Sotirios A. Tsaftaris, Aggelos K. Katsaggelos,
Thrasyvoulos N. Pappas, *and* Eleftherios T. Papoutsakis

How Can DNA Computing Be Applied to Digital Signal Processing?

In our previous article [15], we offered a brief introduction to DNA computing from a signal processing perspective. To assist in our presentation we also provided a very short overview of molecular biology and tools commonly used in DNA computing. In this article we are now focusing on the use of DNA computing for solving digital signal processing problems. Although digital signals have been used as inputs in some DNA computing applications (for example, in databases [6], [10]), there has not been any previous work on applying DNA computing principles in solving DSP problems. This article offers a first step towards filling this gap and thus strengthening the ties between biology and signal processing. We will present one specific application and point to several directions for future research. By focusing our attention to a specific domain, we believe that new exciting applications of DNA computation can be discovered.

We start out with a traditional signal processing problem that we believe can be used as a proof of concept for DNA based DSP. In the next section the most important aspects of this application are subsequently highlighted and applied to the implementation of DNA databases of digital signals. Finally, we conclude with future directions of research, of particular interest to the signal processing community.

DNA Computing Applied to the Solution of Matching Type Problems in Signal Processing

A number of applications in signal and image processing require determining the similarity of either two signals or segments of the two signals, as expressed by the shifts required for one signal to match the other. Among the various types of such problems, the disparity estimation problem is an example that demonstrates the basic ideas.

Much work has been done on matching type problems in general [5] and disparity estimation in particular. The most often used criteria for matching are photometric similarity and spatial consistency, which confines the search space for the displacement or disparity vector.

An additional constraint imposing spatial consistency on the disparity field is the *ordering constraint*, according to which if pixel A is to the left of pixel B in the left image, this order should be preserved in the right image by the two pixels A and B are mapped into. What makes the disparity estimation problem particularly challenging is the presence of occlusions, i.e., parts of the objects visible by one camera but not the other.

To reduce the complexity of disparity estimation, the cameras are usually arranged in a parallel-axis configuration reducing to an intrascanline pixel-matching problem, that is, matching of two one-dimensional (1-D) signals. In most DSP applications the available data are corrupted by random noise commonly modeled as an additive and wide sense stationary process, with Gaussian distribution.

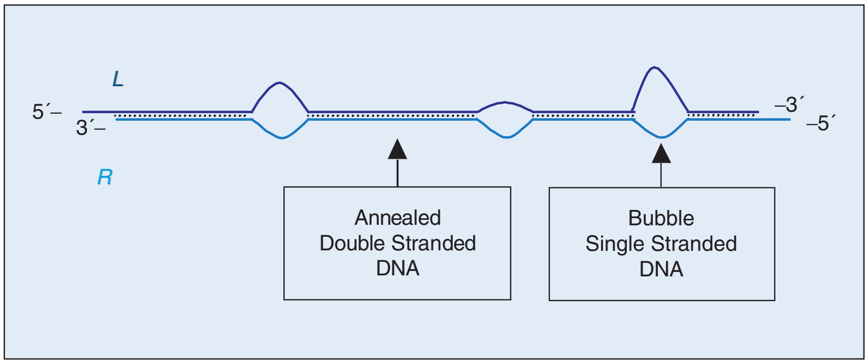
The DNA composition and aqueous-solution chemistry can be used to solve matching type problems [13]. By correctly formulating the disparity estimation problem, it can in principle be solved using DNA molecules by exploiting the computational power of DNA annealing [17].

The first step towards a DNA-based solution is the coding of the digital data into a sequence of DNA bases. Both left and right signals (image lines at the same vertical location) have to be encoded in DNA.

