COMPUTATIONAL BIOLOGY IN THE CLOUD: METHODS AND NEW INSIGHTS FROM COMPUTING AT SCALE

PETER M. KASSON

Departments of Molecular Physiology and Biomedical Engineerng,
University of Virginia, Box 800886
Charlottesville, VA 22908, USA

The past few years have seen both explosions in the size of biological data sets and the proliferation of new, highly flexible on-demand computing capabilities. The sheer amount of information available from genomic and metagenomic sequencing, high-throughput proteomics, experimental and simulation datasets on molecular structure and dynamics affords an opportunity for greatly expanded insight, but it creates new challenges of scale for computation, storage, and interpretation of petascale data. Cloud computing resources have the potential to help solve these problems by offering a utility model of computing and storage: near-unlimited capacity, the ability to burst usage, and cheap and flexible payment models. Effective use of cloud computing on large biological datasets requires dealing with non-trivial problems of scale and robustness, since performance-limiting factors can change substantially when a dataset grows by a factor of 10,000 or more. New computing paradigms are thus often needed. The use of cloud platforms also creates new opportunities to share data, reduce duplication, and to provide easy reproducibility by making the datasets and computational methods easily available.

1. Challenges and opportunities of massive data

In recent years, large-scale datasets have become increasingly common in many biological fields. There have been tremendous strides in the throughput capacity and affordability of genomic sequencing. RNAi and similar techniques have allowed broad surveys of genetic regulation and host-pathogen interaction [1-3]. Multiple techniques for protein-protein association have gone large-scale, leading to “interactome” scale analyses of HIV infection [4] and other processes. “Brain atlas” projects are making available datasets that combine gene expression profiles with detailed anatomic and localization data [5]. And structural genomics projects have steadily increased the number of high-resolution macromolecular structures available for the proteins involved in all these processes. In addition to statistical analysis of all these datasets, more compute-intensive approaches such as large-scale simulation have made it possible to simulate many mutants of a drug target or combine data sources to examine the structure and dynamics of large subcellular structures. All these advances offer the possibility for tremendous insight into biology but pose challenges for effective analysis to maximize this insight.

The particulars involved in analyzing each of these domain-specific datasets have been treated elsewhere; we will discuss some of the common themes, particularly as they relate to cloud computing. Dataset size has increased greatly. Simply transmitting and storing many sequencing datasets is non-trivial. Sharing access to these data is yet more complicated when they are stored on individual researchers’ or centers’ computer systems. The challenges of sharing are compounded when patient data are concerned and access should be restricted and monitored. Analyzing these large datasets can also be very computationally intensive. Most analyses are at best case O(N); anything that leverages the comprehensive nature of large-scale datasets to examine pairwise or higher-order association often scales substantially worse. So the “classical” paradigms of storing datasets on local clusters and running analysis there are challenged three times: by transfer and archival capacity, by storage capacity, and by compute capacity.

2. Cloud solutions for problems of scale

We will briefly discuss how cloud-computing paradigms offer new solutions for these challenges; the workshop will illustrate a number of these as well as give an opportunity to discuss challenges and future directions. This subject has also been reviewed in [6-8] and elsewhere.

2.1. *Flexible capacity: the utility model of computing*

One substantial advantage of cloud-computing solutions is the ability to adjust computational resources according to requirements. Doubling the size of a traditional cluster is both expensive and time-consuming; in a cloud model, it simply involves requesting (and paying for) twice the capacity. The capacity limit of services such as Amazon's EC2 and Google Compute Engine has not been made public, but a recent demonstration showed the Institute for Systems Biology Genome Explorer running on 600,000 cores (Google IO 2012). In terms of raw compute capacity, this alone would likely rank among the top 5 supercomputers in the world. An additional advantage for cloud computing is the ability to dynamically adjust utilization. Most analyses do not run at a constant rate over time—one would like to request a large amount of resources to run a calculation and then release those resources while the results are evaluated and the next analysis is designed. It is thus rare to see a cluster at 100% utilization all the time, and like spare airline seat capacity or spare hotel rooms, the spare cycles are wasted money. In current cloud paradigms, a user pays only for the jobs (or virtual machines) that are running. This provides a much better fit to a fluctuating usage pattern.

2.2. *Storage*

Cloud storage solutions such as Amazon's S3 or Google Cloud Storage offer similar scalability and flexibility to the matching compute solutions. More importantly, they allow large and potentially shared datasets to be stored on the same infrastructure where large-scale analyses are run. This can obviously be achieved if one has a copy of a dataset on one's local cluster, but such an approach quickly becomes redundant when the dataset is held in common to many disparate users—the NCBI Short Read Archive is a good example. Caching a copy of this dataset on each cluster where analysis is run quickly becomes an expensive and redundant exercise. Companies such as DNAnexus have utilized cloud resources to offer storage, access to shared datasets, and transparent sharing of data. Cloud storage also provides enhanced reliability, as the data are backed up in several geographical locations.

2.3. *Parallel analyses*

When both computation and storage are performed in large distributed data centers, one can leverage technologies such as MapReduce[9, 10] and Dremel[11] for performing analyses and database queries in a much more efficient manner.

2.4. *Sharing data, tools, and algorithms*

One important and oft-overlooked benefit of the cloud model is how it can facilitate sharing. Most obviously, cloud data storage allows easy sharing with access control lists and monitoring of access for sensitive data. However, the ability to package and distribute tools and analyses on cloud platforms offers a new transparency in tool sharing and reproducibility. Furthermore, it helps overcome the problem of web server tools that are often overloaded or impose an undue burden on the host resources. If one imagines an analysis program running on a cloud front-end (such as Google App Engine) where any user can design a job to access a shared or private dataset and be presented with a virtual machine to run on his or her account for a common cloud service provider (Amazon EC2, Google Compute Engine, Microsoft Azure, or other), this allows much greater scalability for any public service and also reduces the cost to an individual researcher of making his or her methods publicly accessible.

