Eurographics Conference on Visualization (EuroVis) (2014)
N. Elmqvist, M. Hlawitschka, and J. Kennedy (Editors)

Redesigning the Sequence Logo with Glyph-based Approaches to Aid Interpretation

Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone, and Min Chen

Oxford e-Research Centre, University of Oxford, UK

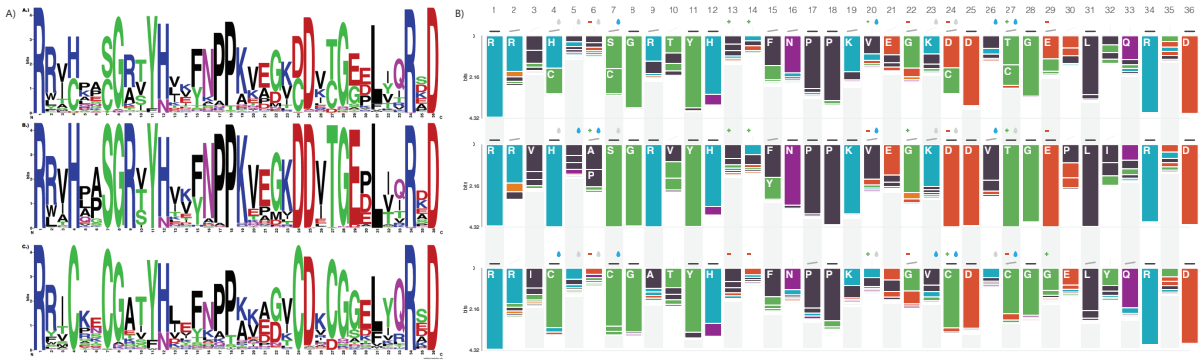

Figure 1: *A) A 'traditional' sequence logo from [Bio13] showing: top - the consensus across 1809 protein sequences; middle - gram negative bacteria; and bottom - gram positive bacteria. B) Our sequence logo visualizing the same data as in A. This approach keeps in place the idea of the original sequence logo but removes the dominance of letters for space filling and adds glyphs to visualize overall properties of the dominating amino acids at a sequence location (e.g. charge and hydropathy).*

Abstract

Sequence logos have been a prominent visualization tool in biological research since their inception two decades ago. Their primary use is in communicating conservation of biological sequences (protein, DNA or RNA), to indicate largest conservation at particular positions - namely places where only ever one or two possible residues (nucleotides or amino acids) are observed. Conservation is indicative of functional importance, as changes, being selected against, reveal a loss of fitness for living organism or cells. Criticism of the sequence logo has long existed, largely directed towards perception problems caused through use of letter height to indicate frequency. Here, we present a solution for use as a static image in publications or interactively on the web to address the reported flaws of the sequence logo. In addition to our improvements, we propose glyph based enhancements, to highlight qualitatively relevant chemical insights resulting from residue substitution between sequences.

1 Introduction

Advances in sequencing technologies have turned a time consuming and expensive process into a quasi main stream commodity. Next generation sequencing allows exploration of the blueprints of life in fascinating new ways, making the genome of thousands of species available to computational analysis. Of particular interest to scientists is the detection of regions of genomes under selective pressure and to ascertain which regions of DNA, RNA and proteins are functionally and/or structurally important across species (e.g., the conserved regions of a protein in mice, human, rat and horse).

Although our ability to process this data has increased, the methods used to visualize and report this conservation in scientific publications has not moved on in over 20 years. This technique, named the named 'sequence logo' [SS90, Sch02]

DOI: 10.2312/eurovisshort.20141159

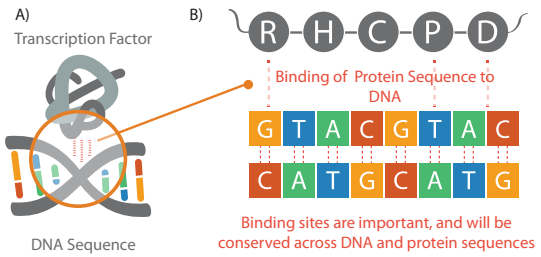

Figure 2: *A) An abstraction of a transcription binding site where a protein binds to a region of DNA. B) Certain residues, amino acids for the protein and DNA nucleotides for DNA are important in creating a bond between the DNA and protein to enable some function to occur.*

is comprised of a series of letters stacked on top of each other. It is these letters and their arrangement which are cause for concern due to their perception problems [Bio13] when too stretched or squashed to interpret. We now introduce a new sequence logo design aimed at addressing their perception issues. This has been implemented in a JavaScript library that can be easily customised and incorporated in to online resources for interactive exploration, or exported as a scalable vector graphic (SVG) for inclusion in traditional journal publications.

