Complete citation: S.A. Tsaftaris, V. Hatzimanikatis, A.K. Katsaggelos, "*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", 24(4), 268-276, December 2005.

In silico estimation of annealing specificity of query searches in DNA databases

S. A. Tsaftaris[1], V. Hatzimanikati [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 consider DNA implementations of databases for digital signals with retrieval and mining capabilities. Digital signals are encoded in DNA sequences and retrieved through annealing between query DNA primers and data carrying DNA target sequences. The hybridization between query and target can be non-specific containing multiple mismatches thus implementing similarity-based searches. In this paper we examine theoretically and by simulation the efficiency of such a system by estimating the concentrations of query-target duplex formations at equilibrium. A coupled kinetic model is used to estimate the concentrations. We offer a derivation that results in an equation that is guaranteed to have a solution and can be easily and accurately solved computationally with bi-section root-finding methods. Finally, we also provide an approximate solution at dilute query concentrations that results in a closed form expression. This expression is used to improve the speed of the bi-section algorithm and also to find a closed form expression for the specificity ratios.

1 Introduction

Baum [1] first proposed building a DNA database capable of storing and retrieving digital information. He demonstrated that such database would allow data (in contrast to conventional address) based searches by exploiting the annealing property of DNA and utilizing common laboratory techniques for database access and management.

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 have equal concentrations.

As with any DNA computing application, the first step towards designing a database and retrieval mechanism is to find reliable methods for encoding digital signals into DNA sequences, also known as the codeword design problem. In our problem, the encoding has to be such that it enables content-based searches and at the same time limits the possibility of errors during retrieval. Furthermore, in the case of digital signals where perfect matches are almost impossible due to noise or the nature of the signals, the encoding scheme needs to allow for and account for similarity based retrieval. To accomplish this we introduced a new constraint, the Noise Tolerance Constraint (NTC) [9, 11].

The second step is to design the structure of the storage elements each of which has unique properties and characteristics. DNA databases consist of a collection of elements. Usually each element has a unique address (or index) block, which uniquely identifies (and therefore enables the retrieval of) a usually larger information-carrying block (data block).

The third step is to decide on the host environment of the database. It can be a test tube, a polymer, or even a living organism. Each host offers unique capabilities but at the same time imposes constraints on the information capacity, database longevity, and ease of use. Finally, input and output methods need to be chosen. Some input (information storage) and output (information retrieval) methods are faster than others.

The steps mentioned above are not independent. For example, the choice of host and database element has a large impact on the codeword design problem and the capabilities of the database.

This paper begins with a brief description of the overall system and its structure (section 2). Subsequently in section 3 we provide the mathematical framework that describes the behavior of the system and present a computational bi-section method, and an approximate closed form solution that is valid for low query concentrations for es-

timating the concentrations of query-target duplex formations at equilibrium. In section 4 we offer simulation results and compare the above two methods at various concentrations. Finally in section 5 we conclude this paper and provide some directions for future improvement.

2 System Description

The problem can be described as follows. Consider a set of digital signals V, with elements the $l \times 1$ vectors v_i, each entry of each is a $k-$bit integer. This set V is hereafter termed as *digital database*. Assume also a vector q that contains l_q, $k-$bit integers, hereafter termed as *digital query*. The problem at hand is to find out whether q 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. Similarly there are different ways of searching within the database. Traditional correlation and convolution techniques can be used, although the computational cost is an important consideration for such choice. The complexity of the problem in the best case scales linearly with the size of V, but in most implementations the complexity is a polynomial function of the size 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.

2.1 DNA Equivalent System

Our research is centered on providing a DNA-based alternative to the above problem. 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 an extraordinary tool that allows for the medium to be part of the computational platform as well, since it allows us to perform data searches using the annealing property of DNA.

A double helix of DNA (Deoxyribo-Nucleic Acid) is made from two single strands of DNA, each of which is a chain of nucleotides A, G, T, and C. Nucleotides can be joined together in a linear chain to form a single strand of DNA. A short single strand of DNA consisting of up to 100 or so nucleotides is called an oligonucleotide or oligo and is also characterized by polarity, originating from the two distinct ends, the 5' end and the 3' end. 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 oligonucleotide has a complementary sequence with opposite polarity; for example, the complementary sequence of 5′-ATG-3′ is 3′-TAC-5′. If two complementary sequences meet in a solution under appropriate conditions, they will attract each other and form a double stranded helical structure. This process is called hybridization. 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 equivalent system to a digital system is to map the digital information into DNA. The problem is also known as the codeword or library or word design problem. In our case the problem translates into finding DNA sequences (words) $x_i, i = 1...N$, of length l bases, capable of encoding N, $k-$bit integers.

