Barrier heights between ground states in a model of RNA secondary structure

1998 J. Phys. A: Math. Gen. 31 3153

(http://iopscience.iop.org/0305-4470/31/14/005)

J. Phys. A: Math. Gen. **31** (1998) 3153–3170. Printed in the UK PII: S0305-4470(98)86996-5

Barrier heights between ground states in a model of RNA secondary structure

Steven R Morgan and Paul G Higgs

School of Biological Sciences, 2.205 Stopford Building, University of Manchester, Oxford Rd, Manchester M13 9PT, UK

Received 20 August 1997

Abstract. Secondary structure formation is an important factor influencing the behaviour of many types of naturally occurring RNA molecules. RNA secondary structure also provides an example of a disordered system showing frustration and a rugged energy landscape with many alternative ground states. We use a numerical method to estimate energy barrier heights between a set of alternative ground-state structures of a given sequence. Both the mean barrier height and the maximum barrier height for a given sequence scale with a sequence length N approximately as $N^{1/2}$. The matrix h of barriers is exactly ultrametric, which means that structures form an exactly hierarchical set of clusters based on barrier heights. The matrix of distances d between structures is correlated with h, but with large fluctuations. Hence, the d matrix is approximately ultrametric, but the clustering of structures based on distances is not perfectly hierarchical. Certain base pairs are observed to be present in every ground-state structure. These 'frozen' pairs divide the molecule up into mutually inaccessible pieces. All of the separate pieces contribute to determining the distance between structures, but only the largest piece will contribute to the barrier height. The length of the largest piece varies between sequences, and scales in proportion to N on average. We compare these results with studies of other disordered systems, and discuss the consequences for the folding of naturally occurring RNAs.

1. Introduction

RNA is a molecule of fundamental biological importance, which plays a large number of roles in living cells. Molecules consist of sequences of A, C, G and U bases which can fold to form secondary structures by forming base-paired helical regions between complementary parts of the same sequence. Possible complementary pairs are GC, AU and the weaker pair GU. Many alternative low-energy folded configurations are possible for any given sequence. Algorithms are available to predict the minimum free energy (MFE) secondary structure of a sequence using dynamic programming [1–5]. These algorithms build up the MFE structure and the partition function recursively starting from small subsections of the sequence. The free energy parameters for all the possible structural motifs are derived from experimental measurements on short RNA oligomers [6].

Our previous work has focused on the thermodynamic properties of RNA [7, 8] and on the influence of kinetics in the folding process [9, 10]. In this paper we study the statistical physics of random RNA sequences. Random heteropolymers have attracted the interest of statistical physicists, since they are examples of disordered systems with interesting thermodynamic behaviour at low temperatures [11–15]. Recently we studied an RNA folding model with deliberately simplified energy rules which has many different

0305-4470/98/143153+18$19.50

ground states [16]. The distribution of overlaps between states was found to be broad at low temperatures, and when triplets of states were considered there was evidence for an ultrametric correlation in the distances between states. These properties are similar to those seen in spin-glass models [17].

Ultrametric clustering of states (as measured by distances or overlaps) is usually thought to indicate the presence of large energy barriers between different low-energy states. In most disordered systems it is very difficult to measure these barrier heights directly. It is instead necessary to perform Monte Carlo simulations for very long time periods and attempt to estimate relaxation times. The largest barrier height crossed during relaxation, scales as the log of the longest relaxation time. Rugged landscape problems are likely to be glass-like at low temperature by their very nature, and this will make simulations of this type difficult and time-consuming. Here we show that in the simplified RNA folding model which we use, it is possible to obtain a good estimate of the distribution of barrier heights between ground states and the way in which heights scale with sequence length. Barriers are estimated directly rather than by the use of kinetic simulation.

2. Definition of the model

We study random RNA sequences composed of A, C, G and U bases with equal probability. For simplicity only AU and CG pairs are permitted. Each pair contributes an energy of -1 unit, irrespective of its position in the structure, and there is no penalty for loops. This model was originally considered as a model for RNA secondary structure prediction before more sophisticated energy parameters were introduced [1]. The topological rules that determine which structures are allowed are the essential feature that create the rugged landscape in this model, causing frustration between competing base pairs. 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 also pair. There are three non-equivalent possibilities for the order: $i < j < k < l$, $i < k < l < j$, and $i < k < j < l$. In the first two cases the pairs are permitted to occur simultaneously, in which case we call them compatible. The third case is known as a pseudoknot, and is disallowed here and in most other work on RNA, since these structures are relatively rare in real RNAs, since current recursive MFE algorithms cannot deal with pseudoknots. Any base pair must also satisfy $|j - i| \geqslant 4$, 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 of the Boltzmann factors of all structures satisfying the above rules.

Let Z_{ij} be the partition function for the section of chain from bases i–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 $j > i$,

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

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

Table 1 shows the way in which several properties of this model depend on the chain length N. The ground-state energy E_0 is proportional to N, whilst the total number of states Ω and the number of degenerate ground states ω both increase exponentially with N.

Table 1. Dependence of the mean ground-state energy, the total number of states per sequence and the mean number of ground states per sequence on sequence length N. The overbar indicates an average over random sequences. These figures were estimated by using linear regression on data obtained numerically with N in the range 50–400.

Quantity	Fitting function	Parameters
Ground state energy	$\overline{E_0} = C_1 + \varepsilon N$	$C_1 = 2.9(\pm 0.2)$ $\varepsilon = -0.368(\pm 0.001)$
Total number of states	$\overline{\ln \omega} = C_2 + \alpha N$	$C_2 = -5.6(\pm 0.4)$ $\alpha = 0.533(\pm 0.001)$
Number of ground states	$\overline{\ln \omega} = C_3 + \beta N$	$C_3 = 1.75(\pm 0.2)$ $\beta = 0.068(\pm 0.001)$

A typical chain of length 200 has approximately 5×10^6 ground states and there are about 70 base pairs in each ground state.

In a previous paper [16] it was shown that one the partition functions Z_{ij} are known for each subsection of the chain, it is possible to generate a configuration at random with a probability equal to its Boltzmann weight. If this method is applied at zero temperature then only ground states are generated. In this paper we use the same method to generate a random set of ground states, and barriers will be estimated between every pair of states in this set. Clearly we cannot generate every possible ground state but since the set we use is generated at random the properties of the states examined should be representative of the complete set of ground states.

We note that structure A can be represented by a set of integers $\{b_i^A\}$, where $b_i^A = j$ and $b_j^A = i$ if bases i and j are paired, and $b_i^A = 0$ if base i is unpaired. As a measure of the distance between two structures A and B of the same sequence we define d_{AB} to be the number of bases for which $b_i^A \neq b_i^B$ (hence d_{AB} is an integer in the range 0–N).

Having calculated the partition function Z_{ij} the probability that bases i and j are paired at equilibrium is

$$p_{ij} = \frac{a_{ij} Z_{i+1,j-1} Z_{ij}^{\mathrm{ends}}}{Z_{1,N}} \tag{2}$$

where Z_{ij}^{ends} is the partition function for the sum of all states which can be constructed from the two ends of the sequence, from 1 to $i-1$ and from $j+1$ to N. This can be calculated from the following recursion:

$$Z_{ij}^{\mathrm{ends}} = Z_{i-1,j}^{\mathrm{ends}} + \sum_{h=1}^{i-5} Z_{hj}^{\mathrm{ends}} a_{h,i-1} Z_{h+1,i-2} + \sum_{h=j+1}^{N} a_{i-1,h} Z_{i-1,h}^{\mathrm{ends}} Z_{j+1,h-1} \tag{3}$$

using the initial conditions $Z_{1j}^{\mathrm{ends}} = Z_{j+1,N}$ and $Z_{iN}^{\mathrm{ends}} = Z_{1,i-1}$. The probability that base i is unpaired is

$$p_{i,0} = 1 - \sum_{j=1}^{N} p_{ij}. \tag{4}$$

The mean Boltzmann weighted distance between all pairs of structures for a given sequence is related to the p_{ij}'s in the following way:

$$\langle d \rangle = N - \sum_{i=1}^{N} \sum_{j=0}^{N} p_{ij}^2. \tag{5}$$

Applying this equation at $T = 0$ gives the mean distance between alternative ground states.