The code-word design constraints listed in [15] are adequate for most existing biocomputing applications, for which nonspecific hybridization is not desirable. In contrast, nonspecific hybridization can be very useful (and is actually necessary) in signal processing applications, where an exact match is typically not possible due to the nature of the signals and noise, and a match in the mean-squared-error (MSE) sense is desired. As presented in [8], base mismatches affect the thermodynamic stability of duplexes. By carefully placing mismatches and assigning signal values to DNA code words, we can inherently control the imperfect matches by adjusting the temperature of the hybridizations.

DNA annealing can provide a global (optimal) solution to the

1053-5888/04/$20.00©2004IEEE



▲ *1. Example of annealing between two DNA sequences that encode the corresponding left (L) and right (R) image lines. Matching parts will anneal and lead to double stranded formations while nonsimilar parts will result in the formation of wobbles or "bubbles."*

problem of disparity estimation. We assume that the left image line is encoded into its DNA sequence and synthesized (in 5′ to 3′ direction) and the right line is encoded into DNA but its Watson-Crick complement is synthesized (in 3′ to 5′ direction). Introducing both DNA sequences in a solution with appropriate conditions will lead to annealing of parts of the sequences that are similar to each other thus forming double stranded DNA.

Nonsimilar parts will create wobbles or "bubbles," due to the nonspecific hybridization. Nonsimilar intensity values will represent nonmatching regions, which were earlier defined as occlusions. An illustrative example is shown in Figure 1. By carefully designing the mapping of signal values to DNA molecules, their annealing can be a powerful tool towards the extraction of disparity information. The elasticity of macromolecules provides an excellent mechanism for implementing the ordering and smoothness constraints of disparity estimation theory. Globally optimal solutions can therefore be obtained using DNA.

Measuring the length and position of each bubble in both the 5′ to 3′ and 3′ to 5′ directions will result in the sought after disparity estimates. Laboratory application of the above idea will result in many formations, similar to the ones in Figure 1. By applying nondenaturalizing gel electrophoresis we can retrieve the formation that has the highest percentage of double stranded sections (smaller mobility compared to other formations with smaller percentage), which represents the optimal solution. However, measuring the position and length of the bubbles in the formation does not seem to be realizable with any of the currently available laboratory techniques. To provide a proof of concept and overcome the current limitations that make the use of the previous optimal approach impractical, a polymerase chain reaction (PCR) procedure combined with gel electrophoresis was proposed to provide a solution to the problem of disparity estimation [13], [14].

We denote by $T_M(x_l, w_l^C)$ the melting temperature of the duplex formed by two DNA sequences, x_l in the 5′ to 3′ direction and w_l^C, the Watson Crick complement of w_l, in the 3′ to 5′ direction. If we consider two signal values p_x, p_w with corresponding DNA representations $x_l = C(p_x)$ and $w_l = C(p_w)$, where $C(\cdot)$ is the function of code-word assignment, they will have melting temperature $T_M(x_l, w_l^C)$. Ideally we want to design the DNA code words in such a way that the melting temperature between DNA code words is *inversely proportional to the absolute difference* between the encoded signal values. In the case when $p_x = p_w$, $T_M(x_l, w_l^C)$ must be maximized. This will enable us to allow a desirable number of nonspecific hybridizations by controlling the temperature of the experiment.

To accomplish this, we have introduced a new constraint, the *noise (or inexact match) tolerance constraint (NTC)* that addresses the problem of assigning DNA code words to neighboring signal values [13], [14]. According to this constraint, duplexes formed by code words assigned to neighboring signal values must have (similar) high melting temperatures, while duplexes outside this neighborhood must have melting temperatures lower than a threshold T. That is,

$$\text{for } x_l = C(p_x) \text{ and } w_l = C(p_w)$$

$$T_M\left(x_l, w_l^C\right) = \begin{cases} \text{maximum} & \text{if } p_x = p_w \\ \propto \frac{1}{f(|p_x - p_w|)} & \text{if } |p_x - p_w| \le T_p \\ <T & \text{if } |p_x - p_w| > T_p \end{cases}$$

where $f(x)$ represents a monotonically increasing function of x, which needs to be specified by the designer along with the values of the parameters T_p and T.