In most DNA computing applications only perfect hybridizations are acceptable. In our case, we want to design the DNA codewords such that the melting temperature between DNA words is inversely proportional to the absolute difference between the encoded signal values. To accomplish this, we have introduced a new constraint, the Noise (or inexact match) Tolerance Constraint (NTC) [9, 10, 11]. 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 hybridizations.

For simplicity let us assume that DNA sequences inside a database are constructed as in **Fig. 1**, although many other different structures can be found [13]. For clarity we term these structures *database elements*. For each database element S_j, with concentration C_j, the double-lined gray part is the index that identifies the data, which appear on the right in solid black lines. Data are concatenations of DNA words. The index part of different elements

should be very dissimilar. Assume that we have M digital signals and hence M database elements. For simplicity lets assume that each database element is $L+IN$ bases long where IN is the index length and L is the data length.

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

(a) M database elements, of concentration C_j and sequence information s_j of length L (in the analysis that follows we ignore the index without loss of generality), $j=1...M$.

(b) A query Q, shown in **Fig. 1** as solid gray line, of concentration $\left|Q\right|_o$ and sequence information s_Q of length $q<L$.

(c) Reaction parameters: temperature T and salt concentration $\left|Na^{++}\right|$

To ease our analysis we introduce the notion of fragments. A fragment F_{ip}^j represents the sequence information of a database element 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, it is apparent that in our case the initial concentration of each fragment is $\left|F_{ip}^j\right|_o = C_j$.

Subsequently we can model the query fragment complexes as QF_{ip}^j. Such complexes are illustrated in **Fig. 1** in various locations. The complexes can have rather elaborate geometric structures and predicting such formations is a well-studied field [2, 8, 16]. Without loss of generality, to ease our analysis and reduce the complexity we will assume that (i) $p=q$, (ii) complexes will have only internal mismatches and no dangling ends, and (iii) the number of complexes is limited to N_T. The goal of this article is to estimate $\left|QF_{ip}^j\right|$ for a given set of inputs and parameters.

3 Modeling the hybridization reactions

The problem of estimating the concentrations $\left|QF_{ip}^j\right|$ can be rather complicated. The following reaction equation describes the formation of those complexes

$$Q + F_{ip}^j \underset{K_r}{\overset{K_f}{\rightleftharpoons}} QF_{ip}^j, \quad \forall i,j,p \ . \tag{1}$$

The parameters K_f and K_r are called the forward and reverse rate constants. These rates depend on environmental parameters and laboratory settings. They are usually difficult to estimate since they require a plethora of laboratory experiments and are not universal. These disadvantages make them unattractive, and usually an equilibrium analysis that does not model the dynamic behavior is sought after.

3.1 Equilibrium Solution

Under an equilibrium assumption, $K_{eq} = \dfrac{K_f}{K_r}$, and the differential equations that describe the mass action equations that satisfy equation (1) become polynomial equations. For simplicity we will drop the "eq" from K_{eq} and since $p=q$ equation (1) becomes

$$Q + F_i^j \underset{K_r}{\overset{K_f}{\rightleftharpoons}} QF_i^j, \quad \forall i,j \ . \tag{2}$$

In equilibrium

$$K_i^j = \exp\left(-\frac{\Delta G_i^j}{R\cdot T}\right), \tag{3}$$

where ΔG is the Gibbs free energy, R is the Boltzman constant, and T is the temperature in Kelvin. There have been many proposals on estimating concentrations at equilibrium [4, 5, 6, 7, 15]. The Gibbs free energy for DNA complexes can be estimated using nearest neighbor thermodynamics parameters, which are available in the literature [2, 8]. We should note that the equilibrium constants depend on the reaction temperature and salt concentration and hence they have to be readjusted for each experiment.

Also,

$$K_i^j = \frac{\left|QF_i^j\right|}{\left|Q\right| \cdot \left|F_i^j\right|}. \tag{4}$$

The conservation equation on the query is given by

$$\left|Q\right|_o = \left|Q\right|_{free} + \left|Q\right|_{bound} = \left|Q\right| + \sum_{i,j} \left|QF_i^j\right|, \tag{5}$$

where the sum is over N_T terms which is the number of complexes in the system.

The parameter α is defined as the percentage of bound query strands and is given by

$$\alpha = \frac{\left|Q\right|_o - \left|Q\right|}{\left|Q\right|_o} = \frac{\sum_{i,j} \left|QF_i^j\right|}{\left|Q\right| + \sum_{i,j} \left|QF_i^j\right|}. \tag{6}$$

After substituting equation (4) into (6) we get

$$\alpha = \frac{\sum_{i,j} K_i^j \cdot \left|F_i^j\right|}{1 + \sum_{i,j} K_i^j \cdot \left|F_i^j\right|}. \tag{7}$$

Utilizing the conservation equations for each species we have

$$\left|F_i^j\right| + \left|QF_i^j\right| = \left|F_i^j\right|_o \Rightarrow \left|F_i^j\right| + K_i^j \left|Q\right| \left|F_i^j\right| = \left|F_i^j\right|_o, \tag{8}$$

where $\left|F_i^j\right|_o$ is the initial concentration of each fragment. Note that in our case $\left|F_i^j\right|_o = C_j$.

