JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 277–292

Metrics on RNA Secondary Structures

VINCENT MOULTON,[1] MICHAEL ZUKER,[2] MICHAEL STEEL,[3] ROBIN POINTON,[4]
and DAVID PENNY[5]

ABSTRACT

Many different programs have been developed for the prediction of the secondary structure of an RNA sequence. Some of these programs generate an ensemble of structures, all of which have free energy close to that of the optimal structure, making it important to be able to quantify how similar these different structures are. To deal with this problem, we define a new class of metrics, the mountain metrics, on the set of RNA secondary structures of a fixed length. We compare properties of these metrics with other well known metrics on RNA secondary structures. We also study some global and local properties of these metrics.

Key words: RNA secondary structure, pseudoknot, suboptimal secondary structures, metric, distance, similarity.

1. INTRODUCTION

A CHALLENGING PROBLEM IN BIOINFORMATICS is the prediction of the *tertiary* (or three dimensional) structure of an RNA molecule (and also, of course, that of a protein molecule) using only its *primary structure* (i.e., its sequence of nucleotides) (see Zuker (1986) for a comprehensive review). Although accurately predicting tertiary structure is currently very difficult, significant progress has been made in predicting the *secondary structure* of RNA molecules (that is, the structure that can be represented as a planar graph) which should, in turn, lead to deep insights into their tertiary structure. There are efficient algorithms for associating a secondary structure to an RNA primary structure which work by minimizing free energy (see Zuker (1989b), Zuker *et al.* (1984) for example). However, due to difficulties in measuring bond energies precisely and also to uncertainties inherent in the models used for folding, searching for a unique structure is an ill-conditioned problem (Zuker, 1986). Thus, it is important not to predict just a single structure when folding a sequence but to predict an *ensemble* or collection of structures, all of whose free energies are close to that of the folding with minimal free energy. These ensembles are usually referred to as collections of *suboptimal structures*, and various algorithms have been designed for their prediction (see McCaskill (1990), Zuker (1989a) for two of the most popular approaches).

In general, given a sequence of fixed length, an ensemble may contain thousands of distinct structures (Uhlenbeck *et al.*, 1995; Zuker (1989b, p.179)). This lack of close correspondence between primary and secondary structure for RNA is expected to have been especially problematic during the origin of life before

[1] FMI (Physics and Mathematics Department), Mid-Sweden University, S 851-70 Sundsvall, Sweden.
[2] Institute for Biomedical Computing, Washington University, Campus Box 8036, 700 S. Euclid Avenue, St. Louis, MO 63110.
[3] University of Canterbury, Private Bag 4800, Christchurch, New Zealand.
[4] Computer Science Department, Massey University, PO Box 11-222, Palmerston North, New Zealand.
[5] Institute of Molecular BioSciences, Massey University, PO Box 11-222, Palmerston North, New Zealand.

proteins evolved and helped stabilize secondary structures. It is necessary to be able to compare secondary structures in an ensemble in order to find conditions of temperature and nucleotide composition that may favor a closer relationship between sequence and structure. This problem is of particular importance to biologists, and is related to the study of *landscapes*, on which much progress has been made recently (see Fontana *et al.* (1993), Boenhoeffer *et al.* (1993) for example).

Here we deal with one aspect of this problem, that of defining *similarity measures* or, more specifically, *metrics* on ensembles (see Dress *et al.* (1996) for an abstract study of similarity theory). In particular, after reviewing two well known classes of secondary structure metrics, the *base-pair* and *tree metrics*, we define a new class, the *mountain metrics*, which have the advantage that they are easy to compute, and to use. We then study some "*global*" and "*local*" properties of these metrics, where, by global we mean properties of the metric over a diverse range of structures, while local refers to those properties of structures that are similar to a selected structure.

In particular, we compute the diameter of certain mountain metrics, and then simulate their distributions, a technique that has proven useful in the study of tree metrics (Steel *et al.*, 1993). We compare these distributions with those of the base-pair and tree metrics. Understanding the way in which the metrics behave locally can be important: *mfold* (MFOLD) uses a base-pair metric in order to filter out locally similar structures from ensembles. Here we study a particular local property of the mountain metrics: we give recursive formulas for computing the number of secondary structures that are within a certain fixed mountain metric distance of a given fixed structure.

2. SECONDARY STRUCTURES

A *secondary structure* S for an RNA sequence of length n, is a simple graph, that is, a graph with only one edge between each pair of vertices, with vertex set $[n] := \{1, \dots, n\}$, whose edge set consists of the edges $\{\{i, i+1\} \mid 1 \leq i \leq n-1\}$, together with a further collection of edges B_S, such that if $\{i, j\}, \{k, l\} \in B_S$ with $i < j$ and $k < l$, then

(i) $i = k$ if and only if $j = l$, and

(ii) $k \leq j$ implies that $i < k < l < j$.

An edge $\{i, j\}$ contained in B_S is called a *base pair*, and – in case $i < j$ – we also denote it by $i \cdot j$. Those vertices not contained in a base pair are called *unpaired*. Condition (i) implies that each vertex (i.e., nucleotide) is allowed to belong to at most one base pair. Condition (ii) excludes the formation of *pseudoknots*. See Figure 1 (a) for an example of a secondary structure (where base pairs are represented by dotted lines).

Note that in this definition we do not allow base pairs of the form $i \cdot i + 1$. In some cases one can also define secondary structures so that they do not have base pairs of the form $i \cdot i + k$ for all $k \in \{1, \dots, r\}$ for a fixed r, to make the model more realistic (Zuker, 1986). In Section 5 we consider such structures.

Let $\mathcal{S}_n$ denote the set of all possible secondary structures on an RNA sequence $[n]$. On reflection, it is not hard to see that the cardinality of $\mathcal{S}_n$ grows extremely rapidly with n (for calculations of this cardinality see Stein *et al.* (1978)), thus making it important to find ways for comparing secondary structures efficiently.

There are many different ways to represent secondary structures (see Figure 1), however, three representations interest us here:

Bracket Representation: This is a compact representation which is obtained for each element $S \in \mathcal{S}_n$, by creating a sequence of length n consisting of parentheses and dots, through replacing each base pair $i \cdot j \in B_S$ in the sequence $[n]$ by a "(" and a ")" in the ith and jth positions, respectively, and replacing those $i \in [n]$ which are unpaired by a "·" (Hofacker *et al.*, 1994, p.171) (see Figure 1 (c)).

Mountain Representation: The bracket notation leads naturally to the *mountain representation* for the secondary structure, which has been used for the "graphical" comparison of secondary structures (Hogeweg *et al.*, 1984; Konings *et al.*, 1989). Basically, each base pair is represented by a horizontal line over the primary sequence at a height that is dictated by its position in the sequence (see Figure 1 (d)).

Tree Representation: There are also various ways of representing secondary structures as trees (Hofacker *et al.*, 1994; Shapiro, 1988; Schmitt *et al.*, 1994). These representations differ in that some compress substructures into single labeled vertices. A tree representation which is directly related to the bracket notation is illustrated in Figure 1(e). Here the secondary structure is represented by an ordered rooted tree:

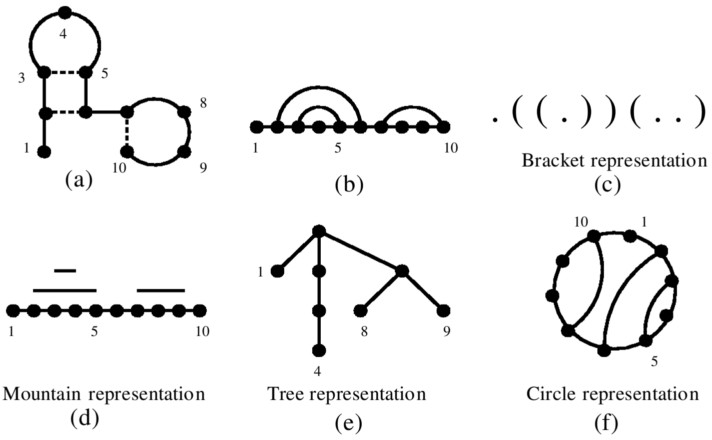

FIG. 1. Representations of RNA secondary structure.

