Published online 1 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W85–W90***
doi:10.1093/nar/gkn220

NOBAI: a web server for character coding of geometrical and statistical features in RNA structure

Vegeir Knudsen[1,2] and Gustavo Caetano-Anollés[1,*]

[1]Department of Crop Sciences, University of Illinois at Urbana-Champaign, IL 61801, USA and
[2]Center for Information Technology Services, University of Oslo, 0316 Oslo, Norway

Received January 14, 2008; Revised March 31, 2008; Accepted April 10, 2008

ABSTRACT

The Numeration of Objects in Biology: Alignment Inferences (NOBAI) web server provides a web interface to the applications in the NOBAI software package. This software codes topological and thermodynamic information related to the secondary structure of RNA molecules as multi-state phylogenetic characters, builds character matrices directly in NEXUS format and provides sequence randomization options. The web server is an effective tool that facilitates the search for evolutionary history embedded in the structure of functional RNA molecules. The NOBAI web server is accessible at 'http://www.manet.uiuc.edu/nobai/nobai.php'. This web site is free and open to all users and there is no login requirement.

INTRODUCTION

Functional RNA molecules are highly diverse and they are mainly structured by a set of short A-form helices that are typically about 10 base pairs in length. These helices define elements of secondary structure which are then arranged in space through tertiary contacts and delimit motifs identifiable at the sequence or structure levels (1,2). Since these elements define the overall fold of the molecule, evolutionary history can be studied directly from secondary structure (3–6). Evolutionary studies have used both geometrical and statistical features (characters) describing topological and thermodynamic attributes of RNA molecules to infer rooted phylogenetic trees (branching histories of inheritance) (5–11). Such trees reveal evolutionary patterns that are generally overlooked by traditional phylogenetic methods that focus on sequence. Geometrical and statistical characters gave trees that were congruent and similarly rooted, and generally matched trees reconstructed from sequences. The approach is robust and it has been applied to a wide variety of molecules at different taxonomical levels, from the subspecies/species levels to the universal tree. For example, patterns in spacer rRNA structures matched diversification patterns of phytopathogenic fungi following continental pathogen introduction (5) or habitat adaptation (6), and resolved intrageneric relationships in Caribbean gorgonian corals (12). At higher taxonomical levels, analysis of rRNA structure helped resolve rapid radiations in cladoceran orders of arthropodes (13) and deep phylogenetic relationships in the grasses (7). Trees of life were even reconstructed from the structure of several molecules, including the small and large subunits of rRNA (6,9) and tRNA (11). Finally, structural evolution was traced in ribosomes (9) and the origins and evolution of eukaryotic retrotransposable elements (8) and tRNA (10) was established.

In order to facilitate phylogenetic analysis, we developed NOBAI (Numeration of Objects in Biology: Alignment Inferences), a software package that automates the task of coding information in the secondary structure of folded RNA sequences. NOBAI consists of four separate modules, which are written in either C or C++: MARTEN (Molecular Analysis and Recording Tool for Evolutionary Numeration), STOAT (Statistics Of Architectural Topology), STRINGGEN and OMROKGEN. MARTEN takes as input sequence and structural information in DCSE format (14), codes geometrical features in paired and unpaired regions of the molecules as linearly ordered multistate characters and produces a file in NEXUS format (15) for analysis with standard phylogenetic software. STOAT takes as input nucleotide sequences in FASTA format, folds the molecules (with or without constraints) using routines in RNAfold (16,17), calculates four normalized morphospace parameters that describe molecular mechanic properties of the molecules and produces a file in NEXUS format. Finally, STRINGGEN and OMROKGEN are sequence randomization tools that read sequence strings and either generate all possible recombinations or rearrange sequences by single nucleotide permutation. All four applications in the NOBAI software package are

*To whom correspondence should be addressed. Tel: 217 333 8172; Fax: 217 333 8046; Email: gca@uiuc.edu

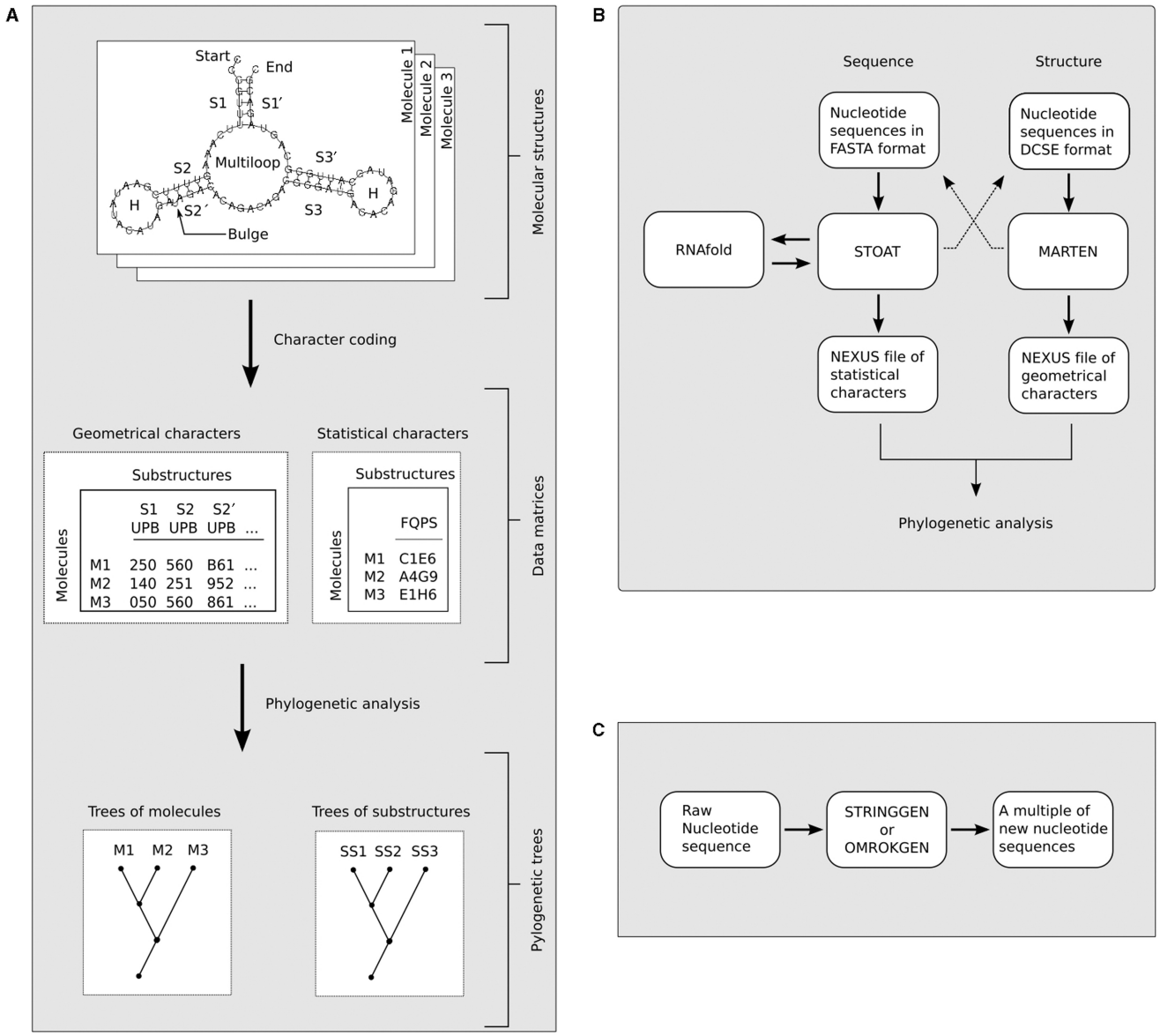

Figure 1. An illustration of the NOBAI software package. (**A**) A sketch of how the secondary structure of RNA molecules can be used to reconstruct the evolutionary history of organisms. H, hairpin; S_x, stem number x; SS_x, substructure number x; M_x, molecule number x; F, the Frobenius norm; Q, Shannon entropy; P, base-pairing propensity and S, mean stem length. (**B**) A sketch of the processing performed by MARTEN and STOAT (the two main programs in the NOBAI package). (**C**) A sketch of the processing performed by STRINGGEN and OMROKGEN.

UNIX command line based and they do not have a graphical user interface (GUI).

