Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 1635 – 1638

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Kepler for 'omics bioinformatics

Mark Bieda*

Department of Biochemistry and Molecular Biology, University of Calgary, 3330 Hospital Dr NW, Calgary, AB T2N 4N1

Abstract

There has been a massive increase in the number of large scale biological datasets during the past twenty years, producing new challenges and complexities for analysis. Many of these new datasets are in the 'omics fields, involving analysis of the genome, transcriptome, and proteome among others. Here, we review 'omics community-specific factors affecting use of bioinformatics workflow systems. We identify the characteristics of the audience for scientific workflow systems in this community, the existence of a large amount of prewritten software, the use of large amounts of data in a typical analysis, and the growing complexity of analyses as important factors in considering workflow design criteria in this field and also future development of Kepler. Generally, many factors favor much increased use of Kepler in bioinformatics in the future, in particular its advantages in comprehensibility, extensibility, and modifiability of bioinformatics pipelines. We suggest concrete steps to enable further use of this flexible workflow system in 'omics analyses.

Keywords: scientific workflows, bioinformatics, microarrays, sequencing, genomics

1. Introduction

During the past two decades, there has been an enormous explosion in the power and number of technologies for deriving large scale biological data on a molecular level (e.g. gene expression microarrays, next-generation sequencing) [1]. Within the world of molecular biology, these high throughput areas are refered to as " 'omics" as in genome. Just as the genome refers to all of the DNA in a cell/tissue, other 'omics terms refer to the overall set of components in a cell. For example, the proteome (and by extension proteomics) refers to large scale determination of protein abundance/identity in a cell type/tissue type. Current additional prevalent uses of omics terms include "metabolome" (for all metabolic products) and "transcriptome" (all RNA transcripts), among others.

New 'omics technologies have led to new challenges for bioinformatics analyses of the resulting data sets. There has been a large amount of development of new bioinformatics algorithms and also of appropriate metadata resources (e.g. pathway analysis systems for gene expression). It has become apparent that even closely related platforms within an 'omics technology (e.g. Affymetrix microarrays versus Illumina microarrays for gene expression) require development of individualized algorithms and software tools for accurate analysis. In addition,

* Corresponding author. Tel.: 403-210-6157; fax: 403-210-8115.
E-mail address: mbieda@ucalgary.ca.

doi:10.1016/j.procs.2012.04.180

development of new technologies and new platforms is ongoing.

Here, we focus on the needs of the 'omics community for scientific workflows. We highlight four basic factors that will affect development of Kepler workflows and Kepler itself for this community. First, the audience of the workflows must be taken into account. We argue that experienced bioinformaticians are the primary audience of the system, but that the great majority of end-users will have relatively few computer skills. Second, the existence of a large and continuously growing set of bioinformatics tools means that Kepler's ability to easily and intelligently integrate external tools is critical. Third, the large size of individual datasets argues for computation primarily based on local resources and local programs, at least for initial steps. Finally, data analyses in this area are growing increasingly complex, which provides clear opportunities for large growth in use of scientific workflow systems generally. We conclude by considering steps for the Kepler community to address these needs.

2. Kepler in the 'omics laboratory

2.1. Audience for Kepler in the 'omics laboratory

Scientific workflow systems are sometimes seen as systems in which the end-user would reconfigure workflows or even develop de novo workflows from a library of actors. This does not fit the realities of molecular biology environments. Instead, usage of scientific workflow systems will currently be divided into two fundamental, and very different, groups. Bioinformaticians, who generally will possess good understanding of bioinformatics algorithms, parameter usage, and general programming, will develop actors and full workflows and, in some cases, also act as final end-users (performing actual data analysis). Classical molecular biologists, generally possessing relatively low amounts of understanding of algorithms or programming, will simply use the software under instruction from the bioinformatician.

