Theoretical predictions of the melting
temperature for DNA using the
stochastic matrix method

L. Dagdug
Departamento de Física, Universidad Autónoma Metropolitana-Iztapalapa,
México, Apartado Postal 55-534, 09340 México, D. F., México.
e-mail: dll@xanum.uam.mx

E. Vázquez-Contreras
Instituto de Química. Departamento de Bioquímica.
Universidad Nacional Autónoma de México.
México, D.F. 04510, México

February 26, 2001

Abstract
In this paper we extend the ideas to describe the glass transition in strong glasses using the stochastic matrix method to obtain the melting temperature for DNA. To carry out our purpose we take the purines and pyrimidines as principal entities, the simplest ones. The main result is the prediction of the melting temperature of poly(dpurine)poly(dpurine) and poly(dpurine-dpyrimidine)poly(dpurine-dpyrimidine), without adjustable parameters and with an excellent agreement with experimental data.

Keywords: DNA; melting temperature, helix-coil transition, nearest-neighbor thermodynamics, stochastic matrix method.

En este artículo se extienden las ideas utilizadas para describir la transición vitrea de vidrios fuertes utilizando el método de la matriz estocástica en DNA para obtener su temperatura de desnaturalización. Para llevar a cabo nuestro propósito utilizamos purinas y pirimidinas como nuestra principal unidad, la más simple posible. El principal resultado que se obtiene es la predicción de las temperaturas de desnaturalización del poly(dpurine)poly(dpyrimidine) y del poly(dpurine-dpyrimidine)poly(dpurine-dpyrimidine), sin utilizar parámetros ajustables y con un excelente ajuste a los valores obtenidos experimentalmente.

Descriptores: DNA, temperatura de desnaturalización, transición, termodinámica de primeros vecinos, método de la matriz estocástica.
Introduction

Nucleic acids consist of a chemical sequence of their fundamental components, nucleotides, analogous to the sequence of linked amino acids comprising a protein. Nucleotides comprise three elements: a heterocyclic ring containing nitrogen, a five-carbon sugar in ring form, and a phosphate group. The bases fall into two classes of pyrimidines (Py) and purines (Pu); the former have a six-membered ring and the latter consist of two fused six- and five-membered rings. Each nucleic acid commonly contains four different types of base, two of each ring form, although the two types of nucleic acid, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), differ in one of the pyrimidines present. Both DNA and RNA contain the same two purines, adenine and guanine, DNA possesses the pyrimidines cytosine and thymine, in RNA the thymine is replaced by uracil. The primary structure of DNA has each nucleoside joined by a phosphodiester from its -hidroxil group to the -hidroxil of one neighbor and by a second phosphodiester from its -hidroxil group to the -hidroxil of its other neighbor [1].

In general, DNA consists of two very long polynucleotide chains wound around each other so as to give a double helix. Watson and Crick [2] suggested that this could be accounted for if a purine from one polynucleotide chain always partners a pyrimidine from the other chain; both purine-purine and pyrimidine-pyrimidine interactions would be prohibited. They suggested that the two strands would be held together by hydrogen bonding, and would not be covalent linked. The building of molecular models explained this by suggesting that hydrogen bonding could occur between the nitrogenous bases such that the purine guanine could bond only to the pyrimidine cytosine, and not to thymine. Similarly, adenine and thymine would form hydrogen bonds only with each other. Thus a thymine of one chain would always oppose an adenine. The
adenine-thymine (A¢T) base pair has two hydrogen bonds, and the guanine-
cytosine (G¢C) pair has three. The two bases comprising each pair are said to
be complementary. This model requires the two-polynucleotide chains to run in
opposite directions, antiparallel. Since pairing occurs in this way, the amount
of adenine and thymine should be equal, and quantities of cytosine and gua-
nine should likewise be equal to each other, also the ratio between adenines and
pyrimidines is one. This last result was an important piece of information for
the development of the structure of DNA and came from the work of Chargax and co-workers [3].

