Sequence Bundles

James King[1], Lydia Nicholas[1] and Marek Kultys*[1]
1. Science Practice Ltd., 115 Bartholomew Road, London NW5 2BJ, United Kingdom;
*corresponding author: marek@science-practice.com

Background

Bioinformaticians use visualisation tools to assist them in investigating multiple sequence alignments. Sequence Logos represent individual positions in these alignments as stacked letters whose size is proportional to their relative frequency in each base. Whilst this method exposes the consensus sequence, it obscures patterns in the relationships between bases within sequences. We interviewed a number of bioinformaticians and found that they were so frustrated by visualisations which removed contextual information, that they preferred to study raw lists of sequence alignments.

Sequence Bundles

We propose a new visualisation method called Sequence Bundles that plots sequences as stacked lines against a Y-axis of letters arranged on a scale representing physiochemical properties. The lines' curved paths expose the conservation of relationships between bases in sequential positions and their place relative to letters on the Y-axis exposes patterns in functionality. The thickness of the stack at each letter in each position indicates of the level of conservation and the consensus sequence is named. Lines representing sequences with different properties (such as being Gram-positive or Gram-negative in case of the competition figure) may be marked with different colours to be compared on the same graph.

Design rationale

Sequence Bundles capture significantly more data than Sequence Logos, but are not overwhelming due to the following design philosophy. Bundles communicate quantitative information in a representational and direct manner, without the use of intermediary symbols or visuals (i.e. bars, slices, symbol heights or arbitrary colour coding). One semi-opaque line represents one protein sequence and its shape represents corresponding amino acids that every protein is made of. The bundled visualisation communicates information in a form suited to human perception—paths, amounts and colour intensity can be intuitively gauged, exposing patterns in the underlying quantitative data. This level of perceived detail is appropriate for global inspection of multiple sequence alignments.

Sequence Bundles can present information clearly and accurately in static form but are also exceptionally well suited to be extended in dynamic software. For instance selecting those sequences which match at one position could reveal that they closely correlate in other distant positions. The Y-axis could be toggled to reorganise the list of residues according to any given principle—physiochemical properties (molecular weight, hydrophobicity, etc.), alphabetically or in a computed order which reduces the tangling in the visualisation. Highlighting groups of sequences which share characteristics whilst dynamically reorganising the Y-axis could assist in exposing patterns which manifest across different aspects of function and structure.

Key developments

A: Shifting the focus of the visualisation from being position-oriented to sequence-oriented; Reason: Residues' functions are associated with their position in relation to one another within proteins. Because Sequence Logos represent residues in isolation without valuable contextual information, their position-oriented focus limits their uses. Sequence-focused Sequence Bundles assist in seeing a string of residues holistically as a functional protein, as well as exposing correlations and motifs, potentially assisting discovery.

B: Using semi-opaque curved paths instead of deformed letters; Reason: Deformed type is hard to read and stacking letters means that highly conserved ones rest on an uneven bed of less conserved ones, which makes them difficult to compare. Unfortunate stacking can lead to letter misinterpretation (e.g. V above I in base 23 of the contest's Logos could be misread as Y). Representing sequences with curved paths allows for their equal and proportional display with strong focus on sequence continuity. Atypical sequences are never removed but are faint enough to be inconspicuous.

C: Reassigning the Y-axis from displaying the amount of information measured in bits to displaying letter-coded amino acids arranged by physiochemical properties; Reason: The bioinformaticians we consulted were uninterested in the level of detail about mutual information shown in protein alignments. For the purpose of protein conformation research, residue physiochemical properties are reportedly a far more important measure and deserve more refined and structured representation than the colour-coding used in Logos (this allows Sequence Bundles to adhere to the best practice of accessibility for users with colour vision deficiency).

D: Integrating three contest Sequence Logos into one combined visualisation; Reason: It is very difficult to compare stacked letters across separate Logos, and we found that users frequently misjudged letters' relative proportions. Looking at the combined Logo to analyse Gram-positive and Gram-negative sequence Logos may be compared to analysing a father and mother by looking at their child—impossible as the child fuses various qualities. Placing the two data sets on the same graph and differentiating by colour allows easy and direct comparison of both groups whilst also offering a general overview of the whole population.

