

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)**ScienceDirect**

Physics Procedia 68 (2015) 37 – 42

Physics

**Procedia**

28th Annual CSP Workshop on “Recent Developments in Computer Simulation Studies in Condensed Matter Physics”, CSP 2015

## Thermodynamics of Denaturation Transition of DNA Duplex Oligomers in the Context of Nearest Neighbor Models: a short review

João Carlos de Oliveira Guerra

*Instituto de Física, Universidade Federal de Uberlândia, Uberlândia, MG, Brazil*

---

### Abstract

In this article, it is shown how additive physical properties of DNA double strand polymers can be expanded in terms of 8 irreducible parameters. This provides self-consistency relations among the corresponding 10 duplex dimer contributions. To allow for oligomer analysis, initiation parameters must be added, and, this adds extra degrees of freedom to the fore mentioned parameters. This modelling permits that nucleation free energies could be calculated as a function of the mean global composition of the chain.

© 2015 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of The Organizing Committee of CSP 2015 Conference

*Keywords:* Thermodynamics of DNA duplex oligomers; Initiation free energy; Nucleation free energy; Irreducible parameters for free energy.

---

### 1. Introduction

Many physical properties of DNA or RNA sequences can be calculated from a number of algorithms in the context of nearest-neighbor (NN) models. NN models give linear representations for experimental measurements on nucleotide chains usually in terms of pairwise (dimer) sequence contributions. However, the notion that NN dimer parameters cannot be assigned from experiments by solving a set of simultaneous linear equations has been given since the development of these models in the context of polynucleotide thermodynamic studies (Gray & Tinoco, 1970). This conclusion is based on the consideration of intrinsic composition closure constraints that effectively reduce the number of degrees of freedom of the model. However, dimer set values fit easily into the theoretical NN model approximation. The extraction of simpler and more direct dimer contributions from such sets has remained an ill-posed problem with no unique solutions, but still embraced by a large community of biochemists (Breslauer et. al.,

1986, Allawi & SantaLucia, 1997, SantaLucia, 1998, SantaLucia et. al., 1986, SantaLucia & Hicks, 2004). To adopt the dimer set formulation further ad hoc regularization hypotheses have been taken by different authors, such as the singular value decomposition method (Doktycz et. al., 1992, SantaLucia, 1998).

In this paper, we present an approach given to this problem based on the analysis of how the nucleotide intrinsic intermolecular symmetries contribute to the structure of NN sets, as proposed by Licinio and Guerra (Licinio & Guerra, 2007). A molecular symmetrical decomposition is found to provide the right number of fundamental properties (irreducible parameters) which is equal to 8, for the case of DNA double strands. Such decomposition is related to the dimer set formulation, and, with that, useful and so far hidden self-consistency relations among dimers are uncovered.

Another point is that in many publications one finds datasets that include experimental values for duplex oligonucleotides, where end effects were believed to be important (Breslauer et. al., 1986, Allawi & SantaLucia, 1997, SantaLucia, 1998, SantaLucia et. al., 1986, SantaLucia & Hicks, 2004). One extension of the irreducible model to investigate how it would accommodate end effects was proposed (Guerra & Licinio, 2011). Specifically for free energy, the result was that such additional parameters would be independent on the compositions of the terminal base pairs. One simple statistical mechanics approach, when applied to the melting transition, shows that the end effects, in the light of the NN approach, should be substituted by two parameters related to the initiation of the double helix (related to two possible base pairings) (Guerra, 2013).

In this short review, all theoretical results we obtained were applied to the analysis of DNA free energy by using the approach based on the double helix initiation parameters, as proposed by Guerra (Guerra, 2013). A self-consistent set has thus been fit to free energy data from 108 short duplex oligomer sequences as available on the literature (Allawi & SantaLucia, 1997). A compact and symmetrical self-consistent set is shown to provide at least as good modelling for oligomer free-energy as standard NN dimer models. The far-reaching strength of the theoretical modelling frame for DNA or RNA sequences as proposed by Licinio and Guerra (Licinio & Guerra, 2007), resides in its compactness and symmetry. One of the immediate and practical consequences of the use of the tetrahedral model is the disclosure of the initially hidden dimer self-consistency relations.