3. Method of estimating barriers

We will define neighbouring configurations as configurations which differ by either the addition or removal of a single base pair. For any given pair of ground states there are many routes through neighbouring configurations by which the molecule can get from one to the other. We will define the barrier ΔE associated with a given route between two ground states A and B as the difference between the ground-state energy and the maximum energy of all the configurations on the route. We are interested in the most probable route from A to B, which for sufficiently low temperature will be the one with the lowest barrier, since the probability of getting over the barrier is proportional to $e^{-\Delta E/kT}$. Let h_{AB} be the minimum value of ΔE_{AB} for all possible routes, i.e. h_{AB} is the highest point on the lowest route.

The number of base pairs in a ground state is P_{tot}, which varies from sequence to sequence, but increases proportionally to N on average. For any two ground states, A and B, we can write $P_{tot} = P_S + P_D$, where P_S is the number of shared pairs (i.e. those which are in both A and B) and P_D is the number of distinct pairs (those which are in A and not in B, or vice versa). We define a direct route from A to B as one which only involves addition and removal of distinct pairs. There will be a set of P_D pairs present in A which must be removed and a different set of P_D pairs present in B which must be added. There are thus $2P_D$ steps along a direct route. We define the direct route barrier h_{AB}^{Dir} between A and B as the minimum of ΔE_{AB} for all possible direct routes.

Indirect routes are those which do not satisfy the above criteria. An indirect route may involve removal of some of the shared pairs along the route, which will then have to be added again as the configuration approaches B. It may also involve addition of extra pairs which are present in neither A or B. These will have to be removed again at a later stage. There are far more indirect routes than direct ones, and indirect routes always have more than $2P_D$ steps.

We begin with the problem of finding the optimal direct route, and its barrier h_{AB}^{Dir}. If we label the pairs in A to be removed from 1 to P_D and those in B to be added from $P_D + 1$ to $2P_D$ then a direct route can be written as a permutation of these integers. Not all of the $(2P_D)!$ permutations are valid routes however, since some will attempt to add pairs before the corresponding incompatible pairs have been removed (see figure 1 route (i)). Also some routes will reach excessively high energies because they remove a lot of pairs early in the route (e.g. route (ii)). A way of creating 'reasonable' routes which are valid and not excessive is to begin with a permutation of the pairs to be added (as shown in (iii) and (iv)), and to slot in the pairs to be removed at the latest possible point, so as not to make the route invalid. It is clear that the optimal direct route is one of these reasonable routes.

There are only $P_D!$ reasonable routes rather than $(2P_D)!$, however, this is still too large for exact enumeration for long sequences since P_D scales as N. The problem of finding the best direct route is similar in nature to the travelling salesman problem [18], for which the number of tours increases as the factorial of the number of cities visited. We did not find an exact solution to the direct route problem, but we used the following heuristic procedure, which we believe finds the optimum route in most cases.

For each of the P_D pairs in B to be added a list is made of those pairs in A which are incompatible with it, as shown in figure 1. An initial estimate of the direct route barrier is obtained as follows.

(i) Find the pair in B which has the least number of incompatible pairs in A. If there are several pairs in B with an equal number, choose one randomly.

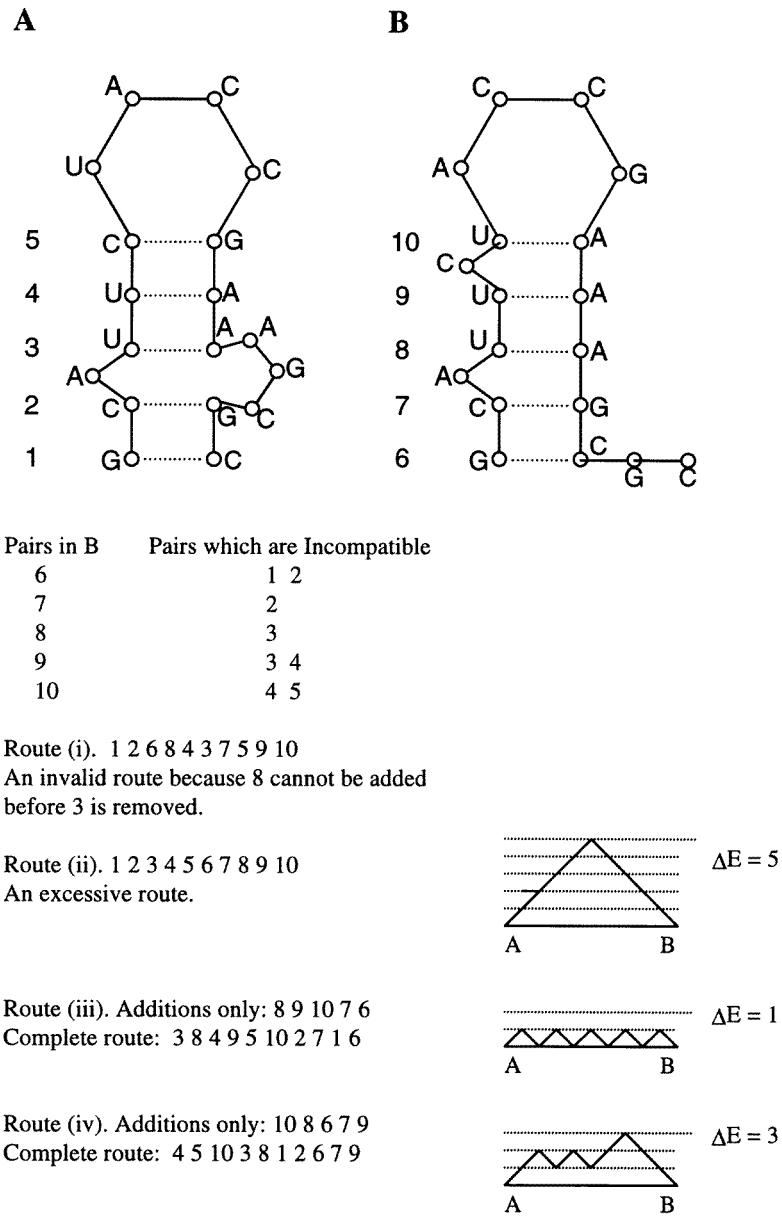

Figure 1. Two ground states for a sequence with $P_{tot} = 5$, $P_S = 0$ and $P_D = 5$. Barriers for several possible routes are calculated. Base pairs in the structures are numbered, and the routes show the order of removal and addition of paris to convert A into B.

(ii) Remove the incompatible pairs from A, and add the new pair to the structure. If there are any other pairs from B that can be added to the structure at the same time, then add these. This creates a new intermediate structure A′.

(iii) Record the energy of the maximum point reached on the route so far.

