Factors Influencing Data Archival of Large-scale Genomic Data Sets

A mathematical formalism to comprehensively evaluate the costs-benefits of archiving large data sets

A RENCI TECHNICAL REPORT

TR-13-03

renci

CONTACT INFORMATION

Kirk Wilhemlsen
Telephone: 919.445.9619
Email: kirk@renci.org

Introduction

Next-generation genomic sequencing technologies and other high-throughput "-omics" technologies have enabled the rapid generation of large-scale data sets (Mardis, 2008; Koboldt et al., 2010). The costs associated with generating and storing these massive data sets have decreased precipitously, while computing power and storage capacity have simultaneously increased (Horvitz and Mitchell, 2010; Kahn, 2011). These new capabilities, coupled with new analytical algorithms and approaches to understanding and interpreting large-scale data (Horvitz and Mitchell, 2010; Koboldt et al., 2010), hold great promise to transform the field of genomics and realize the potential for personalized medicine.

However, the new capabilities raise questions regarding downstream reuse of genomic data and the costs-benefits of data archiving. While investigators fully recognize the drop in sequencing and storage costs, they rarely consider the additional costs associated with the archival of large genomic data sets, as well as secondary factors that may influence decisions related to archiving. These hidden costs and factors include: re-generation of new sources of genomic data (e.g., blood samples derived from lengthy, often expensive, clinical research studies); degradation of stored biological samples; introduction of errors during re-generation of genomic data sources and/or re-sequencing; long-term curation; data compression; data degradation; introduction of new technologies and analytical approaches (thus rendering stored data useless); data reuse needs; and time-related factors such as changes in reuse needs and costs of sequencing and storage.

We describe a mathematical formalism to comprehensively evaluate the costs-benefits of archiving large data sets and thus inform decision-making; we term this formalism the "*archival value criterion.*"

Archival Value Criterion

When deciding what data should be archived, we recommend using an archival value criterion or AVC. To a first approximation, we calculate the AVC metric with the function:

$$AVC = (P_{reuse} \times S')/S$$

where:

S = total cost for the storage and curation of data;

P_{reuse} = estimated probability of reuse; and

S' = cost of re-generation.

Application of the AVC to Guide Decision-Making

As a loose rule of thumb, an AVC metric $>10^2$ suggests data archiving rather than relying on data re-generation. After data have been archived, the criteria for its removal should be much lower, for instance, when the AVC metric is $<10^{-6}$. It is important to note that in practice, the AVC will be influenced by several factors (see next page) and should be computed under multiple scenarios before an informed decision can be made regarding archiving. After the decision is made to archive data, the AVC should

The Team

Kirk C. Wilhelmsen, MD, PhD, RENCI Chief Domain Scientist for Genomics and Professor, Departments of Genetics and Neurology, University of North Carolina at Chapel Hill

Charles P. Schmitt, PhD, RENCI Director of Informatics and Data Science

Karamarie Fecho, PhD, Medical and Scientific Writer, Copperline Professional Solutions on behalf of RENCI

Contact

Kirk Wilhelmsen, 919.445.9619
kirk@renci.org

TR-13-03 in the RENCI Technical Report
Series, September 2013
Renaissance Computing Institute
University of North Carolina at Chapel Hill
919-445-9640
www.renci.org
@renci

About RENCI

RENCI is an institute of the University of North Carolina at Chapel Hill that develops and deploys advanced technologies to enable research discoveries and practical innovations. RENCI partners with researchers, policy makers, and technology leaders to engage and solve the challenging problems that affect North Carolina, our nation and the world. The institute was launched in 2004 as a collaborative effort involving UNC Chapel Hill, Duke University and North Carolina State University. For more information, see www.renci.org.

be recalculated periodically over time to reassess the value of continued data archiving.

Factors that Influence the AVC

Accounting for full storage and curation cost. The total cost of storage and curation will depend heavily on the approaches that are taken. The following mathematical formalism includes cost estimates for factors typically associated with the data storage services offered by cloud providers such as Amazon:

$$S = C_{curate} + (1 + S_{curate} \times C_{curate}) \times C_{comp} \times (S_{ingest} + S_{stable} + P_{reuse} \times S_{retrieval})$$

where:

C_{curate} = fixed cost of curation;

S_{curate} = scaling factor for curation cost based on storage activities and size;

C_{comp} = estimated compression factor if storing data compressed;

S_{ingest} = cost of ingesting/moving data into the storage system;

S_{stable} = cost of storage; and

$S_{retrieval}$ = cost of retrieving data from the storage system.

