BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2622–2623
doi:10.1093/bioinformatics/btq488

Sequence analysis

Advance Access publication August 24, 2010

Deep and wide digging for binding motifs in ChIP-Seq data

I. V. Kulakovskiy[1,7,*], V. A. Boeva[1,2,3,4,5], A. V. Favorov[1,6] and V. J. Makeev[1,7]

[1]Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow 117545, Russia, [2]Institut Curie, 26 rue d'Ulm, [3]INSERM, U900, [4]INSERM, U830, Paris F-75248, [5]Mines ParisTech, Fontainebleau F-77300, France, [6]Johns Hopkins University, Baltimore, MD 21205, USA and [7]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia

Associate Editor: John Quackenbush

ABSTRACT

Summary: ChIP-Seq data are a new challenge for motif discovery. Such a data typically consists of thousands of DNA segments with base-specific coverage values. We present a new version of our DNA motif discovery software ChIPMunk adapted for ChIP-Seq data. ChIPMunk is an iterative algorithm that combines greedy optimization with bootstrapping and uses coverage profiles as motif positional preferences. ChIPMunk does not require truncation of long DNA segments and it is practical for processing up to tens of thousands of data sequences. Comparison with traditional (MEME) 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.

Availability and implementation: ChIPMunk is freely available within the ru_genetika Java package: http://line.imb.ac.ru/ChIPMunk. Web-based version is also available.

Contact: ivan.kulakovskiy@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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.*, 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 DNA segment, some longer than 1000 bp. Classic tools like Weeder (Pavesi *et al.*, 2001), Gibbs Sampler (Lawrence *et al.*, 1993) or MEME (Bailey *et al.*, 2009) were tested for motif discovery in ChIP-Seq segments, but proved to be inefficient for the whole amount of data. Therefore, a fraction of top DNA segments was usually taken (Chen *et al.*, 2008) or/and the segments were truncated around the peak maxima (Ji *et al.*, 2008). Tools like cERMIT (Georgiev *et al.*, 2010) and HMS (Hu *et al.*, 2010) have been developed for motif discovery from ChIP-Seq data. In this note, we present a motif discovery algorithm that can extract the single optimal motif from large datasets like ChIP-Seq without any preprocessing, while taking into account the peak shape.

*To whom correspondence should be addressed.

2 METHODS

The basic ChIPMunk implementation (Kulakovskiy and Makeev, 2009, see also the Supplementary text section 1, STS1) was extended with three features, necessary for motif discovery from ChIP-Seq data: accounting for the background nucleotide composition (via KDIC concept), zero-or-one-occurrence-per-sequence (ZOOPS) mode and positional profiles in the initial data (necessary to consider the base coverage in ChIP-Seq data).

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 Kullback–Leibler divergence (Kullback DIC, KDIC) (Kullback and Leibler, 1951):

$$\mathrm{KDIC} = \mathrm{DIC} - \sum_{j=1}^{w} \sum_{\alpha \in \{A,C,G,T\}} x_{\alpha,j} \log q_\alpha$$

where $x_{\alpha,j}$ are elements of the position count matrix (i.e. letter counts), w is the motif length and q_α 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' sublists. 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 words count). 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 W_i for the i-th sequence is its normalized maximal peak value; by normalization the sum of W_i 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:

$$score(\mathrm{PWM},\mathrm{word}) = \sum_{j=1}^{w} S_{\mathrm{word}[j],j} \cdot \mathrm{profile}[j]$$

where w is the word length and $S_{\mathrm{word}[j],j}$ is the PWM element for the j-th letter in the word. Thus, the positions with a larger profile values contribute more

