

# 研究報告書

## 「発生を制御するヒストン修飾動態の in silico 解析」

研究タイプ: 通常型

研究期間: 平成22年10月～平成27年3月

研究者: 夏目 やよい

### 1. 研究のねらい

DNA はヒストンと呼ばれるタンパク質によって巻き取られており、ヌクレオソームと呼ばれるコンパクトな構造を作って核内に格納されている。ヒストンは単体ではなく、4種類のヒストン (H2A, H2B, H3, H4) が2つずつ集まって形成した8量体を一つの単位として機能しているが、このヒストンのアミノ酸残基が修飾を受けることで転写などの生命現象が制御されていることが明らかになってきている。例えば、H3K4me3 (H3 の4番目のリジンがトリメチル化を受けている) や H3K27ac (H3 の27番目のリジンがアセチル化を受けている) は、転写がおこなわれている領域で見られるのに対して、H3K27me3 (H3 の27番目のリジンがトリメチル化を受けている) は転写が不活化されている領域で見られる。しかし、ヒストン修飾は動的な変化であり、また種類や修飾部位も豊富であるため、未だ不明な点が多い。

近年における次世代シーケンサーの普及は生物学研究に大きな影響を与えた。例えば、クロマチン免疫沈降法と次世代シーケンサーを組み合わせた ChIP-seq によって、興味のあるタンパク質がゲノム上のどこにあるのかを検出することができる。また、RNA の大量並列シーケンシングによって、どの転写産物がどのくらいあるのかを定量的に検出することができる。これらの実験手法の普及に伴って大量のデータが公開されるようになった。modENCODE プロジェクト(ショウジョウバエを含むモデル動物であらゆる配列に関わる情報を収集することを目指している)によってショウジョウバエが卵から成虫になるまでの発生段階でどのヒストン修飾がゲノム上のどこで起きているのか、そのとき転写産物はどのくらいあるのかといった実験結果が入手可能となった。しかし、全ゲノム解析は一つの実験データでも膨大な情報量を含み、複数の異なる実験データを統合して一つの仮説を得るプロセスは困難を極める。そこで本研究では、大量のデータから意味のある情報を探し出すことを目的として発展した情報科学の機械学習と呼ばれる分野に注目した。modENCODE プロジェクトの ChIP-seq データ、RNA-seq データ、更に他のデータベースといった異なる性質の情報統合に機械学習の手法を適用し、ヒストン修飾がショウジョウバエ発生に伴ってどのように変化していくのか(ヒストン修飾の動的変化パターンの検出)、その生物学的意義は何なのか(ヒストン修飾の動的変化に関与する非コード RNA の探索)を明らかにすることを目標に定めた。

### 2. 研究成果

#### (1) 概要

Spectral learning を用いた隠れマルコフモデル (Daniel H et al., *Journal of Computer and System Sciences* (2012) 78 (5): 1460-1480) を応用して、ChIP-seq の時系列データからヒストン修飾の動的変化パターンを検出する解析手法を構築した。更に、ASH1 標的遺伝子における既知の動的変化を検出することに成功した。この手法を用いたゲノムワイド解析をおこなう

ためには、あらかじめ類似パターンを持つと考えられる遺伝子を大まかに分類する必要がある。そこで、その分類基準としてコアプロモーターエレメントの組み合わせを選択した。コアプロモーターとは転写に最低限必要な DNA 配列を指し、この領域内に存在する生物間でよく保存された DNA モチーフをコアプロモーターエレメントという。ショウジョウバエのコアプロモーターエレメントのデータベースから、コアプロモーターエレメント(TATA box、DPE)の有無に従って分類されたコアプロモーターの DNA 配列を入手し、ChIP-seq データからヒストン修飾(H3K4me3、H3K27ac、H3K27me3)が起きている割合をコアプロモーター領域およびその転写領域で計算した。このヒストン修飾の割合が時系列でどのように変化するかを比較したところ、TATA box を持たないグループでは経時的变化が小さいのに対して、TATA box を持つグループでは変化が大きいことが示された。また、RNA-seq データと統合して回帰分析をおこなった結果、TATA box を持たないグループではヒストン修飾の割合で RNA 量を説明できることが示されたのに対して、TATA box を持つグループではヒストン修飾で RNA 量を予測できない群が混ざっていた。更に、各グループ間でヒストン修飾の相対的な頻度を比較した結果、DPE を持つグループでは H3K27me3 の使用頻度が高いことが示された。この結果をもとに、既に構築した解析手法を用いて、TATA box を持ちヒストン修飾が起きているコアプロモーターにおけるヒストン修飾の動的変化パターンの解析を進めようとしている。

