Noncoding DNA Can Regulate Gene Transcription by it’s Base Pair’s Distribution.

U. Sandler & A. Wyler
Jerusalem College of Technology
P.O.B. 16031, Jerusalem 91160, Israel.

Abstract
In this paper we show that the gene transcription velocity can depend exponentially on the first moments of the base pair distribution in DNA. As the major fraction of the DNA of eucaryotes consists of noncoding sequences of base pairs and these determine the statistical characteristics of the base pair’s ordering, noncoding DNA plays an important role in the functioning of the genetic apparatus. In particular, it enables a new independent mechanism for the gradual regulation of gene expression in eucaryotes.

1 Introduction
There are many hypothesis regarding the role of noncoding DNA in the genetic apparatus of eucaryotes (see, for example, Li & Graur, 1991 and the bibliography there), but with the exception of the known role of the introns in a specific case of the immunoglobulin production and the possible advantage in the DNA evolution, the general role of the noncoding DNA is still unclear.

In this article we want to show that the gene transcription depends on the statistical properties of the distribution of base pairs. As the major fraction of DNA of eucaryotes consists of noncoding sequences of base pairs (Alberts at al, 1994), it is just this part of DNA that dominates the statistical sequence characteristics and is therefore important for the regulation of gene expression.

The process of gene transcription in vivo involves a number of chemical and physical interactions that determine the forming of a transcription bubble and its progressing through a chromatin template. For example, a
structural change in the nucleosome complex induced by histones acetylation (Wolffe, 1992, 1996) or deformation of a DNA template (Ausio, 1992) influences the process of the progressing transcription bubble. The real picture of gene functioning is so intricate, that it is unlikely, that the existing level of our knowledge allows to develop a complete theory of gene activity. Nevertheless, it is possible to clarify important details with the help of a few plausible assumptions:

1. Each step of gene transcription needs to overcome energy barriers.

2. The energy barriers depend on the local concentration of AT (or GC) base pairs. In chromatin DNA these energy barriers depend also on the state of the DNA and on the interaction between DNA templates and the other components of chromatin.

3. Thermal fluctuations contribute significantly to the overcoming of the energy barriers.

With these assumptions we show, that the velocity of the transcription depends exponentially on the first moments of the base pair distribution. This theoretical result can be verified by experiments.

In order to make this article more readable we have presented the reasoning and the main results in Section 2, while the physical arguments and mathematical details appear in Section 3. The discussion and conclusions are given in Section 4.

2 The time of gene transcription.

In the interphase cell the major fraction of DNA exists in the condensed chromatin form. In this form the genes cannot be transcribed, because the high density of chromatin prevents the forming and progressing of a transcription bubble (Wolffe, 1992, 1996). This means that the transcription process involves at least a local decondensation of DNA (Ausio, 1992; Wolffe, 1992). In fact, thermal fluctuations sometimes transfer a portion of DNA into the decondensed state, so that the chromatin globule is surrounded with various loops of decondensed DNA. As chromatin is a highly dynamic and quite inhomogeneous complex its varying constituents spend different times in the decondensed state.

The volume fraction of the decondensed part of a gene is proportional to the probability of the corresponding fluctuations (Landau & Lifshits, 1980):

\[ \rho_d \sim e^{N\Delta\mu/T} \]  

(1)
where $\rho_d$ is the volume fraction, $\Delta \mu$ is the difference of the free energies per base pair in the condensed and decondensed state, $N$ is the number of base pairs involved and $T$ is the temperature. $\Delta \mu$ depends on the chromatin state, including chromatin acetylation, the deformation of a DNA chain and interactions between DNA and the bonding proteins. Note, that because DNA in chromatin form is covered by the bonding proteins, the interaction between different pieces of DNA is mainly determined by the interaction between covering proteins. The characteristic time needed for the local decondensation - $\tau_d$ is reversely proportional to $\rho_d$, i.e.:

$$\tau_d \sim \rho_d^{-1} \sim e^{-\frac{N\Delta \mu}{T}}.$$  \hspace{1cm} (2)

