Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

INTERPIN: A repository for intrinsic transcription termination hairpins in bacteria

Swati Gupta, Namrata Padmashali, Debnath Pal*

Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, 560012, Karnataka, India

ARTICLE INFO

Article history:
Received 9 March 2023
Received in revised form
19 July 2023
Accepted 25 July 2023
Available online xxx

Handling Editor: J.L. Mergny

Keywords:
Rho-independent transcription termination
Hairpin
Database
Bacterial genomes

ABSTRACT

The large-scale detection of putative intrinsic transcription terminators is limited to only a few bacteria currently. We discovered a group of hairpins, called *cluster* hairpins, present within 15 nucleotides from each other. These are expected to work in tandem to cause intrinsic transcription termination (ITT), while the *single* hairpin can do the same alone. Therefore, exploring these ITT sites and the hairpins across bacterial genomes becomes highly desirable. INTERPIN is the largest archived collection of *in silico* inferred ITT hairpins in bacteria, covering 12745 bacterial genomes and encompassing ten bacterial phyla for ~25 million hairpins. Users can obtain details on operons, individual *cluster,* and *single* ITT hairpins that were screened therein. Integrated Genome Viewer (IGV) software interactively visualizes hairpin secondary and tertiary structures in the genomic context. We also discuss statistics for the occurrence of *cluster* or *single* hairpins and other termination alternatives while showing the validation of predicted hairpins against *in vivo* detected hairpins. The database is freely available at http://pallab.cds.iisc.ac.in/INTERPIN/.

INTERPIN (database and software) can make predictions for both AT and GC-rich genomes, which has not been achieved by any other program so far. It can also be used to improve genome annotation as well as to get predictions to improve the understanding of the ITT pathway by further analysis.

1. Introduction

The transfer of information from DNA to RNA through the process of transcription is essential for the formation of protein, which ensures routine biochemical activities in the cell. In bacteria, the transcription process works concurrently with translation to regulate correct protein synthesis. The termination of transcription is *intrinsic/Rho-independent* when the nascent transcript created by the RNA polymerase (RNAP) self-terminates the nucleotide addition process or *Rho-dependent* when assisted by additional protein factors like Rho protein, Nus factors, etc. The formation of a hairpin-folded structure in nascent RNA promotes the intrinsic termination process [1]. These hairpin structures can also be found in the ribosomal RNAs [2], in the 5' UTR region of transcription units, where they cause antitermination [3], and in recognition motifs that facilitate RNA-protein binding [4,5].

The vital role of the hairpin in intrinsic transcription termination (ITT) has been extensively studied both *in silico* [6–11] and *in vitro*/*in vivo* [12–20]. The latter mainly focuses on understanding the mechanistic aspect of the ITT process [21–24]. The experimentally derived mechanistic premise became the basis for the *in silico* approaches [12,25]. The experimental systems suffer from the lacunae that the investigations cover only a limited diversity of bacterial organisms [6,26–29]. Consequently, the paradigm laid for *in silico* approaches was based on identifying strong hairpin structures and the presence of a poly A/U trail in the immediate downstream of the hairpin. These criteria succeeded in screening only a minor (1/3) fraction of ITT sites in the genomes and were biased for cases with GC-rich composition [29,30]. Given that ITT mechanisms are more straightforward than the *Rho-dependent* transcription termination, the ability to locate ITT for a majority of the operons is, therefore, important.

We have recently bridged that gap through an exhaustive *in silico* study, covering operons from 13 representative genomes from 6 diverse phyla of the bacterial kingdom [31]. We have been able to locate ITT for 72% of the ITT sites for operons which the RNA-seq data could corroborate. The coverage could be improved to 97%

* Corresponding author.
E-mail address: dpal@iisc.ac.in (D. Pal).

https://doi.org/10.1016/j.biochi.2023.07.018
0300-9084/

when we considered missed termination at the first predicted ITT site from the stop codon and looked at alternate termination sites predicted further downstream. The results showed that about two-thirds of the operons have an ITT site that deploys a group of hairpins called *cluster* hairpins. These *cluster* hairpin components are expected to work in tandem to cause transcription termination. The constituent hairpins in the *cluster* hairpin are present at <15 nucleotides (nt) from each other and could work akin to a *single* strong hairpin owing to kinetic and thermodynamic considerations [32,33]. No bias for AT and GC rich genomes was noted for hairpin occurrence, nor was the poly A/U pattern found essential for ITT.

We have since extended our analysis to cover 25 million hairpins from 12,745 bacterial genomes spanning ten phyla. This paper presents the largest collection of inferred intrinsic terminators through a public repository called INTERPIN (INtrinsic transcription TERmination hairPIN; http://pallab.cds.iisc.ac.in/INTERPIN/). In the database, the genomes have been organized into respective phyla for search by phylum, organism name, and NCBI ID. The web interface allows the results to be easily retrieved, interactively. In the database, 58% of all the inferred hairpins are *cluster* hairpins. For ITT prediction in new bacterial genomes or customized local runs, we provide a software suite with the same name, downloadable from the GitHub repository (https://github.com/swati375/interpin). The program can process multiple genomes in parallel. A job for a single genome of size approximately 3.8 × 10^7 nt with 1800 operons takes about 10 h to process, with >90% of the time expended to obtain the initial RNA-hairpin information from the MFold program [34].

2. Methods

2.1. Data retrieval

We downloaded all bacterial genomes present in NCBI in September 2019. The data encompassed ten phyla - Acidobacteria, Actinobacteria, Proteobacteria, Firmicutes, Fusobacteria, Planctomycetes, Spirochaetes, Chlamydiae, Cyanobacteria, Thermodesulfobacteria. It was ensured that the genomes downloaded had proper names conforming to the Bacteriological Code. The bacterial phyla classification for the dataset was obtained from the LSPN database [[35,36]; https://lpsn.dsmz.de/] as of May 15, 2020. All included bacterial phyla and the corresponding number of bacteria taken in the analysis are shown in Table S1. The same ordered by their phylogenetic closeness are shown in Fig. S1, drawn using iTOL (tree of life [37]). For each genome, FASTA and Genbank annotated sequences are obtained from NCBI and stored on our local servers.

2.2. Construction of the database

The database has been constructed using the method described in our previous study [31]. At first, operons are found for the input organism using Molquest. After that, the Interoperonic Regions (IR) and sequence at position 0 to around 270 with reference to the stop codon of the corresponding operon are extracted. IR is the region between the 3′ end of the coding region of one operon and the 5’ end of the coding region of the next operon. These sequences are used for finding hairpin folded structures using MFold [34]. After the removal of outliers using hairpin features like stem and loop length, *cluster* and *single* hairpins are identified. The definitions of *single* and *cluster* hairpins are explained in Fig. 1, with the help of a schematic.

