SOFTWARE

Enrichment Depletion Logo plots

Kushal K Dey[1*], Dongyue Xie[1] and Matthew Stephens[1,2]

[*]Correspondence:
kkdey@uchicago.edu
[1]Department of Statistics,
University of Chicago,
60637,Chicago, USA
Full list of author information is available at the end of the article

Abstract

Background:
Sequence logo plots have developed into a standard graphical tool for identifying sequence motifs in DNA, RNA or protein sequences, largely because of its ease of interpretation and the visual appeal. However standard logo plots tend to be biased towards highlighting enrichment of symbols, thereby occasionally missing out on finer motif patterns.

Results:
In this article, we propose a new logo representation that highlights both enrichment as well as depletion of symbols at each position, resulting in a more parsimonious visualization. We show the benefits of this representation over the standard information content based logo plot through applications in displaying transcription factor binding site motifs, protein sequence alignments and mutational signature profiles.

Conclusion:
We present an easy-to-use and highly customizable R package *Logolas* that allows the user to plot such enrichment depletion logo plots where the characters in the logo plot can be any string symbol, consisting of alphabets, numerics, punctuations, dots, dashes etc.

Keywords: Logo plots; Enrichment Depletion; EDLogo; String symbols

Background

Ever since their introduction in early 90's by Schneider and Stephens [1], sequence logos have found extensive use in identifying short conserved patterns, also called sequence motifs, in multiple alignment of DNA, RNA and protein sequences. In the standard sequence logos, for each position in the aligned sequences, symbols representing the sequence are stacked on top of each other and the height of each symbol is proportional to its positional frequencies. The stack height is determined by the information content at that position. Several packages in R such as *seqLogo* [2] (exclusive to DNA, RNA sequence alignment), *RWebLogo* (Wagih 2014), *ggseqlogo* (Wagih 2017) [3] and web servers like *WebLogo* (Crooks et al 2004) [4], *Seq2Logo* (Thomsen and Nielsen 2012) [5], *iceLogo* (Coalert el al 2009) [6] etc have been developed for sequence logo visualization of aligned DNA, RNA and protein sequences.

The standard sequence logo visualization based on information content tends to primarily show the enrichment of the symbols (nucleotides or amino acids) at each position. Though *seq2Logo* allows the user to plot position specific scores that account for both enrichment and depletion, the representation is not parsimonious [5]. We introduce here a logo visualization package, *Logolas*, which allows the user to highlight both the enrichment as well as the depletion of symbols in a logo plot, but in a parsimonious and visually appealing way. We call this representation the *Enrichment Depletion Logo* or *EDLogo* plot. Additionally most logo plotting softwares are mainly limited in their applications to DNA, RNA and protein sequence alignment compositional data. *Logolas* provides the user the flexibility to plot logos for any alphanumeric strings and not just English alphabets as in standard packages, which extends the applicability of logo plots to more generic compositional data with string labels. In this article, we demonstrate various applications of the *EDLogo* representation and also highlight several features of the *Logolas* package.

Implementation

In **Supplementary Figure 1**, we illustrate the main intuition behind the *EDLogo* plot. Say for a specific position in a set of aligned DNA sequences, the relative frequencies are $p = (p_A, p_C, p_G, p_T) = (0.33, 0.33, 0.33, 0.01)$. A standard logo will show three equally high symbols A, C, G stacked vertically along the positive Y axis with T at the bottom having negligible height. The standard sequence logo plot is biased towards highlighting base enrichments. So, when this position is flanking by highly enriched bases as in panel (a), its stack height would be relatively smaller compared to that of the neighboring positions which would make the depletion of T even harder to see (panel (b)). *EDLogo* provides an alternative parsimonious and perhaps more meaningful representation by highlighting the depletion of T along the negative Y axis. This representation is designed to highlight large enrichments as well as large depletions, as observed in panel (c). We present the algorithm behind computing the enrichment and depletion of characters for the *EDLogo* representation below.

Let $p_n = (p_{n1}, p_{n2}, \ldots, p_{nB})$ be the weights (normalized position frequencies of aligned sequences) of B characters at site n and $q_n = (q_{n1}, q_{n2}, \ldots, q_{nB})$ be the corresponding background probabilities. B is equal to 4 for DNA, RNA sequence examples and equal to 20 for protein sequences. Typically we encounter q_n to be same for all positions n ($q_n \equiv q$).

