J. theor. Biol. (2001) **210,** 411–423
doi:10.1006/jtbi.2001.2318, available online at http://www.idealibrary.com on IDEAL®

Local Thermodynamic Stability Scores Are Well Represented by a Non-central Student's *t* Distribution

SHU-YUN LE*†‡, WEI-MIN LIU§‡, JIH-H. CHEN‖ AND JACOB V. MAIZEL JR*

**Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, NIH, Bldg 469, Room 151, Frederick, MD 21702, U.S.A., §Department of Computer and Information Science, Indiana University, 723 W. Michigan St., Indianapolis, IN 46202-3216, U.S.A. and ‖Frederick Biomedical Supercomputing Center, SAIC-NCI/FCRDC, P.O. Box B, Frederick, MD 21702, U.S.A.*

(*Received on* 11 *May* 2000, *Accepted in revised form on* 20 *March* 2001)

Local folding in mRNAs is closely associated w ith biological functions. In this study, we reveal the whole distribution of local thermodynamic stability in the complete genome of the poliovirus P3/Leon/37 and the single-stranded RNA sequences that corresponds to the nucleotide sequence of the complete genome sequence (1 667 867 bp) of *Helicobacter pylori* (*H. pylori*) strain 26695. Local thermodynamic stability in the RNA sequences is measured by two standard *z*-scores, significance score and stability score. To estimate the distribution of thermodynamic stability, a model based on the non-central Student's *t* distribution has been developed. Significant patterns of extremes that are either much more stable or unstable than expected by chance are detected. Our results indicate that the highly stable and statistically more significant folding regions are predominantly in non-coding sequences in the two genome sequences. Moreover, the highly unstable folding regions, on the contrary, are predominantly in the protein coding sequences of *H. pylori*. The observed differences across the complete genomic sequences are statistically very significant by a χ^2-test. These extreme patterns may be useful in searching for target sequences for long-chain antisense RNA and for locating potential RNA functional elements involved in the regulation of gene expression including translation, mRNA localization and metabolism.

Introduction

Functional prediction is an important goal of genomic sequence analysis. Computational analysis of the whole genomic sequence is highly desirable for understanding biological properties that may be useful in drug discovery and vaccine development. Until now, numerous computational approaches have been proposed (Karlin *et al*., 1998; Koonin *et al*., 1996; Bell & Forsdyke, 1999; Andrade & Sander, 1997; Herzel *et al*., 1999; Borodovsky *et al*., 1995; Snyder & Stormo, 1995; Gish & States, 1993; Gelfand *et al*., 1996; Uberbacher *et al*., 1996; Badger & Olsen, 1999; Brutlag, 1998; Bucher, 1999; Frech *et al*., 1997). There has been a steady progress in computational analysis for gene identification; however, the computational methods of predicting the gene regulatory elements in genome sequences are still underdeveloped. It is important to develop

†Author to whom correspondence should be addressed. E-mail: shuyun@orleans.ncifcrf.gov

‡Both authors contributed equally to this work.

0022–5193/01/000411 + 13 $35.00/0

efficient computational methods for elements that are related to the mediation of transcription and translation as well as mRNA localization and metabolism. It is also important to explore the potential biological functions in non-coding sequences since only about 3% of human DNA is involved in protein coding.

Previous studies indicated that some RNA functional elements involved in post-transcriptional regulation are correlated with unusual folding regions (UFRs) in either the coding or non-coding sequences, where the folding free energies of the functional elements are significantly more stable than expected by chance (Malim *et al*., 1989, 1990; Le *et al*., 1988, 1993, 1996; Philips *et al*., 1992; Wang *et al*., 1995). It is desirable to have a suitable statistical model to represent the distribution of thermodynamic stability of local segments in order to reliably detect statistical extremes (distinct UFRs) that are either significantly more stable or unstable than expected by chance in a complete genome sequence or a large set of RNA sequences.

The thermodynamic stability of local segments has been studied in various mRNA sequences (Le *et al*., 1988, 1989; Le & Maizel, 1989; Phillips *et al*., 1992; Forsdyke, 1995; Patzel & Sczakiel, 1997; Walton *et al*., 1999; Seffens & Digby, 1999). In this study, we analyse the complete RNA genome sequence of Poliovirus P3/Leon/37 (PV3L) by two standard *z*-scores, significant score (Sigscr) and stability score (Stbscr) (Le *et al*., 1990). In the analysis of a large number of sample observations for these two scores, we developed a sound statistical model to describe their distributions and to estimate statistical extremes by means of a non-central Student's *t* distribution theory. Our statistical tests indicate that the derived, linearly transformed non-central Student's *t* distribution (LTNSTD) is a good statistical model to describe the distributions of the two scores computed in the PV3L genome sequence. Based on this model, distinct UFRs are inferred in the complete genome sequence. The extremes of statistically significant, stable RNA folding fragments are predominantly in the 5′ non-coding sequences of the PV3L genome. Translational control elements involved in internal ribosome binding in poliovirus and other picornaviruses correlate well with predicted UFRs (Pelletier & Sonenberg, 1988; Le & Zuker, 1990; Le *et al*., 1996).

We also explored the use of this approach to analyse the entire genomic sequence of *Helicobacter pylori* (*H. pylori*) strain 26695. *H. pylori* is one of the most common chronic human pathogens. The genome of *H. pylori* strain 26695 (Tomb *et al*., 1997) consists of a circular chromosome having a size of 1 667 867 nucleotides (nt) that includes about 1590 open reading frames, representing 91% of the genome. It also includes 36 tRNA genes, two separate sets of 23S–5S and 16S rRNA genes, along with one orphan 5S gene and one structural RNA gene, which represent 0.7% of the genome. In this study, single-stranded RNA sequences that correspond to the nucleotide sequence of the complete genome were processed using the same procedure that was used in the PV3L. Our aim in the extended study is to verify if the derived, linearly transformed non-central Student's statistical model from the single-stranded RNA sequence is also suitable for the 1.67 million bp genome sequence.

The thermodynamic stability of DNA folding is different from that of RNA folding. In general, the thermodynamic stability for the stacking of Waston–Crick base pairs in single-stranded DNA is less than that of the corresponding base pair stacking in the single-stranded RNA (SantaLucia, 1998). However, the thermodynamic stability and the statistical significance of the sequence folding as measured by Sigscr and Stbscr are computed from the differences of folding energies rather than absolute values. Therefore, the influence of a few DNA pieces that may be misrepresented as RNAs is very limited. Moreover, we demonstrated that the Sigscr computed in *E. coli* 16S rRNA is not sensitive to different sets of energy rules for computing the lowest free energy of the local folding (Le & Maizel, 1989).