## (2) 詳細

### 研究テーマ A「modENCODE データのプロセッシング」

第1年度実施予定であり、研究テーマ B の準備にあたる。用いた modENCODE データセット(GSE15292)には、810 の ChIP-chip データ、114 の ChIP-seq データ、12 の RNA-seq データが含まれている。これらのうち、H3K4me3、H3K9ac、H3K9me3、H3K27ac、H3K27me3、RNA polymerase II の全 ChIP-seq データと RNA-seq データの生データをダウンロードし、リードのクオリティチェック(ChIP-seq、RNA-seq)やマッピング(ChIP-seq、RNA-seq)、ピーク検出(ChIP-seq)、転写産物の定量(RNA-seq)をおこなうための条件検討をおこなった。また、ChIP-chip データについてはプロセッシング済みデータをダウンロードし、ChIP-seq データとの統合を図ったが、条件検討の末に ChIP-seq データのみを用いることとした。

### 研究テーマ B「発生段階におけるヒストン修飾の動的変化の検出」

まず、ChIP-seq の時系列データを用いて、ヒストン修飾の動的変化パターンを検出する解析手法を構築した。初めに、既に遺伝子発現の動的変化パターンを検出する目的で開発・報告されている変法隠れマルコフモデル(Yoneya T and Mamitsuka H, *Bioinformatics* (2007) 23 (7): 842-849.)を用いて、ASH1 標的遺伝子における既知の動的変化を検出することができることを確認した。次に、更に精度を改良するため、spectral learning を用いた隠れマルコフモデル(Daniel H et al., *Journal of Computer and System Sciences* (2012) 78 (5): 1460-1480)を応用して新規の手法を開発した。これについても既知の動的変化を検出することに成功し、精度が上がったことを確認することができたため、ゲノムワイド解析に取りかかることにした。これに際して、あらかじめ類似パターンを持つと考えられる遺伝子を大まかに分類する必要がある、その分類基準として RNA-seq の時系列データにおける遺伝子発現の変動を予定していたがクラスタリングが期待通りの結果にならなかった。そこで、別の分類基準の候補としてコア

プロモーター要素の組み合わせを選択した。コアプロモーターとは転写に最低限必要な DNA 配列を指し、この領域内に存在する生物間でよく保存された DNA モチーフをコアプロモーター要素という。ショウジョウバエのコアプロモーター要素のデータベースから、コアプロモーター要素 (TATA box、DPE) の有無に従って分類されたコアプロモーターの DNA 配列を入手し、ChIP-seq データからヒストン修飾 (H3K4me3、H3K27ac、H3K27me3) が起きている割合をコアプロモーター領域および転写領域で計算した。得られたグループは、i) Inr グループ: TATA box も DPE も持たず、Inr のみを有するコアプロモーターのグループ、ii) DPE グループ: DPE と Inr を持ち、TATA box は持たないグループ、iii) TATA グループ: TATA box と Inr を持ち、DPE は持たないグループ、iv) TATA-DPE グループ: TATA box、DPE、Inr の全てを持つグループ、の4通りである。これらのグループにおいてヒストン修飾の割合が時系列でどのように変化するかを比較したところ、TATA box を持たないグループでは経時変化が小さいのに対して、TATA box を持つグループでは変化が大きいことが示された。

