BIOINFORMATICS

Vol. 16 no. 7 2000
Pages 583–605

Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs

*Elena Rivas and Sean R. Eddy**

Department of Genetics, Washington University, St. Louis, MO 63110, USA

Received on August 4, 1999; revised on December 15, 1999; accepted on December 21, 1999

Abstract

Motivation: *Several results in the literature suggest that biologically interesting RNAs have secondary structures that are more stable than expected by chance. Based on these observations, we developed a scanning algorithm for detecting noncoding RNA genes in genome sequences, using a fully probabilistic version of the Zuker minimum-energy folding algorithm.*

Results: *Preliminary results were encouraging, but certain anomalies led us to do a carefully controlled investigation of this class of methods. Ultimately, our results argue that for the probabilistic model there is indeed a statistical effect, but it comes mostly from local base-composition bias and not from RNA secondary structure. For the thermodynamic implementation (which evaluates statistical significance by doing Monte Carlo shuffling in fixed-length sequence windows, thus eliminating the base-composition effect) the signals for noncoding RNAs are still usually indistinguishable from noise, especially when certain statistical artifacts resulting from local base-composition inhomogeneity are taken into account. We conclude that although a distinct, stable secondary structure is undoubtedly important in most noncoding RNAs, the stability of most noncoding RNA secondary structures is not sufficiently different from the predicted stability of a random sequence to be useful as a general genefinding approach.*

Contact: *eddy@genetics.wustl.edu*

Introduction

One objective of genome sequencing projects is the identification of the complete set of genes contained in the DNA sequence. Traditionally, most efforts have been focused on the detection of protein-coding genes. However, there are an unknown—and possibly large—number of genes that produce functional noncoding RNAs (ncRNAs) (Olivas *et al.*, 1997). Known ncRNAs include transfer and ribosomal RNAs (tRNAs, rRNAs) (Soll and RajBhandary, 1995; Zimmerman and Dahlberg, 1996), many small nuclear and small nucleolar RNAs (Bachellerie *et al.*, 1995; Maxwell and Fournier, 1995; Nilsen, 1998; Tollervey, 1997), and many other RNAs of diverse functions (Baserga and Steitz, 1993; Bovia and Strub, 1996; Brockdorff *et al.*, 1992; Brown *et al.*, 1992; Cech, 1993; Delihas, 1995; Greider and Blackburn, 1996; Muto *et al.*, 1998; Watanabe and Yamamoto, 1994; Willard and Salz, 1997).

To date, most novel ncRNAs have been found by biochemical means. Noncoding RNA genes are typically small and make poor targets for genetic screens. Mature ncRNAs are often not polyadenylated, so they will be underrepresented in polyA-selected, cDNA-based gene expression surveys, such as expressed sequence tag sequencing (Marra *et al.*, 1998). Even in extensively studied organisms such as *Saccharomyces cerevisiae*, ncRNA genes have escaped detection (Olivas *et al.*, 1997). Given the availability of complete genome sequences for a variety of genomes, a computational approach to screening genome sequences for ncRNAs could be advantageous.

Current computational approaches for ncRNA identification are essentially similarity-search algorithms (Dandekar and Hentze, 1995; Lowe and Eddy, 1997, 1999; Woese and Pace, 1993). The advantage of a RNA genefinder algorithm over a similarity-search algorithm is that a genefinder is the appropriate tool to detect novel RNA gene families. A genefinder looks for genes without using homology information.

Genes that produce ncRNAs cannot be detected by protein genefinding algorithms. Noncoding RNA genes carry a much smaller amount of statistical information than protein-coding genes. There is nothing in RNA genes as strong as the codon bias, hexamer frequency, and open reading frame signals exploited by protein genefinders. A RNA genefinder remains elusive because it is difficult to find a statistically significant signal for ncRNA detection.

In a pioneering RNA genefinder attempt, Maizel's group (Chen *et al.*, 1990; Le *et al.*, 1988, 1989, 1990) proposed the use of secondary structure as a statistical signal for

*To whom correspondence should be addressed.

ncRNA gene detection. They proposed that *interesting RNAs will have a more stable secondary structure than expected by chance*. They designed an algorithm that uses the MFOLD RNA secondary structure prediction program (Zuker and Stiegler, 1981) to calculate the energy of a RNA segment, and then performed a large number of shufflings and recomputations of the energy in order to calculate the statistical significance of the MFOLD energy for the given segment. Others have followed the same approach (Seffens and Digby, 1999).

Following Maizel's work, we decided to explore further the use of RNA secondary structure as a statistical signal to detect ncRNA genes. We made two significant modifications to Maizel's approach. First, we implemented an algorithm for smoothly scanning a long genome sequence with a RNA secondary structure prediction algorithm, without having to partition the genome into overlapping windows. Second, instead of using an energy minimization model, we use a stochastic (probabilistic) context-free grammar (SCFG) for RNA secondary structure prediction. In combination with our scanning algorithm, the probabilistic approach has a pleasing advantage; an expected log-odds score gets worse as the length of a scored subsequence increases, so, in a log-odds scoring system, one can do 'local alignment' and identify high-scoring subsequences under a certain maximum target length. The same is not true for free energies because expected thermodynamic stabilities get better with subsequence length; the best-scoring subsequence will always tend to be the longest one, making it difficult to identify meaningful high-scoring subsequences unless one repeats the search with multiple different fixed-length window sizes. In the long run, a probabilistic model should also permit us to include statistical biases that are known to occur in biological RNA structures but are not well described in the current thermodynamic model for RNA structure stability (Ortoleva-Donnelly *et al.*, 1998).

Although our intention is to use these algorithms to detect novel ncRNA genes, they are actually scanning for any significant structured RNA. The genefinder could, for instance, detect *cis*-regulatory regions in mRNAs that involve a significant RNA structure. We use the term 'ncRNA gene' throughout the paper, but it should be understood that this is shorthand for 'a significantly folded subsequence.' It would involve further analysis to distinguish between ncRNA genes and other structural RNA features detected by any algorithm of this type.

We tested the RNA maximum-likelihood scanning algorithm in genomic sequences with known RNA genes. Our algorithm finds significant signals for structured RNA genes, such as tRNAs in *Caenorhabditis elegans*. However, those RNA genes also have a strong base-composition bias with respect to the background *C. elegans* base frequencies. To test how much of a given signal is truly due to secondary structure, instead of just being due to base-composition bias, we also constructed a simple scanning algorithm that only searches for base-composition biases. We have also implemented a scanning algorithm that essentially reproduces the Le and Maizel thermodynamic approach (Le *et al.*, 1988), and systematically evaluated whether it could detect ncRNA genes. The study presented here addresses an important concern affecting secondary structure screening algorithms in general: Does biological RNA secondary structure carry enough statistical signal for that to be a useful ncRNA genefinder?

Methods: three scanning algorithms

Here we describe the different scanning algorithms used in this paper. We start with the probabilistic model for RNA folding that we implement in our maximum likelihood algorithm. Subsequently, we also describe our reimplementation of the thermodynamic scanning algorithm first introduced by (Le *et al.*, 1988), with some attention to the statistical significance of Z-scores. Finally we describe the base-composition scanning algorithm used to compare the two previous structural algorithms.

The probabilistic model of RNA folding

To date, the most accurate algorithms for single-sequence RNA folding—implemented in the programs MFOLD (Zuker and Stiegler, 1981), ViennaRNA (Schuster *et al.*, 1994), and (with pseudoknots) in Rivas and Eddy (1999)—use thermodynamic parameters (Freier *et al.*, 1986; Turner *et al.*, 1987) to describe the different elements of RNA secondary structure. These algorithms calculate the free energy ΔG associated with a given folding structure. However, folding free energies are not good signals for the detection of highly structured RNAs since expected free energies for random sequence generally decrease linearly with the size of the sequence, so it is difficult to find optimal substructures within a longer sequence (such as a genome).

In the search for a statistical signal for RNA structure we turned towards a probabilistic model. There is a correspondence between probabilistic approaches described by SCFGs and the Turner/Zuker thermodynamic model for RNA folding (Durbin *et al.*, 1998), and we have a special interest in the application of probabilistic modeling to biological sequence analysis. Additionally, the scoring system used with probabilistic methods (log-odds scores) is suitable for finding suboptimal sequences within a larger region. Log-odds scores—which compare the likelihood of a sequence being generated by the model versus the likelihood of being generated by a null model—have the power to determine whether a local region fits the model (i.e. has a 'good' folding) or fits better to the null model (i.e. has no significant folding).

A scanning implementation of log-odds scores works in much the same way as Smith–Waterman algorithm scores do for local sequence alignment, where both are able to single out local regions that score better than random expectation.

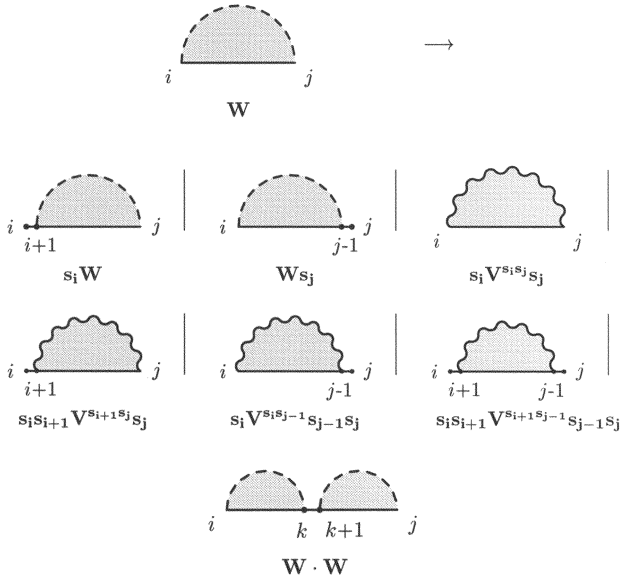
We have, therefore, implemented a probabilistic model for RNA folding using a stochastic context-free grammar that closely reproduces the main features of the current Turner/Zuker thermodynamic model of RNA folding. The main difference is that here we have replaced the thermodynamic scores with probabilistically determined parameters. The model (which has been trained on tRNAs and rRNAs) incorporates some small variations with respect to MFOLD, so we give the exact description of the model below.

Description of the model. To understand this section you should be familiar with context-free grammars (CFGs) (Durbin *et al.*, 1998). These are good models for describing RNA folding (Durbin *et al.*, 1998; Eddy and Durbin, 1994; Lefebvre, 1996; Sakakibara *et al.*, 1994; Searls, 1992), because they allow the correlated emission of two residues. You should also be familiar with the Zuker algorithm (Zuker and Stiegler, 1981)—which is identical to a CFG parsing algorithm—because our model is going to closely reproduce the same folding features. Finally, you should be familiar with the diagrammatic representation introduced in Rivas and Eddy (1999), which gives a convenient visualization of CFGs.

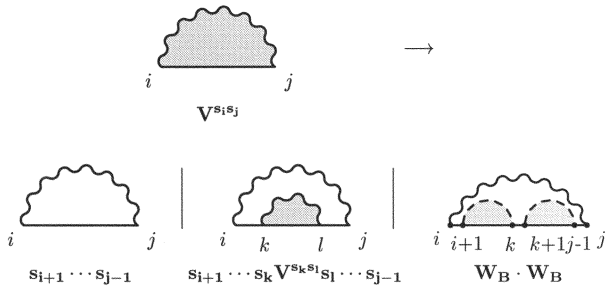
The states of the grammar are labeled W, W_B, V^{ab}. They correspond to the non-gapped matrices wx, wbx and vx of the diagrammatic representation described in Rivas and Eddy (1999). These diagrams constitute a convenient visual representation to enumerate which configurations we take into account in the model. States W and W_B represent a fragment in which the ends (i, j) are either paired or unpaired. State V^{ab} on the other hand, represent a fragment in which the ends (represented by nucleotide a at position i, and nucleotide b at position j) are paired.

Here we provide the grammar rules, together with their equivalent diagrammatic representation. The diagrams (one for every transition of the grammar) are a convenient way to enumerate the different configuration to be taken into account. A wavy line represents a simultaneous pairwise emission of two bases. A dashed line represents the unknown state of i, j. Lower case letters, i, j, k, represent positions, and s_i, s_j, s_k, stand for the terminal (i.e. nucleotide) emitted at those positions.

W acts as the starting state. W and W_B are formally equivalents, but W_B is used exclusively for starting multiloops. The production rules for W are (for W_B, replace W by W_B everywhere in the recursion),



The V^{ab} are the paired states, that is, the states we are in after emitting a pair $a, b \in$ alphabet. We therefore have 16 paired states, one for each pair of possibly emitted nucleotides. This allows us to retain information about a neighboring pair when another one is to be emitted, as in stacking correlations. The recursion for state V^{ab} is (without including hairpin mismatches, which are included in the program),



Here the first transition corresponds to hairpin loops, and is equivalent to function $FH(i, j)$ in Zuker and Stiegler (1981); the second transition corresponds to stems, bulges, and internal loops, and is equivalent to function $FL(i, j, k, l)$ in Zuker and Stiegler (1981); the last transition corresponds to multiloops, that is, loops closed by more than two hydrogen bonds.

