Erschienen in: Journal of Biotechnology ; 261 (2017). - S. 149-156
http://dx.doi.org/10.1016/j.jbiotec.2017.07.028

Journal of Biotechnology 261 (2017) 149–156

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Review

KNIME for reproducible cross-domain analysis of life science data

Alexander Fillbrunn[a,e,*], Christian Dietz[a], Julianus Pfeuffer[b], René Rahn[c], Gregory A. Landrum[d], Michael R. Berthold[a,d,e]

[a] University of Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany
[b] Eberhard Karls Universität Tübingen, Geschwister-Scholl-Platz, 72074 Tübingen, Germany
[c] Freie Universität Berlin, Kaiserswerther Straße. 16-18, 14195 Berlin, Germany
[d] KNIME, Technoparkstrasse 1, 8005 Zurich, Switzerland
[e] Konstanz Research School Chemical Biology, Germany

ARTICLE INFO

Keywords:
KNIME
Workflow systems
Life science

ABSTRACT

Experiments in the life sciences often involve tools from a variety of domains such as mass spectrometry, next generation sequencing, or image processing. Passing the data between those tools often involves complex scripts for controlling data flow, data transformation, and statistical analysis. Such scripts are not only prone to be platform dependent, they also tend to grow as the experiment progresses and are seldomly well documented, a fact that hinders the reproducibility of the experiment. Workflow systems such as KNIME Analytics Platform aim to solve these problems by providing a platform for connecting tools graphically and guaranteeing the same results on different operating systems. As an open source software, KNIME allows scientists and programmers to provide their own extensions to the scientific community. In this review paper we present selected extensions from the life sciences that simplify data exploration, analysis, and visualization and are interoperable due to KNIME's unified data model. Additionally, we name other workflow systems that are commonly used in the life sciences and highlight their similarities and differences to KNIME.

1. Introduction

Graphical workflow systems are used in many areas of research and commerce to model complex processes. Typically a workflow in such a system is composed of different processing nodes that pass data to each other and can be augmented with titles, annotations and descriptions. This makes them a viable alternative to custom scripts which often lack documentation, require the data to be formatted for called tool or have dependencies that are not necessarily in place on every system on which they are supposed to run.

The workflow system KNIME (Berthold et al., 2007) is an open source software that aims to solve these problems by providing a platform that can be extended easily with new tool integrations, has a strongly-typed data system and allows workflow creators to document the steps performed by the workflow in detail. Additionally, old nodes in KNIME are never completely removed from the program but are deprecated so that workflows built with old versions can still be run and produce the same results years later. The user interface of KNIME is shown in Fig. 1.

Apart from the free and open source KNIME Analytics Platform, KNIME also has commercial offerings. The KNIME server provides a platform for sharing workflows. It has a web interface and is connected to a KNIME instance for executing workflows remotely on demand or according to a schedule. Also commercially available are the Big Data Extensions and the KNIME Spark executor.

For data and life sciences these properties of a workflow system are crucial, as workflows can serve as documentation for experiments themselves and facilitate reproducible science. KNIME ships with an extensive catalogue of processing nodes, but its strength is the availability of a multitude of extensions for various research areas. As a platform that embraces the integration of tools and provides a shared data structure, KNIME allows nodes from different research areas to be mixed to create truly cross-domain workflows. In this review paper we present some of the extensions that enable KNIME to be a very capable workflow system for life science research.

Section 2 introduces plugins for data analysis and the integration of popular scripting languages, Section 3 provides an overview over the life science extensions for areas such as cheminformatics, image processing, mass spectrometry and next generation sequencing. In Section 4 we briefly describe other useful additions available for KNIME and in

* Corresponding author.
E-mail addresses: alexander.fillbrunn@uni-konstanz.de (A. Fillbrunn), christian.dietz@uni-konstanz.de (C. Dietz), pfeuffer@informatik.uni-tuebingen.de (J. Pfeuffer), rene.rahn@fu-berlin.de (R. Rahn), greg.landrum@knime.com (G.A. Landrum), michael.berthold@uni-konstanz.de (M.R. Berthold).

http://dx.doi.org/10.1016/j.jbiotec.2017.07.028
Received 17 February 2017; Received in revised form 21 July 2017; Accepted 25 July 2017
Available online 27 July 2017

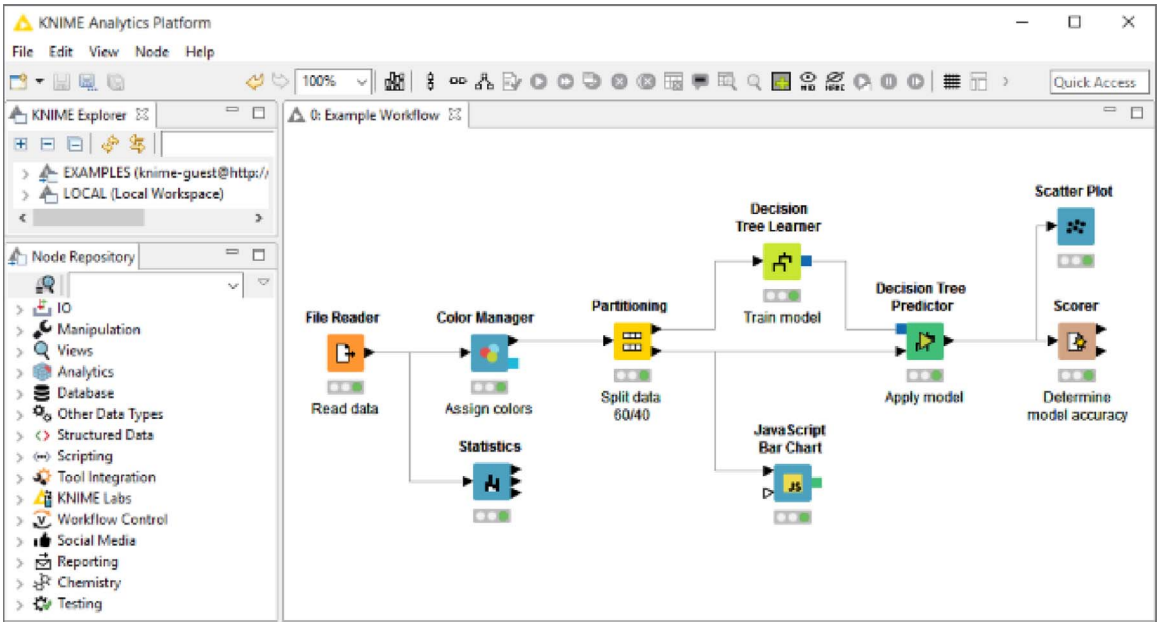

Fig. 1. The user interface of KNIME. The workflow is built in the center of the screen, on the right there are the node and workflow repository views.