Thus, in addition to direct storage costs (i.e., disk space and back-up drives), there are costs associated with curation, compression, and data transfer/retrieval, and these must be considered when evaluating the AVC metric. In the above formalism, C_{curate} represents the fixed cost of assigning staff to annotate the data, prepare the data for submission to an archival storage system, and retrieve the data; S_{curate} reflects the increasing costs associated with staff time as the volume of data increases. S_{stable} represents the actual cost for storage, often provided by commercial data storage providers in terms of "cost per terabyte per year." S_{ingest} and $S_{retrieval}$ represent the costs charged by data storage providers for submission and retrieval of data from the archival storage system, often provided in terms of "cost per terabyte transferred per month with tiers." S_{stable}, S_{ingest}, and $S_{retrieval}$ will depend on the type of storage requested, as cloud storage providers such as Amazon typically offer multiple levels of storage services.

Accounting for full re-generation cost. In accounting for the full re-generation cost (*S'*), one should take into account several factors. The first factor is the actual cost of re-generation, which includes both the cost of maintaining (or newly obtaining) materials for re-generation of data (e.g., blood or saliva derived from lengthy and costly clinical research studies) and the cost to re-generate the data derived from those materials. The second factor is the opportunity lost in allocating resources, such as staff time, to re-generate the data. The third factor is the array of potential complications that necessarily accompany re-generation and influence the cost of re-generation. For instance, stored tissue samples may degenerate or otherwise become corrupt over time (e.g., by freezer failure). Re-generation may also be inaccurate due to the absence of adequate documentation on the original processes used to generate the data (e.g., physical or electronic laboratory notebooks). These considerations may reduce the value of re-generation.

Adjusting for time-dependent parameters. Each of the variables in the AVC formula (*S*, P_{reuse} and *S'*) will likely change over time as technologies, market costs, and reuse needs evolve. A decision should be made as to whether these variables should be modeled as time-dependent parameters, in which case the AVC will need to be calculated as a summation or integral function. P_{reuse} in particular, must be considered carefully as P_{reuse} is impacted by a range of factors, including the likelihood that non-archival copies of the data, materials, and documentation are lost or corrupted over time and the likelihood that new advances in technologies (e.g., new genomic sequencing technologies) render the data less valuable.

Accounting for compression. The AVC may be greatly influenced by data compression, in large part because compression methods vary in their lossyness. The AVC should be estimated using multiple levels of lossy compression in order to determine the lowest level of compression that meets the AVC, with an understanding that the value of P_{reuse} may need to be reduced if compression is too lossy.

Application of the AVC: An Example

The following example illustrates how the AVC can be applied to determine whether it is cost-effective to archive sequencing data. The raw image files for a single human genome can require >1 terabyte of disk storage. In our experience, the cost of storing image files for even a few years is greater than the cost of re-generating the data, and P_{reuse} for an archived image file is very low (~10^{-6}) such that the AVC metric is <10^{-6}. The conversion of image data to FASTQ format reduces the file size ~100×, but it is very lossy. Nonetheless, Preuse will increase because the lost data are of little value due to the need for additional processing before reuse. Assuming that compression to the FASTQ format increases P_{reuse} to 10^{-2}, then the AVC metric will approach 1. Note that P_{reuse} is expected to fall rapidly as a function of time for FASTQ files because the unique opportunities afforded with stored FASTQ files diminish as a function of time, while the cost of processing the stored FASTQ files to a usable format becomes prohibitive. With further lossy compression to VCF format, the size of the FASTQ files can be reduced ~500× to essentially provide a list of sequence variants with annotation related to deviation from a reference genome. A variant list in VCF format has an initial AVC metric >500× higher than a FASTQ file because of its reduced size. The initial P_{reuse} with VCF files may be a little lower than that with FASTQ files because of additional data loss, but because VCF files can be used in combination with other genomes, this reduction is realistically negligible. Further, P_{reuse} with VCF files is likely to be stable over time because computation is not required to make the files usable. Thus, the AVC metric with VCF files should be much greater than that with FASTQ files.

Conclusion

We have described a relatively simple mathematical formalism to guide decision-making regarding the archiving of large genomic data sets; namely, the archival value criterion or AVC.

While we have developed the AVC for application in genomic decision-making, we emphasize that our formalism and approach can be adapted for more widespread application in decision-making regarding the archiving of any large data set.

References

Horvitz E, Mitchell T. From data to knowledge to action: a global enabler for the 21st century. Computing Community Consortium, v. 11. September 11, 2010. http://www.cra.org/ccc/docs/init/From_Data_to_Knowledge_to_Action.pdf. [Accessed March 11, 2013]

Kahn SD. On the future of genomic data. *Science*. 2011;331(6018):728–728.

Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. *Brief Bioinf.* 2010;11(5):484–498.

Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008;9:387–402.

How to reference this report:
Wilhelmsen, K., Schmitt, C. & Fecho, K. (2013). Factors Influencing Data Archival of Large-scale Genomic Data Sets: A Mathematical Formalism to Comprehensively Evaluate the Costs-Benefits of Archiving Large Data Sets. RENCI Technical Report Series, TR-13-03. University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi:10.7921/G0MW2F25

renci
RESEARCH ENGAGEMENT INNOVATION

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL