Nucleic Acids Research

CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments

Eckart Bindewald, Thomas D. Schneider and Bruce A. Shapiro

Nucleic Acids Res. 34:405-411, 2006.
doi:10.1093/nar/gkl269

Nucleic Acids Research, 2006, Vol. 34, Web Server issue **W405–W411**
doi:10.1093/nar/gkl269

CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments

Eckart Bindewald, Thomas D. Schneider[1] and Bruce A. Shapiro[1,*]

Basic Research Program, SAIC-Frederick, NCI-Frederick, Frederick, MD 21702, USA and [1]Center for Cancer Research Nanobiology Program, NCI-Frederick, Frederick, MD 21702, USA

Received February 14, 2006; Revised March 9, 2006; Accepted March 31, 2006

ABSTRACT

We present an online server that generates a 3D representation of properties of user-submitted RNA or DNA alignments. The visualized properties are information of single alignment columns, mutual information of two alignment positions as well as the position-specific fraction of gaps. The nucleotide composition of both single columns and column pairs is visualized with the help of color-coded 3D bars labeled with letters. The server generates both VRML and JVX output that can be viewed with a VRML viewer or the JavaView applet, respectively. We show that combining these different features of an alignment into one 3D representation is helpful in identifying correlations between bases and potential RNA and DNA base pairs. Significant known correlations between the tRNA 3′ anticodon cardinal nucleotide and the extended anticodon were observed, as were correlations within the amino acid acceptor stem and between the cardinal nucleotide and the acceptor stem. The online server can be accessed using the URL http://correlogo.abcc.ncifcrf.gov.

INTRODUCTION

A sequence logo is a way to visualize the sequence conservation or, more precisely, the information content of a region in a set of aligned sequences (1). Sequence logos have been used to identify DNA-protein binding sites (2,3), splice sites (4,5) as well as functionally important regions in proteins (6,7).

A typical sequence logo is constructed as follows. The composition of each alignment column of the region of interest is visualized by plotting the characters found at that position stacked on top of one another; the plotted height of a character is proportional to its respective frequency multiplied by the information content of that alignment column. The character types found at each respective position are plotted as one stack, sorted in increasing order of character frequency.

Because of their usefulness, several implementations and web resources are currently available to generate sequence logos. One implementation is the Delila package which can be used both for nucleotide and protein sequences (1). WebLogo is a web resource that can generate a sequence logo for a given sequence alignment (8). The user has numerous options to customize the sequence logo with WebLogo. The Biojava initiative (http://www.biojava.org) offers predefined Java classes that help in plotting sequence logos.

A regular sequence logo does not take correlations between different positions into account. The information theory based measure for correlations between positions is the mutual information. It is the difference between the information content of two combined alignment columns and the sum of the information content of those two alignment columns taken separately (4).

Correlations between positions have been computed by χ^2-statistical tests (9), but the information measured in bits has the advantage that it is additive and also that it can be compared between any two biological systems. Several programs have been described that display the mutual information between different positions in an alignment into consideration. Gorodkin *et al*. (10) extended the concept of a sequence logo to a structure logo. The visualization of a structure logo is based on a sequence logo, however, an additional symbol ('M') indicates mutual information between pairs of user-defined nucleotide positions. Other interesting tools are the monochromatic MatrixPlot program (11) and the color-producing Diana and Xyplo programs which are part of the Delila package (4) (http://www.ccrnp.ncifcrf.gov/~toms/delila.html). These programs compute for a given sequence alignment a matrix plot consisting of the mutual information values corresponding to all possible pairs of positions in an alignment. MatrixPlot also shows the information of single alignment positions (the height of a sequence logo) as columns at the border of the 2D matrix.

Workman *et al*. (12) developed the enoLOGOS web resource. The enoLOGOS server generates a monochromatic

*To whom correspondence should be addressed. Tel: +1 301 846 5536; Fax: +1 301 846 5598; Email: bshapiro@ncifcrf.gov

Published by Oxford University Press 2006

plot showing a sequence logo and optionally the 2D mutual information matrix. In addition to computing the information content of a sequence alignment, the program can also use a weight matrix as input. Different types of weight matrices are possible, for example, a matrix of interaction energies of different nucleotide types with a binding partner at different positions.

