A Statistical Mechanical Treatment of Error in the Annealing Biostep of DNA Computation

John A. Rose
Electrical Engineering
The University of Memphis
Memphis, TN 38152
jrose@lear.csp.ee.memphis.edu

Russell J. Deaton
Electrical Engineering
The University of Memphis
Memphis, TN 38152
rjdeaton@memphis.edu

Donald R. Franceschetti
Chemistry & Physics
The University of Memphis
Memphis, TN 38152
dfrncsch@memphis.edu

Max Garzon
Computer Science
The University of Memphis
Memphis, TN 38152
garzonm@msci.memphis.edu

S. Edward Stevens, Jr
Microbiology and Molecular Cell Sciences
The University of Memphis
Memphis, TN 38152
estevens@cc.memphis.edu

Abstract

The relaxation of an ensemble of ss-DNA oligonucleotides to an equilibrium of annealed, or hybridized, double-stranded molecules has been harnessed for massively parallel computation. The annealing reactions, however, have the potential to produce error. Various heuristic methods for generating sets of single-stranded DNA (encodings) with minimal potential for hybridization error have emerged. Nevertheless, an adequate model of error demands not only implementation details that minimize the impact of physical uncertainties on the computational process, but also an unambiguous, quantitative measure of confidence with which to associate computational results. In this study, the principles of equilibrium statistical mechanics are used to derive an expression for the ensemble average equilibrium probability of an error hybridization per annealing event, given a specified set of planned hybridizations, a set of encodings, and a reaction temperature.

1 Introduction

The potential for improper hybridization to generate error during annealing was recognized by Adleman in his initial, ground-breaking work [Adleman, 1994]. Random sequences were considered adequate to encode the small instance graph because of the unlikelihood of an accidental Watson-Crick(WC) complementary match from among the large (4^n) number of DNA codewords of length n. The inadequacy of random sequences for the reliable encoding of general graphs has been previously addressed [Deaton et al., 1996].

Several heuristic methods of constructing codeword sets robust with respect to hybridization errors have been proposed. The utility of restriced DNA alphabets in preventing molecular folding was examined in [Mir, 1997]. The partial elimination of mismatches of length k or greater by using radix-k modified De-Bruijn sequences for vertex encoding was suggested in [Smith, 1996]. Modified Hamming encodings, in which members and their WC-complements are a minimal Hamming distance apart, are useful for preventing proper-frame mismatches [Deaton et al., 1996]. The use of Hamming encodings for the variable regions of anchored encoded molecules has been demonstrated to be effective in preventing error hybridizations during the Mark biostep [Frutos et al., 1997]. Finally, a Monte-Carlo based molecular programming technique has been used to evolve encodings with good overall

stringency properties [Zhang and Shin, 1998]. Other treatments include [Garzon et al., 1997].

In each case the encoding heuristic may be useful for minimizing the potential for hybridization error, given a specific problem instance and computational architecture. A qualitative minimization of error, however, is inadequate. Many physical processes have been used to implement models of computation. In each case, the ability of the physical process to successfully implement a basic computational operation is constrained by the tendency of that process to deviate from ideal behavior. As a result of the basic uncertainty inherent in any physical system, such a process can never be entirely error-free. Therefore, it is important to provide not only an architecture that minimizes the impact of physical uncertainties on the computational process, but also a physically relevant, quantitative measure of confidence with which computational results are to be associated.

As a result of its ease of calculation, the concept of a duplex melting temperature, T_m, has enjoyed wide use as an intuitive, physical measure of duplex stability. An examination of the T_m's of planned versus unplanned hybrids has been demonstated to be a useful diagnostic method for the screening [Hartemink and Gifford, 1997] and design [Frutos et al., 1997] of computational sets of molecules. However, the use of T_m to predict error is heuristic in nature, and is inadequate for a precise characterization. From a theoretical standpoint, the application of T_m as a rigorous measure of duplex stability is out of context. T_m is derived from the consideration of an ensemble of base pairs, as opposed to an ensemble of instances of any particular double stranded (ds) configuration. As such, T_m applies properly as a probabilistic measure of the (stacking) state of a stochastically observed pair of adjacent base pairs, and not to the (bound/unbound) state of a given ds species.

