doi:10.1006/jmbi.2000.3836 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 301, 27–33

JMB

Prediction of Transcription Terminators in Bacterial Genomes

Maria D. Ermolaeva*, Hanif G. Khalak, Owen White, Hamilton O. Smith and Steven L. Salzberg

The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD 20850, USA

**Corresponding author*

This study describes an algorithm that finds rho-independent transcription terminators in bacterial genomes and evaluates the accuracy of its predictions. The algorithm identifies terminators by searching for a common mRNA motif: a hairpin structure followed by a short uracil-rich region. For each terminator, an energy-scoring function that reflects hairpin stability, and a tail-scoring function based on the number of U nucleotides and their proximity to the stem, are computed. A confidence value can be assigned to each terminator by analyzing candidate terminators found both within and between genes, and taking into account the energy and tail scores. The confidence is an empirical estimate of the probability that the sequence is a true terminator. The algorithm was used to conduct a comprehensive analysis of 12 bacterial genomes to identify likely candidates for rho-independent transcription terminators. Four of these genomes (*Deinococcus radiodurans, Escherichia coli, Haemophilus influenzae* and *Vibrio cholerae*) were found to have large numbers of rho-independent terminators. Among the other genomes, most appear to have no transcription terminators of this type, with the exception of *Thermotoga maritima*. A set of 131 experimentally determined *E. coli* terminators was used to evaluate the sensitivity of the method, which ranges from 89 % to 98 %, with corresponding false positive rates of 2 % and 18 %.

Keywords: transcription terminators; bioinformatics; microbial genomics; sequence analysis

Introduction

Bacterial genomes are organized into units of expression that are bounded by sites where transcription of DNA into RNA is initiated and terminated. Regulation of gene expression is often accomplished by influencing the efficiency of these processes. Transcription termination is a product of DNA-protein interactions, destabilization of the transcript complex by structures formed in the RNA transcript, or a combination of these phenomena (Richardson, 1993; Henkin, 1996). Identification of sites at which termination events occur, in concert with promotion sites, can provide a basis for organizing genes into structural and functional operons.

One of the mechanisms of transcription termination in bacteria is rho-independent, or instrinsic, termination (Farnham & Platt, 1981; Platt, 1986; Yager & Hippel, 1991; Wilson & Hippel, 1995; Kroll *et al.*, 1992; Smith *et al.*, 1995). This process involves the formation of secondary structure in the mRNA sequence upstream of the termination site. These structures are distinguished by a common mRNA motif: a stem-loop structure with dyadic stem-pairing high in guanine and cytosine residue content, followed by a uracil-rich stretch of sequence proximal to the termination site (Figure 1). A computational approach to identification of rho-independent terminators on a genomic scale involves the calculation of at least three factors:

(1) Stability of the RNA stem-loop structure.
(2) Composition and proximity of the downstream U-rich region.
(3) Location and orientation of the terminator with respect to neighboring genes.

Present address: H.O. Smith, Celera Genomics Corporation, 45 West Gude Drive, Rockville, MD 20850, USA

E-mail address of the corresponding author: mariae@tigr.org

0022-2836/00/010027–7 $35.00/0

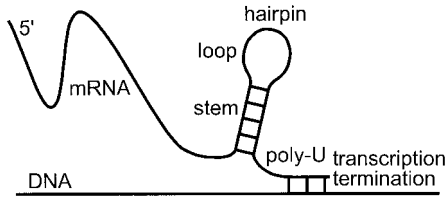

Figure 1. Model of a rho-independent transcription terminator.

Algorithms to detect DNA signals for transcriptional termination developed previously have treated some but not all of these factors (Brendel & Trifonov, 1984; Brendel *et al*., 1986; Carafa *et al*., 1990; Blaisdell *et al*., 1993; Zuker, 1994; Washio *et al*., 1998), and have been applied primarily to *E. coli* genome sequences.

