TFBS Prediction with Stochastic Differential Equation and Time Series

Mohsen Salehi*, Adel Mohammadpour† and Mina Aminghafari**‡

*mohsen_salehi453@aut.ac.ir
†adel@aut.ac.ir
**aminghafari@aut.ac.ir
‡Department of Statistics, Faculty of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic), No. 424, Hafez Ave. Tehran, Iran

Abstract. In molecular biology and genetics, the transcription factor binding sites (TFBS) are the regions on DNA caused a gene is expressed. Prediction of these regions is crucial for them. Several studies have been done, such as applying position weight matrix (PWM) and logistic regression (LR)[1, 2], to distinguish true binding regions from random ones. We considered the Chromosome 1 and tried to use the time series and stochastic differential equation to improve the predictions. We were interested to use the distance of binding sites from each other to predict TFBS regions. In chromosome 1, we dealt with two types binding site, 5’ to 3’ and 3’ to 5’ binding sites. We plotted them and realized that the patterns of them are different, so we considered three features for our study. At first, we worked with total of binding sites, regardless of type of them, then we did on 5’ to 3’ binding sites and finally, on 3’ to 5’ binding sites. We used two approach, time series (TS) [4] and stochastic differential equation (SDE)[3], to forecast better predictions of binding sites rather than applying PWM and LR. In the time series method, we used of Fourier series to find a pattern on distances of BS’s, then in SDE method, we considered the distances of BS’s as a stochastic process to predict them. We compared our results to those using PWM and LR. The results show that SDE method forecast TFBS’s better than TS and applying these two method can predict TFBS regions more successfully than PWM and LR.

Keywords: Stochastic Differential Equation; Time Series; PWM; LR

PACS: 39A50,37M10,92D99

REFERENCES