The threshold T_p is very critical for the noise tolerance of the algorithm. Essentially this threshold is proportional to the variance of the noise, i.e., for low SNRs a relatively larger T_p is desired. The thresholds T and T_p are not independent of each other; T must be set to be considerably smaller than all melting temperatures calculated for pixel values that satisfy $|p_x - p_w| \le T_p$, that is $T \ll 1/f(T_p)$.

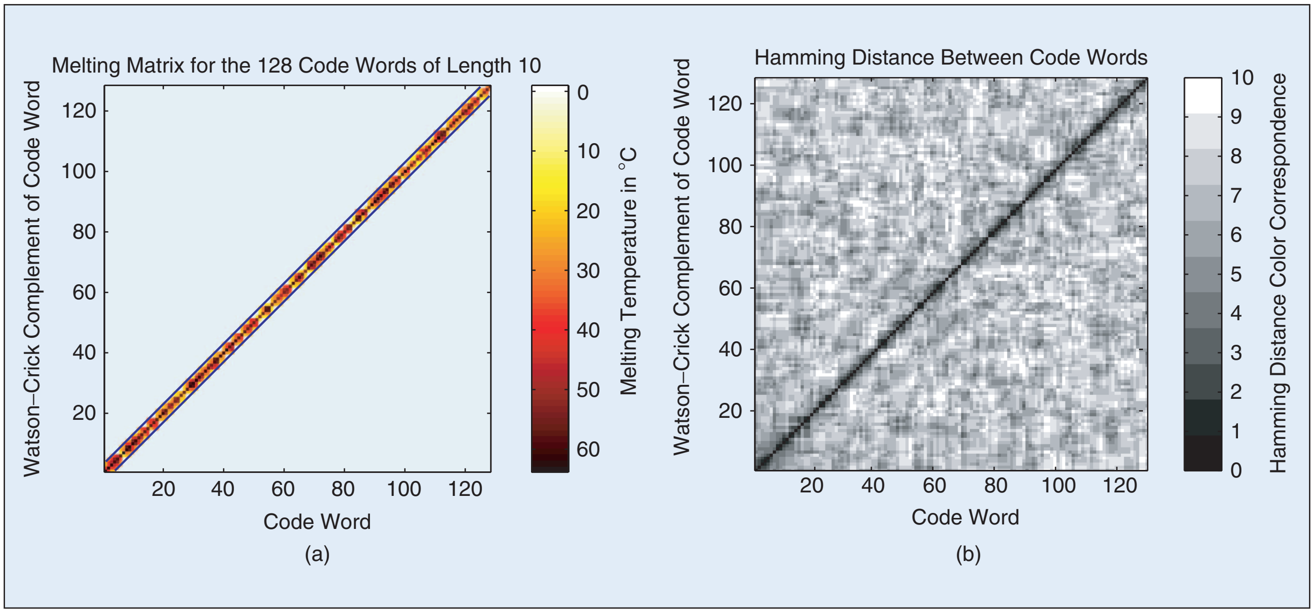
We have developed [13], [14] various algorithms for designing code words that satisfy the NTC. An example of the application of one of the algorithms (a greedy generation and random test algorithm) is shown in Figure 2(a) and (b). The parameters we used for the NTC were $T_P = 3$ and $T = 15$. In addition the consecutive bases constraint $R_B(x_l) \leq 2$ and GC content constraint of 50% were satisfied (for a description of these constraints please see [15]). In calculating the melting temperature the thermodynamic models and parameters of [8] and [11] were used. One solution we obtained consists of 128 code words of length ten, thus enabling the representation of 7-b samples. The melting matrix for these 128 code words is shown in Figure 2(a). In this figure, two lines are plotted in T_P distance parallel to the main diagonal. These lines show the significance of T_P: all allowed melting temperatures must be in the region defined by these lines. In Figure 2(b) the Hamming distance between the code words is shown. It is clear that the Hamming distance alone is not an adequate constraint.