The interaction between DNA and these proteins depends on the local ordering of the base pairs in those places where the proteins are bonded to a template. Because the proteins are bonded with DNA in many different parts the corresponding contribution to $\Delta \mu$ is determined by the statistical characteristics of the AT-GC pair distribution, rather than their exact ordering. A similar argument is valid for the contribution to $\Delta \mu$ from the interaction of DNA with other components of nucleoplasm. However, their dependence on the base pair distribution is likely to be relatively weak as compared to the one mentioned above. Although the contribution of the base pair distribution to $\Delta \mu$ is usually small, it influence on $\tau_d$ can be substantial, because $N$ is large. It is obvious that pieces of a gene not only interact with other parts of a given gene, but with the neighbor genes as well. This may be relevant to the dependence of gene activity on the gene position in the chromosome (the so-called "position effect" in gene expression (Wilson e.a. 1990)).

The shift of the transcription bubble includes the transformation of bonded proteins\(^1\) (Wolffe, 1992), dissociation ("melting") of several base pairs and the elongation of an RNA chain\(^2\). We assume here that each step of the progressing transcription bubble goes through an energy barrier. Then the characteristic time of each shift of the transcription bubble - $\tau_s$ is proportional to (see, for example, Landau & Lifshits, 1980):

$$\tau_s \sim e^{\frac{\delta F}{T}},$$  \hspace{1cm} (3)

where $\delta F$ is the minimal work per base pair needed for overcoming the energy barrier and $n$ is the number of base pairs involved in each shift. The

\(^1\)Note, that in spite of the fact that the connection between a DNA template and a histone core is considerably relaxed during the process of transcription, the DNA remains bonded with the histones.

\(^2\)The elongation of an RNA chain requires local melting of 15-18 base pairs (Kerpola & Kane, 1991)
contribution of RNA polymerase shifting to the time of gene transcription is determined by the sum of all $\tau_s$ and $\tau_d$ over the whole gene.

This time can be estimated as the sum of the times of decondensation and of shifting:

$$
\tau_{\text{transc}} \sim N \left( \tau_{0d} e^{N f_c} + \tau_{0m} e^{n f_s} \right),
$$

where $N$ is the total number of the base pairs in a gene. The functions $f_s, f_c$ depend on the statistical characteristics of base pair distribution - $x$ (see Section 3 for details). The coefficients $\tau_0$ depend foremost on the regulating proteins and are only weakly dependent on base pair distribution.

The most important feature of (4) is that $\tau_{\text{transc}}$ depends exponentially on the base pair distribution characteristics with large multipliers $N$ and $n$ in the exponent of the power. This means that the statistical distribution of AT-GC pairs in the genes and gene neighbors influences the gene transcription apparatus functioning.

3 Physical background and some mathematical details.

3.1 The time of decondensation.

The local volume fraction of the decondensed part of the gene - $\rho_i$ is proportional to the characteristic frequency - $\Omega_i$ of the local decondensation of the gene (De Gennes, 1979).

$$
\rho_i \sim \Omega_i \sim \exp - \frac{\Delta F_i}{T},
$$

where $\Delta F_i$ is the difference of the free energies in the condensed and the decondensed states for this part and $T$ is the temperature. For our purpose we will consider only that part of $\Delta F_i$ which is dependent on the distribution of AT-GC pairs.

The time of gene local decondensation - $\tau_i$ is inversely proportional to the characteristic frequency $\Omega_i$:

$$
\tau_i \sim \Omega_i^{-1} \sim \exp \frac{\Delta F_i}{T}.
$$

In order to get an idea of the explicit form of the dependence of $\tau_i$ on the base pair distribution we consider $\Delta F_i$ in the mean-field-approximation. In
this approximation the effective local free energy needed for decondensation can be written as:

\[
\mathcal{H} \sim \Delta F_0 + H_{\text{def}} + H_{\text{int}}.
\]  

(7)