We first define a score vector $r_n = (r_{n1}, r_{n2}, \ldots, r_{nB})$ for each n.

$$r_{nb} = \log_2 \frac{p_{nb} + \epsilon}{q_{nb} + \epsilon} - median \left(\left\{ \log_2 \frac{p_{nb} + \epsilon}{q_{nb} + \epsilon} : b = 1, 2, \ldots, B \right\} \right) \tag{1}$$

where ϵ is a thresholding parameter controlling the effect of small position weight values p_{nb}. The default choice of ϵ is 0.1.

Once the r_{nb} scores have been defined, we next compute two new scores r^{+}_{nb} and r^{-}_{nb} as follows.

$$r^{+}_{nb} = r_{nb}\mathbf{I}(r_{nb} \geq 0) \qquad\qquad r^{-}_{nb} = r_{nb}\mathbf{I}(r_{nb} < 0) \tag{2}$$

where $\mathbf{I}$ is the indicator function. The above formulation ensures that at each base b, one of r^{+}_{nb} and r^{-}_{nb} is zero.

For each site n, we plot the r^{+}_{nb} values along the positive Y axis and the r^{-}_{nb} values along the negative Y axis. For a base b enriched at position n, r^{+}_{nb} value will be large resulting in large size of the symbol for base b in the positive Y axis of the *EDLogo* plot. For a base b depleted at position n, r^{-}_{nb} value will be large resulting large size of the symbol for base b in the negative Y axis of the *EDLogo* plot.

Results

Sequence logos have been used extensively in visualizing transcription factor binding motifs (TFBSs). Though base enrichment is usually the more prevalent feature in most positions of the binding motif, some transcription factors tend to show depletion of bases at specific positions. In Figure 1(panel (A)), we present the standard logo and the *EDLogo* representation of the Early B cell factor 1 disc 1 (EBF1-disc1) transcription factor. The sequence motif shows strong depletion signals of bases G and C at the center of the sequence. Furthermore, this depletion is also a part of the palindrome TCCCg - cGGGA, where lowercase letters stand for depletion and uppercase case letters stand for enrichment of characters. Note that this depletion signal is hard to see in the standard logo plot (panel (A) *left*) because it is flanked by strong enrichments, but the *EDLogo* representation (panel (A) *right*) shows it clearly. In **Supplementary Figure 2**, we present the *EDLogo* representation of all the members of the EBF1 family and besides EBF1-disc1, the depletion of G and C at the center of the palindromic sequence is also observed in EBF1 - known3 and EBF1 - known4.

In Figure 1 panel (B), we compare the *EDLogo* plot (*right*) with the PSSM profile representation (*left*) of the binding sequence of the Bacterial transcription activator, effector binding domain protein PF06445 (motif 4, Start=153 Length=8). The PSSM visualization is a common alternative to standard logo plots for protein sequences. But here it is evident that the *EDLogo* representation is much more parsimonious and interpretable than the PSSM representation. The main reason behind the sparse visualization in *EDLogo* is due to the median adjustment in determining the enrichment and depletion patterns (see Implementation). Both the position weight matrix (PWM) and position specific scoring matrix (PSSM) for this protein have been fetched from 3PFDB webpage http://caps.ncbs.res.in/3pfdb/ [7] [8].

Another application of *EDLogo* plot is in visualizing mutational signature profiles. These plots also employ another cool feature of *Logolas*, to be able to plot logos for strings. Each mutation signature is usually represented by the type of C and T mutations, ($C \to T$, $C \to A$, $C \to G$, $T \to A$, $T \to C$, $T \to G$) flanked by bases to the left and right. The A and G mutations are clubbed with C and T mutations above to avoid strand bias. In Figure 1 panel (C), we compare the standard sequence logo (*left*) and the *EDLogo* representations (*right*) of the mutational signature profile of lymphoma B cell mutations in Alexandrov et al [9]. We observe that it is much easier to identify the depletion of G on the right flanking base (possibly occurring due to methylated CpG sites being less prone to mutation) in the *EDLogo* plot compared to the standard sequence logo plot. In **Supplementary Figure 3**, we compare the *EDLogo* representation with the *pmsignature* representation due to Shiraishi et al [10] for the same B cell lymphoma mutation type example as in Figure 1 panel (C) and it is evident that the *EDLogo* plot depicts the overall features of the logo plot way more clearly. In **Supplementary Figure 4**, we present the *EDLogo* representations of all 27 cancer mutation signature profiles reported in Figure 4 of Shiraishi et al [10].

