Vol.2011-MPS-86 No.19
Vol.2011-BIO-27 No.19
2011/12/1

Short read シーケンサーデータに対する複次重複処理による結合信頼性向上の検討

大城絢子 †1　岡崎威生 †2
安富祖仁 †1　名嘉村盛和 †2

次世代シークエンサの開発により、読み取り配列の長さを短くし並列処理を行うことで、配列の高速処理が可能になった。またアセンブルにおいて k-mer を利用することで、シーケンサのリードミスにも対処ができている。しかし k の値によってアセンブルの精度が左右されている。そのため未知の配列に対して de novo アセンブルを行う際には配列結合の信頼性が確保できない。そこで本報告では、複数の k-mer を利用することで配列結合の信頼性を向上させることについて検討する。

Accuracy improvement for short read sequencer data by double overlap processing

AYAKO OHSHIRO ,†1 TAKEO OKAZAKI ,†2
HITOSHI AFUSO †1 and MORIKAZU NAKAMURA†2

With development of second generation sequencer technology, it has become possible to speedily get massive short-read sequences. The readmiss sequences are handled by using k-mer in assembly. But the results of assembly depend on k value, therefore, we can't obtain the best result in de novo assembly. In this regard, this research considers accuracy improvement for assembly by a way of multiple k-mer processing.

†1 Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus
†2 Faculty of Engineering, University of the Ryukyus

1. シークエンスアセンブルにおける k-mer の役割と問題点

シークエンスアセンブルでは、リードミスに対応するために k-mer が利用される。k-mer とは定数長の配列のことであり、読み取り配列長より短く設定される。図 1 のように読み取り配列に対して k-mer を 1base ずつずらしたものをハッシュとしテーブルを作成する。ハッシュテーブルには全ての読み取り配列中での、そのハッシュの出現頻度値も格納する。出現頻度の値が著しく小さいハッシュについてはリードミスが含まれると見なされ除外される。ハッシュ間の重複情報を用いて隣接グラフが作成できる。

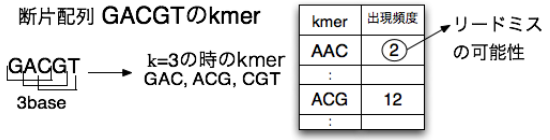

図 1　断片配列からハッシュテーブルを作成 (k=3)

Velvet[?] では k-mer 同士のリンクについて、$(k-1)$base の重複長に対してのみリンクと見なしている。Velvet を用いて k の値がアセンブルの精度に与える影響を比較すると表 1 の結果が得られた。データには大腸菌の配列の一部 10 万 base から擬似的に 2 種類の読み取り配列を作成したものを用いた。表の"Node" は出力された Contig 数を、"max" は出力中で最も長い contig を表す。

Data	k	11	13	15	17	19	21	23	25	27	29	31
read:35base	Node	8996	384	18	2	1	1	1	2	6	54	436
142858 reads	max	132	3426	20835	99979	99977	99975	99973	79560	37295	6079	1061
read:100base	Node	1141	256	205	200	204	211	215	211	210	201	195
2000 reads	max	309	992	1138	1136	1134	907	816	814	812	677	675

表 1　それぞれのデータについて k の値の変化がアセンブルに与える影響

この結果から、k の値が出力される contig 数や長さに影響していること、つまり結果が特定の k の値に依存していることがわかる。de novo アセンブルを行う際には適切な k がわからないため正しく長い contig を生成することが難しいと言える。

Vol.2011-MPS-86 No.19
Vol.2011-BIO-27 No.19
2011/12/1

ノードの内容	ノード一覧
読み取り配列	GACGT AGCCC TAGCC TAGGA CCCAA CCCAA TTAGC AGCCC CCAAG GTTAG
3-mer	GAC ACG CGT AGC GCC CCC TAG AGG GGA CCA CAA TTA AAG GTT
4-mer	GACG ACGT AGCC GCCC TAGC TAGG AGGA CCCA CCAA TTAG CAAG GTTA