## 2. A quantum mechanics formulation for sequence properties

The quantum mechanics expectation for any observable is given in terms of the corresponding operator  $\Theta$  and system state  $|\psi\rangle$  as  $\langle\psi|\Theta|\psi\rangle$ , in Dirac's notation. The state of a sequence composed by  $N$  nucleotides will be expressed as the tensorial product of their nucleotide states  $|b(i)\rangle$ , ( $1 \leq i \leq N$ ):

$$|\psi\rangle = |b(1)\rangle \otimes |b(2)\rangle \otimes \dots \otimes |b(N)\rangle = |b(1); b(2); \dots; b(N)\rangle \quad (1)$$

Considering only sequential NN interactions, the expectation can thus be written simply as

$$\mathbf{E} = \sum_i \langle b(i); b(i+1) | \Theta | b(i); b(i+1) \rangle \quad (2)$$

Here, submatrix elements pertaining to the same component at position  $i$  (diagonal or self-matrixes  $\Theta_{\mu(i)\nu(j)}$ ), which are internal to the sequence ( $i \neq 1, N$ ), should be halved since they are counted twice in this formulation.

A complete and symmetrical representation for the usual DNA (or RNA) four-nucleotide set can be given within a tetrahedral decomposition scheme into a three-dimensional orthonormal base set  $|x\rangle, |y\rangle, |z\rangle$ . The pure nucleotide states  $|b(i)\rangle$  are given as (see, for example, Licinio & Caligiorno, 2004):

$$|A\rangle = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}; |T\rangle = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}; |C\rangle = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}; |G\rangle = \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \quad (3)$$

Note that the four-nucleotide states are not independent, and can be expressed in terms of three independent abstract nucleotide class states. In fact,  $z$ -component discriminates weak (two bridges, AT) versus strong (three bridges, CG) hydrogen bonding for Watson-Crick pairing;  $x$ -component discriminates purines (double-ring, AG) versus pyrimidines (single-ring, CT) nucleotide sizes and  $y$ -component discriminates amino (nitrogen-containing, AC) versus keto (oxygen-containing, GT) nucleotide radicals.

Eq. 2 can be written in a more compact form as

$$\mathbf{E} = \sum_i (S + \langle V|b(i) + \langle b(i)|M|b(i+1) \rangle), \quad (4)$$

In Eq. 4,  $|b(i)\rangle$  are the sequence nucleotide states at coordinate  $i$  given in terms of class states by Eq. 3. To comprehend the physical meaning of each term in Eq. 4 (the term  $S$  and the components of  $V$  vector and  $M$  matrix), the reader is invited to read the paper published by Licinio and Guerra (Licinio & Guerra, 2007).

For measurements concerning double strands, it is well known that complementary strand symmetry further reduces the problem to the statement of only 10 conjugated NN dimer pair values (see the expressions in Eq. 5 below) linked through two independent composition closure relations, which are given by  $\sum_{b=A,T,C,G} (N_{Ab} - N_{bA}) = 0$ , and,  $\sum_{b=A,T,C,G} (N_{Cb} - N_{bC}) = 0$ . Thus, only eight independent parameters should result while the difficulties in defining a 10-dimer set of parameters from a given set of experimental data persist. In that case, complementary strand A/T and C/G pairing symmetry in a dimer, as expressed in Eq. 3, gives the following conjugate NN base component relations:  $b'_x(1) = -b_x(2)$ ,  $b'_x(2) = -b_x(1)$ , and so on. In these relations, primed bases correspond to the complementary dimer and numerals correspond to the first and second nucleotide along 5'–3' direction for each strand, i.e., both order and  $x,y$ -coordinates are inverted for the conjugate pair. The double-strand expansion can be given as a function of a single strand sequence taking into account such implicit symmetries. It is clear in that case that  $V_x = V_z = 0$ ,  $M_{xy} = M_{yx}$ ,  $M_{xz} = -M_{zx}$ , and,  $M_{yz} = -M_{zy}$ , correctly reducing the number of independent elementary tensor set values to 8. From Eq. 4, and, considering these last relations, we obtain for the 10 paired NNs a self-consistent set of expectations obeying

