STATISTICAL THERMODYNAMICS FOR RNA STRUCTURES WITH SIMPLE TERTIARY CONTACTS AND PSEUDOKNOTS

A Dissertation
presented to
the Faculty of the Graduate School
University of Missouri-Columbia.

In Partial Fulfillment
Of the Requirements for the Degree

Doctor of Philosophy

By
ZOIA KOPEIKIN

Dr. Shi-Jie Chen, Dissertation Supervisor

MAY 2006

The undersigned, appointed by the Dean of the Graduate School,
have examined the dissertation entitled

STATISTICAL THERMODYNAMICS FOR RNA STRUCTURES WITH SIMPLE TERTIARY CONTACTS AND PSEUDOKNOTS

Presented by Zoia Kopeikin

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

Professor Shi-Jie Chen

Professor Ioan Kosztin

Professor Peter Pfeifer

Professor Ping Yu

Professor Michael Henzl

ACKNOWLEDGMENTS

I would like to thank the following people for their help, support and contributions to this work: Dr. Shi-Jie Chen, Dr. Wen-Bing Zhang, Dr. Zhijie Tan, Dr. Song Cao, and my committee: Dr. Peter Pfeifer, Dr. Ioan Kosztin, Dr. Ping Yu, and Dr. Michael Henzl.

Contents

Chapter

List of Figures

Figure **Page**

List of Tables

ABSTRACT

RNAs are polynucleotide chains. Despite the widespread importance of RNA folding for cellular function, understanding of the principle of the RNA folding, especially of the tertiary structure folding, is very limited. Reliable prediction for tertiary structural stability and folding pathways is not possible, even for the simplest tertiary folds. In this thesis, we develop statistical mechanical models for the folding of simple RNA tertiary structures.

One of the bottlenecks to model RNA tertiary folding is the entropy problem. The major focus of this thesis is to develop a conformational entropy model for simple RNA tertiary structures, specifically, for *RNA folds with simple tertiary contacts* and *RNA pseudoknotted folds*. A major challenge in the theory is how to account for (1) the nonlocal correlations between different parts of the tertiary structures and (2) the volume exclusion between different nucleotide units of the chain. Our principle approach is “dividing and conquering”: to divide the structure into conformational subunits and to treat the inter-subunit interactions by focusing on the local interactions near the inter-subunit interfaces. The theory is developed based on a two-dimensional lattice model. Extensive tests against exhaustive computer lattice enumerations show that the model is accurate and reliable.

The model developed in this thesis enables predictions for the energy landscapes and conformational transitions for simple RNA tertiary folds and RNA pseudoknots. The model can predict the interplay between the secondary and the tertiary interactions in the conformational changes. The theory has been applied to study the mechanical unfolding of model RNA H-pseudoknots. The information about structural transitions in the unfolding process has been obtained from force-extension curves, computed using the force-dependent parti-

tion functions. The (equilibrium) folding pathways and folding cooperativities have been predicted based on the free energy landscapes.

Though the current form of the model is based on a two-dimensional lattice model, the machinery developed here permits the use of arbitrary chain representation and can be generalized to any off-lattice models. Furthermore, the analytical formulation of the method makes possible the systematic development of the theory for more complex tertiary structures. Therefore, the models developed in this thesis may provide paradigms for modeling more complex RNA tertiary structure folding thermodynamics.

Chapter 1

INTRODUCTION

1.1 What is RNA?

Ribonucleic acids (RNAs) are polymer molecules which perform various critical functions in processes of transmission, expression, and conservation of genetic information. The building blocks of an RNA are nucleotides. Each nucleotide consists of a phosphate, a sugar, and a base (Fig. 1.1). The phosphate groups link the 5' carbon of one ribose to the 3' carbon of the next. Correspondingly, the two ends of the chain are referred as 5' and 3' ends and have different chemical properties. The sequence of a nucleotide chain is defined from the 5'end to the 3' end. Bases are responsible for coding the genetic information. There are four types of bases in RNA: adenine (A), guanine (G), cytosine(C) and uracil (U) (Fig. 1.1). The RNA structure can be described at three different levels. The **primary structure** of the molecule is a sequence of bases, i.e. the covalent chain structure. The conformations of nucleotides depend on the torsion angles for rotation around each covalent bond (Fig. 1.1). There are seven torsion angles that must be specified to characterize the conformation of each nucleotide.

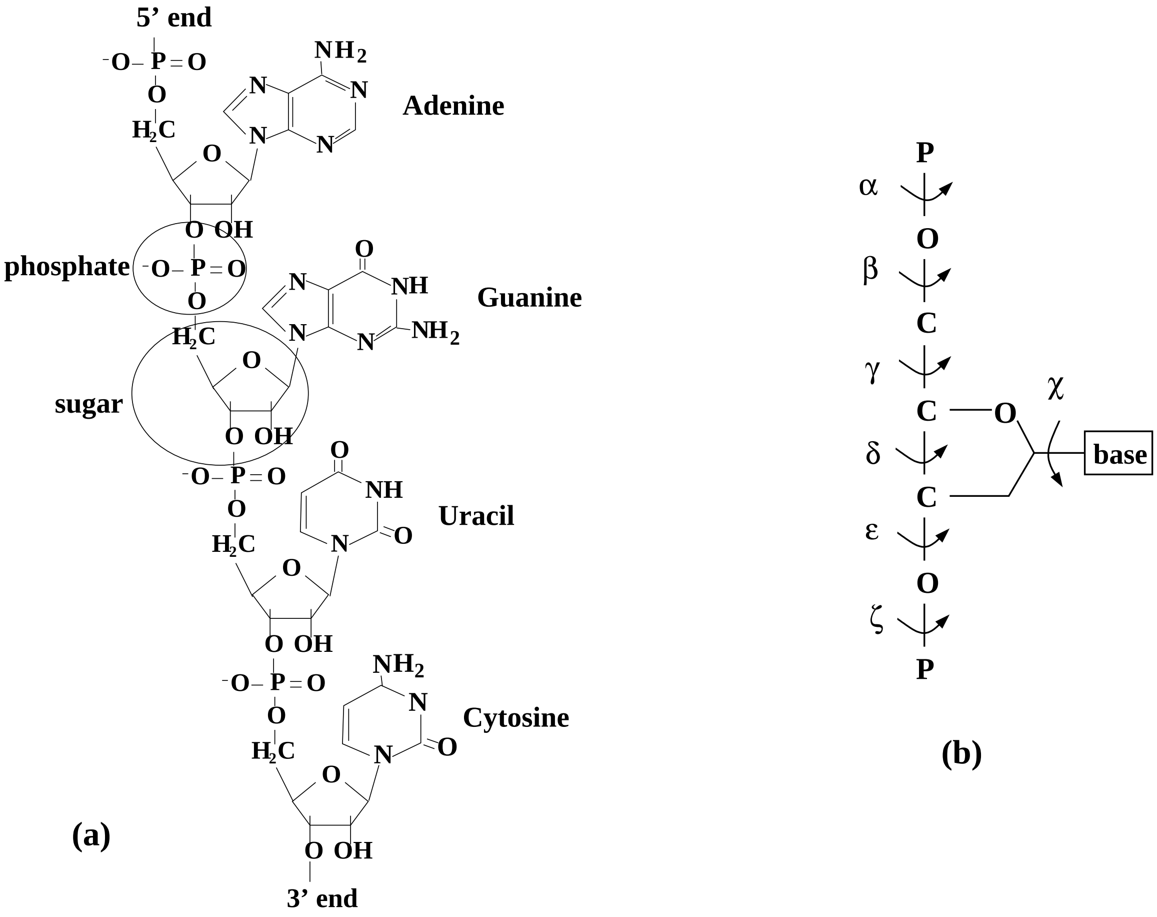


Figure 1.1: (a) RNA nucleotide chain (b) Seven torsion angles are shown which characterize the conformation of each nucleotide and correspond to rotation around covalent bonds.

The **secondary structure** is formed by pairing of bases. J. Watson and F. Crick proposed the idea of specific interactions between complementary bases: A with U and G with C. A weaker base pair is also possible between G and U (so called wobble base pair). As a result of base pairing, a single-stranded RNA molecule folds back onto itself with the formation of double helices (stacked base pairs) of complementary strands. The base pairs are hydrogen-bonded and their geometry is such that they can fit into the helix without it’s distortion.

A secondary structure is formally defined as a list of base pairs (contacts) such that any two contacts (i, j) $(i < j)$ and (k, l) $(k < l)$ are either nested $(i < k < l < j)$ or unrelated $(i < j < k < l)$. A convenient way to visualize intrachain contacts and represent RNA structures is a **polymer graph**. It consists of vertices which represent monomers, and straight line links representing covalent bonds between monomers. Monomers in spatial contact (base pairs in RNA) are also connected by curved links. There are three possible relationships between two contacts: nested, unrelated and crossing linked (Fig. 1.2). *Secondary structures can be represented by graphs which involve only nested and unrelated contacts.*

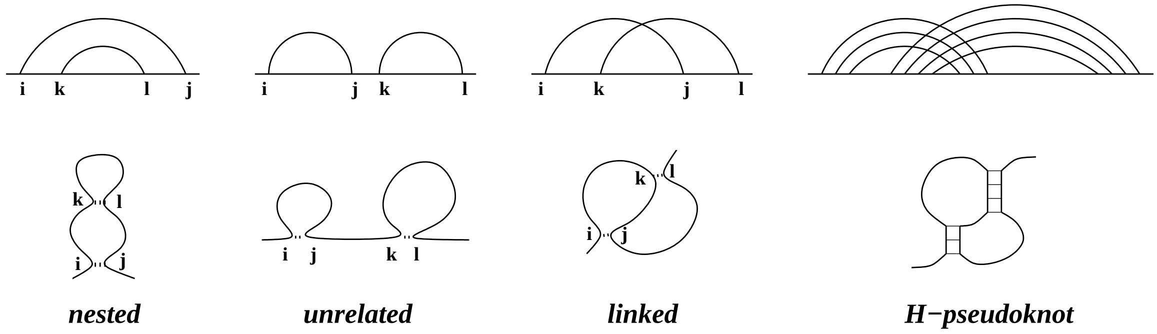


Figure 1.2: Three possible relationships between two links of a polymer graph, and an H-pseudoknot: graphs and chain conformations.

According to the intrachain contacts, we can decompose a secondary structure into relatively independent secondary structural motifs (subunits): stacked base pairs and loops.

Most of RNA molecules adopt A-form double helices which have the following characteristics: they are right-handed, have narrow, deep major (depth 13.5Å, width 2.7Å) and wide, shallow minor (depth 2.8Å, width 11.0Å) grooves, 11 base pairs per turn, translation per residue 2.6Å, helix diameter 19Å and pitch 28Å.

The term “tertiary interactions” is sometimes applied to non-canonical interactions (for example, a hydrogen bond to a phosphate group can often be seen in loops) which are formed in the late stage of folding and are usually weak compared to interactions in secondary structures. Here we use a more restricted definition. If there is a base pair (i, j) $(i < j)$, then any interaction between nucleotides k, l such that $i < k < j < l$ is called tertiary. In the polymer graph, a tertiary interaction is represented by a crossing link (Fig. 1.2). The **tertiary structure** describes how the different parts of the secondary structure are brought together by the tertiary crossing links to form a compact three-dimensional structure. Probably the simplest tertiary structure in an RNA is an H-pseudoknot, which, as shown in Fig. 1.2, contains nucleotides of the hairpin loop paired with tail nucleotides external to the loop.

1.2 RNA functions.

RNA molecules play a variety of critical roles in cell functions. They play a crucial role in the process of protein synthesis, exhibit catalytic activity and store genetic information. There are several types of RNAs according to the functions they have.

1. Messenger RNA (mRNA) molecules are typically several thousand nucleotides long. The DNA molecule, which is usually located in the nucleus of the cell, contains the genetic code which determines the structure of proteins. The genetic code consists of successive triplets of bases; each triplet encodes one amino acid of the polypeptide chain. The function of the messenger RNA is to carry the genetic code to the cyto-

plasm to control the formation of the proteins. The mRNA (as well as other types of RNA) is assembled in a process called *transcription*. The large protein enzyme DNA-dependent RNA polymerase moves along the DNA molecule, temporally unwinding and separating the two DNA strands. Using one of two strands as a template, the RNA polymerase forms the chain of RNA nucleotides which are complementary to DNA nucleotides. In such a way, the sequence of code triplets in the DNA causes the formation of the sequence of complementary code triplets (codons) in the mRNA, which will control the synthesis of a protein molecule. The RNA molecule formed in the transcription is the primary RNA transcript, called pre-messenger RNA, which consists of exons (protein-coding sequences) and introns (non-coding sequences). The primary transcript must undergo processing steps to produce a mature, functional mRNA. Processing includes cutting off introns (splicing) and modification of termini, i.e. formation of untranslated regions (UTR) at 3’ and 5’ ends of the molecule. The functions of UTRs are to protect the molecule from degradation by enzymes called exonuclease, which cleave (break down linkages between nucleotides) at the ends of the molecule, and to regulate the function of the molecule. For example, a stretch of adenine (A) nucleotides (poly(A) tail) is added at the 3’ end of mRNA which protects from degradation and signals translatability (enhances the translation initiation). The formed mature mRNA is released into cytoplasm.

2. Transfer RNA (tRNA) molecules are small (4S) and contain 75-95 nucleotides. They form a very well-defined clover-leaf secondary structure and L-shaped tertiary structure. This structure allows binding of amino acid at one end (acceptor stem) and

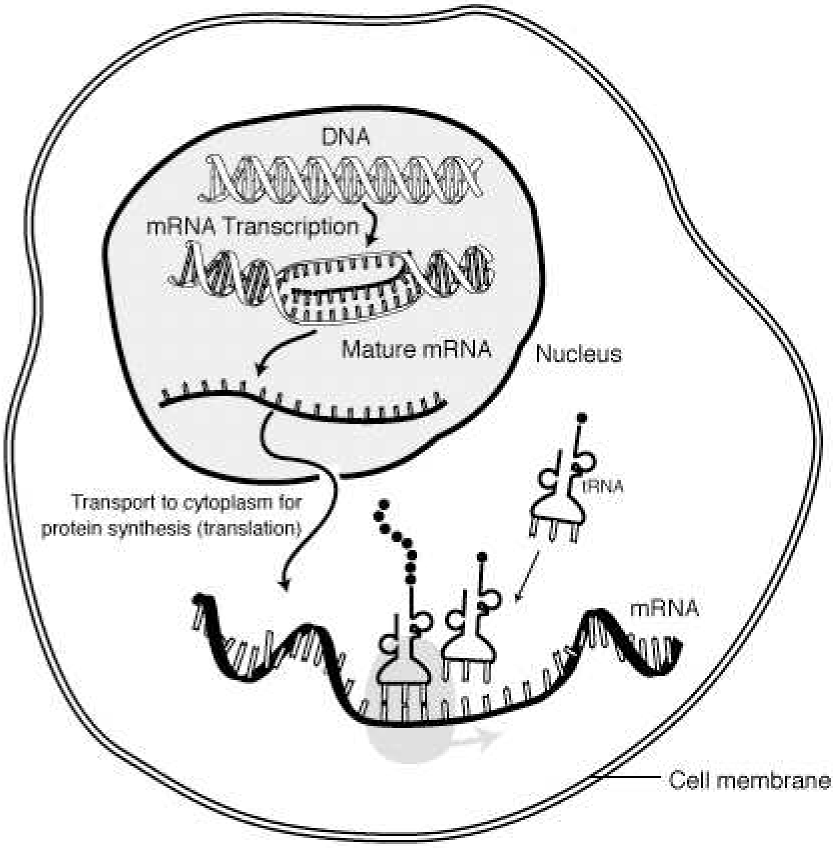


Figure 1.3: Types of RNA and their roles in protein synthesis.

mRNA at the opposite end (anticodon loop). The function of tRNA is to transfer amino acid molecules to protein molecules as the protein is being synthesized. There are more than 20 types of tRNA, each type binds with one of the 20 amino acids and recognizes a corresponding codon on the mRNA. For each amino acid, there is a special enzyme, aminoacyl-tRNA synthetase, which catalyzes its attachment to the acceptor stem of the cognate tRNA. In the anticodon loop, there is a special triplet of bases called anticodon which hydrogen bonds with the complementary codon on the mRNA. This way, the tRNA delivers the appropriate amino acid to the appropriate place on the mRNA.

3. Ribosomal RNAs (rRNAs) along with about 75 different proteins form the ribosomes. The ribosomes are particles on which protein molecules are actually synthesized during the process called *translation*. They function in association with tRNAs which transport amino acids to ribosomes and with mRNAs which provide the information necessary for sequencing the amino acids in proper order. Ribosomes are composed of two sub-units, large and small. In prokaryotes, they are $30S$ and $50S$ subunits. The small subunit contains $16S$ rRNA (approx. 1500 bases). The large subunit contains $23S$ rRNA (approx. 2500 bases) and a smaller $5S$ rRNA (approx. 120 bases). The size is indicated by S numbers which reflect the rate at which the molecules sediment in the ultracentrifuge. There are two sites in the large subunit: the A site accepts a new tRNA bearing an amino acid, and the P site accepts the tRNA with amino acid attached to the growing chain. Two GTP-driven elongation factors (proteins), EF-Tu and EF-G, are responsible for positioning of tRNAs with incoming

and attached amino acids in the appropriate site.

At the beginning of translation, the ribosome comes in contact with the initiation sequence of nucleotides at the 5' end of mRNA. After that, the mRNA travels through the ribosome so that codons are read in a 5' to 3' direction, and a protein molecule forms. When the ribosome meets a chain-terminating codon, the end of a protein molecule is signaled and a protein molecule is freed into the cytoplasm. There are three possible ways (reading frames) of reading a nucleotide sequence as a series of triplets. For example, the nucleotide sequence AGCCCAUGG can be translated in different reading frames as AGC CCA UGG, GCC CAU or CCC AUG. An open reading frame is defined by the start codon (AUG) and triplets are read one after another until the in-frame stop codon (can be UAG, UAA or UGA) is met. The change in translational reading frame is called frameshifting.

4. Small nuclear RNA (snRNA, $\simeq 100 - 300$ nucleotides) are found in the nucleus of cells. They are involved in the processing of other classes of RNA. For example, several snRNAs in complex with proteins form a spliceosome which play an important role in the splicing processing.

5. RNA viruses are particles consisting of one or more RNA molecules surrounded by a protein coat. RNA molecule in this case carries out the role normally played by DNA: it stores genetic information. In DNA viruses genetic material is stored in DNA molecule. RNA viruses can infect bacterial, plant, and animal cells. When the virus enters the host cell, the viral RNA functions as mRNA and is translated by host ribosomes to produce three virus specific proteins, one of which is the enzyme

RNA polymerase (replicase). The RNA polymerase then copies genomic RNA into complementary RNA. This synthesized RNA serves as a template for synthesis of new RNA strands. The newly formed RNA can either serve as a template for more RNA strands, or be packed into new virus, or be translated to produce more proteins.

6. Ribozymes, or RNA enzymes, are catalytic RNA molecules. They catalyze covalent changes in the structure of other (mostly RNA) molecules. Here are several examples. The group I and II introns are able to catalyze their own splicing out of the primary RNA transcript to form mature RNAs. RNase P is an RNA processing endonuclease that specifically cleaves the 5' end of the precursor tRNA. It consists of an RNA of approximately 350 nucleotides bound to a protein of approximately 120 amino acids. The hammerhead and hairpin ribozymes are small catalytic RNA motifs that catalyze self cleavage of the strand. They are involved in the replication of plant viroid and viroidlike satellite RNAs (small, circular RNA molecules of 300-400 nucleotides that infect plant cells. Unlike viruses, they don't have a protein shell. Their precursor RNA contains many repeats of viroid structure, which must be cut out and ligated). Recently, it was shown that the 23S rRNA in the large (50S) subunit of ribosome catalyses the formation of the peptide bond that links each amino acid to the growing polypeptide chain, i.e., it is a ribozyme.

   Before the discovery of ribozymes, only proteins were known to have catalytic activity. The fact that RNA can serve as a catalyst in addition to its ability to store genetic information, provided the support to the theory of "RNA world". It states that there was a time shortly after the origin of life on the Earth when RNA alone carried out all biological functions required for a cell to survive, i.e. served as the genetic material, structural and catalytic molecule.

1.3 The forces that stabilize RNA structures.

In the primary structure of RNA the residues are covalently bonded to each other. The free energy change due to formation of a covalent bond is $\Delta G \simeq 100 k_B T$ ($k_B T \simeq 0.6 kcal/mole$).

The free energy change associated with the change of conformation of the molecule depends on the enthalpy change (heat associated with the reaction at constant pressure), ΔH, and on the entropy change ΔS of the system:

$$\Delta G = \Delta H - T\,\Delta S. \tag{1.1}$$

All thermodynamic variables refer to standard state (1M concentration of species, pressure 1 atm). The ΔH is determined by the potential energy of interaction between nucleotides, $\Delta H = \Delta(E + PV) \simeq \Delta E$ (changes of pressure and volume are almost always negligible for conformational transitions). The energy E is the sum of energies of bonded and non-bonded interactions.

The bonded interactions are stretching and bending of bonds between the nearest neighbors, and rotation of torsion angles. The potential energy of such interactions can be expressed by the formula:

$$E_b = \sum_{\text{bonds}} k_r (r - r_{eq})^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + cos(n\phi)], \tag{1.2}$$

where k_r and k_θ are force constants for bond stretching and bending, $r - r_{eq}$ and $\theta - \theta_{eq}$ are deviations of bond lengths and angles from the equilibrium values, V_n is the energy barrier of rotation, and n is 3 for rotation around the bond linking two tetrahedral atoms (sp^3 bonding) and 2 for sp^2 bonding.

Non-bonded are interactions between nucleotides separated by three or more bonds. Those are: (i) electrostatic Coulomb's interactions; (ii) London attractions (fluctuation dipole-induced dipole interactions) which decays with distance as r^{-6}; (iii) van der Waals short- range repulsions proportional to r^{-12}. The energy of all such non-bonded interactions can be written as

$$E_{nb} = \sum_{\mathrm{i<j}} \left(\frac{q_i q_j}{\epsilon R_{ij}} - \frac{A_{ij}}{R_{ij}^6} + \frac{B_{ij}}{R_{ij}^{12}} \right). \tag{1.3}$$

Here q_i, q_j are charges, R_{ij} is the distance between charges i and j, ϵ is the dielectric constant of the medium, and constants A_{ij} and B_{ij} are positive and depend on the interacting atoms i and j.

When we deal with polymer molecules in aqueous solution, we have to distinguish between translational and conformational entropy. Bringing of two strands together to form a duplex, and binding or release of water molecules and of ions result in change of the translational entropy. If a single-stranded molecule folds to form a compact structure, the conformational entropy decreases because of the restrictions imposed on the torsion angles. The Boltzmann's statistical interpretation of entropy relates entropy S to the number Ω of states of the system with the same energy:

$$S = k_B \, ln\Omega. \tag{1.4}$$

Here k_B is the Boltzmann's constant. In order to reach the stable state, which, according to laws of thermodynamics, is the one with minimal free energy, the system tends to minimize enthalpy and maximize entropy (Eq. 1.1).

The stability of the secondary structure is mainly influenced by three factors: base stacking, base pairing and flexibility of the backbone.

(1) Base stacking is the dominant stabilizing force in an RNA structure. It was found experimentally, that separate bases or nucleosides in water prefer to stack (form columns of several flat bases) rather than to form the hydrogen bonds between complementary pairs. This is an enthalpy driven process, i.e. there exist attractive forces between planar aromatic bases. The bases in stack are held together mainly by London attractions of the polarizable electrons (induced dipole interactions between the π electron clouds of the stacked bases) and Coulombic electrostatic interactions among the net charges on base atoms.

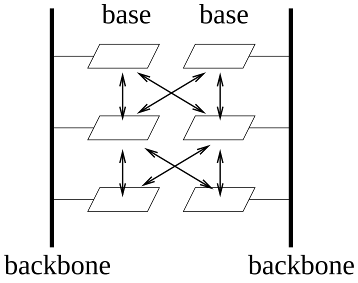


Figure 1.4: The stacking interactions occur between all neighboring bases and significantly stabilize RNA helices.

The stacking interactions significantly stabilize RNA helices. They occur not only between consecutive bases of one strand, but also between bases belonging to neighboring strands (cross-strand stacking, Fig. 1.4). The stacking interactions can also cause so called coaxial (end-to-end) stacking of double helices which is energetically even more favorable then forming of one long helix because of some conformational freedom in the junction region.

(2) Base pairs are formed as a result of hydrogen-bonding interactions between base edges. The hydrogen bonds form between hydrogen atoms with partial positive charges, and oxygen, nitrogen or fluorine atoms with partial negative charges. The nature of hydrogen bonds is mainly electrostatic with a small covalent component. In addition to the standard or canonical Watson-Crick base pairs, many other hydrogen-bonded base pairs (e.g. wobble and Hoogsteen base pairs, see Fig.2) can form; the backbone can interact with bases or with other sugar and phosphate groups. The fact that in usual double helices mostly Watson-Crick base pairs are found is due to their remarkable isostericity, which allows each of four combinations to fit into the A-form helix without it's distortion. The hydrogen-bonding interactions between bases are weak in comparison with base stacking because of the competition with the hydrogen bonding with water.

The formation of the base stack is associated with the free energy change which has enthalpic and entropic contributions. The enthalpic contribution is due to the described non-covalent interactions: stacking forces and hydrogen bonding between bases. The entropy change for base stack formation is caused by two reasons: the "freezing" of the chain conformational entropy because of the restrictions imposed on the (7 per nucleotide) torsional angles of the backbone and of the base, and the change of the solvation entropy, i.e. the entropy of ions and water molecules. In the present research, we focus only on the conformational entropy change. Both, ΔH and ΔS of the stack are sequence-dependent. The sequence-dependence of the stack entropy comes from (i) the dependence of the solvation entropy on the chemical structure of bases, and (ii) the dependence of the degree of the loss of rotational freedom on the strength of the interactions between bases: the larger ΔH is associated with the larger ΔS.

The stacking thermodynamic parameters for different base stacks have been measured by Turner et al [1] . For example, for the stack of $G - C$ and $U - A$ base pairs in $1M$ NaCl and $T = 37°C$,

$$\Delta H = -10.2 \ \ kcal/mol; \ \ \Delta S = -26.2 \ \ cal/(K\, mol);$$

$$\Delta G = \Delta H - T\,\Delta S = -2.1 \ \ kcal/mol.$$

(3) Most of the conformational flexibility of the RNA molecule is induced by it's backbone. The backbone of RNA is a highly charged polyelectrolite, because each phosphate residue bears a unit negative charge. The resulting electrostatic repulsion between two strands (or two parts of one strand) reduces flexibility and acts against the helix formation. However, the presence in aqueous solution of positively charged counterions and their condensation into a small volume around the RNA molecule partly neutralizes the negative backbone charges and can cause the helix formation. Localization of multivalent counterions is more favorable since less particles should be localized to neutralize the same charge which provides significant entropic advantage.

1.4 Structure and conformational changes determine RNA functions.

RNA is extremely versatile and flexible molecule which plays a variety of roles due to its ability to form different structures and interact with many other macromolecules. Many RNA functions are based on specific RNA structures, conformational changes and stability. To demonstrate it, we consider two examples.

The translational repression mechanisms [4, 5]. Translation in prokaryotes is initiated by formation of ternary complex between $30S$ subunit, tRNA which binds with methionine $fmet - tRNA_f^{met}$, and mRNA initiation sequence (initiation codon AUG and a sequence complementary to the $3'$ end of $16S$ rRNA called Shine-Dalgarno sequence) (Fig. 1.5). First, the sequence at the 3' end of $16S$ rRNA binds to Shine-Dalgarno sequence; then, this preinitiation complex undergoes slow isomerization to the final initiation complex in which $fmet - tRNA_f^{met}$ in the ribosome P-site is paired with the initiation codon on mRNA. After that, the process of translation starts, the mRNA molecule moves through the ribosome to synthesize the protein molecule.

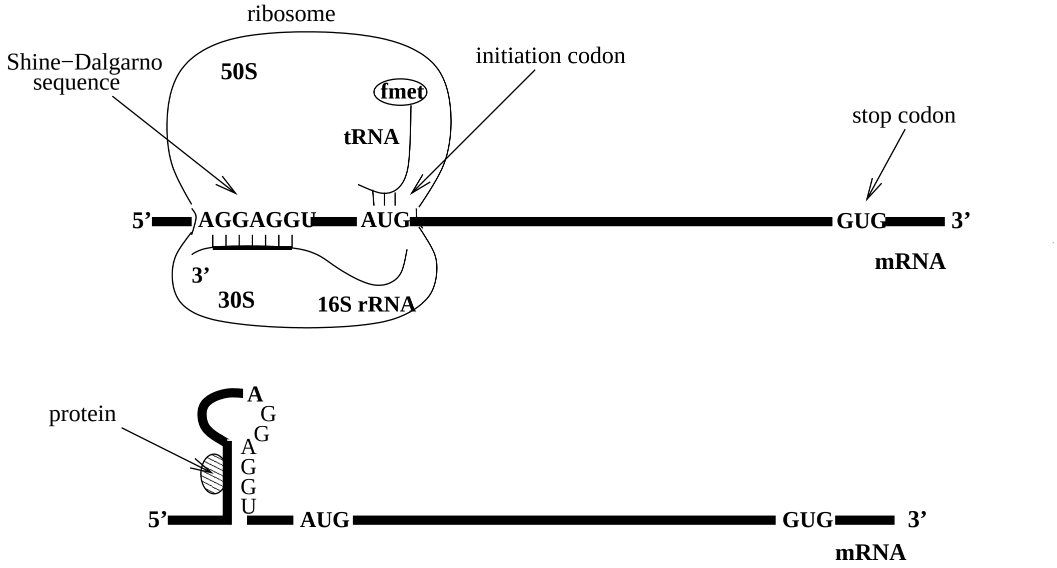


Figure 1.5: Translation is initiated by formation of ternary complex between $30S$ subunit,$fmet - tRNA_f^{met}$, and single-stranded mRNA initiation sequence. Folding of mRNA can prevent the ribosome binding and repress the protein production.

The recognition of a ribosome binding site (RBS) requires the mRNA to be locally single-stranded. Therefore, the binding of the ribosome can be prevented by folding of the messenger RNA, for example, formation of a hairpin containing the mRNA initiation sequence. Whether it will totally prevent the ribosome binding, depends on the stability of the structure. The experiment has been carried out [6] where the stability of the hairpin structure at the coat-gene RBS from bacteriophage MS2 has been varied. The individual base pairs were disrupted and restored by site-directed mutagenesis, and the relative expression (the fraction of mRNA that is bound by a $30S$ subunit) of each mutant was measured. The experiment has clearly shown that translation decreases gradually as the structure is stabilized [6]. The data from other sources [7]-[10] have been compiled [5] which support control of translation by the stability of the mRNA structure at the RBS. The base-pairing at the RBS doesn't totally prevent the ribosome binding because RNA structures don't exist permanently: due to thermal fluctuations the molecule continually and spontaneously folds and unfolds. The fraction of time which each molecule spends in the unfolded state, and therefore the fractional population of the molecules in the unfolded state depends on the structure stability. Since the probability of $30S$ subunit binding depends on the fraction of mRNAs in the unfolded state, it also decreases with the stabilizing of the folded mRNA structure.

Process of splicing of introns [13]. The process is assisted by a ribonucleoprotein complex called spliceosome which consists of five snRNPs: U1, U2, U4, U5 and U6 and other protein factors. The intron splicing occurs in two steps. In the first step, the 2' hydroxil group of a conserved adenosine residue near the 3' end of the intron (so called intron branch site) approaches the 5' end of the intron and causes cleavage of the exon 1 -intron junction. In the second step, the 3' hydroxil of exon 1 disrupts the intron - exon

2 connection and binds exons. The details of the process are shown in Fig. 2 in [13]. Most introns have common consensus sequences at their ends which participate in the spliceosome formation. The process starts from the base-pairing of U1 and U2 snRNPs with two ends of the intron. The U2 duplex bulges the branch-point adenosine. Then, the base-paired U4-U6 and U5 get involved into the conformational changes, namely, duplex U4-U6 unwind, U4 and U1 are displaced, and U6 base-pairs with 5’ splice site and part of U2. The U5 snRNP is believed to base-pair with both exons in order to position them for the second step of splicing. After the process is completed, the spliceosomal components dissociate. Thus, the process of intron splicing is intrinsically associated with multiple structural rearrangements of snRNAs.

To understand how RNA functions and to manipulate its functions, one should have information about (i) specific RNA structures, including interactions with other RNAs and proteins; (ii) thermodynamic stabilities of structures and (iii) kinetics, i.e. how fast is the given structure formed.

Chapter 2

OVERVIEW ON RNA FOLDING EXPERIMENTS.

2.1 Measuring thermodynamic parameters.

The main methods by which thermodynamic parameters for nucleic acid conformational changes and structural stability can be determined, are differential scanning calorimetry (DSC) and temperature dependent optical spectroscopy (melting curves). In DSC, two identical cells with reference solution are taken, one of them containing the sample of interest (RNA molecule). The electrical energy is used to gradually increase the temperatures of both cells from some initial T_i to some final T_f value. The temperature of the transition is supposed to be between T_i and T_f. The energies required to raise the temperature in both cells are measured and their difference found in order to cancel out heat changes due to effects of no interest. Such measured heat capacity is called excess heat capacity (C_p). C_p is associated with the transition (heat absorbed by the conformational change of the molecule). The resulting data is usually plotted as C_p versus T.

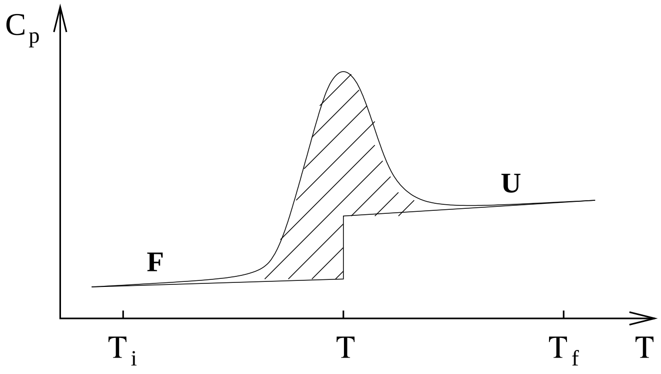


Figure 2.1: The excess heat capacity dependence on temperature for the single transition from state F to state U. The transition temperature T is between initial T_i and final T_f temperatures.

For a single two-state transition from folded state F to unfolded state U the curve will have a peak in the transition region like one shown in Fig. 2.1. The area under the curve for the given temperature interval is the total enthalpy change ΔH_{tot} of the sample system:

$$\Delta H_{tot} = \int_{T_i}^{T_f} C_p dT, \tag{2.1}$$

which includes heat absorption associated with the transition ΔH and heat absorption by molecules in states U and F (for example, as a result of the solvent heat absorption). Subtracting the later heat from ΔH_{tot} gives ΔH (see Fig. 2.1):

$$\Delta H(T) = \Delta H_{tot} - \int_{T_i}^{T} C_p^U dT - \int_{T}^{T_f} C_p^F dT, \tag{2.2}$$

where C_p^U and C_p^F are heat capacities of the molecule in unfolded and folded states. In this way, the ΔH at any temperature T within interval (T_i, T_f) can be calculated. If two peaks are seen in a $C_p(T)$ melting curve, it means that at least two different transitions occur at two different temperatures. Sometimes it is difficult to separate two transitions, especially if their transition temperatures are close to each other.

The advantages of optical methods are the less amount of material required ($< 1mg$ whereas calorimetry requires $1 - 5mg$ per experiment) and more quickly obtained results (because a spectrophotometer equipped with the automatic cell changer can run several samples simultaneously). The optical methods for measuring thermodynamic parameters associated with the molecule conformational changes are based on two equations. The first of them connects the free energy change ΔG with the equilibrium constant K:

$$\Delta G = -RT lnK; \qquad K = \frac{[U]}{[F]}, \tag{2.3}$$

and the second, the van't Hoff equation, describes the relationship between the enthalpy change ΔH for the reaction and the equilibrium constant temperature derivative:

$$\frac{\partial ln\, K}{\partial T} = \frac{\Delta H}{RT^2}. \tag{2.4}$$

Using the above two equations we can obtain ΔH, ΔS and ΔG from K. The absorption spectroscopy is the most common method to determine temperature dependence of the equilibrium constant.

As light passes through the absorbing solution, its intensity decreases exponentially according to Beer-Lambert law:

$$I = I_0\, 10^{-\epsilon c l},$$

where I_0 and I are incident and transmitted light intensities, l is the path length (cm), c is the molar concentration, and ϵ is the molar extinction coefficient ($M^{-1}cm^{-1}$). ϵ is a function of the wavelength of the exciting light. The absorbance is

$$A = log(I_0/I) = \epsilon c l$$

The interactions between nucleotides such as base pairing and stacking lead to the decrease in absorption. The increase in temperature causes the unstacking of base pairs and increase in absorption. The measured absorption melting curve (absorption vs. temperature) provides information about the concentration (population) of different structures from which we can obtain thermodynamic parameters from Eqs. 2.3 & 2.4. The melting curves can be repeated at two different wavelengths: breaking of C-G base pairs produce the largest increase in absorption at 280nm, and A-U base pairs - at 260nm. Comparison of melting curves recorded at different wavelengths can give additional information about the relative amounts of melted C-G and A-U base pairs and can be helpful in the case of multiple close transitions.