The context-free grammar for RNA folding described by the previous production rules is independent of whether we are doing a Turner/Zuker thermodynamic implementation (non-stochastic) or a probabilistic (stochastic) implementation. Each production rule has associated a score in the non-stochastic implementation (the energy 'cost' in the Zuker implementation) that corresponds to a probability in the stochastic counterpart. For example, a stacking

of two pairs (a-b and c-d) is represented by the transition

$$V^{ab} \longrightarrow cV^{cd}d. \tag{1}$$

This production can be interpreted either thermodynamically or stochastically by

$$-\Delta G[FL(ab, cd)] + V^{cd}(i+1, j-1), \tag{2}$$

or

$$\log \frac{P[FL(ab, cd)]}{P^N(c)P^N(d)} + V^{cd}(i+1, j-1). \tag{3}$$

Where $\Delta G[FL(ab, cd)]$ and $P[FL(ab, cd)]$ stand for the free energy and probability respectively of the stem $FL(ab, cd)$. $P^N(c)$ is the probability of nucleotide c being generated by the null model, and $V^{cd}(i, j)$ is the dynamic programming matrix for state V^{cd}. Note the correspondence between 'negenergy' scores ($-\Delta G$) and 'log-odds' scores. For instance, we can compare the scores assigned by the two models for one unfavorable stacking $FL(AU, GG)$, and one favorable stacking $FL(AU, GC)$. While the first has a negenergy of $-\Delta G[FL(AU, GG)] = -0.4$ kcal/mol, the second stacking has a negenergy of $-\Delta G[FL(AU, GC)] = +1.7$ kcal/mol. Similarly, the unfavorable stacking $FL(AU, GG)$ has a worse log-odds score than the favorable stacking $FL(AU, GC)$ (-3.0 versus $+2.1$, in bit units, using a *C.elegans* background null model).

The scanning algorithm. For a given formal context-free grammar, such as the one presented here for RNA folding, we can implement different dynamic programming algorithms. For instance, MFOLD uses the non-stochastic implementation of the grammar to calculate, for a given RNA fragment, the free energy corresponding to the best folding. The equivalent of this MFOLD calculation, using the stochastic version of the RNA folding grammar, is referred to as the Cocke–Younger–Kasami (CYK) algorithm (Durbin *et al.*, 1998).

Similarly, the partition function calculations introduced by McCaskill (1990) to be used in the thermodynamic implementation (in which all possible folding configurations are taken into account) have their counterpart in the Inside algorithm for a SCFG (Durbin *et al.*, 1998). The Inside algorithm calculates the probability of a RNA sequence given a SCFG by summing over all possible foldings (paths) that the model allows:

$$P(\text{sequence} \mid \text{SCFG}) = \sum_{\text{paths}} P(\text{sequence, path} \mid \text{SCFG}). \tag{4}$$

We have implemented the genefinder as an Inside algorithm. In this way we are taking into account suboptimal foldings that could contribute to the stability of the structure almost as much as the 'best path' or optimal folding calculated by the CYK algorithm.

Because we want the algorithm to scan over a large genome of length L, we apply the algorithm to a subsequence of maximum target length w, and we sweep this maximum target window across the whole genome. The algorithm is not limited to scoring regions of fixed length w. The algorithm looks at all subsequences of length $\leq w$, so that it can find the exact $3'$ and $5'$ ends of a high-scoring subsequence (potentially, an interesting RNA) embedded in a given region.

The scanning algorithm requires as many dynamic programming matrices as the grammar has states. The matrices are W, W_B, and $V^{a,b}$. For a given position j in the genome, and a given window size d up to a maximum target length w, we scan from positions $i \equiv j - d$ to j. The (j, d) coordinate system allows us a smooth implementation of the dynamic programming recursions using matrices that have dimension $w \times w$, independent of the target sequence length L.

The recursion for state W is

$$\begin{aligned}
W(j,d) = {} & p^W[s_{j-d}W] \cdot W(j, d-1) && \text{left}\\
& + p^W[Ws_j] \cdot W(j-1, d-1) && \text{right}\\
& + p^W[V^{s_{j-d}s_j}] \cdot V^{s_{j-d}s_j}(j, d) && \text{pair}\\
& + p^W[s_{j-d}V^{s_{j-d+1}s_j}] \cdot V^{s_{j-d+1}s_j}(j, d-1) && \text{left dangling}\\
& + p^W[V^{s_{j-d}s_{j-1}}s_j] \cdot V^{s_{j-d}s_{j-1}}(j-1, d-1) && \text{right dangling}\\
& + p^W[s_{j-d}V^{s_{j-d+1}s_{j-1}}s_j] \cdot V^{s_{j-d+1}s_{j-1}}(j-1, d-2) && \text{left-right dangling}\\
& + p^W[WW] \cdot \sum_{d_1} W(j-d+d_1, d_1) \cdot W(j, d-d_1-1). && \text{bifurcation.}
\end{aligned} \tag{5}$$

For paired state $V^{s_{j-d}s_j}$ we have

$$\begin{aligned}
V^{s_{j-d}s_j}(j,d) = {} & p^V_{s_{j-d},s_j}[FH(s_{j-d+1}\cdots s_{j-1})] && \text{hairpin loops}\\
& + \sum_{d_1,d_2} p^V_{s_{j-d},s_j}[FL(s_{j-d+1}\cdots s_{j-d+d_1}s_{j-d_2}\cdots s_{j-1})] \cdot\\
& \quad V^{s_{j-d+d_1}s_{j-d_2}}(j-d_2, d-d_1-d_2) && \text{stems, bulges, internal loops}\\
& + p^V_{s_{j-d},s_j}[W_BW_B] \cdot\\
& \quad \textstyle\sum_{d_1} W_B(j-d+d_1+1, d_1) \cdot\\
& \quad W_B(j-1, d-d_1-3), && \text{multiloops,}
\end{aligned} \tag{6}$$

for $1 \leq j \leq L$, $0 \leq d \leq \min(j, w-1)$, $0 \leq d_1 \leq d$, $0 \leq d_1 + d_2 \leq d$.

The symbols $p^{\text{state}}[\text{trans}]$ represent the transition probabilities. They are the parameters that characterize the SCFG. For example, $p^W[aW]$ is the transition probability from state W to state W after emitting left nucleotide a. The transition probabilities are calculated as the observed frequencies of the different transitions in a training set of

known structured RNAs, with the conditions

$$\sum_{\text{trans}\in\text{state}} p^{\text{state}}[\text{trans}] = 1, \qquad \forall \quad \text{state}. \tag{7}$$

We use this model for discrimination between structured and not structured regions by calculating log-odds scores, that is, we compare the likelihood that the sequence has been generated by the folding model with the likelihood of being generated by a null model that emits the sequence without structure. The LOD score of a sequence fragment $(s_1 \ldots s_n)$ with $n \leq w$ is given by,

$$\begin{aligned}\text{LOD}(s_1, s_n) &= \log_2 \frac{P(s_1 \ldots s_n \mid \text{SCFG})}{\text{P}(\text{s}_1 \ldots \text{s}_\text{n} \mid \text{Null})} \\ &= \log_2 \frac{W(n, n-1)}{P(s_1 \ldots s_n \mid \text{Null})},\end{aligned} \tag{8}$$

The null model (N) emits the sequence nucleotides according to an estimated ncRNA base composition ($P_A^N = P_U^N = 0.33, P_C^N = P_G^N = 0.17$ throughout this paper, which is the background base composition for *C.elegans* and also for the noncoding regions of *S.cerevisiae*). Finally, $W(n, n-1)$ is given by (5).

Thus the null model is simple enough that the previous quotient (8) can be factored by substituting every transition probability that emits a number of nucleotides $(a_1 \ldots a_n)$ by the probability quotient

$$\begin{aligned}&P^{\text{state}}[\text{trans}(a_1 \ldots a_n)] \longrightarrow \\ &P^{\text{state}}[\text{trans}(a_1 \ldots a_n)]/P^N(a_1) \ldots P^N(a_n).\end{aligned} \tag{9}$$

The training set and parsing algorithm. To train the probabilistic model we have used a set of structural RNAs whose secondary structure is already known by comparative analysis. In the best current implementation, the training set includes 1415 tRNAs from the Sprinzl tRNA database (Steinberg *et al.*, 1993), and 208 small subunit ribosomal RNAs from the de Wachter rRNA structural database (Van de Peer *et al.*, 1994).

When the training set is parsed according to the grammar, it accounts for 231 796 transitions used to estimate 1722 free parameters in the probabilistic model (after some probabilities have been tied to each other; refer to the source code for a detailed description of the tied probabilities). In comparison, MFOLD has 986 free-energy parameters to take care of stacking energies, terminal mismatches for hairpin and interior loops, danglings and hairpin loop destabilizing energies, and some miscellaneous energy parameters.

We estimate transition probabilities t^α from the observed frequencies in the training set. We use Laplace priors, that is, if we find C^α counts for transition α then we estimate the probability of this transition by $t^\alpha = (1 + C^\alpha)/\sum_{\alpha'}(1 + C^{\alpha'})$. The absolute uncertainty for a given t^α, $\sigma(t^\alpha) = \sqrt{C^\alpha}/\sum_{\alpha'}(1 + C^{\alpha'})$, is never worse than 2%, which is similar to the estimated error for free energy change in the thermodynamic parameters (Freier *et al.*, 1986). The relative uncertainty, $\sigma(t^\alpha)/t^\alpha = \sqrt{C^\alpha}/(1 + C^\alpha)$, varies from 50% for almost-forbidden transitions down to 1.2% for the single-emission transitions in state W_B. These variances correspond to our sample of a database of RNAs (the sample variance). How accurately this sample variance estimates the variance of parameters estimated from all structured RNAs (the parametric variance), depends on how well rRNAs and tRNAs reflect general properties of RNA folding, and that is generally unknown (Sokal and Rohlf, 1981). (However we can safely assume that the parametric variance is higher than the sample variance. It would be desirable to train the model on a wider variety of RNAs if secondary structure data were readily available.)

Tested on tRNAs, the accuracy of the folding corresponding to the 'best-path' (CYK algorithm) is less good than that of a thermodynamic implementation (Rivas and Eddy, 1999; Zuker and Stiegler, 1981) (even though tRNAs were part of the training set for the model), but significantly better than a simple base-pair maximization algorithm such as the original Nussinov algorithm (Nussinov *et al.*, 1978). This is an important caveat: our model sacrifices accuracy in RNA folding prediction in return for substantial advantages in the ability to scan a genome and detect high-scoring (significantly folded) subsequences. We may compensate somewhat for this weakness by our choice of an inside algorithm, summing over all possible structures in the region rather than relying on the single maximum scoring structure.

The thermodynamic model of RNA folding

Description of the model. The model for the thermodynamic implementation is the same as the one presented for the probabilistic model, the only difference being that transition scores are not probabilities, but are taken instead from experimentally determined thermodynamic information provided by the Turner group (Freier *et al.*, 1986; Turner *et al.*, 1987).

The scanning algorithm. In terms of its implementation, the thermodynamic algorithm is just a scanning version of Zuker's MFOLD (Zuker and Stiegler, 1981), implemented including coaxial energies [as in Rivas and Eddy (1999)]. One of the differences is the fact that, while the probabilistic model is an Inside algorithm, in the thermodynamic algorithm the score assigned to a given window corresponds to the score of the best possible folding—also referred to as a CYK algorithm. In addition, while the probabilistic algorithm looks at all subsequences of length $\leq w$, the thermodynamic algorithm only scores regions of fixed length w.

Another important difference is that statistical signifi-

cance is evaluated using energy Z-scores. Le *et al.*'s (1988) Z-score calculates the number of standard deviations by which the energy of a fixed-window sequence is different from the average energy of the shuffled sequences. For a given quantity G (such as the free energy) that takes a seq $= \{s_1 \ldots s_N\}$ as argument, we define the Z-score of G for seq as,

$$\text{Z-score}(G; \text{seq}) = \frac{G(s_1 \ldots s_N) - \overline{G}(s_{i_1} \ldots s_{i_N})}{\sigma[G(s_{i_1} \ldots s_{i_N})]}, \quad (10)$$

where $\overline{G}(s_{i_1} \ldots s_{i_N})$ is the average of G over a large number of permutations (shufflings) $\{s_{i_1} \ldots s_{i_N}\}$ of the sequence, and $\sigma[G(s_{i_1} \ldots s_{i_N})]$ is the standard deviation of G over these permutations.

Our thermodynamic scanning algorithm is a close re-implementation of Maizel's RNA genefinder (Le *et al.*, 1988)—apart from a change in the sign in the way we report Z-scores, since we calculate Z-scores of 'negenergies' $(-\Delta G)$.

Approximate estimate of the statistical significance of Z-scores. Assuming that the distribution of negenergy scores on randomized sequences is Gaussian (which is approximately but not exactly true) Z-scores give a direct measure of statistical significance: for example, a Z-score $\geq$4 would only be expected to occur once in every 31 000 samples (i.e. subsequences from the genome that fit the 'random' null hypothesis), according to a Gaussian distribution.