In this study, we describe the use of LTNSTD and show that it is a good model to depict the distribution of the two scores computed in both the single-stranded RNA genome of poliovirus and the *H. pylori* genome. The derived extremes of statistically significant, stable sequences are predominantly in the non-coding sequences, while, statistically significant unstable sequences are predominantly in the protein coding regions.

The difference of observed numbers of extreme patterns in the three separate domains of coding, non-coding and RNA genes is very significant by a χ^2-test.

These extreme patterns can be used in searching gene regulatory elements and potential target sequences for long-chain antisense RNAs. The knowledge of unusual stabilities in RNA may be relevant to interpreting microarray-based gene expression studies. The formation of duplexes occurs by a two-step mechanism (Sczakiel, 1997). First, the structures of complementary regions are opened in a solution in an energy-consuming step. The second step is an energy-releasing process by forming the duplex on the array. Free energy differences between the two steps affect the RNA hybridization/annealing, implying that thermodynamic stability of local folding is important in the binding efficiency.

Methods

Thermodynamic stability of the segments is evaluated by two standard z-scores, significance score (Sigscr) and stability score (Stbscr). Sigscr and Stbscr of an RNA folding are given by

$$Sigscr = (E - E_r)/std_r$$

and

$$Stbscr = (E - E_w)/std_w,$$

where E is the computed lowest free energy of the RNA folding for a given segment, E_r and std_r are the mean and standard deviation, respectively, of the lowest free energies from folding 300 random sequences of the same base composition as the given segment, and E_w and std_w are the mean and standard deviation of the lowest free energies obtained by folding all segments of the same size generated by taking successive, overlapping, fixed length segments stepped successively in one nucleotide (nt) from 5′ to 3′ along the single-stranded RNA sequence (Le *et al.*, 1990). The lowest free energy of the RNA folding is computed for each segment by Zuker's (1994) algorithm using the Turner energy rules (Jaeger *et al.*, 1989; Freier *et al.*, 1986).

In this study, Sigscr and Stbscr are computed by the program SEGFOLD (Le *et al.*, 1990) using a fixed window. To speed up the computation of Sigscr, the mean and standard deviation (std) of the lowest free energies from 300 randomly shuffled sequences, are computed from tabulated coefficients based on window sequence length and base composition if the percent content of base G+C in the window is less than 75% and each base percentage is larger than 3%. Otherwise, E_r and std_r are computed from 100 randomly shuffled versions of the sequences. These tabulated coefficients were derived from the random simulations and least-squares fits (Chen *et al.*, 1990) using the Turner energy rules. Statistical analyses for the Sigscr and Stbscr data are computed using the Statistics Toolbox of MATLAB software package (http://www.mathworks. com).

The sequence and gene structure data of PV3L (accession number K01392) and *H. pylori* (accession number NC_000915) are obtained from the Genome database of the National Center for Biotechnology Information (NCBI). PV3L is a member of the Picornaviridae family that contains a positive sense single-stranded RNA genome of 7431 nucleotides (nt). The complete genome of PV3L contains a 742 nt 5′ non-coding sequence, a 6621 nt protein-coding sequence and a short 3′ non-coding sequence of 68 nt. On the contrary, *H. pylori* genome contains 1553 mRNA-coding regions having a total size of 1 479 387 bp (Tomb *et al.*, 1997). Among them, 25 protein-coding regions are shorter than 100 nt and their total length is 2034 bp. The structural RNA-coding regions contain 43 RNA genes and have a total size of 12 065 bp. Among them, 36 tRNA genes are shorter than 100 nt and their total length is 2729 bp. In the study, other regions that are not listed in the protein- and RNA-coding region are considered as non-coding regions. In the analysis of thermodynamic stability of a local segment, the local segment is defined in the coding and non-coding regions as well as in the RNA gene only if the corresponding window is involved entirely within these regions, respectively.

Results and Discussion

NON-CENTRAL STUDENT'S t DISTRIBUTION

The scores Sigscr and Stbscr computed by the fixed window of 100- nt along the genomic

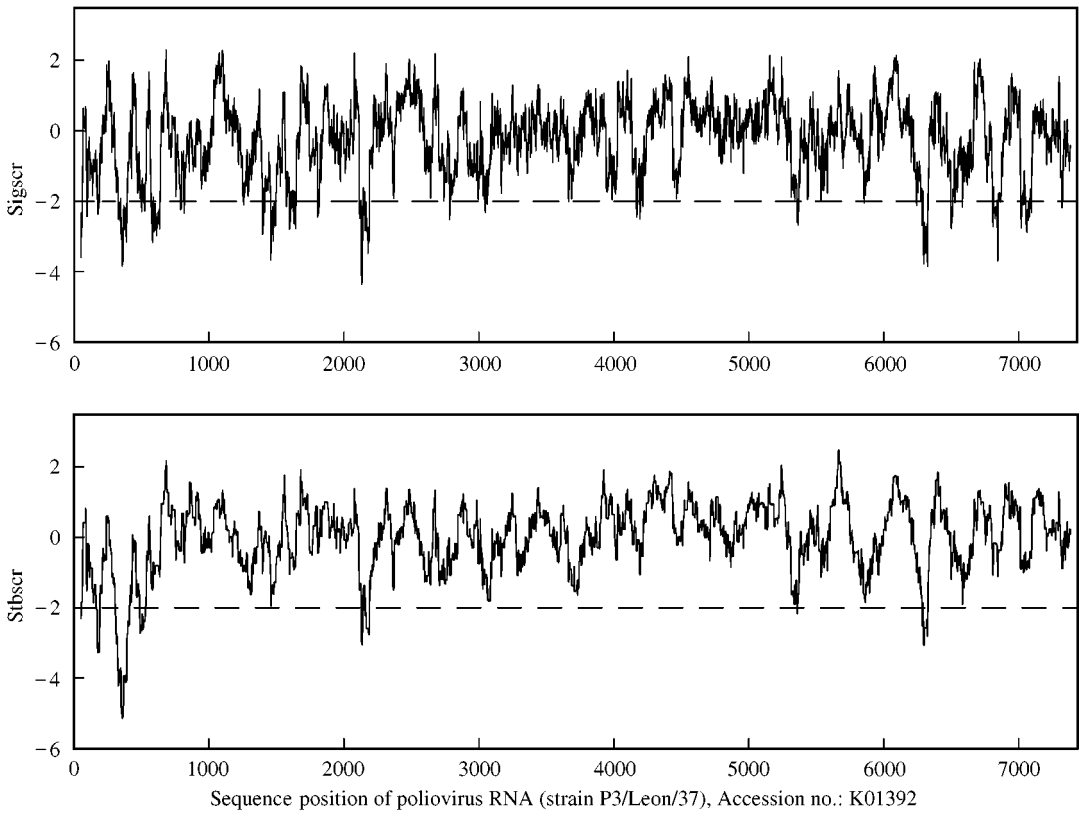

FIG. 1. Significance score (Sigscr) and stability score (Stbscr) computed in the complete poliovirus RNA sequence (PV3L). The profile was produced by plotting the Sigscr (top) and Stbscr (bottom) of 100 nt segments against the position of the middle base in the window as it slides along the sequence. The window size is 100 nt.