The noncovalent forces that stabilize the double helix are disrupted by heat-
ing or by exposure to low salt concentration. The two double strands of double
helix separate entirely when all the hydrogen bonds between them are broken.
The process of strand separation is called denaturation or melting. The mid point
of the temperature range over which the strands of DNA separates is called the
melting temperature, denoted Tm.1

A DNA melting curve is generally a two-dimensional plot displaying some
property of a DNA solution against an external variable producing DNA un-
winding. The external variable is most commonly the temperature, but melting
can also be brought about by extremes of pH, decreases in dielectric constant
of the aqueous medium by alcohol, ketones, etc., exposure to amides, ureas and
similar solvents [4].

Studies of the structure and dynamics of DNA are vital for understanding
the mechanism of how the genetic code is expressed, processes of DNA repli-
cation and transcription, DNA-protein recognition, DNA-drug interactions and
others [5].

Accurate prediction of DNA thermal denaturation is important for several
molecular biological techniques including PCR [6], sequencing by hybridization
[7], antigen targeting [8], and Southern blotting [9]. In this techniques, choice
of a nonoptimal sequence or temperature can lead to amplification or detection
of wrong sequences [10]. Furthermore, knowledge of the sequence dependence
of DNA melting is important for understanding the details of DNA replication,
mutation, repair, and transcription [11],[12].

The cooperative unwinding of two stranded DNA to single strands is also
known as the helix-coil transition. Investigation of this transition have provide
information and insight on the interactions governing DNA unwinding, the in-
fluence of base pair sequence on regional unwinding, and the influence of the
solvent on DNA stability [13],[14],[15],[16].

In this paper we extend the ideas to describe the glass transition in strong
glasses using the stochastic matrix method (SMM) to predict the melting tem-
peratures of DNA [17],[18],[19],[20]. In section 2 we present the stochastic ma-
trix method. In section 3 we use the theoretical ideas presented in section 2
to describe the melting process in DNA, for example, we obtain the Tm of
poly(dpurine)¢poly(dpymidime) and poly(dpurine-dpyrimidine)¢poly(dpurine-

1Tm depend on the specific DNA and solvent and generally fell between 50± C and 100± C.
Finally in section 4 we make some remarks on the nature of these results.

2 The stochastic matrix method

The process of observation of a configuration of a system can be described by matrix (M) acting on a vector (v) if the matrix components are the probability to have a unit (base-pair) presiding for another one, and if the vector components represent the probabilities to have a given unit in the configuration. The matrix transforms this vector into a new one when the next unit is observed, therefore transformation of the vector components depends on the new unit stuck. To obtain the probability to have some configuration of \( n \) base-pairs (bp) is modeled by \( n \) successive applications of the matrix \( M \) into an initial vector \( v_0 \), which characterizes the initial condition on the system. After \( n \) applications, the final configuration of the system can be written as a linear combination of the eigenvectors associated with \( M \), i.e., \( v_n = \sum_{m=1}^{\infty} a_m e_m \), where \( e_m \) is the eigenvector \( M \) with eigenvalue \( \lambda_m \) and \( a_m \) is the projection of \( v_0 \) into \( e_m \).

A matrix with all the columns normalized to one, as \( M^2 \), has the property that at least one eigenvalue is one, while the real part of all others eigenvalues is less than one. This result allows us to assert that only the eigenvectors with eigenvalues equal to one survive after successive applications of \( M \) into \( v_0 \). If we assume that \( M \) has one eigenvector \( e_1 \) with eigenvalue equal to one, then in the limit of large \( n \), \( v_0^1 \) converges to \( e_1 \), with \( a_1 = 1 \); due to conservation of probability. Therefore, this means that the configuration attains a steady statistical regimen presented by \( e_1 \). The explicit form of this eigenvector is obtained by solving the system of equations:

\[
(M - 1)e_1 = 0
\]  

Equation 1 allows us to calculate the probability to have any configuration in the system.

The NN model for nucleic acids assumes that the stability of a given bp depends on the identity and orientation of neighboring base pair. The application of the NN model to nucleic acid was pioneered by Zimm [21] and by Tinoco and coworkers [22][23][24][25][26]. For our theoretical description the NN model allows us to construct the stochastic matrix that describes DNA.

