UNIVERSIDADE NOVA DE LISBOA

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa

Visual representations of consensus sequences

Gil Oliveira
MSc in Biotechnology student

Submitted in partial requirements of the course of
Applied Bioinformatics

Caparica, Portugal

April 2020

Abstract

Consensus sequences are very common themes in Molecular Biology and Bioinformatics, allowing for evolution and molecular interaction studies, to name a few examples. There's a big need for accurate and intuitive sequence visualisation techniques, as they can portray much more relevant information than simply typing the sequence letters. The Sequence Logo was the first sequence visualisation technique, and is the most used nowadays, but new paradigms, such as the ProfileGrids, have been appearing to try to compensate for the limitations of the Sequence Logo. In this paper, I compare the most relevant sequence visualisation tools, evaluate their upsides and downsides and try to match them with their most appropriate use cases.

1 Introduction

Consensus sequences are sequences of nucleotides (on DNA and RNA) or amino acids (in protein), which are comprised of the most commonly encountered letters at that postion [6]. These sequences are generally associated with inter- or intramolecular interactions [3]. A prime example is the Shine-Dalgarno (SD) sequence in prokaryotes, which is involved in the binding of the ribossome to the mRNA.

Even though these sequences are highly conserved, they do present some variations in between organisms and thus, mearly writing out the consensus sequence leaves out information about the frequence and variation of each of the nucleotides or aminoacids in the sequence, which can be very important in Molecular Biology or Bioinformatics analysis.

Let's take the example of the aforementioned consensus sequence. In *E. coli* the SD sequence is **5'-AGGAGG-3'**. This sequence, however, has been shown to have slight variations in different bacteria [4] and thus a bioinformatician, when programming a gene prediction tool, may be misled to believe that that's the full extent of the SD sequence, when it's not. Moreover, one may argue that studying the variety of consensus sequences is important in undertanding intermollecular interactions.

Moreover, one study [1] arrived at the notion that graphical evidence, as opposed to evidence with a text baseline, helps intelligence analysis, such as medical and military decisions, by allowing experts to provide a more balanced and less biased decision. Given the fact that these sequences can have biomedical relevance, we can postulate that these more visual representations of data may help decrease bias in interpretation of disease-related sequences, for instance.

We then arrive at the logical conclusion that the visualisation of consensus sequences plays a very important role in various Molecular Biology and Bioinformatics studies and is therefore relevant to study better ways to display the available data in a way that's informative, visually appealing and easy to understand. Before arriving at a consensus sequence, one has to first perform a multiple sequence alignment (MSA). The tools that will be discussed and compared in this work have the ability to visually display MSAs and can be a very powerful tool in consensus sequence analysis.

2 Sequence Logos

Faced with the challenges of creating a visual representation of consensus sequences that provides more information than just the mere sequence of letters, Schnider and Stephens came forward with a proposal for a new paradigm in consensus sequences visualisation, for which they called Sequence Logos (Fig. 1) [9].

The Sequence Logos are comprised of four elements: the letters, which are shaped to become the "bar" of the chart; the colour, which is used to differentiate each of the symbols, even though a grayscale can also be employed; the height, which represents the conservation level of the residue at a particular alignment column and the axes, the horizontal one representing the location and the vertical

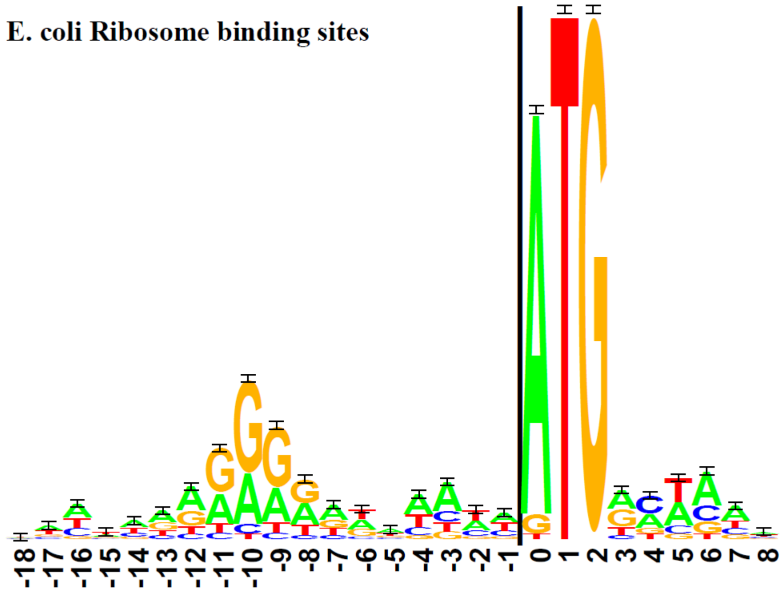

Figure 1: The Sequence Logo representation of the *E. coli* binding site [9, p. 2].

the frequency and conservation level of the symbols [5].

