Regulatory Sequence Analysis

Theoretical distribution of PSSM scores

Jacques van Helden https://orcid.org/0000-0002-8799-8584

Aix-Marseille Université, France Theory and Approaches of Genome Complexity (TAGC)

Institut Français de Bioinformatique (IFB) <u>http://www.france-bioinformatique.fr</u>

Scanning a sequence with a position-specific scoring matrix

- P(S|M) probability for site S to be generated as an instance of the motif.
- P(S|B) probability for site S to be generated as an instance of the background.
- W weight, i.e. the log ratio of the two above probabilities.
 - A positive weight indicates that a site is more likely to be an instance of the motif than of the background.



Score distribution: random expectation

- The theoretical distribution of probabilities for position-weight matrices has been discussed in several articles.
 - Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5, 89-96.
 - Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-77.
- The computation is based on the probability-generating function.
- This function can be used to compute the probability P(W) to obtain exactly a score value of W.
- Each position-weight matrix has its own probability distribution.

$$G_j(x) = \sum f_i x w_{ij}$$

- Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5, 89-96.
- Bailey, T. L. & Gribskov, M. (1997). Score distributions for simultaneous matching to multiple motifs. J Comput Biol 4, 45-59.
- Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-77.

3

Theoretical distribution of matrix scores

- The RSAT program *matrix-distrib* computes the distribution of score probabilities for a given PSSM.
- The distribution is completely determined by
 - Values in the cells of the matrix
 - Prior residue probabilities
- This method can be used to compute
 - $\square P(X=x|M)$
 - The probability to obtain by chance a given score *x*, with a given matrix.
 - $\square \quad P(X \ge x | M)$
 - The probability to obtain by chance a score higher or equal to *x*.
 - The inverse cumulative distribution gives a *P-value*, which indicates the risk of false positive for a given score.
- Computing time increases exponentially with the number of columns, but by rounding values, it is asymptotically linear.
- The original method is based on a Bernoulli assumption for the background model, but we extended it to Markov chains.
- Computing time increases exponentially with Markov order.

Theoretical distribution for the PHO matrix

- The program matrix-distrib (RSAT) computes the complete theoretical distribution of scores for a given PSSM, using the algorithm proposed by Staden (1989), and previously implemented in patser (Hertz, 1990, 1999) and MAST (Bailey, 1994, 1997).
- The theoretical distribution P(S) is quite erratic, because each possible value of score has its own probability, depending on
 - the actual weight values in the matrix, and
 - prior residue probabilities.
- Figure below: probability distribution of weight score according to a Bernoulli model.



P(W=w | M)

probability to obtain by chance a precise weight score *w*, with a given matrix.

$\mathsf{P}(\mathsf{W}{\geq}\mathsf{w}\mid\mathsf{M})$

probability to obtain by chance a weight score higher than or equal to *w*.

This inverse cumulative distribution gives a P-value, which indicates the risk of false positive for a given score.

Theoretical distribution for the PHO matrix

- matrix-distrib also supports computation of P-values with Markov models of any order (algorithm adapted from Touzet & Varré, 2007).
- Computing time increases exponentially with the number of columns, but by rounding values, it is asymptotically linear.
- The original method is based on a Bernoulli assumption for the background model, but we extended it to Markov chains.
- Computing time increases exponentially with Markov order.
- Figures below: probability distribution of the weight score according to a Markov model of order 1.



P(W=w | M)

probability to obtain by chance a precise weight score *w*, with a given matrix.

$\mathsf{P}(\mathsf{W}{\geq}\mathsf{w}\mid\mathsf{M})$

probability to obtain by chance a weight score higher than or equal to *w*.

This inverse cumulative distribution gives a P-value, which indicates the risk of false positive for a given score.

Theoretical distributions

- We used *matrix-distrib* to analyse the theoretical distributions for some matrices according to various BG models.
 - IID (independently and identically distributed) nucleotides (blue)
 - Markov chains of orders 1 to 5, trained on the whole set of upstream sequences of the considered organism.

Yeast PHO4



Mammalian Sp1



Drosophila Kr



Impact of Markov order on the right tail of theoretical distributions

- In *Escherichia coli*, using higher-order background model has a weak effect on the core of the distribution, but affects its right tail, which corresponds to high-scoring sites.
- For TrpR, we observe differences for lower-scoring sites, due to the presence of a particular 4nt in the motif.
 Theoretical Distributions

