Shahjalal University of Science & Technology
Dept. Of Computer Science and Engineering

An Approach to Evaluate and Measure Quality among Multiple Alignments

By
Faisal Ahmed
Reg. No.: 2011331047

Supervisor

Mr. Biswapriyo Chakrabarty
Lecturer, CSE, SUST

An Approach to Evaluate and Measure Quality among Multiple Alignments

By

Faisal Ahmed

Reg. No.: 2011331047

Supervisor

Mr. Biswapriyo Chakrabarty
Lecturer, CSE, SUST

Co-Advisors

Md. Ruhul Amin Shajib
Assistant Professor, CSE, SUST
Md. Eamin Rahman
Assistant Professor, CSE, SUST

Recommendation Letter from Thesis Supervisor

This Student Faisal Ahmed whose thesis entitled **An Approach to Evaluate and Measure Quality among Multiple Alignments** is under my supervision and agree to submit for examination.

Mr. Biswapriyo Chakrabarty
Lecturer,
Dept. of Computer Science & Engineering,
Shahjalal University of Science & Technology

Date : 12 / 04 / 2016

Qualification Form of Bachelor Degree

Student Name : Faisal Ahmed

Thesis Title : An Approach to Evaluate and Measure Quality among Multiple Alignments

This is to certify that the thesis submitted by the student named above in 12 April, 2015. It is qualified and approved by the following persons and committee.

....	....	....
Head of the Dept.	**Chairman,Exam. Committee**	**Supervisor**
Dr Mohammad Reza Selim	Md. Eamin Rahman	Mr. Biswapriyo Chakrabarty
Professor	Assistant Professor	Lecturer

Abstract

Quality measurement of an alignment is an important task in the field of sequence alignment. Due to the advancement in genome sequencing now it is possible to sequence about 18,000 human genome per year. Thus for finding common areas in a genome or aligning coverage is playing a vital role to find common disease genes patterns. Sequence alignment mapping / binary alignment mapping file is a standard format to show result of a specific tool that aligns several reads with a reference genome. There are several tools like IGV, Qualimap to evaluate those files. Qualimap version 2 is the latest version of Qualimap which has outperformed all other tools with various features and statistics giving on input SAM / BAM file. We tried to understand what a SAM file contains and implemented a naïve algorithm named Edit Distance to align multiple short reads with a corresponding reference genome and compared result of those reads with our alignment procedure. For several reads we have been able to outperform our alignment with the dataset of SAM file we are experimenting. But a lot more optimization can be made which will be focused on our future work approach.

Acknowledgments

I would like to thank my supervisor Mr. Biswapriyo Chakrabarty and my co-advisors Md. Ruhul Amin Shajib and Md. Eamin Rahman for their advises and co-operation toward me during my thesis.

List of Figures:

Contents

Chapter 1

Introduction

1.1 Next Generation Sequencing Alignment

Due to the advancement of Next Generation Sequencing methods cost to sequencing a genome has dramatically fallen where on the other hand output data has risen to a whole new level. Scientists are now highly interested to evaluate these sequences and finding various patterns among them. Here comes the alignment concepts. From then various alignment methods have been introduced and they give various types of alignment with finding important patterns between reads and corresponding reference genome. The target is to try to align a read with a reference genome in a way that matches maximum area in a possible way. Thus we can find structural or evolutionary similarity between them. Sequence Alignment Mapping / Binary Alignment Mapping are file format concept to arrange multiple read alignment information in a file. So we can evaluate that file to measure quality of that particular alignment method considering different criteria and factors.

As different methods can create biases among alignment data of same multiple reads and reference genome, it has become obvious to measure quality of them. Therefore a standard file format has been introduced to keep the data of different alignment method based on their biased factors. Now we can evaluate the alignments and measure quality among them with simply using Sequence Alignment Mapping / Binary Alignment Mapping file data.

1.2 Quality Control of Alignments

Quality control is a much more needed criteria of any alignment. Alignment can be done in a numerous way such as following a naïve edit distance or going through noisy areas with skipping or jumping. Such methods create different achievements or failure in fulfilling alignment criteria. Some methods are memory efficient, some can be time efficient. Basically statistical data need to shown to evaluate and measure minimum quality. Several tools are introduced to evaluate Sequence Alignment Mapping / Binary Alignment Mapping file and giving valuable information on them. They lack in fulfilling different criteria or making an alignment better in different ways.

1.3 Tools To Measure Quality

There are several tools have been developed by scientist to evaluate any alignment through Sequence Alignment Mapping or Binary Alignment Mapping file named Picard tools, RNA SeQC, RSeQC, Qualimap. They evaluate SAM or BAM files differently on different factors. The last tool mentioned here Qualimap is the latest edition in all these tools and uses various information gained on SAM or BAM files to get the all of an alignment.

As per we know for now that these tools don't use extra algorithm to make another alignment to make it more standard, we will try to make an exception here by implementing a naïve algorithm to get more standard alignment. A formal algorithm called Edit Distance can be implemented here for the sake of get a standard scale. To make it more linear we have tried to implement a scoring module here. Our target is to get the maximum matching between a read and a reference genome.

In edit distance algorithm there are four cases which can be got. They are- Insertion, Deletion, Matches and Mismatches. A matrix will be made based on scoring and all possible solution module. We have implemented Dynamic Programming (Bottom Up) here to make the matrix more believable and trustable. We will get a path through bottom-left to top-right. That will tell us exact alignment that has been possible so far in the best case. To keep the path linear scoring distribution plays a vital role. So that to get the maximum match the matching score has been given a positive number and others are given negative numbers so that our penalty will be reduced which is the primary goal.

1.4 Approach To Naïve Dynamic Programming

Dynamic Programming is an approach to get the best solution covering all possible best or worst cases. Edit distance can be implemented with dynamic programming in two variations. One is bottom-up and the other one is top-down. We have implemented both these ways and compared. They have given almost same result. Because its' always a linear solution what we will get. As bottom-up don't make stacks like the top-down module it is more time and memory efficient theoretically. So we prefer bottom-up in the most cases.

After implementing edit-distance we traverse through our solution matrix where on one side reference genome is taken and on the other side a sample read is taken. We traversed from top-right to bottom-left to get the actual alignment based on algorithm module. We will describe the algorithm later on the report.

To show differences among the SAM described alignment and the standard scale that we created which is not so memory and time efficient we built some statistical histogram with python. As the implementation of the algorithm was not done completely by that time the histogram shows sample SAM file gives better alignments on maximum reads. For now we have detected some problems about the implementation of the naïve solution as to cut the exact length can be a backward criteria for us to get more standard solution.

How much matches we got through optimal alignments or naïve alignment approaches we must define a scale. We have measured these standards with percentage of identity (POI). It measures percentage of matches considering all the penalties (insertion, deletion, mismatches) it has made. As this is a naïve approach to get the all possible solution and best case by that far we can't consider very lengthy reads that will consume memory and time. Percentage of identity is calculated dividing total matches by total sum of matches, mismatches, insertion and deletions.

Traversing the path might be done with variations. Starting the path from top-right to bottom-left is called global alignment. But there is another kind of alignment here that can also be done. This is called local alignment. We can start the path here from any cell out from the row or column. This will bias the total alignment by cutting a particular part from the read or the reference genome which is not standard but we can observe different things like as what if we try to align a particular part of the read or the read with a particular part of the reference genome. This can give very interesting results like whether a read is more likely to align with a particular part or which part. In pattern finding local alignment plays a vital role more than of the global alignment.

1.5 Evaluating SAM File

Understanding the sequence alignment mapping or binary alignment mapping file is key playing role here in our approach. A better understanding of the file format will make us more comfortable to evaluate it. In short it is called SAM. Different necessary flags indicate different things such as whether the read is revere compliment or not, whether is mapped or not, the chromosome name, read reference name, how this is mapped, number of mismatches, number of matches, number of deletions, number of insertions, position from reference length (this indicates from which position of the reference genome the alignment has been done). Position can be 1-based or 0-based (this is also vaguely included in the file). It also contains the information like whether all the reads are aligned here with the same reference genome or not, next position of the reference or the read. These are some compulsory fields that must be maintained for the file format. There are some other optional fields too like matching scores (where we can know the actual information about the used algorithm and matching criteria on the alignment).

Chapter 2

Background and Related Work

2.1 Next Generation Sequencing

Since 1970 the science of DNA sequencing has passed a long way to come to modern days Next-Generation DNA sequencing methods. In 1977 for the first time scientist gained the ability to make a DNA sequence in a reliable and reproducible manner with the help of Sanger chain termination method. After a decade of that happened invention of first automated, capillary electrophoresis (CE) based sequencing instruments AB370 and AB3730xl. The year was 1998 by then. For human genome projects these instruments became the primary machines. This was called the first generation machine which was considered high-throughput machines for their time. But in 2005 the path took around another dimension. As Genome Analyzer emerged as the second generation machine what made 84 kilobase per run o 1 gigabase per run. The short read and massively parallel technique was a fundamentally different approach that revolutionized sequencing capabilities. The data output of next-generation sequencing (NGS) has surpassed Moor's law that says more than doubling each year.

The Genome Analyser, a single sequencing run could produce roughly one giga-base of data in 2005. This rate climbed to a 1.8 tera-bases of data in a single sequencing run- an astounding 1000 times increase. It is remarkable to reflect on the fact that the first human genome, famously co published in Science and Nature in 2001, required 15 years to sequence and cost nearly 3 billion dollars. In contrast, the HiSeqX Ten, released in 2014, can sequence over 45 human genomes in a single day for approximately 1000 dollar for each sequence.

The introduction of NGS technology has transformed the way scientist thing about genetic information. The 1000 dollar genome enables population-scale sequencing and establishes the foundation for personalized genomic medicine as part of standard medical care. Researchers can now analyse thousands to tens of thousands of samples in a single year. The founding director of the Board Institute of MIT and Harvard and principle leader of the Human Genome Project, states, "The rate of progress is stunning. As costs continue to come down, we are entering a period where we are going to able to get the complete catalog of disease genes. This will allow us to look at thousands of people and see the differences among them, to discover critical genes that cause cancer, autism, heart disease, or schizophrenia."

Next-Generation Sequencing concept is similar to capillary electrophoresis concept that says DNA polymerase catalyzes the incorporation of fluorescently labelled deoxyribonucleotide triphosphates (dTNPs) into a DNA template strand during sequential cycles of DNA synthesis. During each cycle, at the point of incorporation, the nucleotide are identified by fluorophore excitation. The vital difference is that this system doesn't sequence a single fragment of a DNA, instead of that it sequences millions of fragments in a parallel process which is hugely counted. A royal sequencing method named illumine sequencing method is mostly adopted next generation sequencing method so far. It ensures highest accuracy, highest error-free reads and the highest percentage of base calls. This method consists of four steps. They are – Library Preparation, Cluster Generation, Sequencing and Data Analysis.

Library Preparation: The sequences of library are prepared by random fragmented parts of DNA samples. It happens followed by the incident 5' and 3' adapter ligation. Tagmentation process combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

Cluster Generation: For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing.

Sequencing: Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4 reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence context specific errors, even within repetitive sequence regions and homopolymers.

Data Analysis: During data analysis and alignment, the newly identified sequence reads are then aligned to a reference genome. Following alignment, many variations of analysis are possible such as single nucleotide polymorphism (SNP) or insertion-deletion (indel) identification, read counting for RNA methods, phylogenetic or metagenomic analysis and more.