We can also estimate the significance of Z-scores using a second, less common statistical approach: the distribution of extreme values for shuffled sequences. That is, how likely is it that the given Z-score of a biological sequence is *larger than the maximum Z-score* from a collection of randomized versions of the same sequence? When that likelihood is high (say, 99% or higher), the Z-score can be considered significant.

Consider a sequence of length L that has an observed Z-score (z) for which we want to estimate significance. We can look at the distribution of the maximum of the Z-scores of n random sequences

$$P(z_n^{\max} > z^*), \qquad \text{where} \quad z_n^{\max} = \max\{z_1, \ldots, z_n\}, \quad (11)$$

where the set of n random sequences was generated by shuffling the original sequence—thus destroying the secondary structure, but keeping the base composition intact—and have Z-scores $z_1, \ldots, z_n$. If the probability of this maximum being greater than the Z-score of the real sequence is small (such as 0.01 or smaller), then the Z-score will be considered a significant measure of secondary structure.

Assume that the distribution of Z-scores of random sequences is normal with $\mu = 0$ and standard deviation $\sigma^2 = 1$ ($N(0, 1)$). The distribution of maximum $\mathrm{Z}_n^{\max}$ (known as the extreme value distribution) of a normal distribution is known, in the limit of large n, and has the form (Waterman, 1995)[†]

$$P(\mathrm{Z}_n^{\max} > \mathrm{Z}^*) \simeq 1 - \exp\left(-\mathrm{e}^{-a_n(\mathrm{Z}^* - b_n)}\right), \qquad \text{for } n \text{ large}, \quad (12)$$

with

$$a_n = \sqrt{2 \ln n}, \quad (13)$$

$$b_n = \sqrt{2 \ln n} - \frac{1}{2} \frac{\ln \ln n + \ln 4\pi}{\sqrt{2 \ln n}}. \quad (14)$$

Therefore, if we sample $n = 100$ random versions of the original sequence at the time, and demand that the probability that the maximum Z-score for the random set exceeds the Z-score of the actual sequence is no bigger than 0.01, then by (12) Z-score values have to be at least of the order of 3.8 to be considered significant.[‡]

Operationally, therefore, we will use a threshold of Z=4 to define 'significant' hits.

The base-composition model

This is the simplest model able to detect relative CG-biased regions in a genome. While structural RNAs are about 50% CG rich on average, the *C.elegans* genome is quite AT rich (66%). For that reason, the base-composition model we designed—which formally is a stochastic regular grammar with only one state—emits nucleotides with an equally likely uniform distribution (which is approximately the probability distribution of the RNA training set used for the probabilistic model). On the other hand, the null model used to calculate the LOD scores favors AT-rich emissions, as the genomic *C.elegans* background does.

In practice, the emission probabilities of the base-composition model are 0.25 for any nucleotide, and the emission probabilities of the null model are $P^N(A) = P^N(U) = 0.33$ and $P^N(C) = P^N(G) = 0.17$. As a result, the model gives a log-odds score of -0.40 for any A or U found in the sequence, and a log-odds score of $+0.56$ for any C or G. The LOD scores with this base-composition model of a sequence containing n_i

[†] It is often the case that the distribution of Z-scores of the random sequence deviates from a normal distribution towards a larger positive tail (see further results). Thus in these cases, the Z-score cut-off for significance provided here is a lower bound of the actual threshold.

[‡] The general result is this: if we sample n permuted sequences (with n large) and demand that the probability that $\mathrm{Z}_n^{\max}$ exceed Z be no more than P, then we require that Z> Z_* with

$$\mathrm{Z}_* = b_n - \ln(-\ln(1 - P))/a_n. \quad (15)$$

nucleotides for $i = \{A, C, G, U\}$ is given by,

$$\mathrm{LOD(seq)} = -0.40\times(n_A+n_U)+0.56\times(n_C+n_G). \quad (16)$$

The base-composition model is implemented in a scanning version similar to that of the probabilistic structural scanning algorithm: e.g. we detect maximum-scoring subsequences with length $\leq w$.

Implementation

The algorithms have been implemented in ANSI C on Intel/Linux and Silicon Graphics Origin200 platforms. The probabilistic scanning algorithm has a time complexity of $\mathcal{O}(Lw^2)$ (after internal loops have been reduced in one order) and a storage complexity of $\mathcal{O}(w^2)$, for a genome of length L, scanned with a maximum target length w; on an SGI O200 R10K/180 it analyzes 40 bases per s, for $w = 100$ bases, sliding one position at a time. It takes about 87 h to process the whole *S.cerevisiae* genome (12.5 Mb). As for the thermodynamic scanning algorithm, the calculation of free energies has the same complexity in time and memory as that of the maximum-likelihood scanning algorithm; however, the calculation of Z-scores has a worse-case time complexity of $\mathcal{O}(Lnw^3)$ per chosen window length, where n is the number of permutations performed to evaluate Z-scores, and w is the fixed-length scanning window. [In Le *et al.* (1988) and Chen *et al.* (1990), Maizel's group has used a precalculated look-up table approach to remove the factor of n, but here we have not done the necessary precalculations and curve fitting.] In the results presented here we have used $n = 100$ permutations. The complexity in memory of the simple base-composition scanning algorithm is $\mathcal{O}(Lw)$ while the algorithm is linear in memory. The code is available from http://www.genetics.wustl.edu/eddy/software/ncrnascan.

Results: probabilistic CFG model

Preliminary results

The structural probabilistic algorithm produces what at first glance appear to be quite promising results when applied to various biological sequences with known RNA genes. Figure 1 gives an example for a fragment of the *C.elegans* clone C28G1 which contains two tRNAs, between positions 20130–20202 and 20346–20417. Other *C.elegans* RNA genes such as SRP-RNA, U1, U6, etc. give similarly strong scores.

To estimate the statistical significance of the log-odds scores, we generated 10 megabases of random sequence sampled uniformly from a equally likely distribution of nucleotides. We observe that it requires log-odds larger than 9.1 to observe no more than 10 false positives per megabase scanning with a maximum window length of 100 nucleotides, and discarding hits less than 25 nucleotides long. The hits in Figure 1, and for other tested RNAs, are clearly above this significance threshold.

We also applied the structural scanning algorithm to other genomes. In *Methanococcus jannaschii*, the signals for tRNAs were stronger than in *C. elegans*. In *S. cerevisiae*, the tRNA signals were weaker than in *C. elegans*. In *Escherichia coli*, tRNAs did not produce, in general, significant signals. These observations imply a strong correlation between the strength of a tRNA hit and the difference in CG base composition between the tRNA and the background base composition of the given organism. While tRNAs retain a similar CG-rich base composition across species (ranging from 54% CG in yeast to 67% CG in *M. jannaschii*), the organism background base compositions vary tremendously: *M. jannaschii* background is very AT rich (70%); *C. elegans* background is quite AT rich (66%); *S. cerevisiae* is also AT rich but to a lesser extent (61%); however, *E. coli* is not AT rich at all (50%). In general we observe that the larger the relative CG bias of the RNA genes, the stronger the signals appear to be.

Comparison to a simple base-composition model

The strong correlation between RNA gene detectability and the relative CG base-composition bias forced us to carefully analyze the base-composition component of the signals obtained with our algorithm. We constructed a scanning algorithm that only searches for base-composition biases, looking for regions of relatively high CG content. No structural information of any kind is added to the model (see Section **The base-composition model**).

When we apply the base-composition scanning algorithm to the clone fragment C28G1 (see Figure 2), the result is remarkably similar to those obtained with the structural algorithm (cf. Figures 1 and 2). This observation also holds for various other RNA genes.[§]

To elucidate how much of the signal we see with the structural scanning algorithm is due to secondary structure versus base composition (i.e. primary structure) we performed two types of experiments which we describe in the following subsections.

Shuffled sequences

In this type of experiment we try to determine whether the structural algorithm can distinguish a real RNA gene from a shuffled version of that gene. We start with biological sequences containing real known RNA genes for which we find apparently significant hits, and recalculate the score

[§] An estimation of the statistical significance of the log-odds generated by this model, similar to the one used in Section **Preliminary results**, indicates that for log-odds larger than 11 no more than 10 signals per megabase (signals of at most 100 bases, and ≥66% CG rich) are generated by sampling from an equally likely uniform distribution of nucleotides.

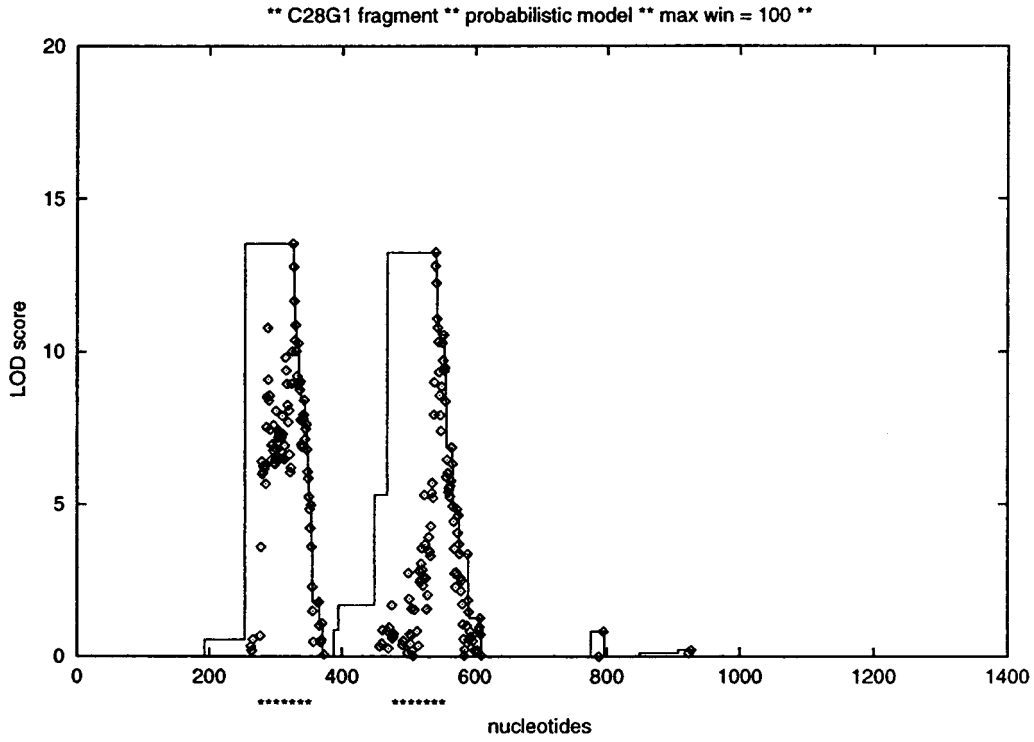


Fig. 1. Results of applying the probabilistic model for the *C.elegans* clone C28G1 fragment (19877–21256). The ordinate represents LODs in bits. This clone fragment contains two tRNAs represented under the abscissa by '****' (local coordinates: 253–325 and 469–540). Scores are represented by a dot placed at the end of the scoring segment. The scoring segment can be of variable length, up to a maximum target length (100 nucleotides in this example). The horizontal bars represent the maximum score that includes a given position.

by shuffling the high-scoring sub-sequence.

The shuffling procedure we propose here is different from the one described by Maizel to estimate statistical significance of energies (Le *et al.*, 1988). In their case, because they work with a fixed scanning window, the shuffling usually covers a region larger than the RNA gene. Therefore, their shuffling procedure not only destroys the secondary structure, it also potentially modifies the base composition of the gene contained within the larger window. Here, because we do not use a fixed window length, we are able to shuffle the exact segment that defines the RNA gene. In that way, our shuffling process maintains the exact base composition of the RNA gene.

If the score of the high-scoring sub-sequence is due to secondary structure, our shuffling process will make the significance of the hit diminish considerably—because the pattern that presumably produced a particular stable folding has been randomized. On the other hand, if the score is merely due to a base-composition bias, there will be little difference between the original score and the shuffled one.

Figure 3 shows the results for the *C.elegans* clone C28G1. As expected, the base-composition model retains most of the scoring shape after the shuffling—after all, base composition contains no structural information. Surprisingly though, the structural algorithm also retains the tRNA hits. This result, which has been reproduced for other RNA genes present in other clones, indicates that the regions of apparent significance are still significant after shuffling and destroying the secondary structure. As expected, the shuffled sequence scores tend to smear out as we expand the shuffled region around the scoring regions in Figures 1 and 2 (calculations not shown).

Chimeric sequences

The above experiments indicate that shuffling does not destroy the statistical signal. However, the structure may still be contributing a significant component of the score. To test this hypothesis we have performed the inverse experiment: embed a real RNA sequence in a random sequence of identical base composition.

