Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME

Pierre Lindenbaum1, Solena Le Scouarne2, Vincent Portero1 and Richard Redon1
1Institut du thorax, Inserm UMR 915, Centre Hospitalier Universitaire de Nantes, 44000 Nantes, France and 2The Welcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have led to an explosion of the amount of data to be analysed. As an example, a VCF (Danecek et al. 2011) file (Variant Call Format—a standard specification for storing genomic variations in a text file) produced by the 1000 Genomes Project contains about 25 million Single Nucleotide Variants (SNV), http://tinyurl.com/ALL2of4intersection (retrieved September 2011), making it difficult to extract relevant information using spreadsheet programs. While computer biologists are used to invoke common command line tools—such as Perl and R—when analysing those data through Unix pipelines, scientific investigators generally lack the technical skills necessary to handle these tools and need to delegate data manipulation to a third party.

Scientific workflow and data integration platforms aim to make those tasks more accessible to those research scientists. These tools are modular environments enabling an easy visual assembly and an interactive execution of an analysis pipeline (typically a directed graph) where a node defines a task to be executed on input data and an edge between two nodes represents a data flow. These applications provide an intuitive framework that can be used by the scientists themselves for building complex analyses. They allow data reproducibility and workflows sharing.

Galaxy (Blankenberg et al. 2011), CyberT2 (Bersier et al. 2009), and Mobyle (Non et al. 2008) are three web-based workflow engines that users have to install locally if computational needs on datasets are very large, or if absolute security is required. Alternatively, softwares such as the KNIME (Konstanz Information Miner) interactive graphical workbench, for the interpretation of large biological datasets. We demonstrate that this tool can be utilized to quickly retrieve previously published scientific findings.

2 IMPLEMENTATION

The java API for KNIME was used to write the new nodes, during the reviewing process of this article another solution based on KNIME, but focusing in FASTQ data files was published by Jagla et al. (Jagla et al. 2011).

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/27/22/3200/196060 by guest on 29 July 2018
We want to thank the Biostar community for its help, Jim Robinson importantly
NOTCH2
www.myexperiment.org/workflows/2320.

As a proof of concept, we tested our nodes to analyse the exomes of six patients from a previously published study (Isidor et al., 2011) can be used visualize the short reads overlapping a variation.

Fig. 1. Screenshot of a Knime4Bio workflow for the NOTCH2 analysis.

In practical terms, a computer biologist was close to our users to help those who require a snippet of java code). At the time of writing, they were able to quickly play with the interface, add some nodes with the construction of a workflow. After this short tutorial, they had their first interactive tool for NGS analysis.

ACKNOWLEDGEMENTS

We want to thank the Biostar community for its help, Jim Robinson and his team for the BigWig java API and Dr Cedric Le Caignec for the NOTCH2 data.

Fig. 1. Screenshot of a Knime4Bio workflow for the NOTCH2 analysis.

3 DISCUSSION

In practical terms, a computer biologist was close to our users to help them with the construction of a workflow. After this short tutorial, they were able to quickly play with the interface, add some nodes and modify the parameters without any further assistance, but the suggestion or the configuration of some specific nodes (for example, those who require a snippet of java code). At the time of writing, Knime4Bio contains 55 new nodes. We believe Knime4Bio is an efficient interactive tool for NGS analysis.

REFERENCES

Addahabi, I.A. et al. (2010) A method and server for predicting damaging missense mutations. Nat. Methods., 7, 248–249.

Berthold, M.R. et al. (2007) Knime: the konstanz information miner. In Prikluschev, C. (eds) GfKl, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, pp. 519–526.

Blankenberg, D. et al. (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. Databases, 2011, b011.

Danecek, P. et al. (2011) The variant call format and VCFtools. Bioinformatics, 27, 2156–2158.

Fiers, M. W.E.J. et al. (2008) High-throughput bioinformatics with the Cyrille2 pipeline system. BMC Bioinformatics, 9, 96.

Hsu, F. et al. (2006) The UCSC Known Genes. Bioinformatics, 22, 1036–1046.

Hull, D. et al. (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res., 34, W729–W732.

Isaak, R. et al. (2011) Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nat. Genet., 43, 306–308.

Jagla, B. et al. (2011) Extending KNIME for next generation sequencing data analysis. Bioinformatics, 27, 2907–2909.

Kent, W.J. et al. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics, 26, 2204–2207.

Liu, X. et al. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat., 32, 863–874.

Li, B. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078–2079.

Li, B. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics, 27, 589–590.

Li, B. et al. (2009) Moby: a new full web bioinformatics framework. Bioinformatics, 25, 3005–3011.

Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. Genome Res., 11, 863–874.

Robinson, J.T. et al. (2011) Integrative genomics viewer. Nat. Biotechnol., 29, 24–26.

von Mering, C. et al. (2007) STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res., 35, D358–D362.