E: Visualising gaps as a separate unit on the Y-axis; Reason: Multiple sequence alignments rely on gaps to optimise alignment. Gaps are never shown in Sequence Logos (exception: HMM Logo) which dissociates visual representations from underlying data. Sequence Bundles display gap locations within each sequence alongside gaps' actual length.

F: Supplementing the legend (Fig. 1.) with meta information; Reason: Our advisors were hesitant to trust Logos as they gave no indication of the number and proportion of compared sequences. Logos based on 9 or 9,000 sequences could look the same, but their credibility could be very different. To avoid cluttering, this simple information has been added to the legend.

Acknowledgments

Our advisors were: Dr Daniel Buchan (University College London, UK), Dr Nick Goldman and Dr Roland Schwarz (European Bioinformatics Institute, Hinxton, Cambridge, UK), and Dr Stathis Sideris—thank you.

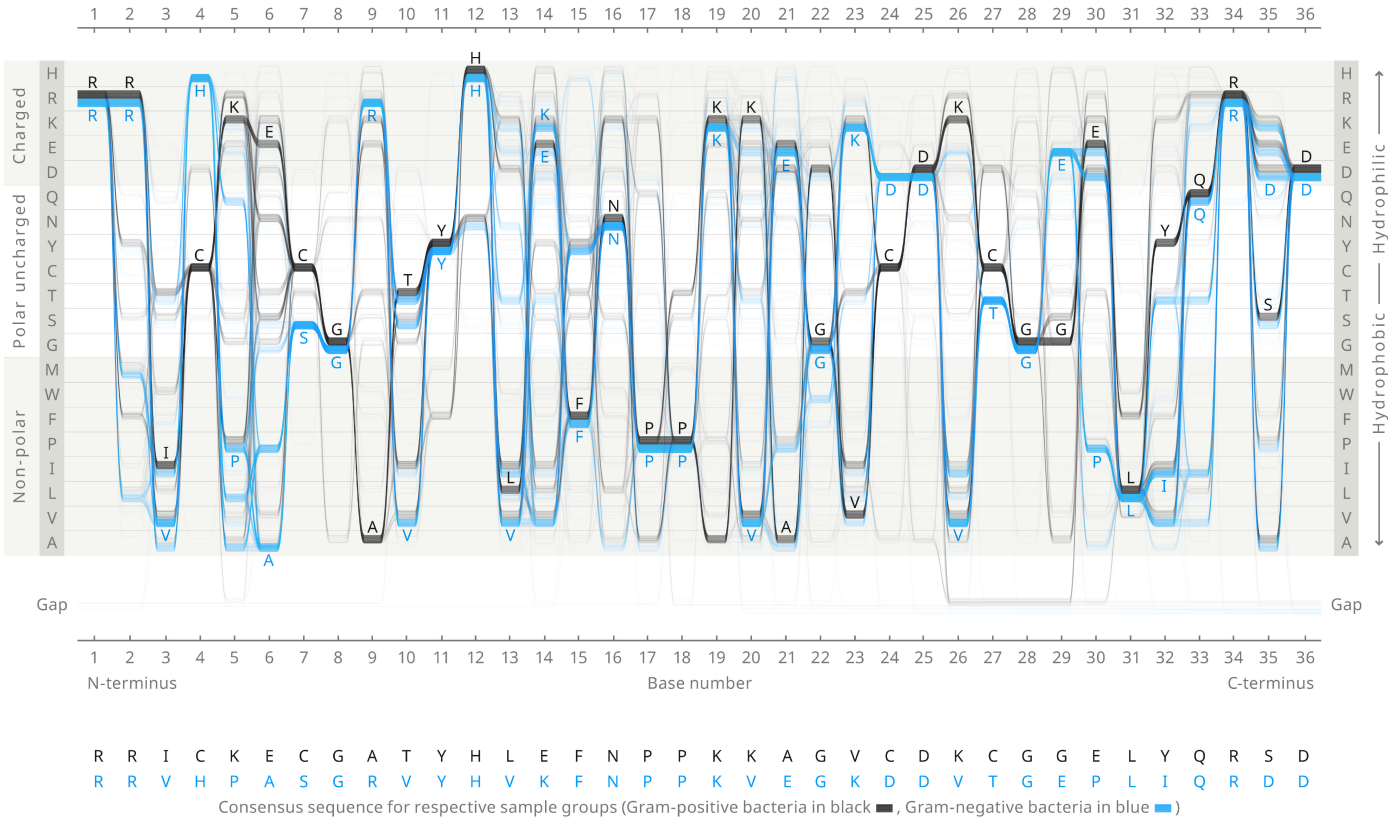
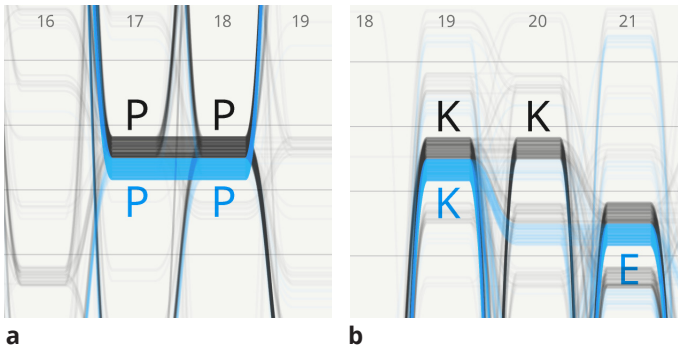