If the structural algorithm is able to detect the RNA motifs above the background then we could say that biological RNAs have enough secondary structure for it to be used as a genefinder signal. Otherwise, we would have

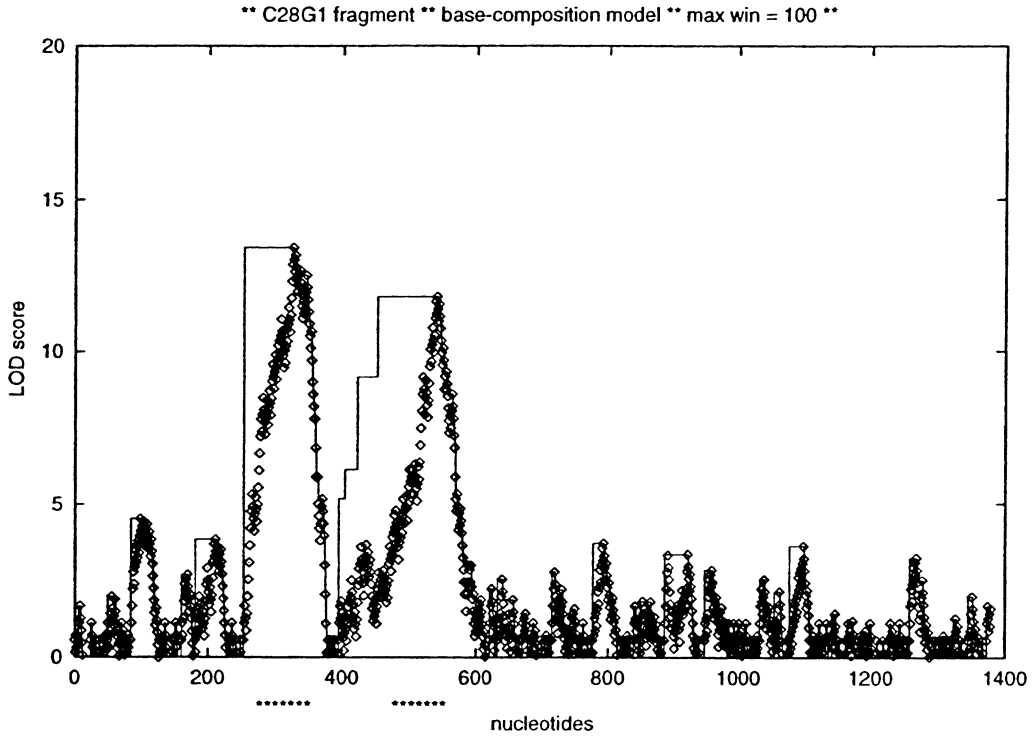


Fig. 2. Result of applying the base-composition model to the *C.elegans* clone C28G1, fragment (19877–21256). As in Figure 1, scores are represented by a diamond placed at the end of the scoring segment, and the solid line represents the maximum score that includes a given position.

to conclude that the secondary structure signal extracted by the algorithm is not sufficient because we cannot distinguish a real ncRNA from something that only shares with it the same base composition. Unfortunately, the latter scenario is the one that holds for most of the examples we have tested, of which we show two examples in Figures 4 and 5.

Results: thermodynamic CFG model

The previous experiments indicate that using the probabilistic model we can detect ncRNAs in many genomes, but simply as the result of base-composition bias and not because of any statistical significance of their secondary structures. We were concerned that our negative results appeared to conflict with the results of Maizel's thermodynamic scanning algorithm. This led us to re-examine the thermodynamic scanning algorithm.

In their thermodynamic implementation, Le *et al.* (1988) used Z-scores to evaluate the significance of the free-energy scores. The Z-score normalizes the sequence energy over shuffled versions of the same sequence; therefore, a Z-score should normalize relative to base-composition content. Are Z-scores a reliable detector of biologically relevant RNA secondary structures? That is, how much secondary structure signal is left once we control for base-composition effects?

In their work (Chen *et al.*, 1990; Le *et al.*, 1988, 1989, 1990) the authors did not provide any test study that could allow us to evaluate whether biologically interesting RNAs *systematically* have significant Z-scores; the evidence they presented was anecdotal in nature. We have re-implemented the Maizel algorithm (see Section **The thermodynamic model of RNA folding**), analyzed the behaviour of the algorithm for known RNA genes, and studied the effect of base-composition bias.

Our results for Z-scores for retrovirus HIV-1 [isolate bh-10, 9748 nucleotides long, Genbank accession: X01762, (Ratner *et al.*, 1985)] given in Figure 6 are similar to those presented in (Le *et al.*, 1990). The signal around position 500 corresponds to the *cis*-acting target sequence (referred to as TAR, positions: 454–514) that interacts with Tat protein. The signal around position 8000 corresponds to the Rev responsive element (referred to as RRE, positions: 7789–8031) that interacts with the Rev protein (Rosen, 1991). The results for retrovirus HIV-1 using Z-scores are comparable with those obtained using the probabilistic scanning algorithm in Figure 7, and for that matter, with those of the base-composition algorithm, Figure 8.

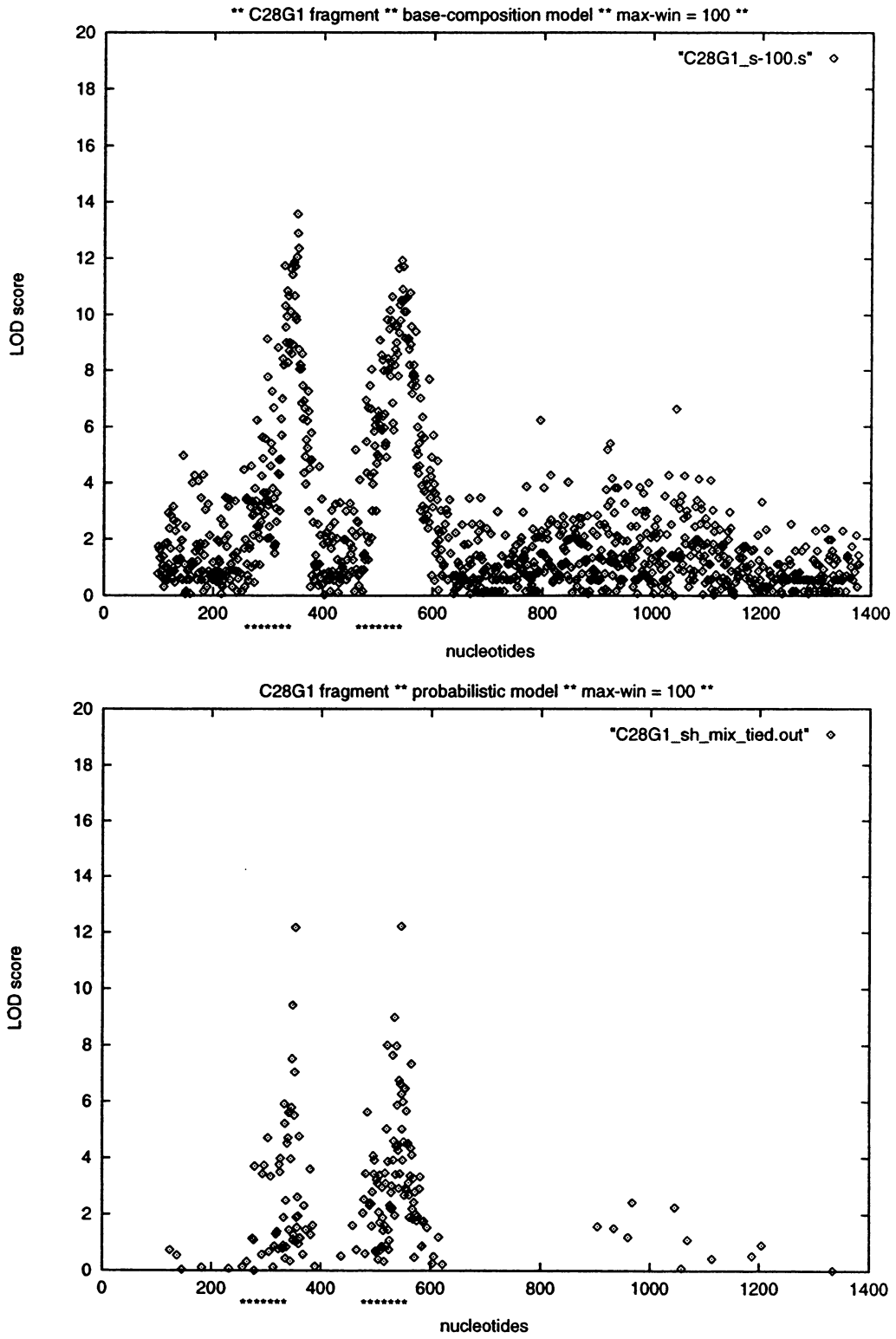


Fig. 3. Results for the C28G1 fragment after shuffling the exact segment that generated every score in Figures 1 and 2. The disappearance of the two tRNA signals would be an indication of secondary structure. The tRNA signals remain for both the base-composition model (top) and the structural model (bottom) after altering the secondary structure pattern.

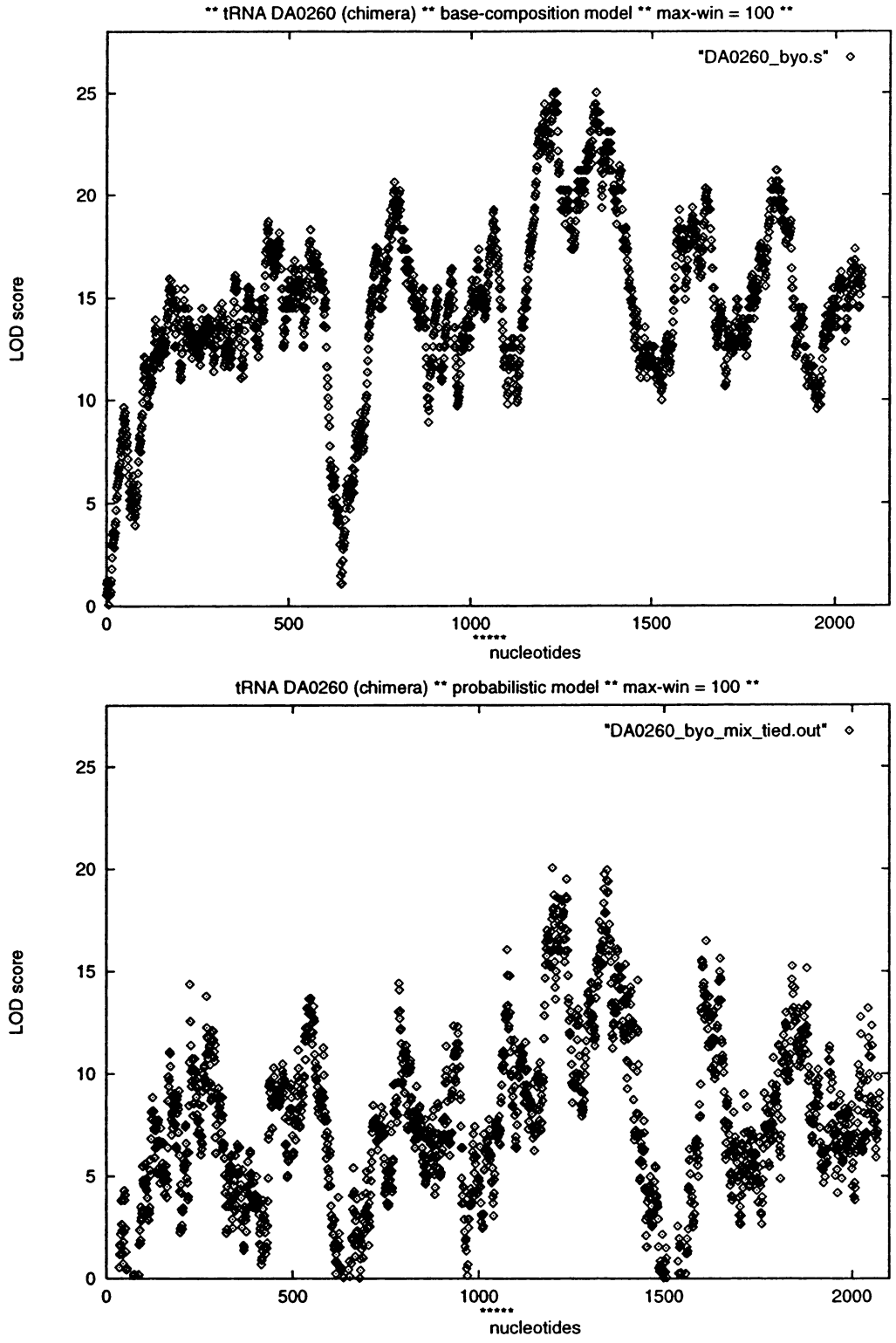


Fig. 4. Chimeric sequence that contains tRNA DA0260 [an alanine coding tRNA from phage T5 from the Sprinzl database (Steinberg *et al.*, 1993)]. The DA0260 tRNA—75 nucleotides, %A = 21.6, %C = 24.4, %G = 32.4, %U = 21.6, and represented by '****'—is flanked at both ends by 1000 nucleotides randomly generated with the same base composition as DA0260. The top figure indicates the base composition of the chimeric sequence. The bottom graph shows the inability of the structural algorithm to distinguish the tRNA from the background. Notice that the log-odds scores are consistently high due to the high CG content of the sequence.

