A workflow system for virtual screening in cancer chemoprevention

Conference Paper · November 2012

DOI: 10.1109/BIBE.2012.6399766

CITATIONS
7

READS
64

8 authors, including:

Christos C. Kannas
Institute of Cancer Research
12 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Kleo Achilleos
University of Cyprus
3 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Zinonas C. Antoniou
University of Cyprus
24 PUBLICATIONS **117** CITATIONS
SEE PROFILE

Christos Nicolaou
Eli Lilly
32 PUBLICATIONS **421** CITATIONS
SEE PROFILE

Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

A Workflow System for Virtual Screening in Cancer Chemoprevention

Kannas C. C., Achilleos K. G., Antoniou Z., Nicolaou C. A., Pattichis C. S.
Department of Computer Science
University of Cyprus
Nicosia, Cyprus

Kalvari I., Kirmitzoglou I., Promponas V. J.
Department of Biological Sciences
University of Cyprus
Nicosia, Cyprus

Abstract — **Computer-aided drug discovery techniques have been widely used in recent years to support the development of new pharmaceuticals. Virtual screening, the computational counterpart of experimental screening, attempts to replicate the results from *in vitro* and *in vivo* methods through the use of *in silico* models and tools. This paper presents the LISIs platform; a web based scientific workflow system for virtual screening that has been implemented primarily for the discovery of chemoprevention agents. We describe the overall design of the system as well as the implementation of its various components. Indicative results from early applications of the system are also presented to illustrate its potential uses and functionalities.**

Keywords-virtual screening; scientific workflow; predictive models; chemoinformatics; chemoprevention; drug discovery

I. INTRODUCTION

Chemoprevention research is the process of finding drugs and natural substances to prevent the occurrence of a particular disease (e.g. cancer) and determining their mechanism of action [1], [2]. This field of biology has only recently begun to attract interest from the life sciences informatics research community. Due to the substantial similarity between chemoprevention and the typical drug discovery process (DDP), applications in the chemoprevention field heavily borrow from applications in DDP. However, some differences do exist, the most important being that chemopreventive compounds must have no toxic effects since they are administered to healthy humans. In contrast, in the case of drugs, some toxic effects may be acceptable based on the severity of the disease they are targeting. In this paper we propose a virtual screening cancer chemoprevention platform based on scientific workflow modelling.

Virtual Screening (VS) is the computational counterpart of biological screening performed in laboratories. Its goal is to decrease the number of compounds physically screened by identifying small subsets of large molecular databases that have an increased probability to be active against a specific biological target. In this respect the method is related to machine learning techniques, such as classification and regression, which prepare predictive models to estimate the behaviour of unknown records based on a set of records with known properties. Typically, VS processes involve substantial numbers of molecules and combine a variety of computational techniques.

Scientific Workflow Management Systems (SWMSs) are powerful tools with enormous possibilities to facilitate the design and execution process of computational experiments. SWMSs enable scientists to plug together problem solving computational components [3] and implement complex *in silico* experiments, such as the analysis of large datasets that arise from sensors or computer simulations and the design and execution of complicated algorithms requiring multiple computationally intensive steps.

Since chemoprevention research and DDP are highly similar, the tools (including software applications) used for drug discovery can also be used for chemoprevention research. These tools can provide to chemoprevention researchers *in silico* models for problems that are common in the two research fields. Moreover, appropriate computational techniques can be used to create specific models for the needs of chemoprevention. Among them, are SWMSs used for VS in DDP, such as Taverna [4], KNIME [5], PipelinePilot [6] and InforSence Suite [7], which can also be used in chemoprevention research.

To this date, few chemoinformatics applications and computational chemistry tools have been reportedly used in the chemoprevention field. Apparently, it is required that cancer chemoprevention researchers have access to a customized, and easy to use, suite of *in silico* tools for handling and analysing relevant data. To fulfil this need we develop the LIfe Science InformaticS (LISIs) system, a SWMS that enables the creation of virtual screening workflows using general tools and methods borrowed from chemoinformatics, as well as components custom designed for cancer chemoprevention research. The LISIs platform is part of GRANATUM, an EU-FP7 project. The aim of this project is to bridge the gap between biomedical researchers ensuring that the biomedical community has access to the globally available information needed to perform complex cancer chemoprevention experiments and to conduct studies on large scale datasets.

The structure of the paper is as follows. In section II background and methodology are given covering the virtual screening process and scientific workflow systems. In section III the modules of the LISIs platform are presented, followed by section IV that describes a showcase and early results. Finally section V gives concluding remarks.