Discussion

Besides the approach shown in Implementation, *Logolas* allows the user to use other options for computing the enrichment and depletion levels of each character in a position of the aligned sequences. We discuss these options in greater detail under Supplementary Methods. In **Supplementary Figure 5** and **Supplementary Figure 6**, we present the different types of *EDLogo* representation for the transcription factor EBF1-disc1 and the protein PF06445 example in Figure 1. The *EDLogo* representation generates heights of the bases along the positive and negative Y-axes representing enrichment and depletion respectively, which can be used as scores for these downstream analysis - like motif matching, comparing motif patterns, regulatory SNP detection etc (see packages *DiffLogo* [11], *motifStack* [12], *atSNP* [13]).

Besides the *EDLogo* representation and string symbols features, *Logolas* provides many other customizable features - different fill and border styles for enriched and depleted symbols in *EDLogo* plot, different approaches of calculating stack heights for a standard logo plot (Renyi entropy at differet scales, Shannon entropy, relative frequnecy based plot), plotting logos in multiple panels and also combining logo plots with external ggplot2 graphics.

The Logolas package is currently released on Bioconductor (`https://bioconductor.org/packages/release/bioc/html/Logolas.html`) and is also under active development on Github (`https://github.com/kkdey/Logolas`). The codes for reproducing the figures in this paper are available on (`https://github.com/kkdey/Logolas-paper`). Vignettes and gallery of logo representation demonstrating various features of Logolas are available at (`https://github.com/kkdey/Logolas-pages`)

Author details

[1]Department of Statistics, University of Chicago, 60637,Chicago, USA. [2]Department of Human Genetics, university of Chicago, 60637, Chicago, USA.

References

1. Schneider, T.D., Stephens, R.: Sequence logos: a new way to display consensus sequences. Nucleic Acids Research **18 (20)**, 6097–6100 (1990)
2. Bembom, O.: seqlogo: Sequence logos for dna sequence alignments. R package version 1.42.0
3. Wagih, O.: ggseqlogo: a versatile r package for drawing sequence logos. Bioinformatics **btx469** (2017)
4. Crooks, G.E.: Weblogo: A sequence logo generator. Genome Research **14 (6)**, 1188–1190 (2004)
5. Thomsen, M.C., Nielsen, M.: Seq2logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Research **40**, 281–287 (2012)
6. Coalert, N., Helsens, K., Martens, L., Vandekerckhove, J., Gevaert, K.: Improved visualization of protein consensus sequences by icelogo. Nature Methods **6**, 786–787 (2009)
7. Shameer, K., Nagarajan, P., Gaurav, K., Sowdhamini, R.: 3pfdb - a database of best representative pssm profiles (brps) of protein families generated using a novel data mining approach. BioData Min. **2(1)**, 8 (2009)

8. Joseph, A.P., Shingate, P., Upadhyay, A.K., Sowdhamini, R.: 3pfdb+: improved search protocol and update for the identification of representatives of protein sequence domain families. Database (Oxford) **bau026** (2014)
9. Alexandrov, L., Nik-Zainal, G., Wedge, D., Campbell, P., Stratton, M.: Deciphering signatures of mutational processes operative in human cancer. Cell Reports **3(1)**, 246–259 (2013)
10. Shiraishi, Y., Tremmel, G., Miyano, s., Stephens, M.: A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genetics **11(12)**, 1005657 (2015)
11. Nettling, M., Treutler, H., Grau, J., Keilwagen, J., Posch, S., Grosse, I.: Difflogo: a comparative visualization of sequence motifs. BMC Bioinformatics **16(387)**, 1188–1190 (2015)
12. Ou, J., Zhu, L.: motifstack: Plot stacked logos for single or multiple dna, rna and amino acid sequence (2015). R package version 1.18.0
13. Zuo, C., Sunyoung, S., K.: atsnp: transcription factor binding affinity testing for regulatory snp detection. Bioinformatics **31(20)** (2015)
14. Koch, C.M., et al.: The landscape of histone modifications across 1in five human cell lines. Genome Research **17(6)**, 691–707 (2007)
15. Kheradpour, P., Kellis, M.: Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. Nucleic Acids Research, 1–12 (2013)

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KKD and MS conceived the idea. KKD implemented the package. KKD and DX tested Logolas on the data applications. KKD, DX and MS wrote the manuscript.

