CABIOS

Vol. 13 no. 6 1997
Pages 583–586

Displaying the information contents of structural RNA alignments: the structure logos

J. Gorodkin, L.J. Heyer[1], S. Brunak and G.D. Stormo[2]

Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206, 2800 Lyngby, DK-2800, Denmark and [1]Department of Applied Mathematics and [2]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Received on March 5, 1997; revised on May 10, 1997; accepted on May 14, 1997

Abstract

***Motivation*:** *We extend the standard 'Sequence Logo' method of Schneider and Stevens (Nucleic Acids Res., **18**, 6097–6100, 1990) to incorporate prior frequencies on the bases, allow for gaps in the alignments, and indicate the mutual information of base-paired regions in RNA.*

***Results**: Given an alignment of RNA sequences with the base pairings indicated, the program will calculate the information at each position, including the mutual information of the base pairs, and display the results in a 'Structure Logo'. Alignments without base pairing can also be displayed in a 'Sequence Logo', but still allowing gaps and incorporating prior frequencies if desired.*

***Availability:** The code is available from, and an Internet server can be used to run the program at, http://www.cbs.dtu. dk/gorodkin/appl/slogo. html.*

***Contact:** E-mail: gorodkin@cbs.dtu.dk*

Introduction

The sequence logo by Schneider and Stephens (1990) is constructed by ranking the letters in terms of their frequency and having the height of the stack on each position being determined by the total amount of Shannon information (see, for example, Cover and Thomas, 1991). We use the same logo program for displaying, but change the computation of the sequence information as suggested in Hertz and Stormo (1995), so that prior frequencies of the bases are included in the information measure and gaps of alignments will also be displayed. Further, to include information governed by RNA structure, we include the mutual information between the base-paired regions. If base pairing occurs together with total sequence conservation, the mutual information is zero, and the logo for that region reduces to the original sequence logo. However, base pairing without sequence conservation means that there is mutual information and a signal is displayed. Further, the specific base pairs are displayed under the logo positions. We have written scripts which take a given alignment file and compute the total information and create a 'symvec' file required for the logo program. Another script modifies the created postscript file and labels the positions of the base-paired regions. These scripts can be freely downloaded on the Web page. Further, it is also possible to paste in the alignment, and a postscript file of the logo will be returned.

From sequence logos to structure logos

In the sequence logo (Schneider and Stephens, 1990), the logo is constructed by calculating the information of each position of the aligned sequences, and then displaying the letters by fractional size on top of each other. However, when calculating this, gaps in the alignment are ignored and the *a priori* base frequencies are not used. We include these by using an extended expression from Hertz and Stormo (1995). Let I_i be the information content of position i of the alignment, so that

$$I_i = \sum_{k \epsilon A} I_{ik} = \sum_{k \epsilon A} q_{ik} \log_2 \frac{q_{ik}}{p_k} \qquad (1)$$

where $A = \{A,C,G,U,-\}$ is the set of bases including gaps, and where q_{ik} is the fraction of 'base' k at position i. When $k \neq$ '$-$', we interpret p_k as the *a priori* distribution of the bases for that genome, or class of functional sequences, etc. We set $p_- = 1$, since $q_{i-} \log_2 q_{i-}$ is zero for q_{i-} equal to zero or one. For the work reported in this paper, we set $p_k = 0.25$, the probability of finding a base randomly. With this value of p_k, $I_i \leq 2$ bits. If I_{ik} is negative, we see fewer of base k at position i than expected, and vice versa if I_{ik} is positive. The terms I_{ik} complete the consensus sequence matrix as described by Hertz *et al.* (1990) and Hertz and Stormo (1995).

Having determined I_i for a position, there are several reasonable methods to set the height for each letter. One option the user can choose, called the 'type 1 logo', follows the method of Schneider and Stephens (1990). In this, the height of letter k at position i is proportional to its frequency:

$$d_{ik} = q_{ik} I_i \qquad (2)$$

© Oxford University Press

Another option, the 'type 2 logo', displays the heights in proportion to their frequencies relative to the expected frequencies:

$$d_{ik} = \frac{q_{ik}/p_k}{\sum_l q_{il}/p_l} I_i \qquad (3)$$

The two logos can be significantly different when the *a priori* base frequencies are quite different. This is important when describing binding motifs as we display what is present beyond what we would expect by chance. For both methods, when I_{ik} is negative, letter k will be displayed upside down to indicate less appearance than expected at random. Note that I_i can be negative if there are enough gaps at that position in the alignment; negative values are not displayed. Note also that displayed gaps (dashes) will always be displayed 'upside down'. The total information content of the sequence alignment becomes $I = \sum_i I_i$, and can be used to evaluate alignments and find common motifs in unaligned sequences (Hertz *et al.*, 1990; Hertz and Stormo, 1995).