The two-state analysis for RNA oligomers has been used by Turner et al [1] to obtain the thermodynamic parameters for base pair stacks and simple loops. Consider the equilibrium reaction when two oligomer strands form a duplex: $S_A + S_B \rightarrow D$. It is assumed that a given strand can exist in only two states: as a single strand or in a duplex, i.e. a two-state model is used. One way to obtain thermodynamic parameters for duplex formation is by fitting the shape of optical melting curve by the model parameters as described above (see Eqs. 2.3 & 2.4). The alternative way is to use the plot of inverse melting temperature $1/T_m$ vs lnC_t. The equilibrium constant for this reaction is $K = [D]/[S_A][S_B]$, total strand concentration $C_t = 2[D] + [S_A] + [S_B]$, and the fraction of strands in duplexes in solution $f = 2[D]/C_t$. Then the equilibrium constant can be expressed as ($[S_A] = [S_B]$):

$$K = \frac{2f}{C_t(1-f)^2}.$$

Solving Eq. 2.3 for T, using the obtained expression for K, and taking into account that at melting temperature $f = 1/2$, we have

$$\frac{1}{T_m} = \frac{R}{\Delta H}\ ln\frac{C_t}{4} + \frac{\Delta S}{\Delta H}.$$

If strands A and B are self-complementary, i.e. the reaction is $2S \rightarrow D$ with $C_t = 2[D]+[S]$, then $C_t/4$ should be replaced by C_t. The plot of lnC_t against $1/T_m$ is a straight line, and the energy parameters ΔH and ΔS for the reaction can be determined.

These two analyzes give similar results if the transition fits the two-state model, and different results otherwise. Calorimetry provides another test for adherence to the two-state model. Most of thermodynamic studies of oligomers are done in 1M NaCl, which is a common physiological salt concentration.

Thermodynamic data have been measured for many different transitions in RNA. It is important to be able to systematize these data so that the results can be extrapolated, and can be used to estimate thermodynamics of other transitions and to predict the conformations with the lowest free energy for other sequences. As a rough approximation, the additive **nearest-neighbor model** has been proposed. It is based on the assumption that the stability of each base pair depends on its nearest neighbors. The assumption is justified by the fact that the short-range interactions, hydrogen bonding and base stacking, mainly contribute to the structure stability. Therefore, the free energy of each structure is estimated by the sum of free energies of its constituent elements: base pair stacks, loops and dangling ends. The thermodynamic parameters are measured for RNAs that contain the different secondary structural elements, and then the contributions from elements determined using the above methods. These parameters are tabulated as a “periodic table” and can be used to compute the free energy for an arbitrary RNA structure. There are ten possible combinations of adjacent base pairs (two base pair stacks) in RNA helices:

5’AA3’ 5’AC3’ 5’AG3’ 5’AU3’ 5’CA3’ 5’CC3’

3’UU5’ 3’UG5’ 3’UC5’ 3’UA5’ 3’GU5’ 3’GG5’

5’CG3’ 5’GA3’ 5’GC3’ 5’UA3’

3’GC5’ 3’CU5’ 3’CG5’ 3’AU5’

For example, the thermodynamic parameter (ΔH, ΔS and ΔG) for the duplex

$$\begin{array}{c} 5'ACUGG3' \\ 3'UGACC5' \end{array}$$

is the sum of the initiation parameter I which accounts for the translational entropy loss due to the bringing two strands together, and of thermodynamic parameters of each stack:

$$I + \begin{array}{c} 5'AC3' \\ 3'UG5' \end{array} + \begin{array}{c} 5'CU3' \\ 3'GA5' \end{array} + \begin{array}{c} 5'UG3' \\ 3'AC5' \end{array} + \begin{array}{c} 5'GG3' \\ 3'CC5' \end{array}$$

To obtain the thermodynamic parameters ΔG, ΔH, and ΔS for each stack, an overdetermined set of oligonucleotides has been studied and the best least-squares values of the parameters found. Experimental data on loop energy parameters are not extensive and partially not reliable. The free energies of some small loops have been determined and extrapolated for larger loops. The obtained by Turner et al thermodynamic parameters for secondary structure elements have been improved by other researchers ([2],[3]) and are collectively referred to as the “Turner rules”.

2.2 Experiments on folding-unfolding kinetics.

There are several ways to fold or unfold an RNA molecule. It can be the change of temperature or ionic concentration, or mechanical force applied to the molecule.

Temperature-jump relaxation spectroscopy.

The kinetics of folding can be measured by the temperature-jump relaxation spectroscopy. The idea is the following. The system is initially taken to be in the equilibrium state. Then the temperature of the system is raised a few degrees very rapidly ($10^{-6}sec$ or less) and the process of the system to approach the new equilibrium state is followed spectroscopically. For the case of stacked and unstacked dinucleoside monophosphate, the concentration of unstacked species is changing exponentially:

$$\Delta[U] = \Delta[U]_0 \; e^{-t/\tau},$$

where $\Delta[U]_0$ is the difference in concentration just before the T-jump and in the new equilibrium state, t is time and τ is the relaxation time. Since for a unimolecular reaction $\tau^{-1} = k_1 + k_{-1}$, the forward and reverse rate constants k_1 and k_{-1} can be derived from experimental values of τ if the equilibrium constant $K = k_1/k_{-1}$ is known. Typically, the rate constants for single strand stacking are about $10^7 s^{-1}$.

Folding/unfolding caused by the change of ionic concentration.

The RNA folding/unfolding kinetics has been extensively studied for the group I ribozyme from *Tetrahymena thermophila*.

The structure and activity of the ribozyme largely depend on the presence of Mg^{2+}

ions, which form a dynamic counterion “atmosphere” around RNA. In the experiments, the folding process was initiated by the addition of ions Mg^{2+}, and was followed by means of time-resolved small-angle X-ray scattering (trSAXS). It was found that the addition of Mg^{2+} causes a rapid compaction of the molecule from an extended secondary fold to a globular state of nearly native dimensions. The compaction was found to be substantially faster than the formation and stabilization of any known tertiary contacts. The collapse occurs in two distinct kinetic phases, with time constants of 10 ms and 100 ms. The time constant of tertiary contacts formation is of the order of seconds. The proposed pathway and physical mechanisms of rapid compaction are the following [14].

At the beginning of the folding, strong Coulomb repulsion between different segments of the molecule causes largely extended RNA conformation, helices push away each other. After the addition of 10 mM Mg^{2+}, the electrostatic repulsion is screened by divalent counterions, and the molecule relaxes to a partially collapsed self-avoiding conformation. The time constant of this electrostatic relaxation, 10 ms, is large compared to the time constant of simple hairpin (secondary structure) formation, with tens of μsecs time scale.

At the second stage, the ribozyme collapses on a 100 ms timescale to a globular state. Because mutations that disable the formation of the long-range tertiary contacts suppress this phase of collapse, it was suggested that the formation of some of the native tertiary contacts is critical for formation of this globular state. However, tertiary contacts formed at this stage appear to be unprotected from solvent in hydroxyl radical footprinting studies. The proposed working model is that the “tertiary collapse” leads to an ensemble of conformations in which formation of tertiary contacts is unsynchronized or transient and therefore not buried from solvent in the majority of conformations.

At the third stage of folding, the globular state rearranges on a timescale of seconds to form states with solvent-protected contacts, the stable tertiary structures.

Mechanical folding-unfolding of RNA.

The idea of the experiment is to attach two ends of an RNA molecule to a force- and extension-measuring device, to apply a small force (in pN range) and to measure the distance between beads. The device can be an atomic force microscope (AFM) or optical tweezers. The elastic properties and kinetics of structural transitions of the single molecule can be studied by recording and analyzing force-extension curves. The advantage of single-molecule experiments is the possibility to follow the individual folding-unfolding trajectories (which is difficult in the bulk studies where multiple species and multiple folding pathways are present) and to stretch the molecule along a well-defined coordinate, end-to-end extension.

In AFM, one end of the molecule is tethered to a flat surface (such as mica, gold or glass) and the other end is attached to the AFM tip. The tip is moving away from the surface, the molecule gets stretched and unfolds. The tip is attached to the end of a cantilever and the force can be measured through it's deflection. In optical tweezers, two beads are attached to the ends of the molecule. One bead is held in a force- measuring optical trap (force is determined by measuring the deflection of the trapping laser beams with position-sensitive photodetectors), the other is linked to a piezoelectric acutator through a micropipette to control the position.

For example, experiments on mechanical unfolding and refolding have been used to study three small RNA molecules (Fig. 1 in [15]): a simple hairpin P5ab, three-helix junction P5abcΔA and the P5abc domain of the *Tetrahymena thermophila* ribozyme (three-helix junction with the bulge) have been performed by Liphardt et al [15].

The molecules were stretched using optical tweezers. For the simple hairpin P5ab the force increased monotonically with extension, but there existed the critical force region at approximately 14pN where the hairpin has shown to be bi-stable and hop between folded and unfolded states in less than 10ms and without intermediates. (Fig. 1A in [15]) When the force was kept at the transition value, the molecule was shown to spend equal time in folded and unfolded states and its end-end distance hopped back and forth by 18nm, indicating the molecule being in folded and unfolded states. The increased/decreased force resulted in increased/decreased time spent in unfolded state. The experiment has been performed in 250mM NaCl, 10mM Mg^{2+} and at 25°C.

The three-helix junction P5abcΔA demonstrated the same hopping between folded and unfolded states, but with a slower kinetics compared to hairpin P5ab due to the presence of two kinetic barriers.

The relatively complex structure of the third molecule, P5abc, is stabilized by Mg^{2+}-dependent tertiary interactions between the P5c helix and the A-rich bulge. The tertiary interactions cause yet even slower kinetics and therefore the absence of the fast hopping between folded and unfolded states in Mg^{2+}.

Chapter 3

STATISTICAL THERMODYNAMICS FOR RNA SECONDARY STRUCTURE FOLDING.

At the center of the statistical thermodynamics is the partition function $Q(T)$, defined as the weighted sum over all the possible conformational states:

$$Q(T) = \sum_{conf} e^{-E/k_B T}, \tag{3.1}$$

where k_B is Boltzmann's constant, T is the temperature, and E is the energy of an individual conformation.

In terms of the polymer graph (see section 1.1), the partition function can be calculated as a sum over all possible graphs instead of all possible conformations, which is computationally more efficient:

$$Q(T) = \sum_{\text{graph}} \Omega \quad e^{-E/k_B T} = \sum_{\text{graph}} e^{-\Delta F/k_B T}, \tag{3.2}$$

where Ω is the number of viable chain conformations that satisfy the constraints on the intrachain contacts depicted by the graph, and $\Delta F = E - k_B T \, ln\Omega$ is the free energy of the ensemble of conformations for the given graph. We assume that the intrachain contacts determine the energy E, as a result, a graph represents an equal-energy conformational ensemble. Strictly speaking, the free energy should be understood as the potential of mean force averaged over solvent configurations as well as chain conformations. Considering the Gibbs free energy $\Delta G = \Delta(F + PV) \simeq \Delta F$, we can rewrite Eq. 3.2 as

$$Q(T) = \sum_{\text{structure}} e^{-\Delta G/k_B T}. \tag{3.3}$$

Here a structure is defined as the macrostate of conformation that satisfy the constraints imposed by the graph and ΔG is the Gibbs free energy of the structure.

3.1 The additive nearest-neighbor model for the free energy ΔG of an RNA secondary structure.

The nearest-neighbor model which has been used to obtain thermodynamic parameters for RNA secondary structural motifs [1] assumes the mutual independence of structural units. As a result, the free energy of the secondary structure is calculated as a sum of free energies of all base pair stacks and loops.

For example, for the fragment of $5S$ rRNA in Fig. 3.1 the folding free energy is calculated as:

$$\Delta G = G_{(folded)} - G_{(unfolded)} = -1.5 - 0.5 - 0.6 - 1.5 + 0.8 - 1.8 - 2.9 + 5.9 = -2.1 kcal/mole$$

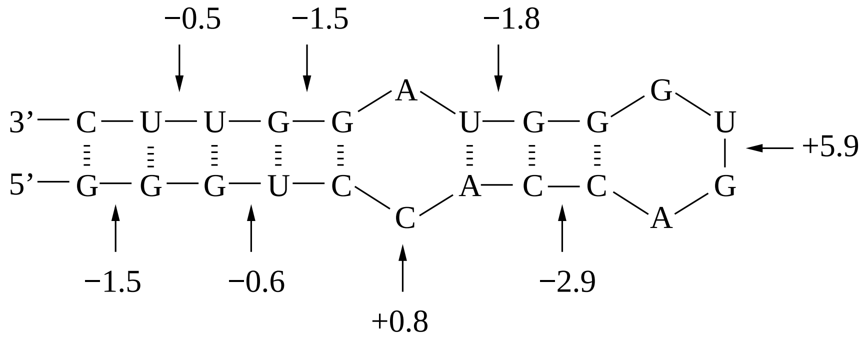


Figure 3.1: Fragment of 5*S* rRNA from *Philosamia cynthia ricini*. The stability of the structure can be calculated using the nearest-neighbor model. The free energy parameters for loops and base pairs are predicted by Turner et al [1].

The given structure is stable if its free energy is lower then the free energy of the unfolded state, i.e. $\Delta G = \Delta H - T\Delta S < 0$. Therefore, for the stability of the structure the low enthalpy (stronger interactions within the structure) and the large entropy (more degrees of freedom, which means weaker intrachain interactions) are favorable. Thus, there exists a competition between stacking, which lowers the enthalpy, and loop formation, which lowers the entropy contribution to the free energy of the structure. The increase in temperature favors unfolding of the structure.

Assuming additivity of free energy and entropy, the nearest-neighbor model neglects the inter-subunit interferences. For example, the excluded volume interactions (impossibility for two monomers to occupy the same site) between different subunits can cause nonadditivity in the free energy and cause the chain entropy to be smaller than the sum of the entropies of all individual structural subunits. For example, for a 58-mer model secondary structure (two-dimensional lattice), neglecting the inter-subunit excluded volume interferences can result in a relative error of 40% in a total entropy [16].

3.2 The nonadditive polymer principle model for RNA secondary structures.

In the recently developed theory [16]-[18], the loop entropies have been evaluated theoretically using the lattice polymer model. The theory is based on graphical representation of intrachain contacts and goes beyond the additivity in the free energy by considering the (excluded volume) interactions between different secondary structural subunits at the junction regions.

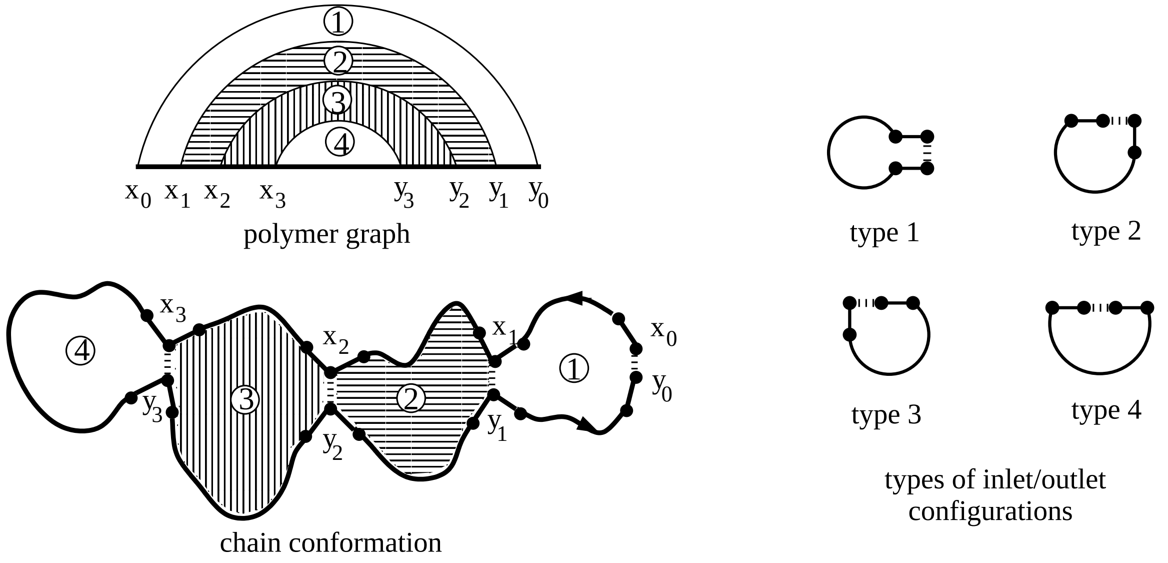


Figure 3.2: A polymer graph and the corresponding firehose-like chain conformation. The shaded regions are subunits of the graph and the conformation. The whole graph (structure) in the figure can be divided into four subunits. Excluded volume interactions between subunits are predominantly localized at junction regions, and on 2D lattice we take care of them considering four types of inlet/outlet configurations.

A key issue in the partition function calculation is the computation of the number of chain conformations Ω in Eq. 3.2 for the given graph. Central to the computation of Ω is the excluded volume effect. In the theory of Chen & Dill [16]-[18] and Zhang & Chen [19] the secondary structure is subdivided into the same subunits as in the nearest-neighbor model. The subunits are defined as regions in the graph enclosed by the links but containing no links in its interior. According to this definition, each subunit corresponds to a base stack or a loop without self-contacts. For example, four subunits connected to each other in a firehose manner are shown in Fig. 3.2. Unlike nearest-neighbor model, the subunits are not considered to be mutually independent, but excluded volume interactions are assumed to occur mainly at junction between two subunits. Therefore, to account for the inter-subunit excluded volume interactions, the subunit conformations have been classified according to their inlet and outlet configurations. On 2D and 3D lattices, the inlet and outlet of each subunit can have one of four or six possible configurations, respectively (Figs. 3.2 and 3.3).

Two matrices have been defined:

1. the structure matrix **S** with matrix elements S_{nm} denoting the number of subunit conformations with the types of inlet and outlet configurations being n and m. The S-matrix have been calculated by exact computer enumeration of self-avoiding walks for small subunits and extrapolated for larger ones. In this way, the intra-subunit excluded volume effect is treated exactly.

2. the viability matrix **Y**, where elements $y_{nm} = 1$ or 0 if connection between type n inlet and type m outlet configurations is viable or not viable, respectively. Thus the **Y**-matrix accounts for the inter-subunit excluded volume.

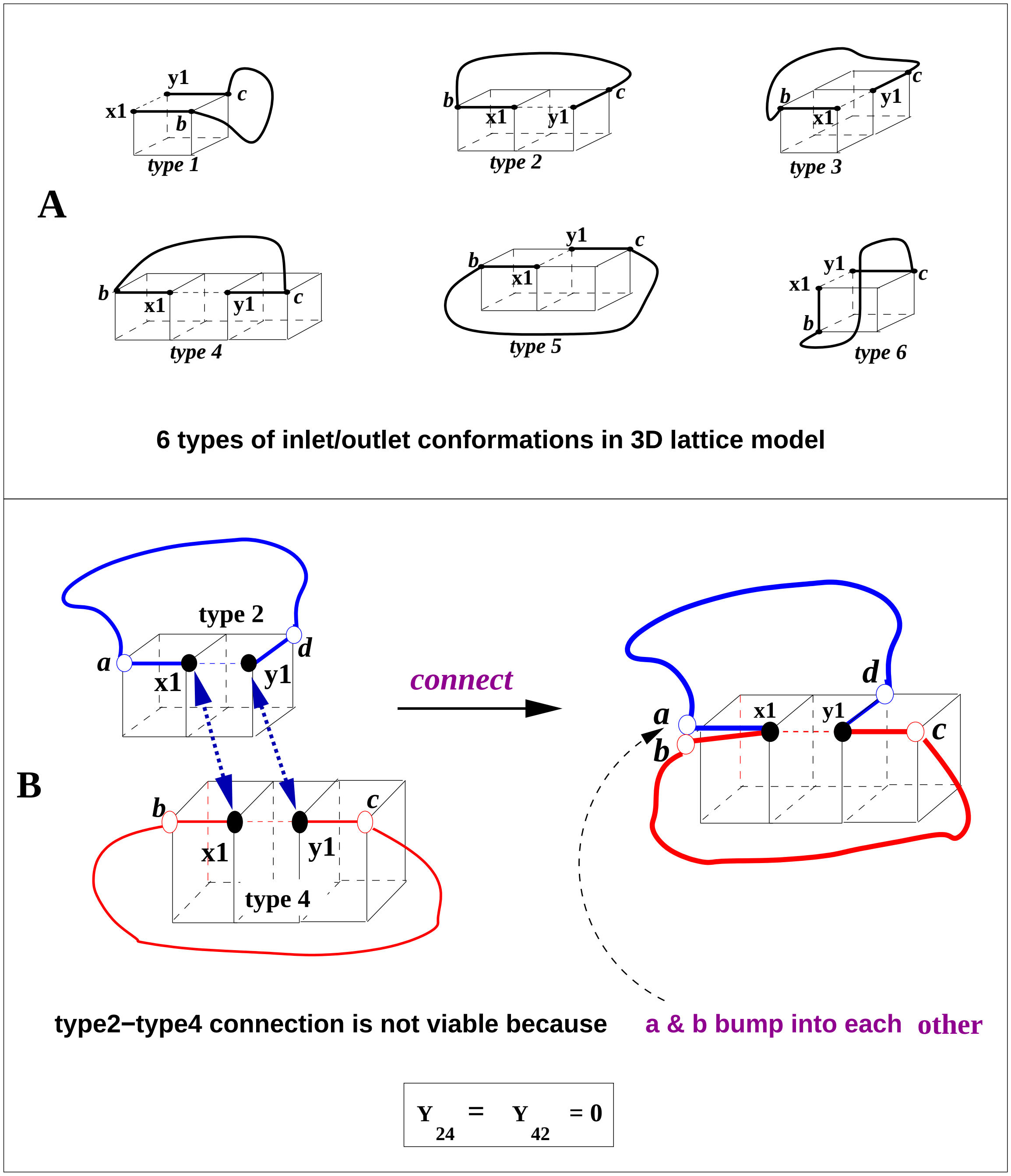


Figure 3.3: (A) On 3D lattice subunits are classified according to six types of inlet/outlet configurations. (B) the viability for the connections between different types of subunits.

Using this matrices, the Ω of the firehose-like structure (as in Fig. 3.2) is estimated by $\mathbf{U} \cdot \mathbf{\Omega}[x_0, y_0] \cdot \mathbf{U}^{\dagger}$, where $\mathbf{U} = [1, 1, 1, 1]$ and $\mathbf{U}^{\dagger}$ is the transpose of $\mathbf{U}$, and matrix $\mathbf{\Omega}[x_0, y_0]$ is obtained by matrix multiplication:

$$\mathbf{\Omega}[x_0, y_0] = \mathbf{S}^{(N)} \cdot [\prod_{j=1}^{N-1} \mathbf{Y} \cdot \mathbf{S}^{(j)}],$$

where $\mathbf{S}^{(j)}$ is the structure matrix of the jth subunit.

The advantages of the method are: (i) division of the graph into subunits and factorability of the partition function into subunit components makes the method both accurate and efficient: computation of the subunit number of conformations requires much less computational time than of the whole chain and can be done with higher accuracy; (ii) the intra- and inter-subunit excluded volume interactions are explicitly accounted for.

In the 3D lattice model, each chain segment is less restricted than in 2D because of the larger degrees of freedom. Therefore, we need to consider the excluded volume interactions between next-nearest-neighbors. It results in 113 types of the inlet/outlet configurations and is unfavorable for the computational efficiency. Therefore, Zhang and Chen [19] proposed a mean-field approach by using the **S** and **Y** matrices averaged over all the possible configurations of the next-nearest-neighbors.

The absence of long-range correlations in secondary structures makes it possible to apply a recursive algorithm to computation of the partition function (Eq. 3.2). In the algorithm, the partition function of the chain of length $(i+1)$ is calculated using the partition function of the chain of length i. The partition function computational time depends on the chain length L as L^6 ([17], [19]).

For the realistic RNA chains the partition function calculations use: (i) the graph energy E calculated as the sum of the experimental enthalpies (Turner Rule) of each base pair stack; (ii) the conformational count Ω obtained by scaling up the lattice model result computed from the above matrix method by an uniform scaling factor.

The heat capacity of the RNA molecule can be computed from the partition function:

$$C(T) = \frac{\partial}{\partial T}[k_B T^2 \frac{\partial}{\partial T} lnQ] \tag{3.4}$$

and compared with the experimental results.

The heat capacities for realistic RNA secondary structures have been calculated using 2D [18] and later - using 3D [19] chain representations. The 3D model makes much better predictions, especially for the width of the melting curves, then the 2D model. The predicted from 2D model melting curve is broader then from 3D model and than the experimental results. It can be explained the following way. For the two-state transition between folded (N) and unfolded (U) states, the sharpness of the transition is known to be proportional to $\Delta S_{UN}^2/\Delta E_{UN}$ (the width of the melting curve is inverse proportional to $\Delta S_{UN}^2/\Delta E_{UN}$). The energy difference between folded and unfolded states ΔE_{UN} is the same for 2D and 3D lattices, and the folded state has a unique conformation in both cases. But for the unfolded state the 3D model allows for much larger diversity of conformations than the 2D model. Therefore, $\Delta S_{UN}^{(3D)} \gg \Delta S_{UN}^{(2D)}$, and the melting curve is sharper in 3D chain representation.

The comparison of melting curves obtained for *E.Coli* 23S rRNA fragment G1051-C1109 using four different models with the experimental one is shown in Fig. 3.4. The McCaskill's recursive algorithm [21] makes unphysical assumptions about loop entropies (loop entropy linearly proportional to the loop length) and ignores the excluded volume effect between structural units.

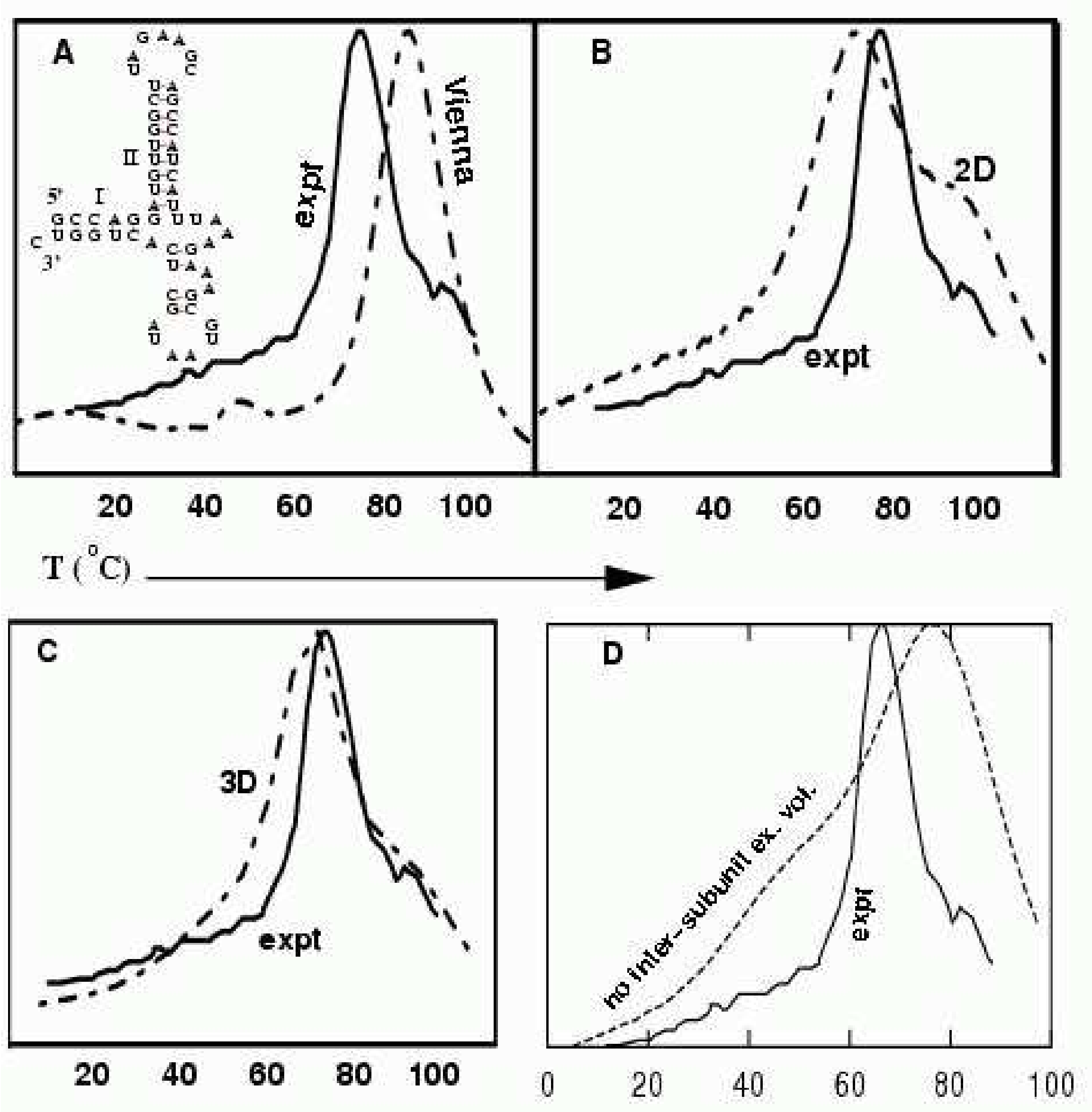


Figure 3.4: Differential melting curves for E.Coli 23S rRNA fragment G1051-C1109 in 1M KCl with 10mM MOPS [20]. The comparisons are between the experimental result (solid line) and the result from (A) Vienna 1.4 package based on the McCaskill's algorithm, (B) the 2D graph-theoretic model, (C) the 3D graph-theoretic model, and (D) a 3D model that neglects inter-subunit excluded volume interactions.

From the comparison of melting curves in Fig. 3.4 it follows that (i) graph-theoretic approach better predicts multiple transitions than McCaskill's algorithm; (ii) 3D model gives better predictions than 2D model; (iii) the excluded volume effect significantly contributes to thermodynamic predictions.

From the partition function, we can obtain the free energy landscape $F(\mathbf{x})$ which is the free energy as a function of the set of parameters $\mathbf{x}$ that characterizes conformational degrees of freedom. The free energy landscape gives information about distribution of free energy of the whole conformational ensemble, from which folding stabilities, the native state(s) with the lowest free energy, and conformational transitions can be predicted. The set of parameters $\mathbf{x}$ should be chosen the optimal way: the number of parameters must be large enough to give the detailed structural information, but small enough so that the landscape is visualizable. Two variables: the number of native (n) and non-native (nn) contacts have been chosen to characterize an RNA secondary structure. The contact (base pair) is called "native" if it is present in the native structure, and "non-native" otherwise.

The free energy landscape has been calculated from the partition function $Q(n, nn, T)$:

$$F(n, nn, T) = -k_B T ln Q(n, nn, T)$$

for an E.coli 23S RNA segment and is shown in Fig. 3.5 [18]. Each point (n, nn) corresponds to the ensemble of conformations that contain n native and nn non-native base pairs. The minima on the free energy landscape correspond to stable and well-populated states, and changes of the minima locations with the temperature identify the structural transitions and degree of their cooperativity. We see (Fig. 3.5) that at T=30°C two almost equally stable conformations coexist: N and Z. The intermediate state Z has many non-native

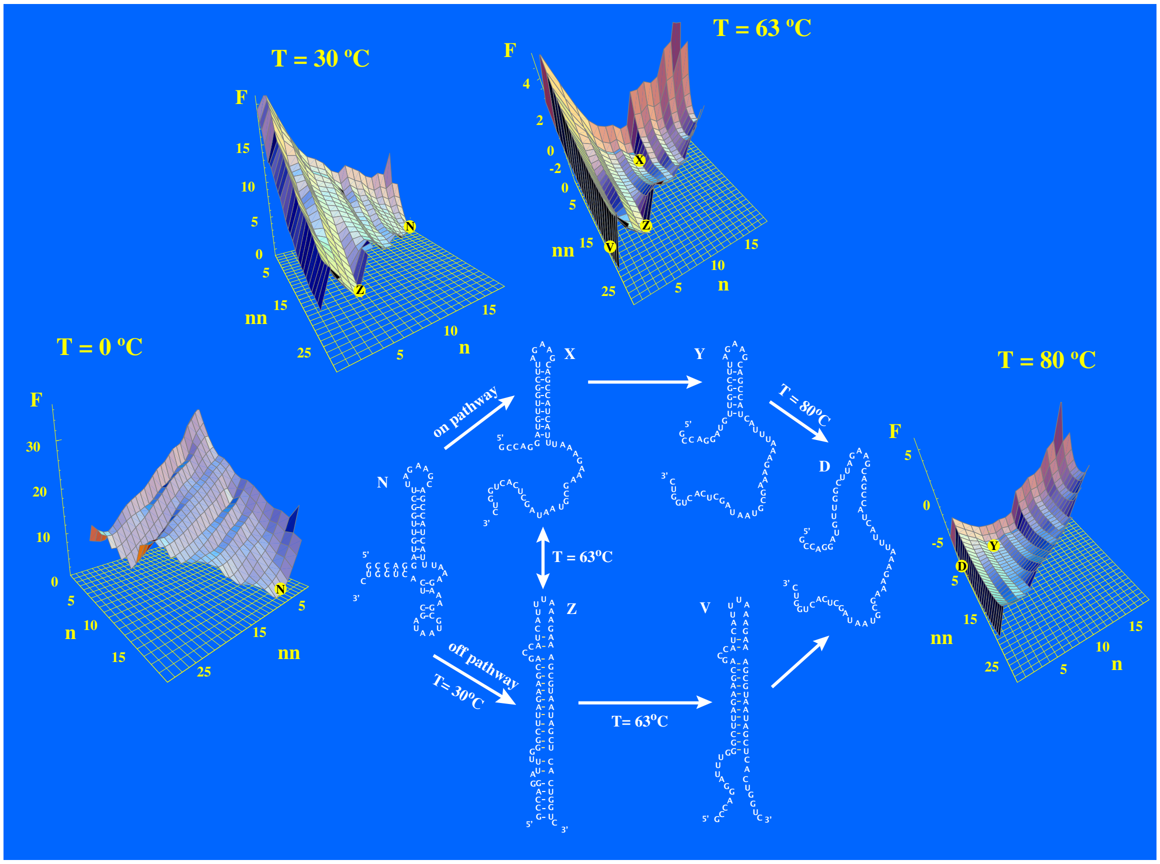


Figure 3.5: Energy landscapes and unfolding pathway for an E.coli 23S RNA segment. The free energies (in kcal/mol) are relative to the native state.

contacts and is a misfolded intermediate. The unfolding of N involves two parallel pathways: through on-pathway intermediates (stable states having a few non-native contacts) and through off-pathway intermediates (can have many non-native contacts).

The theory predicts that RNA secondary structures have a variety of cooperative behaviors: they can have one-state or two-state transitions, stable on-pathway or off-pathway intermediate states. The limitation of the theory is its applicability only to secondary structures.

Chapter 4

STATISTICAL THERMODYNAMICS FOR SIMPLE TERTIARY FOLDS.

This chapter has been published: *Z. Kopeikin and S.-J. Chen. Statistical thermodynamics for chain molecules with simple RNA tertiary contacts. J. Chem. Phys. 122, 094909 (2005)*

4.1 The problems with the modeling of RNA tertiary structure folding thermodynamics.

The tertiary structures is widely occurring class of conformations which is known to play critical structural and functional roles for RNA. But our ability to measure or predict the thermodynamic parameters for such structures is very limited. While the energy parameters for canonical base pairs and simple loops have been obtained from two-state analysis of short RNA duplexes/hairpins melting data, the further attempts to extract the energy parameters for more complex RNA interactions have been hampered by the inability to decipher the melting experiment data. The reason is that the two-state analysis cannot be applied to describe the usually multistate melting transitions of larger secondary and especially tertiary structures. In tertiary structures, the crossing links cause strong correlations between different structural subunits which result in a very convoluted interplay between secondary and tertiary interactions and between different tertiary interactions. The statistical mechanical model is required to enable the deconvolution of the results of melting experiments and the extraction of the energy parameters.

The main problem in theoretical prediction of the folding thermodynamics for tertiary structures is the non-additivity of conformational entropy caused by the inter-subunit excluded volume interactions. The non-additivity effect is much stronger in tertiary structures than in the secondary structures. This is because in secondary structures the inter-subunit interferences mainly occur between neighboring motifs, but in tertiary structures the distant secondary structural motifs are strongly dependent on each other due to long-range tertiary interactions. The calculations based on the additivity assumption and neglecting the inter-subunit correlations lead to a significant inaccuracy in chain entropy and free energy.