Building and Managing Large Scale DNA Databases of Digital Signals

As mentioned above, current laboratory techniques do not allow the use of DNA for obtaining optimal solutions to the disparity estimation problem. The tedious sequential use of PCR experiments provides a proof of concept but is not adequate and does not provide any significant advantage with respect to computations over computer approaches. However, DNA databases may offer a direct application of NTC that is not limited by current laboratory techniques.

Using DNA to build databases and memories has long been envisioned and numerous approaches have been proposed. The scope of this section is not to review in detail all such efforts but to provide the basic ideas in constructing and managing DNA databases and to illustrate how NTC can be used in a very basic DNA database implementation. From the plethora of articles in the area, we briefly mention the following two as an example, due to their use of digital images as data. Reif et al. [10] presented a DNA database capable of storing images. By using a vector quantization (VQ) construction of the DNA indices [9], searches within the database can be performed. Error-correction principles from telecommunication engineering were also used to improve the performance of their system. Mills et al. [6] introduced a model of an analog neural network represented by chemical operations performed on strands of DNA. The proposed chemical operations formed the DNA equivalent of an analog vector computation algebra. They also provided a particular set of DNA operations to achieve the inter-conversion of electrical and DNA data and to represent a Hopfield associative memory capable of storing images.

In principle DNA database elements can either consist of an index and data part or and index and data part captioned between two fixed sequences (primer sites). All parts represent double stranded DNA and their specific form depends on the application. By adding specific



▲ *2. (a) Melting matrix of the designed 128 code words. Melting temperatures under 15 °C have been omitted. The color bar gives the correspondence between color and temperature. (b) Hamming distance between the 128 code words.*

primer sites, an inexpensive database replication mechanism can be implemented using a PCR procedure with primers that will bind to the specified sites.

To illustrate the application of NTC to DNA databases consider the scenario where a collection of audio signals is desired to be stored in a DNA database. The audio signals are encoded with an 8-b level representation. An index-data approach can provide a way to identify the various sequences using the index tag. The collection of the signals is encoded in DNA sequences and random index sequences are attached. The collection of sequences is synthesized and kept in a storage medium. For long-term storage, freeze-dried approaches are a possibility, and recently, even the use of living organisms has been suggested [2] with the risk of introducing errors due to mutations.

Let us now assume that we want to find if a certain sequence (*query*) of digital data exists in the database. A *query strand* is therefore synthesized based on the query. A sample of the solution containing the database will be heated for the double stranded elements to melt and the query strand will be added. The solution will be cooled down to allow hybridization between a database element (single stranded) and the query strand. The detection of the hybridization event can be verified, using one of several spectroscopic techniques, i.e., fluorescent labels attached to the query strand. In this case, a change in the fluorescent response will indicate success.

If, in addition, we want to know in which element the match was found, affinity purification can separate the strands that had a match. De-naturation and solution wash will result in a solution of elements for which a match was found. Assuming the existence of a DNA chip (microarray) that has all the indices installed, placing a sample of the washed solution onto the chip will return the index. Further processing can even return the position of the match. For example, for the elements where a match was found, a PCR step using as primer the query strand followed by gel electrophoresis can result in the desired position.

The design of the indices is quite important. Instead of being assigned at random, they can be smartly designed to carry information about the data, such as, a sequence ID, the used sampling rate or the length of the data segment. The index tags should be completely different from the data to improve the fidelity of the system. Such specificity is feasible, and it falls in the illegal code constraint category described in [15].

If the code-word sets were developed to satisfy the NTC, inexact matches (matches in the MSE-sense) of the query strands with the data will be allowed.

Although the procedure described above is tedious and low in throughput, variations and addition of new technologies such as high-density microarrays [7] and high throughput sequencing technologies can improve the yield of the system.