Here, we describe a system called TransTerm that finds intrinsic terminators in bacterial genomes and evaluates the accuracy of each prediction. Using this algorithm, we conduct a comprehensive analysis of 12 bacterial genomes in an attempt to identify all intrinsic terminators in these genomes.

Methods

The TransTerm algorithm has two stages. First, it scans the entire genome to find all DNA templates of mRNA stem-loops (also referred to as hairpins) with adjacent uracil-rich stretches. For each hairpin, it computes an energy score (characterizing the stability of the hairpin) and a tail score (characterizing how many uracil nucleotides are located at the 3′ end of the hairpin). The second stage of the algorithm analyzes all terminator candidates and calculates a confidence for each one. This confidence is an empirical estimate of the probability that the terminator candidate genuinely functions as a transcription terminator.

Searching a genome for terminator candidates

An energy-scoring function was used to characterize the stability of the stem-loop structure. A hairpin turn is formed by a stem containing complementary nucleotides, capped by a short loop. In the stem, the most stable nucleotide pair is G-C, which is assigned a score *gc*, and A-U pairs are given the score *au*. In RNA there is also a weak interaction between G and U, which is assigned the score *gu*. Other nucleotide pairs do not form hydrogen bonds and weaken the stem; they are given the score *mm* (mismatch). A gap in the base-pairing structure on either side of the stem weakens it even further, and is scored as *gp* (only one gap per hairpin is allowed, given that even one gap usually produces an unstable hairpin). Long loops also destabilize hairpins, and each nucleotide of the loop is assigned score *lp*.

The energy score for a hairpin is computed by combining all these scores:

$$E = gc \cdot x_1 + au \cdot x_2 + gu \cdot x_3 + mm \cdot x_4 + gp \cdot x_5 + lp \cdot x_6 \quad (1)$$

where x_1, x_2 and x_3 are counts of the G-C, A-U and G-U nucleotide pairs in the stem, x_4 and x_5 are the numbers of mismatches and gaps, and x_6 is the number of nucleotides in the loop. This energy function is designed to separate hairpins ($E < E_0$) from non-hairpin structures ($E > E_0$). The optimization problem for the parameters *gc*, *au*, *gu*, *mm*, *gp*, and *lp* can be posed as a linear separability problem in the 6D space defined by these variables. This problem has been well studied in the pattern recognition literature, and decision trees are known to be a very effective solution. We used the OC1 decision tree system (Murthy *et al*., 1994) to obtain the best separation for a training set of 140 sequences, half of which were real terminators and half of which were false. The true (experimentally verified) terminators were taken from Table 2 in the report by Carafa *et al*., 1990), while the 70 false examples were sequences that have similar characteristics, but are highly unlikely to be rho-independent terminators because they are located inside genes. The parameters that give the best separation, classifying 94% of the training examples correctly, were:

$$E = 2.3x_1 - 0.9x_2 + 1.3x_3 + 3.5x_4 + 6.0x_5 + 1.0x_6 - 5.7. \quad (2)$$

These parameters were used in all subsequent calculations of energy scores.

Functional transcriptional terminators are composed of a hairpin with a 3′ poly-U stretch. Most terminators have poly-U tails longer than three base-pairs and the length of poly-U regions has been reported to correlate with termination efficiency (Jeng *et al*., 1997). Our algorithm selects only those hairpins that have at least three consecutive uracil nucleotides near the stem, at a distance of no more than five base-pairs and uses the tail-scoring function from (Carafa *et al*., 1990):

$$T = -\sum_{n=1}^{15} x_n \quad (3)$$

for all U residues in the 15 nucleotide segment where $x_0 = 1$ and

$$x_n = \begin{cases} x_{n-1} \times 0.9 & \text{if the } n\text{th nucleotide is a U} \\ x_{n-1} \times 0.6 & \text{if the } n\text{th nucleotide is other than U} \end{cases} \quad (4)$$

The tail-scoring function T reflects how many U nucleotides are located on the 3′ side of the hairpin

and how close these nucleotides are to the stem. A low value of the tail scoring function corresponds to a U-rich tail.

