Using Phylogenetic Contrast to Visualize Phylogenetic Signals With Sequence Logos

HaoLin Guo

Bachelor Thesis in Computer Science
Stockholm University

May 24, 2025

Introduction: The Challenge of Evolutionary History

- **Sequence Logos**: Visual tools for conserved patterns in biological sequences.
- **Problem**: Traditional logos assume statistical independence, ignoring **phylogenetic non-independence**.
  - Leads to misrepresentation and sampling bias (e.g., "drowning out" effects).
  - Undermines statistical validity.
- **Solution**: Propose a method to account for shared evolutionary history in sequence logo generation.

Background: Traditional Sequence Logos

- Constructed from multiple sequence alignments.
- Stack height indicates positional conservation (R_{seq}).
- Character height proportional to its frequency.
- $R_{seq} = S_{max} - S_{obs}$ (Maximum Entropy - Observed Entropy).
- **Limitation**: Fails to incorporate evolutionary history.

$$\begin{array}{cc}
 & \begin{matrix} A & C & T & G \end{matrix} \\
\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0.5 & 0 & 0.5 \end{bmatrix}
\end{array}$$

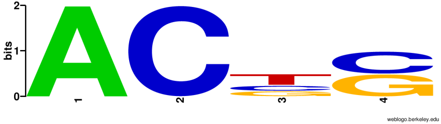


Figure: 2.1: Example of a Traditional Sequence Logo.

Background: Phylogenetic Independent Contrasts (PIC)

- **PIC (Felsenstein, 1985)**: Statistical method to address phylogenetic non-independence.
- Transforms trait values into statistically independent contrasts.
- Assumes **Brownian motion model** for trait evolution.
- **Branch lengths** are critical: quantify divergence, used as weights.

Methodology: Adapting PIC for Sequence Logos

- **Adapting PIC for sequence data**
- Represent each sequence as a **positional frequency matrix**.
- Propagate ancestral symbol probabilities up the tree using branch lengths as weights.
- Aims for a "root-state symbol distribution" reflecting evolutionary conservation.
- Corrects for sampling bias and accounts for sequence relationships.

Methodology: The Iterative Algorithm

- **1. Initialization**: Convert sequences at each tip to positional frequency matrices (M_i).
- **2. Iterative Estimation**: For sister nodes 'i' and 'j' with parent 'k', estimate M_k:

$$M_k = \frac{\frac{M_i}{v_i} + \frac{M_j}{v_j}}{\frac{1}{v_i} + \frac{1}{v_j}}$$

- **3. Branch Length Lengthening**: Update parent branch length v'_k:

$$v'_k = v_k + \frac{v_i v_j}{v_i + v_j}$$

- **4. Completion**: Repeat $n - 1$ times to get root matrix M_R for logo generation.

Experimental Validation: Balanced Tree

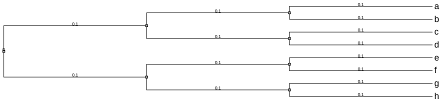


Figure: 4.1: Perfectly Balanced Tree.

Label	Sequence									
S_a	A	A	A	A	A	C	A	A	C	Q
S_b	A	A	R	V	A	C	C	C	A	R
S_c	A	A	V	R	A	C	C	A	C	R
S_d	A	A	E	S	A	C	C	C	A	R
S_e	A	A	S	E	C	A	C	A	C	R
S_f	A	A	T	I	C	A	C	C	A	R
S_g	A	A	I	D	C	A	C	A	C	R
S_h	A	A	D	T	C	A	C	C	A	R

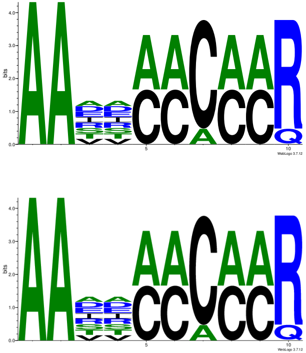


Figure: 4.2: Traditional (top) and PIC (down) sequence logo for the perfectly balanced tree.

- **Outcome**: PIC-generated logo was **identical** to traditional logo.

Experimental Validation: Unbalanced Tree Insights

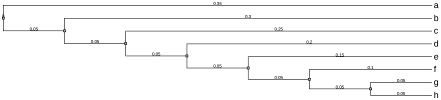


Figure: 4.3: Unbalanced Tree.

- **Outcome**: PIC logo showed clear adjustments (e.g., character heights for sequence corresponding to the tip *a* are adjusted).
- **Implication**: Corrects sampling bias, preserves evolutionary structure.

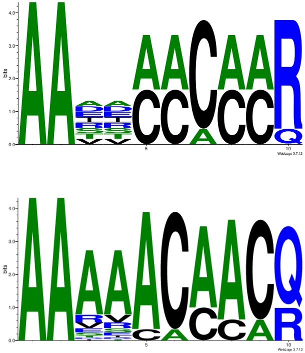


Figure: 4.4: Traditional and PIC Logo for the unbalanced tree.

