A Brief history of sequence logos

Mini Review

A Brief History of Sequence Logos

Kushal K Dey\*

\*Department of Statistics, University of Chicago

5747 S Ellis Ave, Chicago, IL 60637

Correspondence Author

Phone: 312-709-0680

Email: [kkdey@uchicago.edu](mailto:kkdey@uchicago.edu)

**Abstract**

For nearly three decades, sequence logo plots have served as the standard tool for visual representation of aligned sequences of DNA, RNA and proteins. Over the years, a large number of packages and web servers have been developed - some aimed at novel visual representations of these logo plots while others focusing on applying these plots for further downstream analysis of the conserved patterns. Also, over time, we have seen a massive upgrade in the look, flexibility of data handling and the overall scope of these plots in biological applications and beyond. In this paper, I attempt to review some of the popular softwares for making and analyzing sequence logos, with a focus on how these plots have evolved over time and how

I view the future for these plots.

**Introduction**

The seeds of the origin of sequence logos were planted in the early 1980s when researchers, now equipped with a large body of DNA and RNA sequence data, started to develop keen interest in understanding how base compositions at different positions of these aligned DNA/RNA sequences, together with the compositional patterns at neighboring positions, contribute to characterizing structural and functional domains – binding sites, structural RNA, introns etc. For example, a protein or a macromolecule would tend to bind at a site in the DNA that has a recognizable pattern. Initial attempts to determine this pattern focused on building a consensus sequence from the many aligned sequences – however, this approach was criticized for its lack of predictive power (Sadler 1983, Hawley 1983). In 1986 (ref), Schneider et al defined information content of base probabilities at a position as a measure to determine the patterns of relative information across sites. An obvious next step was to visualize succinctly, this flow of information content across the sequences and the proportional contribution of the different bases to this information at each position. This resulted in the origin of sequence logo plot in a seminal work by Schneider and Stephens in 1990 (ref). These sequence logos were also extended to structure logos for aligned RNA sequences with base pairings, that allowed for gaps in the sequences and prior knowledge of base composition. This was followed by sequence logo plots with a known non-uniform underlying background model, which called for the information content to be replaced by the more flexible Kullback-Leibler (KL) divergence.

The current widespread use of sequence logos was largely the result of the development of software packages (in R/python) and web applications that made it extremely easy for the user to create these sequence logos. Possibly the oldest and the most commonly used such tool is the Bioconductor package seqLogo, which is already 11 years old and has been downloaded more than 100,000 times already. seqLogo primarily focused on generating logo plots for DNA sequences from the matrix of base compositions at each position of the aligned sequences. However, later packages proposed the use of logo plots for RNA and amino acid sequences (RWebLogo, seq2Logo, ggseqlogo) and have recently been extended further to sequences of any alphanumeric characters or strings (Logolas). An user currently has access to a plethora of open source softwares that allow for different model assumptions and stylistic configurations in creating the logo plots and enable one to easily interpret the logo representation and perform further downstream analysis – including determining conserved patterns of bases in the sequences called *motifs* and performing motif based comparisons and predictions. In the next section, we discuss the most basic model for creating logo plots and review the functionalities of several of the open source packages available online.

**Discussion**

Consider a set of *n* aligned sequences, with each element of the sequence corresponding to one of a total cohort of *J* symbols. Suppose each aligned sequence is L symbols long. Then the information content at each of these *L* positions can be defined as

*IC(l) = log2 (J) – (Hl + en) Hl = - ∑j plj log\_2 plj ∑j plj = 1 en = (1/log 2) \* (J-1)/2n (1)*

where *plj*  is the composition of element *j* in position *l* and *en* is a correction term adjusting for the number of samples bias (check that it is close to 0 for large *n*). The value of *J* equals 4 for DNA/RNA sequences corresponding to the 4 bases A, C, G, T (U) and equals 20 for protein sequences, corresponding to the 20 amino acids. The height of the stack of logos at position *l* equals to the information content *IC(l)* and the proportional

height of each symbol *j* in the stack equals *plj..* Instead of actual sequences, if only, the positional weights *plj* are available, then the correction term en  is removed from the expression of *IC(l)* in (1).When prior probabilities of symbols *qj* are known, then we replace IC(l) by KL(l; q) reported below for determining stack heights.