Practically, the use of T_m as a strict predictor of duplex stability is equivalent to assuming a nonprobabilistic two-state model of DNA melting. The melting curve of any finite length DNA duplex will necessarily have a substantial width, particularly if at least one member of the hybridized pair is an oligonucleotide [Wetmur, 1997]. This deviation from all-or-none behavior implies that a knowledge of the T_m of a particular duplex is insufficient to predict the state (bound or unbound) of an observed instance at arbitrary reaction temperatures. Relative rankings in the melting temperatures of two duplexes need not correspond to relative rankings in energetic stability at a given temperature [Cantor & Schimmel, 1980]. The appropriate measure of duplex stability is the Gibbs free energy of formation, ΔG, of the configuration at experimental conditions of interest.

The purpose of this work is to provide a physically grounded measure of error that permits a principled comparison of heuristic measures of encoding quality. In the following treatment, an algorithm is described which utilizes the techniques of equilibrium statistical mechanics to estimate the ensemble average probability of an error hybridization per hybridization event, given a specified set of planned hybridizations, a set of encodings, and a reaction temperature. This concept of error is defined to be the *computational incoherence*, ξ, of the annealing reaction. This algorithm has been implemented in the form of a Java system and applied to estimate the error probability vs reaction temperature for Adleman's original encoding set, as well as that of various sets of encodings of Adleman's original 7 vertex graph, each generated using one of the heuristics outlined above.

2 Methods

The principles of equilibrium statistical mechanics have been successfully applied to the problem of reproducing the experimentally observed melting behavior of DNA [Wartell and Benight, 1985, Paner et al., 1990]. A statistical derivation rests on the assumption that all average equilibrium properties of a physical system are derivable from that system's partition function, Z. In DNA duplex formation, the system of interest is a pair of single-stranded DNA molecules. An analysis of duplex formation is concerned primarily with the details of double-stranded (ds) states. Allowed macrostates will, however, include both single stranded (ss) configurations and configurations containing both ds and ss regions. For a particular configuration at constant temperature and pressure, the number of accessible quantum microstates is estimated by the statistical weight, or *Gibbs factor*, Z_i of the configuration. At constant volume, Z_i is related to the Gibbs free energy of the configuration, ΔG_i, by $\Delta G_i = -RTlnZ_i$, where R is the molar gas constant and T is the absolute temperature. ΔG is measured relative to the zero energy state, defined to be equal to that of the unstacked single strands [Cantor & Schimmel, 1980].

The average contribution to ΔG° at $25^\circ C$ made by a pair of stacked ss bases and ds base pairs is approximately -0.04 [Vesnaver and Breslauer, 1991],

and -1.69 kcal/mol [Allawi and SantaLucia, 1997][1], respectively. Contributions from ss regions to the ΔG for bound configurations are therefore typically regarded to be negligible. This assumption reduces Z to the conformal partition function, Z_c, which contains a complete statistical description of the details of duplex formation. Details regarding ss configurations must be considered separately. Z_c is constructed by summing the statistical weights, Z_i, of all configurations, i. The ensemble average probability of occupation of the system in any particular configuration i is given by the ratio of Z_i to Z_c.

There are various models of duplex formation, each of which has its own range of applicability [Cantor & Schimmel, 1980]. For polynucleotides longer than $\approx$ 200 bps, configurations resulting from several independent nucleation events can contribute significantly to Z_c [Benight et al., 1981]. This necessitates use of the general statistical model, in which the Gibbs factor of each configuration is given by the product of three types of stability parameters [Wartell and Benight, 1985]: (1) the equilibrium constant of formation, s_j of each duplex region j, (2) a cooperativity parameter, σ, for each internal loop, and (3) a loop entropy parameter, f(m) for each unbonded loop of m base pairs. The cooperativity parameter, σ, accounts for the fact that two stacking interactions are broken when an internal base pair is opened, whereas only a single stacking interaction is broken on the opening of an end base pair. The loop entropy parameter, $f(m)$, is the probability that two WC-complementary base pairs at the end of a loop of length m are in a spatial relationship that allows bonding.

