SOFTWARE

Logolas : New Frontiers in sequence Logo visualization

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 the ease of interpretation and the visual appeal. However standard logo plots are limited in its applicability owing to limited set of symbols it can plot. Also standard logo plots tend to be biased towards highlight enrichment of symbols, thereby occasionally missing out on finer motif patterns.

Results:
In this article, we present an R package Logolas which allows the user to plot any string comprising of alphabets, numerics, punctuations, dots, dashes etc- which extends the scope of logo plots to most compositional data with labels that are strings. We show applications of string logos for visualizing mutation signature profiles, histone marks composition across different regions of DNA, ecological abundance patterns etc. Also, Logolas provides a new logo representation that highlights both enrichment as well as depletion of symbols at each position, resulting in a more parsimonious visualization of logos. Logolas also provides an adaptive method to scale the position weights based on positional frequency scales, leading to more accurate representation of logos.

Conclusions:
Logolas is easy to use and provides a handful of customizations for generating the logo visualizations. Also, Logolas widens the scope of logos by extending its use beyond the domain of DNA, RNA and protein sequence analysis, and also allows for more accurate and more visually appealing representation.

Keywords: Logo plots; String Logos; Depletion; Sparse Logos; Dirichlet Adaptive Shrinkage; Applications

Background

Sequence motifs are short conserved patterns in DNA, RNA and protein sequences that are believed to have biological significance. Such motifs can be used to identify transcription factor binding sites (TFBS), binding sites of proteins such as nucleases, splice sites etc. Such motifs can be identified through multiple alignment of DNA, RNA and amino acids, which provides a position specific frequency (or weight) for each nucleotide or amino acid that is subsequently stored as a column (per position) in a matrix called the Position Frequency (Weight) Matrix. A Position Frequency Matrix (PFM) or Position Weight Matrix (PWM) is often graphically represented by a sequence logo plot [1]. A sequence logo displays for each position a stack of symbols where each symbols corresponds to a base in a DNA/RNA sequence or an amino acid in a protein sequence. The height of the stack is determined by the information content of the position and the relative sizes of the letters correspond to the frequency of the symbols.

Sequence logos have proved to be an effective tool in informative visualization of sequence motifs in varied applications - for identifying Transcription Factor Binding Site motifs via *TFBStools* (Tan and Lenhard 2016) [2], visualization of sequence variation of protein complexes (Bryson et al 2009) [3], proteolytic cleavage sites (Mahrus et al 2008) [4], splice sites (Emmert et al 2001) [5] and patterns in BLOCKS protein sequences (Henikoff et al 1995, Henikoff et al 1999) [6, 7] etc. One of the first and most extensively used sequence logo visualization packages is *seqLogo* [8], which is exclusively targeted to DNA sequence motif visualization and has a library of only 4 symbols - A, C, G and T - corresponding to the four bases. Web servers like *WebLogo* (Crooks et al 2004) [9], *Seq2Logo* (Thomsen and Nielsen 2012) [10], *WebLogo* python package and *RWebLogo* R package (Wagih 2014) [11] allow the user to plot custom logo plots for both nucleotides and amino acids and have a library comprising of all English alphabets.

Several alternative ways of visualizing sequence logos have been suggested in recent years. R package *motifStack* (Ou and Zhu 2015) [12] proposes novel ways to graphically stack and compare multiple sequence motifs for a DNA, RNA or protein sequence. R package *DiffLogo* (Nettling et al 2015) [13] provides tools to visualize the pairwise differences in motif patterns in case of multiple motifs for a transcription factor or protein domain. *PWMenrich* (Stojnic and Diez 2015) [14] performs motif scanning and enrichment of motifs with subsequent visualization of the enrichment. *iceLogo* (Coalert el al 2009) [15] is a Java based web service that determines the logo stack heights using probability theory instead of information theory. ggseqlogo (Wagih 2017) [16] aggregates ggplot2 graphics with sequence logos to generate publication ready plots.

We introduce here another logo visualization package, *Logolas*, which addresses several limitations of the above packages and makes logo visualization a more generic tool with potential applications in a much wider scope of problems. The standard sequence logo visualization based on Information Content tends to highlight primarily the enrichment of the symbols (bases or amino acids) at each position. seq2Logo allows the user to plot position specific scores instead of position weights which highlight both enrichment and depletion but the representation is not parsimonious [10]. Logolas allows the user to highlight both the enrichment as well as depletion of symbols in a logo plot, but in a more parsimonious and visually appealing way. Also, unlike standard logo making softwares which are limited in their library size to either A, C, G and T or just English alphabets, Logolas allows the user to any string, comprising of alphabets, numerics, punctuations, dots and dashes, which basically encompasses almost all possible strings. As a result, Logolas can be used for graphical representation of almost all compositional data, well beyond the DNA, RNA and protein based PWM models. Standard logo plotting tools plot logos based on the PWM and fail to adapt to the scale of positional frequencies (PFM). Logolas provides a method, called dash (Dirichlet Adaptive Shrinkage) for adaptive scaling of the heights of the logos based on the positional frequencies, thereby providing more reliable logo plots when the alignment frequency for a position is low.