2 Background

Transcription Factors are proteins with a central function: they can, with high specificity recognize and bind to DNA regions and trigger transcription of DNA into messenger RNA. Those genomic regions are known as transcription factor binding sites (TFBS) and their identification is essential to the reconstruction of transcriptional regulatory networks of cells. Sequence comparison across organisms can be used to highlight areas of conservation to pinpoint unknown TFBS. Low variability at some position i generally points to a level of functional importance in that region of DNA or protein.

In order to visualize these binding sites, sequence logos were devised as a visualization technique to view a position weight matrix (PWM), which specifies for each position in a sequence the likelihood of observing a particular residue (DNA/RNA nucleotide or amino acid). From this matrix, a simple frequency-based sequence logo can be drawn by iterating through each position i in the sequence. For each position, the height of each letter a can be calculated as $p(a) \times R$ where $p(a)$ is the measured probability of a occurring at position i, and R is the maximum height at position i.

To determine the maximum height at position i, the principal mechanism is to compute its 'information content', a term derived from Claude Shannon's information theory [Sha48]. Information content represents the level of certainty at a specific position i by measuring how much the probability distribution of the detected letters in that position deviates from a uniform distribution, where all valid letters have an equiprobable chance of occurring. Given s valid letters, $a_j, j = 1,2,\ldots,s$, in an alphabet $\mathbb{Z}$ and their probability distribution function, $p(a_j)$, the level of uncertainty can be defined by the following entropy measure:

$$H(\mathbb{Z}) = -\sum_{j=1}^{s} p(a_j) log_2 p(a_j)$$

where we use the base 2 logarithm, and the unit for the uncertainty H is 'bit'. For an equiprobable distribution, $p(a_j) = 1/s$ for all letters in the alphabet $\mathbb{Z}$. The above entropy measure yields the maximum uncertainty (commonly referred to as the maximum entropy), $H_{max}(s) = log_2(s)$.

For example, for 20 types of amino acids, we have $H_{max}(20) = 4.32$ bits, and for 4 types of nucleic acids, $H_{max}(4) = 2$ bits.

Since it is unlikely that there is an equiprobable distribution of letters at each position i, the actual uncertainty is usually less than H_{max}. Hence the difference between the maximum uncertainty and the measured uncertainty can be defined as the certainty, which is visually encoded as the total height R_i for all letters at position i.

A high stack indicates a high level of certainty with a strong preference for one or at most two nucleotides. A low stack implies a low level of certainty with many possibilities. The idea behind this visual design is that the highly conserved positions 'pop out' more due to their greater height.

3 Related Work

There have been a few attempts to redesign the sequence logo in the past, with most having focused on modifying how such letters are positioned on the y-axis with Kullback Leibler [KL51], position-specific scoring matrix (PSSM) [FZL*04], and berry logos [BHLB06, Ber13]. These in part can alleviate some of the perception issues with the sequence logo, though all continue to use letters to represent size apart from LogoBar by Perez *et al* [PBKB06]. We extend on the work by Perez through support for DNA/RNA sequences, glyphs, an evaluation and a web-based implementation.

4 Motivation

The overall goal of this work is to improve the sequence logo to address their perception issues [Bio13]. Our approach is a three stage process with a domain-expert always in the loop. Those stages are:

1. improving the current *sequence logo* design to address perceptual issues caused by: a) use of letter size to represent value; and b) the arrangement of letters on the y-axis which can make residues look more conserved than others based on the number of letters it is stacked upon;
2. adding *glyphs*, which are visual objects used to depict attributes of a data record [BKC*12] to annotate positions

Figure 3: *The amount of ink used in a letter can influence the perceived 'weight' of the letter. Within a normalized plotting, 'W' and 'M' take up 3 times as much ink as 'I', 'J' and 'L', and twice as much as 'S'.*

when comparing within sets of sequence logos, for instance to highlight qualitative changes of residue; and

3. an Ã§ of the redesigned sequence logo with both advanced and 'naive' sequence logo users to ascertain the benefits of the new design.

5 Design

5.1 Sequence Logo

The frailties of the original sequence logo reside in the following problems:

1. *Using letter size to show value* — letters have an undesirable property in that they are of differing densities. We have quantified this effect (see Figure 3) and found that and 'R', 'H' and 'W' take up to about three times as much ink as 'I' ,'J' or 'L' for example. When used for space filling, this ultimately leads to perception issues as denser letters will be more visible. Take Figure 1 - at position 31, 'L' (Leucine) is the dominant amino acid, however it is harder to see compared with 'R' or 'D' at positions 34 and 36 respectively. Use of letters also means that when there are many positions to display, letters become 'squashed' so it is often impossible to read them;
2. *Placement of the most dominant letter on top of the less dominant letters, leading to possible misinterpretation of the size of a letter depending on where it is placed on the y axis* — this is particularly evident in Figure 1 at positions 2 and 4 of the top sequence logo where we may compare heights between 'R' (Arginine) and 'H' (Histidine). In plot A, 'H' is positioned higher in the chart due to the letters below it. This gives the impression that the conservation of 'H' at position 4 is greater than the conservation of 'R' at position 2. In reality, 'R' is more conserved at position 2, which is evident in plot B.