This work is done within the GRANATUM project, which is partially funded by the European Commission under the Seventh Framework Programme in the area of Virtual Physiological Human (ICT-2009.5.3).

II. BACKGROUND METHODOLOGY

A. Virtual Screening Process

Virtual Screening can be performed on libraries of real or virtual compounds and requires either measured activities for some known compounds or a known structure of the biomolecular target [8]. When only measured activities of compounds are known, virtual screening may employ analog-based library design, classification/regression models or any combination of the above. Each of the methods offers some advantages while it suffers from several shortcomings and so researchers typically design a VS experiment taking into account the specific requirements of each case. If high quality activity measurements about the ligands are available regression methods (in the form of e.g. Quantitative Structure Activity Relationship - QSAR models) can be used to extract rules capturing the essence of ligand similarity, and hopefully binding action, with high confidence. These rules can easily be used to filter untested compounds swiftly. Classification methods have fewer requirements than QSAR but also produce cruder results. Some methods rely on predefined sets of molecular descriptors and this makes them appropriate as general tools. However, such over-dependence on the descriptor set chosen restricts their potential pool of models and general findings. Reports in the literature [9], [10] describe the usage of descriptor sets in the 100's of thousands, a clear attempt to ensure that no significant ligand feature will be missed. Similar issues trouble the usage of 2D analog-based library design methods based on similarity searches.

When the structure of the target receptor is known the VS methods of choice typically rely heavily on docking and small molecule modelling. Initially, they take advantage of the knowledge about the receptor site to model it and then perform docking of molecules from a database in a systematic manner. A number of conformations are usually sampled for each molecule [11] and a score for every possible docking attempt is kept [12], [13]. Due to the costly nature of numerous steps of the process computer clusters are widely employed by the pharmaceutical industry [11], [14]. Additionally, databases of multiple conformers of compounds are prepared to avoid their reproduction for every virtual screening run [15]. As a result, currently, databases with millions of compounds can be screened within a few hours [14].

The key success measurement of VS is the achievement of high enrichment, i.e. getting an experimental hit rate for the subset of compounds it recommends that is considerably increased over that of a random compound set [14]. A successful process with high enrichment results in considerable savings in resources and time, since fewer compounds need to be physically screened while most hits present in the original large database are retrieved. Often, to improve the results of VS, several methods are used and their results are combined to produce a concise, high quality virtual hit list [11], [12]. Furthermore, it is common to perform a pre-processing step where databases of molecules are cleaned by filtering out compounds with undesired properties, such as large size, high flexibility and non-compliance to Lipinski's rule of 5 [16]. During this step compounds containing known unwanted substructures, e.g. known toxicophores, may also be eliminated [14]. However, and despite drastic improvements of various algorithms and steps involved in the process, VS accuracy still varies depending on the pharmaceutical target, the virtual library and the docking and scoring methods used. Thus, a necessary last step to the process is evaluation of the VS experiment results typically via visual inspection by a human expert [17]. A typical set up for a VS experiment is illustrated in Fig. 1.

B. Scientific Workflow Systems

SWMSs accelerate scientific discovery by incorporating data management, analysis, simulation, and visualization tools into a common platform. They provide an interactive visual interface that facilitates the design and execution of workflows. More over SWMS enable remote access as well as data and services sharing, making possible collaborations among geographically distributed researchers. In the next paragraphs a brief overview of the field is given. A more detailed review on SWMS can be found in [18].