To construct the stochastic matrix that allows to describe the melting behavior of DNA we first need to define the units. These units must be given by two combinations: "Pu and Py#" and "Py and Pu#." The bp can be bonding or unbinding, this can be represented as "PuPy#" "PyPu#" and "PuPy#" "PyPu#" where the dot represent the existence of the hydrogen bonding. These four sites give 16 different combination of base stack. For DNA the units are stuck by a phosphodiester from its 5'-hydroxil group to the 3'-hydroxil group of one neighbor and by a second phosphodiester from its 3'-hydroxil to the 5'-hydroxil of its other neighbor.

\(^2\)The columns of matrix \( M \) are normalized to one because each one is a Markov chain.
The 16 different combinations for neighbors can be displayed as a $4 \times 4$ matrix, namely$^3$:

\[
\begin{array}{cccc}
0 & \text{PuPy} & \text{PyPu} & \text{PyPy} \\
\text{PyPu} & \text{PuPy} & \text{PyPy} & \text{PyPy} \\
\text{PyPu} & \text{PyPy} & \text{PyPy} & \text{PyPy} \\
\text{PyPy} & \text{PyPy} & \text{PyPy} & \text{PyPy} \\
\end{array}
\]

where "PuPy" # represents the probability to have a bonding PuPy presiding a bonding PuPy, "PuPy" # represents the probability to have a bonding PuPy presiding an unbonding PuPy, etc. These stack process are in three dimensions and this information must be included in the stacking energy$^4$. The stacking energy also can be determined which kind of DNA we are working$^5$.

Each stack process has a finite probability of occurrence and the statistical weight for each process is one, because only exists one form to stack a bp. Based on NN model each configuration is proportional to its stability constant ($s_i = \epsilon G_i = T \epsilon S_i$ where $i = \text{PuPy}$ or PyPu), inserting all the energetically contributions in matrix 2, the explicit matrix is written as:

$\text{Arrows designate the direction of the sugar-phosphate chain, from } C_0 \text{ atom of a deoxyribose unit to } C_0 \text{ the atom of the next deoxyribose adjacent to and on either side of the phospho diester linkage. Some times nearest-neighbor base pair are represented with a slash separating strands in antiparallel orientation (e.g., AC/TG means } 5'_{\text{AC}} \text{ Watson-Crick base paired with } 3'_{\text{TG}} \text{ or } A^\text{\#}T^\text{\#} \text{ in the notation used throughout this paper)$.}

$\text{Also the information about the solvent, as pH, must be included in the stacking energies.}$

$\text{In the first phase of investigation of DNA secondary structure, diffraction studies on heterogeneous DNA fibers identified two distinct conformations for the DNA double-helix. At low humidity and high salt the favored form is the highly crystalline A-DNA while at high humidity and low salt the dominant structure is B-DNA. B-DNA has now grown into a family of structures encompassing B-, B', C-, C', C''-, D-, E-, and T-DNA's. In 1979 Rich and co-workers solve the x-ray crystal structure of a left-handed helical structure, named Z-DNA. Two particular hallmarks of B-DNA, in contrast to A- and Z-forms, are its flexibility and its capacity to make small adjustments in local helix structure in response to particular base sequences.}$
where $\xi G_{MN}$ is the free energy between the $n^{th}$ $M$th $N$th neighbors, $\xi G_h$ is the free energy due to hydrogen bonds, $k_B$ is the Boltzmann constant and $T$ the temperature.

The eigenvector with eigenvalue equal to one of matrix 3 is a vector with four components that gives us the probability to find any of the following configurations in the system $"PuPy"$, $"PuPy"$, $"PuPy"$, or $"PuPy"$. The sum $"PuPy" + "PuPy"$ gives us the probability to have the denatured bp in the system. Equaling the denatured probability to $1/2$ the $T_m$ for DNA chains for large $n$. Large $n$ means chains larger than 20 base-pairs (bp) and shorter than 350 bp. The shorter regime is imposed because below 20 iterations of $v$ applied to $M$ the probability do not reach his stable value. The lowest value is given because the cooperatively melting regions correspond to 350 bp in length and in our model we don take into account the influence of cooperatively region in the melting process. Although we don take into account this interactions in our theoretical description of denaturation, one of the interesting characteristics of biological macromolecules is the interplay between the cooperative interactions between regions and the independent properties of these regions[4].

