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 been used extensively for identifying short conserved patterns, called *sequence motifs*, in multiple alignment of DNA, RNA and protein sequences. In standard sequence logos, symbols of characters in the sequence are stacked on top of each other at each position of the aligned sequences. The height of the stack is determined by the information content of characters at that position and size of each symbol in the stack is proportional to the relative positional frequency of the corresponding character. Over the years, several softwares such as R packages *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 highlight the enrichment of the symbols at each position. Though *seq2Logo* provides the user several options 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 enrichment and depletion of symbols 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 to visualizing DNA, RNA and protein sequence alignment as they allow the user to plot only English alphabets. *Logolas* provides the user the flexibility to plot logos for any alphanumeric string and hence extends the applicability of logo plots to more generic compositional data with string labels. In this article, we demonstrate the utility of the *EDLogo* representation and also discuss several other 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 represent this position with a vertical stack along the positive Y axis of three equally high symbols for A, C, G on top with a symbol for T at the bottom having negligible height. However, when this position is flanked by highly enriched bases as in the matrix in panel (a), its stack height appears relatively smaller compared to that of the neighboring positions which makes the depletion of T hard to see (panel (b)). *EDLogo* provides an alternative parsimonious and arguably more interpretable representation by highlighting the depletion of T along the negative Y axis as well as the base enrichments at the neighboring positions along the positive Y axis (see panel (c)). We present the algorithm behind computing the enrichment and depletion scores of characters for the *EDLogo* representation below.

Assume that $p_n = (p_{n1}, p_{n2}, \ldots, p_{nB_n})$ denotes the weights or relative frequencies of B_n characters at position n of aligned sequences and $q_n = (q_{n1}, q_{n2}, \ldots, q_{nB_n})$ be the corresponding background probabilities. B_n is equal to 4 for each n in case of aligned DNA/RNA sequences and equal to 20 at each n for amino acid sequences. However in some of our applications, for example in mutation signature modeling, B_n may vary from one position of the sequence to another. Typically we encounter the background probability q_n to be same for all positions n ($q_n \equiv q$).

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

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

where ϵ is a thresholding parameter controlling the effect of small position weight values p_{nb}. The default choice of ϵ is 0.01.

Next, we compute r^+_{nb} and r^-_{nb} as follows.

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

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

Results

We compared the *EDLogo* representation with the standard logo and the weighted Kullback Leibler logo representation (Thomsen et al 2012, [5]) through four example studies - visualization of transcription factor binding sites (TFBS), protein binding sites, mutation signature profiles and histone mark composition (see Figure 1).

Sequence logos have been used extensively in visualizing transcription factor binding motifs (TFBSs) [7–9]. Though base enrichment is typically the more prevalent feature in a TFBS motif, some transcription factors tend to show depletion of bases at specific positions. In Figure 1(panel (a)), we present the logo representations of the Early B cell factor (disc 1 motif as reported in ENCODE `http://compbio.mit.edu/encode-motifs/` [8]). The sequence motif shows a strong signal of depletion of bases G and C at the center of the sequence. Furthermore, this depletion appears to be part of the palindrome TCCCg - cGGGA, where lowercase letters stand for depletion and uppercase case letters stand for enrichment of characters. The depletion signal is hard to see in standard logo since it is flanked by strong enrichments and in the weighted KL logo plot, it is overwhelmed by the depletion at all the other positions. In **Supplementary Figure 2**, we present the *EDLogo* representation of all the reported motifs of EBF1 in ENCODE besides EBF1 (disc1). The depletion of G and C at the center of the palindromic sequence is observed also in EBF1 (known3) and EBF1 (known4) motifs.

Another major field of application of sequence logos has been in visualizing protein sequence binding. In Figure 1 panel (B), we compare the different logo representations with respect to visualizing the binding sequence motif Motif2 (Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)*. The position weight matrix (PWM) for this protein has been fetched from 3PFDB webpage `http://caps.ncbs.res.in/3pfdb/` [10] [11]. The *EDLogo* representation is more parsimonious and arguably more interpretable than the weighted Kullback Leibler representation in displaying both enrichment and depletion patterns.

Besides the *EDLogo* representation, our package *Logolas* allows the user the flexibility to plot logos for alphanumeric strings. The utility of this feature is demonstrated in the applications in panels (c) and (d) in Figure 1. In Figure 1 panel (c), we compare the logo representations of the *mutation signature* profile of lymphoma B cell mutations in Alexandrov et al [12]. A mutation signature is typically represented by the type of C and T mutation, ($C \rightarrow T$, $C \rightarrow A$, $C \rightarrow G$, $T \rightarrow A$, $T \rightarrow C$, $T \rightarrow G$) flanked by the symbols for bases to the left and right. The A and

G mutations are clubbed with C and T mutations to avoid strand bias. From the logo representations in panel (c), we observe that all three logo representations show enrichment of $C \rightarrow T$ type somatic mutations in B cell lymphoma. However, in the *EDLogo* representation, it is easier to identify the depletion of G on the right flanking base, possibly occurring because methylated CpG sites get de-aminated quickly and hence are rare in the genome. In **Supplementary Figure 3**, we present the *EDLogo* representations of cancer mutation signature profiles across various tissue types, including B cell lymphoma, as reported in Alexandrov et al [12]. In **Supplementary Figure 4**, we compare the *EDLogo* representation with the *pmsignature* representation due to Shiraishi et al [13] for the same B cell lymphoma mutation signature example and the *EDLogo* plot arguably depicts the overall features of the mutation signature more clearly.