This paper presents the NOBAI web server. The server is an Apache installation that provides a web interface to the four applications in NOBAI. The communication between the applications, the web server and the user (such as the uploading of an input file and the presentation of the computational results) is done by Perl-CGI scripts. Sample input data for each application is hyperlinked to the application interfaces. The execution time for these samples (when assuming no load on the server) should be less than 10 seconds. The computational results are made available for download from a temporal web page as both individual files and a compressed tar-file.

METHODS

Figure 1A illustrates how the secondary structure of functional RNA molecules can be used to reconstruct evolutionary history in the form of phylogenetic trees. This method requires that we know the geometrical shape of the molecules, either by inference using alignment and/or folding algorithms or experimentally through crystallographic or other methods. The actual secondary structure of a specific RNA molecule can be found in databases [e.g. the EUROPEAN RIBOSOMAL RNA DATABASE (18) or RFAM (19)] or can be obtained using predictive folding software [e.g. the Vienna package (17)].

Geometrical characters describe individual molecular components (substructures, SS), or the entire molecule.

Table 1. Statistical characters of folded RNA sequences used by STOAT. Terms used in the equations: $Q_m = 0.5N\log_2 N$; P_{ij}, the base paring probability; A_k, the length of stem number k (i.e. the number of paired nucleotides in that stem); C, the number of paired bases; N, the number of bases; and B, the number of stems

Shannon entropy	$Q = -\left(\frac{1}{Q_m}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P_{ij} log_2 P_{ij}$
Frobenius norm	$F = \left(\frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(P_{ij})^2\right)^{1/2}$
Base-pairing propensity	$P = \frac{C}{N}$
Mean stem length	$S = \frac{A}{B} = \frac{1}{B}\sum_{k=1}^{B} A_k$

For example, the sample molecule in Figure 1A can be divided into six SS following the DCSE format, one substructure for each stem strand. These SS are further associated with three parameters: the number of unpaired nucleotides (U), the number of paired nucleotides (P) and the number of bulged nucleotides (B). For our sample molecule the first stem strand (S1) has 2 unpaired nucleotides, 5 paired nucleotides and 0 nucleotides in bulges or internal loops. These three numbers can then be used in a geometrical character data matrix to represent S1, as shown in Figure 1A.

Statistical characters define a morphospace that provides global descriptions of elements of secondary structure of folded nucleotide sequences. These morphospace characters include the Shannon entropy of the base-pairing probability matrix (Q), the Frobenius norm (F), the base-pairing propensity (P) and the mean stem length (S) (20,21). Their phylogenetic significance has been previously discussed (7). Table 1 provides equations defining these four statistical parameters. As illustrated in Figure 1A, integers can be assigned to each parameter in a statistical character data matrix.

The generated character data matrices contain characters describing the RNA molecules. Figure 1A shows how data matrices display character states and are used to generate phylogenetic trees. Several software packages, including PAUP* (22), take these matrices in NEXUS format as input for phylogenetic analysis. If the character data matrices are flipped (i.e. rows become columns and vice versa), then the data matrices can be used to create phylogenetic trees of SS in addition to the traditional phylogenetic trees of molecules. These trees of SS are very useful and have been used to define origins and evolution of RNA structure (8–10).

WEB SERVER

The NOBAI web server provides a web interface to the applications in the NOBAI software package. The web server is available at 'http://www.manet.uiuc.edu/nobai/nobai.php', while source code of the applications can be obtained directly from the authors. The service is free and open to all users and there is no login requirement. The web server has one separate web interface for each of the four applications. These web interfaces also contain links to sample input data and a documentation page. The documentation pages are small manuals that give a brief description of both the application and the input parameters. These manuals were originally written with UNIX man-page macros and thereafter converted to *html* by GROFF (23).

The results from the computations performed on the web server are made available for download on a temporary web page after the calculation has ended. This web page presents the computational results as both individual files and a *tar* file (24). The *tar* file is compressed with GZIP (25) and it contains all the result files plus the user provided input. The *tar* file will unpack all the files in the same directory as it is stored. All files related to a specific computation are stored in a separate directory on the server for at least 24 hours. However, the user needs to recall the exact web address to a file in order to download that file after the temporary web page has been lost.

Figure 1 (B and C) gives an overview of the input, output and processing of the four applications in the NOBAI software package. The two main programs are MARTEN and STOAT, which code phylogenetic characters associated with RNA secondary structures. OMROKGEN and STRINGGEN on the other hand, are tools that generate new nucleotide strings in which the positions of the nucleotides in the original sequence are rearranged. All four applications print information to standard out during the computation. On the web server, this information is saved in the stdout.txt file. A brief description of each of the four web interfaces is given below.

MARTEN

MARTEN accepts only input files in the DCSE format. The main output from MARTEN is a geometrical character data matrix written to a NEXUS file. The suffix of the NEXUS formatted file is 'nex'. MARTEN also translates the input sequences from the DCSE format to the FASTA format. The suffix of the FASTA file is 'fas'. The default and maximum number of character states are 64. However, the user can change the number of character states. A capture of the MARTEN web interface and result page are shown in Figure 2.

Note that in certain occasions DCSE files downloaded from the EUROPEAN RIBOSOMAL RNA DATABASE may contain sequences for which the 'helix numbers' are not located directly underneath the stem strands. If the number of stems strands for those sequences are less than the number of 'helix numbers', then MARTEN may fail to assign the stem strands to a 'helix number.' When this situation occurs, MARTEN excludes the sequence from the computation. The user can, however, instruct MARTEN to search for 'helix numbers' outside the stem strands, but it is often better to edit the input file manually instead, if the calculation fails.

STOAT

STOAT reads an input file in the FASTA format and invokes the RNAfold program (version 1.4) from the

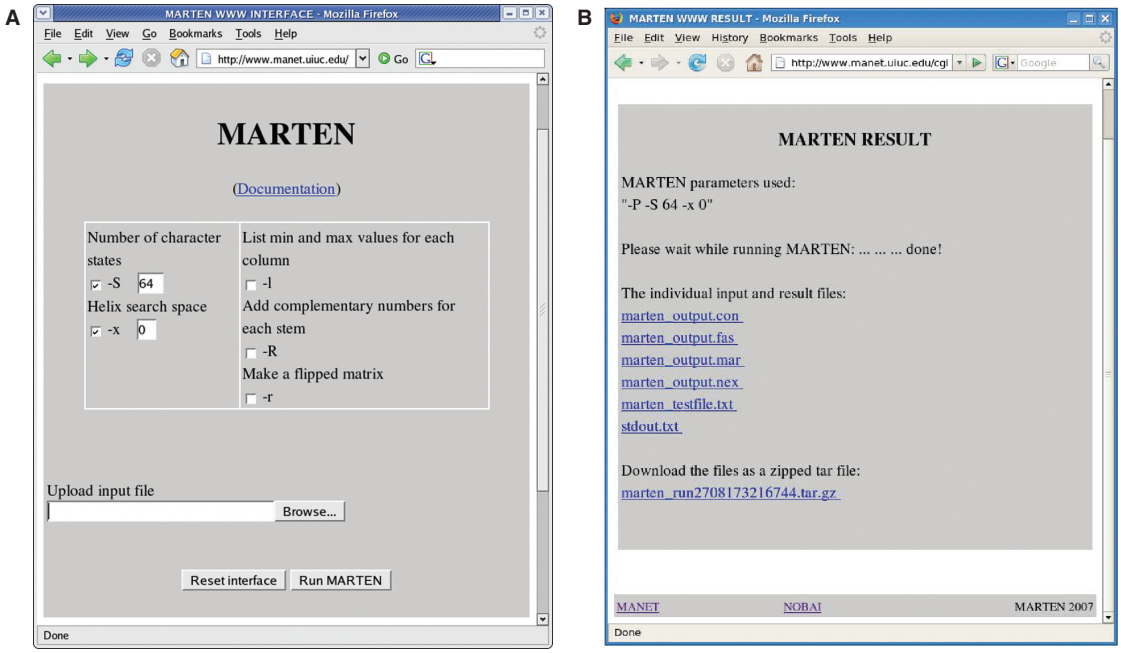

Figure 2. Two examples of web pages of the NOBAI web server. (**A**) Capture of the MARTEN web interface with the default parameters. A brief description of the input parameters can be found under the 'Documentation' hyperlink. (**B**) Capture of a temporary result page generated by the MARTEN web interface. The result files can either be downloaded as separate files or as a compressed tar file.