Implementation

In this section, we discuss the implementation details of the three primary functions in **Logolas** - *logomaker*, used for plotting the standard logo plot, *nlogomaker* used for sparse logo representation and *dash* for performing adaptive scaling of position weights.

Standard Logos

For the standard logo plot, *Logolas* uses information content to determine the height of the stack of symbols per position. This approach is similar to most standard logo plotting softwares like *seqLogo*, *WebLogo* etc. The information content at position n is given by

$$IC(n) := \log_2(B) - H(n), \qquad H(n) := -\sum_b p_{b,n} \log_2 p_{b,n} \tag{1}$$

B here represents the number of possible symbols ($B = 4$ corresponding to 4 nucleotides for DNA sequences and $B = 20$ for protein amino acids). $H(n)$ is the Shannon entropy at position n with $p_{b,n}$ being the probability of symbol b at position n. The inherent assumption with this definition of information content is that all symbols are a priori considered to be equally likely at each position, which is not always true. In case of DNA sequences, some genomes have less GC content (S. Cerevisiae - 38%, Plasmodium falciparum - 19%) compared to humans (41%) and hence are expected to have lower probability of G and C in their background probability compared to humans. For some plants, the background probability is very non-uniform - for example, *Actinidia chinensis* has a background probability of $q = (q_A, q_C, q_G, q_T) = (0.3141, 0.1859, 0.1859, 0.3141)$. When the background probability is not uniform, Kullback-Leibler divergence can be used for determining stack heights.

$$IC(n) := \sum_b p_{b,n} \log_2 \frac{p_{b,n}}{q_{b,n}} \tag{2}$$

where $q_{b,n}$ represents the background probability of base b at position n. *Logolas* allows the user to choose a probability matrix $Q = ((q_{b,n}))$ as background, where the entry at (b, n) is $q_{b,n}$. Alternatively, one can also choose a vector $q = (q_b : b = 1, 2, \ldots, B)$ when the background probability is same at each position, i.e. $q_{b,n} = q_b$ for all n, as proposed by Stormo [17].

A comparison of standard logo plot under uniform background assumption and the species based background information (based on the GC content of the species) for a plant transcription factor Achn021211 in *Actinidia chinensis* is presented in **Supplementary Figure 1**.

Illustration of Sparse Logo Representation

Sequence logos based on information content typically highlight the enriched bases. But in some cases, the bases that are depleted are of greater interest. We present an illustration of the benefits of the sparse logo representation in Figure 1. Say for a specific position, the relative frequencies are $p = (p_A, p_C, p_G, p_T) = (0.33, 0.33, 0.33, 0.01)$ (panel (a)). A standard logo will show three equally high symbols A, C, G stacked vertically with T at the bottom having negligible height (see panel (b)). An alternative parsimonious and probably more meaningful representation is to show the depletion of T instead, since that is the base that has changed from normal while the other bases have remained as is (panel (c)). This problem is further aggravated when a position with depletion is surrounded by highly enriched bases (panel (d)). The information content being biased towards enrichment leads to high stack heights for the enriched bases, which makes the depletion hard to see (panel (e)). Sparse Logo representation shows the enrichment of bases along the positive Y axis and the depletion of bases along the negative Y axis, thereby providing a more accurate yet parsimonious depiction of the sequence logo (panel (f)).

Algorithm for Sparse Logo Representation

There are several options for generating sparse logo representations - *log*, *log-odds*, *ratio* and also information content counterparts of these options, namely *ic-log*, *ic-log-odds*, *ic-ratio*. We discuss the algorithm for computing the heights of enrichment and/or depletion of bases in a sparse logo representation for each of the eight options listed above.

Let $p_n = (p_{n1}, p_{n2}, \ldots, p_{nB})$ be the position weights of the symbols at position n and $q_n = (q_{n1}, q_{n2}, \ldots, q_{nB})$ be the background probability of symbols at position n. Typically we encounter q_n to be same for all positions n ($q_n \equiv q$).

We first define a score vector f_n for each n. Each option varies in its definition of f_n.

