WebLogo: A Sequence Logo Generator

Gavin E. Crooks,1 Gary Hon,1 John-Marc Chandonia,2 and Steven E. Brenner1,2,3

1Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA; 2Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

WebLogo generates sequence logos, graphical representations of the patterns within a multiple sequence alignment. Sequence logos provide a richer and more precise description of sequence similarity than consensus sequences and can rapidly reveal significant features of the alignment otherwise difficult to perceive. Each logo consists of stacks of letters, one stack for each position in the sequence. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino or nucleic acid at that position. WebLogo has been enhanced recently with additional features and options, to provide a convenient and highly configurable sequence logo generator. A command line interface and the complete, open WebLogo source code are available for local installation and customization.

Sequence logos were invented by Tom Schneider and Mike Stephens (Schneider and Stephens 1990; Sharer et al. 1993) to display patterns in sequence conservation, and to assist in discovering and analyzing those patterns. As an example, the accompanying figure (Fig. 1) shows how WebLogo can help interpret the sequence-specific binding of the protein CAP to its DNA recognition site (Schultz et al. 1991). Homodimeric DNA-binding proteins typically display a symmetric double hump in the DNA binding-site logo (Schneider and Stephens 1990), as shown in the figure. Deviations from this basic pattern can indicate additional features; a highly conserved residue in the center of such a pattern may indicate DNA distortion or base flipping (Schneider 2001); an unexpectedly high-sequence conservation may be due to overlapping binding sites (Schneider et al. 1986). Protein logos can illuminate patterns of amino acid conservation that are often of structural or functional importance (Galperin et al. 2001; Rigden et al. 2003). Sequence logos have also been used to display patterns in the BLOCKS protein sequence database (Henikoff et al. 1995), and in DNA-binding site motifs (Robison et al. 1998; Nelson et al. 2002), to analyze splice sites (Stephens and Schneider 1992; Emmert et al. 2001), and in a variety of other contexts. Additional examples, and the raw data for the example presented here, can be found on the WebLogo examples page (http://weblogo.berkeley.edu/examples.html).

The logo generation form (http://weblogo.berkeley.edu/logo.cgi) can process RNA, DNA, or protein multiple sequence alignments provided in either FASTA (Pearson and Lipman 1988) or CLUSTAL (Higgins and Sharp 1988) formats. If the user does not explicitly specify the sequence type, then WebLogo will make a determination on the basis of the symbols found within the sequences. A logo represents each column of the alignment by a stack of letters, with the height of each letter proportional to the observed frequency of the corresponding amino acid or nucleotide, and the overall height of each stack proportional to the sequence conservation, measured in bits, at that position. The letters of each stack are ordered from most to least frequent, so that one may read the consensus sequence from the tops of the stacks. For example, the figure shows that the CAP binding-site consensus sequence is AA-TGTGA------TCACA-TT. Schneider and Stephens (1990) define the sequence conservation at a particular position in the alignment, Rs, as the difference between the maximum possible entropy and the entropy of the observed symbol distribution:

\[ R_{\text{seq}} = S_{\text{max}} - S_{\text{obs}} = \log_2 N - \left( \sum_{n=1}^{N} p_n \log_2 p_n \right) \]

Here, \( p_n \) is the observed frequency of symbol \( n \) at a particular sequence position and \( N \) is the number of distinct symbols for the given sequence type, either four for DNA/RNA or 20 for protein. Consequently, the maximum sequence conservation per site is \( \log_2 4 = 2 \) bits for DNA/RNA and \( \log_2 20 = 4.32 \) bits for proteins. If we neglect the inter-site correlations and assume a uniform background symbol distribution, then the total entropy of the logo, the sum of the sequence conservation at each position, measures the information content of the logo. For binding sites, this total entropy has, in many cases, been shown to be approximately equal to the amount of information needed to locate the binding site within the relevant stretch of DNA (Schneider et al. 1986). For a nonuniform background distribution, such as found in protein sequences or the genomes of many hyperthermophiles, the information content would be given by the relative entropy between the observed and background distributions (Cover and Thomas 1991; Gorodkin et al. 1997; Stormo 1998).

Limited sequence data results in a systematic underestimation of the entropy, which becomes significant if the multiple alignment contains fewer than about 20 nucleotide or 40 protein sequences. By default, WebLogo incorporates a small sample correction (Schneider et al. 1986), which can, in part, ameliorate this bias. In addition, WebLogo can optionally display error bars with heights twice this correction, which gives some idea of the sampling errors made. Note that the error bars may not have uniform height across the logo, as the magnitude of the small sample correction depends on the number of symbols observed at each position. This will vary due to the presence of gaps in the alignment.

A standard sequence logo does not provide any indication of correlations between different positions of the alignment. In general, such inter-site correlations are relatively insignificant in biological sequences (Schneider 1997; Stormo 1998), but there are exceptions, such as base-paired sites in folded RNA structures. Structural logos (Gorodkin et al. 1997), an extension of the sequence logo idea, display part of this additional level of detail.

The symbols that compose the stacks display colors according to the chemical species they represent. The default colors for
nucleotides are G, orange; T and U, red; C, blue; and A, green. Amino acids have colors according to their chemical properties (Lewin 1994); polar amino acids (G, S, T, Y, C, Q, N) show as green, basic (K, R, H) blue, acidic (D, E) red, and hydrophobic (A, V, L, I, P, W, F, M) amino acids as black. The user may customize the coloring scheme, or select a simple black and white option.

WebLogo can create output in several common graphics’ formats, including the bitmap formats GIF and PNG, suitable for on-screen display, and the vector formats EPS and PDF, more suitable for printing, publication, and further editing. Additional graphics options include bitmap resolution, titles, optional axis, and axis labels, antialiasing, error bars, and alternative symbol formats.

The Web site is available to all users without fee. Those who would prefer to run WebLogo on a local server may obtain a command line version with source code (distributed under an Open Source license). We welcome bug reports and suggestions for additional features. Please send these to logo@compbio.berkeley.edu.

ACKNOWLEDGMENTS

WebLogo uses PostScript code and ideas from the programs alpro and makelogo, both part of Tom Schneider’s delila package (Schneider et al. 1982). Many thanks to him for making this software freely available, for encouraging its use, and for feedback on WebLogo. We are also grateful for the enthusiastic encouragement of Michael Galperin. Grants from the NIH (1-K22-HG00056) and the Searle Scholars program (01-L-116) support this work. JMC was supported by NIH grant T-F50-GM62412, and DOE contract DE-AC03-76SF00098. GEC received funding from the Sloan/DOE postdoctoral fellowship in computational molecular biology.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ‘advertisement’ in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


WEB SITE REFERENCES


Received September 26, 2002; accepted in revised form January 6, 2004.