Here \( H_{\text{def}} \) is the energy of deformation

\[
H_{\text{def}} \sim \int_0^{N_i} \frac{\gamma}{2} \left( \frac{\partial s}{\partial l} - \frac{\partial s_0}{\partial l} \right)^2 dl,
\]

(8)

\( N_i \) is the number of the base pairs contained in a given portion of chromatin, \( s (l) \) is a unit tangent vector along the decondensed DNA chain, while \( s_0 (l) \) corresponds to the condensed DNA chain deformation. \( \gamma = \gamma (x) \) is the effective local elasticity of the DNA chain (which depends on the local AT pair concentration - \( x(l) \)). In a first approximation the energy of interaction between the DNA bounded proteins is (De Gennes, 1979)

\[
H_{\text{int}} \sim -T \sum_{\alpha \beta} \chi_{\alpha \beta} \Phi_{\beta} \int_0^{N_i} c_{\alpha i} (l) \phi (l) dl
\]

where \( \chi_{\alpha \beta} \) is the effective parameter of the intermolecular interaction, \( c_{\alpha i} (l) \) is the local concentration of bonding proteins in the part of the gene considered (index \( \alpha \) designates a certain protein), \( \Phi_{\alpha} \) is the relative average volume fraction of the surrounding chromatin, and \( \phi (l) = \phi_{\text{con}} - \phi_{\text{dec}} \sim \phi_{\text{con}} \) is the difference of the DNA volume fraction in the condensed and decondensed state. Then, the free energy of a given portion of chromatin can be found from:

\[
\Delta F_i \sim \Delta F_{0i} - T \ln Z_i,
\]

\( Z_i = \int \int \exp \left( -\frac{\mathcal{H} \{ s \}}{T} \right) Ds \)

(9)

where \( Z \) is the path integral over all possible shapes of the DNA chain under the constraint:

\[
\left| \int_0^{N_i} s_0 (l) dl \right| \ll \left| \int_0^{N_i} s (l) dl \right| = \beta N_i
\]

(10)

where \( \beta \) is some constant. As \( x(l) \) varies much faster than \( s (l) \), in a first approximation we can replace \( \gamma (x) \) by its average value - \( \gamma_i = N_i^{-1} \int \gamma (l) dl \).

Since the effective Hamiltonian - \( \mathcal{H} \) is in a harmonic form, we can evaluate the integral (9) analytically. Then, the minimal energy needed for decondensation of a given portion of chromatin can be estimated as:

\[
\Delta F_i \sim N_i \left( \frac{\alpha \gamma_i}{T \sigma^2} + \sum_{\alpha \beta} \chi_{\beta \alpha} \Phi_{\beta} \langle c_{\alpha i} \rangle \phi \right).
\]

(11)
where
\[ \langle c_{\alpha i} \rangle = \frac{1}{N_i} \int_0^{N_i} c_{\alpha i} dl, \]
\[ a \sim \beta - 4N^{-2} \left( \int_0^N s_0 (l) dl \right)^2 > 0, \]
and \( \sigma \) is a constant that is proportional to a characteristic radius of the nucleosome. In a first approximation \( \gamma_i \) and \( \langle c_{\alpha i} \rangle \) can be expanded into power series as a function of the deviation of the local concentration of AT pairs from the average - \( x \):
\[ \gamma_i \sim \gamma_0 + \gamma_{1i} \Delta x_i + o \left( (\Delta x_i)^2 \right), \]
\[ \langle c_{\alpha i} \rangle = c_0 + c_{1i} \Delta x_i + o \left( (\Delta x_i)^2 \right), \] (12)
where \( \Delta x_i = x_i - \bar{x} \) and \( \gamma_0, \gamma_{1i}; c_0, c_{1i} \) are constants.

