Cost of Sequencing Draft

The real cost of sequencing: higher than you think!

The real cost of sequencing: scaling behaviour

The real cost of sequencing: it's really now computer calc scale

The real cost of sequencing: processing, storage & data transfer

The real cost of sequencing: will it continue to scale?

Scaling of sequencing costs: a fixed & var cost perspective

Scaling of sequencing costs: Are we still in a moore's law regime

The real cost of sequencing: An amortized analysis

** Introduction

[[MG2PM: work on the intro and increase in seq.]]

- Number of species sequenced
- Bases in major journals over time
- Changes in # of sequencers and locations over time (from omicsmaps.com)

* NIH reporter graph

- Cost of sequencing on Genohub

*** TCGA_Data_Universe_2015-07-16 (1).pptx

Low cost sequencing and large databases changing biological research:

The continued decrease in the cost of sequencing and corresponding increase in the size of sequence databases has changed both the biological research landscape and the common modes of research. The establishment and growth of sequencing core facilities has helped increase the accessibility of sequencing technology by mitigating the upfront fixed cost of purchasing machines. [Add detail about omicsmaps.com historical data figure] The per base cost of sequencing has also been falling, allowing investigators to generate more sequence data. Furthermore, the growth of sequence databases has reduced the cost of obtaining useful sequence information for analysis. Data downloadable from databases is ostensibly free. However, costs arise in the need for computational storage and analysis resources as well as the training necessary to handle and interpret the data.

These developments in DNA sequencing are creating a diverse research ecosystem in which consortia are producing and disseminating large standardized datasets and individual investigators are able to contextualize their research by querying the wealth of existing sequence databases. As this research landscape changes we need to reevaluate the way in which research is performed and how the associated costs are calculated.

Illustrations of the dramatic increase in rate and amount of sequencing:
The dramatic drop in sequencing costs has been accompanied by an explosion of sequence data generation. This sequence generation is occurring in part through large consortia generating enormous datasets. Large consortia have taken advantage of sequencing trends to generate population scale genomic data (1000 Genomes) or extensive characterization of cancer genomes by The Cancer Genome Atlas (TCGA).
Meanwhile, an ever expanding set of seq related assays has taken advantage of inexpensive sequencing to serve as a readout in assays investigating a range of biological processes Additionally, ever larger amounts of sequence data are being generated from experimental protocols that utilize DNA sequence

data generated from high throughput platforms as a readout. This data is more lab and experiment specific compared to the standardized creation of general use datasets addressing larger questions such as TCGA, 1000 Genomes, and ENCODE and can serve as a valuable addition to such larger standardized approaches. Individual lab generated and highly question specific datasets can provide valuable additional annotations and context to the larger general datasets.

Online vs. offline science:

In the past, hypothesis driven research provided a clear beginning and endpoint for a given experiment. Data generated during an experiment was designed to test a specific question. The increase in the rate of DNA sequencing and size of sequence databases has spurred a rise in hypothesis generating research. Many sequencing experiments are aimed at providing large standardized datasets for general use. Even in cases where sequencing is used to address a specific question the resultant data may later be employed for a purpose significantly different from that of the original investigator. Under these conditions, the computational storage and analysis component of a sequencing experiment will come to represent an increasing proportion of the costs associated with high throughput sequencing experiments relative to the initial cost of generating sequence data.

The difference between these two paradigms of scientific research is comparable to the distinction in computer science made between offline and online algorithms. Offline algorithms take in all inputs at once and then begin processing. Meanwhile, online algorithms process their input as it arrives without knowledge of all of the subsequent inputs.

The process of hypothesis-driven research is similar to that of the offline algorithm. Data is collected at once and processed afterwards resulting in a conclusion that represents the end of that specific data's use.

As decreased cost and the generation of new seq-based experimental protocols increase the deposition of sequences in public databases it is becoming increasingly difficult to follow the offline algorithm model. The scope of large consortia generated datasets such as TCGA makes waiting for all the data to be generated before performing the analysis prohibitive. The use of sequence from existing databases for secondary analysis creates a situation in which it is impossible to predict the full extent of a dataset's usage. Furthermore, if DNA sequencing and synthesis are to be combined in an iterative cycle to better link genotypic variation to phenotype, then constant analysis of new data in light of the existing sequencing knowledge base is required.

These new modes of biological research necessitate conceptualizing the experimental process as an online algorithm in which both data generation and analysis are viewed as events in a larger research process of unknown duration.

