3 Workflow Systems

Within the last few years a large number of tools and softwares dealing with different computational problems related to HCS have been developed. Incorporating third party or new tools into existing frameworks needs a flexible, modular and customizable workflow framework. Workflow (Pipeline) systems could become crucial for enabling HCS researchers doing large scale experiments to deal with this data explosion. The workflow is termed abstract in that it is not yet fully functional but the actual components are in place and in the requisite order. In general, workflow systems concentrate on the creation of abstract process workflows to which data can be applied when the design process is complete. In contrast, workflow systems in the life sciences domain are often based on a data-flow model, due to the data-centric and data-driven nature of many scientific analyses. A comprehensive understanding of biological phenomena can be achieved only through the integration of all available biological information and different data analysis tools and applications. In general, an ideal workflow system in HCS can integrate nearly all standard tools and software. For example, for an HCS using small molecules, the workflow system must be able to integrate different image processing software and data mining toolkits with flexibility. The possibility that any single software covers all possible domains and data models is nearly zero. No one vendor or source can provide all the tools needed by HCS informatics. So it is suggested that one uses specialized tools from specialized sources. Also not all softwares components can be integrated with all workflow systems.

Workflow environment helps also HCS researchers to perform the integration themselves without involving of any programming. A workflow system allows the construction of complex in silico experiments in the form of workflows and data pipelines. Data pipelining is a relatively simple concept. Visual representation of the workflow process logic is generally carried out using a Graphical User Interface where different types of nodes (data transformation point) or software components are available for connection through edges or pipes that define the workflow process. Graphical User Interfaces provide drag and drop utilities for creating an abstract workflow, also known as "visual programming". The anatomy of a workflow node or component (Fig. 3) is basically defined by three parameters: input metadata, transformation rules, algorithms or user parameters and output metadata. Nodes can be plugged together only if the output of one, previous (set of) node(s) represents the mandatory input requirements of the following node. Thus, the essential description of a node actually comprises only in-and output that are described fully in terms of data types and their semantics. The user can create workflows using any combination of the available tools, readers, writers or database connections in workflow system by dragging/dropping and linking graphical icons. The component properties are best described by the input metadata, output metadata and user defined parameters or transformation rules. The input ports can be constrained to only accept data of a specific type such as those provided by another component. An HCS workflow design is best carried out in phases. In the first phase, a conceptual workflow is generated. A conceptual workflow, as the name suggests, is a sequential arrangement of different components that the user may require to accomplish the given task. It is possible that some of those steps may in turn be composed of several sub components. The next phase converts the conceptual workflow into an abstract workflow by performing a visual drag and drop of the individual components that were figured to be a part of the workflow in the first phase. The workflow is termed abstract in that it is not yet fully functional but the actual components are in place and in the requisite order. In general, workflow systems concentrate on the creation of abstract process workflows to which data can be applied when the design process is complete. HCS screening workflows are based on a dataflow which integrate most of the available, standard software tools (either commercial or public domain) along with different classes of programmable toolkits. As an example, Figure 3 shows a workflow designed to be run by the HCDC-KNIME Workflow Management System (http://hcdc.ethz.ch). This workflow is used by HCS facilities. It obtains RNAi from databases, annotates them, make dilutions steps, barcode handling, split volume. In this case, the tasks, also known as steps, nodes, activities, processors or components, represent either the invocation of a remote Web service (the databases), or the execution of a local recalculation. Data-flows along data links from the outputs of a task to the inputs of another, is prepared according to a pre-defined graph topology. The workflow defines how the output produced by one task is to be consumed by a subsequent task, a feature referred to as orchestration of a flow of data.

Any computational component or node has data inputs and data outputs. Data pipelining views these nodes as being connected together by 'pipes' through which the data flows (Figure 4).