Vienna package (17). RNAfold returns the minimum free energy (MFE) structure and the probability of base-pairing between bases i and j in the sequence (p_{ij}). The base-pairing probability stems from the partition function (26) and is needed to calculate the Shannon entropy and the Frobenius norm. Parameters such as the stem lengths (A_k) and the number of stems (B) are calculated from the secondary structure of the folded nucleotides. Note that STOAT divides stems containing bulges or internal loops into more than one substructure (e.g. stem 2 of the sample molecule in Figure 1A is divided into two SS).

STOAT outputs two statistical character data matrices, each written to a NEXUS file. These two files have the suffixes '_l.nex' and '_g.nex'. The local '_l.nex' file gives the character states relative to a linear scale based on the minimum and maximum values for that run. The global '_g.nex' file gives the character states relative to 0 and 1. The statistical characters used in the matrices are the Shannon entropy, the Frobenius norm, the base-pairing propensity and the mean stem length (Table 1). In the '_g.nex'-file, the mean stem length is given as S^{-1}. The default number of character states is 31 and the maximum number is 64. Figure 3 shows a local NEXUS file generated by STOAT and a consensus phylogenetic tree generated from it using PAUP*.

STOAT also translates sequences from FASTA to DCSE format. This translator does not align the sequences. Consequently, the DCSE format of the sequences is written to separate files.

STRINGGEN and OMROKGEN

STRINGGEN creates all possible recombinations of a given nucleic acid sequence. The application reads a nucleotide sequence and applies a loop to select all possible sequences of that same length and base composition. Because of computational limitations, the maximum sequence length for the web interface is 15 nucleotides.

OMROKGEN also rearranges sequences. However, the rearrangement is performed by a permutation procedure described in refs (20,27). This procedure consists of three perfect shuffles, each swapping nucleotides sequentially at all sites with a randomly chosen site elsewhere in the sequence. On the OMROKGEN web interface the user can specify the number of sequences to be generated.

These randomization tools shuffle sequences of any defined nucleotide composition and can be used to dissect the effects of composition and order of nucleotides in the stability of folded nucleic acids molecules (28). However, they may not be suitable for applications that require dinucleotide shuffling.

CONCLUSIONS

A software package has been developed to code geometrical and statistical phylogenetic characters from the secondary structure of folded RNA sequences. The applications in this software package have been made freely accessible on a web server open for all users. This software package with its web interfaces will facilitate the search for evolutionary patterns embedded in the structure of functional RNA.

ACKNOWLEDGEMENTS

We thank Feng-Jie Sun and Ajith Harish for constructive suggestions regarding the features of MARTEN and

A

```
#NEXUS

[!Data from : STOAT 2007]

BEGIN DATA;
DIMENSIONS  NTAX=91 NCHAR=4;
FORMAT SYMBOLS = "0~9A~U";
CHARSTATELABELS
          1 F,
          2 Q,
          3 P,
          4 S;

MATRIX

Sulfurococcus_mirabilis_1A     N8M8
Sulfolobus_tokodaii_1A         U0UN
Desulfurococcus_mobilis_1A     U1UO
Pyrobaculum_aerophilum_1A      Q4NB
Sulfolobus_acidocaldarius_3A   T1QL
Sulfolobus_acidocaldarius_2A   U0UN
Pyrodictium_occultum_1A        J6TO
Sulfolobus_solfataricus_2A     T1QL
Sulfolobus_sp._1A              T1QL
Sulfolobus_sp._2A              R3NB
Methanococcus_voltae_1A        I9CA
Methanococcus_voltae_2A        FE77
Haloarcula_vallismortis_1A     O4HD
Archaeoglobus_fulgidus_1A      P4HL
Haloarcula_japonica_1A         J94B
Haloarcula_marismortui_2A      O3FC
Haloarcula_marismortui_1A      M6FC
Haloarcula_marismortui_3A      J7FC
Halobacterium_salinarum_1A     HEFI
Halobacterium_salinarum_3A     L7JF
Halococcus_morrhuae_1A         HA4D
Halococcus_morrhuae_2A         HA4D
Haloferax_volcanii_1A          4UA0
Haloferax_volcanii_2A          4UA0
Haloferax_mediterranei_1A      8UF0
Haloferax_mediterranei_2A      7SF0
Halorubrum_distributum_2A      JA9E
Halorubrum_distributum_3A      M6DB
Methanobact_formicicum_1A      P5OE
Methanohalob_evestigatum_1A    O3C8
Methanolobus_tindarius_1A      O3FD
Methanopyrus_kandleri_1A       P5FU
Methanosar_acetivorans_1A      HC6A
........
```

