Overlaps between RNA Secondary Structures

Paul G. Higgs*
Department of Physics, University of Sheffield, Hounsfield Road, Sheffield S3 7RH, United Kingdom
(Received 19 June 1995)

A model for RNA secondary structure is studied that has a rugged energy landscape with many alternative local minima. Several features are observed that are qualitatively similar to the replica theory of spin glasses. Numerical results suggest that the overlap distribution remains broad and non-self-averaging for long random RNA sequences at low temperatures. Significant ultrametric correlations are observed in the distances between triplets of states. Clustering algorithms are used to illustrate the hierarchical arrangement of low energy states in configuration space.

PACS numbers: 87.15.By, 75.10.Nr

Biological macromolecules such as RNA and proteins have many alternative low energy configurations separated by high energy barriers. Ideas taken from the mean field theory of the Sherrington-Kirkpatrick (SK) spin glass model [1,2] have proved useful to describe the thermodynamic properties of these molecules [3]. In the spin glass phase the overlap distribution remains broad in the thermodynamic limit, and is non-self-averaging (sample to sample fluctuations remain in the thermodynamic limit). Theory also predicts that the local minima can be arranged in a hierarchy, and that the matrix of overlaps (or the corresponding distance matrix) will be ultrametric [4]. The suggestion is that many of these features are generic to rugged landscape problems and will occur in many different models. Here we investigate overlaps in a model for RNA secondary structure in the light of spin glass theory.

Algorithms for secondary structure prediction in RNA [5–8] use recursive methods, which calculate the minimum free energy structure, the partition function, and related quantities for a molecule of length N in a time that varies usually as N^3. The full model uses parameters for the free energies of the stacking of base pairs and for formation of loops of different sizes and types that are taken from data measured on small RNA molecules. Here we simplify the parameters to a minimum. We are interested in the thermodynamic behavior of long random sequences, and the precise values of the energy parameters should not affect the conclusions.

We study random RNA sequences composed of A, C, G, and U bases with equal probability. Pairing is permitted only between A and U and between C and G bases. Each pair contributes an energy of –1 unit, irrespective of its position in the structure, and there is no penalty for loops. The topological rules that determine which structures are allowed are the essential feature that creates the rugged landscape in this model. Let i, j, k, and l be the positions of four bases in a sequence numbered from 1 to N, such that i and j can form a pair and k and l can form a pair. There are three nonequivalent possibilities for the order: $i < j < k < l$, $i < k < l < j$, and $i < k < j < l$. In the first two cases the two pairs are permitted to occur simultaneously. The third case is known as a pseudoknot, and is disallowed here and in most other work on RNA. Any base pair must also satisfy $|j - i| \geq 4$, which guarantees that there are at least three unpaired bases in a hairpin loop. We wish to calculate the partition function, which is a sum over all structures satisfying the above rules.

Let Z_{ij} be the partition function for the section of chain from bases i to j inclusive. Let ε_{ij} be the energy of the bond between bases i and j, which is –1 if the pair is AU or CG, and $+\infty$ otherwise, and let $a_{ij} = \exp(-\varepsilon_{ij}/T)$, which is either $\exp(+1/T)$ or zero. The full partition function Z_{1N} can be calculated recursively at any given temperature T, beginning with $Z_{ii} = Z_{i,i-1} = 1$ for all i, and using the following formula for $j > i$.

$$Z_{ij} = Z_{i,j-1} + \sum_{h=i}^{j-4} Z_{i,h-1} Z_{h+1,j-1} a_{ij} .$$

The time required is $O(N^3)$. The recursions for calculating the partition function when the full set of energy parameters are used is more complicated [7], but is still $O(N^3)$.

Any given secondary structure α can be represented as a sequence of integers $b_1^\alpha, \ldots, b_N^\alpha$, such that $b_i^\alpha = 0$ if base i is unpaired in configuration α, and $b_i^\alpha = j$ and $b_j^\alpha = i$ if bases i and j are paired. The overlap $q^{\alpha\beta}$ between two configurations α and β is the fraction of bases in the chain for which $b_i^\alpha = b_i^\beta$. The overlap distribution function is $P(q) = \sum_\alpha \sum_\beta w_\alpha w_\beta \delta(q - q^{\alpha\beta})$, where w_α and w_β are the equilibrium statistical weights of states α and β. It is possible to write a polynomial time recursion for calculating $P(q)$ exactly for a given chain, but both the time and the memory requirements are of high order in N. This approach was only practical for very short chains. There is, however, a way of approximately calculating $P(q)$ using the partition functions Z_{ij}. This consists of obtaining a representative sample of structures for a given temperature, such that the probability of a structure occurring in the sample is proportional to its Boltzmann weight. If the number of structures generated is fairly large, then the distribution of overlaps

 0031-9007/96/76(4)/704(4)$06.00

between pairs of structures in the sample should be a good approximation to the full $P(q)$.

