5J-8

次世代シーケンサーから得られたDNA配列データの高速クラスタリングに関する研究

並木　洋平†　　石田　貴士†　　秋山　泰†

†東京工業大学 大学院情報理工学研究科 計算工学専攻

1　序論

近年，次世代シーケンサーと呼ばれるハイスループットなシーケンサーが登場し，短時間で大量の DNA 配列データを得ることが可能になった．次世代シーケンサーから得られる DNA 配列は，配列長は比較的短く，かつ本数が非常に多いという特徴がある．このように DNA 配列の解読量が年々増加している一方，バイオインフォマティクスの分野ではこれらの大量データを解析するための技術が追い付いていないという問題が発生している．

本研究は，次世代シーケンサーから得られた DNA 配列のクラスタリングの高速化を目的とする．配列クラスタリングとは，配列データセットから類似配列のグループを見つけ，同じクラスタに割り当てていくデータ解析手法である．本研究では最長共通部分列 (Longest Common Subsequence, LCS) の性質を用いることで，クラスタリング処理の大幅な高速化を実現した [1].

2　関連研究

広く利用されている既存の配列クラスタリングツールのひとつに CD-HIT [2] がある．CD-HIT は近似的なクラスタリング手法を用いることによって比較的大量の配列を高速に処理することができ，メタゲノムデータのアノテーションパイプライン中で用いられたり，Uniprot や PDB といった公共データベースにおいて冗長配列を取り除いてデータベースサイズを圧縮するのにも使われている．しかし，1000 万本の配列のクラスタリング処理には数日かかるという具合に，次世代シーケンサーの出力配列ほどの規模のデータをクラスタリングする場合，現実的な時間で処理するのは困難になってくるという問題がある．

A fast clustering method for DNA sequence data from next generation sequencers

†Youhei NAMIKI　†Takashi ISHIDA　†Yutaka AKIYAMA

†Graduate School of Information Science and Engineering, Tokyo Institute of Technology

3　本研究のクラスタリング手法

3.1　配列類似度とクラスタリング基準

DNA 配列をクラスタリングするためには，まず配列間の類似度を定義する必要がある．提案手法では，配列類似度の指標には **sequence identity** を用いている．sequence identity は，2 配列に対して配列アラインメントを行ったときの 2 配列間のマッチ文字数を短い方の配列長で割った値として定義される．

提案手法では，配列間の sequence identity が閾値 s を超える配列同士を同じクラスタにまとめていく．これは CD-HIT と同じクラスタリング基準である．

アラインメント処理は計算量が非常に大きいため，全配列ペアについて sequence identity を計算するのは非現実的である．そのため，不要な sequence identity 計算をできるだけ減らすことが配列クラスタリングを高速化する上で重要である．

3.2　Short word filtering

short word filtering とは，次にクラスタリングするひとつの入力配列に対して，それと共通の k-mer (k 塩基の部分配列) を t 個以上持つ既存クラスタの代表配列を列挙する手法である (図 1)．この処理は index table のデータ構造を用いることで高速に行うことができる．short word filtering を使うことで非類似配列同士の比較処理を大量に枝刈りすることができる．

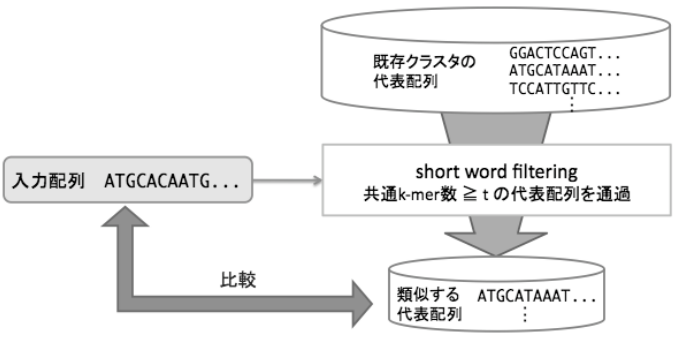

図 1: Short word filtering

short word filtering は CD-HIT でも用いられているが，short word filtering による類似配列ペアのフィルタリングは若干粗く，部分的に偶然一致する非類似配列ペアを通過させてしまうケースも少なくない．そこで，本研究では short word filtering の後により高精度な独自のフィルタリング処理「LCS filtering」を導入している．