- *log* approach

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

- *log-odds* approach

$$f_{nb} = \log_2 \frac{p_{nb}/(1 - p_{nb}) + \epsilon}{q_{nb}/(1 - q_{nb}) + \epsilon} - median \left(\left\{ \log_2 \frac{p_{nb}/(1 - p_{nb}) + \epsilon}{q_{nb}/(1 - q_{nb}) + \epsilon} : b = 1, 2, \ldots, B \right\} \right) \tag{4}$$

- *ratio* approach

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

We next compute $f^+_{nb} = f_{nb}\mathbf{I}(f_{nb} \geq 0)$ 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

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

For each position n, we plot the f^{+}_{nb} values along the positive Y axis and the f^{-}_{nb} values along the negative Y axis. The above formulation ensures that for a base b, one of f^{+}_{nb} and f^{-}_{nb} is zero. For a base b enriched at position n, f^{+}_{nb} value will be large resulting in large size of the symbol for base b in the positive Y axis of the logo plot. For a base b depleted at position n, f^{-}_{nb} value will be large resulting large size of the symbol for the base b in the negative Y axis of the logo plot

Illustration of dash

Both the sequence logo and sparse logo representations are determined by the position weight matrix (PWM) and do not account for the underlying frequency scale of the position frequency matrix (PFM) from which position weights are obtained. Suppose that the background probability is equal for all 4 bases A, C, G, T and in one case, the positional frequencies at a particular positions are (6, 1, 2, 1). In another case, say the positional frequencies at a position are (600, 100, 200, 100). Both these positional frequency vectors will have largely similar representation in the logo plot as they have almost similar positional weights, except for the pseudo-count adjustment. However, in the first case, the total frequency of aligned sequences is only 10 while it is 1000 in the second case. In such a scenario, it is reasonable to shrink the estimated position weights more strongly towards the background probability in the first case compared to the second. At the same time, one would want to determine the amount of shrinkage adaptively. Here we propose an approach called Dirichlet Adaptive Shrinkage (dash) that automatically learns the degree of scaling of position weights for each position based on the underlying scale of the frequencies. This approach is based on the adaptive shrinkage (ash) method due to Stephens (2016) [18] for modeling false discovery rates.

Modeling Framework of dash

We explain the modeling framework of Dirichlet Adaptive shrinkage (dash) for a generic compositional data, including the position frequency matrices for DNA, RNA and proteins.

Suppose we have observe the counts of the constituents for L categories for N compositional samples. For a transcription factor, N would represent the number of binding site positions and L would be equal to 4 corresponding to the bases A, C, G and T.We model these compositional counts vector for each sample n as follows

$$c_n = (c_{n1}, c_{n2}, \cdots, c_{nL}) \sim Mult\left(c_{n+} : p_{n1}, p_{n2}, \cdots, p_{nL}\right), \quad (7)$$

where $n = 1, 2, ..., N$, c_{n+} is the total frequency of the constituents observed for the nth sample and p_{nl}, $l = 1, 2, ..., L$, represents the compositional probabilities such that

$$p_{nl} >= 0 \qquad \sum_{l=1}^{L} p_{nl} = 1. \tag{8}$$

In the *dash* model, we assume a mixture Dirichlet distribution for $p_n = (p_{n1}, p_{n2}, \cdots, p_{nL})$.

$$(p_{n1}, p_{n2}, \cdots, p_{nL}) \sim \sum_{k=1}^{K} \pi_k Dir\left(\alpha_k \mu_1, \alpha_k \mu_2, \cdots, \alpha_k \mu_L\right), \tag{9}$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_L)$ is the known background mean probability ($\mu_l \geq 0$, $\sum_{l=1}^{L} \mu_l = 1$) and $\alpha_k (> 0)$ determines the scale of the concentration parameter for the k th component Dirichlet distribution. The details of model estimation and model configuration are described under Supplementary Methods.

R package

Logolas is available as an R package on Bioconductor (https://bioconductor.org/packages/release/bioc/html/Logolas.html) and is also under active development on Github (https://github.com/kkdey/Logolas). For drawing the symbols, *Logolas* builds on the skeleton used by *seqLogo* [8] to create logos for all alphabets, numerics, punctuations etc and combine them to form strings. *Logolas* also provides an easy interface for the user to create symbolic logo representations of new characters and even add them to strings. For string logo plots, Logolas provides three different color palettes - string-specific coloring (see Fig 4a), character specific coloring (see Fig 3a) and column or position specific coloring (see Fig 4b). The two core functions of Logolas - *logomaker* and *nlogomaker* plot the standard logo and the sparse representations of the logo on a PFM or a PWM matrix respectively. The function *dash* performs Dirichlet Adaptive Shrinkage on a PFM matrix. Logolas also allows the user to use different fill and border styles for enriched and depleted symbols (see **Supplementary Figure 8**) and determine stack heights in multiple ways - using standard Shannon entropy based information content or Renyi entropy based information content at different scales or just based on relative frequencies as in a stacked bar chart (see **Supplementary Figure 9**). Logolas also enables plotting logos in multiple panels in the graphics window and combining logo plots with external ggplot2 graphics.