The following method generates a structure at random with a probability equal to its Boltzmann weight. It is possible to calculate the probability that base N is paired with any other base h in the sequence, and the probability that base N is unpaired. The state of base N can thus be assigned at random using these probabilities. If base N is unpaired, b_N is set to 0. If bases N and h are paired, b_h is set to N and b_N is set to h. This process can be repeated until all the b_i are assigned. At each stage it is necessary to consider an interval of the sequence such that none of the b_i in the interval is yet assigned, and such that it is already known that pairs are forbidden between bases in the interval and those outside. In the above case where N is unpaired it is now necessary to consider the interval from 1 to $N-1$. In the case where N was paired with h it is necessary to consider two new intervals from 1 to $h-1$, and from $h+1$ to $N-1$. Given a general interval from i to j it is always possible to determine the probability that the last base in the interval (i.e., j) is paired with any other base h in the interval. We denote this by $p(j,h;i,j)$. The probability that j is unpaired, given the interval i to j, is denoted $p(j,0;i,j)$. These probabilities are given by

$$p(j,h;i,j) = a_{hj} Z_{i,h-1} Z_{h+1,j-1} / Z_{ij} ,$$

$$p(j,0;i,j) = 1 - \sum_{h=i}^{j-4} p(j,h;i,j) .$$

Base j can now be either paired with some base h or left unpaired by generating a random number and comparing with the probabilities above. The procedure can be repeated in smaller and smaller intervals until all the b_i are assigned. It is always necessary to consider the last base in the interval, since the probability of forming a general pair h,k in an interval i,j is not calculable from the known partition functions. This is an exact method, which is guaranteed to generate states with their correct equilibrium probabilities, hence there is no need for approximate methods such as Monte Carlo with simulated annealing.

Figure 1 shows overlap distributions averaged over many random sequences. At $T = 0.5$ the width of the distribution scales as $N^{-1/2}$ and $P(q)$ becomes a delta function in the thermodynamic limit, as would be expected for a high temperature phase. At $T = 0.1$ the distribution is much broader and appears to narrow only very slowly as N increases, suggesting that $P(q)$ may tend to a nontrivial limit for large N. The width of $P(q)$ was plotted against N on a log-log plot at several temperatures (not shown). Clear $N^{-1/2}$ behavior was observed for $T = 0.5$ and 0.3. For $T = 0.2$, 0.1, and 0.0 the width decreased much less rapidly than $N^{-1/2}$ but a stationary value had not yet been reached for the maximum length considered ($N = 1200$). Note that the ground state is multiply degenerate, therefore $P(q)$ is broad at zero temperature.

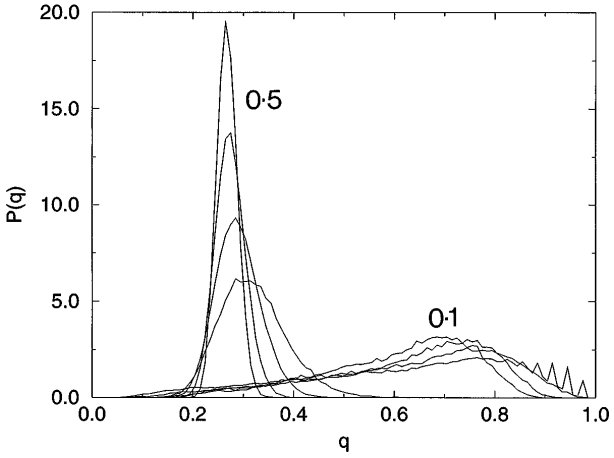

FIG. 1. Mean overlap distributions at $T = 0.1$ and 0.5 for $N = 100$ (lowest of each set), 200, 400, and 800 (highest of each set).

The mean q was measured for each sequence and the standard deviation of the means, δq, was obtained as a measure of sample to sample fluctuations. Figure 2 shows δq as a function of N. For normal thermodynamic averaging δq should decrease as $N^{-1/2}$ for large N, whereas if $P(q)$ is non-self-averaging it should tend to a constant. In Fig. 2, at $T = 0.5$ and 0.3 δq is decreasing at least as rapidly as $N^{-1/2}$, as indicated by the broken line, whereas for $T = 0.2$, 0.1, and 0.0 δq decreases only very slowly. These results suggest the presence of a low temperature phase with non-self-averaging $P(q)$, with a transition temperature somewhere around $T = 0.2$. However, from numerical evidence alone we cannot rule out the possibility that δq continues to decrease very slowly in the thermodynamic limit

