Sequence analysis

Deep and wide digging for binding motifs in ChIP-Seq data
I. V. Kulakovskiy1,7,*, V. A. Boeva1,2,3,4,5, A. V. Favorov1,6 and V. J. Makeev1,7
1Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow 117545, Russia, 2Institut Curie, 26 rue d’Ulm, 3INSERM, U900, 4INSERM, U830, Paris F-75248, 5Mines ParisTech, Fontainebleau F-77300, France, 6Johns Hopkins University, Baltimore, MD 21205, USA and 7Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia

Received on May 25, 2010; revised on August 7, 2010; accepted on August 18, 2010

1 INTRODUCTION

ChIP-Seq (Valouev et al., 2008) is an efficient technology for the identification of DNA sites of a specific protein binding. Being coupled with peak finding algorithms (Fejes et al., 1993) or MEME (Bailey et al., 2001), or CHIP-Munk (Georgiev et al., 2008), it yields a set of DNA segments with each sequence position having a weight reflecting how often DNA nearby was cross-linked with the protein of interest during ChIP stage (the so-called peaks). There can be tens of thousands of data sequences. Comparison with traditional (MIME) or ChIP-Seq-oriented (HMS) motif discovery tools shows that ChIPMunk identifies the correct motifs with the same or better quality but works dramatically faster.

2 METHODS

2.1 KDIC concept

The basic version of the ChIPMunk algorithm searches for the motif with the highest discrete information content (DIC) (Kulakovskiy et al., 2009). DIC does not take into account the background nucleotide composition. To allow for the background, we use the discrete analog of Kulback–Leibler divergence (Kulback and Leibler, 1951):

$$
KDIC = DIC - \sum_{j=1}^{n} \sum_{G,T,C,A} n_{j} \alpha \log \alpha
$$

where $n_{j}$ are elements of the position count matrix (i.e. letter counts), $n$ is the motif length and $\alpha$ are the background nucleotide probabilities. Please refer to the STS 2 for the mathematical details related to KDIC.

2.2 ZOOPS mode

ChIPMunk searches for the gapless multiple local alignment that has the maximum KDIC value. Then a positional weight matrix (PWM) is constructed, and all the aligned words are sorted by their PWM scores. They are classified into ‘the signal’ and ‘the noise’ subsists. To find the boundary word we construct a series of PWMs, made from top 1, 2, …, $n$ words. The idea is that for ‘the signal’ the score of $n$-th word calculated with the $n$-th PWM should be visibly greater than that calculated with $N$-th PWM (where $N$ is the total number of words). See STS 3 for details.

2.3 Positional profiles

There are two types of weights assigned to the initial sequence data. Sequences are weighted as a whole; also the weights are assigned to the each sequence position forming the sequence profiles. The source of all weights is the peak shape data. A weight $Wi_{j}$ for the $i$-th sequence is its normalized maximal peak value; by normalization the sum of $Wi_{j}$ over all $i$ is equal to the total number of sequences. The sequence profiles are normalized to make them fit in [0,1] interval.

PWM optimization includes two alternating steps: (i) the alignment is rebuilt from words with maximal PWM scores in each data sequence and (ii) the PWM is rebuilt from new motif occurrences. PWM scores for putative hits are weighted from sequences profiles:

$$
\text{score}(PWM, word) = \sum_{j=1}^{n} n_{word,j} \alpha \log \alpha
$$

where $n$ is the word length and $n_{word,j}$ is the PWM element for the $j$-th letter in the word. Thus, the positions with a larger profile values contribute more
When the segment length was increased MEME and SeSiMCMC, We thank Biobase and personally Alexander Kel for providing us. We took three ChIP-Seq datasets: NRSF (Johnson et al., 2007), GABP (Valouev et al., 2008) and EWS-FLI1 (Guillon et al., 2009). We used FindPeaks (Fejes et al., 2008) to obtain enriched regions (the peaks). Segments with strict GGAA-type repeats were excluded from the EWS-FLI1 dataset. Motif lengths were fixed at 21, 12 and 11 for NRSF, GABP and EWS-FLI1, respectively. 

Top 100 (NRSF, GABP) and top 500 (EWS-FLI1) peaks were taken to test several motif discovery tools, including MEME (Bailey et al., 2009), SeSiMCMC (Favorov et al., 2005) and HMS, the novel ChIP-Seq-oriented Gibbs sampler (Hu et al., 2010). We used sets of segments that were truncated from 10% to 100% of the peak lengths and were centered at the peak maxima.

The fraction of peaks truncated to 10% of initial length centered at the peak maximum with motif hits having scores greater than the mean + 3 SD of score distribution over all w-mers was used as the measure of motif quality. We assessed the time efficiency of motif discovery and the resulting quality of motifs identified from 500 EWS-FLI1 peaks (Fig. 1). For short segments covering only 10% of the peaks, all the tested tools performed equally well. Nevertheless, in both modes, ChIPMunk clearly outperformed HMS that did not take the peak shape into account, identified incorrect motifs. In contrast, ChIPMunk (in peak mode) always identified the correct motif. In the mode that did not take peak shapes into account, identified incorrect motifs. In contrast, ChIPMunk failed to identify the correct motif in longer segments.

Motif discovery time and resulting quality of the motifs identified in top 500 EWS-FLI1 peaks. Refer to the text for more details. The correct motif consensus gacaGGAAatg is similar to that in Guillon et al. (2009).

Funding: Russian Federal Agency for Science and Innovation State Contract [02.531.11.9003, 02.740.11.5008]; Russian Fund for Basic Research Project [10-04-92663 to V.J.M.].

Conflict of Interest: none declared.

REFERENCES


(see the STS 8) even on a personal computer, taking advantage of the modern multi-core processors.

ACKNOWLEDGEMENTS

We thank Biobase and personally Alexander Kel for providing us with the free access to the TRANSFAC database.

Motif discovery in ChIP-Seq data