The equilibrium constant of duplex formation, s_j, for a duplex j is related to the total difference in the free energies between the stacked and unstacked states of the duplex, ΔG_j, by the expression

$$s_j = e^{\frac{-\Delta G_j}{RT}}. \qquad (1)$$

ΔG_j for duplex j may be estimated by summing the free energy changes of stacking each successive pair of adjacent hydrogen-bonded base pairs $\Delta G_{i,i+1}$ during formation of the duplex from the totally unstacked state, in accordance with the nearest-neighbor model of duplex formation. An estimate of the nearest-neighbor parameters $(\Delta H^0_{i,i'}, \Delta S^0_{i,i'})$ for each of the 10 WC nearest-neighbor pairs has been reported in [Allawi and SantaLucia, 1997]. This single set of parameters has been demonstrated to adequately predict the Gibbs free energy of formation for longer duplexes [SantaLucia, 1998], as well. The total free energy of formation of a duplex j is then:

$$\Delta G_j = \sum_i \Delta G_{i,i+1} + \Delta G_{init} + \Delta G_{sym}, \qquad (2)$$

where the sum is perfomed over adjacent pairs of hydrogen bonded base pairs i and $i+1$, ΔG_{init} is the free energy of initiation at each end, and ΔG_{sym} is a symmetry correction.

Restricting attention to short (length $\leq$ roughly 200 bps) DNAs allows an enormous simplification in the form of Z_c. Configurations resulting from two or more distinct nucleation events may be regarded as sufficiently improbable that contributions to Z_c may be neglected [Benight et al., 1981]. This assumption results in the statistical, or *staggered zipper model* [Cantor & Schimmel, 1980], in which configurations resulting from the hybridization of ssDNA's in any frame are considered, but those containing internal loops are discarded. In this model, Z_i for each configuration is equal to the equilibrium constant of formation, s_j, of the single duplex formed.

The annealing phase of a DNA computation consists of a massive number of identical copies of each member of a *set* of DNA single strands, which have been encoded to hybridize according to a set of hybridization rules. In order to assure that duplex formation proceeds in accordance with the principles of equilibrium statistical mechanics, this mixture must be allowed to reach equilibrium prior to the addition of any secondary enzymes (i.e.: ligase). Reaction products consist of a set of chains of hybridized single strands, in equilibrium with a set of unhybridized ssDNA. At equilibrium, this hybridized subcomponent may be regarded as a massive ensemble of independently stable DNA hybridizations. The most elementary computational event in an annealing biostep may then be identified to be the single hybridization. The remainder of this discussion will be focused on characterizing the error probability of this basic event within an annealed ensemble, at equilibrium.

From the ensemble of DNA hybrids present in the annealing mixture, let a *measurement* be the random selection and observation of the error state of a pair of hybridized ssDNA's. This pair may be imbedded within a longer chain. From a computational perspective, this stochastic system will be observed to lie in *exactly* one of two distinct states: it will be hybridized in a manner either consistent (state $\overline{e}$), or inconsistent (state e) with the hybridization rules of the computation. The probability that this hybridization represents an "error", from the perspective of the DNA

[1] Data normalized to pH 7, 1.0 M $[Na^+]$, and $25°C$.

computation, is equal to the equilibrium ensemble average probability that an observed DNA duplex lies in state e. Let this probability be termed the *computational incoherence*, ξ, of the annealing reaction.

The simplest case of annealing is an ensemble of instances of a single ssDNA species. Given the random observation of a single hybridized pair of ssDNA's, the ensemble average probability that this duplex will be found to be in an error binding mode is given by the ratio of Z_e to Z_c, where Z_e is the sum of the Gibbs factors of all error configurations.

In the most general form of annealing reaction, there will exist many distinct ssDNA species, each of which may be allowed to assume an arbitary initial concentration. The impact of differences in the relative abundance of ssDNA species upon the probability of an error hybridization may be accounted for using the following reasoning. Given the observation of a random, hybridized pair of oligonucleotides, the probability that the pair will be occupy state e may be written as the sum:

$$p(E|h) = \sum_{i,j \geq i} p(i \wedge j) p(e|i \wedge j), \tag{3}$$

where $p(i \wedge j)$ is the joint probability that the duplex observed is between members of the species i and j, and $p(e|i \wedge j)$ is the conditional probability of an error given this event. $p(e|i \wedge j)$ is given by the expression:

$$p(e|i \wedge j) = \frac{Z_e^{i,j}}{Z_c^{i,j}}. \tag{4}$$

An expression for $p(i \wedge j)$ may be constructed combinatorically, as follows. Let $W_{i,j}$ be the number of distinct ways in which a pair of instances of ssDNA species i and j could be assembled. The equiprobability of each manner of assembly follows from the indistinguishability of identical members of a given species and the axiom of equiprobability of consistent quantum microstates. Let n_i^H refer the number of instances of species i involved in hybridizations of any molecularity. Then $W_{i,j}$ is given by the product,

