The real cost of sequencing: higher than you think!
The real cost of sequencing: processing, storage & data transfer

Introduction:

~~+H~~ MORE

The contemporaneous development of Sanger sequencing and the personal computer started a digital revolution in the biosciences. Prior to the introduction of the PC, the use of computers in biological research was sparse. This was due in part to the expense and inaccessibility of mainframe computers. Additionally, some historians of science have argued that the lack of computers in biology was partially due to the incompatibility of computational approaches and biological research. The data generated by biological experiments was often not in a form that benefited from computational processing power. However, this changed with the advent of Sanger sequencing and generation of ever greater amounts of sequence data. Large amounts of sequence data are easily stored in computational databases and conceptualized in a computational framework. As computational and biological have developed together they have spurred and reacted to innovations in each other.

The computing technologies used in the analysis of sequence data have helped shape how researchers approach the analysis and the structure of biological research more generally. The PC era in which Sanger sequencing developed left its imprint on how sequence data is analyzed. In the 1980's sequence databases were developed and filled with ever larger amounts of sequence. However, most of the data relevant to an investigator could be downloaded and processed on local client. The rise of the internet enabled new bioinformatics approaches in which analysis programs could be hosted on websites and data could then be uploaded onto these sites for analysis. The internet further encouraged sharing of sequence data. The completion of the human genome project coincided with the rise of the internet. These conditions created an environment in which researchers could better query the existing sequencing knowledgebase and situate their work within it.

ILLUSTRATIONS FIGS? ?.

The advent of high throughput sequencing has led to datasets that are too large for such sharing and analysis. However, the development of high throughput sequencing has coincided with the rise distributed and cloud computing which provide promising avenues for handing the vast amounts of sequence data being generated and stored in databases. These two technologies are increasingly intertwined and have a significant impact on both the scale, scope, and methods of biological research.

RLX Ed & VAR

Backdrop of the computer industry & Moore's law:

Semiconductor technology has dramatically stimulated the development of integrated circuits for more than the last half century, which has led to the development of the personal computer and the Internet era. People have made observations of various laws which model and predict the rapid developmental progresses in these high-tech areas that are driven by the progress in semiconductor technology. For instance, the well-known Moore's law accurately predicted that the number of transistors integrated in each square inch would double every two year \cite{}. The semiconductor industry has also used the Moore's law to plan its research and development progress. Besides Moore's law, various other corresponding predictive laws have also been proposed for related 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 (also called Moore's second law) was proposed to predict the cost of a semiconductor chip fabrication plant doubles around every four years. Similarly, Kryder's law describes the related roughly yearly doubling of the area storage density of hard drives over the last few decades.

The roughly yearly doubling scaling of these described by these laws over the period of multiple decades is not simply the scaling behavior of a single technology but the superposition of the S-curve behavior of different technology over the life of the scaling behavior (see figure 1). The S-curve behavior of an individual technology is due to the three main phases (development, expansion and maturity). For example, Kyder's Law, yearly doubling scaling behavior over the last two and a half decades is the superposition of the S-curves of fives different technologies. This behavior is also true for sequencing based technologies.

The success of the predictive laws in high tech areas in last half century have encouraged the development of laws to forecast trends in related emergent technologies including sequencing based technologies. The cost of sequencing did roughly follow a Moore's law behavior in the decade before 2008 \cite{NIH cost-seq figure}. However, the sequencing cost has not followed a Moore's like law since 2008 after the introduction of new high throughput sequencing technologies \cite{NIH cost-seq figure}. Instead, the cost of sequencing has dropped faster than would be expected using Moore's law as a guide. In recent five years, the cost of sequencing a personal genomics dramatically dropped to XXX in 2014 from XXXX in 2008. This departure from Moore's law is due to the dramatically different S-curve slopes for Sanger sequencing and NGS. Consequently, transition between these technologies represented a new cost scaling regime. Thus, we think that the development of sequencing technology at this stage is far away from following a predictive trajectory.

[[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]]

Innovations underlying scaling in alignment algorithms:

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 (see figure 2). Underlying this improved performance are a series of discrete algorithmic advances. In the very early Sanger sequencing age, Smith-Waterman and Needleman-Wunsch algorithm used dynamic programming to find a local or global optimal alignment. But the O(n^2) time and space complexity of these approaches make it impossible to map sequences to a large genome. FASTA, BLAST and BLAT, as the successor, introduce a hash table based method that uses a seed-and-extend paradigm with an exact-matched K-mer as the seed. However, the original FASTA approach, which simply combines the K-mer and Smith-Waterman algorithm, cannot make sure best alignments are seeded. BLAST uses a heuristic statistical method to find high-scoring segment pairs (HSPs) 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.

CHK 2

Now, the challenge has turned into rapidly aligning millions of short sequences (reads) to a reference genome for next generation sequencing (NGS) aligners. Among tools selected in our analysis, MAQ and Novoalign are based on hash table, STAR is based on suffix array, BWA and Bowtie are both based on BWT. Gapped-kmer is used by MAQ to improve the sensitivity of seed-and extension schema. And a category of data structure, such as suffix array and FM-index (Ferragina–Manzini index or Full-text index in Minute space) adopted by STAR and BWA respectively, 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 turn out to be sacrificing optimal alignment and error tolerance for extremely fast retrieval of perfect matches.

