Vol.2012-MPS-87 No.27
2012/3/2

# 近隣リードを考慮したショートリードクラスタリングによる塩基配列構造情報の有向非循環グラフ表現

上田大介[†1]　瀬尾茂人[†1]
竹中要一[†1]　松田秀雄[†1]

本研究では高速シーケンサが出力するショートリードを用いて DNA 塩基配列構造の有向非循環グラフ表現を提案する．解析対象となる DNA を断片的に読み取ったショートリードはサンプル特有の塩基配列構造情報を内在している．それらを抽出するためにショートリードクラスタリングが行われるが，従来のクラスタの表示法では構造情報の抽出が難しい．そこで我々は近隣リードを考慮したショートリードクラスタリングを行い，クラスタを有向非循環グラフで表現することで，構造情報の効果的な抽出を目指す．

# Direct Acyclic Graph for Sequence Structures from Short Read Clustering with Neighboring Reads

Daisuke Ueta,[†1] Shigeto Seno,[†1] Yoichi Takenaka[†1]
and Hideo Matsuda[†1]

We propose a method to describe DNA sequence structures as direct acyclic graphs from short reads generated by high-throughput sequencer. The sequenced DNA fragments are called the short reads and they inherent sequence structures. Although short read clustering has developed to extract the sequence structures, current description of the clusters is less applicable to it. Therefore we first operate clustering with neighboring reads, and convert the clusters into direct acyclic graph in order to achieve effective extraction of sequence structures.

†1 大阪大学大学院情報科学研究科バイオ情報工学専攻
Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University

## 1. 研究背景

生物の細胞中には A，T，G，C の塩基により構成された DNA と呼ばれる物質が含まれている．DNA の塩基の並びはその生物の性質を表す重要な要素であるが，類似部分の存在や多倍体における複数染色体の存在など複雑な構造をとることが知られている．また DNA 配列は非常に大きく，現在の技術では一度にすべてを読み取ることができない．そのため DNA 配列解析では，解析対象となる DNA のランダムな位置から数十〜数百塩基の配列を読み取る作業が大量に繰り返される[1)]．冗長に読み取られた断片的な DNA から，解析対象となる DNA の全体像を知る試みがなされている．

読み取った断片的な塩基配列をショートリード (または単にリード) と呼ぶ．リードの集合は解析対象となる DNA を読み取った結果なので様々な構造情報が含まれている．本研究における構造情報とは，対になる染色体に基づく DNA 配列の多型[2)] や進化の過程での類似配列の出現，塩基の読み取りエラー[3)] など塩基の並びに基づくデータ特有の構造のことである．配列解析のひとつの目的にこれら構造情報を抽出することがあり，既存手法としてショートリードクラスタリング[4)] が存在する．

ショートリードクラスタリングは一定ハミング距離以下のリード対をリード集合の中から発見する手法である．類似するリード対の発見は構造情報の抽出への応用が期待されているが，類似リード間の距離を用いた木構造によるクラスタリング結果の表現は塩基配列の共通部分と相違部分が自明でないため構造情報の抽出が難しい．さらに，単純なクラスタリングではサンプル DNA の同じ位置から読み取られた配列しかクラスタリングできないため冗長度を十分に活かすことができない．

本研究では読み取り位置の少し違う近隣リードも含めてクラスタリングし，結果を配列ベースで表現するモデルを提案する．また本モデルの構造情報抽出能力についても考察する．

## 2. 提案手法

### 2.1 目　的

本研究の目的はクラスタリング結果を構造情報抽出のために表現するモデルの作成である．構造情報を表現するモデルを作成するためには近隣リードを考慮したクラスタリングを行う必要があるためその方法を含めて考える．

あるリードと DNA 上の読み取り開始点が $w$ 塩基異なるリードを $w$-近隣リードと呼ぶことにする．近隣リードを考慮したクラスタリングが構造情報抽出に必要である第一の理由は

Vol.2012-MPS-87 No.27
2012/3/2

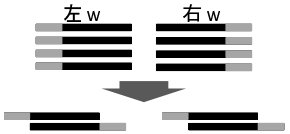


**図 1** 両端を削除したリード集合と推定された近隣リード
Fig. 1 Ends cut reads and inferred neighboring reads.

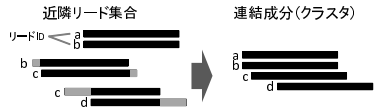


**図 2** 連結成分は近隣リード集合から作成
Fig. 2 Connected components from neighboring reads.

読み取りの冗長度を十分に活用することができるためである．DNA 配列の各塩基は複数回読まれ，複数のリードとして出力されているが，読み取り開始点は異なっている．そのため近隣リードを同一のクラスタにまとめてクラスタリングできなければ冗長度を活用することができない．第二の理由は近隣リードを同一のクラスタにまとめることで特徴的な構造の周辺の塩基配列を知ることができるためである．1 本のリードよりも広い範囲の塩基配列を知ることで，特徴的な構造が DNA 配列上のどの位置に存在するか特定しやすくなる．