The root does not correspond to a part of the RNA secondary structure, internal nodes correspond to base pairs, and leaves correspond to unpaired vertices (as labeled) (see Fontana *et al.* (1993, p.1397) for a detailed description).

3. SECONDARY STRUCTURE METRICS

In this section, we recall two well known classes of metrics defined on secondary structures, namely, base-pair and tree metrics, which are based on the bracket and tree representations of secondary structures respectively, and a new class of metrics called the mountain metrics, which are based on the mountain representation of secondary structures.

3.1. Base pair metrics

One of the simplest metrics that one can define on $\mathcal{S}_n$ is to set the distance between a pair $S_1, S_2 \in \mathcal{S}_n$ equal to the cardinality of the symmetric difference of B_{S_1} and B_{S_2}. However, this metric is clearly very coarse, in that it does not capture much of the secondary structure information.

A more refined metric defined using base pairs was introduced by Zuker (1989b, p.180). For $S_1, S_2 \in \mathcal{S}_n$, define the distance between two base pairs $i \cdot j \in B_{S_1}$ and $i' \cdot j' \in B_{S_2}$ to be

$$d_0(i \cdot j, i' \cdot j') := \max\{|i - i'|, |j - j'|\}.$$

Define $d_Z(S_1, S_2)$ to be the smallest $d \in \mathbb{N}_0$ such that for every base pair $b_1 \in B_{S_1}$ there is a base pair in B_{S_2} within distance d_0 at most d of b_1, and (to ensure symmetry) for every base pair $b_2 \in B_{S_2}$ there is a base pair in B_{S_1} within distance d_0 at most d of b_2. Note that d_Z is defined on the set $\mathcal{S}_n - \{S_o^n\}$ for $n \geq 3$, where S_o^n denotes the secondary structure of length n with no base pairings. It can easily be seen that d_Z is a metric. For example, d_0 induces a Hausdorff metric on the power set of $\mathcal{B} := \{i \cdot j \mid 1 \leq i < j \leq n\}$, and we can consider any secondary structure as a subset of the set $\mathcal{B}$. The metric d_Z was developed to filter out similar suboptimal structures generated by the original *mfold* program (Zuker, 1989b, 1989a; MFOLD).

Note that d_Z operates by finding the *maximal distance* between any two base pairs in any pair of secondary structures. However, this can pose problems as we see in the following example: Define $S_1 :=$ $\cdots(\cdot\cdot)\cdots\cdots\cdots\cdots((((\cdot\cdot\cdot))))$ and $S_2 := \cdots\cdots\cdots\cdots\cdots\cdots((((\cdot\cdot\cdot))))$ both of which are in $\mathcal{S}_{30}$. Then $d_Z(S_1, S_2) = 20$, even though S_1 and S_2 only differ in one base pairing. Thus d_Z can sometimes be insensitive to local changes in structure.

Motivated by this problem, a variant of d_Z was defined by Zuker *et al.* (1991, p. 2708), which is used in the latest version of the program *mfold* (MFOLD): If $S_1, S_2 \in \mathcal{S}_n - S_o^n$, then S_1 is said to be within distance d of S_2, if for every base pair $b_1 \in B_{S_1}$, with the possible exception of d base pairs, there is a base pair $b_2 \in B_{S_2}$ such that the distance between b_1 and b_2 is at most d. Symmetrizing this measure, we define $d_P(S_1, S_2)$ to be the maximum of the two possible distances between S_1 and S_2. In the above example, we see that $d_P(S_1, S_2) = 1$. However, the distance function d_P is not a metric as it fails to satisfy the triangle inequality. For example, if $S_1 = ((((\cdot\cdot(((((\cdot\cdot\cdot\cdot\cdot\cdot)))))\cdot\cdot))))$, $S_2 = \cdot\cdot\cdot\cdot\cdot\cdot\cdot((((\cdot\cdot\cdot\cdot\cdot\cdot))))\cdot((\cdot\cdot\cdot))$, and $S_3 = \cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot(((\cdot\cdot\cdot\cdot\cdot)))$ then $d_P(S_1, S_2) = 4$, $d_P(S_2, S_3) = 4$, but $d_P(S_1, S_3) = 9$.

3.2. Mountain metrics

In this section we define a new class of metrics which is based on the mountain representation of secondary structures (see Section 2 and Hogeweg *et al.* (1984)). These metrics have the advantage that they can be computed very quickly, and have properties which make them easy to handle theoretically. For each $S \in \mathcal{S}_n$ define a vector f_S in $(\mathbb{N}_0)^n$, by setting its ith coordinate $f_S(i)$ equal to the number of "(" brackets minus the number of ")" brackets encountered when traversing the bracket notation from the first position up to, and including, the i-th position so that $f_S = (f_S(1), f_S(2), \dots, f_S(n))$. It is now easy to check that the l_p-norm induces a metric d_M^p on $\mathcal{S}_n$ in the usual way, viz:

$$d_M^p(S_1, S_2) := \|f_{S_1} - f_{S_2}\|_p := (\Sigma_{i=1}^n |f_{S_1}(i) - f_{S_2}(i)|^p)^{\frac{1}{p}},$$

for all $S_1, S_2 \in \mathcal{S}_n$.

Note that the metric d_M^p weights base pairings differently: for example, when $p = 1$ if $S_1 := (\cdot\cdot\cdot\cdot\cdot\cdot)$ and $S_2 := \cdot\cdot(\cdot\cdot)\cdot\cdot$, then $d_M^1(S_1, S_o^8) = 7$, and $d_M^1(S_2, S_o^8) = 3$. Hence, it may be preferable to scale d_M^1 as follows. If $S \in \mathcal{S}_n$ and $k \in [n]$, then set

$$w_S(k) := \begin{cases} \frac{1}{(l-k)} & \text{if } k \cdot l \in B_S \\ \frac{-1}{(k-l)} & \text{if } l \cdot k \in B_S \\ 0 & \text{otherwise,} \end{cases}$$

$f_S'(i) := \Sigma_{k=1}^i w_S(k)$, and $d_M(S_1, S_2) := \|f_{S_1}' - f_{S_2}'\|_1$. Note that if $S \in \mathcal{S}_n$ has a single base pair in any position then $d_M(S, S_o^n) = 1$ (as with d_T, where $d_T(S, S_o^n) = 2$).

Note that it is straightforward to extend the definition of the mountain metrics to the set of *pseudoknots* on n bases $\mathcal{P}_n$, that is, graphs for which Condition (ii) in the above definition of secondary structure is relaxed to the condition (ii)$'$ $i \cdot j, k \cdot l \in B_S$ implies $j \neq k$ (Dam *et al.*, 1992). The algorithms used in *mfold* and VIENNA to predict secondary structures depend heavily on the exclusion of pseudoknots (Zuker, 1989b; Hofacker *et al.*, 1994). The definition of tree metrics is dependent on the tree representation of secondary structure, which cannot be used for pseudoknots. However, the metrics d_Z, d_P, and d_M^p can all be easily extended to $\mathcal{P}_n$.

