

LiSIs: An Online Scientific Workflow System for Virtual Screening

Christos C. Kannas^{*1}, Ioanna Kalvari², George Lambrinidis³, Christiana M. Neophytou², Christiana G. Savva², Ioannis Kirmitzoglou², Zinonas Antoniou¹, Kleo G. Achilleos¹, David Scherf⁴, Chara A. Pitta², Christos A. Nicolaou¹, Emanuel Mikros³, Vasilis J. Promponas², Clarissa Gerhauser⁴, Rajendra G. Mehta⁵, Andreas I. Constantinou² and Constantinos S. Pattichis¹

¹Department of Computer Science, University of Cyprus, Nicosia, Cyprus

²Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus

³Department of Pharmaceutical Chemistry, Faculty of Pharmacy, National & Kapodistrian University of Athens, Athens, Greece

⁴Cancer Chemoprevention and Epigenomics Workgroup, German Cancer Research Center, Heidelberg, Germany

⁵Illinois Institute of Technology (IIT) Research Institute, Chicago, Illinois, USA



Abstract: Modern methods of drug discovery and development in recent years make a wide use of computational algorithms. These methods utilise Virtual Screening (VS), which is the computational counterpart of experimental screening. In this manner the *in silico* models and tools initially replace the wet lab methods saving time and resources. This paper presents the overall design and implementation of a web based scientific workflow system for virtual screening called, the Life Sciences Informatics (LiSIs) platform. The LiSIs platform consists of the following layers: the input layer covering the data file input; the pre-processing layer covering the descriptors calculation, and the docking preparation components; the processing layer covering the attribute filtering, compound similarity, substructure matching, docking prediction, predictive modelling and molecular clustering; post-processing layer covering the output reformatting and binary file merging components; output layer covering the storage component. The potential of LiSIs platform has been demonstrated through two case studies designed to illustrate the preparation of tools for the identification of promising chemical structures. The first case study involved the development of a Quantitative Structure Activity Relationship (QSAR) model on a literature dataset while the second case study implemented a docking-based virtual screening experiment. Our results show that VS workflows utilizing docking, predictive models and other *in silico* tools as implemented in the LiSIs platform can identify compounds in line with expert expectations. We anticipate that the deployment of LiSIs, as currently implemented and available for use, can enable drug discovery researchers to more easily use state of the art computational techniques in their search for promising chemical compounds. The LiSIs platform is freely accessible (i) under the GRANATUM platform at: <http://www.granatum.org> and (ii) directly at: <http://lisis.cs.ucy.ac.cy>.

Keywords: Chemoinformatics, docking, drug discovery, predictive models, QSAR, scientific workflow, virtual screening.

INTRODUCTION

Virtual Screening (VS) can be the first step prior to biological screening. The objective of VS is to select the most promising compounds that will be subsequently scanned in a laboratory setting. In this manner a subset of a large dataset is being tested increasing the probability to identify lead compounds against specific biological targets [1, 2]. In this respect the method is related to machine learning and statistical techniques, such as classification and regression. These methods target to develop predictive models for the identification of the properties of unknown compounds based on a set of compounds with known properties. Typically, VS processes involve substantial num-

bers of molecules and combine a variety of computational techniques, often organized in complex computational pipelines [3].

Scientific Workflow Management Systems (SWMS) are powerful tools with immense potential to expedite the design, development and execution processes of computational experiments. SWMS can be applied by scientists for the solution of complex computational problems [4] and also to design complex *in silico* experiments [5].

In this paper we propose a VS platform based on scientific workflow modelling. A preliminary version of this study was presented in [6].

The Life Sciences Informatics (LiSIs) platform [7] is a part of GRANATUM [8], an EU FP7 project. The aim of GRANATUM [8] is to provide biomedical researchers access to state of the art computational tools to perform

*Address correspondence to this author at the Department of Computer Science, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus; Tel: +357-22892685; Fax: +357-22892701; E-mail: kannas.christos@ucy.ac.cy

complex cancer chemoprevention experiments and to conduct studies on large-scale datasets.

Cancer chemoprevention is defined as the use of natural, synthetic, or biologic chemical agents to reverse, suppress, or prevent carcinogenic progression to invasive cancer [9]. The experimental approach for the discovery of cancer chemopreventive agents is similar to the typical drug discovery process (DDP) [10].

