Structural bioinformatics

Modeling the interplay of single-stranded binding proteins and nucleic acid secondary structure

Robert A. Forties\textsuperscript{1} and Ralf Bundschuh\textsuperscript{1,2,3,}\textsuperscript{*}

\textsuperscript{1}Department of Physics, \textsuperscript{2}Center for RNA Biology and \textsuperscript{3}Department of Biochemistry, The Ohio State University, Columbus, OH, USA

Received on September 10, 2009; revised on October 28, 2009; accepted on November 1, 2009

Advance Access publication November 4, 2009

1Department of Physics, 2Center for RNA Biology and 3Department of Biochemistry, The Ohio State University, Columbus, OH, USA

To whom correspondence should be addressed.

ABSTRACT

Motivation: There are many important proteins which bind single-stranded nucleic acids, such as the nucleocapsid protein in HIV and the RecA DNA repair protein in bacteria. The presence of such proteins can strongly alter the secondary structure of the nucleic acid molecules. Therefore, accurate modeling of the interaction between single-stranded nucleic acids and such proteins is essential to fully understand many biological processes.

Results: We develop a model for predicting nucleic acid secondary structure in the presence of single-stranded binding proteins, and implement it as an extension of the Vienna RNA Package. All parameters needed to model nucleic acid secondary structures in the absence of proteins have been previously determined. This leaves the footprint and sequence-dependent binding affinity of the protein as adjustable parameters of our model. Using this model we are able to predict the probability of the protein binding at any position in the nucleic acid sequence, the impact of the protein on nucleic acid base pairing, the end-to-end distance distribution for the nucleic acid and FRET distributions for fluorophores attached to the nucleic acid.

Availability: Source code for our modified version of the Vienna RNA package is freely available at http://bioinf.mps.ohio-state.edu/Vienna+P. implemented in C and running on Linux.

Contact: bundschuh@mps.ohio-state.edu

1 INTRODUCTION

Nucleic acids have been recognized to play a major role in many biological functions (Alberts et al., 2007; Conon, 2002). Frequently they do not act alone but rather in conjunction with proteins. One important class of such proteins associated with nucleic acids are proteins that bind to single-stranded nucleic acids. For example, the nucleocapsid protein from HIV binds and helps to package its RNA genome (Cosa et al., 2004; Guzman et al., 1998) and is involved in strand transfer during the initial steps of HIV’s replication cycle as well. Other examples include the RecA protein which binds single-stranded DNA and aids in strand exchange in bacteria (Story et al., 1992) and the deactivation of vascular endothelial growth factor through binding with a small DNA aptamer (Taylor et al., 2008).

An important aspect for the biological functions of nucleic acid molecules is the secondary structure of these molecules, i.e. the base pairings formed between their monomers. For an individual nucleic acid molecule by itself the secondary structure can be predicted from its specific sequence relatively reliably (McCaskill, 1990; Zuker and Stiegler, 1981). However, the presence of proteins that directly interact with a nucleic acid molecule will change this secondary structure. In order to understand the biological functions of a nucleic acid molecule and its interactions with nucleic acid binding proteins, it is thus important to fully understand the interplay between the nucleic acid’s secondary structure and the proteins. Here, we develop a quantitative model that predicts secondary structures in the presence of proteins that bind to the single-stranded segments of a nucleic acid molecule, as well as the probability that protein is bound at any given position within the nucleic acid sequence.

Our model is implemented as an extension of the Vienna RNA package. The Vienna RNA package can predict the distribution of secondary structures for RNA in O(n^3) time, if n is the number of bases in the molecule. Through the use of a different energy parameter set the Vienna RNA package can also compute the secondary structure of single-stranded DNA (Mathews et al., 2004; Hofacker et al., 1994). When a protein binds to a single-stranded nucleic acid, there is a free energy gain due to the favorable interaction between the nucleic acid and the protein. However, there is also a free energy cost because the region of the nucleic acid bound by the protein can no longer form base pairings. We incorporate the competition between these two interactions into the Vienna RNA package to determine the impact of the binding protein on nucleic acid secondary structure. Using this model, we can not only predict the secondary structure of nucleic acids in the presence of binding proteins, but also compare with many experimental observables. These include end-to-end distance and melting of the nucleic acid, binding of the protein to the nucleic acid and FRET (fluorescence resonance energy transfer) between dyes attached to the nucleic acid.

