DNA-BASED MATCHING OF DIGITAL SIGNALS

S. A. Tsaftaris[1], A.K. Katsaggelos[1], T.N. Pappas[1] and T.E. Papoutsakis[2].

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

ABSTRACT

Adleman with his pioneering work set the stage for the new field of bio-computing research [1]. His main idea was to use actual chemistry to solve problems that are either unsolvable by conventional computers, or require an enormous amount of computation. The main focus of our research is to consider the application of molecular computing to the domain of digital signal processing (DSP). In this paper we consider matching problems that arise in signal processing applications and are amenable to a DNA-based solution. Digital data are encoded in DNA sequences using a sophisticated codeword set that satisfies the Noise Tolerance Constraint (NTC) that we introduce. NTC, one of the main contributions of our work, takes into account the presence of noise in digital signals by exploiting the annealing between non-perfect complementary sequences. We propose an algorithm to map binary values into DNA codewords by satisfying a number of constraints, including the NTC. Using that algorithm we retrieved 128 codewords that enables us to use a DNA based approach to digital signal matching.

1. INTRODUCTION

Matching of digital signals is a fundamental problem that arises in many signal-processing applications. It can be used as a search or classification mechanism by quantifying the similarity between signals. For example, a search mechanism is essential for database retrieval, and signal classification is a key system component in data mining applications. Furthermore, a number of important problems in signal and image processing, like motion and disparity estimation, are matching type problems.

One of the key characteristics of the digital signal-matching problem is the fact that the matches are typically imprecise, due to the nature of the problem or the presence of noise in the data, which is almost always inevitable. Thus, to solve the matching problem it is necessary to quantify the similarity between signals.

Another key characteristic of the digital signal-matching problem is the enormous amount of computation that is typically required to find the optimal solution. It is thus of critical importance to find efficient implementations that can provide optimal or near optimal solutions. An additional requirement is the scalability of the solution.

In this paper we demonstrate the feasibility of using DNA techniques to solve digital signal matching problems. We address for the first time the presence of noise and inexact matches in digital data, and provide the necessary framework that utilizes the characteristics of the DNA molecules to overcome the associated problems.

The constraints routinely used in the field of bio-computing are adequate for most existing applications, for which non-specific hybridization is generally not desirable. In contrast, non-specific hybridization can be very useful (and is actually necessary) in signal processing applications, where an exact match is typically not possible. Therefore, there is a need to develop a codeword design scheme that will successfully address the problem of imperfect matching of digital signal data. As presented in [4], base mismatches affect the thermodynamic stability of duplexes. By carefully placing mismatches and assigning signal values to DNA codewords, we can inherently take care of the imperfect matches by controlling the temperature of the hybridizations. To model hybridization efficiency we use the melting temperature instead of Hamming distance, since for the size of codewords we are interested in, melting temperature is a better descriptor.

2. DNA-BASED SIGNAL MATCHING

The DNA composition and aqueous-solution chemistry can be used to solve matching type problems [1],[2]. DNA sequences are formed using four nucleotides (bases), adenine (A), thymine (T), guanine (G), and cytosine (C). The chemical structure of these nucleotides allows for unique pairing between A-T and G-C. Every

DNA sequence has chemically distinct ends known as 5' and 3'.

A DNA sequence (single strand) has its unique Watson-Crick complement (single strand). This can be done by replacing every A with a T and vice versa, and every G with a C and vice versa. If two complementary sequences meet in a solution under appropriate conditions they will hybridize and form a double stranded sequence. This hybridization or annealing is called specific, in contrast to non-specific, which takes place when two sequences not completely complementary (thus containing mismatches) hybridize.

Solution of matching problem can be achieved utilizing the annealing or hybridization property of DNA, which provides the matching mechanism. DNA annealing is one of the characteristic properties of DNA that is indispensable for the function of all biological life.

What makes DNA annealing particularly suited for the matching problem is the fact that it depends only on the concentration of elements and is independent of the total number of elements. This is a very important difference with traditional digital databases, in which the search time depends on the size of the database. It is this property of DNA annealing, combined with its high compactness, which makes DNA an excellent information storage medium.

Although in [5] we considered a more complicated matching scenario for simplicity the following matching problem will be considered. Assuming the presence of two signals S_1 and S_2 of length N and $K<N$ respectively, with signal values in the range *[0,255]* we want to find 'best' matches (locations) of the signal S_2 in S_1. In signal processing this has been a common problem and most commonly the mean squared error is used as a minimization criterion to identify the 'best' matches.