We propose a design that retains the ideas of the original sequence logo, so as to ease uptake of a new representation, whilst subtly overcoming its perception issues. Our design, shown in Figure 4 uses filled bars to represent size in place of the letters similar. The top 'residues' (amino acids or nucleotides) are positioned at the top or bottom of the plot so that the conserved sequence can be read more easily than in the original logo. The colours of the bars are a function of the type of amino acid, however these are not fixed and can be changed depending on how the user wishes to visually group residues.

Figure 4: *Redesign of the sequence logo layout from the original (left) to the new version.*

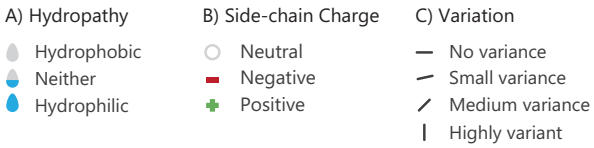

Figure 5: *Glyphs indicating dominating properties of residues at each position. A) Hydropathy (relevant for amino acids). B) Side-chain charge (relevant for amino acids only). C) Variation: indicated using 'GestaltLines'.*

5.2 Glyphs

Amino acid residue side-chain charge and hydrophobicity strongly affect protein folding and its final 3D conformation, so any significant change in those qualities may result in a significant functional change. Therefore, providing a visual aid to remind users of those dimensions presents opportunities to enhance interpretation.

Additionally, we experiment with another glyph-based approach, named 'Gestalt Lines' [BNRS13] to provide an additional indicator of variance.

The glyphs will be positioned above each position to give an overall impression of whether or not there are any dominating characteristics of the amino acids present at a position. These glyphs are shown in Figure 5. Glyphs for hydropathy and side-chain charge are only present in amino acid sequence logos.

6 Implementation

We have implemented an open source JavaScript library (https://github.com/ISA-tools/SequenceLogoVis), built using RaphaelJS [Rap14]. This library renders protein, DNA or RNA sequences as seen in Figures 1B and 6 and a set of parameters allows customization of the visualization, as seen in Figure 6. The web version supports interactivity, providing detail on demand [Shn96] for each position to show the distribution of letters at each position in the sequence alongside contextual information about the amino acids or nucleotides. Additionally, the sequence logo can be saved as a scalable vector graphic (SVG) for inclusion in publications.

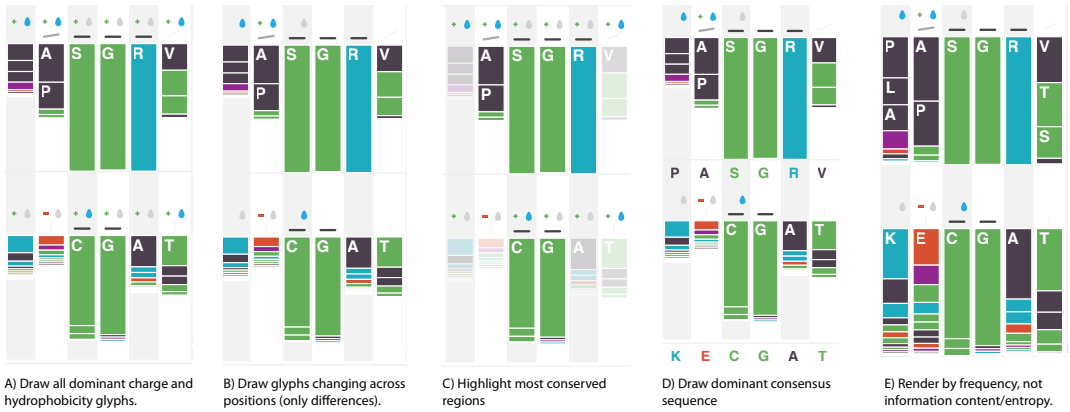

Figure 6: *The rendering library contains many options allowing users to configure the sequence logo.*

7 Evaluation

In order to assess if our changes helped biologists read sequence logos, we devised a survey focused on evaluating the ability to compare the size of letters in the old version of the sequence logo (see Figure 1A) and the new version (Figure 1B). We used the same data to create the two versions of the sequence logo and asked participants to determine which letter between two was the largest. Our hypothesis, as stated earlier, is that it is harder to compare the letters than it is to compare blocks. 41 scientists (15 bioinformaticians, 23 biologists and 3 computer scientists) with varying levels of familiarity with sequence logos, took part in the survey.