The base code for the webserver design has been written using Python 3.0. The figures are drawn using MATPLOTLIB library version 3.4. The database is constructed using the Django framework version 3.2. After running the program on all 12,745 bacterial genomes on a 64-bit Linux machine with 32 cores, the subsequent results obtained in binary format are converted to HTML table format for ease of users.

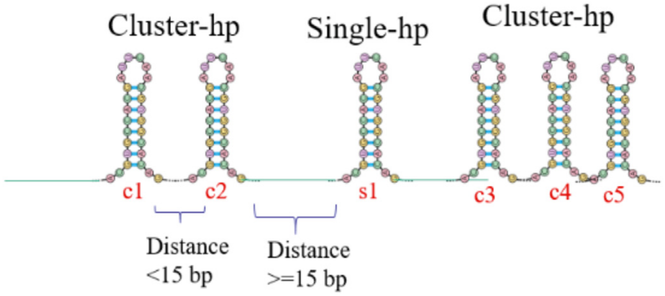


Fig. 1. Diagram showing the criteria for classifying the *cluster* and *single* hairpins (hp). Hairpins <15 nt from each other are clubbed in a group called the *cluster*. These hairpins are expected to work together to cause ITT. The other hairpins lying at larger distances from each other are called *single*. The c1 and c2 form one *cluster*, while the c3, c4, and c5 form another. s1 is a *single* hairpin.

3. Results

The overview of the data in the INTERPIN repository can be seen in Table 1. The data covers all published microbial genomes available as of September 2019. Only chromosomes have been considered, and data about plasmids are excluded. The data covers both AT and GC-rich genomes.

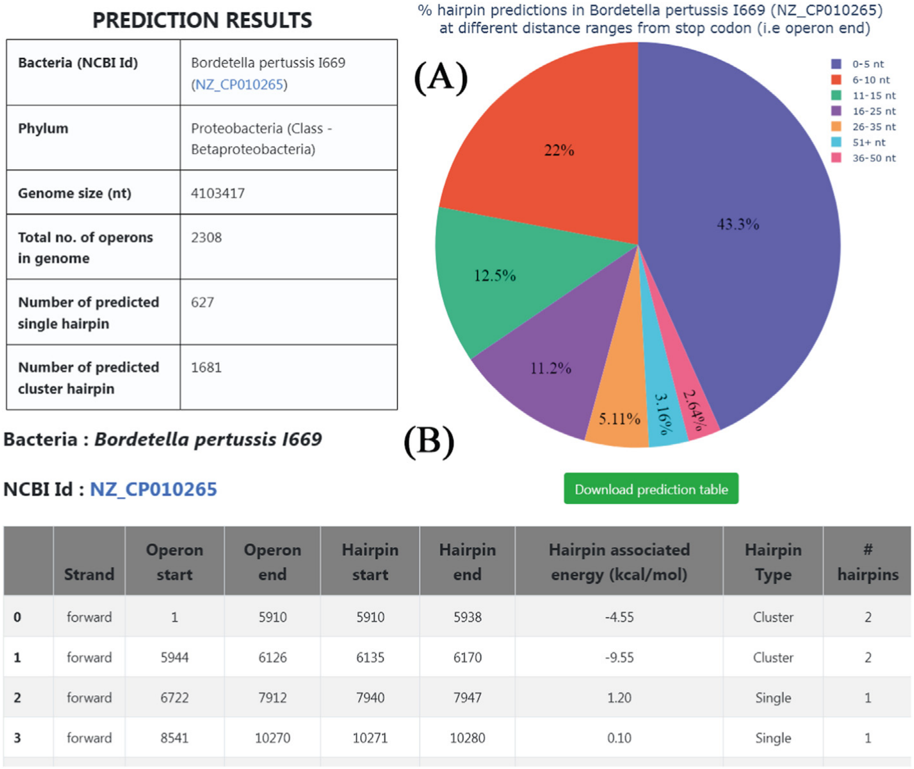
3.1. Access to the database and information retrieval

The INTERPIN homepage provides a general introduction to the repository. The menu is provided through the tabs on the top of the page to enable navigation through the database. The “Phyla” tab provides a link to the individual phylum, as listed in Table 1. Since bacteria within the same phyla are closely related to each other compared to genomes in other phyla, this search tab should be used for browsing data from closely related species. The other tab, named “NCBI”, gives direct access to the data at the genome level through the NCBI ID. A list of all bacteria with their corresponding NCBI ID and phylum is also given to the user. The “Phyla” and the “NCBI” tabs take the user to the results page, which displays hairpin predictions. The “Help” tab gives a detailed description of the database and its usage to assist the user in case of difficulty. The page explains the steps for accessing different features available, along with examples. Each prediction for a bacterium provides information on the operons, frequency of the inferred *cluster* and *single* hairpins, and their distance from the stop codon (Fig. 2A). Raw information about all hairpin predictions, including their location, energies, etc., can also be downloaded by the ‘click here to view prediction file’ button given on the same page. The tabular output (Fig. 2B) shows the hairpin data for the following fields: (i) Hairpin predicted on which strand (forward, reverse), (ii) Operon start position, (iii) Operon end position, (iv) Hairpin start position, (v) Hairpin end position, (vi) Hairpin associated energy (in case of a *cluster,* the average energy of constituent hairpins), (vii) Type of hairpin (*cluster*, *single*), (viii) Number of the constituent hairpin (1 for *single*, else the number is specified). The same information can be downloaded in the CSV format by using the download tab, as shown in Fig. 2B.

The alternate termination sites for each genome are also given for each bacterial genome on the results page. The ‘click here to view alternative prediction sites’ button takes users to a table with all predicted alternate termination sites, giving details about the

Table 1
Summary of contents available from the INTERPIN repository.

Number of Bacterial Genomes	12,745
Number of Phyla	10
Phyla names	Firmicutes, Chlamydiae, Actinobacteria, Spirochaetes, Planctomycetes, Fusobacteria, Cyanobacteria, Thermodesulfobacteria, Acidobacteria, Proteobacteria (α-, β-, γ-, δ-, ε-proteobacteria, other proteobacteria)
Number of Operons	26,784,334
Number of hairpins	24,802,708
Ratio of *cluster*:*single* hairpins	58:42
Percentage of GC rich genomes	55%


(A)
(B)
% hairpin predictions in Bordetella pertussis I669 (NZ_CP010265)
at different distance ranges from stop codon (i.e operon end)
0-5 nt
6-10 nt
11-15 nt
16-25 nt
26-35 nt
51+ nt
36-50 nt
43.3%
22%
12.5%
11.2%
5.11%
3.16%
2.64%
Download prediction table

PREDICTION RESULTS

Bacteria (NCBI Id)	Bordetella pertussis I669 (NZ_CP010265)
Phylum	Proteobacteria (Class - Betaproteobacteria)
Genome size (nt)	4103417
Total no. of operons in genome	2308
Number of predicted single hairpin	627
Number of predicted cluster hairpin	1681