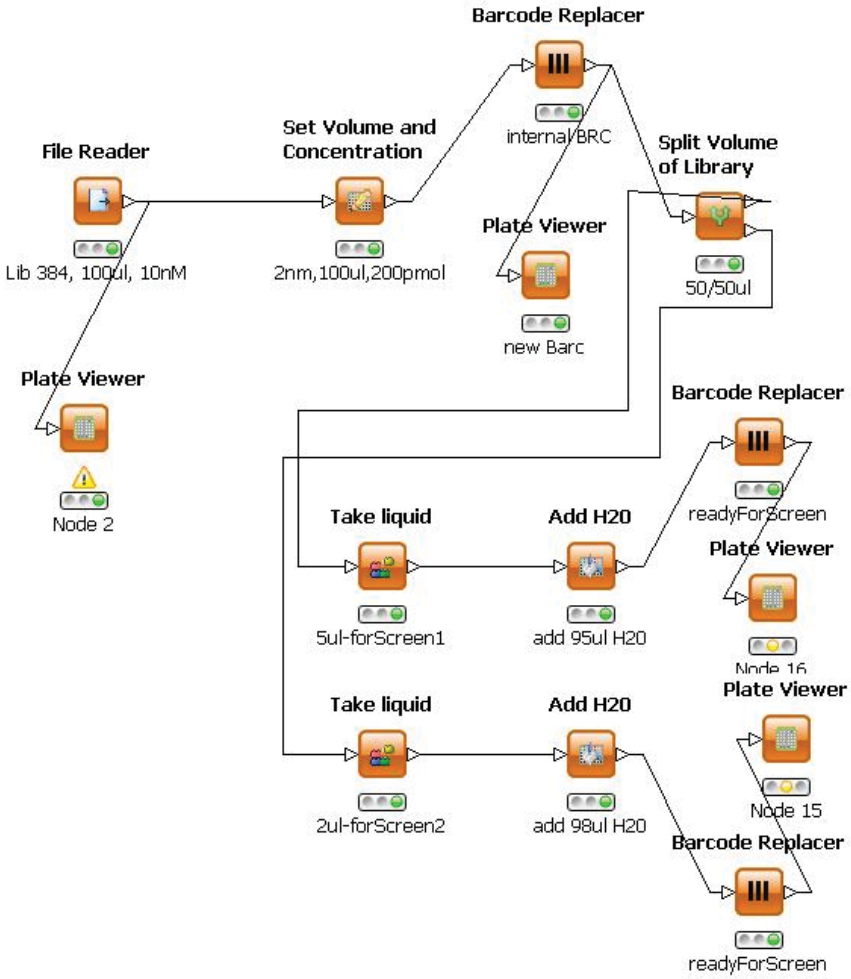

Fig 3: A HCDC-KNIME workflow that simulates the library handling process of multiwell plates including barcode handling

Workflow technology is a generic mechanism to integrate diverse types of available resources (databases, microscopes, servers, software applications and different services) which facilitates data exchange within screening environment. Users without programming skill can easily incorporate and access diverse instruments, image processing tools and produced data to develop their own screening workflow for analysis. In this section, we will discuss the usage of existing workflow systems in HCS and the trends in applications of workflow based systems.

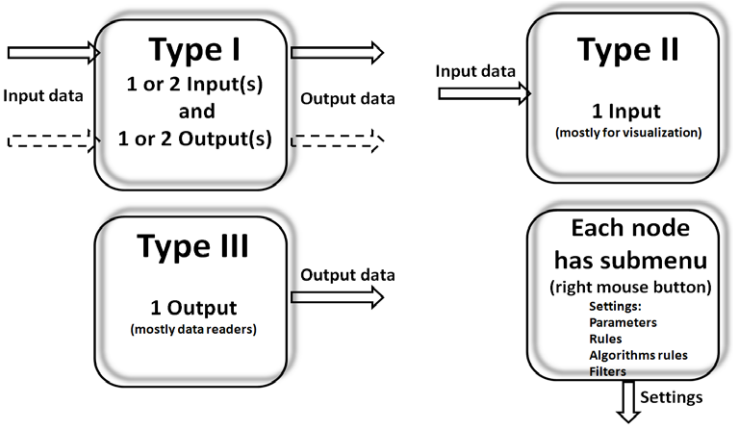

Figure 4: General concept of a pipeline node. The component properties are described by the input metadata, output metadata and user defined parameters or transformation rules. The input and output ports can have one or more incoming or outgoing metadata or images.

