BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT

Thomas A. Hall
Department of Microbiology, North Carolina State University, Raleigh, NC 27695, USA

ABSTRACT

BioEdit is a user-friendly sequence alignment editor and analysis package that is offered free of charge for Windows 95/98/NT systems. BioEdit is a full-featured nucleic acid/protein alignment editor that offers several modes of easy hand-alignment, split-window views, user-defined colors, information-based shading, auto-integration with ClustalW (1), local/internet BLAST (2), restriction-mapping, annotatable plasmid-drawing, box-shading with full color-capability, several built-in analysis options, and a graphical interface for configuring further interfaces to automatically run external analysis programs. BioEdit is also customizable to user preferences with user-defined menu shortcuts and correct handling of all fonts. Among the built-in analyses offered are a set of RNA comparative analysis tools including covariation (3, 4), potential-pairings (3, 4) and mutual information (5) analysis. BioEdit offers the tools required to create and manipulate an alignment, run comparative analyses from the edit window, and view and analyze the data through interactive 2-D graphical matrix plots, area plots, and a rich-text editor. The following note describes a rough sample analysis of the secondary structure of bacterial RNase P RNA (excluding the high G+C Gram-Positive group) by mutual information probing. An initial scanning of mutual information data via the graphical analysis tools reveals all major helices that exist in the *E. coli* structure (6, 8).

INTRODUCTION

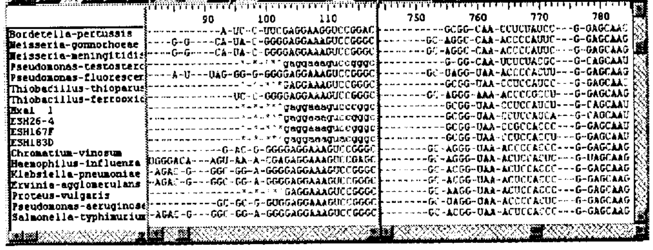
Biological sequence data is becoming available at a quickly increasing rate. It is now commonplace for researchers to analyze biological sequences and sequence alignments on a daily basis. Lagging behind the availability of sequence data, however, is software available to academic researchers for manipulating, formatting and analyzing that data in a convenient manner on their own desktop computers. Particularly lacking in this area is comprehensive software for PC-compatible computers (barring packages that cost $500+ per user license). BioEdit offers a user-friendly interface with wide variety of features, including comparative analysis tools for probing RNA secondary structure (using methods developed by Robin Gutell, 3, 5).

SAMPLE ANALYSIS

Described herein is an example of a mutual information analysis of ribonuclease P RNA from Bacteria. Mutual information refers to the amount of information shared by two positions in an alignment. It is a measure of the degree to which the residues at two positions in an alignment are *not independent of eachother* (5). A high degree of mutual information suggests a high degree of interdependence and may be a sign that two bases interact. Combined with an analysis such as "potential pairings" (3, 4), this is a method for quickly identifying probable base pairs for constructing a secondary structure model of an RNA. Mutual information (M(x,y)) describes concerted variation between positions. M(x,y) is implemented as described in Gutell, 1992 (5). Briefly, M(x,y) = H(x) + H(y) − H(xy). H(x) referes to the "entropy" (variability) of position x. $H(x) = -\Sigma fb_x * \ln(fb_x)$, $H(y) = -\Sigma fb_y * \ln(fb_y)$ and $H(xy) = -\Sigma fb_x b_y * \ln(fb_x b_y)$, where fb_x and fb_y refer to the frequencies of each base b at positions x and y, respectively, and $fb_x fb_y$ is the frequency of each base combination at positional pair x,y. In addition to M(x,y), BioEdit also calculates R1(x,y) and R2(x,y) (for a description, see 5).

