Hairpin Formation and Elongation of Biomolecules

Andrea Montanari[1] and Marc Mézard[2]

[1]*Scuola Normale Superiore and INFN–Sezione di Pisa, I-56100 Pisa, Italy*

[2]*Laboratoire de Physique Théorique et Modèles Statistiques, Université Paris Sud, Batiment 100, 91 405 Orsay Cedex, France*

(Received 12 June 2000)

We introduce a model of thermalized conformations in space of RNA—or single stranded DNA—molecules, which includes the possibility of hairpin formation. This model contains the usual secondary structure information, but extends it to the study of one element of the ternary structure, namely the end-to-end distance. The computed force-elongation characteristics are in good agreement with some recent measurements on single stranded DNA molecules.

DOI: 10.1103/PhysRevLett.86.2178 PACS numbers: 87.15.–v, 05.10.Gg, 61.41.+e, 87.14.Gg

Recent progress in the manipulation of single biomolecules is making gradually accessible a wealth of interesting physical information. One of the basic investigations concerns the force-elongation characteristics: its measurement in double stranded DNA (dsDNA) molecules has provided very interesting results in the last few years, going from a detailed characterization of the elastic properties of the molecules to the existence of new phases of dsDNA in various regimes of tension and overcoiling [1–11].

While the force-elongation characteristics of dsDNA is rather well understood, the corresponding knowledge on single stranded DNA (ssDNA) is poorer: although in some ionic conditions it may be characterized by a simple freely jointed chain (FJC) with elastic bonds [6], this description is not valid when one changes the ionic concentrations [12]. This discrepancy is probably due to the formation of secondary structures in the ssDNA molecule [12], which can bend back onto itself and form local helices where complementary bases A-T and G-C are paired, gaining an energy of several kT per pair.

The formation of secondary structures is a crucial step in the folding of single stranded nucleic acid polymers. Its importance stems from the rather large values of the binding energy involved in this formation, compared to the much smaller energy scale of the interaction between secondary structures which govern the final three-dimensional shape of the molecules (the ternary structure). As discussed recently [13–15], the formation of secondary structures in RNA (which is very similar to the one in ssDNA) provides a wonderful laboratory for detailed studies of some of the basic mechanisms at work in heteropolymer folding.

In this paper we modify and extend the previous studies on RNA or ssDNA secondary structures in order to include one simple aspect of the ternary structure, namely the thermal fluctuations of the end-to-end distance, and its dependence on the pulling force. Our model can be solved exactly using generating function techniques. It involves three parameters: the persistence length of the molecule, the elastic constant characteristic of bond stretching, and the pair binding energy. It predicts the existence of two phases. At low force the polymer is folded and its elongation per bond vanishes. At some critical force there is a second order phase transition: when the force is increased the polymer elongates, the fraction of paired bases decreases, and the behavior eventually approaches the one found in absence of pairing.

In the simplest approximation, the backbone of the polymer is described by a FJC with N elastic bonds. At thermal equilibrium, the probability distribution of a bond to be equal to the vector $\vec{r}$ is given by

$$\mu(\vec{r}) = C \exp\left(-\frac{(|\vec{r}| - b)^2}{2\ell^2}\right), \tag{1}$$

where b is the persistence length, ℓ is a length which characterizes the elasticity of the bond, and $\mu(\vec{r})$ is assumed to be normalized to 1. For RNA or ssDNA, one expects b to be of the order of a few times the distance between successive bases, and ℓ/b to be much smaller than one. The spatial conformation is thus described by the positions $\vec{r}_i$ ($i \in \{1, \ldots, N+1\}$) of the $N+1$ nodes which are the articulation points of the chain. The attraction between complementary bases creates an effective potential $\epsilon_{ij}(\vec{r}_i - \vec{r}_j)$ between nodes i and j (arbitrarily far away from each other along the backbone) which involves a short ranged attraction and a core repulsion. We perform a standard virial expansion of the partition function in terms of the quantities $f_{ij}(\vec{r}) = \exp[-\epsilon_{ij}(\vec{r})/kT] - 1$ which vanish for $|\vec{r}| > a$, where a is the range of the interaction. The secondary structure is characterized by the set of node pairs i, j such that $f_{ij} \neq 0$.

Our main approximation for describing the secondary structure is the standard one in which one keeps only the nested diagrams [13–19], which are defined as follows: (i) Each node can be paired to at most one other node. (ii) Two node pairs $i < j$ and $k < l$ (with, say, $i < k$) can coexist only if they are either independent ($i < j < k < l$) or nested ($i < k < l < j$). This condition neglects the formation of pseudoknots. This is thus the simplest approximation, one in which one adds to the basic elastic model (here, for instance, the FJC) the possibility

 0031-9007/01/86(10)/2178(4)$15.00

of formation of hairpins, consisting of helices, and helices within helices organized in a hierarchical way.