This paper describes a 3D representation of the properties of an RNA or DNA alignment. These 3D sequence logos provide valuable additional information compared to conventional 2D sequence logos. Also, there are currently few online servers in the area of bioinformatics that generate 3D graphics. The online resource presented in this paper could serve as an important example of how 3D content can be utilized as a powerful means to communicate biological information.

MATERIALS AND METHODS

Theory

The purpose of the presented visualization is to depict several different properties of a nucleic acid sequence alignment in one 3D model. Properties of interest are the mutual information of two alignment columns and the information content of individual alignment columns. In the following we describe how we compute the mutual information based on Claude Shannon's information theory (13,14).

For a given set of symbols belonging to an alphabet A_1 that consists of s_1 different symbols, the uncertainty H is defined as follows:

$$H = -\sum_{i=1}^{s_1} p_i \log_2 p_i \quad \text{[bits per symbol]}, \qquad \mathbf{1}$$

where p_i is the probability of a symbol of type i. In the case of sequence alignments, a symbol can, for example, represent a character in an alignment column. The information content R of some data (like a string of characters or a column in a sequence alignment) is defined as the difference in uncertainty before and after processing those data (called H_{before} and H_{after}, respectively):

$$R = H_{\text{before}} - H_{\text{after}} \quad \text{[bits per symbol]}. \qquad \mathbf{2}$$

The mutual information of a sequence alignment is defined as follows:

$$M_{ij} = R_{ij} - R_i - R_j \quad \text{[bits per symbol pair]}, \qquad \mathbf{3}$$

with R_i (or R_j) being the information content of the alignment column i (or j), and R_{ij} being the information content of both alignment columns read simultaneously using an alphabet A_2 with $s_2 = s_1^2$ different symbols (4). The symbols of the alphabet A_2 are all possible pairs of characters taken from A_1. For DNA and RNA one obtains $s_1 = 4$ and $s_2 = s_1^2 = 16$.

For probabilities p_i, we substitute the relative frequency of occurrence $\hat{p}_i = m_i/n$ of the symbols in the available data (with n being the number of sequences that do not have a gap at the alignment column(s) currently under consideration) and m_i being the number of symbols of type i. The approximation of using frequencies as probabilities leads to a bias in the uncertainty H due to small sample size (15,16). We use several different strategies in order to correct for the small sample bias, depending on the number of available sequences and the size of the alphabet. This is described in more detail in section A of the Supplementary Data. These methods allow us to compute an approximate small sample correction term and also an estimate of the standard deviation for the information. A similar strategy for computing the mutual information was used for a secondary structure prediction algorithm (17).

The 3D model

We developed a C++ program that generates a 3D model showing various properties of an alignment (Figures 1 and 3 and Supplementary Figures S2 and S3). Vertical bars, $b(i, j)$, are arranged in a square region and indicate by their height the mutual information of the corresponding pair of positions (i, j). The bar $b(i, j)$ is composed of a stack of smaller bars $b(C, i, j)$, with C being one of the 16 possible base pair*s*. The height of the bar $b(C, i, j)$ is proportional to the frequency of the base pair C multiplied by the mutual information at that pair of positions (i, j). Each bar in a stack is drawn in the order of increasing frequency of the corresponding base pair. The bases on the two axes can be projected onto the corresponding faces of the bar. Because the bar $b(i, j)$ is identical to the bar $b(j, i)$, the square matrix region is symmetric.

The bars $b(C, i,j)$ that stand for a certain type of base pair C at position (i, j) are color-coded according to a user-specified color scheme. Currently two color codes are available. The default color-coding assigns to all complementary base pairs (GC, AU and GU) the colors red, orange and yellow, respectively. Pairs of bases consisting of the same base type (AA, CC, GG or UU) are mapped to a shade of green (ranging from light green for AA to dark green for UU), the remaining pairs (AG, AC and CU) receive a shade of blue (dark, medium and light blue, respectively). The color is independent of the order of the nucleotides in the pair. The second available color scheme (called 'rainbow' in the user interface) colors the entire stack from blue to red based on the mutual information value. A transparent bar at the bottom of the stack represents the set of base pairs that have a frequency below 5%.