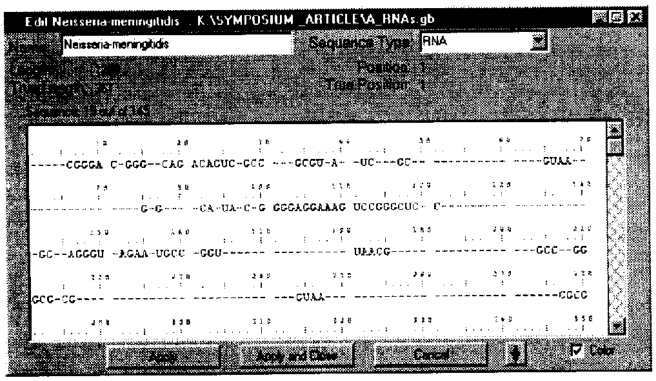
The most important starting point for phylogenetic comparative analysis is the alignment of homologous positions (4). The starting alignment for this sample analysis was the bacterial type A RNase P RNA alignment from the RNase P Database (7, http://www.mbio.ncsu.edu/RNaseP/). Recently BioEdit was used to add 40-50 sequences to this alignment (later refined by J.W. Brown) to bring the total number of sequences to 145. The BioEdit alignment window (Fig 1) offers functions to select and drag blocks of residues, dynamically grab and drag residues (residues slide in real-time with the mouse), insert or delete gaps in or around a sequence, and/or edit on screen or in an single-sequence edit box (Fig 2).

Prior to running a mutual information analysis, user preferences are set (Fig 3). Here, the mutual information preferences have been set to generate an M(x, y) matrix as well



```
                              90        100       110              750       760       770       780
Bordetella-pertussis   ------------A-UC-C-UUCGAGGAAGGUCCGGAC -----------GCGG-CAA-CCUCUAUCC---G-GAGCAAC
Neisseria-gonnorhoeae  ----G-G----CA-UA-C-GGGGAGGAAAGUCCGGGC --------GC-AGGC-CAA-ACCCCAUUC---G-GAGCAAG
Neisseria-meningitidis ----G-G----CA-UA-C-GGGGAGGAAAGUCCGGGC --------GC-AGGC-CAA-ACCCCAUUC---G-GAGCAAG
Pseudomonas-testosterc ---------------*-*-***gaggaaaguccgggc -----------G-GG-CAA-UCUCUACGC---G-CAGCAAU
Pseudomonas-fluorescen ----A-U---UAG-GG-G-GGGGAGGAAAGUCCGGGC --------GC-UAGG-UAA-ACCCCACUU---G-GAGCAAG
Thiobacillus-thioparus ---------------*-*-***GAGGAAAGUCCGGGC -----------GCGG-UAA-CCUCCAUCC---G-GAGCAAC
Thiobacillus-ferrooxic --------------UC-C-GGGGAGGAAAGUCCGGGC --------GC-AGGG-AAA-ACCCCGCCU---G-GAGCAAG
Exal  1                ---------------*-*-***gaggaaaguccc ggc -----------GCGG-UAA-CCUCCAUCU---G-CAGCAAU
ESH26-4                ---------------*-*-***gaggaaaguccgggc -----------GCGG-UAA-CCUCCAUCA---G-CAGCAAU
ESH167F                ---------------*-*-***gaggaaaguccgggc -----------GCGG-UAA-CCGCCACCC---G-GAGCAAU
ESH183D                ---------------*-*-***gaggaaaguacgggc -----------GCGG-UAA-CCUCCACCU---G-GAGCAAU
Chromatium-vinosum     ------------G-AC-G-GGGGAGGAAAGUCCGGGC --------GC-AGGG-UAA-ACCCCACCC---G-GAGCAAG
Haemophilus-influenza  UGGGACA---AGU-AA-A-CGAGAGGAAAGUCCGAGC --------GC-AGGG-UAA-ACUCCACUC---G-UAGCAAG
Klebsiella-pneumoniae  -AGAC-G---GGC-GG-A-GGGGAGGAAAGUCCGGGC --------GC-ACGG-UAA-ACUCCACCC---G-GAGCAAG
Erwinia-agglomerulans  -AGAC-G---GGC-GG-A-GGGGAGGAAAGUCCGGGC --------GC-ACGG-UAA-ACUCCACCC---G-GAGCAAG
Proteus-vulgaris       ---------------*-*-***GAGGAAAGUCCGGGC --------GC-AAGG-UAA-ACUCCACCC---G-GAGCAAG
Pseudomonas-aeruginosa -----------GC-GC-G-GUGGAGGAAAGUCCGGGC --------GC-UAGG-UAA-ACCCCACUC---G-GAGCAAG
Salmonella-typhimurium -AGAC-G---GGC-GG-A-GGGGAGGAAAGUCCGGGC --------GC-ACGG-UAA-ACUCCACCC---G-GAGCAAG
```

Figure 1. the basic BioEdit alignment window. Shown above is part of the type A bacterial RNase P RNA alignment (http://www.mbio.ncsu.edu/RNaseP/). The window is split vertically. Residues may be slid back and forth with the mouse, or sequences may be edited on-screen like in a word-processor.



Edit Neisseria-meningitidis . K:\SYMPOSIUM _ARTICLE\A_RNAs.gb

Name: Neisseria-meningitidis

Sequence Type: RNA

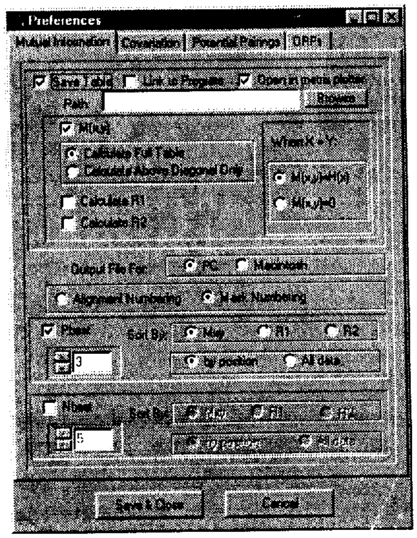
Position: 1

True Position: 1

```
-----CGGGA C-GGG--CAG ACACUC-GCC ---GCGU-A- -UC---GC-- ---------- -----GUAA--
---------- -----G-G--- -CA-UA-C-G GGGAGGAAAG UCCGGGCUC- C--------- ----------
-GC--AGGGU -AGAA-UGCC -GGU------ ---------- UAACG----- ---------- ---GCC--GG
GCG-CG---- ---------- ---------- ---GUAA--- ---------- ---------- ------CGCG
```

☑ Color

Figure 2. Single sequence edit window. If a sequence or alignment is saved in GenBank or BioEdit format, GenBank field data will be retained (LOCUS, DEFINITION, ACCESSION, PID or NID, DBSOURCE, KEYWORDS, SOURCE, REFERENCES, COMMENT and FEATURES fields will be stored). Pressing the red arrow expands the view to include these fields.



Preferences

☑ Save Table ☐ Link to Program ☑ Open in matrix plotter

Path:

☑ M(x,y)

● Calculate Full Table
○ Calculate Above Diagonal Only

☐ Calculate R1
☐ Calculate R2

When X = Y:
● M(x,y)=H(x)
○ M(x,y)=0

Output File For: ● PC ○ Macintosh

○ Alignment Numbering ● Mask Numbering

☑ Pbest Sort By: ● M(xy) ○ R1 ○ R2

3

● by position ○ All data

☐ Nbest Sort By: ● M(xy) ○ R1 ○ R2

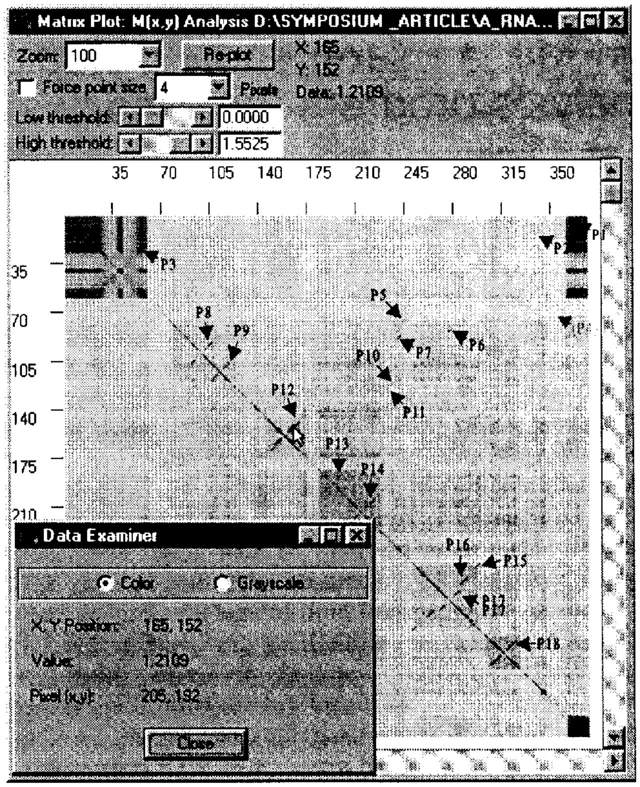
5

● by position ○ All data

Figure 3. The preferences interface.

as a list of the top 5% of M(x,y), values for each position (P-best), with associated R1(x,y) and R2(x,y) values (5, 8). *Escherichia coli* is specified as the "sequence mask" (positions existing in *E. coli* are analyzed). *E. coli* is also specified as the "numbering mask" (data are referred to by *E. coli* positions). The matrix is set to be opened automatically in the matrix plotter (fig 4, 5). The matrix plotter will plot any two-dimensional numerical matrix that is tab- or comma-delimited and has a top row and left column containing positional numbers. Positional numbers (not absolute row or column) are retained and reported in the plotter. For example, a part of an RNA may be analyzed using the full sequence as the numbering mask. The true positions in the RNA are then reported for convenient reference to an existing sequence and/or structure.

After running an M(x,y) analysis, the output is opened in the matrix plotter (Fig 4). Probable helices are seen as strings of high information running perpendicular to the diagonal. Helices labeled in Figure 4 are shown in the *E. coli* RNase P RNA secondary structure in Figure 5. The data



Matrix Plot: M(x,y) Analysis D:\SYMPOSIUM _ARTICLE\A_RNA...

Zoom: 100

☐ Force point size 4 Pixels

X: 165
Y: 152
Data: 1.2109

Low threshold: 0.0000

High threshold: 1.5525

Data Examiner

● Color ○ Grayscale

X,Y Position: 165, 152

Value: 1.2109

Pixel (x,y): 205, 192

Figure 4. Mutual information plot of the type A bacterial RNase P RNA alignment showing the positions of the 18 helices found in the E. coli structure (Fig. 5). Helix labels have been added only to illustrate the structure compared to Figre 5. BioEdit does not add helix labels

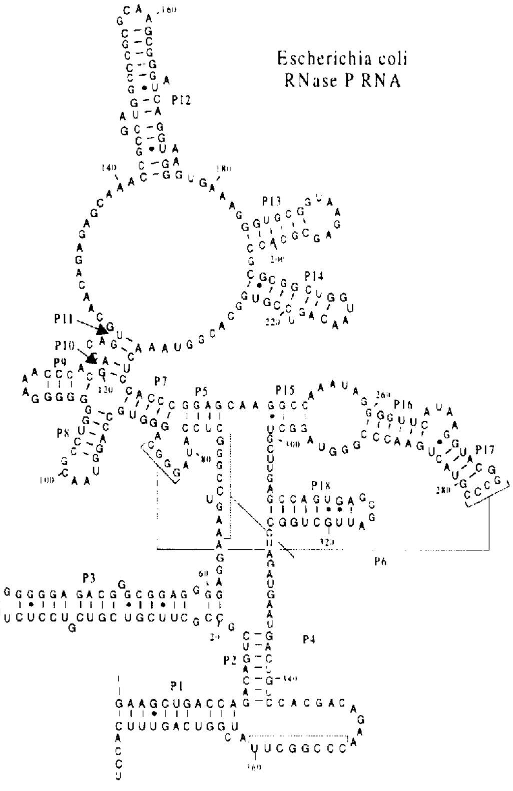


Figure 5. Secondary structure of *E. coli* RNase P RNA. Compare helix locations to the labeled strings of high mutual information in Figure 4.

examiner may be used to view the data interactively with the mouse (Fig. 4). Also, a point may be selected on the plot with the mouse, and the X, Y position and data value are displayed on the control bar. The threshold scrollers may be used to set low or high thresholds. Any point below the low threshold is shaded as "0.0000" (background), while any point above the high threshold is shaded light blue. The zoom function allows zooming in from 25% to 800% (max depends on original size, actually). Figure 6 shows a zoomed view with the low threshold set to bring out helices P8, P9, P12, P13 and P14. The plotted image may be saved to disk or copied to the clipboard as a bitmap.

Individual rows of matrix data may be quickly scanned for high scoring pairs by area plots (Fig 7). A position may be selected on the matrix, and the area plot will reflect the currently selected row. Likewise, when the rows are scanned via area plots, the current selection on the matrix is updated accordingly. The data for any pair may be viewed by clicking the mouse over the corresponding peak on the plot. Figure 7 shows a base triple interaction where nucleotide 316 from L18 interacts with base-pair 94/104 of P8 (8, 9). The area plot may also be zoomed.

A detailed summary of information may be obtained for any pair from the alignment window using the "mutual information examiner" (Fig. 8, concept stolen directly from J.W. Brown, 4). The examiner may also also be directed to reflect the positional numbering of a mask sequence. A printable, copyable text summary may be obtained from this window as well.

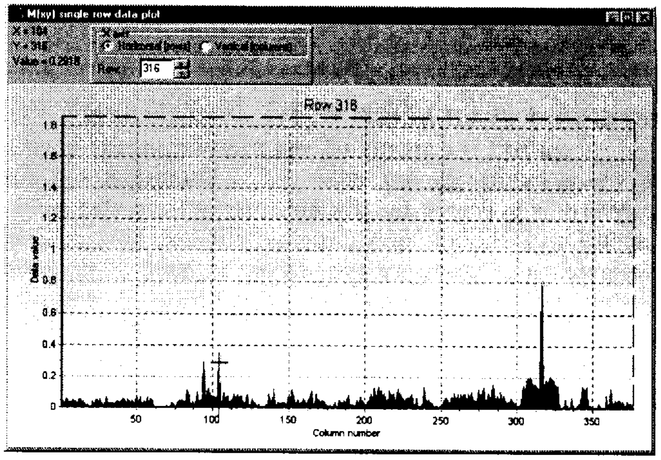
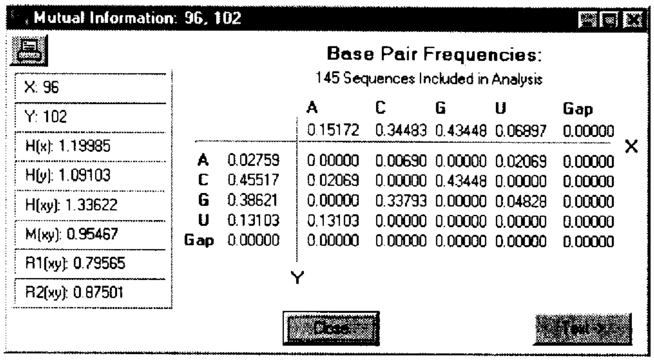


Figure 7. An area plot of matrix row 316 of the M(x,y) matrix. A base triple interaction can be seen between nucleotide 316 and base pair 93/105. This base triple has previously been proven by an in-depth comparative analysis (8) and by intramolecular chemical cross-linking (9).



X: 96
Y: 102
H(x): 1.19985
H(y): 1.09103
H(xy): 1.33622
M(xy): 0.95467
R1(xy): 0.79565
R2(xy): 0.87501

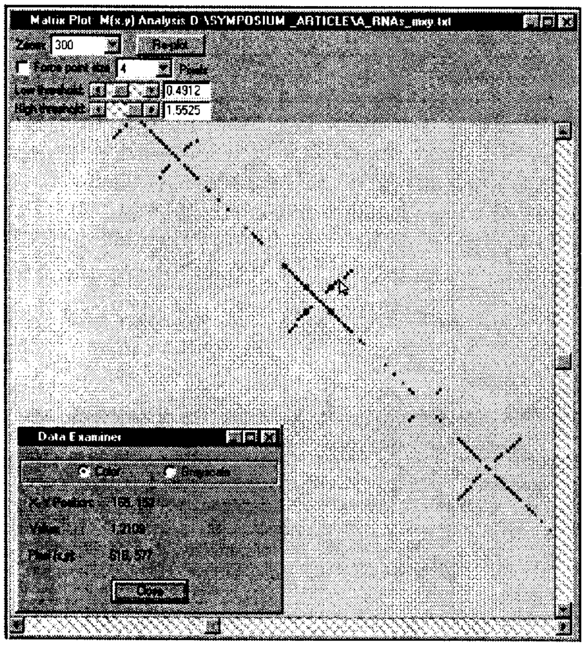
Base Pair Frequencies:
145 Sequences Included in Analysis

		A	C	G	U	Gap
		0.15172	0.34483	0.43448	0.06897	0.00000
A	0.02759	0.00000	0.00690	0.00000	0.02069	0.00000
C	0.45517	0.02069	0.00000	0.43448	0.00000	0.00000
G	0.38621	0.00000	0.33793	0.00000	0.04828	0.00000
U	0.13103	0.13103	0.00000	0.00000	0.00000	0.00000
Gap	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Figure 8. Summary of data for positional pair 96, 102. Pressing the button labeled "Text->" causes a formatted text summary to be displayed in the text editor.