Section 5 other workflow tools for the life sciences are presented. We summarize the main points of this review in Section 6.

2. Data analysis extensions

One of KNIME's strengths is its multitude of nodes for data analysis and machine learning. While its base configuration already offers a variety of algorithms for this task, the plugin system is the factor that enables third-party developers to easily integrate their tools and make them compatible with the output of each other. In this section, we present tool integrations that are of special interest for users in the data sciences.

R is an important tool, not only for data scientists but also for scientists in other research areas. It is a free software environment for statistical computing and graphics (R Development Core Team, 2008). Its greatest strength is its extensive package ecosystem, created by a large community of developers and scientists. Notable libraries include such for building classification and regression models (Venables and Ripley, 2002; Jed Wing et al., 2016), for visualization (Wickham, 2009) and for computations on large datasets (Dowle et al., 2015). Among scientists, R is a popular environment for analysis and visualization. KNIME therefore provides multiple nodes that integrate this environment into the platform, allowing interoperability between the two systems and thus the reuse of existing scripts and packages. These nodes allow KNIME users to process their data tables and to generate visualizations in R; all embedded in a workflow and with results in a format that can be processed further with other nodes.

Recently, the programming language Python has become a strong contender for the data analysts programming language of choice. The development of open-source libraries such as SciPy (Jones et al., 2001) and SciKit-Learn (Pedregosa et al., 2011) as well as the language's ability to use native libraries easily, especially for efficient algorithms, has made it a popular tool for data scientists. While R's style of programming is focused on statisticians, Python is a general purpose programming language with a large set of packages for developers who want to delve into data analysis. KNIME includes Python in various processing nodes for data processing, model learning and prediction, and the generation of visualizations.

Another integration worth mentioning is WEKA (Hall et al., 2009), a set of machine learning algorithms for data mining in Java programs. WEKA provides various models and algorithms for regression and classification, such as neural networks, self-organizing maps, Naive Bayes and many more. Bindings for R and Python are also available. KNIME's nodes for WEKA 3.7 algorithms can execute the libraries' algorithms on KNIME data tables and make generated data and models available on general purpose ports, facilitating further processing with other tools.

It is especially thanks to the work of Yann LeCun and Yoshua Bengio (LeCun et al., 2015) that the application of deep neural networks has boomed in recent years. The technique, which utilizes neural networks with many layers and enhanced backpropagation algorithms for learning, was made possible through both new research and the ever increasing performance of computer chips. The design of a suitable architecture for a deep neural network is an important part of acquiring a well-performing predictor and therefore is of special importance for a tool that learns deep neural networks. In KNIME this requirement is met by providing a collection of nodes for modeling and assembling such a network in a workflow, adding new layers by connecting nodes with each other and feeding the final structure into a learner node that tunes the model's parameters on the training data. A predictor completes the package and enables the application of deep neural network predictors on new, unseen data. By utilizing the Deeplearning4j library[1] for model representation, learning and prediction, KNIME builds upon a well performing open source solution with a thriving community. An example workflow for the integration is depicted in 2 , where Deeplearning4j and KNIME Image Processing are used to predict the font of a character in an image. The workflow can also be found on the KNIME example server.

3. Life science extensions

3.1. RDKit

RDKit (2013) is an open source toolkit for cheminformatics – working with molecules in the computer. The core data structures and algorithms are written in C++ for performance and maximum portability. Wrappers around the C++ are provided for programming

[1] https://deeplearning4j.org/.

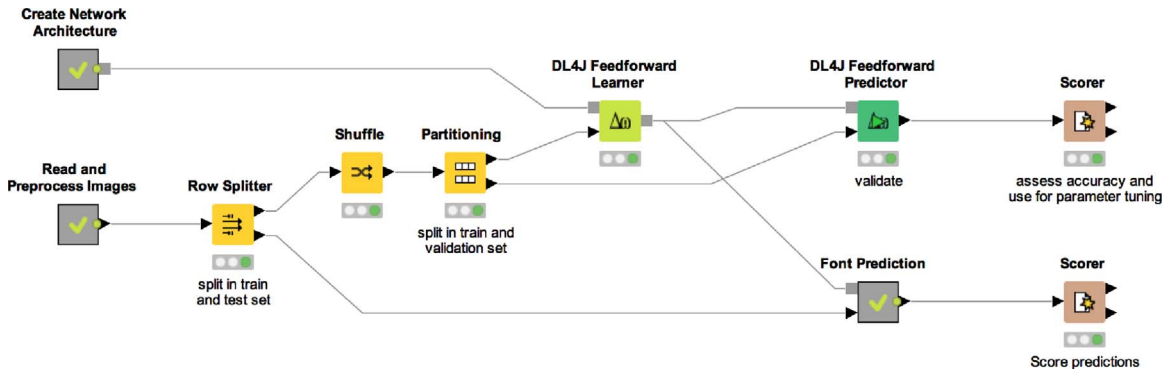

Fig. 2. Workflow for predicting the font in images of individual characters using Deeplearning4j and KNIME Image Processing nodes.

languages like Python, C#, and Java to allow the toolkit to be used in many different environments. The toolkit provides a broad range of functionality for working with molecules as both 2D and 3D objects. This includes support for conformation generation, chemical reactions, a range of chemical file formats, etc. RDKit has a large and active user community and has been integrated into a number of other open-source and commercial projects.