Figure 1.

Sequence Bundle comparing the amino acid distribution and correlation in the adenylate kinase lid (AKL) domain between two groups of organisms: Gram-positive (black ▬) and Gram-negative bacteria (blue ▬).

The ADK lid domain structure is universally conserved, but is stabilized in the Gram-negatives by a hydrogen bonding network between residues 4, 7, 9, 24, 27, and 29 (and several other residues in some organisms), while the Gram-positives are stabilized by a bound metal ion, tetrahedrally coordinated by the Cysteines at 4, 7, 24 and 27. The identities of several other positions (eg 5, 8, 30, 32) are differentially constrained in each subfamily as well, apparently due to steric requirements of the stabilizing residues. [...]

The visualisation is generated from a total of 1809 AKL protein sequences. Disproportion in the number of samples is 932 sequences (Gram-negative) against 886 (Gram-positive), which is the ratio of 100:95.06. No information about the distribution of the samples on phylogenetic tree is available.



a **b**

Figure 2. (a & b).

Comparison of two details from Fig. 1. showing a data feature exposed by Sequence Bundles. The consensus for Gram-positives (black) in bases 17-20 is: ...PPKK... These two repetitions are shown in the contest Logo, but isolated from sequence-wide context this information is misleading. Bundles reveal much more—that while there is a repetition of Prolines (PP) in Gram-positives, there is no repetition of Lysines (KK). Most sequences have two consecutive Prolines in bases 17-18 (hence a bridge between P-P in panel **a**), while only about half have a Lysine in base 19, and the other half have it in base 20. Note that very few black lines bridge the gap between K-K in Bundle visualisation (**b**) because most sequences incluce only one of the Lysines.

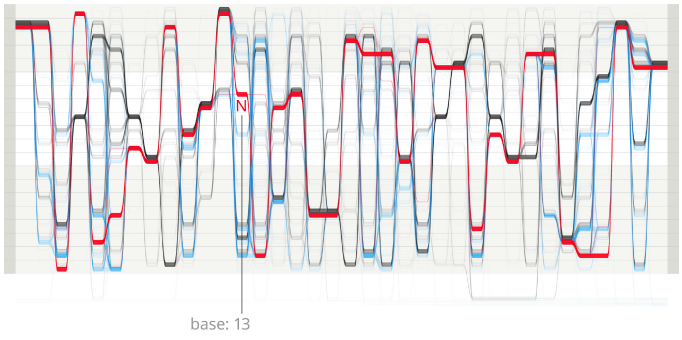


Figure 3.

Demonstration of the potential for Sequence Bundles to assist discovery when used in a dynamic (interactive) software tool.

Sequence Bundles have great potential for being freely explored as direct representations of all aligned sequences. Hovering over threads with a GUI cursor could highlight individual sequences and allow for instant visual comparison. Selecting a group of sequences (as shown in Figure 3.) can reveal even more curious patterns, e.g.: that any protein with Asparagine (N) in base 13 is extremely well conserved throughout the entire sequence length—which can inform further extrapolations in respect to protein's structure or its folding and functioning behaviour.