Amortized analysis:

Amortized analysis is used by computer scientists to evaluate the worst-case scenario over a sequence of operations. This type of analysis is commonly applied to online algorithms. The analysis method relies on the idea that some computationally expensive processes may pay off over a series of operations because they enable speed improvements in subsequent operations. This approach contrasts with less holistic approaches in which the worst-case scenario of each operation in a process is determined independently. These two analysis methods may suggest different series of events as optimal.

It is becoming increasingly important to account for the entire lifespan of a sequence dataset as repeated use of sequence data in different analyses extend the life of a given dataset. As sequencing prices drop and the scale of generated sequence data increases the relative importance of initial data generation decreases. This situation is conducive to the holistic approach taken by amortized analysis.

Considerations in the sequencing pipeline:

[[PM: These next paragraphs might be useful to link between the different sections of the perspective. More work needs to be done applying an amortized analysis to each step in the sequencing pipeling. The amortized analysis might be a good analogy to follow through the perspective as it integrates the fixed and variable cost idea with a more CS perspective.]]

The first question an investigator is faced with is whether to purchase a sequencer or utilize a sequencing core facility or company. Often the costs involved in purchase, maintenance, and operation of a sequencer as well as the requisite expertise required of its operators makes it unfavorable for individual investigators to purchase a sequencer. However, this could change with the advent of low cost sequencer alternatives such as Oxford Nanopore's MinION.

[[The following links to cost on Genohub as well as elements in the intro and conclusion]]

The investigator then must decide on the protocol for sample preparation and which sequencing platform best conforms to their needs. Here a focus on the constraints of the scientific question being asked and the experimental setup available must be taken into account while optimizing for collection of the most informative sequence dataset.

[[The following links to data size, network bandwidth, and computational processing power scaling]]

Next the investigator must decide in where and in what format to store the sequence data. At this step it is important to consider both the volume of sequence generated as well as the frequency and mode of data access required for downstream analyses. Options at this step range from downloading the data on a local machine where all analyses will also be performed to uploading the data to the cloud and similarly performing analyses in the cloud.

[[The following links to alignment algorithms]]
After storing the data, the investigator must decide what algorithms to use in order to initially process the data. These alignment algorithms can be viewed as a microcosm for the types of questions now being asked about sequencing more generally. The initial alignment algorithms developed in the 1970s would be hopelessly slow if confronted with the scale of modern datasets. Over time algorithmic innovations have enabled alignment algorithms to keep up with the dramatic increase in size and scale. These newer algorithms utilize practices encouraged by amortized analysis. They devote a significant amount of their runtime to a computationally expensive indexing operation that later provides significant improvements in the performance of the alignment operation.
[[The following links to the conclusion]]
Once the data has been initially processed, downstream analyses and interpretation must be performed to obtain scientific insights and knowledge. The search for biological meaning in these datasets will be helped by two trends. The first trend is an increase in the size and diversity of sequence-based datasets allowing for ever greater statistical power and new comparisons between datasets. This will help with large scale analyses of both labelled and unlabeled biological datasets. The second trend is a decrease in the cost of DNA synthesis combined with improvements in genome engineering at single nucleotide resolution. This will enable investigators to more easily experimentally follow up on findings derived from sequence analysis.

** Computers , backdrop of the computer industry & Moore's

[[MG2DW : edit the computer background section]]

Moore law accurately predicted the development of integrated circuits; e.g., the number of transistors integrated in each square inch almost double every year \cite{}. The semiconductor industry has also used the Moore law to guide its R&D progress. After Moore's law, various predictive laws were also proposed for the high-tech development (http://sourcetech411.com/2012/12/engineering-laws-moores-rocks-butters-and-others/). For instance from an economic point of view, Rock's law was proposed to predict the cost of a semiconductor chip fabrication plant doubles around every four years. Because the success of the predictive laws in high tech areas in last half century, ones always tend to use those laws to forecast the new technologies including the sequencing technology. For example, the sequencing cost indeed roughly followed the Moore's law before 2008 \cite{NIH cost-seq figure}. The sequencing technology, however, does not simply copy the trajectory of computer industry. It even runs faster than expected. In recent five years, the cost of sequencing a personal genomics dramatically dropped to XXX in 2014 from XXXX in 2008.

