Search algorithms for dinucleotide Position Weight Matrices

Marie Mille, Julie Ripoll, Bastien Cazaux, Eric Rivals

Search algorithms for dinucleotide Position Weight Matrices

Marie Mille[1,2], Julie Ripoll[1] and Eric Rivals[1]

[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier - UMR 5506 CNRS, Univ. Montpellier, 860 rue de St Priest, 34095, Montpellier, France

[2] Master Sciences et Numérique pour la Santé, Faculté des Sciences, Univ. Montpellier, Campus Triolet Place Eugène Bataillon, 34095, Montpellier, France

Corresponding author: eric.rivals@lirmm.fr; marie.mille01@etu.umontpellier.fr

Transcription regulation is an important cellular process. Specialized proteins, called Transcription Factors (TF), bind on short, specific, DNA sequences to regulate the expression of nearby genes. The sequences recognized by a TF in the vicinity of different genes are not identical, but similar. One captures the similarity of those binding site in different representations, which are generally called *motifs*. The most widely used sort of motifs are Position Weight Matrices (PWMs) (also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM)). A PWM is built from a multiple alignment of observed binding sequences and capture the observed variation of nucleotides at the different positions. Several databases (JASPAR, TRANSFAC, etc.) collect PWMs for known TFs. Those PWMs are used to scan new DNA sequences to find putative binding sites and possibly to annotate them. In the case of complete genomes, the scanning procedure for many PWMs may last a long time [1].

PWMs assume that the distinct positions of the binding sequence are independent of each other. However, several studies have observed that a mutation at a given position influences the probability of mutation at neighboring positions. To overcome this limitation of PWMs, Kulakovskiy et al. have proposed a more complex sort of motifs, called di-PWMs, which model the frequency of occurrence of dinucleotides in the binding sites (instead of mononucleotides for PWMs) [2]. Their studies show that di-PWMs improve in sensitivity compared to PWMs, and thus produce less false positives when scanning a sequence.

Our aim is to design new search algorithms for di-PWMs, either online or offline, and to implement them. Our online scanning algorithm computes a partial score for some positions in the current window, and estimates the maximum achievable score for the whole window. If this score does not match the user defined threshold, the window can be discarded. Our offline approach works in two steps. As for read mapping, the genome is first preprocessed to produce an index data structure. Then, for any given motif and a threshold score, we enumerate potential matching words (i.e. words whose score lies above the threshold) and search their occurrences in the index in optimal time. The difficulty is to design an efficient enumeration algorithm. Such an approach was developed for PWMs in the MOTIF software [3]. If time suffices, we plan to compare experimentally our implementations to that of an existing search tool for di-PWMs, called SPRY-SARUS (github) [1].

Acknowledgements

We thank the GEM Flagship project funded from Labex NUMEV (ANR-10-LABX-0020) for the internship of M. Mille. JR is supported JR by project FluoRib (INCa grant n°2018-131).
The participation of M. Mille at JOBIM 2021 is granted by the Faculté des Sciences of the University of Montpellier, Master Sciences et Numérique pour la Santé, parcours Bioinformatique, connaissances et données.

References

[1] Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. Finding significant matches of position weight matrices in linear time. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):69–79, January 2011.

[2] Ivan Kulakovsky, Victor Levitsky, Dmitry Oshchepkov, Leonid Bryzgalov, Ilya Vorontsov, and Vsevolod Makeev. From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004, February 2013.

[3] David Martin, Vincent Maillol, and Eric Rivals. Fast and accurate genome-scale identification of dna-binding sites. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 201–205, 2018.