The TransTerm algorithm searches a complete bacterial genome sequence by calculating for all RNA subsequences whether they satisfy the above criteria; that is, the energy score is below the cutoff, and the tail contains at least three consecutive U nucleotides. In addition, stem length is constrained to be in the range 4 to 20 nucleotide pairs, and loop length must range from 3 to 10 nucleotides. A pseudocode summary of TransTerm is:

```
for the current position C in the genome
      if (there are >= 3 consecutive Us upstream from C
               at a distance of <=5 nucleotides) then
               for (stem length from 4 to 20) do
                        for (loop length from 3 to 10) do
                              for (all possible gap positions including no gap) do
                                 calculate value of the energy scoring function E;
                                 if (E < E0) then
                                        calculate the tail scoring function T;
                                        if (T < T0) then
                                               output a terminator candidate.
```

For each hairpin found, the following data are recorded:

(1) Genomic coordinates of the hairpin.
(2) Directionality (forward or reverse strand).
(3) The energy score E, where a low value corresponds to a stable hairpin.
(4) The tail score T, where a low value corresponds to a U-rich 3′ tail.

This search identifies all hairpins with U-rich tails, with associated energy and tail scores. At this point the algorithm does not yet have any relative scores ranking these potential terminators; some will look like model terminators, while others will have mismatches or gaps in the stem. The next phase of the algorithm calculates a confidence value for each hairpin. The confidence value is an empirical estimate of the specificity of a given prediction; i.e. *C*% of the predictions with confidence *C* are expected to be true terminators.

Calculating confidences of terminator candidates

In order to calculate confidence, TransTerm analyzes two types of genome regions: intragenic (''inside genes'') and intergenic, as shown in Figure 2. In order to make sure that intragenic regions really fall within genes, they are defined to include only those sequences beginning 100 nucleotides after the annotated start codon and ending 100 nucleotides before the stop codon. No transcription terminators are expected to be found in these regions. Intergenic regions are defined as DNA sequences between genes plus 50 nucleotides inside each flanking gene. Most (or all) transcription terminators should occur in intergenic regions. We distinguish two types of intergenic regions, defined by the direction of the flanking genes: ''tail-to-tail'' and ''head-to-tail''.

There should be no real transcription terminators inside genes. However there are sequences within genes that have low values of the energy and tail-scoring functions. We call these structures ''false terminators.'' A key point in calculating a confidence for terminator predictions is estimating the frequency of these false terminators in any given region of the genome.

Rho-independent transcription terminators (as well as the false terminators) have a GC-rich stem, usually at least four to five G-C nucleotides long, and a U-rich tail. Therefore the frequency of false terminators should depend to some extent on the sequence composition. This dependence is modeled in our program by a second order polynomial approximation (Figure 3).

Consider that $N^{inside\ gene}(E, T)$ hairpins are found inside genes with energy E and tail scoring function T. (Here ''inside genes'' means within the set of all intragenic regions of the genome.) All hairpins found in these intragenic regions are ''false'' terminators: structures that are not transcription terminators but that exceed the computationally defined threshold scores. Using our polynomial model of the frequency of false terminators based on AU-content, the number of false terminators in tail-to-tail regions can be calculated as:

$$N_{\text{false}}^{\text{tail-to-tail}}(E, T) = 2kN^{\text{inside gene}}(E, T)\frac{L_{\text{tail-to-tail}}}{L_{\text{inside gene}}} \quad (5)$$

where $L_{tail\text{-}to\text{-}tail}$ is the sum of all lengths of ''tail-to-tail'' genome regions and $L_{inside\ gene}$ is the sum of all lengths of ''inside gene'' regions. $N^{inside\ gene}(E, T)$ is multiplied by 2 because the intragenic regions are searched only on the same strand as the gene, while the tail-to-tail regions are searched on both strands. The value k is computed using the function f(%AU), which is the polynomial function approxi-

