

# Hybridization Error for DNA Mixtures of N Species

J. A. Rose

*Department of Electrical Engineering, The University of Memphis, Memphis, TN, 38152,  
johnrose@memphis.edu*

R. J. Deaton

*Department of Electrical Engineering, The University of Memphis, Memphis, TN, 38152,  
rjdeaton@memphis.edu*

D. R. Franceschetti

*Departments of Chemistry and Physics, The University of Memphis, Memphis, TN, 38152,  
dfrncsch@memphis.edu*

M. Garzon

*Department of Mathematical Sciences, The University of Memphis, Memphis, TN, 38152,  
garzonm@msci.memphis.edu*

S. E. Stevens, Jr.

*Department of Microbiology and Molecular Cell Sciences, The University of Memphis, Memphis,  
TN, 38152, estevens@cc.memphis.edu*

(December 16, 1999)

## Abstract

The statistical behavior of a diverse single stranded DNA (ssDNA) mixture is assuming increasing importance to biotechnology and molecular computing. The ensemble average quantity of interest is defined to be the probability of unwanted hybridization per annealing event, or the computational incoher-

ence ( $\xi$ ). An expression for  $\xi$  is derived in the low dissociation, high  $[Na^+]$  limit, and used to assess the fidelity of various encodings of the annealing biostep of DNA computation.  $\xi$  is then used to demonstrate the feasibility of DNA computations of excellent reliability.

87.14.Gg, 87.15.Aa, 89.70.+c

In 1994, Leonard Adleman demonstrated the feasibility of biocomputing by solving an instance of the Hamiltonian Path Problem (HPP) *in vitro*, using a mixture of single-stranded (ss) DNA molecules and a series of protocols adapted from molecular biology [1]. The tendency for the hybridization reactions to deviate from the set which forms the basis of the computation, however, is recognized to be a primary source of error [2]. The duplex configurations available to a mixture of ssDNA molecules is determined by their specific nucleotide sequences [3]. Several *ad hoc* methodologies for generating sets of ssDNA sequences, or encodings, robust to unplanned modes of hybridization have been proposed [1, 2, 4 – 7]. Although these schemes should offer a means of producing encoding sets with a reduced potential for unplanned, error hybridization, obtaining an accurate estimate of either the absolute or relative rate of hybridization error for such encoding sets requires a more detailed physical-chemical treatment.

The statistical theory of DNA duplex formation may be used to characterize hybridization error in a mixture composed of an arbitrary number of ssDNA species, given a set of encodings, desired hybridizations, and reaction conditions. The perspective chosen is that of a single molecular observer. As this observer will assess the error state of the DNA mixture one hybridized pair at a time, the equilibrium average quantity of interest is the probability that a single hybridization, randomly selected by the molecular observer, is in error. Let this quantity be termed the computational incoherence,  $\xi$ , of the computation.

Although the analysis of duplex formation is concerned primarily with estimating the distribution of ssDNA molecules among accessible ds configurations, complications from ss stacking and intramolecular folding should be considered. The mean free energy of a ss stack) and a ds stack is roughly  $-0.04$  kcal/mol [8] and  $-1.69$  kcal/mol [9], respectively, at pH 7, 1.0 M  $[Na^+]$ , 25°C. Therefore, variations in ds energetics due to differences in ss stacking are typically considered negligible. In addition, all unfolded ssDNAs are assumed to occupy a single configuration of zero free energy [3]. A complete description of the relative occupancies of ds configurations for an ensemble of hybridizing ss species  $i$  and  $j$  is contained in the conformational partition function,  $Z_c^{ij}$ , defined to be the sum of the Gibbs factors ( $Z_k^{ij}$ ) of

all hybridized configurations,  $\{k\}$ :

$$Z_c^{ij} = \sum_{\{k\}} Z_k^{ij} = \sum_{\{k\}} \exp \left[ -\frac{\Delta G_{ijk}^\circ}{RT} \right], \quad (1)$$

where  $\Delta G_{ijk}^\circ$  is the standard Gibbs free energy of formation for configuration  $k$  [3].