Advances in Sequencing Technology:

Paired-End-Sequencing

A major advance in NGS technology occurred with the development of paired-end sequencing involves sequencing both ends of the DNA fragments in a sequence library and aligning forward and reverse reads as read pairs. In addition to producing twice the number of reads for the same time and effort in library preparation, sequences aligned as read pairs enable more accurate read alignment and detect indels, which is simply not possible with single-read data. Analysis of differential read-pair spacing also allows removal of PCR duplicates, a common artifact resulting from PCR amplification during library preparation. Furthermore, paired-end sequencing produces a higher number of SNV calls following read-pair alignment. While some methods are best served by single-read sequencing, such as small RNA sequencing , most researchers currently use the paired-end approach.

Tunable Coverage and Unlimited Dynamic Range:

The digital nature of NGS allows a virtually unlimited dynamic range for read-counting methods, such as gene expression analysis. Microarrays measure continuous signal intensities and the detection range is limited by noise at the low end and signal saturation at the high end, while NGS quantifies discrete, digital sequencing read counts. By increasing or decreasing the number of sequencing reads, researchers can tune the sensitivity of an experiment to accommodate various study objectives. Because the dynamic range with NGS is adjustable and nearly unlimited,: researchers can quantify subtle gene expression changes with much greater sensitivity than traditional microarray based methods. Sequencing runs can be tailored to zoom in with high resolution on particular regions of the genome, or provide a more expansive view with lower resolution. The ability to easily tune the level of coverage offers several experimental design advantages. For instance, somatic mutations may only exist within: a small proportion of Cells in a given tissue sample. Using mixed tumor-normal cell samples, the region of DNA harboring the mutation must be sequenced at extremely high coverage, often upwards of 1000 times, to detect these low frequency mutations within the mixed cell population. On the other side of the coverage spectrum, a method like genome-wide variant discovery usually requires a much lower coverage level. In this case, the study design involved sequencing many samples (hundreds to thousands) at lower resolution, to achieve greater statistical power within a given population.

Advances in Library Preparation

Library preparation methods for NGS are more rapid and straightforward than for traditional CE-based Sanger sequencing. With lllumina NGS, library preparation has undergone rapid improvements in recent years. The first NGS library prep protocols involved random fragmentation of the DNA or RNA sample, gel-based size-selection, ligation of platform-specific oligonucleotides, PCR amplification, and several purification steps. While the 1-2 days required to generate these early NGS libraries were a great improvement over traditional cloning techniques, current NGS protocols, such as Nextera XT DNA Library Preparation, have reduced the library prep time to less than 90 minutes. PCR-free and gel-free kits are also available for sensitive sequencing methods. PCR-free library preparation kits result in superior coverage of traditionally challenging areas such as high AT/GC-rich regions, promoters, and homopolymeric regions. To advance the process even further, lllumina has combined the precision of digital microfluidics with its ease-ofuse principles to create NeoPrep Library Prep System—a complete, fully automated library preparation instrument. Automation of library preparation will reduce opportunities for error, increase reproducibility, and reduce the amount of hands-on time requited for a process that is often a bottleneck in the sequencing workflow.

Multiplexing:

In addition to the fiSi of data output per run, the sample throughput per run in NGS has also increased over time. Multiplexing allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run. With multiplexed libraries, unique index sequences are added to each DNA fragment during library preparation so that each read can be identified and sorted before final data analysis. With PE sequencing and multiplexing, NGS has dramatically reduced the time to data for multi-sample studies and enabled researchers to go from experiment to data faster and easier than ever before. Flexible, Scalable Instrumentation While the latest NGS platforms can produce massive data output, NGS technology is also highly scalable. Sequencing systems are available for every method and scale of study, from small laboratories to large genome centers . lllumina NGS instruments range from the desktop MiSeq Series, with output ranging from 0.3-15 Gb for small genome, amplicon, or targeted sequencing studies, to the colossal HiSeqX Ten fleet, which can generate an impressive, 16-18 Tb per run for population-scale studies. Flexible run configurations are also engineered info the design of lllumina NGS sequencers. For example, the HiSeqff 2500 System offers 2 run modes and single or dual flow cell sequencing while the NextSeqff Seres offers 2 flow ctll types to accommodate different throughput requirements. The HiSeq 3000/4000 Series uses the same patterned flow cell technology as the HiSeqX instruments for cost-effective production-scale sequencing. This flexibility allows researchers to configure runs tailored to their specific study requirements, with the instrument of their choice.

NGS Methods

Next-generation sequencing platforms enable a wide variety of methods, allowing researchers to ask virtually any question related to the genome, transcriptome or epigenome of any organism. Sequencing methods differ primarily by how the DNA or RNA samples are obtained (eg, organism, tissue type, normal vs. affected, experimental conditions) and by the data analysis options used. After the sequencing libraries are prepared, the actual sequencing stage remains fundamentally the same regardless of the method. There are a number of standard library preparation kits that offer protocols for whole-genome sequencing, mRNA-Seq, targeted sequencing (such as exome sequencing or 16S sequencing), custom-selected regions, protein-binding regions, and more. Although the number of NGS methods is constantly growing, a brief overview of the most common methods is presented here.

Genomics:

Whole-Genome Sequencing:

Microarray-based, genome-wide association studies (GWAS) have been the most common approach for identifying disease associations across the whole genome. While GWAS microarrays can interrogate over 4 million markers per sample, the most comprehensive method of interrogating the 3,2 billion bases of the human genome is with whole genome sequencing (WGS). The rapid drop in sequencing cost and the ability of WGS to rapidly produce large volumes of data make it a powerful tool for genomics research. While WGS is commonly associated with sequencing human genomes, the scalable, flexible nature of the technology makes it equally useful for sequencing any species, such as agriculturally important livestock, plant genomes, or disease-related microbial genomes. This broad utility was demonstrated during the recent E. coli outbreak in Europe in 8011, which prompted a rapid scientific response. Using the latest NGS systems, researchers quickly sequenced the bacterial strain, enabling them to better track the origins and transmission of the outbreak as well as Identify genetic mutations conferring the increased virulence.

Exome Sequencing:

Perhaps the most widely used targeted sequencing method is exome sequencing. The exome represents less than 2% of the human genome, but contains a majority of known disease-causing variants, making whole-exome sequencing a cost-effective alternative to whole-genome sequencing. With exome sequencing, the protein-coding portion of the genome is selectively captured and sequenced. It can efficiently identify variants across a wide range of applications, including population genetics, genetic disease, and cancer studies.

De Novo Sequencing

De novo sequencing refers to sequencing a novel genome where there is no reference sequence available for alignment. Sequence reads are assembled as contigs and the coverage quality of de novo sequence data depends on the size and continuity of the contigs (ie, the number of gaps in the data). Another important factor in generating high-quality de novo sequences is the diversity of insert sizes included in the library. Combining short-insert paired-end and long-insert mate pair sequences is the most powerful approach for maximal coverage across the genome. The combination of insert sizes enables detection of the widest range of structural variant types and is essential for accurately identifying more complex rearrangements. The short-insert reads, sequenced at higher depths, can fill in gaps not covered by the long inserts, which are often sequenced at lower read depths. Therefore, using a combined approach results in higher-quality assemblies. In parallel with NGS technology improvements, many algorithmic advances have emerged in sequence assemblers for short-read data. Researchers can perform high quality de novo assembly using NGS reads and publicly available short-read assembly tools. In many instances, existing computer resources in the laboratory are enough to perform de novo assemblies. For example, The E. coli genome can be assembled in as little as 15 minutes using a 32- bit Windows desktop computer with 32 GB of RAM.

Targeted Sequencing

With targeted sequencing, a subset of genes or regions of the genome are isolated and sequenced. Targeted sequencing allows researchers to focus time, expenses, and data analysis on specific areas of interest and enables sequencing at much higher coverage levels. For example, a typical WGS study achieves coverage levels of 30x-50x per genome, while a targeted re-sequencing project can easily cover the target region at 500x-1000x or higher. This higher coverage allows researchers to identify rare variants that would be too rare and too expensive to identify with WGS or CE-based sequencing. Targeted sequencing panels can be purchased with fixed, preselected content or can be custom designed. A wide variety of targeted sequencing library prep kits are available, including kits with probe sets focused on specific areas of interest such as cancer, cardiomyopathy, autism, or custom probe sets. With custom designs, researchers can target regions of the genome relevant to their specific research interests. Custom targeted sequencing is ideal for examining genes in specific pathways, or for follow-up studies from GWAS or WGS. Illumina currently supports 2 methods for targeted sequencing— target enrichment and amplicon generation. With target enrichment, specific regions of interest are captured by hybridization to bio-tinylated probes, then isolated by magnetic pull down. Target enrichment captures between 20 kb-62 Mb regions depending on the library prep kit parameters. The second method, amplicon sequencing, involves the amplification and purification of regions of interest using highly multiplexed. PCR ollgpsets. Amplicon sequencing allows researchers to sequence 28-’ 536 targets at a time, spanning 150 bp-1.5 kb per target, depending on the library prep kit used. This highly multiplexed approach enables a wide range of applications for the discovery, validation, or screening of genetic variants. Amplicon sequencing is particularly useful for the discovery of rare somatic mutations in complex samples (eg, cancerous tumors mixed with germline DNA). Another common amplicon application is sequencing the bacterial 16S rRNA gene across multiple species, a widely used method for phylogeny and taxonomy studies, particularly in diverse metagenomic samples.

Transcriptomics:

Library preparation methods for RNA sequencing (RNA-Seq) typically begin with total RNA sample preparation followed by a ribosome removal step. The total RNA sample is then converted to cDNA before standard NGS library preparation. RNA-Seq focused on mRNA, small RNA, noncoding RNA, or microRNAs can be achieved by including additional Isolation or enrichment steps before cDNA synthesis.

Total RNA and mRNA Sequencing:

Transcriptome sequencing is a major advance in the study of gene expression because it allows a snapshot of the whole transcriptome rather than a predetermined subset of genes. Whole-transcriptome sequencing provides a comprehensive view of a cellular transcriptional profile at a given biological moment and greatly enhances the power of RNA discovery methods. As with any sequencing method, an almost unlimited dynamic range allows identification and quantitation of both common and rare transcripts. Additional capabilities include aligning sequencing reads across splice junctions, as well as detection of isoforms, novel transcripts, and gene fusions. Library preparation kits that support precise detection of strand orientation are for both total RNA-Seq and mRNA-Seq methods.

Targeted RNA Sequencing:

Targeted RNA sequencing is a method for measuring transcripts of interest for differential expression, allele-specific expression, as well as detection of gene-fusions, isoforms, cSNPs, and splice junctions. Illumina TruSeq Targeted RNA Sequencing kits included preconfigured, experimentally validated panels focused on specific cellular pathways or disease states such as apoptosis, cardio toxicity, NFkB pathway, and more. Custom content can also be designed and ordered for the analysis of specific genes of interest. Targeted RNA sequencing is a powerful method for the investigation of specific pathways of interest or for the validation of gene expression microarray or whole transcriptome sequencing results.

Small RNA and Noncoding RNA Sequencing:

Small, noncoding RNA, or microRNA s are short, 18-22 bp nucleotides that play a role in the regulation of gene expression often as gene repressors or silencers. The study of microRNAs has grown as their role in transcriptional and translational regulation has become more evident.