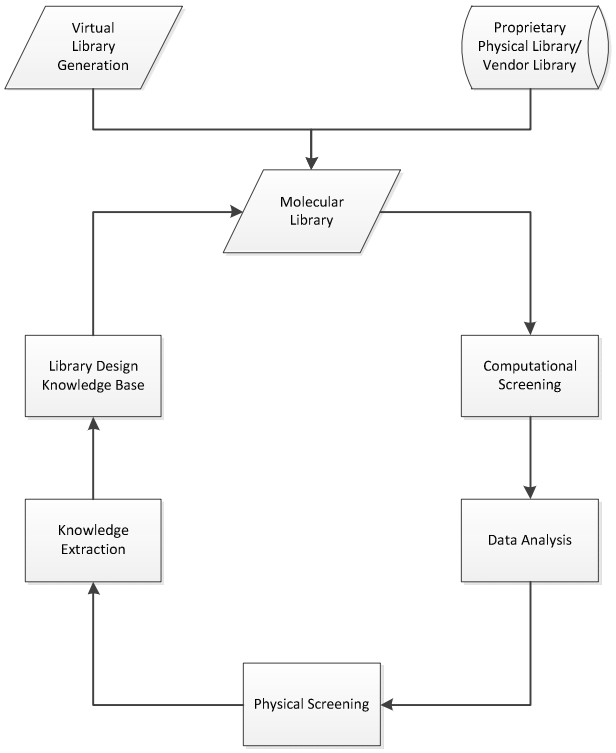

Figure 1. A typical set-up for a virtual screening experiment

Essentially, scientific workflows (SW) technology provides tools that automate the execution of a class of *in silico* experiments, offering multiple benefits for all the phases of an experiment's lifecycle. During the design and implementation phase, a repository of tried and tested workflows can be available to the scientists to choose from. During the execution phase, as experimenting is by definition a repeatable process, workflows can relieve the scientists of repetitive tasks, while at the same time enable keeping track of all the intermediary steps and data (provenance). These traces

can be used at a later stage to enable the reproducibility of the experiment. Provenance information [19] is also useful during the analysis phase to see the evolution of the research, trace the origin of an error or go back on a previous stage and change the direction of the research. Visualization tools are provided for this phase as well for assisting in the evaluation of the results.

The recent popularity of SWMS is partially owed to the emergence of the computational science paradigm, which promotes collaboration between scientists both within and across disciplines. Through the use of SWs, interdisciplinary teams can collaborate closely, share workflows and computational components and jointly undertake research initiatives requiring end-to-end scientific data management and computational analysis. Advances in grid technologies allow workflows to exploit parallel executions enabling large-scale data processing. In this case, workflows are used as a parallel programming model for data-parallel applications. Web services allow ease of access to local and distributed data sources as well as data aggregation from highly heterogeneous environments.

III. GRANATUM LISIs PLATFORM

LISIs aims to provide a set of tools to create, update, store and share Scientific Workflows for the discovery of new chemopreventive agents to chemoprevention experts. The system is available via a web interface through a password protected, tiered login process. Specifically, the login process provides different level access to platform functionalities based on the user profile. The user is able to assemble SWs utilizing available *in silico* models and tools loaded into the platform. Depending on the user profile and associated permissions, users may also construct new models and tools through the development of custom workflows made available by the system for this purpose. Workflows execute on the system server. The execution results are stored on the user's GRANATUM workspace, where the user is able to access, manipulate or share them with other users.

In general, the GRANATUM platform provides access to (i) Registered Users, which are Team Members (Senior Researcher, Junior Researcher) and Collaborators, (ii) Public Users and (iii) Administrators. LISIs platform is accessible only to registered users of the GRANATUM platform.

The LISIs platform is comprised of five (5) major modules, i.e. input, pre-processing, processing, post-processing and results/outputs as illustrated in Fig. 2. Each module hosts a collection of component categories essentially implementing a variety of functionalities. A component category may implement different variations of the same functionality.

A. Input Module

The input module consists of two categories of components: (i) Data File Input, which provides components for loading information from chemical and biological data files, and (ii) Linked Biomedical Space Input, which provides components for retrieving information via other functionalities of the GRANATUM platform [20].

1) Data File Input

A set of components which support parsing different chemical and biological data files. File formats currently supported include: (i) .sdf (Structure Data File), .smi (SMILES - Simplified Molecular Input Line Entry Specification), .pdb (Protein Data Bank) for chemical data files and (ii) .csv (Comma Separated Values) for biological data files.

2) Linked Biomedical Space Input

A set of components which support requesting and parsing data provided/retrieved from the GRANATUM Linked Biomedical Data Space.

B. Pre-Processing Module

The pre-processing module has four component categories: (i) Standardization/Normalization, (ii) Format Transformation, (iii) Chemical Descriptor Calculation and (iv) Compound Fragmentation.

1) Standardization/Normalization

This component category provides tools for standardizing and normalizing features of input data.

2) Format Transformation

This component category provides tools for converting specific file formats to others accepted and processed by the LISIs platform.

3) Chemical Descriptor Calculation

This component category provides tools for the calculation of various chemical descriptors of chemical compounds. Such descriptors include: molecular weight, hydrogen bond donors/acceptors and various 2D fingerprint representations (e.g. Morgan and MACCS fingerprints).

4) Compound Fragmentation

This component category provides tools to identify chemical substructures present in compounds through the *in silico* fragmentation of chemical compound structure. The use of various chemical compound fragmentation methods is available, such as Retrosynthetic Combinatorial Analysis Procedure (RECAP) [21], [22], Ring System Decomposition (RSD) and Molecular Frameworks [22], [23].