Solving equation (8) for $\left|F_i^j\right|$, and utilizing equation (6) we obtain

$$\left|F_i^j\right| = \frac{\left|F_i^j\right|_o}{1 + K_i^j \left|Q\right|_o (1-\alpha)}, \quad \forall i, j. \tag{9}$$

According to [5] the system of equations (7) and (9) can be solved iteratively and provide a solution for the sought after concentrations $\left|QF_i^j\right|$.

Alternatively, a solution to this system of equations can be obtained as follows. Equation (9) can be rewritten as

$$\left|F_i^j\right| = \frac{\left|F_i^j\right|_o}{1 + K_i^j \left|Q\right|}. \tag{10}$$

In this case we can write equation (5) as

$$\left|Q\right|_o = \left|Q\right| + \sum_{j,i} \frac{K_i^j \cdot \left|F_i^j\right|_o \cdot \left|Q\right|}{1 + K_i^j \cdot \left|Q\right|} = \left|Q\right| + \sum_i \frac{\left|F_i^j\right|_o \cdot \left|Q\right|}{\frac{1}{K_i^j} + \left|Q\right|}. \tag{11}$$

Setting

$$q = \left|Q\right| / \left|Q\right|_o, \quad q \in [0,1], \tag{12}$$

and rearranging the terms in (11) we get:

$$\sum_{j,i} \frac{\left(\left|F_i^j\right|_o / \left|Q\right|_o\right) \cdot q}{\left(1 / \left(\left|Q\right|_o K_i^j\right)\right) + q} + q - 1 = 0. \tag{13}$$

Hence we are looking for the solution $q_s \in [0,1]$ that satisfies the above equation.

If we set,

$$f(q) = \sum_{j,i} \frac{\left(\left|F_i^j\right|_o / \left|Q\right|_o\right) \cdot q}{\left(1 / \left(\left|Q\right|_o K_i^j\right)\right) + q} + q - 1, \tag{14}$$

the problem is equivalent to finding the roots of $f(q)$.

With fundamental calculus we obtain that (i) $f(0) = -1 < 0$, (ii) $f(1) > 0$, and (iii) $f'(q) > 0$. Therefore according to the intermediate value theorem there exists a unique solution $q_s \in [0,1]$ such that $f(q_s) = 0$. Finding the solution analytically is infeasible since the equation $f(q_s) = 0$ leads to a $(N_T + 1)$-th order polynomial.

Comparing equations (6) and (12) we see that

$$q = 1 - \alpha \,. \tag{15}$$

3.2 Computational solution

The unique characteristics of equation (13) make it an ideal candidate for a computational bi-section method [14] to find an approximate solution $a \le q_B \le b$, where $a < b$ are given boundaries. Such method can guarantee that q_B is within ε from q_s, where $\varepsilon \le (b-a)/2^{n+1}$ and n is the number of iterations. The concentration

$$|Q_B| = q_B \cdot |Q|_o \,, \tag{16}$$

will henceforth correspond to the solution by bi-section of equation (13). Since bi-section methods are known to be problematic when roots lie close to zero (as in our case), instead of finding the roots of (13) we find the solution of $\log\left(f(q)+W\right) - \log\left(W\right) = 0$, where $W \gg 1$, which avoids numerical precision errors and leads to better convergence.

3.3 Approximate solution when query concentration is diluted

In the following for simplicity, we will assume that $\left|F_i^j\right|_o = C_j = C$ and $C/|Q|_o = \rho$, which results in simplified and more tractable equations. Similar derivations can be made without the above assumption. Using the above assumptions we can rewrite equation (13) as:

$$\frac{\overbrace{1-q}^{h(q)}}{\rho} = \overbrace{\sum_{j,i} \frac{q}{\left(1/\left(|Q|_o K_i^j\right)\right) + q}}^{g(q)} \,, \tag{17}$$

or

$$f(q) = \sum_{j,i} \frac{q}{\left(1/\left(|Q|_o K_i^j\right)\right) + q} + \frac{q}{\rho} - \frac{1}{\rho} = g(q) - h(q) \,. \tag{18}$$

We are interested in roots of equation (18).