Bacteria : *Bordetella pertussis I669*

NCBI Id : NZ_CP010265

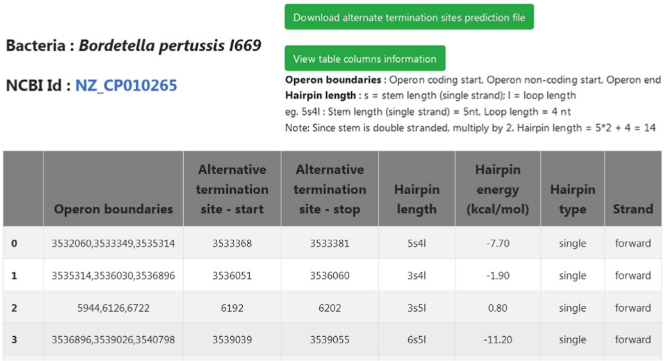
	Strand	Operon start	Operon end	Hairpin start	Hairpin end	Hairpin associated energy (kcal/mol)	Hairpin Type	# hairpins
0	forward	1	5910	5910	5938	-4.55	Cluster	2
1	forward	5944	6126	6135	6170	-9.55	Cluster	2
2	forward	6722	7912	7940	7947	1.20	Single	1
3	forward	8541	10270	10271	10280	0.10	Single	1

Fig. 2. Example output from the INTERPIN database. (A) The figure shows the inferred terminators for *B. pertussis* I669; NCBI Id: NZ_CP010265. The table shows the genome size, number of predicted operons, *clusters*, and *single* hairpins. The NCBI ID is linked to take the user to the NCBI→nucleotide page of the organism. The pie chart on the right shows the distribution of the hairpin frequency with distance from the stop codon. The table in (B) shows all hairpin predictions for the bacterium. Details like hairpin boundaries, energy, type, and corresponding operon boundaries are given. Only the first four rows of the output are shown as an example.

associated operon, hairpin boundaries, energies, lengths, as well as type and strand (Fig. 3). A summary table for all genomes is given under the "NCBI" tab as well.

3.2. Visualization of predictions

The Integrated Genome Viewer (IGV) Java applet [38] has been used to aid the visualization of hairpins. This can be activated by selecting the 'click here to view genome browser' button on the results page discussed above. The output window displays three tracks - a hairpin track that shows the predicted hairpins on the genome, the operon track shows the operons predicted by Molquest (http://molquest.com/molquest.phtml?topic=reference_license (accessed May 17, 2021)) for that genome, and the annotation track shows gene annotations and features information available in NCBI. Each track can be hidden or configured independently by using the settings symbol beside each track.


Download alternate termination sites prediction file
View table columns information

Bacteria : *Bordetella pertussis I669*

NCBI Id : NZ_CP010265

Operon boundaries : Operon coding start, Operon non-coding start, Operon end
Hairpin length : s = stem length (single strand); l = loop length
eg. 5s4l : Stem length (single strand) = 5nt, Loop length = 4 nt
Note: Since stem is double stranded, multiply by 2. Hairpin length = 5*2 + 4 = 14

	Operon boundaries	Alternative termination site - start	Alternative termination site - stop	Hairpin length	Hairpin energy (kcal/mol)	Hairpin type	Strand
0	3532060,3533349,3535314	3533368	3533381	5s4l	-7.70	single	forward
1	3535314,3536030,3536896	3536051	3536060	3s4l	-1.90	single	forward
2	5944,6126,6722	6192	6202	3s5l	0.80	single	forward
3	3536896,3539026,3540798	3539039	3539055	6s5l	-11.20	single	forward

Fig. 3. Snapshot from the INTERPIN webpage showing the alternate termination sites predicted for *Bordetella pertussis* I669 (NZ_CP010265). The table gives details of the associated operon, hairpin boundaries, energy, type, and well as strands. Only the first four rows of the output are shown.

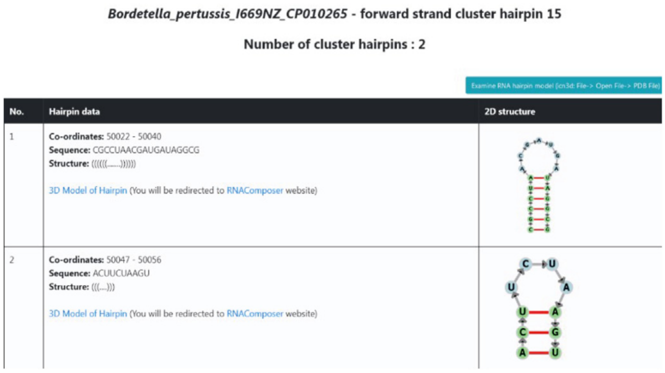
The users can zoom in to visualize the location of the hairpin with respect to any gene or operon. Predictions for forward and reverse strands are segregated for both hairpin and operon tracks. An example is shown in Fig. 4. Interactive zooming at appropriate locations or finding hairpins around any gene by searching using gene names gives flexibility to the user to easily retrieve information. Clicking any hairpin/operon/gene annotation gives more information about that entity.

The hairpin secondary and tertiary structures can also be visualized after they have been located within the IGV window. When the user clicks on any marked hairpin, it gets selected, as seen on the left-hand side in Fig. 4B). On clicking the hyperlink, the user is taken to a new tab. This page displays the organism name, hairpin prediction number, strand information, secondary structure sequence, and its 2D diagram using ViennaRNA [39] (Fig. 5). A statistic of the type and frequency of base pairs present in the predicted hairpin stems is given in Table S2. Additionally, users can click the “3D model of hairpin” link given along with each hairpin, which is redirected to the RNAComposer [40] web service and shows the three-dimensional (3D) model generated for the selected hairpin. After obtaining the 3D structure and downloading the corresponding PDB file from the RNAComposer result page, the user can interactively view and analyze the same using icn3d [41], which is also linked on the top right side of the page (Fig. 5).

3.3. Various statistics on terminators from the INTERPIN repository

INTERPIN contains hairpin data for ITT in 93% of the total operons. 61% of these hairpins are predicted to be *cluster* hairpins, while the rest are *single* hairpins. We eliminate outliers before taking the statistics. Briefly, we first found the operon boundaries and defined IR using them. We used the sequence −20 to +270 to find RNA secondary structure folding and selected hairpins from them. A two-dimensional histogram is used to find the counts of various stem and loop length combinations for the predicted hairpins. The lowermost counts for the combinations constituting 5% of the total cases are rejected, retaining only those hairpin loop-stem length combinations satisfying the threshold.

Fusobacteria has the highest percentage of the predicted *single* hairpin (55%), while Actinobacteria has the highest percentage of predicted *cluster* hairpins (65%). (More details in section R1 in the supplementary file). Most of the hairpins are found close to the stop codon with ~75% of hairpins lying within 25 nt from the stop codon.