C. Processing Module

The processing module consists of five component categories: (i) Attribute Filtering, (ii) Compound Similarity, (iii) Substructure Matching, (iv) Docking Experiments and (v) Predictive Models.

1) Attribute Filtering

This component category provides tools for implementing filters for selecting compounds based on their chemical and biological attributes. Specifically, these components allow

users to enter ranges of acceptable values on available compound properties (including properties calculated by the Chemical Descriptors component and properties provided externally via the Data Input Module).

2) Compound Similarity

This component category provides tools for implementing filters for selecting compounds based on chemical structure similarity to other compounds indicated by the user.

3) Substructure Matching

This component category provides tools for implementing filters for selecting compounds based on whether they contain (or not) the chemical substructure(s) indicated by the user.

4) Docking Experiments

This component category provides tools for implementing filters for selecting compounds based on predicted binding affinity of a compound to a target protein using *in silico* docking experiments. Our platform currently uses two popular docking applications, which are free for use to the academic research community, namely Chil2 GlamDock [24] and AutoDock Vina [25].

AutoDock Vina attempts to find the best protein-ligand docking pose by employing a scoring function that takes into consideration both intramolecular and intermolecular contributions, as well as an optimization algorithm [26]. GlamDock employs a simple Monte Carlo (MC) and a gradient-based minimization procedure, in order to refine the initial MC placement [27].

5) Predictive Models

The Predictive Models component enables the usage of data-driven models designed to predict biochemical properties of interest to the user for the selection of compounds with acceptable predicted properties. The primary aim of this component is to: (i) provide the user with the tools to construct predictive models based on available information on a set of compounds, and (ii) use existing models to predict the attributes of new compounds to select those with an acceptable profile. The constructed models fall into the category of Quantitative Structure - Activity Relationship (QSAR) and Quantitative Structure - Property Relationship (QSPR) models used by the drug discovery community to predict relevant properties of molecules [28–31].

In order to drive the model-building process, a "Hierarchy of Cancer Chemoprevention Properties/Activities" is developed within the GRANATUM consortium. In brief, this light ontology-like effort aims to make available to modelling experts possible (independent) ways by which a substance may act as a cancer chemopreventive agent. The main idea is that this hierarchical structure will facilitate the development of predictive models aiming at different levels of granularity. Those more detailed models give more information than the generic model (Activity + Mechanism of Activity); however, they are more difficult to be constructed, since more data for each specific class should be available. Importantly, these hierarchies are dynamic in nature, i.e. more subclasses may be added at later stages in the light of new data.

During the model construction phase, the input to this component consists of a list of training set compounds with their respective property values and the settings for the predictive modelling application to prepare. The output of the component includes the predictive model and a log file containing measures of the quality of the model estimated by cross-validation and other appropriate techniques.

During the model usage phase the input to this component consists of a list of new, test compounds and the specific predictive model to use. The output of the component is a list of compounds with the predictions of the model for each compound and a log file documenting the results.

The component currently makes use of four popular predictive modelling algorithms widely used by the chemoinformatics community, namely Decision Trees (DT) [32], Random Forests (RF) [33], Support Vector Machines (SVM) [34] and k-Nearest Neighbours (k-NN) [35]. The modular architecture of the system will also enable future extensions, with the possibility of adding other appropriate predictive modelling algorithms.

D. Post-Pprocessing Module

The Post-Processing module contains the Cleaning/Formatting component. The main functionality of this component is the manipulation of the results taken from components described above, for the final reporting and visualization steps.

E. Output Module

The Output module contains components for: (i) Reporting, (ii) Visualization and (iii) Storage.

1) Reporting

This component provides tools for the formatting of the processing results from one or more *in silico* experiments and for basic visualization.

2) Visualization

This component provides tools for creating 2D figures of the resulting compounds and 3D representations of docking results.

3) Storage

This component provides tools to store results for future reuse and sharing.

F. Third Party Tools used by LISIs

The LISIs platform uses several, freely available to the research community tools to expedite development and maximize resources. Specifically, the following 3rd party tools are used: (i) Galaxy [36], an open, web-based platform for data intensive biomedical research, for the implementation of the SWMS; (ii) RDKit [37], an open source chemoinformatics

toolkit, used to support all the chemoinformatics related functionalities; (iii) R [38], a statistical environment used to support data mining, machine learning and statistics related functionalities for the generation of e.g. Predictive Models; (iv) Chil2 GlamDock [24] and AutoDock Vina [25] docking applications used to support docking experiments functionalities.

