Appendices

A. Derivation of the Jacobson Stockmayer model

Even when a polymer is *unstructured*, it still possesses long range features that resemble a random walk. In a random walk, one observes that the distance between the initial starting point of the walker and the last step tend to gradually drift away as a function of the number of interim steps. For a pure random walk, the drift would be seen to occur at a rate $cN^{1/2}$, where c is the distance between each step and N is the number of steps. It is actually rather difficult to walk in a truly random fashion, but the image should be easy to visualize.

Applied to polymers, the step length becomes the distance between monomers and the space is usually 3D rather than 2D. As the length of the polymer increases, the likelihood that the two ends will come within the same vicinity of each other gradually decreases. In principle, this property is true whether the sequence length is 20 nt or 20 sextillion nt. Though perhaps somewhat counterintuitive, it is as though the polymer appears to *know* its ends. In fact, it is because it doesn't care, yet ironically, that means we can predict its behavior. Therefore, the negative entropy for loop formation in Equation (1) reflects the force that is required to constrain the two ends of the polymer to a fixed-distance arrangement. The entropy will become more negative as more diverse parts of the chain are held in close proximity of each other. The Jacobson Stockmayer (JS) equation was an attempt to model this property of polymers.

To derive the JS-model, JS began with an approximation of the freely jointed polymer chain (FJPC) using a Gaussian distribution function where the probability of finding the two ends of the polymer chain within a distance r and $r + \Delta r$ is

$$p(r)\Delta r = \left(\beta^2 / \pi\right)^{3/2} \exp\left(-\beta^2 r^2\right) 4\pi r^2 \Delta r \qquad \text{(A1)}$$

where $\beta = (3/2\xi N b^2)^{1/2}$, N is the number of mers, and ξ is the Kuhn length.

The Kuhn length is a measure of the stiffness of the polymer and will be discussed in detail in Part II of this series. The original derivation of Equation (A1) does not explicitly include the Kuhn length, rather, in the main text of their study, JS implicitly allude to the point that b *is the effective link length of the polymer chain* (p. 1602, col. 1, and Part II of same journal, pp.1611-2). By saying so, JS implicitly recalibrate the mer-to-mer separation distance to $b' = \xi b$ and the number of mers (N) to $N' = N / \xi$. Hence, $N'(b')^2 = \xi N b^2$.

Now, when the two ends of the chain are closed at position 1 and N, the probability that the two ends of the polymer chain will be simultaneously located within the same volume element (v_s) is

$$p(v_s) = \int_{v_s} p(r)dr$$
$$= \left(\beta^2/\pi\right)^{3/2} \int_{v_s} \exp\left(-\beta^2 r^2\right) 4\pi r^2 dr$$

and for a small volume v_s (called the *bond volume*, p. 1606 col 2), this can be approximated as

$$= \left(\beta^2/\pi\right)^{3/2} \int_{v_s} \left\{1 - (\beta^2 r^2) + \frac{1}{2}(\beta^2 r^2)^2 \ldots\right\} 4\pi r^2 dr$$
$$\approx \left(\beta^2/\pi\right)^{3/2} v_s \qquad \text{(A2)}$$

For this temperature independent probability model, the cyclization entropy due to ring closure is simply

$$\Delta S = k_B \ln(p), \qquad \text{(A3)}$$

where k_B is the Boltzmann constant.

Therefore

$$\begin{aligned}\Delta S &= k_{\mathrm{B}} \ln\left[\left(\beta^2/\pi\right)^{3/2} v_s\right] \\ &= k_{\mathrm{B}}\left\{\ln\left(v_s/b^3\right) - \frac{3}{2}\ln\left(\frac{2\pi\xi}{3}\right) - \frac{3}{2}\ln\left(N\right)\right\}.\end{aligned} \tag{A4}$$

In the formulation, it is assumed that $N >> 1$ and moreover, that $N >> \xi$. Note that the assumed product of this cyclization entropy is a ring of n mers with an effective symmetry of $\mathbf{C}_n$. In short, once the ring of monomers is closed, there is no way, even in principle, to find the joining point short of selective isotope labeling.