The first RDKit nodes for KNIME, developed in a collaboration between KNIME.com AG and the Novartis Institutes for BioMedical Research (NIBR) and released in 2011, provided a set of standard functionality to allow working with molecules in KNIME workflows. At the core of this was the addition of an RDKit molecule type that allows efficient information exchange between nodes. In the 6 years since the initial release, NIBR has continued to expand the set of nodes and adapt them as new features like Python integration and streaming execution have become available in KNIME Analytics Platform. Today the RDKit nodes support molecular operations running the gamut from parsing standard molecule formats to generating customized 2D renderings to conformational analysis and 3D molecule alignment. The RDKit KNIME nodes are distributed as part of the KNIME trusted community nodes and the source code is available in github.[2]

The core RDKit functionality used in the RDKit nodes is also being utilized in two other sets of community nodes: those from Vernalis[3] and the Erl Wood nodes from Eli Lilly.[4]

3.2. OpenMS

In recent years, mass spectrometry coupled to liquid chromatography (LC–MS) has become a vital part of experiments conducted in the scientific fields of proteomics and metabolomics. Separating proteins, peptides or metabolites in a chromatographic column over long periods of time, and acquisition of high-resolution mass spectra of the intact entities as well as their fragments results in a significant amount of data. Additionally, new types of instruments and experimental techniques are developed constantly and soon after applied in laboratories to gain insights on complex biological questions. The mass of data and its variety requires efficient and adaptable computational solutions to be processed. One type of software fulfilling these needs are workflow systems. They combine independent, reusable constituents into larger analysis workflows and are therefore able to address experiment-specific requirements.

A rich set of such building blocks for mass-spectrometry data analysis is provided by the standalone, cross-platform command line tools of The OpenMS Proteomics Pipeline (TOPP) (Kohlbacher et al., 2007) which are built on top of the OpenMS C++ library (Rost et al., 2016) (an open-source software solution and programming framework for computational MS). It features more than 150 tools executing a wide range of computational mass spectrometry related tasks. For instance, it provides tools for spectra filtering, identification, and quantification of peptides, proteins, or metabolites. External software, like database search engines (such as MASCOT Perkins et al., 1999, X!Tandem Craig and Beavis, 2004 or MSGF+ Kim et al., 2008), is tightly integrated and consistently used via adapters.

Initial workflow construction capabilities were added with the graphical user interface TOPPAS (TOPP Assistant) (Junker et al., 2012) in 2012. However, designed for fully automated processing, TOPPAS shows severe limitations in downstream analysis and visualization capabilities. These inadequacies led to efforts towards a seamless integration into a full-featured, well-established workflow system which was found in KNIME. To achieve this goal, two missing features had to be designed:

a) a thorough machine readable description of command-line tools, and
b) a method for automatic generation of KNIME nodes based on these descriptions.

Finally, the implementations of the above mentioned features are the following:

a) the common tool description (CTD) format,[5] an XML-based description of a tools function, inputs, outputs and parameters, and
b) the GenericWorkflowNodes project[6] that automatically generates source code for a KNIME node for any command line tool that provides such a CTD.

All tools provided by OpenMS naturally offer exporting their own interface as CTD for even more streamlined integration.

A major difference between regular KNIME nodes and generically generated OpenMS nodes is due to their origin from command line tools (potentially written in any language). Therefore they accept files instead of tables as input. To still allow interaction between them, we included a set of nodes to load the content of mass spectrometry data files (e.g., in the community-driven mzTab Griss et al., 2014 format) into KNIME tables. Every connection between OpenMS nodes is checked for matching file types and prevents connecting incompatible nodes. The parameters of the tools/nodes and their documentation are available via a generic interface for the known KNIME configuration dialog. To allow more complex workflows a set of standard KNIME nodes were rewritten to allow the input/output, looping, merging, or splitting files and file lists.

Ultimately, the integration significantly extends OpenMS' data

[2] https://github.com/rdkit/knime-rdkit.
[3] https://tech.knime.org/book/vernalis-nodes-for-knime-trusted-extension.
[4] https://tech.knime.org/community/erlwood.
[5] https://github.com/WorkflowConversion/CTDSchema.
[6] https://github.com/genericworkflownodes.

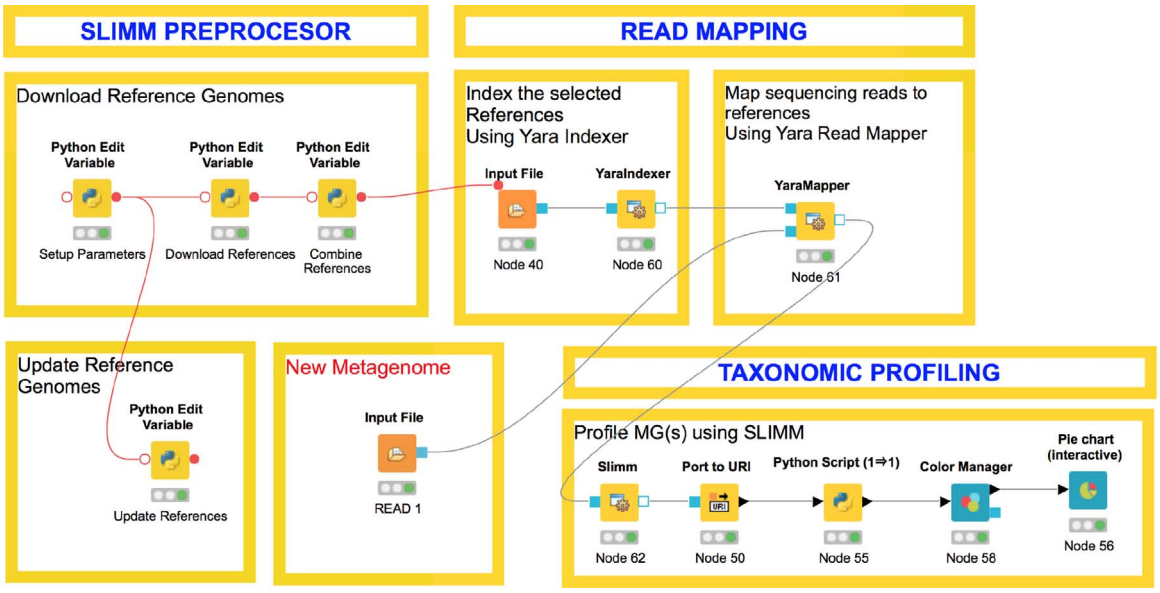

Fig. 3. Workflow for the identification and quantification of microorganisms within microbial communities using combination of SeqAn nodes and KNIME python and color nodes.

processing capabilities allowing for sophisticated downstream analysis and visualization. The user can build a single, automated workflow that processes the initial raw MS data, generates peptide, protein, or metabolite quantities and allows extracting biological knowledge leveraging KNIME's data-mining and visualization capabilities. Integrating with well-known R packages for proteomics data analysis (e.g., isobarBreitwieser et al., 2011, MSqRob Goeminne et al., 2016, or MSStats Choi et al., 2014) or by utilizing existing KNIME plugins for cheminformatics (e.g., to visualize chemical structures of metabolites) further extends the methodological repertoire of bioinformaticians.