3.1 Why is a workflow system important?

Many free and commercial software packages are now available to analyse HCS data sets using statistical method or classification, although it is still difficult to find a single off-the-shelf software package that answers all the questions of HCS analysis. Statistical open source software packages such as BioConductor (www.bioconductor.org) provide large collections of methods suitable for HCS data analysis.

However, their command-line usage can be too demanding for users without adequate computer knowledge. As an alternative, software packages where users can upload their data and receive their processed results are becoming increasingly common: Weka[25], CellAnalyzer[4], CellHTS[3], TreeView[21] have all been published within the last year. Unfortunately, these services often allow only limited freedom in the choice and arrangement of processing steps. Other, more flexible tools, such as Eclipse[6], KNIME[13], JOpera[2], operate either stand-alone or require considerable computer knowledge and extra software to run through the web. In order to make use of the vast variety of data analysis methods around, it is essential that such an environment is easy and intuitive to use, allows for quick and interactive changes to the analysis process and enables the user to visually explore the results. To meet these challenges data pipelining environments have gathered incredible momentum over the past years. These environments allow the user to visually assemble and adapt the analysis flow from standardized building blocks, which are then connected through pipes carrying data or models. An additional advantage of these systems is the intuitive, graphical way to document what has been done.

In a workflow controlled data pipeline, as the data flows, it is transformed and raw data is analyzed to become information and the collected information gives rise to knowledge. The concept of workflow is not new and it has been used by many organizations, over the years, to improve productivity and increase efficiency. A workflow system is highly flexible and can accommodate any changes or updates whenever new or modified data and corresponding analytical tools become available.

3.2 Visualization in workflow systems

Visualization tools are one type of workflow systems that provide a quick and effective means to interrogate HCS data and images stored in a secure repository. Users want to view the data, share it with colleagues, and compare results. Visualization tools in workflow system should provide powerful search and navigation tools to rapidly locate plate, well, cell, and image data. Rich search functions should be available to find data based on various metadata and derived data parameters (e.g., user name, dates/times, assay type, features, and so on). The most basic form of any HCS data visualization node should provide interactive tools for reviewing data with drill-down capabilities from the plate, well, and cell level together with links to images, and any graphical image overlays. Various forms of viewing the data should be provided including tables/spreadsheets and graphs (bar charts, scatter plots, and so on). Various views should also be provided for different types of users (e.g., managers, scientists, operators, IT personnel, and so on). Capabilities should be provided for comparing data within a plate, across plates, and so on. Additional capabilities should also be provided for generating statistics on groups of data (e.g., groups of wells, cells, and so on). The data should be displayed in ways that allow the user to explore patterns and recognize patterns and outliers. Users want to be able to save their analyses and visualizations as well as build reports and save these. Making annotations on the data is also very important. Common uses for visualization in HCS include assessing the quality of the dataset (e.g., identifying outliers and false positives), and identifying hits. There are many possibilities for visualization of HCS data and one important visualizer is a plate viewer.