The binding of any single-stranded binding protein may be characterized by its affinity for the nucleic acid and its footprint, the number of nucleotides bound by one protein. For some proteins, the binding affinity is dependent on the nucleotide sequence, while for others sequence dependence may not be important. In some cases, proteins may bind cooperatively, or may form dimers. We therefore make our model as general as possible, such that it may accommodate any type of protein binding to single-stranded nucleic acids. By fitting experimental data with our model we can gain insight into the physical parameters of a specific binding protein, such as binding affinity, footprint and cooperativity, which give rise...
to the experimental observations and are ultimately responsible for the biological consequences of the interaction between the protein and the nucleic acid. Once the interaction parameters have been determined by comparison with specific experiments, our model allows the prediction of the effect of the protein binding on the secondary structure formation and vice versa for arbitrary sequences of the nucleic acid and arbitrary concentrations of the protein.

The remainder of the article is organized as follows. We first explain the methods we use to calculate nucleic acid secondary structures in the presence of binding proteins. Then, we give an overview of some physical observables which may be calculated using the model presented here. Finally, we discuss the significance of such modeling to the study of nucleic acids in the presence of binding proteins.

2 METHODS

A single-stranded nucleic acid may form secondary structures by looping back on itself, allowing pairing between complementary bases. The probability of any such secondary structure forming can be predicted by calculating the partition function for a nucleic acid's secondary structure. In general, the partition function is

$$Z = \sum_s e^{-\frac{E(s)}{k_BT}},$$

(1)

where the sum is taken over all possible states $s$ of the system, $E_s$ is the free energy of each state, $k_B$ is Boltzmann's constant and $T$ is the temperature. In this formulation, the probability of state $s$ is given by $e^{-\frac{E(s)}{k_BT}}/Z$. In our model, the possible states are all different nucleic acid secondary structures, and the free energy for a secondary structure is given by the sum of the free energies for all loops and base pairings formed in that structure according to the well-established Turner energy model (Mathews et al., 1999, 2004; Xia et al., 1998).

The partition function for nucleic acid secondary structure may be calculated efficiently by using recursive algorithms (McCaskill, 1990). Several quantities related to the partition function are computed for successively larger segments of the nucleic acid until the entire partition function for the whole length of the nucleic acid is built up. We incorporate binding of single-stranded binding proteins in this recursive framework. Here, we will illustrate our approach for a simplified model of nucleic acid secondary structure, in which the free energy to form a base pairing between nucleotides $i$ and $j$, $E(i,j)$, only depends on the identities of $i$ and $j$, and not on the secondary structure formed by the rest of the nucleic acid. Our actual calculation is a modified version of the Vienna RNA package (Hofacker et al., 1994), which implements the full Turner energy model (Mathews et al., 1999, 2004; Xia et al., 1998) in which the free energies are associated with base pair stacking and depend on the secondary structure of the region in which they form. However, since the method to add in binding proteins is essentially the same in the simplified model and in the full model, we explain it here in the more easily accessible case of the simplified model and give the full details for the Turner energy model in the Appendix A.

We represent the partition function for a nucleic acid segment $(i,j)$ with no base pairings and no proteins bound as $Q_{ij}$. Since we define all energies with respect to unpaired nucleic acids, $Q_{ij} = 1$ for any $i$ and $j$. Including $Q_{ij}$ in our formulation allows us to explicitly keep track of unpaired regions, which will be useful later when we allow proteins to bind to these unpaired regions. We also represent the partition function for a segment $(i,j)$ where $i$ is paired with an (unspecified) base $k$, $k+1$ through $j$ are unpaired and the state of $(i+1,k-1)$ is unconstrained as $W_{ik}$.

