DNA HYBRIDIZATION AS A SIMILARITY CRITERION FOR QUERYING DIGITAL SIGNALS STORED IN DNA DATABASES

S. A. Tsaftaris[1], V. Hatzimanikatis[2], and A. K. Katsaggelos[1]

[1] Department of Electrical Engineering and Computer Science
E-mail: {stsaft, aggk}@ece.northwestern.edu
[2] Department of Chemical and Biological Engineering
E-mail: vassily@northwestern.edu
Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208 USA

ABSTRACT

We demonstrate via simulation that hybridization of DNA molecules can be used as a similarity criterion for retrieving digital signals encoded and stored in a synthesized DNA database. After introducing some necessary DNA terminology, we briefly explain how digital signals are transformed to DNA sequences. 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, similar 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.

1. 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 [3]. 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.

Our research is centered on providing a DNA-based alternative to the problem of storing a digital database. It is the compact nature of the DNA molecule that renders it an attractive long-term and compact 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], [10]. In addition there is plethora of tools available to retrieve information from the database [10].

This article's main contribution is the analysis and simulation of a DNA-based database and retrieval mechanism, which mirrors 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 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.

2. DNA EQUIVALENT

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) [7], [8]. 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 Fig. 1, although many other different structures can be found [10]. 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.

The system can be described as follows:

- M DEs (we have M digital signals), each of concentration C_j and sequence information s_j of length L, j=1…M.
- A query Q of length L_Q<L, shown in Fig. 1 as solid gray line, of concentration $\left|Q\right|_o$.
- Temperature T and salt concentration (presently ignored).

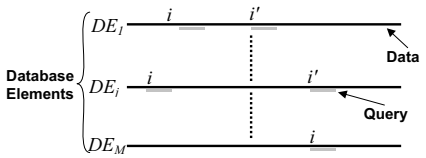

Fig. 1. Illustration of hybridizations between query and *DEj*.

The DNA database can be constructed by first mapping the digital signals into DNA sequences, then chemically synthesizing them. Finally the sequences are placed in a test tube in a soluble state [10]. When a search is needed, Q is the DNA mapping of the digital query q_d. Once Q is inserted in the test tube it will try to hybridize to the most favorable and stable locations to form complexes [10]. The event of stable complex formations can be easily detected in the laboratory to give a yes/no answer. 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 subsequence of DE_j at location i of length p with concentration $\left|F_{ip}^j\right|$. It is clear that $F_{ip}^j \subseteq s_j$. 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 Fig. 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 $N_T = M \cdot L / L_Q$. As a result, we will drop p 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.

3. MODELING HYBRIDIZATION REACTIONS

3.1. 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

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

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

[2]. In equilibrium the following relation between the reaction rates and concentrations holds

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

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 [2], [6], which are a function of the sequences Q and F_i^j.

The mass-conservation equation on the query is

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

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

$$\left|F_i^j\right|_o = \left|F_i^j\right| + \left|QF_i^j\right|, \quad \forall i, j. \quad (4)$$

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:

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

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 [11]

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

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 $\left|Q\right| = q \cdot \left|Q\right|_o$ in Eq. (5). Based on the intermediate value theorem, we have shown that there exists a unique q_s in [0,1] such that $f(q_s)=0$ [11]. 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 [4].

3.2. Query selectivity

Selectivity annealing is defined as

$$SA_i^j = \left|QF_i^j\right| \Big/ \sum_{i,j}\left|QF_i^j\right|. \quad (7)$$

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. It can also be used as an indication of matching efficiency as we will see in section 4.

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**

Fig. 2. The values of database V.

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

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

essentially indicating the percentage of a particular retrieved database element.

3.3. Signal to Noise Ratio

Similar to [5] we can define the signal to noise ratio as

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

In our case desired hybridizations QF_i^j are those for which the difference 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 $E=(1+SNR)^{-1}$.

4. 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 Fig. 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 [9] 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 (section 2) 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 [7], [9] for more details). In all simulations the initial concentration of fragments was $\left|F_i^j\right|_o = C = 10^{-5}\ {}^{mol}/_{Liter}$. The query concentration $\left|Q\right|_o$ was varying multiples of C.