Chemoprevention research (CPR) and DDP are highly similar processes, therefore the tools used for drug discovery can also be applied to CPR. In this context, computational tools can be used to develop specific models for the needs of chemoprevention. For example, SWMS used for VS in DDP, such as Taverna [11-13], KNIME [14, 15], PipelinePilot [16] and IDBS InforSence Suite [17], can also be adapted for use in CPR. Researchers in the chemoprevention field do not generally use these computational tools. Apparently, the field of cancer chemoprevention can be advanced by customised, and easy to use *in silico* tools for data handling and analysing.

It should be emphasized that to the best of our knowledge there are no similar tools to LiSIs in cancer chemoprevention. Moreover, it is noted that the proposed platform features: (i) free and open access to the research community, (ii) it is integrated on the GALAXY platform that is familiar to molecular biologists, (iii) it aims to have a user friendly interface, (iv) it is part of a larger integrated project, GRANATUM, offering access to semantic web technologies, text mining tools, and collaborative environment for sharing data sets, models, etc. [8]. Furthermore, the paper covers two case studies that are used to evaluate and validate the system: (1) a case study on QSAR model for mutagenicity, and (2) on estrogen receptor (ER) binding.

VIRTUAL SCREENING PROCESS

VS process is carried out on libraries of real or virtual compounds and requires known measured activities of control compounds or a known structure of the biomolecular target [18]. When only measured activities of compounds are known, virtual screening uses analogue-based library design, classification and regression models or any combination of these.

Since no method is generally applicable to all cases, a VS experiment takes into account the requirements of each case. For example, if high quality ligand activity measurements are available, regression methods (Quantitative Structure Activity Relationship - QSAR modelling) can be used to extract with confidence rules predicting ligand similarity, and binding action [19].

When the structure of the target receptor is known the VS method typically relies on protein-ligand docking and small molecule modelling. Initially, it takes the advantage of the knowledge about the receptor site to model it and then perform docking from a database. A number of conformations are usually sampled for each molecule [20] and a score for every possible docking is computed [21]. Due to the computationally demanding processes computer clusters are employed by the pharmaceutical industry [20, 22]. In addition, databases of

multiple conformers of compounds are prepared in advance to avoid their reproduction for every VS run [23]. Often, multi-objective methods may be used that enable the use of numerous objectives, analogue or target-based, to identify compounds that simultaneously meet multiple criteria relevant to the virtual screening experiment pursued [24].

The key measure for validating the success of VS is the achievement of high enrichment targeting in an experimental hit rate for the subset of compounds it recommends which is significantly better over that of a random compound set [22]. 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. In practice, to enrich the results of VS, several methods are tried and their results are combined to produce a concise, high quality virtual hit list [20, 21]. Furthermore, it is also a common practice to perform a pre-processing step where databases of molecules are cleaned by filtering out compounds with undesired properties. These properties include, a large size, high flexibility and non-compliance to Lipinski's rule of 5 [25]. During this step compounds containing known unwanted substructures, e.g. known toxicophores, may also be eliminated [22]. Although significant algorithmic improvements have been achieved in the VS process, accuracy still varies depending on the pharmaceutical target, the virtual library and the docking and scoring methods used. The last step is the evaluation of the VS experiment results typically *via* visual inspection by a human expert [26].

SCIENTIFIC WORKFLOW MANAGEMENT SYSTEMS

SWMS target in accelerating scientific discovery by incorporating in their processing steps, data management and analysis, simulation, and visualization tools into a single platform. Most importantly SWMS provide an interactive visual interface that facilitates the design and execution of workflows. A brief overview of the field is given whereas a more detailed review on SWMS can be found in [27].

Scientific workflow (SW) based platforms provide tools that automate the execution of a class of *in silico* experiments, offering significant benefits for all the phases of an experiment's life-cycle. In the context of 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 [28] is also useful during the analysis phase to see the evolution of the research, trace the origin of an error or go back to a previous stage and change the direction of research. Visualization tools are provided for this phase as well for assisting in the evaluation of the results.

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. Moreover, recent advances in grid technologies allow workflows to exploit parallel executions enabling large-scale data processing.

MATERIALS AND METHODS

LiSIs [7] aims to provide a set of tools to create, update, store and share SWs for the discovery of active compounds for biomedical researchers. Access to LiSIs can be achieved *via* a web interface through a password protected login process either from the GRANATUM portal [8] (preferred access point) or directly from the LiSIs portal [7]. The login process provides different levels of access to platform functionality 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 can also be stored on the user's GRANATUM workspace [8], where the user is able to access, manipulate or share them with other users.