$$W_{i,j} = n_i^H n_j^H Z_c^{i,j}. \tag{5}$$

$p(i \wedge j)$ is the ratio of $W_{i,j}$ to $\sum_{i,j} W_{i,j}$, or

$$p(i \wedge j) = \frac{C_i^H C_j^H Z_c^{i,j}}{\sum_{i',j' \geq i'} C_{i'}^H C_{j'}^H Z_c^{i',j'}}. \tag{6}$$

where C_i^H is the number of moles/volume of hybridized species i. Combining expressions 3, 4, and 6 yields:

$$\xi = \frac{\sum_{i,j \geq i} C_i^H C_j^H Z_e^{i,j}}{\sum_{i,j \geq i} C_i^H C_j^H Z_c^{i,j}}. \tag{7}$$

Given that the initial concentration of species i is given by C_i^0, C_i^H is related to the equilibrium concentrations of ss i by the conservation relation, $C_i^H = C_i^0 - C_i$. Exact evaluation of ξ thus requires a complete knowledge of the equilibrium concentrations of all ssDNA species, C_i. A substantial simplification may be achieved by restricting attention to conditions under which the effects of dissociation are negligible for all i. In this case, $C_i^H C_j^H Z_c^{i,j} \approx C_i^0 C_j^0 Z_c^{i,j}$, and ξ reduces to:

$$\xi \approx \frac{\sum_{i,j \geq i} C_i^0 C_j^0 Z_e^{i,j}}{\sum_{i,j \geq i} C_i^0 C_j^0 Z_c^{i,j}}. \tag{8}$$

Negligibility of dissociation ($C_i << C_i^0, \ \forall i$) may be established by a proper choice of reaction temperature ($T_{reaction}$). In order for the annealing reaction to have a reasonable probability of completion for all planned modes of hybridization, $T_{reaction}$ must be held low enough to allow substantial formation of the least favorable planned duplex (e.g.: $T_{reaction} \leq Inf\{T_m\}$, where $\{T_m\}$ is the set of $T_m's$ for planned duplexes). This condition is not, however, sufficient to guarantee the negligibility of dissociation. For duplexes shorter than about 300 bps it is essential to account for the impact of dissociation in order to assure agreement with experiment, even at temperatures marginally beneath T_m [Benight et al., 1981]. Since the initial 20% of the transition is governed completely by Z_c [Wartell and Benight, 1985], a reasonable criterion is to restrict $T_{reaction}$ to lie beneath a maximum defined by $T_{reaction}^{max} \leq Inf\{T_m - \frac{1}{2}\Delta T_m\}$, where ΔT_m is the full width at half maximum of the differential melting curve [Wetmur, 1997]. For the annealing reaction in which the C_i^0 are approximately identical, ξ reduces further to:

$$\xi \approx \frac{\sum_{i,j \geq i} Z_e^{i,j}}{\sum_{i,j \geq i} Z_c^{i,j}} \tag{9}$$

3 Results

The above methodology for calculation of ξ has been implemented in the form of a Java package, and applied to the following sets of encodings of Adleman's graph: (a) Adleman's original encoding set, (b) an encoding set derived from the DeBruijn sequence cited in [Smith, 1996], (c,d) codeword sets evolved to be good and bad, respectively, under the Hamming encoding scheme suggested in [Deaton et al., 1996], (e) an encoding set randomly generated under the use of the a restricted DNA alphabet as suggested in [Mir, 1997], (f) an encoding set evolved to have good stringency properties in [Zhang and Shin, 1998], and (g) a randomly generated set of encodings. ξ is reported for each of these encoding sets for $T_{reaction}$ ranging from