$$\begin{aligned} \mathbf{E}_{TA} &= S + V_z - M_{xx} - M_{yy} + M_{zz} - 2M_{xy} - 2M_{xz} - 2M_{yz} \\ \mathbf{E}_{AT} &= S + V_z - M_{xx} - M_{yy} + M_{zz} - 2M_{xy} + 2M_{xz} + 2M_{yz} \\ \mathbf{E}_{AA-TT} &= S + V_z + M_{xx} + M_{yy} + M_{zz} + 2M_{xy} \\ &\vdots \end{aligned} \quad (5)$$

and so on (Licinio & Guerra, 2007), while the symmetrical set of eight tensor parameters can be inferred from the inverse relations (Licinio & Guerra, 2007)

$$\begin{aligned} S &= \frac{1}{16} [2(\mathbf{E}_{AA-TT} + \mathbf{E}_{AG-CT} + \mathbf{E}_{GA-TC} + \mathbf{E}_{AC-GT} + \mathbf{E}_{CA-TG} + \mathbf{E}_{GG-CC}) + (\mathbf{E}_{TA} + \mathbf{E}_{AT} + \mathbf{E}_{CG} + \mathbf{E}_{GC})] \\ V_z &= \frac{1}{8} [2(\mathbf{E}_{AA-TT} - \mathbf{E}_{GG-CC}) + (\mathbf{E}_{TA} + \mathbf{E}_{AT} - \mathbf{E}_{CG} - \mathbf{E}_{GC})] \\ &\vdots \\ M_{xz} &= \frac{1}{8} (-\mathbf{E}_{TA} + \mathbf{E}_{AT} + \mathbf{E}_{CG} - \mathbf{E}_{GC}) = \frac{1}{4} (-\mathbf{E}_{AG-CT} + \mathbf{E}_{GA-TC}) \\ M_{yz} &= \frac{1}{8} (-\mathbf{E}_{TA} + \mathbf{E}_{AT} - \mathbf{E}_{CG} + \mathbf{E}_{GC}) = \frac{1}{4} (-\mathbf{E}_{AC-GT} + \mathbf{E}_{CA-TG}) \end{aligned} \quad (6)$$

This decomposition enlightens the meaning of the composition-free  $S$  term as the 16-dimer ensemble mean expectation value and of  $V_z$  as the half-differential expectation between AT containing and CG containing dimers. Most important, the double determination of  $M_{xz}$  and  $M_{yz}$  values in the last two expressions in Eq. 6 should coincide for a self-consistent set of dimer values. Explicitly, self-consistency introduces links relating composition order symmetry among dimer properties as  $\mathbf{E}_{AT} - \mathbf{E}_{TA} + \mathbf{E}_{CG} - \mathbf{E}_{GC} = 2(\mathbf{E}_{GA-TC} - \mathbf{E}_{AG-CT})$ , and,  $\mathbf{E}_{AT} - \mathbf{E}_{TA} + \mathbf{E}_{GC} - \mathbf{E}_{CG} = 2(\mathbf{E}_{CA-TG} - \mathbf{E}_{AC-GT})$ , which produces the dimer expectation self-consistency relations below, analogous to the composition closure relations presented in the discussion preceding Eq. 5:

$$\begin{aligned} \sum_{b=A,T,C,G} (\mathbf{E}_{Ab} - \mathbf{E}_{bA}) &= 0, \\ \sum_{b=A,T,C,G} (\mathbf{E}_{Cb} - \mathbf{E}_{bC}) &= 0. \end{aligned} \quad (7)$$