Epigenomics:

While genomics involves the study of heritable or acquired alterations in the DNA sequence, epigenetics is the study of heritable changes in gene activity caused by mechanisms other than DNA sequence changes. Mechanisms of epigenetic activity include DNA methylation, small RNA-meditated regulation, DNA-protein interactions, Irstone modification and more.

Methylation Sequencing:

A critical focus in epigenetics is the study of cytosine methylation (5-mC) states across specific areas of regulation such as promoters or heterochromatin. Cytosine methylation can significantly modify temporal and spatial gene expression and chromatin remodelling. While there are many methods for the study of genetic methylation, methylation sequencing leverages the advantages of NGS technology and genome-wide analysis while assessing methylation states at the single-nucleotide level. Two methylation sequencing methods are widely used: whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). With WGBS-Seq, sodium bisulfite chemistry is converts non-methylated cytosine to uracil, which are then converted to thymine in the Sequence reads or data output. In RRBS-Seq, DNA is digested with Mspl—a restriction enzyme unaffected by methylation status. Fragments in the 100-150 bp size range are isolated to enrich for CpG and promoter containing DNA regions. Sequencing libraries are then constructed using the standard NGS protocols.

ChIP Sequencing:

Protein-DNA or protein-RNA interactions have a significant impact on many biological processes and disease states. These interactions can be surveyed with NGS by combining chromatin immune-precipitation (ChIP) assays and NGS methods. ChIP-Seq protocols begin with the chromatin immune-precipitation step (ChIP protocols vary widely as they must be specific to the species, tissue type, and experimental conditions).

Ribosome Profiling:

Ribosome profiling is a method based on deep sequencing of ribosome-protected mRNA fragments. Purification and sequencing of these fragments provides a “snapshot” of all the ribosomes active in a cell at a specific time point. This information can determine what proteins are being actively translated in a cell, and can be useful for investigating translational control, measuring gene expression, determining the rate of protein synthesis, or predicting protein abundance. Ribosome profiling enables systematic monitoring of cellular translation processes and prediction of protein abundance. Determining what regions of a transcript are being translated can help define the proteome of complex organisms. With NGS, ribosome profiling allows detailed and accurate in vivo analysis of protein production.

2.2 Illumina NGS Solutions

Illumina DNA-to-Data NGS Solutions

The Illumina NGS Workflow

Illumina offers a comprehensive, end-to-end solution for every step of the NGS sequencing workflow, from library preparation to final data analysis (Figure 10). Library preparation kits are available for all NGS methods including WGS, exome sequencing, targeted sequencing, RNA sequencing, and more. Illumina library preparation protocols can accommodate a range of throughput needs, from manual protocols for smaller laboratories to fully automated library preparation workstations for larger laboratories or genome centers. Likewise, Illumina offers aill portfolio of sequencing platforms, from the desktop MiSeq Series to the factory-scale HiSeq, X Series that deliver the right level of speed, capacity, and cost for various laboratory or sequencing center requirements. For the last, step in the NGS workflow, Illumina offers biologist-friendly bioinformatics tools that are easily accessible through the web, on instrument, or through onsite servers.

Integrated Data Analysis

BaseSpace is a bioinformatics software solution for analysing, storing, and sharing NGS data. BaseSpace can be accessed through the internet for data analysis and storage in the Illumina cloud or through an installed local server for data analysis and storage onsite. A major advantage of working ip the BaseSpace environment is that it is directly integrated with Illumina sequencing systems. On-instrument access to BaseSpace enables the integration of many workflow steps, including library prep planning with BaseSpace Prep, run set-up and chemistry validation, and real-time automatic data transfer to the BaseSpace computing environment. The NGS workflow then proceeds seamlessly through alignment and subsequent data analysis steps with BaseSpace Apps. BaseSpace Apps offer a wide variety of analysis pipelines including analysis for de novo assembly, SNP and indel variant analysis, RNA expression profiling, 16S metagenomics, tumor-normal comparisons, epigenetic/ gene regulation analysis, and many more. Illumina collaborates closely w th commercial and academic software developers to create a full ecosystem of data analysis tools that address the needs of various research objectives. In the final stages of the NGS workflow, data can be shared with collaborators or delivered instantly to customers around the world.

2.3 Sequence Alignment

After the revolution in Genome Sequencing finding pattern or same partition of several genome has become another important part of research. This is called alignment. Sequence alignment has become a vital part of bioinformatics as it tells the pattern or viability of a specific gene which causes several diseases. Amino acid sequence alignment and analysis is central to most biochemical and molecular biology applications. Although it should be possible to retrieve all the information we need about a protein directly from its sequence, looking at a sequence without prior knowledge and experience is like reading a text in a foreign language: we may recognize the letters, but we do not understand the meaning and are unable to extract the information. Still, when proteins are concerned, we have learned to extract a substantial part of the information from detailed sequence analysis, using for example multiple sequence alignment. In a multiple sequence alignment a given sequence is compared to a group of evolutionary related sequences from other organisms. The pleasant fact is that we will always find a related protein from some other organism. When we say ”related” we mean that they belong to the same family, the members of which usually perform a similar function in different organisms. We know that in such cases the main characteristic features of a protein sequence and the protein tertiary structure are conserved. Since conservation of function assumes that a certain number of amino acid residues within a protein family are conserved, we need to have some instruments to assess the degree of conservation of each sequence. To assist in the process, alignment techniques and scoring schemes for sequence alignment have been developed. Here I will discuss the basic concepts behind these techniques and will provide two examples to guide you in making sequence alignment using resources available on the Internet. Since we focus here on structural bioinformatics, the alignments we make will be interpreted in terms of the three-dimensional structure. We will also discuss what structural information may be identified in a sequence alignment, how to relate sequence and structural information and how to make use of available structural data to make better sequence alignment.

When making a sequence alignment we need to take into account several factors. For example, we need to understand the effect of replacements of one amino acid by another (amino acid substitutions) in different sequences. Thus, some substitutions are conservative, i.e., they will not introduce any substantial disturbances in the protein structure, while others may have dramatic effect on the structure and function of the protein in question and they are normally rather rare. To account for the different types of substitutions, there are specially designed so called substitution matrices, which can be used for making a correct alignment and for calculating the score of the alignment. Structural information may also be used to assist us in making a correct alignment, for example in understanding the effect of amino acid substitution.

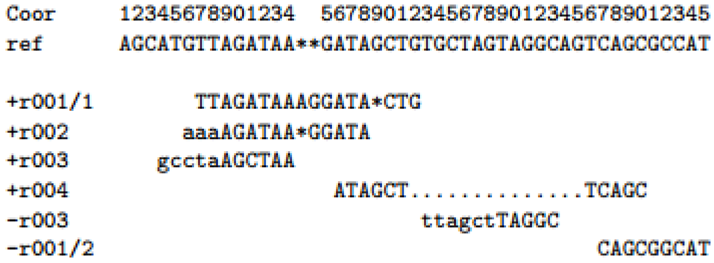
2.4 Tools To Make Alignment

There are several tools to make alignment between two or multiple sequence and to calculate the coverage among them. Example – Blast, CS-Blast, Cudasw++, Diamond, Fasta, GGSearch / GLSearch, Genoogle, HMMER, HHpred / HHsearch, IDF, Infernal, Klast, Userach, parasail, PSI-Blast, PSI-Search, ScalaBlast, Sequilab, SAM, SSearch, Swaphi, Swaphi-LS, SIMM, Swipe, Acana, AlignMe, Bioconductor, BioPerl, Blastz, Lastz, CUDAlign, DNADot, DOTLET, FEAST, G-Pas, GapMis, GGSearch, GLSearch, JAligner, K*Sync, LALIGN, NW-align, mAlign, matcher, McAlign2, MUMmer, needle, Ngila, NW, parasail, Path, PatternHunter etc etc.

To know how much quality they ensured in their tool we need to analysis the alignment result file. This result previously was contained in several doc or other database systems. But nowadays a standard format is maintained. It is called Sequence Alignment Mapping or Binary Sequence Alignment mapping file.

The SAM Format Specification
SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