Fig. (1) is an illustration of the cheminformatics tools available on LiSIs. Below is a brief description of each tool category.

INPUT LAYER

The Input Layer consists of the following two component categories:

Data File Input: provides tools which support parsing different chemical and biological data files. File formats currently supported include Chemical Data Files, which are sdf (SDF - Structure Data File), smi (SMILES - Simplified Molecular Input Line Entry Specification), pdb (PDB - Protein Data Bank), pdbqt (AutoDock Protein and Ligand data files) and Biological Data Files which are csv (CSV - Comma Separated Values), tab (Tab Separated Values) and also text files.

These tools get as input ASCII files and create output files which are pickled Python objects, which we reference them as binary files.

GRANATUM File Input: A component which provides GRANATUM's platform users to upload on LiSIs files located at GRANATUM workspace [8].



Fig. (1). Cheminformatics tools available on LiSIs, each tool is under a specific layer.

ChemSpider Molecule Retrieval: A component which given a file with molecules common name it uses ChemSpider API to retrieve information available for those molecules, and returns the result as a SMILES file.

PRE-PROCESSING LAYER

The Pre-Processing Layer consists of the following four component categories:

Descriptors Calculation: This component provides tools for calculating various descriptors of chemical compounds. Currently the platform enables calculation of whole-compound descriptors with the use of RDKit [29]. Example descriptors include molecular weight, number of hydrogen bond donors and acceptors, polar surface area, number of rings, calculated octanol - water partition coefficient (cLogP), molecular complexity based on the method proposed by Barone [30] and molecular flexibility, as well as molecular fingerprints which can be one of Morgan (circular) fingerprints [31], MACCS [32], Atom-Pair [33], Topological Torsion [34], and topological fingerprints, a Daylight like fingerprint based on hashing molecular sub-graphs [35].

Docking Preparation: This component provides the following tools:

- *3D Coordinate Calculator:* A tool for preparing compounds for docking experiments, by calculating their 3D coordinates and creating the appropriate files required by the docking software used by LiSIs.
- *Protein Cleaner:* A tool provided by AutoDock, which is used to automate the process of cleaning a protein to create the required files used by AutoDock Vina [36, 37].

All the tools of this layer use as input and output binary files.

PROCESSING LAYER

The Processing Layer consists of the following five component categories:

Attribute Filtering: This component provides tools for implementing compound selection filtering tuned on the compounds chemical and biological attributes. These components enable the user to pre-select ranges of acceptable values on available compound properties (including properties calculated by the Chemical Descriptors component and properties provided externally from the Data Input Layer). Three tools are available under this category, “Chemical Properties Filter”, “GRANATUM Ro5 Filter” and “Lipinski Ro5 Filter”.

Compound Similarity: This component provides tools for implementing filters for selecting compounds based on their chemical structure similarity to other compounds designated by the user. Two tools are available under this category, “Similarity Filter” and “Diversity Filter”.

“Similarity Filter” requires two input datasets; the one is used as the reference dataset and the other as the query

dataset. The results are two datasets, which are subsets of the initial reference dataset, where the first one contains the compounds that are similar to the query dataset and the second one contains those that are not similar.

“Diversity Filter” on the other and requires only one input dataset and it generates two datasets where the first contains the compounds that are not similar among them and the second contains compounds that are similar among them.

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

This tool requires as input one dataset of compounds and at least one substructure in SMiles ARbitrary Target Specification (SMARTS) format.

Docking Prediction: This component provides tools for implementing filters for selecting compounds based on predicted binding affinity of a compound to a target protein using *in silico* docking prediction. The LiSIs platform currently uses AutoDock Vina, a popular docking application, freely available to the academic research community. AutoDock Vina attempts to find the best receptor-ligand docking pose by employing a scoring function that takes into consideration both intramolecular and intermolecular contributions, as well as an optimization algorithm [36].

Predictive Modelling: This component aims to provide the user with the tools to construct data-driven predictive models based on available information on a set of compounds. These models are used to predict biochemical properties of interest of new compounds and to select those with an acceptable profile. Our platform uses system’s underlying R installation to support the creation and reuse of predictive models.

This component makes use of four widely used predictive modelling algorithms by the chemoinformatics community: Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and k-Nearest Neighbours (k-NN) [38].

Molecular Clustering: This component provides a unified interface to different molecular clustering methods such as Agglomerative Hierarchical Clustering, Divisive Hierarchical Clustering, k-Means Partitional Clustering and k-Medoids Partitional Clustering. The provided molecular clustering is based on fingerprint similarity or distance.