3.3. Tree metrics

In Section 2, we saw a way to represent secondary structures by trees. There are various metrics defined on trees (see Steel *et al.* (1993) for example), however, in the setting of secondary structures the most popular class of metrics is defined as follows. Any (labeled) tree T_1 can be transformed into any other tree T_2 via a series of *editing operations*, that is, a sequence of deletions, insertions or relabelings of edges and vertices. By assigning a cost to each of these operations, we define a *tree metric* by setting the distance between T_1 and T_2 equal to the smallest sum of the costs along all editing paths between T_1 and T_2 (note that the distance between two secondary structures of different lengths can be defined using this measure (Shapiro, 1988)—here we restrict to the case where the lengths are equal, although the ability to compare structures of different lengths has proven useful in other applications (see, e.g., Collins *et al.*, 1999)). One of the great advantages of this approach is that by varying the tree representation and editing costs one has the flexibility of tailoring metrics to either global or local analysis. For example, the tree representation of secondary structure given in Figure 1(e) is very fine, making it better for local analysis. However, as we have seen, there are coarser tree representations of secondary structures (Hofacker *et al.*, 1994; Shapiro, 1988), which (with appropriate editing costs) are better suited to global analysis.

Here we use the tree representation described in Section 2, which is one of the representations used in the VIENNA RNA folding package (Hofacker *et al.*, 1994; VIENNA) and assign the following editing costs that are used by VIENNA: the deletion or insertion of an unpaired base, or the exchange of an unpaired base for a base pair has cost one, and the deletion or insertion of a base pair has cost two.[1] We denote the resulting metric on $\mathcal{S}_n$ by d_T. Note that $d_T(S_1, S_2) \geq 2$, for any distinct $S_1, S_2 \in \mathcal{S}_n$.

Various algorithms have been developed for the computation of tree metrics defined via editing costs (see Shapiro (1988), Sankoff *et al.* (1983) for example), the most efficient being a dynamical programming algorithm due to Shapiro and Zhang (Shapiro *et al.*, 1990). This algorithm is currently used in the VIENNA package (Hofacker *et al.*, 1994). Tree editing can be regarded as a generalization of *sequence alignment*,[2] which can also be used to define metrics on $\mathcal{S}_n$ (see, e.g., Konings *et al.* (1989)): Representing elements in $\mathcal{S}_n$ in bracket notation, that is, as strings in the symbols (,) and $\cdot$, one can define costs for insertions and deletions and thereby obtain the alignment distance between any pair of elements of $\mathcal{S}_n$ in the usual way, e.g., see Waterman (1995, Chapter 9). However, this has the disadvantage that in resulting alignments base pairings will not necessarily be matched up correctly (for example two ('s may be aligned, one from each string, however their corresponding)'s may not be).

A problem with tree metrics in the setting of suboptimal structures is that they require a complicated algorithm for their computation which, as a consequence, makes them difficult to analyze formally, and time consuming to compute for large numbers of structures (more on this last problem later).

4. A GLOBAL PROPERTY OF SECONDARY STRUCTURE METRICS

In this section we compute the diameter, a global property, of some of the metrics that we have defined in this paper. For a metric d on a finite set X denote its *diameter* by $diam(X, d)$, that is, the maximum value of $d(x, y)$ for all $x, y \in X$. Let $\mathcal{S}'_n$ denote the space of secondary structures where base pairs of the form ( ) are allowed (so, in particular, we have $\mathcal{S}_n \subseteq \mathcal{S}'_n$). Also, define S^n_* to be the secondary structure ((, ... , ((··)), ... ,)) and ((, ... , ((·)), ... ,)) of length n, where n is even and odd respectively.

Theorem 4.1. *The following equalities hold:*

(i) $diam(\mathcal{S}_n - \{S^n_o\}, d_Z) = n - 3$, *for* $n \geq 6$.

(ii) $diam(\mathcal{S}'_n, d_T) = 2(n - 1)$, *for* $n \geq 3$.

(iii)

$$diam(\mathcal{S}_n, d^p_M) = \begin{cases} (2\sum_{m=1}^{k-1} m^p + (k-1)^p)^{\frac{1}{p}} & \textit{if } n = 2k \\ (2\sum_{m=1}^{k} m^p)^{\frac{1}{p}} & \textit{if } n = 2k + 1, \end{cases}$$

for all $k \geq 1$.

(iv) $diam(\mathcal{S}_n, d^\infty_M) = k - 1$, *for* $n = 2k, 2k + 1$, *and all* $k \geq 1$.

Proof. (i) Consider the structures $S_1 := (\cdot)\cdot\cdot, \ldots, \cdot\cdot$ and $S_2 := \cdot\cdot, \ldots, \cdot\cdot(\cdot)$ in $\mathcal{S}_n - \{S^n_o\}$. Then $d_Z(S_1, S_2)$ is equal to $n - 3$. Moreover, it is clearly not possible to place two base pairs in a pair of secondary structures further apart than in this example.

(ii) As mentioned in the footnote in Section 3.3, the metric d_T is defined on the set $\mathcal{S}'_n$. Note that since any structure in $\mathcal{S}'_n$ can be edited to give a structure in $\mathcal{S}'_{n-1}$ and visa versa, with edit cost 1, we immediately see that

$$diam(\mathcal{S}'_{n+1}, d_T) \leq diam(\mathcal{S}'_n, d_T) + 2,$$

from which it immediately follows (using a simple induction argument on n) that $diam(\mathcal{S}'_n, d_T) \leq 2(n-1)$.

We now claim that this upper bound is, in fact, sharp. Set $n = 2k + 1$, for $k \geq 1$. We will show that $d_T(S^n_*, S^n_o) \geq 4k$, thus proving the claim for n odd. Note that on any edit path from S^n_* to S^n_o we must

[1] The VIENNA package allows base pairs of the form "()" in the bracket notation, which we have excluded in our definition of secondary structure.

[2] It was recently pointed out to us in a private communication from R. Giegerich that the tree edit distance is in some sense the incorrect generalization of string edit distance to trees. More details will appear on this elsewhere.

remove, at some stage, the k pairs of brackets from S_*^n. There are two possible ways in which to remove these brackets:

(a) Remove a pair with cost 2, in which case we must add two unpaired bases at some other point in the path with cost at least 2, so as to restore the structure to its original length.

(b) Exchange a pair of the form ( ) with an unpaired base, cost 1, in which case we must add a base at some other point in the edit path which has cost at least 1. However, for this case to arise, we must, at some stage, remove the middle base in S_*^n with cost at least 1, which, in turn, means that we must add in a base with cost at least 1 at some point in the edit path.

In either case (a) or (b) the total cost of removing the base pair (and then restoring the structure to its original length) is at least 4, so that $d_T(S_n(F), S_o^n) \geq 4k = 2(n-1)$ as required.

The case where $n = 2k$ for $k \geq 1$, i.e., n is even, is similar except that we have a pair of brackets ( ) in the middle of S_*^n, which can be removed, and a base added at some stage in the edit path, with cost 2. However, all other base pairs must be removed with cost at least 4 as for n odd, so that $d_T(S_n(F), S_o^n) \geq 4(k-1) + 2 = 2(n-1)$.

(iii) First, note that for $i \in [n]$ and any $S \in \mathcal{S}_n$, we have $f_{S_o^n}(i) \leq f_S(i) \leq f_{S_*^n}(i)$. Hence it follows that $d_M^p(S_1, S_2) \leq d_M^p(S_o^n, S_*^n)$, for all $p = 1, \ldots, \infty$. Thus, (iii) follows from a routine calculation of $d_M^p(S_o^n, S_*^n)$. ■

Corollary 4.2. *We have* $diam(\mathcal{S}_n, d_T) = 2(n-1)$ *for n odd, and* $2(n-2) \leq diam(\mathcal{S}_n, d_T) \leq 2(n-1)$ *for n even,* $n \geq 3$.