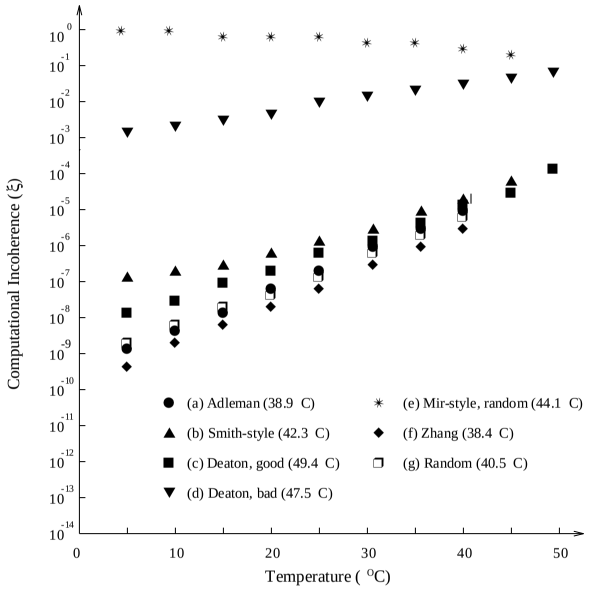

Figure 1: ξ for heuristically generated encodings (a-g), reported from $5°C$ to $T^{max}_{reaction}$, in $5°C$ increments.

$5°C$ to the estimated $T^{max}_{reaction}$ of the encoding set, as listed, in $5°C$ intervals (Figure 1). All ξ values are computed for neutral pH, 1.0 M $[Na^+]$, and uniform C^0_i.

4 Discussion

In addition to the set of hybridizations in which a non-palindromic species may participate, a species with a high level of inverse repeat symmetry may additionally form both a monomolecular hairpin and an associated bimolecular dimer. Although the impact of the dimer is considered in the derivation of ξ, the concentration of folded ssDNAs, C^{hp}_i, has been assumed to have negligible impact on each C^H_i ($C^{hp}_i << C^H_i$, $\forall i$). This assumption is not valid for arbitrary reaction conditions.

The potential for hairpin formation is a function of reaction conditions, ss concentration, and the level of inverse repeat symmetry of each species. Studies on the duplex-hairpin transition of fully and partially palindromic oligonucleotides have demonstrated that formation of the bimolecular duplex is strongly favored over the hairpin at the combination of low temperature and high salt concentration ($[Na^+]$ = 1.0 M) characteristic of an annealing reaction [Marky and Blumenfeld, 1983]. As counterion-condensation theory predicts the change in $\frac{d\Delta G}{dln[Na^+]}$, as a function of length, to be nearly identical for the duplex and hairpin species for DNAs longer than oligonucleotide length [Record and Lohman, 1978], the relative favorability of duplex formation should be observed for poly-, as well as oligonucleotides. The tendency for sufficiently dilute ss concentrations ($C^0_i < 10^{-4}M$) to equalize the favorability of duplex and hairpin formation if low $T_{reaction}$ and high salt concentration [Ross et al., 1991] are not rigorously maintained ($[Na^+]$ = 0.1-0.5 M @ $T >= T_{room}$), however, should be recognized.

The ability to apply reaction conditions which enforce the negligibility of hairpin formation does not imply that encoding sets containing palindromic members are suitable for annealing reactions. In actuality, such an encoding set will always be characterized by a poor ξ, due to the large error contribution, $Z^{i,i}_e$, made by each palindromic dimer.

A discussion of the relative merits of the various heuristic encoding schemes is beyond the scope of this work. However, several general observations on the overall state of error encoding in DNA computation can be made. No encoding scheme has yet produced an encoding set with a ξ smaller than Adleman's original randomly generated encoding set by more than a single order of magnitude. The encoding scheme, (f), which did marginally exceed the performance of random encodings did so at great computational expense. Encoding schemes designed to prevent a specific form of error, such as false positives (c), or hairpins (e), are observed here to do so at the expense of overall hybridization fidelity relative to randomly generated encodings (a, g).

Within the performance limits of standard DNA ligases ($12°C - 30°C$ for T4 DNA ligase and $10°C - 25°C$ for E coli DNA ligase) the minimum error rate achieved by any encoding set is roughly 1 in 10^{-9}. At first glance, this appears to be an impressively small error rate. However, the huge number of hybridizations occuring during annealing (roughly 10^{14} in Adleman's original experiment) indicates the presence of numerous error events.

A model of error appropriate for annealing should not be misconstrued with a complete model of error for DNA computation. Annealing reaction products are often intended to serve as a precursor to a subsequent secondary biostep (e.g. ligase and ssDNA exonuclease in ligate and mark, respectively). Viewed from the perspective of the DNA computation as a whole, the impact of each error mode of hybridization on reliability will not be uniform, but rather will aquire an enzyme-dependent context sensitivity. Only a certain subset of occupied error binding modes may actually be transmitted as error events at the end of a secondary biostep. For instance, error hybridiza-

tions which occupy a configuration suitable for ligation with an adjacent non-error hybridization will have a greatly enhanced probability of ligation, compared to hybridizations in other error configurations. Similarly, an error hybridization will only provide spurious protection from ssDNA exonuclease if it occupies a configuration containing a double-stranded DNA region within a single base pair of the 3' end of an anchored molecule. A complete model of error must therefore evaluate the impact of enzymatic context sensitivities on the computational process.