- S-shaped curves contribution to scaling behavior

Comparison of sequencing technology's trajectory to the growth of the computer industry,

which has experienced a similar if less dramatic scaling in its capabilities, can yield insights into the future of sequencing. The exponential scaling of the number of transistors in a microprocessor reshaped both the computer industry and a host of other industries. This rate of technological improvement enabled increases in computer performance and decreases in cost. Higher performance machines allowed computers to address ever more challenging problems while decreases in cost drove their widespread adoption. [[STL: distributed computing cuts cost. a single beefy node is much expensive than 100 mediocre nodes]] Additionally, the development of intuitive interfaces and research on human-computer interaction helped harness these technological improvements.[[Moore's is baked into the computer industry.... will it become baked to illumina? Cern thing - how has moore's law affected sci - & Moore's 2nd law]]

[[
paradigms for sci computing
mainframe era
pc era
web era [bioinformatics grew up in]
cloud era
how things scale differently !
]]

** Computational component of sequencing - what's happening in bioinformatics

[[MG2STL : edit the bioinformatics section from a more CS perspective]]

The decreasing cost of sequencing and increasing amount of sequence reads generated are placing greater demands on the computational resources and knowledge necessary to handle sequence data. Scalable storage, query and analysis technologies are necessary to handle the increasing amounts of genomic data being generated and stored. For example, distributed file system greatly increases the storage I/O bandwidth, making distributed computing and data management possible. Another example is NoSQL database provides excellent horizontal scalability, data structure flexibility and support for interactive queries.

Changing computing paradigms such as cloud computing are playing a role in managing the flood of sequencing data. HIPAA compliant cloud resources are being developed so that datasets can be stored and shared on remote servers. Analysis scripts are then uploaded to the cloud and the analysis is performed remotely. This greatly reduces the data transfer requirements since only the script and analysis results are transferred to and from the cloud. [[STL: also democratized research...no fixed/sunk cost]] [Include download statistics for datasets]

[[STL(Aug. 8): Should we mention commercial cloud server versus in-house hpc?]]

Traditional scientific computing paradigm is aggressively optimized on linear algebra. This is not of much benefit to nowadays bioinformatics research, which heavily uses statistical learning algorithms, user defined functions and semi-structured data. Moreover, today the parallel programming paradigm has evolved from fine-grained MPI/MP to robust, highly scalable frameworks such as MapReduce and Apache Spark. This situation calls for customized paradigms specialized for bioinformatical study. We have already seen some exciting work in this field (cite ADAM from AMP Berkeley)

** Alignment algorithms

[[MG2SKL : edit the alignment section]]
[[
the algorithmic effect -
dynamic programming
hashing - '92 - blast/fasta hash query
blat - hash the db

BWT - suffix tree

tradeoff from memory & speed - put in fixed v marginal cost
sacrificing of optimal alignment

http://bib.oxfordjournals.org/content/11/5/473.full.pdf

]]

- Alignment algorithms

Beyond structural improvements in data storage, alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing. The running time fulfill Moore's Law and decrease by half every 18 months. In the very early Sanger sequencing age, Smith-Waterman and Needleman-Wunsch algorithm use dynamic programming to find a local or global optimum alignment. Both the time and space complexity is O(mn) and make it impossible to map sequences to a large genome. FASTA, BLAST and BLAT, as the successor, introduce seed-and-extend paradigm and use exact-matched K-mer as the seed. However, the original FASTA simply combine the K-mer and Smith-Waterman algorithm, can not make sure best alignment are seeded. BLAST use heuristic statistical method to find high-scoring segment pair by hashing the query sequence and scan it against sequence database. In contrast, BLAT build index for the genome and scan the K-mer against query sequence, which can achieve 50 times faster than BLAST. For the NGS aligner, the challenge is to rapidly align millions of short sequences (reads) to a reference genome. MAQ adopt gapped-kmer to improve the sensitivity of seed-and extension schema. A category of data structure, such as suffix array,suffix tree and FM-index are used to find perfect match

instead of dynamic programming. In particular, Burrows- Wheeler Transform (BWT) can link suffix array/tree with FM-index to find exact match by enumerating all combinations of possible mismatches and gaps in the query sequence. The result is to sacrifice optimal alignment and error tolerance for extremely fast retrieval of perfect matches.

On the other hand, more and more alignment tools try to reduce the mapping cost by building a index data structure. Though it take a long time to build the FM-index for BWA and Bowtie, and to maintain a uncompressed suffix array for STAR, the index time cost are fixed and the marginal cost for reads alignment can be dramatically reduced.