Additionally, the displays shown in Figures 1 and 3 show conventional 2D sequence logos on two sides of the square matrix region. The black lines above the logos indicate 2 bits. The color-coding is the same as the one typically used in 2D sequence logos (3) with the colors green, cyan, orange and red representing the bases A, C, G and T (or U), respectively. The fraction of gap characters within an alignment column is shown as brown bars on the remaining two sides of the square matrix. Above the brown bars is a black line whose height indicates a gap fraction of 100% and which also corresponds to 1 bit. The two axis bars (drawn in black by default) can be drawn in three alternating colors (red, green and blue) to keep track of potential reading frames.

A secondary structure can also be superimposed on the 3D model. This secondary structure will appear in the resulting 3D model as black cubes floating above the individual stacks showing the mutual information of two alignment positions (shown in Figure 1 and Supplementary Figure S2). This is useful for comparing a secondary structure model with the mutual information matrix. Finally, 1.0 SDs of the estimated

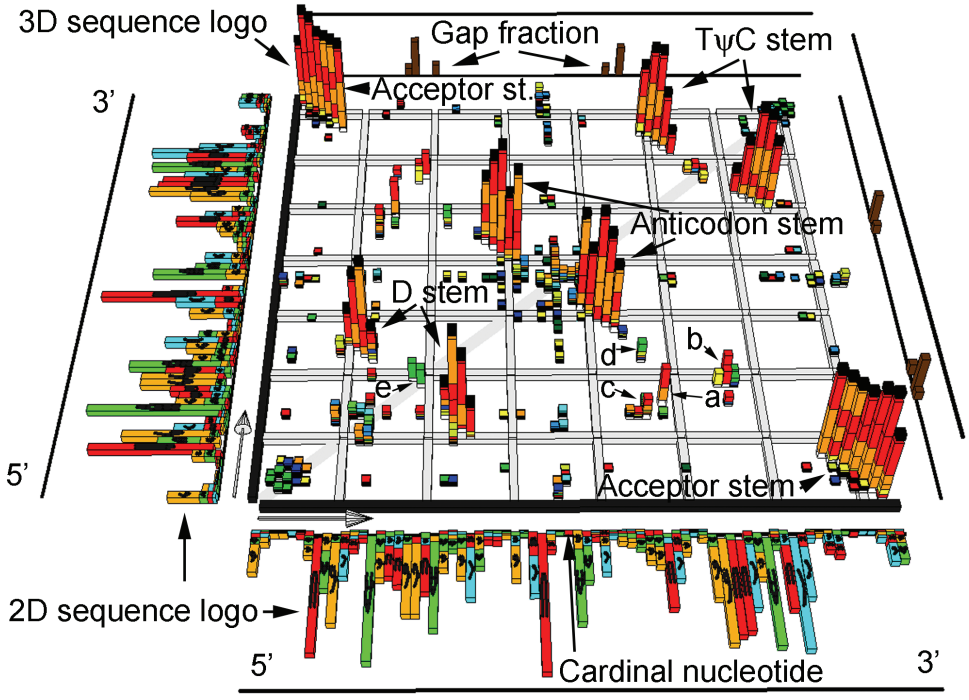

Figure 1. An annotated screenshot of the graphics output corresponding to an alignment of 1114 tRNA sequences showing 2D and 3D sequence logos. The 3D arrows point from the 5′ end to the 3′ end of the alignment. The labels 'a' to 'e' correspond to the base pairs G15:C48, G19:C56, C13:G46, G22:G46 and G18:G19, respectively. The black cubes correspond to base pairs in a reference consensus secondary structure. Only stacks corresponding to mutual information values $M_{ij} \geq 0.01$ bits and $M_{ij} \geq 2$ SDs are shown. The sequences correspond to the 'seed' alignment of RFAM entry RF00005, RFAM database version 7 (20,21). The consensus secondary structure provided by RFAM originated from (36).

mutual information can be visualized in the 3D model by the height of capped gray bars that are located above the 3D stacks of the individual position pairs (Figure 3).