POST-PROCESSING LAYER

The Post-Processing Layer consists of the following two component categories:

Output Reformatting: This component provides a tool to convert results in various formats supported by OpenBabel [39].

Binary File Merging: This component provides a tool for merging binary files, containing chemical structure objects with processing component results, into one binary file.

OUTPUT LAYER

The Output Layer consists of the following component category:

Storage: This component covers the storage of results in various formats for future reuse and sharing. The tools available under this component category convert, binary files containing *in silico* molecules, to various file formats such as SMILES, SDF, CSV and Tabular.

THIRD PARTY TOOLS USED BY LISIS

The LiSIs platform uses the following 3rd party tools that are freely available:

Galaxy [40-42], a web-based platform widely used in the biomedical community for intensive data processing and analysis, used for the customized SWMS platform;

RDKit [29], an open source chemoinformatics toolkit;

Pybel [43], a Python wrapper for the OpenBabel chemoinformatics toolkit [39], used for chemical file format transformations;

R [44], a statistical environment that supports data mining, machine learning and statistics based functionalities; caret (Classification and Regression Training) package [38] is used for the generation and reuse of Predictive Models and for Molecular Clustering;

AutoDock Vina [36, 37] docking application used to support docking experiments functionality.

RESULTS

Comparison with other open source Scientific Workflow Management Systems

Taverna

Taverna is an open-source, grid-aware workflow management system [11-13]. It has found wide application in the bioinformatics, chemistry, data- and text-mining and astronomy communities although the system is domain independent. It is comprised of the Taverna Workbench graphical workflow authoring client, a workflow representation language, and an enactment engine. Taverna is implemented as a service-oriented architecture, based on web service standards. From the advent of its design Taverna was an application that applied web services technology to workflow design. That meant that tools created using different programming languages (e.g. Java, Perl, Python, etc.) or platforms (UNIX, Windows, etc.) could now be accessed *via* a web service interface eliminating any need for integration. The same applied to the databases available on the web. As a result, researchers could design and execute a pipeline of web services, with little programming knowledge. Its architecture supports parallelism, both intra-process and inter-process, asynchronous service support and separation of data and process spaces to support scaling to arbitrary data volumes.

A vital component of Taverna's open architecture is the plug-in functionality. Various plug-ins have been developed for accessing online bio-catalogues or for integrating

chemoinformatics processing services. Provenance also plays an integral part in Taverna, allowing users to capture and inspect details such as who conducted the experiment, what services were used, and what results were produced. An additional strong feature of Taverna is workflow sharing. The users have direct access to the myExperiment [45] social collaboration site where they can upload or download workflows as needed.

KNIME

Konstanz Information Miner (KNIME) is a modular environment that supports operations such as data integration from various sources, processing, modelling, analysing and mining, as well as parallel execution [14,15]. KNIME is primarily used in pharmaceutical research with some applications reported in other areas like customer resource management and data analysis, business intelligence and financial data analysis. It is an open-source platform free for non-profit and academic use. It is available as a local desktop application but additional features such as user authentication, web services integration, web browser interface, remote server and cluster execution are available in (and restricted to) the professional package.

The platform enables the user to visually assemble and execute data pipelines providing an interactive view of the results. KNIME pipeline(s) consist of modular independent components that combine different projects in a single pipeline. At the same time its expandable architecture enables the easy integration of newly developed tools.

One highlight of KNIME's latest additions is the ability to support Predictive Model Markup Language (PMML) [46]. The PMML is an XML-based markup language that enables applications to define models related to predictive analytics and data mining and to share those models between PMML-compliant applications [47]. As a result a model developed by KNIME can be exported and then used in another data mining engine. Another characteristic is the addition of database ports that are JDBC-compliant that work directly in the database enabling even preview of the actual data inside the database tables [46].

Although written in Java, KNIME, permits running Python, Perl and other code fragments through the use of special scripting nodes. This is extremely useful as the majority of scientific work is currently under the form of Python or Perl scripts.

KNIME functionality is enriched by integrating functionality of different data analysis open source projects for machine learning and data mining, for statistical computations and visualizations as well as many chemoinformatics plug-ins.

Galaxy

Galaxy is a web-based platform for data intensive biomedical research [40-42]. It provides a framework for integrating computational tools and an environment for interactive data analysis, reuse and sharing. As stated in [40, 41] the primary design considerations of Galaxy were accessibility, reproducibility and transparency. Galaxy is accessible to scientists with no programming knowledge through the use of Galaxy tools. It produces reproducible

computational analysis results by generating metadata for each analysis step through the automated production of Galaxy History items. It also promotes transparency by enabling the sharing of data, tools, workflows, results and report documents.

A structured well-defined interface allows the wrapping of nearly any tool that can be run from the command-line into a Galaxy tool. The platform is open source and has been designed specifically to meet the needs of bioinformaticians supporting sequence manipulation with built in libraries. It does not support any control flow operations or remote services. Additionally it does not use a workflow language but rather a relational database. The Galaxy workflow system allows for analysis using multiple tools incorporated to the system which may be built and run or extracted from past runs, and rerun.

Pages are a unique feature to Galaxy. They are online documents used to describe the analysis performed but also to provide links to the Galaxy objects that were used in the analysis, i.e. Histories, Workflows, and Datasets. This enables the reader of the document to have direct access to the dataset used, to import the workflow and reproduce the experiment himself. It also makes it even easier for another scientist to continue and build upon reported previous work.

A recent Taverna-Galaxy integration allows the generation of Galaxy tools from Taverna 2 workflows [48]. The tools can then be installed in a Galaxy server and become part of a Galaxy pipeline. Moreover, Galaxy workflows can be directly shared through the myExperiment site [49]. Galaxy can also be instantiated on cloud computing infrastructures and interfaced with grid clusters [50].

CONCLUDING REMARKS

KNIME is considered among the top open source software for chemoinformatics. Taverna is a prominent web service oriented platform employed in more than 350 organizations around the world with frequent enhancements.

Two recent ones being Taverna Mobile and Taverna On-line (under development) [51]. Galaxy is a promising platform where the online features prevail as unique among the three systems.

Galaxy offers additional benefits due to its online nature. There is no need to set up installations on local machines or remote servers, no downloads, no conflicts and no updates to worry about. All tools are available at any personal computer from anywhere in the world provided that they are connected to the internet. The same applies to data. A scientist can import data in the system and process them with the appropriate workflow or design a new one. Moreover the data and work are secure and can be backed up and protected depending on user preferences and specific system specifications. Importantly, all data and work can be shared with other collaborators in real time. Galaxy can even offer the more advanced features such as transparent access to grid services or the cloud, thus, offering speed and efficiency for scientific processes that are computationally expensive and/or data intensive.

Table 1 is a comparison of LiSIs (Galaxy) to KNIME and Taverna platforms. This comparison is focused on their system level and their deployment details.

CASE STUDY: QSAR MODEL FOR MUTAGENICITY

LiSIs was used to create a QSAR model for Mutagenicity to predict mutagenic and non-mutagenic compounds.

The process of creation and validation of QSAR models in LiSIs can be summarized in four distinct steps:

Datasets Loading on LiSIs Platform

Two datasets are needed for training a QSAR model. A dataset containing chemical information of the compounds either in SMILES or SDF format (see “SMI/SDF File Reader”), and a dataset containing the biological information of the compounds (see “Property File Reader”).

Table 1. Comparison of free scientific workflow management systems used in virtual screening process.

	KNIME	Taverna	LiSIs
System Details			
Software Platform	KNIME	Taverna + myExperiment	Galaxy (Modified)
OS Requirements	Cross Platform	Cross Platform	Linux
Web based	No (Desktop based)	No (Desktop based)	Yes
Cluster deployment difficulty	Moderate (Need license)	Moderate	Moderate
Cloud deployment difficulty	High	Moderate	Low
Open Source	Yes	Yes	Yes
Tool development difficulty	Moderate	Moderate	Low
Tools Details			
Chemoinformatics Packages	CDK, RDKit, OpenBabel, Indigo, EMBL-EBI, Vernalis, Enalos, etc.	CDK	RDKit, In-house tools
Machine Learning Tools	Weka, R	R	R
Docking Tools	Available with commercial license	Not Available	AutoDock Vina
2D/3D Visualization Tools	Available	Available	Not Available
Community size	Very Large	Very Large	Large (Galaxy)

Chemical and/or Structural Descriptors Calculation

Chemical and structural descriptors are the features of the compounds. LiSIs provides a tool to calculate a specific set of chemical descriptors (see “Descriptor Calculator”) and a tool for calculating a specific set of structural descriptors (fingerprints) (see “Fingerprint Calculator”).