The potential for intramolecular folding is typically neglected for quasirandom DNAs [10]. A complication results if the n-ary DNA mixture contains species with a high degree of inverse repeat symmetry. In this case, significant fractions of a palindromic ssDNA species may occupy both a folded hairpin configuration and an associated dimer [11, 12]. The presence of a highly stable, folded configuration will not effect the fractional occupancies of the dsDNA configurations available to a given pair of ssDNA species. The total occupancy of all dsDNA configurations for the pair, however, will be reduced because of competition with the folded ssDNA structure. An equilibrium treatment which neglects the presence of a highly stable hairpin configuration will therefore overestimate the absolute concentrations of dsDNA structures. Reaction conditions can, however, be selected which minimize the relative occupancy of folded configurations. For oligonucleotide solutions of highly palindromic DNAs, the bimolecular duplex is favored over the associated hairpin by a combination of high  $[Na^+]$  (0.5-1.0 M), low  $T_{rx}$  ( $\leq 25^\circ\text{C}$ ), and high initial ssDNA concentration ( $\geq 10^{-4}M$ ) [11]. As counterion-condensation theory predicts that the  $[Na^+]$ -induced difference between duplex and hairpin thermostability is maintained with increasing molecular length [13], this relative favorability of the bimolecular duplex over the hairpin is anticipated for solutions composed of short polynucleotides, as well. In any case, the presence of a palindromic ssDNA species in an n-ary DNA mixture of interest will result in a large error rate, due to the occupancy of the associated bimolecular duplex structure, and should be avoided unless formation of either the hairpin or associated dimer is specifically desired.

It is well known that the number of ds configurations potentially accessible to a ssDNA pair is an exponentially increasing function of length [14]. If attention is restricted to short polynucleotides ( $\leq 100$  nucleotides (nt)), however, configurations arising from multiple nucleation events are sufficiently improbable that their contributions to  $Z_c^{ij}$  may be neglected

[15, 16], effectively reducing the number of configurations to a manageable (polynomial) function of length. In the resulting staggered zipper model,  $Z_k^{ij}$  for each configuration is identical to the statistical weight of the configuration's single duplex. In the special case where the ssDNAs are identical, a doubling of each Gibbs factor for staggered species is required by symmetry [9, 15]. This may be accomplished in practice by considering those configurations which differ only by the interchanging of strands to be distinct species, even though they are in fact indistinguishable [15].

According to the nearest-neighbor model of duplex thermostability, an estimate of  $\Delta G_{ijk}^\circ$  may be obtained by summing the free energies of the nearest-neighbor pairs contained in the duplex, adding an initiation parameter for each duplex terminus, and applying a symmetry correction if the duplex is palindromic [17]. An estimate of the standard free energy of formation for each of the 10 Watson-Crick nearest-neighbor pairs has been reported in [9], and has been demonstrated to adequately predict the thermostability of both oligonucleotides and polynucleotides [17].

Let the computational incoherence,  $\xi$ , be the average probability, at equilibrium, that a hybridization randomly observed within the annealed mixture is hybridized in a manner inconsistent with the rules of the computation. This hybridization may be embedded within a longer chain. For the general hybridization reaction, which is composed of many distinct ssDNA species,  $i$ ,  $\xi$  may be written as

$$\xi = \sum_{i,j \geq i} p(i \wedge j) p(e|i \wedge j), \quad (2)$$

where  $p(i \wedge j)$  is the joint probability that the duplex observed is between DNA species  $i$  and  $j$ , and  $p(e|i \wedge j)$  is the conditional probability that the hybridization is also an error.

Let the function  $\delta_{ijk}$  equal 1 if configuration  $k$  between ssDNA species  $i$  and  $j$  is forbidden by the set of hybridization rules of the computation, and 0 otherwise.  $p(e|i \wedge j)$  is the ensemble average value of  $\delta(ijk)$  for fixed  $i$  and  $j$ , which is given by

$$p(e|i \wedge j) = \frac{\sum_k \delta_{ijk} Z_k^{ij}}{\sum_k Z_k^{ij}} = \frac{Z_e^{ij}}{Z_c^{ij}}. \quad (3)$$

where  $Z_e^{ij}$  is equal to the sum of the statistical weights of all configurations between  $i$  and  $j$  which are forbidden.