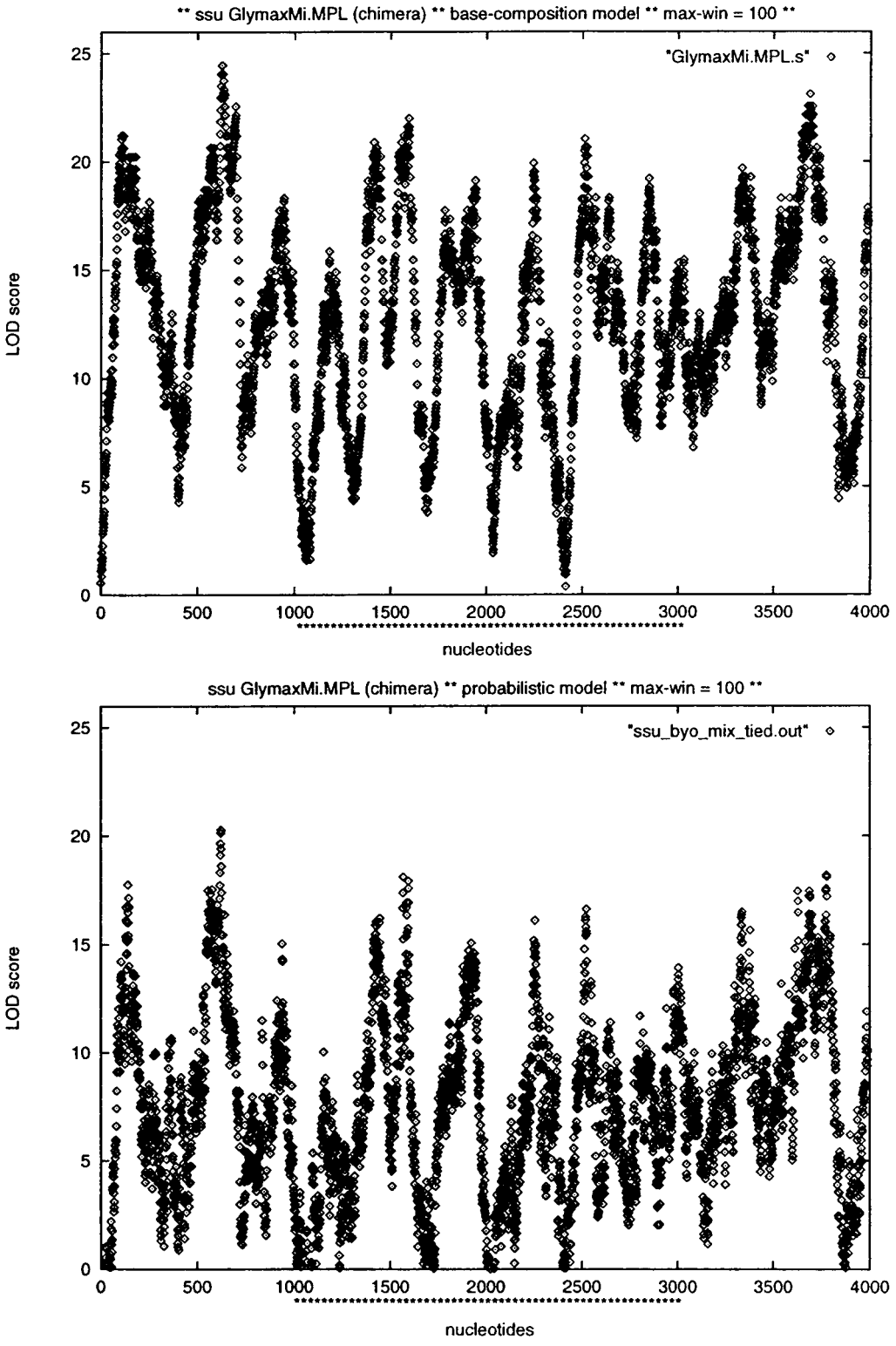


Fig. 5. Chimeric sequence that contains the soy bean mitochondrial small-subunit (SSU) rRNA (GlymaxMi.MPL from the De Wachter database (Van de Peer *et al.*, 1994)). This SSU—1990 nucleotides, %A = 24.4, %C = 23.3, %G = 31.5, %U = 20.8, and represented by '****'—is flanked at both ends by 1000 nucleotides randomly generated with the same base composition as the SSU. As in Figure 4, the SSU signal generated by the structural algorithm is buried in the background.

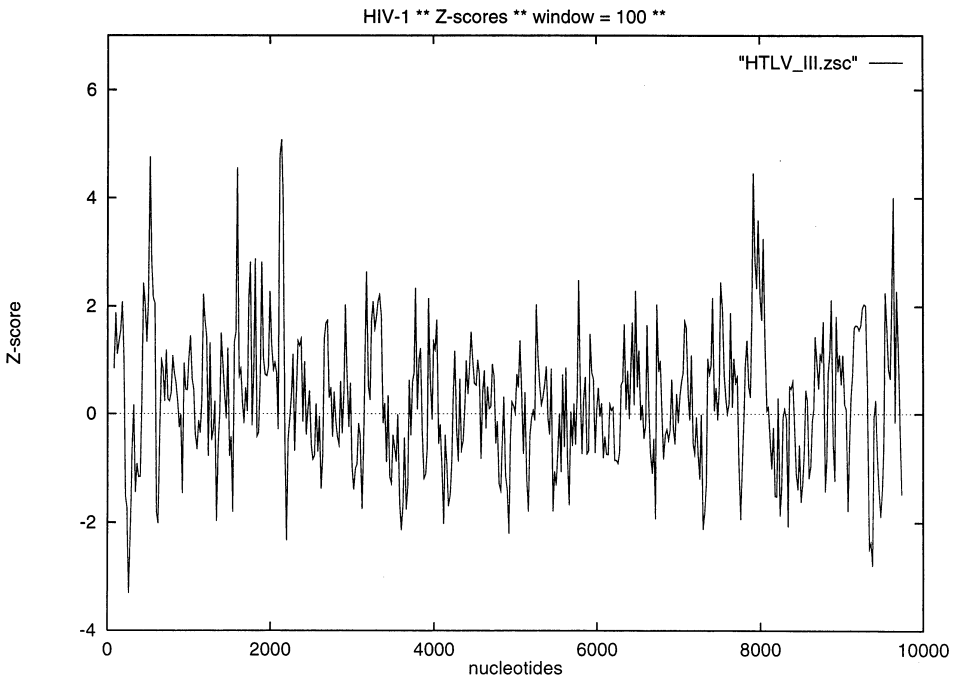


Fig. 6. Negenergy Z-score plot for HIV-1 retrovirus (isolate bh-10) generated using our implementation of the thermodynamic algorithm. This figure closely reproduces the results presented in Le *et al.* (1990, Figure 1).

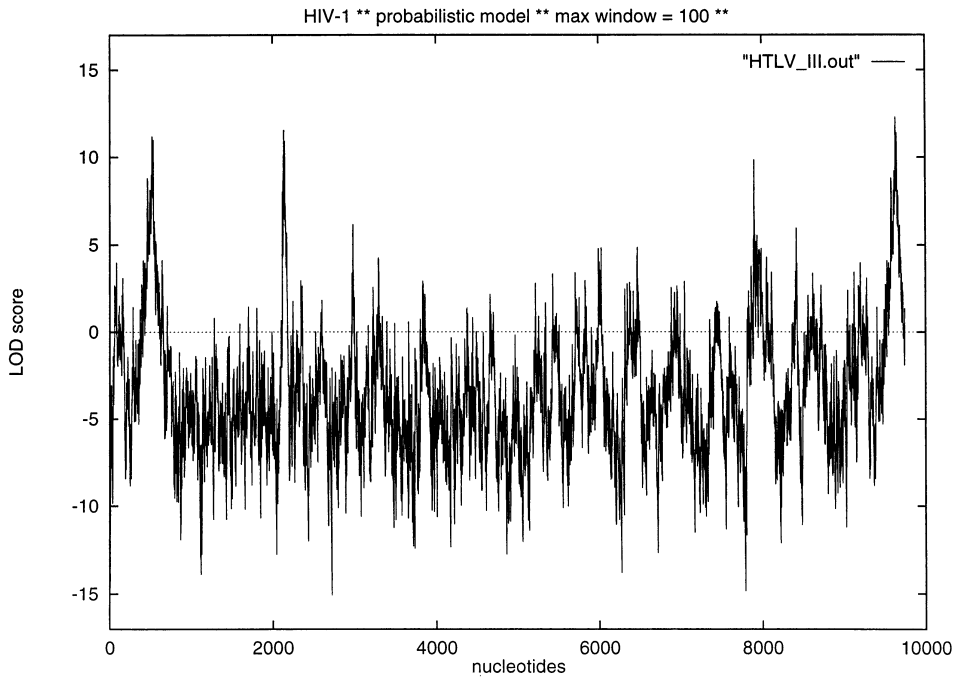


Fig. 7. Log-odds scores plot for HIV-1 retrovirus (isolate bh-10) generated using our probabilistic algorithm for secondary structure. Results are quite similar to those generated with the thermodynamic algorithm.

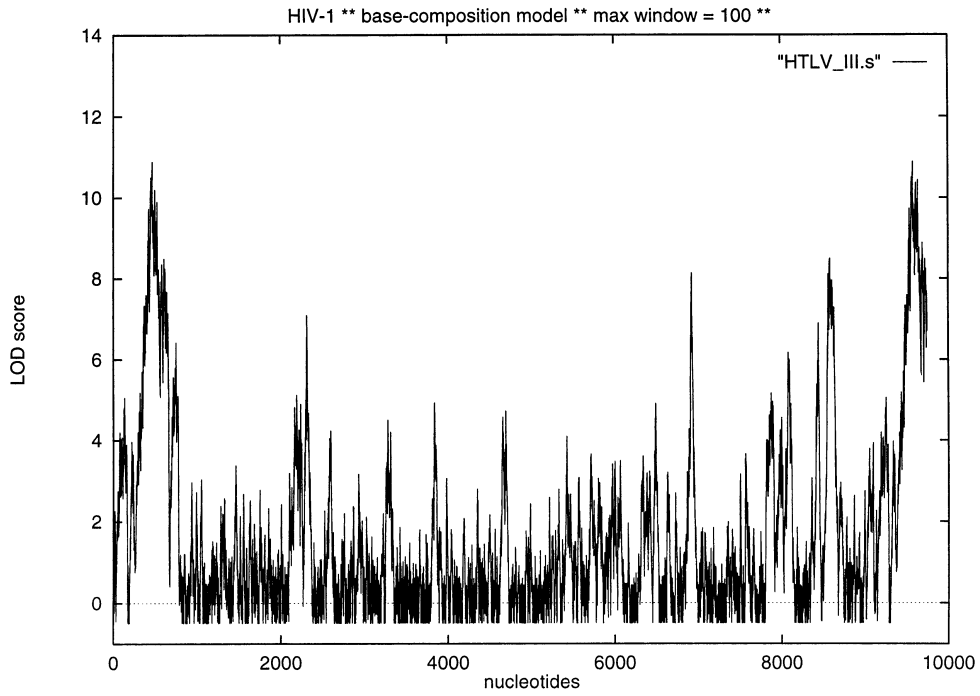


Fig. 8. Log-odds scores plot for HIV-1 retrovirus (isolate bh-10) generated using the base-composition bias algorithm. Results are comparable to those generated with either of the structural algorithms (Figs 6 and 7).

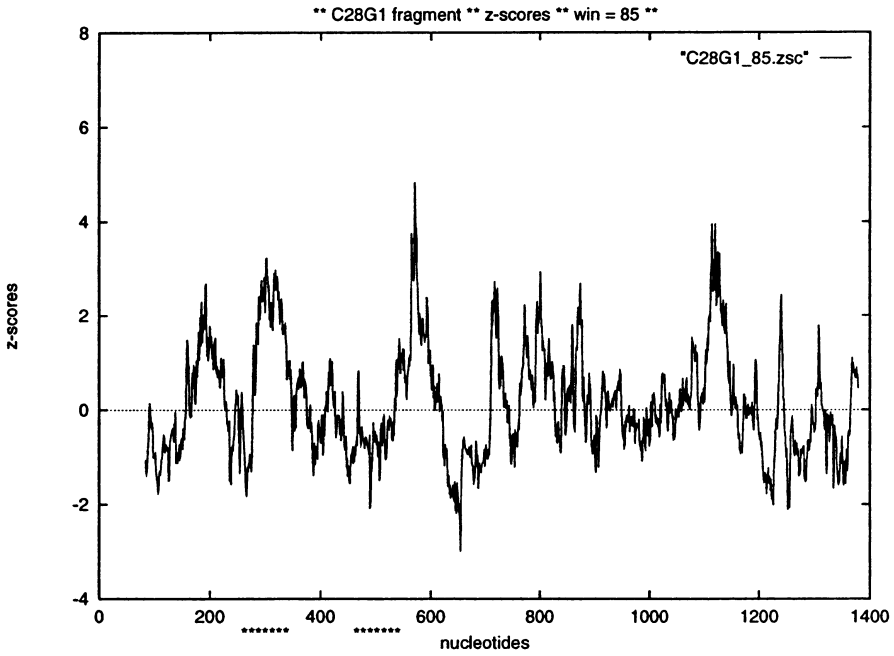


Fig. 9. Negenergy Z-scores for the *C.elegans* clone C28G1, fragment (19877–21256). The scanning window is 85 nucleotides, and we slide one nucleotide at the time. The '****' represent the location of the two tRNAs present in this clone fragment. We calculate Z-scores with respect to 'negenergies' ($-\Delta G$), so that a more positive Z-score indicates allegedly higher thermodynamic stability. Only the Z-score of the second tRNA stands above the background, but the score is at best marginally significant.