When Equation (A4) was transferred to problems in double-stranded DNA, it was applied to small *bubbles* (unpaired bases forming an interior loop along the double-stranded helix) that occurred in mismatched regions of sequences. The *bubbles* tend to involve unpairing and therefore, in Figure 1C, n_1 and n_2 would tend to be equal: $n_L = n_1 + n_2 + d$, where d is an adjustment for nucleation, etc. These *bubbles* were thought to be reasonably large, though that is not always exactly clear [90]. Given n_L is sufficiently large, it is then a simple task to assume a shorter segment of length n_L, where $N \rightarrow n_L$, which yields

$$\Delta S = -\left\{A_{JS} + (3k_{\mathrm{B}}/2)\ln(n_L)\right\} \tag{A5}$$

where

$$A_{JS} = k_{\mathrm{B}}[(3/2)\ln\left(2\pi\xi/3\right) - \ln(v_s/b^3)]. \tag{A6}$$

It is not clear to what extent this transformation $N \rightarrow n_L$ is allowed; however, JS comment that $N > 15$ is acceptable. For folded single-stranded RNA, the same approximation began to be used. Equation (A5) is identical to Equation (1) and is essentially how the original expression entered into these

types of calculations. For particularly small values of n_L, Eqn (A5) is usually substituted with experimentally obtained weights [41] to improve the accuracy of this expression. In this respect, the issues of Equation (A5) for small n_L can somewhat be taken as moot.

Clearly, the original model set $\gamma = 3/2$. The current value of $\gamma = 1.75$ entered the equation from a study done in Ref [21] that used this value based upon a theoretical calculation by Fisher [39]. Fisher determined this value by calculating a self avoiding random walk on a 3D lattice. In Ref [21], the value $\gamma = 3/2$ was simply replaced with $\gamma = 1.75$, and because it lead to some improvements, it has been retained ever since. The non-integer dimension is necessary because a real polymer cannot occupy the same point in space with another part of its own self but a statistical model like the Gaussian polymer chain (GPC) ignores this fact. As a result, one (very improbable) state for the GPC is a polymer chain that folds back and forth on its own self. In effect, one might say that the *statistical spatial dimensions* of a real polymer are larger than its integer spatial dimensions. This phenomenon is known as the self avoiding random walk. It will be shown in Part II of this Series (in the fifth section) that we can infer some genuine physical significance out of this fractal value.

In RNA secondary structure prediction, a further correction has been to use experimentally obtained values for the free energy when evaluating loop sizes of $n_L \leq 9$ nt. Therefore, small values of n_L are in principle corrected for and, since in loop regions $\xi \sim 3$ nt, the rule of $n_L >> \xi$ is, to some extent, sufficiently satisfied to within experimental error. Granted, it is far from an ideal treatment, but the errors do not appear to affect these problems severely, or, at least, the error bars are too large to identify a definite problem.

It will be shown in Part II (in the sixth section), that the value used for A_{JS} in current RNA structure calculations cannot satisfy Equation (A6) for any existing set of experimental parameters attributable to known RNA. Hence, the value used in practice for A_{JS} is actually an empirical constant.

B. Distinction between RNA and protein secondary structure

Because the meaning of *secondary structure* is so easily confused or muddled when discussing RNA secondary structure and protein secondary structure in the same sentence, it is important to explain the difference.

Protein secondary structure refers to the tendency of a protein to form regular conformations in the form of alpha helices (α-helix), beta-strands (β-strand), and a variety of turns like beta turns. Beta strands are also known as extended structures, particularly in bioinformatics software that predict or evaluate protein secondary structure. The secondary structure is contrasted with coil regions where the arrangement of the amino acids does not fit one of these regular patterns sufficiently. Protein secondary structure makes no reference to the relative spatial arrangement of these regular structural elements of the polypeptide chains with respect to each other. Therefore, there is no topology information and no indication of how the α-helix or β-strands of the protein are arranged spatially. This topological information is encoded in the *tertiary structure* for proteins. Although there are examples of long range influence on the protein secondary structure, at least 80% of the protein secondary structure can be predicted based upon neighboring amino acid or next nearest neighbor interactions. It is, therefore, a largely nearest neighbor (i.e., local) feature of amino acids.

RNA secondary structure refers to the arrangement of base pairing in the RNA structure. The base pairing defines a relative spatial arrangement between different nucleic acids in the RNA sequence. Therefore, RNA secondary structure does provide essential information on the topology and the spatial arrangement of the RNA sequence.

The definition of RNA secondary structure has special restrictions of its own in that the base pairing arrangement is usually restricted to a narrow subset of possible base pairing patterns. The rule is as follows. Let a given RNA sequence be numbered from 1 to N, where N is the total number of nucleotide (nt) in the sequence, and let a given set of distinct base pairs (i, j) and (i', j') be defined such that the indices satisfy the following conditions (1) $i \neq j \neq i' \neq j'$, (2) $i < j$, and (3) $i' < j'$. Then an RNA structure is called RNA

secondary structure if every base pair combination (i, j) and (i', j') in the set satisfies one of the following four conditions: (1) $i < i' < j' < j$, (2) $i', j' < i$, (3) $j < i', j'$ or (4) $i' < i < j < j'$. Hence, the arrangements of base pairs in RNA pseudoknot structures, which involve base pairing of the form $i < i' < j < j'$ or $i' < i < j' < j$, are not considered in the standard definition of RNA secondary structure.