HASH →
BWT
TRANS

CONT W/ WHAT?

On the other hand, more and more alignment tools try to reduce the mapping cost by building an index data structure, and generally the index time and alignment time are highly negative correlated (see figure 2). The hash table based tools: BLAT, MAQ and Novoalign take less time to build index structure, but require significant more time to do alignment. Though it take a long time to build the FM-index for BWA and Bowtie, and to construct an uncompressed suffix array for STAR, the index time cost is fixed and the marginal cost for reads alignment can be dramatically reduced, and make them become much more popular to handle progressively rising NGS data.

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

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. 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. 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 the NoSQL database which provides excellent horizontal scalability, data structure flexibility, and support for interactive queries.

INTRO EARLIER W/ CHG PARA-DIGMS

INTEG W EARLIE

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]]

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 bioinformatics study. We have already seen some exciting work in this field (cite ADAM from AMP Berkeley)

The explosion of sequencing data has posed a need of efficient methods for storage and transmission. General algorithms like gzip offer great compatibility, good compression speed and acceptable compression efficiency on sequencing data and are thus widely used. However, to further reduce storage footprint and transmission time, customized algorithms are needed. Many researchers SAM/BAM (Sequence/Binary Alignment/Map) format to store reads. An extensively accepted compression method, CRAM, is able to shrink BAM file by ~30% losslessly and more if lossy on quality score (\cite 21245279). CRAM only records the differences between reads and the reference genome and applies Huffman coding. Developing new and better compression algorithms is an active research field. We believe excellent compatibility and balance between usability and compression ratio are the keys for compression methods. With the latter depending heavily on specific research purposes, there is perhaps no one-size-fit-all algorithm. Besides compression, there is also work on data representation format to improve scalability in parallel computation and achieve better compatibility by defining an explicit data schema (\cite Massie: EECS-2013-207).

[[STL: distributed computing cuts cost. a single beefy node is much expensive than 100 mediocre nodes]]

Illustrations of the dramatic increase in rate and amount of sequencing:

A key component of the sequence data infrastructure is the sequence read archive (SRA), which was created to store and organize high throughput sequencing data generated for research purposes. The database has grown significantly since its creation in 2007. It now contains approximately $3.9*10^{15}$ bases with approximately half of these being open access. The size and growth rate of the SRA highlight the importance of efficiently storing sequence data for access by the broader scientific community. The SRA's centrality in the storage of DNA sequences from next generation platforms means that it also serves as a valuable indicator of the scientific uses of sequencing.

OTHER

A more detailed analysis of the SRA illustrates the pace at which different disciplines adopted sequencing. Plots depicting the cumulative number of bases deposited in the SRA and linked to by papers appearing in different journals provide a proxy for sequencing adoption. More general journals such as Nature and Science show early adoption. Meanwhile, SRA data deposited by articles from more specific journals such as Cell remained low for a significantly longer time before dramatically increasing (see figure 3).

Additionally, it is interesting to look at the contribution of large sequence depositions compared to smaller submissions. This provides an indication of the size distribution of sequencing projects.
At one end of this size spectrum are large datasets generated through the collaborative effort of many labs. These include projects that 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). On top of generating vast amount of sequencing data to better understand human variation and disease, high throughput sequencing has dramatically expanded the number of species whose genomes are are documented. The number of newly sequenced genomes has exhibited an exponential increase in recent years.

The cost of sequencing and the changing biological landscape:

The decrease in the cost of sequencing that has accompanied the introduction of new high throughput machines and the corresponding increase in the size of sequence databases has changed both the biological research landscape and the common modes of research. The amount of sequence data generated by the research community has exploded over the past ten years. This data has come from a variety of sources. In some cases, the decreasing cost has enabled ambitious large-scale projects aimed at measuring human variation in large cohorts and profiling cancer genomes. On the other hand, 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. 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 figure 4).

However, such trends also run the risk of fragmenting the genomics research community. If the sequence data generated by individual labs is not processed uniformly 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.

BUT DID YOU INTRO EARLIER

Meanwhile, the growth of sequence databases has reduced the cost of obtaining useful sequence information for analysis. Sequence 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. The analysis of sequence data has lower fixed costs but higher variable costs compared to sequence generation. The training and salary of bioinformatics analysts is a key fixed cost in sequence analysis. Variable costs associated with data transfer, storage, and processing all scale with the amount of sequence data being analyzed. The combination of costs in sequence data analysis doesn't provide the same economy of scale seen in the generation of sequence data.

In an era of squeezed budgets and fierce competition, job prospects for scientists with training in computational biology remain strong (\cite Explosion of Bioinformatics Careers Science 2014). Universities have increased the number of hires in the areas of computer science, and specifically in bioinformatics (see figure 5).

DANGLING