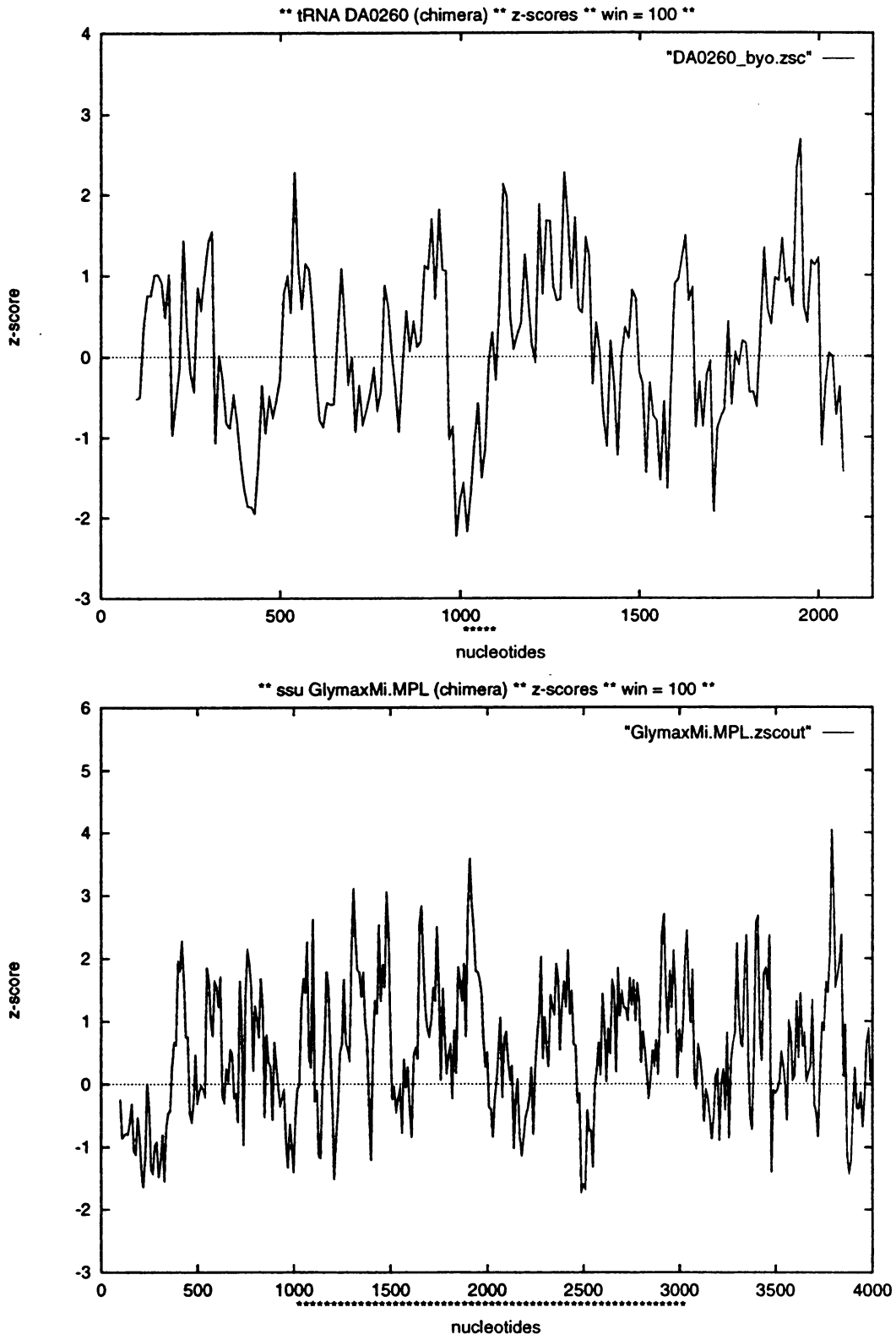


Fig. 10. Negenergy Z-score results for embedding RNA genes in random sequences of identical base composition. These two examples, phage T5 tRNA DA0260 and soy bean mitochondrial SSU, correspond to the examples presented for the probabilistic model, also with negative results.

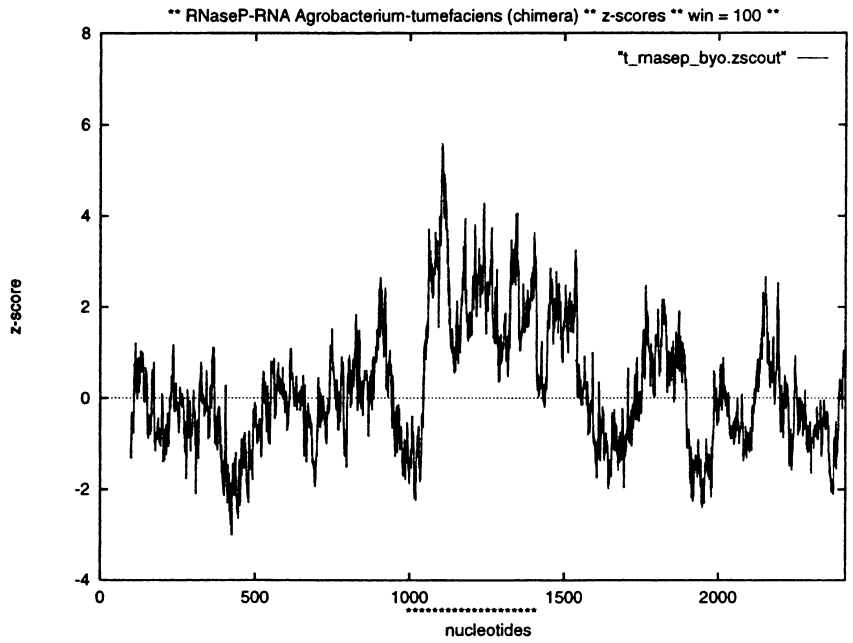


Fig. 11. Negenergy Z-score results for embedding the RNaseP-RNA in *Agrobacterium tumefaciens*—402 nucleotides, %A = 23.6, %C = 27.6, %G = 35.5, %U = 13.4—in random sequence of identical base composition. This figure illustrates a case in which parts of the embedded ncRNA produce Z-scores larger than 4 units, which can be detected above the background of the same base composition. This result is consistent with our estimated lower limit of Z-score 4 units for significance.

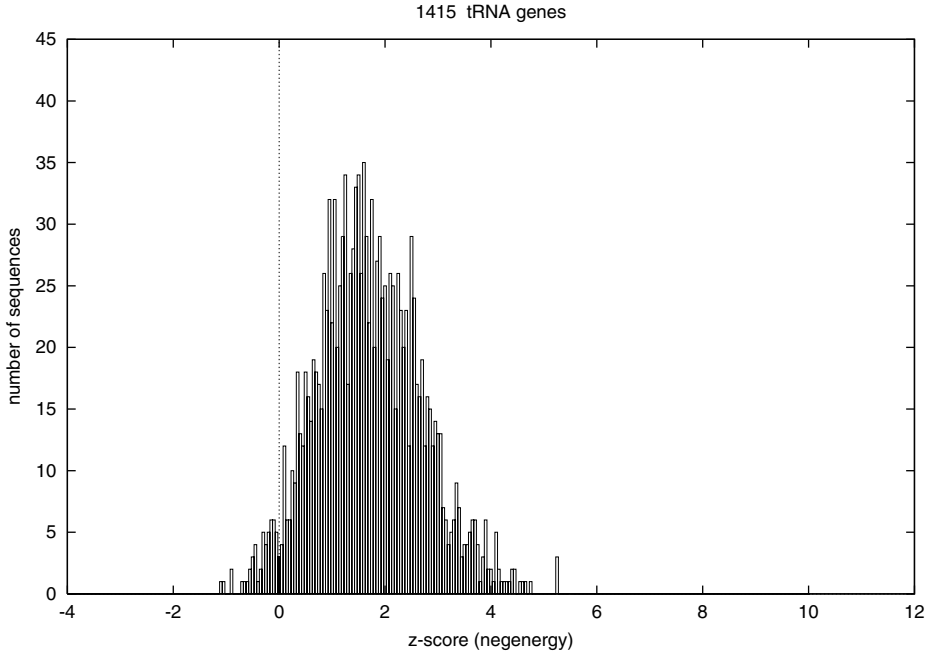


Fig. 12. Histogram of negenergy Z-scores for a collection of 1 415 tRNA genes [Sprinzl database (Steinberg *et al.*, 1993)]. The distribution has a median of $\simeq$ 1.65. Only 1.8% of the tRNAs have Z-scores larger than 4.0.

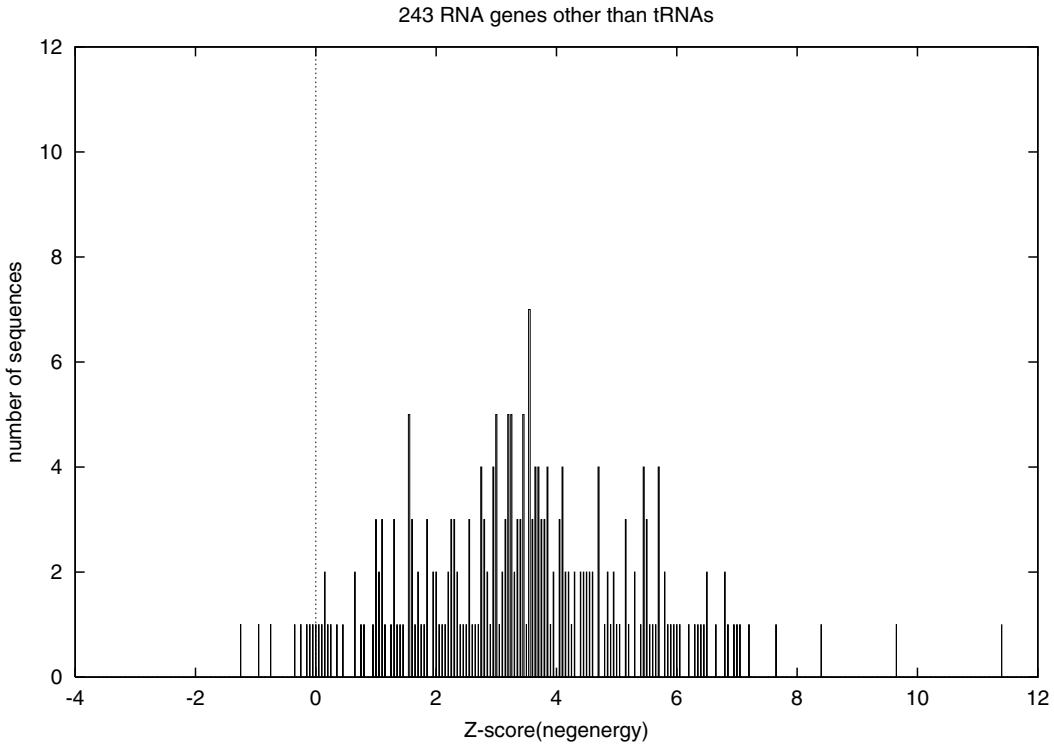


Fig. 13. Histogram of negenergy Z-scores for a collection of 243 RNA genes other than tRNAs. The set includes: 67 SRPs, 18 group I introns, 36 U2s, 104 RNaseP-RNAs, and 18 telomerase RNAs. The density distribution has a median of $\simeq 3.35$. In this case 29.7% of the RNA genes have Z-scores larger than 4.0.

Z-scores for energies

Our tests seem to indicate that most biologically relevant RNA structures are at best marginally detectable by energy Z-scores. For instance, if we apply the thermodynamic algorithm to the *C.elegans* clone C28G1, using a window of 85 nucleotides, the two tRNAs (which have Z-scores $\sim$ 3–4) are hardly detected above the background signals (see Figure 9). If we repeat the chimeric sequence experiment of Section **Chimeric sequences** we observe (similar to the probabilistic model) that relevant RNA genes cannot be distinguished from a background of the same base composition (see Figure 10). It requires Z-scores of about 5 units, such as in the RNaseP-RNA of *Agrobacterium tumefaciens*, for the real signal to start to be reliably detected above the background (Figure 11).

An approximate statistical estimate of the significance of Z-scores indicates that we should not trust any Z-score that is lower that $\sim$4 (see Section **Approximate estimate of the statistical significance of Z-scores**). This conservative estimate of the Z-score cut-off is consistent with the previous results, in which we could not discriminate above background the negenergy Z-scores for RNA genes that were below that lower limit of 4 units.

We have also calculated the Z-scores corresponding to a collection of known RNA genes. Unfortunately, as we observe in Figure 12, the majority of tRNA genes ($\sim$98%) have lower Z-scores than the cut-off. For a collection of other ncRNAs, only 30% show significant secondary structure above the cut-off (Figure 13). Similarly, results presented for mRNA Z-scores (Seffens and Digby, 1999) also fall for the most part below the significance cut-off of Z-score 4 units. Therefore, although there seems to be a slight bias towards real ncRNAs having somewhat more stable structures than randomized sequences of the same base composition, the effect is not strong enough for most ncRNAs to be reliably distinguished from the background of shuffled sequences.