The whole time of gene decondensation is:
\[ \tau_{\text{dec}} \sim \sum_{\text{over gene}} \exp N_i \left( \frac{a \gamma_i}{T \sigma^2} + \sum_{\alpha \beta} \chi_{\beta \alpha} \Phi_{\beta} \langle c_{\alpha} (\Delta x_i) \rangle \phi_{\text{con}} \right). \] (13)

For \( N/N \gg 1 \) we can consider \( \Delta x \) as a random variable with a distribution function - \( P (\Delta x) \). Then, for \( N \gg 1 \) the sum (13) can be estimated as:
\[ \tau_{\text{dec}} \sim \frac{N \int P (\Delta x)}{N} \exp N \left( \frac{a \gamma (\Delta x)}{T \sigma^2} + \sum_{\alpha \beta} \chi_{\beta \alpha} \Phi_{\beta} \langle c_{\alpha} (\Delta x) \rangle \phi_{\text{con}} \right) d\Delta x \]
\[ \sim \frac{N P (\Delta x^*)}{N} \exp N \left( \frac{a \gamma (\Delta x^*)}{T \sigma^2} + \sum_{\alpha \beta} \chi_{\beta \alpha} \Phi_{\beta} \langle c_{\alpha} (\Delta x^*) \rangle \phi_{\text{con}} \right), \] (14)
where \( \Delta x^* \) is the saddle points of the expression under the integral. It is obvious that \( \langle c_{\alpha i} \rangle \) and \( \gamma_i \) depend on the composition of the given gene, as well as \( \Phi_{\alpha} \) on the composition of the other portions of DNA, which can interact with the gene. This means that \( \Delta F_i \) depends on the gene position in a DNA chain. Note, that for large \( N \), even relatively small decreases of \( \sigma \) (for example, due to deacetylation) may substantially suppress the process of chromatin decondensation.

The volume fraction of condensed chromatin can be reasonably estimated as \( \phi_{\text{con}} \sim \Phi_{\beta} \sim 0.1 - 0.3 \). The relative concentration of bonding proteins varies substantially but as the upper limit of \( \langle c \rangle_g \) we can take \( \langle c \rangle_g \leq 0.1 \). Because \( \chi_{\beta \alpha} \) is of the order of the Flory parameter, which is about 1, we find that for \( N \gg 10^2 \) the exponent in (14) will be large so that the base pair distribution influences \( \tau_{\text{dec}} \).
3.2 The time of a bubble moving.

Consider first the transcription of DNA without bonded proteins (this case corresponds to the transcription of a free DNA \textit{in vitro}). Let us assume that the shift of a transcription bubble requires the dissociation of \( n \) pairs. The quantity \( \delta F \) from (3) can be estimated as

\[
\delta F \simeq \varepsilon + nx \delta \mu, \tag{15}
\]

where \( \Delta \mu_{AT,GC} \) is the energy barrier between the two states of an AT or GC pair inside a DNA chain, \( x \) is the local concentration of GC-pairs and \( \varepsilon \) corresponds to the base pair independent part of \( \delta F \).

\( \tau_{\text{mov}} \) is the sum of the \( \tau_s \) from (3) over the whole gene:

\[
\tau_{\text{mov}} \sim \sum_{\text{over gene}} \tau_s. \tag{16}
\]

As above, if we consider \( x \) as a random variable with the distribution function \( P(x) \), (16) is equivalent to averaging of \( \tau_s \):

\[
\tau_{\text{mov}} \sim \sum_{\text{over gene}} e^{\frac{\delta \mu_{nx}}{T}} \simeq N \tau_0 \int e^{\frac{\delta \mu_{nx}}{T}} P(x) dx = N \tau_0 \langle e^{n \nu x} \rangle, \tag{17}
\]

where \( \nu = \frac{\delta \mu}{T} \) and \( \tau_0 \) is a constant.

If \( n \nu \gg 1 \) we can estimate the asymptote of the average \( \langle e^{n \nu x} \rangle \). Let us make a reasonable assumption that the distribution \( P(x) \) has a maximum at the total concentration of the AT pairs in gene - \( \tau \) : \( P(x) = \rho \left( \frac{(x-\bar{x})^2}{2\Delta^2} \right) \).