Workflows, once created in KNIME, can also be run from the command line. This enables fully automated processing of large numbers of files and provided flow variables allow to control important parameters in the workflow. The same variables may alternatively be exposed through input nodes and presented in the user-friendly web interface of a (usually more powerful) KNIME Server, essentially providing "bioinformatics as a service" with automatic reports and digested data as results.

With KNIME being able to export complete, preconfigured workflows into self-contained ".knwf" files, also a large step towards supporting reusable research is made. Workflows can easily be shared with collaborators, uploaded to web archives, or otherwise be made accessible to the scientific community. Given an available copy of the OpenMS plugin with the same version (e.g., through sharing a Docker container) and the use of connector nodes from the (de.NBI/CIBI extension) to proteomics data stored on PRIDE (Vizcaino et al., 2016) this combination additionally ensures full reproducibility of the results.

3.3. SeqAn

In the field of computational biology sequence analysis is a fundamental research area focussing on different topics to expand the knowledge-base about causes and treatments of genetic disorders and other diseases. The rapidly evolving sequencing technologies enabled researches to conduct studies and analysis for entire genomes or even population of genomes from eucaryotes and procaryotes at higher resolutions. However, due to the enormous amount of data large-scale analysis algorithms that require considerable computing resources are becoming increasingly important and so do their implementations as efficient tools. SeqAn (Döring et al., 2008) is an efficient and generic template C++-library addressing these problems by providing state-of-the-art algorithms and data structure implementations in this field. In addition SeqAn implements various applications that can be used for different tasks for example to map reads, apply read error correction, conduct protein searches, run variant detection and many more. However, analysts are not interested in a single execution of one tool but design and execute entire pipelines using different tools for different tasks contained in the pipeline. Often they also require some downstream analysis steps, e.g. computing some statistics, generating reports and so on. Hence it was desirable to add SeqAn applications to the KNIME workflow engine, which offers many additional analysis and data mining features.

Similar to OpenMS we use the common tool description format, to provide an XML-based description of the command line arguments for each tool. A CTD file is used to generate a generic KNIME node from the GenericWorkflowNodes project, which is later deployed as a KNIME plugin and made available through KNIME's update mechanism.

In SeqAn we incorporated a CTD-writer into our argument parser, which provides an easy way to parse command line arguments and options, supporting file types, as well as various option types, such as *STRING* and *INTEGER* types, tuples, and many more. Thus, any program using SeqAn's argument parser, including all the applications developed in SeqAn itself, can write a CTD file and hence be integrated into KNIME. Since most applications developed in the bioinformatics community read files of special formats and standards containing either raw or preprocessed data and write the result to an output file of again various formats, we implemented special input/output file ports for the CTD generated KNIME nodes. These input/output file ports allow seamless integration of file based nodes, where the output file port of one node (application) can be connected to the input file port of a successor node (application) as long as the file formats are compatible. This significantly improves the usability when designing and executing pipelines as KNIME workflows with our applications. Users also benefit significantly from the ability to transfer the tool's help page into the KNIME node from the CTD. Thus the documentation of the node's arguments and options are available in the UI of KNIME.

With the integration of SeqAn tools into KNIME it became a lot easier to implement entire sequence analysis pipelines including preprocessing and post-processing steps that were otherwise mostly handmade and error-prone due to the usage of many unmaintained scripts. Moreover, the workflows can be executed head-less on a server or in the cloud and also may support large compute clusters for an even faster execution.

In Fig. 3 we show a simple workflow[7] for the identification and

quantification of microorganisms within microbial communities (Dadi et al., 2016). The workflow is composed of KNIME python nodes for some pre- and postprocessing steps and SeqAn nodes for read mapping and for applying the program SLIMM on the mapped reads in the taxonomic profiling section.

3.4. ImmunoNodes

ImmunoNodes (Schubert et al., 2017) is another KNIME integration that relies on the Generic KNIME Nodes framework to call command line tools from within a work-flow. The applications embedded in this plugin are written in Python and the FRED 2 (Schubert et al., 2016) framework and enable users to perform (neo-)epitope, cleavage site, and transporter associated with antigen processing (TAP) prediction, as well as human leukocyte antigen (HLA) genotyping, and epitope-based vaccine design. In order to make the plugin platform independent, the python scripts are wrapped in Docker images. These Docker images are then called by a special executor available in the Generic KNIME Nodes framework.

3.5. Image processing

Automated state-of-the-art microscopes allow the acquisition of high resolution, multidimensional images with only minimal user interaction required. This means they can generate a plethora of heterogeneous image datasets. Investigators often combine classical image processing techniques with leading-edge algorithms from the field of machine learning and data mining as well as established statistical analysis and visualization methods in order to solve the great diversity of image analysis problems. Typically, scientists have to transfer the data among different tools manually or come up with highly-customized scripts in order to automate their analysis process. Both solutions are often error-prone, inflexible, opaque and irreproducible; consequently, in many cases, previous work is shelved without reexamination, and is sometimes lost completely (Dietz and Berthold, 2016).