### 3. The Modelling Based on Double Helix Initiation Parameters

The free energy of a duplex sequence of  $N$  bases, in the NN approximation, can be calculated as a pair wise sum, including parameters related to the initiation free energy, as (Guerra, 2013):

$$\Delta G_T = \Delta G_{init} + \sum_{i=1}^{N-1} \Delta G(b_i b_{i+1}) + \Delta G_{sym} \quad , \quad (8)$$

where  $\Delta G(b_i b_{i+1})$  is the free energy contribution given by one of the 10 possible duplex dimers as listed in Eq. 5 (treating the property **E** as the free energy),  $\Delta G_{sym} = 0.43 \text{ kcal/mol}$  is a symmetric correction term applicable to self-complementary duplexes, and, the observable initiation free energy is given by:

$$\langle \Delta G_{init} \rangle = \chi_{A/T} \Delta G^\circ(A/T) + \chi_{C/G} \Delta G^\circ(C/G) \quad , \quad (9)$$

where  $\chi_{A/T}$ , and,  $\chi_{C/G}$ , are, respectively, the compositions of A/T and C/G base pairs along the sequence, and,  $\Delta G^\circ(A/T)$ , and,  $\Delta G^\circ(C/G)$  are the first base pairing free energies.

Writing Eq. 9 in this form, Guerra uses the hypothesis that the double helix nucleation can occur at any site along the strand with equal probability, independently of the local composition.

Simultaneous least mean square deviations fit of this model to the 108 sequence dataset compiled by Allawi and SantaLucia (Allawi & SantaLucia, 1997) gave the values for the free energy parameters listed in Table 1 (Guerra, 2013). As calculated by Guerra, values listed in this table produced a standard deviation equal to that obtained by using the modelling based on end effects (Guerra & Licinio, 2011), and, also, equal to that obtained using the unified set proposed by SantaLucia, namely, 0.14 kcal/mol per dimer (Guerra, 2013).

Since  $\Delta G^\circ(A/T)$  and  $\Delta G^\circ(C/G)$  are essentially similar, the initiation free energy is, in the light of the NN modelling, not dependent on the mean global composition of the DNA duplex oligomer. Observable nucleation free energies are, in turn, dependent on the mean global composition of the double strand, and, an approximate result for them is (Guerra, 2013):

$$\langle \Delta G_{nuc} \rangle = -S - V_z(\chi_{ww} - \chi_{ss}) - M_{zz}(\chi_{ww} + \chi_{ss} - \chi_{ws} - \chi_{sw}) + \langle \Delta G_{init} \rangle \quad (10)$$

Using Eq. 10, we can see that observable nucleation free energies vary within a range which goes from  $\Delta G_{nuc}^{poly A-T} = 2.54 \text{ kcal/mol}$  for a *poly A-T* homopolymer to  $\Delta G_{nuc}^{poly C-G} = 3.80 \text{ kcal/mol}$  for a *poly C-G* homopolymer. This result is relevant since the difference between the nucleation free energies, which is  $\sim 1.3 \text{ kcal/mol}$ , is greater than the bar of errors, whose length is  $\sim 0.7 \text{ kcal/mol}$  (Guerra, 2013).

The values obtained for the irreducible set can be used to get the dimer free energy contributions, using Eq. 5. Thus, using the data in Table 1, and, applying Eq. 5, we obtain for the duplex dimers AG-CT, and, GA-TC, respectively,  $\Delta G_{AG-CT} = (-1.21 \pm 0.06) \text{ kcal/mol}$ , and,  $\Delta G_{GA-TC} = (-1.33 \pm 0.06) \text{ kcal/mol}$ . Note that, as a consequence of self-consistency, values for free energy contributions by the duplex dimers AG-CT and GA-TC are essentially different (in fact, their ranges of allowable values have only an unique common intercept).