Proof. The case where n is odd follows directly from part (ii) of the theorem. In the case where n is even it can be seen, using similar arguments to those presented in the proof of (ii), that $d_T(S_*^n, S_o^n) = 2(n-2)$, which completes the proof. ■

5. LOCAL PROPERTIES OF THE MOUNTAIN METRICS

In this section we give recursive formulae for computing the number of secondary structures that are exactly some fixed mountain metric distance from some fixed structure. This gives us local information on the mountain metrics.

Given a secondary structure $S \in \mathcal{S}_n$, we can associate to it the sequence $x_0, x_1, \ldots, x_n$, defined by $x_0 := 0$, and $x_i := f_S(i)$ for each $1 \leq i \leq n$, arising from the mountain representation of S (see Section 2). In this way it is straight forward to see that secondary structures correspond to sequences $x_0, x_1, x_2, \ldots, x_n$ which satisify:

(1a) $x_i \in \{0, 1, 2, \ldots\}$ for all $i \geq 0$;

(1b) $x_0 = x_n = 0$;

(2) if $x_i > x_{i+1}$, then $x_{i-1} \geq x_i$;

(3) $|x_i - x_{i-1}| \leq 1$ for all $i > 0$.

Note that Condition (2) is equivalent to the requirement that adjacent positions are not paired, and the additional value $x_0 = 0$ excludes certain forbidden configurations. In addition, note that the Conditions (1 a,b) and (3) imply that $x_i \leq \min\{i, n-i\}$ for all $i \geq 0$.

Let $x'_0, x'_1, \ldots, x'_n$ represent a fixed secondary structure S' of length n, and let $N_n^p(s)$ denote the number of secondary structures S of length n with $d_M^p(S', S) = s^{\frac{1}{p}}$, for $s \in \{0, 1, 2, \ldots\}$. We now describe recursions for computing $N_n^p(s)$, for $p \in \{1, 2, \ldots, \infty\}$.

5.1. Recursions for computing $N_n^p(s)$, $p < \infty$

For $t \in \{0, +1, -1\}$, let $N_k^t(s, l)$ denote the number of sequences $x_0, x_1, \ldots, x_k$, which satisfy (1 a), (2), and (3), and for which

$$\sum_{i=1}^{k} |x_i - x'_i|^p = s,$$

and in addition,

$$x_0 = 0, x_k = l, x_{k-1} = l - t,$$

(which replaces Condition (1b)). Thus, for $l > 0$, the number $N_k^t(s, l)$ can be thought of those "partial" secondary structures S whose d_M^p distance to S' restricted to the first k bases is $s^{\frac{1}{p}}$ and which satisfy $f_S(k) = l$ and $f_S(k-1) = l - t$ (see Figure 2).

Now, define

$$N_k(s, l) := N_k^{-1}(s, l) + N_k^0(s, l) + N_k^{+1}(s, l),$$

and

$$N_k^{\leq 0}(s, l) := N_k^{-1}(s, l) + N_k^0(s, l).$$

Then we clearly have

$$N_n^p(s) = N_n(s, 0). \tag{1}$$

We now describe a simultaneous one-step recursion for the pair $(N_n^{+1}, N_k^{\leq 0})$ which allows us to calculate $N_n^p(s)$ via Equation (1), in polynomial time. For $k > 1$ we have the recursion:

$$N_k^{+1}(s, l) = N_{k-1}(s - |x_k' - l|^p, l - 1) \tag{2}$$

$$N_k^{\leq 0}(s, l) = N_{k-1}(s - |x_k' - l|^p, l) + N_{k-1}^{\leq 0}(s - |x_k' - l|^p, l + 1) \tag{3}$$

subject to the boundary conditions,

$$N_k^{+1}(s, l) = N_k^{\leq 0}(s, l) = 0, \text{ if } s < 0, \text{ or } l \notin [0, k],$$

and the initial conditions

$$N_1^{+1}(s, l) = \begin{cases} 1 & \text{if } (x_1', s, l) \in \{(1, 0, 1), (0, 1, 1)\}, \\ 0 & \text{else,} \end{cases}$$

and

$$N_1^{\leq 0}(s, l) = \begin{cases} 1 & \text{if } (x_1', s, l) \in \{(0, 0, 0), (1, 1, 0)\}, \\ 0 & \text{else.} \end{cases}$$

A few words of explanation are in order for these recursions. Note that Equation (2) says that the number $N_k^{+1}(s, l)$ of partial secondary structures S with $f_S(k) = l$ *and* $f_S(k-1) = l - 1$ whose d_M^p distance to S' restricted to the first k bases, is simply the number of incomplete secondary structures with d_M^p distance $s - |x_k' - l|^p$ to S' restricted to the first $k - 1$ bases. In contrast, Equation (3) has two terms on the right hand side, so as to avoid problems with Condition (2), which forces us to exclude the possibility of the last three terms in the sequence $x_0, x_1, \ldots x_k$ representing the partial secondary structure S being $l, l+1, l$. The

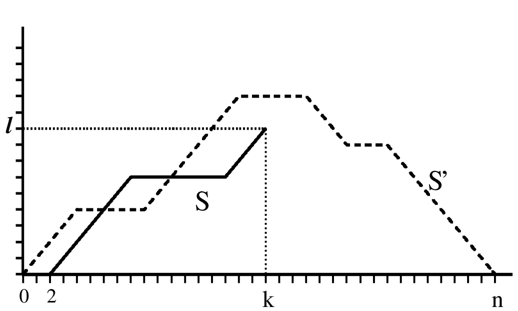

FIG. 2. One of the "partial" secondary structures S that is counted when obtaining $N_k^1(s, l)$, together with the fixed secondary structure S'.

initial conditions are justified as follows: $N_1^{+1}(s,l)$ is the number of sequences $x_0 = 0, x_1 = 1$ satisfying Conditions (1a), (2) and (3), with l_1-distance $s^{\frac{1}{p}}$ from the sequence $x'_0 = 0, x'_1$, where $x'_1 \in \{0,1\}$. Hence, we see immediately see that $N_1^{+1}(s,l)$ is either zero or one, and that for it to be nonzero we must have $s \in \{0,1\}$. This leaves only two possibilities; either $x'_1 = 0$, in which case $s = 1$, or $x'_1 = 1$, in which case $s = 0$. The reasoning for the second initial condition is similar.

Now, to calculate $N_n(s,0)$ we simply compute the matrices:

$$M_k^{+1} = [N_k^{+1}(s',l) : 0 \le s' \le s, 0 \le l \le min\{k, n-k\}]$$

and

$$M_k^{\le 0} = [N_k^{\le 0}(s',l) : 0 \le s' \le s, 0 \le l \le min\{k, n-k\}]$$

by starting with $M_1^{+1}, M_1^{\le 0}$ and applying the above recursions to derive $(M_k^{+1}, M_k^{\le 0})$ from $(M_{k-1}^{+1}, M_{k-1}^{\le 0})$. Note that $M_k^{+1}, M_k^{\le 0}$ are both $O(s \times min\{k, n-k\})$ matrices, and so we need to recursively construct only

$$\sum_{k=1}^{n} s \times min\{k, n-k\} = O(sn^2)$$

entries in order to compute $N_n^p(s)$.

