様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１３９０１
研究種目：若手研究(B)
研究期間：2013～2015
課題番号：２５７３０１７３

研究課題名（和文）高次エピゲノムが生み出す生命情報を読み解く統計解析法の開発

研究課題名（英文）Development of a statistical approach to decipher information from high-order chromatin structure and the epigenome

研究代表者

島村　徹平（Shimamura, Teppei）

名古屋大学・医学（系）研究科（研究院）・特任准教授

研究者番号：００６２３９４３

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究課題では、クロマチン相互作用を網羅的に解析するChIA-PET法により得られた高次エピゲノムデータから、２領域間のクロマチン交互作用を予測するベイズモデルを構築し、高次エピゲノムによる遺伝子発現制御機構を解明するための統計解析法を開発した。具体的には、以下の成果が得られた。（１）ChIA-PETシークエンスタグのリファレンスゲノムへの確率的アライメント、（２）ポアソンモデルによるバイアスの補正、（３）負の二項分布の混合ベイズモデルによる相互作用領域の検出、（４）細胞特異的、シグナル伝達下で変化する転写ファクトリーの同定、（５）ChIA-PETデータ解析パイプラインの開発。

研究成果の概要（英文）：In this study, we develop statistical methods for predicting chromatin interaction regions from high-order epigenome information obtained from chromosome conformation capture techniques such as ChIA-PET and hi-C.
The outline of research results is described below. 1) probabilistic alignment for ChIA-PET sequencing tags to reference genome; 2) bias correction of ChIA-PET count data with Poisson regression model; 3) detection of chromatin interaction regions with Bayesian mixture model of negative binomial distributions; 4) detection of cell-type specific or signal dependent transcription factories; 5) development of pipelines for analyzing ChIA-PET sequencing data.

研究分野：バイオインフォマティクス

キーワード：バイオインフォマティクス　ベイズモデル　次世代シークエンサー　高次クロマチン構造　エピゲノム

１．研究開始当初の背景

染色体はゲノム DNA とその上で機能する数千ものタンパク質から構成される複合体であり、さまざまな生命現象における遺伝子発現制御機構において、染色体の化学的制御と物理的制御が重要な役割を果たしている。この制御を担う本態は、DNA メチル化、ヒストン修飾、DNA とタンパク質の複合体であるクロマチンによって形成されるエピゲノムである。近年の高速シークエンス技術の画期的な向上によって、エピゲノムの網羅的な解析が可能となり、DNA メチル化やヒストン修飾による化学的制御機構やクロマチン一次構造による物理的制御機構の一部については徐々に明らかにされつつある。しかしこれらの一連の解析によって得られる情報の多くはゲノム一次配列に紐付けられた一次元情報であり、エピゲノムによる遺伝子発現制御の全体像に迫るためには、より高次のエピゲノム情報を解析する必要がある。

２．研究の目的

本研究では、クロマチン相互作用を網羅的に解析する ChIA-PET 法により得られた高次エピゲノムデータから、２領域間のクロマチン交互作用を予測するベイズモデルを構築し、高次エピゲノムによる遺伝子発現制御機構を解明するための統計解析法を開発する。具体的には、下記の目標を設定し、これを達成する。

（１）．２領域間のクロマチン相互作用を予測するベイズモデルの構築

ChIA-PET 法から得られたリード配列をリファレンス配列へマッピングする際には、アダプタ処理後のリード配列が短いことに起因して、リード配列がゲノム上の複数にヒットするマルチヒットの問題が生じる（図１）。この場合、リード配列がゲノム上の配列にユニークにヒットする領域に限定して解析を行う方法が考えられるが、相当数のリード配列を取り除いてしまうため、検出できない偽陰性のクロマチン相互作用の数が増加する。逆にマルチヒットするリード配列をそのまま含めて解析を行うと、実際には結合しない偽陽性のクロマチン相互作用の数が増加する。この問題を改善するため、リード配列がマッピングされる過程において、その結合確率を取り入れた確率モデルで表現する。モデルのパラメータは、一次構造体のエピゲノム情報（タンパク結合、ヒストン修飾、クロマチン一次構造など）のほか、共局在情報など事前確率として取り入れた MAP 推定で行う。構築されたベイズモデルの事後確率に従い、擬陽性の高いクロマチン相互作用の数を減らす。

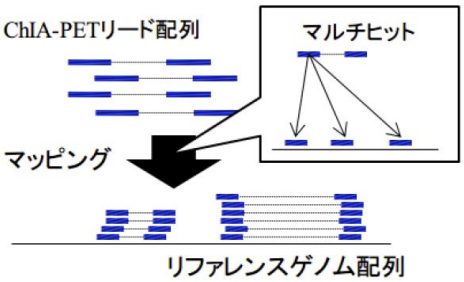

図１：マルチヒットの問題

（２）．クロマチン隣接行列からクロマチン高次構造を推定する統計解析法の開発

クロマチン高次構造の可視化は、クロマチンループなどの二次構造体、転写ファクトリーやヘテロクロマチンなどの三次構造体に関するエピゲノム機構を解明する上で極めて重要である。本研究では、ChIA-PET 法によって得られるクロマチン隣接行列を基に、クロマチン三次構造を多次元尺度構成法によって推定する方法を開発する。

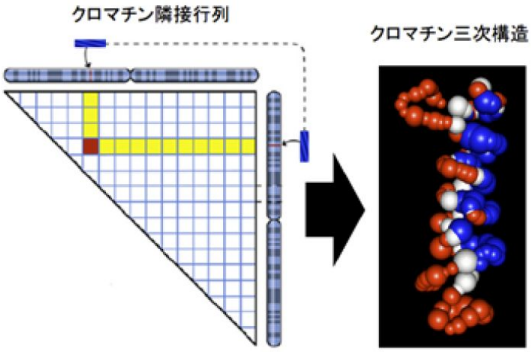

図２：クロマチン高次構造の推定

（３）．転写ファクトリーの同定

染色体は核のなかでクロムソームテリトリーとよばれる領域を形づくって局在し、その間にあるインタークロマチンコンパートメントには活発な転写が生じる構造体があり転写ファクトリーとよばれる RNA ポリメラーゼ II を組織化している巨大な複合体が存在することが 3C アッセイ技術（核内で 3 次元的に近接する領域を網羅的に検出する技術）によって明らかになりつつある（図３）。本研究では、ChIA-PET データと遺伝子発現データを組み合わせることによって転写ファクトリーを同定し、構造の違いがどの程度転写ファクトリーに存在する遺伝子の発現に影響を与えるかを調べる統計解析手法を開発する。

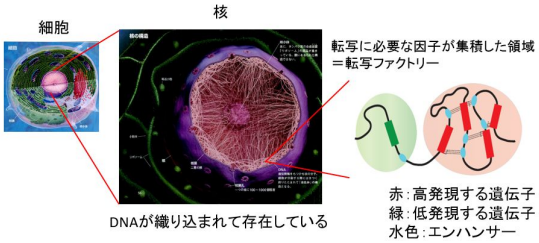

図３：転写ファクトリー

上記（1）、（2）、（3）を推進するにあたって、Encode や Human Epigenome Atlas といった国際プロジェクトで公開されたシークエンスデータを取得し、エピゲノムとトランスクリプトーム情報をデータベース化する。

３．研究の方法

（1）．２領域間のクロマチン相互作用を予測するベイズモデルの構築

リード配列が確率的に結合する要素を取り入れた、クロマチン相互作用過程の確率モデルで考える。ここでは、ChIA-PET から得られる相互作用情報以外に、ChIP-seq から得られるタンパク（RNA ポリメラーゼや転写因子）の結合やヒストン修飾などの情報、DNase-seq、FAIRE-seq から得られるクロマチン開閉領域の情報、結合領域のゲノム情報を組み込めるように、２つのクロマチン領域が結合する確率がこれらの事前情報に依存する形で定式化を行う。このモデルに含まれるパラメータをデータから推定するため、ベイズの枠組みにおける MAP 推定を行う。次に、シミュレーションデータおよび実データを用いて、構築されたモデルの有用性を検証する。

２）．クロマチン隣接行列からクロマチン高次構造を推定する統計解析法の開発

クロマチン隣接行列からクロマチン高次構造を推定する方法として、多次元尺度構成法を用いる方式について検討する。ここで行う詳細な検討から、この推定方式の推定精度・限界を探るとともに、推定方式の拡張すべきコンポーネントを明らかにする。また、距離行列から三次元座標への付置に関して、ChIA-PET 以外に、ChIP-seq、DNase-seq、FAIRE-seq などの情報を組み込めるように定式化する。シミュレーションデータおよび実データを用いて、推定された３次元構造の精度を検証する。

（3）．転写ファクトリーの同定

I の推定結果とトランスクリプトームデータを組み合わせることによって、転写ファクトリーを同定するとともに、構造の違いによってどの程度転写ファクトリーに存在する遺伝子の発現に影響を与えるかを調べる統計的仮説検定手法を開発した。

４．研究成果

（1）ChIA-PET シークエンスタグのリファレンスゲノムへの確率的アライメント

スタート位置 S_k におけるマッピングの信頼度を S_k 周辺でのマッピングしたカウント数で表現するとき、ChIA-PET シークエンスペアタグ t_i と t_j がリファレンスゲノム上の座標 S_k と S_l にそれぞれマッピングする確率を EM アルゴリズムで推定する解析手法を開発した。