(iv) Repeat this procedure with A′ instead of A, until the intermediate structure is

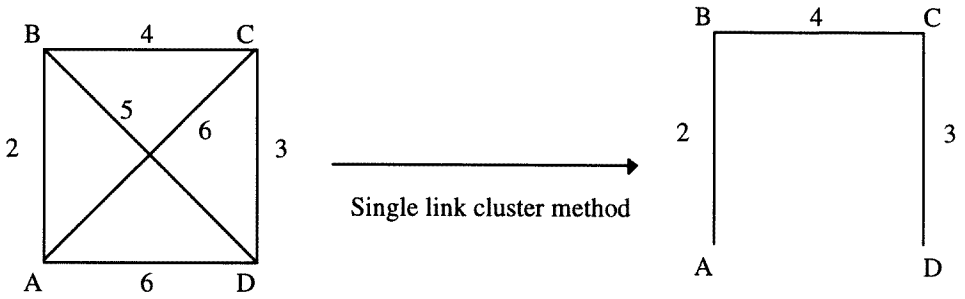

Figure 2. Effect of clustering the barriers using the SLC method. The final set of barriers is ultrametric.

transformed into the final structure B.

In figure 1, route (iii) is constructed by this method. Since step (i) of the method sometimes involves random choices, it is necessary to repeat the method a number of times, making different random choices, and following each new route only as long as the barrier height along the route so far is less than the best estimate for h found so far. In most cases there are only a few choices at each step, and the estimate for h converges after relatively few iterations.

The method above is a 'greedy' algorithm, similar to those used in problems such as the travelling salesman problem. In that problem a greedy algorithm is usually found to give a good first estimate of the minimum route length between cities, but is not guaranteed to find the optimum. Here we also found that the greedy algorithm did not always give the lowest estimate for the barrier height. Therefore, after using the greedy algorithm above, we used the following modified method in an attempt to improve the result. The modified method consists of simply choosing a pair from B at random and removing those pairs from A which are necessary in order to add this pair, otherwise it is the same as before. Since there are no constraints on the route chosen, this method is guaranteed to find the lowest barrier route eventually. The modified method was carried out a few hundred times for each pair of structures. Each trial route was followed only as long as the barrier reached so far did not exceed the minimum h found from previous routes. In this way, time was not wasted following very poor routes.

As stated above, indirect routes involve either adding pairs which are later removed or removing pairs which are later added again. In practice, it was found that indirect routes often had lower barriers than direct ones, and therefore it was necessary to consider indirect routes. A set of ground-state structures was generated and h_{AB}^{Dir} was estimated by the above method for every pair. We then considered indirect routes that were formed by linking direct routes. For example in figure 2 indirect routes from A to D include ABD, ACD and ABCD. The barrier for one of these indirect routes is the largest barrier encountered along the linked direct routes. The optimum indirect route in this case is ABCD which has a barrier of $h_{AD} = h_{BC}^{Dir} = 4$. The algorithm for calculation of the optimum indirect routes is known as the single link cluster (SLC) method in numerical taxonomy, and is well described by Sneath and Sokal [19].

The matrix of true barriers between ground states h is ultrametric by its definition, that is it satisfies $h_{AB} \leqslant \max(h_{AC}, h_{CB})$ for every triplet of states A, B, and C [20]. The matrix of direct barriers is symmetric ($h_{AB}^{Dir} = h_{BA}^{Dir}$) but is not necessarily ultrametric. By using h^{Dir} as the input to the SLC method we obtain an output matrix h^{est} which is exactly ultrametric

and which is our best estimate of the true h matrix. Clearly h^{est} is an upper bound on the true h, because:

(i) the input values of h^{Dir} are themselves upper bounds which were estimated by a heuristic procedure;

(ii) only a finite set of ground states can be considered in the matrix (typically 200), whereas the total number of ground states increases exponentially with N;

(iii) there are possibly some indirect routes which cannot be decomposed into linked direct routes.

Nevertheless, we believe that in practice our method gives a good approximation to the true h values since the SLC method is very tolerant of errors in the input. For example if $h^{\mathrm{Dir}}_{\mathrm{AB}}$ had been overestimated as 7 instead of 6 in figure 2 then it would make no difference to the outcome ($h^{\mathrm{est}}_{\mathrm{AD}} = 4$). The procedure for estimating h^{Dir} is unlikely to make mistakes with low barriers, since it converges quickly to a value which is not significantly improved on by further trial routes. It is the pairs for which $h^{\mathrm{Dir}}_{\mathrm{AB}}$ is estimated as being unusually high which have a greater possibility of being overestimates. These high values tend not to contribute to the final h_{est} as shown above. We performed several tests on our results in order to check if the method was giving good estimates of h. These are described in the results section.

4. Results

4.1. Barrier heights

In this section we estimate the average barrier between ground states, defined as the mean of h_{AB} for all pairs of ground state structures, and the maximum barrier h_{max}, which is the largest element of the h_{AB} matrix. Since these quantities fluctuate from one random sequence to another it is necessary to average them over sequences. These sequence averages will be denoted $\overline{\langle h \rangle}$ and $\overline{h_{\mathrm{max}}}$. This was done for 30 sequences each for various sequence lengths from 50 to 850 bases. For each base sequence used, two sets of 200 ground-state structures were generated, and for each of these two sets the best estimate of the h matrix was found using the clustering algorithm. Two separate sets were considered in order to estimate the reliability of the method. An estimate of the random error in $\langle h \rangle$ was calculated using the difference between $\langle h \rangle$ for the two sets of structures used, and found to be about 2% on average. As a further check on statistical error it is possible to measure the mean distance between all pairs of ground states in the set, and compare this with the exact answer which is calculable from the pair probabilities (see equation (5)). These were found to agree within the same error margin as before.

If more ground-state structures are included in the clustering method there will be a larger number of indirect routes considered, thus the chance of finding low barrier routes (if these exist) will increase. Using 400 structures instead of 200 increases the number of direct routes considered by a factor of 4, and enormously increases the number of indirect routes. However, it was found that the average barrier was only decreased by about 5% for a few examples at the largest sequence length $N = 850$ (it was not possible to test more than a few examples due to the computer time that would be required). The decrease will be lower for smaller sequence lengths because the direct routes are not so likely to be overestimated, as there are not so many possible routes. Hence we believe that using 200 structures is sufficient to obtain a reasonable estimate, even for the largest N value.

Figure 3 shows the values of $\overline{\langle h \rangle}$ and $\overline{h_{\mathrm{max}}}$ against the sequence length N on a log–log scale. For each of the individual sequences there was no significant difference between the

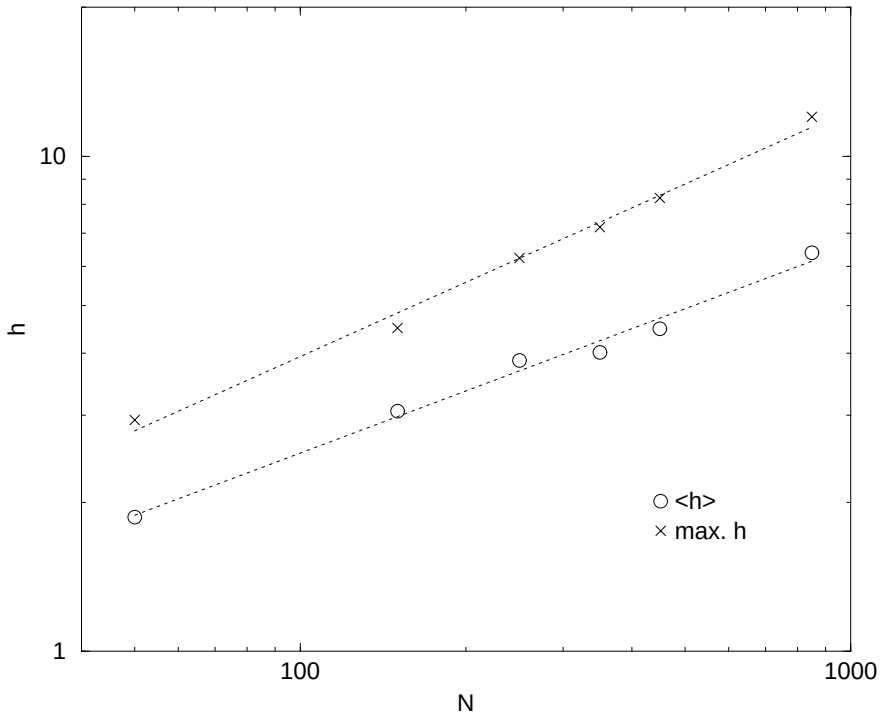

Figure 3. The bottom curve shows the average barrier between all pairs of structures in the set of 200, averaged over all sequences, for a range of sequence lengths. The top curve is the maximum barrier averaged over all sequences for each length. The dotted lines are best fit straight lines from which the scaling laws for the average and maximum barriers are obtained.