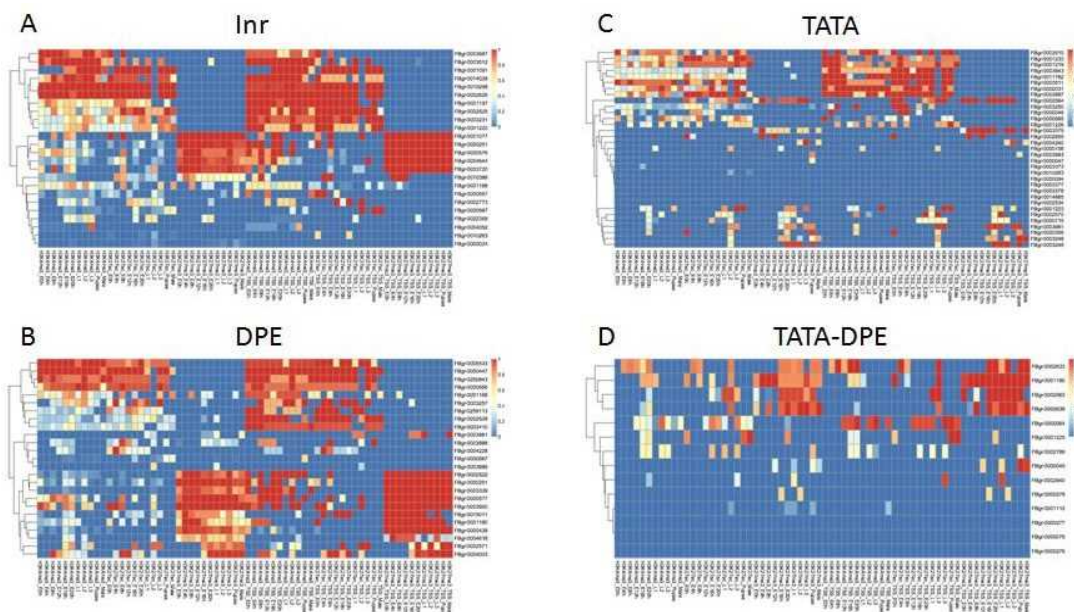


図: 各コアプロモーター要素グループにおけるヒストン修飾の割合

縦軸にコアプロモーターがアサインされている遺伝子、横軸にヒストン修飾 (左から順に H3K4me3 (転写領域) の胚 0-4 時間～成虫雄、H3K27ac (転写領域) の胚 0-4 時間～成虫雄、H3K27me3 (転写領域) の胚 0-4 時間～成虫雄、H3K4me3 (コアプロモーター領域) の胚 0-4 時間～成虫雄、H3K27ac (コアプロモーター領域) の胚 0-4 時間～成虫雄、H3K27me3 (コアプロモーター領域) の胚 0-4 時間～成虫雄)。赤いほどヒストン修飾が起きている領域の割合が高い。Inr グループと DPE グループでは発生段階の進行に伴う変化が小さいことがわかる。

また、RNA-seq データと統合して回帰分析 (目的変数: RNA 発現量、説明変数: ヒストン修飾の割合) をおこなった結果、TATA box を持たないグループではクロスバリデーションにおいて測定値 (RNA-seq の実験結果) と予測値 (回帰式にヒストン修飾の割合を代入して得られた値) の間に期待通り正の相関があり、ヒストン修飾の割合で RNA 量を説明できることが示されたのに対して、TATA box を持つグループでは上記のような群の他に、ヒストン修飾が検出

されていないにも関わらず測定値が予測値から大幅に外れる群が混ざっていた。更に、各グループ間でヒストン修飾の相対的な頻度を比較した結果、DPE を持つグループではH3K27me3の使用頻度が高いことが示された。この結果をもとに、既に構築した解析手法を用いて、TATA box を持ちヒストン修飾が起きているコアプロモーターにおけるヒストン修飾の動的变化パターンの解析を進めようとしている。

#### 研究テーマ C「ヒストン修飾の動的变化に関与する非コード RNA の探索」

残念ながらさがけ研究期間中に研究テーマ C に至ることができなかった。研究テーマ B で開発した解析手法はこのテーマも念頭に入れてつくったものであるため、研究テーマ C についても今後の目標にしたい。

