

A Statistical Analysis of RNA Folding Algorithms Through Thermodynamic Parameter Perturbation

D.M. Layton and Ralf Bundschuh

Department of Physics, Ohio State University, 174 W 18th Av., Columbus OH 43210-1106

Thermodynamic parameters for RNA folding algorithms were perturbed and the effects on secondary structure prediction were analyzed. Natural sequences were found to be typically more stable to parameter perturbation than generated sequences. The use of the thermal ensemble as a measure of the reliability of a structure prediction was also studied and found to be flawed in some cases. A new measure of stability using parameter perturbation is proposed, and its limitations discussed.

In an endeavor to understand the workings of any organism, one cannot ignore the importance of ribonucleic acids (RNA). Via RNA, genetic information is transmitted through the cell. The function of a given RNA is determined by its physical structure, and determining that structure in the laboratory is a laborious, and often unsuccessful, undertaking. It has become an interdisciplinary task to determine these structures a priori.

RNA is a long polymer of chemical bases denoted as A, U, C, G, known as nucleotides. Upon first forming, an RNA is a long strand of these bases. However, soon after the RNA is formed, these bases react chemically with one another forming a new structure; this process is known as folding. The formation of a bond between two nucleotide bases is called a base pairing—A–U and G–C being the possible base pairs known as Watson-Crick pairs. With the formation of each base pairing, the Gibbs free energy of the structure is lowered, and thus, the structure's stability is increased. Since the sequence of bases that defines the RNA is finite and a base can bond with only one other base, the number of possible structures into which a given RNA can fold is finite. Thus, the most thermodynamically likely structure to be formed, the structure with the lowest free energy known as the minimum free energy structure, can be calculated by exhaustion.

Because the process is tedious and time consuming, structures are predicted using computer algorithms. The problem with these algorithms lies in the calculation of the free energies of the structures. The contribution to or detractor from the free energy attributed to a base pairing or the formation of various substructures is measured experimentally and used as parameters in the RNA folding algorithm. In the laboratory, each free energy contribution is measured at a different temperature for practical reasons. These values are then extrapolated to room temperature. Due to this extrapolation and various other reasons, these parameters contain experimental error which the folding algorithm cannot take into account. The goal of this paper is not to discuss the causes of these experimental errors or the validity of the extrapolations and approximations involved in determining the parameters used in calculating the free energy of a structure, but simply, to shed light on the consequences of these errors.

If the predicted minimum free energy (mfe) structure, also referred to as the ground state, has a far lower free

energy than any alternative structure, the ground state is said to be thermodynamically stable. For the purposes of this paper, one must define another type of stability. Should the predicted mfe structure require a strict adherence to one or more thermodynamic parameters in order to remain the predicted mfe structure, that structure/prediction is said to be unstable with respect to parameter perturbation. To study this instability, one must also devise some method of comparison for structures.

Using three widely accepted methods for the comparison of two structures: tree distance, string distance, and profile distance. The free energies and profile energies will also be compared. The tree distance is based on a metric that views a structure as being defined by a tree diagram. Likewise, the string distance views the structure as a string of "."'s, "("'s, and ")"'s representing an unpaired base, a paired base, and its base pair partner, respectively. The profile distance, however, compares structures through differences in their thermal ensemble. The thermal ensemble is a list of the probability of each particular base pair being formed. Since base pairings that are not present in the ground state still have a nonzero probability of occurring, the profile distance reflects differences in possible alternative structures as well. Likewise, possible, but not actually occurring, base pairings contribute to the before mentioned profile energy. In this paper, all distances are scaled by the length of the sequence. This scaling allows one to compare the stability of different sequences and permits a more intuitive interpretation of the data. For example, a distance of 0.20 is analogous to a 20% difference in structure since the distance between any structure and itself is zero, and by the way the distances are defined, the distance between any two structures can not exceed the sequence length.

In order to study the dependence of structure prediction to thermodynamic parameters, one must perturb the parameters. The assumption that the error in the parameters is roughly Gaussian spread. In accordance, parameters were perturbed using a random number generator with a Gaussian probability density with mean zero and standard deviation, ϵ , such that,

$$\rho(x) = \frac{1}{\sqrt{2\pi}\epsilon} e^{-\frac{x^2}{2\epsilon^2}} \quad (1)$$

Great care was taken to preserve inherent symmetry in

the parameters(e.g. the enthalpy of A pairing with U being equal to the enthalpy of U pairing with A).