This allows us to calculate the unconstrained partition function $Q_{ij}$ for a segment $(i,j)$ starting at base $i$ and ending at base $j$ by the recursion relation

$$Q_{ij} = O_{ij} + W_{ij} + \sum_{k=1}^{j-i-k-1} Q_{ik} W_{kj},$$

(2)

where $Z_{\text{min}}$ is the minimum number of nucleotides for the nucleic acid to bend back on itself and form a hairpin loop. No terms $Q_{ij}$, where $k < i$ are included as the sum over $k$ is restricted to $k \geq i$. Therefore, there is no ambiguity in the definition of $Q_{ij}$. The restricted partition function $W_{ij}$ is then given by

$$W_{ij} = \sum_{k=1}^{j-i} Q_{ik} W_{kj} + Q_{ij}.$$

(3)

Note that we use the convention $O_{ij} = 1$ even for $i > j$, which corresponds to a substrand of negative length. Therefore, there is no issue with the sum in Equation (3) including the term $O_{ij}$ when $k = j$. We show an arch diagram representation of the decomposition of the partition function given by Equations (2) and (3) in Figure 1.

Using these recursion relations, we are able to calculate $W_{ij}$ and $Q_{ij}$ for any segment. We begin by setting $Q_{ij} = 1$ and $W_{ij} = 0$ for all segments where $j-i < Z_{\text{min}}+1$. These segments are too short to allow any base pairing to occur, so their unconstrained partition function $Q_{ij}$ is equal to $O_{ij} = 1$. We then calculate $W_{ij}$ and then $Q_{ij}$ for the segments with length $j-i = Z_{\text{min}}+1$ using Equations (2) and (3), respectively. We then continue to work our way up to larger and larger segments using these recursions. Finally, we obtain the total partition function for the whole sequence by calculating $Q_{ij}$, where $n$ is the number of nucleotides in the sequence.

To include the impact of binding proteins, we make use of a chemical potential for proteins binding to unpaired regions of the nucleic acid. A chemical potential gives the energy for a system to accept one particle from its environment. In this case, each protein from the surrounding solution which binds to the nucleic acid lowers the free energy of the system by the chemical potential $\mu_i(i)$, where $i$ indicates the position of the first nucleotide the protein binds to. This helps to stabilize secondary structures of the nucleic acid which have unpaired regions where proteins may bind. The protein may bind more tightly to some nucleotide sequences than others, so the chemical potential can vary with position within the nucleotide sequence $i$. Additionally, as protein concentration is increased, it becomes more favorable for proteins to bind to the nucleic acid. For dilute solutions the chemical potential depends on protein concentration $c$ following $\mu(c,i) = \mu_i(c_0) + k_BT \ln(c_i/c_0)$, where $c_0$ is an arbitrary reference concentration (Schroeder, 2000).

As we now allow proteins to bind to unpaired regions of nucleic acid, we no longer represent these regions as $O_{ij} = 1$. We instead use the sum over the contributions from all possible combinations of proteins which may bind to the nucleic acid segment

$$Q_{ij} = \sum_k \mu(c,k) + \sum_{j-i}^{j-i-k-1} \mu(c,k) T_{ij} \cdots T_{ik},$$

(4)

where the sums over $k$ and $l$ must be restricted to only allow combinations of binding proteins which fit on the segment without overlapping. The sum over $k$ in the first term is taken over all positions in which a single protein may bind. The sums over $k$ and $l$ in the second term are taken over all combinations of two proteins which may simultaneously bind to the segment. Larger terms,
which are not shown, would need to be included for each number of proteins which can simultaneously bind the segment. While cooperativity could easily be included in this formulation, the proteins are assumed here to bind to the nucleic acid independently, so the chemical potential for multiple proteins binding is just the sum of their individual chemical potentials.

