Chapter 10

A Review on the Thermodynamics of Denaturation Transition of DNA Duplex Oligomers in the Context of Nearest-Neighbor Models

João C. O. Guerra

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/62574

Abstract

In this review, we show that additive physical properties of DNA double strands can be written in terms of eight (polymeric) irreducible parameters. This results in self-consistency relations constraining the 10 duplex dimer contributions. Studies of thermodynamic stability of duplex oligomers are feasible, adding extra degrees of freedom, and this is performed, initially, considering the influence of end parameters on the thermodynamic stability of oligomers. Hence, we connect a statistical mechanics approach to the nearest-neighbor (NN) approach in the framework of the two-state model. This provides one correlation between end effects and initiation phenomena. Because of that, inside the framework of the NN modeling, the role played by end effects could not be so well defined. Thus, we propose a new model that permits to provide the nucleation free energies. The power of this model is relating the nucleation free energy to the mean composition of the chain, permitting to obtain a good estimate for the free energy associated only to the Watson–Crick base pairings.

Keywords: Thermodynamics of DNA duplex oligomers, Initiation free energy, Nucleation free energy, Irreducible parameters for free energy

1. Introduction

Many DNA biotechnological applications, such as PCR or cDNA expression profiling, depend on thermodynamic parameters, which are sequence dependent. We could cite the strand melting temperature as an example of such thermodynamic parameters. In a general way, physical properties of DNA or RNA sequences can be calculated, in a very simple form, from algorithms in the context of nearest-neighbor (NN) models, whose core characteristic is

providing linear representations for experimental measurements on nucleotide chains always in terms of pairwise (dimer) sequence contributions.

However, NN dimer parameters cannot be assigned from experiments by solving a set of simultaneous linear equations. This is known since the beginning of the development of these models in the context of polynucleotide thermodynamic studies [1]. In fact, when we consider intrinsic composition closure constraints, the number of degrees of freedom of the model is effectively reduced.

Dimer occurrence relations are well known, thus allowing for decomposition of sequence properties into dimer contributions. Many authors, because of that, have preferred to use dimers as fundamental units because they provide the most straightforward decomposition scheme [2–6]. Although the dimer set values fit easily into the theoretical NN model approximation, the dimer composition is overstated. In fact, the dimer set size, which is equal to 16 (in the case of a simple chain) and 10 (in the case of double chains) [2–7], is greater than the number of degrees of freedom of the problem. However, the extraction of dimer set contributions has remained an ill-posed problem. To accomplish this task further, ad hoc regularization hypothesis has been used so far. As a corollary, so-far-unknown constraints must also link the full dimer set properties in some hidden way to restore full set unity. Alternative approaches have considered decompositions into irreducible and hence smaller sets of short sequences or dimer combinations [8–11]. Comparison between different laboratory sets and physical interpretation of set values becomes a difficult task due to the arbitrariness of possible renderings. 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 [2–6]. To adopt the dimer set formulation further, ad hoc regularization hypotheses have been taken by different authors, such as the singular value decomposition method [4, 12].

In this review, among other objectives, we present an approach 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 [13]. Therefore, to achieve that, initially, it is introduced to a general quantum mechanics statement, giving physical properties for a sequence of heterogeneous molecules treated as subsystems assuming any of a given complete set of molecular states. The four-nucleotide set has a corresponding four-state representation. At this point, a careful choice of the number of degrees of freedom is made in order to project the representation into a three-dimensional molecular class space. Luckily, the three independent molecular classes are readily associated to the main biochemical classification of nucleotides as comprising purine–pyrimidine, amino–keto, and strong–weak bases. The representation of the four-nucleotide set as a tetrahedron in the three-dimensional space is at the heart of the approach, as proposed by Licinio and Guerra [13]. This representation has been used to generate DNA walks for sequence composition analysis or display. The corresponding proper space metrics have also been recently used for phylogenetic sequence comparisons [14]. In the following, we proceed to contract the original quantum mechanics statement into an irreducible formulation using the four-nucleotide tetrahedron representation. This molecular symmetrical decomposition is found to provide the right number of fundamental properties

(free parameters), which is equal to 8, for the case of DNA double strands. We shall refer to these fundamental properties as constituting a symmetrical set of irreducible tensorial parameters. Next, we relate this decomposition to the dimer set formulation. The comparison uncovers useful and so far hidden self-consistency relations among dimers.

However, an important point still would need to be clarified. In fact, in many publications, one finds datasets that include experimental values for duplex oligonucleotides, where end effects were believed to be important [2–6]. Nevertheless, such initiation and termination parameters would seem to be very sensitive to the modeling and have changed a lot even inside the same research group [3–6]. In fact, Xia et al. had already argued that data from melting experiments of RNA duplexes are of insufficient accuracy to distinguish end effects [15]. With this motivation, as a second step in the development of the approach proposed by us and presented in this review, we proposed to extend the irreducible model to investigate how it would accommodate end effects. Guerra and Licinio in fact performed such extension and calculated the irreducible parameters for free energy, entropy, enthalpy, and the respective end contributions [16]. Later, a detailed algorithm for performing such calculations is described. However, at this point, it is necessary to anticipate some conclusions. For example, Guerra and Licinio obtained values for the end effects with relatively large errors. In addition, specifically for free energy, they could not distinguish between the weak and strong terminal base pairs. In the light of their finding, one simple statistical mechanics approach, when applied to the melting transition, shows that the approach based on end effects, according to the NN approach, proves to be naive, even heuristic. In fact, since the end effects were initially (wrongly) identified as the nucleation free energies, they should be dependent on the mean global composition of the chain. However, an only slightly more detailed statistical mechanics approach can show that, summed to the eight (polymeric) irreducible parameters for free energy, as already mentioned, there are other two parameters related to the initiation of the double helix (related to two possible base pairings). That is, in the light of the NN approach, there are 10 parameters, which expand the free energy of any DNA oligomers [17].

Before we continue our discussion throughout the forthcoming sessions, it is important to inform the reader that all theoretical results we obtained were applied to the analysis of DNA free energy by introducing, initially, the formulation of end contributions to the model, which will be presented later in this chapter. A simple statistical mechanics approach is then applied to the problem. As a result, a second set of parameters, including this time the initiation parameters, will be obtained. Anyway, a self-consistent set has thus been fit to free energy data from 108 short duplex oligomer sequences as available in the literature. We will show that, using both the modeling, the first based on end effects and the second based on the use of double helix initiation parameters, the more compact and symmetrical self-consistent set is shown to provide at least as good modeling for oligomer free energy as standard NN dimer models. The far-reaching strength of the theoretical modeling frame for DNA or RNA sequences as proposed by us resides in its compactness and symmetry. As will be discussed later in this review, 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

Complexity in biological phenomena represents an enormous challenge and a rich field for the application and development of physical methods. To unfold simple biopolymer phenomena, we start by a biochemical meaningful nucleotide representation into molecular classes and count on tools provided by the quantum mechanics. Here, we shall use the quantum mechanics formulation based on the matrix representation. What is needed from start is some base set for the description of the states of the system, which, for us, is a DNA or RNA sequence. The ensemble of sequence states is given by allowable sequence composition alone. We want to describe and isolate gross composition states. Inner electronic states or molecular conformation contributions, which would require a much finer level of quantum description, are so far intrinsically averaged. State transitions are of course forbidden if one neglects mutations. The sequence state will be given in terms of its molecular constitution, and a nucleotide set representation will condition the sequence representation.

The quantum mechanics expectation for any observable is given in terms of the corresponding operator $\boldsymbol{\Theta}$ and system state $|\Psi\rangle$ as $\langle\psi|\boldsymbol{\Theta}|\psi\rangle$, in Dirac's notation. The state of a system comprising N particles or molecules is usually expressed as the tensorial product of their component states $|b(i)\rangle$, $(1 \leq i \leq N)$:

$$\left|\Psi\right\rangle = \left|b(1)\right\rangle\left|b(2)\right\rangle \otimes \cdots \otimes \left|b(N)\right\rangle = \left|b(1);b(2);...;\left|b(N)\right\rangle\right. \tag{1}$$

For d-dimensional component states, this would lead a priori to the specification of $(Nd)^2$ operator matrix elements $\mu(i)\nu(j)$. If interaction range is limited, however, then many off-diagonal matrix elements become null, and a reduced formulation can be sought. Considering only sequential NN interactions, the expectation can thus be written simply as

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

Here, submatrix elements pertaining to the same component at position i (diagonal or self-matrices $\boldsymbol{\Theta}_{\mu(i)\nu(j)}$), which are internal to the sequence ($i \neq 1, N$), should be halved because they are counted twice in this formulation (see Fig. 1). We hope further reduction of this development can be obtained considering implicit symmetries of the Hermitian $\boldsymbol{\Theta}$ matrix and its invariants under orthonormal base representations.

