Lecture 1: Introduction

Monday 28 March 2016

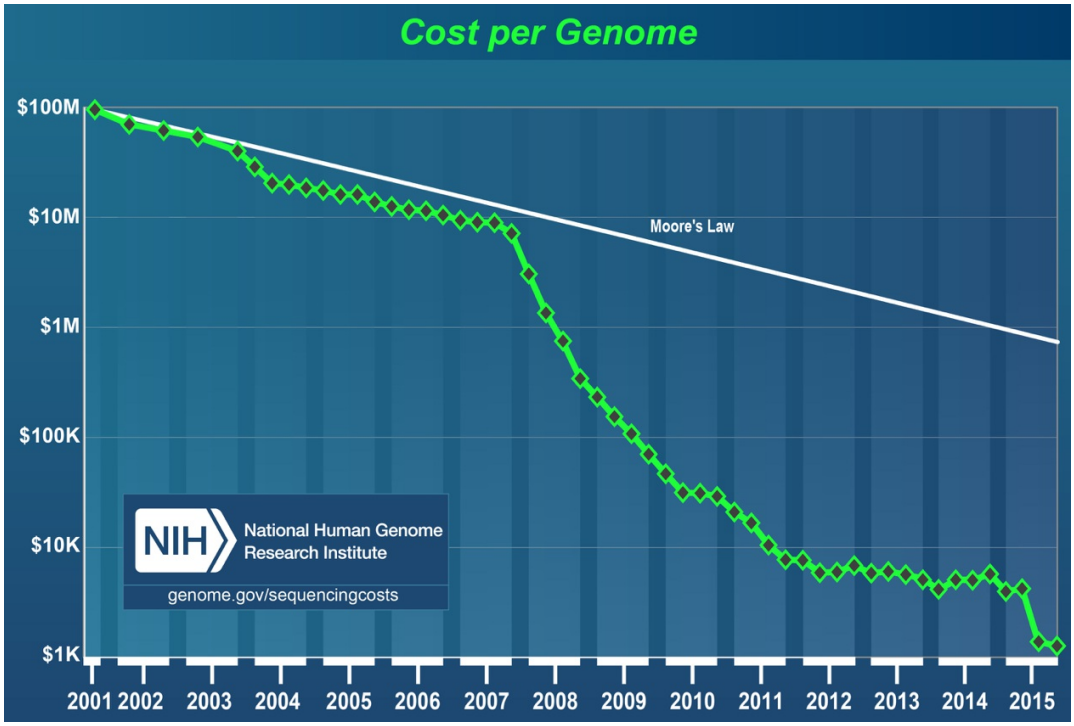
Scribed by Anja Brandon and revised by the course staff

Topics

What is high-throughput sequencing?

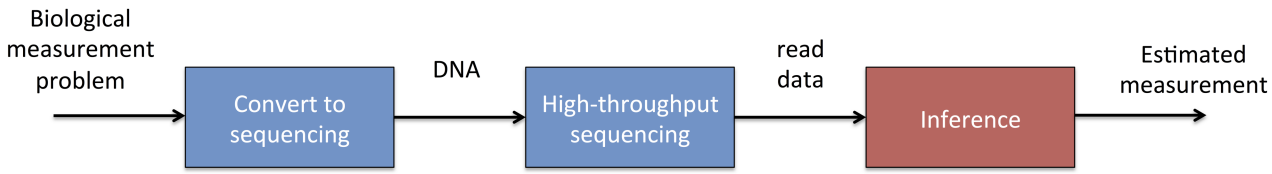
A main object of interest in this course is thegenome of a organism, which is made of*deoxyribonucleic acid* (DNA). All computational methods we discuss will be related to deducing the sequence of the genome or some property closely related to the genome. High-throughput sequencing refers to modern technologies used to identify the sequence of a segment (or "strand") of DNA. Only 15 years ago, sequencing technologies were around 6 orders of magnitude slower than they are today in terms of both cost and throughput.

The first major sequencing project was the Human Genome Project. A big consortium began collaborative efforts in 1990 to sequence the entire human genome. The project was nominally completed in 2003, costing $2.7 billion and 13 years of work by labs around the world. In 2015, the cost of sequencing a genome was approximately $1000. This is testament to how far the technology has evolved. As shown below, the cost of DNA sequencing has been falling at a rate faster than Moore's law over the last 15 years.