Since the entropy is related to the conformational count Ω by formula $S = k_B ln\Omega$, the additivity of entropy would mean that the number of conformations of the system equals the product of numbers of conformations of each subunit, which is true only for mutually independent subunits. That it is not true for tertiary structures, we demonstrate for the simple case of two crossing links, namely, graph and representative conformation shown in Fig. 4.1. The constituent secondary structural motifs are two loops. By means of exact computer enumerations on 2D lattice we find the numbers of conformations for loops

$\Omega_A = 52$ and $\Omega_B = 210$ and for the graph $\Omega = 560$, and see, that $\Omega_A \cdot \Omega_B = 52 \cdot 210 = 10920 \gg 560$. It corresponds to the 47% overestimation of entropy.

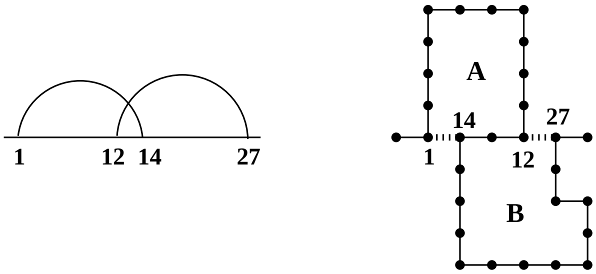


Figure 4.1: The graph and corresponding 2D chain conformation chosen to demonstrate the non-additivity of free energy of chain conformations with crossing links in graphical representation.

It follows from the above example that because of the non-additivity effect we cannot use the energy parameters for secondary structural motifs to compute the entropy and free energy of the tertiary structure.

We present here the thermodynamic model for chains with simple tertiary contacts. The model is based on a graphical representation of intrachain contacts, so the general methodology does not depend on any specific chain representation. To illustrate the theory, the simplest two-dimensional lattice representation of the chain conformations has been used in the present research. The theory explicitly takes into account the excluded volume interactions within and between subunits.

In the following sections, we develop the theory for the calculation of the conformational count Ω by systematically increasing the structural complexity. We start from the simplest structure - conformations with two crossing links. This simplest case would provide a useful paradigm for the treatment of more complicated tertiary folds. We will also apply the theory to investigate how the interplay between secondary and tertiary interactions determine the folding stability, cooperativity and pathway for model RNA chains.

4.2 Conformations with two crossing links.

We first treat conformations that can be represented by graphs that contain only two crossing-linked contacts. Such type of graphs represent a large class of conformations because the graphs can contain an arbitrary number of nested or unrelated contacts and so in general can form complex secondary structures. In the conformational space, the crossing links provide linkers for the different secondary structures.

4.2.1 Basic graphs with two crossing links.

We start with the simplest elementary graphs that contain only two contacts, which are crossing-linked; see contacts $(1, i)$ and $(j, N-1)$ in Fig. 4.2a. Our goal in this section is to develop a theory to count chain conformations for such a graph. To focus on how the curved links affect the conformational statistics, we neglect the dangling tails, which involve no curved links. We use only single monomers (labeled as 0 and N in Figs. 4.2a & b) to account for the excluded volume interactions between the tails and the linked part of the conformation. The full tails will be added back in the final partition function calculation to keep the completeness of the conformational ensemble.

As illustrated in Fig. 4.2b, the whole chain conformation can be decomposed into two loops, A and B. Loops A and B are correlated through (1) the interface from monomer j to monomer i, namely, the (common) chain segment $j \rightarrow i$ should have exactly the

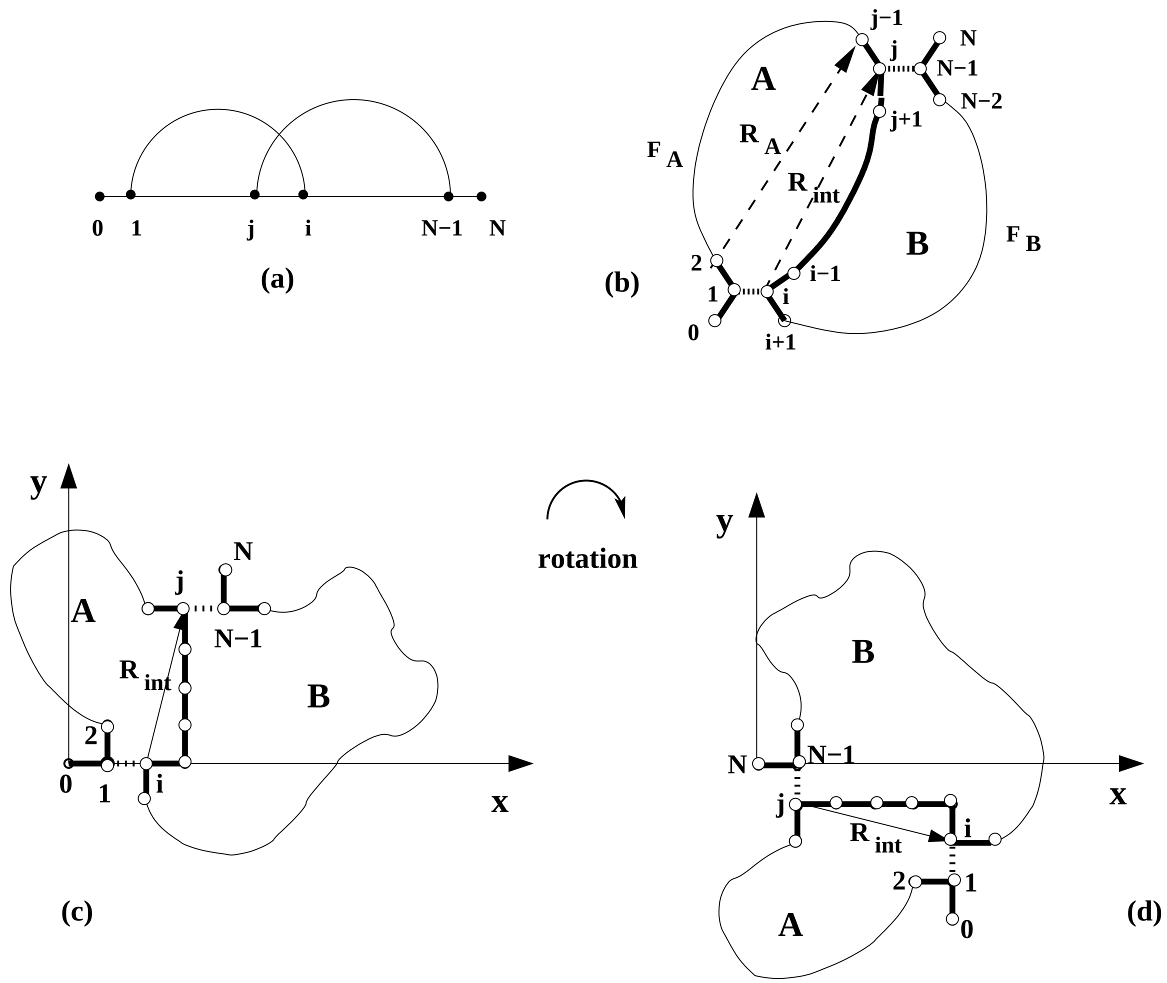


Figure 4.2: (a) The simplest graph with two crossing links. (b) The crossing linked chain conformation consists of two loops *A* and *B*, having the common chain segment (interface). It can be divided into relatively independent subunits: enlarged interface (shown bold) and two free single-stranded segments (F_A and F_B). The number of conformations of segment F_A can be considered to be a function of its end-end vector R_A, or, as a further approximation, of interfacial end-end vector R_{int}. (c) & (d) Two orientations of the enlarged interface on two-dimensional lattice with the given coordinate axes. The coordinate systems in (c) and (d) define the arguments of the function $\omega_f(y_{\text{int}}, l_{\text{int}}, f)$ for the numbers of conformations of segment F_A and F_B, respectively.

same conformation in both loops, and (2) the steric hindrance (the excluded volume effect) between the two loops. As a result, the total number of the chain conformations is usually much smaller then the product of the number of conformations of each individual (isolated) loop.

In general, exact calculation for the excluded volume interactions between A and B is not viable. However, since the excluded volume interferences between A and B occur mainly near the interface, we can focus on the interface, which is a much simpler and a much more manageable system. To account for the excluded volume in the vicinity of the interface, we define an "enlarged interface" as the system consisting of the interface and its neighboring monomers. Specifically, as shown in Fig. 4.2b, the enlarged interface $\mathbf{I}$ consists of monomers 0, 1, 2 and $N-2, N-1, N$ in addition to the chain segment from monomer $j-1$ to monomer $i+1$. The enlarged interface can approximately account for the correlations between A and B. To describe the conformations for the other parts of the chain, we define "free loop segment" F_A for chain segment from monomer 2 to $j-1$ for loop A and free loop segment F_B for chain segment from monomer $i+1$ to $N-2$ for loop B. F_A and F_B have chain lengths of $f_A = j-3$ and $f_B = N-i-3$, respectively. For each given conformation $\mathbf{I}$ of the enlarged interface, we use $\Omega_f(\mathbf{I}, f_A)$ and $\Omega_f(\mathbf{I}, f_B)$ to denote the number of conformations for F_A and F_B for a given $\mathbf{I}$.

With the separation of the free loop segments F_A and F_B from the enlarged interface $\mathbf{I}$, the computation of the number of accessible conformations Ω for the two-contact graph in Fig. 4.2a becomes tractable, and Ω can be computed as

$$\Omega = \sum_{I=1}^{\omega_I} \Omega_f(\mathbf{I}, f_A) \cdot \Omega_f(\mathbf{I}, f_B), \tag{4.1}$$

where $I = 1, 2, ..., \omega_I$ denotes the viable conformations of the enlarged interface, ω_I is the number of the viable conformations of the enlarged interface. Fig. 4.3 shows all the $\omega_I = 21$ conformations for an enlarged interface with a 5-mer (4-bonds) interface in a two-dimensional lattice. The number of the enlarged interface conformations ω_I increases exponentially with the chain length of the interface.

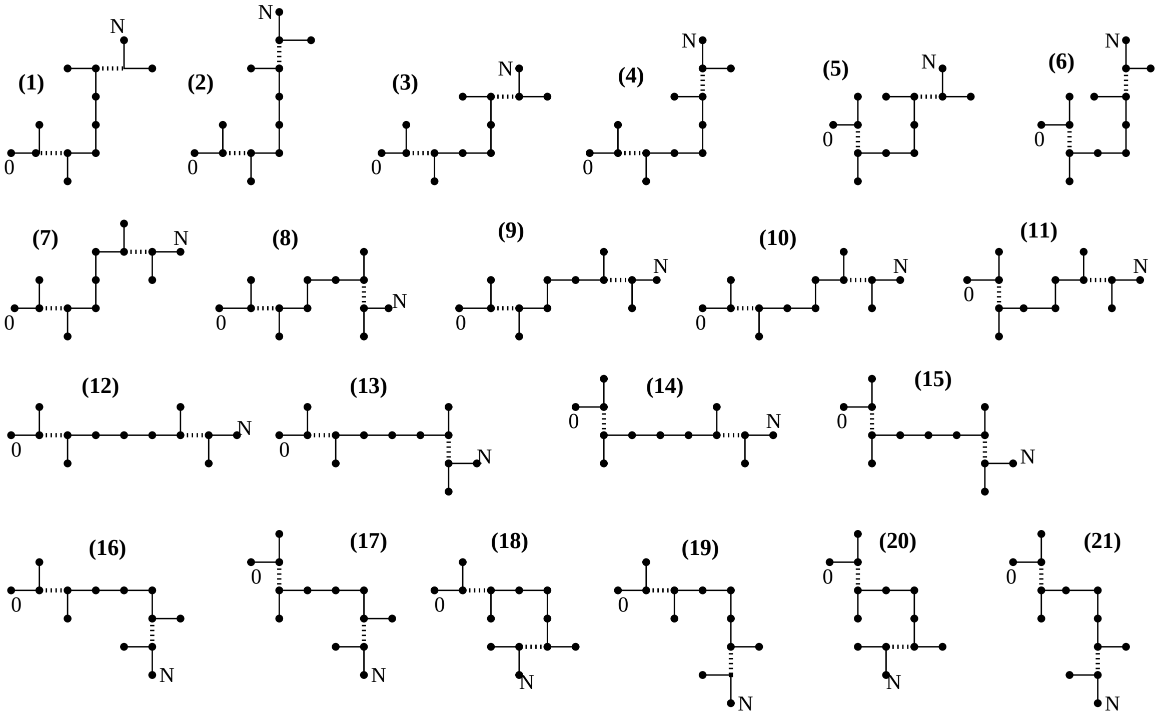


Figure 4.3: The viable conformations of the 5-mer interface, numbered from 1 to 21.

Central to the computation of Ω from Eq. 4.1 is the calculation of $\Omega_f(\mathbf{I}, f_A)$ and $\Omega_f(\mathbf{I}, f_B)$. In the following, in order to be specific, we use $\Omega_f(\mathbf{I}, f_A)$ to illustrate the methodology. Due to the large number of possibilities for $(\mathbf{I}, f_A)$, it is impractical to have a precomputed table that lists all the values of $\Omega_f(\mathbf{I}, f_A)$ for all the possible $(\mathbf{I}, f_A)$'s. However, as we will show in the following, through successive approximations, we can transform the calculation for $\Omega_f(\mathbf{I}, f_A)$ into a much simpler and computationally viable problem.

General methodology

First, as shown in Fig. 4.2b for the free loop segment F_A and the enlarged interface $\mathbf{I}$, $\Omega_f(\mathbf{I}, f_A)$ mainly depends on the enlarged interface conformation $\mathbf{I}$ through the positions of the two ends of the enlarged interface (i.e., monomers 2 and $j-1$) and the excluded volume interactions between F_A and $\mathbf{I}$ in loop A. Therefore, for $\Omega_f(\mathbf{I}, f_A)$, we can approximately represent the conformation $\mathbf{I}$ of the enlarged interface by the chain length of the interface l_{int} $(= i - j)$ and the end-end vector $\mathbf{R}_A$ = the vector from monomer 2 to monomer $(j - 1)$ (see Fig. 4.2). We note that $\mathbf{R}_A$ is also equal to the end-end vector of F_A. A larger end-end distance corresponds to more stretched conformations of $\mathbf{I}$ and F_A, and gives less accessible conformations, i.e., a smaller $\Omega_f(\mathbf{I}, f_A)$. Through this approximation, we can compute $\Omega_f(\mathbf{I}, f_A)$ as a function of $(\mathbf{R}_A,\ l_{\text{int}},\ f_A)$.

Second, for more complex graphs with multiple crossing links and multiple interfaces, it is hard to track the conformations for each of the enlarged interfaces. So it would be much more convenient to use the interfaces rather than the enlarged interfaces. Therefore, we simplify the $\mathbf{R}_A$-dependence by the $\mathbf{R}_{\text{int}}$-dependence, where $\mathbf{R}_{\text{int}}$ is the end-end vector of the interface (see Fig. 4.2b) instead of the enlarged interface. To account for the excluded volume effect, which is originally represented by the enlarged interface, we approximate the number of the conformations of the free loop segment $\Omega_f(\mathbf{I}, f_A)$ for a given enlarged interface conformation $\mathbf{I}$ by the average $\overline{\Omega_f(\mathbf{I}, f_A)}$ over all the possible enlarged interface conformations that have the same end-end vector of the interface $\mathbf{R}_{\text{int}}$. The resultant approximate value of $\Omega_f(\mathbf{I}, f_A)$ would be a function of $\mathbf{R}_{\text{int}}$ rather than of the enlarged interface conformation **I**.

Third, because no intrachain contact exists for the interface chain segment, its conformation is largely extended. As an approximation, we assume that the interfacial chain does not double back. As a result, for $\mathbf{R}_{\text{int}} = (x_{\text{int}}, y_{\text{int}})$, where $x_{\text{int}} = x_j - x_i, y_{\text{int}} = y_j - y_i$ in Fig. 4.2b, we have

$$l_{\text{int}} = |x_{\text{int}}| + |y_{\text{int}}|. \tag{4.2}$$

The above relation shows that the x and y components of $\mathbf{R}_{\text{int}}$ are not independent of each other for a given l_{int}. Therefore, we can further simplify the dependence of $\Omega_f(f_A, \mathbf{I})$ on the *vector* $\mathbf{R}_{\text{int}}$ as the dependence on a single *scalar* variable, say, y_{int}. The y_{int}-dependence of $\Omega_f(\mathbf{I}, f_A)$ is much more manageable than the original **I**-dependence because of the much less possibilities for y_{int} than for **I**.

For a given vector $\mathbf{R}_{\text{int}}$ the actual values of the x_{int} and y_{int} components depend on the coordinate system. The coordinate system should be defined in such a way that it gives consistent treatment for loops A and B. The conformation of the enlarged interface defines the directionality of the coordinate system. For $\Omega_f(\mathbf{I}, f_A)$, we define the coordinate system by fixing the coordinates of the first three monomers, labeled as 0, 1, and 2 in Fig. 4.2c, to (0,0), (1,0), and (1,1). While for $\Omega_f(\mathbf{I}, f_B)$, we use the equivalent monomers, namely, monomers N, $N-1$, and $N-2$, to define the coordinate system. The correspondence between these two sets of monomers becomes obvious if we apply a rotational transformation to the two loops, as shown in Fig. 4.2d. For the enlarge interface conformation shown in Fig. 4.2c & d, $\mathbf{R}_{\text{int}} = (x_{\text{int}}, y_{\text{int}})$ is equal to (1, 4) for loop A (see Fig. 4.2c) and (4, -1) for loop B (see Fig. 4.2d).

Finally, by averaging over all the possible interface chain conformations that have the same y_{int}, we can represent $\Omega_f(\mathbf{I}, f_A)$ as a function of the y_{int} of the interface chain conformation instead of $\mathbf{R}_{\text{int}}$. The resultant $\Omega_f(\mathbf{I}, f_A)$ for a given enlarged interface conformation $\mathbf{I}$ is simplified as a function of y_{int} (= the y-component of the end-end vector of the interface $\simeq$ that of the free loop chain segment), l_{int} (= the chain length of the interface), and f_A (= the length of the free loop chain segment):

$$\Omega_f(\mathbf{I}, f_A) \simeq \omega_f(y_{\text{int}},\ l_{\text{int}},\ f_A). \tag{4.3}$$

For example, for the enlarged interface conformation in Fig. 4.2c & d, l_{int} is equal to 5, and (y_{int}, f_A) is equal to $(4, j-3)$ for loop A and $(-1, N-i-4)$ for loop B. Therefore, $\Omega_f(\mathbf{I}, f_A) \simeq \omega_f(4, 5,\ j-3)$; $\Omega_f(f_B, \mathbf{I}) \simeq \omega_f(-1, 5,\ N-i-4)$.

With Eq. 4.3, we have

$$\Omega \simeq \sum_I \omega_f(y_{\text{int}}^{(A)},\ l_{\text{int}}^{(A)},\ f_A)\ \omega_f(y_{\text{int}}^{(B)},\ l_{\text{int}}^{(B)},\ f_B), \tag{4.4}$$

where $(y_{\text{int}}^{(A)},\ l_{\text{int}}^{(A)},\ f_A)$ and $(y_{\text{int}}^{(B)},\ l_{\text{int}}^{(B)},\ f_B)$ are the respective parameter sets for a given conformation $\mathbf{I}$ of the enlarged interface.

In the limit of very short interface chain segment, i.e., if the interface is much shorter than the free loop chain segment: $l_{\text{int}} << f$, where f is the chain length of the free loop segment, $\Omega_f(\mathbf{I}, f)$ would be only weakly dependent on the enlarged interface conformation $\mathbf{I}$. As a result, we can neglect the y_{int}-dependence of $\omega_f(y_{\text{int}},\ l_{\text{int}},\ f)$ and approximate it by an y_{int}-independent function $\omega_0(\ l_{\text{int}},\ f)$:

$$\omega_f(y_{\text{int}},\ l_{\text{int}},\ f) \simeq \omega_0(\ l_{\text{int}},\ f) = \frac{\Omega_{loop}}{\omega_l}, \tag{4.5}$$

where Ω_{loop} is the number of conformations of the loop and ω_l is the number of conformations for the part of the enlarged interface within the loop. For example, in Fig. 4.2, for the free loop chain segment F_A from monomer 2 to monomer $j-1$, $f_A = j-3$ and $l_{\rm int} = i-j$. Ω_{loop} is for the loop $1 \to i \to 1$, and ω_l is for the chain segment $j-1 \to i \to 1 \to 2$, which is part of the enlarged interface. In terms of the $\omega_0(\ l_{\rm int},\ f)$ function, for $l_{\rm int} << \ f_A,\ ,f_B$, we can compute Ω in Eq. 4.4 as

$$\Omega \simeq \omega_l\ \ \omega_0(\ l_{\rm int},\ f_A)\ \ \omega_0(\ l_{\rm int},\ f_B). \tag{4.6}$$

Illustrative calculations and tests in two-dimensional lattice model

We choose a short (25-mer) chain to illustrate the method. The chain makes two crossing linked contacts, as specified by the graph in Fig. 4.4a. For the given graph, the length of the interface $8 \to 12$ is $l_{\rm int} = 4$, and the lengths of the free loop segments are $f = 5, 9$ for $2 \to 7$ and $13 \to 22$, respectively.

To compute the number of conformations Ω from Eq. 4.4, for each of the 21 enlarged interface conformations of the 5-mer interface in Fig. 4.3, we need to know $\omega_f(y_{\rm int},\ l_{\rm int},\ f) = \omega_f(y_{\rm int},\ 4,\ 5)$ and $\omega_f(y_{\rm int},\ 4,\ 9)$ for the respective free loop segments. Following the step-by-step procedure presented in the previous section for the calculation of $\omega_f(y_{\rm int},\ l_{\rm int},\ f)$ function, we performed exact computer enumeration in a two-dimensional lattice and obtained $\omega_f(y_{\rm int},\ l_{\rm int},\ f)$ for all the possible parameter sets for $1 \leq\ l_{\rm int} \leq 12$ and $1 \leq\ f \leq 24$. Table 4.1 shows the results for small $l_{\rm int}$ and f values. For each conformation **I** of the enlarged interface in Fig. 4.3, we determine the $y_{\rm int}$ value of the interface in the coordinate system defined by the positions of monomers 0, 1, 2 for $y^{(A)}_{\rm int}$ and by the positions of monomers N, $N-1$, $N-2$ for $y^{(B)}_{\rm int}$. Summing over all the 21 conformations gives the

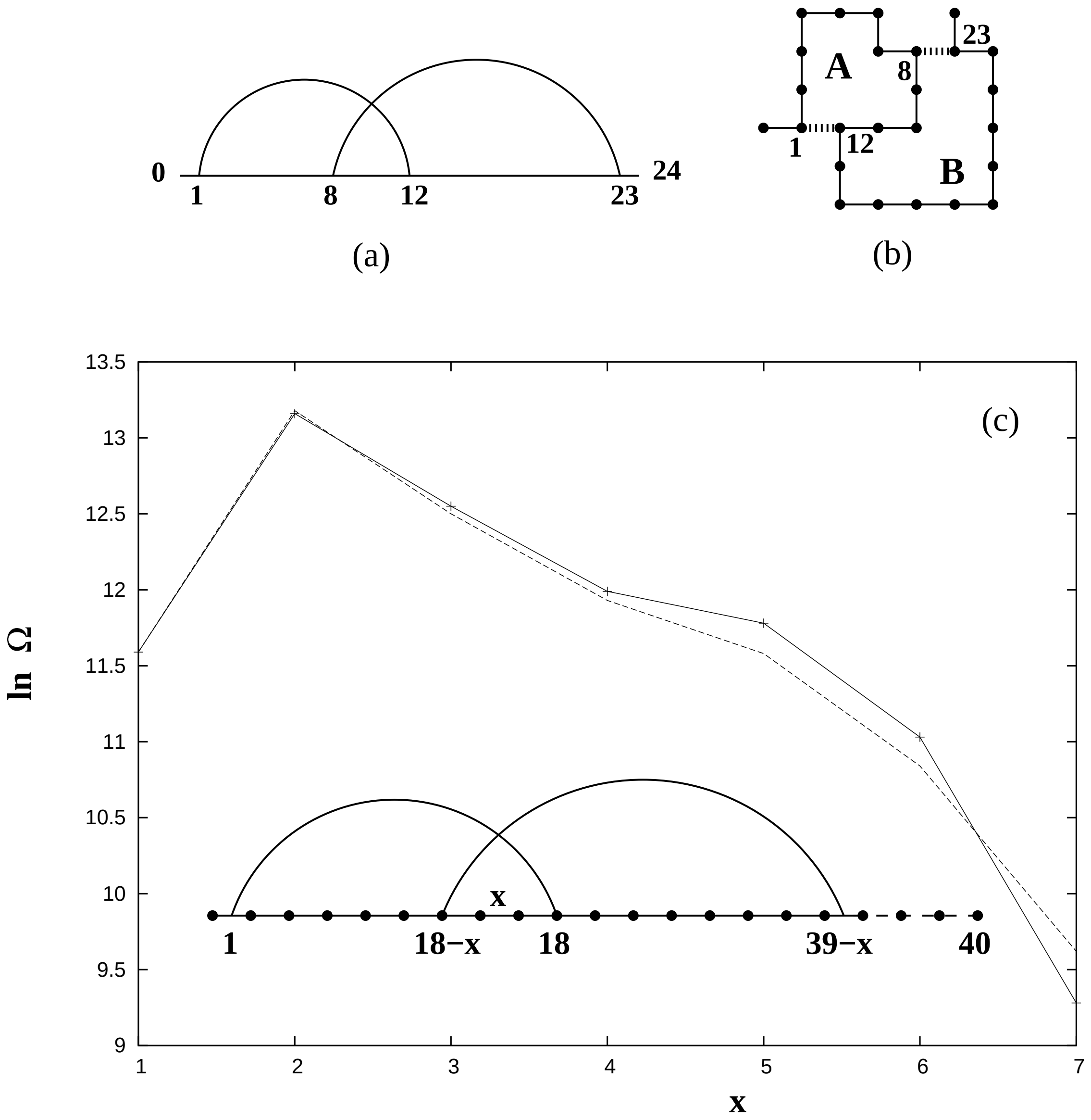


Figure 4.4: (a) and (b) A simple graph with two crossing links is used to illustrate the method of calculation of values of function $\omega_f(y_{\rm int}, l_{\rm int}, f)$ (see Table 4.1) and how to obtain the total number of conformations for the given graph (Table 4.2). (c) The test of the theory (dashed line) against exact enumeration (solid line) using the graph with variable position of the middle contact.

Table 4.1: Numbers of free loop segment conformations for some values of l_{int} and f

l_{int}	y_{int}	f = 1	2	3	4	5	6	7	8	9	10
1	0		0.00		1.00		1.00		1.00		5.00
2	1	0.00		0.00		0.00		0.00		0.00	
	0	0.00		0.50		1.00		2.50		6.25	
3	2		1.00		0.00		0.00		0.00		0.00
	1		0.00		0.00		2.00		4.00		9.00
	0		0.00		0.50		2.00		6.25		17.75
	-1		0.00		0.00		0.00		1.00		7.50
4	3	0.00		1.00		0.00		0.50		1.00	
	2	0.40		0.40		0.20		0.60		1.60	
	1	0.00		0.00		0.50		2.75		8.50	
	0	0.00		0.00		0.50		3.50		13.50	
	-1	0.00		0.00		0.00		0.00		1.00	
	-2	0.00		0.00		0.00		0.00		1.00	

number of chain conformations for the graph:

$$\Omega = \sum_{I} \omega_f(y_{\text{int}}^{(A)},\ 4,\ 5)\ \omega_f(y_{\text{int}}^{(B)},\ 4,\ 9) = 45.12. \tag{4.7}$$

The details of calculations are shown in Table 4.2. The exact value for Ω obtained from the exact computer enumeration is 49, which is quite close to the estimated result.

In Fig. 4.4c, as a test for the method, we compute the number of conformations for a series of graphs with two crossing links. We again find good agreement between the analytical calculation and the exact computer enumeration.

4.2.2 More complex graphs with two crossing links.

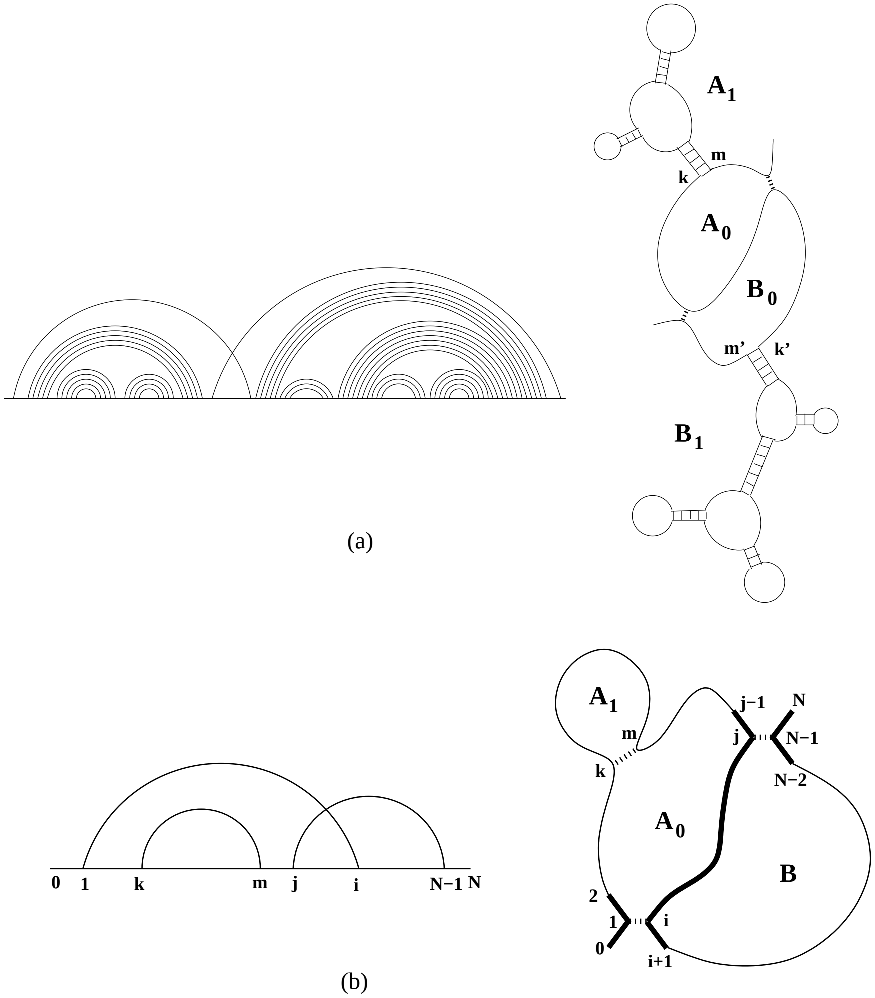
The above theory can be generalized to treat more complex graphs that contain multiple non-crossing links in addition to two crossing links; see Fig. 4.5a for an example. The non-crossing links bear either nested or unrelated relationships with the crossing links. Since a cluster of the nested and unrelated links form a secondary structure, the type of graphs in Fig. 4.5a can be regarded as (tertiary) crossing-linked secondary structures.

We use the graph in Fig. 4.5b to illustrate the theory. The difference between the graph/conformation in Fig. 4.5b and Fig. 4.2 comes from the additional loop A_1 attached to loop A_0 in Fig. 4.5b. We treat the composite loop $A_0 + A_1$ as an effective loop A. The effective "free loop segment" F_A for "loop A" is the chain segment from monomer 2 to monomer $j - 1$, and the free loop segment F_B for loop B is from monomer $i + 1$ to monomer $N - 2$. The interface is from monomer j to monomer i of length $l_{\text{int}} = i - j$, and the enlarged interface is shown as the thick lines in Fig. 4.5b.

We again use Eqs. 4.1 & 4.4 to treat the graph. The key is how to compute $\Omega_f(\mathbf{I}, f_A)$ (=

Table 4.2: Illustrative calculation of the number of conformations of the graph shown in Fig. 4.4a

I	$y_{\text{int}}^{(A)}$	$y_{\text{int}}^{(B)}$	$\omega_f(y_{\text{int}}^{(A)}, 4, 5)$	$\omega_f(y_{\text{int}}^{(B)}, 4, 9)$	product
1	3	-1	0.0	1.0	0.0
2	3	-1	0.0	1.0	0.0
3	2	-2	0.2	1.0	0.2
4	2	-2	0.2	1.0	0.2
5	2	-2	0.2	1.0	0.2
6	2	-2	0.2	1.0	0.2
7	2	2	0.2	1.6	0.32
8	1	1	0.5	8.5	4.25
9	1	1	0.5	8.5	4.25
10	1	1	0.5	8.5	4.25
11	1	1	0.5	8.5	4.25
12	0	0	0.5	13.5	6.75
13	0	0	0.5	13.5	6.75
14	0	0	0.5	13.5	6.75
15	0	0	0.5	13.5	6.75
16	-1	3	0.0	1.0	0.0
17	-1	3	0.0	1.0	0.0
18	-2	2	0.0	1.6	0.0
19	-2	2	0.0	1.6	0.0
20	-2	2	0.0	1.6	0.0
21	-2	2	0.0	1.6	0.0
					sum=45.12



(a)

(b)

Figure 4.5: (a) Two crossing- linked secondary structures. The number of conformations of each secondary structure can be computed from the previously developed matrix method [16, 17] to obtain $S^{(\nu)}_{A_1}$ in Eq. 4.8. (b) To demonstrate the method, the simplest representative of the secondary structure, loop A_1, has been attached to the loop A_0.

the number of conformations for F_A for a given conformation $\mathbf{I}$ of the enlarged interface). To account for the conformational constraint imposed by the additional pair (k, m) and the excluded volume interactions between A_0 and A_1, we classify four types of conformations for the contact (k, m) in a two-dimensional lattice [16] (see Fig. 3.2c). For a given μ-th ($\mu = 1, 2, 3, 4$) type conformation of the (k, m) contact, we use $S_{A_1}^{(\mu)}$ to denote the number of conformations for loop A_1, and use $\omega_f^{(\mu)}(y_{\text{int}}^{(A_0)},\ l_{\text{int}},\ f_{A_0}) = \omega_f^{(\mu)}(y_j - y_i,\ i - j,\ k + j - m - 2)$ to denote the number of conformations for the free loop segment F_{A_0} which consists of the chain segment $2 \rightarrow k$, the contact (k, m), and the chain segment $m \rightarrow j - 1$. F_{A_0} has chain length of $f_{A_0} = k + j - m - 2$.

The sum over the four types of the (k, m) contact conformations gives $\Omega_f(\mathbf{I}, f_A)$:

$$\Omega_f(\mathbf{I}, f_A) = \sum_{\mu=1}^{4} \sum_{\nu=1}^{4} \omega_f^{(\mu)}(y_j - y_i,\ i - j,\ k + j - m - 2)\ Y_{\mu\nu}\, S_{A_1}^{(\nu)}, \tag{4.8}$$

where $Y_{\mu\nu} = 1$ and 0 for a viable and non-viable connection between a type μ and a type ν conformation, respectively [16]. For example, $Y_{12} = 1, Y_{24} = 0$. Furthermore, through exact computer enumeration for $\omega_f^{(\mu)}$'s, we find that in a two-dimensional lattice,

$$\alpha^{(\mu)} = \frac{\omega_f^{(\mu)}(y_{\text{int}},\ l_{\text{int}},\ f_{A_0})}{\omega_f(y_{\text{int}},\ l_{\text{int}},\ f_{A_0})} \simeq 0, 0.15, 0.15, 0.18 \ \text{ for } \ \mu = 1, 2, 3, 4, \ \text{ respectively.}$$

Here we assume that the total length of loop A_0 is longer than 4.

Combining the above results, we obtain the following simplified expression for ω_{F_A}:

$$\Omega_f(\mathbf{I}, f_A) = \omega_f(y_{\text{int}}^{(A_0)},\ l_{\text{int}},\ f_{A_0}) \sum_{\mu=1}^{4} \sum_{\nu=1}^{4} \alpha^{(\mu)}\, Y_{\mu\nu}\, S_{A_1}^{(\nu)}. \tag{4.9}$$

For F_B, from Eq. 4.3, we have $\Omega_f(\mathbf{I}, f_B) = \omega_f(y_{\text{int}}^{(B)},\ l_{\text{int}},\ f_B)$. With the above results for $\Omega_f(\mathbf{I}, f_A)$ and $\Omega_f(\mathbf{I}, f_B)$ we obtain the conformational count Ω from Eq. 4.1 for the graph.

We can further generalize the above method to treat more complex crossing-linked secondary structures, for example, the graph and structure shown in Fig. 4.5a, where RNA secondary structures are brought into contact through the crossing-linked loops A_0 and B_0. For such complex graphs, we need to replace the $S_{A_1}^{(\nu)}$ vector in Eq. 4.9 for the loop A_1 by the corresponding vector for the complex secondary structure attached to A_0. The computation of such vector for an arbitrary secondary structure is quite straightforward with the previously developed matrix method[12], by replacing $S_{A_1}^{(\nu)}$ with a product of matrices for the secondary structural units.