5 Conclusion

In this work, the principles of equilibrium statistical mechanics have been used to characterize the error behavior of an annealing reaction. An expression for the equilibrium ensemble average probability of an error hybridization per hybridization event has been derived. Reaction conditions appropriate and inappropriate for its use have been discussed. The error propensities of various sets of encodings, each generated using a currently accepted heuristic methodology, have been evaluated. Results suggest the inability of these methodologies to improve substatially the reliability of DNA computations over that expected of a randomly generated encoding.

References

[Allawi and SantaLucia, 1997] H. T. Allawi, J. SantaLucia, Jr., Biochemistry **36**, 10581 (1997).

[Adleman, 1994] L. M. Adleman, Science, **266**, 1021 (1994).

[Benight et al., 1981] A. Benight, R. Wartell, and D. Howell, Nature **289**, 15 (1981).

[Cantor & Schimmel, 1980] C. R. Cantor, P. R. Schimmel, *Biophysical Chemistry, Part III: The Behavior of Biological Macromolecules* (Freeman, New York, 1980).

[Deaton et al., 1996] R. Deaton, M. Garzon, R. E. Murphy, J. A. Rose, D. R. Franceschetti, S. E. Stevens, Jr., Physical Review Letters, **80**, 417 (1998).

[Frutos et al., 1997] A. Frutos, Q. Liu, A. Thiel, A. Sanner, A. Condon, L. Smith, R. Corn, Nucleic Acids Research **25**, 4748 (1997).

[Garzon et al., 1997] M. Garzon, P. Neathery, R. Deaton, R. Murphy, D. R. Francescetti, S. E. Stevens, Jr., *Proceedings of the Second Annual Genetic Programming Conference, Stanford University, 1997*, (Morgan Kauffman, San Francisco, 1997).

[Hartemink and Gifford, 1997] A. Hartemink, D. Gifford, *Preliminary Proceedings of the Third Annual Meeting on DNA Based Computers, University of Pennsylvania, 1997*, DIMACS Proc. Series (American Mathematical Society, Providence, RI, 1997).

[Marky and Blumenfeld, 1983] L. Marky, K. Blumenfeld, Biopolymers **22**, 1246 (1983).

[Mir, 1997] K. U. Mir, *DNA Based Computers II, Princeton University, 1999*, DIMACS Proc. Series (American Mathematical Society, Providence, RI, 1997).

[Paner et al., 1990] T. Paner, M. Amaratunga, M. Doktycz, A. Benight, Biopolymers **29**, 1715 (1990).

[Record and Lohman, 1978] M. T. Record, T. Lohman, Biopolymers **17**, 159 (1978).

[Ross et al., 1991] P. Ross, F. Howard, M. Lewis, Biochemistry **30**, 6269 (1991).

[SantaLucia, 1998] J. SantaLucia, Jr., Proc. Natl. Acad. Sci **95**, 1460 (1998).

[Smith, 1996] W. D. Smith, *DNA Based Computers, Princeton University, 1996*, DIMACS Proc. Series (American Mathematical Society, Providence, RI, 1996).

[Vesnaver and Breslauer, 1991] G. Vesnaver, K. Breslauer, Proc. Natl. Acad. Sci., **88**, 3569 (1991).

[Wartell and Benight, 1985] R. M. Wartell, A. S. Benight, PHYSICS REPORTS (Review Section of Physics Letters) **126**, 67 (1985).

[Wetmur, 1997] J. G. Wetmur, *Preliminary Proceedings of the Third Annual Meeting on DNA Based Computers, University of Pennsylvania, 1997*, DIMACS Proc. Series (American Mathematical Society, Providence, RI, 1997).

[Zhang and Shin, 1998] B-T Zhang, S-Y Shin, *Proceedings of the Third Annual Genetic Programming Conference, University of Wisconsin at Madison, 1998*, (Morgan Kauffman, San Francisco, 1998).