value of $\langle h \rangle$ estimated from the two separate sets of states, although fluctuation between sequences was considerable. In some cases, the value of $h_{\max}$ differed between the two separate sets for the same sequence. This is noticeable because $h_{\max}$ is always an integer. When this occurred, the lower of the two values was taken as the $h_{\max}$ for that sequence. An average over sequences was then calculated. From the gradient of the graphs we can estimate that $\overline{\langle h \rangle}$ scales as $N^{0.42}$, and the maximum barrier $\overline{h_{\max}} \sim N^{0.50}$. There are several sources of error in the estimation of the exponents. Statistical error is of course present due to the limited number of sequences examined at each length, and the limited number of structures considered for each sequence. The estimates of the statistical errors in the exponents obtained from the least-squares fitting routine were quite small, however (about ± 0.02). For reasons which we discuss below, we expect these two exponents to be equal. The apparent difference between the two estimates suggests that there are uncertainties of the order of ± 0.05. These are due to finite-size effects in the $N = 50$ point (removal of this point leads to slight changes in the estimate), and a possible systematic overestimation in h values (which would affect the large N points most).

In previous work on this model [16] the distribution of distances between structures was investigated. It was found that this distribution was broad at low temperatures even for long sequences, and that there was some evidence for ultrametric clustering of states based on the distance values. In models of disordered systems it is usually assumed that clusters of states which are far apart in distance will be separated by a high-energy barrier, and hence that transitions between these states will be very slow. In many simulations

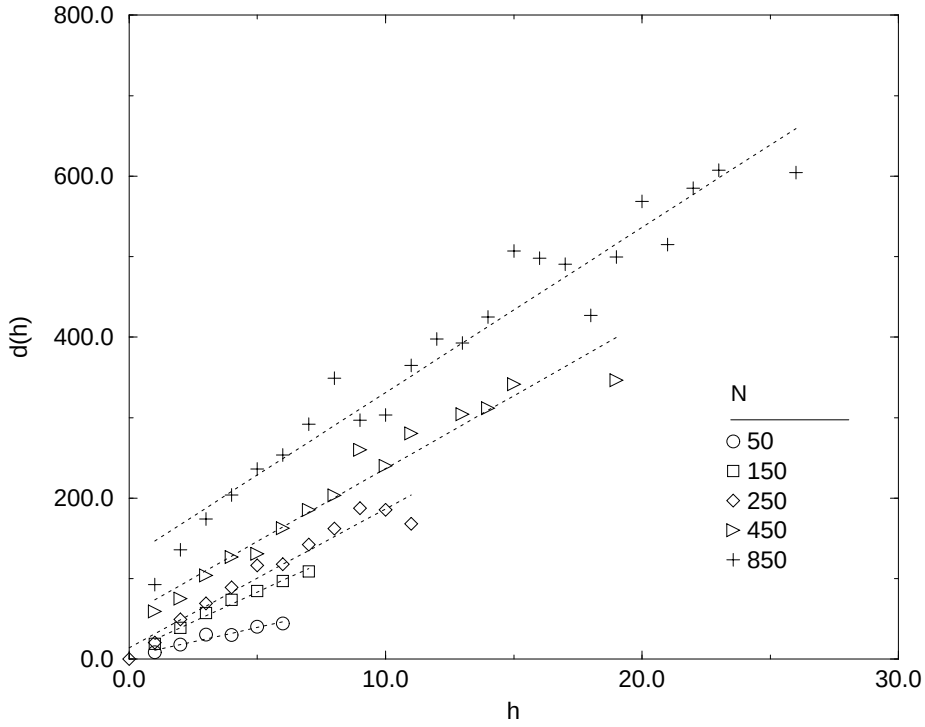

Figure 4. The average distance $d(h)$ between all pairs of structures which have an energy barrier of h. It can be seen that there is a strong correlation between h and $d(h)$. The dotted lines are best fit straight lines which have been added as a guide to the eye.

of disordered systems it is easy to look at distances between sets of states generated by a technique such as Monte Carlo simulation or simulated annealing, as was done in the SK model [21] and a protein folding simulation [22] for example, but it is not straightforward to define barrier heights. Alternative models assume the presence of a tree-like hierarchy of states and consider diffusion of a particle between these states [23–25]. In this case there is a clear definition of barrier heights and dynamics, but there is no notion of distance between states, or of what structure is actually represented by the states. In the RNA model studied here we have a clear definition of both distances and barriers between alternative states, and both are accessible numerically. It is therefore of interest to ask how the barriers and distances are correlated.

Figure 4 shows the function $d(h)$, defined as the mean distance between pairs of structures having a barrier h, after averaging over sequences and over pairs of structures for each sequence, i.e.

$$d(h) = \frac{\sum_{\text{Sequences}} \sum_{\text{A}} \sum_{\text{B}} d_{\text{AB}} \delta(h_{\text{AB}}, h)}{\sum_{\text{Sequences}} \sum_{\text{A}} \sum_{\text{B}} \delta(h_{\text{AB}}, h)} \tag{6}$$

where $\delta(h_{\text{AB}}, h) = 1$ if $h_{\text{AB}} = h$, and zero otherwise.

For any sequence length N it can be seen that $d(h)$ increases with h, indicating a clear correlation between distances and barriers. The dotted lines on the figure are best-fit straight lines shown as a guide to the eye. The relationship does seem to be roughly linear. Note that $d(h)$ is only defined over the range of h which actually occurs in the sequences of

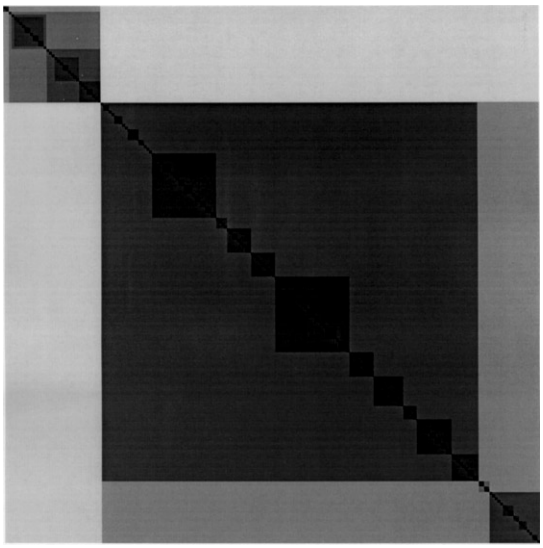

Figure 5. Matrix representation of the barriers between each structure in the set for a sequence of length 450. The lighter the shade of grey, the higher the barrier between two structures. The structures are arranged in order on each axis so that the clusters of similar barriers obtained from the SLC algorithm can be clearly seen. This matrix is exactly ultrametric.

each of the lengths studied. The points at high h values for each of the curves in figure 4 represent average distances between very unusual pairs of structures which happen to have much higher h and d values than normal. Only a small fraction of sequences will have structures separated by these high h values.