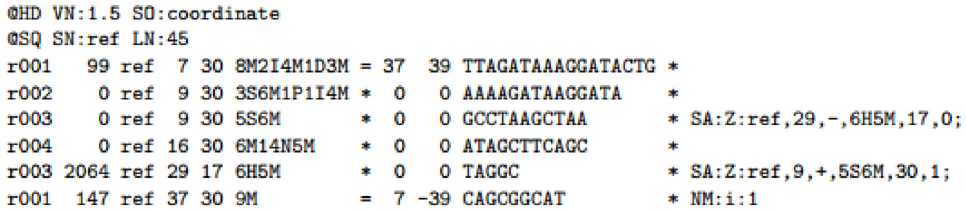
Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment



```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

2.5 SAM File Format Terminologies

Terminologies and Concepts:

Template:

A DNA/RNA sequence part of which is sequenced on a sequencing machine or assembled from raw sequences.

Segment:

A contiguous sequence or subsequence.

Read:

A raw sequence that comes off a sequencing machine. A read may consist of multiple segments. For sequencing data, reads are indexed by the order in which they are sequenced.

Linear alignment:

An alignment of a read to a single reference sequence that may include insertions, deletions, skips and clipping, but may not include direction changes (i.e. one portion of the alignment on forward strand and another portion of alignment on reverse strand). A linear alignment can be represented in a single SAM record.

Chimeric alignment:

An alignment of a read that cannot be represented as a linear alignment. A chimeric alignment is represented as a set of linear alignments that do not have large overlaps. Typically, one of the linear alignments in a chimeric alignment is considered the "representative" alignment, and the others are called "supplementary" and are distinguished by the supplementary alignment flag. All the SAM records in a chimeric alignment have the same QNAME and the same values for 0x40 and 0x80 flags. The decision regarding which linear alignment is representative is arbitrary.

Read alignment:

A linear alignment or a chimeric alignment that is the complete representation of the alignment of the read.

Multiple mapping:

The correct placement of a read may be ambiguous, e.g. due to repeats. In this case, there may be multiple read alignments for the same read. One of these alignments is considered primary. All the other alignments have the secondary alignment flag set in the SAM records that represent them. All the SAM records have the same QNAME and the same values for 0x40 and 0x80 flags. Typically the alignment designated primary is the best alignment, but the decision may be arbitrary.

1-based coordinate system:

A coordinate system where the first base of a sequence is one. In this coordinate system, a region is specified by a closed interval. For example, the region between the 3^{rd} and the 7th bases inclusive is [3, 7]. The SAM, VCF, GFF and Wiggle formats are using the 1-based coordinate system.

0-based coordinate system:

A coordinate system where the first base of a sequence is zero. In this coordinate system, a region is specified by a half-closed-half-open interval. For example, the region between the 3rd and the 7th bases inclusive is [2, 7). The BAM, BCFv2, BED, and PSL formats are using the 0-based coordinate system.

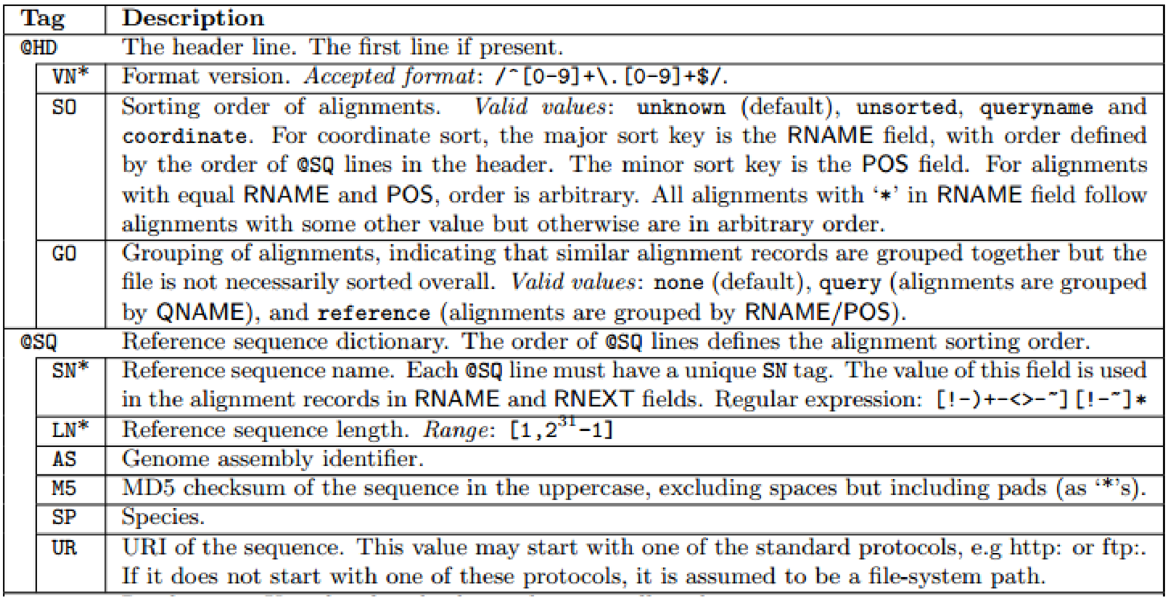
Phred scale:

Given a probability 0 ¡ p ≤ 1, the phred scale of p equals −10 log10 p, rounded to the closest integer.

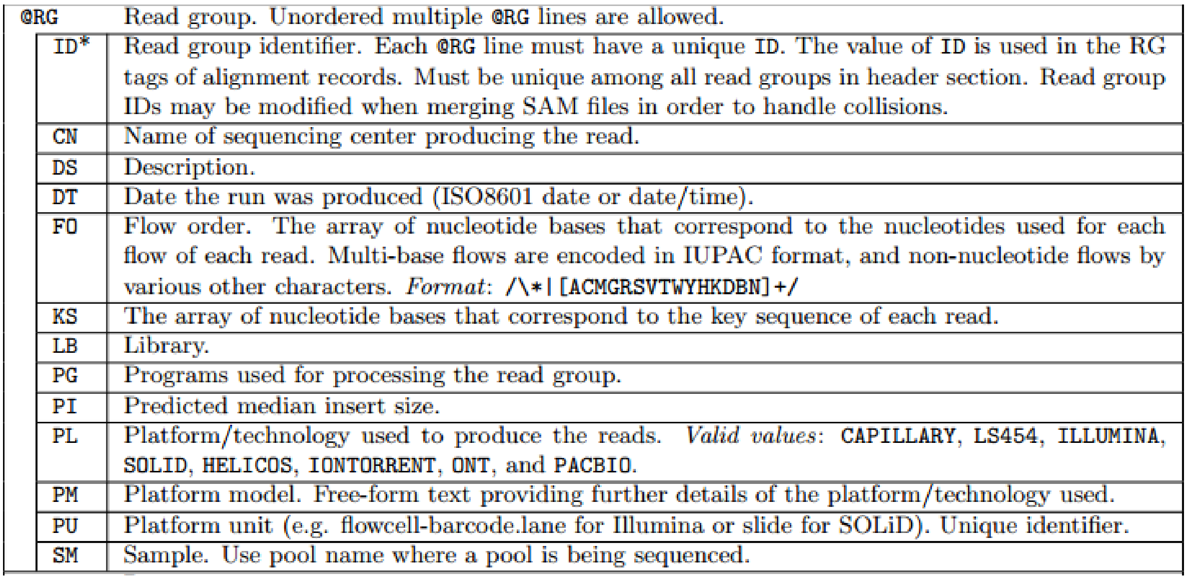
The header section:

Each header line begins with the character '@' followed by one of the two-letter header record type codes defined in this section. In the header, each line is TAB-delimited and, apart from @CO lines, each data field follows a format 'TAG:VALUE' where TAG is a two-character string that defines the format and content of VALUE. Thus header lines match /ˆ@[A-Z][A-Z](“t[A-Za-z][A-Za-z0-9]:[-˜]+)+$/ or /ˆ@CO“t.*/.

The following table describes the header record types that may be used and their predefined tags. Tags listed with '*' are required; e.g., every @SQ header line must have SN and LN fields. As with alignment optional fields, you can freely add new tags for further data fields. Tags containing lowercase letters are reserved for local use and will not be formally defined in any future version of this specification.



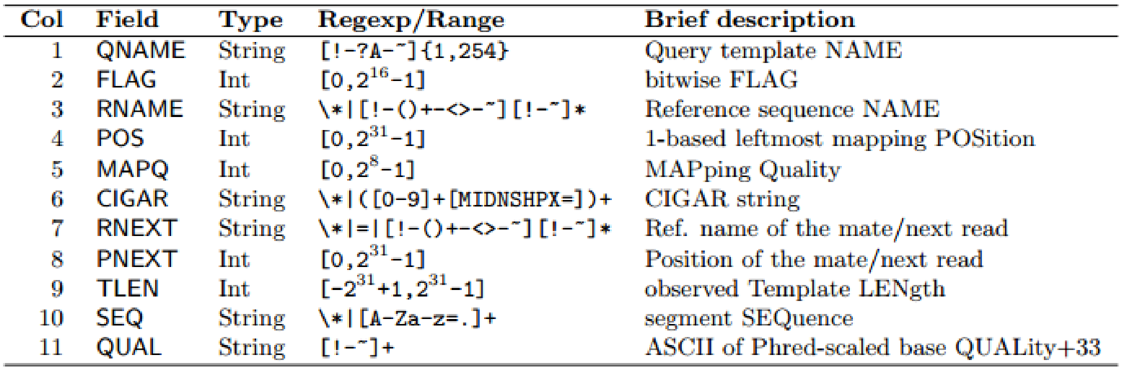
Tag		Description
@HD		The header line. The first line if present.
	VN*	Format version. *Accepted format*: `/^[0-9]+\.[0-9]+$/`.
	SO	Sorting order of alignments. *Valid values*: `unknown` (default), `unsorted`, `queryname` and `coordinate`. For coordinate sort, the major sort key is the RNAME field, with order defined by the order of `@SQ` lines in the header. The minor sort key is the POS field. For alignments with equal RNAME and POS, order is arbitrary. All alignments with '`*`' in RNAME field follow alignments with some other value but otherwise are in arbitrary order.
	GO	Grouping of alignments, indicating that similar alignment records are grouped together but the file is not necessarily sorted overall. *Valid values*: `none` (default), `query` (alignments are grouped by QNAME), and `reference` (alignments are grouped by RNAME/POS).
@SQ		Reference sequence dictionary. The order of `@SQ` lines defines the alignment sorting order.
	SN*	Reference sequence name. Each `@SQ` line must have a unique `SN` tag. The value of this field is used in the alignment records in RNAME and RNEXT fields. Regular expression: `[!-)+-<>-~][!-~]*`
	LN*	Reference sequence length. *Range*: `[1,2`^{31}`-1]`
	AS	Genome assembly identifier.
	M5	MD5 checksum of the sequence in the uppercase, excluding spaces but including pads (as '`*`'s).
	SP	Species.
	UR	URI of the sequence. This value may start with one of the standard protocols, e.g http: or ftp:. If it does not start with one of these protocols, it is assumed to be a file-system path.



@RG		Read group. Unordered multiple `@RG` lines are allowed.
	ID*	Read group identifier. Each `@RG` line must have a unique `ID`. The value of `ID` is used in the RG tags of alignment records. Must be unique among all read groups in header section. Read group IDs may be modified when merging SAM files in order to handle collisions.
	CN	Name of sequencing center producing the read.
	DS	Description.
	DT	Date the run was produced (ISO8601 date or date/time).
	FO	Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Multi-base flows are encoded in IUPAC format, and non-nucleotide flows by various other characters. *Format*: `/\*|[ACMGRSVTWYHKDBN]+/`
	KS	The array of nucleotide bases that correspond to the key sequence of each read.
	LB	Library.
	PG	Programs used for processing the read group.
	PI	Predicted median insert size.
	PL	Platform/technology used to produce the reads. *Valid values*: `CAPILLARY`, `LS454`, `ILLUMINA`, `SOLID`, `HELICOS`, `IONTORRENT`, `ONT`, and `PACBIO`.
	PM	Platform model. Free-form text providing further details of the platform/technology used.
	PU	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier.
	SM	Sample. Use pool name where a pool is being sequenced.

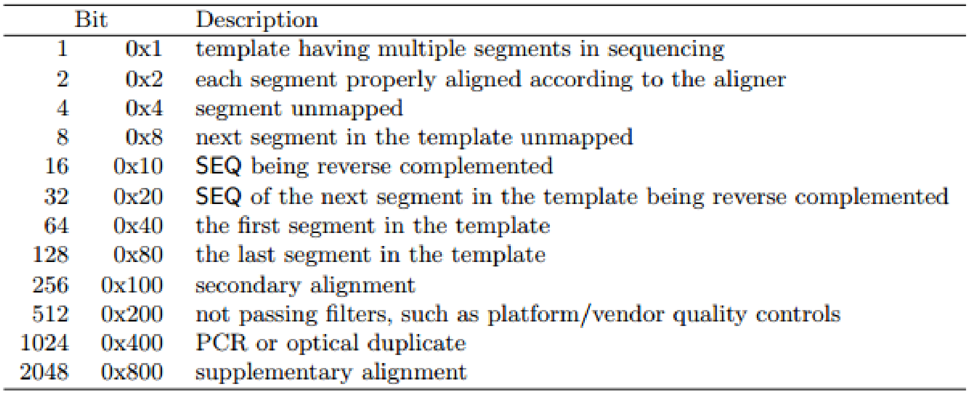
The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:



<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>[0,2^{16}-1]</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-&lt;&gt;-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>[0,2^{31}-1]</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>[0,2^{8}-1]</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-&lt;&gt;-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>[0,2^{31}-1]</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>[-2^{31}+1,2^{31}-1]</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

1. QNAME: Query template NAME. Reads/segments having identical QNAME are regarded to come from the same template. A QNAME '*' indicates the information is unavailable. In a SAM file, a read may occupy multiple alignment lines, when its alignment is chimeric or when multiple mappings are given.
2. FLAG: Combination of bitwise FLAGs. Each bit is explained in the following table:



Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

• For each read/contig in a SAM file, it is required that one and only one line associated with the read satisfies 'FLAG & 0x900 == 0'. This line is called the *primary line* of the read.

• Bit 0x100 marks the alignment not to be used in certain analyses when the tools in use are aware of this bit. It is typically used to flag alternative mappings when multiple mappings are presented in a SAM.

• Bit 0x800 indicates that the corresponding alignment line is part of a chimeric alignment. A line flagged with 0x800 is called as a *supplementary* *line*.

• Bit 0x4 is the only reliable place to tell whether the read is unmapped. If 0x4 is set, no assumptions can be made about RNAME, POS, CIGAR, MAPQ, and bits 0x2, 0x100, and 0x800.

• Bit 0x10 indicates whether SEQ has been reverse complemented and QUAL reversed. When bit 0x4 is unset, this corresponds to the strand to which the segment has been mapped. When 0x4 is set, this indicates whether the unmapped read is stored in its original orientation as it came off the sequencing machine.

• If 0x40 and 0x80 are both set, the read is part of a linear template, but it is neither the first nor the last read. If both 0x40 and 0x80 are unset, the index of the read in the template is unknown. This may happen for a non-linear template or the index is lost in data processing.

• If 0x1 is unset, no assumptions can be made about 0x2, 0x8, 0x20, 0x40 and 0x80.

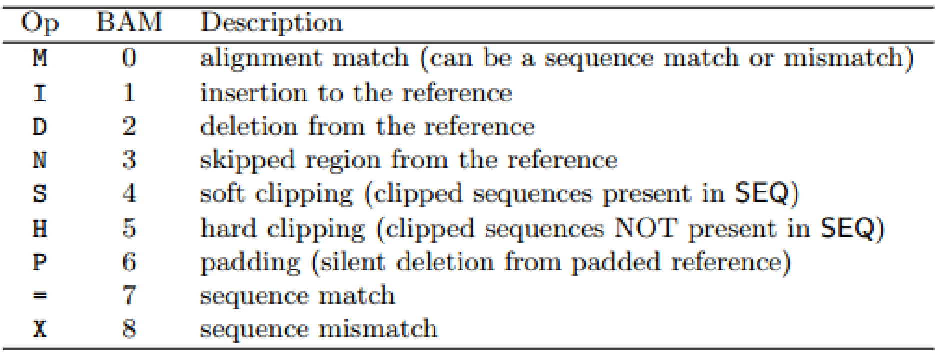
• Bits that are not listed in the table are reserved for future use. They should not be set when writing and should be ignored on reading by current software.

3. RNAME: Reference sequence NAME of the alignment. If @SQ header lines are present, RNAME (if not '*') must be present in one of the SQ-SN tag. An unmapped segment without coordinate has a '*' at this field. However, an unmapped segment may also have an ordinary coordinate such that it can be placed at a desired position after sorting. If RNAME is '*', no assumptions can be made about POS and CIGAR.

4. POS: 1-based leftmost mapping position of the first matching base. The first base in a reference sequence has coordinate 1. POS is set as 0 for an unmapped read without coordinate. If POS is 0, no assumptions can be made about RNAME and CIGAR.

5. MAPQ: Mapping Quality. It equals −10 log10 Pr*(*mapping position is wrong*)*, rounded to the nearest integer. A value 255 indicates that the mapping quality is not available.

6. CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

• H can only be present as the first and/or last operation.

• S may only have H operations between them and the ends of the CIGAR string.

• For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.

• Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