The hierarchical structure of the retained diagrams makes it possible to write a recursion relation for the partition function $Z_{j,i}(\vec{r})$ which describes the set of nodes $k \in \{i, i+1, \ldots, j\}$, with an end-to-end distance $\vec{r}_j - \vec{r}_i = \vec{r}$. The recursion is explained in Fig. 1 which shows that, when $j - i \geq 2$ (the last sum drops out when $j - i = 2$),

$$Z_{j,i}(\vec{r}) = \int d\vec{u}\, \mu(\vec{u}) Z_{j-1,i}(\vec{r} - \vec{u}) + f_{ji}(\vec{r}) \int d\vec{u}_1\, \mu(\vec{u}_1)\, d\vec{u}_2\, \mu(\vec{u}_2) Z_{j-1,i+1}(\vec{r} - \vec{u}_1 - \vec{u}_2) + \sum_{k=i+1}^{j-2} \int \prod_{m=1}^{3} d\vec{u}_m\, \mu(\vec{u}_m)\, d\vec{v}\, f_{jk}(\vec{v}) Z_{j-1,k+1}(\vec{v} - \vec{u}_2 - \vec{u}_3) Z_{k-1,i}(\vec{r} - \vec{u}_1 - \vec{v}). \quad (2)$$

This recursion relation provides the definition of our model for RNA or ssDNA folding. It modifies the recursions which have been written previously in the studies of RNA secondary structures [14] in two aspects. On the one hand, it includes the spatial structure, i.e., the positions of the nodes. Second, it uses the virial expansion in which the interaction term between i and j is given by $f_{ij}(\vec{r})$. This is needed in order to get back the usual FJC in the limit where the interaction potential ϵ vanishes.

As a first step in the study of this model, we investigate in the following the case where the interaction energy $\epsilon_{ij}(\vec{r})$ is independent of the pair i, j. This amounts to using an effective interaction, averaged over the several bases included within the persistence length b, in which the only effect of the sequence which is kept is the global concentration in the various base pairs. The effect of sequence heterogeneities, which is crucial for dynamical properties, is left for future studies.

In the homogeneous case the partition function $Z_{j,i}(\vec{r})$ depends only on $n = j - i$, and is denoted by $Z_n(\vec{r})$. We introduce the Fourier transform of the generating function of the Z_n's,

$$\Xi(\zeta, \vec{p}) = \int d\vec{r} \left(\sum_{n=0}^{\infty} Z_n(\vec{r}) \zeta^n \right) e^{i\vec{p}\cdot\vec{r}}, \quad (3)$$

which is expressed in terms of the Fourier transforms:

$$\sigma(\vec{p}) = \int d\vec{r}\, \mu(\vec{r}) e^{i\vec{p}\cdot\vec{r}}$$

$$f(\vec{p}) = \int d\vec{r} [\exp(-\beta\epsilon(\vec{r})) - 1] e^{i\vec{p}\cdot\vec{r}}. \quad (4)$$

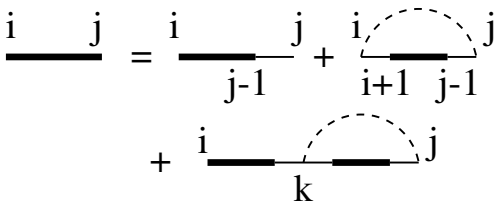

FIG. 1. Recursion relation for the partition function. A thin line denotes one bond, which in the elastic FJC is a vector chosen with probability (1). A dashed line between i and j corresponds to an interaction term $\exp[-\epsilon_{ij}(\vec{r})/kT] - 1$. The full line is the partition function which adds up the effect of all nested or independent interaction lines.