In **Fig. 2** we see a graphical illustration of equation (17). The gray heavy line represents $h(q)$ the left part of equation (17). The black curve, illustrates each term inside the sum, which is asymptotic to 1 as $q \to 1$. The black heavy curve represents the whole sum $g(q)$, which asymptotically reaches N_T. When $q \to 0$ the quantity $1/\left(|Q|_o K_i^j\right)$ of each term inside the sum, is much larger than q, thus can be approximated as a linear function close to zero. For linearity to hold we have to assume that $\rho \gg 1$, or equivalently $|Q|_o = C$, thus the query concentration is in dilute. With that observation we can approximate each term in the sum as

$$\frac{q}{\left(1/\left(|Q|_o K_i^j\right)\right) + q} \approx q \cdot K_i^j \cdot |Q|_o \,. \tag{19}$$

Then equation (17) becomes:

$$\frac{1-q}{\rho} \approx \sum_{j,i} q \cdot K_i^j \cdot |Q|_o = q \cdot |Q|_o \cdot \sum_{j,i} K_i^j \,, \tag{20}$$

and subsequently we can solve for q to obtain

$$q_a \approx \frac{1}{1 + \rho \cdot |Q|_o \cdot \sum_{j,i} K_i^j} \,, \tag{21}$$

or

$$|Q_a| \approx \frac{|Q|_o}{1 + \rho \cdot |Q|_o \cdot \sum_{j,i} K_i^j} \,. \tag{22}$$

We can see that equation (21) turns out to be an independent function of ρ if $C/|Q|_o = \rho$. The equation then simplifies to $q_a \approx \left(1 + C \cdot \sum_{j,i} K_i^j\right)^{-1}$. This can be further simplified to

$$q_a \approx \frac{1}{C \cdot N_T \cdot \overline{K}}, \tag{23}$$

or

$$|Q_a| \approx \frac{1}{\rho \cdot N_T \cdot \overline{K}}, \tag{24}$$

where N_T and $\overline{K}$ are the number of complexes and the mean value of equilibrium constants respectively.

3.4 Finding concentrations $|QF_i^j|$ and query selectivities

From equations (4) and (10) we have:

$$|QF_i^j| = |F_i^j|_o - \frac{|F_i^j|_o}{1 + K_i^j \cdot |Q|} = \frac{|F_i^j|_o \cdot K_i^j \cdot |Q|}{1 + K_i^j \cdot |Q|}. \tag{25}$$

The concentrations therefore can be found by substituting $|Q_B|$ from equation (16) or $|Q_a|$ from equation (24) in the above equation.

After finding the concentrations the selectivity percentages can be found according to

$$SA_i^j = \frac{|QF_i^j|}{\sum_{i,j} |QF_i^j|}. \tag{26}$$

This dimensionless expression can be seen as the percentage of the complex $|QF_i^j|$ within all the bounded complexes. In our case equation (26) can be termed as query selectivity or query specificity, which is commonly found in the literature of database design.

3.5 Finding K_{system}

K_{system} corresponds to the equilibrium constant of a lumped system where the individual targets are lumped and treated as a single species A. The concentration of A is essentially equivalent to the sum of $|F_i^j|$.

$$Q + A \underset{K_r}{\overset{K_f}{\rightleftharpoons}} QA \tag{27}$$

Our goal is to find the equilibrium constant of the above reaction.

Using conservation equations we have the following system of equations:

$$\begin{aligned} |Q|_o &= |Q| + |QA| \\ |A|_o &= |A| + |QA| \\ |QA| &= K_{system} \cdot |Q| \cdot |A| \end{aligned} \tag{28}$$

The solution of the above system of equations is

$$|Q|_o = |Q| + K_{system} \cdot |Q| \cdot \frac{|A|_o}{1 + K_{system} \cdot |Q|}. \tag{29}$$

Solving for K_{system} results in:

$$K_{system} = \frac{|Q|_o - |Q|}{|Q|^2 + |Q| \cdot |A|_o - |Q| \cdot |Q|_o}. \tag{30}$$

In order to find K_{system} we need to know $|A|_o$ and the concentration of the free query $|Q|$. We can argue that $|A|_o = \sum_{i,j} |F_i^j|_o = C \cdot \sum_{i,j} 1 = C \cdot N_T$. The concentration $|Q|$ is given either by the numerical solution $|Q_B|$ of equation (13) or by the approximate $|Q_a|$ of equation (22).

If we assume $\rho > 1$, we can substitute $|Q| = |Q_a|$ from equation (24) into (30) and we can show that

$$K_{system} \approx \overline{K} \,. \tag{31}$$

3.6 Estimating selectivity ratios

An ideal outcome of a laboratory experiment in DNA computing (and even in biotechnological applications) is to have great separation between wanted and unwanted hybridizations. One way to express that and evaluate performance is to find the ratio of the concentration of a wanted hybridization versus an unwanted hybridization.