3. Nucleotide class-state representation

The most straightforward representation for a four-nucleotide set is, obviously, a four-dimensional vector. This "independent-nucleotide" representation has been implicitly

http://dx.doi.org/10.5772/62574

adopted by many authors and leads to 4 × 4 matrices or 16 parameter sets when considering nucleotide pairwise properties [11]. This representation, however, already overstates the nucleotide composition problem from the beginning. The set representation should be more concisely established in a three-dimensional space. Thus, 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 follows [14]:

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

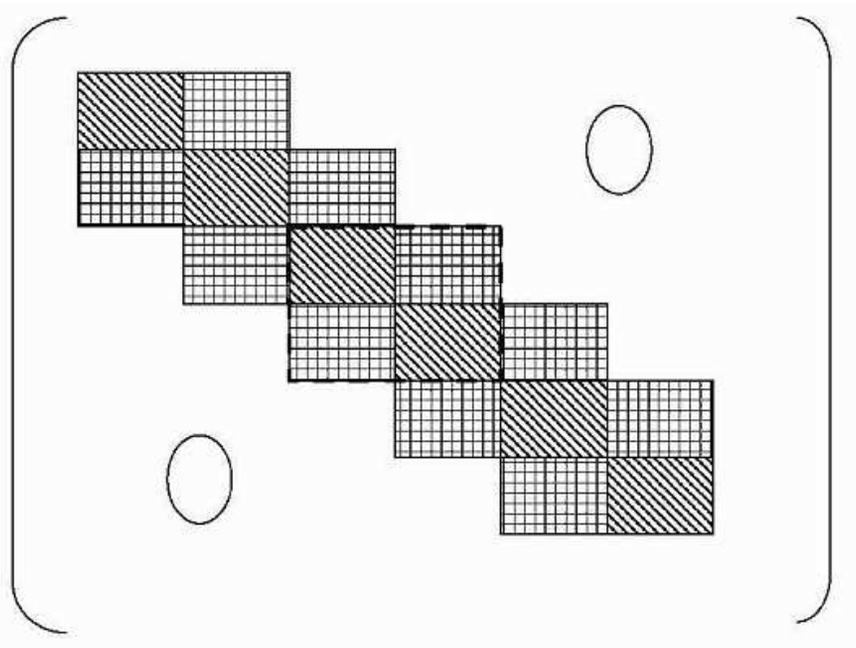

Figure 1. Structure of an expectation matrix for a sequence of $n = 6$ identical components (molecules in arbitrary states). The components have d degrees of freedom represented through d orthogonal base states, which result in $3n–2 = 16$ submatrices of size d^2. In this case, only nearest-neighbor interactions are considered. The matrix above corresponding to the quantum mechanics formulation of Eq. 1 is Hermitian and periodic, allowing for a more synthetic representation. One periodic module of four submatrices implicit in Eq. 2 has been distinguished by a dashed line. Observe that internal submatrices in the diagonal are counted twice according to the formulation of Eq. 2 [13].

The nucleotides themselves are represented as a nonorthogonal (tetrahedral) $\sqrt{3}$ -modulus vector set (Fig. 2). The four-nucleotide states are not independent and can be expressed in terms of three independent abstract nucleotide class states. Due to this decomposition, z-component discriminates weak (two bridges, AT) versus strong (three bridges, CG) hydrogen bonding for Watson–Crick (WC) pairing; x-component discriminates purine (double ring, AG) versus pyrimidine (single ring, CT) nucleotide sizes; and y-component discriminates amino (nitrogen containing, AC) versus keto (oxygen containing, GT) nucleotide radicals.

In quantum mechanics language, a $|x\rangle$ base state, for example, is a ring number or purine–pyrimidine class state, whereas $|A\rangle = |x\rangle + |y\rangle + |z\rangle$ is an adenine molecular state decomposed in terms of proper nucleotide class subspaces. Any pure nucleotide state can thus be represented in terms of molecular class states.

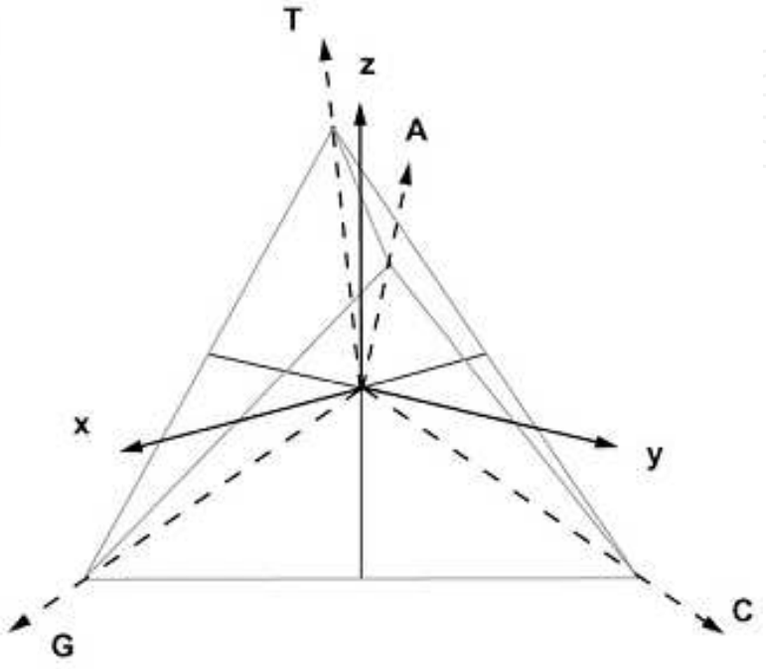

Figure 2. Orthonormal x–y–z base set and tetrahedral DNA-nucleotide set representation. Each of the three axes distinguishes a specific molecular class feature. Purines are distinguished from pyrimidines through x-coordinate. Amino is distinguished from keto through y-coordinate. And, finally, weak WC hydrogen-bridge binding is distinguished from stronger binding through z-coordinate [14].

Each possible nucleotide pair shares one of its fundamental molecular structural characteristics as a group in a given class, which differs from the complementary pair as another group in the same class. This is latent when we observe Eq. 3, which translates perfectly well the intrinsic cubic symmetry of the tetrahedron. From now, we proceed to construct our approach, which will use a complete nucleotide representation, and, then, having seen based on this representation, it will provide properties associated to each molecule decomposing them in terms of three differential affinity groups or classes. Therefore, the choice of a tetrahedral set is thus natural and convenient for its intrinsic orthogonality and symmetry properties, which are related to common molecular group classifications. Nevertheless, its main advantage is to fulfill the necessity for a three-dimensional bijective representation of a four-set composition.

4. Irreducible representation

Returning to the quantum mechanics formulation, our intention is to exploit remaining invariants and redundancies from the structure of the matrix operator present in Eq. 2 in order

http://dx.doi.org/10.5772/62574

to further reduce its number of parameters. The three-dimensional nucleotide basis should be kept in mind. The sequence-dependent states of an observable will then assume discrete values given by a most compact expansion of its expectation as follows:

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

in substitution to Eq. 2; in Eq. 4, $|\, b(i)\rangle$ are still the sequence nucleotide states at coordinate i, which are given in terms of class states by Eq. 3.

The bracket notation indicates vector and dyadic contractions as usual. The expansion in Eq. 4 is quite intuitive, in the sense that the first two terms represent linear contributions to a property from the sequence composition, whereas the third term comprises nonlinear effects due to NN interference or differential stacking interactions. Comparison with Eq. 2 allows the identification of its components. The first term is a constant or mean contribution to the observable, given as the invariant trace of the square expectation periodic matrix $S = Tr(\boldsymbol{\Theta})$. The trace represents a molecular state independent contraction of the self-matrix diagonal, where, by construction, any pure nucleotide component (Eq. 3) equally squares to one ($b_\mu^2 = 1$). The remaining cross terms of the self-matrix similarly contract to a vector because all pure nucleotide states $|\, b(i)\rangle$ also have cyclically multiplicative class components ($b_x = b_y b_z$, etc.). This contraction gives the second term as an order-independent or global-composition contribution, with components $\langle V \,| = 4\ Re\ (\boldsymbol{\Theta}_{y(1)z(1)},\ \boldsymbol{\Theta}_{x(1)z(1)},\ \boldsymbol{\Theta}_{x(1)y(1)})$. The third term is an NN or first-order sequence stacking contribution to the observable. The stacking matrix M is a second-rank tensor and has its elements given from the cross expectation matrix as $M_{\mu\nu} = 2\ Re(\boldsymbol{\Theta}_{\mu(1)\nu(2)})$. The symmetrical sum of the expectation matrix Hermitian conjugates results in a fully contracted real formulation.

