**Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences**

Martin C. Frith, John L. Spouge, Ulla Hansen, Zhiping Weng

Deal with “clusters” of cis-regulatory sequences in DNA. Problem: scoring sequences when there are, at times, long gaps

**A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling**

**Gert Thijs, Magali Lescot,...**

Creating a background from Markov-chain that predicts upcoming nucleotides in a sequence based on the m-previous one (for a background of m-th order)

**On counting position weight matrix matches in a sequence, with**

**application to discriminative motif ﬁnding**

Saurabh Sinha

1. Counting vs. asking “does the PWM occur?”; one is continuous, the other discrete
2. What counts as an “occurrence”? threshold

w-score: quantifies the number of occurrences of a PWM in a sequence.

Why is simply multiplying and adding not enough?

There are “high sampling probability”-motifs

*Do they mean with respect to the nucleotides contained? A and G are more frequent than C and U?*

**Efficient and accurate P-value computation for Position Weight Matrices**

Hélène Touzet and Jean-Stéphane Varré

Why do I need the p-value to set a score threshold?

every PWM has a different probability of arising by chance, depending on composition, scoring scheme and length

**Pairwise Statistical Significance of Local Sequence Alignment Using Sequence-Specific and Position-Specific Substitution Matrices**

Ankit Agrawal and Xiaoqiu Huang

Scoring the fit of a motif to a sequence by merely computing the PSSM is considered a naive approach, as different motifs can arise randomly with differing chances. A shorter sequence will have higher chances than a longer one, and dinucleotide frequencies suggest that a sequence made up of C and U will be less likely than an A-G one. In order to approach the problem in a statistically rigorous way, a p-value distribution of the motifs in question has to be computed and a cut-off value determined.

Pairwise statistical significance: sequence is fit to another. Seq2 is shuffled around.

Paper describes a different way of calculating statistical significance that doesn’t rely on a large database of sequences.

What do I need the database for? Creating a suitable background nucleotide distribution to estimate how often my motif would REALLY arise randomly given that distribution.

?? Pairwise significance takes a single partner-sequence and shuffles it around enough times to create a wide variety of “random” sequences to compare the motif against.

**Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**

Gerald Z. Hertz and Gary D. Stormo

“A good alignment is assumed to be one whose alignment matrix is rarely expected to occur by chance.”

Information content: variant of log-likelihood ratio

Information content is not an end in itself: they calculate the p-value for obtaining a given information content value