Proteins can form parallel strand arrangements of β-strands or α-helices where two or three strands can be in close proximity. If two strands run in parallel, then the amino acids in the respective chains can form double strand spatial pairing (i, j) interactions. If three strands run in parallel in close proximity, then the amino acids can form triple strand spatial pairing (i, j, k) interactions, where the notation indicates that the amino acid index j shares a direct neighbor with both i and k. RNA occasionally forms tertiary structures containing triple strand interactions and triple helices can be made from RNA; however, pseudoknots are by far the most common form of tertiary structure interaction. Nevertheless, in all such cases, these interactions can be referenced by including a separate link to both pairs: (i, j) and (j, k). It follows that in as much as the notation is adequate for describing RNA interactions, the strand notation used for RNA structures can be applied equally to proteins. Likewise, the entropy of folding β-strands and α-helices into some specific spatial arrangement in a protein structure can be evaluated using similar methodology as is done for RNA pseudoknots.

C. On defining what is the *denatured state*

This Appendix assumes that the reader has read most of this work, or is familiar with the topics addressed here.

Because of traditions and conceptual issues as well, it is not as easy as it might seems to define this "denatured state" mathematically. Basically, there appear to be two possible ways to define this state

- case 1: the point where $f_{\text{int}}(R_s) = 0$.
- case 2: in terms of r_{rms}: $r_{\text{rms}}^2 = \xi N b^2$.

Case 1: The point $f_{\text{int}}(R_s) = 0$ represents a lone stationary point on the force-extension curves (see Figures 5D and 5E) where the force vanishes and is a uniquely defined location on the force-extension curve. Traditional mechanics would favor this position as the equilibrium position (x_s) of a spring equation: $f(x) = -k(x - x_s)$.

Going forward, the generalized force for these problems is

$$f_{\text{int}}(r) = T\left(\frac{\partial S}{\partial r}\right)_T = \delta k_{\text{B}} T \left\{ \frac{\gamma}{r} - \frac{\vartheta_{\xi N}}{b}\left(\frac{r}{b}\right)^{\delta - 1} \right\} \tag{C1}$$

therefore

$$R_s = \left(\frac{\gamma}{\vartheta_{\xi N}}\right)^{1/\delta} b \tag{C2}$$

where, in this work, we have focused on $\delta = 2$, so that $\vartheta_{\xi N} = (\gamma + 1/2)/(\xi N)$. From here, we can generalize things like the normalization constant in Eqn (23) to

$$C_{\gamma\delta\xi N} = \frac{\delta \vartheta_{\xi N}^{\gamma + 1/\delta} b^{\gamma\delta + 1}}{4\pi\Gamma(\gamma + 1/\delta)} \tag{C3}$$

and the general form for $\vartheta_{\xi N}$ is

$$\vartheta_{\xi N} = \left(\frac{\Gamma(\gamma + 3/\delta)}{\Gamma(\gamma + 1/\delta)} \frac{1}{\xi N} \right)^{\delta/2} = \left(\frac{\Gamma(\gamma + 3/\delta)}{\Gamma(\gamma + 1/\delta)} \frac{b^2}{\langle r^2 \rangle} \right)^{\delta/2} . \qquad \text{(C4)}$$

where $\langle r^2 \rangle = r_{\text{rms}}^2 = \xi N b^2$. This well defined structure in the equations seems to militates against the use of r_{rms} as the denatured position.

However, this is exactly where we run into trouble. There is a circular definition between $C_{\gamma\delta\xi N}$, R_s and $\vartheta_{\xi N}$ that forces us to define the quantity $\langle r^2 \rangle = \xi N b^2$. Therefore, these other quantities are only defined once r_{rms} (or some equivalent parameter) is given.

Case 2: There are also several things going for using the definition of the denatured state as r_{rms}. The radius of gyration is proportional to r_{rms} and therefore it represents a measurable quantity. This makes it more reproducible. It is also less dependent on the collection of parameters γ, δ and there are others we can invoke as well. However, it certainly does not lead to $f_{\text{int}}(R_s) = 0$, which leave one questioning whether it is adequate for a given equation such as Eqn (44b), where we combined a hybrid Gaussian and worm like chain result together with some arbitrary definitions for the crossover point (γ_w in particular).

As a result, we are not off the hook in either case. In the first publication we did on the CLE model, we sided with case 1. However, in later studies, we switched to case 2.