This quantity can again be calculated recursively by

$$O_j = O_{j-1} + e^{-\Delta E_{j-1}/k_BT} O_{j-1}$$

where $F$ is the footprint, or number of nucleotides bound by one protein. For $j=1$, $F=1$, we still use $O_0=1$, as such sequences are not long enough to accommodate even one protein binding. For longer sequences, the above recursion allows for all possible combinations of proteins which may bind to the sequence. Since $\mu^{\text{ext}}(x,j)$ depends on both sequence and protein concentration, $O_j$ and thus $W_j$ and $Q_j$ do as well.

As indicated earlier, this yields the partition function for a simplified model of nucleic acid secondary structure, which we use to illustrate the method by which we model the impact of binding proteins on secondary structure. Our full model is identical to the calculation here except that more details of different types of base pairings are included when calculating $W_j$, as required by the Turner energy model (Mathews et al., 1999, 2004; Xia et al., 1998). The full calculation of $W_j$ can be found in the Appendix A.

3 OBSERVABLES

To compare this model with experiments we must be able to extract quantities that can be observed in experiments. By calculating restricted partition functions, we may predict melting probabilities, end-to-end distances, FRET values for fluorophores attached to the nucleic acids and the probability that a protein is bound at any position within the nucleic acid sequence as a function of protein concentration. The dependence of the observables on protein concentration depends on the interaction parameters, i.e. the footprint size $F$ and the binding free energies $\mu^\text{ext}(x,j)$. Thus, these parameters can be extracted from comparing model predictions with actual experiments.

3.1 Melting probability

Nucleic acid melting refers to breaking of base pairings between nucleotides. The probability that a given base is paired may be calculated directly using the partition function method outlined above, generalizing McCaskill’s approach (McCaskill, 1990) to our extended model.

We wish to calculate $Q_{\text{pair}}(p,q)$, the partition function for the full sequence with the constraint that base $p$ is paired to base $q$. In order to do this, we must first calculate the partition function for a sequence twice the length of the original. This will allow us to compute all possible base pairings between segments at the ends of the sequence while excluding those in the middle. We construct this sequence by setting bases (1,1) equal to the original sequence, and then repeating this sequence again for bases ($n+1$,2n). We next calculate the partition function for this doubled sequence, $Q_{\text{pair}}^\prime$, using Equation (2). We may then calculate the constrained partition function using

$$Q_{\text{pair}}(p,q) = Q_{\text{pair}}(p+1,q-1)e^{-E(p,q)/(k_BT)}Q_{\text{pair}}^\prime(q+1,n+1)$$

Since when we calculated $Q_{\text{pair}}^\prime$ we duplicated the sequence, $Q_{\text{pair}}(p,q)$ actually gives the partition function for the union of segments (1, p-1) and (q+1, n).

Using this constrained partition function, we may then calculate the probability that base $p$ is paired with $q$ via

$$P_{\text{pair}}(p,q) = Q_{\text{pair}}(p,q)/Q_{1,n}$$

By calculating this pairing matrix as a function of protein concentration, we may then examine the impact of binding proteins on DNA melting. (Note that for pedagogical reasons we ignore the fact that the minimum hairpin constraint $L_{\text{min}}$ should not apply to base pairs which bracket the connection between the two copies of the molecule. Again, our code contains the full calculation of the pairing probabilities.)

3.2 End-to-end distance

The end-to-end distance for a state of a nucleic acid is proportional to the number of exterior bases present, or the equivalent length of single-stranded nucleic acid which may be used to represent the state. Exterior bases are defined here as the number of unpaired bases which are not part of any loop, plus a factor $\Delta$ (the average width of a stem divided by the width of 1 nt) times the number of stems not contained in any loop. To calculate end-to-end distance and FRET (see below), we need to first find the restricted partition function for the segment $(1,j)$ having exactly $m$ exterior bases (Gerland et al., 2001)

$$Q_{1,j}(m) = \sum_{i=j-L_{\text{min}}}^{j} W_{1,k-1}(m-(j-k+1))O_{k,j} + W_{1,k}(m) + O_{1,j}$$

where $O_{j,m} = 1$ for $j=m$ and zero otherwise. Therefore, the first term in Equation (8) is simply an unpaired segment of length $m$. $W_{1,k}(m)$ is the restricted partition function for $(1,j)$ with $m$ exterior bases and the additional restriction that $j$ is paired to an unspecified base in $(1,j-1)$. It is calculated from

$$W_{1,k}(m) = \sum_{k=m-\Delta+1}^{j-L_{\text{min}}} Q_{1,k-1}(m-\Delta e^{-E(x,j)}/k_BT)Q_{1,k+1,j-1}$$