The online server

An online server based on the Java Servlet technology was developed. The user specifies the sequence alignment in FASTA format. Optionally, a corresponding secondary structure can be specified in another text field.

Several options are currently available to the user. Two different color-coding schemes are offered. Three different cutoff options specify the displayed subset of the mutual information stacks. The only stacks drawn are those that correspond to a mutual information larger than the 'Lower Cutoff', smaller than the 'Upper Cutoff' and larger than a certain number of standard deviations above zero (specified by the 'Standard Deviation Cutoff'). This can be used to reduce the number of bars being drawn thus improving interpretability, increasing the rendering speed and avoiding memory limitations that are specific to the viewer program. It is also possible to 'collapse' an alignment with respect to one of its sequences: this means that all columns of the alignment that correspond to a gap in the chosen sequence are removed. The drawing of characters on the surface of the 2D sequence logos and the 3D sequence logo can be switched on or off. The 2D sequence logos can be drawn parallel ('flat'-mode) or perpendicular with respect to the square matrix region. Gray bars indicating 1 SD of the estimated mutual information can optionally be drawn (Figure 3).

After submitting the sequence data, the user receives a result page that offers three different output formats:

(i) JVX/JavaView: the server generates XML-style output intended for the JavaView applet (http://www.javaview.de) (18,19). The JavaView XML dialect is called JVX. The applet displays a 3D model within a web page. The user can modify the camera position and angle. The height of the drawn sequence characters is adjusted to be half the height of the bars they are projected onto. The figures in this paper that show 3D sequence logos are screenshots generated with JavaView.
(ii) VRML output: the generated 3D model is available in VRML 2.0 format. A click on this link will launch a VRML viewer. For technical reasons, the letters indicating the nucleotide compositions in the VRML representation are not 'stretched' proportionally as in the JavaView output shown in Figures 1 and 3.
(iii) A list in text format of the computed information values of the 2D and 3D sequence logos.

RESULTS

3D logo of tRNAs

As an example, we use an alignment of tRNA sequences provided by RFAM [1114 sequences of the RFAM 'seed' alignment RF00005 obtained in September 2005 (20,21)]. A screenshot from a 3D model generated with the CorreLogo

server is shown in Figure 1. Alignment columns that correspond to gaps in one of its sequences [yeast (*Saccharomyces cerevisiae*) Phe-tRNA, GenBank accession no. K01553.1] were removed prior to submission to the web server (alternatively one could have used the server's 'collapse' option). This makes it easier to compare the resulting data with a previously published secondary structure (22). The RFAM consensus secondary structure mapped onto this sequence is shown in Figure 2.

How does a Watson–Crick based paired helix appear in a 3D sequence logo? We define an anti-diagonal group of length L as a set of pairs of the form $[(i + k, j - k) \mid k = 0, 1, ..., L - 1]$. The pairs can be canonical or represent other kinds of correlations. Commonly, one expects pairs of positions in an alignment that have a high amount of mutual information and a high fraction of complementary bases to correspond to base pairs in a secondary structure. Also, one expects those columns to be aggregated in anti-diagonal groups, corresponding to stems of the secondary structure, that, because they are complementary, show up as 'tall' bars in yellow, orange and red. One can clearly identify in each triangular half of the matrix in Figure 1 four different anti-diagonal groups of red, orange and yellow bars with high mutual information, corresponding to the four stems of the tRNA cloverleaf structure.

The consensus secondary structure according to RFAM is depicted by a set of black cubes on top of the 3D sequence logo. This makes it feasible to compare a model of a secondary structure to the information contained in the sequence alignment. The anti-diagonal groups of bars with high mutual information are in agreement with the consensus secondary structure provided by RFAM.