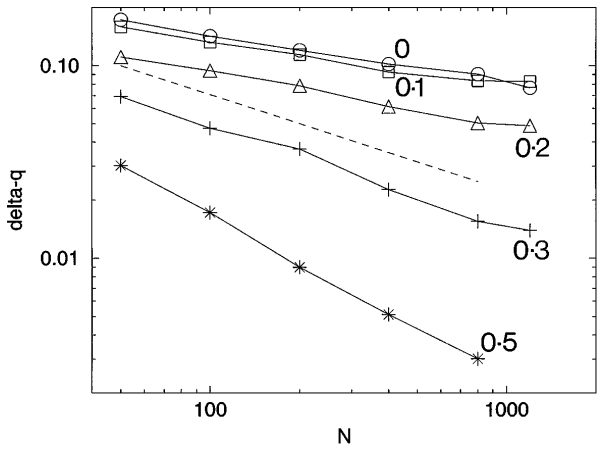

FIG. 2. The standard deviation δq of the mean overlap shown as a function of N for temperatures between 0 and 0.5. The broken line has a slope of $-1/2$.

even at low temperatures. In the SK model the spin glass transition is signaled by a cusp singularity in the magnetic susceptibility. We have looked for singularities in the RNA model using recursion formulas for various derivatives of the partition function, but this has proved inconclusive. It is difficult to locate cusp singularities because they are easily masked by finite size effects, hence it is difficult to prove the presence of a phase transition in this model, and the possibility remains that there is no transition.

The other key feature of the replica theory of the SK model is that the valleys in the energy landscape (or pure states) are predicted to be arranged in a hierarchical way, so that the matrix of distances between the pure states is ultrametric. For any three pure states α, β, and γ there is a triangle with sides equal to the distances $d^{\alpha\beta}$, $d^{\beta\gamma}$, $d^{\alpha\gamma}$. The ultrametric property is that the two longest sides of the triangle must be equal. In the RNA model the distance will be defined as $d^{\alpha\beta} = 1 - q^{\alpha\beta}$, which is just the fraction of the b_i variables that are different in the two structures. These distances satisfy the triangle inequality $d^{\alpha\gamma} \leq d^{\alpha\beta} + d^{\beta\gamma}$, for any α, β, and γ. We will let s denote the mean value of the difference between the longest and second longest sides of triangles in the measured data. The average is taken over many triangles per sequence and over many sequences. Values of s are shown as a function of N at two representative temperatures in Table I. It is found that s is quite small at all temperatures, and always decreases as N increases.

However, there are two reasons why s should be small that do not indicate a significant ultrametric structure. First, the width of the distribution of d scales as $N^{-1/2}$ in the high temperature phase, hence s must also scale as $N^{-1/2}$ (see data in Table I for $T = 0.5$). Second, the triangle inequality also limits the maximum size of s. In order to demonstrate that the real triangles are closer to being Isosceles than would be expected merely from the triangle inequality we generated random uncorrelated triangles in the following way. Groups of six structures $\alpha,\beta,\gamma,\delta,\mu,\nu$ were considered, from which three independent distances $d^{\alpha\beta}$, $d^{\gamma\delta}$, and $d^{\mu\nu}$ were obtained. If these three distances satisfied the triangle inequalities, then they were included in the sample of uncorrelated triangles, otherwise they were rejected and a new group of six generated. The uncorrelated triangles have the same distribution of lengths of sides as the real triangles formed from triplets of states, but there is no correlation between the lengths of the two longest sides other than that introduced by the triangle inequality. Table I shows the ratio of s (for real triangles) to s_{unc} (for uncorrelated triangles). This is always less than 1. At $T = 0.5$ the ratio appears to increase towards 1 as N increases. This means that the data are ultrametric only in a trivial way: s is small mostly due to the narrow $P(q)$ and the triangle inequality (cf. example of trivial ultrametricity in [9]). By contrast, at $T = 0.1$ the ratio s/s_{unc} is smaller, and decreases with N. Thus there are real ultrametric correlations that become more significant with increasing N.

TABLE I. Mean deviation from ultrametricity s per triangle, and the ratio of s for real and uncorrelated triangles shown for two temperatures.