The probability that a nucleic acid of length $n$ has exactly $m$ exterior bases is then simply

$$P(m) = Q_{1,n}(m)/Q_{1,n}$$

If we then make use of a polymer model which predicts that an unpaired single-stranded nucleic acid of length $m$ has a probability $P_{\text{poly}}(m)$ of having end-to-end distance $x$, the predicted end-to-end distance distribution for a nucleic acid including base pairing and protein interactions is

$$P(x) = \sum_{m=\Delta}^{n} P(m)P_{\text{poly}}(x,m)$$

One such polymer model is the worm-like chain model (Kratky and Porod, 1949). The end-to-end distance function for this model $P_{\text{poly}}(x,m)$ is given in Thirumalai and Ha (1998), and depends on the contour length of single-stranded DNA and its persistence length, the length over which orientations of the polymer become decorrelated. This model provides a good fit to experimental measurements of single-stranded DNA flexibility with a persistence length of 2.5 nm (Murphy et al., 2004), and a contour length of 0.63 nm per nucleotide (Olson, 1975; Saenger, 1984).
3.3 FRET distribution

FRET (Fluorescence resonant energy transfer) experiments are a convenient way to experimentally measure the end-to-end distance of biological molecules (Ha et al., 1996, 2002). To this end, two different fluorophores are attached to the ends of a nucleic acid molecule. The donor fluorophore is excited by a laser tuned to its absorption maximum, and then may either decay or transmit its excitation to the acceptor fluorophore via FRET. The emission of the system is measured at the characteristic wavelengths of both the donor and the acceptor. Emission by the acceptor corresponds to an instance where FRET excitation transfer occurred, while emission by the donor corresponds to no FRET. The FRET efficiency is defined as $E_{\text{FRET}} = I_A / (I_D + I_A)$ (Ha et al., 1996), where $I_D$ is the emission intensity of the donor and $I_A$ is the emission intensity of the acceptor. Therefore, the FRET efficiency is the fraction of the total emission which comes from the acceptor, after having undergone FRET transfer from the donor. The rate at which FRET occurs is highly dependent on distance, and is given by

$$k_{\text{FRET}} = k_0 \left( \frac{R_0}{R} \right)^6,$$  \hspace{1cm} (12)

where $k_0$ is the rate of fluorescent decay for an isolated donor fluorophore, $R_0$ is the FRET radius, which is a property of the fluorophores used and their environment and $R$ is the distance between the donor and the acceptor, which in this case is the end-to-end distance for the nucleic acid. Therefore, FRET gives a measurement of the end-to-end distance of the nucleic acid.

Since $k_{\text{FRET}}$ depends non-linearly on the end-to-end distance, it is non-trivial to convert the end-to-end distance distribution to a FRET distribution. Therefore, we use a Monte Carlo simulation to sample the end-to-end distance distribution and predict when FRET will occur, as described in Murphy et al. (2004).

For completeness, we summarize this approach here. The Monte Carlo simulation proposes random steps in the distance between the two fluorophores, and accepts or rejects these steps based on the Metropolis algorithm using energies calculated from the worm-like chain model. We use a step size $\Delta r = 0.55$ nm and a time step $\Delta t = 0.01$ ns, which together correspond to a spatial diffusion constant of $3 \times 10^{-4} \text{ m}^2/\text{s}$ (Murphy et al., 2004).