sequence of PV3L is displayed in Fig. 1. Statistics of local thermodynamic stability in the PV3L RNA sequence are listed in Table 1. The asymmetric distributions of Sigscr and Stbscr in the PV3L sequence are shown in Fig. 2. It is clear that the distributions of Sigscr and Stbscr do not follow the normal distribution because of the skewness of samples for Sigscr and Stbscr data in PV3L viral RNA sequence. Since nearby scores in these data are not fully independent, we take two random samples with size of 200 and 500 observations so that the distance between two neighboring observations in the corresponding random sample is larger than or equal to 12 and 30 nt, respectively. The sample means, sample standard deviations (std), and sample coefficients of skewness of these random observations are also listed in Table 1.

These scores should be described by an asymmetric continuous distribution with range $(-\infty, \infty)$. Simple normal distributions are unsatisfactory. The non-central Student's t

TABLE 1

Statistics of RNA folding scores computed in the complete RNA genome of the poliovirus P3/Leon/37 (PV3L). Sigscr and Stbscr were computed from the complete viral RNA of PV3L using a fixed window of 100 *nt. The two random samples were constructed by randomly selecting* 500 *and* 200 *observations from the complete* 7332 *observations of PV3L so that the distance between the two neighboring points is larger than or equal to* 12, *and* 30 *nt, respectively*

Score	Sample size (N)	Mean	Std.	Skewness
Sigscr	7332	− 0.269	1.081	− 0.476
Sigscr	500	− 0.256	1.097	− 0.399
Sigscr	200	− 0.276	1.131	− 0.362
Stbscr	7332	0.000	1.000	− 1.118
Stbscr	500	− 0.021	1.024	− 1.073
Stbscr	200	+ 0.013	0.980	− 0.934

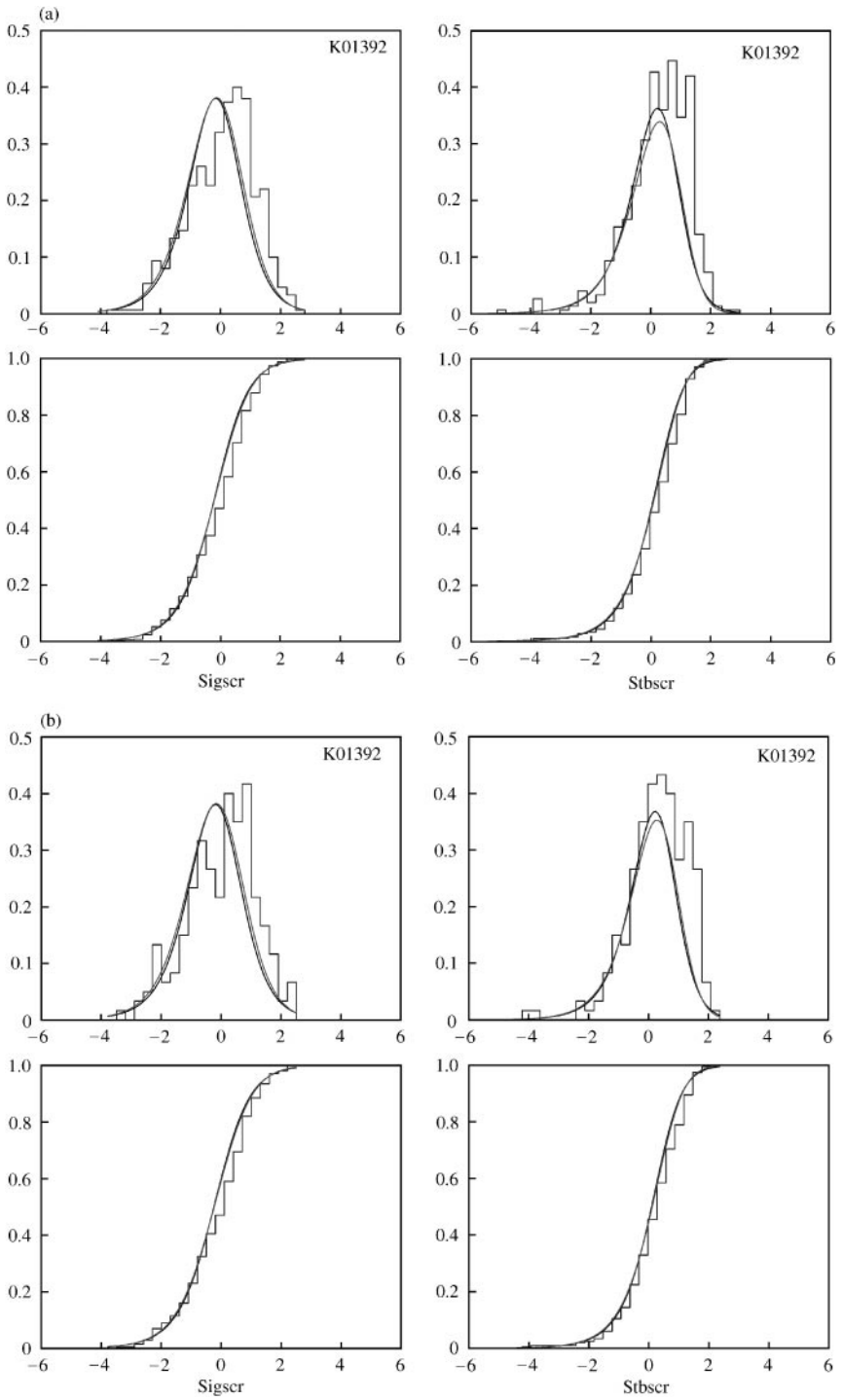

FIG. 2. Empirical probability density functions (top) and empirical distribution functions (bottom) plotted together with linearly transformed, theoretical probability density functions and cumulative distribution functions of non-central t distribution. (a) Sigscr and Stbscr data from the random sample ($N = 500$) of PV3L. (b) Sigscr and Stbscr data from the random sample ($N = 200$) of PV3L. For the selection of the two random samples see the text. In the plot, the horizontal axis represents `Sigscr` on the left, and `Stbscr` on the right. The vertical axis is for probability density functions (top) or distribution functions (bottom). The empirical step functions are plotted with step size 0.3. The theoretical curves with the degree of freedom $f = 6$ almost coincides with the curve derived by $f = 8$, as shown in the plot.