1. Model(s) training and validation:

During the training process the algorithm used strives to correlate the calculated descriptors for each compound with its biological or chemical property/activity. Training is usually followed by validation; the process by which the robustness and prediction performances of the QSAR model(s) are established. LiSIs performs those two processes in tandem, while at the same time tries to optimize the algorithm that does the prediction for its main tuning parameter. To achieve this LiSIs employs learning algorithms provided by the R environment. LiSIs currently supports four tools to create a QSAR model based on different algorithms:

a. k-Nearest Neighbours:

Description: A compound is classified by a majority vote of its k nearest neighbours.

Tuning variable: Number of neighbours (k).

b. Support Vector Machines:

Description: An SVM model is a representation of the training data as points in space, mapped so that the different compound classes are divided by a gap that is as wide as possible. New compounds are projected into that same space and predicted to belong to a class based on which side of the gap they are placed.

Tuning variable: Soft margin (C).

c. item Decision Trees:

Description: Recursive partitioning builds a decision tree that uses several dichotomous dependent variables to try and classify chemical compounds.

Tuning variable: Complexity (cp).

d. Random Forests:

Description: Random forests are an ensemble method that, during training, builds a large number of decision trees that utilize a specified number of random descriptors. The final prediction is the mode of the predictions by the individual trees.

Tuning variable: Number of variables randomly sampled as candidates at each split (mtry).

Cross-validation is a model validation technique to assess how the results of a predictive model will generalize to an independent data set. In general, cross-validation partitions the sample into training and test sets using the former to build the model and the latter to assess its performance. The procedure is performed multiple times and the final validation results are the average of the repeats. Alternatively, LiSIs offers the option to create bootstraps out of the original data as training sets while using the original data as the test set. The available algorithms provide different cross-validation options, such as:

- a. Bootstrapping
- b. 0.632+ bootstrapping: An improvement of the classic bootstrapping designed to correct the bias introduced by including data points of the test set into the training set.
- c. k-Fold Cross-Validation
- d. k-Fold Cross-Validation done multiple times
- e. Leave One out Cross-Validation
- f. Leave Group out Cross-Validation: This method is repeated splitting of the data into training and test sets (without replacement).

Best model selection can be performed by using one of the available tuning metrics:

- a. Accuracy
- b. Area Under the Curve
- c. Cohen’s Kappa
- d. Sensitivity
- e. Specificity

1. Model(s) Annotation and Publishing:

Once the QSAR model is created and validated the user has the option to make it publicly available to the rest of the LiSIs’ users. The model can then be utilized by the “Property Predictor” tool to filter-out compounds predicted to lack or possess specific properties. The user can mix and match several models in a side-chain fashion to substantially reduce the number of compounds that will be tested *in vitro* and to enhance the enrichment ratio of the original data set. Prior to publishing the user is encouraged to annotate the QSAR model with some essential info such as the data source, validation performance, the classifying algorithm and the value of its tuning parameter, etc.

The procedure described above was used to train and validate four QSAR models for Mutagenicity, using DT, kNN, RF and SVM algorithms respectively, from which we later selected the best.

The input dataset used was the AMES Mutagenicity dataset available at [52] that consists of 6512 molecules, out of which 3503 are mutagenic (positive) and 3009 non-mutagenic (negative).

Fig. (2) illustrates the workflow that has been used to train the QSAR model. This workflow creates four QSAR models using the all available algorithms. The models use the same train and test datasets, and specific configuration depending on the algorithm used in each case. At the end the workflow provides an aggregated report in order to identify the best QSAR model. When the user decides which of the four models is the best then he/she can annotate it accordingly.

Fig. (3) illustrates the workflow that should be used when you want to predict mutagenicity using this Mutagenicity QSAR model. The key point in this workflow is that the steps of “Descriptor Calculator” and “Fingerprint

