Supervised learning method for predicting chromatin boundary associated insulator elements

Pawel Badnarz and Bartek Wilczynski

Institute of Informatics, Warsaw University, Poland

Abstract

In eukaryotic cells, the DNA material is densely packed inside the nucleus in the form of a DNA-protein complex structure called chromatin. Since the actual conformation of the chromatin fibre defines the possible regulatory interactions between genes and their regulatory elements, it is very important to understand the mechanisms governing folding of chromatin. In this paper, we show that supervised methods for predicting chromatin boundary elements are much more effective than the currently popular unsupervised methods. Using boundary locations from published Hi-C experiments and modEncode tracks as features, we can tell the insulator elements from randomly selected background sequences with great accuracy. In addition to accurate predictions of the training boundary elements, our classifiers make new predictions. Many of them correspond to the locations of known insulator elements. The key features used for predicting boundary elements do not depend on the prediction method. Because of its miniscule size, chromatin state cannot be measured directly, we need to rely on indirect measurements, such as ChIP-Seq and fill in the gaps with computational models. Our results show that currently, at least in the model organisms, where we have many measurements including ChIP-Seq and Hi-C, we can make accurate predictions of insulator positions.