Toivonen, Jarkko; **Taipale, Jussi**; **Ukkonen, Esko**

Seed-driven learning of position probability matrices from large sequence sets. (English) Zbl 1443.92147

Schwartz, Russell (ed.) et al., 17th international workshop on algorithms in bioinformatics, WABI 2017, Boston, MA, USA, August 21–23, 2017. Proceedings. Wadern: Schloss Dagstuhl – Leibniz Zentrum für Informatik. LIPIcs – Leibniz Int. Proc. Inform. 88, Article 25, 13 p. (2017).

Summary: We formulate and analyze a novel seed-driven algorithm SeedHam for PPM learning. To learn a PPM of length ℓ, the algorithm uses the most frequent ℓ-mer of the training data as a seed, and then restricts the learning into the ℓ-mers of training data that belong to a Hamming neighbourhood of the seed. The PPM is constructed from background corrected counts of such ℓ-mers using an algorithm that estimates a product of ℓ categorical distribution from a (non-uniform) Hamming sample. The SeedHam method is intended for PPM learning from large sequence sets (up to hundreds of Mbases) containing enriched motif instances. A variant of the method is introduced that decreases contamination from artefact instances of the motif and thereby allows using larger Hamming neighbourhoods. To solve the motif orientation problem in two-stranded DNA we introduce a novel seed finding rule, based on analysis of the palindromic structure of sequences. Test experiments are reported, that illustrate the relative strengths of different variants of our methods, and show that our algorithm outperforms two popular earlier methods. A C++ implementation of the method is available from `https://github.com/jttoivon/seedham/`.

For the entire collection see [Zbl 1372.68022].

MSC:

92D20 Protein sequences, DNA sequences
92-04 Software, source code, etc. for problems pertaining to biology

Keywords:

motif finding; transcription factor binding site; sequence analysis; Hamming distance; seed

Software:

DECOD; DREME; GitHub; JASPAR; SeedHam; TRANSFAC

Full Text: DOI

References:

[1] Timothy L. Bailey. Dreme: motif discovery in transcription factor chip-seq data. {\it Bioin-} {\it formatics}, 27(12):1653, 2011.

[2] Michael F. Berger, Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston W. Estep, and Martha L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. {\it Nature Biotech.}, 24(11):1429-1435, 2006.

[3] Mathieu Blanchette and Saurabh Sinha. Separating real motifs from their artifacts. {\it Bioin-} {\it formatics}, 17(SUPPL. 1), 2001.

[4] Peter Huggins, Shan Zhong, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives, Marcel H. Schulz, Itamar Simon, and Ziv Bar-Joseph. DECOD: fast and accurate discrim inative DNA motif finding. {\it Bioinformatics}, 27(17):2361, 2011.

[5] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.{\it Genome Res.}, 20(6):861-873, 2010.

[6] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, et al. DNA-binding spe cificities of human transcription factors. {\it Cell}, 152(1-2):327-339, 2013.

[7] Edward M. McCreight. A space-economical suffix tree construction algorithm. {\it J. ACM}, 23(2):262-272, 1976. · Zbl 0329.68042

[8] Arnold R. Oliphant, Christopher J. Brandl, and Kevin Struhl.Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligo nucleotides: analysis of yeast GCN4 protein. {\it Mol. Cell. Biol.}, 9(7):2944-2949, 1989.

[9] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. In silico representation and discovery of transcription factor binding sites. {\it Brief. Bioinformatics}, 5(3):217-236, 2004.

[10] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L. Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. {\it Nat. Methods}, 4(8):651-657, 2007.

[11] Albin Sandelin, Wynand Alkema, Par Engstrom, Wyeth W. Wasserman, and Boris Len hard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. {\it Nucleic Acids Res.}, 32(Database issue):D91-94, 2004.

[12] Gary D Stormo. DNA binding sites: representation and discovery. {\it Bioinformatics}, 16(1):16- 23, 2000.

[13] Gary D. Stormo, Thomas D. Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. {\it Nucleic Acids} {\it Res.}, 10(9):2997-3011, 1982.

[14] Martin Tompa, Nan Li, Timothy L. Bailey, George M. Church, Bart De Moor, Eleazar Eskin, Alexander V. Favorov, Martin C. Frith, Yutao Fu, W. James Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. {\it Nature biotech-} {\it nology}, 23(1):137-144, 2005.

[15] Craig Tuerk and Larry Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. {\it Science}, 249(4968):505-510, 1990.

[16] Esko Ukkonen. On-line construction of suffix trees. {\it Algorithmica}, 14(3):249-260, 1995. · Zbl 0831.68027

[17] Peter Weiner. Linear pattern matching algorithms. In {\it Switching and Automata Theory,} {\it 1973. SWAT'08. IEEE Conference Record of 14th Annual Symposium on}, pages 1-11, 1973.

[18] Edgar Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. {\it Brief. Bioinformatics}, 9(4):326-332, 2008.

[19] :13