If loop B_0 also has a complex secondary structure attached, similar to Eq. 4.9, we have

$$\Omega_f(\mathbf{I},\ f_B) = \omega_f(y_{\text{int}}^{(B_0)},\ l_{\text{int}},\ f_{B_0})\ \sum_{\mu=1}^{4}\sum_{\nu=1}^{4}\alpha^{(\mu)}\,Y_{\mu\nu}\,S_{A_1}^{(\nu)}, \tag{4.10}$$

where $S_{A_1}^{(\nu)}$ is for the secondary structure attached to B_0. With Eqs. 4.9, 4.10, and 4.1, we can compute the number of conformations for any crossing-linked arbitrary secondary structures.

4.2.3 Graphs with multiple crossing links in series.

We can apply the above approach to treat graphs with a series of crossing links; see Fig. 4.6a. We use Ω_n, $\omega_2(n)$, and $\omega_1(n)$ ($n = 1, 2, ...$) to denote the number of conformations for the graph in Fig. 4.6b (with n crossing links), Fig. 4.6c, and Fig. 4.6d, respectively. Using the following approximation

$$\frac{\Omega_{n+1}}{\Omega_n} \simeq \frac{\omega_2(n)}{\omega_1(n)},$$

we have

$$\Omega_n \simeq \Omega_2 \prod_{r=2}^{n-1} \frac{\omega_2(r)}{\omega_1(r)}. \tag{4.11}$$

As shown in Fig. 4.6e, tests against exact computer enumeration in two-dimensional lattice model shows that Eq. 4.4 can give a good estimation for Ω_n.

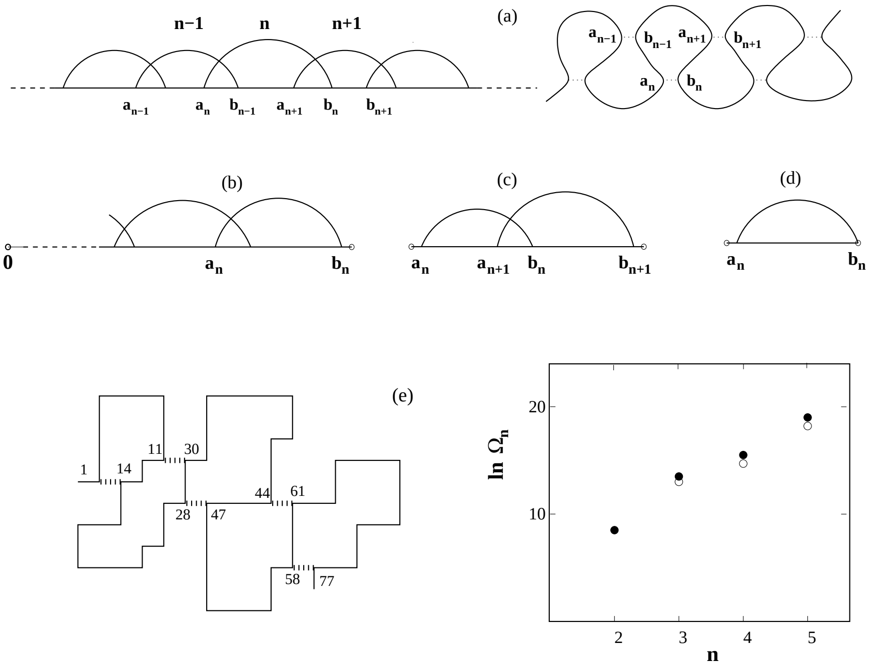


Figure 4.6: (a) Multiple crossing links in series, the graph and the conformation. The conformational count for graph in (b) with n crossing links can be computed using conformational counts for subgraphs (c) and (d). (e) Test against exact computer enumeration (filled circles) shows that the theory (empty circles) gives the good estimation of the number of conformations for graphs with crossing links in series.

4.3 Graphs with multiple crossing links.

4.3.1 Graphs with a tertiary contact added to a set of nested contacts.

In this section we treat graphs which contain, as shown in Fig. 4.7a, a tertiary contact $(j, N-1)$ that crosses two nested links $(1, i)$ and (m, k). In contrast to the graph in Fig. 4.2, an additional contact is established between monomers m and k in Fig. 4.7a, resulting in two nested loops A_1 and A_2.

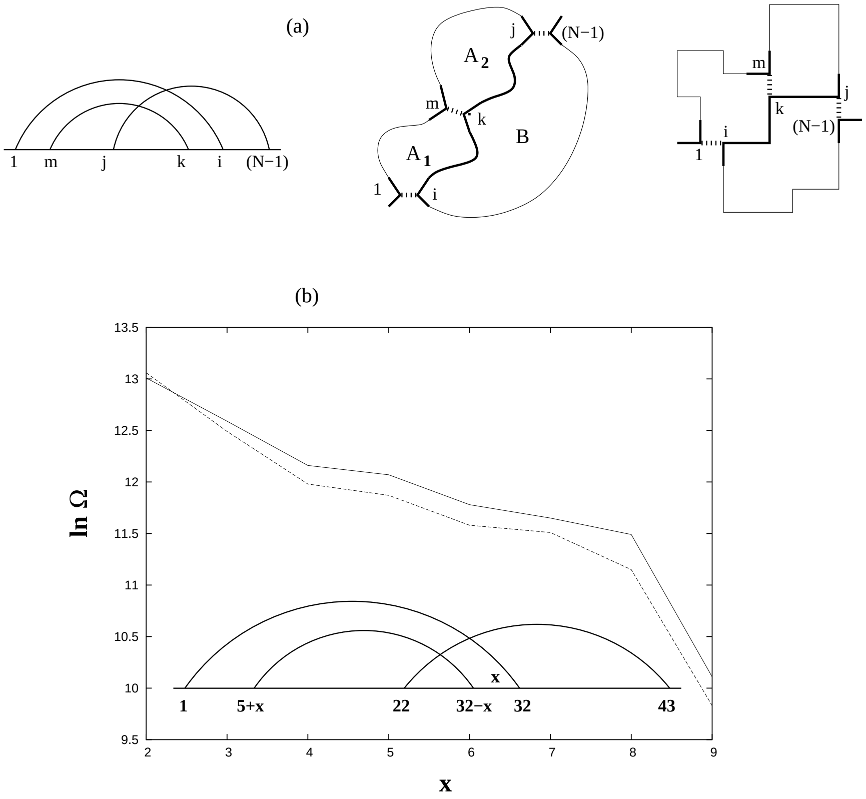


Figure 4.7: (a) The conformation with three crossing links and the corresponding graph. The additional contact is established between monomer of free loop segment m and an interfacial monomer k. The thick lines denote the enlarged interface **I**. (b) The test of the theory (dashed line) against exact enumeration (solid line) using the graph with variable position of the middle contact.

To account for the conformational constraint arising from the contact (k, m), we include the conformation of the (k, m) contact in the enlarged interface **I**; see the thick lines in Fig. 4.7a. Correspondingly, we define the free loop segments F_{A_1}, F_{A_2}, and F_B as the chain segments from monomer 2 to monomer $m - 1$, from $m + 1$ to $j - 1$, and from $i + 1$ to $N - 2$, respectively. The sum over all the possible conformations of **I** gives the number of conformations for the graph:

$$\Omega = \sum_{I} \Omega_f(\mathbf{I},\, f_{A_1})\, \Omega_f(\mathbf{I},\, f_{A_2})\, \Omega_f(\mathbf{I},\, f_B), \tag{4.12}$$

where $\Omega_f(\mathbf{I},\, f_{A_1})$, $\Omega_f(\mathbf{I},\, f_{A_2})$, and $\Omega_f(\mathbf{I},\, f_B)$ are the numbers of conformations of the respective free loop segments, and can be given by the ω_f-function defined in Eq. 4.3 and tabulated in Table 4.1:

$$\Omega_f(\mathbf{I},\, f_{A_1}) \;=\; \omega_f(y_{\rm int}^{(A_1)},\; l_{\rm int}^{(A_1)},\; f_{A_1}) = \omega_f(y_k - y_i, i - k, m - 3); \tag{4.13}$$

$$\Omega_f(\mathbf{I},\, f_{A_2}) \;=\; \omega_f(y_{\rm int}^{(A_2)},\; l_{\rm int}^{(A_2)},\; f_{A_2}) = \omega_f(y_j - y_k, k - j, j - m - 2); \tag{4.14}$$

$$\Omega_f(\mathbf{I},\, f_B) \;=\; \omega_f(y_{\rm int}^{(B)},\; l_{\rm int}^{(B)},\; f_B) = \omega_f(y_i - y_j, i - j, N - i - 3). \tag{4.15}$$

Using the above three equations, for a given enlarge interface conformation $\mathbf{I}$, we can obtain $\Omega_f(\mathbf{I},\, f_{A_1})$, $\Omega_f(\mathbf{I},\, f_{A_2})$, and $\Omega_f(\mathbf{I},\, f_B)$ from Table 4.1. Fig. 4.7b shows that the method for the calculations for Ω is reliable as tested against the exact computer enumeration.

We can generalize the above approach to treat graphs with more crossing links; see Fig. 4.8a. Similar to Eq. 4.12, we have

$$\Omega = \sum_{I} \left(\prod_{r=1}^{n} \Omega_f(\mathbf{I},\, f_{A_r}) \right) \Omega_f(\mathbf{I},\, f_B), \tag{4.16}$$

where $\mathbf{I}$ is the conformation of the enlarged interface (shown as thick lines in Fig. 4.8a), and $\Omega_f(\mathbf{I},\, f_{A_r})$ $(r = 1, 2, ..., n)$ and $\Omega_f(\mathbf{I},\, f_B)$ are the numbers of conformations of the free loop segments in loops A_r and B (see Fig. 4.8a). Using Eqs. 4.13-4.15, for each given $\mathbf{I}$, we can calculate $\Omega_f(\mathbf{I},\, f_{A_r})$ $(r = 1, 2, ..., n)$ and $\Omega_f(\mathbf{I},\, f_B)$ in terms of the ω_f-function tabulated in Table 4.1.

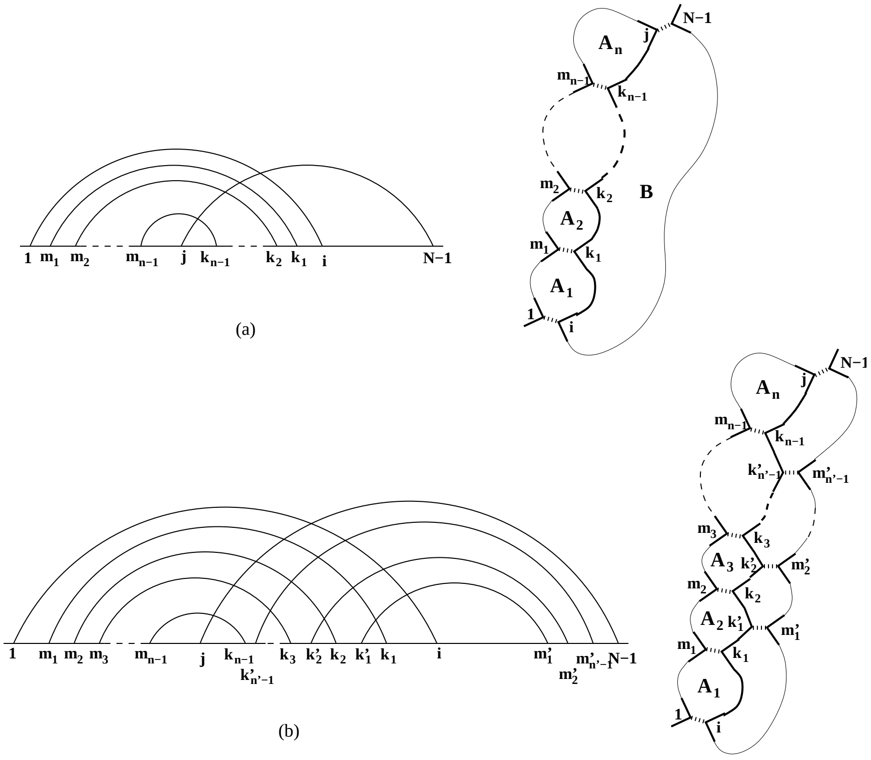


Figure 4.8: The method can be generalized to treat graphs with more crossing links (a) and two crossing sets of nested links (b). The enlarged interfaces are shown with thick lines.

4.3.2 Graphs with a tertiary contact added to a secondary structure.

Using Eqs. 4.17-4.18, we can treat more complex graphs with complex secondary structures attached to loops A_r ($r = 1, 2, ..., n$) and B in Fig. 4.8a. To account for the conformations of the secondary structures attached to the loops, we use Eq. 4.9 to calculate $\Omega_f(\mathbf{I}, f_{A_r})$ and $\Omega_f(\mathbf{I}, f_B)$ in Eq. 4.16:

$$\begin{aligned}\Omega_f(\mathbf{I}, f_{A_r}) &= \omega_f(y_{\text{int}}^{(A_r)}, l_{\text{int}}^{(A_r)}, f_{A_r}) \sum_{\mu=1}^{4} \sum_{\nu=1}^{4} \alpha^{(\mu)} Y_{\mu\nu} S_{A_r}^{(\nu)}; \quad (4.17)\\ \Omega_f(\mathbf{I}, f_B) &= \omega_f(y_{\text{int}}^{(B)}, l_{\text{int}}^{(B)}, f_B) \sum_{\mu=1}^{4} \sum_{\nu=1}^{4} \alpha^{(\mu)} Y_{\mu\nu} S_{(B)}^{(\nu)}, \quad (4.18)\end{aligned}$$

where $S_{A_r}^{(\nu)}$ and $S_B^{(\nu)}$ are the conformational counts for the secondary structures attached to A_r and to B, respectively. Substituting the above results for $\Omega_f(\mathbf{I}, f_{A_r})$ and $\Omega_f(\mathbf{I}, f_B)$ in Eq. 4.16 yields the number of conformations Ω for the graph.

4.3.3 Multiple crossing links between nested contacts.

The graph in Fig. 4.8b consists of two crossing-linked sets of nested links. The enlarged interface $\mathbf{I}$ is shown as thick lines in Fig. 4.8b. The sum over all the possible conformations for the enlarged interface gives the number of conformations for the graph:

$$\Omega = \sum_{I} \prod_{r=1}^{n} \Omega_f(\mathbf{I}, f_{A_r}) \prod_{s=1}^{n'} \Omega_f(\mathbf{I}, f_{B_s}), \quad (4.19)$$

where F_{A_r} and F_{B_s} are the free loop segments from monomer $m_{r-1} + 1$ to monomer $m_r - 1$ in loop A_r and from monomer $m'_{s-1} + 1$ to monomer $m'_s - 1$ in loop B_s, respectively. Here monomers $1, i, j, N - 1$ are regarded as m_0, k_0, k'_n, m'_n, respectively. $\Omega_f(\mathbf{I}, f_{A_r})$ and $\Omega_f(\mathbf{I}, f_{B_s})$ are the numbers of conformations of F_{A_r} and F_{B_s}, which can be obtained through the ω_f-

function, as shown in Eqs. 4.13, 4.14, and 4.15:

$$\Omega_f(\mathbf{I}, f_{A_r}) = \omega_f(y_{\text{int}}^{(A_r)},\ l_{\text{int}}^{(A_r)},\ f_{A_r}) \tag{4.20}$$

$$\Omega_f(\mathbf{I}, f_{B_s}) = \omega_f(y_{\text{int}}^{(B_s)},\ l_{\text{int}}^{(B_s)},\ f_{B_s}) \tag{4.21}$$

where $y_{\text{int}}^{(A_r)}$ is the y-component of the end-end vector $k_{r-1} \to k_r$ for the interface from monomer k_{r-1} to monomer k_r, $l_{\text{int}}^{(A_r)} = k_{r-1} - k_r$ is the chain length of the interface, and $f_{A_r} = m_r - m_{r-1} - 2$ is the chain length of F_{A_r}, and $y_{\text{int}}^{(B_s)}$ is the y-component of the end-end vector $k'_s \to k'_{s-1}$ for the interface from monomer k'_{s-1} to monomer k'_s, $l_{\text{int}}^{(B_s)} = k'_{s-1} - k'_s$ is the chain length of the interface, and $f_{B_s} = m'_s - m'_{s-1} - 2$ is the chain length of F_{B_s}.

4.4 Illustrative calculation for the partition function.

The partition function is defined in Eq. 3.2 as a sum over all the possible graphs. Therefore, the first step toward it's calculation is to enumerate graphs. We will enumerate all the graphs involving up to four crossing links formed by a tertiary contact. We assign the interaction energy $-\epsilon_2$ for each secondary contact and $-\epsilon_3$ for each tertiary contact. ϵ_3 can be different from ϵ_2. The energy of a graph is equal to $-\epsilon_2\cdot$ [the number of secondary contacts]$-\epsilon_3\cdot$ [the number of tertiary contacts]. Though the present theory can treat the sequence-dependence of the chain, for the purpose of an illustrative calculation, here we do not take into account the sequence and temperature dependence of ϵ_2 and ϵ_3. Effectively, we consider a homopolymer. To simplify the calculation, we consider relatively short chains of less than 35 monomers (nucleotides). For longer chains, we need to include more complex graphs with more tertiary contacts and crossing links.

4.4.1 Secondary and tertiary structure elements.

We can classify the crossing linked graphs into three groups according to the number of the crossing links (two, three, and four); see Fig. 4.9a for representative examples for each group of the graphs. For a given chain length, we exhaustively enumerate all the possible arrangements of the crossing links and shuffle secondary links around all possible positions on the graph so that the total number of links do not exceed four. For each generated graph, we compute the number of accessible chain conformations Ω. Fig. 4.9b shows the total number of (two-dimensional lattice) conformations for each group of the graphs for different chain lengths. Also plotted in the figure is the result from the exact computer enumeration. We find good agreement between the two sets of results, especially for two- and three-crossing links graphs, the theory and the computer enumeration give nearly identical results.

The graphs so far considered do not contain tails, and we call them the structure elements. Chains which are not longer than 35 nucleotides and have up to four contacts can fold into a larger number of different secondary and tertiary structure elements. A secondary structure from monomer a to monomer b is an element if an outermost contact $(a+1, b-1)$ is formed. The examples of secondary structure elements are shown in Fig. 4.10. For a chain segment between monomers a and b, we enumerate all the possible tertiary and secondary elements. For each tertiary element, we use the theory developed above to compute the number of chain conformations and to calculate the energy in terms of $-\epsilon_2$ and $-\epsilon_3$. From Eq. 3.2, the sum over all the possible elements gives the "element" partition function for the segment between a and b. For a homopolymer with given ϵ_2 and ϵ_3, such

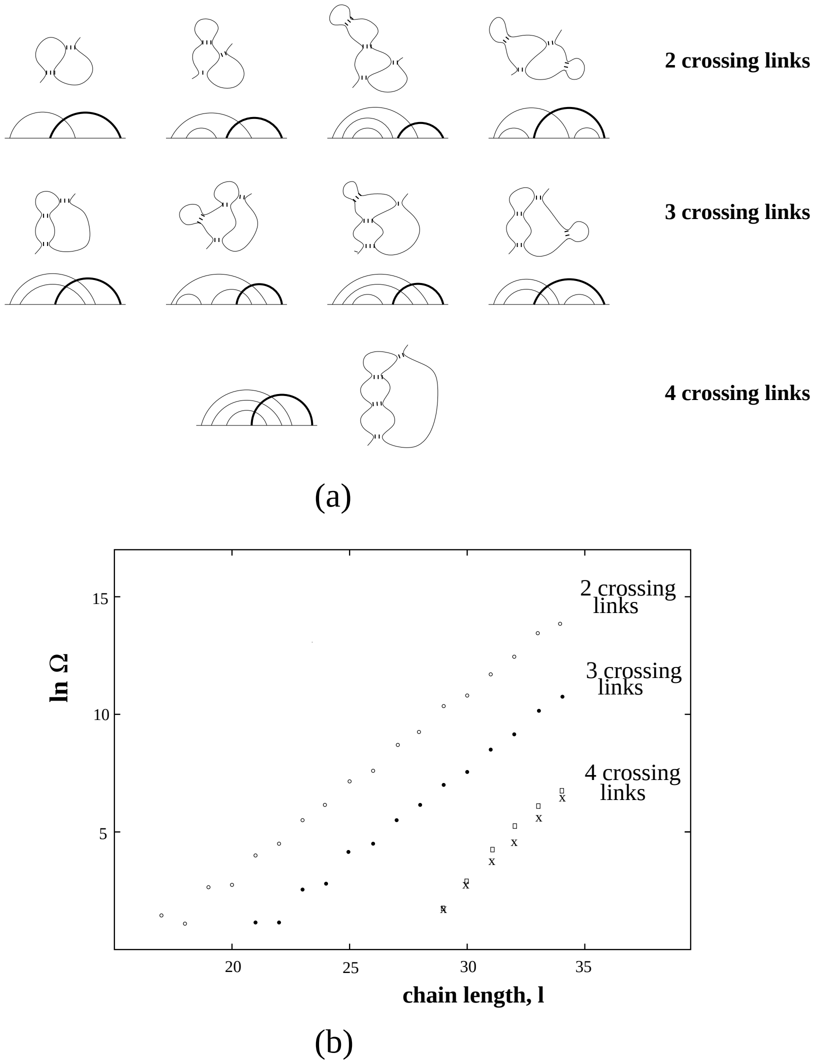


Figure 4.9: (a) The three groups of the graphs (with two, three and four crossing links) and their representative conformations. The tertiary contact is shown bold. The structures presented are called tertiary structure elements because they do not contain tails. (b) The total number of conformations of each group as a function of the chain length *l*. The total number of contacts is restricted to be ≤4. The results of approximate and exact calculations are not distinguishable for two and three crossing linked graphs. For the case of four crossing links, squares correspond to exact calculations and crosses represent results from our theory.

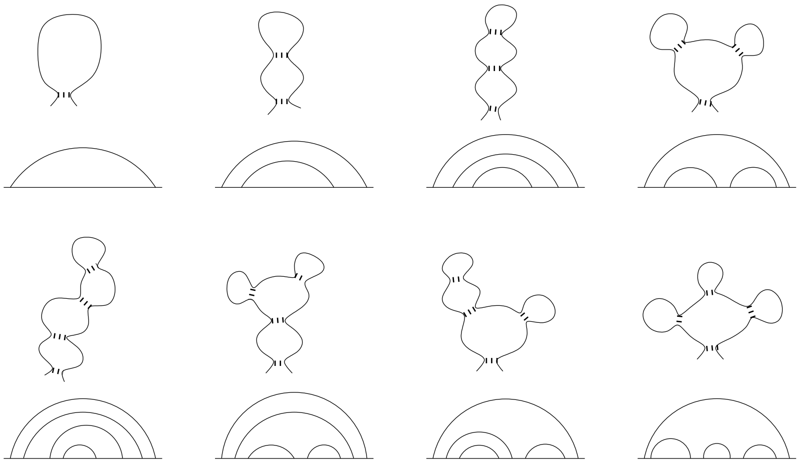


Figure 4.10: The graphs and representative conformations of secondary structure elements with the same restrictions as for the tertiary structure elements: there are only up to four single contacts.

partition function is a function of the length $l = b - a$ only. So we denote the element partition function as $Q_0(l)$.

4.4.2 Graphs consisting of one or more structure elements and single-stranded segments.

To obtain the full partition function of the chain, we need to add the contributions from the tails. For a given chain length L, the graph generally can contain multiple tertiary or secondary structure elements and the tails (Fig. 4.11). To calculate the number of conformations for such structures, one can, roughly speaking, multiply the numbers of conformations of each element and of each single-stranded segment.

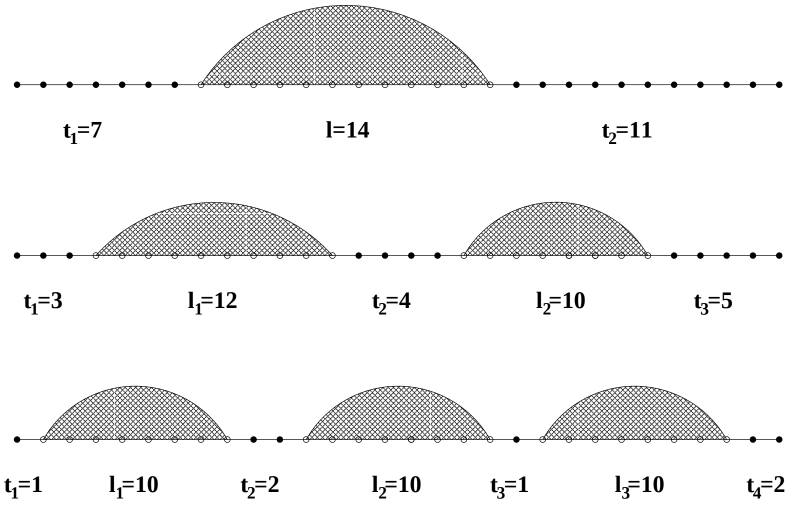


Figure 4.11: To calculate the density of states, all possible sequences of secondary and tertiary structure elements and of single-stranded segments should be taken into account. The shadowed regions of the graphs represent secondary and tertiary structure elements, dots denote nucleotides of single-stranded chain segments.

Table 4.3: Numbers of tail conformations obtained by exact computer enumeration on 2D lattice.

t	1	2	3	4	5	6	7
ω_t	1	2	4	9	21	50	118
t	8	9	10	11	12	13	14
ω_t	281	666	1 584	3 743	8 877	20 934	49 522
t	15	16	17	18	19	20	21
ω_t	116 579	275 205	646 909	1 524 458	3 579 101	8 418 185	19 768 268

Graphs with a single secondary or tertiary structure element.

The partition function of one (tertiary or secondary) structure element with tails is given by: $Q_0(l)\,\omega_t(t_1)\,\omega_t(t_2)$, where $\omega_t(t_i)$ is the number of the tail of length t_i (see Table 4.3). In fact, the number of tail conformations depends mainly on the total length of tails $t = t_1 + t_2$, and as an approximation, we have $\omega_t(t_1)\,\omega_t(t_2) \simeq \omega_t(t-1)$. For a given total chain length L, there are $t - 1 = L - l + 1$ possible positions for a single structure element of chain length l. Therefore, the total partition function for all such graphs is

$$Q_1 \simeq Q_0(l)\,\omega_t(L - l + 1)\,(L - l + 1). \tag{4.22}$$

Graphs with two structure elements.

In this case, the chain has three single-stranded segments, each with chain length denoted by t_1, t_2, and t_3. The total length of the single-stranded segments is can be determined from the total chain length L and lengths l_1, l_2 of each structure element: $t = t_1 + t_2 + t_3 = L - l_1 - l_2 + 4$ and the partition function can be approximated by $Q_0(l_1)\,Q_0(l_2)\,\omega_t(L - l_1 - l_2 + 1)$.

The number of graphs containing two such elements and single-stranded segments of total length t can be found from the following considerations. t_1 can have $t-2$ possible values: $t_1 = 1, 2, ...(t-2)$, t_2 can be chosen from $(t - t_1 - 1)$ possible values, and $t_3 = t - t_1 - t_2$. Therefore, the number of graphs for given l_1 and l_2 is : $\sum_{i=2}^{t-1}(t-i) = (t-2)(t-1)/2$ and the sum over all such graphs gives the partition function for graphs with two structure elements:

$$Q_2 \simeq \frac{1}{2}(L - l_1 - l_2 + 2)\,(L - l_1 - l_2 + 3)Q_0(l_1)\,Q_0(l_2)\,\omega_t(L - l_1 - l_2 + 1). \qquad (4.23)$$

Graphs with three structure elements.

The maximum number of structures in sequence which is allowed by restriction imposed on the total chain length is three. We estimated the corresponding partition function by

$$Q_3 \simeq Q_0(l_1)\,Q_0(l_2)\,Q_0(l_3)\,\omega_t(L - l_1 - l_2 - l_3 + 1)\sum_{i=1}^{L-l_1-l_2-l_3+3} i\,(i+1)/2, \qquad (4.24)$$

where the coefficient $\sum_{i=1}^{L-l_1-l_2-l_3+3} i\,(i+1)/2$ is the number of possible sets (t_1, t_2, t_3, t_4), obtained from considerations similar to that described above.

4.4.3 Density of states and partition function.

To compute the density of states $g(E)$ (= the total number of all the conformations that have energy E) for the chain of the given length, we consider all the secondary and tertiary structure elements along with the tails, such that the total chain length is fixed and the total number of contacts doesn't exceed four. The result for a 32-mer homopolymer is shown in Fig. 4.12 for three different values of (ϵ_2, ϵ_3).

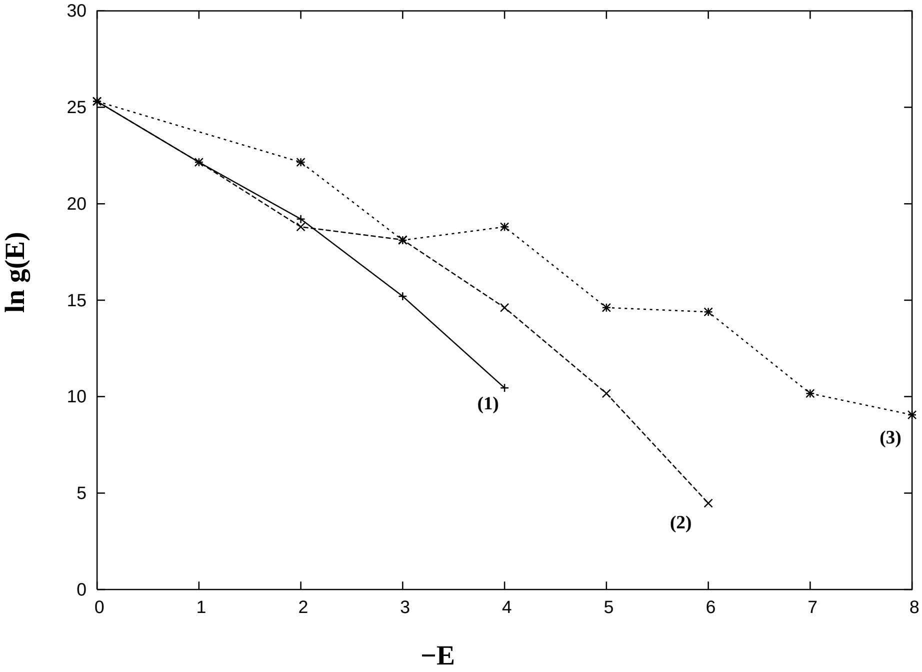


Figure 4.12: The density of states of a 32-mer homopolymer, calculated for $(\epsilon_2, \epsilon_3) = (\epsilon, \epsilon)$ for line (1), $(\epsilon, 2\epsilon)$ for line (2), $(2\epsilon, \epsilon)$ for line (3).

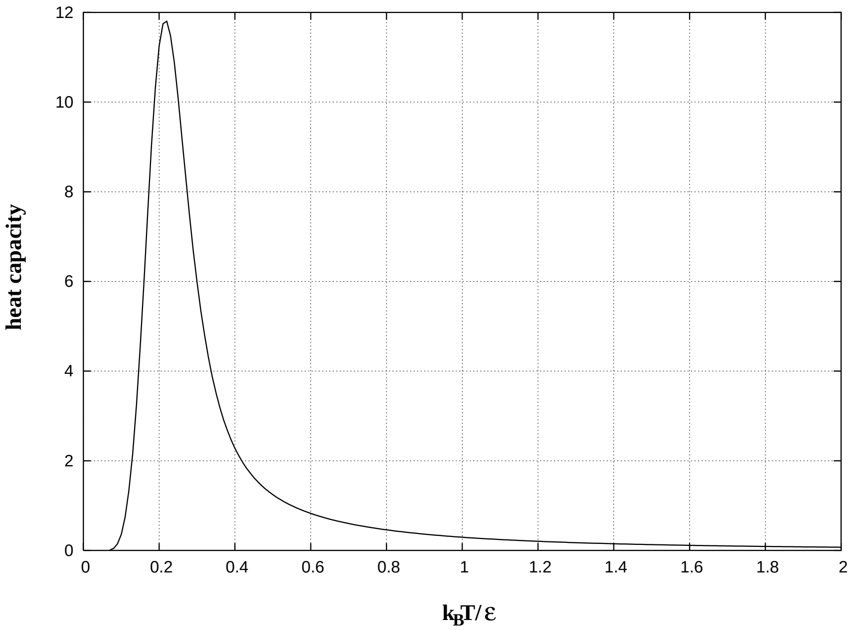


Figure 4.13: The temperature dependence of the heat capacity C(T) calculated from the model 30-mer chain. The transition takes place at the melting temperature $k_BT/\epsilon = 0.22$.

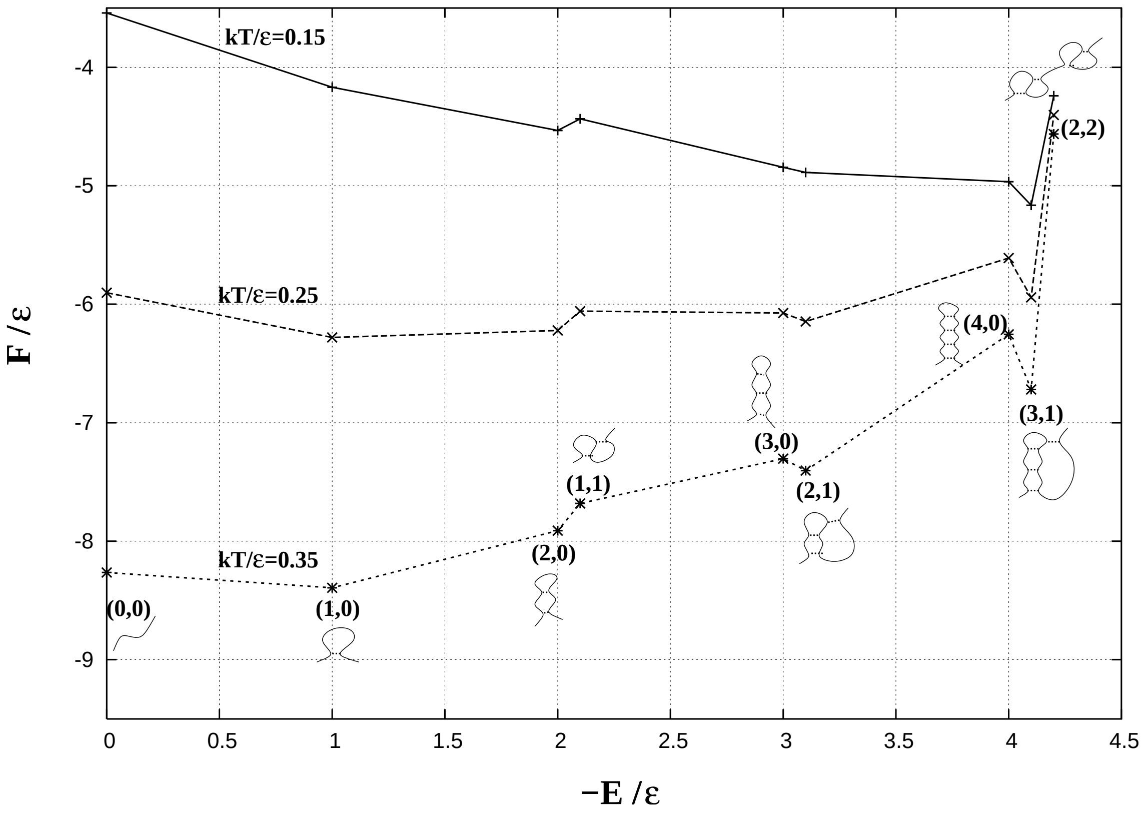


Figure 4.14: The free energy as a function of the energy for different $k_B T$ and $\epsilon_2 = \epsilon_3 = \epsilon$. Since the energies of secondary and tertiary contacts are taken to be equal, the secondary and tertiary conformations would be indistinguishable in the plot. To separate them, the free energies of conformations with one and two tertiary contact(s) are shown by points shifted by 0.1 and 0.2 to the right along the x-axis, respectively. The points on the graph are marked (i, j), where i is the number of secondary contacts and j is the number of tertiary contacts. For each point, a representative conformation is shown. For example, states (3, 0) and (2, 1) have the same energy 3ϵ, but with 0 and 1 tertiary contact, respectively.