Figure 7 shows survey results (upper part) and test images (lower part). In three out of the four questions given to users, a higher number of correct answers were recorded with the new sequence logo. The gain was significant for questions 2 and 3, with up to twice as many correct answers with the new representation. The exception was in question four, where there was a small (8%) advantage gained by using the original version. The greater density of 'G' over 'L' may have lead to more people choosing correctly by chance, explaining the observation. More thorough experimentation would be needed to validate this hypothesis however.

The feedback from users regarding the glyphs was also largely positive with: 80% of respondents agreeing that showing hydropathy was useful; 83% agreeing that showing side-chain charge was useful; and 59% indicating that the variance glyph using GestaltLines was useful. The feedback has led to the removal of the GestaltLines since the level of variance can be adequately determined using bar height.

The approval of our redesign was also measured with users asked to give their preference between Figure 1A and B. 95% of respondents said that they preferred the new representation over the original, citing amongst others, 'cleanliness' and 'clarity' as the major factor in their decision.

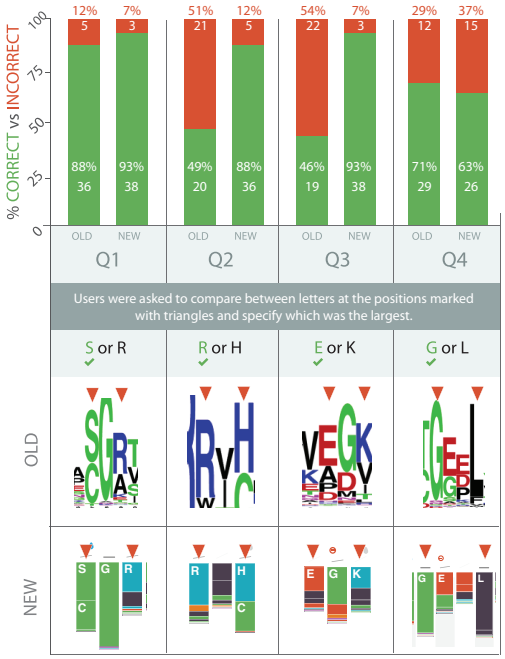

Figure 7: *Evaluation responses showing the ability to discriminate between conservation levels in the two designs of the sequence logo, old/original and new.*

8 Conclusion

We have presented a new design for the sequence logo that incorporates glyph-based techniques to aid interpretation. We have also provided an implementation of this new logo that can be immediately incorporated in to the workflows of scientists for interactive use or inclusion in publications. Our usability tests showed that users generally found consensus sequence reading tasks easier with the new sequence logo. Users, on the whole, agreed that the new representation did a better job of improving the display of salient information.

References

[Ber13] BERRYLOGO: https://github.com/leipzig/berrylogo, 2013. 2

[BHLB06] BERRY C., HANNENHALLI S., LEIPZIG J., BUSHMAN F. D.: Selection of target sites for mobile dna integration in the human genome. *PLoS computational biology* 2, 11 (2006), e157. 2

[Bio13] BIOVIS: http://www.biovis.net/year/2013/info/redesign-contest, 2013. 1, 2

[BKC*12] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics 2013-State of the Art Reports* (2012), The Eurographics Association, pp. 39–63. 2

[BNRS13] BRANDES U., NICK B., ROCKSTROH B., STEFFEN A.: Gestaltlines. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 171–180. 3

[FZL*04] FUJII K., ZHU G., LIU Y., HALLAM J., CHEN L., HERRERO J., SHAW S.: Kinase peptide specificity: improved determination and relevance to protein phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America 101*, 38 (2004), 13744–13749. 2

[KL51] KULLBACK S., LEIBLER R. A.: On information and sufficiency. *The Annals of Mathematical Statistics* (1951), 79–86. 2

[PBKB06] PÉREZ-BERCOFF Å., KOCH J., BÜRGLIN T. R.: Logobar: bar graph visualization of protein logos with gaps. *Bioinformatics 22*, 1 (2006), 112–114. 2

[Rap14] RAPHAELJS: http://raphaeljs.com/, 2014. 3

[Sch02] SCHNEIDER T. D.: Consensus sequence zen. *Appl Bioinformatics. 1* (2002), 111–119. 1

[Sha48] SHANNON C.: A mathematical theory of communication. *Bell Systems Technical Journal. 27* (1948), 379–423. 2

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on* (1996), IEEE, pp. 336–343. 3

[SS90] SCHNEIDER T. D., STEPHENS R. M.: Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res. 18* (1990), 6097–6100. 1