KL (l ; q) = - *∑j plj log\_2 ( plj /qj)*

The first breakthrough in terms of software for creating logo plots took place in 2004 with the web application WebLogo designed to generate logo plots from multiple sequence alignment of DNA, RNA and proteins following the same model as (1). WebLogo provided various style options to the user to customize the plots and save the results in multiple formats. However, as an web application, WebLogo required the user to upload the data, chose options and then save the results for each example set of sequences, making it labor intensive to run over many examples. In 2014, WebLogo was wrapped in an R package RWebLogo to address this limitation.

The R package that made creating logo plots a one function step, with the flexibility to loop over many examples with minimal user intervention, was seqLogo. This package assumed the input to be the position weight probabilities obtained from aligning DNA sequences and would follow the model assumptions in (1) sans the correction factor. The succeeding tools like the web application seq2Logo addressed these limitations of seqLogo. Seq2Logo was in fact flexible in handling both sets of sequences (as in WebLogo, RWebLogo), as well as position specific scoring matrices for both DNA/RNA and protein sequences, allowed for non-uniform prior probabilities for the symbols, as well as different methods to determine stack heights besides (1) and (2).

Several packages have also been developed suggesting improvements on the model in (1). K-mer probability logo uses positional interdependencies of aligned sequences to visualize conserved patterns (motifs). dagLogo can visualize conserved amino acid sequence patterns in groups defined by charge, hydrophobicity etc. pLogo uses a probability based method where stack heights are scaled based on statistical significance of the alignment with respect to a background model. The iceLogo web server also uses probability theory to scale the heights even in the absence of a background model. R package Logolas uses median adjustment of log positional weights to highlight both enrichment and depletion of symbols in the logo plot.

Besides the modeling framework, recently we have also observed stylistic changes and improved flexibility in the logo plotting mechanism. Some recent packages gglogo and ggseqlogo have integrated ggplot2 graphics with logo plots, thereby generating highly improved and publication ready logo plot visualizations. While the initial packages were restricted to using only English alphabets as symbols in a logo plot, ggseqlogo has the flexibility to use numbers and Greek letters as symbols as well. R package Logolas has further extended this flexibility to have any alphanumeric strings as symbols, thereby extending the scope of logo plots beyond DNA, RNA and protein sequences to visualizing mutation signature profiles and more general compositional data.

One primary application of logo plots has been in detection of conserved patterns or motifs in a set of aligned sequences. Several packages have been developed for identifying and comparing these motifs across many sets of aligned sequences. motifStack R package uses a distance metric based on STAMP to calculate distance between motifs and subsequently visualize the alignment of multiple motifs on a tree or circle based on the similarity scores., with an aim at identifying motifs in these difference patterns.

R package DiffLogo visualizes pairwise differences in motifs corresponding to multiple sequence logos, with an intuitive visualization of the difference patterns. R package Two sample Logo also uses statistical tests to compare between two aligned sets of sequences and visualize the differences between them. R package motifcounter matches a sequence with previously known motifs and then uses the number of motif hits based on match scores, with that of random DNA sequences for enrichment.

Various databases currently host aligned sequence data or positional frequency (weight) data for different transcription factors and proteins, with the data generated from wet lab experiments such as HT-SELEX, Chip-Seq, Chip-chip etc. Motif prediction algorithms used to generate the PFM (PWM) include MEME, ChiPMunk etc. TRANSFAC, JASPAR are popular databases for transcription

1. Smith, T.F., Waterman, M.S. and Sadler, J.R., 1983. Statistical characterization of nucleic acid sequence functional domains. *Nucleic acids research*, *11*(7), pp.2205-2220.
2. Hawley, D.K. and McClure, W.R., 1983. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic acids research*, *11*(8), pp.2237-2255.