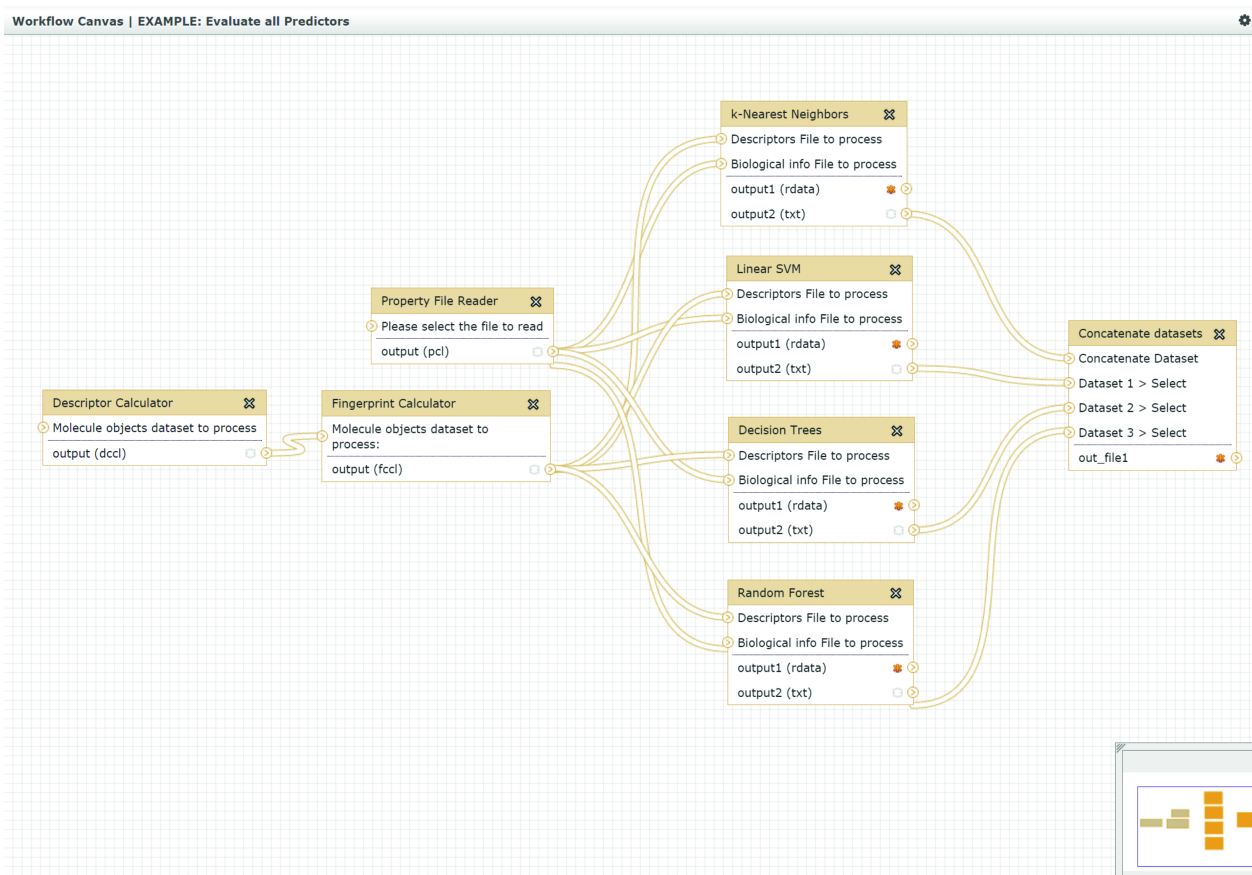


Fig. (2). LiSIs workflow for case study: “QSAR model for Mutagenicity”.