Cost of DNA sequencing over the years.
DNA is a very important biomolecule, but it's only one of many important biomolecules. Other important biological molecules include ribonucleic acids (RNA) and proteins. Some innovative bio-chemistry has allowed the use of DNA sequencing technology for measuring properties of various other biological molecules (and there are even proposals on how to use DNA sequencing for detecting dark matter).

High-throughput sequencing can be thought of as a microscope that can be used to measure a variety of quantities. The basic paradigm (shown below) is to reduce the estimation problem of interest to a DNA sequencing problem, which can be handled using high-throughput sequencing. This is similar in principle to the reduction used to solve many mathematical problems or to show NP-hardness of various problems.



The *-seq paradigm: Convert the problem of interest to a DNA sequencing problem.
For the biochemist, the challenge is in determining how to convert the problem of interest to a problem which can be tackled using high-throughput sequencing. This is similar to how biologists design experiments such that the results can be observed under a microscope. For the computational biologist, the challenge is in performing the relevant type of inference on the data observed using high-throughput sequencing. Some important sequencing assays are:

- RNA-Seq: RNA is an important intermediate product for producing protein from DNA. While every cell in an organism has the same DNA, an individual cell's RNA content may be very different. RNA in cells can also vary depending on temporal and environmental factors. RNA-Seq is an assay that "measures" RNA, and this was the first assay in which high-throughput sequencing was used to measure a molecule other than DNA. The assay was developed in 2008 by Mortazavi *et al*.
- ChIP-Seq: Different cells express different RNA because of *epigenetic* factors or molecules that influence how the genome is packed in the cell. DNA in cells are bound to proteins called histones, and for different cells, different parts of the genome are bound to histones. DNA wrapped around histones are harder to access and are not converted to RNA. ChIP-Seq is an assay which was developed measure the regions of the genome that are bound to histones. This assay was developed in 2007 by Johnson *et al*.. Another recent assay called ATAC-seq measures regions of the genome that are *not* bound to histones.
- Hi-C-Seq: This assay measures the 3D structure of molecules and was developed by Belton *et al*. in 2012.

One of the most interesting and important problems in genomics is predicting *phenotype* (physical characteristics such as a person's height or a person's favorite color) from *genotype* (DNA sequence). In medicine, understanding the relationship between phenotype and genotype can allow researchers to predict a patient's susceptibility to certain diseases by sequencing the patient's genome. A big success-story here is the discovery that presence of a particular mutation in the gene BRCA1 increases the risk of breast cancer to around 45%.

Another important application of high-throughput sequencing is cancer. Cancer is a "disease of the genome." It is caused by rearrangements of the genome, which are sometimes very large. By sequencing cancer cells, one gets information about the nature of the cancer-causing mutation and can tailor treatment.

Non-invasive pre-natal testing for genetic birth defects is another powerful application of high-throughput sequencing. Traces of fetal DNA can be found in the blood of the mother. The main idea here is to sequence

the maternal blood and infer fetal genetic birth defects from the sequence. High-throughput sequencing has been used successfully for detecting Down syndrome.

Sequencing methods

Science progresses by the invention of measuring methods. High-throughput Sequencing is one such measurement tool; however, high-throughput Sequencing is different from many measurement tools because it has a significant computational component. High-throughput sequencing (also called *shotgun sequencing*) takes the DNA sequence as input, breaks it into smaller fragments or *reads*, and returns a noisy version of these smaller fragments. We note that the length of reads range from 50-50000 while the human genome is of length 3 billion. Fortunately, these small noisy subsequences also contain information about the genome. While a single read contains very little information about the entire sequencing, a typical experiment generates a few hundred million reads (and hence is called "high-throughput"). Extraction of the information contained within reads requires clever computational processing, and this is the flavor of problems we will discuss in this class. We also note that the sequencing process can be very noisy. Each of the reads can be potentially different from the original subsequence of the DNA the read came from.

The sequencing revolution arose due to the rapid evolution of sequencing technologies. Sequencing began with Fred Sanger, who first came up with Sanger sequencing technology. This was a relatively low-throughput technology and was the dominant technology until the late 1990s. Second generation sequencing is most heavily represented by Illumina and is currently the dominant technology. Recent developments in Illumina sequencing have allowed scientists perform single-cell sequencing or the sequencing of individual cells. Companies like PacBio and Oxford Nanopore have led recent developments in third and fourth generation sequencing technologies.

