International Journal of Theoretical Physics, Vol. 29, No. 8, 1990

CG Methylation in DNA Transcription

J. Chela-Flores[1,2] and R. L. Migoni[1,3]

Received January 18, 1990

A simple model of DNA is considered in which the nucleotides cytosine (C) and guanine (G) are not assumed to be identical, and in which macroscopic thermodynamic quantities may be calculated exactly. The H bonds between the C and G nucleotides are assumed to be Morse potentials. We discuss the statistical mechanics of the DNA molecule in the configuration (5′... GGG... 3′; 3′... CCC... 5′), which may be copied by RNA polymerase into a messenger RNA (mRNA) strand (5′... CCC... 3′). This model suggests that replacements of C by 5-methylcytosine (5mC) may be a secondary effect in the inhibition of genetic expression, not interfering directly with the formation of an open state. An experimental test is suggested. The implications of this result are discussed for a related system, in which the enzyme methylase is known to methylate almost exclusively those Cs that are followed by Gs as a regulatory strategy employed by some eukaryotes.

1. INTRODUCTION

Recently two studies of nonlinear systems have considered models of molecular chains in which exact calculations were performed of the dynamical excitations in terms of equilibrium statistical mechanics (Dorby *et al.*, 1988; Peyrard and Bishop, 1989) (the latter paper is referred to below as PB). The partition function was evaluated rigorously in one dimension by means of transfer-operator techniques (Scalapino *et al.*, 1972). We believe that these approaches make it plausible to begin to consider the problems underlying the statistical mechanics of the control of gene expression at the level of gene transcription. Higher levels of regulation of gene expression, such as mRNA processing, translation, and posttranslational mechanisms, lie beyond our present understanding of statistical mechanics.

[1]International Centre for Theoretical Physics, Trieste, Italy.
[2]Permanent address: Departamento de Fisica, Universidad Simon Bolivar, Caracas 1080-A, Venezuela, Also at Institute Internacional de Estudios Avanzados, Caracas 1015-A, Venezuela.
[3]Permanent address: Instituto de Fisica Rosario (IFIR), 2000 Rosario, Argentina.

0020-7748/90/0800-0853$06.00/0

One aspect of the control of gene expression at the level of DNA transcription is the problem of cytosine-guanine (CG) methylation (Adams and Burdon, 1985; Razin *et al.*, 1984). It has not been possible to give a definitive answer to this question either from the point of view of biochemistry or molecular genetics. Methylation of a given base in a DNA molecule consists of certain modifications that are introduced into chromatin by the addition of a methyl ($-CH_3$) group at the carbon-5 position of cytosine that is immediately followed by guanine (CG methylation). This strategy is observed to yield an inverse proportionality between the extent of methylation of a given base and its activity in transcription (Felsenfeld and McGhee, 1982).

The main insight provided by studies of DNA sequences rich in unmethylated dinucleotides CG is that methylation may be a *secondary* event following primary inactivation by other mechanisms, serving therefore to imprint inactivity (Bird, 1986). In invertebrates, on the other hand, the strategy may be different: The nematode *Caenorhabditis elegans* is reported to have no 5mCs in its DNA. One possible way to follow up the consequences of methylation is in terms of molecular genetics (Chela-Flores, 1987).

The purpose of the present work is to present an alternative approach, namely, we propose to study the problem of the control of genetic expression in terms of statistical mechanics. Thus, in Section 2 we review the essential steps in the work of PB in order to present, in Section 3, a simple DNA model in which we allow the possibility of having both purine as well as pyrimidine nucleotides. We show in this section how the partition function is calculated in the model, allowing us to write, in equation (3.11), an analytic expression for the free energy, as well as an exact expression for the mean stretching of the CG bonds. In Section 4 we show that the Hamiltonian of the model admits small-amplitude motions. The corresponding dispersion relations are obtained, and the acoustic and optic branches are shown to lie within given bounds. Finally, in Section 5 we discuss how the bound on the optic phonons suggests that in DNA helices where methylation does occur, it is only a secondary process in the inhibition of genetic expression. We conclude with some comments on the possible further problems that may be envisaged in the statistical mechanics approach to molecular genetics.

