

データマイニングによるヒストンの機能解析

The functional analysis of histones by data mining

京都大学化学研究所バイオインフォマティクスセンター 夏目 やよい

背景と目的

本研究の目的は、ショウジョウバエの発生時においてヒストン修飾のパターンがどのように変化していくのか、その変化が発生をどのように調節しているのかを明らかにすることである。現在、Drosophila Regulatory Elements modENCODE project によってショウジョウバエの卵から成虫までの12段階においてヒストン修飾と遺伝子発現を網羅的に調べた巨大なデータセット（以下 modENCODE 発生データ）を用いた解析をすすめている。このデータセットを解析するための手法を構築し、適した解析条件を探索する上で、ゲノムワイドな解析を始める前により小さなデータセットを用いて既知のヒストン修飾変化パターンの検出を試みることは有用であると考えられた。そこで、まず初めにヒストンメチル基転移酵素 Ash1 に注目することにした。

検討内容

①modENCODE 発生データ (ChIP-seq) 前処理

NGS によって検出されるフラグメントのマッピングとピーク検出のためのソフトウェアは SICER (Bioinformatics (2009) 25 (15): 1952-1958) を用いた。設定したパラメータは以下の通り。ウィンドウサイズ: 200bp (デフォルト、ヌクレオソーム1単位に相当)、ギャップサイズ: 0bp (ショウジョウバエの遺伝子密度が高いため)。

②Ash1 の結合領域

Schwartz YB らによる Ash1 の ChIP-chip データ (Sg4 細胞、GSE18176) より、Ash1 の結合領域を検出した。ソフトウェアは rMAT (R Bioconductor for Affymetrix tiling array, Bioinformatics (2010) 26 (5): 678-679.)を用い、FDR 0.01 を閾値として Ash1 の結合の有無を判定した。

③隠れマルコフモデル (HMM) による学習

Yoneya らによる変法隠れマルコフモデル (Bioinformatics (2007) 23(7):842-9.)を基に解析をおこなった。②で検出した領域を①から抜き取り、Ash1 の結合領域1つを1単位としてヒストン修飾の度合いを数値化した。このとき ChIP-seq データにおいてフラグメント数を定量的に扱うことは適切ではないと考え、i) $-\log(\text{p-value})$ をヒストン修飾の度合いのスコアとして ChIP-seq の各ウィンドウで計算し (p-value: ChIP-seq のピークが偶然に起こる確率)、Ash1 の結合領域におけるスコアは $-\log(\text{p-value})$ の平均をとる方法と、ii) ChIP-seq の各ウィンドウの p-value が閾値を下回る場合にはヒストン修飾有としてスコアを1、上回る場合にはヒストン修飾無としてスコアを0.001として、Ash1 の結合領域における平均スコアをとる方法を検討した。なお、ii)の閾値には $1\text{E-}05$ を用いた。

結果・考察

Ash1 結合領域は 531 か所検出された。modENCODE 発生データに含まれる Embryo 0-4h、4-8h、8-12h、12-16h、16-24h、Larvae 1、L2、L3、Pupae の 10 段階のデータを用いて解析をおこなったところ、i)の方法ではスコアの変動による影響が大きく、遷移状態数（ヒストン修飾が変化する段階の数）を変えても予想した結果を得ることができなかった。そこで、よりスコアそのものの変動に頑健であることが期待される ii)の方法を用いたところ、遷移状態数 3 のときに①H3K27Ac のピーク後に H3K27me3 のピークが現れる、②H3K27me3 のピーク後に H3K4me3 のピークが現れるといった予想結果と近い結果が得られた。特に①の結果は遷移状態数を 4 に変えても同様に得られた。②については遷移状態数を変えると異なる傾向が得られたことから、ii)の前処理でより感度・精度の高い解析手法に改良する余地があると言える。