From $g(E)$, we can compute the partition function as a sum over possible energy values:

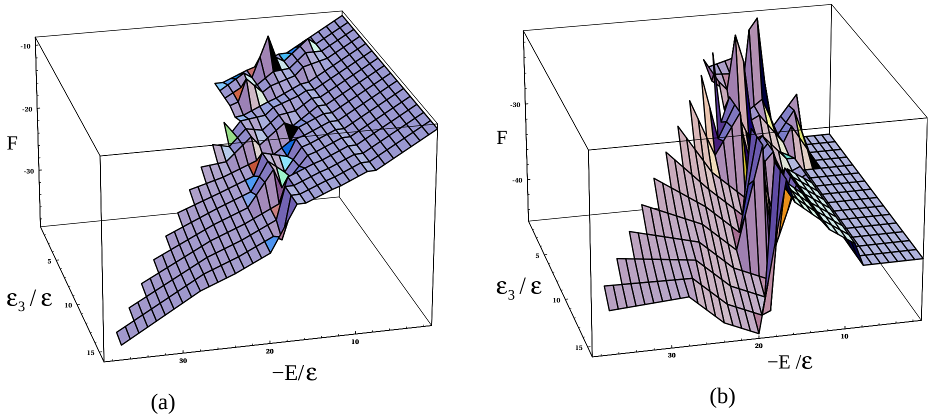
$$Q(T) = \sum_E g(E) \quad e^{-E/k_B T}. \tag{4.25}$$

From the partition function $Q(T)$, we can compute the heat capacity $C(T) = \frac{\partial}{\partial T}[k_B T^2 \frac{\partial}{\partial T} lnQ]$. The temperature-dependence of the heat capacity (the melting curve shown in Fig. 4.13) for a 30-mer homopolymer and $\epsilon_2 = \epsilon_3 = \epsilon$ shows a single transition at the melting temperature $k_B T_m/\epsilon = 0.22$. The energy-dependence of the microcanonical ensemble free energy $F(E) = E - k_B T \ln g(E)$ for different temperatures is plotted in Fig. 4.14. In order to separate conformations with and without tertiary interactions, we show them as separate points shifted along the x-axis by $0, 0.1$ and 0.2 for conformations with zero, one, and two tertiary contact(s). For example, for $E = -4\epsilon$, we have three sets of points, corresponding to conformations that have (4 secondary contacts), (3 secondary and 1 tertiary contacts), and (2 secondary and 2 tertiary contacts), respectively; see Fig. 4.14. We define a contact as a secondary structural contact if it is part of a secondary structure (= set of nested or unrelated contacts), and a tertiary contact if breaking it would cause the crossing-linked (tertiary) conformation become a secondary structure. For some simple conformations, the distinction between secondary and tertiary contacts is not unambiguous. For example, in the conformations marked $(1, 1)$ in Fig. 4.14, both contacts can be either secondary or tertiary. The free energy plot reveals that the most stable state at low temperature, i.e., the lowest energy state, is the one with 3 secondary structural contacts and one tertiary contact ((3,1) in Fig. 4.14). The unfolding transition from this “native” state to an ensemble of unfolded and partially unfolded states occurs around T_m when they have about the equal free energies.

4.5 Interplay between secondary and tertiary interactions.

In order to study the interplay between the secondary and tertiary contact energies, we fix the secondary energy parameter $\epsilon_2 = 4\epsilon$ and change the energy of the tertiary contact ϵ_3. The native structure and the melting temperature will obviously change with the changing of ϵ_3. Figure 4.15 shows the free energy landscapes F as a function of E and ϵ_3 for a 30-mer chain at different temperatures. We find that at low temperature T, independent of the tertiary energy parameter $\epsilon_3 \leq 15\epsilon$, the native state (minimum free energy) is always the state with the lowest energy E. As the temperature is increased, there exists a critical tertiary energy parameter ϵ_3^* such that the minimum free energy state is the highest E unfolded state for $\epsilon_3 < \epsilon_3^*$ and shifts to a low E state for $\epsilon_3 > \epsilon_3^*$. As the tertiary interaction is strengthened, the folding-unfolding melting temperature $T_m(\epsilon_3)$ would increase. So for a given temperature T, there exist a critical ϵ_3^* determined from $T = T_m(\epsilon_3^*)$. For $\epsilon_3 < \epsilon_3^*$, $T_m(\epsilon_3) < T$, so the chain is in the unfolded state, and for $\epsilon_3 > \epsilon_3^*$, $T_m(\epsilon_3) > T$, so the chain is in the folded state.

For a given set of the (ϵ_2, ϵ_3) parameter, we can calculate the heat capacity melting curve from the partition function. From the melting curve, we can identify the temperatures at which the conformational transitions occur. From the transition temperatures, we divide the temperature range into several pre-transition and post-transition regimes. In each regime, we find the most stable state. By performing the analysis for the melting curves and the free-energy landscapes for different (ϵ_2, ϵ_3) parameter sets, we are able to obtain the phase diagram for different parameters (T, ϵ_3) (Fig. 4.16). The chain have different stable structures in different (T, ϵ_3) regions. The stable state in the phase diagram is marked with (n, m)



(a)

(b)

Figure 4.15: The free-energy landscape at temperatures (a) $k_BT/\epsilon = 0.4$ and (b) $k_BT/\epsilon = 1.8$ for a 30-mer homopolymer with $\epsilon_2 = 4\epsilon$ and $0 \leq \epsilon_3 \leq 15\epsilon$.

for structures with n secondary and m tertiary contacts. The energy of the corresponding structure is $n\epsilon_2 + m\epsilon_3$.

From the phase diagram, we find that in the region where ϵ_3 is comparable with ϵ_2, the chain undergoes multiple transitions in the melting process. Overall speaking, the melting is less cooperative due to the interplay between the secondary and tertiary interactions when ϵ_3 is comparable with ϵ_2, and more cooperative when $\epsilon_3 \gg \epsilon_2$ or $\epsilon_3 \ll \epsilon_2$. In the $\epsilon_3 \ll \epsilon_2$ limit, the melting involves the secondary structural changes only, such as $(4, 0) \rightarrow (3, 0) \rightarrow (2, 0) \rightarrow (1, 0) \rightarrow (0, 0)$ for $\epsilon_3 = 1$, here (m, n) denotes states with m secondary and n tertiary contacts. In the $\epsilon_3 \gg \epsilon_2$ limit, the melting transitions mainly involve the breaking of the tertiary contacts (e.g. $(2, 2) \rightarrow (1, 1) \rightarrow (0,)$ for $\epsilon_3 = 14$). In the intermediate range of ϵ_3, the melting transitions involve the change of either the secondary or the tertiary structural contacts (e.g., $(3, 1) \rightarrow (2, 1) \rightarrow (1, 1) \rightarrow (1, 0) \rightarrow (0, 0)$ for $\epsilon_3 = 7$).

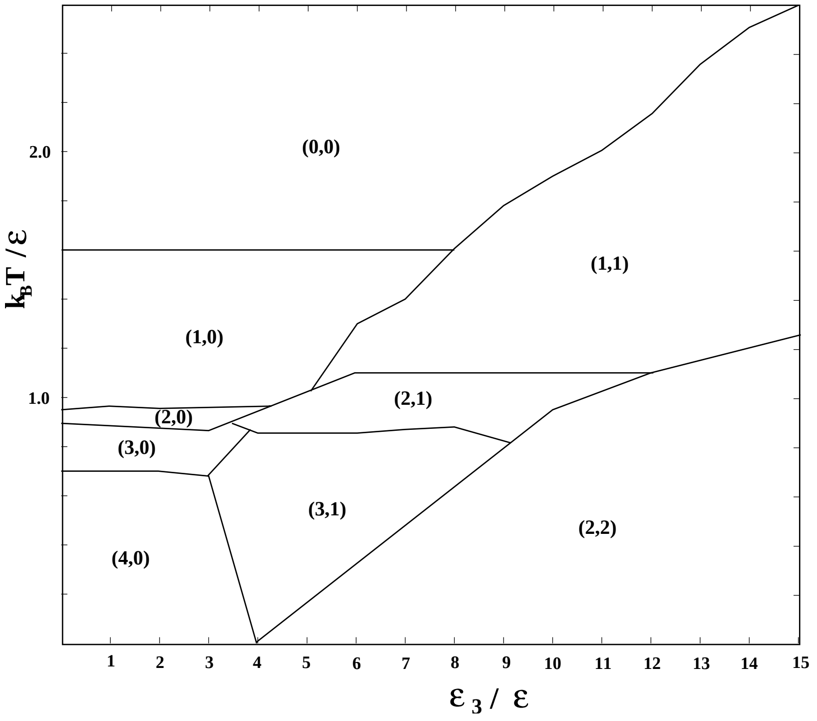


Figure 4.16: The phase diagram for a 30-mer chain. (m, n) denotes conformations with m secondary and n tertiary contacts.

4.6 Discussion.

In the present study, we have established a statistical mechanical machinery for simple RNA tertiary contacts to treat the nonadditive chain entropy and the partition function, from which the thermodynamic properties can be predicted. The method that we have developed enables the calculation for the number of chain conformations and the partition functions of the RNA-like molecules with simple tertiary interactions. The key idea of the method is to use the intrachain contacts to subdivide the conformation into different loops, and to assume that the excluded volume interferences between the loops predominantly come from the excluded volume of the monomers near the interfaces between the loops. The method has been shown to give accurate results for two-dimensional lattice test systems. Several simple types of tertiary folds are considered in the present work. These types of tertiary folds represent a large class of RNA tertiary structures. Applications to an illustrative simple model suggest that the interplay between the secondary and tertiary interactions can cause rugged free energy landscape and noncooperative melting transitions. Moreover, the generality of the above basic idea for the method suggests that the method may be extended to treat more complex tertiary topologies that involve multiple crossing-linked tertiary contacts. In addition, the method is developed based on the graphic representation of the structure and is thus general in terms of chain representation. The method can be implemented in more realistic off-lattice chain representations. Tertiary structure thermal stability is strongly dependent on the ionic solution condition. The electrostatic effect is not the focus here. But the model developed here would provide a more complete statistical mechanical framework for the modeling of the electrostatic interactions.

Chapter 5

STATISTICAL THERMODYNAMICS FOR RNA PSEUDOKNOTS.

The major part of this chapter (section 5.4) has been accepted for publication: *Z. Kopeikin and S.-J. Chen. Statistical thermodynamics for RNA pseudoknots. J. Chem. Phys. In press, scheduled issue: April 2006*

5.1 RNA pseudoknots - structure and functions.

RNA pseudoknots are simple tertiary structures composed of single-stranded loop segments and helical stems. They *are formed by base pairing of nucleotides of a loop (hairpin, internal, bulge or bifurcated) with nucleotides outside that loop*. The pseudoknots can be classified into several types depending on the types of loops. The simplest and most general is the H(airpin)-pseudoknot which includes only two loops and two stems (Fig. 5.1). The stems $S1$ and $S2$ combine to form a quasi-continuous helical structure of $S1+S2$ base pairs with one continuous and one discontinuous complementary strands. The loops $L1$ and $L2$ are not equivalent since they cross the major and minor helical grooves of stems $S2$ and $S1$, respectively. The structural diversity of H-pseudoknots is due to the differences in the helix-helix junction, such as the number of nucleotides on the continuous strand between stems, extent to which stems are coaxially stacked, bending and rotation (with respect to the A-helix geometry) angles at helical junction. The full coaxial stacking takes place if the single-stranded segment between stems consists of only one bond. Tertiary interactions play the dominant role in establishing the global fold of the molecule. They tie together

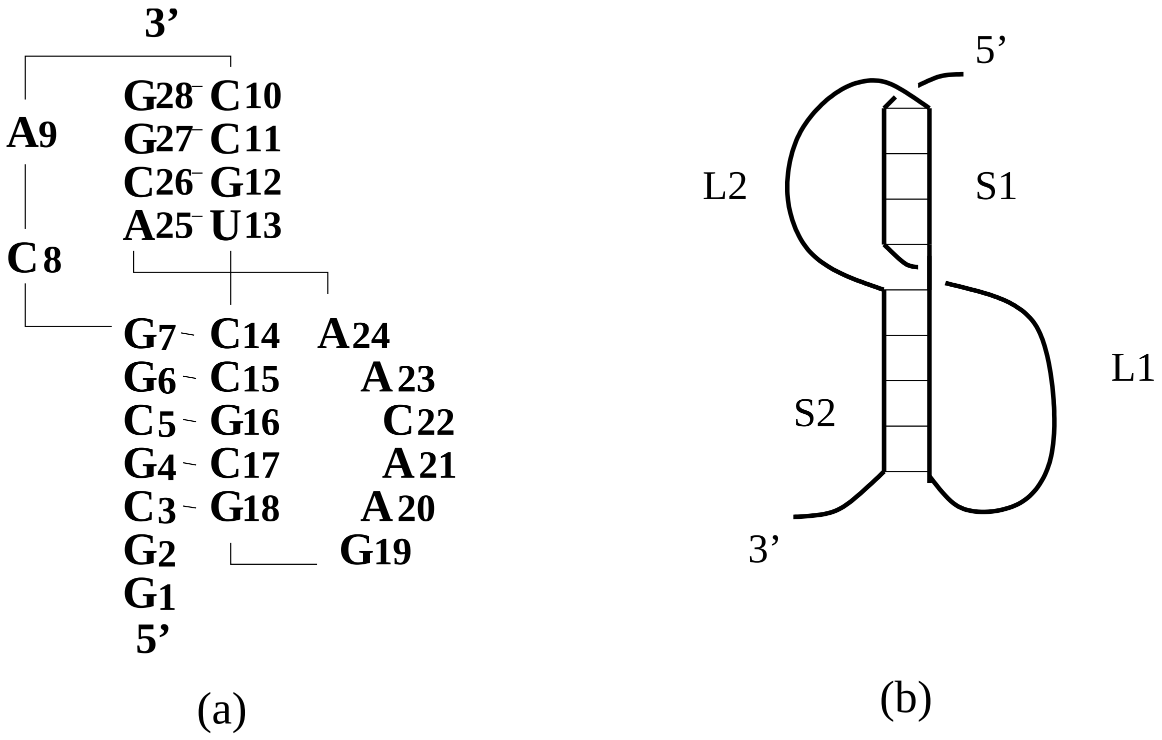


Figure 5.1: (a) The pseudoknot at the site of ribosomal frameshifting of Beet western yellows virus. (b) The schematic structure of an H-pseudoknot with stems coaxially stacked.

the otherwise weakly related loops of branched secondary structure to produce the specific, rigid, and functional three-dimensional structure.

Pseudoknot stability depends on the presence of divalent cations (Mg^{2+}), and is only marginally greater then stability of the constituent hairpins, with the gain in free energy being only 1-2 kcal/mol at 37^oC [22], [23]. On this basis, the role of pseudoknots as conformational switches has been suggested: the input/output of a small amount of energy may be sufficient to open/close a pseudoknot.

Pseudoknots are found in a wide variety of functional roles in RNAs [24]. Here are several examples.

(1) Some mRNAs contain pseudoknots which are involved in the regulation of translation. For *initiation translation regulation*, pseudoknots are usually positioned in the non-protein coding leader sequence, or in the sequence containing the ribosomal binding site. For example, in two mRNAs, encoding ribosomal proteins S15 and S4, the translation initiation site (Shine-Dalgarno sequence and initiation codon) is located within a pseudoknot structure [4, 5]. The formation of the fmet-tRNA-ribosome-mRNA initiation complex requires the pseudoknot to be at least partially unfolded. For mRNA encoding S15, it has been proposed that the pseudoknot is in a conformational equilibrium with the alternative hairpin structure. Binding of S15 increases the pseudoknot stability and represses its production. For ribosomal protein S4, the allosteric mechanism has been suggested for translational repression: the protein binding induces the conformational change which prevents initiation of translation. Another pseudoknot has been found within the gene 32 mRNA of bacteriophage T2. It is located in the 5’ non-coding sequence, upstream of the ribosome binding site and has been shown to serve as a binding site for the gene 32 protein. At low

concentrations, protein binds first to the pseudoknot, and then, as the protein concentration increases, the region of mRNA coated by the protein extends 3' until the ribosome binding site is bound and translation is shut off.

The role of RNA pseudoknots present within the protein coding regions of mRNAs can be *stimulation of ribosomal frameshifting or of translational readthrough* [29]. The majority of retroviral mRNAs contain the overlapping reading frames of the *gag*, *pro* and/or *pol* genes (encoding the proteins that form the viral capsid, protease and reverse transcriptase, respectively). To translate these genes, the purposeful -1 (one nucleotide to the left) shift in the reading frame is programmed into the RNA. The effective frameshifting requires the presence of two signals: the so-called slippery sequence, which is the actual frame shift site, and a structural element, stem-loop or pseudoknot located downstream of a slippery sequence. The slippery sequence is a heptonucleotide X XXY YYZ, where X can be any base, Y is A or U, and Z is A, U, or C. The simultaneous slippage model proposes that two ribosome-bound tRNAs (at sites A and P) simultaneously slip one nucleotide in 5' direction from the zero frame XXY YYZ to the -1 frame XXX YYY, so that at least two codon-anticodon base pairs can be formed after the slippage occur. The efficiency of frameshifting varies from 1-5% to 50% and regulates the relative concentrations of structural (*gag* gene) and catalytic proteins (*pro, pol* genes). The structural proteins are needed in much larger amounts than the catalytic proteins (polymerase and protease) for the efficient viral assembly and replication.

The presence of the downstream pseudoknot has been observed to stimulate frameshifting, but the precise mechanism is unknown. It has been shown that the change in the pseudoknot position with respect to the slippery sequence or replacement of the pseudoknot by the hairpin greatly diminishes frameshifting and readthrough efficiency. One of the most probable assumptions is that the ribosome pauses or stalls over the slippery sequence upon encountering the pseudoknot, which increases the probability of frameshifting or readthrough to occur. Moreover, the topological constraints of the pseudoknot, in which the 5' (closest to the ribosome) and the 3' structural boundaries of the pseudoknot, when compared to an hairpin, are on opposite sides of the molecule, may be important.

(2) Three pseudoknots have been found in the small subunit 16S ribosomal RNA [30]. Though the detailed functional roles of these pseudoknots are not known at present, the mutational analysis has shown that they are essential for the ribosome proper organisation, stability and functioning. It seems likely that one of pseudoknots is important for the binding of tRNA to the ribosomal A site.

(3) The probing experiments with $Fe^{II} - EDTA$ revealed that the catalytic core of most ribozymes (e.g. RNase P, group I introns, hammerhead ribozyme) is constructed from independently folded secondary structural domains which are brought together by tertiary interactions [31]. The individual domains have been shown to be unable to carry out the catalytic function. The assembling of a fully functional ribozyme requires the presence and binding of divalent metal ions (Mg^{2+}), which are known to be the necessary condition for the formation of stable tertiary contacts. The pseudoknots in the catalytic core are formed as a result of interactions between distant segments of the molecule.

5.2 Experiments on RNA pseudoknot folding thermodynamics.

As it is becoming clear that RNA pseudoknots play a number of important functional roles, intensive attempts have been made to experimentally study the folding and unfolding of pseudoknots. The studies are aimed at the obtaining information about the contributions of different structural elements to the pseudoknot stability, prediction of the equilibrium unfolding pathway and drawing correlations between the pseudoknot structure, stability and function.

The unfolding of RNA molecules usually can be modeled as a series of sequential two-state unfolding steps. Each two-state transition is characterized by an unfolding enthalpy and melting temperature. The enthalpies and melting temperatures for canonical base pair stacks can be estimated using Turner rules. It allows the identification of the melting steps of the secondary structure elements (helices). But the absence of non-canonical tertiary thermodynamic parameters makes it difficult to draw reliable conclusions about details of tertiary structures. Only some information about the presence, strength and location of non-canonical tertiary interactions can be obtained.

Though some pseudoknots fold at low concentrations of monovalent ions, the presence of high concentrations of monovalent or moderate concentrations of divalent ions was shown to strongly stabilize the pseudoknot by associating as weakly or partially localized ions in the regions of higher affinities, for example, at pseudoknot helix-helix junctions.

Here is the brief description for some experiments on pseudoknot unfolding.

(1) Gluick & Draper [32] proposed the folding pathways and estimated thermodynamic

parameters of the α mRNA pseudoknot, which plays a role in translational repression by ribosomal protein S4. The α mRNA pseudoknot is depicted in Fig. 5.2 and has been shown to undergo an allosteric conformational transition that regulates translational efficiency.

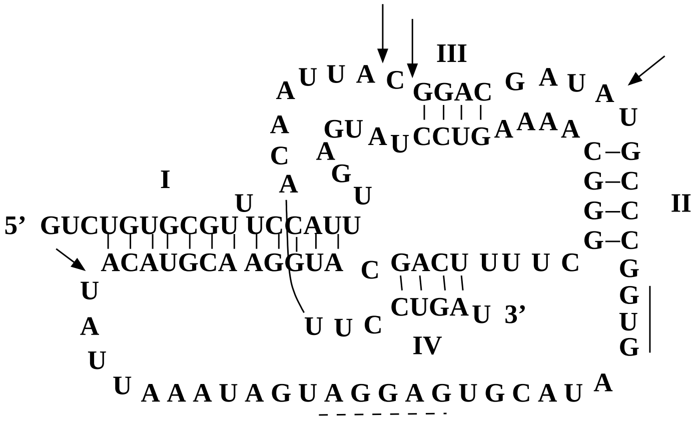


Figure 5.2: The α mRNA pseudoknot involved in control of translational initiation. Melting experiments were carried out on mRNA fragments with disruptions (at points shown with arrows) and compensatory mutations in helices. The initiation codon is underlined with solid, and the Shine-Dalgarno sequence-with dashed lines. The helices are labeled by roman numerals.

Two strategies has been used in studying the pseudoknot melting behavior. The first is to synthesize the RNA fragments which have a common 5' terminus and 3' termini at positions denoted by arrows in Fig. 5.2. The second strategy is to introduce compensatory mutations which result in disruption of individual helices. The obtained both ways RNA molecules include different sets of helices, and the comparison of the results of melting experiments should reveal the contribution of each helix to pseudoknot stability.

The melting of the pseudoknot involves multiple structural transitions. The following most probable unfolding pathway has been proposed for the α mRNA pseudoknot at 5 mM

Mg^{2+} and 100 mM KCl:

1. The lowest temperature transition requires presence of moderate concentrations of Mg^{2+} or high concentrations of K^{+} and very likely corresponds to the unfolding of non-canonical tertiary contacts.

2. Helix IV unfolds next, but the measured enthalpy change is substantially larger than predicted for helix IV alone from Turner rules (by 65 kcal/mol). This suggests that helix IV makes additional, non-canonical contacts with other parts of the RNA.

3. Helix I unfolds after helix IV.

4. Helices II and III melt last in a single, cooperative transition. The canonical base-pairing alone does not provide any rationale for coupling of this two helix units. Therefore, it seems likely that some additional non-canonical structure causes their linking.

The analysis of the melting experiment suggests that the pseudoknot is in fact quite stable and that there are multiple unfolding intermediates, some of which involve substantial non-canonical interactions.

(2) Theimer and Giedroc [33] carried out an experiment on unfolding of the frameshifting pseudoknot of mouse intracisternal A-type particles (mIAP), an endogenous retrovirus. The pseudoknot and its proposed equilibrium unfolding pathway are shown in Fig. 5.3.

The unfolding thermodynamics has been studied by comparison of melting data for the native pseudoknot, and compensatory base-pair substitution and deletion mutants. The unfolding pathway includes four optically and calorimetrically defined steps. Stem 2 melts

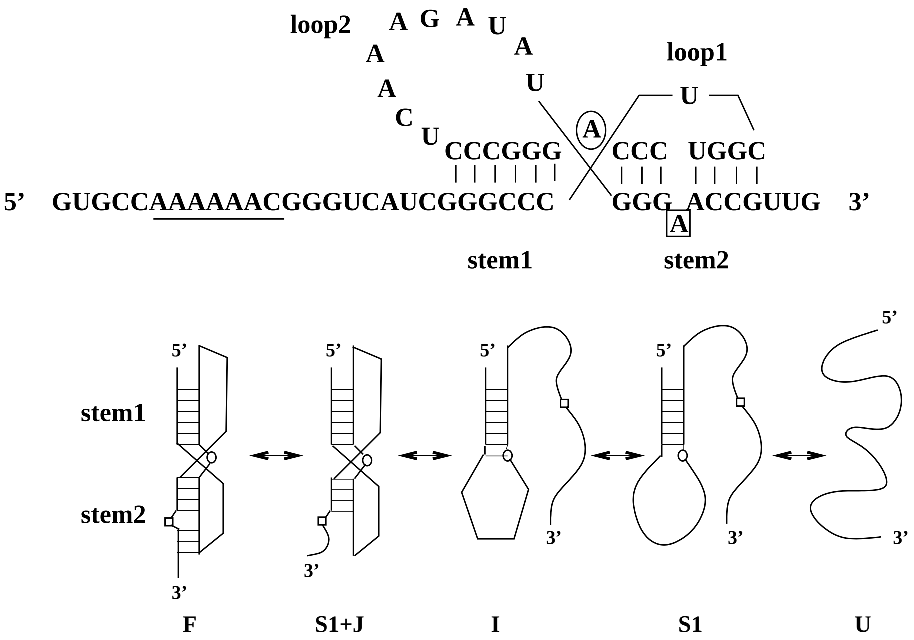


Figure 5.3: The frameshifting mIAP pseudoknot and its unfolding pathway based on a multiple, interacting, sequential two-state transition analysis of melting data. The slippery sequence is underlined.

first in two closely coupled low-enthalpy transitions at low melting temperatures (F → S1+J → I). The intermediate state I was shown to consist of the stem 1 hairpin and an unknown non-canonical tertiary structure in the hairpin loop. The tertiary structure unfolds at the third step to give the stem 1 hairpin (S1).

The experiments were carried out in 50 mM KCl in the absence of divalent cation, and the observed van't Hoff enthalpy was found to be comparable to the predictions based on the nearest-neighbor model. The stability of non-canonical interactions formed in state I does not influence the overall stability because they melt before the melting of the stem 1. With the increasing of Mg^{2+} concentration, no additional folding or energetically significant loop-stem interactions has been found.

(3) The crystal structure of another -1 frameshifting pseudoknot, from beet western yellows virus (BWYV), reveals many loop-stem non-canonical tertiary interactions. In particular, nucleotide C8 in loop 1 forms a well-defined base-triple contact with the G12-C26 base pair in stem 2 (Fig. 5.4). Another structural feature revealed by X-ray crystallography, is the loop 2-stem 1 interactions where loop 2 forms a series of non-canonical hydrogen-bonding contacts with the minor groove of stem 1.

The studies of the unfolding thermodynamics of BWYV pseudoknot [34] has shown that this non-canonical tertiary interactions are strongly pH-dependent and make a substantial enthalpic contribution to the pseudoknot stability. It was found that mutations resulting in the disruption of either C8·G12-C26 triplex or the loop 2-stem 1 interactions, greatly destabilize the pseudoknot.

The equilibrium unfolding of BWYV pseudoknot includes three distinct transitions. The first transition is attributed to the melting of non-canonical tertiary contacts (the base

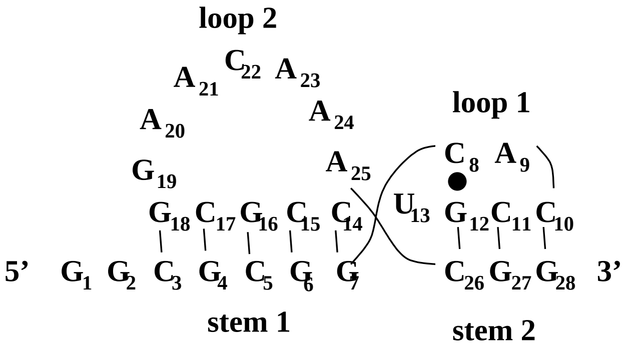


Figure 5.4: The BWYV frameshifting pseudoknot.

triplet); it is followed by unfolding of stem 2, and stem 1 unfolds last. The non-canonical tertiary structure contributes at pH 6.0 nearly 30 kcal/mol in unfolding enthalpy and ≃4 kcal/mol in stability in addition to ΔH and ΔG calculated from the nearest-neighbor model and accounting for the unfolding of canonical base pair stacks. The pH-dependence of the pseudoknot stability is totally attributed to the protonation of N3 of C8 in loop 1, which enables the formation of a third hydrogen bond to the G12-C26 base pair of stem 2 and results in higher stability of the non-canonical tertiary structure. The experiments have shown that non-canonical loop-stem interactions are absolutely required for stabilization of the BWYV pseudoknot, because the intrinsic stability of stem 2 is low and doesn't benefit from coaxial stacking with stem 1.

From the analysis of melting experiments it becomes obvious that the pseudoknot folding-unfolding is a complex process which often involves multiple sequential transitions. To correctly interpret the experimental data, it is important to have (1) the energy parameters for non-canonical interactions, and (2) the theoretical model which can predict unfolding thermodynamics and account for all possible intermediates. Two bottlenecks for modeling tertiary folding are the tertiary conformational entropies and the ion effects in tertiary interactions. The first problem is a focus in this research.

5.3 Previous models for RNA pseudoknot thermodynamics.

5.3.1 Model based on a modified Jacobson-Stockmayer approximation.

The first polymer principle-based theoretical evaluation of free energy parameters for H-pseudoknots has been done by Gultyaev et al [35]. The authors considered H-pseudoknots with not more than one nucleotide at the junction between stems (Fig. 5.1b). The free energy of such a structure $\Delta G = \Delta H - T\Delta S$ has been approximated by the sum of free energies of stems and loops. Whereas the stacking energy of stems ΔH can be obtained from the nearest-neighbor model of helix propagation [36], the loop entropies ΔS which mainly contribute to the loop free energies, need to be estimated. It has been done using the Jacobson-Stockmayer [37] approximation for the entropy of N-mer loop:

$$S(N) = R(Nln\Omega - [A + \frac{3}{2}lnN]),$$

where R is universal gas constant, $Rln\Omega$ is the conformational entropy of the free chain per monomer, and A is the constant depending on the loop closure. The intra-loop excluded volume effect has been taken into account by replacement of $\frac{3}{2}$ by 1.75 [38]. Thus, the free energy change associated with the formation of the N-mer loop can be estimated by:

$$\Delta G = RT(A_{loop} + 1.75 lnN)$$

with the assumption that there is no enthalpic contribution to the loop stability ($\Delta H = 0$). The constant A_{loop} depends on the loop type. The Jacobson-Stockmayer approximation is valid for secondary structure (hairpin, internal, and bulge) loops. To apply the formula to pseudoknot loops, some specific features of pseudoknot topology should be taken into account. It results in the following expressions for free energies of loops $L1$ and $L2$:

$$\Delta G_{L1} = A_{deep}(S2) + 1.75RT ln(1 + N - N_{mindeep}(S2));$$

$$\Delta G_{L2} = A_{shallow}(S1) + 1.75RT ln(1 + N - N_{minshallow}(S1)),$$

where A parameters depend on whether the loop spans deep or shallow groove and on the stem length, $N_{mindeep}(S2)$ and $N_{minshallow}(S1)$ are lengths of the shortest possible loops for the given groove type and stem length, and unity is added to make the logarithm equal to zero for the minimal value of N. The parameters in the formulas for ΔG have been estimated for several known pseudoknots obtained from experiments and/or philogenetic comparisons. In particular, the upper limits of parameters have been derived from the requirement that the pseudoknot is more stable then alternative structures (hairpins formed by stems $S1$ and $S2$). As a result of the above considerations, the set of the free energy parameters have been proposed.

Such estimations are rather crude due to the neglected sequence-dependence of the loop entropy as well as the simplified treatment for the excluded volume interactions between loops and stems and within the loops. Nevertheless, the estimation is more accurate than the previously used single value of 4.2 kcal/mol for free energy of all pseudoknot loops [39] since it accounts for the dependence of pseudoknot free energy on lengths of loops and stems. The estimated free energy parameters have been used for computer predictions of RNA structures and were proven not to overestimate pseudoknot stabilities significantly. Moreover, the proposed thermodynamic parameters have also been tested on two model pseudoknot systems, for which the melting experiments have been conducted. The melting temperatures for each of two pseudoknots have been estimated from the condition of co-existence of the pseudoknot and its alternative hairpins and found to be within an error of 5°C as compared with the experimental results.

5.3.2 The 3D lattice model for H-pseudoknots

The three-dimensional lattice model for H-pseudoknots has been developed by Lucas & Dill [40]. There are no restrictions imposed on the number of monomers at the junction between stems. The model can predict the density of states and the partition function for all possible pseudoknot conformations.

The key problem is the calculation of the number of conformations for a given pseudoknot. The following method has been proposed. The pseudoknot without tails is decomposed into two pseudoknot core units, each consisting of a stem and a loop (Fig. 5.5).

The pseudoknot core unit is denoted $U(n, m)$, where n is the number of base pairs in the stem (double-stranded hairpin), and m is the number of monomers in the loop. If $n = 0$, the pseudoknot core unit is an $m - step\ polygon$ (denoted $U^{3D}(m)$), i.e. an m-step neighbor-

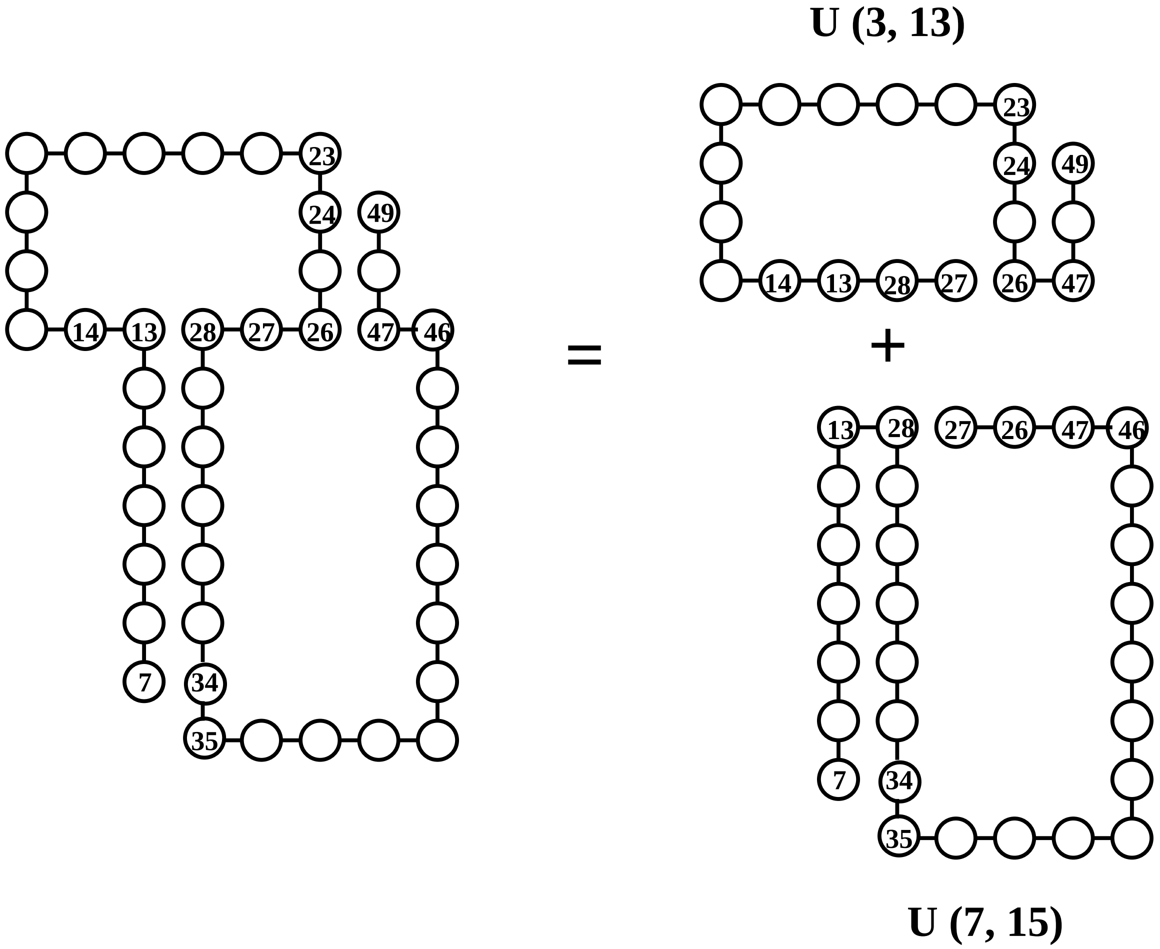


Figure 5.5: The H-pseudoknot without tails is decomposed into two pseudoknot core units U(3,13) and U(7,15). Monomers 13, 28, 27, 26, and 47 are superimposed in the two pseudoknot core units to form the pseudoknot.

avoiding walk with the only contact formed between the first and the last monomers. The case $m = 0$ corresponds to a planar n-step polygon ($U^{2D}(n)$) because in the model the stem conformations are assumed to lie in the plane which is perpendicular to the stem base pairs. This assumption is not physical because there is no special plane in which helical stem can bend freely. The numbers of conformations for each type polygons (denoted as $\Omega_{U^{3D}(m)}$ and $\Omega_{U^{2D}(n)}$) have been computed by exact enumeration for n (or m)< 20 and asymptotic expressions [41] have been used for longer chains. Then the number of conformations of the pseudoknot core unit is:

$$\Omega_{U(n,m)} \simeq \frac{2\Omega_{U^{2D}(n)}}{\Omega_{U^{3D}(n)}}\Omega_{U^{3D}(m+n)},$$

where the quantity $\Omega_{U^{2D}(n)}/\Omega_{U^{3D}(n)}$ gives the fraction of planar three-dimensional n-mer loops.