2. THE NONLINEAR MODEL FOR DNA DENATURATION

Recently the process of DNA transcription has been studied by PB in the framework of statistical mechanics. The denaturation of DNA was studied in terms of soliton propagation that might account for a pulse of local denaturation moving along the helix at some steady rate.

Nucleotide pairs linking up the two DNA strands are assumed to interact by means of hydrogen bonds represented by Morse potentials $V_M(x_n)$ given by

$$V_M(w_n) = D(e^{-aw_n} - 1)^2 \tag{2.1}$$

In this notation a and D are constants and w_n denotes the out-of-phase motion of the bases:

$$w_n = (u_n - v_n) \tag{2.2}$$

The degrees of freedom u_n and v_n correspond to the displacement of the bases from the equilibrium positions along the direction of the hydrogen bonds connecting the nucleotide pairs. If m denotes the mass of the single nucleotide assumed in this particular model, then the complete Hamiltonian may be written as

$$H_{PB} = \sum_n \{\tfrac{1}{2}m(\dot{u}_n^2 + \dot{v}_n^2) + \tfrac{1}{2}k[(u_n - u_{n-1})^2 + (v_n - v_{n-1})^2] + V_M(u_n - v_n)\} \tag{2.3}$$

where k denotes the interaction strength for the harmonic potentials that are assumed to represent the stacking interactions; this contribution to the potential energy arises from a combined effect of hydrophobic interactions as well as electronic interactions between stacked bases (...). The various molecular differences between purines (G, A) and pyrimidines (C, T) are not taken into account (Lehninger, 1982).

By considering the equations of motion for the out-of-phase displacements w_n, we have assumed that the enzymatic activity of RNA polymerase (in bacterial cells, since in eukaryotes more than one RNA polymerase is needed) may be understood in terms of an open state propagating along the DNA molecule as a soliton wave (Englander *et al.*, 1980).

An important aspect of this DNA model is that the mean stretching of the hydrogen bonds may be studied as a function of temperature. It is very pleasing that the increments of the mean value of the $\langle y \rangle$ parameter as a function of temperature is essentially in agreement with the hyperchromic effect well known in biochemistry, i.e., unstacking of base pairs occurs with rising T, resulting in increased ultraviolet absorption (Stryer, 1988).

3. A PURINE–PYRIMIDINE MODEL OF DNA

In order to be able to discuss the effects of methylation on DNA, we consider the Hamiltonian for the model to be the following:

$$H = \sum_n \{\tfrac{1}{2}[(m\dot{u}_n^2 + M\dot{v}_n^2) + k_C(u_n - u_{n-1})^2 + k_G(v_n - v_{n-1})^2] + V_M(u_n - v_n)\} \tag{3.1}$$

In this notation m denotes the C mass and M the G mass; u_n and v_n denote, respectively, the C and G transverse displacements at site n of the strands of the DNA molecule. Finally, as in Section 2, V_M denotes the Morse potential that represents the hydrogen bond between the nucleotide pairs. This idealized model of the double helix with purines and pyrimidines is illustrated in Figure 1.

For a chain consisting of N purines (G) on a strand paired to N pyrimidines on the opposite strand, we are able to write the classical partition function in terms of the Hamiltonian (3.1), following the work of Dorby *et al.* (1988):

$$Z = Z_{KE} Z_U Z_W \tag{3.2}$$

where, for a chain of N sites, we have

$$Z_{KE} = [2\pi k_B T(mM)^{1/2}/\hbar^2]^N \tag{3.3}$$

$$Z_U = [2\pi k_B T/(k_C + k_G)]^{N/2} \prod_{n=1}^{N-1} |2\sin(n\pi/N)| \tag{3.4}$$