(either positively or negatively) into the PWM score. Then, the positional count weights $x_{\alpha,j}$ for letter α at motif position j are calculated:

$$\alpha \in \{A,C,G,T\}: \; x_{\alpha,j} = \sum_i W_i \cdot \text{profile}_i[j] \cdot \delta(\alpha, \text{word}_i[j])$$

$$\alpha = \mathbf{N}: \; x_{\mathrm{N},j} = \sum_i W_i \cdot (1 - \text{profile}_i[j])$$

$$\forall j: \; \sum_{\alpha \in \{A,C,G,T,\mathbf{N}\}} x_{\alpha,j} = \sum_i W_i \qquad \delta(a,b) = \begin{cases} \text{if } a = b \text{ then } 1 \\ \text{if } a \neq b \text{ then } 0 \end{cases}$$

Here profile$_i$[j] is the value of profile of sequence i at position j, and **N** denotes 'unknown nucleotide' letter. After collecting the data from all words, $x_{N,j}$ values are uniformly split between $x_{\alpha,j}$ ($\alpha \in A,\ C,\ G,\ T$) so the total sum is equal to the number of words.

The convergence depends on the data. If high profile regions cannot be aligned or their composition is similar to the background, different motifs can have very close KDIC values, rendering the convergence poor. Yet, for the real ChIP-Seq data, convergence improves after profile smoothing by substituting the profile values for their maxima within a sliding window with the length equal to the motif length.

3 RESULTS

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. When the segment length was increased MEME and SeSiMCMC, 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, ChIPMunk failed to identify the correct motif in longer segments. Nevertheless, in both modes, ChIPMunk clearly outperformed HMS that took into account peak shapes, both in speed and quality. For NRSF and GABP, the correct motifs were extracted by the all tools in a wider range of sequence lengths, which probably indicates better data quality (see the STS 4).

All in all, here we presented an extended version of our motif discovery tool ChIPMunk that allows using information from strong but also from weaker sites (see STS 10). ChIPMunk is reasonably fast and practically can process datasets of thousands of sequences (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.

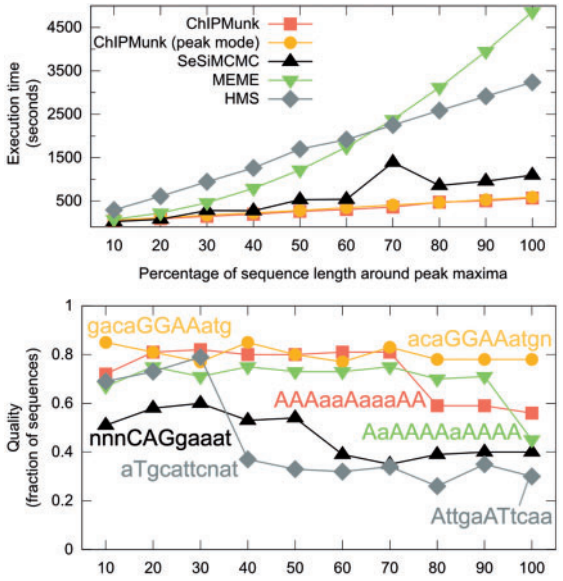

Fig. 1. The time efficiency of motif discovery and resulting quality of the motifs identified in top 500 EWS-FLI1 peaks. Refer to the text for more details. The correct motif consensus gacaGGAAatg is similar to that in Guillon *et al.* (2009).

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

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202– W208.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Favorov,A.V. *et al.* (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, **21**, 2240–2245.

Fejes,A.P. *et al.* (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.

Georgiev,S. *et al.* (2010) Evidence-ranked motif identification. *Genome Biol.*, **11**, R19.

Guillon,N. *et al.* (2009) The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLoS ONE*, **4**, e4932.

Hu,M. *et al.* (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kulakovskiy,I.V. and Makeev,V.J. (2009) Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics*, **54**, 667–674.

Kulakovskiy,I.V. *et al.* (2009) Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics*, **25**, 2318–2325.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Lawrence,C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Pavesi,G. *et al.* (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17** (Suppl. 1), S207– S214.

Valouev,A. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, **5**, 829–834.