We are inclined to argue in favor of r_{rms} for the following reasons.

- The error introduced by using r_{rms} instead of R_s amounts to the difference of a constant. For example, in the case of Gaussian-like equations of state, $\delta = 2$ and $r_{\text{rms}} \geq R_s$. Let us compare R_s and r_{rms} relative to some other state like r_b. Then

$$\Delta S(r_b \to R_s) = k_B \left\{ \gamma \ln\left(\frac{R_s^2}{r_b^2} \right) - \frac{\zeta}{r_{rms}^2} \left(R_s^2 - r_b^2 \right) \right\},$$

and simplifying using the relationship $R_s^2 = (\gamma / \zeta) r_{rms}^2$ and $\zeta = \gamma + 1/2$, we obtain

$$\Delta S(r_b \to R_s) = k_B \left\{ \gamma \ln\left(\frac{\gamma}{\zeta} \frac{r_{rms}^2}{r_b^2} \right) - \left(1 - \frac{\zeta r_b^2}{r_{rms}^2} \right) \right\}.$$

Likewise,

$$\Delta S(r_b \to r_{rms}) = k_B \left\{ \gamma \ln\left(\frac{r_{rms}^2}{r_b^2} \right) - \left(\gamma + 1/2 - \frac{\zeta r_b^2}{r_{rms}^2} \right) \right\}.$$

The difference $\Delta\Delta S = \Delta S(r_b \to R_s) - \Delta S(r_b \to r_{rms})$ is

$$\Delta\Delta S = k_B \left\{ \gamma \ln\left(\frac{\gamma}{\gamma + 1/2} \right) + \gamma - 1/2 \right\}. \quad \text{(C5)}$$

Hence, any error this might incur amounts to an additive correction constant and therefore only contributes to the baseline in the free energy per each effective cross link (or contact).

- Second, since r_{rms} is a *measurable* quantity from the radius of gyration, this is where the system spends most of its time, even if effectively, R_s is the "true" denatured state.
- Third, for any single cross link interaction, the curve in Figure 5D is rather flat over a wide range of R_s. The more visible effects in Figure 5E result from the *collective* contribution of all the cross links as one unit. This means that any error introduced by using r_{rms} should be small. Moreover, in many applied cases, $R_s \le r_{rms}$, and therefore, the value of r_{rms} represents an upper bound.
- Finally, r_{rms} is a single value (measurable quantity) that does not depend on many complicated (and not so easily determined) parameters.

Therefore, we continue to assign r_{rms} to the denatured state in this work.

Supplementary Bibliography:

Hagerman PJ (1997) Flexibility of RNA. Annual Review Biophysics and Biomolecular Structucture 26: 139-156.

Destainville N, Manghi M, Palmeri J (2009) Microscopic mechanism for experimentally observed anomalous elasticity of DNA in two dimensions. Biophys J 96: 4464-4469.

Palmeri J, Manghi M, Destainville N (2008) Thermal denaturation of fluctuating finite DNA chains: the role of bending rigidity in bubble nucleation. Phys Rev E Stat Nonlin Soft Matter Phys 77: 011913.

Rapti Z, Smerzi A, Rasmussen KO, Bishop AR, Choi CH, et al. (2006) Healing length and bubble formation in DNA. Phys Rev E Stat Nonlin Soft Matter Phys 73: 051902.

Lesk AM (2001) Introduction to protein architecture : the structural biology of proteins. Oxford ; New York: Oxford University Press. xii, 347 p. p.

Richardson JS (1981) The anatomy and taxonomy of protein structure. Advances in Protein Chemistry 34: 167-339.

King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5: 2298-2310.

Holbrook SR (2008) Structural principles from large RNAs. Annu Rev Biophys 37: 445-464.

Higgs PG (2000) RNA secondary structure: physical and computational aspects. Quarterly Rev Biophys 33: 1999-1253.

Chen G, Chang KY, Chou MY, Bustamante C, Tinoco I, Jr. (2009) Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. Proc Natl Acad Sci U S A 106: 12706-12711.

Silverman SK, Cech TR (1999) Energetics and cooperativity of tertiary hydrogen bonds in RNA structure. Biochemistry 38: 8691-8702.

Tabaska J, Cary R, Gabow H, Stormo G (1998) An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14: 691-699.

Zhang Z, Schwartz S, Wagner L, and Miller W (2000) A greedy algorithm for

aligning DNA sequences. J Comput Biol 7:203-214. (BLASTN 2.2.26+)

Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic acids research 2010 Jul, 38 Suppl: W695-9.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007) ClustalW and ClustalX version 2. Bioinformatics 2007 23:2947-2948