5.2. *Recursions for computing $N_n^\infty(s)$*

We now describe an analogous recursion to that given in the previous section for the case where $p = \infty$. For $t \in \{0, +1, -1\}$, let $T_k^t(s,l) := \sum_{s'=0}^{s} N_k^t(s',l)$, and $T_k(s,l) := T_k^{-1}(s,l) + T_k^0(s,l) + T_k^{+1}(s,l)$. Then we have the recursions:

$$T_k^{+1}(s,l) = \begin{cases} T_{k-1}(s, l-1) & \text{if } |x'_k - l| \le s, \\ 0 & \text{else,} \end{cases}$$

and

$$T_k^{\le 0}(s,l) = \begin{cases} T_{k-1}(s,l) + T_{k-1}^{\le 0}(s, l+1) & \text{if } |x'_k - l| \le s, \\ 0 & \text{else,} \end{cases}$$

subject to the boundary conditions, $T_k^{+1}(s,l) = T_k^{\le 0}(s,l) = 0$, if $s < 0$, or $l \notin [0,k]$ and the initial conditions

$$T_1^{+1}(s,l) = \begin{cases} 1 & \text{if } (x'_1, l) = (1,1), s \ge 0; \text{ or } (x'_1, l) = (0,1), s \ge 1, \\ 0 & \text{else,} \end{cases}$$

and

$$T_1^{\le 0}(s,l) = \begin{cases} 1 & \text{if } (x'_1, l) = (0,0), s \ge 0; \text{ or } (x'_1, l) = (1,0), s \ge 1, \\ 0 & \text{else.} \end{cases}$$

In this case, s is constant in the recursions so we need only calculate the $O(n^2)$ entries in the vectors

$$t_k^{+1} := [T_k^{+1}(s,l) \; : \; l = 0, 1, \ldots, min\{k, n-k\}]$$

and

$$t_k^{\le 0} := [T_k^{\le 0}(s,l) \; : \; l = 0, 1, \ldots, min\{k, n-k\}]$$

to compute $T_n(s,0)$ which is the number of secondary structures at distance at most s from the input sequence, under the mountain metric with $p = \infty$. In case we explicitly require $N_n^\infty(s)$ we can use the identity: $N_n^\infty(s) = T_n(s,0) - T_n(s-1,0)$.

5.3. Recursions for restricted structures

If we require that loops within the secondary structures at a fixed distance from a fixed structure should contain at least $r \geq 2$ bases within them (see Section 2), then the previous analysis can be modified as follows. First, we replace (2) by a natural extension which ensures that the requirement is satisfied:

(2)* if $x_i > x_{i+1}$, then $x_{i-1}, \ldots, x_{i-r} \geq x_i$.

Note that (2)* reduces to (2) when $r = 1$.

Now, for the case $p < \infty$ we replace the above two simultaneous recursions by the two simultaneous recursions: for $k > 1$,

$$N_k^{+1}(s, l) = N_{k-1}(s - |x'_k - l|^p, l - 1),$$

$$N_k^{\leq 0}(s, l) = N_{k-1}(s - |x_k^0 - l|^p, l) + N_{k-1}^{\leq 0}(s - |x'_k - l|^p, l + 1) - \sum_{j=2}^{r} N_{k-j}^{+1}(s^{(j)}, l + 1),$$

where $s^{(j)} = s - |x'_k - l|^p - \sum_{t=1}^{j-1} |x'_{k-t} - (l + 1)|^p$, and the boundary conditions and initial conditions are as described earlier for the $r = 1$, $p < \infty$ case, together with the condition $N_k(s, l) = 0$ if $k \leq 0$, $l \geq 1$.

For the case $p = \infty$ we have the recursions:

$$T_k^{+1}(s, l) = \begin{cases} T_{k-1}(s, l - 1) & \text{if } |x'_k - l| \leq s, \\ 0 & \text{else,} \end{cases}$$

and

$$T_k^{\leq 0}(s, l) = T_{k-1}(s, l) + T_{k-1}^{\leq 0}(s, l + 1) - \sum_{j=2}^{r} T_{k-j}^{+1}(s, l + 1) V(j, k, s, l + 1),$$

where

$$V(j, k, s, l + 1) = \begin{cases} 1 & \text{if } |x'_t - (l + 1)| \leq s, \text{ for all } t : k - j < t < k, \text{ and } |x'_k - l| \leq s, \\ 0 & \text{else,} \end{cases}$$

and where $T_k^{\leq 0}$, T_k^{+1} are subject to the same boundary and initial conditions as were described for the $r = 1$, $p = \infty$ case, together with the condition $T_k(s, l) = 0$, if $k \leq 0$, $l \geq 1$.

6. COMPUTATIONAL RESULTS

To investigate more fully properties of the secondary structure metrics that we have defined, we implemented routines for computing d_Z, d_P, d_M^p ($p = 1, 2, \ldots, \infty$), and d_M. We used the routine provided in the VIENNA package (VIENNA) for computing d_T. We also implemented the recursion formulae that are given in Section 5. Secondary structures were generated using the VIENNA and *mfold* packages.

All programs were written in GNU C, and then run under Solaris 2.5 on a Sun SPARC Server 1000 with 4x 40Mhz CPU's, and 512Mb RAM. An example experiment was run with 200 structures of various length, in which every structure was compared to all others, giving 19900 metric calculations. For the mountain metrics the overall running time was relatively quick, and it was only the longer sequences (of length greater than about 600) which needed to be left for at most a few hours. The calculations for the tree metric (from the VIENNA package) required considerably more time and we were forced to abandon some experiments after the program had run for a week.

6.1. Distributions

In order to compare distributions of metrics we first generated a set of random sequences of a fixed length over the alphabet $\{A, U, C, G\}$, and then folded these sequences using the RNAfold routine from the VIENNA package, which associates a secondary structure of minimal energy to an RNA sequence. We then computed the distance between each pair of these structures using the various metrics. In Figure 3, a representative of the results that we obtained, we present the distributions of $d_T, d_Z, d_P, d_M^{\infty}, d_M^1$ and d_M, which were obtained by generating 200 random sequences of length 100. We normalized each of the distances by the diameter (except the d_M distance which we normalized by $d_M(S_*^{100}, S_o^{100})$).

In general, we found that the pairs of metrics d_T and d_M, d_Z and d_M^{∞}, and d_Z and d_M^1, had similarly shaped distributions. The metrics d_T and d_M had a similar spread of values, which in part leads us to believe that d_M will provide an attractive alternative to d_T. Note that the values of d_P in Figure 3 are generally lower than d_Z and less spread out. This is in accordance with the fact that d_P was designed to ignore outlying base pairs (see Section 3.1). It is surprising to us that d_T gives so few low values as compared with d_P and d_M^1. Moreover, to date we have not been able to find an explanation for the 'oscillation' of the distribution for d_M^{∞} between the values of 20 and 40 in Figure 3.

In general, we advocate the production of distributions such as those in Figure 3 when making an analysis as they indicate the possible range of values for the metrics in practice; thus, for example, if a value of 65 were obtained between two structures of length 100 then Figure 3 would indicate that these structures are, relatively speaking, far apart. Such techniques have already been used in studying tree comparison metrics (Steel *et al.*, 1993). More work needs to be done in this direction, for example, improving the method that we use to generate random structures.

We also measured the distances between each pair of structures in ensembles of suboptimal structures generated by *mfold* that, using window value 0 to ensure that the program did not prefilter the structures, was set to produce suboptimal structures that had free energy within ten percent of the optimal structure.