7. RNEXT: Reference sequence name of the primary alignment of the NEXT read in the template. For the last read, the next read is the first read in the template. If @SQ header lines are present, RNEXT (if not '*' or '=') must be present in one of the SQ-SN tag. This field is set as '*' when the information is unavailable, and set as '=' if RNEXT is identical RNAME. If not '=' and the next read in the template has one primary mapping (see also bit 0x100 in FLAG), this field is identical to RNAME at the primary line of the next read. If RNEXT is '*', no assumptions can be made on PNEXT and bit 0x20.

8. PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. This field equals POS at the primary line of the next read. If PNEXT is 0, no assumptions can be made on RNEXT and bit 0x20.

9. TLEN: signed observed Template LENgth. If all segments are mapped to the same reference, the unsigned observed template length equals the number of bases from the leftmost mapped base to the rightmost mapped base. The leftmost segment has a plus sign and the rightmost has a minus sign. The sign of segments in the middle is undefined. It is set as 0 for single-segment template or when the information is unavailable.

10. SEQ: segment SEQuence. This field can be a '*' when the sequence is not stored. If not a '*', the length of the sequence must equal the sum of lengths of M/I/S/=/X operations in CIGAR. An '=' denotes the base is identical to the reference base. No assumptions can be made on the letter

cases.
11. QUAL: ASCII of base QUALity plus 33 (same as the quality string in the Sanger FASTQ format). A base quality is the phred-scaled base error probability which equals −10 log10 Pr–base is wrong˝. This field can be a '*' when quality is not stored. If not a '*', SEQ must not be a '*' and the length of the quality string ought to equal the length of SEQ.

As Sequence Alignment Mapping says a lot story about an alignment and how a particular tool / software aligned multiple reads, our task here is to evaluate and make a standard scale to measure how that alignment worked.

Chapter 3

Current Technology

There are several tools to evaluate a Sequence Alignment Mapping. Such as Picard tools, RNA-SeQC, RSeQC, Qualimap, Integrative Genomic Viewer (IGV tools). Qualimap (version 2) is the latest advancement in this field. So we will describe about this tool here.

As next generation sequencing has risen to a new extent a lot of novel method has been developed to evaluate sequence alignment data and making alignment among several sequences. The sequence alignment/map (SAM) and the binary alignment/map (BAM) formats have become the standards used for representation of nucleotide sequence alignments for these algorithms (Li et al., 2009). The results from such alignments can be used in subsequent analyses, such as genome-wide comparative studies, to drive conclusions concerning a variety of biological processes, such as gene expression and epigenomic modifications. SAM/BAM files usually contain the information from tens to hundreds of millions of reads, and the quantity of data contained in these files is continuously increasing. Unfortunately, SAM/BAM data files frequently contain biases that are introduced by sequencing technologies, during sample preparation, (Harismendy et al., 2009; Metzker, 2009) and/or the selected mapping algorithm (Flicek and Birney, 2009). Therefore, one of the fundamental requirements during analysis of these data is to perform quality control, i.e. to get an idea of how reliable mapping data are and how well data fit with the expected outcome. To these ends, we have developed Qualimap, a Java application that aims to facilitate the quality-control analysis of mapping data. Recently, some efforts have been made to facilitate this task, for example see SAMStat (Lassmann et al., 2011), RNA-SeQC (DeLuca et al., 2012) or Picard; Qualimap advances this area by providing some additional features.

Qualimap is a multi-threaded application built in Java and R that provides a graphical user interface to perform the quality control of alignment sequencing data. A command-line interface has been also implemented, so Qualimap can be incorporated in particular analysis pipelines. The first step in the program is to select the type of analysis to be run, which can be: BAM QC for alignment data—with optionally a set of regions of interest—or Count QC for count data. When dealing with alignment data, the main input for Qualimap is the BAM file to be analyzed. The application processes information by splitting the reference genome into a given number of windows (400 by default), collecting the information and parallelizing the process where appropriate. The results are summarized in a dedicated panel and graphically represented in different charts, which can be exported for further analysis. The user can also concentrate the analysis to specific regions of interest by including a general feature format (GFF) or a browser extensible data (BED) file. In this case, the information is shown separately for reads that are mapped inside or outside of the defined regions. Qualimap also provides insights into mapping performance by studying the read counts overlaps with genomic features of interest. These read counts can be loaded into the program as a text file that can be directly computed by using a dedicated tool in Qualimap. For example, the saturation rate for the detected features can be analysed with respect to the sequencing depth, unveiling whether more features could be detected by increasing the sequencing depth. This is of particular interest, for example, in RNA-seq assays. Likewise, it is possible to analyse the read counts separately in user-defined groups of features.

In order to show the performance of Qualimap with read-count data, we made use of the study by Marioni et al. (2008) in which the authors estimated differences in gene expression profiles between human liver and kidney RNA samples using multiple sequencing replicates. For the sake of clarity, in this work, we considered data from the kidney sample only. To begin, we mapped the reads using TopHat (Trapnell et al., 2009) with default parameters; next, we obtained the read counts for each gene using the tool provided in Qualimap, discarding non-uniquely mapped reads and using the Ensembl 64 annotation (Flicek et al., 2011). Figure 1b shows a per-biotype detection plot. There is an enrichment of the protein-coding biotype in the mapped reads, as expected, but also a significant number of other detected biotypes such as pseudogene, lincRNA and processed-transcript. Likewise, the saturation point was not reached, as more genes were detected when sequencing depth was increased.

The simultaneous comparison of multiple samples allows examination of consistency between samples and visual detection of outliers. To estimate the variability between analysed datasets, Qualimap performs a principal component analysis based on specific features derived from the alignment, including coverage, GC content, insert size and mapping quality. Qualimap 2 also introduces a novel analysis mode called RNA-seq QC. This mode allows computation of metrics specific to RNA-seq data, including per transcript coverage, junction sequence distribution, genomic localization of reads, 50–30 bias and consistency of the library protocol. A detailed comparison of Qualimap to RSeQC and RNA-seq QC tools that are focused on a similar goal can be found in. The most significant difference to other tools is the subsequent RNA-seq QC analysis step that Qualimap performs after computation of read counts. The mode Counts QC was completely redesigned to allow processing of multiple samples. Normally, this mode estimates the quality of the read counts that are derived from intersecting sequencing alignments within genomic features. Counts are usually applicable for analysis of differential gene expression from RNA-seq data. Having multiple biological replicates per condition is common in RNA-seq experiments; therefore, it is beneficial to be able to analyze counts data from all generated datasets simultaneously. Multisample analysis of read counts allows inspection of sample grouping, as well as discovery of outliers and batch effects. Similar to the previous version, the Counts QC mode estimates the saturation of sequencing depth, read count densities, correlation of samples and distribution of counts among classes of selected features. Additionally, new plots that explore the relationship between expression values and GC-content or transcript lengths are available for users. Counts QC is based on the NOIseq package for gene expression estimation. The analysis results include a combined overview of the counts from all samples along with a QC report for each individual sample. Moreover, the analyzed datasets can have different conditions, e.g. treated and untreated. In this case, plots comparing groups of sample counts corresponding to particular conditions are generated.

Qualimap 2 is an application for exploratory analysis and QC of HTS alignment data written in Java and R. The major enhancement over the previous version lies in the ability to perform multi-sample analyses. Additionally, a large number of bug fixes and enhancements have been implemented since the initial release. In the present version, we have kept the concept of a simple, user-friendly application that follows an ‘open-source’ path. Qualimap 2 has gathered a community of users who frequently suggest new features and contribute their code. Notably, most of the novel features in BAM QC mode were proposed and tested by users.

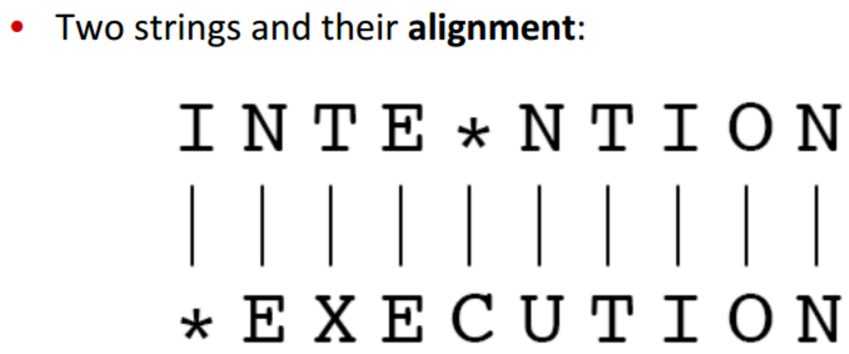
Chapter 4

Methodology

Here we will discuss what we have implemented so far to construct a standard scale.

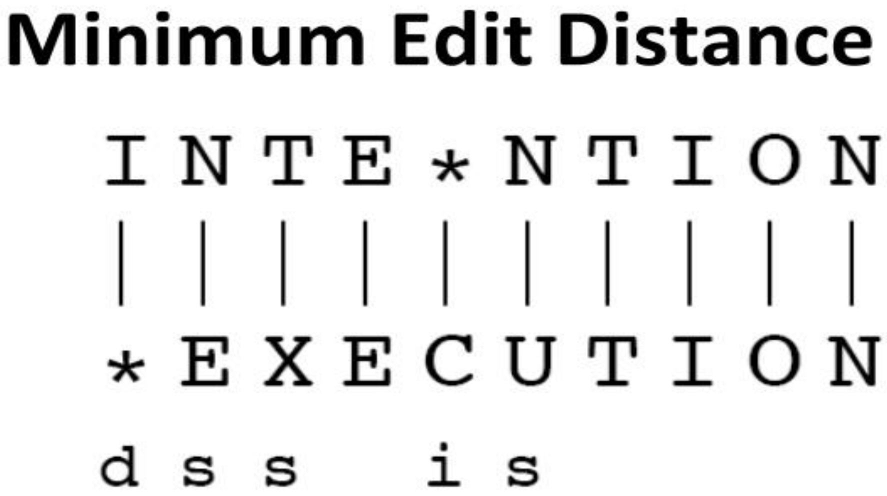
Firstly we have implement a well-known string matching algorithm called "Edit Distance". This tells us how it will match or mismatch between a read and a reference genome. This algorithm can be implemented in two ways. As we have to find the optimal result we have to implement it in such a way that ensures all possible solutions and get the best out of it. Basically this algorithm can be implemented two ways. One is bottom up approach and the other one is top down approach. Top down approach is a recursive method. On the other hand bottom up doesn't need to be recursive so it is more memory efficient as it doesn't take stacks to find the solutions. We have implemented it both way but the result was almost same as the scoring system ensured it would be a linear solution.

For example we can see below two figure that has aligned a read with a reference string. The upper one is reference and the other one is read.



- Two strings and their **alignment**:

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
```