This model has important implications for use of general purpose workflow systems in the bioinformatics community and for the appropriate development of workflows for this community. Critically, the workflow systems must offer primary value to the bioinformatician; end-users treating workflows as "black box" applications may not see special value in the Kepler approach. The low level of computer skills of most end-users means that simplicity and clarity of actual workflow operation must be emphasized in Kepler development and workflow development. We suggest that Kepler's ability to add annotations to the workflow canvas with control over position, text font, size, and color offers a major advantage and should be heavily employed in 'omics workflows.

2.2 Large existing and rapidly growing body of software

The current phase of 'omics features both rapid introduction of new experimental platforms and great expansion in the number of laboratories using these technologies. Many bioinformaticians work on development of new primary data analysis algorithms with implementation and testing using traditional programming approaches (e.g. Java programming). Scientific workflow systems are sometimes viewed as environments for development of these primary data analysis algorithms. Under this model, actors focusing on low-level functions are important. However, it seems highly unlikely that Kepler will be used in this way given the bioinformatics community's long history of success of algorithm development using traditional approaches and the community's needs for algorithm speed and ease of distribution.

Currently, there are large numbers of existing software packages and many are being produced for both primary data analysis and secondary analysis. Importantly, many 'omics software packages offer well-written, pretested implementations of complex algorithms, including handling of special cases. The often-used R/BioConductor set features 516 packages as of 2011 [2], with nearly all aimed toward 'omics analysis. These packages typically consist of 10 or more modules, potentially leading to >5000 different modules, each of which could be implemented as an actor. Moving these ~5000 or more modules into Kepler as individual actors would currently be very time-consuming. Hence, the Kepler community should seek to develop approaches to automate or at least semi-automate the conversion of R/BioConductor packages to sets of Kepler actors.

Given the large body of high-quality tools, the experienced bioinformatician is most interested in using existing programs as components in pipelines. Under these conditions, it is important that efficient creation of Kepler actors based on external programs is specifically supported. Intermediate tools that allow rapid, guided production of these

actors (perhaps based on a simple external utility) would greatly enhance use of Kepler in this community. Specific scientific questions often require customized 'omics pipelines. Generally, given the rapidly changing landscape of 'omics technologies and software, there are also great needs for modifiability and extensibility of pipelines, clear strengths of Kepler. Finally, the intrinsically graphical nature of Kepler workflows can enable rapid comprehension of pipeline functionality and components, which is also favored in a situation with rapidly changing demands and potential pipeline components.

2.3 Large data files

Although scientific workflow systems are often phrased in terms of use of remote web services [3], this is not an appropriate usage model in this community. 'Omics analyses increasingly involve files (or sets of files) that are large in total (100s of megabytes to several gigabytes) [4]. Furthermore, total data set sizes are increasing over time, mostly due to increases in the size of individual data sets as high-throughput techniques increase in scale. Moving these large sets of data from one web service to another is slow and can be problematic, so use of local resources (local data and local programs) is heavily favored. In total, realistic workflows must rely on local resources in most cases.

However, at later stages of data processing, there is often (but not always) relevant and significant data reduction. For example, a large ChIP-seq (chromatin immunoprecipitation followed by sequencing) dataset is usually reduced to a much smaller dataset of binding site locations in the genome. Use of remote web services for the later processing stages is reasonable and appropriate. In particular, use of remote data annotation resources may be beneficial.

2.4 Data analyses are growing increasingly complex

The increasing size of the 'omics world has led to a large increase in demands on bioinformatics analyses. To begin, analyses often involve multiple data types. For example, ChIP-seq data is often combined with gene expression data to gain insight into functions of transcription factors. Second, the existence of different platforms aimed at the same type of data (e.g. different formats of gene expression microarrays) means that cross-platform analysis pipelines must be developed. Third, the development of high quality metadata resources and new types of resources (e.g. gene pathway resources) produces demands for more complex, better metadata analyses [5]. Finally, the existence of large, well-described and publically available archives of data (e.g. [6]) has led to increasing interest in meta-analyses of already published data.

