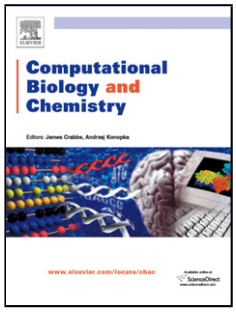

Title: Optimal hybrid sequencing and assembly: Feasibility conditions for accurate genome reconstruction and cost minimization strategy

Author: Chun-Chi Chen Noushin Ghaffari Xiaoning Qian Byung-Jun Yoon

PII: S1476-9271(17)30199-8
DOI: http://dx.doi.org/doi:10.1016/j.compbiolchem.2017.03.016
Reference: CBAC 6674

To appear in: *Computational Biology and Chemistry*

Received date: 29-3-2017
Accepted date: 30-3-2017

Please cite this article as: Chun-Chi Chen, Noushin Ghaffari, Xiaoning Qian, Byung-Jun Yoon, Optimal hybrid sequencing and assembly: Feasibility conditions for accurate genome reconstruction and cost minimization strategy, *<![CDATA[Computational Biology and Chemistry]]>* (2017), http://dx.doi.org/10.1016/j.compbiolchem.2017.03.016

Optimal hybrid sequencing and assembly: Feasibility conditions for accurate genome reconstruction and cost minimization strategy

Chun-Chi Chen[a], Noushin Ghaffari[b], Xiaoning Qian[a,*], Byung-Jun Yoon[a,*]

[a] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA*
[b] *AgriLife Genomics and Bioinformatics, Texas A&M AgriLife Research, Texas A&M University, College Station, TX 77845, USA*

Abstract

Recent advances in high-throughput genome sequencing technologies have enabled the systematic study of various genomes by making whole genome sequencing affordable. Modern sequencers generate a huge number of small sequence fragments called reads, where the read length and the per-base sequencing cost depend on the technology used. To date, many hybrid genome assembly algorithms have been developed that can take reads from multiple read sources to reconstruct the original genome. However, rigorous investigation of the feasibility conditions for complete genome reconstruction and the optimal sequencing strategy for minimizing the sequencing cost has been conspicuously missing. An important aspect of hybrid sequencing and assembly is that the feasibility conditions for genome reconstruction can be satisfied by different combinations of the available read sources, opening up the possibility of optimally combining the sources to minimize the sequencing cost while ensuring accurate genome reconstruction. In this paper, we derive the conditions for whole genome reconstruction from multiple read sources at a given confidence level and also introduce the optimal strategy for combining reads from different sources to minimize the overall sequencing cost. We show that the optimal read set, which simultaneously satisfies the feasibility conditions for genome reconstruction and minimizes the sequencing cost, can be effectively predicted through constrained discrete optimization. Through extensive evaluations based on several genomes and different read sets, we verify the derived feasibility conditions and demonstrate the performance of the proposed optimal hybrid sequencing and assembly strategy.

Keywords: assembly, sequencing, whole genome reconstruction

1. Introduction

Modern high-throughput shotgun sequencing devices sequence genomes using proprietary techniques to generate a large number of relatively short sequence fragments. Depending on the technology used, the sequence fragments, typically called reads, have different lengths. The desired read length affects the choice of sequencing technology and the overall cost of the sequencing experiments. In genome assembly studies, assembly algorithms go through multiple steps to reconstruct the original genome from the numerous tiny reads, where conditions on minimum read length and coverage need to be met to distinguish repeats and faithfully reconstruct the original genome. At present, there are various high-throughput sequencing platforms [1, 2], where the major commercially available technologies for next-generation sequencing (NGS) include Illumina HiSeq, Roche 454, and Life Technologies SOLiD. Additionally, third generation technologies such as PacBio have emerged, which are based on single-molecule sequencing and generate long reads. Depending on the technology used, different sequencing platforms generate reads of different length and quality at different cost. In general, the cost of generating long reads is substantially higher than that of obtaining short reads, while longer reads make the assembly more accurate, particularly when repeated regions and gaps are present in the genome. It is possible to reduce the average sequencing cost by combining reads with different length and cost from multiple sources obtained through different sequencing technologies. This is referred to as hybrid assembly, and hybrid assemblers have been developed to assemble genome sequences based on reads from multiple sources [3–7], which include widely-used algorithms such as CABOG [3] and ALLPATHS-LG [4].

Although there exist various hybrid assemblers that can assist with genome assembly from multiple read sources, there is still a pressing need for rigorous investigation of the *feasibility* of complete genome reconstruction and the overall *sequencing cost* for such hybrid approaches. In recent years, there have been research efforts to examine the minimum requirements for complete genome reconstruction [8] and to derive a lower bound for the read length and the coverage [9] for the case of genome assembly based on a single read source. The increasing popularity of hybrid assembly, as well as the potential quality improve-

*Corresponding author
Email addresses: xqian@ece.tamu.edu (Xiaoning Qian), bjyoon@ece.tamu.edu (Byung-Jun Yoon)