The same problem can be solved with DNA chemistry if we assume the presence of a function that maps digital data into DNA sequences. The *target* is the DNA sequence representation of S_1 and the *probe* is the complementary DNA sequence representation of S_2. When the *target* and the *probe* meet in a solution they will hybridize at the position that has the maximum thermodynamic stability. The stability depends on the reaction temperature and the solution's salt concentration. Although the benefit of using a DNA approach may not seem clear now, imagine the case of different target sequences in the solution and the problem of detecting the presence of a unique probe in the whole set. The benefit of using DNA is obvious in this case since the "execution time" will not change because such a reaction depends only on the concentration of the sequences and not on their number. It is therefore easy to see the extension of this work to DNA based databases of digital data where one needs to search for a pattern (probe) in a whole set of data (targets).

2.1. Codeword design

The first challenge towards a DNA-based solution of any computational problem is the coding of the digital data into DNA sequences. Thus, codeword design for DNA computation has been a very active research topic [2]. During the codeword design process, certain constraints need to be enforced for improving the fidelity of DNA based computation. DNA hybridization depends on parameters such as salt and DNA concentrations and temperature. If the temperature is not appropriately chosen, then two strands that are expected to hybridize will not, resulting in an error. Also an error results if two sequences hybridize when they are expected not to. In order to describe these constraints we first introduce the appropriate notation.

We denote a DNA sequence of length l in a *5'* to *3'* direction taking values from the alphabet *{A,T,G,C}* by x_l. x_l^C is the Watson-Crick complement of the sequence in the *3'* to *5'* direction. Reading a sequence from right to left can form the reverse of a sequence, x_l^R.

The following (standard) constraints on codewords have been proposed and used extensively [2]. All constraints refer to codewords of the same length l. The constraints are divided in two groups. The self-constraints and the group constraints:

Self-Constraints depend only on the codeword under examination and are:

• **Consecutive bases**. In some applications consecutive occurrences of the same base increase the number of errors

• **Self-Complementarity.** A codeword must not be self-complementary.

• **The GC content constraint**. The ratio of the sum of occurrences of G and C bases in a codeword and the length of the codeword must lie in a certain range.

Group constraints depend on the codeword and the rest of the codewords and are:

• **Hamming distance**. For all possible distinct pairs of codewords x_l, w_l the Hamming distance $d_H(x_l, w_l)$ (the number of base differences between two words) must be greater than a threshold E_H .

• **Reverse Complement**. For all possible distinct pairs of codewords x_l, w_l the following must be true $d_H(x_l^R, w_l^C) \geq E_{RC}$.

• **Frame-Shift**. If we denote the concatenation of two codewords x_l and w_l, by $x_l w_l$, then no other codeword $z_l \neq \{x_l, w_l\}$ can be found in the concatenation. In other words no codeword must result from the suffix of one codeword and the prefix of another.

• **Melting temperature** T_M. The melting temperature T_M of a duplex is defined as the temperature at which half of the

strands are in the double-stranded state. T_M for perfect and non-perfect (mismatch containing) duplexes can be estimated under some constraints based on a nearest neighbor model [4].

Melting temperature is typically used to estimate duplex stability and hybridization efficiency under given conditions. $T_M(x_l, w_l^C)$ denotes the melting temperature of the duplex formed by two DNA sequences, x_l in the 5' to 3' direction and w_l^C in the 3' to 5' direction. If we consider two signal values p_x, p_w with DNA representations x_l, w_l respectively, they will have melting temperature $T_M(x_l, w_l^C)$. Ideally we want to design the DNA codewords in such a way that the melting temperature between DNA codewords must be inversely proportional to the absolute difference of the encoded signal values. In the case when $p_x = p_w$, $T_M(x_l, w_l^C)$ must be maximized. To accomplish this, we introduce a new constraint, the *Noise (or inexact match) Tolerance Constraint (NTC),* that addresses the problem of assigning DNA codewords to neighboring signal values:

$$for \quad x_l = C(p_x) \quad and \quad w_l = C(p_w)$$

$$T_M(x_l, w_l^C) = \begin{cases} \text{maximized} & if \quad p_x = p_w \\ \propto \dfrac{1}{f(|p_x - p_w|) + \varepsilon} & if \quad |p_x - p_w| \le T_p \\ < T & if \quad |p_x - p_w| > T_p \end{cases}$$

where $f(x)$ represents any monotonically increasing function of x.