### 2.2 クラスタリング手順

近隣リードを考慮したクラスタリングは読み取り開始点を揃えたリードを仮想的に作成し，類似リードのペアを検索することで実現する．

まず $w$ の上限を $W$ と定め，各リードの左端 $w$ 塩基を削除したリードと右端 $w$ 塩基を削除したリードを作成する．次に左端を削除したリード集合と右端を削除したリード集合の間で類似するリード (ハミング距離が 1 以下のリード) を SlideSort[4] によってすべて検索する (図 1)．検索されたリードのペアを推定上の $w$-近隣リードとすることで真の近隣リードを推定する．$w = 0, 1, \cdots, W$ について上記の手順を行うことで全近隣リードを検索することができる．互いに近隣であるリード対が求められれば，それらを連結成分ごとに分類することでクラスタを作成する (図 2)．

### 2.3 クラスタのグラフ表現

グラフは各クラスタに 1 個作成される．リードの部分配列をノードで表現し，2 つの部分配列間を接続するリードが存在する場合にエッジを引く．すなわちクラスタに属するリード集合 $R = r_1, r_2, \cdots, r_N$ を用いてグラフ $G = (V, E)$ を構成する．$v \in V$ は長さ任意の塩基配列を表す．ノード $v_s$ からノード $v_t$ に至るすべての経路を $path(v_s, v_t)$ で表現し，経路 $l$ が表す塩基配列を $seq(l)$ で表すと，構成するグラフが満たすべき条件は $\forall k, \exists s, t \ \ seq(l) = r_k$ ただし $l \in path(v_s, v_t)$ となる．すなわちグラフ上のエッジ集合は $E = \{e_{v_i,v_j} \mid \exists k, l \in path(v_s, v_t)$ について $seq(l) = r_k \cap (v_i, v_j) \in subpath(l)\}$ となる．ここで $subpath(l)$ は経路 $l$ のすべての部分経路を表す．

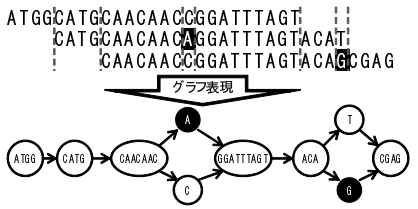


**図 3** 近隣リードから作成される有向グラフの例
Fig. 3 Directed graph from neighboring reads.

このようなグラフ表現は各リードを 1 ノードとしたグラフを初期条件とし，近隣リードを順次追加することで構成可能である．近隣リードを追加する際に，塩基配列が部分一致するノードは分割し，塩基配列が不一致もしくはノードが存在しない場合は新しく追加する．新しく追加されるリードは対になる近隣リードにより位置が特定されるため，構成されるグラフは有向非循環グラフとなる．配列を用いた例を図 3 に示す．

## 3. 考　　察

本手法はクラスタリング結果を塩基配列ベースで表現できるため特徴的な構造の抽出に応用しやすい形であるといえる．特徴的な構造は主に DNA の変異や，特定の配列の繰り返し，部分配列の組み合わせであることが多い．これらはグラフ表現中の分岐により解析できることが期待できる．さらに近隣リードを考慮したクラスタリングを行ったことで各ノードをサポートするリード数がわかる．それをノードの重みで表現すれば定量的な解析も可能である．これらの利点により本手法は塩基配列構造解析の進展に貢献するものと考えている．

## 参　考　文　献

1) Metzker, L. M. : Sequencing technologies - the next generation, *Nature Reviews Genetics*, Vol.11, NO.1, pp.31–46 (2009)
2) Yue, P. and Moult, J. : Identification and Analysis of Deleterious Human SNPs *Journal of Molecular Biology*, Vol.356, NO.5, pp.1263–1274 (2006)
3) Erlich, Y. et al.: Alta-Cyclic: a self-optimizing base caller for next-generation sequencing *Nature Methods*, Vol.5, NO.8, pp.679–682 (2008)
4) Shimizu, K. and Tsuda, K.: SlideSort: All Pairs Similarity Search for Short Reads *Bioinformatics*, Vol.27, NO.4, pp.464–470 (2010)