Experimental Validation: Further Unbalanced Cases

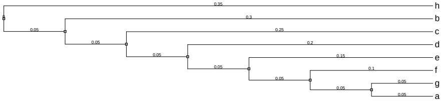
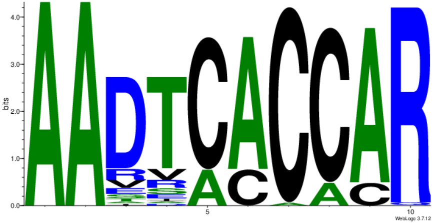


Figure: 4.5: Modified Unbalanced Tree and the corresponding PIC logo.



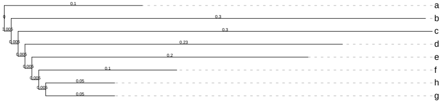
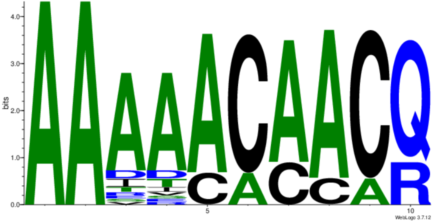


Figure: 4.7: Unbalanced Tree with Different Branch Lengths and the corresponding PIC logo.



- **Implication**: Adjustments tied to phylogenetic position; sensitive to quantitative evolutionary divergence.

Results: Biological Dataset PS00027

- 1362 protein sequences.

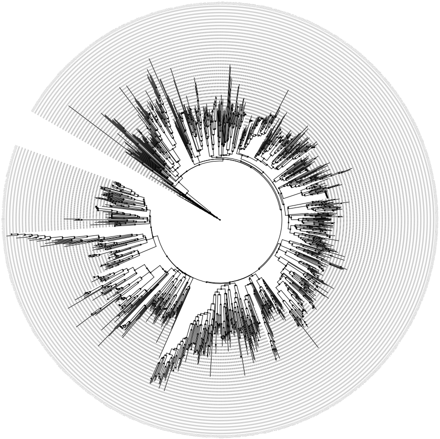


Figure: 5.1: PS00027 Phylogenetic Tree.

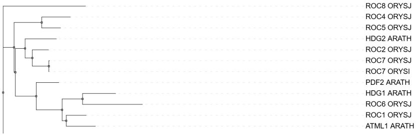


Figure: 5.2: The two smaller sub-trees formed by the child node of the root.

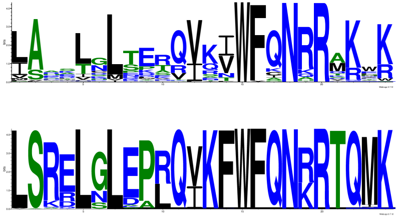


Figure: 5.3: Traditional and PIC Logo for PS00027.

- **Observation**: PIC logo heavily influenced by these two smaller sub-trees (with 1 and 11 tips).
- **Implication**: Can lead to "over-representation" of basal clades; rooting choice is critical.

Results: Biological Dataset PS00673

- 16 protein sequences.

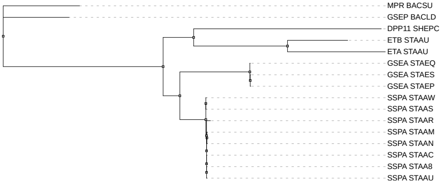
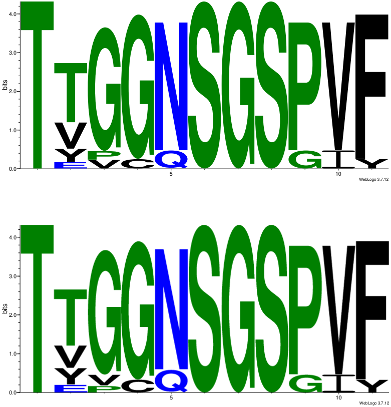


Figure: 5.4: PS00673 Phylogenetic Tree.



- **Observation**: Less pronounced differences; subtle adjustments (e.g., position 3).

Figure: 5.5: Traditional vs. PIC Logos for PS00673.

Discussion: Advantages and Limitations

Advantages

- Preserves underlying evolutionary structure.
- Biologically informed, nuanced visualization.
- Principled alternative for phylogenetically significant datasets.

Limitations

- Possible ”Over-representation” of basal sub-trees with short branches.
- Finding the right root.

Conclusion and Future Directions

- A modified PIC algorithm for sequence logos.
- Provides a more accurate visualization of evolutionary conservation.
- Resolves biases that can obscure true evolutionary signals.

Future Directions

- Investigate **rerooting phylogenetic trees** to mitigate over-weighting.
- Explore different rooting strategies or adaptive weighting schemes for balanced representation.

Questions?

Thank you for your attention!