4th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2013

A next-generation sequence clustering method for *E. coli* through proteomics-genomics data mapping

Mikang Sim\(^a\), Ho-Sik Seok\(^a\), Jaebum Kim\(^{a,b,*}\)

\(^a\)Department of Animal Biotechnology, Konkuk University, Seoul 143-701, Korea
\(^b\)UBITA Center for Biotechnology Research (CBRU), Konkuk University, Seoul 143-701, Korea

**Abstract**

Recent publications of various ‘omics’ data have provided new challenges and opportunities to the development of novel approaches to the assembly of next-generation sequences. As an attempt to improve the quality of assembled sequences, we developed a next-generation sequence clustering method by using the interdependency between genomics and proteomics data, which has not been well utilized so far in this field. Given a set of next-generation read sequences with a number of protein sequences, our method clusters the read sequences by mapping to the protein sequences. As a preliminary research, we selected *Escherichia coli* (*E. coli*) as our target species and simulated next-generation reads of *E. coli* to evaluate our method by analyzing the actual adjacency of the clustered reads in the *E. coli* genome. We found that (i) read base matching (RBM) ratio, which represents the amount of bases in a read that are mapped to a protein sequence, higher than 50–70% is a useful criterion for effective read clustering and (ii) higher RBM ratio does not always lead to better quality of clusters in the case of *E. coli*. These preliminary results demonstrate that the integrative approach is simple yet has great potential for clustering adjacent reads in a genome.

© 2013 The Authors. Published by Elsevier B.V. Open access under [CC BY-NC-ND license](http://creativecommons.org/licenses/by-nc-nd/).

Selection and peer-review under responsibility of the Program Committee of CSBio2013

**Keywords:** next-generation sequence assembly; *E. coli*; multiomics mapping; clustering

* Corresponding author. Tel.: +82-2-450-0456.
**E-mail address:** jbkim@konkuk.ac.kr.
1. Introduction

Current limitations of next-generation sequencing (NGS) technologies lead to the generation of the huge number of very short read sequences. To recover original genome sequences, the NGS read sequences should be assembled. Sequence assembly generally relies on the simple assumption that highly similar DNA fragments originate from the same position within a genome, which does not hold in some cases. For example, in metagenomic analyses which are defined as the direct genetic analysis of all genomes contained in an environmental sample, nearly identical sequences may originate from genomes of different species in the sample or genomic repeats yield similar fragments originated from different places in the genome of the same species. Because most genome sequence assembly technologies concatenate short NGS read sequences based on their overlaps, the resulting assemblies may be fragmented and error-prone, especially when the read sequences are generated from complex genomic samples, such as a metagenome. In addition, the need for checking the overlaps between all possible pairs of read sequences makes these ‘read-centric’ methods computationally unfeasible.

In this paper, we present a novel NGS read clustering method to construct more accurate groups of neighboring reads by taking advantage of multiomics data, especially amino acid sequences and protein-protein interactions (PPIs). The recent accumulation of multiomics data has provided us new opportunities to address the above problems of sequence assembly. For example, if a protein interaction network is given and some properties of interest of particular proteins are measured, then it is possible to propagate the similar properties of unmeasured proteins through the inter-connections of those proteins. Here, the protein interaction network and the properties of interest are analogous to the PPI networks and correspondence between a protein and read sequences respectively. Therefore the co-occurrence of read sequences in the same cluster can be predicted based on the mapping of read sequences to proteins that are collected from inter-connections in the PPI network. This idea is particularly useful for dealing with non-overlapping read sequences originated from the same genomic locus because their adjacency cannot be observed from genomic similarity.

Here, we report the following preliminary results: (i) the use of proteomics data does have an impact on enhancing the quality of read clusters, (ii) in determining optimal clusters of reads, the minimum base matching ratio between read sequences and proteins of 50~70% is the most effective, and (iii) higher matching ratio does not always mean higher quality clusters in the case of E. coli. The method described here has great potential for creating more reliable sequence clusters to help the production of more reliable sequence assemblies.

