DNA-MOTIF IDENTIFICATION UNDER VARIOUS KINDS OF ORDER RESTRICTION USING MULTIPLE CONTRASTS

X. Mi , L.A. Hothorn

Biostatistics Unit, Hannover, Germany

Email : mi@biostat.uni-hannover.de

In DNA sequences, different positions have different degree of conservation. The DNA-binding proteins are bounded to some very conservative base pairs, called motif (Lawrence and Reilly, 1990). Usually a multinomial distribution of the four bases ACTG on the DNA sequences is supposed. Recently, vanZwet et al. (2005) constrained the order of the entropy $H_j(X\ ) = \mathrm{E}_p\{\log \frac{1}{p(X_j)}\}$ in the columns of the position specific weight matrix $\{X_{ij}\}$ which characterizes the motif being sought. Following this idea, we use a multiple contrast approach which transfers the position specific weight matrix for the multinomial model into binomial ($\pi_j = \frac{\max\limits_i (X_{i,j})}{sum(X_{i,j})}$), asymptotically binomial or entropy model. For the specific order restricted alternative, a multiple contrast test (contrast C=$\sum\limits_j c_j \pi_j , \sum\limits_j c_j = 0$) is proposed. The reliability of the algorithm increases with the number of fragments, but the computations increase only linearly if we choose a suitable contrast matrix. Analogously to the vanZwet et al. (2005) approach a global decision against the null hypothesis is possible. However, using a parametric bootstrap, model selection of the most likely pattern is possible. A priori, power estimation is possible for different expected proportion profiles for different motif configurations.

Examples with both simulated and real data show that this extension helps discover motifs as the data become noisier. A related R-package is proposed for a simple analysis of real data with quite different a priori motif definitions.