$\xi$  is concerned primarily with the statistical details of the individual hybridizations present in an annealed mixture, and not with the statistical details of the set of longer molecular species in which each hybridization may be embedded. A first estimate of  $p(i \wedge j)$  may therefore be obtained by considering an abstract mixture that is identical to the physical mixture, but is constructed by restricting the formation of molecular species to the set of allowable bimolecular duplexes (i.e.: no longer chains). This is equivalent to postulating that the distribution of hybridized species is determined primarily by the initial formation of bimolecular duplexes. Departures observed in the physical mixture from the ratios of hybridized concentrations calculated using this approximation should be small, since the same set of conformal partition functions,  $Z_c^{ij}$ , which governs bimolecular duplex formation, also governs the subsequent formation of longer chains. The quantity  $p(i \wedge j)$  is then approximated by:

$$p(i \wedge j) \approx \frac{C_{ij}}{\sum_{i',j' \geq i'} C_{i'j'}} = \frac{C_i C_j K_a^{ij}}{\sum_{i',j' \geq i'} C_{i'} C_{j'} K_a^{i'j'}}, \quad (4)$$

where the bimolecular duplex ( $C_{ij}$ ) and ssDNA ( $C_i, C_j$ ) concentrations refer to the abstract mixture, and  $C_{ij} = C_i C_j K_a^{ij}$  has been invoked for each association reaction. The equilibrium constant of association,  $K_a^{ij}$  is given by the expression:

$$K_a^{ij} = \beta_{ij} Z_c^{ij}, \quad (5)$$

where  $\beta_{ij}$  is the ratio of the external degrees of freedom (i.e., rotational and translational) of ds to ss configurations. If each association equilibrium is modelled by the simple process, 2 spheres  $\rightarrow$  rod, then  $\beta_{ij}$  may be approximated by the expression

$$\beta_{ij} = \kappa N^{-3}, \quad (6)$$

where  $\kappa$  is the association constant and all ssDNAs are of length  $N$  nt [14]. Substitution yields:

$$p(i \wedge j) \approx \frac{C_i C_j Z_c^{ij}}{\sum_{i',j' \geq i'} C_{i'} C_{j'} Z_c^{i'j'}}. \quad (7)$$

Combining Eqs 2, 3, and 7 yields:

$$\xi \approx \frac{\sum_{i,j \geq i} C_i C_j Z_e^{ij}}{\sum_{i,j \geq i} C_i C_j Z_c^{ij}}. \quad (8)$$

An exact evaluation of equation 8 requires the explicit calculation of the equilibrium concentrations ( $C_i$ ) of each ssDNA species,  $i$ . This calculation requires the simultaneous solution of a large number of coupled equilibrium equations. An expression for  $p(i \wedge j)$  in the limit of low dissociation, however, may be obtained as follows. Assuming the negligibility of folding, the concentration of associated instances of each ssDNA species  $i$  is equal to

$$C_i^H = C_i^0 \Theta_a^i = C_i^0 - C_i, \quad (9)$$

where  $C_i^0$  and  $C_i$  are the initial and equilibrium concentrations of species  $i$ , respectively, and  $\Theta_a^i$  is the fraction of associated stands for species  $i$ . Substitution yields

$$\xi \approx \frac{\sum_{i,j \geq i} C_i^0 C_j^0 (1 - \Theta_a^i) (1 - \Theta_a^j) Z_e^{ij}}{\sum_{i,j \geq i} C_i^0 C_j^0 (1 - \Theta_a^i) (1 - \Theta_a^j) Z_c^{ij}}. \quad (10)$$

The behavior of  $\xi$  in the low dissociation limit is obtained by allowing  $\Theta_a^i \rightarrow 1$  for all ssDNA species. This yields:

$$\xi \approx \frac{\sum_{i,j \geq i} C_i^0 C_j^0 Z_e^{ij}}{\sum_{i,j \geq i} C_i^0 C_j^0 Z_c^{ij}}. \quad (11)$$

If all  $C_i^0$  are identical, this expression reduces to:

$$\xi \approx \frac{\sum_{i,j \geq i} Z_e^{ij}}{\sum_{i,j \geq i} Z_c^{ij}}. \quad (12)$$

Each ssDNA species is encoded to participate in at least one strongly favorable mode of hybridization. Adherence to the low dissociation limit may therefore be established by requiring that the reaction temperature ( $T_{rx}$ ) be sufficiently low to allow substantial formation of each planned duplex. The *melting temperature* ( $T_m$ ) of a DNA duplex is defined as temperature at which the average fraction of stacked base pairs in an ensemble of identical

instances of the duplex is equal to  $\frac{1}{2}$  [14]. In addition, the melting transition of any short polynucleotide will necessarily have a substantial width  $\Delta T_m$  [18]. Let  $\{T_m\}$  and  $\{\Delta T_m\}$  be the set of melting temperatures and melting curve widths of the set of planned ds duplexes, respectively. For duplexes shorter than about 300 bps, the the initial 20% of the melting transition is governed completely by  $Z_c$  [10], indicating the negligibility of dissociation in this regime. It is therefore reasonable to require that  $T_{rx}$  lie below a maximum defined by the minimum value of  $T_m - \frac{1}{2}\Delta T_m$  for the set of planned modes of hybridization.