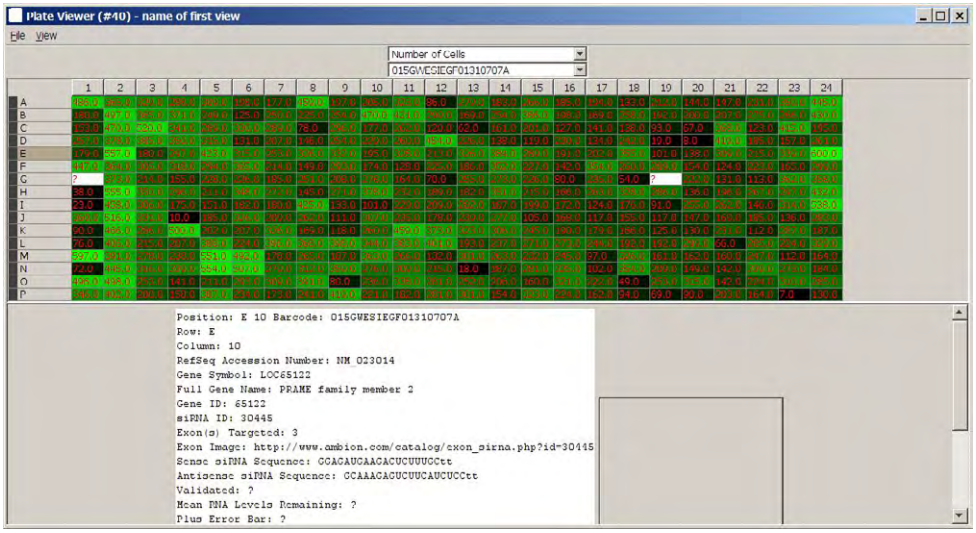

Fig 5: Plate viewer plug-in. Visualization of image processing parameters in a heatmap with access to library metadata.

Plate Viewer (PV) guarantees the identification of library and well position of a specific compound on a plate. The history of the location of each compound in the screen, run and replicate along with reformatting information are recorded and reconstructed by PV. Within the GUI the user may select the library, plate and if desired, compounds data derived from a specific 96, 384 or 1536-well plate. Once a plate is selected, a window is opened in a plate viewer that provides functions for easy navigation within the plate that helps extracting comprehensive information about particular compounds (Figure 5).

3.3 Architecture of workflow systems

The design of a typical architecture of workflow system is based mostly on plugin framework. The entire application is a functional set of nodes, working together. For example, a plug-in for opening and processing HCS files (library, numeric results and images) in HCDC-KNIME was developed within the KNIME environment. All those open source components (Eclipse environment, KNIME, R-Project, Weka and ImageJ) were chosen for their platform-independence, simplicity, and portability. The pipeline model describes the exact behaviour of the workflow when it is executed. The nodes are usually designed with the following main principles:

- *Resource type for primary data:* The source of data can be collection of high level images familiar to the user or single image. Software should support as many as possible image types.
- *Computation*: Dataflow pipelines dictate that each process be executed as soon as its input data are available. Node processes that have no data dependencies amongst each other can be executed concurrently. They are used for analysis pipelines, data capture, integrating data from different sources, and populating scientific models or data warehouses. Control flows directly dictate the flow of the process execution, using loops, decision points etc.

- *Interactivity*: Node execution could be fully automatic or interactively steered by the user. Data flows are combined by a simple drag&drop process from a variety of processing units. Customized applications can be modeled through individual data subpipelines.
- *Adaptivity*: The nodes and workflow design or instantiation can be dynamically adapted "in flight" by the user or by automatically reacting to changed environmental circumstances.
- *Modularity*: Processing units and containers should not depend on each other in order to enable an easy distribution of computation and allow for independent development of different image processing algorithms.
- *Easy expandability*: There should be easy ways to add new hardware (e.g. microscope), data analysis, or image processing software nodes or views. The distribution of new item should be easy, through a simple plug-in mechanism without the need for complicated install/ reinstall procedures.

Figure 6 shows in a schematic way an example of an HCS data analysis flow and the corresponding nodes used in the HCDC-KNIME workflow system.