According to this constraint, duplexes formed by codewords assigned to neighboring signal values must have (similar) high melting temperature, while duplexes outside this neighborhood must have melting temperatures lower than a threshold T. The neighborhood is defined by the parameter T_P, which is a function of the amount of mismatch or noise in the data. The threshold T and the neighborhood range $[-T_P, +T_P]$ are the main parameters of the constraint that must be specified. The NTC replaces the Hamming distance and melting temperature constraints, and can be combined with the other constraints.

Another important aspect in codeword design and experimental setup is scalability (e.g., with increasing database size and number of codewords). Scalability has been a great challenge in all DNA computation efforts, and therefore, addressing it is critical.

3. ALGORITHMS FOR CODEWORD DESIGN

With the introduction of the NTC, the codeword design problem changes drastically. As we saw above, it must be reformulated so that it takes into account the new NTC constraint.

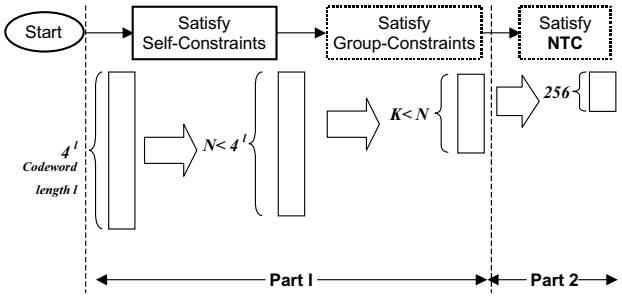

Figure 1. Block diagram of the algorithm.

Another very important design parameter is the codeword length, which must be kept small to reduce the overall DNA sequence length, and thus, cost.

The algorithm we developed accepts as input all the desired design parameters, such as, codeword length, desired constraints, and noise tolerance parameters. The output is an *n-bit* set of codewords that best satisfy the input constraints. As can been seen from Figure 1 the algorithm breaks down the problem into two independent tasks: the generation of codewords that satisfy the standard constraints listed above and the assignment of codewords to signal values.

In *Part 1* all possible codewords of a given length *l*, that satisfy the *self-constraints* are generated and placed in the *input* pool. To initiate the next step a codeword is randomly selected from the *input* pool and moved to the *output* pool. (Every time a codeword is selected from the *input* it is removed from it as well). Implementing the frame shift constraint a pruning step follows, according to which, all codewords from the *input* pool that have as suffix or prefix part of this codeword are eliminated.

After initialization the algorithm continues by randomly picking codewords from the input pool. If the word is not reverse-complementary to all codewords in the *output* pool, the word is moved to the *output* pool and the pruning step is repeated. The algorithm continues with the remaining codewords in the *input* pool, until none are left or a certain number of random draws have taken place. The elements in the pool at the end of the algorithm depend on the starting codeword and the order in which the codewords were chosen. All random draws follow a uniform distribution.

In order to assign pixel values to DNA codewords an algorithm that implements the NTC is used, identified as *Part 2*. In principle a codeword is assigned to a value such that the thermodynamic stability with codewords assigned to values in the neighborhood defined by T_P are monotonically decreasing and are above the threshold T while codewords further away are less than T. Ideally the melting temperature between a codeword assigned to a value and codewords assigned to adjacent ones will have a 'bell' shaped profile. This process is implemented utilizing a combination of stochastic search and a greedy