Results

Sequence logos have been used extensively in visualizing transcription factor binding motifs (TFBSs). Typically such representations tend to highlight the enrichment of bases at different positions of the sequence, which is indeed the more common feature. However some transcription factors tend to show signals of depletion of bases at specific positions. In Figure 2, we present the standard logo and the sparse logo representations (log approach) of the Early B cell factor 1 disc 1 (EBF1-disc1) transcription factor. Based on the sequence logo, EBF1-disc1 seems to recognize a palindromic sequence of bases TCCCg - cGGGA to get activated, where lowercase letters are used to represent depletion and uppercase case letters stand for enrichment. Not only does the sequence motif show strong depletion signals of bases G and C in the center but the depletion is also a part of the palindrome. Note that this depletion signal is hard to see in the standard logo plot (panel (a)) because it is flanked by strong enrichments, but the sparse logo representation (panel (b)) highlights it clearly. In **Supplementary Figure 2**, we present the sequence logos of all the members of the EBF1 family and the signal depletion of G and C at the center of the palindromic sequence is also observed in EBF1 - known3 and EBF1 - known4, besides EBF1-disc1 transcription factor.

As described in the previous section, the stack composition and stack heights for a sparse logo representation can be defined in a number of ways. In Figure 3, we present the different sparse logo representations of the binding sequence of the Effector domain protein. We compare the sparse logo representations in Figure 2 with the PSSM profile representation of the same protein in **Supplementary Figure 3**. It is evident that the sparse logo representation is much more parsimonious and interpretable than the PSSM representation. 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/` [19] [20]. In **Supplementary Figure 4**, we compare how the depletion signal in the middle of its binding sequence motif of EBF1-disc1 transcription factor gets differentially highlighted by the different sparse logo representations.

In Figure 4, we present two applications of string logo representations of *Logolas*. Koch *et al* (2007) [21] recorded the number of different types of histone modifications at sites that overlap with an intergenic sequence, intron, exon, gene start and gene end for a lymphoblastoid cell line, GM06990 using ChIP-chip data. Figure 4 panel (a) presents the standard logo plot representation of the histone marks compositional data, where histone mark names - for example H3K4ME1 - are alphanumeric strings and therefore, are ideal fit for string logo representation. In Figure 4 panel (b), we present the logo plot representation of the Himalayan bird species abundance compositional data due to White *et al*, restricted to three regions of the Himalayas. Here we use the bird species names, which are strings, as the symbols in the logo plot.

One important application of both the string logo feature and the sparse logo representation of *Logolas* is in visualizing mutation signature profiles. Each mutation signature is usually represented by the type of mutation at the center ($C \rightarrow T$, $C \rightarrow A$, $C \rightarrow G$, $T \rightarrow A$, $T \rightarrow C$, $T \rightarrow G$) flanked by bases to the left and right. The string logo feature is used here to plot the symbols for the mutation types ($X \rightarrow Y$). In Figure 5, we present the sparse logo representation (ratio) of 24 mutational signature profiles across different tissues, analyzed using mutational data from 7042 cancers by Alexandrov et al (2013) [22]. This plot is a *Logolas* version of the Supplementary Figure 3 plot from Shiraishi et al (2015) [23]. Shiraishi et al [23] proposed a grade of membership model to show that each mutational signature arises from one of K distinct signature profiles, where K was chosen to be 27. In **Supplementary Figure 5**, we present the sparse logo representations of all the 27 signature profiles. In **Supplementary Figure 6**, we compare the *Logolas* representation of signature profile 16 with the *pmsignature* visual representation in Shiraishi et al [23] and it is evident that the logo plot representation depicts the overall features of the mutation signature more clearly. For example, 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 logo plot representation compared to the *pmsignature* plot.

Many current transcription factor databases for animal and plant transcription factors like HOCOMOCO (`http://hocomoco11.autosome.ru/`) [24] [25], PlantTFDB v4.0 (`http://caps.ncbs.res.in/3pfdb/`) [26], [27] [28], JASPAR (`http://jaspar.genereg.net/`) [29], [30] etc store the transcription factor binding site models in terms of positional frequencies (PFM). In some cases, the number of matched sequences used to generate the PFM matrix for a transcription factor is quite small. Some transcription factors with low frequency scales in the total number of matched sequences for generating the PFM matrix are `EGR3_HUMAN.H10MO.D` (frequency scale = 8), `EPAS1_HUMAN.H10MO.D` (frequency scale = 12), `EGR4_HUMAN.H10MO.D` (frequency scale = 6), `ERF_HUMAN.H10MO.D`(frequency scale = 7) etc (`http://hocomoco11.autosome.ru/final_bundle/hocomoco11/core/HUMAN/mono/HOCOMOCOv11_core_HUMAN_mono_jaspar_format.txt`).