A further illustration of the correlation between d and h is shown in figures 5 and 6. For one typical sequence of length 450 the best estimate of the h matrix was obtained using the SLC method. The SLC method gives a natural ordering of states so that closely related states are consecutive. The h matrix is plotted in figure 5, with the states in the appropriate order, using a grey scale so that lighter shades of grey represent higher barriers. The result is a well-defined hierarchical system of clusters. Figure 6 shows the matrix of distances between the same set of clusters as in figure 5, and with the same ordering of structures. Many of the main features of the h matrix can also be seen in the distances, although the cluster pattern is much less clear. Figure 6 looks like figure 5 with 'noise' added.

In a previous study [16] clustering patterns were shown for the d matrix, but the h matrix was not then available. If the d matrix is clustered directly (as was done in [16]) then a slightly different pattern results because the ordering of structures which optimizes clustering of the d matrix is not precisely the same as that which optimizes the clustering of the h matrix. It is the clusters which emerge from the h matrix which are the more meaningful because they are precisely ultrametric. The ultrametric inequality (see section 3 and [20]) leads to the isosceles triangle property: that the two largest of the three h values for any triplet of states must be equal. It is clear that this must be true for the energy barriers as we have defined them, since one possible route from A to B is to go via C, and the barrier associated with this result is just $\max(h_{AC}, h_{CB})$. This is because the route returns to the ground-state energy on reaching the intermediate state C (or any other intermediate ground state). There is no reason why the d matrix should be precisely ultrametric, but since d and h values are quite strongly correlated, the d matrix still appears to be approximately

ultrametric. This confirms the results of [16], where triplets of distances were tested for the isosceles triangle property. It was found that the difference between the two largest sides was smaller than would be expected by chance, indicating an approximate degree of ultrametricity in the d values.

We wish to stress that even though the h matrix is ultrametric by definition, the hierarchical cluster pattern seen in figure 5 is non-trivial. The results of figures 3–5 taken together show that there is a broad range of barrier heights for any N, and that the mean barrier increases with N. A trivial example of ultrametricity would be obtained by setting all the barriers to be equal, in which case the inequality would be satisfied, but there would be no hierarchy of clusters. The distance matrix becomes trivially ultrametric in this way for most disordered systems at high temperature [16, 22]. It is also worth pointing out that whatever input matrix is used for the SLC method, the output matrix is always exactly ultrametric. However, this does not mean that the SLC method can create a hierarchical structure 'out of nothing'. If the input matrix has no hierarchical structure within it (e.g. a random matrix) the output matrix will be trivially ultrametric.

A further feature which is seen in the $d(h)$ measurements in figure 4 is that for any given h, $d(h)$ increases with N. In other words it is possible for ground-state structures of larger sequences to differ by a greater amount than shorter sequences without having any greater energy barrier between them. This was the observation that led us to consider the idea of effective length discussed below.

4.2. Pairing probability and effective sequence length

The probability that bases i and j are paired can be obtained from the partition function (equation (2)). Examination of the matrix p_{ij} indicates that certain positions in the sequence are much more variable in structure among all the possible equilibrium states than others. If a base i is in a structurally variable region there will be several bases j for which p_{ij} is non-zero, all of which are of the same order of magnitude. If a base is almost always paired with one particular partner, then one of the p_{ij}'s will be close to 1 whilst all the others are almost 0. We define the maximum pairing probability for each base i as

$$p_{\max}(i) = \max_{1 \leqslant j \leqslant N} (p_{ij})$$

and the integrated distribution of these probabilities $F(p_{\max})$, to be the fraction of bases for which $p_{\max}(i)$ is greater than or equal to $p_{\max}$. The function $F_(p_{\max})$, averaged over all sequences of length 250, is shown in figure 7 at various temperatures. The $T = 0$ curve gives important information about the properties of the ground states. First, this curve covers the whole of the range 0–1, which indicates that there are many bases with intermediate values of $p_{\max}$. This is because there are many degenerate ground states. The $T = 0$ curve intercepts the right-hand axis at approximately 0.18, indicating that there is a fraction 0.18 of bases which are always paired to the same other base in every ground-state configuration ($p_{\max} = 1$). The curve intercepts the left axis at about 0.92, showing that there is a fraction 0.08 of bases which are always unpaired in every ground state ($p_{\max} = 0$). The curves for the other temperatures are smooth, and we expect that they are continuous at the two boundaries, such that they do not intercept the axes. The curves shift toward lower $p_{\max}$ values at higher temperature since a greater number of alternative configurations become accessible to each base on the sequence. A quantity analogous to $F(p_{\max})$ was studied by Derrida *et al* [26] in a model of directed walks with random self-interactions. There it was shown that the curves behaved as a power law close to the limit. This behaviour was termed 'weak freezing'. An important difference between the directed walk model and the

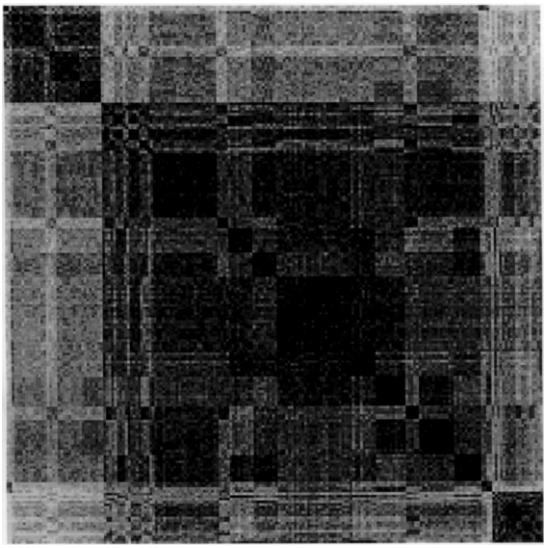

Figure 6. Matrix representation of the distances between the same set of structures as in figure 5, with the states arranged in the same order. A clear correlation between the distances and the barriers of figure 5 is visible.

RNA model studied here is that the former has a non-degenerate ground state whereas the latter has multiple degeneracy. This means that the $F(p_{\max})$ curve is non-trivial even in the limit $T = 0$ which we are principally interested in here.

The practical consequences of figure 7 are that there are certain base pairs which are present in all ground-state structures. We will call a base pair frozen if $p_{ij} \geqslant 0.99$ at $T = 0$. Frozen base pairs divide the sequence into mutually inaccessible regions because of the constraints which disallow pseudoknots. A single frozen base pair ij gives rise to two regions in which further base pairing may occur independently: the loop from $i + 1$ to $j - 1$, and the two ends combined (from 1 to $i - 1$ plus from $j + 1$ to N). We define the effective length N_{eff} of the sequence as the length of the largest region of the molecule in which further pairing may occur without disturbing the frozen pairs. The concept of effective length is illustrated in figure 8 for a sequence with $N = 250$. Structures B–D are three ground-state structures, and A is a structure containing only the frozen base pairs common to the ground states. Frozen pairs are marked with a thick line and the effective length of the sequence is the length of the largest loop in structure A. In B–D, different structures are seen in the free, unfrozen regions, but the frozen structure is the same in each case. This allows us to explain why the distance $d(h)$ between states with the same barrier height increases with the system size N. Separate regions created by these frozen pairs are independent of one another as far as determining barrier heights is concerned. When the system overcomes the energy barrier from ground states A to B, each region in A can be transformed independently into the new structure within the same region in B. The overall energy barrier between the two structures will just be the largest of the barriers from the independent regions, which will usually come from the region of the largest size N_{eff}. Hence we may expect that barrier heights will scale more closely with N_{eff} than with N. On the other hand, the distance between the structures is the sum of the distances for each region, which explains why $d(h)$ increases with N for any given h. Also the 'noise' seen in the distance matrix plot of figure 6 is caused by the variability in the structure in