In Figure 4 we present the results obtained for part of the mRNA from Schistosoma mansoni. This sequence was obtained from the paRNAss server (paRNAss), and is an example of a RNA molecule that may fold into two distinct conformational structures (Giegerich *et al.*, 1994). In Figure 4 it can be seen that each metric indeed picks up two distinct classes of structures, as is indicated by the bimodal behavior of the distributions. However, it is also clear that the d_M^1 metric is not very sensitive at either long or short range (i.e., for structures that are either quite different or similar, respectively). Also the d_Z metric picks up a significant number of structures that are distance 25–30 apart; a result that needs to be investigated further. The d_T and d_M distributions are very similar and both metrics appear sensitive at long and short range.

6.2. Recursions

To investigate the recursions given in Section 5, we considered various naturally occurring sequences with the property that *mfold* either folded them in a well-defined way (i.e., with very few suboptimal structures that were all "similar" to the optimal structure), or in a poorly-defined way (i.e., produced many varied suboptimal structures). In particular, we took the optimal secondary structure generated by *mfold* and then computed the number of secondary structures within a given d_M^1 distance s of this structure for various values of the parameter r (see Section 5.3). We made the same computations for d_M^{∞}, but the results were similar, so we do not include them here.

In general, we found that, independent of whether the secondary structure was well-defined or poorly-defined, the number of secondary structures grew very rapidly with increasing s. For example, in Table 1 we present the results obtained for two RNA sequences that fold in a well-defined and a poorly-defined way (the length of these sequences were 167 and 359 respectively, and the number of suboptimal structures within 10% of optimal generated by *mfold*—using window value 0—for each sequence were 10 and 253, respectively). Moreover, in general, although for increasingly large values of r there were virtually no secondary structures close-by for small values of s, once s became sufficiently large the number of structures again grew very rapidly with increasing s (see Table 1).

We conclude that the space of secondary structures equipped with the mountain metric d_M^1 is in some sense homogeneous (i.e., that the number of secondary structures within a certain mountain metric distance of a certain fixed structure does not heavily depend upon the fixed structure in question). Thus it is

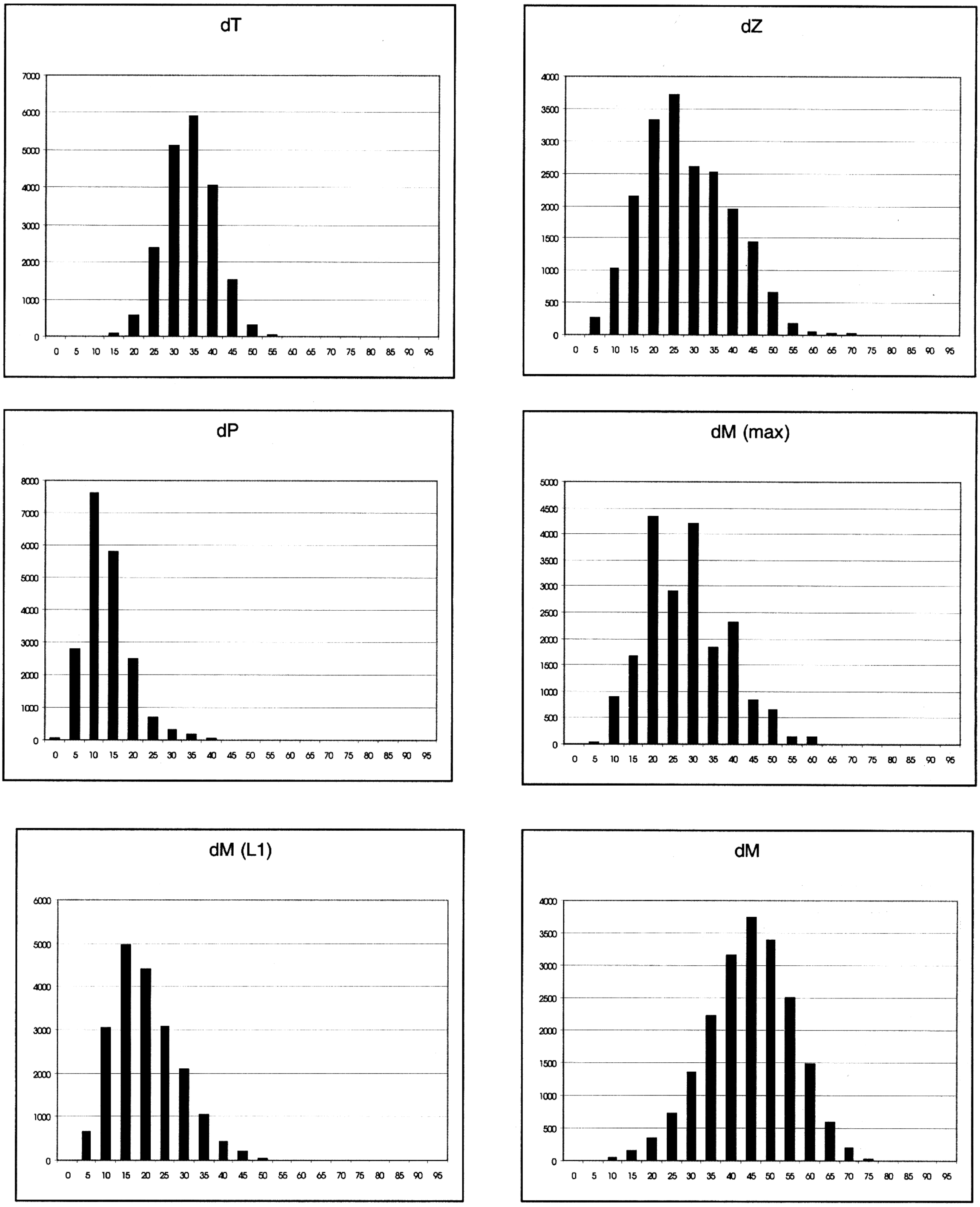

FIG. 3. Distributions of metrics; distances (as percentage of diameter, x-axis) between randomly generated structures vs. frequency (y-axis). dT = d_T, dZ = d_Z, dP = d_P, dM(max) = d_M^{∞}, dM(L1) = d_M^1, and dM = d_M