In the development of  $\xi$ , several approximations are used to facilitate tractability. Use of the staggered zipper model avoids calculating a statistical weight for each member of the set of looped configurations. Although reasonable for general DNAs of length  $\leq 100$  nt, and for quasi-random DNAs of length  $\leq 200$  nt, this approximation is invalid for longer polynucleotides. Restricting attention to the low dissociation limit avoids a calculation of the set of equilibrium ssDNA concentrations. This approximation will, however, apply only to an annealing reaction performed in the low dissociation (low  $T_{rx}$ ) limit. Furthermore, for the n-ary DNA mixture which contains palindromic ssDNAs,  $\xi$  applies only to conditions of high  $[Na^+]$  (0.5-1.0 M) and high initial ssDNA concentrations ( $C_i^0 \geq 10^{-4}$  M). The range of conditions for which  $\xi$  assumes a tractable form corresponds roughly to typical experimental conditions [1, 9, 11].

$\xi$  is shown (Figure 1) for the following sets of ssDNA for Adleman’s original experiment: (a) Adleman’s set [1], (b) an encoding set derived from the modified deBruijn sequence suggested in [4], (c, d) sets with good and bad modified Hamming-distance properties [2], and (e, f) sets with good and bad stringency properties [5]. Equal initial ssDNA concentrations ( $C_i^0 = 4 \times 10^{-4} \text{M}, \forall i$ ), pH 7.0, and 1.0 M  $[Na^+]$  conditions are assumed. In order to estimate the impact of nearest-neighbor parameter uncertainty upon  $\xi$ , the uncertainty in  $\Delta G_{ijk}$  for each configuration  $ijk$  was calculated as suggested in [9]. The resulting uncertainty in  $\xi$  was then estimated by standard error propagation. In each case  $\frac{\Delta \xi}{\xi} \leq 0.01$ .

Encoding set (g) (Figure 2) was evolved using a standard genetic algorithm with  $-\log_{10} \xi$  as the applied measure of fitness. This set is predicted to have an error rate at  $10^\circ C$  of less

than  $10^{-14}$  errors/hybridization. This compares favorably with the predicted error frequency estimated by Intel for the Pentium chip (roughly  $10^{-9}$  errors/division) [19]. This result demonstrates that despite a reputation for infidelity, computations of excellent reliability should in principle be implementable in DNA. As a case in point, consider the implementation of encoding (g) under the experimental conditions used in [1]. Since the resulting annealing ensemble will contain roughly  $10^{14}$  hybrids, fewer than 1 error hybridization, on average, is predicted to exist in solution, at equilibrium.

$\xi$  has immediate applicability as a tool for the analysis and design of encodings for DNA computation. The ability to estimate the statistical probability of unwanted interaction in a diverse annealing mixture also has implications for other areas of biotechnology. An immediate application to DNA chip design is the assessment and development of sets of DNA molecules whose members exhibit minimal mutual hybridization potential.  $\xi$  may also be used to estimate the interaction potential between members of sets of ssDNA probes, a consideration of relevance to antisense technology.

In conclusion, the principles of statistical thermodynamics have been applied to characterize the equilibrium probability of error for the diverse mixture of annealing ssDNA molecules. The resulting measure of error, the computational incoherence ( $\xi$ ) has been used to estimate the fidelity of various sets of extant encodings. Finally,  $\xi$  has been used to demonstrate the theoretical feasibility of implementing computations of excellent fidelity using DNA.

## ACKNOWLEDGEMENTS

We are grateful to A. Parrill from the University of Memphis Department of Chemistry, to M. Hagiya from the University of Tokyo Department of Information Sciences, to A. Suyama from the University of Tokyo Institute of Physics, and to J. McCaskill from the German National Research Center for Information Technology, for their helpful comments during preparation of this manuscript.