Decomposition of nucleotide sequence observable expectation as given in Eq. 4 naturally leads to an irreducible 13-parameter description of physical properties (S, V_μ, and $M_{\mu\nu}$), which we call the symmetrical set, within the NN approximation. Note that a traditional description of stacking-dependent properties is often stated in terms of the NN dimer composition, that is, as a linear combination of the 16-ordered 5′-3′ NN dimer set E_{ij} :

$$\mathbf{E} = \sum_{i,j=A,\,T,\,C,\,G} N_{ij} \mathbf{E}_{ij} \tag{5}$$

However, the NN dimer set is overspecified, that is, only a smaller set of NN combinations can be a priori obtained from inversions of Eq. 5 because Eq. 5 is supplemented by independent composition closure relations. For implicit circular sequences (or for very long sequences, i.e., polynucleotides), these can be taken as any three of the following:

$$\begin{gathered}\sum_{b=A,\,T,\,C,\,G}\left(N_{Ab}-N_{bA}\right)=0\\ \sum_{b=A,\,T,\,C,\,G}\left(N_{Tb}-N_{bT}\right)=0\\ \sum_{b=A,\,T,\,C,\,G}\left(N_{Cb}-N_{bC}\right)=0\\ \sum_{b=A,T,C,G}\left(N_{Gb}-N_{bG}\right)=0,\end{gathered} \tag{6}$$

reducing the number of independent dimers in the set to arbitrary 13. Similar arguments hold for linear oligomers. In comparison, the decomposition of physical properties in the symmetrical set proposed here is in a fundamental level; since from the beginning, it includes only a priori linearly independent terms and gives contributions to the observable in the hierarchic form of three expectation tensors of increasing rank, corresponding to different levels of analysis. The 16-NN expectations can otherwise be easily obtained as a linear combination of the 13 symmetrical-set tensor components. In that case, it is useful to rewrite Eq. 4 in a form appropriate for NN dimer decomposition as follows:

$$\mathbf{E}_{b(1)b(2)} = S+\left\langle V\middle|\frac{b(1)+b(2)}{2}\right\rangle+\left\langle b(1)\middle|M\middle|b(2)\right\rangle, \tag{7}$$

where, to correctly account for additivity, as given by Eq. 5 for each dimer in a sequence, the two nucleotide linear contributions are halved. Explicitly, one has applying Eq. 3 to Eq. 7:

$$\begin{aligned}\mathbf{E}_{TA} &= S+V_z-M_{xx}-M_{xy}-M_{xz}-M_{yx}-M_{yy}-M_{yz}+M_{zx}+M_{zy}+M_{zz}\\ \mathbf{E}_{AT} &= S+V_z-M_{xx}-M_{xy}+M_{xz}-M_{yx}-M_{yy}+M_{yz}-M_{zx}-M_{zy}+M_{zz}\\ \mathbf{E}_{CA} &= S+V_y-M_{xx}-M_{xy}-M_{xz}+M_{yx}+M_{yy}+M_{yz}-M_{zx}-M_{zy}-M_{zz}\\ \mathbf{E}_{TG} &= S-V_y-M_{xx}+M_{xy}+M_{xz}-M_{yx}+M_{yy}+M_{yz}+M_{zx}-M_{zy}-M_{zz}\end{aligned} \tag{8}$$

and so on. Tensor elements can be either conversely determined from reported dimer values or self-consistently derived from fits to raw polynucleotide data using Eqs. 8 and 5, or directly from Eq. 4, while from a theoretical point of view, molecular symmetry arguments or ab initio calculations could be used to guess tensor structure and values.

4.1. Double strands

For measurements concerning double strands, aside end effects, 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. 12) linked through two independent composition closure relations as follows:

http://dx.doi.org/10.5772/62574

$$\begin{aligned}\sum_{b=A,\,T,\,C,\,G}\left(N_{Ab}-N_{bA}\right)=0,\\ \sum_{b=A,\,T,\,C,\,G}\left(N_{Cb}-N_{bC}\right)=0,\end{aligned} \tag{9}$$

so that 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 conjugate NN base component relations as follows:

$$\begin{aligned}b'_x(1)=-b_x(2);\; b'_x(2)=-b_x(1),\\ b'_y(1)=-b_y(2);\; b'_y(2)=-b_y(1),\\ b'_z(1)=b_z(2);\; b'_z(2)=b_z(1),\end{aligned} \tag{10}$$

where primed bases correspond to the complementary dimer and numerals correspond to the first and second nucleotides along 5′-3′ direction for each strand, that is, 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 the aforementioned implicit symmetries (by adding contributions from both strands to Eq. 7 taking into account Eq. 10 and then redefining the tensor set, that is, $\mathbf{E}'_{b1b2}=\mathbf{E}_{b1b2}+\mathbf{E}_{b1'b2'}$). It is clear in that case that

$$V_x=V_z=0,\; M_{xy}=M_{yx},\; M_{xz}=-M_{zx},\; M_{yz}=-M_{zy} \tag{11}$$

correctly reducing the number of independent elementary tensor set values to 8. From Eqs. 7 and 11, the decomposition for the 10 paired NNs gives a self-consistent set of expectations obeying

$$\begin{aligned}\mathbf{E}_{TA}&=S+V_z-M_{xx}-M_{yy}+M_{zz}-2M_{xy}-2M_{xz}-2M_{yz}\\ \mathbf{E}_{AT}&=S+V_z-M_{xx}-M_{yy}+M_{zz}-2M_{xy}+2M_{xz}+2M_{yz}\\ \mathbf{E}_{AA-TT}&=S+V_z+M_{xx}+M_{yy}+M_{zz}+2M_{xy}\\ \mathbf{E}_{AG-CT}&=S+M_{xx}-M_{yy}-M_{zz}-2M_{xz}\\ \mathbf{E}_{GA-TC}&=S+M_{xx}-M_{yy}-M_{zz}+2M_{xz}\\ \mathbf{E}_{AC-GT}&=S-M_{xx}+M_{yy}-M_{zz}-2M_{yz}\\ \mathbf{E}_{CA-TG}&=S-M_{xx}+M_{yy}-M_{zz}+2M_{yz}\\ \mathbf{E}_{GG-CC}&=S-V_z+M_{xx}+M_{yy}+M_{zz}-2M_{xy}\\ \mathbf{E}_{CG}&=S-V_z-M_{xx}-M_{yy}+M_{zz}+2M_{xy}+2M_{xz}-2M_{yz}\\ \mathbf{E}_{GC}&=S-V_z-M_{xx}-M_{yy}+M_{zz}+2M_{xy}-2M_{xz}+2M_{yz}\end{aligned} \tag{12}$$

while the symmetrical set of eight tensor parameters can be inferred from the inverse relations

$$
\begin{aligned}
S &= \frac{1}{16}\left[2\left(\mathbf{E}_{AA-TT}+\mathbf{E}_{AG-CT}+\mathbf{E}_{GA-TC}+\mathbf{E}_{AC-GT}+\mathbf{E}_{CA-TG}+\mathbf{E}_{GG-CC}\right)+\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}+\mathbf{E}_{CG}+\mathbf{E}_{GC}\right)\right] \\
V_Z &= \frac{1}{8}\left[2\left(\mathbf{E}_{AA-TT}-\mathbf{E}_{GG-CC}\right)+\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}-\mathbf{E}_{CG}-\mathbf{E}_{GC}\right)\right] \\
M_{xx} &= \frac{1}{16}\left[2\left(\mathbf{E}_{AA-TT}+\mathbf{E}_{AG-CT}+\mathbf{E}_{GA-TC}-\mathbf{E}_{AC-GT}-\mathbf{E}_{CA-TG}+\mathbf{E}_{GG-CC}\right)-\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}+\mathbf{E}_{CG}+\mathbf{E}_{GC}\right)\right] \\
M_{yy} &= \frac{1}{16}\left[2\left(\mathbf{E}_{AA-TT}-\mathbf{E}_{AG-CT}-\mathbf{E}_{GA-TC}+\mathbf{E}_{AC-GT}+\mathbf{E}_{CA-TG}+\mathbf{E}_{GG-CC}\right)-\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}+\mathbf{E}_{CG}+\mathbf{E}_{GC}\right)\right] \\
M_{zz} &= \frac{1}{16}\left[2\left(\mathbf{E}_{AA-TT}-\mathbf{E}_{AG-CT}-\mathbf{E}_{GA-TC}-\mathbf{E}_{AC-GT}-\mathbf{E}_{CA-TG}+\mathbf{E}_{GG-CC}\right)+\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}+\mathbf{E}_{CG}+\mathbf{E}_{GC}\right)\right] \\
M_{xy} &= \frac{1}{16}\left[2\left(\mathbf{E}_{AA-TT}-\mathbf{E}_{GG-CC}\right)-\left(\mathbf{E}_{TA}+\mathbf{E}_{AT}-\mathbf{E}_{CG}-\mathbf{E}_{GC}\right)\right] \\
M_{xz} &= \frac{1}{16x-8}\left(-\mathbf{E}_{TA}+\mathbf{E}_{AT}+\mathbf{E}_{CG}-\mathbf{E}_{GC}\right)=\frac{1}{4}\left(-\mathbf{E}_{AG-CT}+\mathbf{E}_{GA-TC}\right) \\
M_{yz} &= \frac{1}{16x-8}\left(-\mathbf{E}_{TA}+\mathbf{E}_{AT}-\mathbf{E}_{CG}+\mathbf{E}_{GC}\right)=\frac{1}{4}\left(-\mathbf{E}_{AC-GT}+\mathbf{E}_{CA-TG}\right)
\end{aligned}
\tag{13}
$$

