P39

Modeling DNA affinity landscape through two-round support vector regression with weighted degree kernels

Xiaolei Wang 1,2, Hiroyuki Kuwahara 1,2, and Xin Gao 1,2

1 Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia

2 Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia

Abstract

Background

A quantitative understanding of interactions between transcription factors (TFs) and their DNA binding sites is key to the rational design of gene regulatory networks. Recent advances in high-throughput technologies have enabled high-resolution measurements of protein-DNA binding affinity.

Importantly, such experiments revealed the complex nature of TF-DNA interactions, whereby the effects of nucleotide changes on the binding affinity were observed to be context dependent. A systematic method to give high-quality estimates of such complex affinity landscapes is, thus, essential to the control of gene expression and the advance of synthetic biology.

Results

Here, we propose a two-round prediction method that is based on support vector regression (SVR) with weighted degree (WD) kernels. In the first round, a WD kernel with shifts and mismatches is used with SVR to detect the importance of subsequences with different lengths at different positions. The subsequences identified as important in the first round are then fed into a second WD kernel to fit the experimentally measured affinities. To our knowledge, this is the first attempt to increase the accuracy of the affinity prediction by applying two rounds of string kernels and by identifying a small number of crucial k-mers.

The proposed method was tested by predicting the binding affinity landscape of Gcn4p in *Saccharomyces cerevisiae* using datasets from HiTS-FLIP. Our method explicitly identified important subsequences and showed significant performance improvements when compared with other state-of-the-art methods. Based on the identified important subsequences, we discovered two surprisingly stable 10-mers and one sensitive 10-mer which were not reported before. Further test on four other TFs in *S. cerevisiae* demonstrated the generality of our method.

Conclusion

We proposed in this paper a two-round method to quantitatively model the DNA binding affinity landscape. Since the ability to modify genetic parts to fine-tune gene expression rates is crucial to

the design of biological systems, such a tool may play an important role in the success of synthetic biology going forward.