Data storage and algorithmic improvements also need to be packaged in intuitive and easily navigable formats to spur the wider adoption of sequencing information amongst the biological research community. Illumina's BaseSpace takes a promising step in this direction by creating an environment that integrates everything from data transfer out of the sequencers to the app-like options for analysis programs.

** How have reduced costs changed biological research:

[[MG2DS: work on the end]]

- Bioinformatics jobs
- P/E ratio of illumina vs. other tech
- Use of datasets by secondary analysts
- ???? Cost of sequencing on Genohub

*** how will this shape sci. what does it mean for bioinformatics jobs, big v little

The dramatic drop in sequencing costs has changed the biological research landscape and spurred increased generation of sequencing data. However, to what extent are the increases in sequencing data due to large sequencing centers and established projects producing ever more sequencing data as compared to adoption of sequencing approaches by labs which did not previously use sequencing data?

As sequencing has become less expensive it has become easier for individual labs with smaller budgets to undertake sequencing projects. These developments have helped democratize and spread sequencing technologies and research, increasing the diversity and specialization of experiments. Using Illumina sequencing alone, nearly 150 different experimental strategies have been described (Ref of the poster "For all your Seq needs) yielding information about nucleic acid secondary structure, interactions with proteins, spatial information within a nucleus, and more. For example, natural products research has seen a resurgence as it has become clear that the biosynthetic potential of organisms greatly exceeds what has been observed to be produced; efforts to heterologously express and characterize "silent biosynthetic clusters" holds much promise for the discovery of new antibiotics and other medicines. Perhaps unsurprisingly, the market continues to expect growth from Illumina; their stock valuation outperforms other small-cap biotech, as well as similarly sized companies from other sectors (See P/E ratio of illumina vs. other tech).

In an era of squeezed budgets and fierce competition, job prospects for scientists with training in computational biology remain strong (Explosion of Bioinformatics Careers Science 2014). Universities have increased the number of hires in the areas of computer science, and specifically in bioinformatics (See Bioinformatics jobs).

Wrap up sentences after seeing the whole doc.

======vv Cut-out vv ======

As sequencing has become less expensive it has become easier for individual labs with smaller budgets to undertake sequencing projects. These developments have helped democratize and spread sequencing technologies and research. However, such trends also run the risk of fragmenting the genomics research community. If the sequence data generated by individual labs is not processed properly and made easily accessible and searchable then analysis of integrated datasets will become increasingly challenging. In addition to posing technical issues for data storage, the increasing volume of sequences being generated presents a challenge to integrate newly generated information with the existing knowledge base.

It is critically important that as the amount of sequencing data continues to increase it is not simply stored but done so in a manner that is easily and intuitively accessible to the larger research community. In the case of consortia, there are often required to ensure that their data is uniformly processed and easily accessible to the public. [[STL: probably we could talk a little about distributed database systems and how it realizes interactive querying in large scale]]

A recurring theme in the topic of high throughput sequencing is that of fixed and variable costs. The initial purchase of sequencing machines is a large initial fixed cost. However, this cost is often largely shouldered by sequencing core facilities and not directly by individual investigators. The fixed cost is amortized in accounting, and affects pricing. Nevertheless, as newer sequencing machines are able to produce more reads, the average total cost of sequencing decreases. Moreover, if the number of sequencing facilities increases, creating greater competition, economic theory predicts that the price of sequencing should be driven down and approach marginal cost. In an environment of perfect competition, the cost of sequencing should be equal to the marginal cost, and the fixed cost of purchasing a sequencing machine should not enter into the pricing function; rather, it should impact only the decision of whether or not to operate. If we think about the use of previously generated sequencing information there are almost no fixed costs in obtaining sequence information. This condition would suggest a significant increase in market (sequence-based research) entry. What is keeping researchers out of this area? The variable costs of computational resources and training.

====

Possible Figures:

- S-shaped curves contribution to scaling behavior
- Alignment algorithms
- Cost of sequencing on Genohub
- Bioinformatics jobs
- Number of species sequenced
- P/E ratio of illumina vs. other tech
- Bases in major journals over time
- Changes in # of sequencers and locations over time (from omicsmaps.com)
- Use of datasets by secondary analysts (Can we split this into reuse of consortia generated data vs data generated from individual labs)