**Table 1** Irreducible Parameters for free energy at Standard Conditions (37 °C and 1 M salt and DNA)

Irreducible parameters for free energy	Values in kcal/mol
$\Delta G^\circ(A/T)$	1.7±0.03
$\Delta G^\circ(C/G)$	1.8±0.02
$S$	-1.38±0.02
$V_z$	0.58±0.04
$M_{xx}$	0.04±0.01
$M_{yy}$	-0.02±0.01
$M_{zz}$	-0.05±0.01
$M_{xy}$	-0.07±0.01
$M_{xz}$	-0.03±0.01
$M_{yz}$	-0.03±0.01

#### 4. Conclusions

Physical properties of nucleotide sequences may be calculated with an optimal set of tensor coefficients (Eq. 4) assuming projections within this tetrahedral representation. The coefficients are expressed in hierarchical differential form, so lower levels of approximation are explicitly embodied in the description. The symmetrical set is shown to provide a frame for the analysis of DNA duplex free energy fully compatible with experimental data, is shown to give a proper irreducible representation for dimer properties (Eqs. 4 and 5), and, solves an old indeterminacy of dimer sets by establishing self-consistency relations among dimer coefficients (Eqs. 6 and 7).

The modelling based on the double helix initiation parameters substitutes the end effects by the initiation parameters (Guerra & Licinio, 2011, Guerra, 2013). The values obtained for the free energies of the first base pairing were essentially similar leading to the invariance of the initiation free energy with the composition of the chain. Nevertheless, the nucleation free energy, which can be calculated from the initiation free energy, by using Eq. 10, depends on the composition, once that the difference between the maximal and minimal values which it can assume is larger than the errors bar. The modelling based on the double helix initiation parameters uses a set of ten parameters which is constituted by the eight polymeric irreducible parameters already known plus two parameters related to two possible base pairings (the initiation free energy parameters). With this set one calculates free energies for DNA oligomers at least as well as standard models considering a larger set of parameters do.

#### Acknowledgements

We thank Brazilian agency CNPq and FAPEMIG for financial support.

#### References

- Allawi, H. T., SantaLucia, J., Jr., 1997. Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochemistry* 36, 10581 – 10594.
- Breslauer, K. J., Frank, R., Blocker, H., Marky, L. A., 1986. Predicting DNA duplex stability from the base sequences. *Proc Natl Acad Sci USA* 83, 3746 – 3750.
- Doktycz, M. J., Goldstein, R. F., Paner, T. M., Gallo, F. J., Benight, A. S., 1992. Studies of DNA dumbbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T4 endloops: evaluation of the nearest-neighbor stacking interactions in DNA. *Biopolymers* 32, 849 – 864.
- Gray, D. M., 1997a. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA.RNA hybrids and DNA duplexes. *Biopolymers* 42, 795 – 810.
- Gray, D. M., 1997b. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. I. Simple sets of independent sequences and the influence of absent nearest neighbors. *Biopolymers* 42, 783 – 793.
- Gray, D. M., Tinoco, I., Jr., 1970. A new approach to study of sequence-dependent properties of nucleotides. *Biopolymers* 9, 223 – 244.
- Guerra, J. C. O., 2013. Calculation of Nucleation Free Energy for Duplex Oligomers in the Context of Nearest Neighbor Models. *Biopolymers* 99, 538 – 547.
- Guerra, J. C. O., Licinio, P., 2011. Terminal Contributions for Duplex Oligonucleotide Thermodynamic Properties in the Context of Nearest Neighbor Models. *Biopolymers* 95, 194 – 201.
- Licinio, P., Caligiore, R. B., 2004. Inference of phylogenetic distances from DNA-walk divergences. *Phys A Stat Theor Phys* 341, 471 – 481.
- Licinio, P., Guerra, J. C. O., 2007. Irreducible Representation for Nucleotide Sequence Physical Properties and Self-Consistency of Nearest-Neighbor Dimer Sets. *Biophys J* 92, 2000 – 2006.
- SantaLucia, J., Jr., 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95, 1460 – 1465.
- SantaLucia, J., Jr., Allawi, H. T., Seneviratne P. A., 1986. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry* 35, 3555 – 3562.
- SantaLucia, J., Jr., Hicks, D., 2004. The thermodynamics of DNA structural motifs *Annu Rev Biophys Biomol Struct* 33, 415 – 440.
- Vologodskii, A. V., Amirikyan, B. R., Lyubchenko, Y. L., Frank-Kamenetskii, M. D., 1984. Allowance for

heterogeneous stacking in the DNA helix-coil transition theory. *J. Biomol. Struct. Dyn* 2, 131 – 148.