Bordetella_pertussis_I669NZ_CP010265 **- forward strand cluster hairpin 15**

Number of cluster hairpins : 2

Examine RNA hairpin model (icn3d: File-> Open File-> PDB File)

No.	Hairpin data	2D structure
1	**Co-ordinates:** 50022 - 50040 **Sequence:** CGCCUAACGAUGAUAGGCG **Structure:** ((((((.......)))))) 3D Model of Hairpin (You will be redirected to RNAComposer website)	
2	**Co-ordinates:** 50047 - 50056 **Sequence:** ACUUCUAAGU **Structure:** (((....))) 3D Model of Hairpin (You will be redirected to RNAComposer website)	

Fig. 4. Integrated visualization of the intrinsic transcription termination hairpins, associated operon, and annotation on the genome of *B. pertussis* I669; NCBI ID: NZ_CP010265. (A) Hairpin track with the hairpin locations marked green and red for forward and reverse strands, respectively. The red and yellow colors indicate the operons on the forward and reverse strands, respectively, which separate the hairpins. (B) The maximally zoomed inset from (A) covers the ~47–52 kb genome region. To obtain a desired zoomed inset, the gene name or the location boundaries can be input in the search bar, as shown.

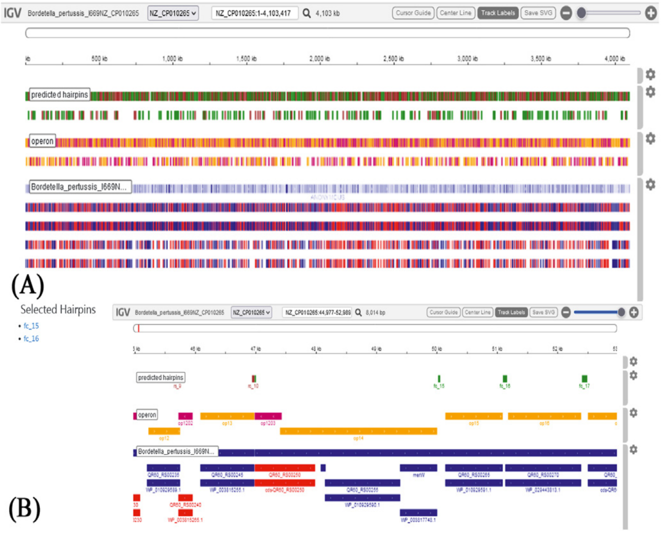


Fig. 5. Details of selected *cluster* hairpin (fc_15) from the IGV viewer page. The location, sequence, structure, and link to generate the 3D model using RNAComposer are given for each hairpin. ‘icn3d’ buttons on the top right take the user to the integrated PDB file analysis tool.

Out of all bacteria analyzed, 55% are GC-rich. The GC content and frequency of *cluster* hairpins found in a genome are highly correlated with a correlation coefficient of 0.94, showing that GC-rich genomes prefer *cluster* hairpins over *single* hairpins for termination. Energy scores in the range of −5 to 5 kcal/mol are the most favored energies for *cluster* and *single* hairpins, with 83% of total hairpin energy scores in this range. The minimum and maximum *cluster* hairpin sizes are 2 and 18, respectively. 93% of *cluster* hairpins constitute 2–5 smaller hairpins (see Table S3). On average, 86% of *cluster* hairpins and 85% of *single* hairpins have stem lengths between 4 and 18 nt (2–9 base pairs) and loop lengths between 4 and 11 nt. The methods for obtaining all the data and analysis have previously been described [31].

3.4. Comparison of INTERPIN predicted hairpins with in vivo validated hairpins

A total of 1400 published terminator sequences for *B. subtilis* (NC_000964) [16,17], and 599 for *E. coli* K-12 MG1655 (NC_000913/U000960) [14,15] were taken for the study. We labeled these as “experimental hairpins”. The statistics show that the hairpins partially or fully overlap in 64% of the operons from *B. subtilis*, and 45% in *E. coli* (Fig. 6; all except Case1 and Case4). Case 1 presents instances where the experimental hairpin is before the first predicted hairpin from the stop codon (*identified* hairpin). Statistics reveal that 25/91/50/7 and 2/23/57/11 experimental hairpin cases are present at 1–4/5-9/10–24/25-50 base distance, respectively, from the stop codon corresponding to *B. subtilis* and *E. coli*. If we check for alternate sites for Case 4, 74 IRs out of 174 have alternate termination units overlap with experimental hairpins for *B. subtilis* and 83 IRs out of 195 IRs in *E. coli*. Taking these additional overlapping cases, the coverage goes up to 72% and 61%, corresponding to *B. subtilis* and *E. coli*, respectively (Table 2). For the remaining, where no overlap was seen, the distance between the two hairpins is shown in Table 3. About 16% and 36% of non-overlapping INTERPIN hairpins from *B. subtilis* and *E. coli*, respectively, lie within the next 14 nt of the experimental hairpin indicating that they could be part of a *cluster* hairpin arrangement. Notwithstanding, the remaining non-overlapping INTERPIN hairpins having

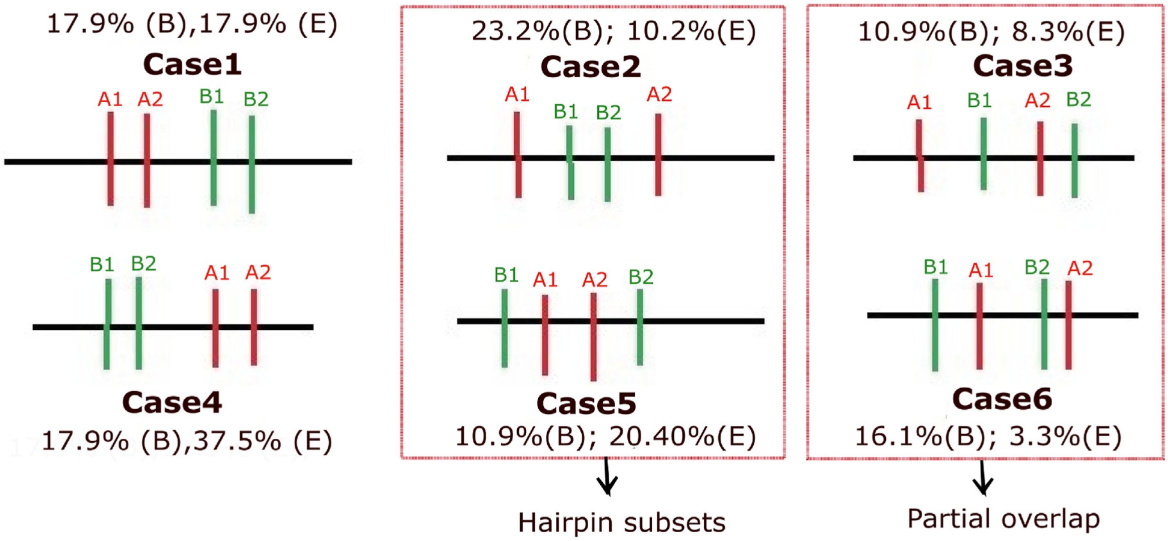


Fig. 6. Diagram comparing experimental hairpins and hairpins identified by INTERPIN for *B. subtilis* (B) and *E. coli* (E). The labels A1 and A2 indicate the start and end of the experimental hairpin, while B1 and B2 point to the start and end of the INTERPIN hairpin. The percentages have been calculated with respect to the total experimental hairpins found in the IR: 967 for *B. subtilis,* and 520 for *E. coli*. Out of this, 30 and 13 cases in *B. subtilis* and *E. coli*, respectively, did not have a prediction by INTERPIN.

