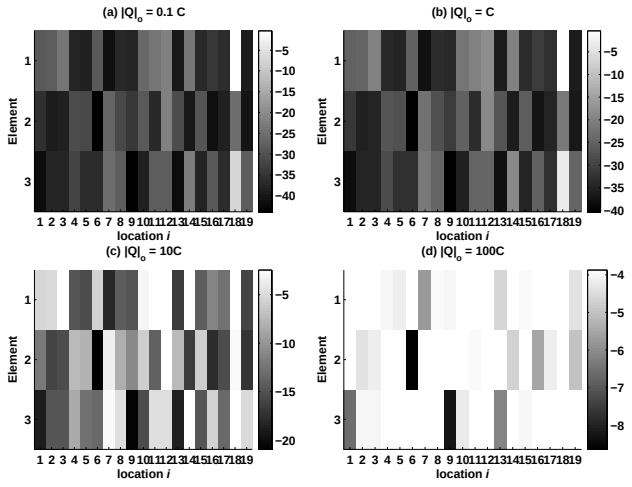

Figure 8: SA_i^j (j, i as y- and x-axis respectively) as pseudo images for four query concentrations. ('White' indicates large selectivity, while 'black' indicates small.).

Actually DE_2 contains the fragment F_{12}^2 (undesired hybridization MSE=10) that when hybridized with the query it creates a more stable bond than the bond formed between the query and F_{14}^3 (desired hybridization MSE=6.5). In other words it has a larger equilibrium constant ($K_{12}^2 > K_{14}^3$). This creates a false positive, whose effects are obfuscated due to the presence of other more stable pairs. On the contrary let us consider an example database where only two DEs are present (DE_1, DE_2) and DE_1 includes the fragment represented by F_{12}^2 while DE_2

(a)

	$\lvert Q\rvert_o$			
j'	$\frac{1}{10}C$	C	$10C$	$100C$
$SA^{j'}$ 1	9.99E-01	9.75E-01	5.12E-01	3.52E-01
$SA^{j'}$ 2	2.85E-10	1.04E-08	1.94E-01	3.20E-01
$SA^{j'}$ 3	6.74E-04	2.40E-02	2.92E-01	2.92E-01

(b)

SNR	2.11E+09	5.79E+07	6.46E-01	9.46E-02
E	4.71E-10	1.72E-08	6.07E-01	9.13E-01

Figure 6: (a) Eq. (8), (b) *SNR* and *E* for various $\lvert Q\rvert_o$ with query {21 11}.

includes the fragment represented by F_{12}^2. If F_{14}^3 is the only desired hybridization and then we would have had a false positive since DE_2 would have been retrieved.

Such errors are expected since the codewords are designed for single word interactions (so that the codeword design problem is computationally tractable) and not multiple word ones.

Conclusions

In this article we have shown that DNA hybridization can be used as the DNA equivalent of a digital matching criterion when developing DNA databases capable of storing digital signals. Such databases offer significant advantages over digital databases since they are much more compact and require less maintenance. Our simulations showed that at low query concentrations hybridizations are capable of retrieving data from the database that are similar to the query. Furthermore we can control the sensitivity and accuracy of the database by adjusting the concentration. We showed also that larger queries could be used. At the same time we highlighted the need for developing codeword design algorithms that are capable of reducing false positives when multiple word queries are used.

Our goal is to offer a demonstrational small scale *in vitro* DNA database, but in order to reduce laboratory costs we have developed the simulators employed here. Evidence of the simulations leads us to believe that such a system may not be long from becoming a reality. Our work is in fact inspired by nature since we are utilizing hybridization, one of the most fundamental properties of DNA. All the information needed to create an organism is stored into DNA. Evolution has created this remarkable molecule for that purpose. It is well worth investigating the potential of using the same molecule to store information other than biological.

This work can also be applied in designing, or estimating the performance of sequences used as primers or probes in Polymerase Chain Reactions or Microarray experiments. These techniques are commonly used, for example, in amplifying and separating genomic material and in identifying genetic diseases.

Acknowledgements

Mr. Tsaftaris would like to thank the *Alexander S. Onassis* public benefit foundation for their financial support.

References

[1] Baum, E. B. 1995. Building an associative memory vastly larger than the brain. Science 268(210): 583-585.

[2] Dieffenbach, C.W.; and Dveksler, G. S. 1993. Setting up a PCR Laboratory. PCR Methods Applications 3(2):S1-S7.

[3] Dragon, E. A. 1993. Handling reagents in the PCR laboratory. PCR Methods Applications 3(2):S8-S9.

[4] Hammes, G. G. 2000. Thermodynamics and Kinetics for the Biological Sciences. New York: John Wiley & Sons.

[5] Haralick, R. M.; and Shapiro, L. G. 1993. Computer and Robot Vision. Reading, Mass: Addison-Wesley, vol. II, Ch. 16.

[6] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; and Flannery, B. P. 1992. Numerical Recipes in C. Cambridge Univ. Press, Ch. 9.

[7] Rose, J. A.; Deaton, R. J.; and Suyama, A. 2005. Statistical thermodynamic analysis and design of DNA-based computers. International Journal of Natural Computing 3: 443-459.

[8] SantaLucia, Jr. J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest–neighbor thermodynamics. Proceedings National Academy Sciences. USA 95(4): 1460–1465.

[9] Tsaftaris, S. A.; Katsaggelos, A. K.; Pappas, T. N.; and Papoutsakis, E. T. 2004. DNA based matching of digital signals. In Proceedings IEEE International Conference on Acoustics Speech and Signal Processing Montreal, Quebec, Canada. 581-584.

[10] Tsaftaris, S. A.; Katsaggelos, A. K.; Pappas, T. N.; and Papoutsakis, E. T. 2004. How can DNA-computing be applied in digital signal processing? IEEE Signal Processing Magazine. 21(6):57-61.

[11] Tsaftaris, S. A.; and Katsaggelos, A. K. 2005. A New Codeword Design Algorithm for DNA-Based Storage and Retrieval of Digital Signals. In Pre-Proceedings of the 11th International Meeting on DNA Computing, June 6-9, London, Canada, 2005.

[12] Tsaftaris, S. A.; and Katsaggelos, A. K. 2005. On Designing DNA Databases for the Storage and Retrieval of Digital Signals. Lecture Notes in Computer Science 3611: 1192-1201.

[13] Tsaftaris, S. A.; Hatzimanikatis, V.; and Katsaggelos, A. K. 2005. *In silico* estimation of annealing specificity of query searches in DNA databases. Journal of Japan Society of Simulation Technology (JSST) special issue "Application and Simulation of DNA Computing". in press.

[14] Vessman, J. *et al.* 2001. Selectivity in analytical chemistry. Pure Applied Chemistry. 73(8): 1381–1386.