IV. Showcase And Results

LISIs has been used for the implementation of a VS experiment for screening molecules from a combined dataset, consisted of known estrogen receptor alpha (ER-alpha) inhibitors, probable DNA methyltransferase (DNMT) inhibitors and molecules obtained from the Indofine catalogues, which should have druglike features, satisfy a cytotoxicity prediction model and have high docking affinity against the ER-alpha protein. Fig. 3 is the graphical illustration of the showcase described in detail below.

At the Input Module level, we used a dataset consisting of 42 known ER-alpha inhibitors retrieved from PubChem [39], ~2400 compounds taken from Indofine's [40] online catalogues and 43 DNMT inhibitors listed in [41].

At the Pre-Processing Module level, a set of physiochemical molecular descriptors were calculated including molecular weight, hydrogen bond donors, hydrogen bond acceptors, and LogP.

At the Processing Module level, the following models were used:

a) A custom made Rule of Five (Ro5) filter: specifically the parameters used where: (i) Molecular Weight between 160 and 700, (ii) Hydrogen Bond Donors less or equal to 5, (iii) Hydrogen Bond Acceptors less or equal to 10, and (iv) LogP less or equal to 5.

b) A Toxicity Prediction Model: trained with the dataset of PubChem [39] Bioassay with id "AID 464"[1], consisted of 706 compounds, 331 were listed as active and 375 were listed as inactive. The model used an SVM implementation, specifically the Nu-Support Vector Classification (Nu-SVC) with linear kernel function and nu[2] value equal to 0.5. As a validation method Stratified K-fold[3] method with K equal to 10 was used. The resulting sensitivity and specificity were 0.6061 and 0.6486 respectively.

c) Docking experiments against ER-alpha: using Chil2 GlamDock [24] and AutoDock Vina [25]. ER-alpha's 3ERT[4] conformation was used and was retrieved from RCSB PDB[5], since we were looking for antagonists to ER-alpha protein. Docking tools were setup to provide us with the best docking affinity score.

At the Post-Processing Module level, a merge of the results from the individual processing components was made. Finally, at the Output Module level, a selection of molecules highly ranked was handpicked; a small sample of those satisfying the filters applied are shown in Table I. These molecules are currently undergoing further investigation by our expert chemoprevention colleagues.

V. Concluding Remarks

The LISIs platform aims to fill a current void in chemoprevention, and in general life sciences, research. Its successful deployment will have a major impact on enabling chemoprevention researchers to utilize state of the art computational techniques to search for promising chemical compounds that may lead to the discovery of novel agents with chemopreventive properties. These initial results show the need and the potential of such a platform for the chemoprevention research community.

Acknowledgment

This work has been partially supported through the EU-FP7 GRANATUM[6] project, contract number ICT-2009.5.3. Close collaboration with our partners from the laboratory of Cancer Biology and Chemoprevention[7] at Department of Biological Sciences[8], University of Cyprus and Cancer Chemoprevention and Epigenomics Workgroup[9] at German Cancer Research Center[10], members of the GRANATUM consortium and their crucial role in accomplishing the work described is acknowledged.

TABLE I. Sample of Highly Ranked Molecules

Canonical SMILES	Toxicity Predict.	Docking Affinity	ER-alpha Inhibitor
O=C(c1ccc(OCCN2CCCCC2)cc1)c1c2ccc(O)cc2sc1-c1ccc(O)cc1	Inactive	-11.707346	Y
O=c1oc2cc(O)ccc2c(-c2ccccc2)c1-c1ccc(O)cc1	Inactive	-10.768855	N
COc1ccc(-c2c(=O)oc3cc(O)ccc3c2-c2ccccc2)cc1	Inactive	-10.283143	N
O=c1oc2cc(O)ccc2c(-c2ccccc2)c1-c1ccccc1	Inactive	-10.216389	N
O=c1oc2ccccc2c(-c2cc(O)cc(O)c2)c1-c1ccccc1	Inactive	-10.150722	N
C#CC1(O)CCC2C3CCc4cc(O)ccc4C3C(OC)CC21C	Active	-9.380188	Y
O=C([O-])CNC(=O)Cc1c2c(oc(=O)c1)cc(O)cc2	Inactive	-7.711231	N

1 http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=464

2 Upper bound of the fraction of training error and lower bound of the fraction of support vectors. Should be in the interval of (0, 1].