At each step, the donor may either decay with probability $k_0 \Delta t$, or transmit its excitation to the acceptor with probability $k_{\text{FRET}} \Delta t$. These probability distributions are sampled using the GNU scientific library (Galassi et al., 2009). If either type of deexcitation occurs, we record whether a FRET transfer or decay was observed for this excitation, and then wait for the donor to become excited again. The end-to-end distance continues to fluctuate according to the worm-like chain model, and each step has a probability to become re-excited of $k_0 \Delta t$. To be specific, we use for our illustration a rate for isolated donor fluorophore decay of $k_0 = 1 \text{ ns}^{-1}$ (Murphy et al., 2004), and FRET radius of $R_0 = 6$ nm (Ha et al., 2002), which is appropriate for the Cy3/Cy5 donor/acceptor pair often used in such experiments. The rate of donor excitation depends on the irradiation intensity, and therefore can vary with the experimental apparatus. We choose a rate for donor excitation of $k_D = 1 \text{ ns}^{-1}$, which is a typical value that could be achieved in such experiments.

We run this Monte Carlo simulation for 1 000 000 donor excitations at each different number of exterior bases $m$. This allows us to accurately sample the probability distribution of the FRET intensity, $P_F(E_{\text{FRET}} | m)$, as a function of the number of exterior bases $m$ and the FRET efficiency $E_{\text{FRET}}$. We may then predict the observed FRET distribution for a nucleic acid using

$$P(E_{\text{FRET}}) = \sum_{m=0}^{\Delta} P(m) P_F(E_{\text{FRET}} | m).$$ \hspace{1cm} (13)

The distribution $P_F(E_{\text{FRET}} | m)$ does not depend on the nucleic acid sequence used or any binding proteins present. Therefore, if we calculate $P_F(E_{\text{FRET}} | m)$ once for a given set of experimental conditions and polymer model parameters, we then may calculate $P(E_{\text{FRET}})$ for any combination of nucleic acids and binding proteins.

As an example, we apply this method to simulate a single-stranded DNA containing 38 nt, shown in Figure 2a. We calculate the FRET distribution for this sequence at many different concentrations of single-stranded binding protein, for a protein with a chemical potential for binding of 10 kcal/mol at 1 M concentration and a footprint of 6 nt. This distribution is shown in Figure 2b. At low protein concentration, we only see FRET values very close to 1. This corresponds to the sequence forming almost all of the possible base pairings shown in Figure 2a. As protein concentration is increased, we start to see a second FRET peak form between $E = 0.4$ and 0.7, which corresponds to melting of the 6 bp stem closest to the ends of the DNA sequence. Finally, at very high protein concentration we see a third peak, which corresponds to all bases being unpaired, form between $E = 0.1$ and 0.4. This example shows that a relatively small single-stranded DNA in the presence of single-stranded binding proteins can result in a complicated FRET distribution, without any cooperativity or sequence specificity in protein binding. For instance, Figure 2c shows that at a protein concentration of 500 nM all three FRET peaks coexist, which is a relatively complex distribution for such a small molecule. The details of the distribution arise purely from the promotion of DNA melting by adding increasing amounts of binding protein.

3.4 Protein binding

Using the methods described above, we may calculate the total partition function for the nucleic acid secondary structures in the presence of single-stranded binding proteins. The interaction between the nucleic acid and the binding proteins is governed by the chemical potential for protein binding $\mu(c,i)$. Therefore, by varying the value of $\mu(c,i)$ at a given nucleotide $i$, we may examine the strength of the binding interaction between the nucleic acid and protein at nucleotide $i$. We may then repeat this at every position within the nucleic acid sequence to find the probability that a protein is bound at each possible position. With this information we can compute experimentally measurable quantities such as the average number of proteins bound per nucleic acid, or the probability that at least one protein is bound to the nucleic acid.

Specifically, the probability that a protein is bound to the nucleic acid starting at nucleotide $i$ is (Schroeder, 2000)

$$P_B(c,i) = \frac{k_B T}{Q_{1,n}} \frac{\delta Q_{1,n} |_{\delta \mu(c,i)}}{\delta \mu(c,i)}$$ \hspace{1cm} (14)
Since we can calculate $Q_{1,n}$ for any choice of $\mu^i(c,i)$ by Equation (2), we can obtain the partial derivative $\delta \mu^i(c,i)$ and hence $P(c,i)$ numerically by calculating $Q_{1,n}$ for two nearby values of $\mu(c,i)$.

