compariSeq: Rethinking Sequence Logos

Sean McKenna, Philip S Quinan, Alex Bigelow

Abstract—Comparing Sequence Charts (compariSeq) is a redesign of traditional sequence logos for the task of comparing multiple biological sequences. All relevant data encoded in traditional sequence logos is conserved, however, attention is directed to the most important data, colors are more perceptually accessible, and the task of direct comparison at particular locations is supported.

1 INTRODUCTION

One of the most common methods to display biological sequences are the sequence logo [2]. At each location in a sequence, both the relative frequency of amino acids and information content (in bits) are encoded. Information content indicates both conservation across species and provides some information about the reliability of the consensus sequence presented. In addition, residues are typically colored according to their chemical properties or some other classification.

Sequence logos have several major flaws, however. Comparing residue frequencies is extremely difficult as the letters are stacked vertically. It is even more difficult to accurately compare lengths at non-neighboring locations, which is confounded by the scaling of relative frequencies according to information content.

To better understand the space of tasks associated with sequence logos, we conducted informal interviews with three biologists of varying expertise. These interviews yielded several specific observations, which inspired our new design: compariSeq.

2 COMPARISEQ

The layout of compariSeq is separated into components based on an overview+detail design pattern [3]: an overview for general analysis, and a comparison view for location-specific analysis across sequences.

2.1 Overview

The overview provides users with the entire sequence so that they can find general trends and locate interesting features. Sequence locations are arranged vertically, a rotation of the original design. At each location, conservation is encoded in the length of a horizontal bar. Within each bar, the relative frequencies of all residues at that location are displayed as a pie chart, with the highest frequency residues additionally listed in decreasing order from left to right. The benefit of this particular design is that the consensus sequence can quickly and easily be extracted by a vertical scan, while also also providing an indication of locations where the consensus is less pronounced. Selection of a specific sequence location highlights that location across all sequences, producing a linked comparison view.

2.2 Comparison View

The comparison view extracts the location-specific data and vertically arranges each sequence, creating an effective layout for comparing across sequences. Conservation is represented similarly, but the relative frequency of each residue is now encoded by the position of colored dots along a horizontal-axis, which is widely accepted as a superior type of encoding channel for accurate retrieval of quantitative information [1]. The dot size encodes frequencies scaled by conservation, thereby decreasing the visual salience of less important residues. Residues above a certain threshold get labeled in both views.

2.3 Justification of Design Choices

Our interviews with domain experts indicated that the most interesting value is the conservation of a location across species. Locations with low conservation are largely ignored. As such, conservation is the largest and most salient component of the visualization. To support this mental filtering task in a static plot, we retained the bar chart component of sequence logos.

Retaining the color encoding for amino acid properties is appropriate in that a color convention already exists for sequence logos. However, we observed that this convention is only loosely followed, so we took liberties with the exact colors to provide a more perceptually accessible palette. Hue-based color families still encode categorical data, but now each amino acid also has its own color, arguably enabling more effective search and comparison [4].

Pie charts are appropriate for showing frequencies in the overview for several reasons. As observed in interviews, frequencies are secondary in importance to conservation. If conservation is low, frequencies should be easy to ignore in a static plot. In some scenarios, using a pie chart as a small multiple would make them somewhat difficult to see, but this is desirable in that the size offsets the visual salience that the colors impart to specific amino acids. Furthermore, we observed that only a rough understanding of the frequency distribution is necessary in the overview, so the angle and area comparison problems inherent in pie charts are less relevant. The blending effect for small values is not a concern, since low-frequency residues are only important if their combined distribution is radically different from the high-frequency residues.

The primary task that we have identified is the *comparison* of specific locations across sequences. While heavily task-dependent, we observed that domain knowledge, coupled with conservation and approximate frequencies, is enough to determine which locations are most interesting. After interesting locations have been identified, detailed comparison of conservation and frequencies across sequences is performed spatially along a common axis, as in the comparison view.

By intelligently filtering and presenting the data, compariSeq offers an effective enhancement over sequence logos for comparison tasks.

ACKNOWLEDGMENTS

The authors wish to thank Heidi Schubert, Matthew Sdano, Venkatesh Rajamanikkam, and Miriah Meyer from the University of Utah.

REFERENCES

[1] J. Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[2] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100, 1990.

[3] B. Shneiderman. The Eyes Have It : A Task by Data Type Taxonomy The Eyes Have It : A Task by Data Type Taxonomy for Information Visualizations. *IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[4] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

- *Sean McKenna, E-mail: sean@cs.utah.edu.*
- *Philip S Quinan, E-mail: psq@cs.utah.edu.*
- *Alex Bigelow, E-mail: abigelow@cs.utah.edu.*