In Fig. 3 we show as a pseudo-gray image $1/(MSE+1)$, between q_d and V. In Fig. 4 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 Fig. 3 with each of the sub-plots in Fig. 4 we can see that for low query concentrations the selectivities resemble the inverse MSE of Fig. 3, specifically graph (c) is close to Fig. 3. As $\left|Q\right|_o$ increases the separation between the elements is not

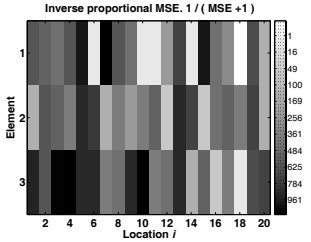

Fig. 3. Inverse proportional MSE for *V*. ('White' indicates MSE (min) = 0, while 'black' indicates MSE (max) =961.)

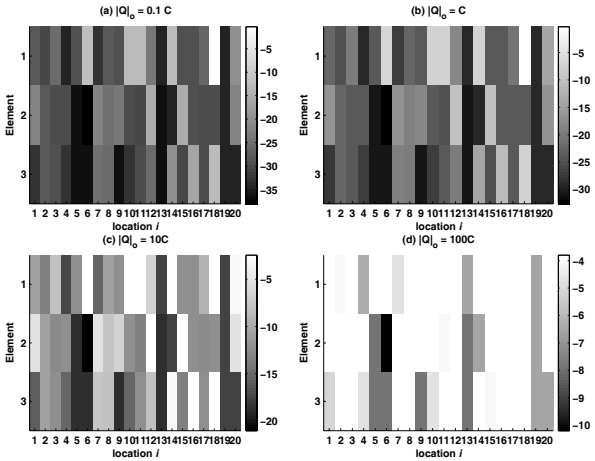

Fig. 4. SA_i^j (*j, i* as y- and x-axis respectively) as pseudo images for four query concentrations. ('White' indicates large selectivity, while 'black' indicates small.)

adequate. Furthermore, we can see that F_{18}^1 (=21 in *V*) corresponds to a digital value equal to q_d (=21) and hence the *MSE*=0. 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 Fig. 4(d).

The per-DE selectivities found by Eq. (8) are shown in Fig. 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*=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 Fig. 5(b), that is *SNR* decreases while *E* increases as the query concentration increases.

5. CONCLUSION

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. This leads us to believe that such a system may not be long from fruition. Such databases offer significant advantages 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. 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.

(a)

		$\|Q\|_o$			
	j'	$1/100\,C$	$1/10\,C$	C	$10C$
$SA^{j'}$	1	1.00E+00	9.99E-01	5.43E-01	3.80E-01
	2	4.70E-07	1.11E-04	2.19E-01	3.42E-01
	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

Fig. 5. (a) Eq. (8), (b) *SNR* and *E* for various $|Q|_o$.

6. REFERENCES

[1] E. B. Baum, Building an associative memory vastly larger than the brain, *Science*, vol. 268, no. 5210, pp.583-585, 1995.

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

[3] R. M. Haralick and L. G. Shapiro, "Image Matching," in *Computer and Robot Vision*, Reading MA: Addison-Wesley, vol. II, Ch. 16,1993.

[4] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Root Finding and Nonlinear Sets of Equations" in *Numerical Recipes in C*, Cambridge Univ. Press, ch. 9, 1992.

[5] J. A. Rose, R. J. Deaton, and A. Suyama, "Statistical thermodynamic analysis and design of DNA-based computers," *Int. J. of Nat. Comput.*, vol. 3, pp. 443-459, 2005.

[6] J. SantaLucia, Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest–neighbor thermodynamics," *Proc. Natl. Acad. Sci.*, USA, vol. 95, no. 4, pp. 1460–1465, 1998.

[7] S. A. Tsaftaris, A. K. Katsaggelos, T. N. Pappas, and E. T. Papoutsakis, "DNA based matching of digital signals," in *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing,* vol. 5, Montreal, Quebec, Canada, pp. 581-584, 2004.

[8] S. A. Tsaftaris, A. K. Katsaggelos, T.N. Pappas, and E.T. Papoutsakis, "How can DNA-computing be applied in digital signal processing?", *IEEE Signal Processing Mag.*, vol. 21, no. 6, pp. 57-61, 2004.

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

[10] S. A. Tsaftaris and A. K. Katsaggelos, "On Designing DNA Databases for the Storage and Retrieval of Digital Signals," *Lecture Notes in Comput. Sci.*, vol. 3611, pp. 1192-1201, Jul 2005.

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