Kurzfassung

Das ultimative Ziel eines Sequenzierprojektes besteht darin, die gesamte DNA-Sequenz des untersuchten Organismus zu erhalten. Mit Sequenziertechnologien der nächsten Generation ist es nicht möglich, das komplette Genom auf einmal zu sequenzieren. Kleine Teile werden sequenziert und in der Assemblierungsphase wird das originale Genom rekonstruiert. Für gewöhnlich erhält man als Ergebnis ein Gerüst des Genoms, ein sog. Scaffold, das aus Contigs und Lücken besteht. Diese Lücken können durch repetitive Sequenzen im Genom oder durch lückenhafte Abdeckung der Genomsequenz durch Reads entstehen. Lücken, die durch lückenhafte Abdeckung entstanden sind, können nur durch zusätzliche Sequenzierung geschlossen werden.

Das Ziel dieses Berufspraktikums war die Entwicklung eines Programms zum Schließen von Lücken mit Hilfe von PacBio Sequenzierdaten. Der Vorteil dieser Sequenziertechnologie liegt in den außergewöhnlich langen Reads. Leider weist diese Technologie aber eine hohe Fehlerrate auf (bis zu 19%). Deshalb besteht der Lückenschließungsprozess aus mehreren Schritten um sehr genaue Lückensequenzen zu generieren. (1) PacBio Reads und Scaffold Contigs werden aligniert. (2) Reads relevant für Lücken werden extrahiert. (3) Für alle Lücken werden Multiple Sequenzalignments mit den zugehörigen Reads durchgeführt. (4) Mit Hilfe des ReAligners wird das Alignment verbessert und eine Konsensus-Sequenz berechnet. (5) Die generierten Konsensus-Sequenzen werden in das Genomgerüst integriert.

Das oben beschriebene Konzept war grundlegend als R Prototyp implementiert worden. Die Ziele dieses Berufspraktikums umfassten die Erweiterung des R Prototypen, die Implementierung einer Java Applikation und die Anwendung des Algorithmus auf zwei bakterielle Genomgerüste. Die wichtigsten Ergebnisse beinhalteten die Fertigstellung der beiden Applikationen, Tests zum Vergleich von verschiedenen Algorithmen und Methoden und die erfolgreiche Anwendung auf die beiden bakteriellen Genome wobei 187 von 189 Lücken geschlossen werden konnten.

Abstract

The ultimate goal of every sequencing project is to obtain the complete DNA sequence of the organism under investigation. Using next generation sequencing technologies it is impossible to sequence the entire genome at once. Small pieces will be sequenced and reconstructed to the original genome in an assembly phase. The common output is a scaffold consisting of contigs and gaps. These gaps can be caused by repetitive sequences in the genome or by missing sequence coverage by reads. Gaps caused by missing sequence coverage can only be closed by additional sequencing.

The aim of this internship was the development of a program for closing gaps with PacBio sequence data. This sequencing technology's advantage is its exceptional long reads. However, it has a high error rate (up to 19%). Therefore the closing process consists of several steps to generate highly accurate gap sequences. (1) PacBio reads are aligned to the scaffold contigs. (2) Reads related to a specific gap are extracted. (3) Multiple sequence alignment for all extracted reads belonging to one gap is performed. (4) The alignment is refined and a consensus sequence is computed using ReAligner. (5) The generated gap sequences are integrated into the draft genome.

The concept described above had been implemented in a basic R prototype. The goals of this internship included the extension of the R prototype, the implementation of a Java application and the application of the algorithm to two bacterial draft genomes. The most important results included finishing both applications, tests to compare various tools and methods and the successful application to both bacterial genomes whereby 187 of 189 gaps could be closed.