（2）ポアソンモデルによるバイアスの補正

ChIA-PET データには、ライブラリ調製やシークエンス段階で生じうるバイアス（GC 含有率、マッピングのしやすい領域・しにくい領域による違い）、コピー数によるバイアスの他に、相互作用する 2 領域間の距離よるバイアスなどが含まれる。そこで、ポアソン回帰モデルに基づくこれらバイアスの補正法を開発した。

（3）負の二項分布の混合ベイズモデルによる相互作用領域の検出

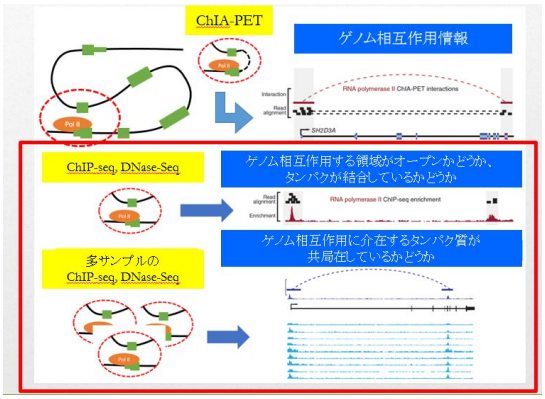

図４：相互作用に関する事前情報

補正済みの ChIA-PET カウント数に対し、相互作用する・しないを未観測の潜在変数としてもつ負の二項分布の混合モデルを考え、事前情報が与えられたもとで相互作用する確率としない確率の比をモデル化するとともに、EM アルゴリズムによるこれらのモデルのパラメータ推定法を開発した。これにより、相互作用するかしないかの事後確率として評価することができ、さまざまな情報を事前情報として取り入れることが可能となった（図４）。

（4）細胞特異的、シグナル伝達下で変化する転写ファクトリーの同定

混合ベイズモデルから得られた交互作用情報とトランスクリプトームデータを組み合わせて使用することによって、二種類の細胞株（A549、HUVEC）の ChIA－PET データから、細胞の特異性を決める転写ファクトリーを同定した（図５）。また、構造の違いがどの程度転写ファクトリーに存在する遺伝子発現に影響を及ぼすかを検定する仮説検定法を開発した。

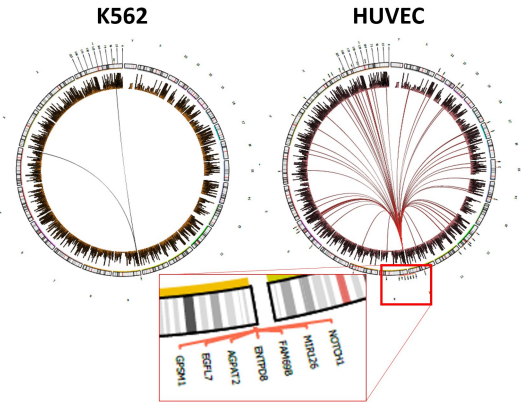

図５：細胞特異的転写ファクトリー

（５）ChIA-PET データ解析パイプラインの開発

ChIA-PET シークエンスデータのリファレンスゲノムへの確率的マッピング、データに含まれるバイアスの補正、他のオミックスデータの統合、混合負の二項分布モデルによる相互作用領域の検出、転写ファクトリーの同定を行うための解析パイプラインを統計言語 R の解析パッケージとして開発した。

５．主な発表論文等

〔雑誌論文〕（計　２件）（共に査読あり）

① Hasegawa T, Niida A, Mori T, Shimamura T, Yamaguchi R, Miyano S, Akutsu T, Imoto S, A likelihood-free filtering method via approximate Bayesian computation in evaluating biological simulation models, Computational Statistics & Data Analysis 94, 2015, 63-74.

② Hasegawa T, Mori T, Yamaguchi R, Shimamura T, Miyano S, Imoto S, Akutsu T, Genomic data assimilation using a higher moment filtering technique for restoration of gene regulatory networks, BMC Systems Biology, 9, 2015, 14.

〔学会発表〕（計　３件）

① Shimamura T, Bayesian integrated analysis of chromatin interaction maps, Pacific Symposium on Biocomputing 2015, 5th Jan 2015, The Fairmont Orchid, Hawaii, USA.

② Shimamura T, Accrate detection of chromatin interactions from ChIA-PET sequencing data with a hierarchical Bayes model, Clinical Genomics and Informatics Europe, 5th Dec 2013, Sheraton Hotel and Spa, Lisbon, Portugal.

③ Shimamura T, Hierarchical Bayes model-based analysis of chromatin interaction maps, Genome Informatics, 2nd Nov 2013, Cold Spring Harbor Laboratory, NY, USA.

〔その他〕

ホームページ
http://www.nagoya-sysbiol.info/

６．研究組織

(1)研究代表者

島村　徹平（SHIMAMURA, Teppei）

名古屋大学・大学院医学系研究科・特任准教授

研究者番号：００６２３９４３