This emphasizes the need to scale the position weights appropriately taking the frequency scale into account, which is accounted for in *dash*. In order to validate the dash approach, we first take the positional frequency matrix of the Aryl hydrocarbon receptor (`AHR_HUMAN.H11MO.0.B` in HOCOMOCO), which has a frequency scale of 154, a number large enough to trust the position weights computed from it. We use the PWM matrix for this transcription factor as the base and generate two random subsamples of sequences from this position weight matrix - one of size 5 and the other of size 30. In Figure 6, we show how the *dash* scaling produces a sequence logo plot (panel (c)) which is a closer approximation to the original logo plot (panel (a)), compared to the pre-dash scaled version (panel (b)), in particular for the first subsample of size 5.

So far, we have considered applying *dash* each PFM separately. However, owing to the small number of positions observed in general in a TFBS model for a particular transcription factor, the amount of shrinkage the method learns from data is limited. However, one way to bypass this problem is to combine the position frequency data across all positions for all transcription factors in a species and run *dash* on the pooled data to learn the mixture proportions in Eqn 9 and then update the position weights of each transcription factor based on the fitted model from the pooled data. We apply this combined *dash* approach on 290 transcription factors of *Actinidia chinensis*, with the data collected from PlantTFDB [27] [26] [28]. The position weight scaling for each position of the SBP family protein transcription factor *Achn185971* with combined *dash* and uncombined *dash*, corresponding to the specified GC content (37%) based background probability for *Actinidia chinensis*, is compared in **Supplementary Figure 7**.

Discussion

The *Logolas* package builds on top of existing softwares like *seqLogo*, *WebLogo* [9] [8] to introduce novel features like the sparse logo representation, string logos and the adaptive scaling of positional weights based on the frequency scale in *dash*. The aim here is to not only improve the visual informativeness of a logo plot but also to make logo plots a more generic tool applicable in viewing general compositional data beyond the DNA, RNA and protein sequence position weight matrices - as depicted in our example applications like mutational signature profile, histone marks composition visualization etc.

A string logo can be viewed as an alternative visual representation to a stacked bar chart where each color stack is plotted instead as a symbol logo. The string logo plot is more preferable to stacked bar chart representation when the number of columns or bars are relatively small and the number of stacks per bar are moderate or large, in which case the string symbols look more appealing than colors and color legends. However, when the number of columns or bars are very large compared to the number of stacks per bar, as in case of the STRUCTURE plot representation [31] [32], colors and color legends are more preferred to symbols.

We have suggested a number of options for generating the sparse logo representation of a position weight matrix. 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. The *ratio* approach seems to be a bit more balanced in representing both enrichment and depletion. Also, we have mostly found the *log* and *log-odds* (also *ic-log* and *ic-log-odds*) to be largely equivalent in its representation. We suggest the user to present the sparse logo representations for *log* (or *log -odds*), *ic-log* (or *ic-log-odds*), *ratio* and *ic-ratio* methods for a fair comparison.

Various softwares focus on downstream analysis of the position weight matrix by using it for motif discovery and motif matching (R package *motifcounter* [33]), comparing motif patterns across multiple motifs (R packages *motifStack* [12] and DiffLogo [13]), regulatory SNP detection for testing for transcription factor binding affinity (R package *atSNP* [34]) etc. *Logolas* provides the heights of the bases along the positive and negative Y-axes, which can be used as scores for many of these downstream analysis. These scores will contain signals for both enrichment and depletion and may improve the accuracy of some of these motif based analyses. This is one of the future directions we would like to pursue.

Conclusion

We present a new R/Bioconductor package named *Logolas*, an easy to use and flexible tool which opens some new avenues in logo visualization. We propose a new parsimonious representation of logos aimed at highlighting both enrichment and depletion of symbols or bases at different positions, unlike the standard information content based approach which is biased towards the enrichment of bases. *Logolas* also allows the user to plot strings as symbols in logo plots and we show various applications of such string logos in viewing mutational signature profile, histone marks composition and ecological composition data. We also propose a technique called Dirichlet Adaptive Shrinkage (*dash*) that adaptively scales the positional weights of a PWM matrix based on the underlying frequency scales thereby providing more robust logo plots.

Logolas 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 website for the *Logolas* project is hosted on https://kkdey.github.io/Logolas-pages/ and all the codes for plotting the figures in this paper are available on https://kkdey.github.io/Logolas-pages/Paper.

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.

