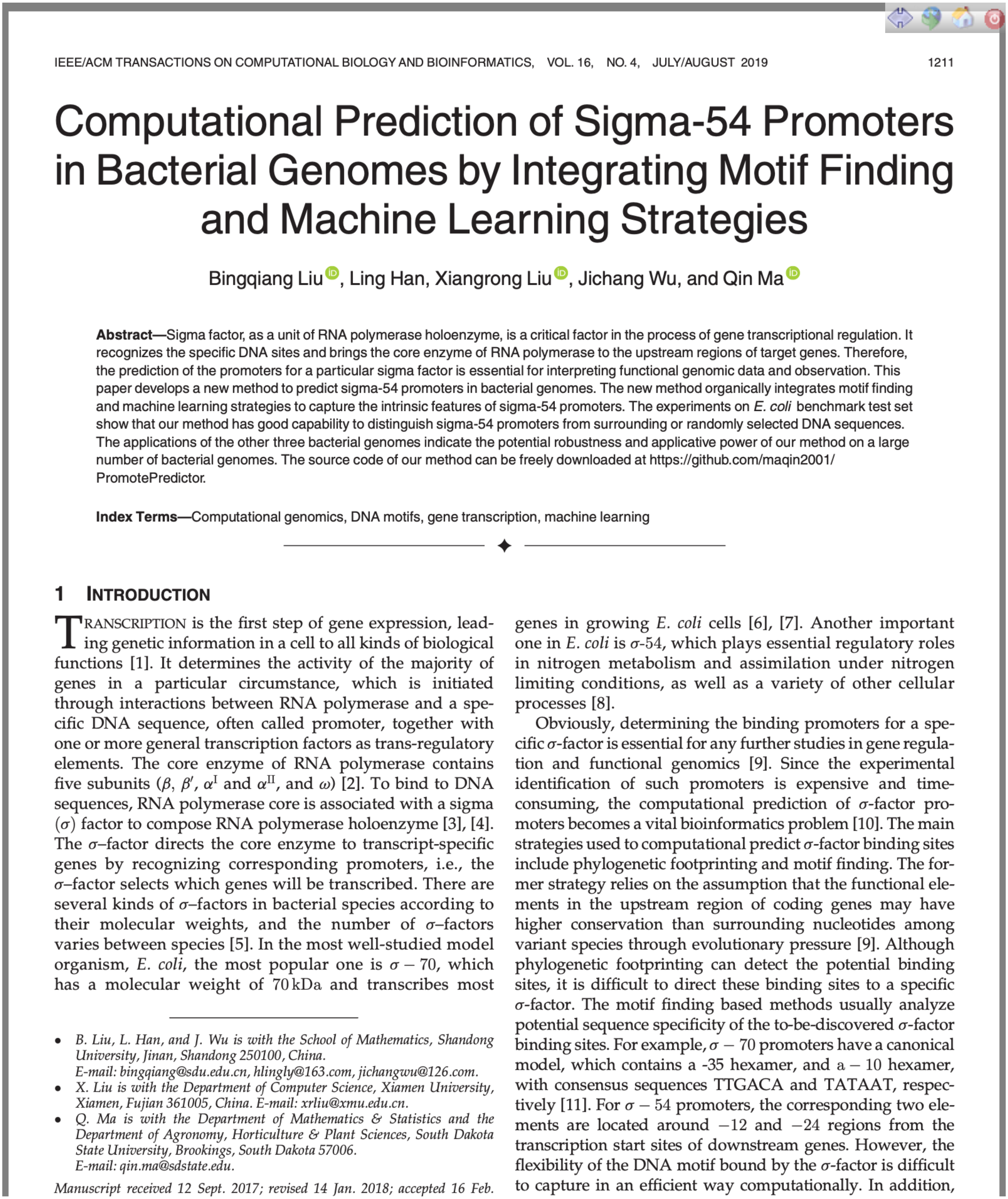

# Computational Prediction of Sigma-54 Promoters in Bacterial Genomes by Integrating Motif Finding and Machine Learning Strategies