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 importantly, the double determination of M_{xz} and M_{yz} values in the last two expressions in Eq. 13 should coincide for a self-consistent set of dimer values. Explicitly, self-consistency introduces links relating to composition order symmetry among dimer properties as follows:

$$
\begin{aligned}
\mathbf{E}_{AT}-\mathbf{E}_{TA}+\mathbf{E}_{CG}-\mathbf{E}_{GC} &= 2\left(\mathbf{E}_{GA-TC}-\mathbf{E}_{AG-CT}\right) \\
\mathbf{E}_{AT}-\mathbf{E}_{TA}+\mathbf{E}_{GC}-\mathbf{E}_{CG} &= 2\left(\mathbf{E}_{CA-TG}-\mathbf{E}_{AC-GT}\right).
\end{aligned}
\tag{14}
$$

Note that, analogous to the composition closure relations (Eq. 9), the dimer expectation self-consistency relations (Eq. 14) may also be combined to read as follows:

$$
\begin{aligned}
\sum_{b=A,\,T,\,C,\,G}\left(\mathbf{E}_{Ab}-\mathbf{E}_{bA}\right) &= 0, \\
\sum_{b=A,\,T,\,C,\,G}\left(\mathbf{E}_{Cb}-\mathbf{E}_{bC}\right) &= 0.
\end{aligned}
\tag{15}
$$

5. The modeling based on end effects

From now, we proceed to extend the irreducible model to investigate how it accommodates end effects. For the case of circular DNA, or even, for a DNA polymer, knowing the eight (polymeric) irreducible parameters (S, V_z, and the six elements of the M matrix) is sufficient

http://dx.doi.org/10.5772/62574

for the prediction of additive physical properties. For an oligomer, additional end effects would become important and would need to be accounted for. Thus, to correctly account such effects for, consider the following duplex sequence as follows:

$$\begin{array}{l} E\ b_1 b_2 b_3 \cdots b_N E \\ E b'_1 b'_2 b'_3 \cdots b'_N E, \end{array} \tag{16}$$

where, according to the notation introduced by Gray [10, 11], E is a pseudo-base indicating the terminations of the sequence. Pseudo-base E simply would represent one of the NNs to the end base pairs, and, under this viewpoint, it indicates interactions between the end base pairs and the surrounding solvent. Following the reasoning line suggested by Licınio and Guerra [13] and introduced in Section 3 of this review, $|A\rangle$, $|T\rangle$, $|C\rangle$, and $|G\rangle$, in Eq. 3, would correspond to the 3D part of 4D vectors with the fourth component equals to zero, and $|E\rangle$ would be a new molecular state, linearly independent with $|A\rangle$, $|T\rangle$, $|C\rangle$, and $|G\rangle$, and written as follows:

$$|E\rangle = \begin{pmatrix} \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{17}$$

Then, applying Eq. 7, and, considering Eq. 11, for the duplex dimer $Eb_1 - b'_1 E$, we obtain the contribution of the end base pair b_1 / b'_1 for the thermodynamical stability of the sequence, and analog reasoning can be applied for the end base pair b_N / b'_N. Thus, for the pseudo-duplex dimer $Eb_1 - b'_1 E$,

$$\mathbf{E}\left(Eb_1\right) = A + Bx_1 + Cy_1 + Dz_1, \tag{18}$$

where A, B, C, and, D are parameters that determine the property under consideration. And for the pseudo-duplex dimer $b_N E - Eb'_N$, :

$$\mathbf{E}\left(b_N E\right) = A + Bx'_N + Cy'_N + Dz'_N, \tag{19}$$

where, in Eqs. 18 and 19, x_k is the x-component of the vector $|b_k$, and so on. According to Eqs. 18 and 19, the orientation of the end base pair would be important; for example, one A/T end

base pair would not produce the same effect as one T/A end base pair. Therefore, at least in theory, it would be necessary to discriminate four-end pairings, which are listed in the following:

$$\begin{matrix} EA, & ET, & EC, & EG \\ ET, & EA, & EG, & EC. \end{matrix} \tag{20}$$

Finally, we can conclude that the four possible end base pairs in Eq. 20 can be expanded in terms of four parameters, namely *A*, *B*, *C*, and *D*. Consequently, for a duplex oligomer, the additional four parameters related to the ends should be added to the eight polymeric parameters already known, producing a total of 12 irreducible parameters, in the light of the modeling based on the end effects.

6. Results and discussion for the modeling based on the end effects

From now on, the thermodynamical property **E** will be, for us, the free energy of the duplex formation. According to the model based on the end effects, the free energy of a duplex sequence of *N* bases in the NN approximation could be calculated as the pairwise sum including end effects as a function of 12 irreducible parameters from Eqs. 12, 18, and, 19, as follows:

$$\Delta G_T = \Delta G\left(Eb_1\right) + \sum_{i=1}^{N-1} \Delta G\left(b_i b_{i+1}\right) + \Delta G\left(b_N E\right) + \Delta G_{\text{sym}} \tag{21}$$

where $\Delta G_{\text{sym}} = 0.43$ kcal/mol is a symmetric correction term applicable to self-complementary duplexes.

Simultaneous least-mean-square-deviation fit of this model to the 108 sequence data compiled by Allawi and SantaLucia [12] gave the values for the free energies, which are listed in Table 1. Guerra and Licinio [16] calculated irreducible parameters for the thermodynamic properties of free energy, entropy, and enthalpy but, in Table 1, only the irreducible parameters for free energy are shown. In Table 1, $\Delta G(ET)$ is the contribution for the free energy of the sequence from the T/A end base pair, and so on. Here, we prefer to use the irreducible parameters $\Delta G(EA)$, $\Delta G(ET)$, $\Delta G(EC)$, and $\Delta G(EG)$ in the place of the parameters *A*, *B*, *C*, and *D* as defined in Eq. 18 or 19. In fact, there is no loss of generality in the use of the firsts once that they are linear combinations of the seconds.

For comparison, we performed another calculation, supposing that the contributions from the ends do not depend on the orientation of the end base pairs, that is, an A/T end pair would

http://dx.doi.org/10.5772/62574

contribute in the same way as a T/A end pair, as it is usually found in the literature [2–6]. As a result, we obtained Table 2.

Irreducible parameters for free energy	Values (kcal/mol)
$\Delta G(ET)$	0.94± 0.07
$\Delta G(EA)$	0.87± 0.07
$\Delta G(EG)$	0.82± 0.07
$\Delta G(EC)$	0.87± 0.06
S	−1.37± 0.02
V_z	0.57± 0.01
M_{xx}	0.04± 0.01
M_{yy}	−0.01± 0.01
M_{zz}	−0.05± 0.01
M_{xy}	−0.07± 0.01
M_{xz}	−0.02± 0.01
M_{yz}	−0.02± 0.01

Table 1. Irreducible Parameters for Free Energy at Standard Conditions (37 ℃ and 1 M Salt and DNA)

Irreducible parameters for free energy	Values (kcal/mol)
$\Delta G(EA,\ ET)$	0.91± 0.07
$\Delta G(EC,\ EG)$	0.84± 0.06
S	−1.37± 0.02
V_z	0.57± 0.01
M_{xx}	0.04± 0.01
M_{yy}	−0.01± 0.01
M_{zz}	−0.04± 0.01
M_{xy}	−0.07± 0.01
M_{xz}	−0.02± 0.01
M_{yz}	−0.03± 0.01

Table 2. Irreducible Parameters for Free Energy at Standard Conditions (37 ℃ and 1 M Salt and DNA)

Considering the values obtained for the irreducible parameters for free energy presented in Tables 1 and 2, some observations must be carried out:

1. Mean values and errors are essentially the same, independently of the modeling.
2. Defining the root-mean-square deviation per dimer χ [13, 16] as follows:

$$\chi = \sqrt{\sum_{i=1}^{108}\left[\Delta G_{\text{exp}}(i) - \Delta G_{\text{theor.}}(i)\right]^2 / \sum_{i=1}^{108} N(i)}, \tag{22}$$

the free energy irreducible parameters in Tables 1 and 2 are such that they minimize χ. The quantityχ defines a global minimal deviation, between the theoretical values calculated from the irreducible parameter set for the free energies of the 108 sequences and the experimental values. In Eq. 22, $\Delta G_{\text{exp}}(i)$ is the experimental value for the free energy for the ith sequence, $\Delta G_{\text{theor.}}(i)$ is its corresponding theoretical value, and $\sum_{i=1}^{108} N(i)$ is the total number of duplex dimers for the ensemble of 108 sequences. The value obtained forχ considering 10 (or 12) parameters is precisely the same, namely 0.14 kcal/mol per dimer [16], which also coincides with the 12-parameter model using values reported by SantaLucia for the free energies for the 10 duplex dimers [3–6]. This means that, considering only the overall data ensemble quality, there is no practical reason to prefer a model with a greater number of parameters.