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. Tan, G., Lenhard, B.: Tfbstools: an r/bioconductor package for transcription factor binding site analysis. Bioinformatics **32(10)**, 1555–1556 (2016)
3. Bryson, S., Julien, J.P., Hynes, R.C., Pai, E.F.: Crystallographic definition of the epitope promiscuity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2f5: Vaccine design implications. Journal of Virology **83 (22)**, 11862–11875 (2009)
4. Mahrus, S., Trinidad, J.C., Barkan, D.T., Sali, A., Burlingame, A.L., Wells, J.A.: Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein n-termini. Cell **134 (5)**, 866–876 (2008)
5. Emmert, S.: The human xpg gene: gene architecture, alternative splicing and single nucleotide polymorphisms. Nucleic Acids Research **29(7)**, 1443–1452 (2001)
6. Henikoff, S., Henikoff, J.G., Alford, W.J., Pietrokovski, S.: Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene **163 (2)**, 17–26 (1995)
7. Henikoff, S., Henikoff, J.G., Pietrokovski, S.: Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics **15 (6)**, 471–479 (1999)
8. Bembom, O.: seqlogo: Sequence logos for dna sequence alignments. R package version 1.42.0
9. Crooks, G.E.: Weblogo: A sequence logo generator. Genome Research **14 (6)**, 1188–1190 (2004)
10. 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)
11. Wagih, O.: Rweblogo: plotting custom sequence logos. R package version 1.0.3
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. 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)
14. Stojnic, R., Diez, D.: Pwmenrich: Pwm enrichment analysis (2015). R package version 4.10.0
15. Coalert, N., Helsens, K., Martens, L., Vandekerckhove, J., Gevaert, K.: Improved visualization of protein consensus sequences by icelogo. Nature Methods **6**, 786–787 (2009)
16. Wagih, O.: ggseqlogo: a versatile r package for drawing sequence logos. Bioinformatics **btx469** (2017)
17. Stormo, G.D.: Dna binding sites: representation and discovery. Bioinformatics **16 (1)**, 16–23 (2000)
18. Stephens, M.: False discovery rates: a new deal. Biostatistics **18 (2)**, 275–294 (2016)
19. 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)
20. 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)
21. Koch, C.M., et al.: The landscape of histone modifications across 1in five human cell lines. Genome Research **17(6)**, 691–707 (2007)
22. 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)
23. 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)
24. Kulakovskiy, I.V.e.a.: Hocomoco: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Research **41**, 195–202 (2013)
25. Kulakovskiy, I.V.e.a.: Hocomoco: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Research **44**, 116–125 (2016)
26. Jin, J.P., He, K., Tang, X., Li, Z., Lv, L., , Zhao, Y., Luo, J.C., Gao, G.: An arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. Molecular Biology and Evolution **32(7)** (2015)
27. Jin, J.P., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J.C., Gao, G.: Planttfdb 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research **45(D1)** (2017)
28. Jin, J.P., Zhang, h., Kong, L., Gao, G., Luo, J.C.: Planttfdb 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research **42(D1)** (2014)
29. Sandelin, A., Wynand, A., Engstrom, P., W.W., W., Lenhard, B.: Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research **32 (Database issue)**, 91–94 (2004)
30. Mathelier, A., et al.: Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiless. Nucleic Acids Research **42 (D1)**, 142–147 (2014)
31. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., A., Z.L., Feldman, M.W.: The genetic structure of human populations. Science **298** (2002)
32. Dey, K.K., Hsiao, C.J., M., S.: Visualizing the structure of rna-seq expression data using grade of membership models. PLOS Genetics **13(3)** (2017)
33. Kopp, W.: motifcounter: R package for analysing tfbss in dna sequences. (2017). R package version 1.0.0
34. Zuo, C., Sunyoung, S., S., K.: atsnp: transcription factor binding affinity testing for regulatory snp detection. Bioinformatics **31(20)** (2015)
35. Kheradpour, P., Kellis, M.: Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. Nucleic Acids Research, 1–12 (2013)

Figures

Figure 1 Illustration of the sparse logo representation. We present an illustration of how sparse logo representation accounts for depletion signal and provides a more informative visualization of the sequence motif. In panel (a), we present a positional weight vector with a depletion of the base T while the other bases have similar position weights. In panel (b), we present the standard logo representation of this positional vector that highlights the abundance of A, C and G at this position. In panel (c), we present the sparse logo representation (ratio approach) that highlights the depletion of T instead. In panel (d), we present a position weight matrix with the position weight vector in panel (a) at the second position flanked by enrichments around it. In panel (e), we present the corresponding standard logo plot representation of the PWM matrix at panel (d). The signal at the second position gets swamped by the bias towards enrichment signals flanking it. In panel (f), we present the sparse logo representation of the PWM matrix in (d), where both the enrichment signals as well as the depletion signal at position 2 are clearly observed.