3.3 LCS filtering

LCS filtering とは，short word filtering を通過したクラスタ代表配列と入力配列のペアに対して行うフィルタリング処理であり，2 配列間の最長共通部分列 (the Longest Common Subsequence, LCS) の長さを求めることで，2 配列が類似しているかどうかを判別する．2 配列間の最長共通部分列の長さ LLCS と配列長 n，および 2 配列間の sequence identity の間には以下の式が成立する．

$$\frac{\mathrm{LLCS}}{n} \geq \text{sequence identity} \quad (1)$$

つまり，2 配列間の最長共通部分列の長さ LLCS を配列長 n で割った値は，2 配列間の sequence identity が取り得る値の上界となる．この性質から，LLCS/n の値を計算し，sequence identity の閾値 s 以上ならば同じクラスタに割り当てられる可能性のある類似配列ペアとしてアラインメント処理に進み，閾値 s 未満ならば捨てて処理を打ち切ることで，非類似配列ペアに対するアラインメント処理を省くことができる (図 2)．

LCS filtering は short word filtering よりも高い精度で類似配列ペアをフィルタリングでき，また LLCS はビット演算で高速計算が可能である [3]．そのため，LCS filtering を導入する方が低速なアラインメント処理を何度も行うよりも結果的に高速になる．

4 評価実験

本研究で提案する手法を用いたクラスタリングプログラムを C++で実装し，CD-HIT と提案手法の間で速度の比較実験を行った．評価実験の入力データセットは MetaSim [4] で生成したメタゲノムの配列データを用いた．入力データセットの DNA 配列の配列長は 100b，150b の 2 通り (固定長)，配列数は 100 万本，500 万本の 2 通りで，計 4 通りの入力データセットを用いた．また，クラスタリングの際の閾値やパラメータについては，sequence identity の閾値 s は 0.9 (90%)，short word filtering の k の値は 9 とし，共通 k-mer の個数の閾値 t は 1 とした．計算機環境は SUSE Linux 10 のワークステーションで，AMD Opteron 2.8GHz のシングルコアを用いた．

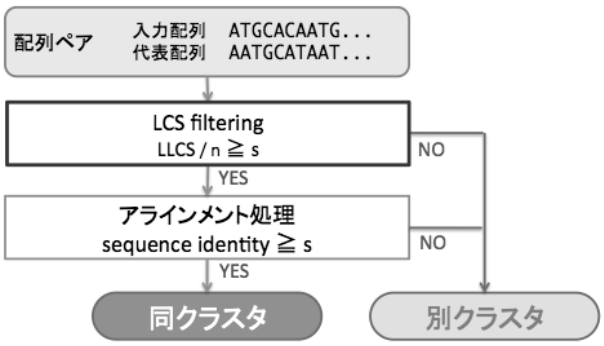

図 2: Short word filtering 後の処理の流れ

表 1: CD-HIT と提案手法の計算時間

	配列数	100b		150b	
CD-HIT	100 万本	41m40s		45m15s	
提案手法	100 万本	7m10s	(5.81)	13m45s	(3.29)
CD-HIT	500 万本	11h17m22s		11h28m17s	
提案手法	500 万本	2h11m09s	(5.16)	3h04m43s	(3.73)

CD-HIT と提案手法のクラスタリングの所要時間を表 1 に示す．また，表 1 の提案手法の計算時間の横には CD-HIT に対する提案手法の速度向上率を括弧書きで示している．表から分かる通り，全ての場合において CD-HIT よりも提案手法の方が速いという結果が得られた．

5 まとめ

本研究では，最長共通部分列の性質を用いて次世代シーケンサーから得られた DNA 配列のクラスタリングを高速化した．評価実験では，固定長の DNA 配列 500 万本について，配列長 100 塩基の場合は CD-HIT に対して約 5.2 倍，150 塩基の場合は約 3.7 倍の速度向上を実現した．本手法を用いることで，今までのクラスタリングツールでは難しかった大規模な DNA 配列データのクラスタリングが現実的な時間内で可能になると考えられる．

参考文献

[1] 最長共通部分列に基づく DNA 配列の高速クラスタリング, 並木 洋平, 石田 貴士, 秋山 泰, 情報処理学会研究報告バイオ情報学 (BIO), 2010-BIO-23(24): 1-7.

[2] Li W. and Godzik A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, 22(13): 1658-1659 (2006).

[3] Hyyrö H., Bit-parallel LCS-length computation revisited, *Proc. 15th Australasian Workshop on Combinatorial Algorithms (AWOCA)*, pp.16-27 (2004).

[4] Richter D.C., Ott F., Auch A.F., Schmid R. and Huson D.H., MetaSim: a sequencing simulator for genomics and metagenomics, *PLoS ONE*, 3(10): e3373 (2008).