表 2 断片配列、k-mer についてのノード一覧

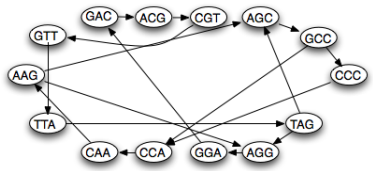

図 2 隣接グラフ (3-mer)

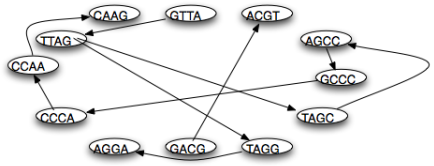

図 3 隣接グラフ (4-mer)

2. 複次重複処理による隣接グラフの生成

そこで、多様なデータに対して頑健な配列結合を行うために、複数長の k-mer を用いた結合方法を検討する。複数の k-mer について隣接グラフをそれぞれ作成し、k-mer 同士のリンクが、その k-mer が含まれる読み取り配列同士のリンクでもあるのかを、読み取り配列をノードとした隣接グラフに帰着させることで結合の信頼性をあげると考えられる。

例えば表 2 の読み取り配列に対して $k = 3, k = 4$ の場合を考えると隣接グラフは図 2、図 3 のようになる。

これらを元の読み取り配列に対応させ隣接グラフを生成すると図 4 のようになる。 エッジにはどの k-mer の隣接グラフから帰着してきたかの情報を格納している。

以上をまとめると、複数の k-mer を用いた隣接グラフは次のようにして生成できる。

＜入力＞読み取り配列集合

＜出力＞読み取り配列をノードした隣接グラフ

1 {$k_1, \ldots, k_n$}について n 個のハッシュテーブルを作成する。
出現頻度値が小さいハッシュについてはリードミスを含むとみなし除外する。

2 各 k_i-mer について隣接グラフを作成する。

3 読み取り配列をノードとする隣接グラフは、各読み取り配列に対応する k-mer の隣接情報をもとに生成する。

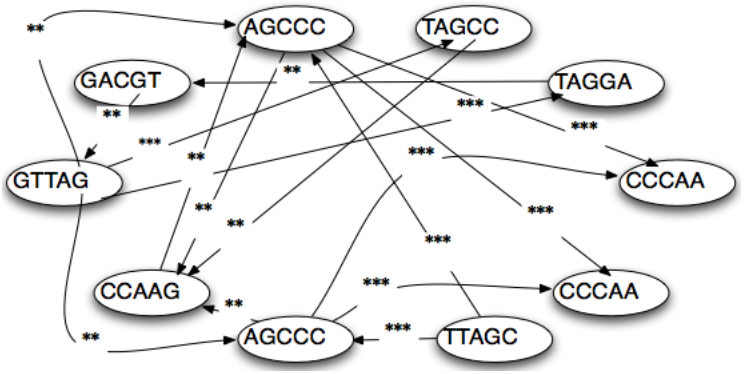

図 4 読み取り配列をノードとした隣接グラフ

手順 3 において、各 k-mer の隣接グラフが有する結合情報の信頼性を指標化し統合したうえで、読み取り配列の結合を判断する必要がある。また読み取り配列にはリードミスが含まれるため、結合する際に不整合部分が重複部に発生する可能性があり、不整合塩基の補正について検討しなければならない。

参 考 文 献

1) ”Daniel R. Zerbino and Ewan Birney ” Velvet:Algorithms for de novo short read assembly using de Bruijn graphs, Genome Research, vol.18, pp.821- 829, (2008)
2) ”Jared T. Simpson, kim Wong, Shaun D. Jackman, et al”ABySS: A parallel assembler for short read sequence data, Genome Research, vol.19, pp.1117- 1123, (2009)
3) ”Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman” An Eulerian path approach to DNA fragment assembly, PNAS, vol.98, pp9748-9753, (2001)
4) ”Rene L. Warren , Granger G. Sutton , Steve J. M. Jones and Robert A. Holt” Assembling millions of short DNA sequences using SSAkE, Bioinfomatics, vol.23,pp500-501, (2007)
5) ”Mihai Pop” Genome assembly reborn: recent computational challenges BRIEFINGS IN BIOINFORMTICS, vol.10, NO 4, pp354-366