

---

## Computing exact p-values for DNA motifs (Part I)

Jing Zhang<sup>1</sup>, Bo Jiang<sup>2</sup>, Ming Li<sup>3,1</sup>, John Tromp<sup>4</sup>, Xuegong Zhang<sup>2</sup>, Michael Q. Zhang<sup>5,2</sup>

<sup>1</sup>State Key Laboratory of Intelligent Technology & System, Department of Computer Science and Technology, Tsinghua University, 100084, China.

<sup>2</sup>MOE Key Laboratory of Bioinformatics, Department of Automation, Tsinghua University, 100084, China.

<sup>3</sup>School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

<sup>4</sup>CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

<sup>5</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11274, USA.

---

### ABSTRACT

**Motivation:** Many heuristic algorithms have been designed to approximate p-values of DNA motifs described by position weight matrices, to evaluate their statistical significance. They often significantly deviate from the true p-value by orders of magnitude. Exact p-value computation is needed for ranking the motifs. Furthermore, surprisingly, the complexity of the problem is unknown.

**Results:** We show the problem to be NP-hard, and present MotifRank, software based on dynamic programming, to calculate exact p-values of motifs. We define the exact p-value on a general and more precise model. Asymptotically, MotifRank is faster than the best exact p-value computing algorithm, and is in fact practical. Our experiments clearly demonstrate that MotifRank significantly improves the accuracy of existing approximation algorithms.

**Availability:** MotifRank is available from <http://bio.dlg.cn>

**Contact:** mzhang@cshl.edu, mli@uwaterloo.ca