Using DNA to store digital signals or data in general offers significant advantages when compared to other media. The DNA molecule, especially in its double stranded form, is very stable, compact, and inexpensive. PCR is an economical and efficient way to replicate databases. Querying the database can be implemented with a plethora of techniques. Given a query molecule, DNA annealing provides a very efficient way to search for similar molecules in the database compared to digital databases. In digital databases the query time increases proportionally to an increase in the size of the database. However, in DNA databases when annealing is used as a search mechanism, the querying time is not dependent on the size of the database. This is because DNA kinetics are dependent on relative concentrations and not on the number of different molecules.

Future Directions

Bioinformatics research and computational biology can potentially benefit considerably from the application of signal processing techniques and vice versa. Chen et al. [3] provided a very comprehensive presentation of such potential benefits.

Cook et al. [4] provided an essay on the self-assembling computation of the Sierpiński triangle, a triangular fractal pattern named after Waclaw Sierpiński. In their paper they also described the correlation between the Sierpiński triangle and the binary version of wavelet and Fourier transforms, as well as, the Hadamard transform. Specifically, it is stated that the above have self-similar matrices closely related to the Sierpiński triangle. Self-assembly and tiling can also be used to study Markov fields which have been extensively used in image processing. In the future, one can imagine a self-assembly approach to image processing following similar principles.

We saw earlier that error correction and VQ methods can be used to improve the fidelity of molecular processes. Among the first to examine their application to DNA computing was Wood [18]. In some cases codeword design for molecular computing has similar properties to communication over noisy channels. The frame shift problem in DNA computing is very similar to the bit erasure problem most commonly seen in quantum and optical channels. Signal processing principles can be directly applied to develop such codeword sets or to develop new error correcting codes more suitable for molecular computing applications.

Digital signals are an integral part of modern technologies, and there is a growing need for their efficient storage. In the previous section we addressed the problem of designing DNA databases of digital content. The efficient storage and retrieval of the data are open problems. They are currently in their embryonic state, far from perfection.

As is well known, the building blocks of a DSP processor are addition, multiplication, and buffering or delay. One can envision the possibility of designing molecular circuits that behave like digital filters. Molecular circuits are part of nanoscale research but can be seen as DNA computing devices to some extent. Various research groups have demonstrated molecular circuits that behave like transistors [1] and adders [12], [16] and illustrate the future in which, complex molecular circuits acting as digital filters are realizable.

Conclusions

As an illustration of how DNA computing can be applied to DSP, a matching type of problem was considered. DNA-based DSP requires new code-word designs that can account for imperfect matches and the presence of noise. Towards this end, the NTC was presented and the benefits of its use were analyzed. DNA databases of digital signal data were presented as an immediate application of the NTC. The benefits of using DNA to store digital data, especially signals, include: information compactness, economical database replication, efficient query mechanisms, and query time that is independent of database size. All of the above make DNA an excellent candidate for storage compared to traditional magnetic or optical approaches.

A secondary goal of this article was to offer to the signal processing community future directions in a new unexplored area of research and to add another research domain to the biocomputing community. Although the technology today is limited and the idea of re-educating the engineer in biology seems scary, we hope that this publication will inspire new scientists and lead to the foundation of a new discipline, that of DNA-based digital signal processing.

Acknowledgments

S.A. Tsaftaris would like to thank the Alexander S. Onassis Public Benefit Foundation for their continuing financial support via their scholarship program.

Sotirios A. Tsaftaris, *Aggelos K. Katsaggelos*, and *Thrasyvoulos N. Pappas* are with the Department of Electrical and Computer Engineering, Northwestern University. *Eleftherios T. Papoutsakis* is with the Department of Chemical Engineering, Northwestern University.

References

[1] E. Ben-Jacob, Z. Hermon, and S. Caspi, "DNA transistor and quantum bit element: Realization of nano-biomolecular logical devices," *Phys. Lett. A*, vol. 263, no. 3, pp. 199–202, 1999.