3. The intrinsic errors obtained for the contributions by the ends are sensibly larger than the errors for the other irreducible parameters. In this way, in all the decomposition schemes, the contributions of the ends are not so well defined, that is, we could not differentiate its orientation (for example, we could not differentiate A/T from T/A). Thus, or the available experimental data are not still sufficiently precise or even this modeling is still inadequate to account for end effects.
4. It is also verified that the C/G or G/C end pairing is only slightly more stable than the A/T or T/A end pairing. However, the intrinsic errors in data shown in Tables 1 and 2 are considerable, allowing for portions of the ranges of possible values of the end parameters to coincide. Thus, strictly speaking, in the modeling based on end effects, there is no differentiation between the terminal base pairs.
5. The errors of the irreducible parameters for free energy were estimated in the following way: Guerra and Licinio selected 100 sets of 80 sequences chosen randomly and then calculated the mean deviation for the parameters obtained from each set [16].

As shown, end contributions are fit with large errors to experimental data, as compared to the fits of other NN or dimer contributions. Besides A/T from T/A as well as C/G from G/C, ending contributions could not be respectively differentiated. More than that, we could not distinguish between the weak and the strong terminal base pairs. However, using both the sets, one can calculate free energies for DNA oligomers at least as well as standard models considering a larger set of parameters do [3–6]. Guerra and Licinio [16] also extended their analysis and obtained equivalent sets of irreducible parameters for enthalpy and entropy. By simultaneously minimizing the deviations from melting temperatures and entropies of the chains, they obtained the most precise set, which is capable of predicting melting temperatures for DNA

chains with a standard deviation of 2.2°C for sequence against a deviation of 2.5°C for previous parameters found in the literature [3–6].

In the light of our finding, the formulation based on the use of end effects, according to the NN approach, proves to be naive, even heuristic. The extra parameters (up to now, the end parameters), which must be summed to the eight (polymeric) irreducible parameters for predicting thermodynamical properties of duplex oligomers, seem not to depend on the composition of the terminal base pairs. From now, we will invoke a new hypothesis, which will be detailed later in this review. With base on this hypothesis, we will conclude that, in the light of the NN model, 10 is the number of parameters expand the free energy of any DNA oligomers: eight (polymeric) irreducible parameters for free energy, already described, plus two parameters related to the initiation of the double helix (related to two possible base pairings).

7. The modeling based on double helix initiation parameters

Equation 21 establishes how to calculate the total free energy of a sequence of length N, according to the NN model, using the methodology based on the modeling by end effects.

On the other hand, in the statistical mechanics viewpoint, the free energy of the duplex formation ΔG_T relates to the equilibrium constant K_{eq} as follows:

$$\Delta G_T = -RT \ln K_{\text{eq}}. \tag{23}$$

Whenever nucleation is the limiting process, the two-state model establishes that once the process is initiated, the helix extends to both ends of the chain [7]. The partition function or the equilibrium constant K_{eq} for the duplex formation can then be calculated as follows:

$$K_{\text{eq}} = \sigma \prod_{i=1}^{N} s_i, \tag{24}$$

where σ is the nucleation equilibrium constant and s_i is the propagation equilibrium constant, which refers to the addition of the *i*th base pair to the preexisting duplex. For heteropolymers, σ and s_i depend on the composition of the chain. Inserting Eq. 24 into Eq. 23, we obtain:

$$\Delta G_T = -RT \ln \sigma - RT \sum_{i=1}^{N} \ln s_{i,} \tag{25}$$

that is,

$$\Delta G_T = \Delta G_{\text{nuc}} - RT\sum_{i=1}^{N}\ln s_i. \tag{26}$$

Equation 26 can be conveniently rewritten as follows:

$$\Delta G_T = \Delta G_{\text{nuc}} - RT\ln s_k - RT\sum_{\substack{i=1\\ i\neq k}}^{N}\ln s_i. \tag{27}$$

Eqs. 26 and 27 have the same signification, but when writing Eq. 27 in the form shown, we suppose that the formation of the first base pair of the duplex occurs in the *k*th site. Therefore, we can see that, by comparing Eq. 27 with Eq. 21, the nucleation free energy corresponds to the end effects in the NN approach, except by the term $-RT\ln s_k$, that is,

$$\Delta G_{\text{nuc}} - RT\ln s_k = -RT\ln\sigma s_k = \Delta G\left(Eb_1\right) + \Delta G\left(b_N E\right). \tag{28}$$

Quantity $\Delta G_{\text{nuc}} - RT\ln s_k$, as shown in Eq. 28, in another way corresponds the initiation free energy, ΔG_{init}, and, correspondently, σs_k is the initiation equilibrium constant associated to the formation of the first base pair of the duplex. Furthermore, to the light of the NN modeling, the initiation free energy plays the role of the end effects. Finally, the sum of the propagation free energies corresponds, also to the light of the NN model, to the sum of the dimer free energies with the following equation:

$$-RT\sum_{\substack{i=1\\ i\neq k}}^{N}\ln s_i = \sum_{i=1}^{N-1}\Delta G\left(b_i b_{i+1}\right). \tag{29}$$

Recently, Guerra and Licinio connected to the two approaches, namely the NN and the statistical mechanics approaches, and they calculated the equilibrium constants and free energies for nucleation and propagation of a double helix in the following transition reactions [16]:

$$\begin{aligned}\text{poly } A + \text{poly } T &\rightleftharpoons \text{poly } A\cdot T\\ \text{poly } C + \text{poly } G &\rightleftharpoons \text{poly } C\cdot G.\end{aligned} \tag{30}$$

For the above homopolymers, they obtained the following nucleation free energies, at standard 1 mol concentration:

$$\begin{aligned}\Delta G_{\text{nuc}}\left(\text{poly } A\cdot T\right) &= 1.81 \text{ kcal / mol}\\ \Delta G_{\text{nuc}}\left(\text{poly } C\cdot G\right) &= 1.69 \text{kcal / mol.}\end{aligned} \tag{31}$$

http://dx.doi.org/10.5772/62574

These values were obtained using values obtained for end effects calculated from the simultaneous least-mean-square-deviations fit of the NN model to the 108-sequence data compiled by Allawi and SantaLucia [2] and listed in Tables 1 and 2, and values experimentally obtained for A/T and C/G base pairings compiled by the Frank-Kamenetskii Group [18]. Once they obtained intrinsically large errors for the end effects, the nucleation free energies for poly $A\Delta T$ and poly $C\Delta G$ homopolymers could be considered essentially similar. This result seemed strange because nucleation free energies would depend on the oligomer composition as a whole. This could indicate that end effects, as usually accounted in the NN models, could have an improper representation, having as consequence, poor fitting parameters, and an incoherent interpretation of the nucleation. Thus, the usual modeling by end effects must be seen as a didactic and heuristic approximation for DNA properties, but a better modeling needs to be discussed.

As a more appropriate modeling is a necessity, we will look for a more precise interpretation for the nucleation free energy term in the expansion of the free energy of a duplex oligomer. For this, initially, we will write the free energy for the formation of a duplex oligomer as found in some approaches in the literature [4, 6, 19]:

$$\Delta G_T = \Delta G_{\text{init}} + \sum_{i=1}^{N-1} \Delta G\left(b_i b_{i+1}\right) + \Delta G_{\text{sym}}, \tag{32}$$

where, according such references, ΔG_{init} is the "initiation" or "nucleation" free energy. Such quantity, in accordance with these referred references, is related to the difficulty of aligning the two strands and forming the first WC base pair "nucleating" the double helix which, after this step, will propagate to the ends of the chain. Specially in the work of Manyanga et al. [19], ΔG_{init} is indiscriminately called the initiation or nucleation free energy. However, the term ΔG_{init} in Eq. 32 is the initiation free energy, as can be verified by returning to the discussion that follows Eq. 28. In fact, Eq. 28 shows that nucleation free energy ΔG_{nuc} is obtained from the initiation free energy ΔG_{init} by adding a term related to the "propagation" of the WC first base pair, $-RT\ln s_k$. Therefore, it becomes clear, from now, that the terms of initiation and nucleation free energies are effectively different. It is also clear that Eq. 32 has significance if and only if ΔG_{init} is the initiation free energy. Thus, we can establish the problem: How does the term ΔG_{nuc} depend on the sequence composition? Answering to this question will help us to understand why the modeling by end effects that have been used is theoretically incorrect.

