平成２５年度　京都大学化学研究所スーパーコンピュータシステム　利用報告書

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胚から成虫までの 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（GSE15292 他、以下 modENCODE データ）を用いた解析をすすめている。これまでに、Ash1 と Pol II の結合領域におけるヒストン修飾の動的変化に絞ったサブ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 preliminary analysis をおこない、期待される結果を得ることに成功した。一方で、充分な感度で検出ができなかった変化パターンがあったことから、より感度・精度の高い解析手法に改良する必要があると判断した。そのため、近年機械学習の分野で注目を集めている spectral learning (*Journal of Computer and System Sciences* (2012)78(5):1460-1480)によって各観測値の出現確率と各種パラメータ数学的に計算するアプローチをとることにし、人工データを用いて動作確認をおこなった。

検討内容

離散値を扱う HMM（dHMM）

解析手法、人工データは前回の領域会議で報告したものを用いた。HMM を用いた解析をおこなうに際して、下記のような解析条件で検討をおこなった。設定した閾値は、パラメータとして調節可能である。

① データ 1、データ 2 を閾値（ここでは 1）以上であれば 1、閾値未満であれば 0 と変換してバイナリ行列を得た。

②ヒストン修飾の変化パターンが観測される確率を計算する際に用いる計算式において、分母が閾値（ここでは 0）未満であればその観測値は信頼できないとして無視した。

③発生段階が次の段階に進む際にヒストン修飾の変化が起こった場合、条件確率は低くなる。ヒストン修飾の変化が起きたとみなす閾値を 0.3 とした。

連続値を扱う HMM（cHMM）

① データ 1、データ 2 を平均 2、分散 1 の分布をとるように変換した。

② カーネル密度推定（カーネル関数 : ガウシアン）と呼ばれる方法を導入して、バイナリ変換することなく計算を続けた。

③ その他の閾値などは dHMM と同じにした。

比較結果

SD1、SD2 は、上記のバイナリ変換後のピーク例のように data1 のピーク（赤）が data2 のピーク（青）よりも先に現れるように作成されている。dHMM では cHMM と比較してより明確に data1 と data2 のピーク差を検出することができたため、dHMM を採用することとした。

ChIP-seq データのプロセシング手順

ヒストン修飾の有無をピーク検出ソフトウェアで決定したのちに、ヒストン修飾が起きていることを示すピークを①一番 TSS が近い遺伝子、または②オーバーラップしている遺伝子（必ずしも一番 TSS が近い遺伝子ではない）にアサインすることにより、各遺伝子およびその近傍でヒストン修飾の有無を 0/1 で表した行列（以後、ヒストン修飾カタログ）をあらかじめ作成した。更に、Ash1 および PolII の ChIP-chip データ（GSE18176）から MAT を用いて Ash1 および PolII の結合部位を検出した際にも、同じ基準で結合部位を遺伝子にアサインし、得られた遺伝子リスト分をヒストン修飾カタログから抜き出して以降の解析に用いた。

結果・考察

ChIP-chip のデータ解析の結果、Ash1 結合部位は 149 個の遺伝子に、PolII 結合部位は 5643 個の遺伝子にアサインされた。Ash1 結合部位近傍遺伝子 149 個のうち、PolII 結合が検出された 123 個の遺伝子分をヒストン修飾カタログから抜き出し、dHMM で解析した。その結果、前回同様に H3K4me3 のピークが H3K27me3 よりも先に現れるパターンが検出されたほか、以前検出することができなかったパターン（H3K27ac→H3K27me3）を検出することができた。