Deep learning reveals the general rules shaping the potential energy landscape of DNA methylation patterns

Dohoon Lee[1], Sun Kim[2,3,4]

[1]Bioinformatics Institute, Seoul National University, Seoul, 08840, Republic of Korea,
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, 08840, Republic of Korea,
[3]Department of Computer Science and Engineering, Seoul National University, Seoul, 08840, Republic of Korea and
[4]Institute of Engineering Research, Seoul National University, Seoul, 08840, Republic of Korea

DNA methylation (DNAm) is one of the core epigenetic modifications involved in several fundamental cellular processes such as development, aging, and carcinogenesis. Recently, the cell-to-cell variation of DNAm, which occurs due to the imperfect somatic inheritance of DNAm states, has been actively studied for its biological and clinical implications. For example, the intratumoral heterogeneity of DNAm patterns (DMPs) is associated with adverse outcomes through increased cellular diversity and adaptive potential of a tumor. However, most of the precise mechanisms regulating the extent of the heterogeneity of DMPs remains unclear. One promising approach is to introduce the concept of 'potential energy' for DMPs, defined by the genomic and epigenomic context, and suppose that the distribution of DMPs follows Boltzmann-Gibbs distribution upon the potential energy landscape. Having the problem reduced to find a general function that translates the local biological context to potential energy levels, we train a deep neural network model that takes genome sequence, histone marks, and chromatin accessibility as input features and produces the potential energy levels of DMPs as output. Dissecting and interpreting the trained model lets us prioritize the genomic or epigenomic elements, such as DNA motifs and histone marks, based on their contribution to shaping the potential energy landscape of DMPs. Furthermore, we develop a web application that allows experimental biologists to freely conduct in silico mutagenesis experiments using our model. We therefore expect that our model will serve as a powerful, versatile, and unbiased hypothesis-generating tool for DNAm studies.