



INTRODUCTION

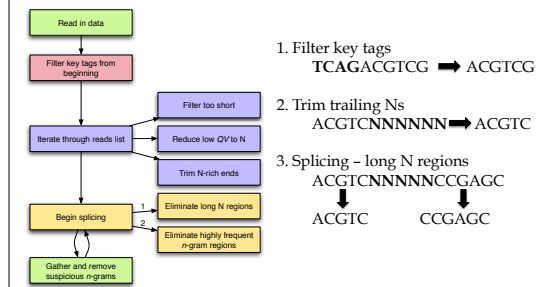
For next-generation sequencers, reconstructing an organism's genome from millions of reads is a computationally expensive task. Our algorithm solves this problem by organizing and indexing the reads using n -grams, which are short, fixed-length DNA sequences of length n . These n -grams are used to efficiently locate putative read joins, thereby eliminating the need to perform an exhaustive search over all possible read pairs.

We have developed a novel n-gram method for the assembly of genomes from next-generation sequencers.

Specifically, a probabilistic, iterative approach was utilized to determine the most likely reads to join through development of a new metric that models the probability of any two arbitrary reads being joined together.

PRE-PROCESSING

Pre-processing is used to filter and remove low quality reads from the data set and prevent them from being added to the dictionary.



DATA STRUCTURES

Genome
ACTGCTAT

Reads

1. ACTGC
2. GCTAT
3. CTGCTA

Dictionary (4-gram)

ACTG: (1,0,1)
CTGC: (1,1,0), (3,0,2)
GCTA: (2,0,1), (3,2,0)
CTAT: (3,0,2)
TGCT: (3,1,1)

ASSEMBLY

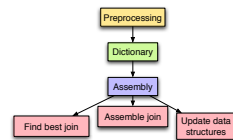


Figure 1 The major steps of the assembly program

Find Best Join

The n -gram index serves to substantially reduce the reads to consider for joining to those that share a common n -gram. This function uses a simple probabilistic model to assess the likelihood that these two reads came from the same origin in G. The resulting score is used to judge this join against all other possible joins being considered for this n -gram.

Maximize Overlap and InDel Determination

CGGTTAGAACTACAGATTGAAAAAGAAA%TGAACCTAA|TGAAGTCG
 GAACAAT|GATTGAAAAAGNN-%TGAACCTAA|TGAAGTCG
 3333344443**2**333333333333**333**3333333333 33333333
 3333344443**3**333333333333**333**3333333333 37777777
 3333333333333333 %3333333333 33333333
 0167777777777666674 %6654422777 77777777

Assemble Join

The assembly function combines the two reads with the highest probability score calculated in the previous section and forms a new contig. If the two nucleotides are:

- Same $\Rightarrow$ add nucleotide to contig.
 - Different $\Rightarrow$ add nucleotide with the higher quality value to contig.
 - Either nucleotide = 'N' $\Rightarrow$ add other nucleotide.
 - Both nucleotides = 'N' $\Rightarrow$ add 'N' to contig.
- $$\begin{array}{l}
 X_i = \overline{\text{CTGCTACN}} \\
 X_j = \overline{\text{GCTATA}}
 \end{array}
 \begin{array}{c}
 \text{A} \quad \text{B} \quad \text{C} \\
 | \quad | \quad | \\
 \text{GCTACN} \\
 \text{GCTATA}
 \end{array}
 \Rightarrow \text{Contig: CTGCTACAA}$$

Update Data Structures

Reads

1. CTGCTACN
2. GCTATAA
3. CCGCTA
4. CTGCTATAA

Dictionary (Before)

CTGC: (1, 0, 4)
TGCT: (1, 1, 3)
GCTA: (1, 2, 2), (2, 0, 3)
CTAC: (1, 3, 1)
TACN: (1, 4, 0)
CTAT: (2, 1, 2)
TATA: (2, 2, 1)
ATAA: (2, 3, 0)
CCGC: (3, 0, 2)
CGCT: (3, 1, 1)
GCTA: (3, 2, 0)

Dictionary (After)

CTGC: (4, 0, 5)
TGCT: (4, 1, 4)
GCTA: (4, 2, 3)
CTAC:
TACN:
CTAT: (4, 3, 2)
TATA: (4, 4, 1)
ATAA: (4, 5, 0)
CCGC: (3, 0, 2)
CGCT: (3, 1, 1)
GCTA: (3, 2, 0)

RESULTS

	Assembly Software	N50 (kbp)	Average length (kbp)	Max length (kbp)	Number of contigs
<i>Drosophila melanogaster</i>	ngAssem	> 32	-	52.2	-
	SHARCS	> 30	-	> 50	-
<i>Helicobacter pylori</i>	ngAssem	76.6	16.6	216.2	126
	PACB454	28.6	10.9	80.7	144
	Newbler 2.5	140.7	62.5	327.9	25
<i>Staphylococcus aureus</i>	ngAssem	2.1	1.8	23.6	2008
	Velvet	5.4	2.5	22.9	1095
	SSAKE	2	1.2	12.6	2432
	SHARCS	1.2	760 bp	8.6	3635
<i>Brucella suis</i>	ngAssem	8.4	9.3	31.3	316
	Euler	19	22	82	280
	Minimus	32	30	89	203
	PHRAP	80	53	198	88

Genome	Accession number	Length (bp)	Num. of reads (simulated)	Avg. read length (bp)	Std. (bp)	Coverage
<i>Drosophila melanogaster</i>	ACC006575	79,745	500,000 (simulated)	30	0	188
<i>Helicobacter pylori</i>	SRP001115	1,588,278	72,412	267	-	39.14
<i>Staphylococcus aureus</i>	NC003923	2,820,462	3,860,000	35	0	48
<i>Brucella suis</i>	AE014291	3,315,173	36,276	895.8	44.1	9.8

Tables 1 & 2 Assembly results and genome details for four organisms

CONCLUSION

This work has shown that a probabilistic n -gram based method can be used to successfully assemble small length genomes. *Drosophila melanogaster* resulted in a largest contig size of 52 kbp and an N50 of ≥ 32 kbp, *Helicobacter pylori* resulted in a 216 kbp contig and an N50 of 76 kbp, *Staphylococcus aureus* resulted in a 23.6 kbp largest contig and an N50 of 2.1 kbp, and *Brucella suis* resulted in a 31.3 largest contig and an N50 of 8.4 kbp. All four assemblies resulted in greater than 99% accuracy and identity.

Future work includes the addition of paired-ends, parallelization, and local alignments in the scoring mechanism to aid in determining the optimal join.

Acknowledgements

Part of this work was performed through the Bucknell University Program for Undergraduate Research.

References

1. Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95, 315-327.
2. Narzisi, G., & Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PLoS one*, 6(4), e19175.
3. Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research*, 1165-1173.
4. Segar, Matthew W. (2012). Utilization of Probabilistic Models in Short Read Assembly from Second-Generation Sequencing. *Honor's Theses*. Paper 97.