In an environment of small scale and relatively simple analyses, scientific workflow systems are more of a luxury than a necessity. However, just over the past two years, bioinformatics analyses have reached a size and complexity at which the attributes of scientific workflow systems are now requirements instead of attractive options. Hence, there is a current opportunity for Kepler to play a much larger role in the bioinformatics community. Of special note is that bioinformatics analyses are already parameter-rich. The ability of Kepler, in particular, to provide a clear representation of parameters on the workflow canvas will be increasingly valued under these conditions of increasingly complex analyses.

2.5. Two paths forward for use of Kepler in the laboratory

How can Kepler play a larger role in the complex ecosystem of bioinformatics tools? The disadvantages of Kepler should not be underestimated, in that there can be a significant learning stage and the concepts underlying the Kepler model will not be familiar to most bioinformaticists. As bioinformatics analyses become more complex, Kepler (or other similar systems) has the promise of providing robust solutions to difficult problems.

There are two clear potential models to ease Kepler acceptance in the world of bioinformatics. First, the construction of specific 'omics workflows and actors - especially those embodying the most used tools in the field - may be critical for adoption of Kepler in this community. Second, development of "unique" workflows in the Kepler system would stimulate usage. We suggest that these workflows should focus upon poorly addressed issues in bioinformatics or areas that have attracted less attention than mammalian genomics (e.g. proteomic analyses). Two other approaches are useful but less impactful. Large workflows that embody complex analyses can serve as

easy demonstrations of the power of the Kepler model and should in themselves be useful (e.g. WATERS for metagenomics [7]). Finally, production of workflows within Kepler that use existing algorithms but provide important functionality addressing critical current issues will induce some usage. For example, workflows combining data types would be very attractive to some bioinformaticians.

3. Conclusions

We suggest that the nature of bioinformatics analyses places certain constraints on development of Kepler workflows and the probable use of Kepler in bioinformatics contexts. First, bioinformatics workflows should emphasize use of local resources (data, programs) particularly for the first stages of computation. This follows from the large size of primary data files and the fact that many bioinformatics applications are not available as web services. However, for secondary analysis when files are smaller, development of a library of actors for connecting to popular bioinformatics web services, especially those concerned with annotation (metadata) issues, should be emphasized. Second, because bioinformaticians will control adoption of this system in 'omics fields, ease of development of workflows focusing primarily on external programs should be emphasized. In particular, special provision should be made to ensure that R/BioConductor workflows can be easily developed in the Kepler framework. For example, R/BioConductor syntax highlighting and rapid R/BioConductor actor testing would be advantageous in future Kepler versions. Third, because most end-users will be classical molecular biologists with few computer skills, continued development of graphical aspects of Kepler is a priority for this community. Changes should be user-tested. Fourth, 'omics workflow design should take advantage of Kepler's abilities to promote extensibility, modifiability, and comprehensibility of workflows. In particular, the ability of Kepler to organize parameters on the workflow canvas should be emphasized. 'omics is growing and bioinformatics education is becoming more standardized and widespread, so comprehensibility will become more important as more users desire deeper understanding of bioinformatics pipelines. Finally, due to this community's use of mostly external applications that produce data files as output, data provenance for 'omics workflows should develop a focus on tracking external files and use a format designed for bioinformatician accessibility.

Acknowledgements

We thank Thomas Stropp and Timothy McPhillips for discussions relevant to this topic.

References

1. A. Kowald and C. Wierling, Standards, tools, and databases for the analysis of yeast ’omics data. Methods Mol. Biol. 759 (2011) 345-365.
2. M. Morgan, Bioconductor Annual Report 2011. - http://www.bioconductor.org/about/annualreports/AnnRep2011.pdf.
3. V. Curcin and M. Ghanem, Scientific workflow systems - can one size fit all?. in Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, 2008, pp. 1-9.
4. J. Rougemont and F. Naef, Computational analysis of protein-DNA interactions from ChIP-seq data. Methods Mol. Biol. 786 (2012) 263-273.
5. D. W. Huang, et al, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protocols, 4 (2008) 44-57.
6. E. W. Sayers, et al, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 40 (2012) D13-25.
7. A. L. Hartman, et al, Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. *BMC Bioinformatics*, 11 (2010) 317.