High-throughput sequencing is a fast changing area with new technologies emerging constantly. All these technologies give us reads, but each uses different chemical processes to generate the reads. There are two main properties of reads that are important from a computational perspective:

1. *Read lengths*: The longer the reads are the more information they contain. Ideally, a read is simply the entire genome. Unfortunately, a read of such length is not achievable by chemistry today or in the foreseeable future. Illumina reads are around 100bp-200bp long, and PacBio reads are over 10000 bp long. While PacBio reads are longer than Illumina reads, they are still much shorter than genome lengths.
2. *Error rates and types of errors*: Illumina has relatively low error rates of 1-2%, and errors here are mostly substitution errors (*i.e.* a base being replaced by some other base). PacBio reads have higher error rates of 10-15%, and errors here are insertions and deletions.

The figure below shows some characteristics of different sequencing technologies.

Sequencer	Sanger 3730xl	454 GS	Ion Torrent	SOLiDv4	Illumina HiSeq 2000	Pac Bio
Mechanism	Dideoxy chain termination	Pyrosequencing	Detection of hydrogen ion	Ligation and two-base coding	Reversible Nucleotides	Single molecule real time
Read length	400-900 bp	700 bp	~400 bp	50 + 50 bp	100 bp PE	>10000 bp
Error Rate	0.001%	0.1%	2%	0.1%	2%	10-15%
Output data (per run)	100 KB	1 GB	100 GB	100 GB	1 TB	10 GB
Approx cost per GB		10,000	1000	100	10	1000

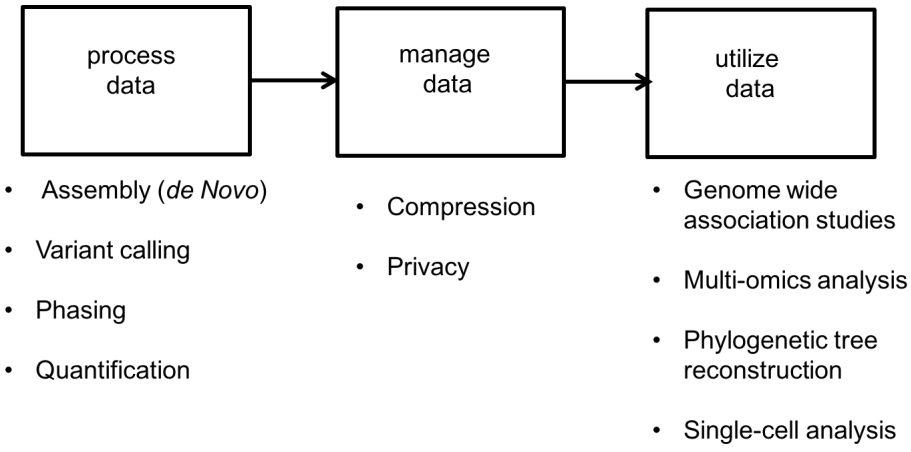
Characteristics of different sequencing technologies.

Data science of high-throughput sequencing

The success of high-throughput sequencing is mainly due to the creative use of read data to solve various problems. For this course, data science problems can be categorized into one of three types:

1. *Data processing*:
2. Assembly or *de novo* assembly: Recovering the DNA or RNA from short noisy reads.
3. Variant calling: Individuals of the same species have very similar genomes. For example, any two humans share 99.8% of their genetic material. Because a reference human genome is available, scientists are often interested in the differences of an individual's genome from this reference genome. Variant calling is the problem of inferring these differences.
4. Phasing: The chromosomes in humans (and other higher animals) come in pairs but are crushed and sequenced together. Often scientists want to separate the sequence on the two chromosomes. This is called the phasing problem.
5. Quantification: RNA is an important biological molecule in cells, as discussed above. There are potentially 10000s types of RNA molecules observed in an individual cell. Quantification is a counting problem; scientists are interested in estimating how many copies of each type of RNA are in a cell or population of cells.
6. *Data management*: With large databases, natural problems that arise include
7. Privacy
8. Compression
9. *Data utilization*: Here we use the data to make useful inferences. These problems include
10. Single-cell analysis: Properties like diversity in cell populations are inferred from single-cell datasets.
11. Genome Wide Association Studies (GWAS): This problem looks at the association between genomes and various characteristics of individuals.
12. Multi-omics data analysis: Methods for combining DNA, RNA, and protein measurements to make predictions.

These different problems are illustrated below:



Data science of High-throughput sequencing.