distribution satisfies this condition. Moreover, its moments can be calculated in closed analytic forms. Therefore, it is possible to fit the distribution to the data with the moment computations. The non-central t distribution has two parameters (Evans *et al.*, 1993): the degree of freedom, f (a positive integer), and the non-centrality parameter, δ (a real number). It is denoted by $\mathbf{t}: f, \delta$. Its probability density function (p.d.f.) (Lehmann, 1959) can be expressed as

$$p(x; f, \delta) = \frac{1}{2^{(f+1)/2}\Gamma(f/2)\sqrt{\pi f}} \int_0^\infty y^{(f-1)/2} \times \exp\left[-\frac{1}{2} y - \frac{1}{2}\left(x\sqrt{\frac{y}{f}} - \delta \right)^2 \right] \mathrm{d}y. \quad (1)$$

Its rth moment about the origin (Evans *et al.*, 1993) is

$$\mu'_r = \left(\frac{f}{2}\right)^{r/2} \frac{\Gamma((f-r)/2)}{\Gamma(f/2)} \sum_{j=0}^{r/2} \binom{r}{2j} \times \frac{(2j)!}{2^j j!} \delta^{r-2j} \quad (f > r). \quad (2)$$

To simplify the notation, we introduce

$$g(f) = \frac{\Gamma((f-1)/2)}{\Gamma(f/2)}. \quad (3)$$

Using these formulae, we can write the mean $\mu(f, \delta)$, variance $\sigma^2(f, \delta)$, and coefficient of skewness $\eta_3(f, \delta)$ as

$$\mu(f, \delta) = \sqrt{\frac{f}{2}}\, g(f)\, \delta \quad (f > 1), \quad (4)$$

$$\sigma^2(f, \delta) = \frac{f}{f-2}\left[1 + \delta^2 \left(1 - \frac{f-2}{2} g^2(f) \right) \right] \quad (f > 2), \quad (5)$$

$$\eta_3(f, \delta) = \frac{\mu_3(f, \delta)}{\mu_2^{3/2}(f, \delta)} = \sqrt{\frac{f-2}{2}} \times \frac{g(f)\,\delta\,[3 + (f-2)\delta^2((f-3)g^2(f) - (2f-7)/(f-2))]}{(f-3)[1 + \delta^2(1 - (f-2)/2\ g^2(f))]^{3/2}}, \quad (f > 3). \quad (6)$$

In the above expressions, μ_r is the rth (central) moment about the mean, i.e.

$$\mu_r(f, \delta) = \int_{-\infty}^{\infty} (x - \mu(f, \delta))^r\ p(x; f, \delta)\, \mathrm{d}x.$$

The third central moment may also be obtained from the formula (Evans *et al.*, 1993)

$$\mu_3(f, \delta) = \mu'_3(f, \delta) - 3\mu'_2(f, \delta)\,\mu(f, \delta) + 2\mu^3(f, \delta), \quad (7)$$

where

$$\mu'_2(f, \delta) = \frac{f}{f-2}(1 + \delta^2), \quad (8)$$

$$\mu'_3(f, \delta) = \sqrt{\frac{f}{2}} \frac{f g(f)}{f-3} \delta(3 + \delta^2). \quad (9)$$

ESTIMATION OF PARAMETERS

Although eqn (4) shows that the non-centrality parameter δ is proportional to the mean of the non-central t distribution, we should not directly use the sample mean to estimate δ, because the mean can be easily changed with a linear transformation. To catch the essence of asymmetry of the sample distribution, we use the sample coefficient of skewness that is invariant under any linear transformation.

Let the observed data be $\{y_i\}$ $(i = 1, \ldots, n)$. Consider a linear transformation:

$$y_i = a x_i + b, \quad a > 0, \quad (10)$$

Let x_i be distributed as $\mathbf{t}: f, \delta$. We estimate the parameters a, b and δ by assuming that the sample mean, sample variance and sample coefficient of skewness are equal to the mean, variance and coefficient of skewness of the distribution for a given degree of freedom f. We can then vary f and choose the values satisfying the Kolmogorov–Smirnov test. Let the sample mean of y_i be $\bar{y}$, the sample standard deviation be s_y, and the sample coefficient of skewness, k, be

$$k = \frac{\sum_{i=0}^{n} (y_i - \bar{y})^3}{n s_y^3}. \quad (11)$$

The three equations are

$$\frac{\bar{y}-b}{a}=\sqrt{\frac{f}{2}}\,g(f)\,\delta, \tag{12}$$

$$\frac{s_y^2}{a^2}=\frac{f}{f-2}\left[1+\delta^2\left(1-\frac{f-2}{2}\,g^2(f)\right)\right], \tag{13}$$

$$k=\sqrt{\frac{f-2}{2}}$$

$$\times\frac{g(f)\,\delta[3+(f-2)\delta^2((f-3)\,g^2(f)-(2f-7)/(f-2))]}{(f-3)\,[1+\delta^2(1-(f-2)/2\;\;g^2(f))]^{3/2}}. \tag{14}$$

For a given degree of freedom f, we solve eqn (14) to get δ, then substitute the value of δ into eqn (13) to obtain a and then substitute the values of a and δ into eqn (12) to obtain b.

We remark that according to our computation, the right-hand side of eqn (14) is an increasing function of δ. When δ approaches infinity, it tends to the limit

$$L(f)=\sqrt{\frac{f-2}{2}}$$

$$\times\frac{g(f)\,(f-2)((f-3)\,g^2(f)-(2f-7)/(f-2))}{(f-3)\,(1-(f-2)/2\;\;g^2(f))^{3/2}}. \tag{15}$$

When δ approaches negative infinity, the right-hand side of eqn (14) tends to the limit $-L(f)$. Therefore, when the sample coefficient of skewness is strictly between $-L(f)$ and $L(f)$, there is a unique solution of eqn (14) for δ. The sample coefficient of skewness and δ have the same sign.

Moreover, for fixed δ, the absolute value of the right-hand side of eqn (14) is a decreasing function of f. Thus, for a given sample coefficient of skewness, when f increases, the absolute value of δ satisfying eqn (14) will increase. We will see that this property leads to the useful conclusion that the quantile displacement from the mean of a sample is not sensitive to the choice of degree of freedom f.

KOLMOGOROV–SMIRNOV TEST

The Kolmogorov–Smirnov test can show the goodness of fit between a theoretical cumulative distribution function $F(x)$ and an empirical distribution function $F_n(x)$, where n is the sample size. In our case, $F(x)$ is the non-central t distribution function. Let the data be $y_1, y_2, \ldots, y_n$. Sort them in the ascending order to obtain $z_1 \leqslant z_2 \leqslant \cdots \leqslant z_n$. The empirical distribution function (Hogg & Tanis, 1997) is defined as

$$F_n(x)=\begin{cases}0, & x<z_1,\\ i/n, & z_i\leqslant x<z_{i+1},\\ 1, & z_n\leqslant x.\end{cases} \tag{16}$$

The Kolmogorov–Smirnov statistics (KS) is defined to be

$$D_n=\sup_x\,(|F_n(x)-F(x)|). \tag{17}$$

If D_n is sufficiently small, we may consider the sample is well described by the proposed distribution function $F(x)$.