The second step is to assemble the pseudoknot from two pseudoknot core units, $U(n1, m1)$ and $U(n2, m2)$ (Fig. 5.5), and tails. The two pseudoknot core units have several common monomers (their number denoted o). Therefore, the number of pseudoknot conformations can be estimated as a product of numbers of pseudoknot core units conformations divided by the number of conformations $\omega(o)$ of the o-mer single-stranded segment. To take care of local excluded volume interactions between pseudoknot core units, an excluded volume term π has been introduced. π has been found empirically to depend on the length of the nonoverlapping single-stranded section of the loops and assigned to be $\pi = 1/70$ if $m1 - o > 2$ and $m2 - o > 2$, and $\pi = 1/35$ otherwise. The number of pseudoknot conformations is given then by

$$\Omega_{pseud} \simeq \pi\frac{\Omega_{U(n1,m1)}\Omega_{U(n2,m2)}}{\omega(o)}.$$

When tails of lengths $t1$ and $t2$ are added to the pseudoknot, the number of pseudoknot conformations should be multiplied by the numbers of tails conformations (three-dimensional neighbor-avoiding walks of lengths $t1 - 1$ and $t2 - 1$). To account for the excluded volume interactions between tails and the rest of the structure, another empirically found excluded volume term $\pi_t = 2/3$ has been introduced for each tail.

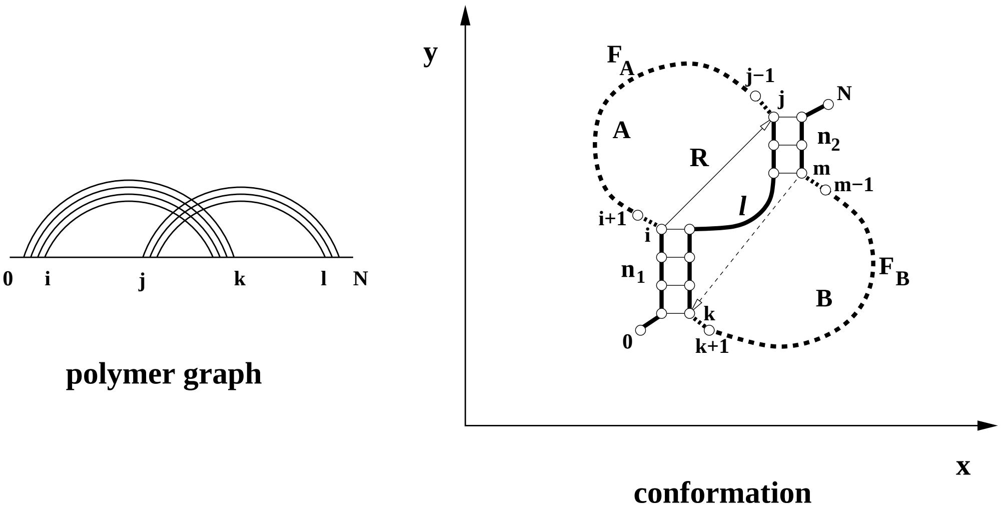
The advantages of the method are that (a) it is much more efficient than the exact computer enumeration, (b) the loop-loop and loop-stem correlations are considered from polymer principle instead of empirical estimations, and (c) the excluded volume effect is rigorously considered.

However, the model is restricted to the 3D lattice conformation and cannot treat realistic RNA pseudoknots, disallows the formation of possible partially unfolded and misfolded intermediate states in the pseudoknot folding process, and employs an unphysical assumption about the bending of the helix.

In the following section, we develop a statistical mechanical model for RNA pseudoknots that can treat the excluded volume effect and the nonadditivity arising from the correlation between different structural subunits. Though our major focus here is on the H-pseudoknot and its partially unfolded states, the methodology developed in this work is general and can be extended to treat more complex pseudoknotted structures. In addition, the method uses graphical representation for intrachain contacts and is thus independent of any specific chain representation. For illustrations, we use two-dimensional (2D) lattice chain conformations, where the excluded volume effect is accounted for by configuring the chain conformations as self-avoiding walks in a 2D lattice. An advantage of the present theory is its ability to treat the complete conformational ensemble for the pseudoknotted structures (and the secondary structures), including all the partially folded and misfolded states.

5.4 A polymer statistical mechanical model for RNA pseudoknots.

For the calculation of the partition function $Q(T)$ using the graph theoretic approach (Eq. 3.2), it is necessary to calculate the conformational count Ω for the given graph. We start from the simple H-pseudoknot, and then generalize the method to treat more complex pseudoknots. The H-pseudoknot which we will consider is shown in Fig. 5.6.



polymer graph

conformation

Figure 5.6: The H-pseudoknot is subdivided in our theory into enlarged interface (solid bold) and two free loop segments (dashed bold). The coordinate system is chosen so that the *y*-axis is along the first pseudoknot stem. The number of free loop segment conformations is approximated by the function of the free loop segment's length and of coordinates of the vector **R**.

For mathematical convenience, we cut off dangling tails, leaving only one bond of each tail to account for the excluded volume interactions between the tails and the rest of the structure. We call such tail-free structure “structure element”. A full structure consists of two parts: structure element and the tails (in the 5’ and 3’ terminal regions of the nucleotide chain). Loops A and B share the common single-stranded (interfacial) chain segment and are also constrained by the helical stems, therefore, the conformations of loops A and B are strongly correlated with each other. The basic idea in our pseudoknot conformational entropy theory is to divide a pseudoknot into three components, namely, the enlarged interface (shown as solid lines in Fig. 5.6), and the two non-interfacial free loop segments F_A and F_B (shown as dashed lines in Fig. 5.6) of lengths f_A and f_B, respectively. The enlarged interface includes two stems of lengths n_1 and n_2, the single-stranded interfacial segment of length l, and the tail monomers 0 and N in Fig. 5.6.

A great challenge in the entropy calculation is how to treat the excluded volume effect. According to the decomposition of the pseudoknot structure, we classify the excluded volume interactions into two types: (a) between the interfacial monomers (i.e., monomers in the enlarged interface) and (b) between the free loop monomers (i.e., monomers in each free loop segment) and between the free loop monomers and the interfacial monomers. We treat the former (type a) excluded volume effect by considering all the viable self-avoiding enlarged interface conformations (denoted as $\mathbf{I}$) that *(i)* do not make self-contacts other than those specified by the graph and *(ii)* allow viable positioning of monomers $i+1$, $j-1$, $k+1$, and $m-1$ without causing additional contacts. In terms of $\mathbf{I}$, we compute the number of conformations Ω_P of the pseudoknot as a sum over all the possible viable conformations $\mathbf{I}$

of the enlarged interface (Eq. 4.1):

$$\Omega_P = \sum_{\mathbf{I}} \Omega_{\mathrm{f}}(\mathbf{I}, f_A) \cdot \Omega_{\mathrm{f}}(\mathbf{I}, f_B),$$

where $\Omega_{\mathrm{f}}(\mathbf{I}, f_A)$ and $\Omega_{\mathrm{f}}(\mathbf{I}, f_B)$ are the numbers of conformations of the free loop segments F_A and F_B for a given $\mathbf{I}$, respectively. The later (type b) excluded volume effect plays a crucial role and should be taken into account in the computation of $\Omega_{\mathrm{f}}(\mathbf{I}, f_A)$ and $\Omega_{\mathrm{f}}(\mathbf{I}, f_B)$.

In Eq. 4.1, we neglect the volume exclusion between the free loop segments F_A and F_B. This is because, first, the free loop segments are spatially separated by the enlarged interface and, second, the steric hindrance between each free loop segment and the interface is accounted for in the calculation of Ω_{f} (see next section). For very large loops and long free loop segments, the volume exclusion between F_A and F_B may become important. However, in that case, the chain entropy is large, so the error caused by neglecting the $F_A - F_B$ volume exclusion is relatively small as compared with the (large) chain entropy.

If one has a table for $\Omega_{\mathrm{f}}(\mathbf{I}, f_A)$ and $\Omega_{\mathrm{f}}(\mathbf{I}, f_B)$ for all the possible free loop segment lengths f_A and f_B and enlarged interface conformations $\mathbf{I}$, the computation of Ω_P for the given graph (pseudoknot) from Eq. 4.1 would be efficient and straightforward. However, the number of possible parameter sets $(f_A, f_B, \mathbf{I})$, grows exponentially with the length of the interface, which makes the tabulation of Ω_{f} for all the possible parameter sets practically impossible. In the next section, we develop a method to approximate the conformational count Ω_{f} for the free loop segments. The derived expression for Ω_{f} would be mathematically convenient and computationally efficient.

5.4.1 Number of conformations of a free loop segment.

Because of the obviously equal roles of loops A and B in Fig. 5.6, instead of calculating Ω_{f} for the two loops separately, we focus on the calculation for one of the loops, namely, loop A. In what follows, we derive an approximate expression for the number of conformations Ω_{f} for the free loop segment F_A in loop A. The derived result for loop A would be equally applicable to loop B.

Strictly speaking, the number of conformations $\Omega_{\mathrm{f}} = \Omega_{\mathrm{f}}(\mathbf{I}, f_A)$ for the free loop segment F_A (in Fig. 5.6) is a function of the enlarged interface conformation $\mathbf{I}$, which is dependent on the interface length l and the length of the two helical stems (n_1 and n_2), etc. However, as an approximation, we replace the $\mathbf{I}$-dependence by the end-end vector $\mathbf{R}$-dependence; see the vector from i to j in Fig. 5.6. The end-end vector for F_B is shown as the dashed arrow in Fig. 5.6. We further define a two-dimensional Cartesian coordinate system by choosing the y-axis along stem 1 (directed upward), as shown in Fig. 5.6. Vector $\mathbf{R}$ can be described through the components: $\mathbf{R} = (x, y)$, where $x = x_j - x_i$ and $y = y_j - y_i$ in Fig. 5.6. In terms of the components of $\mathbf{R}$, we reduce the Ω_{f} function from a large parameter space (for the interfacial chain conformations $\mathbf{I}$) to a much smaller two-variable parameter space (for $\mathbf{R} = (x, y)$):

$$\Omega_{\mathrm{f}}(\mathbf{I}, f) \simeq \Omega_{\mathrm{f}}(\mathbf{R}, f) \to \Omega_{\mathrm{f}}(x, y, f).$$

Since the information about $\mathbf{I}$ is now embedded in $\mathbf{R}$, for an f_A-mer free loop segment F_A with fixed end-end vector $\mathbf{R} = (x, y)$, we compute the conformational count $\Omega_{\mathrm{f}}(\mathbf{R}, f_A)$ as an average over all the possible enlarged interface conformations $\mathbf{I}$ with the fixed end-end vector **R**:

$$\Omega_{\mathrm{f}}(\mathbf{R}, f_A) \simeq \frac{\sum_{\mathbf{I}} \Omega_{\mathrm{f}}(\mathbf{I}, f_A)}{\Omega_{\mathbf{I}}},$$

where $\Omega_{\mathbf{I}}$ is the number of all the viable enlarged interface conformations with the given $\mathbf{R} = (x, y)$, and $\sum_{\mathbf{I}}$ is the sum over all these conformations. Physically, the calculated $\Omega_{\mathrm{f}}(x, y, f_A)$ function is the average conformational count of F_A for each given enlarged interface conformation **I**.

In fact, the first helix stem length n_1 affects the conformational count of F_A only through weak excluded volume interactions between F_A, the first stem and the tail attached to the first stem (which is so far one monomer long). For the helix stem length $n_1 \geqslant 2$ base stacks, the conformational count for F_A is nearly independent of n_1. In our calculation, we choose $n_1 = 3$ base stacks and fix the orientation of helix stem 1 to be upright (in the y direction). We then enumerate all the possible conformations of the enlarged interface for all the possible values of l and n_2 such that $1 \leqslant l + n_2 \leqslant 10$. For each enlarged interface conformation **I** and the given free loop segment length f_A ($f_A \leqslant 21$), we calculated the number of free loop segment conformations $\Omega_{\mathrm{f}}(\mathbf{I}, f_A)$ by means of exact computer enumerations of self-avoiding walks on 2D lattice. As an illustration, in Fig. 5.7a, we show all the viable conformations of the enlarged interface with fixed $n_1 = 3$ (base stacks) and end-end vector $\mathbf{R} = (x, y) = (3, 4)$. For the given end-end vector **R**, the length n_2 of stem 2 has only four possible viable values: $n_2 = 1, 2, 3, 4$, and the length of the single-stranded interfacial segment is restricted by $l \leqslant 10 - n_2$. The possible interface conformations are represented as self-avoiding random walks on the dashed grid (with one of the conformations drawn as solid line). Fig. 5.7b shows how the volume exclusion between the free loop segment and

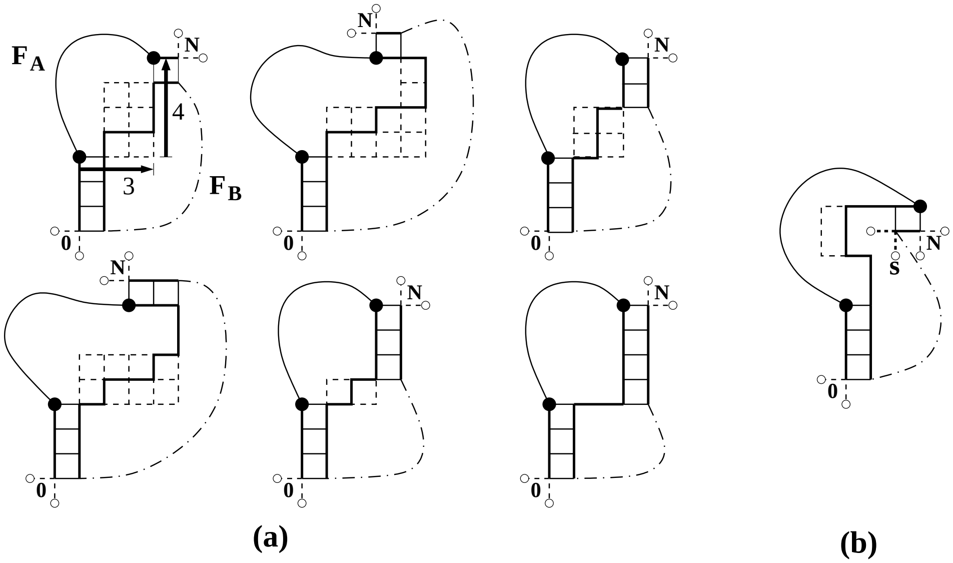


Figure 5.7: (a) All the viable conformations of possible enlarged interfaces that give end-end vector $\mathbf{R} = (x, y) = (3, 4)$ for the free loop segment F_A. The position and length of the first stem are fixed. Big black circles denote the end monomers of the free loop segment F_A. Part of the 2D lattice which can be covered by the viable conformations of the interfacial segment for possible length and position of the second stem is shown with dashed lines. The dash-dotted line denotes the F_B segment. Two possible positions for 0 and N are shown for each interfacial conformation. (b) An example of the conformation of the enlarged interface which is not viable because all the possible positions of the monomer s would make an additional contact with the interfacial segment.

the interface determine the viability of conformations. The depicted conformation of the enlarged interface in Fig. 5.7b is not viable because it is impossible to position monomer s in the square lattice without making additional contacts with the single-stranded interfacial segment.

5.4.2 Number of conformations of pseudoknots.

Pseudoknot with two stems.

With the reduced function $\Omega_{\mathrm{f}}(\mathbf{R}, f)$ for free loop segments F_A and F_B, we obtain the conformational count of the pseudoknot from Eq. 4.1 by replacing $\Omega_{\mathrm{f}}(\mathbf{I}, f)$ by $\Omega_{\mathrm{f}}(\mathbf{R}, f)$:

$$\Omega_P \simeq \sum_{\mathbf{I}} \Omega_{\mathrm{f}}(\mathbf{R}_A,\ f_A) \cdot \Omega_{\mathrm{f}}(\mathbf{R}_B,\ f_B),$$

where $\mathbf{R}_A$ and $\mathbf{R}_B$ are the end-end vectors of the free loop segments F_A and F_B, respectively, for a given conformation $\mathbf{I}$ of the enlarged interface. As shown in Fig. 5.6, $\mathbf{R}_A$ is the vector $i \to j$ and $\mathbf{R_B}$ is the vector $m \to k$. The coordinate system for vector $\mathbf{R_B}$ is defined by stem 2 in the same way as the coordinate system for vector $\mathbf{R_A}$ is defined by stem 1 (the y-axis is along the corresponding stem, directed upright).

The method can be generalized to treat more complicated pseudoknots. Here we demonstrate how to extend the method to treat pseudoknots with an internal loop in the stems (Fig. 5.8a). Such “pseudoknot” conformations are important because they may emerge as partially folded or misfolded intermediates in the pseudoknot folding process.

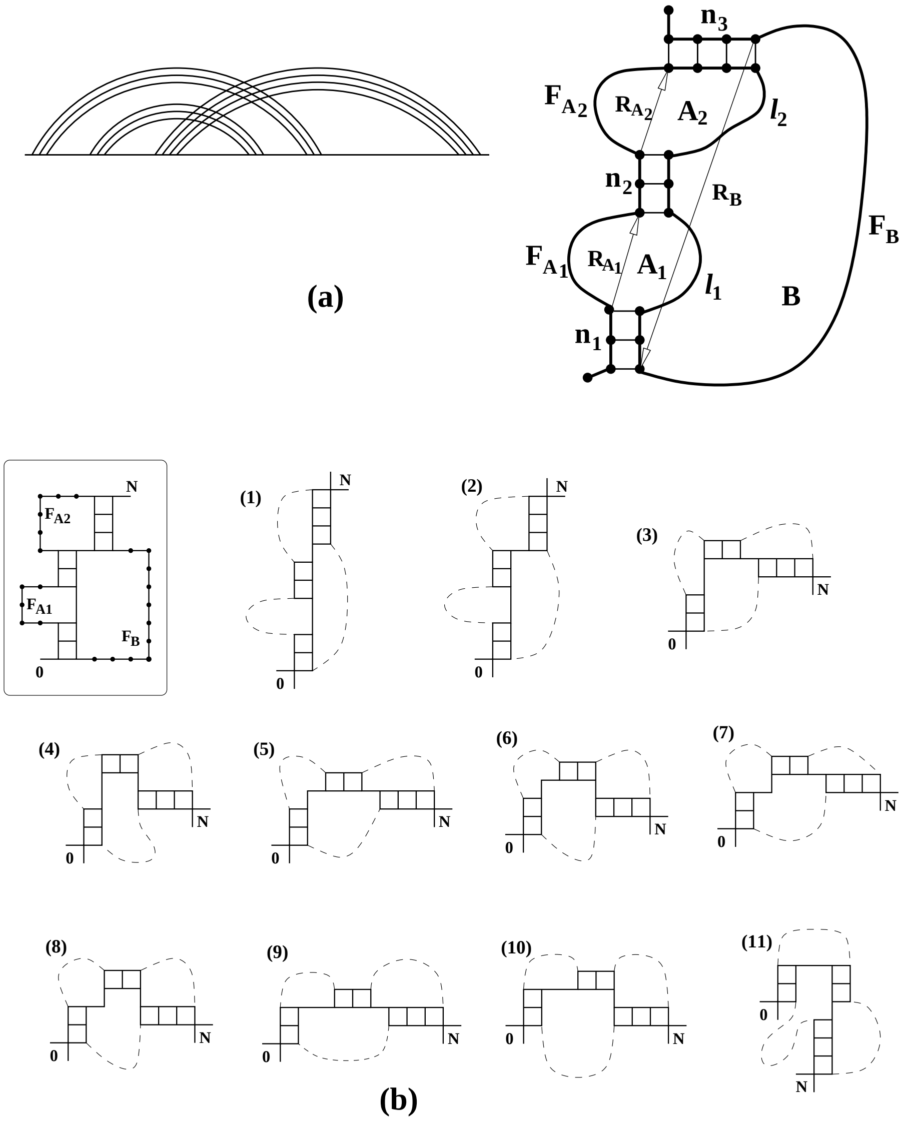


Figure 5.8: (a) The polymer graph and structure of a “pseudoknot” with an internal loop (A_1) formed in a helix stem. $\mathbf{R_{A_1}}, \mathbf{R_{A_2}}$, and $\mathbf{R_B}$ shown in the figure are the end-end vectors of the free loop segments F_{A_1}, F_{A_2} and F_B, respectively. (b) The structure of the pseudoknot with three stems chosen to illustrate the method and the viable conformations of the enlarged interface. Two positions are shown for each of the end monomers 0 and N.

Generalized pseudoknots with three stems.

As shown in Fig. 5.8a, the enlarged interface for such a structure consists of three stems (n_1, n_2, n_3), two single-stranded interfacial segments (l_1, l_2), and two end monomers $(0, N)$. For a given conformation **I** of the enlarged interface, the end-end vectors $\mathbf{R}_{A_1}$, $\mathbf{R}_{A_2}$, and $\mathbf{R}_B$ of the free loop segments F_{A_1}, F_{A_2}, and F_B, respectively, can be unambiguously determined. The conformational count of the "pseudoknot" in Fig. 5.8a can be calculated from the following sum over all the possible conformations **I** of the enlarged interface:

$$\Omega_P \simeq \sum_{\mathbf{I}} \Omega_{\mathrm{f}}(\mathbf{R}_{A_1},\ f_{A_1}) \cdot \Omega_{\mathrm{f}}(\mathbf{R}_{A_2},\ f_{A_2}) \cdot \Omega_{\mathrm{f}}(\mathbf{R}_B,\ f_B). \tag{5.1}$$

Illustrative calculation for the number of conformations of pseudoknots with three stems.

In Fig. 5.8b, we show the structure of a pseudoknot with an internal loop formed in one of the helix stems so the "pseudoknot" is partially folded (unfolded) and contains three helix stems. Also shown in the figure are all the 11 viable conformations of the enlarged interface numbered from 1 to 11 (excluding the end monomers 0 and N). Since monomers 0 and N each can have two possible positions, from Eq. 5.1, we calculate the number of pseudoknot conformations as the following:

$$\Omega_P \simeq (2 \times 2) \sum_{\mathbf{K=1}}^{11} \Omega_{\mathrm{f}}(\mathbf{R}_{A_1},\ f_{A_1}) \cdot \Omega_{\mathrm{f}}(\mathbf{R}_{A_2},\ f_{A_2}) \cdot \Omega_{\mathrm{f}}(\mathbf{R}_B,\ f_B), \tag{5.2}$$

where **K** denotes an interface conformation shown in Fig. 5.8b. The lengths of the free loop segments for the given pseudoknot are: $f_{A_1} = 5$ for F_{A_1}, $f_{A_2} = 6$ for F_{A_2}, and $f_B = 11$ for F_B. The details of calculations are given in Table 5.1, where we show, for each interfacial

Table 5.1: Illustrative calculation for the number of conformations of the pseudoknot in Fig. 5.8b

K	(x_{A_1}, y_{A_1})	$\Omega_f^{A_1}$	(x_{A_2}, y_{A_2})	$\Omega_f^{A_2}$	(x_B, y_B)	Ω_f^B	$\Omega_f^{A_1} \cdot \Omega_f^{A_2} \cdot \Omega_f^B$
1	(0,2)	1.0	(1,4)	2.6	(1,7)	23.2	60.3
2	(0,2)	1.0	(2,3)	3.0	(2,6)	55.5	166.5
3	(1,3)	1.8	(1,4)	2.6	(-3,3)	21.6	101.1
4	(1,3)	1.8	(2,3)	3.0	(-2,2)	8.0	43.2
5	(2,2)	1.1	(1,4)	2.6	(-2,4)	32.1	91.8
6	(2,2)	1.1	(2,3)	3.0	(-1,3)	17.0	56.1
7	(2,2)	1.1	(1,4)	2.6	(-2,4)	32.1	91.8
8	(2,2)	1.1	(2,3)	3.0	(-1,3)	17.0	56.1
9	(3,1)	0.5	(1,4)	2.6	(-1,5)	52.5	68.2
10	(3,1)	0.5	(2,3)	3.0	(0,4)	37.3	55.9
11	(4,0)	0.0	(1,4)	2.6	(-1,1)	0.0	0.0
							sum=791.1

conformation **K**, the coordinates (x, y) of the end-end vectors and the approximate numbers of conformations Ω_f for each free loop segment (F_{A_1}, F_{A_2}, and F_B).

As shown in Table 5.1, Eq. 5.2 gives $\Omega_P^{\text{appr}} = 4 \times 791.1 = 3164.4$, which is close to the result from the exact computer enumeration: $\Omega_P^{\text{exact}} = 2799$.

5.4.3 Pseudoknot partition function calculation. Thermal unfolding of pseudoknots.

Central to the folding thermodynamics is the partition function. The calculation of partition function (Eq. 3.2) for a given nucleotide sequence requires the enumeration of all the possible polymer graphs and the counting of conformations accessible to each graph. In the present study, we consider the complete ensemble of pseudoknotted structures as well as the secondary structures. For the secondary structure partition function, we use a previously developed statistical mechanical theory [16], [17].

For the pseudoknotted conformations, in the preceding sections, we ignored the full tails and took into account only the first monomer of each tail closest to the pseudoknot structure element to represent the volume exclusion effect. We now add the full tails back to the calculation.

To take into account tails, we have used pre-calculated [17] tables for the numbers of tail conformations $\Omega_T(t)$ for tail length $t \leqslant 24$ and the following fitted formula for longer tails [17]:

$$ln\ \Omega_T(t) \simeq -2.62208 + 0.83927\ t + 0.30984\ ln\ (t).$$

This formula has also been used to obtain the number of conformations of the open (fully unfolded) chain. With the conformational count of the tails, we can calculate the number of conformations of the pseudoknot with tails as the product of the number of conformations of the pseudoknot structure element Ω_P and the numbers of conformations of tails: $\Omega = \Omega_P\ \Omega_T(t_1)\ \Omega_T(t_2)$, where t_1 and t_2 are lengths of the two tails at the 5'- and the 3'-terminal, respectively.

The Go-model pseudoknots.

We consider a 33-mer chain with a fully folded native state shown in Fig. 5.9a. As a simplified (Go-type) model, we assume that no other contacts besides those (native contacts) depicted in Fig. 5.9a can form. The interaction energy of each base pair stack is assigned to be -3 or -1 as shown in the figure. The energy of an isolated (unstacked) base pair is assumed to be 0. Since in the Go-model, only the native contacts can form or disrupt in the folding/unfolding process, the conformational ensemble can be generated from the different ways to break the native contacts. Examples of partially unfolded states are also shown in Fig. 5.9a. The unzipping of either helix stem can occur at the top, bottom or internal base pair of the stem. The partially unfolded states can be conveniently represented by two parameters: p denotes the state of stem 1 and q - of stem 2. The partially unfolded states of a stem depend on the stem length and can be enumerated. The possible states of stems of our native pseudoknot (5- and 2-base stack) are shown in Fig. 5.9b. In this way, each possible state can be described by a pair of numbers (p, q), which defines the set of contacts. For example, $(p, q) = (1, 1)$ is the native pseudoknot in Fig. 5.9a, $(2, 4)$ is a hairpin with an internal loop, and $(2, 3)$ is a pseudoknot with an internal loop in stem 1.

The parameters p and q are used as the labels for the states of the stems. Each (p, q) pair unambiguously defines a pseudoknotted structure. The p's and q's shown in Fig. 5.9b are exhaustive for the short stems shown in the figure. As a caveat, we note that for illustrative purpose, we here use the two-dimensional lattice model, which excludes some conformations due to the lattice constraint (e.g., a 6-mer loop is not possible in a two-dimensional square lattice). In addition, in this section, in order to focus on the stem-loop interplay,

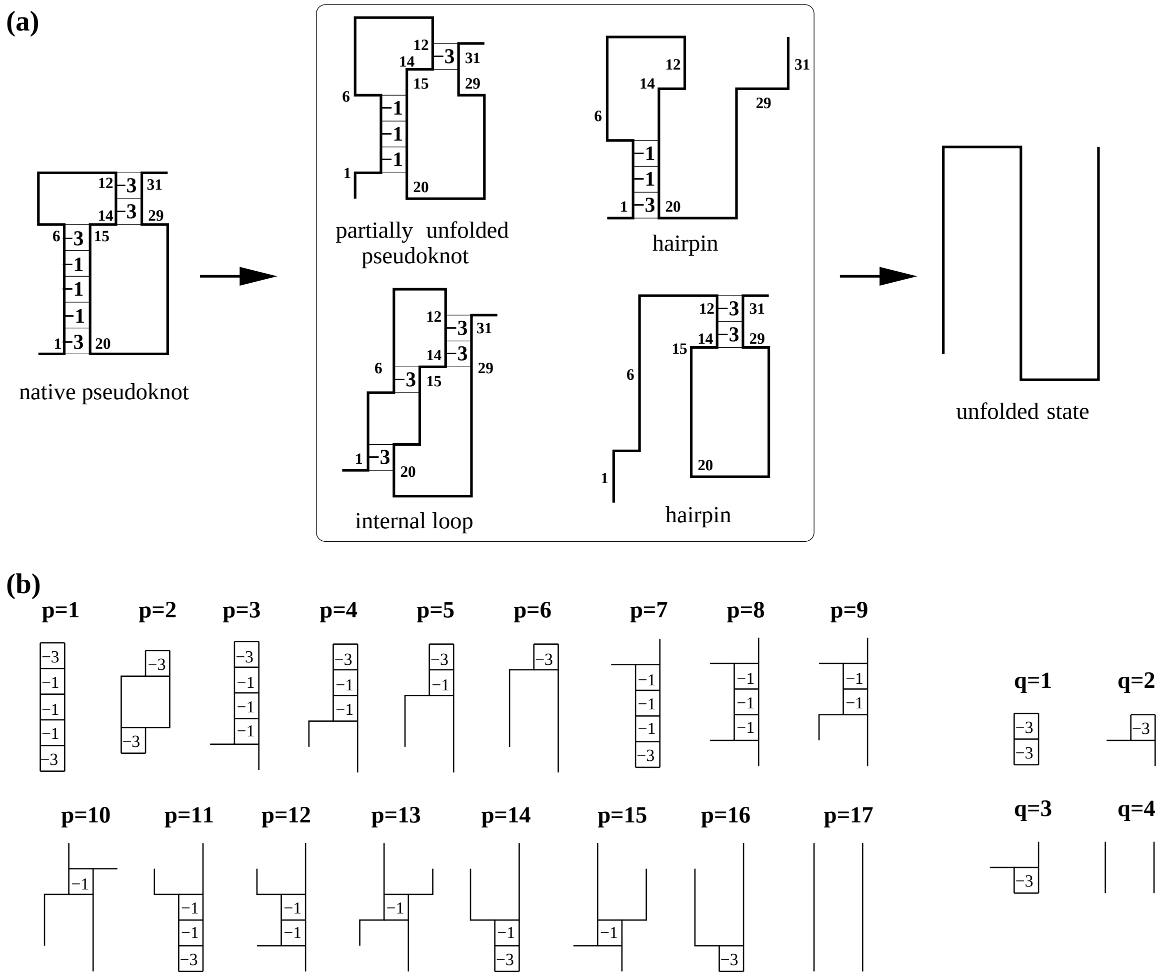


Figure 5.9: (a) The 33-mer pseudoknot-forming chain is used to illustrate the calculations for the density of states and partition function. The numbers in boxes (stacks) denote the energies of the corresponding (native) base stack. The four types of the representative partially unfolded conformations considered in the partition function calculation are shown. (b) The partially unzipped states of the (two) helix stems are labeled with parameters p and q, respectively. As a result, each state (folded, partially folded, and unfolded) can be represented by the parameter pair (p, q).

we use the Go-model, which disallows the formation of the misfolded states. For longer stems, more complex multiple internal loops can be formed. In the next section, we will go beyond the Go-model by using the complete conformational ensemble, including all the possible misfolded structures.

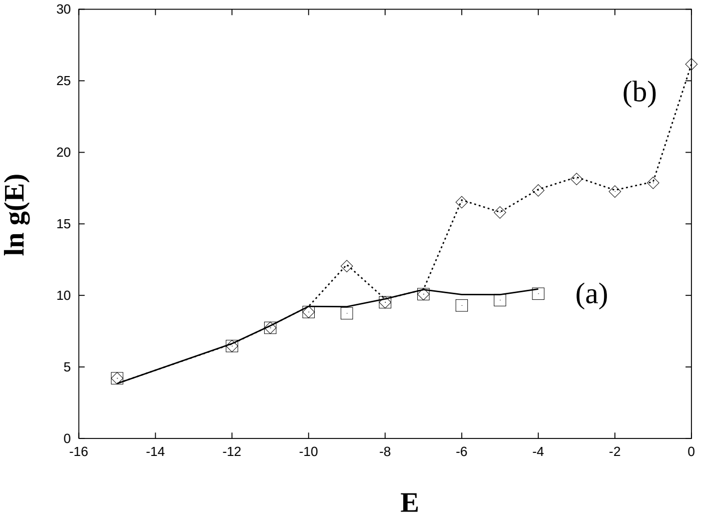


Figure 5.10: Test of the theory (line) against exact computer enumeration (symbols) for the density of states (a) for all the pseudoknotted conformations and (b) for the complete pseudoknot/hairpin/open conformational ensemble for the 33-mer chain shown in Fig. 5.9a.

To test the accuracy of the theory, we compute the density of states for all the possible pseudoknots (including all the partially unfolded states) and for the complete (pseudoknot/hairpin/open chain) conformational ensemble (including all the possible partially unfolded states) using both the theory developed here and the exact computer enumeration. Fig. 5.10 shows the test results. We find good accuracy of the theoretical prediction.

To study the folding thermodynamics for the model pseudoknot in Fig. 5.9a, we calculate the free energy landscape $F(p,q)$, which is the free energy of the macrostate for all the possible secondary structures and pseudoknots described by the conformational state (p, q). $F(p, q)$ is computed from the following equation:

$$F(p, q) = E(p, q) - k_B T \ ln \ \Omega(p, q),$$

where $E(p, q)$ and $\Omega(p, q)$ are the energy (sum of the energies of the stacks) and the number of conformations of the macrostate described by (p, q). The free energy minima correspond to stable states. $F(p, q)$ is temperature T-dependent and temperature change causes the change in the free energy landscape $F(p, q)$ and the transitions between different stable states.

Fig. 5.11A shows the free energy landscape $F(p, q)$ for the model pseudoknot in Fig. 5.9a. The landscape shows single pronounced minimum at $(p, q) = (1, 1)$ (= the native state shown in Fig. 5.9a) at low temperature and $(17, 4)$ (= fully unfolded state) at high temperature. At $k_B T = 0.672$, the landscape shows that the transition between the native state and the unfolded state involves two hairpin conformations as the intermediate states: $(17, 1)$ and $(1, 4)$. Each hairpin intermediate is formed through the disruption of a helix stem of the native pseudoknot.