In the next section we use matrix 2 to obtain the $T_m$ for two simply systems, $\text{poly}(dPu) \text{poly}(dPy)$ and $\text{poly}(dPu-dPy) \text{poly}(dPu-dPy)$.

### 3 Results and comparison with experiment

In this section we use matrix 2 to calculate $T_m$ for $\text{poly}(dPu) \text{poly}(dPy)$ and $\text{poly}(dPu-dPy) \text{poly}(dPu-dPy)$. To obtain the $T_m$ we proceeded as follows: obtaining the eigenvector with eigenvalue equal to one we can find the probability to have native and denatured bp. Using the definition of $T_m$ we equal the probability of denatured to $1/2$ and solving the equation for $T_m$ we can find the melting temperature.
3.1 Calculation of $T_m$ for poly(dpurines)poly(dpyrimidines)

For this particular case only the terms "PuPu" are conserved in matrix 2, and this matrix is reduced to a 2 × 2 matrix, namely:

$$
\begin{pmatrix}
0 & "PuPy" \\
"PuPy" & 1
\end{pmatrix}
$$

Inserting the energetically contributions in matrix 4 we find that,

$$
\begin{pmatrix}
0 & e^{-\frac{\epsilon_{GPuPu} + \epsilon_{GhPuPy}}{k_B T}} \\
e^{-\frac{\epsilon_{GPuPu} + \epsilon_{GhPuPy}}{k_B T}} & 1
\end{pmatrix}
$$

(4)

where $\epsilon_{GPuPu}$ is the free energy for the dimer "PuPu" and $\epsilon_{GhPuPy}$ is the free energy for the hydrogen bond between Pu and Py. If the hydrogen bond of a bp is broken in a base pair we have $\epsilon_{GPuPu}$ and if the hydrogen bonds of a dimer are broken we have $\epsilon_{GPuPu} + 2\epsilon_{GhPuPy}$.

After normalizing each column of matrix 5 we obtain:

$$
\begin{pmatrix}
0 & e^{-\frac{\epsilon_{GPuPu} + \epsilon_{GhPuPy}}{k_B T}} \\
e^{-\frac{\epsilon_{GPuPu} + \epsilon_{GhPuPy}}{k_B T}} & 1
\end{pmatrix}
$$

(5)

The explicit form of the eigenvector with eigenvalue one is obtained solving equation 1, which for the present case yields the following vector,

$$
\begin{pmatrix}
0 \\
1 + e^{-\frac{\epsilon_{GPuPu} + \epsilon_{GhPuPy}}{k_B T}} + e^{-\frac{2\epsilon_{GPuPu} + 2\epsilon_{GhPuPy}}{k_B T}}
\end{pmatrix}
$$

(6)

Vector 7 gives us the probability to find "PuPu" and "PuPy" base-pairs in poly(dPu)poly(dPy) at any temperature. Now, if we want to calculate the $T_m$ we only need to equal the second component of vector 7 to one half,

$$
\frac{1 + e^{Gh=k_B T_m}}{1 + 2e^{Gh=k_B T_m} \pm e^{2Gh=k_B T_m}} = \frac{1}{2}
$$

(7)
Equation 8 implies that $\xi \ G_h = 0$ and this is the condition to find the melting temperature. $\xi \ G_{hp_{py}}$ can also be calculated from $\xi \ H_{hp_{py}}$ and $\xi \ S_{hp_{py}}$ by using the equation:

$$\xi \ G_{hp_{py}} = \xi \ H_{hp_{py}} - T_m \xi \ S_{hp_{py}}$$  \hspace{1cm} (9)

where $\xi \ H_{hp_{py}}$ is the enthalpy due to a hydrogen bonding and $\xi \ S_{hp_{py}}$ his respective entropy.

