Abstract - is a short description of the field of investigation, methods used, and its conclusions. ///// what you did and what you found out.

The advent of next generation sequencing (NGS) assemblers have allowed many researchers to explore many new areas of research. These NGS assemblers pave way for major breakthrough in genomic research. Despite the many advantages of next-generation sequencers, e.g., the high-throughput sequencing rate and relatively low cost of sequencing, the assembly of the reads produced by these sequencers still remains a major challenge. In this paper, the author presents Strand Aligning Sequence Assembler (SALSA), a sequence assembler where reads are stored in a TRIE data structure and constructed using a hybrid approach of an overlap-based and suffix/tree construction.

We’re able to perform better

Background - would say why this field was chosen, what is already known about it, and why further investigation would be useful. // why you did it. Motivation:: // Sequence Assembly // Reference Geneome is unavailable // Importance

Currently, Sanger sequencing has been dominant for building reference genomes and is used in most research [6], but it is expensive and only a few life forms have been fully sequenced.

Short read sequence assembly is important because current sequence technologies are unable to sequences a whole genome [3]. Short read sequencing is significantly cost effective against its counterpart due to the readily available high throughput NGS.

2. Allows more research and analysis for plants and animals that have not yet been fully sequenced.

Several approaches to sequence assembly shotgun approach.

1. K-Spectrum-Based Construction

Often uses the de Bruijn graph structure [4] and is often used to read short read overlaps (ideal for high coverage, very short read (25–50 bp) data sets.). Uses graphs and edges to perform its construction. K-mers are illustrated as vertex and the read reads are the edges and the graph becomes proportional to the size of the sequences read. There are several read assemblers using this approach VELVET[5].

2. Overlap-Based Construction

The goal of Overlap-Based layout is to find the shortest Hamiltonian distance.[3]

3. Greedy-Based Construction

The Greedy-Based algorithm seeks after the greatest immediate sequences to the construction of the

There’s no reference genome.

The De novo sequence assembly is the process where individual sequences from the original DNA sequence are combined to form a larger sequence based on the information available of similar nucleotides to reconstruct the original sequence each individual sequences were derived from [1].

There are many challenges for short sequence assembly applications.

Short reads.

Errors due to less unique end-pair reads

Repeated sequences larger than the read lengths

Problem

The assembler accepts a data file that houses DNA sequences that have been split into multiple pieces of smaller sequences. Once the reads have been read, it will then attempt to reconstruct the smaller sequences and assemble the genome it once represented.

Development Environment.

For the test, an Intel Core i7-3537U CPU @ 2.0 GHZ with 4GB RAM laptop computer was used.

Application Parameters

SALSA operates with three specific parameters that determine its results: A sample data file, the read length, and number of reads generated. The sample data file can be generated using an optional parameter or be set by the user manually in the resource file. The read length is a pivotal value that determines the length of each read sequence split from the aforementioned sample data file. Finally, the number of reads generated when increased will increase the chance of an overlap (coverage).

Input:

Reads taken from genome

Output:

Assemble Genome

Baseline Method:

In order to compare the performance of SALSA, a Greedy-Based baseline algorithm was used.

The baseline algorithm begins its preparation stage similar to

Surprisingly the Greedy method

Application Parameters

Sample Data File

Frag Length

Reads

Baseline Method

Preparation

Get sample data

Chop it up

Put it all into a set

Method:

Choose a

My Method

SALSA:

In the Graph-based SALSA approach

TRIE datastructure

The TRIE data structure is a special tree that

The worst case for O(M) where M is the length of the search key string.

Preparation:

Based on the sample data parameter, a random sequence is generated or a user defined sequence is used. Once the application’s parameter has been defined the splicer creates the set number of reads that are assigned based on the read length defined and output into an external resource file separated by a new line.

In the second phase, each spliced sequences are read into the SALSA application and stored into a list where each unique sequence is stored as a graph node. Once a complete node set is created, the application will create an Eulerian path between all the vertexes and calculate a weight value based on the similarity score of the two nodes in question. Explain the scoring. During this loop, the max edge weight is kept for safe keeping

Once the graph and edges have been constructed with the corresponding weights, the reconstruction process begins. Here, the sequences are merged with the best fitting edge scores to form larger contigs. This process continues until there’s only one sequence, or the sequences are unable to match pair ends.

Read in sample data

Construct Graph | Edge

Evaluate Weight

Reconstruction of sample data

Measure of Performance

How similar

Error in assembled genome

Speed of computation

Memory usage

Compare above vs baseline

Performances

Future works

Statistics for repeats

Conclusion

Real Data construction much more difficult than simulated data? Why? Real Data has a lot of repeated sequences making the construction really difficult. Not a surprise because of Article in[1][2]

References

[1] Paszkiewicz K., Studholme D. De novo assembly of short sequence reads. Briefing in Bioinformatics vII;5: 2010

[2] Kingsford C., Schatz MC., Pop M. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 2010; 11:21

[3] El-Metwally S., Hamza T., Zakaria M, Helmy M., Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLOS Computational Biology. Vol 9;12 Dec 2013

[4] Li R., Zhu H., Ruan J., Qian W. et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Research December, 2009.

[5] Zerbino D., Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Cold Spring Harbor Laboratory Press March, 2008.

[6] Sanger. Wellcome Trust Sanger Institute http://www.sanger.ac.uk/research/areas/bioinformatics/grc/ Cambridge, UK. June 2014