In Figure 1 panel (d), we present the logo representations of the relative abundance data of 5 different histone marks across various genomic regions in the lymphoblastoid cell line GM06990, as reported in Table S2 (*upper*) in Koch et al 2007 [14]. We used as background the randomly simulated data reported in Table S2 (*lower*) of Koch et al as background. The standard logo representation is dominated by the higher variation in histone marks at the intergenic, exon and intron regions. However, the *EDLogo* and weighted Kullback Leibler logo both highlight the enrichment of the histone marks H3AC and H3K4me3 in the gene start and gene end regions. H3K4me1 and H4AC have broad distributions across the genome (as shown in Koch et al 2007 [14]) and hence appear to be enriched with respect to other marks in the intergenic, exon and intron regions in all logo representations.

Discussion

We propose a new technique of logo visualization called *EDLogo* which by design, highlights both large enrichments as well as large depletion of characters at each position. For each site n, *EDLogo* computes a median adjusted log ratio of the probability at that position with respect to a specified background for each character b, called r_{nb}, to determine the enrichment or depletion of the character b. Apart from this log ratio score, *EDLogo* provides other options for determining the scores r_{nb} based on ratio, log odds ratio, probability adjusted log ratio and some information theory scaled versions of all these methods (see details in Supplementary Methods). In **Supplementary Figure 5**, we demonstrate a comparison of the different *EDLogo* scoring schemes for visualizing the protein binding motif data from Figure 1 panel (b). The log ratio method is simple, easy-to-use, allows for unconstrained range of values for the scores r_{nb} and also shows demonstrates the "mirror property" when the probabilities p and background q in Equation1 are flipped (see **Supplementary Figure 6**).

When applied to the transcription factor or protein binding data examples, the *EDLogo* representation outputs the strength of enrichment r^{+}_{nb} and depletion r^{-}_{nb} for each base b at position n. These scores can be used for further downstream analysis, like motif matching, comparing motif patterns, regulatory SNP detection etc (see packages *DiffLogo* [15], *motifStack* [16], *atSNP* [17]).

Besides the *EDLogo* representation and the flexibility of using string symbols, *Logolas* provides many other features - various customizable styles and color palettes for the enriched and depleted logos, various methods of calculating information content for determining stack heights in standard logo (Renyi entropy at differet scales, Shannon entropy, relative frequnecy based plot), ease of integrating logo plots with external graphics like ggplot2 etc.

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. Tan, G., Lenhard, B.: Tfbstools: an r/bioconductor package for transcription factor binding site analysis. Bioinformatics **32(10)**, 1555–1556 (2016)
8. Kheradpour, P., Kellis, M.: Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. Nucleic Acids Research, 1–12 (2013)
9. Zhao, X., et al.: Jaspar 2013: An extensively expanded and updated open-access database of transcription factor binding profiles. TBA **TBA**(TBA), (2013)
10. 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)
11. 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)
12. 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)
13. 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)
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. 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)
16. 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
17. Zuo, C., Sunyoung, S., S., K.: atsnp: transcription factor binding affinity testing for regulatory snp detection. Bioinformatics **31(20)** (2015)

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, Yang Li and Hussein Al-Asadi for their valuable feedback and helpful discussions.

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 [12]. 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 relative abundance distribution of histone modification sites across various genomic regions in the lymphoblastoid cell line GM06990 (Table S2 in Koch et al 2007 [14]). The EDLogo representation is more interpretable, in particular at the gene start and gene end regions, compared to the standard logo 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. **EDlogo representation of the different motifs of the EBF1 transcription factor**: We present the *EDlogo* representation of the 6 reported motifs of the transcription factor Early B cell Factor 1 (EBF1) in ENCODE project http://compbio.mit.edu/encode-motifs/ [8]. EBF1-known4 and EBF1-disc1, and to some extent EBF1-known3 showed the depletion of G and C in the middle of the binding site.

S3 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) [12]. 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.

S4 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) [13] for mutation signature profile of lymphoma B cell from Alexandrov et al 2013 [12]. 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.

S5 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).

S6 Fig. **Mirror Property of EDLogo - log ratio approach**: We compare the EDLogo representation of the EBF1 transcription factor (disc1 motif) against a background of uniform probabilities at each position in panel (a) with an EDLogo representation of uniform probabilities at all positions against a background given by the EBF1 (disc1) motif probabilities in panel (b). We show that the figure in panel (b) is a mirror image of the figure in panel (a).

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 scoring schemes offered by *EDLogo* are *log-odds*, *ratio* and the information content based counterparts of the above approaches, namely *ic-log*, *ic-log-odds*, *ic-ratio*. Additionally we also implemented a 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

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

- *ratio* approach

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

- *probKL* approach

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

As in the *log* approach, we define $r^+_{nb} = r_{nb}\mathbf{I}(r_{nb} \geq 0)$ and $r^-_{nb} = r_{nb}\mathbf{I}(r_{nb} < 0)$ 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 r^+_{nb} and r^-_{nb} scores

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

These r^+_{nb} values are plotted along the positive Y axis, while the r^-_{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.