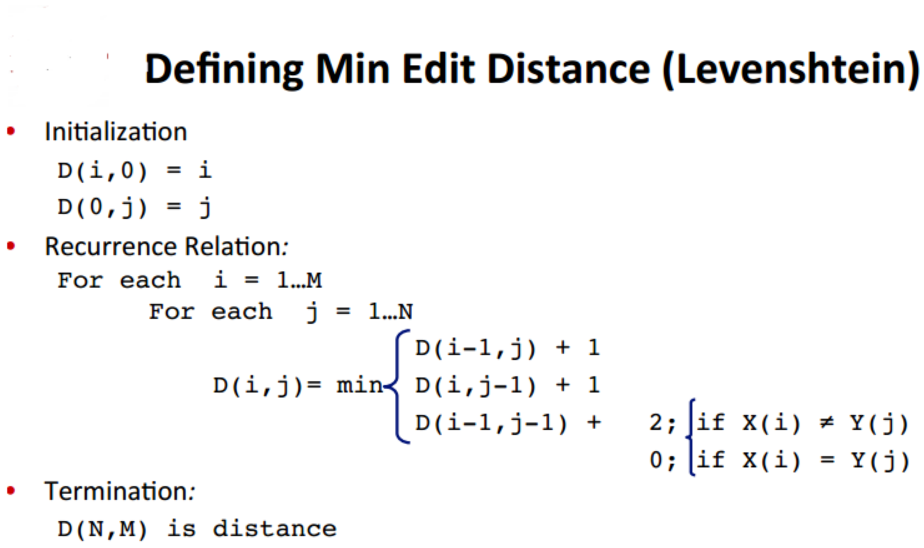


Minimum Edit Distance

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

Here d means deletion, i means insertion, s means mismatches and the other ones are matching.

Alignment is basically of two types with edit distance with respect to path finding. The below figure shows the sudocode of the algorithm.



Defining Min Edit Distance (Levenshtein)

- Initialization

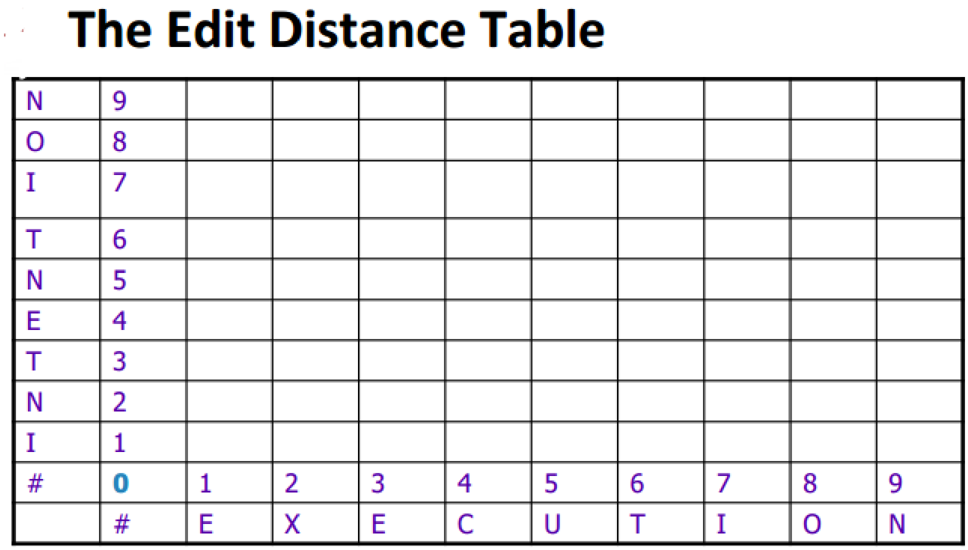
```
D(i,0) = i
D(0,j) = j
```

- Recurrence Relation:

```
For each  i = 1…M
      For each  j = 1…N
                      ⎧ D(i-1,j) + 1
          D(i,j)= min ⎨ D(i,j-1) + 1
                      ⎩ D(i-1,j-1) +   2; ⎧if X(i) ≠ Y(j)
                                       0; ⎩if X(i) = Y(j)
```

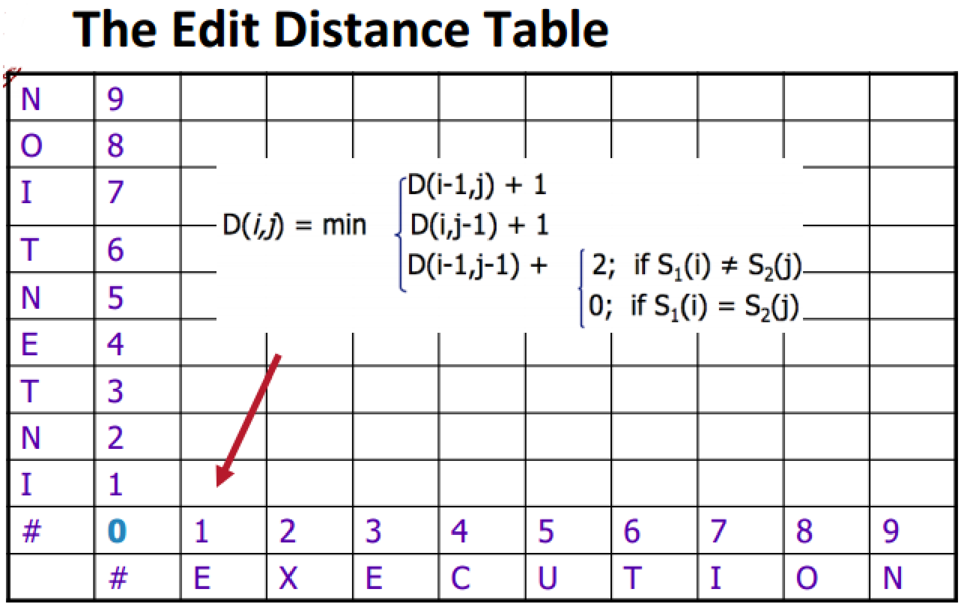
- Termination:

```
D(N,M) is distance
```



The Edit Distance Table

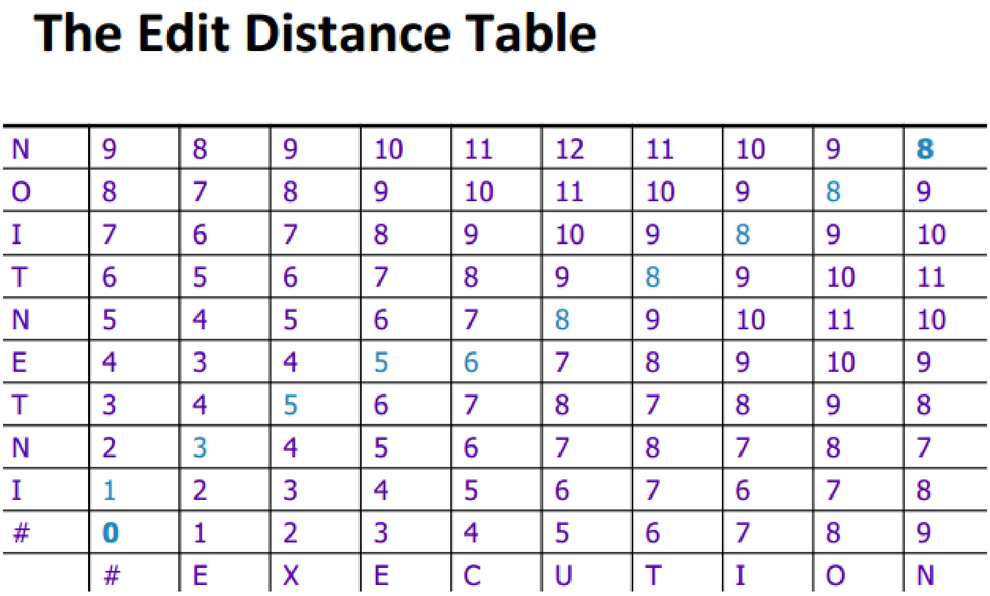
N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N



The Edit Distance Table

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

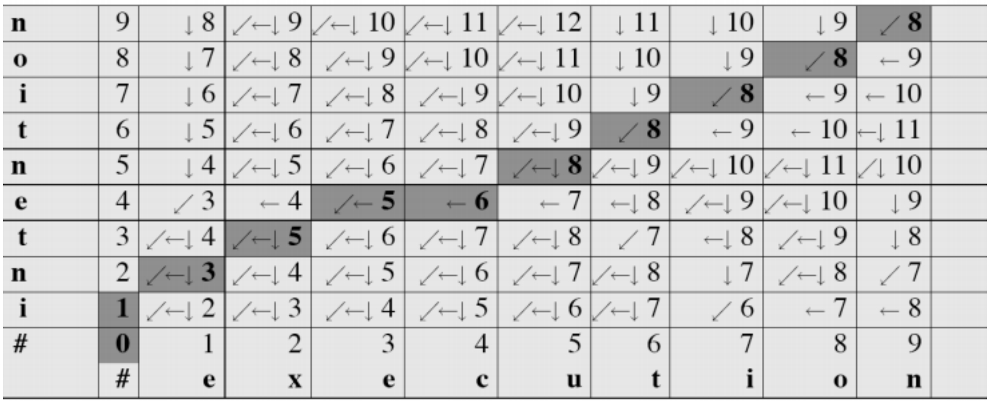
$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$