[2] Y. Benenson, R. Adar, T. Paz-Elizur, Z. Livneh, and E. Shapiro, "DNA molecule provides a computing machine with both data and fuel," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2191–2196, 2003.

[3] J. Chen, H. Li, K. Sun, and B. Kim, "How will bio-informatics impact signal processing?" *IEEE Signal Processing Mag.*, vol. 20, no. 6, pp. 16–26, 2003.

[4] M. Cook, W.K. Rothemund, and E. Winfree, "Self-assembled circuit patterns," in *Preliminary Proc. 9th Int. Meeting on DNA Based Computers*, 2003, pp. 99–110.

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

[6] A.P. Mills, B. Yurke, and P.M. Platzman, "Article for analog vector algebra computation," *Biosystems*, vol. 52, no. 1-3, pp. 175–180, 1999.

[7] E.F. Nuwaysir, W. Huang, T. Albert, J. Singh, K. Nuwaysir, A. Pitas, T. Richmond, T. Gorski, J.P. Berg, J. Ballin, M. McCormick, J. Norton, T. Pollock, T. Sumwalt, L. Butcher, D. Porter, M. Molla, C. Hall, F. Blattner, M.R. Sussman, R.L. Wallace, F. Cerrina, and R.D. Green, "Gene expression analysis using oligonucleotide arrays produced by maskless photolithography," *Genome Res.*, vol. 12, no. 11, pp. 1749–1755, 2002.

[8] N. Peyret, P.A. Seneviratne, H.T. Allawi, and J. SantaLucia, Jr., "Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A-A, C-C, G-G, and T-T mismatches," *Biochemistry*, vol. 38, no. 12, pp. 3468–3477, 1999.

[9] J.H. Reif and T.H. LaBean, "Computationally inspired biotechnologies: Improved DNA synthesis and associative search using error-correcting codes and vector-quantization," in *Revised Papers 6th Int. Workshop on DNA-Based Computers: DNA Computing, Lecture Notes in Computer Science*, Springer-Verlag, vol. 2054, pp. 145–172, 2001.

[10] J.H. Reif, T.H. LaBean, M. Pirrung, V. Rana, B. Guo, K. Kingsford, and G. Wickham, "Experimental construction of very large scale DNA databases with associative search capability," in *Proc. 7th Int. Meeting DNA Based Computers*, Tampa, FL, 2001, pp. 231–247.

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

[12] M.N. Stojanovic and D. Stefanovic, "Deoxyribozyme-based half-adder," *J. Amer. Chem. Soc.*, vol. 125, no. 22, pp. 6673–6676, 2003.

[13] S.A. Tsaftaris, "DNA-based digital signal processing," M.S. thesis, Northwestern Univ., Dept. Elec. Computer Eng., 2003.

[14] S.A. Tsaftaris, A.K. Katsaggelos, T.N. Pappas, and E.T. Papoutsakis, "DNA based matching of digital signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, Montreal, Quebec, Canada, May 17–21, 2004, pp. 581–584.

[15] S.A. Tsaftaris, A.K. Katsaggelos, T.N. Pappas, and E.T. Papoutsakis, "DNA computing from a signal processing viewpoint," *IEEE Signal Processing Mag.*, vol. 21, no. 5, pp. 100–106, 2004.

[16] B. Yurke, A.P. Miller, and S.L. Cheng, "DNA implementation of addition in which the input strands are separate from the operator strands," *BioSystems*, vol. 52, no. 1-3, pp. 165–174, 1999.

[17] E. Winfree, "On the computational power of DNA annealing and ligation," in *DNA Based Computers: Proc. DIMACS Workshop*, April 4, 1995, pp. 199–221.

[18] D.H. Wood, "Applying error correcting codes to DNA computing," in *Proc. 4th Int. Meeting on DNA Based Computers*, 1998, pp. 109–110.