Similar problems have been identified previously in other scientific communities and the huge impact of KNIME Analytics Platform inspired these scientific communities to develop KNIME Image Processing (KNIP) (Dietz and Berthold, 2016), extending the KNIME Analytics Platform by integrating previously isolated tools and algorithms from bioimaging and beyond. In the spirit of integration, the core of KNIP is built on the same technologies that have driven the latest versions of the well-established image analysis software ImageJ2/Fiji (Rueden et al., 2017; Schindelin et al., 2012; Pietzsch et al., 2012; Linkert et al., 2010). Furthermore, collaborative efforts between KNIP, ImageJ2 (University of Wisconsin) and Fiji (MPI-CBG Dresden) drive the continuous development of these libraries. Based on this technical infrastructure, several other well-known open source projects, for example ImageJ2 and FIJI themselves, Ilastik (Sommer et al., 2011), CellProfiler (Carpenter et al., 2006), ClearVolume (Royer et al., 2015) and OMERO (Allan et al., 2012) have been integrated in KNIP. Most notably, integrating with ImageJ2 and FIJI allows scientists to easily turn ImageJ2 plugins into KNIME nodes, without having to be able to script or program a single line of code.

In order to further foster this “write once, run anywhere” framework, several independent projects collaborated closely in order to create *ImageJ-Ops*, an extensible Java framework for image processing algorithms. ImageJ-Ops allows image processing algorithms to be used within a wide range of scientific applications, particularly KNIME and ImageJ and consequently, users need not choose between those applications, but can take advantage of both worlds seamlessly.

The success of this workflow-based and integrative approach for the field of bioimage analysis has already been proven several times: a diversity of projects are using KNIME Image Processing in multiple application scenarios, ranging from medium to large-scale high-throughput bioimage analysis (Saha et al., 2013; Lodermeyer et al., 2013; Aligeti et al., 2014) to automated microscopy (Gunkel et al., 2014).

In conclusion, the KNIME Image Processing extensions not only enable scientists to easily mix-and-match image processing algorithms with tools from other domains (e.g. machine-learning), scripting languages (e.g. R or Python) or perform a cross-domain analysis using heterogenous data-types (e.g. molecules or sequences), they also open the doors for explorative design of bioimage analysis workflows and their application to process hundreds of thousands of images.

4. Other extensions

Apart from plugins that were developed specifically for data and life sciences, KNIME also provides extensions for other use cases and areas of research. Here we want to mention the text processing (Thiel, 2009) and the network mining (Thiel et al., 2009) plugins as well as the various ways in which data can be im- and exported into/from KNIME. The example workflows for the text and network mining extensions can also be found on the KNIME example server, which is available from every KNIME installation.

Research of basically all areas produces a huge amount of textual artifacts such as papers, books, Internet discussions, and lab reports. Extracting valuable information from those sources is usually a tedious task due to the sheer amount of text. Automated methods for information extractions can therefore be a useful tool for augmenting other data with information from the above mentioned sources. However, text is not an easy data source to tap, as it may contain ambiguities, sarcasm, complex and long sentence structures, rhetorical questions, and context sensitive meaning. Usually text processing follows a clear cut pipeline and is therefore well suited for integration into a workflow system. Tasks from this pipeline include:

- Inferring the document structure (title, abstracts, references, tables, etc.)
- Tagging words with a *Part-of-Speech Tagger*
- Inferring word stems
- Creating *Bags-of-Words*
- Filtering stop words that are irrelevant for the analysis
- Calculating the sentiment of a sentence or document

For all those and many more operations the KNIME text processing extension provides nodes that operate on a custom data format that can be embedded into standard KNIME tables. Fig. 4 depicts a workflow that loads documents retrieved from PubMed, performs preprocessing and then learns a model to distinguish between documents about AIDS in humans and cancer in mice.

Another research area – one that has received a surge of attention through the rise of social networks – is network mining. It is concerned with extracting useful information from a network of connected nodes to find substructures, patterns or nodes with many in- or outgoing connections. The KNIME plugin for network mining is built for handling large networks and provides nodes to create, generate, manipulate, analyze, and visualize those structures. An example for network mining can be found in Fig. 5. Here data about publications from PubMed[8] is parsed into a network and the most important authors for a given keyword are extracted.

KNIME's network mining extension also has an integration with the open source bioinformatics software platform Cytoscape,[9] which can be used to visualize molecular interaction networks and biological

[7] https://github.com/seqan/slimm.

[8] https://www.ncbi.nlm.nih.gov/pubmed/.

[9] http://www.cytoscape.org/.

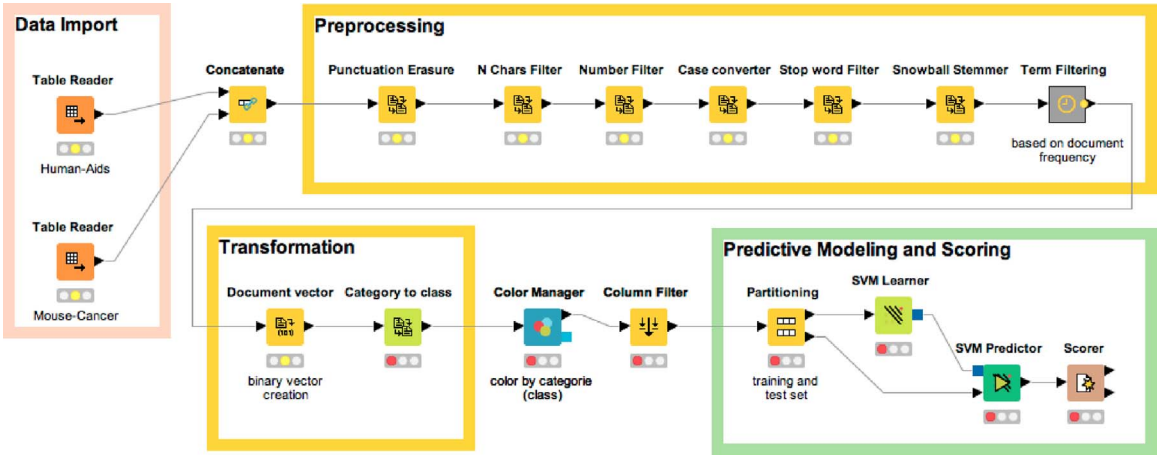

Fig. 4. A workflow using the text processing plugin to learn a model that can distinguish between documents about AIDS in humans and cancer in mice.

pathways. Installing the KNIME Connector plugin in Cytoscape enables users to exchange networks between the two tools.