Acknowledgements

The authors would like to acknowledge Yuichi Shiraishi, John Blischak, Peter Carbonetto and Hussein Al-Asadi for their valuable feedback and helpful discussions. The authors would also like to thank Alexander White and Trevor Price for the ecological data application.

Figures

Figure 1 Comparison of standard logo, weighted KL logo and EDLogo representations for various studies. We present a comparative study of the *EDLogo* representation with respect to the standard logo and the weighted KL logo representation due to seq2Logo software [5], through various examples. In (panel (A)), we present the logo representation of the transcription factor binding site of the EBF1-disc1 transcription factor. We observe that the *EDLogo* plot captures the depletion of G and C in the middle of the sequence and the overall palindromic nature of the enrichment and depletion in the binding motif much better than the standard logo and the weighted KL logo. In panel (B), we compare the three approaches with respect to visualizing the binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)*. We observe that the *EDLogo* representation is visually more parsimonious and detailed than the weighted KL logo. For the plots in panels (C) and (D), we use the special feature of Logolas of plotting logos for string symbols. For example in panel (C), we present the logo representation of the mutational signature profile of the all mutations in lymphoma B cells, with data taken from Alexandrov et al 2013 [9]. We observe that the depletion of G to the right of the mutation - possibly occurring due to the rarity of CpG sites owing to de-amination of methylated cytosines - much more clearly in the *EDLogo* representation compared to the other approaches. In panel (D), we present the logo representations of the distribution of histone modification sites across various genomic regions in the lymphoblastoid cell line GM06990 (Table S2 in Koch et al 2007 [14]). The data in this case is not compositional and we use the scoring scheme in Equation ?? for determining the relative heights of the symbols. In the standard logo representation, the patterns of variation at gene end and gene start regions are hard to see, whereas the weighted KL logo representation is largely dominated by the discrepancy in the frequency of histone modifications between the observed table and the background. The EDLogo representation is more interpretable and reflects patterns in histone marks across various regions along expected lines.

Supplementary Figures

S1 Fig. **Illustration of the EDLogo representation.** We present an illustration of how *EDLogo* representation accounts for depletion signal and provides a more informative visualization of the sequence motif. In panel (a), we present a position weight matrix with the position weight vector at the second position having a depletion of T, but is flanked by enrichments around it. In panel (b), we present the corresponding standard logo plot representation of the PWM matrix in panel (a). The signal at the second position gets swamped by the bias towards enrichment signals flanking it. In panel (c), we present the *EDLogo* representation of the PWM matrix, where both the enrichment signals as well as the depletion signal at position 2 are clearly observed.

S2 Fig. **Mechanism behind EBF1-disc1 transcription factor binding**: We present a demo of how the loss of binding affinity of the transcription factor EBF1-disc1 to the binding site in presence of G and C in the middle of the site is reflected as a depletion signal in the *EDLogo* representation.

S3 Fig. **EDlogo representation of the members of the EBF1 family of transcription factors**: We present the *EDlogo* representation for the binding sites of 6 transcription factors in the EBF1 family. EBF1-known4 and EBF1-disc1, and also to some extent EBF1-known3 seem to show the depletion of G and C in the middle of the binding site. The PWM data for all the transcription factors have been obtained from the ENCODE TF Chip-seq datasets and are hosted on the webpage http://compbio.mit.edu/encode-motifs/ [15].

S4 Fig. **Different options for EDLogo representation - Protein example**: We present the *EDLogo* representation of the binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* under several other scoring schemes (*log-odds*, *ratio*, *ic-log*, *ic-ratio*, *ic-log odds* and *probKL*) besides the log based scoring used in Figure 1 Panel (B).