3 A variation of K-fold which returns stratified folds, i.e. create folds preserving the same percentage for each target class as in the complete set.

4 http://www.rcsb.org/pdb/explore/explore.do?structureId=3ert

5 http://www.rcsb.org

6 http://www.granatum.org

7 http://www.ucy.ac.cy/goto/biosci/el-GR/andreasc.aspx

8 http://ucy.ac.cy/goto/biosci/en-US/HOME.aspx

9 http://www.dkfz.de/en/tox/cancer_chemoprevention.html

10 http://www.dkfz.de/en/index.html

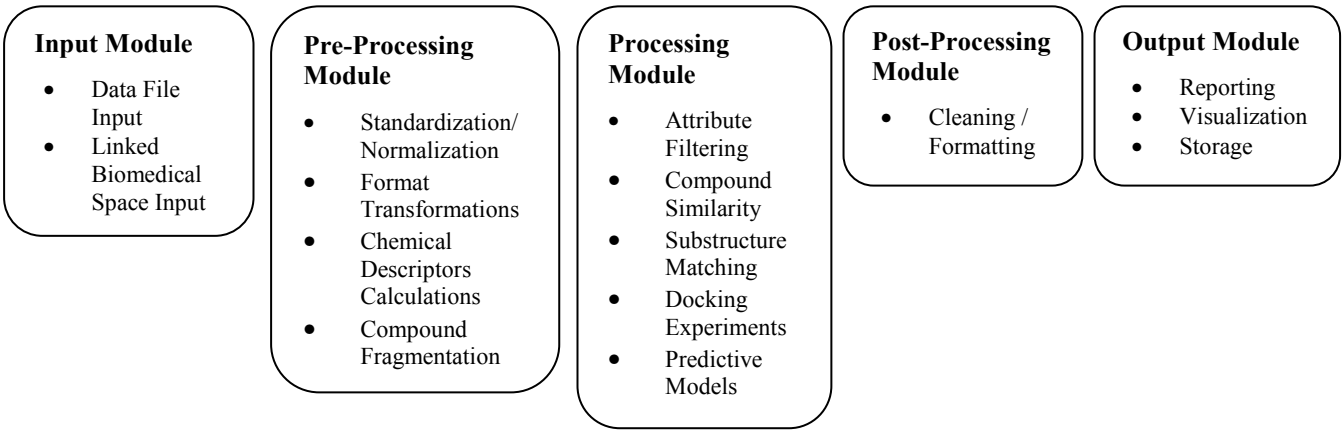

Figure 2. LISIs Modules and Components

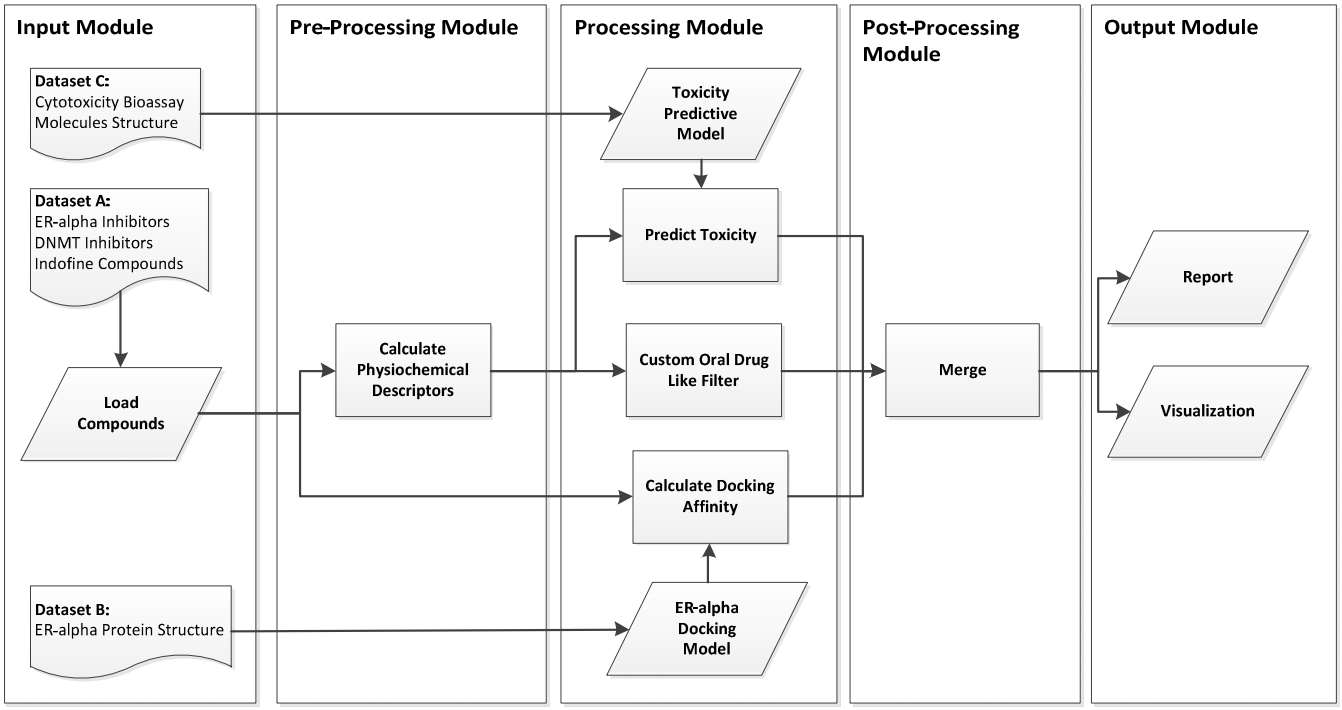

Figure 3. Showcase workflow

REFERENCES

[1] G. J. Kelloff, C. C. Sigman, and P. Greenwald, "Cancer chemoprevention: progress and promise," *Eur. J. Cancer*, vol. 35, no. 14, pp. 2031–2038, Dec. 1999.

[2] G. D. Geromichalos, "Importance of molecular computer modeling in anticancer drug development," *J BUON*, vol. 12 Suppl 1, pp. S101–118, Sep. 2007.

[3] A. Barker and J. V. Hemert, "Scientific Workflow: A Survey and Research Directions," in *Proceedings of the 7th international conference on Parallel processing and applied mathematics*, Berlin, Heidelberg, 2008, pp. 746–753.

[4] "Taverna - open source and domain independent Workflow Management System." [Online]. Available: http://www.taverna.org.uk/. [Accessed: 16-Jul-2012].

[5] "KNIME | Konstanz Information Miner." [Online]. Available: http://www.knime.org/. [Accessed: 16-Jul-2012].

[6] "Pipeline Pilot is Accelrys' scientific informatics platform." [Online]. Available:

http://accelrys.com/products/pipeline-pilot/. [Accessed: 16-Jul-2012].