The question posed in the last paragraph will guide us throughout this section. To answer it, consider, initially, the general reaction of formation of a double helix of length N. Such duplex is formed from two separated and complementary strands S and S'. This process is the chemical reaction $S + S' \rightleftarrows S \cdot S'$. Figure 3 shows a scheme of the status of the two strands before and after the nucleation of the double helix. Before the nucleation, all the bases in each one of the two strands occupy the single strand state, and the two strands are sufficiently distant one from the other. Thereafter, during the nucleation, all the bases continue in the single strand

state, but the strands are approaching one to the other via juxtaposition between the bases b_k and $b_k^{'}$ ($1 \leq k \leq N$). We suppose, with this, that the nucleation occurs in the kth site of the double chain. Finally, after the nucleation, the formation of the WC first base pair occurs, that is, the base pair $b_k / b_k^{'}$ is formed. Succeeding the nucleation event and the formation of the first base pair, we have the propagation of the double helix to both the directions, extending to the two ends of the chain, if the transition is a two-state process. As shown in Figure 3, the formation of the first WC base pair is constituted by one nucleation step followed by one propagation step. Therefore, the equilibrium constant σs_k refers to the formation of the first WC base pair, through the establishment of hydrogen bonds between the bases b_k and $b_k^{'}$. If the free energy associated to the formation of the first base pair is ΔG_{init}, then we can write the equilibrium constant σs_k as

$$\sigma s_k = \exp\left\{-\frac{\Delta G_{\text{init}}}{RT}\right\} \tag{33}$$

that is,

$$\Delta G_{\text{init}} = -RT \ln \sigma s_k = \Delta G_{\text{nuc}} - RT \ln s_k. \tag{34}$$

In order to consider the propagation of the double helix from the nucleating base pair $b_k / b_k^{'}$ and extending to both the ends, Eq. 24 can be modified for to produce:

$$K_{\text{eq}} = \left(\prod_{i=1}^{k-1} s_i^{\leftarrow}\right) \sigma s_k \left(\prod_{i=k+1}^{N} s_i^{\rightarrow}\right) \tag{35}$$

In Eq. 35, σ is the nucleation equilibrium constant, $\kappa = \sigma s_k$ is the initiation equilibrium constant (which is evidently related to the process of formation of the WC first base pair $b_k / b_k^{'}$), and $s_i^{\leftarrow}$ ($i < k$) and $s_i^{\rightarrow}$ ($i > k$) are the propagation equilibrium constants related to the propagation of the double helix, by stacking of the base pair $b_i / b_i^{'}$ on the preexistent duplex, respectively, in the 3′–5′ (downward) and 5′–3′ (upward) directions. Thus, substituting Eq. 35 into Eq. 23, we obtain

$$\Delta G_T = -RT \ln \left[\left(\prod_{i=1}^{k-1} s_i^{\leftarrow}\right) \sigma s_k \left(\prod_{i=k+1}^{N} s_i^{\rightarrow}\right)\right] = -\sum_{i=1}^{k-1} RT \ln s_i^{\leftarrow} - RT \ln \sigma s_k - \sum_{i=k+1}^{N} RT \ln s_i^{\rightarrow}. \tag{36}$$

http://dx.doi.org/10.5772/62574

As the propagation equilibrium constant depends on the local composition, we associate to the propagation equilibrium constant for the addition of the *i*th base pair, in downward direction, a value such that

$$-RT\ln s_i^{\leftarrow} = \Delta G\left(b_i b_{i+1}\right) \tag{37}$$

.

Analogously, the propagation equilibrium constant for the addition of the *i*+1th base pair, in upward direction, assumes a value such that

$$-RT\ln s_{i+1}^{\rightarrow} = \Delta G\left(b_i b_{i+1}\right) \tag{38}$$

Thus, from Eqs. 37 and 38, the propagation equilibrium constants would be, to the light of the NN approach, given by

$$s_i^{\leftarrow} = s_{i+1}^{\rightarrow} = \exp\left[-\frac{\Delta G\left(b_i b_{i+1}\right)}{RT}\right] \tag{39}$$

The first summation in Eq. 36, $-\sum_{i=1}^{k-1} RT\ln s_i^{\leftarrow}$, refers to the sum of the free energies of all the duplex dimers in downward direction related to the nucleating base pair $b_k/b_k^{'}$. In another words, such term is the total free energy related to the propagation of the double helix, starting from the nucleating base pair $b_k/b_k^{'}$ and propagating in downward direction. From Eq. 37, we have clearly that $-\sum_{i=1}^{k-1} RT\ln s_i^{\leftarrow} = \sum_{i=1}^{k-1} \Delta G\left(b_i b_{i+1}\right)$. Now speaking about the second summation in Eq. 36, $-\sum_{i=k+1}^{N} RT\ln s_i^{\rightarrow}$, it refers to the free energies of all the duplex dimers in upward direction related to the base pair $b_k/b_k^{'}$, that is, it is the total free energy related to the propagation of the double helix, starting from the base pair $b_k/b_k^{'}$ and propagating in upward direction. Applying Eq. 38, we have $-\sum_{i=k+1}^{N} RT\ln s_i^{\rightarrow} = \sum_{i=k}^{N-1} \Delta G\left(b_i b_{i+1}\right)$. Thus, Eq. 36 can be rewritten as follows:

$$\begin{aligned}\Delta G_T &= -RT\ln\left[\left(\prod_{i=1}^{k-1} s_i^{\leftarrow}\right)\sigma s_k\left(\prod_{i=k+1}^{N} s_i^{\rightarrow}\right)\right] \\ &= \sum_{i=1}^{k-1}\Delta G\left(b_i b_{i+1}\right) - RT\ln\sigma s_k + \sum_{i=k}^{N-1}\Delta G\left(b_i b_{i+1}\right),\end{aligned} \tag{40}$$

that is,

$$\begin{aligned}\Delta G_T &= -RT\ln\left[\left(\prod_{i=1}^{k-1}s_i^{\leftarrow}\right)\sigma s_k\left(\prod_{i=k+1}^{N}s_i^{\rightarrow}\right)\right] \\ &= -RT\ln\sigma s_k+\sum_{i=1}^{N-1}\Delta G\left(b_ib_{i+1}\right),\end{aligned} \tag{41}$$

where $\Delta G_{\text{init}}=-RT\ln\sigma s_k$ is the initiation free energy, and $\sum_{i=1}^{N-1}\Delta G\left(b_ib_{i+1}\right)$ is the sum of the dimer free energies. Defining $\Delta G\left(Ob_k\right)=-RT\ln s_k$ the free energy change associated to the process of the "propagation" of the first WC base pair, we can rewrite Eq. 41 as

$$\begin{aligned}\Delta G_T &= \Delta G_{\text{init}}+\sum_{i=1}^{N-1}\Delta G\left(b_ib_{i+1}\right) \\ &= \Delta G_{\text{nuc}}+\Delta G\left(Ob_k\right)+\sum_{i=1}^{N-1}\Delta G\left(b_ib_{i+1}\right).\end{aligned} \tag{42}$$

Equation 42 shows that the free energy for the duplex formation can be written in terms of the initiation or the nucleation free energy, producing two approaches completely equivalent (the two equalities in Eq. 42). We will prefer, however, the first because it permits to obtain directly the initiation free energy for the duplex formation, as it will be shown in the next section. In addition, the nucleation free energy can be calculated from the initiation free energy, as shown in Eq. 34. Then, for applying Eq. 42, we will assume that the event of nucleation can occur by approaching the strands to each other via juxtaposition between any bases b_k and $b_k^{'}$ $(1\le k\le N)$, with equal probability. The "nucleating" base pair, in turn, can be an A/T or C/G base pair. Thus, if the event of the formation of the first base pair can occur at any site along the double chain with the same probability, we can write the observable initiation free energy as follows:

$$\left\langle\Delta G_{\text{init}}\right\rangle=p_{A/T}\Delta G^{\circ}\left(A/T\right)+p_{C/G}\Delta G^{\circ}\left(C/G\right). \tag{43}$$

In Eq. 43, $\langle\Delta G_{\text{init}}\rangle$ is the observable initiation free energy, and $p_{A/T}$ and $p_{C/G}=1-p_{A/T}$ are, respectively, the probabilities with which the first base pair formed in the DNA double chain is A/T and C/G base pair. Finally, $\Delta G^{\circ}(A/T)$ and $\Delta G^{\circ}(C/G)$, are the free energy changes associated to the formation of the first base pair if it is an A/T or C/G base pair, respectively. As our approach is built on the hypothesis that the event of the formation of the first base pair can occur at any site along the chain with equal probability, the probabilities $p_{A/T}$ and $p_{C/G}$ are simply the compositions of A/T and C/G base pairs. Then, we have that $p_{A/T}=\chi_{A/T}=n_{A/T}/N$, and $p_{C/G}=\chi_{C/G}=n_{C/G}/N$, where $\chi_{X/Y}$ and $n_{X/Y}$ are, in turn, respectively, the relative occurrence number (composition) and the number of X/Y base pairs

http://dx.doi.org/10.5772/62574

occurring along the duplex oligomer in question. Equation 34 shows how the nucleation free energy can be calculated from the initiation free energy. Therefore, the observable nucleation free energy can be written as

$$\left\langle \Delta G_{\text{nuc}} \right\rangle = \left\langle \Delta G_{\text{init}} \right\rangle + \left\langle RT \ln s_k \right\rangle. \tag{44}$$