We add the contribution from base pairing to the logo because binding sites on RNA are often composed of a combination of sequence and structure constraints. For example, two positions may each contain all four bases in equal amounts, and therefore have zero information content, but also be constrained so that they are always complementary because they base pair in an RNA structure. We include an indication of such constraints in the logo because they can contribute significantly to the accurate description of the binding motif. Mutual information is a convenient measure of the non-independence between two positions (Gutell *et al.*, 1992). The mutual information, M_{ij}, between two positions, i and j, is symmetrical, so we assign half of it to each position. That is, we define $M_i = M_j = M_{ij}/2$. Then the total information plotted at each position is the sum of the independent information content and the mutual information for that position:

$$J_i = I_i + M_i \qquad (4)$$

The M_i contribution is displayed with the letter M on top of the I_i contribution.

The mutual information is only calculated for pairs of positions that are indicated in the input file to be base paired and, because we are only interested in showing the extra information due to base pairs, we use a modified form of mutual information (although the standard form can be easily substituted). We define $\tilde{q}_{ij}$ to be the fraction of sequences that have complementary bases at positions i and j, which includes G-U base pairs. We call $E[\tilde{q}_{ij}]$ the expected value of $\tilde{q}_{ij}$, which is based on the composition of the two positions independently. That is, if the two positions were not base paired (and therefore correlated), we would expect to see that many complementary pairs by chance alone. In general, we have:

$$E[\tilde{q}_{ij}] = \sum_{(k,l) \in B} C_{kl}\, q_{ik}\, q_{jl} \qquad (5)$$

where $B = (A\backslash\{-\})^2$, which is the set of all pairs of bases, and C_{kl} is a symmetrical matrix with $C_{kl} = 1$ when k and l are complementary and zero otherwise. [Note that in general we could set the values of C_{kl} to indicate the 'degree' of complementarity between any two bases so that some were given more weight than others, such as G-U pairs (Gorodkin *et al.*, 1997a,b). We would then also have to use the weights in calculating $\tilde{q}_{ij}$] The definition of C_{kl} can also be altered to describe any type of correlation one might be interested in, and hence the mutual information defined below can 'measure' this feature.

The mutual information we determine is then the log-likelihood ratio of the observed to expected frequency of complementary bases, also known as the Kullback–Leibler distance between two distributions (Cover and Thomas, 1991):

$$M_{ij} = \tilde{q}_{ij} \log_2 \frac{\tilde{q}_{ij}}{E[\tilde{q}_{ij}]} + (1 - \tilde{q}_{ij}) \log_2 \frac{(1 - \tilde{q}_{ij})}{(1 - E[\tilde{q}_{ij}])} \qquad (6)$$

If all of the sequences in the alignment are complementary at positions i and j (as in the examples shown below), then:

$$M_{ij} = -\log_2 E[\tilde{q}_{ij}] \qquad (7)$$

The total information of the alignment becomes $J = I + M$, $M = \sum_i M_i$.

Some examples

In Figure 1, we show two examples of published SELEX (Tuerk and Gold, 1990) data for which we have plotted the type 2 structure logos.

Figure 1a and b shows an example from RNA sequences that were selected to bind to the R17 coat protein (Schneider *et al.*, 1992). We included only data which had stems of at least six base pairs. Figure 1a is the sequence logo alone, and Figure 1b contains the mutual information from the base pairs. Note that if the M symbols are ignored, the two plots are identical. However, the Ms indicate that there are significant constraints on the sequences that are not evident from the sequence logo alone. For instance, positions 2, 15 and 16 have nearly zero sequence information, but because the structure is conserved in all of the examples, there is considerable mutual information. For some paired positions, such as 6, 7, 12 and 13, there is complete conservation of structure as well as significant preference for particular bases. This shows up as some degree of sequence informa-