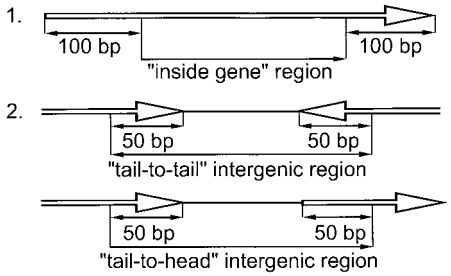

Figure 2. Regions of genome scanned for rho-independent terminator candidates. Top, regions inside genes (intragenic), where no terminators are expected. Center and bottom, intergenic regions, where terminators (if any) are expected.

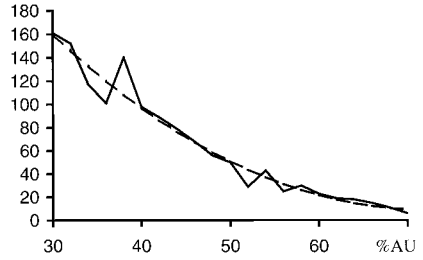

Figure 3. The number of terminator candidates (*y*-axis) found in Mb of random sequence as a function of *AU*-content (*x*-axis) is shown by the solid line. In the program, this is approximated by the broken line, which is a second-order polynomial fit using the least squares method.

mating the frequency of false terminators:

$$k = \frac{f(\%AU_{\text{intergenic regions}})}{f(\%AU_{\text{inside gene}})} \tag{6}$$

The number of true terminators found by TransTerm in tail-to-tail regions is the number of all terminator candidates in these regions minus the false terminators:

$$N_{\text{real}}^{\text{tail-to-tail}}(E,T) = \max(N^{\text{tail-to-tail}}(E,T) - N_{\text{false}}^{\text{tail-to-tail}}(E,T), 0) \tag{7}$$

The probability that a terminator candidate in a tail-to-tail region with energy E and tail scoring function T is real is the ratio of number of real terminators to the number of all terminator candidates:

$$C(E,T) = \left(\frac{N_{\text{real}}^{\text{tail-to-tail}}(E,T)}{N^{\text{tail-to-tail}}(E,T)}\right) \times 100\% \tag{8}$$

We will call this probability the confidence of a terminator.

Combining equations (5), (7), and (8) gives an equation where the confidence of a terminator candidate in a tail-to-tail region depends only on the number of hairpins with the same characteristics found in all tail-to-tail and intragenic regions:

$$C(E,T) = \max\left(\left(1 - \frac{2kN^{\text{inside gene}}(E,T)\cdot L^{\text{tail-to-tail}}}{N^{\text{tail-to-tail}}(E,T)\cdot L^{\text{inside genes}}}\right), 0\right) \times 100\% \tag{9}$$

Both $N^{inside\ gene}(E, T)$ and $N^{tail\text{-}to\text{-}tail}(E, T)$ are two-dimensional functions. For both intergenic and intragenic regions, $N(E, T)$ is the number of hairpins with energy $E \pm \Delta E/2$ and tail scoring function $T \pm \Delta T/2$. This term can be related to the two-dimensional distribution of energy and tail scores $q(E, T)$ in the following way:

$$q(E,T) \times \Delta E \times \Delta T = \frac{N(E,T)}{N} \tag{10}$$

where N is the total number of hairpins in the region. Since the number of hairpin sequences is insufficient for building reliable two-dimensional distributions $q(E, T)$, the algorithm approximates $q(E, T)$ by assuming that values of the energy scoring function and tail scoring function are independent:

$$q(E,T) = q_E(E) \times q_T(T) \tag{11}$$

This assumption is probably not strictly correct, and Table 1 shows average values of $q(E, T)$ and