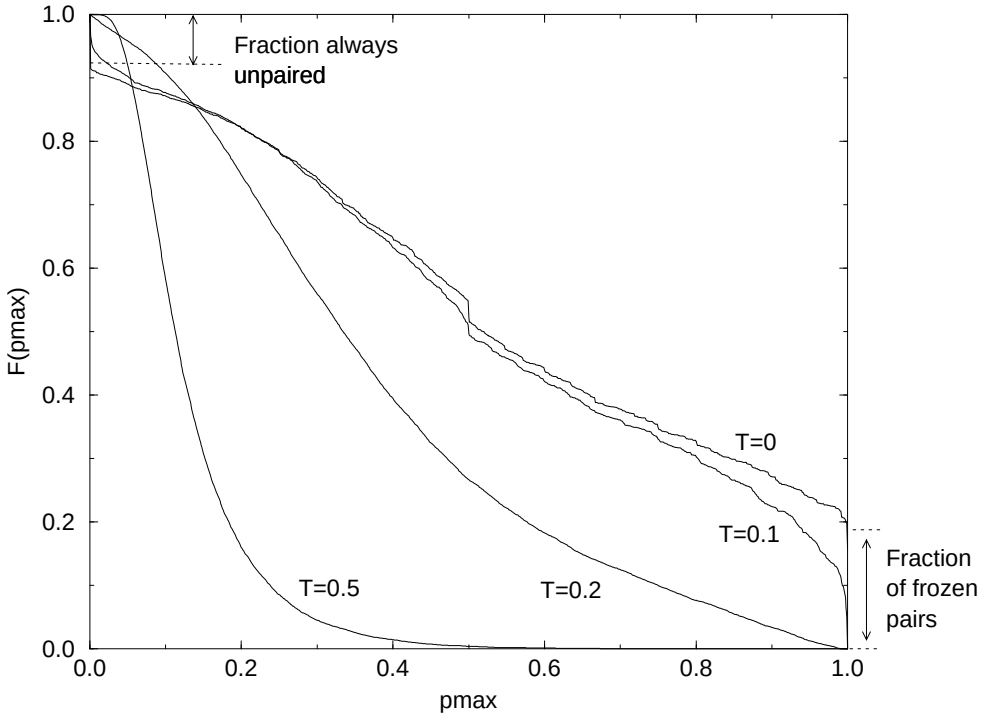

Figure 7. The integrated distribution $F(p_{\max})$ of the largest pairing probability for all bases from all of the sequences for length 250. The distributions are shown for temperatures from 0 to 0.5. At very low temperature about 20% of bases in all the structures are frozen. Also about 8% of bases are never paired. At higher temperatures any base can pair with many other bases with a fairly high probability, hence most $P_{\max}$ values are small.

the smaller regions which do not contribute to h.

We calculated the effective length of all structures for each of the sequences previously used for different values of N. Graphs of the average and the maximum barrier for each sequence against its effective length are shown in figures 9 and 10, plotted on a log–log scale. Each symbol represents a single sequence, with different symbols used for each N. For each N the effective length values vary greatly with sequence, but on average increase in proportion to N. We estimate that the mean of $N_{\text{eff}} \sim 0.30\ N$ for large N. For individual sequences the average and maximum barriers correlate closely with N_{eff} as expected. From these graphs we estimate $\langle h \rangle \sim N_{\text{eff}}^{0.47}$, and $h_{\max} \sim N_{\text{eff}}^{0.54}$.

We now have four estimates for the scaling exponent which we would expect to all be equal. Since $N_{\text{eff}} \sim N$, we would expect the scaling of barriers with N in section 4.1 to be the same as the scaling with N_{eff}. Also, if the highest level branch point in the clustering hierarchy splits the states into two groups of roughly equal size, then roughly half the elements of the h matrix will equal $h_{\max}$, hence we expect the exponent for $h_{\max}$ to be the same as that for $\langle h \rangle$. The differences between the four exponents measured therefore reflects uncertainty in the measurement. The four values are close to $\frac{1}{2}$ and are consistent with $\frac{1}{2}$ within the error range of the measurement, which is in the order of ± 0.05.

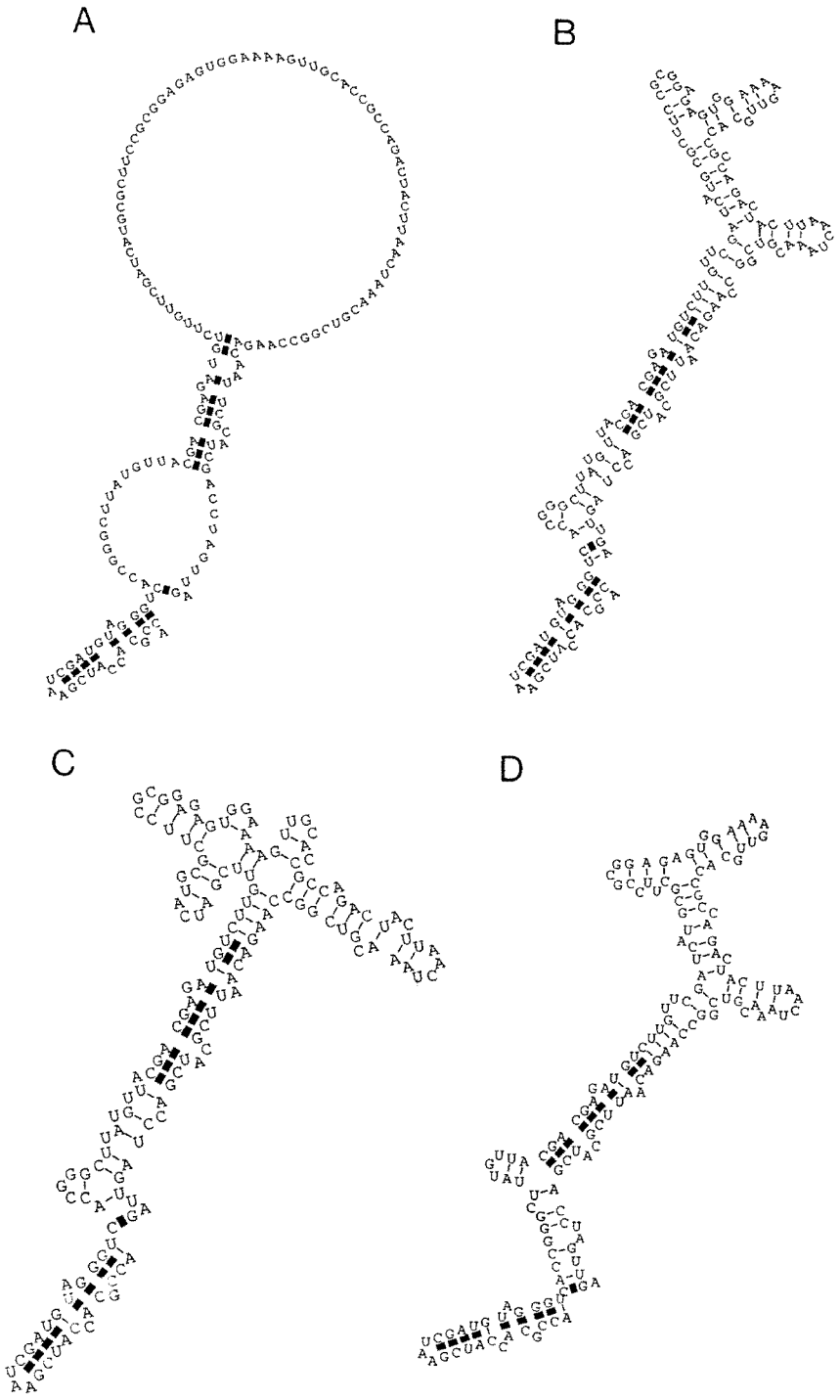

Figure 8. Illustration of frozen structure at low temperature in a sequence of length 250. Structure A contains only frozen pairs, the loops representing unconstrained regions. Structures B–D have the same frozen pairs and helices, but the unconstrained regions are paired differently, with the largest variation occurring in the largest unconstrained region. The effective length of this sequence is the length of the largest loop in A.