An important feature for any data processing tool is the capability to read data from a variety of heterogeneous sources and also write results back to them. In KNIME data are imported using nodes that usually have no input port. Integrating a new data source is only a matter of implementing such a node and passing the read data to the next node. Most data importing nodes directly transform the data into KNIME's native table format so that so that they can be processed easily by the many other processing nodes KNIME offers.

Apart from the natively available file import nodes for various formats and the JDBC database connectors, there are many KNIME extensions that handle connections to more specialized data sources. One of these extensions, for example, can read data from MongoDB,[10] one of the leading NoSQL databases.

Especially important nowadays is the support for connections to cloud storage providers, as the huge amounts of data that scientists deal with are not usually stored locally. With the right extensions installed, KNIME can load files via FTP (File Transfer Protocol), Amazon S3 (Simple Storage Service) and Microsoft Azure Blob Store. A commercial extension provides access to data stored on HDFS (Hadoop Distributed File System).

Another source of data that scientists often tap into are web services. KNIME offers generic nodes for downloading data from such services, but also has dedicated nodes for the most important ones. The specialized nodes require less technical expertise from the user, as they provide easy to use configuration interfaces that are tailored to the service. For the life sciences KNIME provides specialized nodes for access to biodata and compound databases, such as PRIDE, PubChem, BARD and PeptideAtlas. Additionally there are nodes for drawing data from Twitter, Google and SPARQL web services.

5. Comparison with other workflow tools

There are many workflow systems apart from KNIME that are widely used in life science research. In this section we want to present some of them and, if available, list literature that reviews and compares them with KNIME and one another.

A quite extensive review of the six workflow management systems Discovery Net, Taverna, Triana, Kepler, Yawl, and BPEL can be found in Curcin and Ghanem (2008). Here, the authors analyze the different systems regarding their handling of data and control flow and provide a high-level framework for such comparisons.

The commercial tool Pipeline Pilot[11] provides a graphical interface for authoring scientific workflows. Originally the tool was meant to orchestrate the execution of cheminformatics tools, but has gained capabilities for other domains through a variety of extensions. Such plugins for Pipeline Pilot are grouped into so-called collections, of which many are suitable for use in the life sciences. Among others, Pipeline Pilot offers collections for cheminformatics, bioinformatics, imaging and statistics. For a comparison between Pipeline Pilot and KNIME we refer the reader to Warr (2012).

Another commonly used workflow tool for the life sciences is the open source software Galaxy (Goecks et al., 2010). This web-based tool allows users to start jobs on a computing backend from their browser, enabling multiple users to access a single instance over the network. The Galaxy system's strengths lie in its tight integration of tools and data sources from the life sciences, which are available in the Galaxy Tool Shed, a central repository for tools that are compatible with the workflow system. At the time of writing the tool shed contains 4807

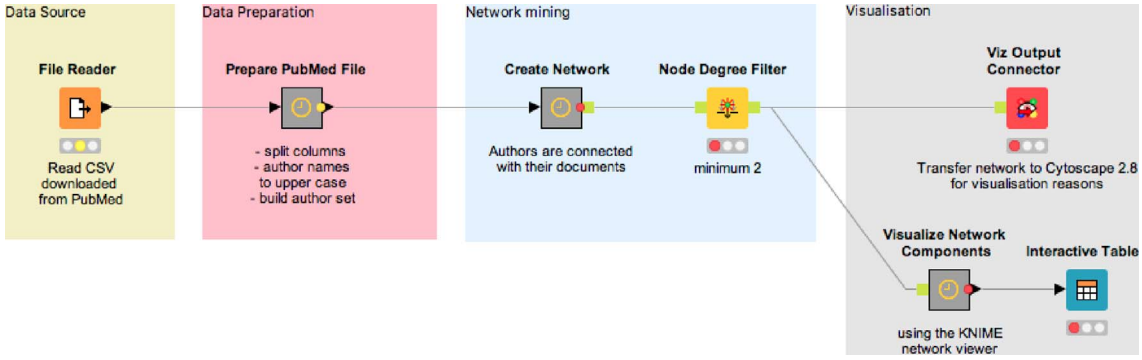

Fig. 5. A network mining workflow in KNIME. The workflow extracts documents for certain keywords from PubMed and finds the most important authors in that area.

tools, most of which are for domain specific analysis of life science data. OpenMS, SeqAn and RDKit, for example, are available for installation. Machine learning can be integrated into Galaxy via WEKA, but generally KNIME provides a wider variety of general purpose statistics and machine learning tools, which come bundled with every installation. While Galaxy can be used in a single user setting, it works best for collaboration when it is installed on a central server and accessed over the network, using the browser as an interface. KNIME provides a similar setup through the commercial KNIME Server, which provides sharing and remote and scheduled execution of workflows.

Another tool for designing and executing workflows is Taverna (Oinn et al., 2004). Though it was originally devised as a tool for building bioinformatics workflows, it has matured into a workflow management system for many domains and recently transitioned to be a project in the Apache Incubator.[12] Taverna consists of multiple parts: a graphical workbench, a command-line, a server and a player. First workflows are composed locally in the workbench, where the user has access to various processing nodes and data sources. Then the workflow can either be executed from the workbench directly, or saved in Taverna's workflow definition format SCUFL2. Workflows exported in this format can be executed from the Taverna Command-line or remotely on the Taverna Server. Once a workflow is deployed to the server, users can open them using the Taverna Player in order to run them interactively. The Taverna workbench comes with an integration to myExperiment,[13] a website for publishing and sharing scientific workflows. KNIME offers this integration as part of the de.NBI/CIBI plugin.[14]

The tools presented above are all used in various areas of the life sciences, but their main task is the orchestration of external tools that exchange files with each other. Natively, KNIME goes a different way by encouraging a deep tool integration that is compatible with KNIME's table format. With this approach data are embedded into table cells allowing for easy tool interoperability without the need for file conversions. The Generic KNIME Nodes framework that is used by OpenMS and SeqAn more closely resembles the tool orchestration paradigm of Galaxy and Taverna, though nodes for transforming data between both worlds are available.