	$T = 0.5$		$T = 0.1$	
N	s	s/s_{unc}	s	s/s_{unc}
50	0.058	0.80	0.058	0.73
100	0.042	0.84	0.055	0.67
200	0.030	0.87	0.048	0.62
400	0.021	0.89	0.044	0.55
800	0.015	0.90	0.040	0.52
1200	0.012	0.91	0.037	0.48

A large number of hierarchical clustering algorithms [10] have been developed for the analysis of distance matrices. Figure 3 shows a representation of the distance matrix for a set of 200 alternative structures for one single random RNA sequence of length 400. The structures were obtained at $T = 0$, therefore they are all degenerate ground states. The single link clustering method [10] was used to order the structures so that similar structures are consecutive. Clusters appear as blocks along the diagonal of the matrix; 20 grey levels have been used, with darker shades indicating smaller distances. A pattern of clusters is seen in Fig. 3 with at least three hierarchical stages of division being apparent at different grey levels. The structure gradually disappears as the temperature is raised. At $T = 0.5$ there is just a black line along the diagonal,

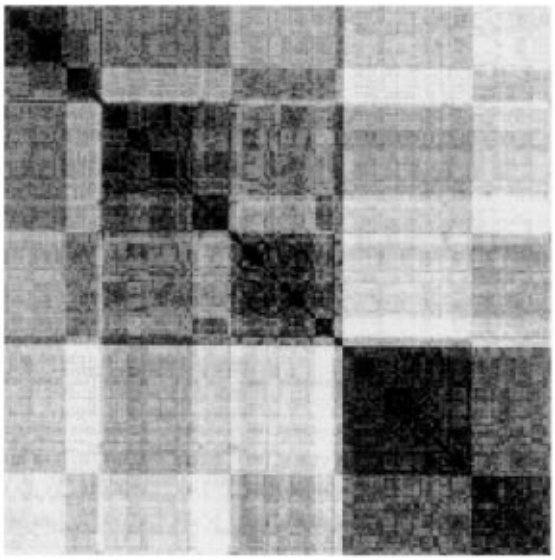

FIG. 3. Matrix of distances between alternative structures for a random sequence of $N = 400$ at $T = 0$. Darker shades indicate shorter distances. A hierarchical pattern of clusters is visible.

since all the off-diagonal elements are approximately equal.

The pattern in Fig. 3 is what would be expected for a matrix that is approximately ultrametric [2,4,11]. In an exactly ultrametric matrix all the elements in any off-diagonal block would be equal, and hence appear the same grey shade. In the figure, structure is apparent within the off-diagonal blocks, indicating some deviation from ultrametricity. It is possible that these deviations are finite size effects. On the other hand, the replica theory of spin glasses only applies for the mean field SK model. The RNA model studied here is neither mean field nor a spin glass, and we have no analytical theory for this model. We therefore have no reason to expect exact ultrametricity. It may be that the deviations from ultrametricity are real deviations from the mean field theory, not finite size effects. Rammal and co-workers [11,12] have considered ways of quantifying the degree of ultrametricity in a distance matrix. More work along these lines may help to distinguish finite size effects from true deviations from the mean field theory.

Evidence for ultrametricity has been seen in numerical simulations of the SK model [13], in solutions of the traveling salesman problem [14] and the graph bipartitioning problem [15], in a lattice model with asymmetric couplings [16], and in simulations of protein folding [9]. This article gives another example. This model is very suitable for numerical work, since the partition function and related quantities can be generated exactly with no need to resort to Monte Carlo techniques. Even though this is not a mean field model many features have been observed that are qualitatively similar to those of the mean field SK spin glass.

*Present address: School of Biological Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK.

[1] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[2] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[3] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[4] M. Mézard, G. Parisi, N. Sourlas, G. Toulouse, and M. Virasoro, J. Phys. (Paris) **45**, 843 (1984).
[5] M. Zuker, J. A. Jaeger, and D. H. Turner, Nucl. Acids Res. **19**, 2707 (1991).
[6] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatshefte für Chemie **125**, 167 (1994).
[7] J. S. McCaskill, Biopolymers **29**, 1105 (1990).
[8] P. G. Higgs, J. Phys. I (France) **3**, 43 (1993); J. Chem. Soc. Faraday Trans. **91**, 2531 (1995).
[9] B. Velikson, J. Bascle, T. Garel, and H. Orland, Macromolecules **26**, 4791 (1993).
[10] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy* (W. H. Freeman and Co., San Francisco, 1973).
[11] R. Rammal, G. Toulouse, and M. A. Virasoro, Rev. Mod. Phys. **58**, 765 (1986).
[12] R. Rammal, J. C. Angles d'Auriac, and B. Doucot, J. Phys. Lett. **46**, L945 (1985).
[13] R. N. Bhatt and A. P. Young, J. Phys. Condens. Matter **1**, 2997 (1989).
[14] S. Kirkpatrick and G. Toulouse, J. Phys. (Paris) **46**, 1277 (1985).
[15] J. R. Banavar, D. Sherrington, and N. Sourlas, J. Phys. A **20**, L1 (1987).
[16] A. T. Bernardes and H. J. Herrmann, Int. J. Mod. Phys. C **4**, 765 (1993).