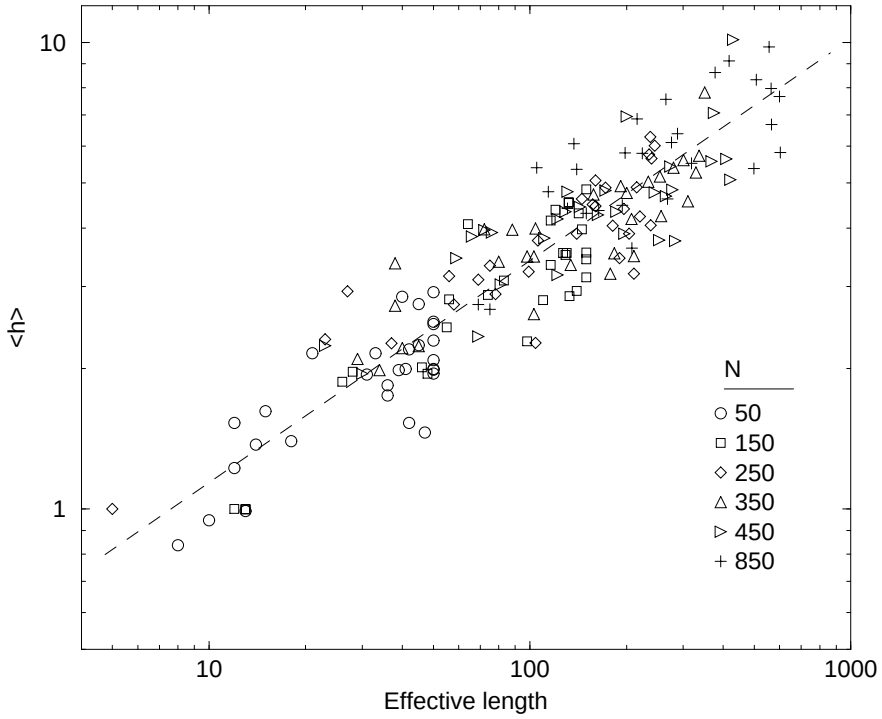

Figure 9. Log–log plot of the average barrier height against effective length. The symbols represent different sequence lengths as labelled. For each sequence length the range of effective length is large, and correlates closely with the average barrier measured for each sequence. The dotted line is a best fit of the data.

5. Discussion and conclusions

We have used a heuristic numerical procedure to estimate the energy barriers between ground states in a model of RNA secondary structure. The model is a simplified version of the real energy rules for RNA base pairing, but still displays disorder and frustration, necessary features for rugged energy landscapes.

The most widely studied system displaying these properties is the Sherrington–Kirkpatrick (SK) spin-glass model. The replica theory [17] predicts that the lowest energy configurations of the SK model are arranged in ultrametric clusters using a distance metric in configuration space. Numerical evidence appears to support this theory [21]. There is also evidence for this in other complex optimization problems, such as the travelling salesman problem [18], the graph colouring problem [27] and the graph bipartitioning problem [28], as well as in the configuration space of proteins [22]. It has been conjectured that ultrametricity may lie at the heart of many problems of this sort, where there are many possible competing and nearly equally good solutions. This work shows explicitly that there are hierarchical clusters of states in the RNA folding model which we considered. The clusters based on barrier heights are exactly ultrametric, and those based on distances are approximately ultrametric. None of the previous studies was able to examine both distances and barrier heights, as we have done here.

Mackenzie and Young [29, 30] have used numerical simulations of relaxation in the SK model to estimate the equilibrium relaxation times and hence energy barriers between

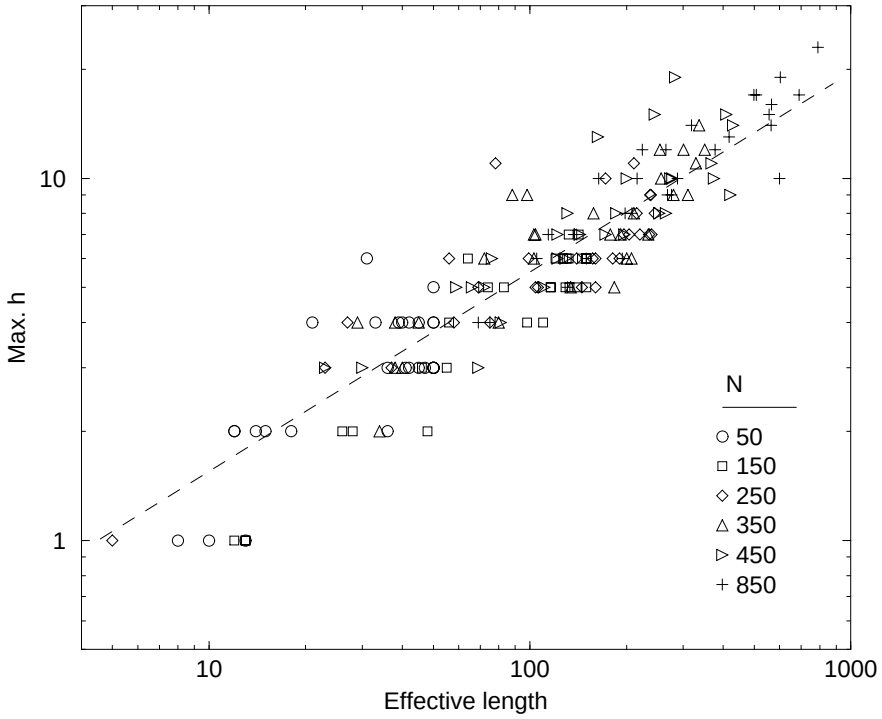

Figure 10. Log–log plot of the maximum barrier height against effective length. The symbols represent different sequence lengths as labelled. The dotted line is a best fit of the data. The fit is poorer at the lower end of the line because the barriers are integers.

ground states. They obtain two distinct scaling laws for the barriers in the model, $N^{1/4}$ which they call 'non-ergodic', and $N^{1/2}$ or 'ergodic' barriers. The ergodic barriers are observed to exist between a ground state and its inverse obtained by flipping all the spins, and non-ergodic barriers are found between states with a larger overlap. The symmetry of the SK model under reversal of all spins does not have an equivalent in the RNA model, although the correlation between d and h which we see shows that structures which are very dissimilar to each other tend to have the largest barriers. We have shown here that the barriers between ground states in the RNA model are consistent with an $N^{1/2}$ scaling law (within a fairly large error margin).