$$Z_W = \int_{-\infty}^{\infty} d^N w \exp\left\{-\beta \sum_{i=1}^{N} [\tfrac{1}{2}\tilde{k}(w_i - w_{i-1})^2 + V_M(w_i)]\right\} \tag{3.5}$$

where we have used the notation $\int d^N w = \int dw_1 \cdots dw_N$, and where

$$\tilde{k} = k_C k_G/(k_C + k_G) \tag{3.6}$$

Here the variables k_C and k_G are, respectively, the stacking interaction coupling constants for the nucleotides labeled by the subscript (see Figure 1).

The Z_W factor of the partition function coincides with the Z_Y factor of PB, provided that k and a of Z_Y are replaced by $\tilde{k}$ and $a/\sqrt{2}$, respectively. Z_W has been evaluated by means of the transfer integral operator technique. In the thermodynamic (large-N) limit, the calculation of PB yields

$$Z_W = e^{-N\beta E_0} \tag{3.7}$$

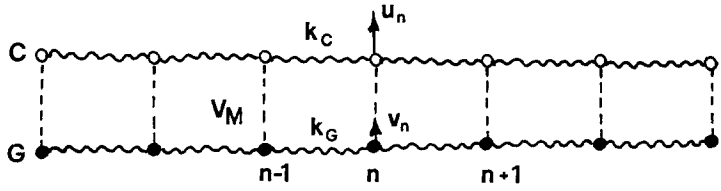

Fig. 1. An idealized DNA molecule with pairs of unequal nucleotides (one purine and one pyrimidine). The open circles denote cytosine bases, and the filled circles denote guanine bases.

where E_0 denotes the ground-state energy of the Schrödinger-like equation associated with the continuum limit of the transfer operator. In the present case we have

$$E_0 = [\ln(\beta\tilde{k}/2\pi)]/2\beta + a(D/2\tilde{k})^{1/2}/\beta - a^2/8\beta^2\tilde{k} \tag{3.8}$$

The corresponding eigenfunction is

$$\phi_0(w) = [a^{1/2}(2d)^{d-1/2}/\Gamma^{1/2}(2d-1)] \exp[\tfrac{1}{2}aw - d(aw + e^{-aw})] \tag{3.9}$$

where the d parameter is given by

$$d = (2\tilde{k}D)^{1/2}/ak_BT \tag{3.10}$$

It has been previously remarked that the condition for the existence of a discrete eigenvalue spectrum is $d > 1/2$, a result which follows from equation (3.9).

From the above results it follows that the free energy density per nucleotide pair is given by

$$\begin{aligned} f &= -(\ln Z_{KE}Z_U + \ln Z_W)/\beta N \\ &= k_BT\{\ln[(\hbar/k_BT)^2(k_Ck_G/mM)^{1/2}] + a(D/2\tilde{k})^{1/2} \\ &\quad - a^2(k_BT)^2/8\tilde{k}\} \end{aligned} \tag{3.11}$$

The expression for the free energy illustrates how the calculation of the partition function in the purine–pyrimidine model for the DNA macromolecule yields essentially analytic results for the macroscopic thermodynamic quantities. This is an important result of PB, which is not lost when we allow the bases to differ.

A result which is particularly interesting for the analysis of mass or modifications in the coupling constant of the stacking interactions $\tilde{k}$ is the expression for the mean stretching $\langle w \rangle$ of the CG bonds. The expression for $\langle w \rangle$ in the present purine–pyrimidine model is also given by:

$$\langle w \rangle = \int \phi_0^2(w)\, w\, dw \tag{3.12}$$

Since $\phi_0(w)$ does not depend on mass, then the mean value $\langle w \rangle$ is also independent of mass. This is a general result for any classical displacement average in a system with a Hamiltonian which does not contain terms mixing coordinates with momenta. It therefore follows that methylation effects on $\langle w \rangle$, if any, can only arise on the average through changes in the stacking interaction coupling constants k_C [alternatively, the same applies on the constant $\tilde{k}$; cf. equation (3.6)]. Methylation effects on the mean value $\langle w \rangle$

due to changes in the Morse potential parameters a or D seem unlikely, since the Morse potential represents the hydrogen bond in a nucleotide pair and is therefore unlikely to be affected by such an enzymatic process. In Section 5 we shall return to the relative contributions of the interaction parameters to the process of methylation.

