Genome Medicine

COMMENTARY

Computational ecosystems for data-driven medical genomics

Jonas S Almeida*

Abstract

In the path towards personalized medicine, the integrative bioinformatics infrastructure is a critical enabling resource. Until large-scale reference data became available, the attributes of the computational infrastructure were postulated by many, but have mostly remained unverified. Now that large-scale initiatives such as The Cancer Genome Atlas (TCGA) are in full swing, the opportunity is at hand to find out what analytical approaches and computational architectures are really effective. A recent report did just that: first a software development environment was assembled as part of an informatics research program, and only then was the analysis of TCGA's glioblastoma multiforme multi-omic data pursued at the multi-omic scale. The results of this complex analysis are the focus of the report highlighted here. However, what is reported in the analysis is also the validating corollary for an infrastructure development effort guided by the iterative identification of sound design criteria for the architecture of the integrative computational infrastructure. The work is at least as valuable as the data analysis results themselves: computational ecosystems with their own high-level abstractions rather than rigid pipelines with prescriptive recipes appear to be the critical feature of an effective infrastructure. Only then can analytical workflows benefit from experimentation just like any other component of the biomedical research program.

Anduril

A report by Ovaska *et al.*, recently published in Genome Medicine [1], describes the use of an integrative computational infrastructure, Anduril, to analyze glioblastoma multiforme data in The Cancer Genome Atlas (TCGA). The logical and logistic consistency of the framework allowed the assembly of a large analytical workflow made of hundreds of individual processes, while avoiding the reproducibility and traceability pitfalls that currently plague complex analyses for biomarker identification. Consequently, instead of being limited to a specific molecular signal, the study was able to approach the integrated analysis of a wide variety of data. Specifically, data on 338 patients with primary glioblastoma multiforme, with clinical annotations, hybridization arrays, SNP, exon, gene expression and microRNA, were analyzed together. Tellingly, the authors were not only able to associate novel genomic alterations with glioblastoma multiforme progression, but also, and contradicting the criticism of mechanistic inconclusiveness of 'fishing expeditions', proceeded to explore novel roles for moesin in cell proliferation. As the study illustrates, hypothesis-generating functional analysis still can, and maybe should, be the corollary of integrative data driven analytical workflows. After reading this report by Ovaska *et al.* it is not hard to imagine that even mechanistic hypothesis testing may one day be treated as an extension of a broader integrative workflow.

Integrative computation for personalized medicine

The promises of personalized molecular medicine are increasingly driving large-scale associative genomics projects that bring together distributed teams involving multiple disciplines. The unprecedented size and scope of initiatives such as TCGA inevitably come with new types of growing pains. The problem of integration, as described in a recent report by The National Academy [2], is quickly becoming the central challenge for the life sciences. Only a few years ago this was still the province of the visionary [3]. However, the clamor for better formal knowledge representation frameworks is now coming from all corners, including the critical contribution of the storage infrastructure community [4]. In that regard, the study by Ovaska *et al.* may be revealing of what is in store for the data analysis of large-scale genomic data generation initiatives. Anduril is not the first integrative framework proposed, and the report compares with existing frameworks such as GenePattern, Ergatis and Taverna [5-7]. In fact, GenePattern is the

*Correspondence: jalmeida@mdanderson.org
Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA

framework where many of the analytical workflows of the TCGA initiative itself are deployed. What is particularly interesting about Anduril in this regard is not so much what it does as to what extent it successfully reflects the relationship between its architecture and the team that uses it. Before dwelling on that, it is worth recalling that this framework has pushed the idea of component modularity all the way to a shared input/output (I/O) bus (that is, a set of logical connections that can be shared by multiple software components in order to communicate with one another). As a result, a computational ecosystem is enabled where, instead of workflows made of components designed to define a pipeline, one has components with application programming interfaces designed for scaled-up re-usability. Whereas in the conventional pipeline approach each component is designed of as a piece of a specific analytical puzzle, in the ecosystem approach the application programming interface of each module is made sufficiently abstract as to be treated like an autonomous, generic element of many possible workflows.

The Anduril framework [8] was devised with a specific team of users in mind. This team comprises three roles: molecular biologists at both the data acquisition and the interpretation ends of the workflow, computational statisticians developing specialized data analysis modules in a variety of programming environments, and, finally, dedicated analysts assisting and articulating both groups. The command line operation of the framework suits the analyst group as an environment to make full use of the component-based workflow framework designed from maximum re-usability and minimum administration load. The execution of individual components by the core engine of Anduril is automatically triggered by I/O dependencies that point to filenames in a shared file system. It is also telling that the ensuing high-level abstraction led the developers of Anduril to identify their own domain-specific language, releasing the whole initiative from having to choose between the many actual programming languages used for the individual components. Even if it is far from certain that Anduril will find a broader community of users, it is clear that this computational framework was the critical resource that enabled this particular group to act as a team. Therefore, it appears that integrative multidisciplinary teams may respond better to computational frameworks (plural) designed to match them, instead of forcing existing collaborative teams into a shared workflow mold. The latter remain the primary impulse of large-scale genomics initiatives, with very mixed results.

Multidisciplinary collaboration in a distributed world

Another provocative observation is that the authors of this study, and of the supporting computational framework, are not themselves involved in the TCGA initiative. This may be the beginning of a trend towards computational integration between unrelated research groups. This may actually be the better way for large-scale genomics initiatives to be translated into biomedical applications. If that is the case, then the global reach would become a priority feature of such initiatives, with a critical attention to streamlined programmatic access to the data generated.

Some features of the integrative framework reflect the collaborative team work in ways that are less relevant to this commentary. The physical co-location of the computational components at universities in the Helsinki-Turku area allow for an architecture tied together by a shared file system. At a time of widening availability of Hypertext Transfer Protocol (HTTP)-mediated cloud computing resources and convergence towards semantic web formalisms, the reliance of Anduril on I/O via read/write of files may be an unreasonable proposition for distributed deployments. A more distributed computational ecosystem may be better served by web services, potentially extending component execution, not just reporting, to any machine connected to the Web. Nevertheless, the design of Anduril as a platform able to host and sustain abstract workflow representations that call arbitrary components is novel and compelling beyond the specific details of its architecture.

Abbreviations

I/O, input/output; SNP, single nucleotide polymorphism; TCGA, The Cancer Genome Atlas.

Competing interests

The author's research is partially funded by The Cancer Genome Atlas initiative through a Genome Data Analysis Center award: 1U24CA143883-01.

Published: 20 September 2010

References

1. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen A-M, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S: **Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme.** *Genome Med* 2010, **2:**65
2. National Research Council of The National Academies: *A New Biology for the 21st Century*. Washington, DC: The National Academies Press; 2009.
3. Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner DJ: **Computer science. Creating a science of the Web.** *Science* 2006, **313:**769-771.
4. Hey T, Tansley S, Tolle K (Eds): *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research; 2009.
5. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P: **Taverna: a tool for the composition and enactment of bioinformatics workflows.** *Bioinformatics* 2004, **20:**3045-3054.
6. Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV: **Ergatis: a web interface and scalable software system for bioinformatics workflows.** *Bioinformatics* 2010, **26:**1488-1492.
7. Reich, M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0.** *Nat Genet* 2006, **38:**500-501.
8. **Anduril** [http://csbi.ltdk.helsinki.fi/anduril]

doi:10.1186/gm188
Cite this article as: Almeida JS: **Computational ecosystems for data-driven medical genomics.** *Genome Medicine* 2010, **2:**67.