Table 1. Comparison of average values of $q(E, T)$ and $q_E(E)\cdot q_T(T)$ for *E. coli* intergenic terminator candidates

	$-20.2 \leqslant E < -12$	$-12 \leqslant E < -10$	$-10 \leqslant E < -6$
$-4 \leqslant T < 2$	$q(E,T) = 1.05 \times 10^{-2}$ $q_E(E) \times q_T(T) = 1.31 \times 10^{-2}$	$q(E,T) = 2.66 \times 10^{-2}$ $q_E(E) \times q_T(T) = 2.06 \times 10^{-2}$	$q(E,T) = 2.48 \times 10^{-2}$ $q_E(E) \times q_T(T) = 2.69 \times 10^{-2}$
$-5 \leqslant T < 4$	$q(E,T) = 1.15 \times 10^{-2}$ $q_E(E) \times q_T(T) = 1.31 \times 10^{-2}$	$q(E,T) = 2.94 \times 10^{-2}$ $q_E(E) \times q_T(T) = 3.35 \times 10^{-2}$	$q(E,T) = 2.75 \times 10^{-2}$ $q_E(E) \times q_T(T) = 2.68 \times 10^{-2}$
$-6.4 \leqslant T < 5$	$q(E,T) = 8.15 \times 10^{-3}$ $q_E(E) \times q_T(T) = 7.99 \times 10^{-3}$	$q(E,T) = 2.06 \times 10^{-2}$ $q_E(E) \times q_T(T) = 2.65 \times 10^{-2}$	$q(E,T) = 1.93 \times 10^{-2}$ $q_E(E) \times q_T(T) = 1.67 \times 10^{-2}$

$q_E(E) \times q_T(T)$ for *Escherichia coli* intergenic terminator candidates, for different E and T. The maximum difference between $q(E, T)$ and $q_E(E) \cdot q_T(T)$ is about 30 %. At a confidence value of 98 %, the error in the confidence estimate introduced by the independence assumption for E and T is less than 0.5 % (see equations (9)-(11)). This error increases to 15 % for a confidence value of 50 %. If the confidence cut-off for genome annotation is set relatively high (we recommend using 98 % so as to avoid false positives), then the error introduced by this assumption will be much smaller than the one that would result from building a two-dimensional distribution on a small amount of data. Combining equations (9)-(11) gives:

$$C(E,T) = \max\left(\left(1 - \frac{2kN^{\text{inside genes}} \times L^{\text{tail-to-tail}} \times q_E^{\text{inside genes}}(E) \times q_T^{\text{inside genes}}(T)}{N^{\text{tail-to-tail}} \times L^{\text{inside genes}} \times q_E^{\text{tail-to-tail}}(E) \times q_T^{\text{tail-to-tail}}(T)}\right), 0\right) 100\% \quad (12)$$

Even the one-dimensional distributions $q_E(E)$ and $q_T(T)$ are not smooth (Figures 4-5), especially for regions with a small number of hairpins. To simplify calculations and to make them more reliable in the case of a small number of hairpins, $q_E(E)$ and $q_T(T)$ are approximated with the piece-wise linear functions $Q_E(E)$ and $Q_T(T)$. These functions are built by dividing the range into a small number of intervals, calculating an average value for each interval, and connecting the averages. Other approximations for $q_E(E)$ and $q_T(T)$ did not significantly change our results.

Some terminators in tail-to-tail regions have poly-U tails on both of the DNA strands, allowing them to function bi-directionally, terminating both of the surrounding genes (Postle & Good, 1985). The confidence value of a bi-directional terminator is calculated as the probability that it functions as a transcription terminator in at least one of the directions:

$$C_{\text{bi-directional}} = \left(1 - \left(1 - \frac{C_+}{100\%}\right)\left(1 - \frac{C_-}{100\%}\right)\right) \times 100\% \quad (13)$$

where C_+ and C_- are the confidences of the unidirectional terminators.