Figure 2 Application of sparse logo in detecting depletion patterns in Transcription Factor Binding sites (TFBS). We present the standard logo and the sparse logo representation of a transcription factor EBF1-disc1. The logo standard logo plot in panel (a) seems to indicate that the transcription factor binds in a dimerized form to its binding site. However, it fails to capture the depletion of G and C in the two positions in the middle of the dimer, which is apparently captured by the sparse logo representation in panel (b). The stack heights in the sparse logo representation in this plot has been determined by the *log* approach. The PWM data for EBF1-disc1, computed from the ENCODE TF Chip-seq datasets, is hosted on the webpage http://compbio.mit.edu/encode-motifs/ [35]

Figure 3 Various sparse logo representations of protein sequence motif. The sparse logo representation under various stack height and stack composition methods - *log*, *log-odds*, *ratio*, *ic-log*, *ic-log-odds* and *ic-ratio* for the Bacterial transcription activator, effector binding domain protein PF06445 (motif 4, Start=153 Length=8). The data is fetched from the 3PFDB website http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/get_entry.cgi?id=PF06445.

Figure 4 Example applications of string logo plots. We present two example biological applications of string logo plots. In panel (a), we present the logo plot representation of the composition of histone modification types that overlap with an intergenic region, intron, exon, gene start or gene end for the lymphoblastoid cell line GM06990 as reported in Koch et al 2007 [21]. In panel (b), we present the logo plot representation of the abundance compositions of bird species families in three different regions of the Himalayas - colored red, green and blue. The data has been taken from White et al 2017 [ref].

Figure 5 Logolas representation of cancer mutational signature profiles 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) [22]. Each mutational signature profile has the mutation type at the center and is flanked by two bases to the left and two bases to the right.

Figure 6 Subsampling experiment to validate the performance of Dirichlet Adaptive Shrinkage (dash). In panel (a), we present the standard logo plot representation of the position frequency matrix (PFM) of the Aryl hydrocardon receptor. This PFM matrix is then used to define the position weights and two random subsamples, A of size 5 and B of size 30 are generated from this position weight matrix. Panels (b) and (c) demonstrate the logo plot representation estimated position weight matrix from the 5 symbols in the subsample A before applying dash and post dash scaling. Panels (d) and (e) show the same results for subsample B, with size 30. We notice that for subsample A case, the dash scaled PWM in panel (c) is a closer approximation of the original PWM in panel (a) compared to the pre-dash version in panel (b). However, the effect of dash is comparatively lower for subsample B, since the subsample size (30) is much larger and both the pre-dash and post-dash PWMs of subsampled symbols in panels (d) and (e) are good approximations to the original PWM.

Supplementary Figures

S1 Fig. **Standard Logo plot comparison under uniform and non-uniform background base probabilities .** Logo plot representation of the plant transcription factor Achn021211 (MYB family protein) in *Actinidia chinensis*. The background probability for this species based on GC content is $q = (q_A, q_C, q_G, q_T) = (0.3141, 0.1859, 0.1859, 0.3141)$. The PWM matrix is obtained from PlantTFDB site (http://planttfdb.cbi.pku.edu.cn/tf.php?sp=Ach&did=Achn021211). In panel (a), we present the standard logo plot of the PWM matrix with uniform background for all 4 bases. In panel (b), we present the standard logo plot with the above specified background probability.

S2 Fig. **Sparse logo representation of the members of the EBF1 family of transcription factors**: We present the sparse logo 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/ [35].

S3 Fig. **PSSM logo plot for protein sequence motif**: The logo representation of the position specific scoring matrix (PSSM) for the Bacterial transcription activator, effector binding domain protein PF06445 (motif 4, Start=153 Length=8). The data is fetched from the 3PFDB website http://caps.ncbs.res.in/cgi-bin/mini/databases/3pfdb/get_entry.cgi?id=PF06445.

S4 Fig. **Various approaches of sparse logo representations for a transcription factor** : The sparse logo representation under various stack height and stack composition methods - *log*, *log-odds*, *ratio*, *ic-log*, *ic-log-odds* and *ic-ratio* for the Early B cell factor 1 disc 1 (EBF1-disc1) transcription factor. The data is fetched from the CompBio website of MIT http://compbio.mit.edu/encode-motifs/.

S5 Fig. **Logolas plots for the mutational signature profiles for 27 clusters in Shiraishi et al (2015)**: We present the sparse logo representations (ratio) method for the 27 cluster signature profiles obtained from fitting a grade of membership model on the cancer mutational signature data across 30 cancer types by Shiraishi et al (2015) [23]. This plot is an alternative logo plot based representation of Figure 4 in Shiraishi et al (2015) [23].