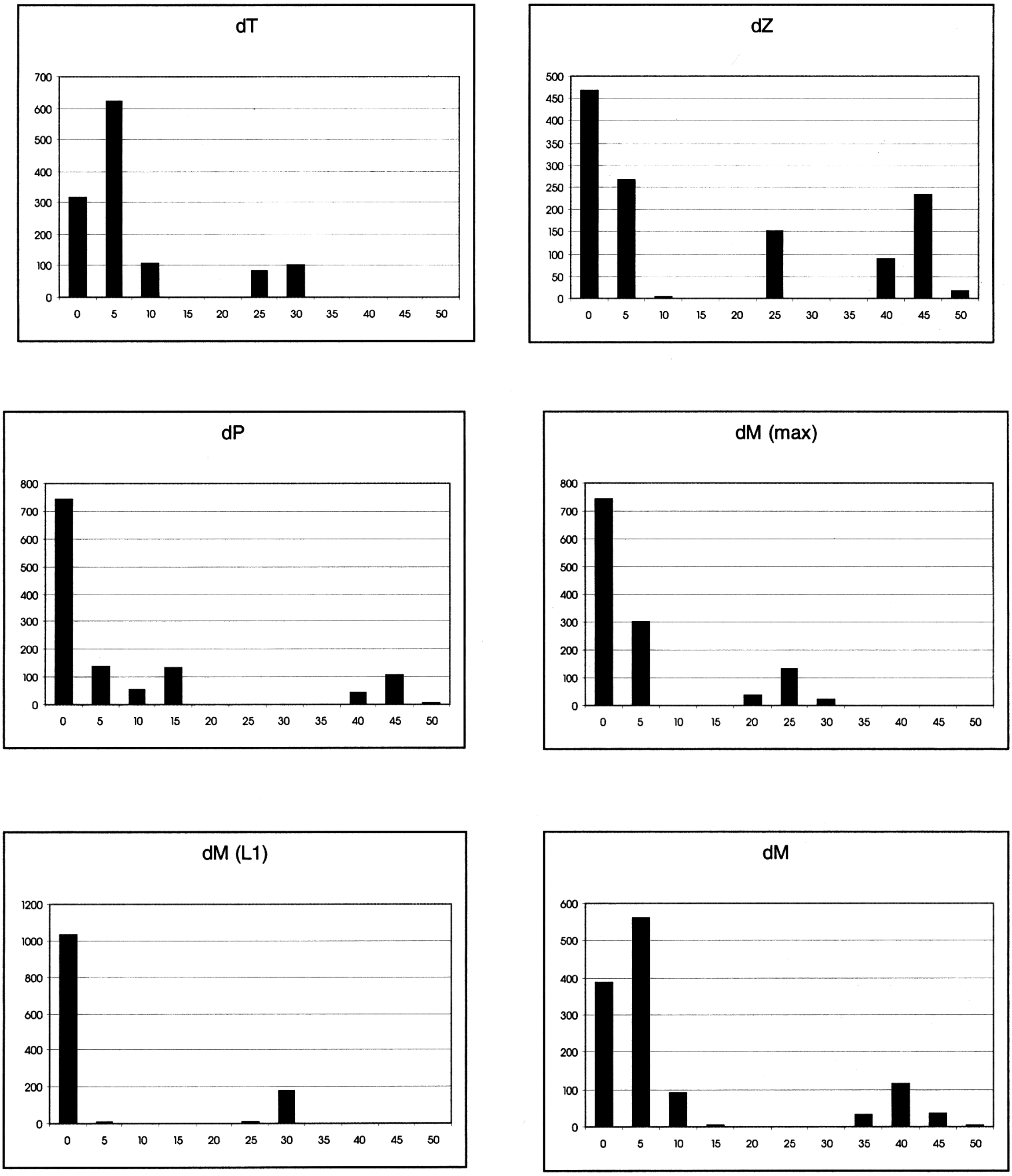

FIG. 4. Distances (as a percentage of diameter, x-axis) between an ensemble of structures generated by M-fold for a fixed RNA sequence vs. frequency (y-axis). dT = d_T, dZ = d_Z, dP = d_P, dM(max) = d_M^∞, dM(L1) = d_M^1, and dM = d_M

Table 1. Number of Foldings within Mountain Distance s of the Optimal Folding for a Well-Defined (Top) and Poorly-Defined Folding (Bottom) RNA Sequence, for Various Values of r

	r				
s	*1*	*2*	*3*	*4*	*5*
0	1	1	1	1	0
1	45	36	35	28	0
2	1049	685	641	420	0
3	16832	9125	8155	4453	0
4	208575	95241	80793	37261	0
5	2124297	827339	662919	261002	0
6	18488985	6209557	4681071	1587289	0
7	141221612	41300677	29198277	8590438	0
8	965031133	247905020	163936311	42116013	0
9	5986290657	1361416696	840362529	189572280	0
10	34095324485	69134630169	3976863553	791666722	0
11	1,79937E+11	32740667577	17528593300	30931126800	0
12	8,86519E+11	1,45605E+11	72482776384	11384826658	0
13	4,10311E+12	6,11585E+11	2,82895E+11	39704097505	0
14	1,79347E+13	2,43804E+12	1,04746E+12	1,3184E+11	0
15	7,43717E+13	9,2624E+12	3,69543E+12	4,18588E+11	3
16	2,93741E+14	3,3656E+13	1,24696E+13	1,2754E+12	92
17	1,10883E+15	1,17332E+14	4,03773E+13	3,74132E+12	1496
18	4,01272E+15	3,9353E+14	1,2583E+14	1,05964E+13	17066
19	1,39593E+16	1,27294E+15	3,78366E+14	2,90506E+13	152762
20	4,67948E+16	3,97976E+15	1,10035E+15	7,72689E+13	1139354
0	1	1	1	1	0
1	118	104	99	93	0
2	7027	5477	4972	4404	0
3	281481	194666	168799	141455	0
4	8529676	5251255	4355859	3464060	0
5	208507618	114640101	91088552	68937488	3
6	4281832947	2109090704	1607235925	1160564768	276
7	75959761157	33623174968	24602713397	16990350806	12946
8	1,18806E+12	4,74018E+11	3,33403E+11	2,20686E+11	412342
9	1,66398E+13	6,00184E+12	4,06194E+12	2,58232E+12	10023247
10	2,1126E+14	6,90859E+13	4,50331E+13	2,75491E+13	198174775
11	2,45548E+15	7,30077E+14	4,5878E+14	2,70552E+14	3317255802
12	2,63413E+16	7,14045E+15	4,32949E+15	2,46533E+15	48321647339
13	2,6259E+17	6,50715E+16	3,81016E+16	2,09823E+16	6,24924E+11
14	2,44665E+18	5,55714E+17	3,14483E+17	1,67733E+17	7,2851E+12
15	2,14133E+19	4,46934E+18	2,44637E+18	1,26547E+18	7,74716E+13
16	1,76804E+20	3,39952E+19	1,80118E+19	9,04827E+18	7,58774E+14
17	1,38244E+21	2,45469E+20	1,25985E+20	6,15371E+19	6,89865E+15
18	1,02707E+22	1,68814E+21	8,39891E+20	3,99355E+20	5,86075E+16
19	7,27199E+22	1,10899E+22	5,35222E+21	2,4801E+21	4,6784E+17
20	4,92002E+23	6,97727E+22	3,26875E+22	1,47764E+22	3,5259E+18

unlikely that we can distinguish between secondary structures that are either well- or poorly-defined using the number of structures that are within a certain d_M^1 distance. In this respect, it would be interesting to develop recursions which would include primary structure information (see Hofacker *et al.* (1998) for more on this possibility): for example to filter out those secondary structures that are close to a given secondary structure that could not have the required primary structure.

7. DISCUSSION

There are now a range of metrics available for comparing folded RNA structures. For most of these metrics some properties, such as the maximum diameter, are known and the number of structures close to one that is specified is known for certain mountain metrics. In addition there are some empirical results for both global and local properties. We do not advocate the use of any particular metric because the choice will depend on the application (and generally speaking, it is probably safest to try as many metrics as possible). For example, the metric d_T has some useful properties when comparing smaller numbers of trees, but the computational time would be considered excessive for data sets with thousands of structures based on long sequences, and the metric d_Z is used in *mfold* to filter out similar structures, in part due to the fact that this metric is straight-forward to implement.

In general it has been observed that d_T and d_M behave similarly (for example, see Figures 3 and 4). Thus, d_M should provide a useful alternative to d_T for analyzing ensembles of structures. We would not recommend the use of d_M^1 for such a task (see, e.g., Figure 4). An analysis of ensembles is required when studying the energy landscape of structures for a given RNA molecule. For example, we are currently developing tools for landscapes which should shed light on possible RNA properties before the evolution of protein synthesis, and hence on the origin of life (Moulton *et al.*, 1999).