Table 2
Statistics showing alternate termination sites in *E. coli* and *B. subtilis*.

Alternate sites					
Organism	**Case 2**	**Case 3**	**Case 5**	**Case 6**	Total
B. subtilis	38	8	9	19	74
E. coli	39	15	24	5	83

Table 3
Statistics showing the number of cases where the location of the experimental hairpins and INTERPIN predicted hairpins having no overlap and the distance in nt between the two hairpins is measured.

Experimental hairpin first	**Distance in base pairs between experimental and INTERPIN hairpins**					
	1–14	**15–29**	**30–49**	**50–99**	**100–200**	**Total**
B. subtilis	94	59	14	6	0	173
E. coli	38	36	13	5	1	93
INTERPIN hairpin first	**1–14**	**15–49**	**50–99**	**100–200**	**>200**	**Total**
B. subtilis	16	18	6	4	56	100
E. coli	40	19	11	12	30	112

experimental cases from Case1 and Case4 could be due to the selection criteria of ITT sites based on termination efficiency where additional parameters such as ($3 \leq$ loop length ≤ 10; $\Delta G < -7$ kcal/mol and $4 \leq$ stem nt ≤ 17) or high transcript expression levels have been imposed. This may be why 17% and 11% of the experimental hairpins in *B. subtilis*, and *E. coli*, respectively, are detected in the operonic regions.

Many studies compare experimentally derived hairpins to predictions and provide an analysis using the percentage of sequence matches and mismatches to understand the ITT efficiency of terminators [8,9,19,42]. For stochastic grammar [43,44] or statistical profiling [45,46] based methods, the parameters for prediction are derived from the experimental hairpins themselves, which show promising results on some bacteria while not for others (for phylogenetically different bacteria, for example). We have taken the *in vivo* validated set of 1400 and 599 hairpins from *B. subtilis* and *E. coli*, respectively, to find their genomic-average energy values. Now, for all predicted hairpins in the two bacteria which matched in location to the *in vivo* set, the ratio of energy to average energy was calculated to estimate terminator efficiency. The hairpins with a ratio >1 are expected to have higher termination efficiency. In *E. coli*, 89.7% of total predicted hairpins and 95% in *B. subtilis* have a ratio >1, suggesting that most of the predicted hairpins are efficient terminators, estimated in the background of experimentally determined values.

Looking for the presence or absence of poly A/U pattern, we saw that only 10% of operons have a pattern within four nt from the hairpin 3′ end when we search for a pattern immediately after the hairpin (hairpin formed inside the RNA exit channel). This number improves to 29% if a distance up to 10 nt is seen. If we consider the case where the hairpin is formed just outside of the exit channel; i.e., a minimum of 7 nt gap (5 nt RNA exit channel and 2–3 nt spacer) exists between poly A/U pattern and the hairpin 3’ end, 11% operons lie within 10 nt from the hairpin end while 26% within 15 nt (see section R2 in supplementary file), showing that poly A/U pattern is not a necessary feature of ITT.

Our results are also corroborated by findings of the GeSTer database, where almost 90% of terminators in mycobacterial species have a suboptimal (<3 Us) or no U-tract following the hairpin stem [19](see section R3 in supplementary file). Although a few of these terminators were shown to fail in causing intrinsic termination *in vitro* and *in vivo* [47], Ahmad et al. reconfirmed that the U-tract is dispensable for ITT in bacteria, both by *in vitro* and *in vivo* experiments, in their recently published work [48].

3.5. Missed ITT

3.5.1. Alternate termination sites

We found several *cluster* and *single* hairpins for each operon in a bacteria using the method explained in Ref. [31]. After that, we chose the hairpin closest to the stop codon (defined as the *identified* hairpin), which would be the earliest to form and competent to cause ITT. If a *cluster* or *single* hairpin cannot cause ITT at the first instance from the stop codon, the *cluster* or *single* hairpins at the downstream locations can act as alternate termination sites to prevent a readthrough to enforce ITT.

We looked for other hairpins downstream of the first ITT site to check whether the next in line would be a *cluster* or a *single* hairpin. Out of all 24,802,708 operons, 12% have a single termination site, whereas 80% are *cluster* hairpins. About 27% of the operons have one additional termination site, and 31% have two additional sites (Table 4). One operon has a maximum of 10 termination sites. We further checked the type of hairpin for operons with one and two additional sites (Table 5). It can be seen that the presence of a *cluster* hairpin is predominant — CC and CCC constitute 37%, while SS and SSS are restricted to 7%, where C denotes '*cluster*' and S denotes '*single*' hairpin. Mixed combinations (two Cs in three alternate sites: 26%; two Ss in three alternate sites: 16%) also indicate a higher occurrence of the *cluster* hairpins.

Table 4
Distribution of the number of ITT sites for operons.

# ITT sites	Cases	%
1	2,971,446	12
2	6,808,757	27
3	7,651,794	31
4	5,198,585	21
5	1,868,833	8
6	285,863	1
7	17,143	<1
8–10	287	<1

Table 5
Statistics of hairpin type for operon with two and three ITT sites. The percentages have been calculated with respect to the total operons with 2 and 3 predicted termination sites.

Hairpin combination[a]	Frequency (%)
CC	3,422,196 (24)
SC	1,590,504 (11)
CS	1,154,593 (8)
SS	641,464 (4)
CCC	1,843,366 (13)
SCC	1,418,872 (10)
CSC	1,066,316 (7)
SSC	810,657 (6)
CCS	986,215 (9)
SCS	640,218 (4)
CSS	495,883 (3)
SSS	390,267 (3)

[a] C = *cluster*, S = *single*.

3.5.2. Co-occurrence with Rho-termination

Rho-dependent transcription termination is a major transcription termination process in some bacterial species. The presence of hairpins is used as a regulatory mechanism for *Rho-dependent* termination in the leader region in operons. One such regulation is proved by the Rho-antagonizing RNA element (RARE). The RARE is a sequence in leader RNA that can form alternating structures and regulate Rho protein. When it is part of the stem-loop of a hairpin, it is inaccessible to Rho. In open confirmation, it traps Rho in an inactive form and allows operon expression (suppressing Rho termination) [49]. The hairpin inactivates the RNAP elongation complex, leading to inefficient *Rho-dependent* termination [50].