The equilibrium constant s_k is associated to the first propagation step, that is, to the formation of the first WC base pair, which can be an A/T or C/G base pair. Invoking newly our simplifying hypothesis, which establishes that the formation of the first base pair can occur with equal probability in any site along the chain, we can write that

$$\left\langle RT \ln s_k \right\rangle = -\sum_{\{b_1 b_2\}} p_{b_1 b_2} \Delta G\left(b_1 b_2\right), \tag{45}$$

where the summation is over all the possible duplex dimers occurring along the chain, that is, b_1 and b_2 can be anyone of the four nucleotides A, T, C, or G. In Eq. 45, $p_{b_1 b_2}$ is the probability with which the base pair $b_2 / b_2^{'}$ is preceded by the base pair $b_1 / b_1^{'}$. Obviously, such probabilities are equal to the compositions of dimers along the double chain, that is, $p_{b_1 b_2} = \chi_{b_1 b_2}$, where $\chi_{b_1 b_2}$ is the composition of the duplex dimer $b_1 b_2 - b_2^{'} b_1^{'}$. Therefore,

$$\left\langle RT \ln s_k \right\rangle = \left\langle \Delta G\left(Ob_k\right) \right\rangle = -\sum_{\{b_1 b_2\}} \chi_{b_1 b_2} \Delta G\left(b_1 b_2\right). \tag{46}$$

Equation 44 can be rewritten as follows:

$$\left\langle \Delta G_{\text{nuc}} \right\rangle = -\sum\nolimits_{\{b_1 b_2\}} \chi_{b_1 b_2} \Delta G\left(b_1 b_2\right) + \chi_{\frac{A}{T}} \Delta G^{\circ}\left(A / T\right) + \chi_{C/G} \Delta G^{\circ}\left(C / G\right). \tag{47}$$

From Eq. 47, it becomes clear that the nucleation free energy depends on the composition of the DNA double strand due to the presence of the terms $\chi_{A/T}$, $\chi_{C/G}$, and $\chi_{b_1 b_2}$, in the right side of the equation. According to Eq. 47, as there are 10 possible duplex dimers, $\left\langle \Delta G_{\text{nuc}} \right\rangle$ must be a function of 10 parameters: the already known eight polymeric irreducible parameters plus two parameters related to the formation of the first base pair, as defined in Eq. 43. We can simplify the approach contained in Eq. 47, discriminating the bases b_1 and b_2 only according to the weak–strong classification criteria. In this way, Eq. 47 becomes

$$\begin{aligned} \left\langle \Delta G_{\text{nuc}} \right\rangle = &- \chi_{ww} \Delta G\left(ww\right) - \chi_{ws} \Delta G\left(ws\right) - \chi_{sw} \Delta G\left(sw\right) - \\ &- \chi_{ss} \Delta G\left(ss\right) + \chi_{w} \Delta G^{\circ}\left(A / T\right) + \chi_{s} \Delta G^{\circ}\left(C / G\right). \end{aligned} \tag{48}$$

where $\Delta G(ww)$ is the mean free energy of a stack of two weak base pairs, χ_{ww} is its composition, and so on. Using Eq. 12, we obtain the following values for the mean dimer free energies:

$$\begin{aligned} \Delta G(ww) &= S + V_z + M_{zz} \\ \Delta G(ss) &= S - V_z + M_{zz} \\ \Delta G(ws) &= \Delta G(sw) = S - M_{zz}. \end{aligned} \tag{49}$$

Inserting Eq. 49 into Eq. 48, we can obtain:

$$\begin{aligned} \langle \Delta G_{\text{nuc}} \rangle &= -S - V_z(\chi_{ww} - \chi_{ss}) - M_{zz}(\chi_{ww} + \chi_{ss} - \chi_{ws} - \chi_{sw}) + \\ &+ \chi_w \Delta G^{\circ}(A/T) + \chi_s \Delta G^{\circ}(C/G) \end{aligned} \tag{50}$$

Equation 42 can be used to predict the free energy of any duplex oligomer if we know the values of all the polymeric irreducible parameters for free energy plus the free energy changes associated to the formation of the first base pair. Now, we can return to the set of 108 sequences compiled by Allawi and SantaLucia to obtain the set of eight polymeric irreducible parameters together with these two additional parameters. This will be done in the following section.

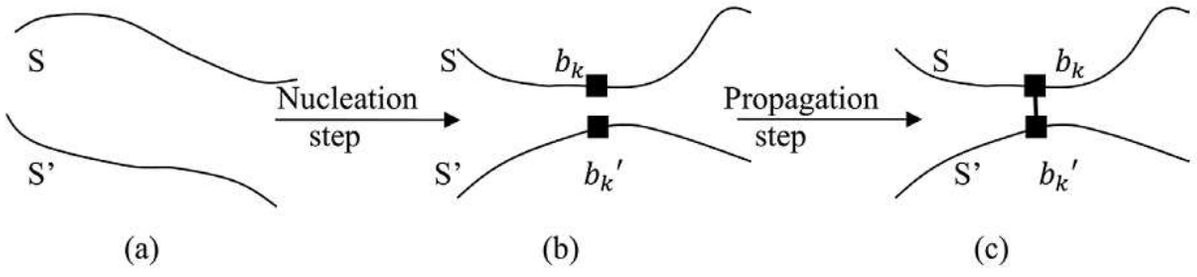

Figure 3. The formation of the first WC base pair. (a) Strands *S* and *S′* are sufficiently distant one from the other. All the bases in both the chains are in the single strand state. (b) It occurs an approximation between strands *S* and *S′*. However, all the bases are still in the single strand state. (c) It is formed the first base pair, namely the base pair $b_k / b_k^{'}$, through the establishment of H bonds between the bases b_k and $b_k^{'}$. After that, the double helix propagates in both the directions extending to the ends of the chain [17].

8. Results and discussion for the modeling based on double helix initiation parameters

The free energy for a duplex sequence of *N* bases in the NN approximation can be calculated as the pairwise sum, using the initiation free energy, as a function of the 10 parameters for free energy from Eqs. 12 and 43 as follows:

http://dx.doi.org/10.5772/62574

$$\left\langle \Delta G_{\text{init}} \right\rangle + \sum_{i=1}^{N-1} \Delta G\left(b_i b_{i+1}\right) + \Delta G_{\text{sym}} \tag{51}$$

Simultaneous least-mean-square-deviation fit of this model to the 108 sequence data set compiled by Allawi and SantaLucia [2] gave the values for the free energy parameters, as listed in Table 3 [17].

Irreducible parameters for free energy	Values (kcal/mol)
$\Delta G^{\circ}(A/T)$	1.7± 0.3
$\Delta G^{\circ}(C/G)$	1.8± 0.2
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

Table 3. Irreducible Parameters for Free Energy at Standard Conditions (37°C and 1 M Salt and DNA)

Given the root-mean-square deviation per dimer, as defined in Eq. 22, the parameters for free energy in Table 3 are those that minimize χ. The value obtained for χ was 0.14 kcal/mol per dimer [17], which coincides precisely with that obtained for the 12 parameter models using values reported by SantaLucia for the free energies of the 10 duplex dimers [2, 4–6]. Thus, how it happened for the modeling by end effects, from the overall data ensemble quality, there would not be practical reason to prefer a model with a greater number of parameters. The mean values and the errors of the parameters for free energy, as listed in Table 3, were estimated by Guerra in the following way [17]: he selected 1000 sets of 70 sequences chosen randomly, and then he calculated the mean and the deviation for the parameters obtained from each set. Some immediate conclusions can be done with respect to the data contained in Table 3. First, the intrinsic errors of the free energies related to the formation of the first base pairings are only a little larger than the errors of the other irreducible parameters for free energy. Second, considering only the bar of errors, the free energy changes for the initiation of a double chain through the formation of an A/T or C/G base pair are essentially similar. Thus, if it is correct the hypothesis of that the duplex formation can be initiated by the formation of a base pair at any site along the double helix with equal probability (independently of the local composition), then, the initiation free energy is essentially independent on the local composi-

tion. Finally, once we have obtained the initiation free energy parameters, as listed in Table 3, we are ready for to estimate the nucleation free energy of any duplex oligomer, using Eq. 50. Equation 50 establishes that observable nucleation free energies depend on the mean global composition of the DNA double strand and vary within a range that goes from

$$\Delta G_{\text{nuc}}^{\text{poly } A\cdot T} = 1.38 - 0.58 + 0.04 + 1.7 = 2.54 \frac{\text{kcal}}{\text{mol}},$$

for a poly $A\cdot T$ homopolymer to

$$\Delta G_{\text{nuc}}^{\text{poly } C\cdot G} = 1.38 + 0.58 + 0.04 + 1.8 = 3.80 \frac{\text{kcal}}{\text{mol}},$$

for a poly $C\cdot G$ homopolymer. Observe that the difference between these values for nucleation free energies, which is ~1.3 kcal/mol, is greater than the bar of errors estimated for nucleation free energies, which is ~0.7 kcal/mol. On the another hand, the results obtained above, for the poly $A\cdot T$ and poly $C\cdot G$ homopolymers, are in total discordance with results obtained previously using the modeling by end effects [16], as was to be expected. In fact, heteropolymers must have one value for the nucleation free energy that must be inside such interval, and it must depend on their composition. Finally, the mean observable nucleation free energy is $\Delta G_{\text{nuc}}^{\text{mean}} = \left(\Delta G_{\text{nuc}}^{\text{poly } A\cdot T} + \Delta G_{\text{nuc}}^{\text{poly } C\cdot G}\right)/2 = 3.17 \text{ kcal/mol}$. This value is only a little lower than that obtained by Manyanga et al. [19].