If $\xi \ G_{hp_{py}} = 0$, then:

$$T_m = \frac{\xi \ H_{hc}}{\xi \ S_{hc}}$$  \hspace{1cm} (10)

It is very important to remark that equation 10 only depends on hydrogen parameters and has not influence of staking free energies, as it was shown by Wartell and Benight[4].

If we want to calculate $T_m$ for poly(dG)poly(dC), guanine is the purine and cytosine is the pyrimidine. Experimentally was founded that $\xi \ H_{gc} = \pm 5.8 \text{kcal/mol}$[28] and that $\xi \ S_{hc} = \pm 16 \text{e.u.}$, substituting these values in equation 10 we found that melting temperature for poly(dG)poly(dC) is 362.5 K, experimentally 360.8 K[29] is obtained; only 1.7 K of difference. Also equation 10 implicate that the process is given in one step, this means that all the chain is denatured at the same time at $T_m$.

Using equation 10 we are able to calculate for poly(dA)poly(dT) $\xi \ S_{bt}$, since experimentally is founded that $\xi \ 2 \text{kcal/mol}$ is released per mol hydrogen bonding formed and $T_m = 53 \text{K}$[36]. Taking into account that between A and T exist two hydrogen bonding (around $\xi \ 4 \text{kcal/mol}$), using equation 10 we found $\xi \ S_{bt} \pm 12.26 \text{e.u.}$; between $\pm 11 \text{e.u.}; \pm 13 \text{e.u.}$, the range of values obtained experimentally.

### 3.2 Calculation of $T_m$ for poly(dpurine-dpyrimidine)poly(dpurine-dpyrimidine)

In this subsection, as in the preceding subsection, using matrix 2 we find the $T_m$ for poly(dPu-dP py)poly(dPu-dPy).

For this particular case in matrix 2 only the terms " Pu P y" are conserved, obtaining:

$$\begin{pmatrix}
0 & 0 & " Pu P y" \\
" Pu P y" & 0 & " Py Pu" \\
" Py Pu" & 0 & 0
\end{pmatrix}$$  \hspace{1cm} (11)

Abbreviations: e.u., entropy units (cal/K mol)
Inserting the energetically contributions in the last matrix we obtain that,
\[
\begin{array}{cccc}
0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} \\
e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 \\
0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} \\
e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 \\
0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m} & 0 & e^\frac{-\epsilon G_{P\text{Py}}}{k_B T_m}
\end{array}
\]

After normalizing each column of matrix 12, solving equation 1 with \( M \) given by equation 12 and adding \( "P_u P_y + "P_y P_u = 1 = 2 \), we find the following condition to obtain the \( T_m \):
\[
e^\frac{-\epsilon G_{P\text{PuPy}} + \epsilon G_{P\text{PyPu}} + \epsilon G_{P\text{PyPy}}}{k_B T_m} + 2e^\frac{-\epsilon G_{P\text{PuPy}} + \epsilon G_{P\text{PyPu}}}{k_B T_m} + e^\frac{-\epsilon G_{P\text{PuPy}}}{k_B T_m} + e^\frac{-\epsilon G_{P\text{PyPu}}}{k_B T_m} = 0
\]

Equation 13 depends on the hydrogen parameters as the staking ones. Condition given in equation 13 is more complicated than \( \epsilon G_{H_{P\text{PuPy}}} = 0 \), the condition obtained to calculate \( T_m \) for poly(dG)poly(dC). In this case the stacking energies place a fundamental role, and the competition between \( \epsilon S \) and \( \epsilon H \) govern the melting behavior.