Making Sequence Logos nowadays is very easy, given the fact that there are free tools available. One such example is WebLogo, that automatically generates Sequence Logos from MSAs [2].

As we can see, this graphical method of representing data is a major step forward in relation with what existed before, as these logos allowed us to get a better feel for the relative frequencies of the residues at every position, which is crucial for a more complete understanding of such sequences.

It should then come as no surprise that these logos have become ubiquitous in many Molecular Biology and Bioinformatics studies, being the preferred method of visually displaying consensus sequences.

3 Shortcomings of Sequence Logos

Of course, like most visual representations of data, Sequence Logos are not without its drawbacks. Many of which are derived from the fact that, when converting raw data to a some form of visual display, there is some information that is lost in that process. There can be instances in which that information can have some relevance in the study, so researchers should beware of relying solely on these logos to study consensus sequences.

3.1 Biases in Sequence Logo analysis

Even though graphical tools, in general, have a tendency to minimize biases in decision-making [1], one should tread carefully in asserting that Sequence Logos are ideal ways to analyse biological data.

A 2012 study by Nung Kion and Yin Bee [5] has uncovered that researchers have a tendency to misuse Sequence Logos in computational transcription factor analysis. The authors arrived at that conclusion by analysing published articles and studying the validity of the conclusions derived from Sequence Logo analysis.

That analysis yielded that, in biological assays, the use of these logos tends to be more accurate because researchers complement their studies with statistical testing and additional biological data, rather than only relying on the logo itself. In sharp contrast, there is a significant tendency for biases in computational analysis, mainly confirmation bias [5].

These judgement errors usually occurred when researchers compared their algorithm with another by visual comparison of the two Sequence Logos [5]. These comparisons, according to the authors, are unfair because the Sequence Logo does not provide enough information to objectively access the quality difference of the two methods, adding that these depictions were not designed to allow easy comparison of similarity between them without expert knowledge on the motif being studied.

Perhaps one of the main reason why comparing Sequence Logos is so prone to bias is because scientific findings rely on reliable and systematic evidence to be corroborated, whereas evidence that relies solely on visual interpretations, like in these cases, is not solid enough to meet the criteria associated with the scientific method [5].

3.2 Readability and usability

The design of the Web Logo is far from ideal, as the letters are distorted to represent the conservation and frequency of the residue. That means readability takes a bit hit, as it's harder to distinguish the letters if they're very distorted, particularly the ones that have less significance. One could argue that those, because they're less frequent, are less important, but in protein Sequence Logos, that have many more letters than nucleic acid Sequence Logos, important information may be missed due to poor readability, especially if it's a small figure in a printed paper.

On that note, the fact that Sequence Logos are static and non-interactive also limits the amount of information that they can carry and how they involve the reader. These are, as stated by its original authors, simple tools [9], so one can not expect much user intractability from them, as they were designed to be placed on printed materials and to provide the most relevant information at a glance.

Some authors [5] also argue that the use of colours can be a drawback as well, as visualisations may appear more convincing than they really are, the colours in a Sequence Logo may impress the reader more than the actual quality of the motifs presented.

4 New approaches

Since the introduction of the Sequence Logo there have been new approaches that attempt to mitigate the shortcomings of Sequence Logos. Many of them have been incremental and use-specific (e.g. WebLogo and HMMLogo) but others, such as ProfileGrids, which will be discussed ahead, have taken bigger strides [8].

4.1 ProfileGrids

In an attempt to overcome the shortcomings of Sequence Logos, and to allow for the visual representation of large data sets, Roca, Almada and Abaijan introduced, in 2008, a new paradigm in sequence aligment visualisation, for which they called ProfileGrids (Fig. 2) [8].

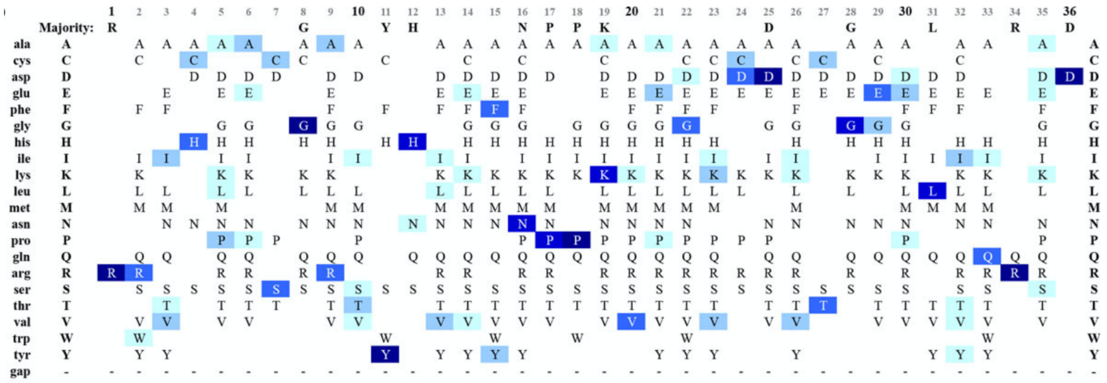

Figure 2: ProfileGrid visualisation of the AKL alignment protein residue distributions. Adapted from the original article [7, p. 4].

