論文の内容の要旨

論文題目

Efficient algorithms for processing genome-wide chromatin information

（ゲノム全域のクロマチン情報を処理する効率的アルゴリズム）

氏名　市川　和樹

<序論>

真核生物のＤＮＡ核内で形成されるクロマチン構造は遺伝子の発現において重要な役割を持つ。クロマチンの基本構造であるヌクレオソームの配置は DNA 配列への転写因子の結合に影響を与える。特に転写開始点周辺のヌクレオソームの配置が遺伝子の転写制御に関わることが明らかになっている。クラスタリングを用いてヌクレオソームポジショニングを分類した研究により、従来知られているのプロモーター下流の+1 ヌクレオソームが安定して位置しているパターン以外に様々な配置が存在することが判明しているが、ヌクレオソームポジショニングがどのように制御されているかの機構はまだ完全には解明されていない。

また、同様に転写制御に関連するクロマチン情報であるヒストンタンパク質への修飾と CpG サイトの DNA メチル化が協調的に働くことが発生分化の過程において重要な役割を持つ。メダカのゲノムでは発生関連遺伝子のプロモーターの多くが数 kb 以上の長さの DNA 低メチル化領域(HMD)で、かつ転写抑制に関連するヒストン H3 のリジン 27 のメチル化(H3K27me3)により修飾された領域である large K27HMD に含まれていることが知られている。分化の過程において K27HMD の長さが縮小することが遺伝子の発現に影響を与えることが示唆されている。

Chip-seq, MNase-seq, や Single-Molecule Real-Time sequencing などの観測技術の発展によりゲノム全域のクロマチン情報を得られるようになったことでクロマチン構造を構成する機構について多くの発見が得られてきたが、同時にそれらの大規模なクロマチン情報を効率よく処理する手法が更なる発見のためにも必要とされている。

そのため、本研究では nucleosome positioning signal などのデータの分類によく用いられる Pearson correlation distance を用いた *k*-means clustering の高速化手法”BoostKCP”と、

K27HMD のような特定のエピゲノム修飾をもつ領域を線形時間で検出する手法”CSMinfinder”の二つのアルゴリズムを新たに考案し実際の生物学的データを使用して性能の検証を行った。またより正確なクロマチン構造の情報を得るためにギャップの少ないゲノムの構築の試みとして Hi-C data を用いたメダカゲノムの De novo assembly を行った。

<結果>

_k_-means clustering の高速化手法 BoostKCP の考案

biological data の解析に広く用いられるクラスタリング手法 *k*-means clustering を高速化するアルゴリズム BoostKCP(Boosting *K*-means Clustering for Pearson correlation distance)を考案した。*k*-means clustering とは各点を最近接のクラスタに分類するステップと分類された点から新たなクラスタの中心を求めるステップを繰り返すことでデータを *k* 個のクラスタに分類する手法である。BoostKCP は一番大きな計算量を持つ各点の最近接のクラスタへの分類の計算過程において、Pearson correlation distance の性質を用い不要な計算を枝刈りし計算時間の短縮を行う。

大規模データでの BoostKCP の性能比較

Transcription starting site 周辺の human nucleosome positioning signal data 及びその他二つの大規模データを使用して BoostKCP と、通常の *k*-means clustering と、高速化手法の Elkan’s algorithm, Hamerley’s algorithm の計算時間を計測し比較を行った。

TSS 周辺の nucleosome positioning signal data では次元数 d=10, 20, 50, 101, 201, 501, 1,001, 2,001 の場合においてクラスタ数 k =10, 20, 30 の条件でクラスタリングを行い、通常の *k*-means clustering に対して BoostKCP を使用した場合の計算時間は 5.37-26.5 倍高速化された。また他の高速化手法と比較して Hamerley’s algorithm の 4.36-24 倍, Elkan’s algorithm の 1.3-8.31 倍の高速化がされた。特に計算時間の長い高次元高クラスタ数において BoostKCP の高速化の効果が高かった。