A link between RNA folding and the SK model has previously been made by Fernandez [31] and Fernandez and Shakhnovich [32] who simulate the folding of the molecule occurring during chain synthesis. Rearrangements of secondary structure occur principally close to the growing end of the chain. They observed barrier heights scaling as $N^{1/4}$, and likened this to the 'non-ergodic' barriers in the SK model [29]. Here we are looking at alternative structures of chains of fixed lengths, and the barriers that we measure should correspond to the largest barriers in the energy landscape. Our results give a scaling law close to $N^{1/2}$, which is the same as that for the 'ergodic' barriers in the SK model, although there is no strong argument why these models should scale in the same way. The difference between our results and those of Fernandez indicates that the molecule only crosses relatively small barriers during the folding process. In other words, the order of formation of different helices is important for determining the final structure, because the molecule does not

necessarily manage to completely rearrange its structure to find the most thermodynamically stable configuration.

A major advantage of the RNA model as an example for the study of disordered systems is that the partition function can be calculated exactly in polynomial time, allowing relatively long sequences to be considered. Previous exact calculations of the partition function for the SK model were limited to $N = 20$ spins [33]. Another major advantage of the model is that we were able to obtain a direct estimate of the energy barriers between ground states without using Monte Carlo simulation. Our algorithm for estimating the direct barriers is similar in spirit to one used by Morgenstern [34] to estimate the energy barriers in a spin-glass model with short-range interactions (the '$\pm J$ model'). In that paper the barrier height between ground states is estimated by following different random paths through configuration space. Spins are flipped and then kept fixed until the final state is reached, while allowing any remaining non-fixed spins to rearrange themselves to the new situation of each configuration along the route.

In a system that is non-ergodic on physical timescales the phase point is effectively confined in one subregion or component of phase space. Theoretical treatments of such systems should compute thermal averages over one component at a time (Palmer [35]). The idea of broken ergodicity and the division of phase space into components has been developed more recently by Stein and Newman [36, 37]. We have also begun to study finite-temperature dynamics in our RNA model. We have run Monte Carlo simulations in which two initially identical structures are allowed to relax simultaneously at low temperature using different sets of random numbers, and the distance between the structures is calculated as a function of time. We are able to calculate the equilibrium average distance between structures from the partition function, and also a restricted average obtained between all pairs of structures within a barrier of 1. Preliminary results show that in many cases the distance obtained by simulation agrees with the restricted average with remarkable accuracy.

The work in this paper has been presented from the point of view of statistical physics, as an example of an interesting disordered system. We conclude by briefly discussing the relevance of energy barriers to the properties of real RNA. First it should be remembered that the energy parameters used for realistic structure prediction are much more complicated than those used here, although the algorithm for the partition function is still $O(N^3)$ [3]. Both energy and entropy parameters are included, so that the Boltzmann weights of different structures are temperature dependent. Interactions in helices are dependent on stacking between neighbouring pairs, hence the pairs tend to occur in long helical regions, whereas in the simplified model used here the pairs are independent of each other. With the more realistic rules ground states will not be exactly degenerate, but there should still be many low-energy states of approximately equal energy. We have not developed algorithms for barrier-height determination using the realistic energy parameters, since the details would be much more complex, but we expect that the results would be very similar. In any case it is clear that there is a large possibility of real RNA molecules becoming trapped in metastable states during the folding process, since stacking free energies in helices can be large compared with the thermal energy kT. This means that finding the minimum free-energy structure may not be the most appropriate way of predicting the structure formed in nature. We have argued that barrier heights become large enough to affect the folding of real RNA for molecules longer than about 100 nucleotides [10], and hence that the kinetics of the folding process is important in many natural RNAs. Examples of RNA folding studies from the biochemical literature are given in [9] and [10], and we will not discuss these further here.

The thermodynamics of RNA secondary structure is important for understanding the

folding behaviour and the function of natural RNAs. The model which we have studied allows the clustering of states and the energy barriers between states to be investigated in a much more direct way than has been done with most other disordered systems in statistical physics. We expect that RNA folding models will continue to be of interest in both the biochemistry and physics communities.

References

[1] Nussinov R and Jacobson A B 1980 *Proc. Natl Acad. Sci., USA* **77** 6309
[2] Waterman M S 1986 *Adv. Appl. Math.* **7** 455
[3] McCaskill J S 1990 *Biopolymers* **29** 1105
[4] Zuker M, Jaeger J A and Turner D H 1991 *Nucleic Acids Res.* **19** 2707
[5] Hofacker I L, Fontana W, Stadler P F, Bonhoeffer L S, Tacker M and Schuster P 1994 *Monatshefte Chem.* **125** 167
[6] Freier S M, Kierzek R, Jaeger J A, Sugimoto N, Caruthers M H, Nielson T and Turner D H 1986 *Proc. Natl Acad. Sci., USA* **83** 9373
[7] Higgs P G 1993 *J. Physique* I **3** 43
[8] Higgs P G 1995 *J. Chem. Soc. (Faraday Trans.)* **91** 2531
[9] Higgs P G and Morgan S R 1995 *Lect. Notes Artif. Intell.* **929** 852
[10] Morgan S R and Higgs P G 1996 *J. Chem. Phys.* **105** 7152
[11] Shakhnovitch E and Gutin A 1990 *J. Chem. Phys.* **93** 5967
[12] Dill K A, Bromberg S, Yue K, Fiebig K M, Yee D P, Thomas P D and Chan H S 1995 *Protein Sci.* **4** 561
[13] Thirumalai D and Woodson S A 1996 *Acc. Chem. Res.* **35** 627
[14] Panchenko A R, Luthey-schulten Z and Wolynes P G 1996 *Proc. Natl Acad. Sci., USA* **93** 2008
[15] Fink T M and Ball R C 1997 *Physica* **107D** 199
[16] Higgs P G 1996 *Phys. Rev. Lett.* **76** 704
[17] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[18] Kirkpatrick S and Toulouse G 1985 *J. Physique* **46** 1277
[19] Sneath P H A and Sokal R R 1973 *Numerical Taxonomy* (San Francisco, CA: Freeman)
[20] Rammal R, Toulouse G and Virasoro M A 1986 *Rev. Mod. Phys.* **58** 765
[21] Bhatt R N and Young A P 1989 *J. Phys.: Condens. Matter* **1** 2977
[22] Velikson B, Bascle J, Garel T and Orland H 1993 *Macromolecules* **26** 4791
[23] Teitel S 1988 *Phys. Rev. Lett.* **60** 1154
[24] Bachas C P and Huberman B A 1987 *J. Phys. A: Math. Gen.* **20** 4995
[25] Knapp E W 1989 *Phys. Rev.* B **39** 4834
[26] Derrida B, Griffiths R B and Higgs P G 1992 *Europhys. Lett.* **18** 361
[27] Bacci S and Parga N 1989 *J. Phys. A: Math. Gen.* **22** 3023
[28] Banavar J R, Sherrington D and Sourlas N 1987 *J. Phys. A: Math. Gen.* **20** L1
[29] Mackenzie N D and Young A P 1982 *Phys. Rev. Lett.* **49** 301
[30] Mackenzie N D and Young A P 1983 *J. Phys. C: Solid State Phys.* **16** 5321
[31] Fernandez A 1990 *Phys. Rev. Lett.* **64** 2328
[32] Fernandez A and Shakhnovich E I 1990 *Phys. Rev.* A **42** 3657
[33] Young A P and Kirkpatrick S 1982 *Phys. Rev.* B **25** 440
[34] Morgenstern I 1983 *Lect. Notes Phys.* **192** 305
[35] Palmer R G 1983 *Lect. Notes Phys.* **192** 234
[36] Stein D L and Newman 1995 *Phys. Rev.* E **51** 5228
[37] Newman C M and Stein D L 1996 *J. Stat. Phys.* **82** 1113