The confidences of terminators in head-to-head regions are calculated in a similar manner, although they are required to be located on the same DNA strand as the surrounding genes.

Results and Discussion

Application of TransTerm to the *E. coli* and *H. influenzae* genomes identified 1111 and 505 (respectively) terminator candidates. The distributions of confidence values are shown in Figure 6. 34 % of the *E. coli* terminator predictions and 62 % of those for *H. influenzae* have confidence values greater than 99.5 %.

The confidence value expresses the specificity of the prediction; i.e. the number of predictions that are expected to be correct. In order to evaluate the sensitivity of the method, i.e. how many true terminators are actually detected, it is necessary to evaluate it on known, experimentally determined terminators. The *E. coli* genome was used to evaluate sensitivity because it is the only genome with a

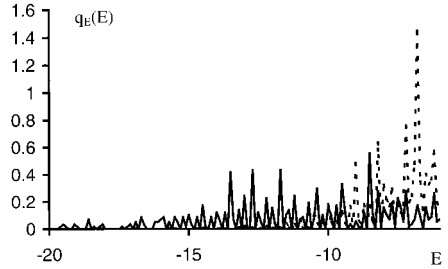

Figure 4. Distribution of the energy scoring function E for hairpins found in intragenic regions of *E. coli* (broken line) and in ''tail-to-tail'' intergenic regions (continuous line).

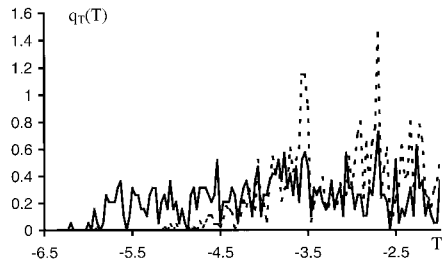

Figure 5. Distribution of the tail scoring function T for hairpins found in intragenic regions of *E. coli* (broken line) and in ''tail-to-tail'' intergenic regions (continuing line).

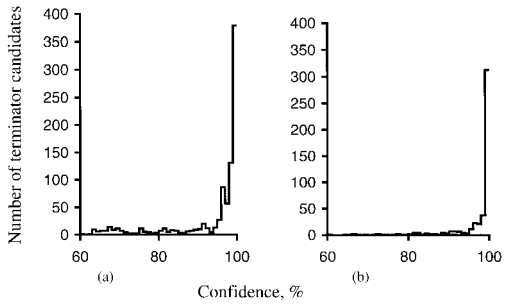

Figure 6. Distribution of confidences of terminator candidates found in intergenic regions of (a) *E. coli* and (b) *H. influenzae* genomes. Numbers of terminator candidates with confidences less than 60 % are 195 for *E. coli* and 34 for *H. influenzae*.

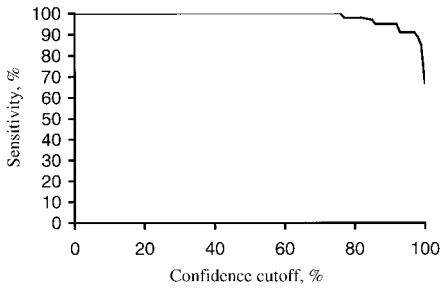

Figure 7. Trade-off between sensitivity and confidence for *E. coli* terminator identification.

substantial number of experimentally determined transcription terminators.

The complete set of 131 experimentally determined rho-independent terminators reported by Carafa *et al*. (1990) was used to measure the algorithm's sensitivity. Of these, 61 were used for testing, while 70 were included in the training set on which the algorithm's parameters were optimized.

TransTerm found all 61 terminators in the test set, although with varying degrees of confidence. Figure 7 shows the sensitivity/specificity trade-off. With the confidence set at 98 % (meaning that 98 % of all predictions will be correct), the sensitivity is 89 %, meaning that 89 % of all true terminators will be found. If the confidence were reduced to 82 %, then sensitivity would rise such that 98 % of all terminators would be found.