GENERAL DESCRIPTION OF THE PROGRAM

In addition to simple comparative analysis tools, BioEdit offers a diversity of simple sequence manipulation and analysis tools for both nucleic acids and proteins, and for both alignments and single sequences. Sequence alignments may be shaded according to user preferences and with a choice of similarity tables to produce print-quality metafile figures suitable for pasting directly into a manuscript or onto a slide background for presentation. BioEdit reads and writes several formats, retains GenBank field data, and offers its own binary file format for fast read and write of very large files (the 6205-sequence prokaryotic 16S rRNA alignment will open and save in only a few seconds once converted to BioEdit format). An HTML 2.0 compliant Web browser is built-in, but BioEdit also auto-links to your favorite browser with its own bookmarks. Included NCBI BLAST tools allow for local database creation and searching, or searching GenBank via a BLAST 2.0 client or the WWW. For more information, see http://www.mbio.ncsu.edu/RNaseP/info/programs/BIOEDIT/bioedit.html.


Matrix Plot M(x,y) Analysis D:\SYMPOSIUM _ARTICLE\A_RNAs_mxy.txt
Data Examiner
Close

Zoom: 300
□ Force point size 4 Pixels
Low threshold 0.4912
High threshold 1.5525

Figure 6. View of the matrix plotter zoomed to 300% and with the low threshold set to 0.4912. The same base pair (512, 165) is being viewed as in Figure 4.

SYSTEM REQUIREMENTS

BioEdit runs on Windows 95/98/NT systems (or on a very fast Macintosh running Virtual PC, but then a little slowly). Any i486+ with Win95+ should run BioEdit, but a Pentium 166+ (or equivalent) with 32 Mb+ of memory is recommended. The full install takes about 13.9 Mb of disk space, but a smaller version (no local BLAST) is available which takes about 8.4 Mb.

PROGRAM AVAILABILITY

BioEdit is currently offered free of charge to anyone anywhere by anonymous download. Installation requires agreement with license and disclaimer terms (presented before installation). BioEdit fully installs itself and no confusing configuration or accessory systems are required of the user. BioEdit may be obtained from the RNase P database: http://www.mbio.ncsu.edu/RNaseP/info/programs/BIOEDIT/bioedit.html. Documentation (in Adobe Acrobat format) may be obtained from the same location.

ACKNOWLEDGEMENTS

I would like to express gratitude first to J.W. Brown, my mentor, for suggestions, patience (for putting up with the frequent distractions from my graduate work) and space on the RNase P database server. I would also like to thank Royden Saah, Andy Andrews, Jim Pannucci, Jay Andrus, Dave Dotson, Laurie Hall, Danny Williams, Roderic Page, Carla Kuiken, Fracois Michel, Francis Durst, Piotr Zimniak, Jens Kriese, Wayne Bolton, Alexey Troitsky, Peter Stogios, Jeroen Raes, Balazs Harrach, John Pawlowski and Brian Gaschen for useful suggestions or reports of bugs/problems.

REFERENCES

1. Thompson, J.D., Higgins, D.G and Gibson, T.J. (1994) *Nucleic Acids Res.*, 22: 4673-4680.

2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* **215**: 403-410.

3. Gutell R.R., Weiser, B., Woese, C.R., Noller, H.F. (1985) *Prog. Nucleic Acid Res. Mol. Biol.* **32**: 155-216.

4. Brown, J.W. (1991) *Comput. Appl. Biosci.* **7**: 391-393.

5. Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J. and Stormo, G.D. (1992) *Nucleic Acids Res.* **20**: 5785-5795.

6. Pace, N.R. and Brown, J.W. (1995) *J. Bacteriol.* **177**: 1919-1928.

7. Brown, J.W. (1999) *Nucleic Acids Res.* **27**: 314.

8. Brown, J.W., Nolan, J.M., Haas, E.S., Rubio, M.A.T., Major, F. and Pace, N.R. (1996) *Proc. Natl. Acad. Sci. USA* **93**: 3001-3006.

9. Harris, M.E., Kazntsev, A.V., Chen, J.-L. and Pace, N.R. (1997) *RNA* **3**: 561-576.