In Table 2, we list the KS statistics for the Sigscr and Stbscr of PV3L for the random sample that includes 500 observations ($N=500$). From Table 2, we can see that the error is minimal when the degree of freedom (f) is 14 and 10, respectively. However, the errors are not very sensitive to f values. In fact, all of these KS statistics are less than 0.0729, the acceptance limit with the significance level 0.01, or 0.0608, the acceptance limit with the significance level 0.05. For data from another random sample

TABLE 2

Kolmogorov–Smirnov statistics computed in the random sample ($N=500$) *for window size of* 100 *nt*

f	KS	f	KS	f	KS
(a) Sigscr					
6	0.0622	7	0.0559	8	0.0515
9	0.0483	10	0.0458	11	0.0439
12	0.0423	13	0.0417	14	0.0412
(b) Stbscr					
6	0.0481	7	0.0429	8	0.0392
9	0.0365	10	0.0344	11	0.0349
12	0.0354	13	0.0359	14	0.0363

($N = 200$), we obtain similar results. The Kolmogorov– Smirnov test indicates that the theoretical noncentral t distribution has a good fit to PV3L data. The linearly transformed, theoretical probability density functions and cumulative distribution functions of non-central t distribution of Sigscr and Stbscr data are displayed in Fig. 2.

QUANTILE DEVIATION AND APPROXIMATION WITH MULTIPLE REGRESSION

Let the quantile, q_α, with probability α in a non-central t distribution be

$$q_\alpha = \mu + c_\alpha \sigma, \quad \text{where } P(x \leqslant q_\alpha) = \alpha, \qquad (18)$$

where μ and σ are, respectively, the mean and std of the distribution. The coefficient c_α in eqn (18) is the quantile deviation from the mean in units of the std. It depends on the probability α, the degree of freedom f, and the sample coefficient of skewness k (which determines the non-centrality parameter δ). When the sample coefficient of skewness k is 0, the quantile q_α and the quantile deviation c_α should be equal to those of symmetric t distribution. Therefore, we have

$$c_\alpha(f,k) = c_\alpha^0(f) + d_\alpha(f,k)\,k. \qquad (19)$$

Using multiple regression for f between 6 and 25, and coefficients of skewness with step size 0.1, we find an approximate formula (not shown). Table 3 shows the theoretical values of c_α where the degree of freedom f is 12 and the coefficient of skewness k is equal to -0.5, as well as the corresponding values calculated with the regression formula.

STATISTICS OF THERMODYNAMIC STABILITY OF LOCAL SEGMENTS IN THE RNA GENOME OF POLIOVIRUS, PV3L

Extreme UFRs, with very low Sigscr values, are much more stable than expected by chance. Similarly, extremely low Stbscr values indicate local sequences where the folded structures are more stable than the average computed from all segments with the same size over the entire genome sequence. We have indicated that the distribution of Sigscr and Stbscr in the PV3L genome can be fitted well by the derived LTNSTD as shown in Fig. 2. Based on the theoretical LTNSTD, we detected the extreme UFRs, type 1 UFR that had very low Sigscr ($\leqslant -3.701$), and type 3 UFR that had low Stbscr ($\leqslant -3.748$). The likelihood of the type 1 and type 3 UFRs occurring in the complete PV3L are less than or equal to 0.005 by chance, respectively. There were six type 1 UFRs and 64 type 3 UFRs in the 5′ non-coding region of the PV3L genome. Eleven type 1 UFRs were detected and no type 3 UFRs were found in the protein-coding region. We can expect that the highly stable and more statistically significant folding patterns are predominantly in the 5′ non-coding region in the PV3L genome. To validate our results from another aspect, we did an exact test for 2×2 table (Bailey, 1995). The relevant probability is very significant, $p = 0.000$. Therefore, we can detect a strong expectation for highly stable UFRs in the 5′ non-coding region of the PV3L genome. It was indicated previously that a UFR in the 5′ non-coding region of poliovirus RNA was correlated to the regulation of translational initiation (Pelletier & Sonenberg, 1988; Le & Zuker, 1990; Le *et al.*, 1996).

TABLE 3

Quantile deviations from the mean in units of std for the non-central t distribution with degree of freedom $f = 12$ and coefficient of skewness $k = -0.5$

α	0.0010	0.0025	0.005	0.010	0.025	0.050
c_α	− 4.24	− 3.63	− 3.19	− 2.75	− 2.17	− 1.73
Predicted c_α	− 4.27	− 3.66	− 3.20	− 2.73	− 2.12	− 1.64
Relative error	0.008	0.008	0.003	0.006	0.026	0.050
α	0.9990	0.9975	0.995	0.990	0.975	0.950
c_α	2.90	2.60	2.36	2.12	1.79	1.51
Predicted c_α	2.97	2.62	2.35	2.08	1.73	1.44
Relative error	0.024	0.008	0.004	0.018	0.034	0.042

STATISTICS OF THERMODYNAMIC STABILITY OF LOCAL SEGMENTS IN *H. PYLORI*

Statistics of local thermodynamic stability in *H. pylori* genomic sequences are listed in Table 4. The distribution of Sigscr and Stbscr computed along the genomic sequence shows an asymmetric distribution (Fig. 3) in which the distribution of Sigscr is biased towards the negative direction more seriously than that of Stbscr. Since nearby scores in these data are not fully independent, we take three random samples with size of 10 000, 5000 and 3000 for data computed by fixed windows of 100, 300 and 500 nt, respectively, so that the distance between two neighboring points in the statistical sample is larger than or equal to the window size. This step is useful for our later application of the Kolmogorov–Smirnov test for independent observations (Hogg & Tanis, 1997). Table 5 lists the sample means, sample standard deviations, and sample coefficients of skewness of these random observations. In the random sample, Stbscr data show asymmetry for window size 100, but no asymmetry, with mean = 0 and std = 1.0, for window sizes 300 and 500.

Using the same procedure as we employed in the PV3L genome, we also derived the theoretical LTNSTD. In Table 6, we list the KS statistics for the Sigscr of *H. pylori* with window size 300. From Table 6, we can see that the error is minimal when the degree of freedom (f) is 13. However, the errors are not very sensitive to f values. In fact, most of these KS statistics are less than 0.0231, the acceptance limit with the significance level 0.01, or 0.0192, the acceptance limit with the significance level 0.05. For other data of *H. pylori* with different window sizes, we obtain similar results. The Kolmogorov–Smirnov test indicates that the theoretical LTNSTD has a good fit to our data. The linearly transformed, theoretical probability density functions and cumulative distribution functions of non-central t distribution of Sigscr data are displayed in Fig. 3.