Tools used

When working with high-throughput sequencing data, we first attempt to model the data. This usually involves many assumptions which are not true in practice. While inaccurate, these models are used to come up with initial interesting algorithms. As real data often does not satisfy these assumptions, some additional effort is required to get working algorithms even when the modeling is reasonable. Some tools we will use in this course are:

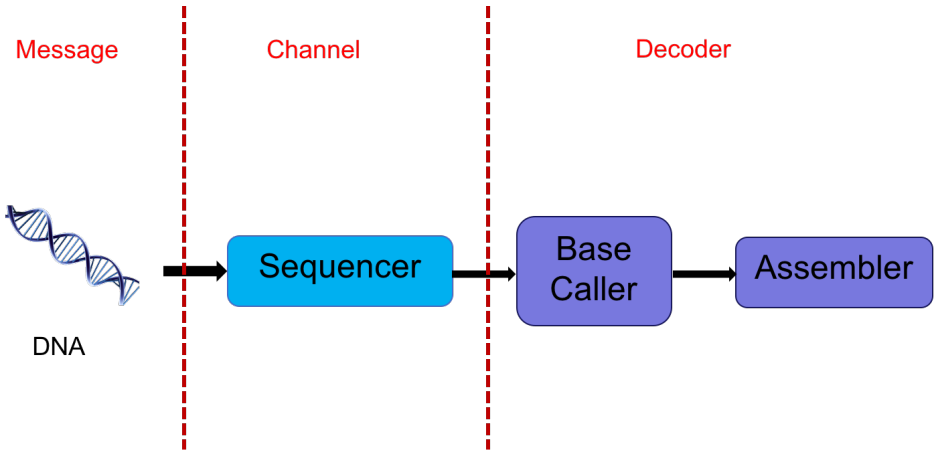
- Combinatorial algorithms: Problems like genome assembly involve working on combinatorial objects like graphs, and combinatorial algorithms naturally follow.
- Statistical Signal Processing: Because the data is noisy, we need signal processing techniques for dealing with the noise.
- Information Theory: When performing inference, this gives a sense of how much data is necessary to achieve "good" estimates, allowing us to design optimal algorithms to achieve such estimates.
- Machine Learning

Two Example Applications

In this section, we discuss two representative problems that will be covered in this course.

DNA-assembly

The DNA sequencer outputs an analog signal (e.g. light intensities or electric signals). We want to process this signal to get the sequence. In essence, one could think of the DNA as a message, the sequencer as a communication channel, and the base caller and assembler as the decoder. This abstraction is shown below:



DNA assembly as a message decoding problem.

This abstraction gives us multiple avenues of exploring potential problems. The extraction of digital information (discrete bases) from analog signals is a statistical signal processing problem. This involves various stochastic models with many parameters which need to be estimated. Furthermore, one often has to account for signals from adjacent bases interfering with one another. Dealing with intersymbol interference is also a signal processing problem.

We can also consider the problem of assembling the genome from the reads obtained after processing the analog signals. We want to first obtain an estimate of the number of reads necessary to be able to assemble with reasonable accuracy. Using tools from information theory, we can identify bottlenecks and design principles to deal with them, allowing us to design efficient algorithms to overcome these bottlenecks. By efficient here we mean linear in the number of reads. In general, the size of data makes any super-linear algorithm unfeasible in most cases; however, there are cases where smart algorithm design and low level optimized software allows one to use algorithms that are quadratic in the number of reads.

Single-cell RNA quantification and analysis

As discussed above, RNA is another important biological molecule. There exists around 10000 RNA sequences (or *transcripts*) floating in each cell, each of which are 1000-10000 bp long. Biologists are interested in the problem of *quantifying* or estimating the number of each RNA transcript in a cell.

Biologists and chemists have figured out ways to convert RNA back into DNA (mainly using an enzyme *reverse transcriptase*), and then sequence the DNA to get reads using shotgun sequencing. The computational problem is trying to estimate the number of transcripts of each type from these reads.

One often uses the reference of known transcripts observed in an organism: the *transcriptome*. Despite there existing a reference, many transcripts have common subsequences and therefore one can not always be sure of where a read originates from. A good algorithm for solving this problem is expectation-maximization (EM).

In a bulk experiment, a biologist crushes the 100s of millions of cells in a tissue together. After shotgun sequencing, the biologist obtains a mixture of the RNA of all cells in the tissue. The transcript counts (or abundances) obtained are therefore an estimate of the sum over all cells. Recent technologies have allowed biologists to sequence biological samples such as tissue at the single-cell resolution. Single-cell technologies allow researchers to observe the diversity of cells within a cell population. Relevant problems here include clustering relevant features to uncover cell types.

Slides