In these representations, each residue occupies a different line, and each position a different column, including the gaps in the sequences.

When compared to Sequence Logos, profile grids offer several advantages: 1) fixed height of residue symbols, allowing for clear reading of all symbols; 2) colour shading of cells, representing the relative frequency of each residues [7].

This paradigm also encompasses some degree of interactivity though a Java program named JProfileGrid, which combines the examining of aminoacid frequences across an MSA, identification of conserved motif regions and allows for comparing species-specific residues against a sequence family [8].

This paradigm builds upon previous the Sequence Logo and its derivations, offering a more suitable and interactive visualisation scheme for MSAs, having been showcased at the 2013 BioVis Redesign Contest. [7]

5 Conclusion

Visual representations of consensus sequences and other MSAs play an important role in giving scientists a better feel for a plethora of aspects surrounding relevant aminoacid and nucleotide motifs and can influence decision-making in related fields.

The more ubiquitous form of sequence visualisation, the Sequence Logo, is a simple approach to visualise motifs, and can be really useful in quickly giving the reader the most conserved residues, but it trades off the ability to convey information about larger data sets and to clearly display aminoacid sequences, being better suited for nucleotide sequences with short sizes.

The spiritual successor to these logos are Profile Grids, which overcome many of the limitations os Sequence Logos, especially for large aminoacid data sets, but are less intuitive and more cluttered, but with much better readability.

The logical conclusion is that visual representations of data should preferably be used in addition to other forms of testing and data analysis and should not be used on their own to compare different assays and that the most appropriate form of visual representation of motifs varies based on factors such as the type of molecule, the size of the data set and the length of the sequences.

References

[1] M. B. Cook and H. S. Smallman. “Human Factors of the Confirmation Bias in Intelligence Analysis: Decision Support From Graphical Evidence Landscapes”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50.5 (Oct. 1, 2008), pp. 745–754. ISSN: 0018-7208. DOI: `10.1518/001872008X354183`. URL: `http://hfs.sagepub.com/cgi/doi/10.1518/001872008X354183` (visited on 04/24/2020).

[2] G. E. Crooks. “WebLogo: A Sequence Logo Generator”. In: *Genome Research* 14.6 (May 12, 2004), pp. 1188–1190. ISSN: 1088-9051. DOI: `10.1101/gr.849004`. URL: `http://www.genome.org/cgi/doi/10.1101/gr.849004` (visited on 04/14/2020).

[3] A. Liljas. “Consensus Sequence”. In: *Encyclopedia of Genetics*. Elsevier, 2001, pp. 457–458. ISBN: 978-0-12-227080-2. DOI: `10.1006/rwgn.2001.0270`. URL: `https://linkinghub.elsevier.com/retrieve/pii/B0122270800002706` (visited on 04/22/2020).

[4] Jiong Ma, Allan Campbell, and Samuel Karlin. “Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures”. In: *Journal of Bacteriology* 184.20 (Oct. 15, 2002), pp. 5733–5745. ISSN: 0021-9193, 1098-5530. DOI: `10.1128/JB.184.20.5733-5745.2002`. URL: `https://JB.asm.org/content/184/20/5733` (visited on 04/22/2020).

[5] Lee Nung Kion and Oon Yin Bee. “Decision Making Biases in Using Sequence Logo Visualization”. In: *2012 Southeast Asian Network of Ergonomics Societies Conference (SEANES), Network of Ergonomics Societies Conference (SEANES), 2012 Southeast Asian* (July 1, 2012), pp. 1–6. ISSN: 978-1-4673-1732-0. DOI: `10.1109/SEANES.2012.6299605`.

[6] Benjamin A. Pierce. *Genetics: A Conceptual Approach*. 4th ed. New York: W.H. Freeman, 2012. 1 p. ISBN: 978-1-4292-3250-0.

[7] Alberto I Roca. “ProfileGrids: A Sequence Alignment Visualization Paradigm That Avoids the Limitations of Sequence Logos”. In: *BMC Proceedings* 8.S2 (Aug. 2014), S6. ISSN: 1753-6561. DOI: 10.1186/1753-6561-8-S2-S6. URL: https://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-8-S2-S6 (visited on 04/14/2020).

[8] Alberto I Roca, Albert E Almada, and Aaron C Abajian. “ProfileGrids as a New Visual Representation of Large Multiple Sequence Alignments: A Case Study of the RecA Protein Family”. In: *BMC Bioinformatics* 9.1 (Dec. 2008), p. 554. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-554. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-554 (visited on 04/27/2020).

[9] Thomas D. Schneider and R.Michael Stephens. “Sequence Logos: A New Way to Display Consensus Sequences”. In: *Nucleic Acids Research* 18.20 (1990), pp. 6097–6100. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/18.20.6097. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/18.20.6097 (visited on 04/14/2020).