Using $Z_0(\vec{p}) \equiv 1$ and $Z_1(\vec{p}) \equiv \sigma(\vec{p})$ one derives from the recursion relation (2) the functional equation

$$\Xi(\zeta, \vec{p}) = \frac{1}{\zeta} \frac{\omega(\zeta, \vec{p})}{1 - \sigma(\vec{p})\omega(\zeta, \vec{p})}, \quad (5)$$

where the kernel ω satisfies the integral equation

$$\omega(\zeta, \vec{p}) = \zeta + \zeta^3 \int \frac{d^3q}{(2\pi)^3} f(\vec{p} - \vec{q}) \sigma(\vec{q})^2 \Xi(\zeta, \vec{q}). \quad (6)$$

The force-elongation characteristics for a chain with N bonds can be deduced from the partition function in the presence of a force:

$$Z_N^{\vec{F}} = \int d\vec{r}\, Z_N(\vec{r}) e^{\beta\vec{F}\cdot\vec{r}}. \quad (7)$$

Its generating function is nothing but $\Xi(\zeta, \vec{p}_F)$, where $\vec{p}_F$ is an imaginary momentum given by $\vec{p}_F = (0, 0, -i\beta F)$ for a force F pulling in the third direction. For a long chain, $N \gg 1$, one expects a partition function behaving as $Z_N^{\vec{F}} \sim A \exp[-\beta N \phi(F)]/N^{\alpha}$. The free energy per bond $\phi(F)$ determines the radius of convergence of the series defining the generating function $\Xi(\zeta, \vec{p}_F)$. It is thus equal to $\phi(F) = (1/\beta) \ln(\zeta^*)$, where ζ^* is the singularity of $\Xi(\zeta)$ which is the nearest to the origin. From the free energy per bond one deduces the elongation L along the axis of the force, $L = -N \partial\phi/\partial F$, as well as the average fraction of pairings n_p (defined as the number of pairings divided by N), $n_p = \partial \ln(\zeta^*)/\partial \ln(\gamma)$.

The integral equation (5) is easily solved in the case where the range of the interaction potential is small compared to b (this approximation is again valid when b is much larger than the interbase distance). One can then neglect the momentum dependence of f and substitute $f(\vec{p})$ by the constant γb^3, where γ is a dimensionless number characteristic of the strength of the pairing and defined by $\gamma = f(\vec{0})/b^3 = \int d\vec{r}/b^3 [\exp(-\beta\epsilon(\vec{r})) - 1]$. The kernel ω is then momentum independent. The relation (6) between ζ and ω can be written as $\omega = \zeta + \zeta^2 A(\omega)$, where the function $A(\omega)$ is monotonously increasing and such that $A'(\omega = 1) = \infty$. One can then show that $\omega(\zeta)$ has a second order branching point at $\zeta_{\rm bp}$ and is analytic for $|\zeta| < \zeta_{\rm bp}$, where $\zeta_{\rm bp}$ is the maximum of the function $[-1 + \sqrt{1 + 4\omega A(\omega)}]/2A(\omega)$.

The singularities of Ξ which control the large n behavior of Z_n are the branching point of $\omega(\zeta)$ at $\zeta_{\rm bp}$ and the pole at $\zeta_{\rm p}(\vec{p})$ determined by the vanishing of the denominator of Eq. (5), when the momentum is equal to $\vec{p}_F$: $\omega(\zeta_{\rm p})\sigma(\vec{p}_F) = 1$. For purely elastic bonds with $\ell \ll b$, one finds $\sigma(\vec{p}) \simeq [\sin(pb)/pb]\exp(-p^2\ell^2/2)$, and the pole is located at

$$\omega(\zeta_{\rm p}) = \frac{\beta F b}{\sinh(\beta F b)}\, e^{-\beta^2 F^2 \ell^2/2}. \tag{8}$$

Each of the two singularities is associated with one phase of the model. As far as we neglect the momentum dependence of ω (i.e., for small interaction radius), the position of the branching point does not depend upon the force. Therefore it corresponds to a folded phase (which we call "hairpinned phase"). The free energy per bond is given by $\phi(F) = (1/\beta)\log\zeta_{\rm bp}$. The length of the polymer is of order N^0 in the long chain ($N \to \infty$) limit. A fraction $n_{\rm p}$ of nodes is paired with $n_{\rm p}$ independent on the applied force. The chain is bent in a few, i.e., $O(N^0)$, hairpins, each one involving $O(N)$ bonds. The "elongated" phase corresponds to the pole singularity. The free energy $\phi(F) = (1/\beta)\log\zeta_{\rm p}(\vec{p}_F)$ is force dependent and the elongation is extensive (proportional to N). This can be written as $L(F) = n_{\rm free}(F)L_{\rm FJC}(F)$ where $L_{\rm FJC}(F)$ is the elongation without interaction (i.e., in the case $\gamma = 0$) and $n_{\rm free}(F)$ the fraction of nodes which do not belong to any hairpin. The fraction of pairings rapidly decreases with the applied force. The number of hairpins is $O(N)$.