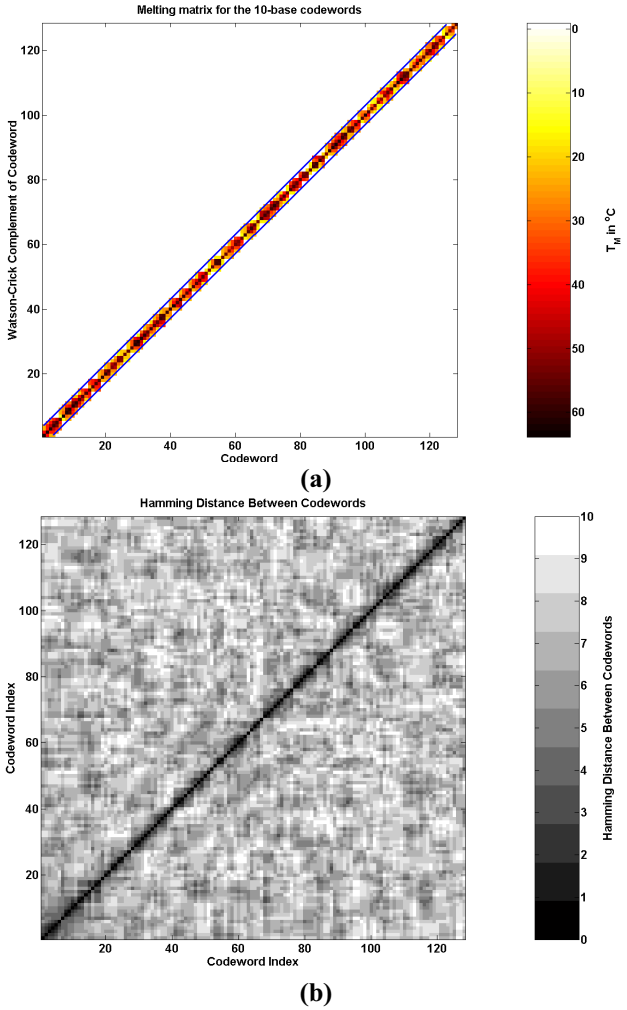

Figure 2. In (a) the melting matrix of the 128 codewords is shown. Melting temperatures under 15 °*C* have been omitted for visualization reasons. The color bar gives the correspondence between color and temperature. In (b) the Hamming distance between the 128 codewords is shown.

optimization algorithm. The convergence of the algorithm depends on the selected thresholds, codeword length, and size of the available pool.

4. SIMULATION RESULTS

We have implemented a thermodynamic simulator to calculate melting temperatures of perfect or non-perfect DNA duplexes based on the models presented in [4]. We have also implemented the algorithm described in the previous section, and succeeded in retrieving 128 codewords of length 10 that satisfy all the imposed constraints. Specifically, the parameters were: consecutive bases 2 GC content 50 %, and for the NTC they were T_P *=3* and *T=15 °C*.

Figure 2(a) depicts in two dimensions the melting matrix for the 128 codewords. The melting matrix is acquired by calculating the T_M of all possible codeword pairs. It is obvious that the maximums occur in the main diagonal since these values correspond to perfect hybridizations. In the same figure, two lines are plotted in T_P distance parallel to the main diagonal that show the significance of the parameter T_P. As the figure shows, all allowed hybridizations of codewords must have T_M higher than 15 °*C* and lie in the region defined by these lines.

To illustrate that melting temperature is a better descriptor of hybridization efficiency compared to Hamming distance, we include in Figure 2(b) the Hamming distance between codewords (white for the maximum Hamming distance of 10, black for 0, and linear gray levels for intermediate values). Notice the variation of the Hamming distance throughout the codeword set and especially outside the main diagonal and its 3-point neighborhood.

5. CONCLUSIONS

The codeword design techniques we propose are the first to utilize imperfect matches. Instead of over constraining the codeword design problem by trying to avoid unwanted hybridizations, we propose an entirely new design approach that incorporates “unwanted” hybridizations in order to account for noise and inexact matches in the data.

We presented a greedy algorithm for finding good codewords utilizing the noise tolerance constraint, although there is huge room for improvement. We have also developed an evolutionary approach to codeword design, according to which all constraints are satisfied simultaneously [5].

Our work is the first to use a DNA computing approach in digital signal processing. Although a simple matching problem was used as a proof of concept, the same mechanism can be used in developing DNA based databases of digital signals.

5. REFERENCES

[1] L. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, pp. 1021-1024, November 1994.

[2] B. Alberts *et al*, *Essential Cell Biology*. New York: Garland Publishing, 1998.

[3] A. Brenneman and A. Condon, "Strand design for bio-molecular computation," (Survey Paper), University of British Columbia, Vancouver, Canada, to appear 2001.

[4] J., SantaLucia Jr, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. Natl. Acad. Sci. USA,* vol. 95, pp. 1460-1465, 1998.

[5] S.A. Tsaftaris, "DNA-based Digital Signal Processing", MSc Thesis, Northwestern University, Dept. of Electrical and Computer Engineering, Evanston, IL, USA, May 2003.