In a real genome, our situation is probably even worse. Using Z-scores as our measure of significance makes the assumption that the genomic background behaves like random sequence, giving Z-scores distributed normally around zero. However, this does not have to be the case. For example, in the course of our experiments we had noticed that many RNA genes have a different base composition than the surrounding genome (cf. C28G1), which made us consider the following kind of artifact: a small CG-rich island, positioned in a larger

(otherwise AT-rich) window. This type of construct (although created with no secondary structure) may have a positive Z-score, just because a Monte Carlo shuffling procedure will destroy the original base-composition inhomogeneity. To test this hypothesis we have generated a set of unstructured sequences with an inhomogeneous base composition. The sequences of this experiment are CG rich in their first half, and AT rich in their second half. We observe in Figure 14 that while sequences with homogeneous base composition (either CG rich or AT rich) have a distribution of Z-scores centered around zero, the set with inhomogeneous base composition is skewed towards positive Z-scores (centered around 0.7). An algorithm which has to use a fixed-length window for a given calculation would be the most sensitive to this heterogeneous base-composition effect (since the ends of a gene are not properly calculated). This fixed-length feature might artificially introduce inhomogeneities within the fixed window used to calculate the Z-scores.

Furthermore, in another recent paper, a second type of statistical artifact is suggested: Workman and Krogh argue that most of the positive Z-scores observed in mRNA stabilities disappear when they are compared with shuffled sequences that preserve dinucleotide composition, rather than simply preserving mononucleotide composition (Workman and Krogh, 1999).

Z-scores for log-odds

After the experiments presented in Section **Results: Probabilistic CFG model** for the probabilistic model, it is clear that most of the signal provided by the probabilistic structural model is due to base composition. Our results indicate that a 'significant' log-odds score (according to the empirical estimation of significance given in Section **Preliminary results**) is only signaling a region that has a base composition bias larger than that expected by sampling from a uniform distribution of nucleotides.

In order to extract any information about secondary structure from log-odds, we have to be able to remove the base-composition bias. A possible way of achieving that goal is to calculate the Z-scores associated with the log-odds scores provided by the model—in the same fashion that we use (10) in the previous section to calculate Z-scores of negenergies.

Similarly to what occurs for Z-scores of negenergies, we observe that log-odds Z-scores of known structural RNAs are only marginally significant (Figures 15 and 16). Furthermore, it requires a Z-score of 5.05 for the CG-rich unstructured sequences (3.45 for the AT-rich sequences) to have only three (out of the 1400) unstructured sequences scoring above that value. We also observe that base-composition heterogeneities shift the curve of log-odds Z-scores for unstructured sequences towards significant values, and that this shift has nothing to do with secondary structure (see Figure 17).

Discussion

Clearly, many functional and catalytic ncRNA genes assume evolutionarily conserved, thermodynamically stable, and well-defined secondary structures in order to fulfill their roles in a cell. However, this is very different from the question of whether these biologically meaningful RNA secondary structures are distinguishable from those spuriously predicted for other sequences (for instance, for randomized sequences, or for nontranscribed sequences in a genome). In this paper we have asked whether RNA secondary structure prediction algorithms could be used for detecting novel noncoding RNA genes against the background of a large genome sequence. We have been reluctantly forced to the general conclusion that, at least when using the current state of the art in secondary structure prediction algorithms, the real RNA secondary structures do not appear to be significantly distinguishable from predicted structures for random sequences, neither by a thermodynamic nor a statistical approach—at least not enough to construct a reliable ncRNA genefinding algorithm based on this idea.

We observed (as had Maizel's group) that base composition inhomogeneities in a genome were a confounding factor in this kind of screen, and a source of artifactually promising scores. Initially, we thought that the probabilistic model would give us a direct measure of statistical significance (the log-odds score), obviating the need for the laborious Monte Carlo estimation of Z-scores in the thermodynamic approach, and thus greatly speeding the scans. Instead, we observed that significant 'RNA structure' log-odds scores resulted from local regions of high CG composition, and we still needed to perform Monte Carlo shufflings to remove the effect of base composition from the log-odds scores.

For both negenergy and log-odds Z-scores, we do observe a slight bias towards positive scores in real ncRNAs. However, the bias is slight, and is only significant when the whole distribution is observed; it is not enough that *individual* RNAs can be reliably distinguished from the background. Furthermore, in Workman and Krogh (1999), it is argued that even this effect mostly goes away when the sequence randomization procedure preserves dinucleotide composition.

The most promising result we obtained is in Figure 13, which indicates that a fraction (about 30%) of a sample of non-tRNA ncRNAs have predicted thermodynamic stabilities that are significant compared to completely randomized sequences. However, our significance threshold ($Z = 4$) is relatively soft; we could expect somewhere around 1–10 false positives

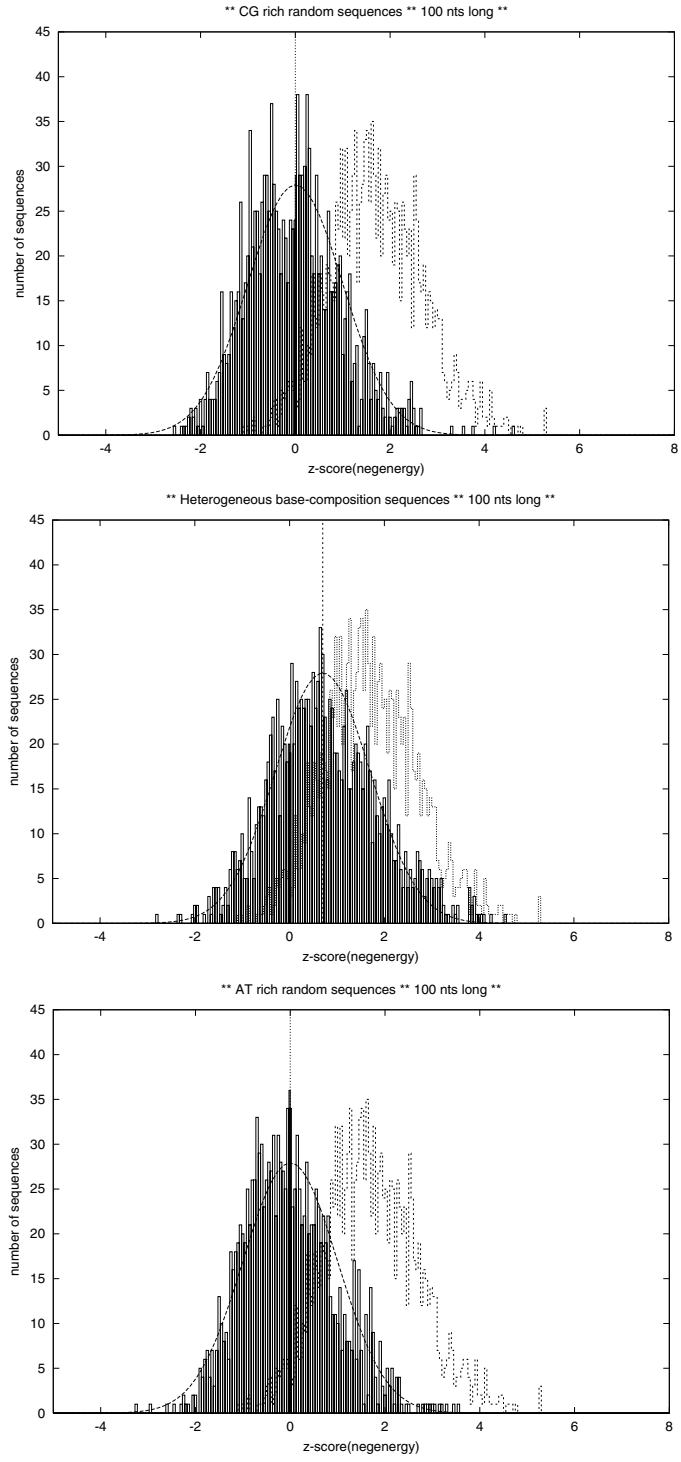


Fig. 14. Comparison of the negenergy Z-score histograms for unstructured sequences of different base compositions (filled area) versus tRNA genes (discontinuous contour). All three cases include $\sim$ 1400 sequences of 100 nucleotides each. The sequences have been generated as follows: top, 68% CG rich, homogeneously distributed; middle, first half is 68% CG rich, and the second half is 68% AT rich; bottom, 68% AT rich, homogeneously distributed. The distribution of Z-scores for unstructured sequences with homogeneous base composition (top and bottom) is centered around 0 (shown fit to a normal $N(0, 1)$ density distribution). In the presence of CG inhomogeneities (middle) the Z-score distribution is shifted towards positive values (shown fit to a $N(0.7, 1)$ density distribution).

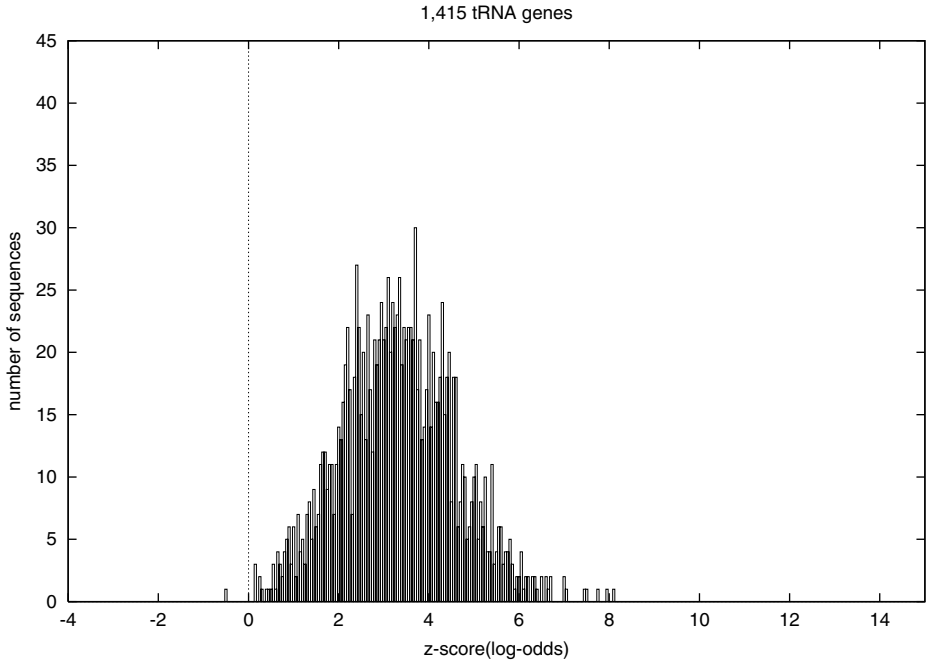


Fig. 15. Histogram of log-odds Z-scores for a collection of 1 415 tRNA genes [Sprinzl database (Steinberg *et al.*, 1993)]. The density distribution has a median of $\simeq$ 3.3. 400 of the 1 415 tRNAs have Z-scores larger than 4.0. Log-odds Z-scores have a better sensitivity to tRNAs than negenergy Z-scores just because tRNAs have been used as part of the training set for the probabilistic model.

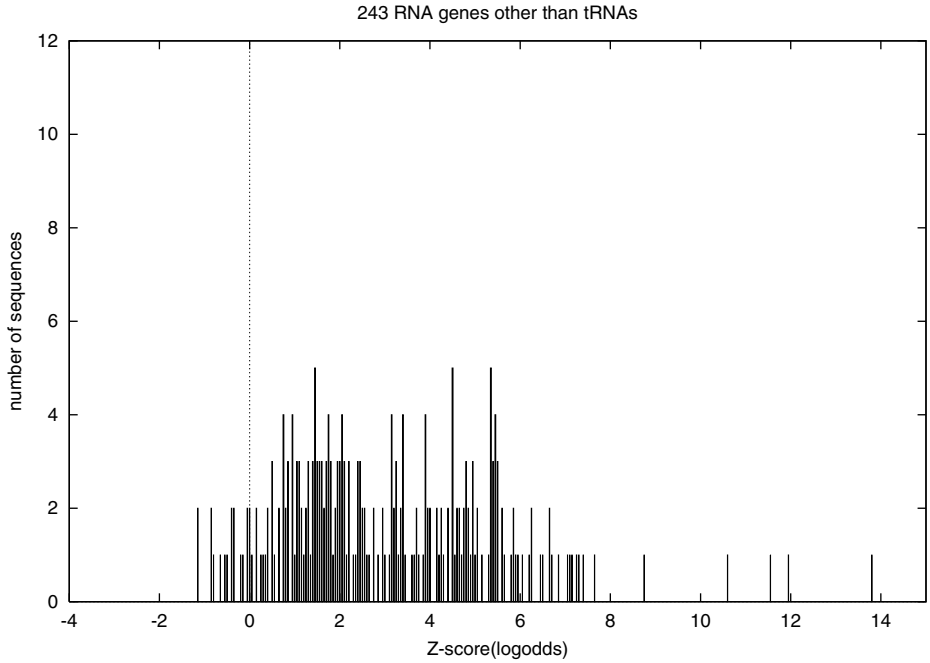


Fig. 16. Histogram of log-odds Z-scores for the collection of 243 non-tRNA RNA genes introduced in Figure 13. In this case 29.6% of the RNA genes have Z-scores larger than 4.0.