3. Challenges in the cloud

Cloud computing offers new computational paradigms to deal with data and analyses at scale. However, simply applying the same algorithms and programming paradigms on 1000x the data often yields poor results. Working at scale generates different limiting factors on performance and cost, both algorithmic and logistical (data locality and transfer speed/cost, network latency, virtual machine start-up). Paradigms and tools such as the aforementioned

MapReduce (available in an open-source implementation from Hadoop) can yield great benefits but require refactoring code and rethinking computational approaches. We will discuss these and other challenges during the workshop session; participants will also share their experiences in overcoming some of these issues. Cloud computing is a classic example of more is different—working on different platforms and at larger scale offers new capabilities but also requires new ways of thinking and generates different performance-limiting problems. Nevertheless, we believe this paradigm offers the possibility for unprecedented insight into biological function.

Acknowledgments

P.K. is a visiting scientist at Google, Inc. and gratefully acknowledges many discussions with colleagues there.

References

1. Konig, R., Y. Zhou, D. Elleder, T.L. Diamond, G.M. Bonamy, J.T. Irelan, C.Y. Chiang, B.P. Tu, P.D. De Jesus, C.E. Lilley, S. Seidel, A.M. Opaluch, J.S. Caldwell, M.D. Weitzman, K.L. Kuhen, S. Bandyopadhyay, T. Ideker, A.P. Orth, L.J. Miraglia, F.D. Bushman, J.A. Young, and S.K. Chanda, *Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication.* Cell, 2008. **135**(1): p. 49-60.
2. Karlas, A., N. Machuy, Y. Shin, K.P. Pleissner, A. Artarini, D. Heuer, D. Becker, H. Khalil, L.A. Ogilvie, S. Hess, A.P. Maurer, E. Muller, T. Wolff, T. Rudel, and T.F. Meyer, *Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication.* Nature, 2010. **463**(7282): p. 818-22.
3. Hao, L., A. Sakurai, T. Watanabe, E. Sorensen, C.A. Nidom, M.A. Newton, P. Ahlquist, and Y. Kawaoka, *Drosophila RNAi screen identifies host genes important for influenza virus replication.* Nature, 2008. **454**(7206): p. 890-3.
4. Jager, S., P. Cimermancic, N. Gulbahce, J.R. Johnson, K.E. McGovern, S.C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G.M. Jang, S.L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A.J. O'Donoghue, A. Todorovic, J.H. Morris, D.A. Maltby, T. Alber, G. Cagney, F.D. Bushman, J.A. Young, S.K. Chanda, W.I. Sundquist, T. Kortemme, R.D. Hernandez, C.S. Craik, A. Burlingame, A. Sali, A.D. Frankel, and N.J. Krogan, *Global landscape of HIV-human protein complexes.* Nature, 2012. **481**(7381): p. 365-70.
5. Hawrylycz, M.J., E.S. Lein, A.L. Guillozet-Bongaarts, E.H. Shen, L. Ng, J.A. Miller, L.N. van de Lagemaat, K.A. Smith, A. Ebbert, Z.L. Riley, C. Abajian, C.F. Beckmann, A. Bernard, D. Bertagnolli, A.F. Boe, P.M. Cartagena, M.M. Chakravarty, M. Chapin, J. Chong, R.A. Dalley, B.D. Daly, C. Dang, S. Datta, N. Dee, T.A. Dolbeare, V. Faber, D. Feng, D.R. Fowler, J. Goldy, B.W. Gregor, Z. Haradon, D.R. Haynor, J.G. Hohmann, S. Horvath, R.E. Howard, A. Jeromin, J.M. Jochim, M. Kinnunen, C. Lau, E.T. Lazarz, C. Lee, T.A. Lemon, L. Li, Y. Li, J.A. Morris, C.C. Overly, P.D. Parker, S.E. Parry, M. Reding, J.J. Royall, J. Schulkin, P.A. Sequeira, C.R. Slaughterbeck, S.C. Smith, A.J. Sodt, S.M. Sunkin, B.E. Swanson, M.P. Vawter, D. Williams, P. Wohnoutka, H.R. Zielke, D.H. Geschwind, P.R. Hof, S.M. Smith, C. Koch, S.G. Grant, and A.R. Jones, *An anatomically comprehensive atlas of the adult human brain transcriptome.* Nature, 2012. **489**(7416): p. 391-9.
6. Schatz, M.C., B. Langmead, and S.L. Salzberg, *Cloud computing and the DNA data race.* Nat Biotechnol, 2010. **28**(7): p. 691-3.
7. Dudley, J.T. and A.J. Butte, *In silico research in the era of cloud computing.* Nat Biotechnol, 2010. **28**(11): p. 1181-5.
8. Schadt, E.E., M.D. Linderman, J. Sorenson, L. Lee, and G.P. Nolan, *Computational solutions to large-scale data management and analysis.* Nat Rev Genet, 2010. **11**(9): p. 647-57.
9. Dean, J. and S. Ghemawat, *MapReduce: A Flexible Data Processing Tool.* Communications of the Acm, 2010. **53**(1): p. 72-77.
10. Dean, J. and S. Ghemawat, *Mapreduce: Simplified data processing on large clusters.* Communications of the Acm, 2008. **51**(1): p. 107-113.
11. Melnik, S., A. Gubarev, J.J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, *Dremel: Interactive Analysis of Web-Scale Datasets.* Communications of the Acm, 2011. **54**(6): p. 114-123.