The algorithm is designed to find typical rho-independent transcription terminators, those that fit the ''hairpin + poly-U'' model. It will not find transcription terminators that have a significantly different structure; for example, there is one terminator in the *E. coli* training set that lacks a poly-U tail and therefore would not be found by the algorithm. To date, reports of such atypical rho-independent transcription terminators are quite rare, but that does not rule out new classes of terminators being discovered in the future.

We compared performance of TransTerm with the Terminator program in the GCG package (Brendel & Trifonov, 1984). Terminator missed 4 out of the 61 terminators in our test set, while TransTerm detected all of them. At the same time, TransTerm apparently produces fewer false positives than Terminator. For example, TransTerm detected no false positives (using a generous confidence cutoff of 50 %) inside the first ten genes on the *E. coli* genome, while Terminator detected 13 false positives. Another benefit of our algorithm is that in addition to finding potential terminators, it assigns a confidence value to each prediction. The user does not need to set thresholds for any of the complex scoring functions described above; these are set automatically once the user determines the appropriate confidence threshold.

The TransTerm program runs under Unix and requires about ten minutes per megabase of input sequence on a 550 MHz Intel processor. The computational time requirement scales linearly with genome size. The program is available from the TIGR website at: www.tigr.org.

The algorithm described here is being used as an annotation tool for defining the locations of rho-independent transcription terminators in bacterial and archaeal genomes.

Table 2 contains the numbers of terminator candidates in 12 bacterial genomes using a confidence threshold of 98 %. As the Table illustrates, for some organisms such as *Treponema pallidum*, few terminators with high confidence were found using our algorithm. We interpret these results to mean these organisms have transcription termination mechanisms that probably use a different structure than those containing a stem-loop and poly-U tail.

Finally, we analyzed one other type of terminator candidates: hairpins that are located inside genes but are anti-directional to the gene (i.e. located on the opposite strand). These are presumably false terminators; however, we found that the numbers of anti-directional and co-directional false terminators were significantly different. For example, using energy and tail score cut-offs of -6 and -4, the number of co-directional intragenic hairpins in *E. coli* is 312, while the number of anti-directional hairpins is 577. The probability that this difference

Table 2. Numbers of terminators with confidence 98 % or higher found in 12 bacterial genomes

Genome	Number of terminators
Escherichia coli	567
Haemophilus influenzae	371
Archaeoglobus fulgidus	2
Borrelia burgdorferi	51
Deinococcus radiodurans	390
Helicobacter pylori	9
Methanococcus jannaschii	14
Mycobacterium tuberculosis	16
Mycoplasma genitalium	14
Thermotoga maritima	118
Treponema pallidum	1
Vibrio cholerae	792

is random is very small:

$$P(312, 889) = (0.5)^{889} \sum_{k=0}^{312} C_{889}^{k} \times 100\%$$

$$= (0.5)^{889} \sum_{k=0}^{312} \frac{889!}{k!(889-k)!} \times 100\%$$

$$< 10^{-16}\% \qquad (14)$$

One possible explanation for this overabundance of anti-directional hairpins is that some of them may, in fact, be real terminators. They may terminate transcription when genes located on opposite strands of DNA overlap, or they may function as additional terminators for a transcript on the reverse strand located nearby. The efficiency of rho-independent terminators is likely to be less than perfect (Reynolds *et al*., 1992). These additional terminators may function as redundant downstream stops. Another possible explanation of the difference in numbers of co-directional and anti-directional hairpins is that both occur as a reset of random mutations, but selection pressures operate against co-directional hairpins because they interfere with transcription of the genes in which they are located. Anti-directional terminators may also have another function; e.g. mRNA transcript stabilization (Guarneros *et al*., 1982; Mott *et al*., 1985; Abe & Aiba, 1996).