## REFERENCES

- [1] L. M. Adleman, *Science*, **266**, 1021 (1994).
- [2] R. Deaton, M. Garzon, R. E. Murphy, J. A. Rose, D. R. Franceschetti, S. E. Stevens, Jr., *Phys. Rev. Lett.* **80**, 417 (1998).
- [3] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry, Part III: The Behavior of Biological Macromolecules*, (Freeman, New York, 1980).
- [4] W. D. Smith, in *DNA Based Computers*, edited by R. J. Lipton and E. B. Baum, (American Mathematical Society, Providence, RI, 1996), 121.
- [5] B-T Zhang and S-Y Shin, in *Proceedings of the Third Annual Genetic Programming Conference, University of Wisconsin at Madison, 1998*, (Morgan Kaufman, San Francisco, 1998), 735.
- [6] A. Hartemink, D. Gifford, in *DNA Based Computers III*, edited by H. R. Rubin and D. H. Wood, (American Mathematical Society, Providence, Rhode Island, 1999), 25.
- [7] A. Frutos, Q. Liu, A. Thiel, A. Sanner, A. Condon, L. Smith, R. Corn, *Nucleic Acids Research* **25**, 4748 (1997).
- [8] G. Vesnaver and K. Breslauer, *Proc. Natl. Acad. Sci.*, **88**, 3569 (1991).
- [9] H. T. Allawi and J. SantaLucia, Jr., *Biochemistry* **36**, 10581 (1997).
- [10] R. M. Wartell and A. S. Benight, *Physics Reports (Review Section of Physics Letters)* **126**, 67 (1985).
- [11] P. Ross, F. Howard, M. Lewis, *Biochemistry* **30**, 6269 (1991).
- [12] L. Xodo, G. Manzini, F. Quadrifoglio, G. van der Marel, J. van Bloom, *Nucleic Acids Research* **19**, 1505 (1991).
- [13] M. T. Record and T. Lohman, *Biopolymers* **17**, 159 (1978).

- [14] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers*, (Academic Press, New York, 1970).
- [15] J. Applequist and V. Damle, *J. Am. Chem. Soc.* **87**, 1450 (1965).
- [16] A. S. Benight and R. M. Wartell, *Biopolymers* **22**, 1409 (1983).
- [17] J. SantaLucia, Jr., *Proc. Natl. Acad. Sci* **95**, 1460 (1998).
- [18] J. G. Wetmur, in *DNA Based Computers III*, edited by H. R. Rubin and D. H. Wood, (American Mathematical Society, Providence, 1999), 1.
- [19] Pratt, V. (1994) [http://boole.stanford.edu/pub /PENTIUM/individual.bugs/bug21](http://boole.stanford.edu/pub/PENTIUM/individual.bugs/bug21).

## FIGURES

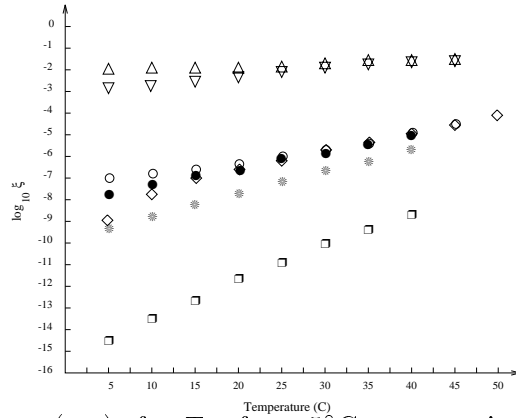


FIG. 1.  $\xi$  for encoding sets (a-g), for  $T_{rx}$  from  $5^\circ\text{C}$  to an estimated  $T_{rx}^{max}$ . (a)  $\bullet$ , Adleman ( $38.9^\circ\text{C}$ ); (b)  $\circ$ , Smith ( $42.3^\circ\text{C}$ ); (c)  $\diamond$ , Deaton, et al.: good ( $49.4^\circ\text{C}$ ); (d)  $\nabla$ , Deaton, et al.: bad ( $47.5^\circ\text{C}$ ); (e)  $*$ , Zhang & Shin: good ( $38.4^\circ\text{C}$ ); (f)  $\triangle$ , Zhang & Shin: bad ( $42.4^\circ\text{C}$ ); (g)  $\square$ , Evolved Set ( $37.3^\circ\text{C}$ ).  $T_{rx}^{max}$  values, listed in parenthesis, are computed for  $C_i^0 = 4 \times 10^{-4}\text{M}$ , for each ssDNA species.

Vertex	ssDNA Encoding (5' to 3')
0	GCAGATAGACAAGAGATTAG
1	TCCTACTGAGCCCTGTAATT
2	CGCCGCACGCTAAGTGGGAT
3	CATATTTTAGAAACGACCAT
4	ACAATTTGATACTATGTTAT
5	CACCCGAGGCAATAAAGGTT
6	CATAATGTGTTGGAATATAG

FIG. 2. Evolved Encoding Set (g).