In our model there exists a second order phase transition between the hairpinned phase at low force and the elongated phase at high force. This phase transition is a robust feature of the model which does not depend on the details of the interaction potential and of the bond stretching potential: the branched point singularity, associated with the hairpinned phase, is present as soon as $\sigma(\vec{p}) \sim 1 - \kappa\vec{p}^2$ for small $|\vec{p}|$, which is the generic situation. The pole singularity, associated with the elongated phase, is always present. The boundary between the two phases occurs at a critical force $F_c(\gamma)$ which increases monotonically with γ. Slightly above the threshold the elongation grows linearly with the force $L(F) \propto F - F_c(\gamma)$. The asymptotic behaviors of the dimensionless critical force are

$$\beta b F_c(\gamma) \sim \frac{1}{4}\log(\gamma) \qquad \text{for } 1 \ll \gamma \ll e^{b^2/\ell^2},$$

$$\beta b F_c(\gamma) \sim \frac{\gamma}{8\pi\kappa^2} \qquad \text{for } \gamma \ll 1\,. \tag{9}$$

Notice that the linear dependence of F_c at small γ is a prediction which is independent of the detailed form of the bond probability distribution (1).

Equations (6) and (8) can be easily solved numerically. We compared our theoretical predictions with the experimental data presented in Ref. [12] on the ln force versus elongation characteristic for a charomid-ssDNA at room temperature under different salinity conditions. Using the elastic model for bond stretching (1), our three fitting parameters are the persistence length b, the elasticity ℓ, and the interaction parameter γ.

As shown in Fig. 2, we obtain a good agreement with the experimental curve at the highest salt concentration (10 mM PB, 5 mM Mg). The small elongation region ($L/L_0 < 0.1$) of the experiment was not considered since the interactions of the molecule with the glass plate cannot be ignored (this forbids a study of the critical force region with the present data). The number of bonds was fixed as in [12] such that $Nb = 1.6875L_0$, where L_0 is the crystallographic length of the double stranded DNA. A least squares fit yields the following results: $b = 19.2$ Å, $\gamma = 1.89$, and $\ell = 1.01$ Å. The orders of magnitudes of the various parameters are correct. The persistence length is of the order of 3 times the interbase distance b_0 (our approximation of a large value of $b > b_0$ is marginally self-consistent and should be improved upon in the future). The value of ℓ, when expressed in terms of the enthalpic elasticity S as in [12], corresponds to $S = b/(\beta\ell^2) \simeq 1000pN$, typical of the values measured at higher forces [6,10,11]; the approximation $\ell \ll b$ is valid. The value of γ is characteristic of the strength of the interaction. For a potential well of width a and depth ϵ_0, one has $\gamma \sim (a/b)^3\exp(\beta\epsilon_0)$, which is compatible with some typical values such as $a \sim 4$ Å, $\epsilon_0 \sim 2.9kT$.

From our computation one can deduce the pairing fraction n_p in the conditions of the experiment. This is plotted in Fig. 3. It is clear that in the region of forces above $10pN$ there is no pairing. This is consistent with the measurements of Ref. [12] which showed that the characteristics

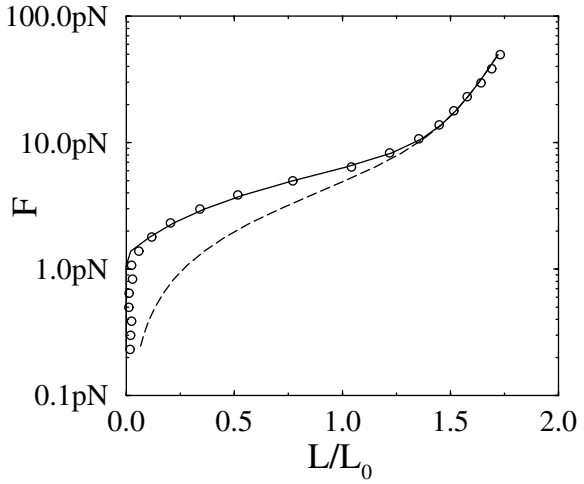

FIG. 2. Fit of the force-elongation characteristics of charomid ssDNA. The circles are the experimental data of Ref. [12]. The continuous line is the best fitting curve obtained with our model. The dashed curve is the FJC characteristics obtained by switching off the interaction. The difference between the two is due to the formation of hairpins.