Acknowledgments

This research was supported by the Merck Genome Research Institute under Grant No. 74. S.L.S. was supported in part by NIH grant R01-LM06845 and NSF grants KDI-9980088 and IIS-9902923.

References

Abe, H. & Aiba, H. (1996). Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization. *Biochimie,* **78**, 1035-1042.

Blaisdell, B. E., Rudd, K. E. & Karlin, M. A. (1993). Significant dispersed recurrent DNA sequences in the *Escherichia coli* genome. Several new groups. *J. Mol. Biol.* **229**, 833-848.

Brendel, V. & Trifonov, E. N. (1984). A computer algorithm for testing potential prokaryotic terminators. *Nucl. Acids Res.* **12**, 4411-4427.

Brendel, V., Hamm, G. H. & Trifonov, E. N. (1986). Terminators of transcription with RNA polymerase from *Escherichia coli*: what they look like and how to find them. *J. Biomolec. Struct. Dynam.* **3**, 705-723.

Carafa, Y. A., Brody, E. & Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* **216**, 835-858.

Farnham, P. J. & Platt, T. (1981). Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucl. Acids Res.* **9**, 563-577.

Guarneros, G., Montanez, C., Hernandez, T. & Court, D. (1982). Posttranscriptional control of bacteriophage λ int gene expression from a site distal to the gene. *Proc. Natl Acad. Sci. USA,* **79**, 238-242.

Henkin, T. M. (1996). Control of transcription termination in prokaryotes. *Annu. Rev. Genet.* **30**, 35-57.

Jeng, S. T., Lay, S. H. & Lai, H. M. (1997). Transcription termination by bacteriophage T3 and 5P6 RNA polymerases at Rho-independent terminators. *Can. J. Microbiol.* **43**, 1147-1156.

Kroll, J. S., Loynds, B. M. & Langford, P. R. (1992). Palindromic haemophilus DNA uptake sequences in presumed transcriptional terminators from *H. influenzae* and *H. parainfluenzae*. *Gene,* **114**, 151-152.

Mott, J. E., Galloway, J. L. & Platt, T. (1985). Maturation of *Escherichia coli* tryptophan operon mRNAs: evidence for 3′ exonucleolytic processing after rho-independent termination. *EMBO J.* **4**, 1887-1891.

Murthy, S. K., Kasif, S. & Salzberg, S. L. (1994). A system for induction of oblique decision trees. *J. Artificial Intelligence Res.* **2**, 1-32.

Platt, T. (1986). Transcription termination and the regulation of gene expression. *Annu. Rev. Biochem.* **55**, 339-372.

Postle, K. & Good, R. F. (1985). A bi-directional rho-independent transcription terminator between the *E. coli* tonB gene and an opposing gene. *Cell,* **41**, 577-585.

Reynolds, R., Bermüdez-Cruz, R. M. & Chamberlin, M. J. (1992). Parameters affecting transcription termination by *Escherichia coli* RNA polymerase: I. Analysis of 13 Rho-independent terminators. *J. Mol. Biol.* **224**, 31-51.

Richardson, J. P. (1993). Transcription termination. *Crit. Rev. Biochem. Mol. Biol.* **28**, 1-30.

Smith, H. O., Tomb, J.-F., Dougherty, B. A., Fleischmann, R. D. & Venter, J. C. (1995). Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science,* **269**, 538-540.

Washio, T., Sasayama, J. & Tomita, M. (1998). Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucl. Acids Res.* **26**, 5456-5463.

Wilson, K. S. & von Hippel, P. H. (1995). Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc. Natl Acad. Sci. USA,* **92**, 8793-8797.

Yager, T. D. & von Hippel, P. H. (1991). A thermodynamic analysis of RNA transcript elongation and termination in *Escherichia coli*. *Biochemistry,* **30**, 1097-1118.

Zuker, M. (1994). Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.* **25**, 267-294.

Edited by F. E. Cohen

(Received 4 January 2000; received in revised form 3 April 2000; accepted 28 April 2000)