In addition to the obvious stem regions, several other stacks appear in the 3D plot (Figure 1); the stacks with mutual information values >0.3 bits are labeled with the letters 'a' to 'e'. The labeled stacks correspond to the positions a, G15:C48; b, G19:C56; c, C13:G46; d, G22:G46 and e, G18:G19 (the base positions are determined with respect to the reference sequence shown in Figure 2). The interactions a–d correspond to well-known tertiary interactions; the interactions c and d are part of the base-triple C13:G22:G46 (22–25). The base pair d, G22:G46, corresponds to a large fraction of non-complementary base pairs, as can be seen in Figure 1 in the form of blue and green bars.

The mutual information of the tRNA base pairs c, C13:G22, and d, G22:G46, has been reported before by Chi and Kolodziejczak (26) based on an alignment of 131 tRNA sequences and by Gutell *et al*. (27) based on an alignment of 1710 sequences. Chi and Kolodziejczak (26) used a monochromatic circular representation to depict the mutual information of base pairs. Gutell *et al*. (27) showed the mutual information of the tRNA alignment using a contour plot as well as a surface plot. The 3D sequence logo, however, gives additional information; for each pair of alignment columns it shows the amount of mutual information, the relative frequencies of occurrence of different base pairs, the amount of complementarity, a confidence interval and base conservation in an associated 2D logo.

The tRNA anticodon stem/loop region has been analyzed in detail by Yarus (28). He found correlations between nt 36 (the 3′ residue of the 3 nt anticodon) and its adjacent nucleotides in

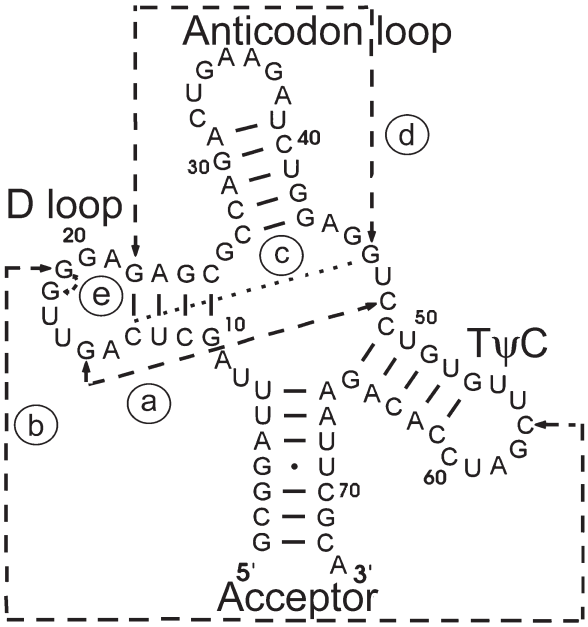

Figure 2. tRNA consensus secondary structure according to RFAM entry RF00005 (20,21,36) mapped onto the sequence of yeast (*S.cerevisiae*) Phe-tRNA (GenBank accession no. K01553.1). The correlations a, b, c, d and e (compare Figures 1 and 2) are indicated by dashed lines.

the anticodon loop and even the anticodon stem. Nt 36 is called the 'cardinal nucleotide' and its adjacent correlated region (positions 27–31, 37–43) is called the 'extended anticodon'. Some of these correlations were also reported in the form of mutual information by Gutell *et al*. (27). In their analysis they report the mutual information between the cardinal nucleotide and the top 10 most correlated nucleotide positions. This corresponds in their case to a somewhat arbitrary cutoff of 0.153 bits, and the correlations between the cardinal nucleotide and the positions 27–30, 40, 41, 42 and 43 were missed because they are not among the top 10 correlating positions. The CorreLogo server on the other hand provides a measure of significance by allowing the user to specify a relative cutoff, meaning that the mutual information has to be x SDs above zero. In Figure 1 and Supplementary Figures S2 and S3 only stacks with mutual information $M_{ij} \geq 0.01$ bits and $M_{ij} \geq 2$ SDs are shown. By using the computed standard deviation to filter the data, the CorroLogo server reveals the correlations of the extended anticodon proposed by Yarus (28) that were missed by Gutell *et al*. (27) (with the exception of the base pairing nt 30 and 40 which apparently have no significant mutual information and were detected neither by the CorreLogo server nor by Gutell *et al*.).