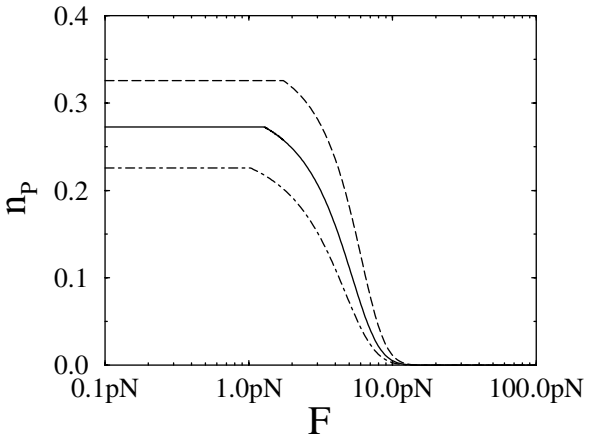

FIG. 3. The fraction of paired nodes in the secondary structure as a function of the external pulling force. The three curves refer to three different values of the interaction parameter γ. From top to bottom $\gamma = 3.9$, 1.9, and 1.0. The other parameters of the model correspond to the experimental situation: they are fixed as in the fit of Fig. 2.

of two different ssDNA's with different G-C concentration merge in that region. One should keep in mind that the two fitting parameters b and ℓ are basically fixed by this high force region where there is no pairing. The low force part is the one which fixes the binding parameter γ.

When the salinity is lowered, some new physical effects become relevant. The electrostatic interactions between the bases are less effectively screened, and probably the FJC is not a good model. One possibility to test it is to use, instead of the elastic FJC, the experimental force-elongation characteristics measured on a molecule exposed first to a chemical treatment (for instance, glyoxal) which decreases the ability of the bases to pair. Our model should then allow us to deduce from two experimental curves (one with glyoxal, the other without) the effect of the secondary structures [20].

In this paper we have introduced a solvable model of the structure of ssDNA or RNA molecules which includes, together with the secondary structure, one important element of its ternary structure. The model gives a general framework for including the effect of hairpin formation in the elongation properties of the molecules. When used with a simple FJC model for the polymer without hairpins, it is in good agreement with the experimental data at high ionic concentration. Several extensions of this study are natural. The description of data obtained at smaller ionic concentration requires one to go beyond the FJC approximation. Another natural extension of our study, also possible within this model, is to study the effects of the disorder in the sequence of bases.

It is a pleasure to thank D. Bensimon and V. Croquette for many useful discussions and for providing us with their data. M. M. is on leave from Laboratoire de Physique Théorique de l'Ecole Normale Supérieure.

[1] S. B. Smith, L. Finzi, and C. Bustamante, Science **258**, 1122 (1992).
[2] T. T. Perkins, S. R. Quake, D. E. Smith, and S. Chu, Science **264**, 8222 (1994).
[3] T. R. Strick, J.-F. Allemand, D. Bensimon, A. Bensimon, and V. Croquette, Science **271**, 1835 (1996).
[4] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 (1994); A. Vologodskii, Macromolecules **27**, 5623 (1994).
[5] P. Cluzel, A. Lebrun, C. Heller, R. Lavery, J. L. Viovy, D. Chatenay, and F. Caron, Science **271**, 792 (1996).
[6] S. B. Smith, Y. Cui, and C. Bustamante, Science **271**, 795 (1996).
[7] T. R. Strick, J.-F. Allemand, D. Bensimon, and V. Croquette, Biophys. J. **74**, 2016 (1998).
[8] T. R. Strick, V. Croquette, and D. Bensimon, Proc. Natl. Acad. Sci. U.S.A. **95**, 10 579 (1998).
[9] J.-F. Allemand, D. Bensimon, R. Lavery, and V. Croquette, Proc. Natl. Acad. Sci. U.S.A. **95**, 14 152 (1998).
[10] U. Bockelmann, B. Essevaz-Roulet, and F. Heslot, Phys. Rev. E **58**, 2386 (1998).
[11] M. Rief, H. Clausen-Schaumann, and H. E. Gaub, Nat. Struct. Biol. **6**, 346 (1999).
[12] B. Maier, D. Bensimon, and V. Croquette, Proc. Natl. Acad. Sci. U.S.A. **97**, 12 002 (2000).
[13] P. G. Higgs, J. Phys. I (France) **3**, 43 (1993); Phys. Rev. Lett. **76**, 704 (1996).
[14] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479 (1999).
[15] A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, Phys. Rev. Lett. **84**, 2026 (2000).
[16] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[17] M. Zuker and D. Sankoff, Bull. Math. Biol. **46**, 591 (1984).
[18] W. Fontana *et al.,* Biopolymers **33**, 1389 (1993).
[19] S.-J. Chen and K. A. Dill, J. Chem. Phys. **103**, 5802 (1995).
[20] M.-N. Dessinges, B. Maier, D. Bensimon, V. Croquette, A. Montanari, and M. Mézard (to be published).