For a series of natural sequences of varying length, that is, RNA sequences which have been observed in biological systems, the mfe structures were predicted by computer algorithm using the excepted experimentally measured parameters and their Gaussian perturbed counterparts. Energies and distances from the ground state for perturbed parameter predicted structures were calculated. For a given epsilon, the energies and distances were averaged over structures predicted from one hundred different parameters files generated by perturbing the excepted parameters. The ϵ 's initially used were $\{ 1, 2, 4, 8, 16, 32, 64, 128, 256 \}$ that is,

$$\epsilon = 2^x, \text{ where } x \in \mathbf{z} \wedge 0 \leq x \leq 8 \quad (2)$$

. Later, the calculations were made using $\epsilon = \{ 40, 48, 56, 72, 80, 88 \}$. The same process was done using generated non-biological sequences. Each natural sequence was compared against three sets of nineteen generated sequences of equivalent length. The generated sequences in the first set were completely random sequences of A's, U's, C's, and G's. The sequences in the second set preserved the composition of the natural sequence to which it was being compared. That is, the generated sequences were permutations of the natural sequence. The third set of sequences additionally preserved dinucleotide pairs. For example, if a G follows a C ten times in the natural sequence, the same holds true in the generated sequence.

Structures for natural sequences were also studied base pair by base pair. As mentioned previously, the thermal ensemble predicts the likelihood of a particular base pair occurring in the structure. It is an excepted practice to assume that if the thermal ensemble predicts no group of base pairs not present in the predicted structure to have a high probability of occurring, then the prediction is stable. To study this assumption, the thermal ensemble was compared to natural sequence structures predicted from 1000 different parameter files. The resulting base pairings were cataloged, and the frequencies at which they occurred was compared to thermal ensemble probability for the individual base pairings. Which structure a particular parameter file predicted was also noted, as well as, the respective distances between every predicted structure an every other structure.

The energies and distances from the ground state calculated from structures predicted by one hundred different perturbed parameter files were averaged. typically, natural sequences seem to be more stable than their non-biological counterparts regardless of the method by which they were generated. However, about one in ten non-biological sequences showed a lower tree, string, and profile distance at one or more ϵ than its natural counterpart. Although the tendency toward stability is obvious, a reliable method for identifying natural sequences could not be ascertained. The energies revealed that although all sequences deviate from there ground state energies,as a function of ϵ in roughly the same manner, the

mfe and the profile energy of the ground state is unique to the sequence, and is depends greatly on the composition of the sequence. Both of the methods of generating sequences which preserved the composition of the sequence showed a clear distinction between the energies of natural sequences and those of generated sequences as can be seen in Fig. 1. Preserving composition also in-

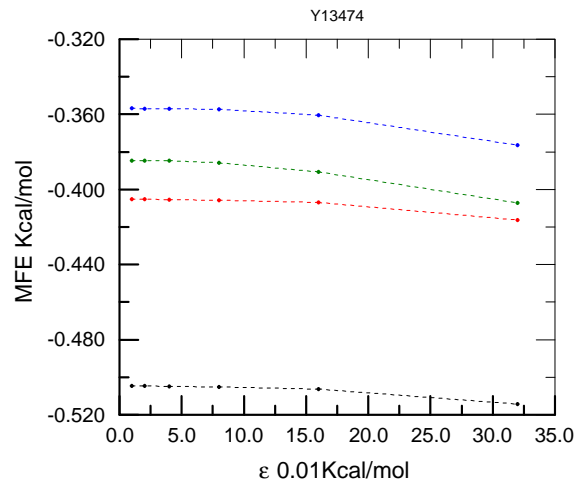


FIG. 1: The mfe at various ϵ for one natural sequence compared against a background of three random permutations of that sequence. Only three generated sequences are displayed to prevent cluttering the graph.

creased stability in regards to distance, but for all five measures, there appeared to be little or no difference between random permutations and dinucleotide preserving sequences. One should also note the overall instability of natural sequences apparent from Fig. 2. Most of the

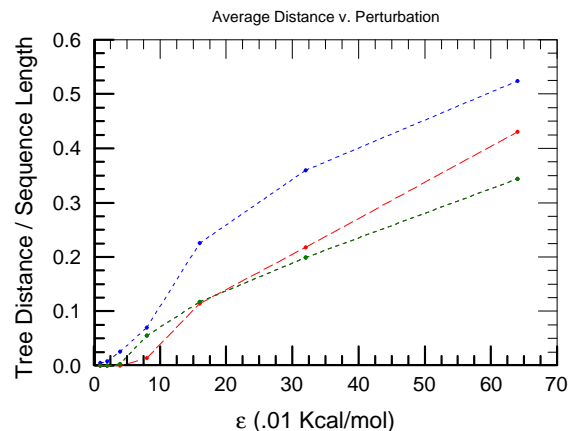


FIG. 2: The average tree distances of various natural sequences are plotted for different ϵ . Since there distance measurement are linearly proportional to sequence length, the scaling allows them to be compared with one another.

thermodynamic parameters have an error of about 0.03

Kcal/mol, which corresponds to $\epsilon \approx 30$. Thus, $\epsilon = 32$ is a of particular interest. As can be seen in Fig. 1, there is already a significant deviation from the ground state structure at that ϵ .

