A STATISTICAL METHOD FOR PWM CLUSTERING

SOLENNE CARAT1,2, REMI HOULGATTE1, JEREMIE BOURDON2

Introduction
Motif discovery is a fundamental problem in molecular biology. It possesses important applications in the study of regulatory signals and transcription factor binding sites discovery. Several motif discovery tools have been proposed (see [1] for a complete review). They all extract significant motifs from sets of sequences. Nevertheless, addressing motif discovery for complex organism is still a challenge. It is thus interesting to take profit of the specificities of every discovery tools with different parameters for extracting several putative interesting motifs. Doing this impose to deal with redundant motifs that must be removed.

Here, we propose a method for comparing several motifs given by their PSSM (Position Specific Scoring Matrices). This method automatically detects periodic motif and redundant motif. It is also possible to compare a final set of motifs with public databases [2,3]. Notice also that palindromic motifs can be detected easily with this method.

Methods
Our method is based on comparison of PSSM. The use of PSSM, rather than PWM, is justified by the exactness of the content, while PWM may require pseudo-count adaptation. These PSSM can be constructed easily from any motif discovery tools. All matrices are compared pairwise. Reverse complements are also taken into account. For each pair of motifs (m,n), comparison is done between m and all possible shift of n. Shifts allow to detect imbricate motifs. The specificity of our comparison method is that its is performed only on bases which frequencies are superior to determinate threshold, like background for example. This limits the effects of noise in the comparison. Finally, a Chi-square test is used to compare the two distributions of frequencies.

This comparison method allows to detect periodic motif, like tandem repeat GC, comparing PSSM to itself with lag of 2 bases. If these two PSSM are similar, motif is periodic (Fig. 1.1). In the same vein, comparing a PSSM with its reverse complement allows to determine if it is a palindrome (Fig. 1.2).

2 LINA, CNRS UMR6241, Nantes, Jeremie.Bourdon@univ-nantes.fr
Fig. 1: Comparison of several motifs

**Optimizations**
Many parts of the process treatment are quite independent. It is thus possible to take advantage of modern computer architectures (multicore computers, clusters, grid) by a parallization of these parts of computation. This allows a huge gain of the time needed to get a full result.

**Discussion**
Motif comparison allows to detect periodic and palindromic motifs, and identify transcription factors that recognize it through public databases. Moreover, by grouping similar motif, it is possible to generate consensus motifs that correspond to a larger number of sequences, and to reduce number of motifs to be studied.

1. G. K, Sandve, F. Drablos (2006), Biology direct, 1:11
2. A. Sandelin et al., (2004), Nucleic Acids Res. 32: D91-94