B

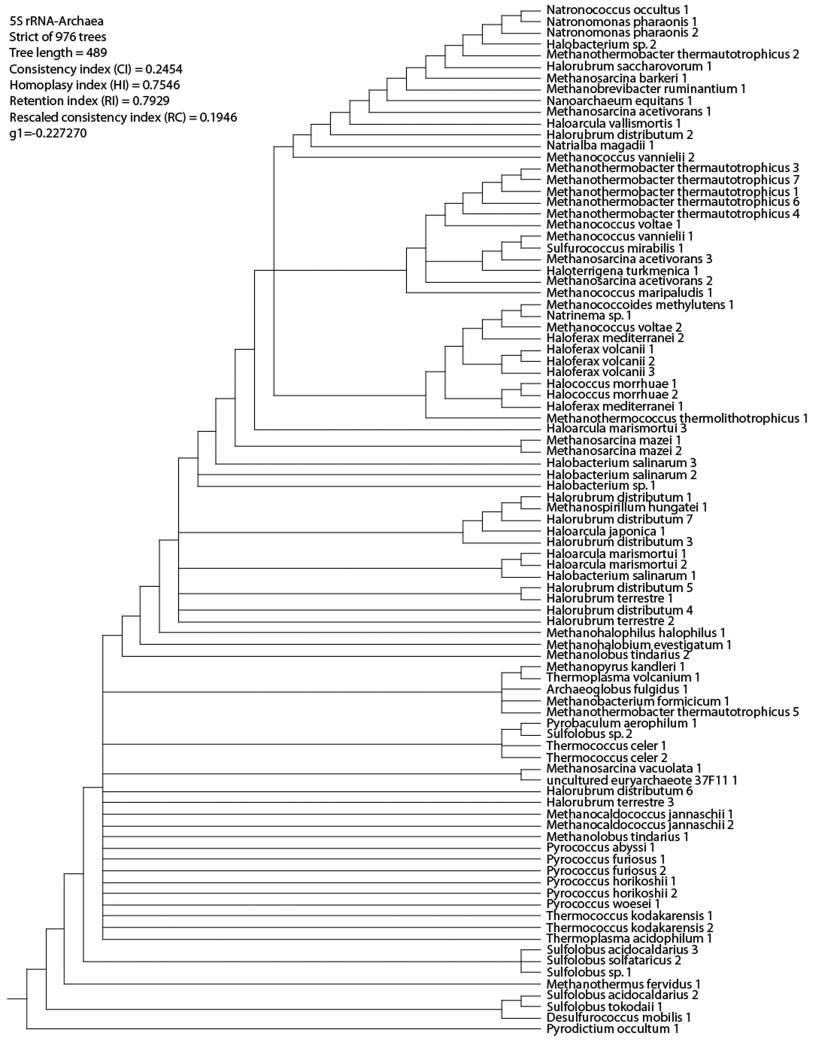

Figure 3. Character data matrix produced by STOAT in NEXUS format (_1.nex file) from 5S rRNA archaeal structures (**A**) and strict consensus phylogenetic tree reconstructed using PAUP* using maximum parsimony as the optimality criterion. (**B**) Note that only a top segment of the NEXUS file is shown and that the 5S rRNA sequences are provided as sample input data in the web server.

STOAT. A preliminary version of the NOBAI web server with limited functionality has been running since year 1999 hosted by the University of Oslo. The server was revamped completely this past summer, the Norwegian web server retired and the service transferred to the University of Illinois and hosted by the MANET database (29). This study was funded by National Science Foundation [MCB-0343126 to G.C.] and Critical Research Initiative of the University of Illinois. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. Funding to pay the Open Access publication charges for this article was provided by Critical Research Initiative.

Conflict of interest statement. None declared.