2. Materials and methods

2.1. Data

Here we aim to build clusters of potentially adjacent reads before attempting whole genome sequence assembly. Therefore our experimental setting is different from typical approaches. Specifically, in addition to a set of read sequences of the target species, we assumed that we can collect a number of proteins related to the target species X. Before applying our method to real situations, we evaluated our method by using E. coli as our target species, which already has enough data that can be used as benchmarks for the evaluation. We collected a genome sequence of E. coli K12 (E. coli K12 substrain MG1655 accession number: U00096) from the National Center for Biotechnology Information (NCBI) database and simulated 13,918,800 reads using the program “ARFl8”. Among the 13,918,800 reads, randomly selected 138,440 reads were used for the evaluation. Proteins related to E. coli were obtained from the Database of Interacting Proteins. Among the 4,000 proteins collected from the Database of Interacting Proteins, we randomly sampled 100 proteins in order to mimic the situation where we can only collected 100 proteins of the target species. In order to enlarge the available protein data, we added additional proteins by using the PPI database. We searched for interacting proteins for the randomly sampled 100 proteins and obtained 141 additional proteins which have known interactions with the above 100 proteins.
2.2. Read clustering and algorithm

Sequence assembly algorithms first create contigs by concatenating read sequences and then construct longer fragments, called scaffolds, by further merging the contigs. Our proposed method attempts to build accurate clusters of read sequences, each of which can be used to create more accurate contigs. To achieve this goal, our method builds the clusters of read sequences from adjacent reads confirmed not by genomic overlapping but by proximity in the mapped protein sequences.

The core idea of the proposed method is the use of proteomics data, such as protein sequences and PPIs. Given a number of proteins and their sequences, we searched for interacting proteins by examining the PPI database and appended the sequences of interacting proteins into the input set of protein sequences. Because of this expansion of protein sequences, more number of reads can be identified as candidates in the same clusters. The proposed method is explained in Fig. 1 and Table 1. From the genome sequence of *E. coli* K12, we generated a set of single-end reads.
using the program ART (coverage: 300x). Because of its high coverage, we randomly selected only 1% of reads. In addition to read selection, we randomly selected 100 proteins for clustering. These randomly sampled sets of reads and proteins constitute an initial set for clustering.

Because the number of selected proteins is likely to be lower than the actual number of proteins comprising the target species, we increased the number of proteins by selecting additional proteins from the PPI database, which interact with proteins obtained at the step III in Table 1. By considering PPI, we added 141 more proteins to the protein set. After data preparation, we aligned the reads to the amino acid sequences of the proteins using blastx.

3. Experimental results

We evaluated our method by using simulated NGS reads and proteins of \textit{E. coli} (see Materials and methods). In this evaluation, we focused on the adjacency of reads and the effective mapping ratio of an individual read. Despite its simplicity, the proposed method produced a set of valid read clusters. Through the blastx mapping (see Materials and methods), 26 proteins were discarded due to poor alignment result. For remaining 215 reads-protein groups, we analyzed the effect of read alignability with the following definition of the mapping ratio of a read.

\begin{equation}
\text{Read Base Matching (RBM) ratio} = \frac{\text{# of successfully mapped bases in a read}}{\text{Length of a read}}
\end{equation}