Let us assume that we have a wanted complex $\left|QF_i^j\right|$ and an unwanted complex $\left|QF_r^{j'}\right|$. Their respective equilibrium concentration equations are:

$$\left|QF_i^j\right| = K_i^j \cdot |Q| \cdot \left|F_i^j\right| = K_i^j \frac{|Q| \cdot \left|F_i^j\right|_o}{1 + K_i^j |Q|}, \tag{32}$$

and

$$\left|QF_r^{j'}\right| = K_r^{j'} \cdot |Q| \cdot \left|F_r^{j'}\right| = K_r^{j'} \frac{|Q| \cdot \left|F_r^{j'}\right|_o}{1 + K_r^{j'} |Q|} \,. \tag{33}$$

Their ratio assuming they have equal initial concentrations is thus

$$\frac{\left|QF_i^j\right|}{\left|QF_r^{j'}\right|} = \frac{K_i^j \dfrac{|Q| \cdot \left|F_i^j\right|_o}{1 + K_i^j |Q|}}{K_r^{j'} \dfrac{|Q| \cdot \left|F_r^{j'}\right|_o}{1 + K_r^{j'} |Q|}} = \frac{K_i^j}{K_r^{j'}} \cdot \frac{1 + K_r^{j'} |Q|}{1 + K_i^j |Q|} \,. \tag{34}$$

The above ratio can be evaluated at $|Q| = |Q_B|$. Alternatively by substituting $|Q| = |Q_a|$ from equation (22), a rather interesting observation can be derived. The second fraction can be further simplified as

$$\frac{1 + K_r^{j'} |Q_a|}{1 + K_i^j |Q_a|} = \frac{1 + \dfrac{K_r^{j'} |Q|_o}{1 + \rho \sum_{i',r'} K_{i'}^{r'} |Q|_o}}{1 + \dfrac{K_i^j |Q|_o}{1 + \rho \sum_{i',r'} K_{i'}^{r'} |Q|_o}} \approx \frac{1 + \dfrac{K_r^{j'}}{\rho \sum_{i',r'} K_{i'}^{r'}}}{1 + \dfrac{K_i^j}{\rho \sum_{i',r'} K_{i'}^{r'}}} \approx \frac{1 + \dfrac{K_r^{j'}}{\rho \cdot N_T \cdot \overline{K}}}{1 + \dfrac{K_i^j}{\rho \cdot N_T \cdot \overline{K}}} \,, \tag{35}$$

and since

$$\frac{SA_i^j}{SA_r^{j'}} = \frac{\left|QF_i^j\right|}{\left|QF_r^{j'}\right|} \,, \tag{36}$$

substituting equation (35)into (34) we obtain

$$\frac{SA_i^j}{SA_r^{j'}} = \frac{\left|QF_i^j\right|}{\left|QF_r^{j'}\right|} = \frac{K_i^j}{K_r^{j'}} \cdot \frac{\rho \cdot N_T \cdot \overline{K} + K_r^{j'}}{\rho \cdot N_T \cdot \overline{K} + K_i^j} \,. \tag{37}$$

Equation (37) illustrates that at dilute concentrations the ratio of concentrations of two query-fragment complexes is analogous to the ratio of their equilibrium constant (which is expected), but it is also analogous to a term that highlights the dependency upon the network of fragments.

4 Simulation results and numerical considerations

For our simulations we developed MATLAB routines to evaluate q_B and q_a as presented above. To find the equilibrium constants we used the set of 32 words of length 19 presented in [12]. We used thermodynamic parameters from [8], and references therein, to estimate the Gibbs free energy [9] for all possible word-word combinations, 1024 in total. This set-up will emulate situations where a single word query is used to search inside a database. The database is assumed to consist of targets that contain concatenations of words. 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, 10, 12] for more details).

The equilibrium constants of the 1024 combinations were divided into two groups: (a) a 'Good' set of 212 pairs of desired hybridizations, and (b) a 'Bad' set of 812 undesired pairs. Random pools of $N_T = 100$ constants were drawn. For the results presented here we chose a pool with the following distribution: 2 from the 'Good' set ($K_1^0 = 5.40e10$, $K_2^0 = 3.23e9$) and 98 from the 'Bad' (maximum 'bad' constant $K_{36}^0 = 1.33\text{e}8$). The corresponding mean value of the set is $\overline{K} = 5.74e8$. In all simulations the initial concentration of fragments was $\left|F_i^j\right|_o = 10^{-5} mol/L$.

We evaluated the approximation error $\varepsilon = \left|(q_B - q_a)/q_B\right|$ for $\rho \in [0,100]$. As shown in **Fig. 3** the error is rather small when ρ is large which strengthens the validity of our approximation. Indicatively the error falls below 8% when $\rho > 10$.