Comparing the results obtained for the eight polymeric irreducible parameters for free energy, as listed in Table 3 of this section, with results obtained recently using the end effects [16], and contained in Tables 1 and 2 of the Section 6 of this review, we can conclude that the irreducible parameters are not essentially affected with the alteration in the modeling. In another words, if we substitute one modeling by another, the end effects, which, obviously, do not depend essentially on the compositions of the two terminal base pairs, are substituted by the initiation free energies, which do not depend essentially on the global composition of the chain. Therefore, dimer free energies, which depend only on the irreducible parameters for free energy, also are not essentially affected.

Free energy changes associated to the formation of the second base pair are given by the following equation:

$$\left\langle \Delta G\left(b_{k\pm1} / b'_{k\pm1}\right)\right\rangle = \left\langle \Delta G\left(b_k / b'_k\right)\right\rangle + \left\langle \Delta G\left(Ob_k\right)\right\rangle \tag{52}$$

depending if the second base pair formed is located at the k+1th site or at the k−1th site of the chain. Using Eq. 43 for $\left\langle \Delta G\left(b_k / b'_k\right)\right\rangle = \left\langle \Delta G_{\text{init}}\right\rangle$, Eq. 46 for $\left\langle \Delta G(Ob_k)\right\rangle$, and also the approximations given by Eq. 49, we obtain the following:

http://dx.doi.org/10.5772/62574

$$\left\langle \Delta G\left(A/T\right)\right\rangle_{\text{base pairing}} = \left(0.7 \pm 0.3\right)\ \text{kcal / mol}$$

and

$$\left\langle \Delta G\left(C/G\right)\right\rangle_{\text{base pairing}} = \left(0.1 \pm 0.3\right)\text{kcal / mol.}$$

The values listed above are just the base pairing contributions for the dimer free energies, which were encountered experimentally by the Frank-Kamenetskii Group [18]. Yakovchuk et al. obtained for the A/T and C/G base pairings, the base pairing free energies of 0.57 kcal/mol and −0.11 kcal/mol, respectively [18]. Therefore, we have obtained values that agree reasonably well with those obtained by the Frank-Kamenetskii Group. In addition, the values for the base pairing free energies are reasonably well defined because their ranges of allowable values have only an unique common intercept.

9. Conclusions

A geometrical representation of four-nucleotide sets as a tetrahedron (Eq. 3 and Fig. 2) allows for the association of the three most distinctive molecular group classifications with corresponding orthogonal cubic axis. 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. This includes an ensemble mean expectation from scalar coefficient S alone and a global composition approximation, as expressed through V-component contributions. The symmetrical set is shown to provide a frame for the analysis of DNA duplex free energy fully compatible with experimental data. Such a symmetrical set of coefficients allows for the translation among different decomposition frames. It also gives a proper irreducible representation for dimer properties (Eqs. 8 and 12). It solves an old indeterminacy of dimer sets by establishing self-consistency relations among the dimer coefficients (Eqs. 14 and 15).

Using the modeling based on end effects, for predicting correctly physical properties of duplex oligomers, we saw that end contributions are fit with large errors to experimental data, as compared to the fits of other NN or dimer contributions. Besides, we could not distinguish between the weak and the strong terminal base pairs. However, using both the sets constituted by two- or four-ending parameters, one calculates free energies for DNA oligomers at least as well as standard models, considering a larger set of parameters do [2, 4–6].

The modeling based on the double helix initiation parameters substitutes the end effects by the initiation parameters. The free energy changes associated to the formation of the first base pair, in the duplex formation, are fit to experimental data with errors only slightly larger than those for the NN or dimer contributions. Furthermore, we obtained that the values for the first

base pairing free energies are essentially similar (because the difference between them had a value smaller than the estimated bar of errors). Thus, this could indicate an invariance of the initiation free energy with respect to the composition of the chain. Nucleation free energy, however, depends on the composition, and it can be calculated from the initiation free energy by using Eq. 34. What supports this statement is the fact of that the difference between its maximal and minimal values is larger than the error bars. The model based on the double helix initiation parameters is constructed by using the simplifying hypothesis, which establishes that the nucleation can occur at any site of the chain with equal probability, independently of the local composition. An important result, which becomes such hypothesis quite reasonable, is the fact of that the base pairing contributions for the dimer free energies seem to agree well with values experimentally obtained by the Frank-Kamenetskii Group. Finally, this modeling uses a set of 10 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.

Author details

João C. O. Guerra*

Address all correspondence to: jcog@infis.ufu.br

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

References

[1] Gray D M, Tinoco I, Jr. A new approach to study of sequence-dependent properties of nucleotides. Biopolymers. 1970; 9: 223–244. DOI:10.1002/bip.1970.360090207

[2] Allawi H T, SantaLucia J, Jr. Thermodynamics and NMR of internal G T mismatches in DNA. Biochemistry. 1997; 36: 10581–10594. DOI:10.1021/bi962590c

[3] Breslauer K J, Frank R, Blocker H, Marky L A. Predicting DNA duplex stability from the base sequences. Proc Natl Acad Sci USA. 1986; 83: 3746–3750.

[4] SantaLucia J, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci USA. 1998; 95: 1460–1465.

[5] SantaLucia J, Jr., Hicks D. The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct. 2004; 33: 415–440. DOI:10.1146/annurev.biophys.32.110601.141800

[6] SantaLucia J, Jr., Allawi H T, Seneviratne P A. Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry. 1986; 35: 3555–3562. DOI: 10.1021/bi951907q

[7] Cantor C R, Schimmel P R. Biophysical Chemistry Part III: The Behavior of Biological Macromolecules. San Francisco: Freeman; 1980.

[8] Vologodskii A V, Amirikyan B R, Lyubchenko Y L, Frank-Kamenetskii M D. Allowance for heterogeneous stacking in the DNA helix-coil transition theory. J Biomol Struct Dyn. 1984; 2: 131–148. DOI:10.1080/07391102.1984.10507552

[9] Goldstein R F, Benight A S. How many numbers are required to specify sequence-dependent properties of polynucleotides? Biopolymers. 1992; 32: 1679–1693. DOI: 10.1002/bip.360321210

[10] Gray D M. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. II. Thermodynamic parameters of DNA RNA hybrids and DNA duplexes. Biopolymers. 1997; 42: 795–810. DOI:10.1002/(SICI)1097-0282(199712)42:7<795::AID-BIP5>3.0.CO;2-O

[11] Gray D M. 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. 1997; 42: 783–793. DOI:10.1002/(SICI)1097-0282(199712)42:7<783::AID-BIP4>3.0.CO;2-P

[12] Doktycz M J, Goldstein R F, Paner T M, Gallo F J, Benight A S. 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. 1992; 32: 849–864. DOI:10.1002/bip.360320712

[13] Licinio P, Guerra J C O. Irreducible representation for nucleotide sequence physical properties and self-consistency of nearest-neighbor dimer sets. Biophys J. 2007; 92: 2000–2006. DOI:10.1529/biophysj.106.095059

[14] Licinio P, Caligiorne R B. Inference of phylogenetic distances from DNA-walk divergences. Phys A Stat Theor Phys. 2004; 341: 471–481. DOI:10.1016/j.physa.2004.03.098

[15] Xia T, SantaLucia J, Jr., Burkard M E, Kierzek R, Schroeder S J, Jiao X, Cox C, Turner D H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson−Crick base pairs. Biochemistry. 1998; 37: 14719–14735. DOI:10.1021/bi9809425

[16] Guerra J C O, Licinio P. Terminal contributions for duplex oligonucleotide thermodynamic properties in the context of nearest neighbor models. Biopolymers. 2011; 95: 194–201. DOI:10.1002/bip.21560

[17] Guerra J C O. Calculation of nucleation free energy for duplex oligomers in the context of nearest neighbor models. Biopolymers. 2013; 99: 538–547. DOI:10.1002/bip. 22214

[18] Yakovchuk P, Protozanova E, Frank-Kamenetskii M D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. 2006; 34: 564–574. DOI:10.1093/nar/gkj454

[19] Manyanga F, Horne M T, Brewood G P, Fish D J, Dickman R, Benight A S. Origins of the "nucleation" free energy in the hybridization thermodynamics of short duplex DNA. J Phys Chem B. 2009; 113: 2556–2563. DOI:10.1021/jp809541m

IntechOpen