Given that the correlations with the cardinal nucleotide were found to be significant by using the estimation of standard deviation, other correlations of the same magnitude can be explored with CorreLogo. First, we note correlations between the anticodon, including the cardinal nucleotide, and the acceptor stem. These correlations are probably related to the recognition of the correct aminoacyl-tRNA synthetase by the tRNA (29–31). In addition, there are two 5 base wide patches of correlations on the 5′ and 3′ ends of the tRNAs that are also correlated through the acceptor stem. Just as Yarus proposed that the correlations between the

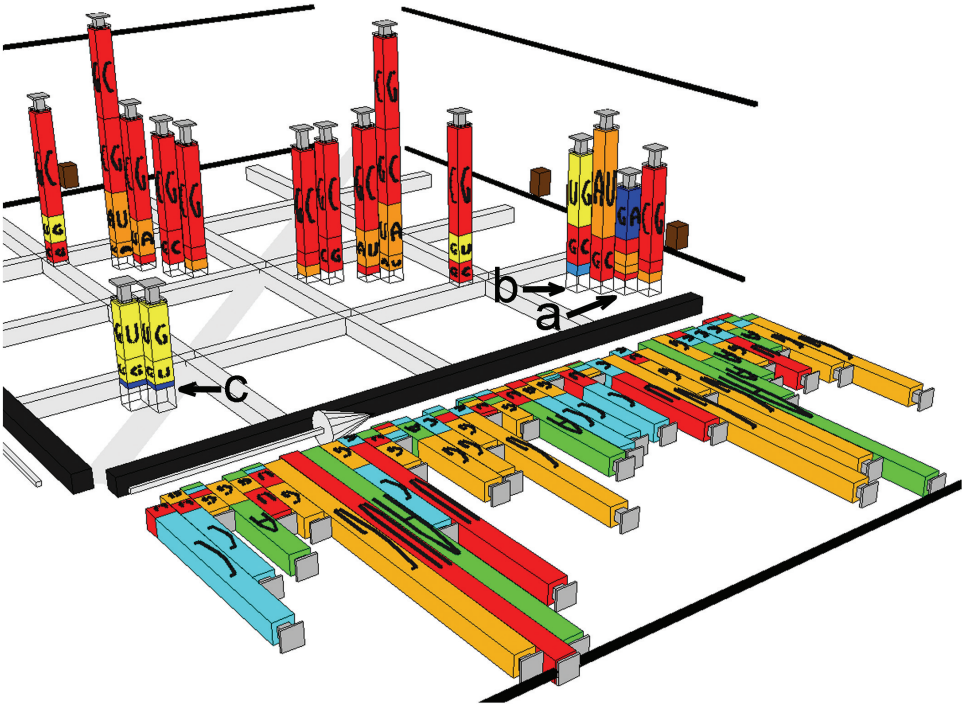

Figure 3. A 3D sequence logo of the loop E region of 5S ribosomal RNA [from RFAM entry RF00001 (20,21,35)]. Individual bars corresponding to base pairs are labeled in this view. The gray bars depict with their height the standard deviation of the estimated information in the 2D logos and the mutual information in the 3D logos. Stacks corresponding to mutual information values of <0.5 bits are not shown. The labels a–c correspond to pairs of positions that are not Watson–Crick base pairs and have a mutual information >0.5 bits.

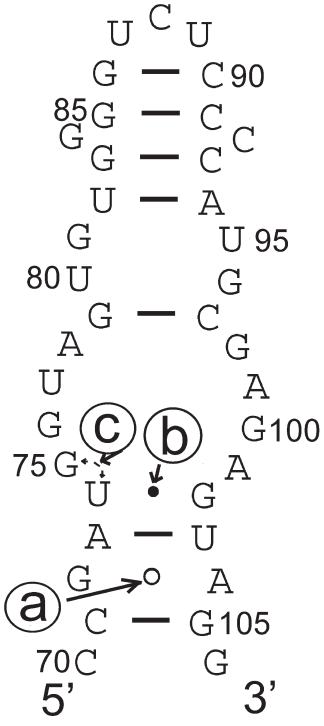

Figure 4. A consensus secondary structure of the loop E region of 5S ribosomal RNA derived from the 3D sequence logo is shown in Figure 3. The consensus structure is mapped onto the loop E region of *E.coli* 5S rRNA sequence (AB03926.1/6056-6092). The labels a–c correspond to pairs of positions that are not Watson–Crick base pairs and have a mutual information >0.5 bits.