REFERENCES

1. Hermann,T. and Patel,D.J. (1999) Stitching together RNA tertiary structures. *J. Mol. Biol.*, **294**, 829–849.
2. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
3. Billoud,B., Guerrucci,M.A., Masselot,M. and Deutsch,J.S. (2000) Cirripede phylogeny using a novel approach: molecular morphometrics. *Mol. Biol. Evol.*, **17**, 1435–1445.
4. Collins,L.J., Moulton,V. and Penny,D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, **51**, 194–2004.
5. Caetano-Anollés,G. (2001) Novel strategies to study the role of mutation and nucleic acid structure in evolution. *Plant Cell Tissue Organ Cult.*, **67**, 115–132.
6. Caetano-Anollés,G. (2002) Evolved RNA secondary structure and the routing of the universal tree of life. *J. Mol. Evol.*, **54**, 333–345.
7. Caetano-Anollés,G. (2005) Grass evolution inferred from chromosomal rearrangements and geometrical and statistical features in RNA structure. *J. Mol. Evol.*, **60**, 635–652.
8. Sun,F.-J., Fleudépine,S., Bousquet-Antonelli,C., Caetano-Anollés,G. and Deragon,J.M. (2007) Common evolutionary trends for SINE RNA structures. *Trends Genet.*, **23**, 26–33.
9. Caetano-Anollés,G. (2002) Tracing the evolution of RNA structure in ribosomes. *Nucleic Acids Res.*, **30**, 2517–2587.

10. Sun,F.-J. and Caetano-Anollés,G. (2008) The origin and evolution of tRNA inferred from phyogenetic analysis of structure. *J. Mol. Evol.*, **66**, 21–35.
11. Sun,F.-J. and Caetano-Anollés,G. (2008) Evolutionary patterns in the sequence and structure of transfer RNA: early origins of Archaea and viruses. *PLoS Comp. Biol.*, **4**, e10000018.
12. Grajales,A., Aguilar,C. and Sánchez,J.A. (2007) Phylogenetic reconstruction using secondary structures of Internal Transcribed Spacer 2 (ITS2, rDNA): finding the molecular and morphological gap in Caribbean gorgonian corals. *BMC Evol. Biol.*, **7**, 90.
13. Swain,T.D. and Taylor,D.J. (2003) Structural rRNA characters support monophyly of raptorial limbs and paraphyly of limb specialization in water fleas. *Proc. R. Soc. Lond. B*, **270**, 887–896.
14. De Rijk,P. and De Wachter,R. (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. *Comp. Applic. Biosci.*, **9**, 735–740.
15. Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) NEXUS: an extendable file format for systematic information. *Syst. Biol.*, **46**, 590–621.
16. Hofacker,I.L., Fontana,W., Bonhoeffer,S. and Standler,P.F. (1994) Fast folding and comparison of RNA secondary structure. *Monatshefte Chem.*, **125**, 167–188.
17. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
18. Wuyts,J., Perrière,G. and Van de Peer,Y. (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.
19. Griffith-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
20. Schultes,E.A., Hrabrer,P.T. and LaBean,T.H. (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76–83.
21. Gardner,P.P., Holland,B.R., Moulton,V., Hendy,M. and Penny,D. (2003) Optimal alphabets for an RNA world. *Proc. R. Soc. Lond. B*, **270**, 1177–1182.
22. Swofford,D.L., (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, Massachusetts.
23. GNU roff (groff) software; typesetting package which reads plain text mixed with formatting commands and produces formatted output. http://www.gnu.org/software/groff/ (21 April 2008, date last accessed)
24. GNU tar (tar) software; utility to create tar archives. http://www.gnu.org/software/tar/ (21 April 2008, date last accessed)
25. GNU zip (gzip) software; compression utility. http://www.gnu.org/software/gzip/ (21 April 2008, date last accessed)
26. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, **29**, 1105–1119.
27. Knuth,D. (1973) *The Art Of Computer Programming*, Vol. 3. Addison Wesley, Reading, MA, p. 237.
28. Forsdyke,D.R. (2007) Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. *J. Theor. Biol.*, **248**, 745–753.
29. Kim,H.S., Mittenthal,J.E. and Caetano-Anollés,G. (2006) MANET: tracing evolution of protein architecture in metabolic networks. *BMC Bioinformatics*, **7**, 351.