4. DYNAMICAL EXCITATIONS

The purine–pyrimidine model of DNA developed in Section 3 by means of the Hamiltonian (3.1) admits small-amplitude motions (phonons). In order to show how to obtain the dynamical excitations, we first restrict our attention to a linearized (L) form of the Morse potential V_M:

$$V_M^{(L)} = Da^2 w^2 \tag{4.1}$$

while the total potential energy in the harmonic approximation is given by

$$\Phi(u, v) = \tfrac{1}{2} \sum_n [k_C(u_n - u_{n-1})^2 + k_G(v_n - v_{n-1})^2 + 2a^2 D(u_n - v_n)^2] \tag{4.2}$$

Therefore, the equations of motion are given by

$$m\ddot{u}_n = k_C(u_{n+1} + u_{n-1} - 2u_n) - 2a^2 D(u_n - v_n) \tag{4.3}$$

$$M\ddot{v}_n = k_G(v_{n+1} + v_{n-1} - 2v_n) + 2a^2 D(u_n - v_n) \tag{4.4}$$

These equations may be solved in the usual manner. We find that the dispersion relations are given by

$$\omega_\pm^2(q) = \omega_0^2/2 + 2\Delta_+ s_q^2 \pm [(\omega_0^2/2)^2 + (2\Delta_- s_q^2)^2 + 2\omega_0^2 \Delta_- \alpha s_q^2]^{1/2} \tag{4.5}$$

where

$$\Delta_\pm = k_C/m \pm k_G/M \tag{4.6}$$

and the ω_0^2 and α parameters are defined, respectively, as $2a^2D/\mu$ and $(M-m)/(M+m)$, where we have used the notation

$$1/\mu = 1/m + 1/M \tag{4.7}$$

$$s_q = \sin(qa/2) \tag{4.8}$$

In equation (4.8), q denotes the modulus of the wavevector.

We remark that the acoustic branch is given by the function $\omega_-(q)$ $[\omega_-(0) = 0]$, while the optic frequency at $q = 0$ is given by

$$\omega_+^2(0) = \omega_0^2 \tag{4.9}$$

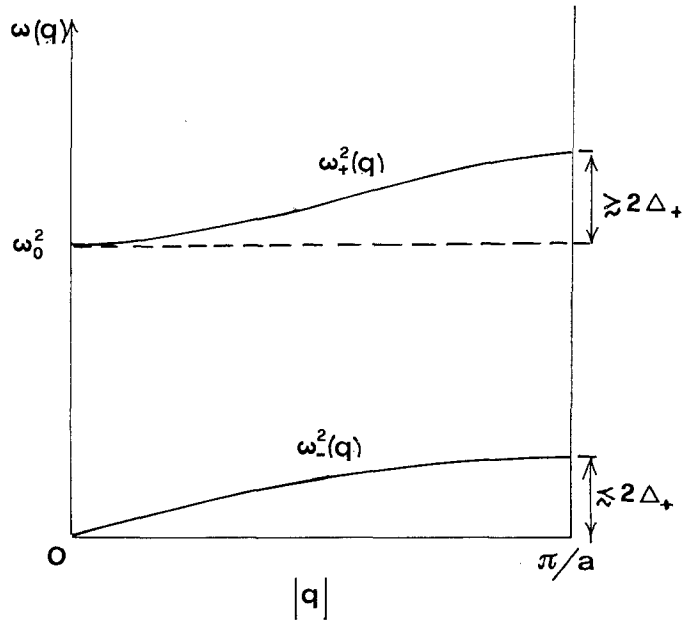

Fig. 2. The phonon dispersion relation for the DNA model described in Section 3. The sequence of nucleotide pairs is shown in Figure 1. The $\Delta_{\pm}$ parameters are defined in the text [cf. equation (4.6)]. The Δ_{+} parameter satisfies the inequality $\Delta_{+} \ll \omega_0^2$.