In **Fig. 4** we plotted the selectivity percentages of complexes QF_1^0 , QF_2^0 , and QF_{36}^0 substituting q_B found for values of $\rho \in [0.001,100]$ in equation (26). From the figure we can see than when $\rho < 1$ (query in excess) the selectivity of the three complexes is rather small and almost equal. On the other hand when $\rho > 1$ the selectivity of the highest equilibrium constant dominates is close to 95% of the total hybridized complexes. The system demonstrates a non-linear behavior at the transition region. All three selectivities increase but at some point the SA_{36}^0 starts decreasing, subsequently SA_2^0 decreases, while SA_1^0 keeps increasing and finally dominates the network. It is clear that around this region the competitive hybridization due to lack of query strands becomes apparent and the most stable complex dominates the network.

This becomes more apparent in **Fig. 5** where we plot the selectivity ratios between the three fragments with the bi-section method in **Fig. 5**(a) and the approximation method **Fig. 5**(b). For the approximation method we used equation (37). The approximation method fails to capture the effect at the transition region since it is not accurate for $\rho < 1$. The ratios are upper-bounded by the ratio of equilibrium constants. That bound is achieved for high ρ .

We also estimated K_{system} for various concentrations, as seen in **Fig. 6**. An important observation arises from observing K_{system} at extremes. We see that K_{system} is rather small when compared to dilute cases. This can interpreted to the effect of differential sensitivity of the systems due to the different constants we have in the pool. As the concentration of the query decreases the distribution of constants becomes more important. On the other hand as the concentration increases and the system becomes saturated this importance weakens. Furthermore we notice that for $\rho > 1$, K_{system} approaches asymptotically $\overline{K}$, which proves by simulation the validity of equation (31).

From analyzing again the selectivities we see that, in dilute, the separation efficiency between wanted and unwanted hybridization is very good. On the other hand, the relative separation between wanted hybridizations may not be representative of the desired design. It is therefore critical to evaluate the deviation from the desired outcome and use that as a guide to fine-tune reaction parameters and concentrations.

5 Concluding Remarks

Our DNA computing research requires hybridizations between query and target that can be non-specific with multiple mismatches. In this paper we examined theoretically and by simulation the efficiency of such a system by estimating the concentrations of query-target duplex formations at equilibrium. A coupled kinetic model was used to estimate the concentrations. Simulation is an integral part of DNA computing research since via simulation the number of needed laboratory experiments can decrease, thus potentially reducing operational cost.

We offered an alternate solution to the system of equations that leads to a single equation that was proved to have a unique solution and can be easily adapted for computational solutions. We provided an approximate solution that is accurate in dilute query concentrations. Using this approximation, we can achieve a closed form expression to the solution of the system of equations. Furthermore, it allowed us to evaluate the selectivity ratios and estimate the performance of the query mechanism.

Our simulation results showed that in dilute query concentration, where competition is favored, the separation efficiency between wanted and unwanted hybridizations is remarkable. Separation efficiency among wanted hybridization is again very high, but not necessarily desirable. It depends on the design of the system and the sought after outcome and accuracy.

On the other hand, in excess there is almost no separation between hybridizations, that is, almost all three complex formations are equiprobable. This is the typical reason why in most PCR experiments the annealing temperature is set high to avoid mishybridizations since the query concentration is very high and the specificity is rather small.

The behavior of most systems is usually controlled by temperature but we illustrated that it can also be controlled by concentration ratios. It remains to be seen whether temperature or concentration control provides better query selectivity.

Appropriate selection of concentrations is important for outcome accuracy and can also decrease the cost of laboratory experiments. It is expected that lower concentrations will need longer reaction times and will have an impact on hybridization speed and hence "query speed". Exactly assessing the impact is infeasible without knowledge of forward and backward reaction rates. On the other hand relaxation time can give an insight on the level of dependency and provide guidelines for system design. In a future article we plan to introduce such analysis.

It is clear that finding the appropriate reaction parameters is very critical since they affect the outcome of the experiment significantly. It is also critical to find similar approximations for excess cases in order to further probe into system behavior and find closed form solutions for selectivity and specificity.

In a future paper we will consider the inverse problem. Namely, we commence by setting desired separation efficiencies and work backwards to find the best concentration parameters, and then to find the equilibrium constants. The constants will give us an estimate on the needed Gibbs free energy from which we can extract sequence information that will eventually give us a DNA codeword library that when used will result in the expected separation efficiency.

Acknowledgements

The authors would like to thank Prof. Osamu Ono for his invitation to contribute to the Journal of the *Japan Society of Simulation Technology* and Professors Suyama and Rose for useful remarks. Finally, Mr. Tsaftaris would like to thank the *Alexander S. Onassis* public benefit foundation for their financial support.