cardinal nucleotide and the extended anticodon provide a uniform presentation of the anticodon to a fixed mRNA (28), these 5′–3′ correlations in the acceptor stem may provide a uniform presentation of the amino acid to the ribosome. An alternative hypothesis is that these correlations represent recognition of the tRNAs by their respective aminoacyl-tRNA synthetases. That is, if there were one distinctive 2D tRNA acceptor stem logo for each synthetase, switching logos would appear as correlations (29). Such classes of logos might also explain other observed correlations.

3D logo of loop E of 5S ribosomal RNA

We also generated a 3D sequence logo using an alignment of 5S ribosomal RNA. Crystal structures containing the loop E region have been published in recent years (32,33). Leontis and Westhof (34) compare the crystal structure data of the loop E region with results obtained previously using chemical probing and phylogenetic analysis. The high amount of non-standard base pairs and the wealth of available experimental data make the loop E structure an interesting test case for 3D sequence logos.

The alignment we used contains 602 sequences and corresponds to entry RF00001 of the RFAM database (20,21,32,35). The alignment was 'collapsed' with respect to gaps in a reference sequence (*Escherichia coli* sequence AB035926.1/5989-6104) prior to submission to the server. Also, the alignment was truncated such that it contains only the 5S loop E region [nt 70–106 of the reference sequence; the nucleotide numbering is in accordance with the bacterial loop E consensus sequence presented in Ref. (34)].

Figure 3 shows a screenshot that was taken from the resulting 3D sequence logo. In this case, correlations below 0.5 bits are not displayed. A secondary structure model corresponding to the displayed correlations is shown in Figure 4. The correspondence between the 3D sequence logo and the secondary structure derived from the crystal structure is good but not perfect. For example, base pair interactions in the loop region (nt 75–78 and 98–101) do not appear in the 3D sequence logo. Part of the reason is that the high amount of sequence conservation of residues 76–78, 98, 99 restricts the observable correlations as computed by mutual information. A 3D sequence logo with a lower cutoff ($M_{ij} \geq 0.01$ bits and $M_{ij} \geq 2.0$ SDs) is provided in the Supplementary Data (Supplementary Figure S3). Other significant correlations are shown in Supplementary Figure S3, but an in-depth discussion of all correlations is beyond the scope of this paper.

One can see in Figure 3 that, although they are part of an anti-diagonal group, the positions labeled 'a' and 'b' (corresponding to bp 72:104 and 74:102, respectively) have a high amount of non-Watson–Crick base pairs (GA and UG are dominant, respectively). This was known before; Leontis and Westhof (34), for example, plot 10 different secondary structures next to each other in order to show the variability of the loop E region. The point is that a single 3D sequence logo effectively summarizes the information that is conveyed when plotting secondary structures of several individual sequences. It highlights regions corresponding to unusual base pairings and displays the base pair preferences of individual positions. This information is not available in a monochromatic 2D plot of the mutual information.

These examples show that a combined plot depicting the mutual information of alignment positions, the color-coded fraction of base pairings, sequence conservation and the fraction of gaps can be valuable in identifying nucleotide base pair interactions. The 3D sequence logos provide a useful tool for analyzing nucleotide sequence alignments and secondary structure models.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank an anonymous reviewer for suggesting that we examine correlations of the tRNA anticodon region. We wish to thank the Advanced Biomedical Computing Center (ABCC) at the NCI for their computing support. This work has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. Funding to pay the Open Access publication charges for this article was provided by the NCI.