We assume that the strength of the hydrogen bond is significantly higher than the stacking interactions. This implies that $\omega_0^2 \gg \Delta_{\pm}$ [cf. Section 5, remark (i)]. At $q = \pi/a$, we find the following bounds for the small-amplitude excitations:

$$2(k_C/m + k_G/M) < \omega_+^2(\pi/a) - \omega_0^2 < 4k_C/m \tag{4.10}$$

$$4k_G/M < \omega_-^2(\pi/a) < 2(k_C/m + k_G/M) \tag{4.11}$$

These results are shown schematically in Figure 2. We may conclude that the optic branch (i.e., the stretching modes of the CG bonds) is lowered with increasing m. This tendency is slightly more enhanced near $q = \pi/a$. On the other hand, the acoustic branch is lowered slightly only near $q = \pi/a$.

5. DISCUSSION AND CONCLUSIONS

From the work in Sections 3 and 4, we can make several qualitative remarks that are relevant from the point of view of biochemistry:

(i) In Figure 2 the optical and acoustic branches may be seen to be very flat due to the weak stacking interactions (k_C, k_G) that have been

assumed in the model, as compared with the coupling constant D of the Morse potential (a numerical estimate of this variable is due to PB).

(ii) As the cytosine mass increases (for instance, by replacing C for 5mC), the optical branch is lowered itself [cf. equation (4.9)].

(iii) As the energy of the optical modes decrease (a phenomenon corresponding to the stretching modes of the CG bonds), the RNA polymerase in the case of prokaryotes, and the RNA polymerases in the case of eukaryotes, can transfer energy more easily to the double helix, so as to break the CG bonds locally, thereby contributing toward the creation of the open state that allows the mRNA precursor to be synthesized.

In spite of the fact that the enzyme methylase cannot induce metylation on the specific DNA macromolecule that we have considered in Section 3, it may be nevertheless significant that changes of the C base for 5mC, which imply an increment in the value of the mass m, have been shown (within the limitations of the model) not to imply a significant impediment to the RNA polymerase activity [cf. comment (iii) above).

On the other hand, our theoretical work suggests an experiment that may deserve attention: The expression (3.12) for the mean stretching of the CG bonds was shown by PB (and the same result holds for the present purine–pyrimidine model) to increase very rapidly in a narrow range around a certain characteristic temperature corresponding to the melting temperature for denaturation T_m. Therefore, the threshold for energy localization (the soliton mode), and hence the condition favorable for the formation of an open state, may be reached for values well above the normal T_m. However, from the above-mentioned work, it follows that T is a very sensitive function of the k parameter. For steric reasons k_C (and in turn k) may be affected by the methylation process. Thus, a possible mechanism for understanding methylation as a primary cause for the inhibition of gene expression can in principle be that methylation may raise T_m to values well above homeostasis. This effect, if verified experimentally, could explain why vertebrates display the inverse proportionality between methylation and normal gene expression.

It should be underlined that *in vivo* methylation cannot occur on a strand of G nucleotides bonded on a strand of C nucleotides. For methylation to occur, it is necessary that the C bases be followed by G bases.

Thus, further research seems to be necessary in the study of the statistical mechanics of the DNA molecule described by the configuration

$$(5'...CGCG...3'; 3'...GCGC...5') \tag{5.1}$$

This problem seems particularly relevant, since the configuration (5.1) consists of a sequence which is base paired to the same sequence, but in the opposite orientation on the opposite strand of the DNA macromolecule.

