Next Generation Sequencing Pipeline Development and Data Analysis

Julian Pierre¹, Jordan Taylor², Amit Upadhyay³, Bhanu Rekepalli³
1 Texas Southern University, 2 Austin Peay State University, 3 University of Tennessee

ABSTRACT

DNA sequencing is one of the most important platforms for study in biological systems today. The High-Throughput-next generation sequencing technologies delivers fast, inexpensive, and accurate genome information. Next Gen Sequencing can produce over 100 times more data than methods based on Sanger Sequencing. The next gen sequencing technologies offered from Illumina/Solexa, ABI/SOLID, 454/Roche, and Helicos has provided unprecedented opportunity for high-throughput functional genomic research. Next gen sequence technology offer novel and rapid ways for genome-wide characterization and profiling of mRNAs, transcription factor regions, and DNA patterns.

Next Generation Sequencing

DNA sequencing was one of the most expensive processes. While the cost of genome sequencing goes down, the cost of analyzing data is still expensive. In the future, the “$1,000 genome will come with a $20,000 analysis price tag.”

Sequence Analysis refers to the process of subjecting a DNA, RNA or protein sequence to a wide range of analytical methods to:
• Compare sequences to find similarities and infer if they are homologous
• To identify the features of the sequence such as gene structure, distribution, introns and exons, and regulation of gene expression
• Identify Sequence differences and variations such as mutations

Data Analysis

In Next-Generation Sequencing, data analysis is one of the most expensive processes. While the cost of genome sequencing goes down, the cost of analyzing data is still expensive. In the future, the “$1,000 genome will come with a $20,000 analysis price tag.”

Sequence Assembly

Sequence Assembly refers to the process of subjecting a DNA, RNA or protein sequence to a wide range of analytical methods to:
• Compare sequences to find similarities and infer if they are homologous
• To identify the features of the sequence such as gene structure, distribution, introns and exons, and regulation of gene expression
• Identify Sequence differences and variations such as mutations

DNA Sequencing

DNA sequencing was one of the most expensive processes. While the cost of genome sequencing goes down, the cost of analyzing data is still expensive. In the future, the “$1,000 genome will come with a $20,000 analysis price tag.”

Downstream Analysis

Next Gen Sequencing Pipeline Development and Data Analysis

When we parsed the BLAST we created a program in Python to group all the nodes related to a specific Contig ID. This grouping is helpful as some Contigs have multiple nodes. Making it easier to search for the desired results.

Comparisons

When we parsed the BLAST we created a program in Python to group all the nodes related to a specific Contig ID. This grouping is helpful as some Contigs have multiple nodes. Making it easier to search for the desired results.

MUMmer is a rapid alignment system used for rapidly aligning entire genomes. It can also align incomplete genomes and can easily handle 100s of contigs from a shotgun sequencing project.

Here is a sample of data where it gets the percentage of each Contig based on the lengths of all the nodes related to that Contig ID.

REFERENCES

• http://www.roche.com/research_and_development/r_d_over view/r_d_sites.htm?id=18
• http://www.illumina.com/technology/solexa_technology.html
• http://blast.ncbi.nlm.nih.gov/Blast.cgi
• https://main.g2b.psu.edu/dan/p/fastq
• http://ori.dhs.gov/education/products/n_illinois_u_datamanagement/datopic.html

CONTACT INFO

Julian Pierre – jut.pierre@yahoo.com
Jordan Taylor – jtaylor74@my.apsu.edu
Amit Upadhyay – aupadhy1@utk.edu
Bhanu Rekepalli – brekupal@utk.edu