In terms of downstream statistical analysis KNIME not only provides an integration of WEKA, like Galaxy also does, but additionally comes with many native tools for machine learning, statistics and visualization.

A tool that is more similar to KNIME is Orange (Demšar et al., 2013), a workflow tool that builds on the large foundation of Python libraries for data analysis and machine learning. Orange has a very easy to use user interface with large icons and lets users annotate their workflows with arrow pointers and text. The data processing nodes, called widgets in Orange, reside in a neatly arranged drawer on the left side of the workflow. The widgets Orange ships with are mainly for data import and preparation as well as machine learning and visualization. New widgets can be downloaded and installed separately. Among others, there exist extensions for bioinformatics, text mining, image processing and network mining. In total Orange has fewer nodes than KNIME, which may be an advantage for beginners who can easily obtain an overview of what is possible with the tool, making it easy to find the widget that is required.

In terms of workflow execution Orange takes a different approach than KNIME and the other tools. Upon insertion into the workflow, widgets are executed immediately, so that the user does not have to worry about starting the process or monitoring its execution. For quickly analyzing small data this is very useful, as results appear almost instantaneously, but when large amounts of data have to be processed the computation may cause the user interface to become unresponsive. The lack of a progress indicator may be confusing for the user in such cases.

In conclusion the choice for the right workflow tool depends on the task one wants to accomplish. Orchestrating the execution of many command line tools is a task for Galaxy, while an analysis of life science data with subsequent statistical analysis and visualization is best carried out in KNIME or Orange. Orange with its “ad-hoc” execution of nodes caters to scientists doing quick analyses on small amounts of data, while KNIME is built from the ground up for large tables and images. Noteworthy is that none of the mentioned tools provide image processing capabilities as extensive as those of the KNIME Image Processing plugin (KNIP). Orange, for example, which has an extension for image analysis, provides only a single widget that can transform images into feature vectors, which can then be processed further by other widgets.

6. Conclusion

Workflow systems as a platform for a multitude of tools provide the unifying framework for data transfer and flow control that custom scripts are often lacking. KNIME, as an open source software, focuses on integration and therefore allows third party developers to easily embed their tools and make them interoperable with each other, independent of their respective domain. We presented a slew of projects from the Life Sciences that have integrated their software into KNIME, giving the scientific community the opportunity to create pipelines that combine data retrieval, exploration, analysis, and visualization. These projects include software libraries for mass spectrometry, next generation sequencing, and image processing as well as tools for cheminformatics. But KNIME also integrates many tools that, even though they are not specifically designed for life sciences, provide important capabilities for data retrieval and visualization and statistical analysis. Various database connectors let users load their data from heterogeneous sources and scripting support enables scientists to quickly write data processing nodes in their favorite scripting language. The integration of WEKA as well as many custom implementations of common data mining algorithms complete the scientists’ toolbox. Compared to other tools KNIME focuses on a deeper integration of tools and tries to manage the data that flows in the workflow by itself. Tools like Galaxy and Taverna, on the other hand, rather orchestrate command line tools that exchange files. Orange is very similar to KNIME in that it has extensive machine learning capabilities, but focuses more on the analysis of smaller data sets. We conclude that there are workflow tools for a variety of different use cases and that it is the scientists task to choose the tool that fits the problem at hand best. While there are certainly overlaps, each tool excels at its intended purpose.

Funding

This work was supported by the German Network for Bioinformatics (de.NBI) [grant number 031A535C] and the Konstanz Research School Chemical Biology.

References

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2007. KNIME: the Konstanz information miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria ISBN:3-900051-07-0.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, New York ISBN:0-387-95457-0.

from Jed Wing, M.K.C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., 2016. caret: Classification and Regression Training,

[10] http://www.mongodb.com.
[11] http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/.
[12] http://incubator.apache.org/projects/taverna.html.
[13] http://www.myexperiment.org.
[14] https://tech.knime.org/denbicibi-contributions.