Then, the saddle point - \( x_* \) of the integral in (17) is found from:

\[
\frac{1}{P(x_*)} \frac{\partial P(x_*)}{\partial x_*} + n \nu \sim -x_* - \bar{x} \frac{\Delta^2}{2} + n \nu = 0, \tag{18}
\]

where \( \Delta \) is the characteristic width of the maximum of \( P(x) \) that can be estimated as \( \Delta^2 \sim \langle x^2 \rangle - \langle x \rangle^2 = \bar{x}^2 \). (\( \bar{x}^2 \) is the average root-mean-square deviation of the local concentration.)

If \( x_* \sim \bar{x} + \Delta^2 n \nu < 1 \) and \( \Delta \ll \min(x_*; 1-x_*) \), the average of \( \exp(n \nu x) \) can be estimated as

\[
\langle e^{n \nu x} \rangle \sim e^{n \nu (\bar{x} + n \nu \Delta^2)} \rho \left( \frac{n^2 \nu^2 \Delta^2}{2} \right). \tag{19}
\]

\( \delta F \) corresponds to the forward shift of a transcription bubble. In fact, there is a possibility for a backward shift of the bubble, but because of an asymmetry of a transcription bubble and therefor the corresponding energy barriers: \( \delta F_{\text{back}} \gg \delta F \), back moving is suppressed.
In the case \( x_\ast > 1 \), this average will be:
\[
\langle e^{n\nu x} \rangle \sim e^{n\nu} \rho \left( \frac{(1-x)^2}{2\Delta^2} \right).
\](20)

so, in this case, \( \tau_{mov} \) depends relatively weak on the base pair distribution.

Consider now the moving of a transcription bubble along a gene in a eucaryote nucleus, where the DNA remains bonded with the histones and with some other proteins. In this situation each shift of a transcription bubble includes the transformation of bounded proteins and the connections with the DNA chain. The bonded proteins change also the values of \( \Delta \mu_{AT} \) and \( \Delta \mu_{GC} \). Thus, instead of (15) we must write:
\[
\delta F_i \simeq \bar{\varepsilon}_i + n x_i \delta \bar{\mu}_i,
\](21)
where the quantity \( \bar{\varepsilon}_i \) includes also the energy of the above mentioned transformation. The index \( i \) denotes the position of the transcription bubble along the DNA chain. Equation (16) is then transformed into:
\[
\tau_{mov} \sim \sum_i \exp \left( \frac{\delta F_i}{T} \right) = \sum_i e^{\phi_i} \exp \left( n \nu_i x_i \right),
\]
\[\phi_i = \bar{\varepsilon}_i \frac{T}{\nu_i}, \quad \nu_i = \delta \bar{\mu}_i \cdot \]
(22)

Let us assume that a given gene interacts with \( m_a \) proteins of type \( a \), \( m_b \) of type \( b \) and so on and that it contains \( m_0 \) protein free places. Then we can rewrite (22) as:
\[
\tau_{mov} \sim \sum_k^{m_a} \sum_l^{m_0} e^{\phi_l} \exp \left( n \nu_l^0 x_{l+k} \right)
+ \sum_k^{m_a} \sum_l^{m_0} e^{\phi_l} \exp \left( n \nu_l^a x_{l+k} \right) + ...
\]
(23)

If \( m_a \) is large enough we can estimate the sums over \( k \) using the same reasons as for (17)\(^4\):
\[
\sum_k^{m_a} \exp \left( n \nu_l^a x_{l+k} \right) \sim \int e^{n \nu_l^a x} P(x) \, dx = m_a \langle \exp \left( n \nu_l^a x \right) \rangle.
\]
(24)