The Edit Distance Table

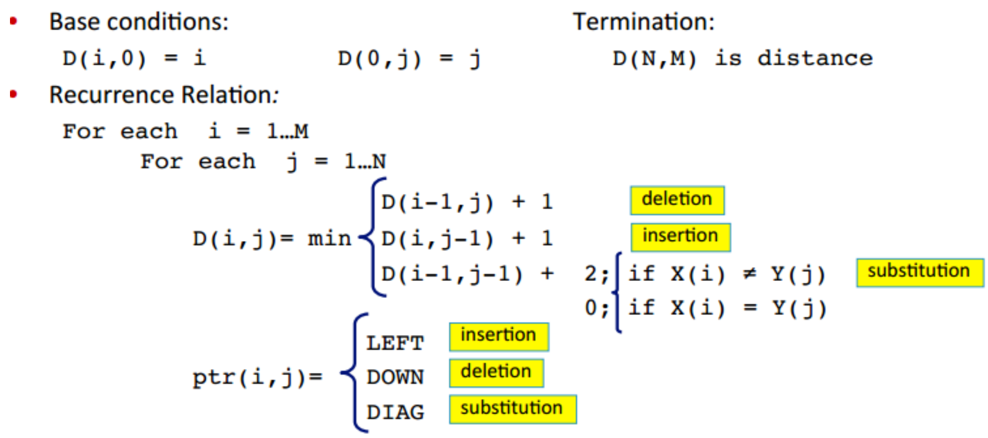
N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Here the blue ones are indication that we should follow this pathway to find the optimally best reference string that has matched.



n	9	↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙←↓ 12	↓ 11	↓ 10	↓ 9	↙ **8**
o	8	↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↓ 10	↓ 9	↙ **8**	← 9
i	7	↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	↙ **8**	← 9	← 10
t	6	↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙ **8**	← 9	← 10	←↓ 11
n	5	↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ **8**	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙↓ 10
e	4	↙ 3	← 4	↙← **5**	← **6**	← 7	←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9
t	3	↙←↓ 4	↙←↓ **5**	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙ 7	←↓ 8	↙←↓ 9	↓ 8
n	2	↙←↓ **3**	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↓ 7	↙←↓ 8	↙ 7
i	**1**	↙←↓ 2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙ 6	← 7	← 8
#	**0**	1	2	3	4	5	6	7	8	9
	#	**e**	**x**	**e**	**c**	**u**	**t**	**i**	**o**	**n**

Following is the path finding back-tracing algorithm.



- Base conditions:
  `D(i,0) = i`   `D(0,j) = j`

Termination:
  `D(N,M) is distance`

- Recurrence Relation*:*

```
For each  i = 1…M
      For each  j = 1…N
```

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 & \text{deletion} \\ D(i,j-1) + 1 & \text{insertion} \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} & \text{substitution} \end{cases}$$

$$ptr(i,j) = \begin{cases} \text{LEFT} & \text{insertion} \\ \text{DOWN} & \text{deletion} \\ \text{DIAG} & \text{substitution} \end{cases}$$

This type of path printing is called global alignment, but if we choose any particular index from row or column then it is called local alignment. We have found some good results from this alignment from our implementation that will be shown in result and discussion chapter.

Chapter 5

Result and Discussion

For data we have got a SAM file which includes thousands of reads. We took 1000 reads from there with maximum length of 1000.

After alignment the result has four things to consider to give it a score. They are number of matches, number of mismatches, number of insertions and number of deletions.

Percentage of identity is a well-known established scale to measure quality of a particular alignment performance.

Percentage of identity = Total number of matches / Summation of total matches, mismatches, insertions and deletions.

Following figures show what we have experimented and what result we have found so far.

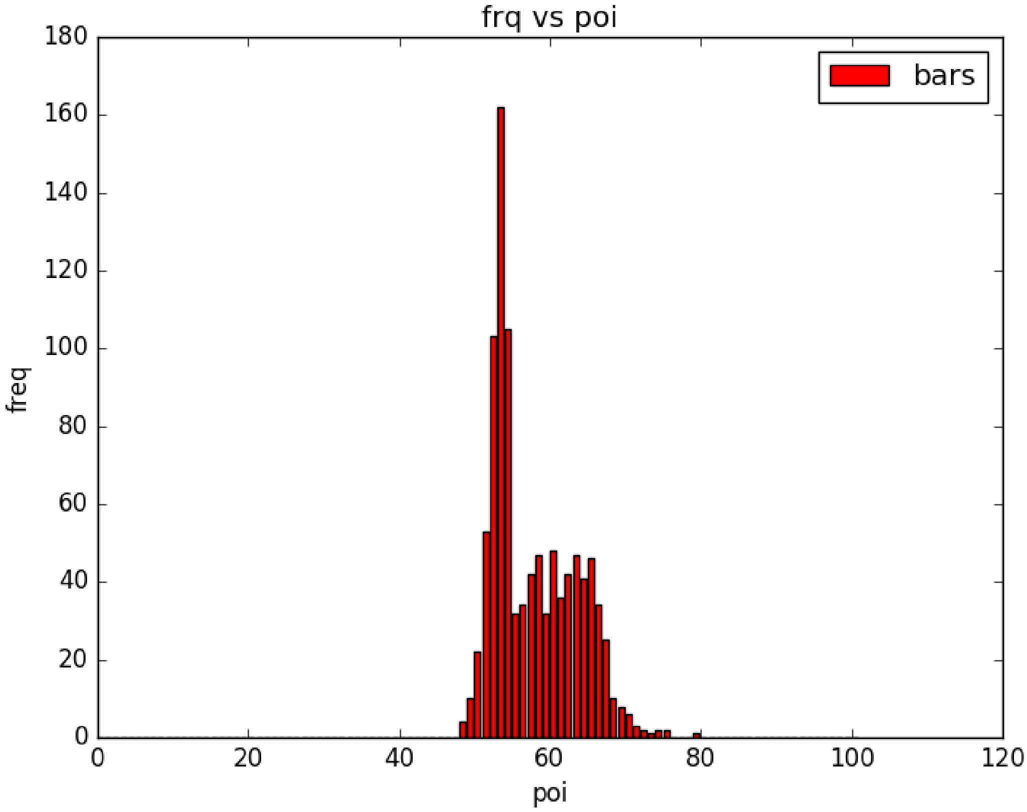


Fig 1: On this figure on x-axis percentage of identity is drawn and on y-axis their frequency is drawn.

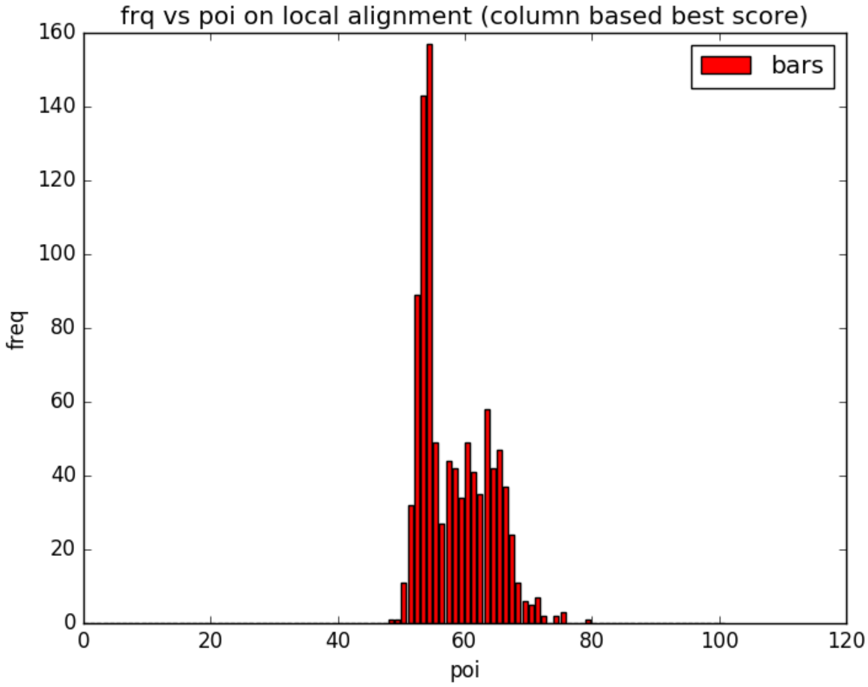


Fig 2: This figure shows local alignment result on column best score finding.

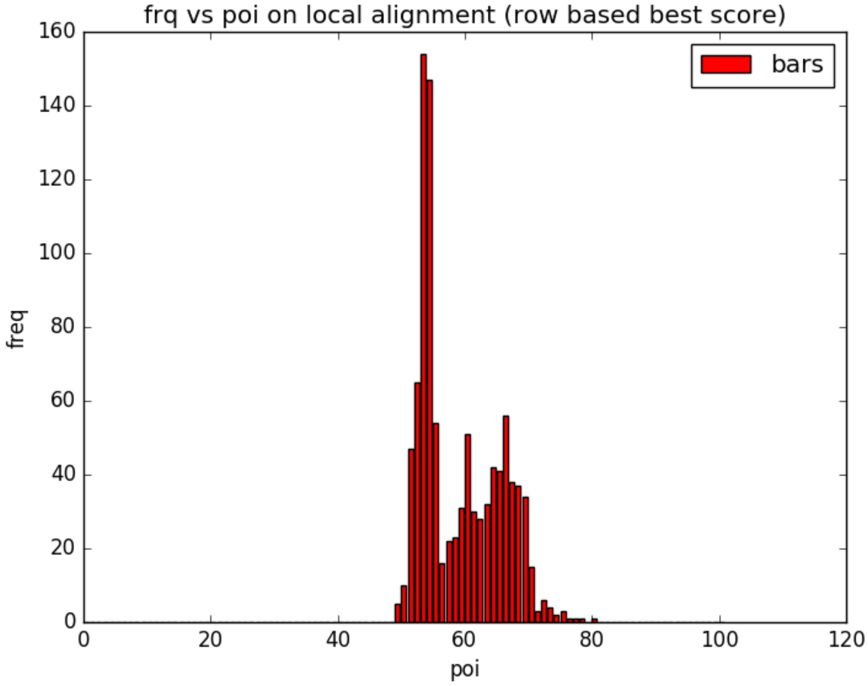


Fig 3: This figure shows local alignment result on row best score finding.