Each protein will bind $F$ consecutive nucleotides, where $F$ is the protein’s binding footprint. If the first nucleotide bound by a protein is $i$, then the protein will bind nucleotides $i$ through $i+F−1$. Therefore, the probability that nucleotide $i$ is bound by a protein (regardless of whether it is the first nucleotide bound by this protein) is

$$P_{\text{occupied}}(c,i) = \sum_{j=i-F+1}^{i} P(c,j),$$

where we have set $P(c,j)=0$ for $j<0$. Figure 3a shows a plot of $P_{\text{occupied}}(c,i)$ as a function of protein concentration for the sequence shown in Figure 2a. As very low concentration, the protein binds to the loop at the middle of the nucleic acid, which does not require any of the base pairings which form in the absence of protein to be broken. The protein next invades the relatively weak stem formed by the ends of the nucleic acid, breaking this base pairing and reaching 50% binding to this region at a concentration of \(\sim 150\) nM. Finally, at high protein concentration the much more stable stem formed near the middle of the nucleic acid is broken and the protein binds to this region, reaching 50% binding at \(\sim 700\) nM.

Multiple proteins may bind to each nucleic acid, up to a limit of $n/F$, where $n$ is the number of nucleotides in the nucleic acid and $F$ is again the protein’s binding footprint. The average number of proteins bound per nucleic acid molecule is

$$P_{\text{avg}}(c) = \sum_{i=1}^{n} P(c,i).$$

Since we can calculate $Q_{1,n}$ for any choice of $\mu^i(c,i)$ by Equation (2), we can obtain the partial derivative $\delta \mu^i(c,i)$ and hence $P(c,i)$ numerically by calculating $Q_{1,n}$ for two nearby values of $\mu(c,i)$.
nucleic acid and allowing two additional proteins to bind. Corresponding to breaking of the more stable stem near the middle of the nucleic acid, allowing two more proteins to bind. At higher concentrations, proteins bound. This corresponds to breaking the weak stem near the ends of the sequence, increasing approximately linearly with concentration from about 1.0 to 3.0 nM concentration. The number of proteins bound then increases as two proteins may bind to the loop in the middle of the DNA sequence without disturbing any base pairings. The number of proteins bound to each nucleic acid as a function of protein concentration calculated from Equation (16). The number of proteins bound jumps up to 1.0 almost immediately at low concentration, as two proteins may bind to the loop in the middle of the DNA sequence without disturbing any base pairings. The number of proteins bound then increases approximately linearly with concentration from about 1.0 to 3.0 proteins bound. This corresponds to breaking the weak stem near the ends of the nucleic acid, allowing two more proteins to bind. At higher concentrations, the number of proteins again increases linearly, but at a much slower rate, corresponding to breaking of the more stable stem near the middle of the nucleic acid and allowing two additional proteins to bind.

We can also compute the fraction of nucleic acid molecules which are bound by at least one protein, which may be compared directly with experimental binding assays. First, we compute that the fraction $P_{\text{unbound}}(c)$ plotted as a function of both position within the nucleic acid sequence $i$ and protein concentration $c$. Darker shades of gray indicate higher binding probability, as indicated in the legend. At low protein concentrations, the protein first binds to the loop in the middle of the sequence. Then, as the concentration is increased the protein next binds near the ends of the sequence, breaking the relatively weak stem formed there. Finally, at very high concentrations the protein binds to the G and C nucleotides near the middle of the sequence, breaking the much stronger pairings between these bases. (b) The average number of proteins bound to each nucleic acid as a function of protein concentration calculated from Equation (16). The number of proteins bound jumps up to 1.0 almost immediately at low concentration, as two proteins may bind to the loop in the sequence, breaking the much stronger pairings between these bases.