Another application of ensemble analysis is the detection of conformal switches between secondary structures (Giegerich *et al.*, 1994). In Figure 5 we present what is known as a morphological plot for the RNA sequence that we used to generate Figure 4 (see Section 6.1). This was generated by plotting the values of the mountain metrics d_M and d_M^∞ against one another for all pairs of structures in a collection of 50 suboptimal structures generated by *mfold*. The plot in Figure 5 compares favourably with the example plots given on the server (paRNAss), which were generated using computationally more expensive metrics (Giegerich *et al.*, 1994). It clearly displays two clusters, which correspond to the possibility for the RNA in question to switch between two stable structural variants. We generated similar results using the tree-metric, however, these took longer to compute (a problem shared by the energy barrier metric used by paRNAss) (Giegerich *et al.*, 1994).

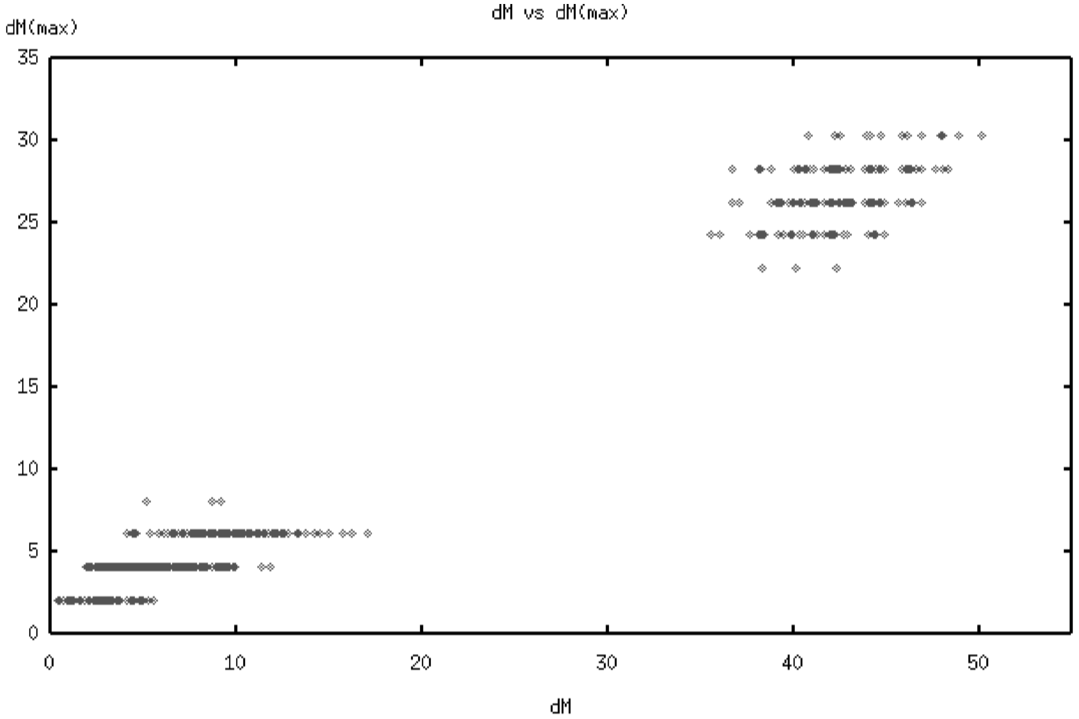

FIG. 5. Morphological plot using mountain metrics.

There is still much work to be done on the problem of understanding the landscape of structures of a given RNA molecule. The availability of good metrics for comparing RNA structures is allowing the quantitative study of this problem.

ACKNOWLEDGMENTS

The authors would like to thank Andreas Dress and Robert Giegerich for their helpful comments on earlier versions of this manuscript. We would also like to thank Ivo Hofacker for providing help with the VIENNA package and also a routine for generating random secondary structures. Finally, we thank the referee for giving some valuable comments.

REFERENCES

Bonhoeffer, S., McCaskill, J., Stadler P., and Schuster, P. 1993. RNA multi-structure landscapes. *European Biophysics Journal* 22, 13–24.

Collins, L., Moulton, V., and Penny, D. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP, submitted.

Dam, E., Pleij, K., and Draper, D. 1992. Structural and functional aspects of RNA pseudoknots. *Biochemistry* 31, 11665–11676.

Dress, A., Moulton, V., and Terhalle W. 1996. T-Theory—an overview. *European Journal of Combinatorics* 17, 161–175.

Fontana, W., Konings, D., Stadler, P., and Schuster, P. 1993. Statistics of RNA secondary structures. *Biopolymers* 33, 1389–1404.

Giegerich, R., Haase, D., and Rehmsmeier, M. 1999. Prediction and visualization of structural switches in RNA. *In* Altman, R., Lauderdale, K., Dunker, A., Hunter, L., and Klein, T., eds., *Biocomputing '99, Proceedings of the Pacific Symposium*, Mauna Lani, Hawaii, January 4–9, 1999, 126–137. World Scientific Press.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125, 167–188.

Hofacker, I., Schuster, P., and Stadler, P. 1998. Combinatorics of RNA secondary structures, *Discrete Applied Mathematics*, 88, 207–237.

Hogeweg, P., and Hesper, B. 1984. Energy directed folding of RNA sequences. *Nucleic Acids Research* 12 No. 1, 67–74.

Konings, D., and Hogeweg, P., 1989. Pattern analysis of RNA secondary structure: similarity and consensus of minimal energy folding. *Journal of Molecular Biology* 207, 597–614.

McCaskill, J. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119.

M-FOLD, `http://www.ibc.wustl.edu/~zuker/rna/`

Moulton, V., Gardner, P., Pointon, R., Creamer, L., Jameson, G., and Penny, D., RNA folding argues against a hot-start origin of life, Mid Sweden University Department of Mathematics Report 11, 1999.

The paRNAss page, `http://bibiserv.techfak.uni-bielefeld.de/parnass/`

Piccirilli, J., Krauch, T., Moroney, S., and Benner, S. 1990. Enzymatic incorperation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* 56, 1420–1424.

Sankoff, D., and Kruskal, J. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Addison-Wesley.

Schmitt, W., and Waterman, M. 1994. Linear trees and RNA secondary structure. *Discrete Applied Mathematics* 51, 317–323.

Shapiro, B. 1988. An algorithm for comparing multiple RNA secondary structures. *CABIOS* 4, 387–393.

Shapiro, B., and Zhang, K. 1990. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS* 6, 309–318.

Steel, M., and Penny, D. 1993. Distributions of tree comparison metrics—some new results. *Systematic Biology* 42(2), 126–141.

Stein, P., and Waterman, M. 1978. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics* 26, 261–272.

Uhlenbeck, O. 1995. Keeping RNA happy. *RNA* 1, 4–6.

Waterman, M. 1995. Introduction to Computational Biology. Chapman and Hall.

THE VIENNA PACKAGE (Version 1.1), `http://www.tbi.univie.ac.at/`

Zuker, M., and Sankoff, D. 1984. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology* 46, No. 4, 591–621.

Zuker, M. 1986. RNA folding prediction: The continued need for interaction between biologists and mathematicians. *Lectures on Mathematics in the Life Sciences* 17, 87–124.

Zuker, M. 1989a. On finding all suboptimal foldings of an RNA molecule. *Science* 244, 48–52.

Zuker, M., 1989b. The use of dynamic programming algorithms in RNA secondary structure prediction, 159–184. *In* Waterman, M., ed. *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton.

Zuker, Z., Jaeger, J., and Turner, D. 1991. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Research* 19, No. 10, 2707–2714.

Address correspondence to:
Vincent Moulton
FMI (Physics and Mathematics Department)
Mid Sweden University
S 851-70 Sundsvall, Sweden