Bingqiang Liu, Ling Han, Xiangrong Liu, Jichang Wu, and Qin Ma

**Abstract**—Sigma factor, as a unit of RNA polymerase holoenzyme, is a critical factor in the process of gene transcriptional regulation. It recognizes the specific DNA sites and brings the core enzyme of RNA polymerase to the upstream regions of target genes. Therefore, the prediction of the promoters for a particular sigma factor is essential for interpreting functional genomic data and observation. This paper develops a new method to predict sigma-54 promoters in bacterial genomes. The new method organically integrates motif finding and machine learning strategies to capture the intrinsic features of sigma-54 promoters. The experiments on *E. coli* benchmark test set show that our method has good capability to distinguish sigma-54 promoters from surrounding or randomly selected DNA sequences. The applications of the other three bacterial genomes indicate the potential robustness and applicative power of our method on a large number of bacterial genomes. The source code of our method can be freely downloaded at https://github.com/maqin2001/PromotePredictor.

**Index Terms**—Computational genomics, DNA motifs, gene transcription, machine learning

## 1 Introduction

Transcription is the first step of gene expression, leading genetic information in a cell to all kinds of biological functions [1]. It determines the activity of the majority of genes in a particular circumstance, which is initiated through interactions between RNA polymerase and a specific DNA sequence, often called promoter, together with one or more general transcription factors as trans-regulatory elements. The core enzyme of RNA polymerase contains five subunits ($\beta$, $\beta'$, $\alpha^{\mathrm{I}}$ and $\alpha^{\mathrm{II}}$, and $\omega$) [2]. To bind to DNA sequences, RNA polymerase core is associated with a sigma ($\sigma$) factor to compose RNA polymerase holoenzyme [3], [4]. The $\sigma$–factor directs the core enzyme to transcript-specific genes by recognizing corresponding promoters, i.e., the $\sigma$–factor selects which genes will be transcribed. There are several kinds of $\sigma$–factors in bacterial species according to their molecular weights, and the number of $\sigma$–factors varies between species [5]. In the most well-studied model organism, *E. coli*, the most popular one is $\sigma - 70$, which has a molecular weight of $70\,\mathrm{kDa}$ and transcribes most genes in growing *E. coli* cells [6], [7]. Another important one in *E. coli* is $\sigma$-54, which plays essential regulatory roles in nitrogen metabolism and assimilation under nitrogen limiting conditions, as well as a variety of other cellular processes [8].

Obviously, determining the binding promoters for a specific $\sigma$-factor is essential for any further studies in gene regulation and functional genomics [9]. Since the experimental identification of such promoters is expensive and time-consuming, the computational prediction of $\sigma$-factor promoters becomes a vital bioinformatics problem [10]. The main strategies used to computational predict $\sigma$-factor binding sites include phylogenetic footprinting and motif finding. The former strategy relies on the assumption that the functional elements in the upstream region of coding genes may have higher conservation than surrounding nucleotides among variant species through evolutionary pressure [9]. Although phylogenetic footprinting can detect the potential binding sites, it is difficult to direct these binding sites to a specific $\sigma$-factor. The motif finding based methods usually analyze potential sequence specificity of the to-be-discovered $\sigma$-factor binding sites. For example, $\sigma - 70$ promoters have a canonical model, which contains a -35 hexamer, and $\mathrm{a} - 10$ hexamer, with consensus sequences TTGACA and TATAAT, respectively [11]. For $\sigma - 54$ promoters, the corresponding two elements are located around $-12$ and $-24$ regions from the transcription start sites of downstream genes. However, the flexibility of the DNA motif bound by the $\sigma$-factor is difficult to capture in an efficient way computationally. In addition,

- *B. Liu, L. Han, and J. Wu is with the School of Mathematics, Shandong University, Jinan, Shandong 250100, China.*
  *E-mail: bingqiang@sdu.edu.cn, hlingly@163.com, jichangwu@126.com.*
- *X. Liu is with the Department of Computer Science, Xiamen University, Xiamen, Fujian 361005, China. E-mail: xrliu@xmu.edu.cn.*
- *Q. Ma is with the Department of Mathematics & Statistics and the Department of Agronomy, Horticulture & Plant Sciences, South Dakota State University, Brookings, South Dakota 57006.*
  *E-mail: qin.ma@sdstate.edu.*

*Manuscript received 12 Sept. 2017; revised 14 Jan. 2018; accepted 16 Feb.*