### 3. 今後の展開

今後は、これまでに見いだしたヒストン修飾の動的变化と CPE の関連性をふまえ、既に構築した解析手法を用いて、ゲノムワイドにヒストン修飾が起きている TATA グループにおけるヒストン修飾の割合からヒストン修飾動的变化パターンの検出とその生物学的意義の解明を目指す。また、TATA box を持つコアプロモーターで見られたヒストン修飾非依存的な転写とも言える現象は実際にどのような意味を持つのかなど、ヒストン修飾と CPE の関係性についてより詳細に調べたい。

### 4. 評価

#### (1) 自己評価

##### (研究者)

当初の研究構想は、研究テーマ A: ChIP-seq データと RNA-seq データのプロセッシング、研究テーマ B: ショウジョウバエ発生段階におけるヒストン修飾の変化を検出、研究テーマ C: B で見いだしたヒストン修飾の変化に関与する非コード RNA の探索となっていた。

現時点までに、研究テーマ B に必要な新規解析手法の開発が終了し、ゲノムワイド解析の条件を見つけ出す段階まで達成した。その過程でコアプロモーターエレメントとヒストン修飾の関連性といった新規性の高い発見をすることができた。この発見については現在論文投稿準備中であり、研究テーマ B のゲノムワイド解析の方向性についても見いだすことができた。さがけ研究期間中に研究目標を達成することはできなかったが、提案した研究テーマが軌道に乗ってきたことについては評価したい。研究費は主にコンピュータと出張費に使用した。受け入れ研究室では情報科学の議論、さがけ領域では生物学の議論の機会に恵まれ、更に海外や国内の学会に参加する機会を多く与えられて情報生命科学の議論もすることができた。本研究のアプローチは類似研究が少なく、期待される研究成果をあげることによって今後コンピュータ解析を積極的に取り入れた研究の有効性を示すことができると考えている。これまでの手厚いサポートにふさわしい研究成果につなげていきたい。



(2) 研究総括評価(本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った)。

(研究総括)

隠れマルコフモデルを用いる前に予め類似パターンを持つと考えられる遺伝子を分類し、その基準としてコアプロモーターエレメントの組み合わせを選択した。その結果、TATA boxを持たないグループではヒストン修飾の経時的変化が小さいのに対し、TATA boxを持つグループでは変化が大きいこと、また、TATA boxを持たないグループではヒストン修飾の割合でRNA量が説明できることなどが分かった。さらに、DPEを持つグループではH3K27me3の使用頻度が高いことも分かった。これらの結果をもとに、TATA boxを持ち、ヒストン修飾が起きているコアプロモーターにおけるヒストン修飾の動的パターンの解析を進めことになる。長い試行錯誤の結果、新規解析手法の開発が終了し、ゲノムワイド解析の条件を見つけ出す段階に達した。隠れマルコフモデルによる機会学習により法則性を見つけたり予測したりするのはまだこれからである。数理解析や情報解析は今後医学生物学初め様々な分野で益々重要になる中、明確な目標を掲げ新規の発見に導くような手法を開発してくれることを期待します。

## 5. 主な研究成果リスト

(1) 論文(原著論文)発表

(2) 特許出願

研究期間累積件数: 0件

(2) その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

- ・ 「*In silico* analysis of histone modification dynamics at Ash1 binding regions in *D.melanogaster*」 13<sup>th</sup> Annual International Workshop on Bioinformatics and Systems Biology 2013年7月31日ー8月2日、京都(日本)
- ・ 「Detection of timing-differences in histone modification at Ash1 target genes」 ECOMB/ISCB 2013 Conference on Regulatory & Systems Genomics with DREAM Challenges 2013年11月8ー12日、トロント(カナダ)
- ・ 「Detection of histone modification dynamics at Ash1 target genes」 Next Generation Sequencing Conference (NGS) 2014 2014年6月2-4日、バルセロナ(スペイン)
- ・ 京都大学化学研究所広報誌「黄檗第39号」研究トピックス(さきがけプロジェクト研究活動紹介)