S6 Fig. **Comparison of Logolas sparse logo plot with pmsignature representation for cancer mutation signatures**: We compare the sparse logo plot representation and the pmsignature representation due to Shiraishi et al (2015) [23] for mutation signature profile of cluster 16 in their paper. 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, the logo plot representation shows the depletion of G at the right flanking base more clearly than the pmsignature plot. Also, overall, the logo plot representation is more interpretable and visually appealing in highlighting the mutation signature patterns compared to the pmsignature plot.

S7 Fig. **Dirichlet Adaptive Shrinkage (dash) training on combined transcription factor data for a species**: We compare the two versions of Dirichlet Adaptive Shrinkage (dash) applied to the SBP protein transcription factor Achn185791 in *Actinidia chinensis*. In one case, the parameters of the *dash* model are learnt from the positional frequency data from Achn185791, while in the other case, the the parameters are learnt from the pooled positional frequency data across all 290 transcription factors of *Actinidia chinensis*, which we refer to as *combo dash* in this plot. The background probability of the bases for this species are $q = (q_A, q_C, q_G, q_T) = (0.3141, 0.1859, 0.1859, 0.3141)$. The transcription factor data for *Actinidia chinensis* along with the background probability information are derived from the PlantTFDB v4.0 database http://planttfdb.cbi.pku.edu.cn/index.php?sp=Ach.

S8 Fig. **Fill and border styles in Logolas.**: A demonstration of how fill and border styles can be used to distinguish between the enrichment and depletion of symbols at a position in a sparse logo plot.

S9 Fig. **Stack heights in Logolas.** : A demonstration of how stack height for a position in a standard logo plot can be determined in various ways in *Logolas*. In panel (a), the standard Shannon entropy based Information content is used to determine the height of the stack of symbols at each position. For panels (b) and (c), we use Renyi entropy based information content for two levels of tuning paramter α, one when $\alpha = 0.1$ is small and the other when $\alpha = 100$ is large. In panel (d), a relative frequency based stacked bar chart representation using logos is implemented. All these options can be passed as input arguments and control arguments to the *logomaker* functon in *Logolas*.

Supplementary Methods

Dirichlet Adaptive Shrinkage - model estimation

We discuss here the model fit of the Dirichlet Adaptive Shrinkage (*dash*) method. The model estimation steps draw heavily from the model fit in the adaptive shrinkage (ash) method due to Stephens 2016 [18].

We first estimate the mixture proportions π_k in Equation 9 by empirical Bayes approach and then obtain the posterior distribution of p_n. The estimator of p_n is the posterior mean.

Let

$$l_{nk} = \frac{c_{n+}!\Gamma(\alpha_k)}{\Gamma(c_{n+} + \alpha_k)} \prod_{l=1}^{L} \frac{\Gamma(c_{nl} + \alpha_k \mu_l)}{c_{nl}!\Gamma(\alpha_k \mu_l)}. \tag{10}$$

Once can then use EM algorithm or convex programming to estimate the mixture proportions π_k by maximizing the following objective function.

$$\log L(\pi) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k l_{nk} \right) \tag{11}$$

Once π_k is estimated, we define posterior weight of the sample n in the component mixture k to be

$$\omega_{nk} = \frac{\hat{\pi}_k l_{nk}}{\sum_k \hat{\pi}_k l_{nk}}. \tag{12}$$

The posterior of p_n is then

$$f(p_n|\hat{\pi}, c_n) = \sum_{k=1}^{K} \omega_{nk} f_k(p_n|c_n), \tag{13}$$

where $f_k(p_n)$ is the posterior component with prior component equal to the k^{th} component of the dash prior and $f_k(p_n)$ is

$$f_k(p_n|c_n) := Dir\left(c_{n1} + \alpha_k \mu_1, c_{n2} + \alpha_k \mu_2, \cdots, c_{nL} + \alpha_k \mu_L\right). \tag{14}$$

Therefore, the posterior mean of p_n is

$$E(p_n|\hat{\pi}, c_n) := \sum_{k=1}^{K} \omega_{nk} \frac{c_n + \alpha_k \mu}{\sum_l^L (c_{nl} + \alpha_k \mu_l)}, \tag{15}$$

where $\mu = (\mu_1, \mu_2, \cdots, \mu_L)$.

Model configuration
We fix the number of mixture components K and the the concentration values α for each component of the mixture Dirichlet distribution.
Ideally K can be chosen as large as possible ideally, however for computational ease, we restrict K to be 10.
We define the vector of concentrations α for the K th components as follows

$$\alpha = (Inf, 100, 50, 20, 10, 5, 2, 1, 0.1, 0.001) \tag{16}$$

$\alpha_k = Inf$ corresponds to point mass at the background mean μ, $\alpha_k = 1$ corresponds to the most uniform looking Dirichlet distribution and $\alpha_k < 1$ allows for components with concentration of mass at the edges of the simplex. *Logolas* allows the user to specify the choice of K and the concentration vector α as well as the background mean probability μ.