[7] "IDBS - InforSense Suite - analytical data management." [Online]. Available: http://www.idbs.com/products-and-services/inforsense-suite/. [Accessed: 16-Jul-2012].

[8] W. L. Jorgensen, "The Many Roles of Computation in Drug Discovery," *Science*, vol. 303, no. 5665, pp. 1813–1818, Mar. 2004.

[9] A. Rusinko 3rd, M. W. Farmen, C. G. Lambert, P. L. Brown, and S. S. Young, "Analysis of a large structure/biological activity data set using recursive partitioning," *J Chem Inf Comput Sci*, vol. 39, no. 6, pp. 1017–1026, Dec. 1999.

[10] X. Chen, A. Rusinko, and S. S. Young, "Recursive Partitioning Analysis of a Large Structure-Activity Data Set Using Three-Dimensional Descriptors1," *J. Chem. Inf. Model.*, vol. 38, no. 6, pp. 1054–1062, Nov. 1998.

[11] E. M. Krovat, T. Steindl, and T. Langer, "Recent Advances in Docking and Scoring," *Current Computer - Aided Drug Design*, vol. 1, no. 1, pp. 93–102, Jan. 2005.

[12] C. Bissantz, G. Folkers, and D. Rognan, "Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations," *J. Med. Chem.*, vol. 43, no. 25, pp. 4759–4767, Dec. 2000.

[13] M. Stahl and M. Rarey, "Detailed analysis of scoring functions for virtual screening.," *J Med Chem*, vol. 44, no. 7, pp. 1035–1042, Mar. 2001.

[14] B. Waszkowycz, T. D. J. Perkins, R. A. Sykes, and J. Li, "Large-scale virtual screening for discovering leads in the postgenomic era," *IBM Syst. J.*, vol. 40, no. 2, pp. 360–376, 2001.

[15] P. Lyne, "Structure-based virtual screening: an overview," *Drug Discovery Today*, vol. 7, no. 20, pp. 1047–1055, Oct. 2002.

[16] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 23, no. 1–3, pp. 3–25, Jan. 1997.