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. D.H. Mathews, M.E. Burkard, S.M. Freier, J.R. Wyatt and D.H. Turner, "Predicting oligonucleotide affinity to nucleic acid targets," *RNA*, vol. 5, pp. 1458-1469, 1999.
5. G.L. Moore and C.D. Maranas, "Predicting Out-of-Sequence Reassembly in DNA Shuffling," *Journal of Theoretical Biology*, vol. 219, pp. 9-17, 2002.
6. J. A. Rose, R. J. Deaton, and A. Suyama, "Statistical thermodynamic analysis and design of DNA-based computers," *Int'l Journal of Natural Computing*, vol. 3, pp. 443-459, 2005.
7. J. A. Rose and A. Suyama, "Physical modeling of biomolecular computers: Models, limitations, and experimental validation", invited paper: Int'l Journal of Natural Computing 3, pp. 411-426, 2005.
8. 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.
9. S.A. Tsaftaris, *DNA-based Digital Signal Processing*, M.S. Thesis, Northwestern University, Dept. of Electrical and Computer Engineering, June 2003.
10. 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.
11. 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.
12. 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 International Meeting on DNA Computing*, June 6-9, London, Canada, 2005.

13. S.A. Tsaftaris and A.K. Katsaggelos, "On Designing DNA Databases for the Storage and Retrieval of Digital Signals," International Conference on Natural Computation, special session on Recent Advances in Biomolecular Computing, Changsha, China, August 26-29, 2005, *Lecture Notes in Computer Science*, vol. 3611, pp. 1192-1201, Jul 2005.
14. 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.
15. J.Y. Wang and K. Drlica, "Modeling hybridization kinetics," *Math Biosci*, vol. 183, issue 1, pp 37-47, 2003.
16. M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res.*, vol. 31 issue 13, pp. 3406-15, 2003.

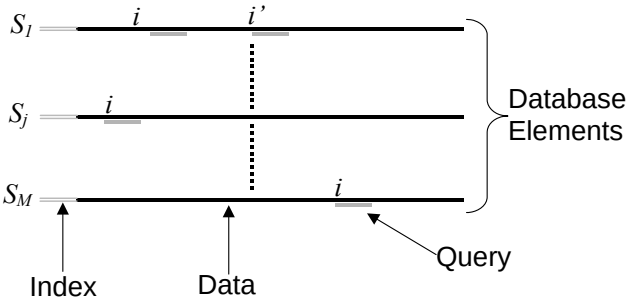


Fig. 1. Illustration of hybridizations between query and database elements.

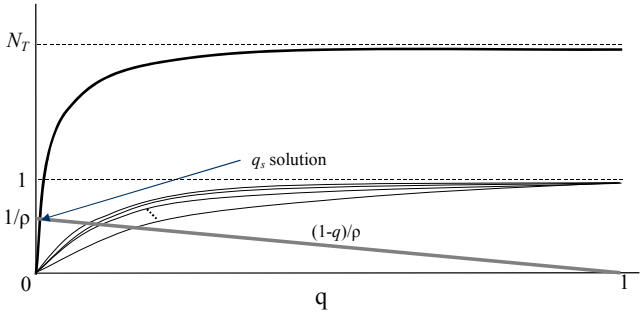


Fig. 2. An Illustration of equation (17).

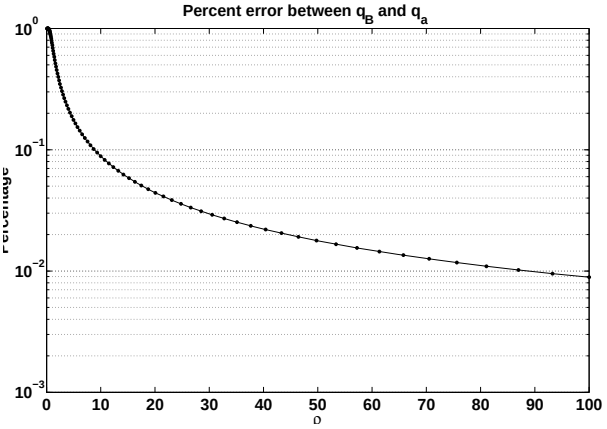


Fig. 3. The percentage error between q_B (bi-section method) and q_a (approximation) for various values of ρ.