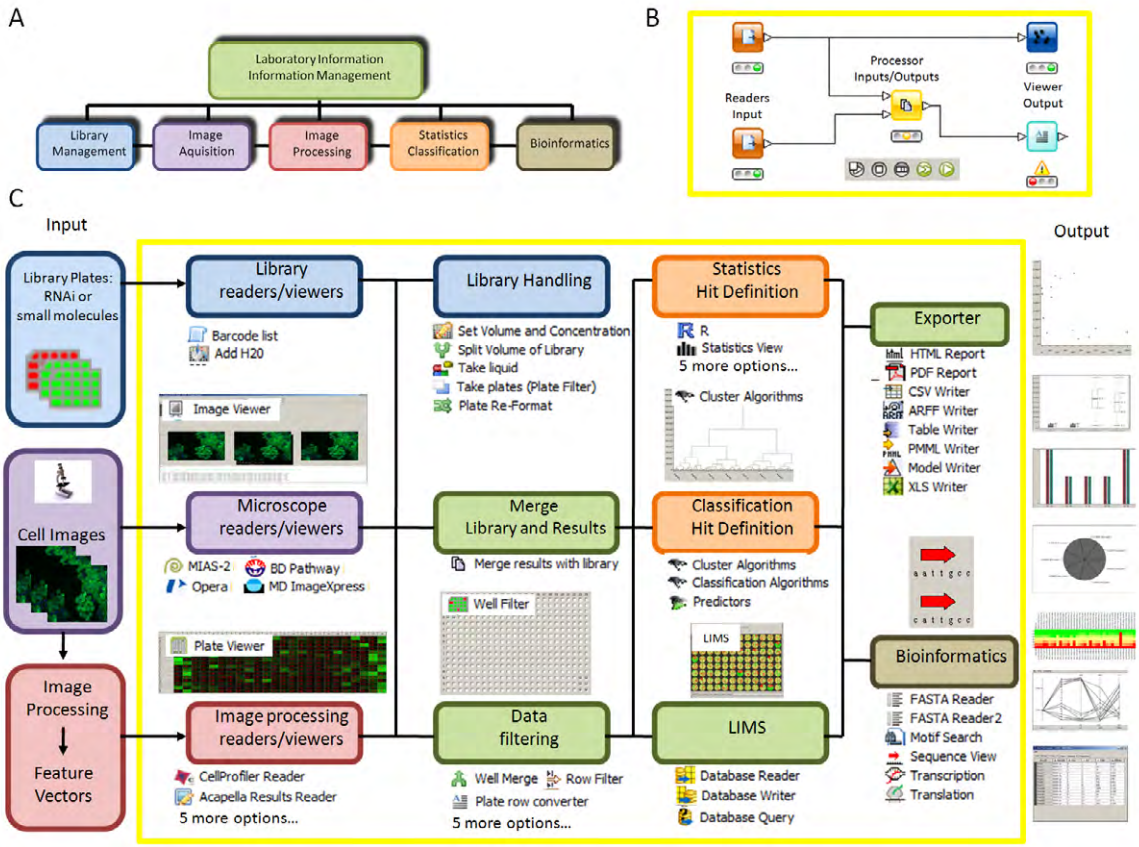

Fig 6: HCDC Platform. **a**: Informatics elements behind High Content Screening. **b**: Illustration of a workflow environment with nodes managing the data flow. **c**: Summary of some functionality of HCDC.

3.4 Public Domain Workflow Systems

The choice of tools optimized for HCS from entire collection of open-source workflow systems in the life sciences domain is limited. Open-source workflow systems provide major advantages for an the academic environment, not just because they are free of license charges, but also because open-source workflow systems are based on community models of development in which people from diverse background actively contribute to the application. Interestingly, many commercial life sciences workflow products make heavy use of open source and publicly available programs for pre- and post-processing analysis of screening data using packages like R-Project, CellHTS, Weka, BioJava, BioPerl, Chemistry Development Kit (CDK), EMBOSS, etc. HCDC-KNIME (http://hcdc.ethz.ch, http://knime.org) from the ETH Zurich and University of Konstanz is based on the Eclipse platform provides an excellent data mining platform for drug discovery informatics, bioinformatics and chemistry research. HCDC-KNIME workflow recently released specific nodes for HCS including library handling, quality controls, connection to microscopes, image processing tools, plate visualizations. For small molecule screening the tool already includes CDK nodes and other plug-ins to incorporate existing data analysis tools, such as Weka, the statistical toolkit R, Python scripting and JFreeChart. Tripos Inc. and Schroedinger support KNIME for chemoinformatics and drug discovery nodes or plug-ins. For chemoinformatics and QSAR studies HCDC-KNIME includes nodes that can be used to visualize and transform molecular structures, compute QSAR descriptors and molecular properties, generated fingerprint(s), perform data mining, implement machine learning algorithms (Support Vector Machines, Regression and Bayesian Modeling, Principal Component Analysis), and search for functional groups (substructure and similarity searching). There are other open-source workflow systems in the life sciences domain which can be used in HCS but they needs optimization and specific node development which will support screening standards, formats, hardware, third-parties software. A list of other available workflow systems for life science which can be used in post-processing or used to find molecule correlation between other large scale technologies is provided below:

- Pegasus (Planning for Execution in Grids[6]) is a workflow mapping engine which maps complex scientific workflows onto distributed resources. The Genome Analysis and Database Update (GADU) system, uses Pegasus to perform high-throughput analysis and annotation of the genomics information.
- Core Kepler tool has General Atomic and Molecular Electronic Structure System (GAMESS) as an ab initio quantum chemistry package.
- Kepler[15] provides full support for computational chemistry (nodes for Babel, OpenBabel, and GAMESS) and related workflow for statistical analysis.
- Pegasys[23] developed at UBC provides a specialized workflow management for high-throughput sequence data analysis and annotation. Pegasys can incorporate new tools into existing frameworks.

- The DiscoveryNet[19] platform is a system that integrates bioinformatic tools based on grid computing technologies. Applications of DiscoveryNet are reported for high throughput genomics, proteomics, chemoinformatics, large-scale genotyping data analysis, realtime drug resistance studies and integrative life science analysis.
- SOMA workflow[14] is used to handle different molecular modeling problems related to computer-aided drug design processes developed at genomic biology techniques such as microarrays with bioinformatics tools such as BLAST to identify and characterize eukaryotic promoters.
- myGrid[24] project is a tool for developing semantically enabled grid middleware for supporting bioinformatics and drug discovery applications and, is regarded as the most powerful workflow system.
- Taverna[18] which addresses problems beyond the capabilities of the present system to improve many areas including Data flow centric model, Scalability and Data streaming. Taverna includes services based on SOAP, BioMoby[26], Biomart[8], Soaplab[22], SeqHound[16] and R for numerical analysis.
- Wildfire[20] is a distributed, grid-enabled workflow construction and execution environment developed at the Bioinformatics Institute (A*STAR).
- Biopipe[12] is a workflow framework based on BioPerl which also allows for execution of workflows across clusters.

3.5 Commercial Workflow Systems

- ChemSense is a commercial package which provides high range of chemoinformatics solutions for HCS using small molecules ranging from the analysis and visualization of chemical libraries to the development of combinatorial chemistry libraries, and includes a wide range of QSAR, ADME-Tox prediction, molecular modeling and evaluation methods.
- With their Open Workflow Partner Network, InforSense provides the best-of-breed tools for specific scientific analytic needs.
- Accelerys Pipeline Pilot [11]is one of the very first workflow systems in life sciences optimized for HCS. Pipeline Pilot is chemically intelligent and possesses a robust and highly scalable environment that can run on a multiprocessing environment. Pipeline Pilot is widely used to process High Content Screening and drug discovery data and it comes with specialized solutions for computational chemistry, chemoinformatics and bioinformatics. Pipeline Pilot covers chemistry (compound library acquisition, combinatorial library design, molecular property calculators, filters, and manipulators), ADME/Tox, Decision Trees, Modeling, R Statistics, Reporting, sSequence Analysis, BioMining, Text Analytics and Integration Collection (flexible mechanisms to link external applications and databases).

Below is a list of other commercial available workflow systems for life sciences which can be used in a post-processing or used to find molecule correlation between other large scale technologies:

- BioLog's BioLib is an open architecture Informatics System that creates a unique set of drug discovery IT tools. BioLib covers Protein Modeling, Small Molecules and Peptide Analysis, Database Access, Sequence Analysis and Data Mining.
- UeberTool is a software system for the integration and analysis of molecular biological data with over 200 types of bioinformatics methods. It also enables access to public biological databases and proprietary data including all UeberTool results.