In order to examine the competition between the helix and loop stability, we further calculate the free energy landscape for pseudoknots with different sizes of loops. For larger loops, as shown in Fig. 5.11B, the free energy landscape reveals 3-state transitions, where the native state, the open chain and the hairpin with the longer stem are equally populated at $k_B T = 0.6113$. The larger loop would destabilize the folded state. As a result, state $(17, 1)$, which emerges as a folding intermediate for pseudoknot with smaller loops, is now absent. This is because the 2-stack short helix stem in the $(17, 1)$ state is not stable enough

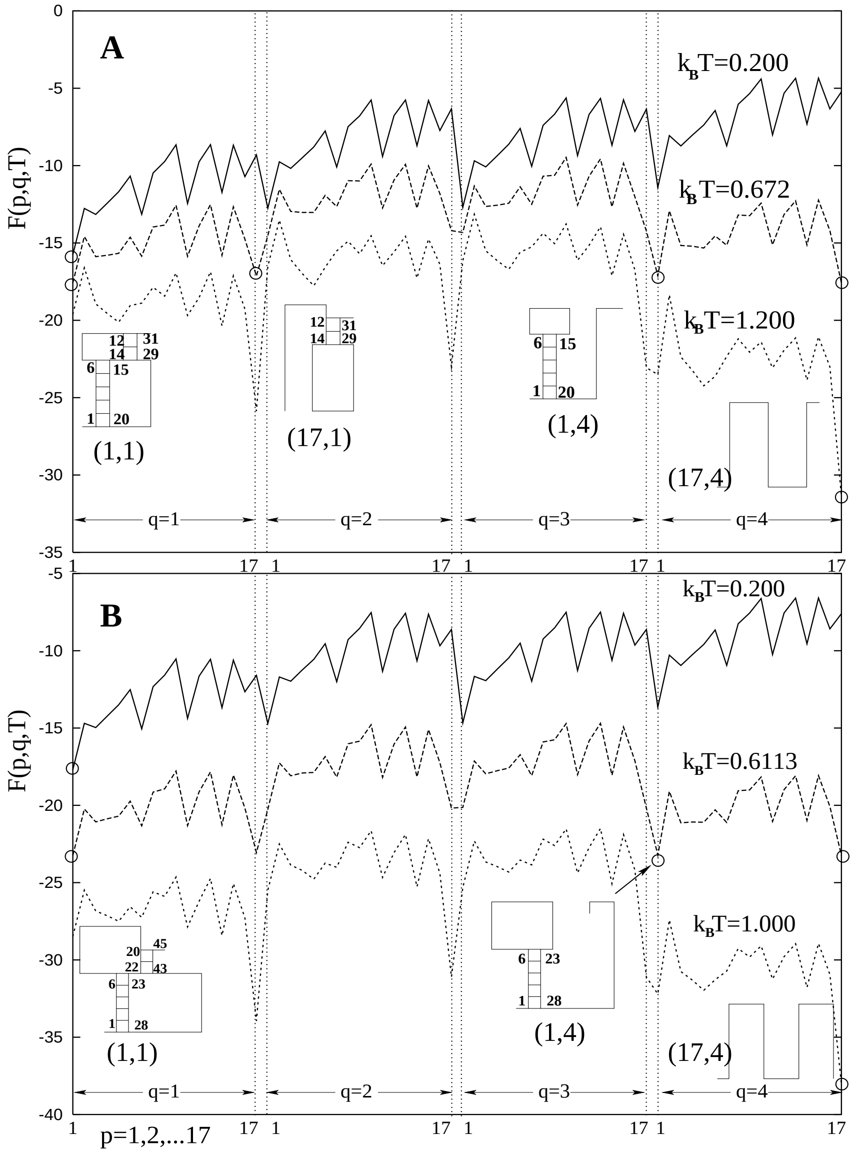


Figure 5.11: The free energy $F(p,q,T)$ as a function of structure denoted by parameters pair (p,q). Free energy minima, i.e. stable states for a given temperature, are circled, and the representative conformations shown. (A) For the pseudoknot with smaller loops, the two hairpins (17, 1) and (1, 4), each formed through the disruption of a native helix stem, emerge as stable intermediate states. (B) For the pseudoknot with larger loops, the transition is three-state: the native state, open chain and hairpin with the longer helix stem, which are equally populated at $k_BT = 0.6113$.

to compete with the destabilizing larger loops. In contrast, the state $(1, 4)$ emerges as an intermediate because the 5-stack long helix stem is sufficiently enough to compete with (the larger) loop.

Pseudoknot folding with misfolded states.

The formation of non-native contacts is not allowed in the above Go-model pseudoknots and therefore many structures, namely, the misfolded states, are excluded from consideration. In this section, we treat complete conformational ensemble, including all the possible misfolded conformations. We choose two (30- and 34-nt) pseudoknot-forming nucleotide sequences. We allow the formation of all the possible $A - U$ and $C - G$ base pairs. The energy of a base stack is equal to -1 for a stack formed by two $A - U$ base pairs, -3 for a stack formed by one $A - U$ and one $C - G$ base pair, and -4 for a stack formed by two $C - G$ pairs.

We enumerate all the possible secondary structures and pseudoknotted structures through the enumeration of all the possible polymer graphs. For each graph, we compute the energy E and the number of accessible conformations Ω. In Figs. 5.12 & 5.13 we show the free energy landscapes ($F = E - k_BT \, ln \, \Omega$) for the 30-nt and the 34-nt sequences, respectively. The free energy landscapes show contrasting folding thermodynamics for the two sequences. The sequence in Fig. 5.12 unfolds through the sequential disruption of the two helix stems. The landscape at $k_BT = 1.0$ shows the emergence of an intermediate state (= minimum in the free energy landscape) which is formed through the breaking of the less stable 2-stack helix in the native pseudoknot. In contrast, the sequence in Fig. 5.13 unfolds through the formation of a misfolded state I, shown as the minimum in the free energy landscape at $k_B T = 0.9$. The misfolded intermediate (hairpin I in Fig. 5.13) is formed through a complete rearrangement of the base pairs from the native pseudoknot.

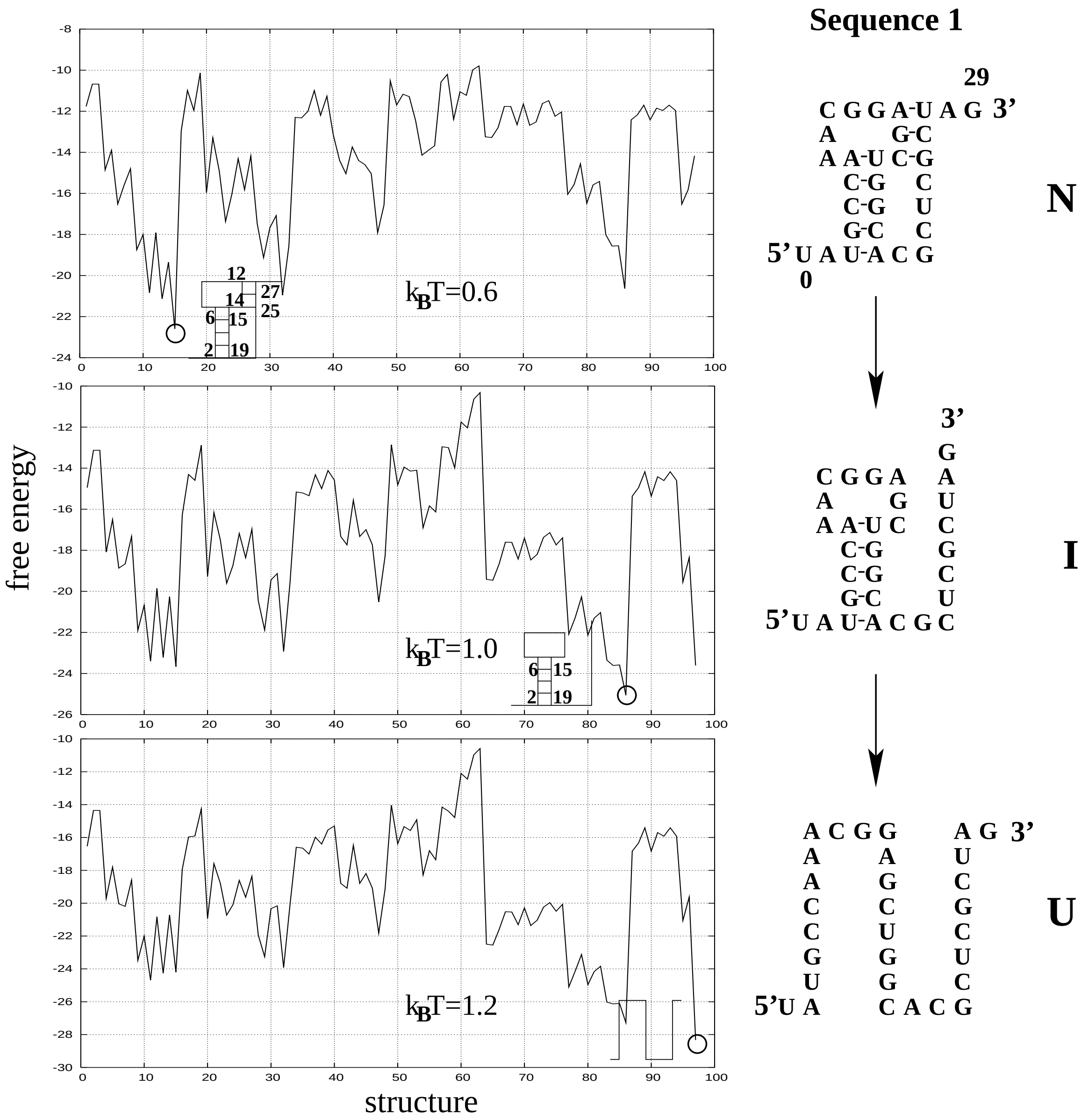


Figure 5.12: The free energy as a function of structure for a 30-mer nucleotide sequence. The unfolding is a two-step process with the (less stable) shorter helix stem disrupted first ($N \rightarrow I$) followed by the breaking of the longer (more stable) helix stem ($I \rightarrow U$).

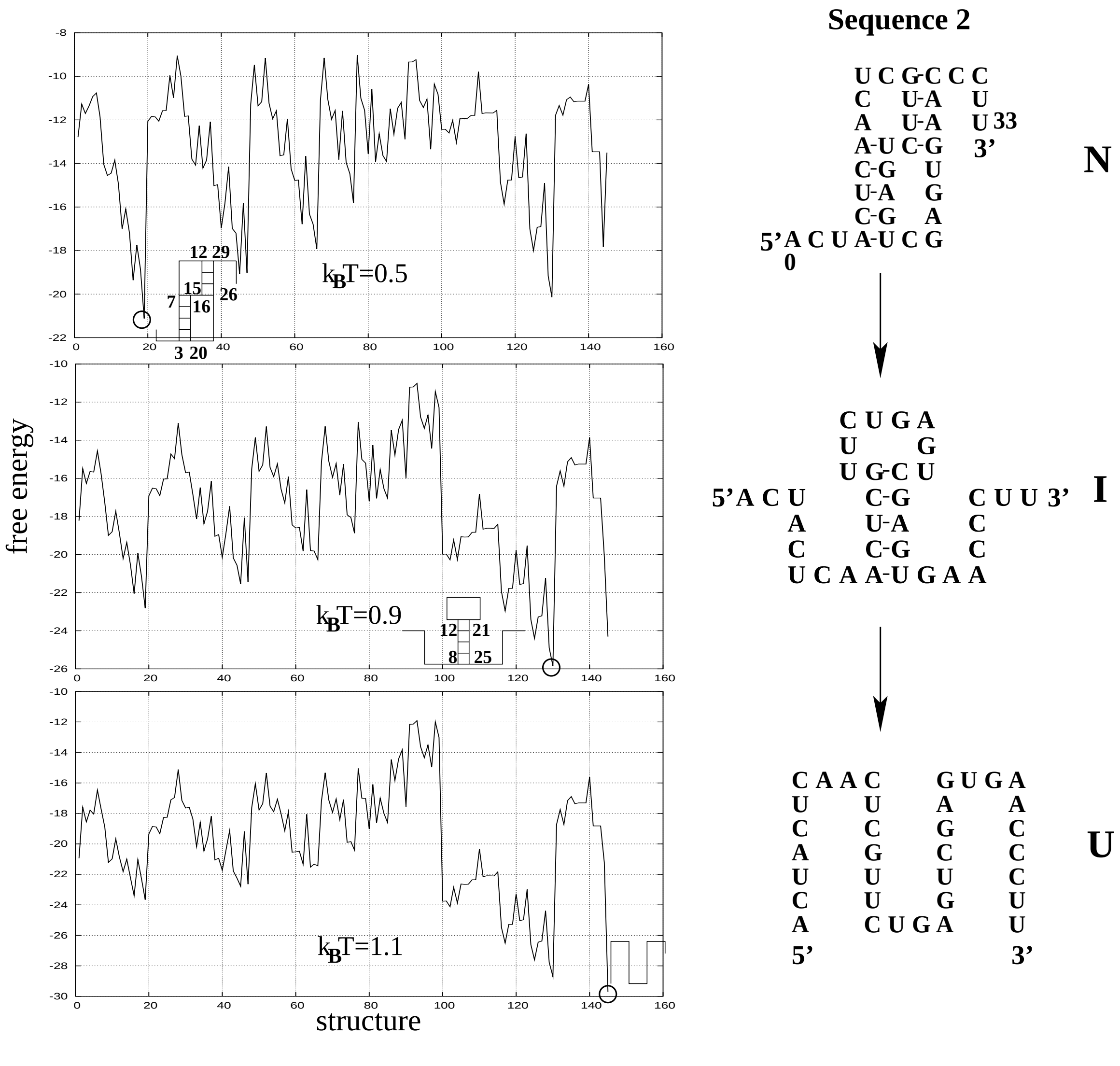


Figure 5.13: Free energy as a function of structure. The melting process for the 34-mer nucleotide sequence involves a misfolded hairpin (*I*) as an intermediate state.

Chapter 6

STATISTICAL THERMODYNAMICS FOR FORCE-INDUCED RNA FOLDING AND UNFOLDING.

The major part of this chapter (section 6.2) has been accepted for publication: *Z. Kopeikin and S.-J. Chen. Statistical thermodynamics for RNA pseudoknots. J. Chem. Phys. In press, scheduled issue: April 2006*

6.1 Force-induced RNA hairpin folding.

6.1.1 Experiments on force-induced RNA hairpin folding.

In the experiment on RNA mechanical folding-unfolding described in Section 2.2 (Liphardt et al [15]), three simple structural units of RNA were stretched using optical tweezers (Fig. 1 in [15]).

The experiment makes it possible to study thermodynamics of the folding-unfolding process. It was shown that the characteristic features in the force-extension curve can be used to obtain the unfolding free energy and the size of the structural element. The force-extension curve of the simple hairpin P5ab shows at 14.5 pN the 18 nm plateau corresponding to the hairpin unfolding. The process is reversible, which suggests the thermodynamic equilibrium. At the critical force the hairpin hopped between folded and unfolded states without intermediates. The corresponding ΔG (Gibbs free energy of the hairpin with respect to the open chain) has been determined in the following ways.

1. The hopping between folded and unfolded states is a stochastic thermally facilitated process. The probability of the hairpin opening versus force (Fig. 2B in [15]) was obtained by summing a normalized histogram of hairpins opened versus force (data from 36 consecutive pulls of one molecule). This dependence can be fit well by the statistics of a two-state system in an external field at finite temperature. The energy of the hairpin-laser trap system is $E(F) = \Delta G(F_{1/2}) - F\Delta x$, where $\Delta G(F_{1/2}) = F_{1/2}\Delta x$ is the free energy change of unfolding and stretching the hairpin at the midpoint of the transition ($F_{1/2}$) and the extension difference between the folded and unfolded

states Δx is assumed to be constant. The probability for the system to have energy E is $p(E) = 1/(1 + e^{E/k_B T})$. From this analysis, $\Delta G = 193 \pm 6 kJ/mol$ has been obtained.

2. The force-extension curve for the hairpin shows at 14.5pN the 18nm plateau corresponding to the hairpin unfolding. The process is reversible, which suggests the thermodynamic equilibrium. The unfolding free energy can be determined as the average area under the plateau, which equals the potential of mean force of folding. It gives $\Delta G = 157 \pm 20 kJ/mol$.

3. Within the critical force range where folding/unfolding hopping is observed (Fig. 2C in [15]), the equilibrium constant $K(F)$ can be obtained as a ratio of the average lifetimes of the molecule in two states. This yields $\Delta G = 156 \pm 8 kJ/mol$.

Thus we see that all three methods give the similar results and therefore are likely reliable.

In general, the results of single-molecule experiments can not be interpreted using conventional thermodynamic rules [42]. The reason is the following. The traditional bulk experiments usually deal with the most populated states of the large ensemble of molecules and yield smoothly changing, averaged over time and population values. The standard thermodynamic theory can be applied to such measurements. The single-molecule experiments, in contrast, yield data with large fluctuations, which describe the states of the individual molecule and individual trajectory, randomly deviating from the average of the population. This fluctuations affect the interpretation of data.

The results of single-molecule experiments were shown [42] to depend on the choice of the statistical ensemble, i.e. on which variables are held constant and which are allowed to fluctuate in the unfolding process. In the experiments where the molecule is stretched by force, the following two types of statistical ensembles are generally considered.

(a) Constant distance ensemble (isometric): the molecule's end-end distance is held constant and the fluctuation of the force is recorded in experiment. Such an experiment can be performed, for example, in the following way. One end of the molecule can be attached to a rigid support, and the other end is attached to the bead held in optical trap. The feedback loop controls the position of the trapped bead with respect to the other end of the molecule and cancels its fluctuations by moving the trap center. The fluctuating force, which is proportional to the displacement of the bead in the optical trap from its equilibrium position, as a function of time can be determined from the movement of the optical trap. In the equilibrium unfolding process, the molecule's end-end distance (extension) D is changed slowly, and the mean force averaged over the appropriate time period $\overline{F}$ as a function of D is plotted as the force-extension curve.

(b) Constant force ensemble (isotensional): the force is held constant, and the molecule's end-end distance fluctuates. In such experiments, the feedback loop makes the optical trap move in such a way that the bead position with respect to the center of the optical trap is fixed (which means the fixed force acting at the bead). The fluctuating end-end distance is then averaged and recorded as $\overline{D}(F)$. The equivalent force-extension curve $F(\overline{D})$ can be obtained by the inversion of the relation $\overline{D}(F)$.

It was shown [42] that the difference between results of isometric and isotensional experiments is essential when the system is small (short molecules). For long and flexible molecules the fluctuations of variables are negligible and the difference between the two statistical ensembles vanishes.

6.1.2 Previous thermodynamic theory for force-induced RNA hairpin folding.

The experiments have motivated the theoretical research on unfolding induced by force [43]-[46].

Gerland et al [46] theoretically studied the mechanical unfolding of RNA secondary structure. The experimental setup is sketched in Fig. 6.1a. The two ends of an RNA molecule are attached to the force-extension measuring device, for example, optical tweezers. The potential of optical tweezers is modeled by the harmonic potential of the linear spring, connected in series with the RNA molecule (Fig. 6.1b). In the experiment, the spring constant λ varies with the laser intensity. The intermediate values of λ amount to working in the mixed ensemble, whereas large and small spring constants correspond to isometric and isotensional experiments, respectively.

The extension of the spring R_s is measured with respect to the minimum of the trapping potential. The total extension R_t is hold constant, whereas the extensions of the linear spring and of the RNA molecule, R_s and R, undergo thermal fluctuations and are averaged over all accessible conformations of the spring and the RNA molecule at fixed R_t. The average force acting on the RNA molecule and its average extension are:

$$\begin{aligned} &< f >= \lambda < R_s > \\ &< R >= R_t - < R_s > \end{aligned} \tag{6.1}$$

In the following, it is assumed that the pulling occurs in the quasiequilibrium regime, i.e. slow with respect to the rate of conformational transitions.

For calculation of the average values, the partition function of the system should be

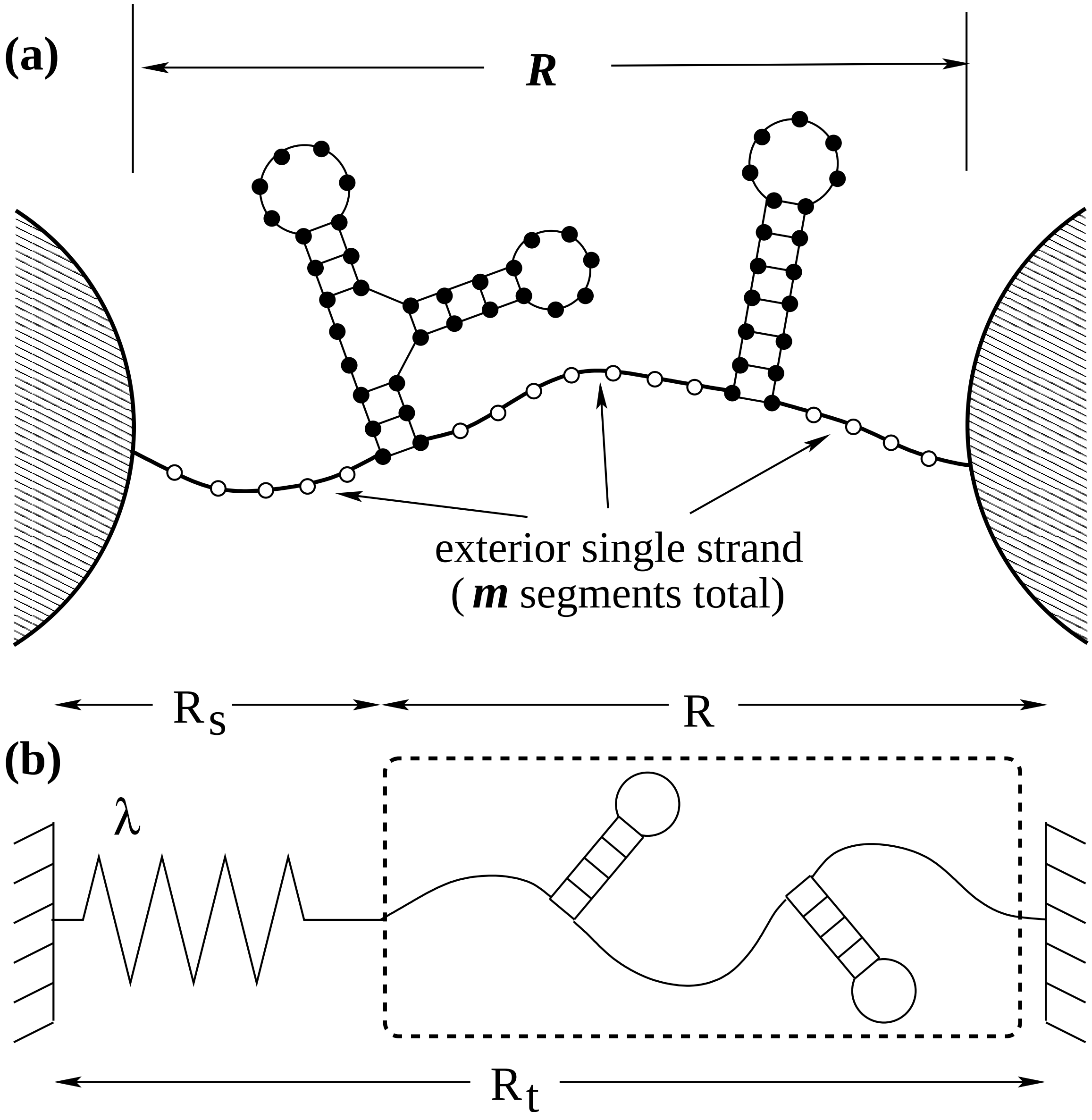


Figure 6.1: Sketch of the system. (a) Two ends of an RNA molecule are attached to the beads of optical tweezers. (b) The potential of the optical trap is modeled by the potential of a linear spring with the spring constant λ.

known. The conformation of the RNA molecule is divided into two parts: the secondary structure (filled circles in Fig. 6.1a) and the single-stranded (ss) segments exterior to the secondary structure (open circles in Fig. 6.1a). This two parts are coupled through the length of the single-stranded segments, i.e. the number m of the exterior bases (open circles in Fig. 6.1a). There is the interplay between the lowering the RNA total base pairing energy which requires the decreasing of m, and gaining the conformational entropy by increasing m. The total partition function of the RNA molecule is the convolution of the partition function of the secondary structure $Q(m)$, and the function $W_{tot}(R_t, m)$ which denotes the total end-end distance distribution of the single-stranded m-mer RNA in series with the spring:

$$Z(R_t) = \sum_m Q(m) W_{tot}(R_t, m). \tag{6.2}$$

The secondary structure partition function $Q(m)$ summed over all RNA secondary structures with m exterior single-stranded bases can be calculated through the recursion methods using the experimentally determined rules for the free energy of secondary structures [47].

The function $W_{tot}(R_t, m)$ can be expressed as:

$$W_{tot}(R_t; m) = \int_0^{R_t} dR W_{RNA}(R, m) \frac{e^{-\beta\lambda(R_t - R)^2)/2}}{\sqrt{2\pi/\beta\lambda}} \tag{6.3}$$

where $W_{RNA}(R, m)$ is the distribution of the ssRNA molecule alone, i.e. the probability that the chain with m exterior open bases has the end-end distance R. It is multiplied by the probability that the spring has the length $(R_t - R)$. $W_{RNA}(R, m)$ can be determined from an elastic freely jointed chain (EFJC) model [44], [46].

From the partition function $Z(R_t)$, the total free energy can be obtained: $G(R_t) = -k_B T \ln Z(R_t)$, and from the free energy - the average force $< f > (R_t) = \partial G(R_t)/\partial R_t$

and the average extension $< R >$ using Eqs. 6.1 can be calculated. The force-extension curve for the mixed statistical ensemble is $< f > (< R >)$. If the spring is very stiff, the fluctuations of R can be neglected and the statistical ensemble is constant distance; for the soft spring, the fluctuations of force are negligible, and the ensemble is constant force.

As a test of the theory, it was applied to the P5ab hairpin studied in the experiment of Liphardt et al [15]. In the experiment, the double-stranded DNA linkers were used to connect the RNA molecule with the beads to avoid surface interactions. The dsDNA part has been modeled as a wormlike chain (WLC), which means a continuous filament with a bending stiffness that exponentially decays over a distance along its contour. The persistence length l_p is defined as the contour length over which segment directions are correlated. As an extreme case, chain segments of length l_p behave as rigid rods. The WLC model gives the following relationship between the force f and the DNA extension [50]:

$$f(R_{DNA}) = \frac{k_B T}{l_p}(\frac{1}{4(1 - R_{DNA}/L)^2} + \frac{R_{DNA}}{L} - \frac{1}{4}) \tag{6.4}$$

where $l_p = 3.57$nm is the persistence length, and R_{DNA} estimates the total extension of DNA linkers. From the above equation, we obtain R_{DNA} for a given force f.

The experimental force-extension curve for the P5ab hairpin is in a good agreement with the theoretical one, with the spring constant $\lambda = 0.2\ pN/nm$ (Fig. 2a in [46]). The characteristic feature of the experimental FEC - a hump indicating the opening of the hairpin is also observed at the theoretical FEC, but the force at which the opening of the hairpin occurs, is overestimated by the theory. The corrections for the free energy for different ionic concentrations [51] result in a better agreement with the experiment.

The FEC for the nonbinding control sequence shown in Fig. 2b in [46] with dotted line, is calculated theoretically for the dsDNA linkers in series with the ssRNA molecule, using WLC and EFJC models. Comparison between FECs for the hairpin and for the control sequence gives information about the total binding free energy of the hairpin (which equals the area between two curves).

6.1.3 Our improved statistical thermodynamic model for force-induced RNA hairpin folding.

Recently, we developed a new statistical mechanical model for the force-induced equilibrium RNA hairpin folding [52]. The same experimental setup sketched in Fig. 6.1 has been used, but the constant force and constant extension ensembles have been considered separately. For the constant force ensemble the spring length R_s is a constant (since it is proportional to the force which is constant), therefore we can consider the RNA molecule separately, without taking the spring into account. In the case of the constant extension ensemble, the extension which is actually hold constant is R_t; it includes the extension of the spring, and therefore the explicit consideration of the spring is necessary in this case.

Constant force ensemble.

Under the assumption that the pulling process is quasistatic, the energy change of the system can be written as:

$$dU = TdS - PdV + FdR \tag{6.5}$$

Then, for the Gibbs free energy defined as

$$G(F) = U + PV - TS - FR, \tag{6.6}$$

the free energy change equals:

$$dG(F) = -S\,dT + VdP - RdF. \tag{6.7}$$

Therefore, at constant temperature and pressure, the change in free energy caused by the applied force f is

$$\Delta G(f) = -\int_0^f < R > (F)dF. \tag{6.8}$$

Similar to Eq. 6.2, the partition function for the RNA molecule with the applied constant force f is the weighted sum over all the possible m (= the length of the single-stranded chain segment, see the open circles in Fig. 6.1a):

$$Z(f) = \sum_m Q(m)W_{RNA}(f, m), \tag{6.9}$$

where $Q(m)$ is the secondary structure partition function, and $W_{RNA}(f, m)$ is the force distribution for the single-stranded molecule with m exterior bases.

To calculate $Q(m)$, the statistical model for secondary structures developed by Chen & Dill [16, 17] has been used, and the total free energy of all possible secondary structures with fixed m has been obtained from $Q(m)$: $\Delta G_0(m) = -k_B T \ln Q(m)$.

The weight function $W_{RNA}(f, m)$ can be expressed as $W_{RNA}(f, m) = e^{-\Delta G_{ss}(f,m)/k_B T}$, where $\Delta G_{ss}(f, m)$ is the change in free energy of the m-mer ssRNA due to the applied force f.

The ssRNA is described by the modified elastic freely jointed chain model [49] which yields the average extension per bond (end-end distance of the whole chain divided by m):

$$r_{ss}(f) = l(\coth(\frac{fb}{k_B T}) - \frac{k_B T}{fb})(1 + \frac{f}{S}), \tag{6.10}$$

where $l = 5.6$Å is the distance between subsequent nucleotides, $b = 15$Å is the Kuhn length (=2× persistence length) and $S = 800pN$ is the stretch modulus of ssRNA.

Now we apply Eq. 6.8 to the single-stranded part of RNA with the average extension $< R > (F) = m\, r_{ss}(F)$ and have

$$\Delta G_{ss}(f, m) = -\int_0^f m\, r_{ss}(F) dF.$$

The total free energy of the RNA with the given m is thus known: it is the sum of the free energies of the folded and the (m-mer) single-stranded parts of the molecule: $\Delta G(f, m) = \Delta G_0(m) + \Delta G_{ss}(f, m)$, and the total free energy of the RNA for all possible values of m can be obtained using Eq. 6.9:

$$\Delta G(f) = -k_B T \ln Z(f) = -k_B T \ln \sum_m e^{-\Delta G(f,m)/k_B T}$$

The mean end-end distance of the RNA for a fixed force f is the sum over possible m of the average extensions for the given m multiplied by the probability that at the given force the number of monomers of ssRNA equals m:

$$< R_{RNA} > (f) = \sum_m m\, r_{ss}(f)\, e^{-(\Delta G(f,m) - \Delta G(f))/k_B T} \qquad (6.11)$$

The extension of the dsDNA linkers connecting RNA with beads also contributes into R_t and should be accounted for. The wormlike chain approximation (Eq. 6.4) has been used to model the DNA linkers of the total length R_l. Finally, for the given force f, with the length of the DNA linker R_{DNA} given by Eq. 6.4, the force-extension curve can be obtained from

$$< R > (f) =< R_{RNA} > (f) + R_{DNA}(f). \qquad (6.12)$$

Constant distance ensemble.

In this case the molecule extension is held constant and the applied force is free to fluctuate. For the Gibbs free energy defined as

$$G(R) = U + PV - TS, \tag{6.13}$$

the free energy change is given by (see also Eq. 6.5)

$$dG(R) = -S\,dT + VdP + FdR. \tag{6.14}$$

At constant temperature and pressure, the free energy change of the molecule as a function of the end-end distance R is equal to the quasi-static work done on the molecule during the extension from 0 to R:

$$V(R) = \int_0^R < F > (x)dx, \tag{6.15}$$

where $V(R)$ is the potential of mean force and $< F > (R)$ is the mean force for the system at fixed R. For the constant distance ensemble, the explicit consideration of the spring is necessary [45]. The partition function for the system consisting of the RNA and the spring is given by Eqs. 6.2 & 6.3 with $Q(m)$ computed from the Chen & Dill model ([16], [17]) and $W_{RNA}(R, m)$ computed from the potential of mean force $V(R)$:

$$W_{RNA}(R, m) = e^{-\frac{V(R)}{k_B T}}. \tag{6.16}$$

Therefore, the procedure of calculating the force-extension curve $< f > (R_t)$ consists of the following steps:

1. The m-mer ssRNA and the DNA linkers are modeled as a freely jointed chain and a wormlike chain, respectively. So, the average force as a function of the end-end

distance of the RNA with DNA linkers $< f > (R)$ can be found from Eqs. 6.4, 6.10, and $R(< f >) = m\, r_{ss}(< f >) + R_{DNA}(< f >)$.

2. From $< f > (R)$, the potential of mean force $V(R)$ and end-end distance probability distribution of the molecule with DNA linkers $W_{RNA}(R, m)$ are determined using Eqs. 6.15 and 6.16, respectively.

3. Substitution of $W_{RNA}(R, m)$ into Eq. 6.3 yields the distribution of the system of RNA, DNA linkers and the spring in series $W_{tot}(R_t, m)$, partition function of the system $Z(R_t)$ (Eq. 6.2), free energy $\Delta G(R_t) = -k_B T \ln Z(R_t)$, and finally the force-extension curve $< f > (R_t) = \partial G(R_t)/\partial R_t$.

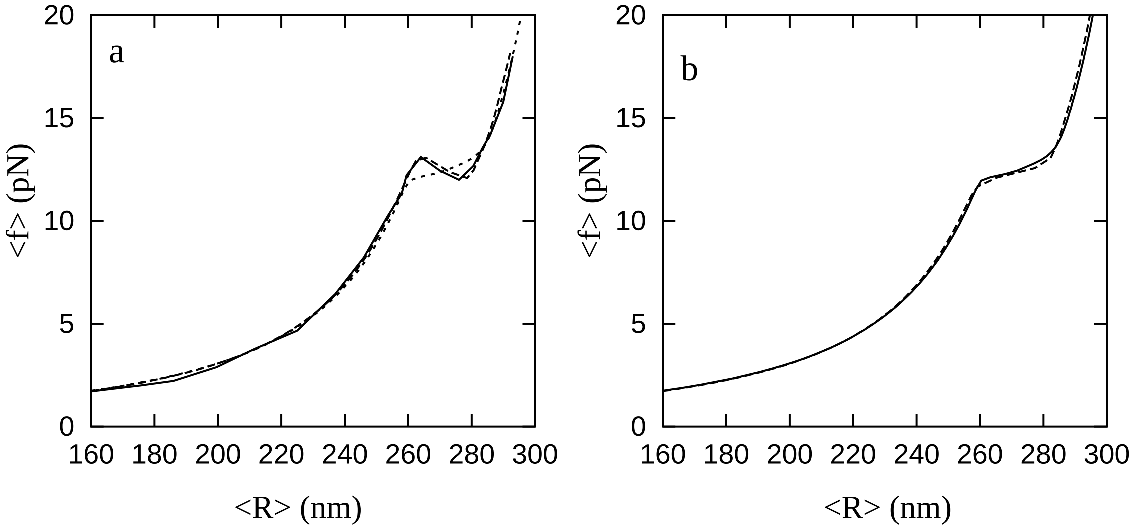


Figure 6.2: The FEC of the P5ab hairpin with added DNA linker: (a) experimentally obtained by Liphardt et al [15] (solid line), and theoretically obtained from our model for $\lambda = 0.2$pN/nm (dashed line) and $\lambda = 0.01$pN/nm (dotted line) at 0.25M NaCl. (b) theoretically obtained for constant force ensemble (solid line) and constant distance ensemble (dashed line) with $\lambda = 0.01$pN/nm.

The force-extension curves have been calculated for the P5ab hairpin (Fig. 6.2). The free energy parameters were adjusted to experimental salt condition ([Na$^+$]=0.25M) by using SantaLucia's corrections [51]. The theoretical curve with $\lambda = 0.2$ has been found to be in a great agreement with the experimental one (Fig. 6.2a), and the FEC for the constant distance ensemble with $\lambda = 0.01$ (soft spring) is very close to FEC for the constant force ensemble.

6.2 Force-induced RNA pseudoknots folding.

Our RNA pseudoknot folding thermodynamics theory developed in Section 5.4 can be further developed to study the thermodynamics of the mechanical unfolding of RNA pseudoknot. We assume that the two ends of the molecule are attached to two beads whose positions and the acting force are controlled by the force-extension measuring device. In our calculations, we consider the fixed force and fixed extension ensembles separately.

We assume that one end of the chain is fixed, and the constant force $\mathbf{f}$ is applied to the other end in the constant force experiments. The average end-end vector $\mathbf{D}$ of the molecule is recorded as a function of $\mathbf{f}$. The work $\mathbf{f} \cdot \mathbf{D}$ done on the molecule by force $\mathbf{f}$ contributes to the energy of the molecule, and the partition function for the constant force ensemble is:

$$Z(T, \mathbf{f}) = \sum_E \sum_{\mathbf{D}} g(E, \mathbf{D}) e^{-(E - \mathbf{f} \cdot \mathbf{D})/k_B T}, \tag{6.17}$$

where $g(E, \mathbf{D})$ is the constrained density of states, i.e. the number of conformations with the energy E and end-end vector $\mathbf{D}$. From $Z(T, \mathbf{f})$, the mean extension of the chain at the given force can be calculated as

$$\overline{\mathbf{D}}(T,\mathbf{f}) = k_B T \frac{\partial}{\partial \mathbf{f}} ln Z(T,\mathbf{f}). \tag{6.18}$$

In the constant extension experiments, the end-end vector **D** of the chain is assumed to be held constant, and the average force **f** acting on the molecule is recorded as a function of **D**. Since **D** is constant, the partition function for the constant extension ensemble is

$$Z(T,\mathbf{D}) = \sum_E g(E,\mathbf{D}) e^{-E/k_B T}. \tag{6.19}$$

If $Z(T,\mathbf{D})$ is known, the mean force $\bar{\mathbf{f}}$ corresponding to the given **D** can be calculated by the formula:

$$\bar{\mathbf{f}}(T,\mathbf{D}) = -k_B T \frac{\partial}{\partial \mathbf{D}}\ ln\ Z(T,\mathbf{D}). \tag{6.20}$$

6.2.1 Density of states

In both the constant force and constant extension models, the key problem is how to calculate the density of states $g(E,\mathbf{D})$, which is the total number of conformations with energy E and end-end vector **D**. We assume that one end of the molecule is attached to the wall, and the force **f** acting on the molecule is directed perpendicular to the wall. Instead of the end-end vector **D**, we consider it's component denoted as D, in the direction of the force. In the partition function calculation, the conformational ensemble includes both secondary structures and pseudoknotted structures. In what follows we give detailed description for the calculation for pseudoknotted and secondary structures, separately.