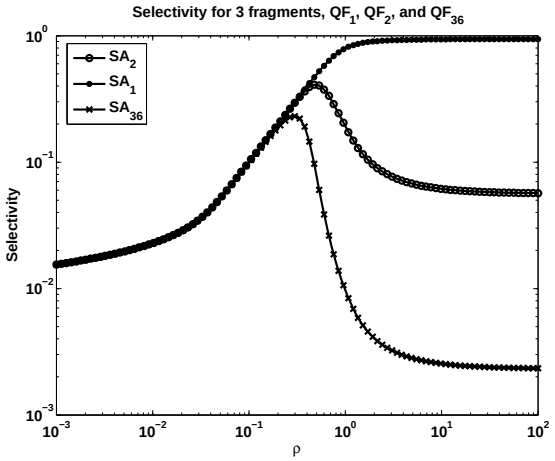
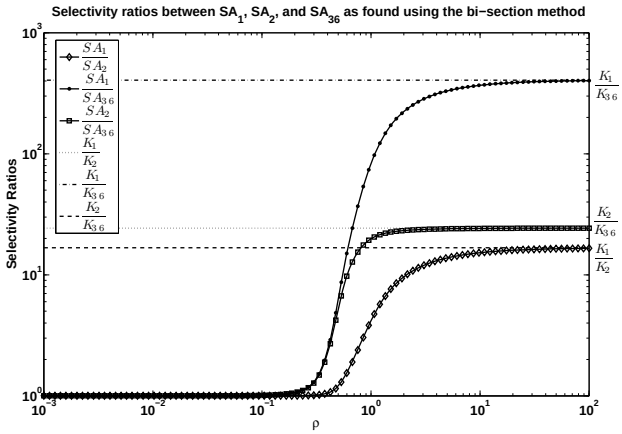
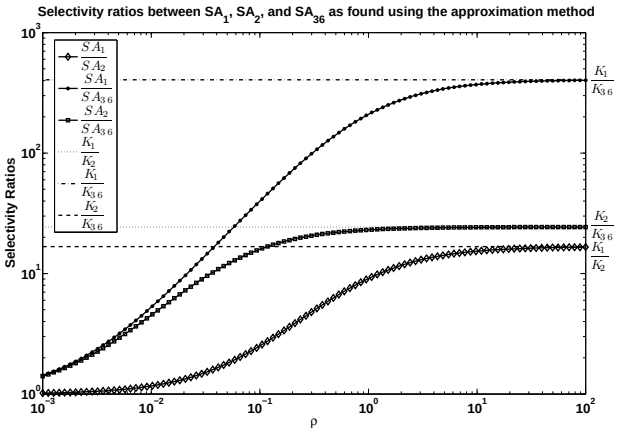


Fig. 4. Selectivity percentage (y-axis) plot for the three fragments under examination for various values of ρ (x-axis) estimated using the bi-section method.



(a)



(b)

Fig. 5. Selectivity ratios (y-axis) for the three fragments under examination for various values of ρ (x-axis) estimated using: *(a)* the bi-section method and *(b)* equation (37). The corresponding equilibrium constant ratios are also plotted using doted lines. It is clear that as ρ increases the selectivity ratios approximate asymptotically the equilibrium constant ratios.

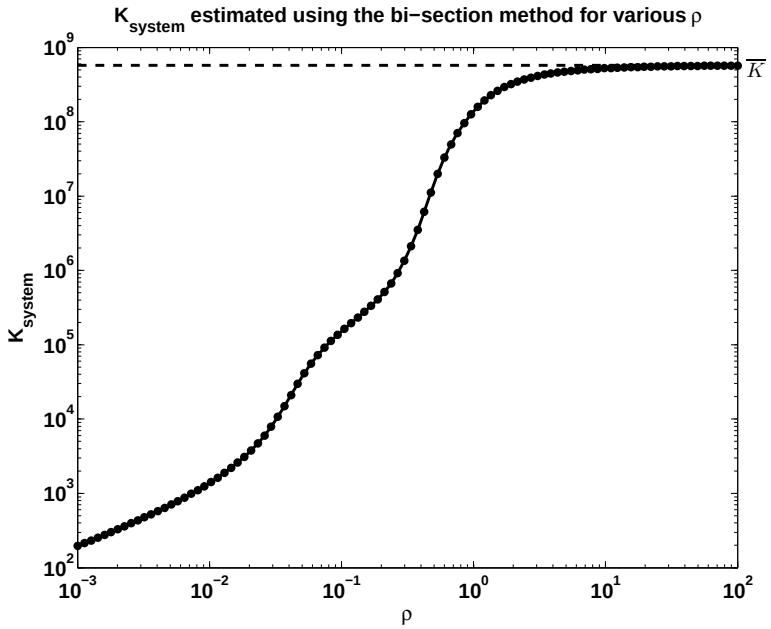


Fig. 6. K_{system} (y-axis) plot for various values of ρ (x-axis) estimated using the bi-section method. $\overline{K}$ is also plotted with a dashed line.