Since we have a large set of observations ($N = \sim 1.67$ million) in the complete *H. pylori* genome, we first focus our attention on the statistical extreme of UFRs. Based on the values of Sigscr and Stbscr for the local segment we define eight types of UFR, termed as type 1–8. For example, a type 1 UFR is defined only if its Sigscr is less than or equal to -4.94. Thus, type 1 UFR consists of more stable folding fragments with respect to their randomly shuffled sequences. Similarly, type 2 UFR refers to the fragment with $Sigscr \geqslant 2.52$. Type 2 UFR signifies less stable folding fragments with respect to their randomly shuffled sequences. Based on the derived LTNSTD, the probabilities of type 1 and type 2 UFRs occurring in the complete *H. pylori* genome are less than or equal to 0.001 by chance, respectively. If both $Sigscr \leqslant -4.94$ and $Stbscr \leqslant -3.70$ the UFR is defined as type 5 UFR. Type 5 UFRs represent fragments that are significantly more stable than both their randomly shuffled sequences and other fragments with the

TABLE 4

Statistics of RNA folding scores computed in H. pylori. *Significance score* (*Sigscr*) *and stability score* (*Stbscr*) *that represent local thermodynamic stability were computed from single stranded RNAs corresponding to the nucleotide* (*nt*) *sequence of the complete* H. pylori *genome by three fixed windows of* 100, 300 *and* 500 *nt. In the computation, the three fixed windows were moved by a nt each time along* H. pylori *sequence so that three samples of two scores were collected, respectively*

Score	Window size (bp)	Sample size	Mean	Std.	Skewness
Sigscr	100	1 667 768	− 0.469	1.110	− 0.655
Stbscr	100	1 667 768	0.000	1.000	− 0.277
Sigscr	300	1 667 568	− 1.125	1.267	− 0.895
Stbscr	300	1 667 568	0.000	1.000	− 0.022
Sigscr	500	1 667 368	− 1.702	1.463	− 1.057
Stbscr	500	1 667 368	0.000	1.000	+ 0.030

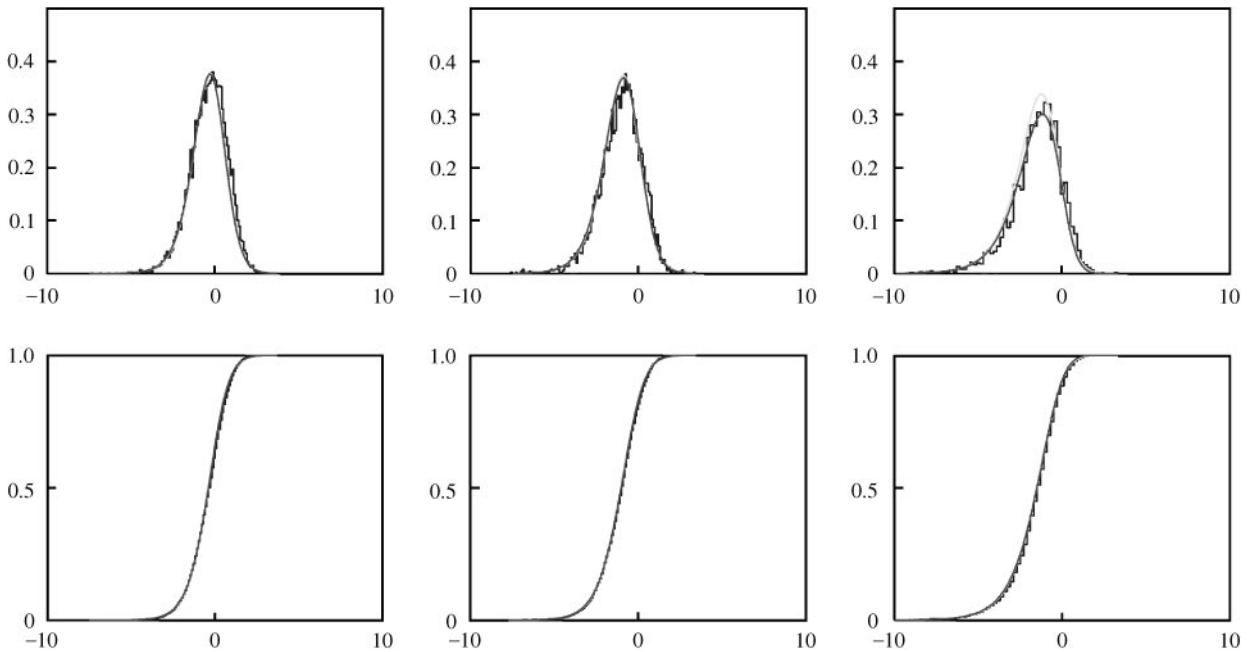

FIG. 3. Empirical probability density functions (top) and empirical distribution functions (bottom) plotted together with linearly transformed, theoretical probability density functions and cumulative distribution functions of noncentral t distribution. The left, middle and right graphs are, respectively, for `Sigscr` data from *H. pylori* with window sizes 100, 300 and 500 nt. The theoretical curves are not very sensitive to the degree of freedom, f. The two curves derived by $f = 8$ and 10, almost coincide as shown in the plot. The empirical step functions are plotted with step sizes 0.1, 0.1 and 0.2 for the left, middle and right graphs. In every figure, the horizontal axis represents `Sigscr`, and the vertical axis is for probability density functions (top) or distribution functions (bottom).

TABLE 5

Statistics of RNA folding scores computed in random samples from H. pylori. *The three random samples corresponding to the three fixed windows of* 100, 300 *and* 500 *nt were generated from the three samples mentioned in Table* 4 *by randomly selecting* 10 000, 5000 *and* 3000 *points, respectively. As a result, the distance between two neighboring points in each random sample is larger than or equal to the window size*

Score	Window size (bp)	Sample size	Mean	Std.	Skewness
Sigscr	100	10 000	− 0.466	1.102	− 0.489
Sigscr	300	5000	− 1.110	1.241	− 0.592
Sigscr	500	3000	− 1.709	1.492	− 1.077
Stbscr	100	10 000	+ 0.003	1.002	− 0.276

same size. The locations of these distinct UFRs from type 1 to type 8 are sorted by means of gene information listed in the Feature Table of *H. pylori* compiled in the database. Table 7 lists counts of the extreme UFR patterns detected in the protein coding, RNA gene, and non-coding regions, respectively, in the complete genomic sequence. As shown in Table 7, 526 type 1 UFRs of length 100 nt are involved entirely within protein coding regions, three type 1 UFRs of length

TABLE 6
Kolmogorov–Smirnov statistics of Sigscr computed in H. pylori *for window size of* 300 *nt*

f	KS	*f*	KS	*f*	KS
6	0.0252	7	0.0196	8	0.0156
9	0.0126	10	0.0103	11	0.0094
12	0.0088	13	0.0084	14	0.0096
15	0.0106	16	0.0114		

100 nt are within RNA genes, and 1006 type 1 UFRs are within non-coding regions. Also, there are 436, three, and 141 type 2 UFRs of length 100 nt in the protein coding, RNA gene, and non-coding regions, respectively.