We examined a few species proposed to majorly use *Rho-dependent* termination. RhoTermPredict [51] was used to predict the Rho utilization sites (RUT) in *E. coli, S. enterica* (proteobacteria), and *B. subtilis* (firmicutes). Table 6 shows the occurrence of operons where a RUT site, hairpin, or both have been detected. When both RUT site and hairpins are detected, only 2.4% of operons in *B. subtilis* and 3.1% in *E. coli* and *S. enterica* have a RUT site first, followed by hairpin within 10–100 nt (in the termination range). If we consider only these operons where the RUT site is closest to the stop codon, 72% of such cases from *B. subtilis* and 66% of cases from *E. coli* and *S. enterica* have a hairpin 10-100 nt downstream, making them the most likely *Rho-dependent* transcription termination site.

Table 6
Statistics of occurrence of hairpins and Rho utilization sites in three bacterial species.

Category	*B. subtilis*	%	*E. coli* + *S. enterica*	%
Hairpin only	105,005	44	1,319,634	34
Rho sites only	12,891	5	125,454	3
Both Hairpin + Rho site	120,411	51	2,451,495	63
a) Hairpin first	113,792		2,270,671	
b) Rho site first	6619 (72)[a]		180,824 (66)*	

[a] Percentage of cases where the hairpins are present in 10–100 nt downstream of the RUT site. This percentage is calculated with respect to the operons where both the RUT site and hairpin are found, but the RUT site lies closer to the stop codon.

It is to be noted that the presence of a RUT site alone does not guarantee *Rho-dependent* transcription termination. It has been found that though ChIP-ChIP data showed that Rho is associated with all nascent RNA, it accounted for only 20% of transcription terminations [52]. A prolonged pause is required for *Rho-dependent* termination in the mgtA leader region. This pause is stabilized by a hairpin in the leader sequence [53] (11 nt upstream from the pause site/Rho termination site). Typically, the termination region in *Rho-dependent* termination is spread over a series of weak pauses ranging from 10 to 20 nt to up to ~100 nt downstream of the RUT site [51]. Our hairpins are predicted after the stop codon. We have looked for the presence of a hairpin in the termination region of Rho (10–100 nt after the Rho site). It should also be considered that a hairpin in the leader region promotes Rho-termination. At the same time, the one at the 3' end may hamper *Rho-dependent* termination since the occurrence of the hairpin at the RUT site may make it inaccessible for Rho-protein binding. The above observations can verify that the bulk of the hairpins presented in INTERPIN are credible ITT sites.

3.6. Matching the location of the predicted hairpin with the RNA-seq derived hairpin

Matching the location of the predicted ITT site with those inferred from RNA-seq data offers an alternate approach to validate prediction against *in vivo* data. This was already used by us on a smaller scale previously [31]. In this work, we obtained all downloadable RNA-seq data (380 non-redundant bacterial genomes) from the SRA website (http://www.ncbi.nlm.nih.gov/sra) as of March 2022. We first aligned the RNA-seq data of these 380 bacteria with corresponding reference genomes. Of these, 18 did not have properly paired data or had <5% alignment with the reference genome and were eliminated from the study.

On average, 36% IRs in each genome have the coverage of RNA-seq reads >90% in the corresponding operon (read depth at a genome location ≥1). We found regions of slope drop in a cumulative frequency plot for RNA-seq reads in all these IRs. Next, we took the first such region and matched its location with each IR's predicted hairpins from INTERPIN. Hairpins closest to or present in this region were determined and found in 84% of these IRs. We called these *RNA-seq derived* hairpins. Now, we matched these hairpins with the predicted hairpins from INTERPIN. On average, 66% of predicted hairpins partially or fully overlapped with the *RNA-seq derived* hairpins. The hairpins matching or partially overlapping with *RNA-seq derived* termination sites are *clusters* in 65.6% of cases and *single* in the rest. Suppose we also include the alternate termination sites predicted for each genome and match their location to *RNA-seq derived* hairpin. In that case, we see that an additional 7.5% of IRs have an exact match, and 24% of IRs have an overlap between an alternate termination site and *RNA-seq derived* hairpin, taking the cumulative overlap to 97.5% of cases.

3.7. Comparison of INTERPIN predictions with SVM based model predictions

A new predictor, iTerm-PseKNC, was published in 2018 [54]. This is an SVM model trained on 280 experimentally confirmed ITT sites and 560 non-terminator sequences in *E. coli*, to identify terminator sequences in bacteria. A comparison of prediction results was made with INTERPIN.

3.7.1. Initial testing on control set of non-terminating sequences

Initially, the model was run on a small set of 308 bacterial genomes. We first extracted the identified hairpin sequences and alternate termination sites (represented by T in Fig. 7) in one file, the non-terminating region in between ITT sites (represented by N in Fig. 7) in the second file, and the 100 nt after +270 from stop codon (represented by A and B in Fig. 7) in the third file. The SVM model calculates the probability of each sequence being a terminator or non-terminator for each of these files.

We take that for a sequence, the event of being a terminator or a non-terminator is mutually exclusive. So, P(terminating) + P(non-terminating) = 1. Using this, the probability of termination is found for each sequence. So, if the SVM outputs a sequence as a non-terminator with probability ‘p’, its probability of termination is ‘1-p’.

If a given sequence is larger than 81 nt its probability is given by creating multiple sequences of size 81 nt, using the sliding window approach. We have averaged all probability to obtain a single value in such cases. After this, we calculated the histogram for the frequency (in percentage) vs. probability of termination for all three cases. The probabilities have been divided into four groups (Table 7). Greater than 71% of hairpins predicted as ITT sites by INTERPIN also have high termination probabilities vis iTerm-PseKNC. However, the non-terminating sequences also show high termination probabilities via the SVM model prediction. The sequences 100 nt from +270 from the stop codon is showing a low probability of termination (~75% of sequences have a probability of termination below 0.5), which is correct because terminators should not be found so far from the stop codon.

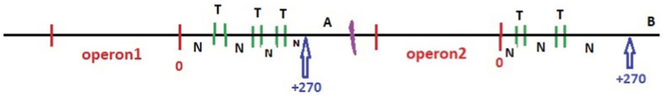


Fig. 7. Example diagram showing the different sequences taken as input for the SVM model. Two operons: operon 1 and operon2, are shown. T represents ITT and alternate termination sites, while N represents non-terminating sites. A and B are sequences 100 nt after +270 end from the stop codon, whose end is marked in purple just before the start of operon2. The non-coding region might extend even beyond 370 nt, but that is not taken for analysis.

Table 7
The probability of ITT for terminating, non-terminating, and sequences 100 nt after +270 from the stop codon for the initial 308 bacteria set and for the control set of non-terminating sequences, obtained from the SVM model for ITT prediction.

