

Average Log-Likelihood Ratio: Formulas

1 Setup

A position frequency matrix (PFM) has 4 rows (A, C, G, T) and W columns (positions). Column j is a probability vector $\mathbf{f}_j = (f_{A,j}, f_{C,j}, f_{G,j}, f_{T,j})$ with $\sum_i f_{i,j} = 1$.

Let $\mathbf{b} = (b_A, b_C, b_G, b_T)$ be a background distribution (default: uniform, $b_i = 0.25$).

2 Column-level ALLR

Given two columns C_1 and C_2 with frequency vectors \mathbf{f}_1 and \mathbf{f}_2 :

$$\text{ALLR}(C_1, C_2) = \frac{1}{2} \left[\sum_{i \in \{A, C, G, T\}} f_{i,1} \ln \frac{f_{i,2}}{b_i} + \sum_{i \in \{A, C, G, T\}} f_{i,2} \ln \frac{f_{i,1}}{b_i} \right]. \quad (1)$$

The first sum asks how well C_2 explains observations drawn from C_1 (relative to background); the second sum is the reverse. Averaging makes the measure symmetric.

Properties:

- $\text{ALLR}(C_1, C_2) = \text{ALLR}(C_2, C_1)$.
- If both columns equal the background, $\text{ALLR} = 0$.
- Identical informative columns yield a positive score; dissimilar columns yield a low or negative score.

3 Matrix-level ALLR with sliding alignment

Let PFM A have width w and PFM B have width $W \geq w$. For offset $k = 0, 1, \dots, W - w$, define the aligned score:

$$S(k) = \frac{1}{w} \sum_{j=1}^w \text{ALLR}(A_j, B_{k+j}), \quad (2)$$

where A_j is column j of A and B_{k+j} is column $k + j$ of B .

The best-match score and offset are:

$$S^* = \max_k S(k), \quad k^* = \arg \max_k S(k). \quad (3)$$

If A is wider than B , their roles are swapped so the shorter matrix always slides over the longer one.

4 P-value by permutation test

The null hypothesis is: the observed similarity S^* is due to chance.

4.1 Procedure

1. Generate N random PFMs of the same width as the target. Each column is drawn independently from $\text{Dirichlet}(\alpha_0 \mathbf{b})$, where $\alpha_0 > 0$ is a concentration parameter (default: $\alpha_0 = 4$).
2. For each random PFM R_n , compute S_n^* by sliding alignment against the input PFM (same procedure as above).
3. The p-value is:

$$P = \frac{\#\{n : S_n^* \geq S^*\} + 1}{N + 1}. \quad (4)$$

The $+1$ in numerator and denominator is the standard pseudo-count correction that avoids $P = 0$ and accounts for the observed score itself being one draw from the distribution.

4.2 Interpretation

Small P (e.g. $P < 0.05$) indicates the match is unlikely under the null model. The permutation approach automatically accounts for motif width and information content without parametric assumptions.

Reference

Wang, T. and Stormo, G.D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380. <https://academic.oup.com/bioinformatics/article/19/18/2369/194379>