The comparison of structures base pair by base pair yielded several interesting results. Comparing the thermal ensemble of the ground state structure to the actual frequency of occurrence of particular base pairings yielded the plots in Fig. 3. Observe that at small ϵ since few or no alternative structures are predicted, the plot appears to be very much a step function; base pairs which the thermal ensemble predicts to have a significant probability ($\approx 30\%$ or more) occur while less likely base pairings do not. As ϵ is increased, more alternative structures begin to appear, and one can see the edges of the step begin to smooth. By $\epsilon = 32$ a strong correlation is apparent. If increased beyond $\epsilon = 32$, the correlations differs in no significant qualitative way. This behavior is typical, however two sequences at epsilons as small as two showed frequently occurring base pairings with an almost zero thermal ensemble probability. Upon further investigation, this discrepancy could be attributed to a difference in the way the sequence is folded to determine the thermal ensemble, called profile folding, and the folding which is used to determine the mfe structure. In profile folding, a base is necessarily allowed to participate in multiple dangles, a previously mentioned substructure. However, the model is improved for mfe structure folding by not allowing such "double counting" to occur when calculating the free energies. Once this improvement is disabled, the anomalous data points disappear, and the typical step behavior is restored.

In search of another measure of stability, the probability that the ground state structure will be the predicted structure at various epsilons was studied. Since the parameter files which produce each structure were cataloged, the frequency at which a structure occurred is known, as well as, the number of alternative structures possible at a given ϵ . As can be seen in table 1, the ground state structure is most likely not the true structure for some sequences for ϵ as small as two, and for all sequences for ϵ of twelve. No correlation between frequency of occurrence of an alternate structure and its distance from the ground state was observed.

Although no method for discerning natural from non-

biological sequences through an analysis of their response to parameter perturbation, it is comforting to find that natural sequences have a tendency toward stability and that nature does not simply select sequences at random. Rather she selects those which have the lowest free energy for a particular composition. The nominal bonus to stability, although helpful, falls short of this scientist's expectations. From the data, one can see that current error in the thermodynamic parameters casts serious doubt upon the structural predictions made by folding algorithms. Moreover, a reduction of error to ≈ 0.02 Kcal/mol, an order of magnitude less than current error, is necessary to make reliable predictions using current methods. Still yet, for some sequences, structural prediction will remain questionable, and prediction should still be checked for reliability using one of the discussed methods.

Perhaps the best "quick and dirty" method is the use of the thermal ensemble probabilities. By searching for likely, but absent, base pairings one can determine if likely alternate structures exist. Conveniently, programs which perform this process are standard in folding algorithm packages, and in most cases give an accurate picture of the structures stability. Should one have the time, one should check the ground state probability using the method outlined in this paper. The ground state probability method gives one not only a probabilistic measure of the accuracy of the prediction, but also all the probable alternative structures and some gauge of their likelihood of being the *true* structure. The amount of time sacrificed for the additional information is wholly dependent upon how accurate one wishes the probabilities to be. The information displayed to make table 1 took 1000 calls to the folding function, easily the most time consuming step in this method, for each epsilon to get a 0.1% accuracy. One may cut the computation time by an order of magnitude and still have the probability within a percent.

With advances in computer chip technology, the difference in computation time may soon become negligible. More importantly, advances in the laboratory and in the model will make folding algorithms more reliable and, therefore, more useful in understanding the mechanisms of all organisms on a cellular level.

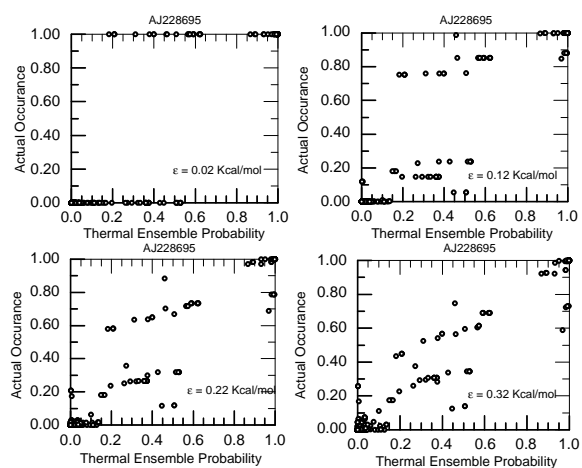


FIG. 3: These four plots show the progressive convergence of the probability of base pairs forming as predicted by the thermal ensemble and the actual probability as predicted by a folding algorithm. Each point represents a base pairing and its position represents its respective probabilities of forming.