Definitely InforSense KDE and SciTegic's Pipeline Pilot are state of the art workflow systems widely used in HCS operation and applied in academic and non-academic organizations. Table 1 summarizes all available workflow systems (open-source and commercial) used in life science data analysis.

Software	License type	Vendor and URL	Features	Integration with other software
HCDC-KNIME	Open source	LMC, ETH Zurich http://hcdc.ethz.ch University Konstanz, http://knime.org	• Provides interactive views of data and models • Based on the Eclipse platform with extensible modular API • Modular data exploration platform • Plate Viewer • Library Handling • QualityConrol • Classification • Statistics • Pattern recognition • Library Investigation • HeatMaps • Molecular structure • Image processing • LIMS integration	CellProfiler Acapella Extrenal tool execution MD Micro microscope BD Pathway microscope Opera microscope MAIA Scientific microscope Matlab R-Project Weka Java API
myGrid	Open Source	UK e-Science http://www.mygrid.org.uk/	• High level, knowledge-enabled middleware based on web services to support personalized *in silico* experiments in bioinformatics on the grid • creation, discovery and enactment form a central feature of myGrid services	

Software	License type	Vendor and URL	Features	Integration with other software
Taverna	Open Source	EBI http://taverna.sourceforge.net	• Built-in support for web services, local Java functions, BioMoby, and Soaplab • workflow language (XScufl)	
MIGenAS	Open Source	Max-Planck-Society http://www.migenas.org/	• Integrated bioinformatics workflow engine for web-based sequence analysis • Focused on research with microbial genomes	
Kepler	Open Source	UC Berkeley http://kepler-project.org/	• Scientific workflow system built on top of the Ptolemy II system • XML based workflow definition – MoML • Actor prototyping tool	
GeneBeans	Open Source	UNC Wilmington http://www.uncw.edu/csc/bioinformatics/	• Uses a three-layer architecture • An engine, with Graphical User Interface, that models bioinformatics queries as dataflow graphs Discovery (Net Imperial College London), http://ex.doc.ic.ac.uk/new/ • System is a middleware that allows service developers to integrate tools based on existing and emerging grid standards • Supports high throughput genomics, proteomics and chemoinformatics • Uses discovery process mark-up language (DPML)	

Software	License type	Vendor and URL	Features	Integration with other software
Pegasys	Open Source	UBC Bioinformatics Centre http://bioinformatics.ubc.ca/pegasys/	• High-throughput sequence data analysis workflow • Tools for pair-wise and multiple sequence alignment, gene prediction, RNA gene detection, masking repetitive sequences in genomic DNA • Easy integration with Atlas biological data warehouse and its API ProGenGrid (University of Lecce), http://datadog.unile.it/progen • Service Oriented Architecture (SOA) • Provides services for drug discovery, access to distributed data and data sharing • Uses gSOAP Toolkit for web services and Globus Toolkit as grid • Middleware	
Wildfire	Open Source	A*STAR http://wildfire.bii.a-star.edu.sg/	• Distributed, grid-enabled workflow construction and execution • Borrows user interface features from Jemboss • Uses GEL as underlying workflow execution engine Orange (University of Ljubljana), http://www.ailab.si/orange/ • Component-based framework • Seamless integration within Python • Components for Functional Genomics	
Biopipe	Open Source	OBF http://www.biopipe.org	• Collection of Perl modules for constructing workflows from BioPerl applications	
Pegasus	Open Source	USC http://pegasus.isi.edu/	• Provide abstract workflow and maps it to the available grid resources • Supports a deferred mode • Well-defined APIs and clients	