This configuration therefore allows the physiological role played by methylation to be inherited directly by a templating process (Alberts *et al.*, 1988).

However, if we assume that the stacking forces in the configuration (5.1) do not significantly change with respect to the configuration discussed in Section 3,

$$(5'...GGG...3'; 3'...CCC...5') \tag{5.2}$$

since these interactions only involve hydrophobic and Coulombic effects, then we may infer that the main conclusion of this work still holds.

The role played by disorder in the problem of methylation lies beyond the scope of the present work, but it does not seem to be beyond present computational capabilities to introduce disordered DNA molecules coding for real proteins, as, for instance, the DNA in the configuration

$$(5'...CGAATGC...3'; 3'...GCTTACC...5') \tag{5.3}$$

Such molecules could be studied from the same point of view of this work, so as to determine the relationship between random methylation yielding, for instance, the configuration

$$(5'...(5mC)GAATGG...3'; 3'...GCTTACC...5') \tag{5.4}$$

and the inhibition of gene expression.

Finally, it would seem reasonable to extend this work to include quantitative calculations such as those performed by Prohofsky and coworkers in the transcription of DNA in the context of the melting of the double helix (Prohofsky *et al.*, 1979; Prohofsky, 1988).

NOTE ADDED IN PROOF

One justification of the hypothesis (p. 859) on the relative strength of the hydrogen bonds and the stacking interactions (SI) is that considerable relaxation of chromatin packing has to occur prior to transcription due to various factors (including SI), without breaking hydrogen-bonded nucleotides, in order to allow some previously silenced enzymatic processes to occur. This may be illustrated by the reactivation of the Barr body in aged female mammals (Wareham *et al.*, 1987).

ACKNOWLEDGMENTS

One of the authors (J.C.F.) would like to thank the Instituto Internacional de Estudios Avanzados for a travel grant and partial financial support, and to the International Centre for Theoretical Physics (ICTP) for partial support. R.L.M. would like to thank ICTP for full support during

the course of this research. Both authors thank Prof. Abdus Salam, the International Atomic Energy Agency, and UNESCO for hospitality at ICTP, Trieste.

REFERENCES

Adams, R. L. P., and Burdon, R. H. (1985). *Molecular Biology of DNA Methylation*, Springer-Verlag, Berlin.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1988). *Molecular Biology of the Cell*, 2nd ed., Garland, New York, p. 583.

Bird, A. P. (1986). *Nature*, **321**, 209.

Chela-Flores, J. (1987). *Journal of Theoretical Biology*, **126**, 127.

Dorby, A. O., Migoni, R. L., and Cecatto, H. A. (1988). *Physical Review B*, **38**, 2801.

Englander, S. W., Kallenbach, N. R., Hegger, A. J., Krumhansl, J. A., and Kitwin, S. (1980). *Proceedings of the National Academy of Science USA*, **77**, 7222.

Felsenfeld, G., and McGhee, J. (1982). *Nature*, **296**, 602.

Lehninger, H. (1982). *Principles of Biochemistry*, Worth, New York, pp. 864–865.

Peyrard, M., and Bishop, A. R. (1989). *Physical Review Letters*, **62**, 2755.

Prohofsky, E. W. (1988). *Physical Review A*, **38**, 1538.

Prohofsky, E. W., Lu, K. C., Van Zandt, L. L., and Putnam, B. F. (1979). *Physics Letters*, **70A**, 492.

Razin, A., Cedar, H., and Riggs, A. D., eds. (1984). *DNA Methylation Biochemistry and Biological Significance*, Springer-Verlag, Berlin.

Scalapino, D. J., Sears, M., and Ferrell, R. A. (1972). *Physical Review B*, **6**, 3409.

Stryer, L. (1988). *Biochemistry*, 3rd ed., W. H. Freeman, San Francisco, California, p. 82.

Wareham, K. A. *et al.* (1987). *Nature*, **327**, 725–727.