Sequences	Probability for ITT			
Initial 308 bacteria set	[0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1)
Terminating	11.51%	9.12%	7.58%	71.78%
Non-terminating	1.95%	1.79%	1.98%	94.28%
100 nt after +270 from stop codon	49.76%	24.91%	12.92%	12.40%
Control set	**[0,0.25)**	**[0.25,0.5)**	**[0.5,0.75)**	**[0.75,1)**
	0.6%	0.6%	1.9%	97.4%

To check for the performance of the SVM model, we ran the SVM model on a set of 567 experimentally known non-terminating sequences from de Hoon and co-workers [6]. The probability of termination was calculated as before (Table 7). Here too we see that >95% of sequences are being assigned high probabilities of termination, which is an incorrect classification. On checking for the relation between the probability of termination and sequence length, we see an inverse correlation of −0.8. This means that when the SVM model encounters longer sequences, it has a higher chance to classify them as non-terminating sequences. It shows that the trained model may be biased for selecting only shorter sequences as terminators.

3.7.2. Results on the whole bacterial dataset

The calculation was extended to find the probability of termination for our whole bacterial dataset, as in the previous section (Table 8). The probability of termination and correlation of these probabilities with sequence length was found to be −0.6. The SVM model has no discriminating power, which may be because it uses sequence-based features alone for training the model.

4. Conclusion

Apart from being the most extensive collection of intrinsic transcription terminators, INTERPIN, provides a one-pit stop for visualizing all hairpin predictions with respect to genes and operons as well as hairpin secondary and tertiary structures. It highlights the presence of *cluster* hairpins in bacterial genomes. Our work provides an initial base for finding potential intrinsic terminators that can be studied and validated *in vivo*. These may also aid studies that find targets for drugs in virulent bacteria, where this has not been done before. Our ability to make predictions for both AT and GC-rich genomes increases the coverage of our repository, which has not been achieved so far by any other program. This can also be used to improve genome annotation and to get predictions

Table 8
The probability distribution of ITT for terminating, and non-terminating sequences by applying the SVM model for ITT prediction.

Sequences	Probability for ITT			
	[0,0.25)	[0.25,0.5)	[0.5,0.75)	[0.75,1)
Terminating	11.53%	8.88%	7.47%	72.12%
Non-terminating	1.84%	1.73%	1.90%	94.54%

for further analysis to improve the understanding of the ITT pathway.

NCBI sourced data

NCBI is used to obtain the genome FASTA file and gene annotation (*.gbk and *.gtt format) files. These files generate tracks in the IGV application on the web server and are stored in internal databases. These tracks are derived from publicly-accessible data in the NCBI RefSeq, Gene, and GenBank databases. The codes for the same have been shared, and links are given in the browser and code availability section below.

Browser and code availability

The database can be freely accessed at http://pallab.cds.iisc.ac.in/INTERPIN/. The program for local use is provided via the GitHub repository and can be accessed using the following link: https://github.com/swati375/interpin. A detailed README file gives information on how to download and run the software, the prerequisites, and also explains the output format.

CRediT author statement

Swati Gupta: Methodology, Software, data curation, Writing- Original draft preparation, Visualization, Investigation. Debnath Pal: Conceptualization, Methodology, Writing- Reviewing and Editing. Namrata Padmashali: Software, Visualization.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

SG thanks the Department of Biotechnology (DBT), New Delhi for a fellowship. DP thanks the DBT for supporting the computational facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biochi.2023.07.018.

References

[1] P.C. Bevilacqua, J.M. Blose, Structures, kinetics, thermodynamics, and biological functions of RNA hairpins, Annu. Rev. Phys. Chem. 59 (2008) 79–103.

[2] H.F. Noller, Structure of ribosomal RNA, Annu. Rev. Biochem. 53 (1984) 119–162.

[3] W.S. Yarnell, J.W. Roberts, Mechanism of intrinsic transcription termination and antitermination, Science 284 (1999) 611–615.

[4] G.W. Witherell, O.C. Uhlenbeck, Specific RNA binding by Q beta coat protein, Biochemistry 28 (1989) 71–76.

[5] H.N. Wu, O.C. Uhlenbeck, Role of a bulged A residue in a specific RNA-protein interaction, Biochemistry 26 (1987) 8221–8227.

[6] M.J. de Hoon, Y. Makita, K. Nakai, S. Miyano, Prediction of transcriptional terminators in Bacillus subtilis and related species, PLoS Comput. Biol. 1 (2005) e25.

[7] M.D. Ermolaeva, H.G. Khalak, O. White, H.O. Smith, S.L. Salzberg, Prediction of transcription terminators in bacterial genomes, J. Mol. Biol. 301 (2000) 27–33.

[8] P.P. Gardner, L. Barquist, A. Bateman, E.P. Nawrocki, Z. Weinberg, RNIE: genome-wide prediction of bacterial intrinsic terminators, Nucleic Acids Res. 39 (2011) 5845–5852.

[9] C.L. Kingsford, K. Ayanbule, S.L. Salzberg, Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake, Genome Biol. 8 (2007) R22.

[10] A. Mitra, A.K. Kesarwani, D. Pal, V. Nagaraja, WebGeSTer DB–a transcription terminator database, Nucleic Acids Res. 39 (2011) D129–D135.

[11] X.-F. Wan, D. Xu, Intrinsic terminator prediction and its application in Synechococcus sp. WH8102, J. Comput. Sci. Technol. 20 (2005) 465–482.

[12] Y.J. Chen, P. Liu, A.A. Nielsen, J.A. Brophy, K. Clancy, T. Peterson, C.A. Voigt, Characterization of 582 natural and synthetic terminators and quantification of their design constraints, Nat. Methods 10 (2013) 659–664.

[13] I. Gusarov, E. Nudler, The mechanism of intrinsic transcription termination, Mol. Cell. 3 (1999) 495–504.

[14] X. Ju, D. Li, S. Liu, Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria, Nat Microbiol 4 (2019) 1907–1918.

[15] J.B. Lalanne, J.C. Taggart, M.S. Guo, L. Herzel, A. Schieler, G.W. Li, Evolutionary convergence of pathway-specific enzyme expression stoichiometry, Cell 173 (2018) 749–761 e738.

[16] Z.F. Mandell, R.T. Oshiro, A.V. Yakhnin, R. Vishwakarma, M. Kashlev, D.B. Kearns, P. Babitzke, NusG is an intrinsic transcription termination factor that stimulates motility and coordinates gene expression with NusA, Elife 10 (2021).

[17] S. Mondal, A.V. Yakhnin, A. Sebastian, I. Albert, P. Babitzke, NusA-dependent transcription termination prevents misregulation of global gene expression, Nat Microbiol 1 (2016) 15007.

[18] J.M. Peters, A.D. Vangeloff, R. Landick, Bacterial transcription terminators: the RNA 3'-end chronicles, J. Mol. Biol. 412 (2011) 793–813.