Furthermore, we can construct a series of contingency tables with only two rows from Table 7. For example, we can make the contingency table using the data listed in the three right most columns in Table 7. Counts of extreme UFRs, type 1 and 2 detected in the protein coding, RNA gene, and non-coding regions are listed in each row of the contingency tables, respectively. We find $\chi^2 = 288.3$ and 310.1, with two degrees of freedom for window size of 100 and 300 nt, and $\chi^2 = 12.61$ with one degree of freedom for window size 500 nt, respectively. The data therefore provides

TABLE 7
Extreme UFRs detected in the protein coding, RNA gene and non-coding regions of H. pylori. *Eight types of UFR (from type* 1 *to* 8*) are defined based on the computed Sigscr and/or Stbscr as listed in the second and third column. Numbers listed in parentheses indicate the probabilities of the eight types of UFR occurring in the complete* H. pylori *genome sequence computed by derived linearly transformed, theoretical non-central t distributions of the two scores, Sigscr and Stbscr. Numbers listed in the right three columns are the counts of these distinct UFRs that are involved entirely within the protein coding, RNA gene and non-coding regions in the genome*

UFR type	Sigscr (P-tail)	Stbscr (P-tail)	Protein	RNA	Non-coding
(a) *UFR counts detected in windows of* 100 *nt*					
1	$\leqslant -4.94$ (0.001)		526	3	1006
2	$\geqslant 2.52$ (0.001)		436	3	141
3		$\leqslant -3.70$ (0.001)	745	91	45
4		$\geqslant 2.65$ (0.0025)	1401	—	452
5	$\leqslant -4.94$ (0.001)	$\leqslant -2.57$ (0.01)	316	3	429
6	$\geqslant 1.19$ (0.05)	$\geqslant 2.65$ (0.0025)	826	—	312
7	$\geqslant 1.19$ (0.05)	$\leqslant -1.71$ (0.05)	1	1	—
8	$\leqslant -2.39$ (0.05)	$\geqslant 1.56$ (0.05)	29	—	43
(b) *UFR counts detected in windows of* 300 *nt*					
1	$\leqslant -6.73$ (0.001)		522	2	773
2	$\geqslant 2.61$ (0.001)		445	—	73
3		$\leqslant -3.10$ (0.001)	148	626	370
4		$\geqslant 3.10$ (0.001)	1479	—	309
5	$\leqslant -6.73$ (0.001)	$\leqslant -2.33$ (0.01)	125	2	314
6	$\geqslant 0.74$ (0.05)	$\geqslant 3.10$ (0.001)	1027	—	152
7	$\geqslant 0.74$ (0.05)	$\leqslant -1.64$ (0.05)	56	6	—
8	$\leqslant -3.25$ (0.05)	$\geqslant 1.64$ (0.05)	232	—	47
(c) *UFR counts detected in windows of* 500 *nt*					
1	$\leqslant -9.35$ (0.001)		514	—	114
2	$\geqslant 1.77$ (0.001)		1535	—	216
3		$\leqslant -3.10$ (0.001)	629	675	69
4		$\geqslant 3.10$ (0.001)	1905	—	233
5	$\leqslant -9.35$ (0.001)	$\leqslant -2.33$ (0.01)	17	—	13
6	$\geqslant 0.32$ (0.05)	$\geqslant 3.10$ (0.001)	1218	—	118
7	$\geqslant 0.32$ (0.05)	$\leqslant -1.64$ (0.05)	18	19	—
8	$\leqslant -4.40$ (0.05)	$\geqslant 1.64$ (0.05)	11	—	23

very strong evidence associating UFR pattern and gene structure, with $p < 0.001$. Thus, we observe that the significantly more stable folding regions (type 1 UFRs) are predominantly in non-coding sequences, and that the significantly less stable folding regions (type 2 UFRs) are predominantly in protein coding regions. The difference of observed counts for extreme UFRs among the RNA gene, coding and non-coding sequences is statistically very significant by a χ^2-test. Similarly, we can make another contingency table of two rows using the data of the types 5 and 6 UFRs in Table 7. Our χ^2-test indicates their $\chi^2 = 176.1$, 540.5 and 36.4 for window of 100, 300 and 500, respectively. The χ^2-test shows strong support to suggest that distinct UFRs with both very low Sigscr and Stbscr are predominantly in non-coding sequences, and the UFRs with both very high Sigscr and Stbscr are predominantly in protein coding sequences.

Where such highly stable and significant UFR patterns are found in the coding or non-coding regions of viral genomes (Wang *et al*., 1995; Malim *et al*., 1990; Phillips *et al*., 1992) and in bacteriophage mRNA (Le *et al*., 1993), they reflect the existence of RNA structures with important regulatory functions. We may likewise expect microbial genomes to have similar, and perhaps different, important functional RNA structures. Knowledge of these structures, or their absence, should be useful for understanding the genome and developing antimicrobial drug strategies. The approach used here is generally applicable. For instance, we also find that the samples of Sigscr and Stbscr data computed from other microbial genomes (such as *H. pylori* strain J99, *Mycoplasma genitalium* strain G37 and *Mycoplasma pneumoniae* strain M129) are also well described by the proposed statistical model (data not shown).

In this study, we have employed an LTNSTD to describe the distribution of Sigscr and Stbscr computed in whole-genomic sequences. For a given sample, we can calculate the sample mean, standard deviation, and coefficient of skewness. Using the formula proposed in this study, we can calculate the non-centrality parameter and quantile. As a result, we can estimate extreme UFRs in the sample using the derived LTNSTD. Approaches such as antisense RNA therapeutics, or the targeting of RNA-binging drugs should particularly benefit from identification of the unique regions of genomic sequences based on sound statistical features.

The assistance in the preparation of the manuscript by John Owens is gratefully acknowledged. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The program SEGFOLD and its modified version SIGSTB used in this study are available via anonymous ftp as /home/ftp/pub/users/shuyun/fcopy/sigfold at ftp.ncifcrf.gov.

REFERENCES

ANDRADE, M. A. & SANDER, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8,** 675–683.

BADGER, J. H. & OLSEN, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16,** 512–524.

BAILEY, N. T. J. (1995). *Statistical Methods in Biology*, 3rd edn. Cambridge U.K.: Cambridge University Press.