We divided the mapping results based on nine intervals of the RBM ratio: intervals with the RBM ratio higher than 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, and 10%. For example, a group of reads showing the RBM ratio of 35% is considered to be in the three intervals of 30%, 20% and 10%. By using each of the above nine intervals as a cutoff threshold for read mapping to a protein, we checked the correctness of the mapped reads by comparing the genomic positions of the reads and mapped proteins. For example, if a read is mapped to a protein that is originated from a gene covering the read in a genome, then the read is considered as correctly mapped. As expected, there was a negative correlation between the number of correctly mapped reads and the RBM ratio (Fig. 2A) because the use of a more stringent cutoff threshold is likely to prevent reads from being mapped to proteins. The number of correctly mapped reads decreased very slowly between the RBM ratio 50%~70%, and it dropped dramatically after 70% of the RBM ratio. We also examined the ratio of correctly mapped reads to all of the reads mapped to a protein in terms of the RBM ratio (Fig. 2B). When we used 40% as the RBM ratio, the ratio of correctly mapped reads increased to 0.435, and it remained almost the same until the RBM ratio 90%.

![Fig. 2. The number and ratio of mapped reads to a protein. (A) The average number of correctly mapped reads to proteins. A read is considered as correctly mapped if the genomic positions of the read and the mapped protein overlap. (B) The ratio of correctly mapped reads to the whole set of mapped reads per protein.](image-url)
We next analyzed the effect of the RBM ratio on the quality of a read cluster. The quality of a read cluster was measured by the positional diversity of mapped reads in a cluster, called the read dispersion (RD) score, which is defined as:

$$RD\ score = \left( \frac{\sum_{i=1}^{N_p} |S_{p,i} - m_p|}{N_p} \right) \times \frac{1}{L}$$

where, \(N_p\) is the number of reads in a cluster \(P\), \(S_{p,i}\) is a start position of the \(i\)th read on a genome sequence, \(m_p\) is a mean value of start positions of reads in a cluster \(P\) on a genome sequence, and \(L\) is the length of a genome sequence.

We found that the RBM ratio alone may not be a perfect indicator of the cluster quality. For example, Fig. 3 shows two opposite trends in the RD score as a function of the RBM ratio. For a set of 115 proteins, there was a positive correlation between the RBM ratio and the RD score (Fig. 3A) in the RBM ratio intervals of 10~80%. However, for a set of different 100 proteins, an opposite pattern was observed (Fig. 3B). These results indicate that we may need additional omics data other than the RBM ratio obtained from mappings to protein sequences to reliably predict the quality of read clusters at least for \(E.\ coli\).

4. Discussion

In this paper, we developed a simple yet promising read clustering method. Instead of relying on genomic overlapping between NGS read sequences, we utilized mapping between NGS reads and amino acid sequences of proteins. There are similar approaches in NGS transcriptome assemblies\(^1\). Jaffe et al. also utilized proteomic data for the primary annotation of a new genome\(^1\). Dalevi et al. also introduced a similar idea of annotating short reads using proxygens\(^1\). But there exists clear difference between the proposed method and other integrative approaches in the sense that the proposed method aims to increase the reliability of the mapped reads through alignment with amino acid sequences. In addition, we assumed that we were able to collect related proteins and enlarged the candidates of the mapped proteins by using protein-protein interaction information. Because we adopted only proteomics data, the proposed method is limited to the protein coding regions or coding DNA sequences (CDSs). In spite of this limitation, our method provides valuable insight for other researchers. Firstly, the proportion of protein coding regions in prokaryotes is relatively very large so the proposed method can deal with a large fraction of a target genome sequence. Secondly, many assembly errors originated from repeat regions limit the size of assembled genomic fragments to just a few thousand bases especially in complex genomes\(^2\). However, the proposed method can produce a set of highly reliable read clusters by using the mapping information to amino acid sequences of
proteins, which may result in reliable genome assemblies that cannot be well constructed by using the genomic information alone.

However, there remain important shortcomings. Contrary to other sequence assemblers designed to assemble whole genomes, the proposed method can only deal with sequences corresponding to protein coding regions. Although it is possible to enlarge the covered regions by considering more PPI databases, we cannot cover the non-protein coding regions in a genome. In the near future, we will address this problem by incorporating many more multiomics data and developing a unified model for them. Furthermore, we will develop a novel assembly method based on the proposed idea.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2013R1A1A2008183).

References