特定のエピゲノム修飾の組み合わせをもつ区間を検出するアルゴリズム -CSMinfinder-

ＤＮＡ配列中から特定のエピゲノム修飾の組み合わせを持つ領域を線形時間で検出するアルゴリズム CSMinfinder(Chromatin State with minimum length finder)を考案した。

CSMinfinder は DNA 配列上の各位置のエピゲノム修飾状態と、検出したい修飾の組み合わせパターンとの類似度を計算し、配列中から類似度の合計値が最大になる重複のない領域

の組み合わせの検出を行う。CSMinfinder の利点として検出する区間の最低長を設定することができ目的の区間をより連続した長い領域として検出することができる。

CSMinfinder を使用したメダカゲノムの large K27HMD の検出

CSMinfinder を使用してメダカのゲノムで large K27HMD の検出を行った　区間の最低長を 4kbp としたとき、CSMinfinder は 911 か所の K27HMD を検出し、そのうち 386 領域にプロモーター領域が含まれていた。他の ad-hoc な K27HMD 検出手法である Nakamura's method が検出した 246 か所の developmental gene のプロモーター領域のうち 242 カ所が CSMinfinder でも検出された

K27HMD の領域長の比較

ヒトゲノムにおいて CSMinfinder と、同じくエピゲノム修飾状態により DNA 配列を区間分けする手法である ChromHMM と Nakamura's method により K27HMD を検出し、区間の長さを比較した。ChromHMM では領域が細かく分断されて検出され、8Kbp 以上の長さの領域として検出された領域が 4 領域のみに対して CSMinfinder と Nakamura's method で検出された領域は 8Kbp 以上の領域は 654 領域、314 領域あり、9Kbp-12Kbp 以上の K27HMD の区間数はほぼ等しかった。

Hi-C data を使用したゲノムアセンブリ

メダカの 3 つの近交系 Hd-rR, HNI, HSOK について SMRT シーケンスのリードを使用し *De Novo* アセンブリによりドラフトゲノムを構築した。genetic marker, BAC-end pairs, fosmid-end pairs から、contig を染色体上にアンカーし、さらにセントロメア領域の反復配列を含む 11 の contig のアセンブリを Hi-C data を使用して行った。最新の ver2.2.4 ゲノムでは Hd-rR, HNI, HSOK の総長は~733.5, ~677, ~744Mbp、ゲノム中のギャップ数は 491, 717, 318 個であり、2007 年に発表された version1 の Hd-rR メダカゲノムのギャップ数 97,933 個と比べて大幅に数が減少した。

<結論>

観測技術の発達により、ゲノム全域にわたってクロマチン情報を得ることが可能となり、クロマチン構造のもたらす影響についての解析が進んできた。しかし、大規模なエピジェネティックデータから有用な情報を検出するためにはより高速で効果的な手法が求められる。

本研究ではそのために二つのアルゴリズムを考案した。Pearson correlation distance を用いた *k*-means クラスタリングの高速化手法”boostKCP”は、nucleosome positioning signal data を使用したクラスタリングにおいて枝刈りを使用しない場合に比べて 5.37-26.5 倍の高速化を行うことができることを示した。様々な次元数クラスタ数において boostKCP の高速化は有用でありデータのクラスタリング条件の検討を短時間で行うことができるため、大規模データから新たな知見を得る手がかりとして有用であると考えられる。また同じく CSMinfinder では最低長を設定して特定のエピゲノム修飾を持つ区間の検出を高速に行うことができる。メダカ、ヒトのゲノムで large K27HMD の検出を行い developmental gene のプロモーター領域を持つ区間を検出した。本研究で作成された手法はどちらも論文中で取り扱ったデータ以外へも適用可能であり、新たに作成された ver2.2.4 のメダカゲノムなどの他のデータのクロマチン情報の解析のためにも応用することが可能である。