Using Taylor expansion at first order in the exponentials in equation 13 we find that:
\[
\epsilon G_{P\text{PuPy}} + \epsilon G_{P\text{PyPu}} = 0
\]

Inserting \( \epsilon G_{P\text{PuPy}} = \epsilon H_{P\text{PuPy}} \) and \( \epsilon G_{P\text{PyPu}} = \epsilon H_{P\text{PyPu}} \), in equation 14 we find,
\[
T_m = \frac{\frac{\epsilon H_{P\text{PuPy}} + \epsilon H_{P\text{PyPu}}}{\epsilon S_{P\text{PuPy}} + \epsilon S_{P\text{PyPu}}}}{\frac{\epsilon S_{P\text{PuPy}} + \epsilon S_{P\text{PyPu}}}{\epsilon S_{P\text{PuPy}} + \epsilon S_{P\text{PyPu}}}}
\]

If \( \epsilon S_{P\text{PyPy}} = \epsilon S_{P\text{PyPu}} \) finally we obtain
\[
T_m = \frac{\frac{\epsilon H_{P\text{PuPy}} + \epsilon H_{P\text{PyPu}}}{\epsilon S_{P\text{PuPy}} + \epsilon S_{P\text{PyPu}}}}{\frac{\epsilon S_{P\text{PyPy}} + \epsilon S_{P\text{PyPu}}}{\epsilon S_{P\text{PuPy}} + \epsilon S_{P\text{PyPu}}}}
\]

Equation 16 reproduce the well known result \( T_m = P T_{MN} \). Our theoretical framework give us the explicit conditions imposed to obtain this last result. In first place, the ratio between the stacking energies and \( k_B T \) should permit us
to apply a Taylor expansion. In second place, $\xi S_{PuPy} / 4 \xi S_{PyPu}$, so that DNA polymers are well accepted.

It is difficult to decide which experimental parameters use in equation 16 because there has been disagreement concerning this issue. Particularly there is a difference between DNA polymer and oligonucleotide MN thermodynamic trends and the salt dependence of nucleic denaturation. The major source of confusion in the literature is that the different studies use different oligonucleotide and polymer design, different methods for determining thermodynamics, different method for analyzing data, different salt conditions, and different formats for presenting MN parameters. Some of the experimental values of nearest-neighbor thermodynamics can be consulted in references: [29],[30],[31],[32],[33],[34] and [35]. Also theoretical efforts have been spent to calculate stacking energies by \textit{ab initio} calculations, for a review consult reference [5]. Equation 13 could help to discern which set of data is the best one since it is only well established that $\xi PuPy / PyPu$ is more stable than $\xi PuPy / PyPu$.

4 Conclusions

In this article we present a theoretical framework that allows us to predict $T_m$ of poly(dG)poly(dC) and poly(dG-dC)poly(dG-dC) without adjustable parameters. This theoretical framework was supported in the description of denaturation of DNA by the stochastic matrix method. For poly(dpurine)poly(dpyrimidine) we obtain an interesting result for the melting temperature, it only depends on $\xi G_{hpyp}$ (the free energy for the hydrogen bonds between purines and pyrimidines), and at $T_m$ it could carry out that $\xi G_{hpyp} = 0$. Using the relation $\xi G_{hpyp} = \xi H_{hpyp} - T \xi S_{hpyp} = 0$ we obtain $T_m = \xi H_{hpyp} / \xi S_{hpyp}$. It is important to remark that this last result does not depend on the stacking energies, and predict that the denaturation of this chain is in one step.

For poly(dpurine-dpyrimidine)poly(dpurine-dpyrimidine) we find the well known relation $T_m = T_{MN}$ but only under specific restrictions between stacking energies and entropies. In one hand the ratio between stacking energies and $k_B T$ must be educated to apply a Taylor expansion, and in other hand $\xi S_{PyPu} / 4 \xi S_{PyPu}$.

Finally, it is important to remark that in this paper we give a theoretical framework that allows us to obtain a theoretical description of DNA denaturation and to calculate his melting temperature without adjustable parameters. In a future work it could be extended this description taking into account that nature of purines and pyrimidines, it means, include adenines, guanines, thymines and cytosines in the stochastic matrix.

\footnote{It should be stressed that the stacking energies calculated by \textit{ab initio} account consider only molecules in vacuo and do not take into account hydrophobic interactions which however contribute significantly to stacking interactions.}
Acknowledgement

L. D gratefully acknowledges the guidance and useful conversations with Professor L. S. Garcia-Colin who contributed the best comprehension of some of the ideas exposed in this work. Also authors acknowledge with gratitude the important discussions with O. Resendiz-Antonio at the beginning of this work.

References