Pseudoknotted conformations.

To find the number of conformations of the pseudoknotted structure with the end of one tail attached to the wall, and the end of the other tail being at the distance D from the wall, we use the following approach. The structure can be divided into three parts: two tails and the structure element. For a given D, we exhaustively enumerate all the possibilities of three numbers: D_1 (= extension of tail 1), D_P (= end-end extension of the pseudoknot element), and D_2 (= extension of tail 2), such that $D_1 + D_P + D_2 = D$. For each set of (D_1, D_p, D_2), we calculate the product of the numbers of conformations of the first tail with extension D_1, of the pseudoknot element with extension D_P, and of the second tail with extension D_2.

To illustrate the principle, we choose the 2D lattice representation for chain conformations. On the two-dimensional lattice, a pseudoknot structure element (without tails) can have four different orientations with respect to the wall (Fig. 6.3). Since we consider the pseudoknot to be the central part of the structure, a convenient choice of the coordinate system (x, y) would be such that the y-axis is along the stem with the tail attached to the wall. This is the same coordinate system as the one used for the calculation of the number of pseudoknot conformations (Fig. 5.6). In such coordinate system, the wall and the force have different orientations with respect to the pseudoknot (Fig. 6.3).

For a fixed orientation of the pseudoknot, we need two different functions for the conformational count of tails: $\Omega_T^{\|}(t, d)$ and $\Omega_T^{\perp}(t, d)$, which denote the numbers of conformations for a tail of length t with end-end distance d in the (positive or negative) direction parallel ($\|$) and perpendicular ($\perp$) to the first (closest to the pseudoknot structure element) bond of the tail, respectively. We have obtained the $\Omega_T^{\|}$ and $\Omega_T^{\perp}$ functions by means of exact

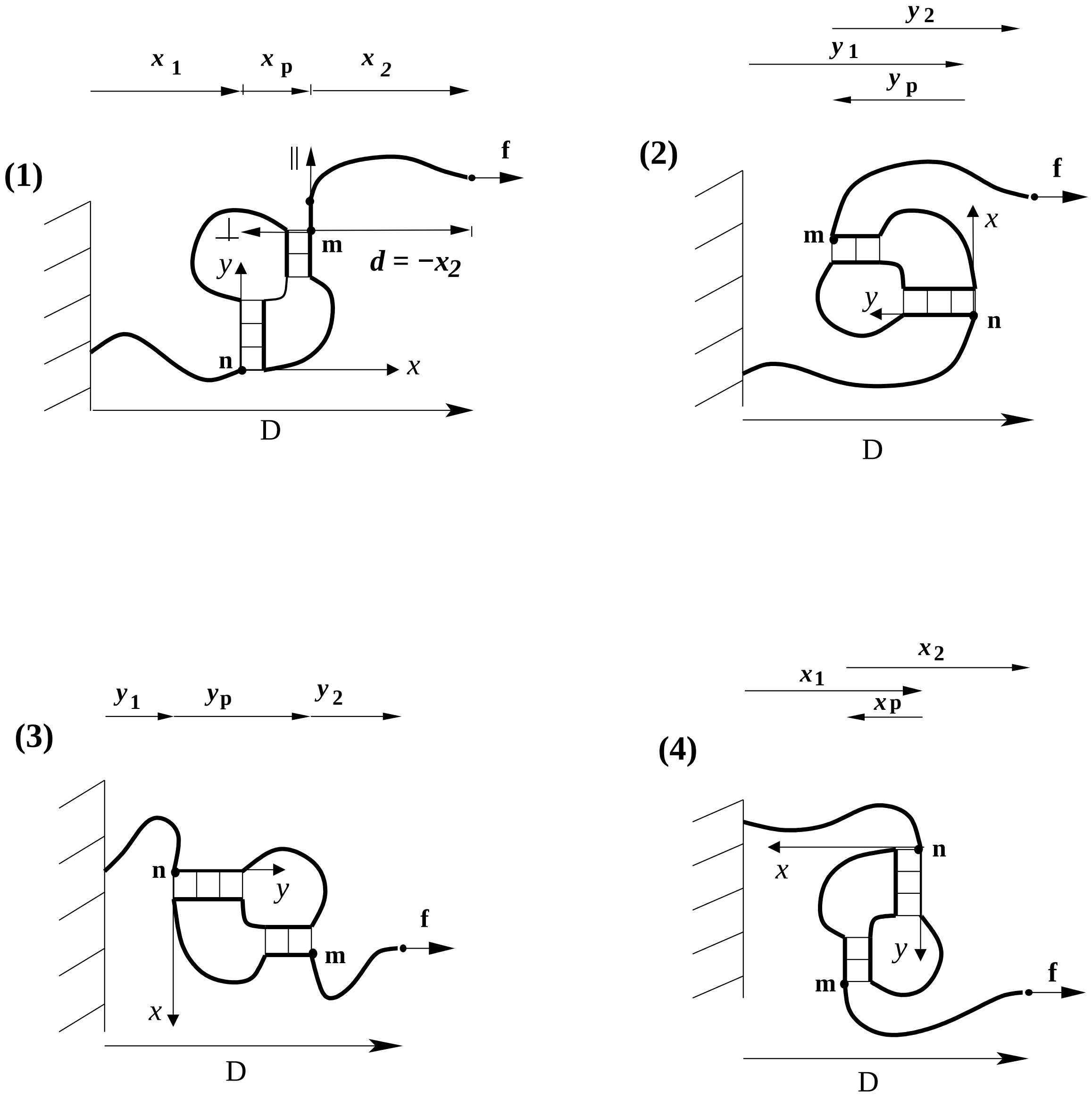


Figure 6.3: In the two-dimensional lattice model, a pseudoknot can have four possible orientations with respect to the wall. $x_p = x(m) - x(n)$ and $y_p = y(m) - y(n)$. The coordinate system can be defined by the position of the first stem of the pseudoknot (y-axis along the stem), while the || and ⊥ axes of the tail is defined by the orientation of the first bond as shown in (1).

computer enumeration for $1 \leqslant t \leqslant 11$.

For a given extension D along the line of the action of force, we follow the following procedure to calculate the the number of conformations of the given pseudoknot with the given end-end extension D.

1. We enumerate the different orientations of the pseudoknot, as shown in Fig. 6.3.

2. For each pseudoknot orientation, we enumerate the conformations of the pseudoknot enlarged interface. For a given enlarged interface conformation, the end-end vector (x_p and y_p in Fig. 6.3) and the extension D_P of the pseudoknot element are fixed.

3. For each enlarged interface conformation, using Eq. 4.1, we calculate the number of pseudoknot element conformations Ω_P.

4. We enumerate all the possibilities of D_1 and D_2 such that $D_1 + D_2 = D - D_P$. The numbers of tail 1 and tail 2 conformations which correspond to the given D_1 and D_2, respectively, are approximated by the function $\Omega_T^{\|}(t, d)$ or $\Omega_T^{\perp}(t, d)$. Which of two Ω_T functions should be used depends on the orientations of the wall and the force with respect to the pseudoknot (cases 1-4 in Fig. 6.3) and on the directions of the first bonds of the tails. For example, for the first bond of tail 2 directed upward in Fig. 6.3, case 1, the number of tail 2 conformations which have the end-end extension D_2 along the x-axis, equals $\Omega_T^{\perp}(t_2, d = -x_2)$, because tail 2 has the end-end extension $d = -D_2 = -x_2$ along (and in the negative direction of) the tail's (local) $\perp$-axis (see Fig. 6.3).

5. For each pair of (D_1, D_2), we calculate the product of the number of conformations

Ω_P of the pseudoknot with the given enlarged interface and the numbers of conformations of the tails $\Omega_T^z(t_1, d_1)$ and $\Omega_T^z(t_2, d_2)$, where t_1 and t_2 are the lengths of the two tails, $z = \|$ or $\perp$, and d_1 and d_2 are required extensions of tails in direction z.

6. The number of conformations of the given pseudoknotted structure with the given end-end extension D is given by the sum over *(i)* all the possible (D_1, D_2), *(ii)* the enlarged interface conformations **I**, and *(iii)* the pseudoknot orientations:

$$\Omega_{ps.str.}(D) = \sum_{\text{orientation}} \sum_{\mathbf{I}} \sum_{D_1} \sum_{D_2} \Omega_T(t_1, d_1) \cdot \Omega_P \cdot \Omega_T(t_2, d_2). \qquad (6.21)$$

Here the values of parameters d_1 and d_2 are determined by the orientation of the pseudoknot with respect to the wall and on the enlarged interface conformation **I**, which defines the direction of the first tail bonds with respect to the pseudoknot.

The density of states $g_P(E, D)$ for pseudoknotted structures with energy E and end-end distance D can be found then by summation of numbers of conformations with the given extension D over all pseudoknotted structures (graphs) with the given energy E:

$$g_P(E, D) = \sum_{\text{graphs with E}} \Omega_{ps.str.}(D).$$

For the number of conformations Ω_T for the single-stranded chain segments, we have exactly enumerated the conformations for different chain lengths $10 \leqslant l \leqslant 28$ and different extensions $D \leqslant l$. We find that we can fit the results by two different functions:

$$ln\, \Omega_T(l, D) = (-2.34782 + 0.803224\, l) + (-0.0239763 + 2.59832/l)\, D \qquad (6.22)$$

for $D \leqslant l/2$ (less stretched chain) and

$$ln\ \Omega_T(l, D) = \left[\frac{3.1649436}{l} - \frac{4.19464}{l^2}\right] D(l - D) \qquad (6.23)$$

for $l/2 \leqslant D \leqslant l$ (more stretched chain). In Fig. 6.4c, we show the comparison between the above approximated $\Omega_T(l, D)$ and that from exact computer enumeration and find good agreement. The above approximations for $ln\ \Omega_T(l, D)$ are used as extrapolations for longer open chains.

Secondary structures.

We first use hairpin conformations to illustrate the methodology. We will then generalize the theory to treat more complex secondary structures. A unique feature of a hairpin structure element is that it's end-end distance (between monomers a and b in Fig. 6.4) is fixed. So the conformational count for a given total extension D of the hairpin element with tails can be calculated as

$$g_2(E, D) \simeq \sum_{\text{hpins with E}} k\ \Omega_H(E) \cdot \Omega_T(l, D), \tag{6.24}$$

where $\Omega_H(E)$ is the number of conformations of the hairpin element (from a to b without tails) with energy E, $\Omega_T(l, D)$ is the number of conformations of the single-stranded chain segment as a function of length $l = t_1 + t_2 + 1$ and the extension D (see Fig. 6.4a), and the pre-factor k accounts for the volume exclusion between the tails and the hairpin structure element. We find that $k \simeq 1/4$ for $D \leqslant l - 4$ and $k \simeq 1$ otherwise.

For secondary structures, which consist of multiple sequentially connected secondary structure elements (e.g. hairpins), we calculate the density of states from the following recursive relation (see Fig. 6.4b):

$$g_2^{(n)}(E, D) = k\Omega_H^{(n)}(E_n) g_2^{(n-1)}(E - E_n, D), \tag{6.25}$$

where $E = \sum_{i=1}^{n} E_i$ is the total energy, E_i is the energy of the ith secondary structure element

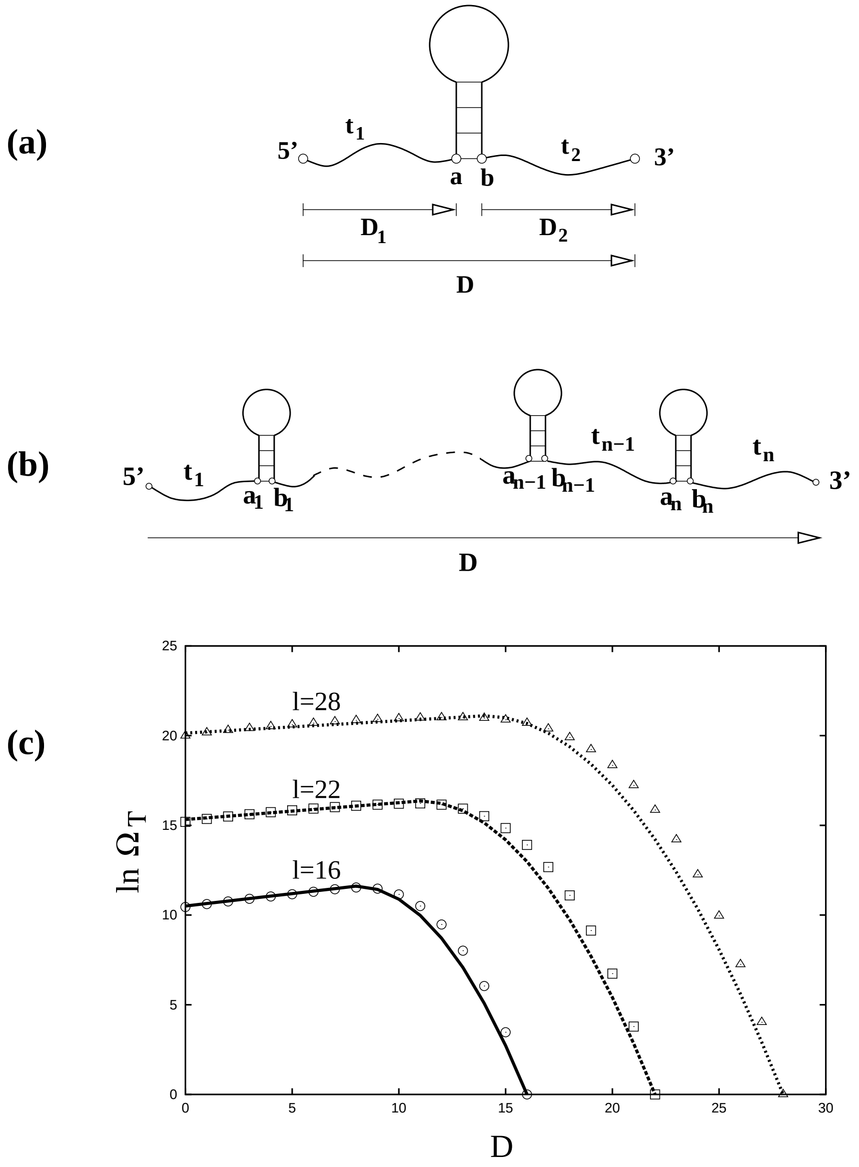


Figure 6.4: (a) A hairpin (with tails) can be divided into hairpin element and the single-stranded segment, which consists of two tails (t_1 and t_2) and the outermost contact (a, b) of the hairpin structure element. (b) A general secondary structure can be decomposed as multiple sequentially connected secondary structure elements. (c) The number of conformations of a single-stranded segment as a function of the segment’s length l and the end-end extension D from exact enumeration (symbols) and the analytical approximation (lines).

(between a_i and b_i), and $g_2^{(n-1)}$ is the density of states of a reduced chain, where the nth structure element (between a_n and b_n) is replaced by a single bond connecting a_n and b_n. $\Omega_H^{(n)}(E_n)$ is the number of conformations of the nth secondary structure element.

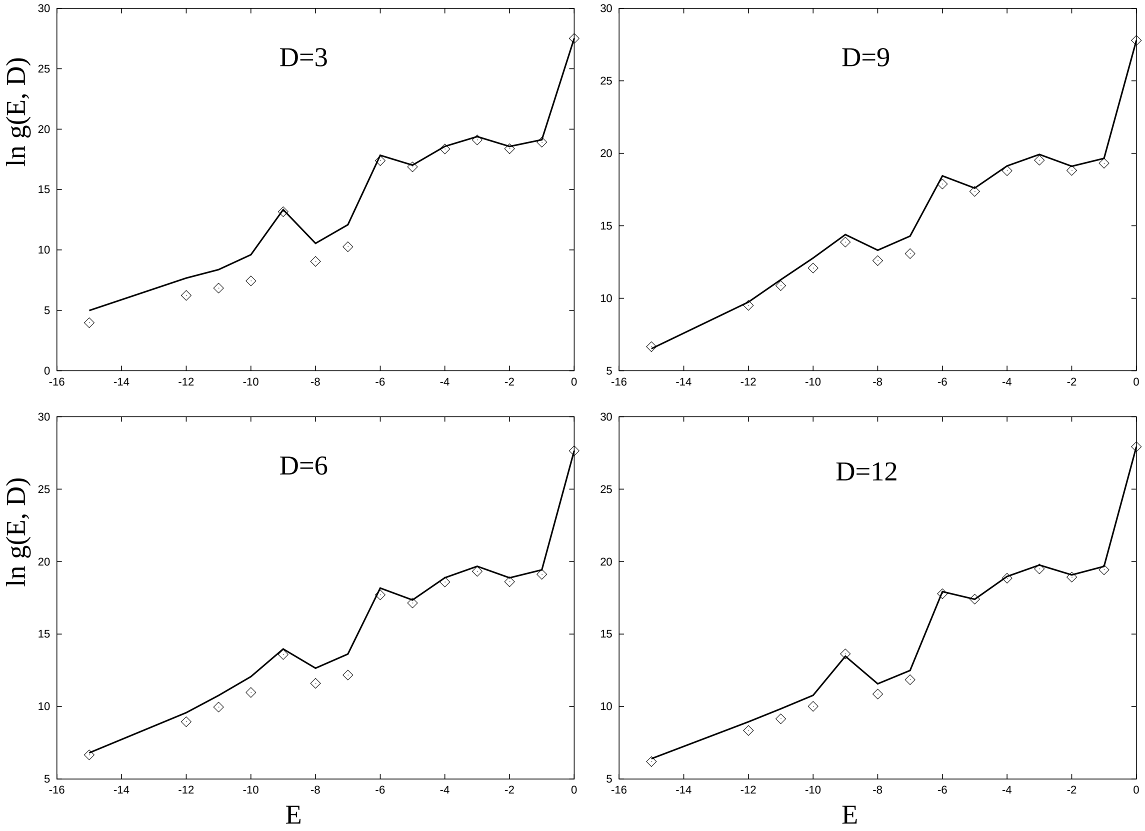


Figure 6.5: The density of states $g(E, D)$ for the pseudoknot/hairpin/open conformational ensemble for the 38-mer pseudoknot (shown in Fig. 5.9a with added tails of lengths $t_1 = 3$ and $t_2 = 4$). Symbols: from exact enumeration; lines: from the theory developed in this study.

The sum of the densities of states for secondary and pseudoknot structures gives the (total) density of states $g(E, \mathbf{D})$ in Eqs. 6.17 & 6.18. In Fig. 6.5, we show the tests for our theory against exact computer enumeration for a 38-mer pseudoknot-forming chain. We find good agreements.

6.2.2 Force-extension curve, misfolded pseudoknots, and folding thermodynamics

We consider two specific pseudoknot-forming nucleotide sequences for which thermal unfolding processes have been studied (Figs. 5.12 & 5.13). In the calculation for the density of states, we enumerate all the possible secondary and pseudoknotted states, including all the possible misfolded states and partially folded states. We show the predicted force-extension curves and the conformational changes for the two sequences in Fig. 6.6.

We find that for the two sequences, both the isometric and isotensional curves show a major transition from the native pseudoknot (N) to a misfolded intermediate state (I_1). A notable feature shown in Fig. 6.6 is that the misfolded intermediate (I_1) emerges when the extension of the molecule is both small and large compared to the average extension of the native pseudoknot N. The formation and the re-formation of the intermediate state I_1 (with both small and large extensions) can be explained in the following way.

The native pseudoknots have shorter tails than the intermediate states. Therefore, the end-end extension of the native pseudoknot is much more restricted than that of the intermediate state I_1. For instance, for sequence 1 in Fig. 6.6A, the end-end extension for the native pseudoknot N and the intermediate I_1 in a two-dimensional lattice are in the range of $[3, 10]$ and $[1, 13]$, respectively. Therefore, for sequence 1, for extension outside the range $[3, 10]$, the native pseudoknot cannot exist and new structure would emerge. In contrast, the intermediate state I_1 can accommodate a wide range of extension and can exist as a stable state outside the range $[3, 10]$.

As shown in Fig. 6.4c, the number of tail conformations quickly decreases as the tail

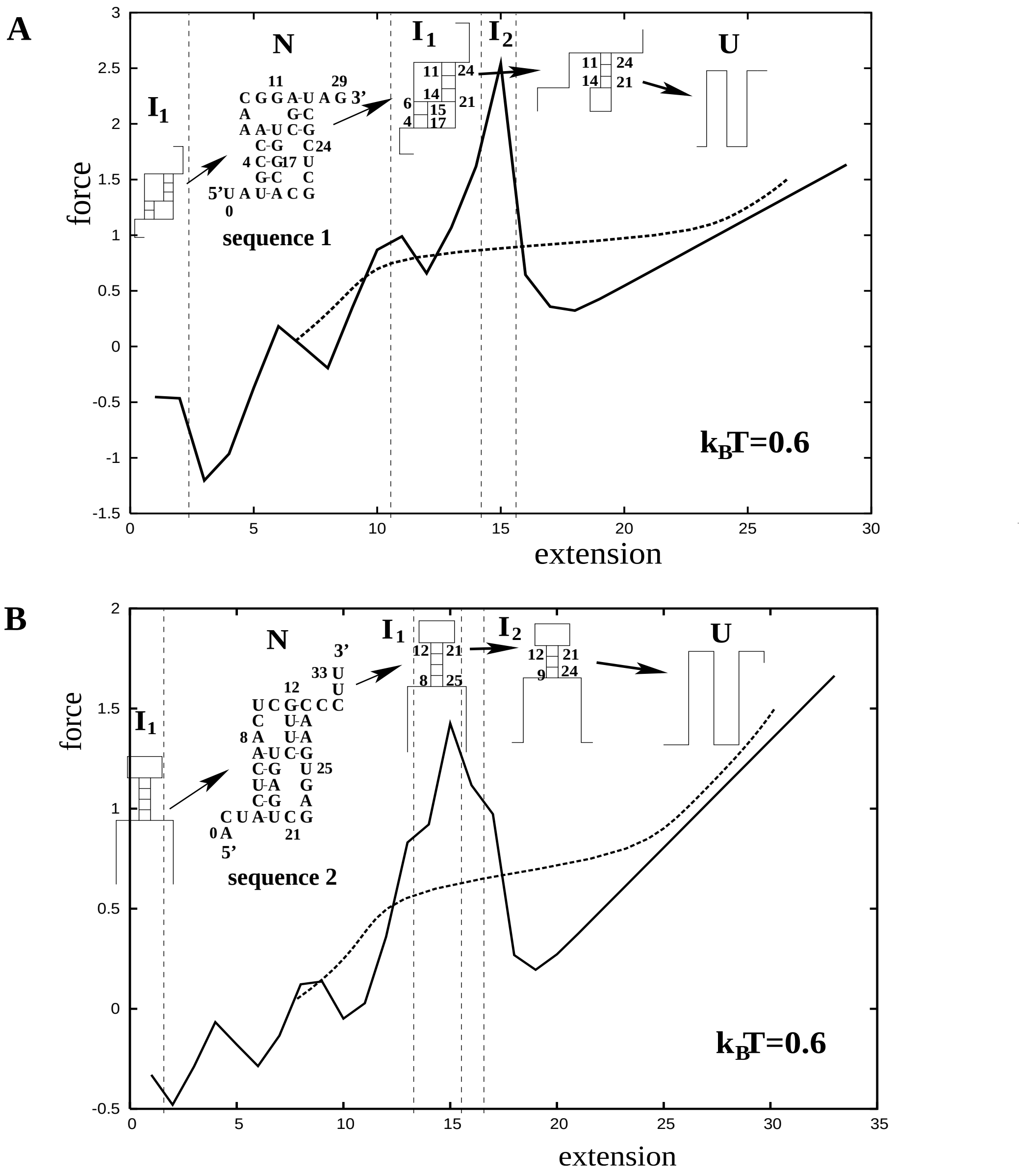


Figure 6.6: The calculated isometric (solid lines) and the isotensional (dashed lines) force-extension curves and the structural transitions for the two nucleotide sequences specified. For both sequences, the misfolded structures (I_1) emerges as intermediate states for those extensions of the molecules at which the overstretched native pseudoknots are less stable.

is stretched. When the native pseudoknot, which has short tails, is restricted to have very small or large extensions, the tails would inevitably be stretched. The entropy of the tails is small and the free energy $F = E - TS$ of the structure is high. Therefore, the native pseudoknot is unstable and a structural transition would occur. For the misfolded structures (I_1), however, the longer tails allow the chains to achieve the required (small or large) extensions without becoming highly stretched. So the intermediate state can be more stable than the native pseudoknots for small and large extensions.

In the limiting regime of very small end-end extension, the chain is in highly compact (and thus low-entropy) states. Large extension would cause an increase in the freedom of such highly compressed chain. So in the small extension limit, larger extension would lead to lower free energy, resulting in an apparent negative equilibrium pulling force.

Another notable feature in Fig. 6.6 is that sequence 1 mechanically unfolds through very different steps of conformational transitions than thermal unfolding. The thermal unfolding involves simple disruptions of native contacts, while in the mechanical unfolding, the native pseudoknot unfolds, refolds, and unfolds again during the pulling process. In contrast, for sequence 2, the intermediate state is the same misfolded hairpin as that in the thermal unfolding.

Chapter 7

CONCLUSION AND DISCUSSION.

We present a statistical mechanical theory for the folding thermodynamics of pseudoknotted chain conformations, including all the possible partially folded and misfolded structures. The model enables the calculation of conformational entropy and the partition function of pseudoknots. The key idea of the theory is *(i)* to decompose a pseudoknot structure into stem-loop subunits, *(ii)* to separate out the interfacial segments between the subunits, and *(iii)* to account for the correlation and volume exclusion between subunits through localized effects within and near the interface. The theory has been shown to give good predictions for the folding thermodynamics as tested against exact computer enumerations. The theory enables predictions for the folding stability, native-like and misfolded folding intermediates, and folding free energy landscapes for simple tertiary, pseudoknotted and secondary structures.

The current form of the model is not without limitations. Possible further development of the model should address the following issues.

1. More realistic off-lattice chain representations can be used within the present graph-theoretic framework. An example of possible off-lattice chain representation is the virtual bond model [53], where the six bonds of each nucleotide backbone (Fig. 1.1b) are replaced by two virtual bonds. The length of the virtual bond is nearly fixed (3.9Å), and the conformation of each nucleotide is described by two torsional and two bond angles of the virtual bonds. The bond angles of the virtual bonds vary between 90° and 120° in the single-stranded chain region. Based on this observation, Cao and Chen [54] employed the diamond lattice to describe the secondary structure chain conformations, which has the bond angle 109°.

2. Non-canonical intra-loop interactions and other tertiary interactions can play important roles in the sequence- and temperature-dependence of the loop entropy and the folding thermodynamics and therefore should be taken into account in the further development of the model. The generality of the basic ideas in the current model suggests the possibility to systematically extend the model to treat more complex tertiary folds, including ones with multiple crossing-linked contacts[8].

3. The helical stems in pseudoknots tend to form the energetically favorable coaxial stacks. Such coaxial stacking is neglected in the present model and should be included in the further development of the model.

4. RNA folding is strongly dependent on the ionic solution condition. The divalent metal ions (Mg^{2+}) play critical roles in coordinating and stabilizing tertiary interactions. To properly take into account this effect, the chain conformational model should be combined with the polyelectrolyte theory to account for the ion electro-

statics.

Thermodynamic parameters (especially the conformational entropy) for tertiary structures are currently very limited, mainly due to the lack of a rigorous statistical mechanical model for tertiary folding. The present model provides a general method for the computation of the conformational entropy for tertiary structures. Moreover, the statistical mechanical framework developed here can also be used to extract the thermodynamic parameters from the experiments.

The generality of the basic idea suggests the possibility to extend the method to treat more complex tertiary folds, including ones with multiple crossing-linked contacts. Before further extending the theory to treat more complex tertiary folding, we would first apply the theory to relatively simple tertiary folds and calibrate and validate the theory through comparisons with the existing experimental data, including the thermodynamic experiments on the force-induced folding-unfolding of single RNA molecule.

BIBLIOGRAPHY

[1] D.H. Turner, N. Sugimoto, S.M. Freier. (1988). RNA structure prediction. *Ann. Rev. Biophys. Chem.* **17**: 167-192.

[2] S.M. Freier et al (1986). Improved free energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* **83**: 9373-9377.

[3] J.A. Jaeger, D.H. Turner & M. Zuker (1989). Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* **86**: 7706-7710.

[4] D.E. Draper, T.C. Gluick, and P.J. Schlax. (1998). Pseudoknots, RNA folding, and translational regulation. In *RNA structure and function* (ed. R.W. Simons and M. Grunberg-Manago), pp 415-436. Cold Spring Harbor Laboratory Press, New York.

[5] M.H. de Smit (1998). Translational control by mRNA structure in eubacteria: molecular biology and physical chemistry. In *RNA structure and function* (ed. R.W. Simons and M. Grunberg-Manago), pp 495-540. Cold Spring Harbor Laboratory Press, New York.

[6] M.H. de Smit and J. van Duin. (1990). Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis. *Proc. Natl. Acad. Sci. USA* **87**: 7668-7672

[7] M.N. Hall et al. (1982). A role for mRNA secondary structure in the control of translation initiation. *Nature* **295**: 616-618

[8] G. Buell et al. (1985). Optimizing the expression in *E. coli* of a synthetic gene encoding somatomedin-C (IGF-I). *Nucleic Acids Res.* **13**: 1923-1938

[9] V.P. Schulz and W.S. Reznikoff (1991). Translation initiation of IS50R read-through transcripts. *J. Mol. Biol.* **221**: 65-80

[10] F.G. Wulczyn and R.Kahmann (1991). Translational simulation: RNA sequence and structure requirements for binding of Com protein. *Cell* **65**: 259-269

[11] D.E. Draper (1993). Mechanisms of translational initiation and repression in prokariotes. In *The translational apparatus* (ed. K.H. Nierhaus et al), pp 197-207. Plenum Press, New York.

[12] M.H. de Smit (1994). "Regulation of translation by mRNA structure." Ph.D. thesis, Leiden University, Leiden, The Netherlands.

[13] A.P. Abhijit and J.A. Steitz. (2003). Splicing double: insights from the second spliceosome. *Nature reviews* **4**: 960-970.

[14] R. Das, L.W. Kwok et al. (2003). The Fastest Global Events in RNA Folding: Electrostatic Relaxation and Tertiary Collapse of the Tetrahymena Ribozyme. *J. Mol. Biol.* **332**: 311-319.

[15] J. Liphardt, B. Onoa, S.B.Smith, I.Tinoco, and C.Bustamante. (2001). Reversible unfolding of single RNA molecules by mechanical force. *Science*. **292**: 733-737.

[16] S.-J. Chen and K.A. Dill (1995). Statistical thermodynamics of double-stranded polymer molecules. *J. Chem. Phys.* **103**: 5802-5813.

[17] S.-J. Chen and K.A. Dill (1998). Theory for conformational changes of double-stranded chain molecules. *J. Chem. Phys.* **114**: 4602-4616.

[18] S.-J. Chen and K.A. Dill (2000). RNA folding energy landscapes. *Proc. Natl. Acad. Sci. USA* **97**: 646-651.

[19] W. Zhang and S.-J. Chen (2001). A three-dimensional statistical mechanical model of folding double-stranded chain molecules. *J. Chem. Phys.* **114**: 7669-7681.

[20] L.G. Laing, T.C. Gluick & D.E. Draper (1994). Stabilization of RNA structure by Mg ions, specific and nonspecific effects. *J. Mol. Biol.* **237**: 577-587.

[21] J.S. McCaskill (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105-1119.

[22] R. Mans, M.H.V. Steeg, P. Verlaan, C. Pleij, & L. Bosch (1992). Mutational analysis of the pseudoknot in the tRNA-like structure of Turnip Yellow Mosaic Virus RNA. Aminoacylation efficiency and RNA pseudoknot stability. *J. Mol. Biol.* **223**: 221-232.

[23] K.A. Theimer, Y. Wang, D.W. Hoffman, H.M. Krisch, & D.P. Giedroc (1998). Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *J. Mol. Biol.* **279**: 545-564.

[24] E. ten Dam, K. Pleij, D. Draper. (1992). Structural and functional aspects of RNA pseudoknots. *Biochemistry* **31**: 11665-11676.

[25] P. Fechter, J. Rudinger-Thirion, C. Florentz, & R.Giege. (2001). Novel features in the tRNA-like world of plant viral RNAs. *CMLS, Cell. Mol. Life Sci.* **58**: 1547-1561.

[26] T.W. Dreher (1999). Functions of 3'-untranslated regions of positive strand RNA viral genomes. *Annu. Rev. Phytopathol.* **37**: 151-174.

[27] R. Mans, C. Pleij, & L. Bosh. (1991). Transfer RNA-like structures: structure, function and evolutionarz significance. *Eur J Biochem*, **201**: 303-324.

[28] D.R. Gallie & V. Walbot. (1990). RNA pseudoknot domain of tobacco mosaic virus can functionally substitute for a poly(A) tail in plant and animal cells. *Genes Dev.* **4**: 1149-1157.

[29] D.P. Giedroc, C.A. Theimer, and P.L. Nixon (2000). Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Mol. Biol.* **298**: 167-185.

[30] C.W.A. Pleij. (1990). Pseudoknots a new motiv in the RNA game. *Trends Biochem Sci* **15**: 143-147.

[31] R.T. Batey, R.P. Rambo, and J.A. Doudna. (1999). Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed.* **38**: 2326-2343.

[32] T.C. Gluick & D.E. Draper. (1994). Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**: 246-262.

[33] C.A. Theimer & D.P. Giedroc. (1999). Equilibrium unfolding pathway of an H-type RNA pseudoknot which promotes programmed -1 ribosomal frameshifting. *J. Mol. Biol.* **289**: 1283-1299.

[34] P.L. Nixon & D.P. Giedroc. (2000). Energetics of a strongly pH dependent RNA tertiary structure in a frameshifting pseudoknot. *J. Mol. Biol.* **296**: 659-671.

[35] A. P. Gultyaev, F. H. D. van Batenburg, C. W. A. Pleij (1999). An approximation of loop free energy values of RNA H-pseudoknots. *RNA* **5**: 609-617.

[36] J.Jr. SantaLucia & D. H. Turner (1998). Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* **44**: 309-319.

[37] H. Jacobson and W. H. Stockmayer (1950). Intramolecular reaction in polycondensations. I.The theory of linear systems. *J. Chem. Phys.* **18**: 1600-1606.

[38] M. E. Fisher (1966). Effect of excluded volume on phase transitions in biopolymers. *J. Chem. Phys.* **45**: 1469-1473.

[39] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij (1990). Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* **18**: 3035-3044.

[40] A. Lucas and K. A. Dill, (2003). Statistical mechanics of pseudoknot polymers. *J. Chem. Phys.* **119**: 2414-2421.

[41] N. Madras and G. Slade, *The Self-Avoiding Walk*, Birkhauser, Boston, 1993.

[42] D. Keller, D. Swigon, & C. Bustamante (2003). Relating single-molecule measurements to thermodynamics. *Biphys. J.* **84**: 733-738.

[43] S. Cocco, R. Monasson & J.F. Marko (2001). Force and kinetic barriers to unzipping of the DNA double helix. *Proc. Natl. Acad. Sci. USA* **98**: 8608-8613.

[44] A. Montanari & M. Mezard (2001). Hairpin formation and elongation of biomolecules. *Phys. Rev. Lett.* **86**: 2178-2181.

[45] U. Gerland, R. Bundschuh & T. Hwa (2001). Force-induced denaturation of RNA. *Biphys. J.* **81**: 1324-1332.

[46] U. Gerland, R. Bundschuh & T. Hwa (2003). Mechanically probing the folding pathway of single RNA molecules. *Biphys. J.* **84**: 2831-2840.

[47] A.E. Walter et al (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA.* **91**: 9218-9222.

[48] P.J. Flory *Statistical mechanics of chain molecules.* Interscience Publishers, New York, 1967

[49] S. B. Smith, L. Finzi, C. Bustamante (1992). Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**: 1122-1126.

[50] C. Bustamante, J. F. Marco, E. D. Siggia, and S. B. Smith (1994). Entropic elasticity of lambda-phage DNA. *Science* **265**: 1599.

[51] J.Jr. SantaLucia (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95**: 1460-1465.

[52] W. Zhang, Z. Kopeikin, and S.-J. Chen (2005). Mechanical folding kinetics of RNA hairpins, to be submitted

[53] W.K. Olson (1980). Configurational statistics of polynucleotide chains. An updated virtual bond model to treat effects of base stacking. *Macromolecules* **13**: 721-728.

[54] S. Cao and S.-J. Chen (2005). Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* **11**: 1884-97.

VITA

Zoia Kopeikin was born August 28, 1960, in Moscow, Russia. She earned M.S. Degree in Physics at the Moscow State University in 1983. She was working in the Moscow State University until 1993. In 2001 she went on to earn the Doctorate in Biological Physics from the University of Missouri-Columbia in 2006. She is married to Sergei Kopeikin.