\(^4\)For large \( m_a \) the dependence of \( P_i(x) \) on \( i \) can be neglected and we can write \( P_i(x) \approx P(x) \).
Thus, for $\tau_{\text{mov}}$ we get:

$$
\tau_{\text{mov}} \sim \sum_{\alpha=0}^{M} m_{\alpha} \sum_{l} e^{\delta \phi_{\alpha}} \langle \exp (n\nu_{\alpha}^{\alpha} x) \rangle \sim \mathcal{N} \sum_{\alpha} \langle c_{\alpha} \rangle \tau_{\alpha} \langle \exp (n\nu_{\alpha}^{\alpha} x) \rangle ,
$$

where $\langle c_{0} \rangle$ corresponds to the concentration of protein free places. The averages $\langle \exp (n\nu_{\alpha}^{\alpha} x) \rangle$ can be estimated as above (see (19) and (20)). Since the concentration of the histones is much larger than of the other proteins, the corresponding term in the sum (25) is dominant and we can write qualitatively:

$$
\tau_{\text{mov}} \sim \mathcal{N} \langle c_{H} \rangle \tau_{H} \langle \exp (n\nu_{H}^{\alpha} x) \rangle .
$$

The expressions (17) and (26) are analogous, thus in both situations the base pair distribution influences $\tau_{\text{mov}}$ qualitatively in the same way.

The upper limit for $n$ is obvious: $n \sim 15 \div 18$, while the lower limit depends on the model of the elongation (Kerpola & Kane 1991). As a reasonable estimate for this limit one can take $n \sim 3 \div 10$. It should be noted, that the values $n$ in (17) and (26) may be different. The order of magnitude of $\delta \mu_{i}^{H}/T$ depends on many factors, but can be roughly estimated from the shift of the melting temperature for AT and GC enriched DNA fragments (Poland & Scherada, 1970). This gives us $\nu_{H}^{*} \sim 1$. Thus the quantity $n\nu_{H}^{*}$ can be large enough, and so the base pair distribution moments considerably influence the transcription time.

4 Discussion and Concluding remarks.

At first we discuss our assumptions. The first assumption (see Introduction) seems reasonable, because in most biochemical processes in the cell there are energy barriers to overcome. Thermal fluctuation obviously contribute to this process but the question is how significant this contribution is. It has been shown here that if the local chromatin decondensation or the transcription bubble shifting are the bottle necks for the transcription process, the thermal fluctuations can lead to an exponential dependence of the transcription velocity on the moments of base pair distribution.

There are other situations where this dependence is much weaker. Consider, for example, the case where RNA elongation is the limiting factor for transcription. Denote as $\omega_{C}$, $\omega_{U}$ the velocities of the elongation per nucleotide for G,C and A,T nucleotides resp. Then the total time of elongation will be:

$$
\tau_{\text{transc}} \sim \tau_{\text{elong}} \sim \mathcal{N} \left\{ (1 - \bar{z})\omega_{U}^{-1} + \bar{z}\omega_{G}^{-1} \right\} .
$$
i.e. a polynomial (instead of exponential) dependence of the transcription velocity on the average gene composition.

If we compare equations (17) and (26) we see a non-trivial result that the consideration of chromatin DNA instead of free DNA leads only to a change in the meaning of the coefficients in the expression for the average velocity of the gene transcription.

In eucaryotes the major part of DNA consists of noncoding sequences, which in fact determine the statistical characteristics of the base pair ordering. Therefore the statistical distribution of base pairs in noncoding sequences of DNA has an effect on gene activity. Because the mutations and the recombination and transposition of the movable elements can continuously change the base pair distribution in noncoding DNA there is a new type of regulation of gene activity in eucaryotic cells.

Such mechanism of regulation can change gene activity without redesigning of the existing system of gene activators and the initiation and termination phases of gene transcription. This is very important for the adaptation of cells in a multicellular organism in the evolution process. The result can shed light on the fact that the world of eucaryotes is much more complicated than the world of procaryotes and why so much biological ”junk” of noncoding sequences remained in existence under pressure of selection. It follows also, that mutations of noncoding DNA which change the statistical features of the base pair’s distribution, influence the functioning of the genetic apparatus.

Experiments show, that $\tau_{\text{transc}}$ may vary three order of magnitude for different genes (Kraevsky, 1996). The proposed theory predicts the dependence of the velocity of transcription on gene composition, but for a direct verification of the theory, data about base pair distribution and the velocity of transcription for genes with different base pair composition are required.

References