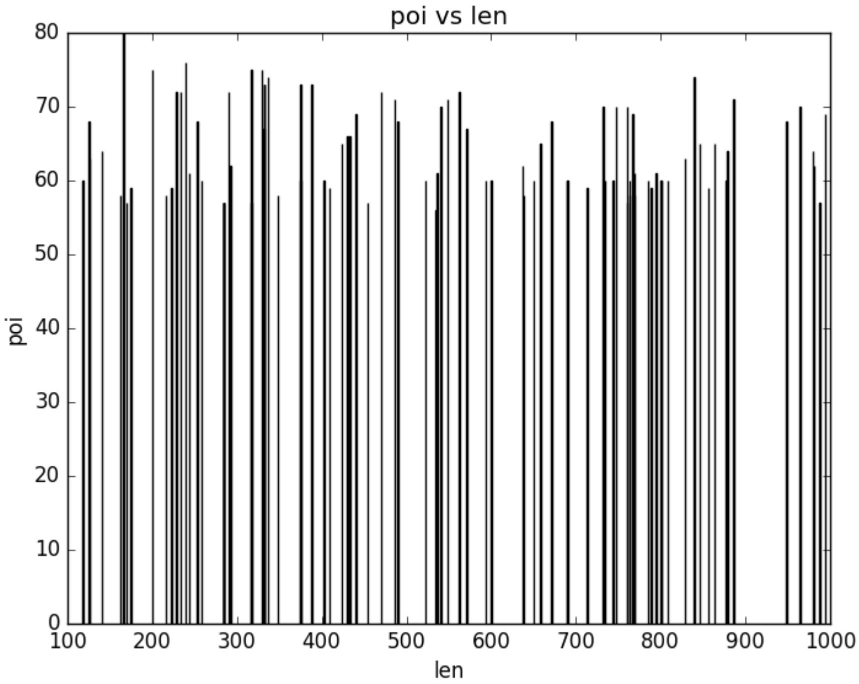


Fig 4: This figure shows what best poi has we got so far with respect to length of reads.

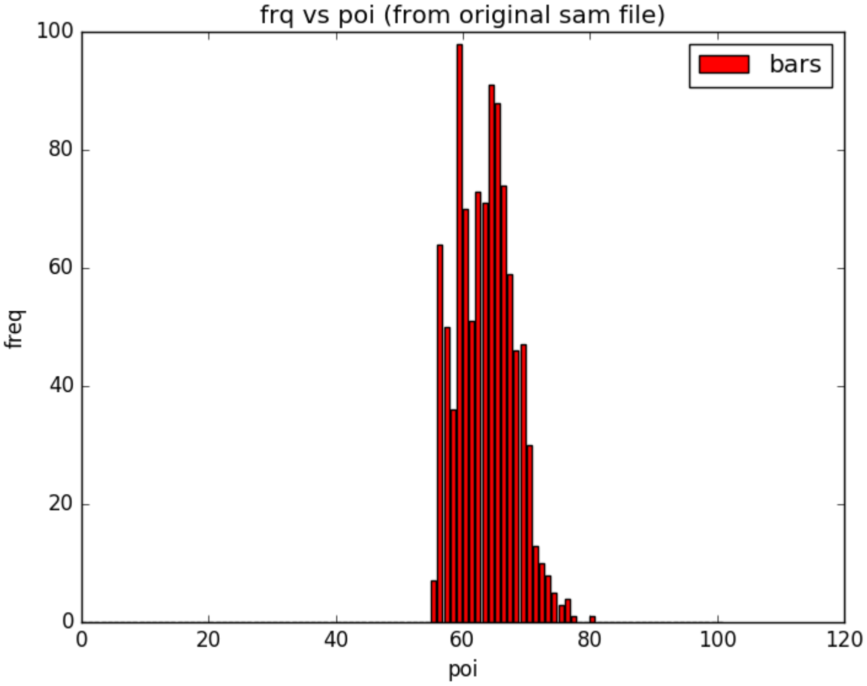


Fig 5: This result is got from original SAM file data.

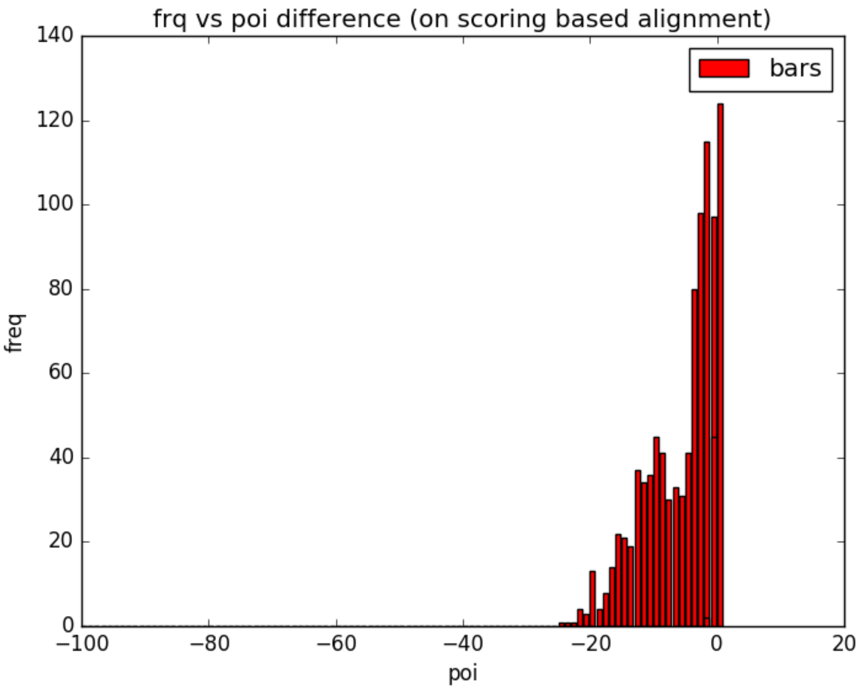


Fig 6: POI difference with global alignment

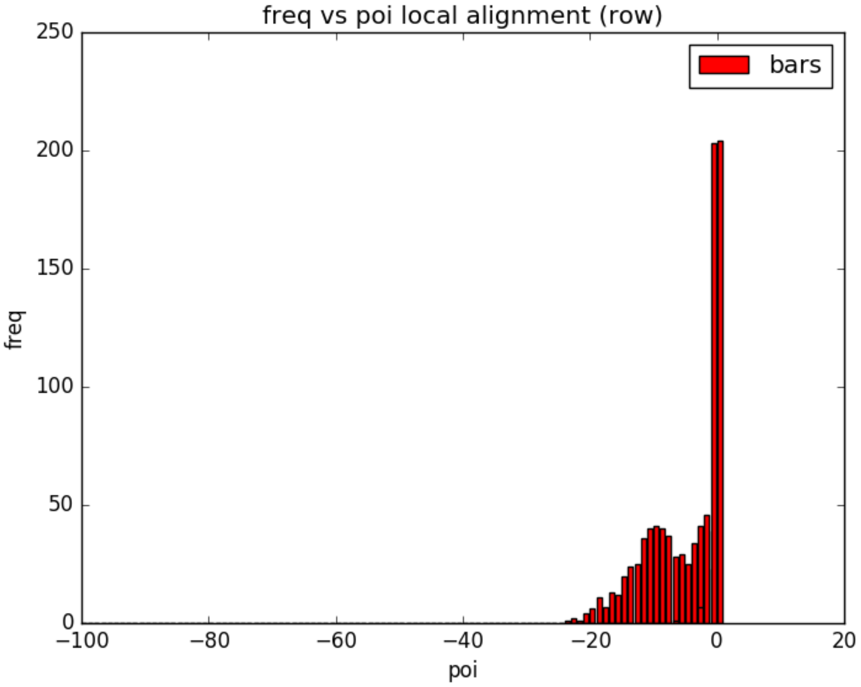


Fig 7: POI difference with local alignment (row)

Chapter 6

Future Work Approach

Our naive approach has not worked well for many reads. We will try for the betterment of that with more optimized window sized alignment. So that will be able to evaluate SAM file results with more error free alignments.

SAM file evaluating tools have many backwardness as Qualimap tool doesn't apply any algorithm to evaluate the reads. It just gives several statistics on a particular SAM file. We will go for the comparison of more than one SAM file on same reads and their alignments on different algorithms. Then we will be able to have a clear view how those reads differ with different positions (deletions, insertions, matches, mismatches) and how it can be improved with having more analysis on particular reads with reference genome.

Conclusion

Quality measurement of alignment is as important as to validate and compare a specific alignment algorithm with others. So that better approaches will get their recognition in a standard scale.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors PNAS 1977;74:5463-5467.

2. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. Science. 2003;300:286-290.

3. Davies K. (2010) 13 years ago, a beer summit in an English pub led to the birth of Solexa. BiolT World (www.bio-itworld.com/) 28 Sep 2010.

4. Illumina (2014) FliSeqX Ten preliminary system specification sheet. (www.illumina.com/documents/products/datasheets/datasheet-hiseq-xten.pdf)

5. Fallows J. (2013) When will genomics cure cancer? A conversation with Eric S. Lander.The Atlantic (www.theatlantic.com/) 22 Dec 2013.

6. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. Gen Biol. 2013;14:R51.

7. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53-59.

8. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. PLoS One. 2013;22;8(10):e77910.

9. Illumina (2014) Nextera DNA Library Preparation Kits data sheet. (www.illumina.com/documents/products/datasheets/datasheet˙nextera˙dna˙sample˙prep.pdf).

10. Illumina (2014) Nextera XT DNA Library Preparation Kit data sheet. (www.illumina.com/documents/products/datasheets/datasheet˙nextera˙xt˙dna˙sample˙prep.pdf).

11. Illumina (2013) TruSeq DNA PCR-Free Library Preparation Kit data sheet. (www.illumina.com/documents/products/datasheets/datasheet˙truseq˙dna˙pcr˙free˙sample˙prep.pdf).

12. Grad YH, Lipsitch M, Feldgarden M, et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. PNAS. 2012;109:3065-3070.

13. McEllistrem MC. Genetic diversity of the pneumococcal capsule: implications for molecular-based serotyping.
Future Microbiol. 2009;4:857-865.

14. Lo YMD, Chiu RWK. Next-generation sequencing of plasma/serum DNA: an emerging research and molecular diagnostic tool. Clin Chem. 2009;55:607-608.

15. Ram JL, Karim AS, Sendler ED, and Kato I. Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the lllumina sequencing platform. Sysf BiolReprodMed. 2011;57:117-118.
16. Wang Y, Kim S, Kim IM. Regulation of metastasis by microRNAs in ovarian cancer. Front Oncol. 2014;10:143.
17. Dior Up, Kogan L, Chill HH, Eizenberg N, Simon A. Emerging roles of microRNA in the embryo-endometrium cross talk. Semin ReprodMed. 2014;32:402-409.
18. Adams,D. et al. (2012) BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotechnol. 30, 224–226.
19. Anders,S. et al. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. 8, 1765–1786.
20. DeLuca,D.S. et al. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics, 28, 1530–1532.
21. Garcıja-Alcalde,F. et al. (2012) Qualimap: evaluating next generation sequencing alignment data. Bioinformatics, 28, 2678–2679.
22. Koeppel,M. et al. (2015) Helicobacter pylori infection causes characteristic DNA damage patterns in human cells. Cell Rep., 11, 1703–1713.
23. Patel,R.K. and Mukesh,J. (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One, 7, e30619.
24. Ramirez-Gonzalez,R.H. et al. (2013) StatsDB: platform-agnostic storage and understanding of next generation sequencing run metrics. F1000Res. 2, 248.
25. Rosenbloom,K.R. et al. (2013) ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res., 41, D56–D63.