Conflict of interest statement. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

REFERENCES

1. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*., **18**, 6097–6100.
2. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol*., **284**, 241–254.
3. Hengen,P.N., Bartram,S.L., Stewart,L.E. and Schneider,T.D. (1997) Information analysis of Fis binding sites. *Nucleic Acids Res*., **25**, 4994–5002.
4. Stephens,R.M. and Schneider,T.D. (1992) Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol*., **228**, 1124–1136.
5. Rogan,P.K., Svojanovsky,S. and Leeder,J.S. (2003) Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. *Pharmacogenetics*, **13**, 207–218.
6. Galperin,M.Y., Nikolskaya,A.N. and Koonin,E.V. (2001) Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol. Lett*., **203**, 11–21.
7. Rigden,D.J., Jedrzejas,M.J. and Galperin,M.Y. (2003) An extracellular calcium-binding domain in bacteria with a distant relationship to EF-hands. *FEMS Microbiol. Lett*., **221**, 103–110.
8. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.
9. Olsen,G.J. (1983) *Comparative analysis of nucleotide sequence data*. Phd Thesis. University of Colorado Health Science Center, Denver, CO.
10. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci*., **13**, 583–586.
11. Gorodkin,J., Staerfeldt,H.H., Lund,O. and Brunak,S. (1999) MatrixPlot: visualizing sequence constraints. *Bioinformatics*, **15**, 769–770.
12. Workman,C.T., Yin,Y., Corcoran,D.L., Ideker,T., Stormo,G.D. and Benos,P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*., **33**, W389–W392.
13. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J*., **27**, 379–423, 623–656.
14. Pierce,J.R. (1980) *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications, NY.
15. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol*., **188**, 415–431.
16. Basharin,G.P. (1959) On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl*., **4**, 333–336.
17. Bindewald,E. and Shapiro,B.A. (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.
18. Polthier,K., Khadem,S., Preuß,E. and Reitebuch,U. (2002) Publication of interactive visualizations with JavaView. In Borwein,J., Morales,M.H., Polthier,K. and Rodrigues,J.F. (eds), *Multimedia Tools for Communicating Mathematics*. Springer, Heidelberg, pp. 314.
19. Polthier,K. and Majewski,M. (2004) Using MuPAD and JavaView to visualize mathematics on the internet. *Proceedings of the 9th Asian Technology Conference in Mathematics*, pp. 465–474.
20. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.
21. Griffiths-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.
22. Saenger,W. (1984) *Principles of Nucleic Acid Structure*. Springer–Verlag New York, NY.
23. Quigley,G.J. and Rich,A. (1976) Structural domains of transfer RNA molecules. *Science*, **194**, 796–806.
24. Gautheret,D., Damberger,S.H. and Gutell,R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol*., **248**, 27–43.
25. Gautheret,D. and Gutell,R. (1997) Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Res*., **25**, 1559–1564.
26. Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci*., **7**, 347–352.
27. Gutell,R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued

development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.
28. Yarus,M. (1982) Translational efficiency of transfer RNA's: uses of an extended anticodon. *Science*, **218**, 646–652.
29. Rodin,S., Rodin,A. and Ohno,S. (1996) The presence of codon–anticodon pairs in the acceptor stem of tRNAs. *Proc. Natl Acad. Sci. USA*, **93**, 4537–4542.
30. Hou,Y.M. and Schimmel,P. (1988) A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature*, **333**, 140–145.
31. Swairjo,M.A., Otero,F.J., Yang,X.L., Lovato,M.A., Skene,R.J., McRee,D.E., Ribas de Pouplana,L. and Schimmel,P. (2004) Alanyl-tRNA synthetase crystal structure and design for acceptor-stem recognition. *Mol Cell*, **13**, 829–841.
32. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science*, **292**, 883–896.
33. Correll,C.C., Freeborn,B., Moore,P.B. and Steitz,T.A. (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, **91**, 705–712.
34. Leontis,N.B. and Westhof,E. (1998) The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *RNA*, **4**, 1134–1153.
35. Szymanski,M., Barciszewska,M.Z., Erdmann,V.A. and Barciszewski,J. (2002) 5S ribosomal RNA database. *Nucleic Acids Res.*, **30**, 176–178.
36. Hou,Y.M. (1993) The tertiary structure of tRNA and the development of the genetic code. *Trends Biochem Sci.*, **18**, 362–364.