[19] S. Unniraman, R. Prakash, V. Nagaraja, Alternate paradigm for intrinsic transcription termination in eubacteria, J. Biol. Chem. 276 (2001) 41850–41855.

[20] K.S. Wilson, P.H. von Hippel, Transcription termination at intrinsic terminators: the role of the RNA hairpin, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 8793–8797.

[21] N. Komissarova, J. Becker, S. Solter, M. Kireeva, M. Kashlev, Shortening of RNA: DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination, Mol. Cell. 10 (2002) 1151–1162.

[22] D. Pörschke, Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix, Biophys. Chem. 1 (1974) 381–386.

[23] T.J. Santangelo, J.W. Roberts, Forward translocation is the natural pathway of RNA release at an intrinsic terminator, Mol. Cell. 14 (2004) 117–126.

[24] P.H. von Hippel, T.D. Yager, Transcript elongation and termination are competitive kinetic processes, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 2307–2311.

[25] V. Brendel, E.N. Trifonov, A computer algorithm for testing potential prokaryotic terminators, Nucleic Acids Res. 12 (1984) 4411–4427.

[26] A.R. Castillo, S.S. Arevalo, A.J. Woodruff, K.M. Ottemann, Experimental analysis of Helicobacter pylori transcriptional terminators suggests this microbe uses both intrinsic and factor-dependent termination, Mol. Microbiol. 67 (2008) 155–170.

[27] Y. d'Aubenton Carafa, E. Brody, C. Thermes, Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures, J. Mol. Biol. 216 (1990) 835–858.

[28] Z.X. Deng, T. Kieser, D.A. Hopwood, Activity of a Streptomyces transcriptional terminator in Escherichia coli, Nucleic Acids Res. 15 (1987) 2665–2675.

[29] A. Mitra, K. Angamuthu, V. Nagaraja, Genome-wide analysis of the intrinsic terminators of transcription across the genus Mycobacterium, Tuberculosis 88 (2008) 566–575.

[30] E.A. Lesnik, R. Sampath, H.B. Levene, T.J. Henderson, J.A. McNeil, D.J. Ecker, Prediction of rho-independent transcriptional terminators in Escherichia coli, Nucleic Acids Res. 29 (2001) 3583–3594.

[31] S. Gupta, D. Pal, Clusters of hairpins induce intrinsic transcription termination in bacteria, Sci. Rep. 11 (2021) 16194.

[32] J.H. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, C.W. Pleij, Structural parameters affecting the kinetics of RNA hairpin formation, Nucleic Acids Res. 34 (2006) 3568–3576.

[33] R.M. Winslow, R.A. Lazzarini, The rates of synthesis and chain elongation of ribonucleic acid in Escherichia coli, J. Biol. Chem. 244 (1969) 1128–1136.

[34] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (2003) 3406–3415.

[35] A.C. Parte, LPSN–list of prokaryotic names with standing in nomenclature, Nucleic Acids Res. 42 (2014) D613–D616.

[36] A.C. Parte, J. Sarda Carbasse, J.P. Meier-Kolthoff, L.C. Reimer, M. Goker, List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ, Int. J. Syst. Evol. Microbiol. 70 (2020) 5607–5612.

[37] F.D. Ciccarelli, T. Doerks, C. Von Mering, C.J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life, Science 311 (2006) 1283–1287.

[38] J.T. Robinson, H. Thorvaldsdóttir, D. Turner, J.P. Mesirov, igv. js: An embeddable javascript implementation of the integrative genomics viewer (IGV), bioRxiv (2020).

[39] R. Lorenz, S.H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA package 2.0, Algorithm Mol. Biol. 6 (2011) 26.

[40] M. Biesiada, K.J. Purzycka, M. Szachniuk, J. Blazewicz, R.W. Adamiak, Automated RNA 3D structure prediction with RNAComposer, Methods Mol. Biol. 1490 (2016) 199–215.

[41] J. Wang, P. Youkharibache, D. Zhang, C.J. Lanczycki, R.C. Geer, T. Madej, L. Phan, M. Ward, S. Lu, G.H. Marchler, Y. Wang, S.H. Bryant, L.Y. Geer, A. Marchler-Bauer, iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures, Bioinformatics 36 (2020) 131–135.

[42] M. Naville, A. Ghuillot-Gaudeffroy, A. Marchais, D. Gautheret, ARNold: a web tool for the prediction of Rho-independent transcription terminators, RNA Biol. 8 (2011) 11–13.

[43] J.W. Anderson, P.A. Haas, L.A. Mathieson, V. Volynkin, R. Lyngso, P. Tataru, J. Hein, Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information, Bioinformatics 29 (2013) 704–710.

[44] B. Knudsen, J. Hein, Pfold: RNA secondary structure prediction using stochastic context-free grammars, Nucleic Acids Res. 31 (2003) 3423–3428.

[45] D. Gautheret, A. Lambert, Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles, J. Mol. Biol. 313 (2001) 1003–1011.

[46] A. Lambert, J.F. Fontaine, M. Legendre, F. Leclerc, E. Permal, F. Major, H. Putzer, O. Delfour, B. Michot, D. Gautheret, The ERPIN server: an interface to profile-based RNA motif identification, Nucleic Acids Res. 32 (2004) W160–W165.

[47] A. Czyz, R.A. Mooney, A. Iaconi, R. Landick, Mycobacterial RNA polymerase requires a U-tract at intrinsic terminators and is aided by NusG at suboptimal terminators, mBio 5 (2014) e00931.

[48] E. Ahmad, S.R. Hegde, V. Nagaraja, Revisiting intrinsic transcription termination in mycobacteria: U-tract downstream of secondary structure is dispensable for termination, Biochem. Biophys. Res. Commun. 522 (2020) 226–232.

[49] A. Sevostyanova, E.A. Groisman, An RNA motif advances transcription by preventing Rho-dependent termination, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E6835–E6843.

[50] D. Dutta, J. Chalissery, R. Sen, Transcription termination factor rho prefers catalytically active elongation complexes for releasing RNA, J. Biol. Chem. 283 (2008) 20243–20251.

[51] M. Di Salvo, S. Puccio, C. Peano, S. Lacour, P. Alifano, RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases, BMC Bioinf. 20 (2019) 117.

[52] J.M. Peters, R.A. Mooney, P.F. Kuan, J.L. Rowland, S. Keles, R. Landick, Rho directs widespread termination of intragenic and stable RNA transcription, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 15406–15411.

[53] K. Hollands, A. Sevostiyanova, E.A. Groisman, Unusually long-lived pause required for regulation of a Rho-dependent transcription terminator, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) E1999–E2007.

[54] C.-Q. Feng, Z.-Y. Zhang, X.-J. Zhu, Y. Lin, W. Chen, H. Tang, H. Lin, iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators, Bioinformatics 35 (2019) 1469–1477.