R Package Version 6.0-64. http://CRAN.R-project.org/package=caret.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
Dowle, M., Srinivasan, A., Short, T., 2015. S.L. with contributions from R Saporta. E. Antonyan, data.table: Extension of Data.frame, r package version 1.9.6.
Jones, E., Oliphant, T., Peterson, P., et al., 2001. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/http://www.scipy.org/ (accessed 09.11.16). http://www.scipy.org/.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11 (1), 10–18.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
RDKit: Open-source cheminformatics, http://www.rdkit.org (accessed 11.04.13).
Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Sturm, M., 2007. Topp – the openms proteomics pipeline. Bioinformatics 23 (2), e191–e197.
Rost, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., Wojnar, D., Wolski, W.E., Schilling, O., Choudhary, J.S., Malmstrom, L., Aebersold, R., Reinert, K., Kohlbacher, O., 2016. Openms: a flexible open-source software platform for mass spectrometry data analysis. Nat. Meth. 13 (9), 741–748. http://dx.doi.org/10.1038/nmeth.3959.
Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20 (18), 3551–3567. http://dx.doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2. http://www.ncbi.nlm.nih.gov/pubmed/10612281.
Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics (Oxford, England) 20 (9), 1466–1467. http://dx.doi.org/10.1093/bioinformatics/bth092. http://www.ncbi.nlm.nih.gov/pubmed/14976030.
Kim, S., Gupta, N., Pevzner, P.A., 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J. Proteome Res. 7 (8), 3354–3363. http://dx.doi.org/10.1021/pr8001244. http://www.pubmedcentral.nih.gov/articlerender.fcgi?.artid=2689316&tool=pmcentrez&rendertype=abstract.
Junker, J., Bielow, C., Bertsch, A., Sturm, M., Reinert, K., Kohlbacher, O., 2012. Toppas: a graphical workflow editor for the analysis of high-throughput proteomics data. J. Proteome Res. 11 (7), 3914–3920.
Griss, J., Jones, A.R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G.G., Salek, R.M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.-W., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaino, J.A., Hermjakob, H., 2014. The mzTab data exchange format: communicating ms-based proteomics and metabolomics experimental results to a wider audience. Mol. Cell. Proteomics. http://dx.doi.org/10.1074/mcp.O113.036681.
Breitwieser, F.P., Müller, A., Dayon, L., Köcher, T., Hainard, A., Pichler, P., Schmidt-Erfurth, U., Superti-Furga, G., Sanchez, J.-C., Mechtler, K., Bennett, K.L., Colinge, J., 2011. General statistical modeling of data from protein relative expression isobaric tags. J. Proteome Res. 10 (6), 2758–2766. http://dx.doi.org/10.1021/pr1012784. http://www.ncbi.nlm.nih.gov/pubmed/21526793.
Goeminne, L.J.E., Gevaert, K., Clement, L., 2016. Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. Molec. Cell. Proteomics: MCP 15 (2), 657–668. http://dx.doi.org/10.1074/mcp.M115.055897. http://www.ncbi.nlm.nih.gov/pubmed/26566788.
Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., Vitek, O., 2014. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30 (17), 2524–2526. http://dx.doi.org/10.1093/bioinformatics/btu305. http://www.ncbi.nlm.nih.gov/pubmed/24794931.
Vizcaino, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.W., Wang, R., Hermjakob, H., 2016. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44 (D1). http://dx.doi.org/10.1093/nar/gkv1145.
Döring, A., Weese, D., Rausch, T., Reinert, K., 2008. SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics 9 (1), 11. http://dx.doi.org/10.1186/1471-2105-9-11. http://www.ncbi.nlm.nih.gov/pubmed/18184432.
Dadi, T.H., Renard, B.Y., Wieler, L.H., Semmler, T., Reinert, K., 2016. SLIMM: Species Level Identification of Microorganisms from Metagenomes.
Schubert, B., de la Garza, L., Mohr, C., Walzer, M., Kohlbacher, O., 2017. ImmunoNodes-graphical development of complex immunoinformatics workflows. BMC Bioinformatics 18 (1), 242.
Schubert, B., Walzer, M., Brachvogel, H.-P., Szolek, A., Mohr, C., Kohlbacher, O., 2016. FRED 2: an immunoinformatics framework for Python. Bioinformatics btw113.
Dietz, C., Berthold, M.R., 2016. KNIME for open-source bioimage analysis: a tutorial. Focus on Bio-Image Informatics. Springerpp. 179–197.
Rueden, C.T., Schindelin, J., Hiner, M.C., DeZonia, B.E., Walter, A.E., Eliceiri, K.W., 2017. ImageJ2: ImageJ for the Next Generation of Scientific Image Data. arXiv:1701.05940.
Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9 (7), 676–682.
Pietzsch, T., Preibisch, S., Tomančák, P., Saalfeld, S., 2012. ImgLib2?.generic image processing in Java. Bioinformatics 28 (22), 3009–3011.
Linkert, M., Rueden, C.T., Allan, C., Burel, J.-M., Moore, W., Patterson, A., Loranger, B., Moore, J., Neves, C., MacDonald, D., et al., 2010. Metadata matters: access to image data in the real world. J. Cell Biol. 189 (5), 777–782.
Sommer, C., Straehle, C., Koethe, U., Hamprecht, F.A., 2011. Ilastik: interactive learning and segmentation toolkit. 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEEpp. 230–233.
Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al., 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 7 (10), R100.
Royer, L.A., Weigert, M., Günther, U., Maghelli, N., Jug, F., Sbalzarini, I.F., Myers, E.W., 2015. ClearVolume: open-source live 3D visualization for light-sheet microscopy. Nat. Methods 12 (6), 480–481.
Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W.J., Neves, C., Patterson, A., et al., 2012. OMERO: flexible, model-driven data management for experimental biology. Nat. Methods 9 (3), 245–253.
Saha, A.K., Kappes, F., Mundade, A., Deutzmann, A., Rosmarin, D.M., Legendre, M., Chatain, N., Al-Obaidi, Z., Adams, B.S., Ploegh, H.L., Ferrando-May, E., Mor-Vaknin, N., Markovitz, D.M., 2013. Intercellular trafficking of the nuclear oncoprotein DEK. Proc. Natl. Acad. Sci. U. S. A. 110 (17), 6847–6852. http://dx.doi.org/10.1073/pnas.1220751110. arXiv:http://www.pnas.org/content/110/17/6847.full.pdf+html, http://www.pnas.org/content/110/17/6847.abstract.
Lodermeyer, V., Suhr, K., Schrott, N., Kolbe, C., Stürzel, C.M., Krnavek, D., Münch, J., Dietz, C., Waldmann, T., Kirchhoff, F., Goffinet, C., 2013. 90K, an interferon-stimulated gene product, reduces the infectivity of HIV-1. Retrovirology.
Aligeti, M., Behrens, R.T., Pocock, G.M., Schindelin, J., Dietz, C., Eliceiri, K.W., Swanson, C.M., Malim, M.H., Ahlquist, P., Sherer, N.M., 2014. Cooperativity among rev-Associated nuclear Export signals regulates HIV-1 gene expression and is a determinant of virus species tropism. J. Virol. 88 (24), 14207–14221.
Gunkel, M., Flottmann, B., Heilemann, M., Reymann, J., Erfle, H., 2014. Integrated and correlative high-throughput and super-resolution microscopy. Histochem. Cell Biol. 141 (6), 597–603. http://dx.doi.org/10.1007/s00418-014-1209-y.
Thiel, K., 2009. The KNIME Text Processing Plugin.
Thiel, K., Kötter, T., Berthold, D.M., Silipo, D.R., Winters, P., 2009. Creating Usable Customer Intelligence from Social Media Data: Network Analytics Meets Text Mining.
Curcin, V., Ghanem, M., 2008. Scientific workflow systems-can one size fit all? Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, IEEE 1–9.
Warr, W.A., 2012. Scientific workflow systems: pipeline pilot and knime. J. Comput.-Aid. Molec. Des. 26 (7), 801–804.
Goecks, J., Nekrutenko, A., Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11 (8), R86.
Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., et al., 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20 (17), 3045–3054.
Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B., 2013. Orange: data mining toolbox in python. J. Mach. Learn. Res. 14, 2349–2353.