The binding probability $P_{\text{bound}}(c,i)$ we are able to compare with a wide variety of experiments. We can predict FRET for dyes attached to the nucleic acid, the probability of the protein binding at every position in the nucleic acid sequence, as well as end-to-end distance distributions and base pairing probabilities for the nucleic acid. This allows interpretation of such experiments in terms of the physical binding parameters of the protein, such as its binding affinity and footprint.

**ACKNOWLEDGEMENT**

We thank Paul Barbara for helpful discussions of the behavior of single-stranded binding proteins.

**Funding**: National Science Foundation Graduate Research Fellowship (to R.F.); the Petroleum Research Fund of the American Chemical Society (Grant No. 42555-G9 to R.B.); National Science Foundation (Grant No. DMR-0706002 to R.B.).

**Conflict of Interest**: none declared.

**REFERENCES**


for each case. This would allow us to consider protein binding to different kinds of loops separately, in case the protein has different affinities for unpaired segments within different kinds of loops or external to loops.

The partition function for an internal loop closed by pairing i with j and k with l, with (i + 1, k - 1) and (l + 1, j - 1) unpaired (where $i < k < l < j$), is

$$B_{ijkl} = \sum_{k=1}^{j-l} e^{-\Delta H_{ijkl} / kT} O_{i,j-1} O_{l,j-1}$$

\hspace{1cm} (A.4)

where $\Delta H_{ijkl}$ is the free energy cost to form this internal loop, bulge or set of stacked base pairs. If $k > i + 1$ and $j > l + 1$, then the structure is an internal loop. If either $k = i + 1$ or $j = l + 1$, then the structure is a bulge. If both $k = i + 1$ and $j = l + 1$, then the structure is a set of stacked base pairs. All needed values of $\Delta H_{ijkl}$ for internal loops, bulges and stacked base pairs have been extrapolated from experiments (Matthews et al., 2004).

We note that while use of Equation (A.4) gives a partition function which scales with $O(n^3)$, the scaling is reduced to $O(n^2)$ by only allowing internal loops smaller than a cutoff size.

The partition function for a multiloop, in which i and j are paired, and there are at least two more stems on (i + 1, j - 1) is $B_{m_{ij}}$. The free energy of a multilooop is given by the free energy of its components plus an additional cost to close the loop

$$C^M = C^C + n_{unpaired}C^U + n_{stem}CS$$

\hspace{1cm} (A.5)

where $C^C$ is the fixed cost for a multiloop closed by pairing bases i and j, $n_{unpaired}$ is the number of unpaired nucleotides which form the multiloop, $C^U$ is a cost per unpaired base in the multiloop, and $n_{stem}$ is the number of stems contained in the multiloop and $CS$ is the cost per stem. The partition function for a multiloop is then

$$B_{m_{ij}} = \sum_{k=i+2}^{j-l} Q_{m_{i+k,j}},$$

\hspace{1cm} (A.6)

where $Q_{m_{i+k,j}}$ is the partition function for a segment (i, j) containing at least one stem with multiloop closure energy included, given by

$$Q_{m_{i+k,j}} = \sum_{l=i+1}^{j-1} W_{m_{i+k,l}} W_{m_{l+j}} e^{-\Delta H_{m_{i+k,j}} / kT}$$

\hspace{1cm} (A.7)

and $W_{m_{i+k,l}}$ is the partition function for (i, j) where i is paired with an (unspecified) base k, and l through j are unpaired, the state of $(i+1, j-1)$ is unconstrained, and multiloop energies are included. It is identical to $W_{ij}$ except for the addition of multiloop energies, and is given by

$$W_{m_{i+k,l}} = \sum_{l=i+1}^{j-1} B_{i+k,l} e^{F_{unpaired}(i) / kT}$$

\hspace{1cm} (A.8)

The above gives the restricted partition function with all contributions for different types of loops found in the Vienna RNA package. We then need only to insert $W_{ij}$ in place of $W_{ij}$ in Equation (2) to calculate the full unconstrained partition function for nucleic acid secondary structure in the presence of binding proteins.