Software	License type	Vendor and URL	Features	Integration with other software
Triana	Open Source	Cardiff University http://www.trianacode.org	• Part of GridOneD project for creating Java middleware for grid applications • Pluggable architecture • Peer-to-peer implementation based on the Sun's JXTA protocols	
WsBAW & BioWBI	Open Source	IBM http://www.alphaworks.ibm.com/tech/wsbaw	• WsBAW is Java client application through which users are able to send batch requests to a specific bioinformatics workflow execution engine BioWBI, by using a web service AdaptFlow (University of Leipzig), http://informatik.uni-leipzig.de • Rule-based dynamic workflow adaptation based consultation system • Supports the handling of the complex trial therapy processes	
BioAgent	Open Source	O2I http://www.bioagent.net	• Supports the biomedical and clinical research • Oncology over Internet (O2I) context	
SOMA	Open Source	Finnish IT Center for Science http://www.csc.fi/proj/drug2000	• Workflow for small molecule property calculations • Supported by the core workflow program Grape • Includes programs: CORINA, ROTATE, BRUTUS, GOLD, SYBYL, • VOLSURF, XSCORE	
Pipeline Pilot	Commercial	Accelrys http://www.scitegic.com/	• Visual programming HCS optimized • plate viewer • HeatMaps	Widely applied in drug discovery and HTS Visual programming Integration of third party applications Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods

Software	License type	Vendor and URL	Features	Integration with other software
ChemSense	Commercial	InforSense http://www.inforsense.com/	• Built on InforSense KDE core provides seamless integration with third party tools • Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods	• Wide range of Bioinformatics, QSAR, molecular modeling and evaluation methods • Full integration of Perl, R/Bioconductor and Matlab • HCS instrumentation support
VIBE	Commercial	INCOGEN http://www.incogen.com/	• Can interface with a variety of environments, including high throughput platforms such as Sun Microsystem's Grid Engine • supports GRID computing used in image processing	• Extensive use of XML for configuration, data exchange, data storage and communications • Extensible Java API
ueberTool	Commercial	Science Factory http://www.science-factory.com/	• can be used in HCS for post-analysis, hits evaluation • Integration and analysis of molecular biological data • Graphical user interface makes constructing bioinformatics workflows • Blast • supports FASTA format	• Integrated programming language for extending core functionalities
BioLib	Commercial	BioLog	• Support Bio-IT data warehousing • Comprehensive Methods Library and a customizable Algorithm Library • Open-architecture Informatics System	

Table 1. Workflow systems in life sciences domain. Green highlighting indicate workflow systems optimized for High Content Screening.

3.6 Summary and Vision

Large-scale HCS data analysis needs flexible workflow based integration of different components and sub-processes from diverse formats (library, image readers, microscope nodes, image processing results, data mining) which can provide *in silico* experimental design through visual programming and execution on grids. A pipeline system in HCS is a very new concept and still evolving. The final goal is a distributed and ubiquitous environment which can integrate all automated microscopes, all available image processing packages, bioinformatics databases and data from other large scale experiments (proteomics, microarray, flow cytometry, new sequencing, etc). Workflow systems can be data-intensive, computation intensive, analysis intensive, visualization-intensive, process-intensive. Problem of service composition is how to compose simple services to perform complex tasks. The scalability of a workflow system is an important factor which helps in large-scale HCS data analysis in a high performance parallel and distributed computing environment[5]. One very important aspect is the option to run workflows in the background on a remote server which is, especially advantageous in case of long running workflows. In this situation only the control of the workflow should be presented in the remote GUI (desktop or web client). Present workflow systems in life sciences which can be applied for HCS need to integrate several resources like web technology (LIMS systems), grid services (for powerful image processing) and web services (access to bioinformatics sources). Web and grid services provide access to distributed resources, while workflow techniques enable the integration of these resources to perform *in silico* experiments. Most of the HCS database systems (LIMS) and very often all external compound information (RNAi or small molecules) are accessible over the web. After finalizing the experiment and retrieving the hits it is necessary to investigate (compare) results with external information. Semantic web services will help accessing this biological knowledge in a distributed, heterogeneous environment by adding semantics, defining common ontologies and applying them to software tools and databases. Semantic web technology can provide more generic solutions that can be re-used between related workflows. “Web services” is a distributed computing technology that provides software services over the web. Over the past few years the evolution of web services in bioinformatics[17] has shown tremendous impact on the sharing of data and tools. With the intention directed towards execution in a heterogeneous and often distributed environments, the interoperability of web services has become much more important.