[17] A. C. Anderson and D. L. Wright, "The Design and Docking of Virtual Compound Libraries to Structures of Drug Targets," *Current Computer - Aided Drug Design*, vol. 1, no. 1, pp. 103–127, Jan. 2005.

[18] K. G. Achilleos, C. C. Kannas, C. S. Pattichis, and C. A. Nicolaou, "Open Source Workflow Systems: A review," presented at the IEEE 12th International Conference on BioInformatics and BioEngineering, Larnaka, Cyprus, 2012.

[19] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *SIGMOD Rec.*, vol. 34, no. 3, pp. 31–36, Sep. 2005.

[20] "GRANATUM - Project Vision." [Online]. Available: http://granatum.org/. [Accessed: 20-Aug-2012].

[21] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann, "RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry," *J. Chem. Inf. Model.*, vol. 38, no. 3, pp. 511–522, May 1998.

[22] C. A. Nicolaou and C. S. Pattichis, "Molecular Substructure Mining Approaches for Computer-Aided Drug Discovery: A Review," in *Proceedings of the 5th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine*, Ioannina, Greece, 2006.

[23] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. Molecular frameworks," *J. Med. Chem.*, vol. 39, no. 15, pp. 2887–2893, Jul. 1996.

[24] "Chil^2 - Molecular Design for Science and Technology." [Online]. Available: http://www.chil2.de/Glamdock.html. [Accessed: 17-Jul-2012].

[25] "AutoDock Vina - molecular docking and virtual screening program." [Online]. Available: http://vina.scripps.edu/. [Accessed: 17-Jul-2012].

[26] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J Comput Chem*, vol. 31, no. 2, pp. 455–461, Jan. 2010.

[27] S. Tietze and J. Apostolakis, "GlamDock: development and validation of a new docking tool on several thousand protein-ligand complexes," *J Chem Inf Model*, vol. 47, no. 4, pp. 1657–1672, Aug. 2007.

[28] A. Z. Dudek, T. Arodz, and J. Gálvez, "Computational methods in developing quantitative structure-activity relationships (QSAR): a review," *Comb. Chem. High Throughput Screen*, vol. 9, no. 3, pp. 213–228, Mar. 2006.

[29] E. Pontiki, D. Hadjipavlou-Litina, G. Geromichalos, and A. Papageorgiou, "Anticancer activity and quantitative-structure activity relationship (QSAR) studies of a series of antioxidant/anti-inflammatory aryl-acetic and hydroxamic acids," *Chem Biol Drug Des*, vol. 74, no. 3, pp. 266–275, Sep. 2009.

[30] R. Perkins, H. Fang, W. Tong, and W. J. Welsh, "Quantitative Structure–Activity Relationship methods: perspectives on drug discovery and toxicology," *Environ Toxicol Chem*, vol. 22, no. 8, p. 1666, 2003.

[31] J. J. Sutherland, L. A. O'Brien, and D. F. Weaver, "A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships," *J. Med. Chem.*, vol. 47, no. 22, pp. 5541–5554, Oct. 2004.

[32] Wikipedia contributors, "Decision tree," *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc., 07-Aug-2012.

[33] Wikipedia contributors, "Random forest," *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc., 07-Aug-2012.

[34] Wikipedia contributors, "Support vector machine," *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc., 01-Aug-2012.

[35] Wikipedia contributors, "*k*-nearest neighbor algorithm," *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc., 30-Jul-2012.

[36] "The Galaxy Project: Online bioinformatics analysis for everyone." [Online]. Available: http://galaxy.psu.edu/. [Accessed: 16-Jul-2012].

[37] "RDKit." [Online]. Available: http://www.rdkit.org/. [Accessed: 17-Jul-2012].

[38] "The R Project for Statistical Computing." [Online]. Available: http://www.r-project.org/. [Accessed: 17-Jul-2012].

[39] "The PubChem Project." [Online]. Available: http://pubchem.ncbi.nlm.nih.gov/. [Accessed: 18-Jul-2012].

[40] "INDOFINE Chemical Company, Inc. | NJ Chemicals | Los Angeles County Chemicals | NJ Molecules | Los Angeles County Molecules | Xian | Shijiazhuang | Shanghai | China | Call 908-359-6778." [Online]. Available: http://www.indofinechemical.com/. [Accessed: 18-Jul-2012].

[41] J. L. Medina-Franco, F. López-Vallejo, D. Kuck, and F. Lyko, "Natural products as DNA methyltransferase inhibitors: a computer-aided discovery approach," *Molecular Diversity*, vol. 15, pp. 293–304, Aug. 2010.