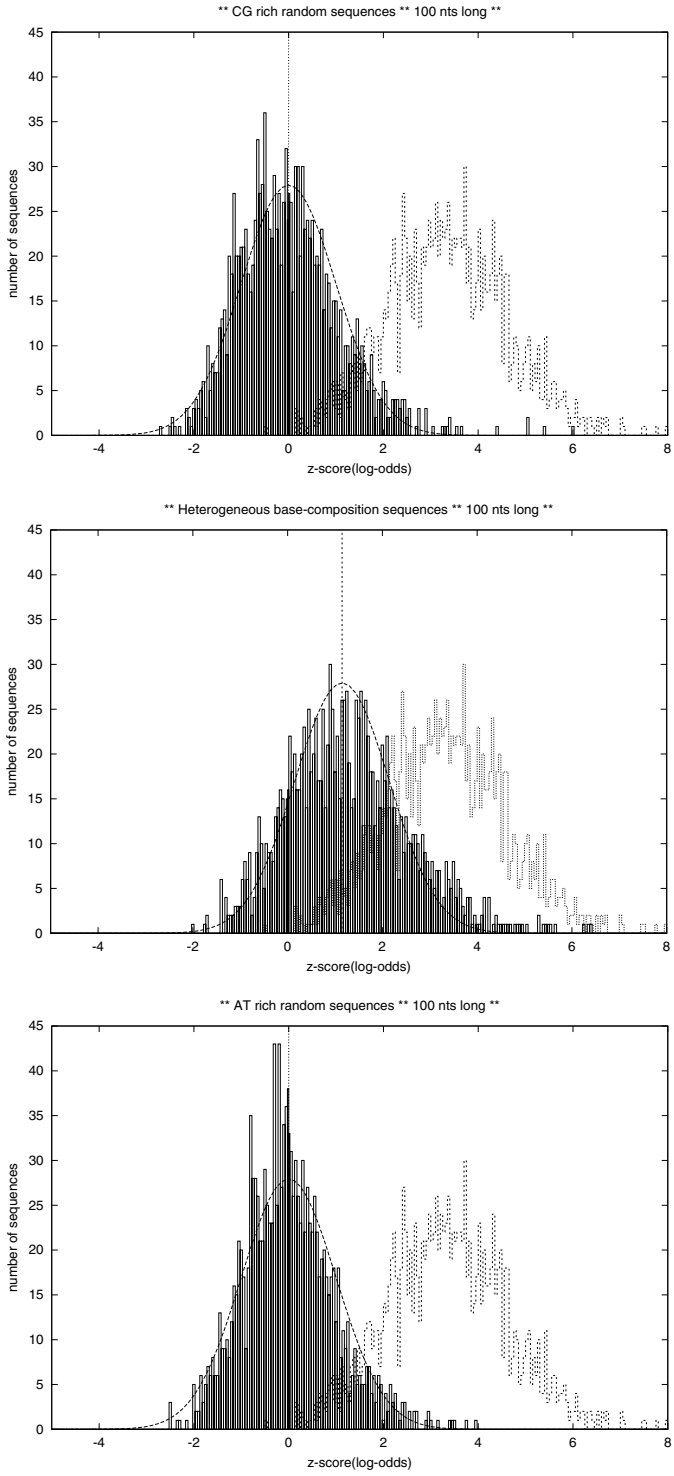


Fig. 17. Comparison of the log-odds Z-score histograms for unstructured sequences of different base compositions (filled area) versus tRNA genes (discontinuous contour). All three cases are identical to the ones described in Figure 14. The distribution of Z-scores for unstructured sequences with a homogeneous base composition (top and bottom) is centered around 0 (shown fit to a normal $N(0, 1)$ density distribution). In the presence of CG inhomogeneities (middle) the Z-score distribution is shifted towards positive values (shown fit to a $N(1.15, 1)$ density distribution).

per megabase at this threshold, depending on how many overlapping windows we examined in a genome scan. Therefore, a Z-score type calculation might still be used to weakly detect a *small subset* of ncRNAs or other biologically interesting RNA structures, consistent with the anecdotal results presented by the Maizel group.

It will require additional sources of statistical information to produce a reliable ncRNA genefinder. One such source of information would be promoter consensus. Even in eukaryotes, where pol II promoters for protein genes are notoriously difficult to recognize, pol II and pol III ncRNA gene promoters are fairly well conserved and information rich. It should be possible to incorporate a probabilistic model of promoter consensus into the our SCFG scanning algorithm.

Ironically, our results suggest the RNA-genefinding potential of simply scanning for CG-rich islands in an AT-rich genome. For AT-rich biased genomes, such as the nematode *C.elegans*, the archaeon *M.jannaschii*, and especially in AT-rich extreme thermophiles, ncRNA genes—probably for reasons of stability—tend to be quite CG rich (Bult *et al.*, 1996). A simple base-composition model that searches for CG-rich regions might extract considerable information about ncRNAs in these biased genomes. We are currently exploring this simple approach.

Acknowledgments

This work was supported by NIH grant HG01363. E.R. acknowledges the support of a postdoctoral fellowship granted by the Sloan foundation.

References

Bachellerie,J.-P., Michot,B., Nicoloso,M., Balakin,A., Ni,J. and Fournier,M.J. (1995) Antisense snoRNAs: a family of nucleolar RNAs with long complementarity to rRNA. *Trends Biochem. Sci.*, **20**, 261–264.

Barrett,C., Hughey,R. and Karplus,K. (1997) Scoring hidden Markov models. *Comput. Applic. Biosci.*, **13**, 191–199.

Baserga,S.J. and Steitz,J.A. (1993) The diverse world of small ribonucleoproteins. In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World.* Cold Spring Harbor Press, New York, pp. 359–381.

Bovia,F. and Strub,K. (1996) The signal recognition particle and related small cytoplasmic ribonucleoprotein particles. *J. Cell Sci.*, **109**, 2601–2608.

Brockdorff,N., Ashworth,A., Kay,G.F., McCabe,V.M., Norris,D.P., Cooper,P.J., Swift,S. and Rastan,S. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, **71**, 515–526.

Brown,C.J., Hendrich,B.D., Rupert,J.L., Lafreniere,R.G., Xing,Y., Lawrence,J. and Willard,H.F. (1992) The human Xist gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized in the nucleus. *Cell*, **71**, 527–542.

Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M. and Venter,J.C. (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science*, **273**, 1058–1073.

Cech,T.R. (1993) Structure and mechanism of the large catalytic RNAs: group I and group II introns and ribonuclease P. In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World.* Cold Spring Harbor Press, New York, pp. 239–270.

Chen,J.-H., Le,S.-Y., Shapiro,B., Currey,K.M. and Maizel,J. (1990) A computational procedure for assessing the significance of RNA secondary structure. *Comput. Applic. Biosci.*, **6**, 7–18.

Dandekar,T. and Hentze,M.W. (1995) Finding the hairpin in the haystack: searching for RNA motifs. *Trends Genet.*, **11**, 45–50.

Delihas,N (1995) Regulation of gene expression by trans-encoded antisense RNAs. *Mol. Microbiol.*, **15**, 411–414.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G.J. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucl. Acids Res.*, **22**, 2079–2088.

Freier,S., Kierzek,R., Jaeger,J., Sugimoto,N., Caruthers,M., Neilson,T. and Turner,D. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, **83**, 9373–9377.

Greider,C.W. and Blackburn,E.H. (1996) Telomeres, telomerase and cancer. *Sci. Am.*, **274**, 92–97.

Gutell,R., Power,A., Hertz,G., Putz,,E. and Stormo,G. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, **20**, 5785–5795.

Le,S.-Y., Chen,J.-H., Currey,K.M. and Maizel,J. (1988) A program for predicting significant RNA secondary structures. *Comput. Applic. Biosci.*, **4**, 153–159.

Le,S.-Y., Chen,J.-H. and Maizel,J. (1989) Thermodynamic stability and statistical significance of potential stem–loop structures situated at the frameshift sites of retroviruses. *Nucl. Acids Res.*, **17**, 6143–6152.

Le,S.-Y., Chen,J.-H. and Maizel,J. (1990) Efficient searches for unusual folding regions in RNA sequences. In Sarma,R.H. and Sarma,M.H. (eds), *Structure and Methods: Human Genome Initiative and DNA Recombination,* Vol 1, Adenine Press, pp. 127–136.

Lefebvre,F. (1996) A grammar-based unification of several alignment and folding algorithms. In Rawlings,C. *et al.*, (eds), *Procceedings, Fourth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, pp. 143–154.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, **25**, 955–964.

Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.

Marra,M.A., Hillier,L. and Waterston,R.H. (1998) Expressed sequence tags–ESTablishing bridges between genomes. *Trends Genet.*, **14**, 4–7.

Maxwell,E. and Fournier,M. (1995) The small nucleolar RNAs.

Ann. Rev. Biochem., **64**, 897–934.

McCaskill,J.S. (1990) The equilibrium partition function and base pair bindings probabilities for RNA secondary structure. *Biopolymers (A5Z)*, **29**, 1105–1119.

Muto,A., Ushida,C. and Himeno,H. (1998) A bacterial RNA that functions as both tRNA and an mRNA. *Trends Biochem. Sci.*, **23**, 25–29.

Nilsen,T.W. (1998) RNA-RNA interactions in nuclear pre-mRNA. In Simon,R. and Grunberg-Manago,M. (eds), *RNA Structure and Function* Cold Spring Harbor Laboratory Press, New York, pp. 279–307.

Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.

Olivas,W.M., Muhlrad,D. and Parker,R. (1997) Analysis of the yeast genome: Identification of new non-coding and small ORF-containing RNAs. *Nucl. Acids Res.*, **25**, 4619–4625.

Ortoleva-Donnelly,L., Szewczk,A.A., Gutell,R.R. and Strobel,S.A. (1998) The chemical basis of adenosine conservation throughout the *Tetrahymena* ribozyme. *RNA*, **4**, 498–519.

Ratner,L., Haseltine,W., Patarca,R., Livak,K., Starcich,B., Josephs,S.F., Doran,E.R., Rafalski,J.A., Whitehorn,E.A., Baumeister,K., Ivanoff,L., Petteway,S.R.Jr, Pearson,M.L., Lautenberger,J.A., Papas,T.S., Ghrayeb,J., Chang,N.T., Gallo,R.C. and Wong-Stall,F. (1985) Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*, **313**, 277–284.

Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

Rosen,C.A. (1991) Regulation of HIV gene expression by RNA-protein interactions. *Trends Genet.*, **7**, 9–14.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, **22**, 5112–5120.

Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structure. *Proc. R. Soc. Lond. B. Biol. Sci.*, **255**, 279–284.

Searls,D.B. (1992) The linguistics of DNA. *Am. Sci.*, **20**, 579–591.

Seffens,W. and Digby,D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucl. Acids Res.*, **27**, 1578–1584.

Sokal,R.R. and Rohlf,F.J. (1981) *Biometry. The Principles and Practice of Statistics in Biological Research* 2nd edn, W.H.Freeman, New York.

Soll,D. and RajBhandary,U.L. (1995) *tRNA: Structure, Biosynthesis, and Function*. ASM Press, Washington.

Steinberg,S., Misch,A. and Sprinzl,M. (1993) Compilation of RNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, **21**, 3011–3015.

Tollervey,D. and Kiss,T. (1997) Function and synthesis of small nucleolar RNAs. *Curr. Opin. Cell Biol.*, **9**, 337–342.

Turner,D., Sugimoto,N., Jaeger,J., Longfellow,C., Freier,S. and Kierzek,R. (1987) Improved parameters for prediction of RNA structure. *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 123–133.

Van de Peer,Y., Van den Broeck,I., De Rijk,P. and De Wachter,R. (1994) Database on the structure of small ribosomal subunit RNA. *Nucl. Acids Res.*, **22**, 3488–3494.

Watanabe,Y. and Yamamoto,M. (1994) *S.pombe* mei2+ encodes and RNA-binding protein essential for premeiotic DNA synthesis and meiosis I, which cooperates with a novel RNA species meiRNA. *Cell*, **78**, 487–498.

Waterman,M.S. (1995) *Introduction to Computational Biology*. Chapman and Hall, London.

Willard,H.F. and Salz,H.K. (1997) Remodeling chromatin with RNA. *Nature*, **386**, 228–229.

Woese,C.R. and Pace,N.R. (1993) Probing RNA structure, function, and history by comparative analysis. In Gesteland,R.F. and Atkins,J.F. (eds), *The RNA World* Cold Spring Harbor Press, New York, pp. 91–117.

Workman,C. and Krogh,A. (1999) No evidence that mRNA have lower folding free energy than random sequences with the same dinucleotide distribution. (in press)

Zimmerman,R.A. and Dahlberg,A.E. (1996) *Ribosomal RNA*. CRC Press, Boca Raton.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, **9**, 133–148.