S5 Fig. **Comparison of Logolas EDLogo plot with pmsignature representation for cancer mutation signatures**: We compare the *EDLogo* plot representation and the *pmsignature* representation due to Shiraishi et al (2015) [10] for mutation signature profile of lymphoma B cell from Alexandrov et al 2013 [9]. The position 0 corresponds to the mutation. Positions -1 and -2 correspond to the the two left flanking bases with respect to the mutation. Positions 1 and 2 correspond to the the two right flanking bases with respect to the mutation. Clearly, *EDLogo* representation shows the depletion of G at the right flanking base more clearly and is more interpretable and visually appealing in highlighting the overall mutation signature patterns compared to the*pmsignature* plot.

S6 Fig. **EDLogo plots for the mutational signature profiles of 30 cancer types in Alexandrov et al (2013)**: We present the sparse logo representations of the cancer mutational signature profiles across a number of tissues where the mutational signature data has been collected from 7042 cancers by Alexandrov et al (2013) [9]. Each mutational signature profile is represented by the mutation type at the center and the two bases flanking it to the left and two bases to the right.

Supplementary Methods

Here we discuss the additional options for creating stacks of symbols in the *EDLogo* plots. We call the method discussed in the Implementation section for computing the scores r_{nb} as the *log* approach. Some other other approaches for *EDLogo* plot scoring schemes are *log-odds*, *ratio* and their information content based counterparts namely *ic-log*, *ic-log-odds*, *ic-ratio*. Additionally we also implemented scoring scheme based on the probability weighted Kullback-Leibler logo (*probKL*) proposed in *Seq2Logo* [5].

Let p_n be the position weights of the symbols at position n and q_n be the background probabilities at that position. We define the score vector f_n for the *log-odds* and *ratio* approaches.

- *log-odds* approach

$$f_{nb} = \log_2 \frac{p_{nb}/(1-p_{nb})+\epsilon}{q_{nb}/(1-q_{nb})+\epsilon} - median\left(\left\{\log_2 \frac{p_{nb}/(1-p_{nb})+\epsilon}{q_{nb}/(1-q_{nb})+\epsilon} : b = 1,2,\dots,B\right\}\right) \tag{3}$$

- *ratio* approach

$$f_{nb} = \frac{p_{nb}+\epsilon}{q_{nb}+\epsilon} - median\left(\left\{\frac{p_{nb}+\epsilon}{q_{nb}+\epsilon} : b = 1,2,\dots,B\right\}\right) \tag{4}$$

- *probKL* approach

$$f_{nb} = p_{nb}\log_2 \frac{p_{nb}/(1-p_{nb})+\epsilon}{q_{nb}/(1-q_{nb})+\epsilon} - median\left(\left\{p_{nb}\log_2 \frac{p_{nb}/(1-p_{nb})+\epsilon}{q_{nb}/(1-q_{nb})+\epsilon} : b = 1,2,\dots,B\right\}\right) \tag{5}$$

As in the *log* approach, we define $f^{+}_{nb} = f_{nb}\mathbf{I}(f_{nb} \geq 0)$ and $f^{-}_{nb} = f_{nb}\mathbf{I}(f_{nb} < 0)$ where $\mathbf{I}$ is the indicator function. For the *ic* based approaches (*ic-log*, *ic-log-odds* and *ic-ratio*), we additionally compute the information content for each position $IC(n)$ and then redefine the f^{+}_{nb} and f^{-}_{nb} scores

$$f^{+}_{nb} \leftarrow IC(n) \times \frac{f^{+}_{nb}}{\sum_b \left(f^{+}_{nb} + f^{-}_{nb}\right)} \qquad f^{-}_{nb} \leftarrow IC(n) \times \frac{f^{-}_{nb}}{\sum_b \left(f^{+}_{nb} + f^{-}_{nb}\right)} \tag{6}$$

These f^{+}_{nb} values are plotted along the positive Y axis, while the f^{-}_{nb} values are plotted along the negative Y axis. In terms of performance, the *log* and *log-odds* tend to highlight the depletion signal more. On the other hand, all the *ic* based options - *ic-log*, *ic-log-odds* and *ic-ratio* are slightly biased towards the enrichment signal.