Algorithms for searching dinucleotidic Position Weight Matrices (di-PWM)

Marie Mille[1], Julie Ripoll[1], Bastien Cazaux[1], Eric Rivals[1]

[1]LIRMM, Montpellier University, CNRS, UMR 5506, Montpellier, France
***Corresponding author**: rivals@lirmm.fr

Abstract

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 Matrix (PWM) (also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM)). A PWM is built from a multiple alignment of "true" 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 PWM may last a long time [1].

PWM assume that the distinct positions of the bound sequence are independent of each other. However, several works have observed that a mutation at given position influence the probability of mutation at neighboring positions. To overcome this limitation of PWMs, Kulakovskiy et al. have proposed a more complex sort of motif, 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. Many search algorithms are available for mononucleotidic PWM, but only one exist for di-PWMs [1].

We propose two search algorithms for di-PWMs: the first one is a scanning window algorithm with some adapted speed up trick, the second one is enumeration based. The 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 requested threshold, the window can be discarded. A new precomputed table is provided and compare to a classic LookAheadTable [3].

The enumeration strategy relies on the observation that searching for exact matches is faster than computing window scores. The underlying idea is to first enumerate all *valid words* (i.e., words whose score lies above the user defined score threshold) and their score, then in a second phase to search for the set of valid words using any algorithm that solves the Set Pattern Matching problem [4]. Here

for this sake, we used a Python module that implements the classical Aho-Corasick automaton [5].

We also conducted running time experiments for searching di-PWMs from the HOCOMOCO database [6] with both algorithms, and compared our Python implementations to a tool written in Java, called SPRY-SARUS [1].

Numerous perspectives of this work can be considered, including off-line search of the set of valid words in a precomputed genome index (as done for PWM within the MOTIF software [7]).

A presentation in French of these algorithms can be found in [8]. The di-PWM search algorithms will soon be available as a Python package entitled `dipwmsearch` (which can be installed with PyPI).

Acknowledgments: The internship of Marie Mille was kindly supported by the GEM Flagship project funded from Labex NUMEV (ANR-10-LABX-0020). Julie Ripoll is supported by the INCA project ”FluoRib”. ER thanks the support from the Marie-Curie ITN ”ALgorithms for PAngenome Computational Analysis” (ALPACA).

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 Kulakovskiy, 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, Feb 2013.

[3] Michael Beckstette, Robert Homann, Robert Giegerich, and Stefan Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7(1), Aug 2006.

[4] Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, 1997.

[5] Alfred Aho and Margaret Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18:333–340, 1975.

[6] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, and et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, Nov 2018.

[7] 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.

[8] Marie Mille. Recherche de motifs probabilistes : le cas des Matrices Poids Position dinucléotidiques (di-PWM). Research report, Université de Montpellier, July 2021.