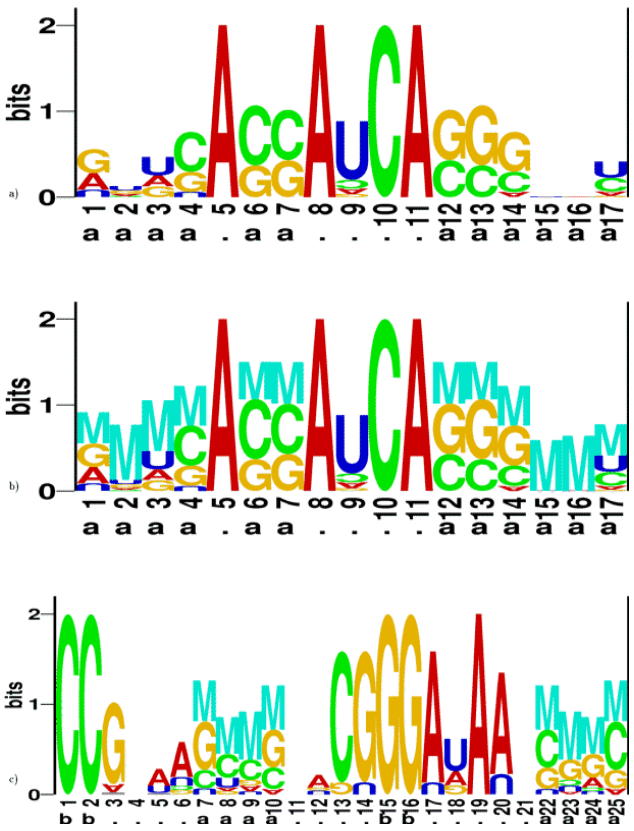

Fig. 1. The logos for the data. The symbols for sequence alignment are A, C, G, U, and –. The letter involving structure is M. The logos are for (**a**) the pure sequence logo for the R17 data set, (**b**) the sequence–structure logo for the R17 data set, and (**c**) the sequence–structure logo for the HIV data set. Only positions with positive information content are shown. The letters 'a' and 'b' indicate basepairing for respective regions.

tion and additional mutual information, so that the total information is nearly as great as in the conserved positions (5, 8, 10 and 11).

Figure 1c is an example of RNA which contains a pseudoknot with specific affinity for HIV-1-RT (Tuerk *et al.*, 1992). The actual alignment shown is the result of the structural alignment performed in Gorodkin *et al.* (1997b) plus adding the knowledge of the pseudoknot involved. Only positions which do not contain gaps are assigned to be involved in base pairing. This example illustrates how completely conserved base pair regions display the same amount of information as in plain sequence logo, but how base pairings for non-sequence conservation obtain additional signal from the mutual information.

Final remarks

We presented structure logos as an extension of sequence logos to display the consensus sequence and secondary structure of structural RNA alignments. We find that the structure logo captures both sequence and structure information of the structural alignments, and is a natural extension of the sequence logo. The illustrations clearly show that including the mutual information for base-paired regions indicates much more clearly the important constraints within the motif. In addition, we altered the plain sequence logo to include gaps in the alignment, to include any prior distribution of the bases, and to display the height of each base relative to its expected frequency, and having it turned upside down in the case of under-representation compared to its prior (or expected) probability. The structure logo may also be used to display correlations between structure and sequence; aligning sequences according to a given common structure might reveal signals in sequence outside the structure regions. Such a correlation is, for example, known in ribosomal frame shifting, where the pseudoknot is correlated to 'slippery' regions (Le *et al.*, 1991).

Further, we suggest that the way to calculate information content for pure sequence alignment be extended to proteins as well. Finally, we emphasize that the method of using mutual information between correlated positions in an alignment can be extended beyond displaying secondary RNA structures.

Acknowledgements

This work was partially supported by NIH grant HG00249 to G.D.S. J.G. and S.B. were supported by the Danish National Research Foundation. We thank Russ Altman and Sean Eddy for useful comments and suggestions.

Note added in proof

A similar page for protein sequence alignments is available at http://www.cbs.dtu.dk/gorodkin/appl/plogo.html

References

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley and Sons, USA.

Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997a) Finding common sequence and structure motifs in a set of RNA sequences. In Gaasterland, T. *et al.* (eds.): *Proceedings of the Fifth International Conference on Intelligent Systems in Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 120–123.

Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997b) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res,* 25, 3724–3732.

Gutell,R.R., Power,A., Hertz,G.Z., Putz,E. 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.

Hertz,G.Z. and Stormo,G.D. (1995) Identification of consensus patterns in unaligned DNA and protein sequences: A large-deviation statistical basis for penalizing gaps. In Lim,H.A. and Cantor,C.R. (eds), *Proceedings of the Third International Conference on Bioinformatics and Genome Research*. WORLD Scientific Publishing Co. Ltd, Singapore, pp. 201–216.

Hertz,G.Z., Hartzell,G.W.,III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Applic. Biosci.*, **6**, 81–92.

Le,S., Shapiro,B.A., Chen,J., Nussinov,R. and Maizel,J.V. (1991) RNA pseudoknots downstream of the frameshift sites of retrovirus. *Genet. Anal. Tech. Applic.*, **8**, 191–205.

Schneider,D., Tuerk,C. and Gold,L. (1992) Ligands to the bacteriophage R17 coat protein. *J. Mol. Biol.*, **228**, 862–869.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

Tuerk,C., MacDougal,S. and Gold,L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl Acad. Sci. USA*, **89**, 6988–6992.