BELL, S. J. & FORSDYKE, D. R. (1999). Accounting units in DNA. *J. theor. Biol.* **197,** 51–61.

BORODOVSKY, M. J., MCININCH, E. V., KOONIN, E. V., RUDD, K. E., MEDIGUE, C. & DANCHIN, A. (1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucl. Acids Res.* **23,** 3554–3562.

BRUTLAG, D. L. (1998). Genomics and computational molecular biology. *Curr. Opin. Microbiol.* **1,** 340–345.

BUCHER, P. (1999). Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9,** 400–407.

CHEN, J.-H., LE, S.-Y., SHAPIRO, B., CURREY, K. M. & MAIZEL JR., J. V. (1990). A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.* **6,** 7–18.

EVANS, M., HASTINGS, N. & PEACOCK, B., eds (1993). *Statistical Distributions*, 2nd edn. New York: Wiley.

FORSDYKE, D. R. (1995). Conservation of stem-loop potential in introns of snake venom phospholipase A genes. An application of FORS-D analysis. *Mol. Biol. Evol.* **12,** 1157–1165.

FRECH, K., QUANDT, K. & WERNER, T. (1997). Software for the analysis of DNA sequence elements of transcription. *Comput. Appl. Biosci.* **13,** 89–97.

FREIER, S. M., KIERZEK, R., JAEGER, J. A., SUGIMOTO, N., CARUTHERS, M. H., NEILSON, T. & TURNER, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. U.S.A.* **83,** 9373–9377.

GELFAND, M. S., MIRONOV, A. A. & PEVZNER, P. A. (1996). Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. U.S.A.* **93,** 9061–9066.

GISH, W. & STATES, D. J. (1993). Identification of protein coding regions by databases similarity search. *Nat. Genet.* **3,** 266–272.

HERZEL, H., WEISS, O. & TRIFONOV, E. N. (1999). 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* **15,** 187–193.

HOGG, R. V. & TANIS, E. A. eds (1997). *Probability and Statistical Inference*, 5th edn. Upper Saddle River, NJ: Prentice-Hall.

JAEGER, J. A., TURNER, D. H. & ZUKER, M. (1989). Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. U.S.A.* **86,** 7706–7710.

JIANG, Q., HIRATSUKA, K. & TAYLOR, D. E. (1996). Variability of gene order in different Helicobacter pylori strains contributes to genome diversity. *Mol. Microbiol.* **20,** 833–842.

KARLIN, S., CAMPBELL, A. M. & MRAZEK, J. (1998). Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32,** 185–225.

KOONIN, E. V., MUSHEGIAN, A. R. & RUDD, K. E. (1996). Sequencing and analysis of bacterial genomes. *Curr. Biol.* **6,** 404–416.

LE, S.-Y. & MAIZEL JR., J. V. (1989). A method for assessing the statistical significance of RNA folding. *J. theor. Biol.* **138,** 495–510.

LE S.-Y. & ZUKER, M. (1990). Common structures of the 5′ non-coding RNA in enteroviruses and rhinoviruses: thermodynamical stability and statistical significance. *J. Mol. Biol.* **216,** 729–741.

LE, S.-Y., CHEN, J.-H., BRAUN, M. J., GONDA, M. A. & MAIZEL JR., J. V. (1988). Stability of RNA stem-loop structure and distribution of random structure in the human immunodeficiency virus (HIV-1). *Nucl. Acids Res.* **16,** 5153–5168.

LE, S.-Y., CHEN, J.-H. & MAIZEL JR., J. V. (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. & MAIZEL JR., J. V. (1990). In: *Structure & Methods*: *Human Genome Initiative and DNA Recombination*. (Sarma, R. H. & Sarma, M. H., eds), Vol. I, pp. 127–136. Schenectady: Adenine Press.

LE, S.-Y., CHEN, J.-H. & MAIZEL JR., J. V. (1993). Identification of unusual RNA folding patterns encoded by bacteriophage T4 gene 60. *Gene* **124,** 21–28.

LE, S.-Y., SIDDIQUI, A. & MAIZEL JR., J. V. (1996). A common structural core in the internal ribosome entry sites of picornavirus, hepatitis C virus, and pestivirus. *Virus Gene* **12,** 135–147.

LEHMANN, E. L., ed. (1959). *Testing Statistical Hypothesis*. New York: Wiley.

MALIM, M. H., HAUBER, J., LE, S.-Y., MAIZEL, JR, J. V. & CULLEN, B. R. (1989). The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338,** 254–257.

MALIM, M. H., TILEY, L. S., MCCARN, D. F., RUSCHE, J. R., HAUBER, J. & CULLEN, B. R. (1990). HIV-1 structural gene expression requires binding of the Rev *trans*-activator to its RNA target sequence. *Cell* **60,** 675–683.

PATZEL, V. & SCZAKIEL, G. (1997). The hepatitis B virus posttranscriptional regulatory element contains a highly stable RNA secondary structure. *Biochem. Biophys. Res. Commun.* **231,** 864–867.

PELLETIER, J. & SONENBERG, N. (1988). Internal initiation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334,** 320–325.

PHILLIPS, T. R., LAMONT, C., KONINGS, D. A., SHACKLETT, B. L., HAMSON, C. A., LUCIW, P. A. & ELDER, J. H. (1992). Identification of the Rev transactivation and Rev-responsive elements of feline immunodeficiency virus. *J. Virol.* **66,** 5464–5471.

SANTALUCIA JR., J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. U.S.A.* **95,** 1460–1465.

SCZAKIEL, G. (1997). The design of Antisense RNA. *Antisense Nucl. Acid Drug Dev.* **7,** 439–444.

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

SNYDER, E. E. & STORMO, G. D. (1995). Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248,** 1–18.

TOMB, J. F., WHITE, O., KERLAVAGE, A. R., CLAYTON, R. A., SUTTON, G. G., FLEISCHMANN, R. D., KETCHUM, K. A., KLENK, H. P., GILL, S., DOUGHERTY, B. A. *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547.

UBERBACHER, E. C., XU, Y. & MURAL, R. J. (1996). Discovering and understanding genes in human DNA sequence using GRAIL. *Meth Enzymol.* **266,** 259–281.

WALTON, S. P., STEPHANOPOULOS, G. N., YARMUSH, M. L. & ROTH, C. M. (1999). Prediction of antisense oligonucleotide binding affinity to a structured RNA target. *Biotechnol. Bioeng.* **65,** 1–9.

WANG, C., LE, S.-Y., ALI, N. & SIDDIQUI, A. (1995). An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5′ noncoding region. *RNA* **1,** 526–537.

ZUKER, M. (1994). Prediction of RNA secondary structure by energy minimization. *Meth. Mol. Biol.* **25,** 267–294.