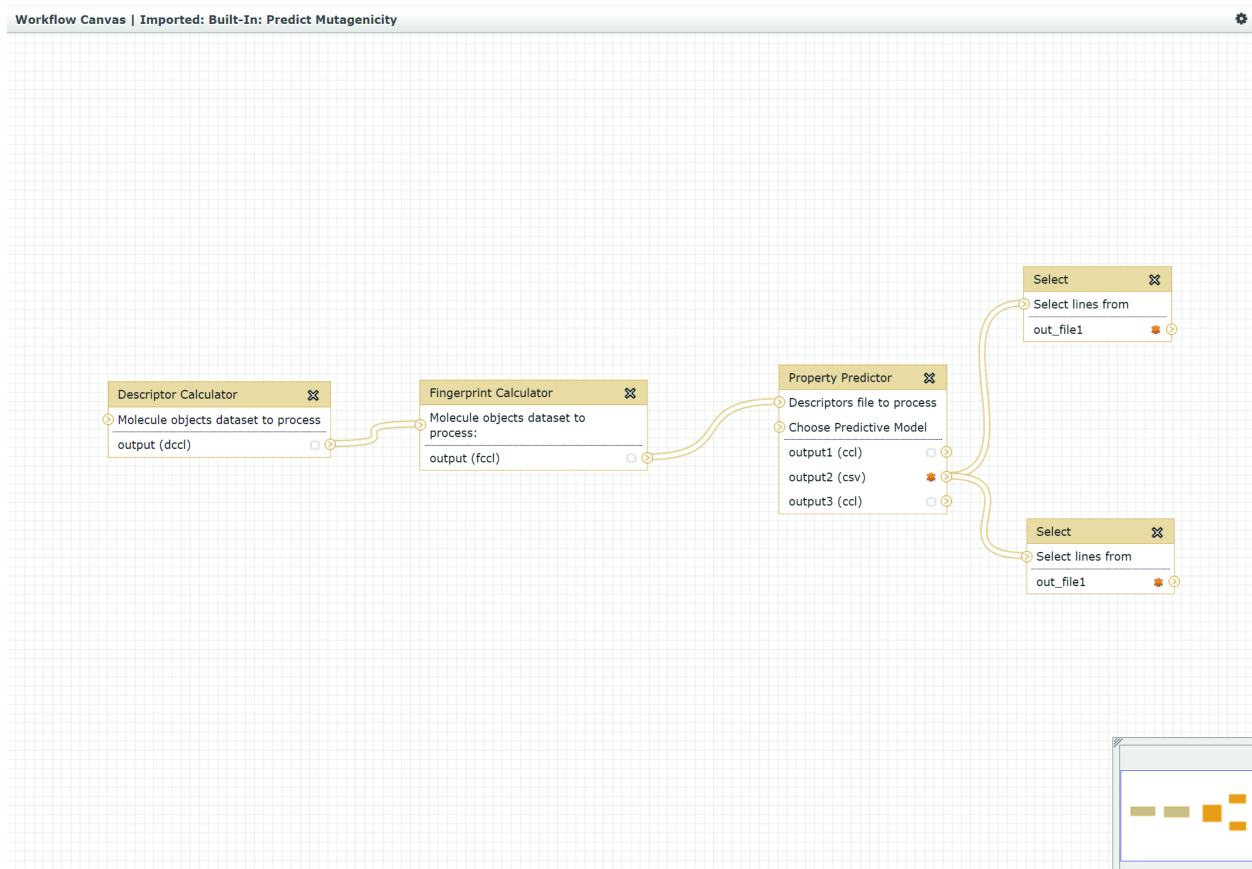


Fig. (3). LiSIs workflow for predicting mutagenicity used in case study: “QSAR model for Mutagenicity”.

Calculator” must have the same input parameters as the ones used at the same steps in the QSAR model training workflow illustrated in Fig. (2). At the step “Property Predictor” the second input should be the previously trained QSAR model for Mutagenicity.

Table 2 shows the final details of the trained Mutagenicity QSAR model using LiSIs alongside with the details of the reference mutagenicity QSAR model by Hansen *et al.* [53].

CONCLUDING REMARKS

LiSIs provides an easy and simple way to train and validate QSAR models using the server’s underlying R installation. Despite the fact that it provides a limited range of classification algorithms, the ones available are the most commonly used.

In this case study we used LiSIs to show its ability and potential in training and using QSAR models. For the purpose of this case study we used the AMES Mutagenicity dataset to train a Mutagenicity QSAR model and compare it with the one proposed by Hansen *et al.* in [53].

As shown in Table 2 the QSAR model for Mutagenicity trained with LiSIs is comparable with the one proposed by Hansen *et al.*, despite the fact that the descriptors used are obtained from different software.

Case Study: Identify Natural Compounds Able to Bind to Estrogen Receptor- α (ER- α) and/or Estrogen Receptor- β (ER- β)

LiSIs has been used for the implementation of a VS experiment in order to identify natural compounds able to

bind to Estrogen Receptor- α (ER- α) and/or Estrogen Receptor- β (ER- β).

Fig. (4) illustrates the complete workflow used by LiSIs for the showcase described. At the Input Layer, parsing of the input datasets takes place. To start with the initial datasets in SMILES format include 2414 compounds from Indofine chemical company [56], 55 compounds characterized by Medina-Franco *et al.* [57] and 21 known ER ligands retrieved from PubChem [58], shown in Table 3, which were used as a positive control dataset for the validation of docking tools. Tools were used to read chemical input files and create compound object structures for further processing by the Pre-Processing and Processing Layers. The total number of unique compounds pushed to the next layer were 2413 from Indofine (one was found to contain erroneous molecular information), 54 from Medina-Franco (two were found to be similar) and 21 from PubChem’s ER agonists and antagonists (one was found to contain two disconnected fragments), datasets for a total of 2488 compounds.

At the Pre-Processing Layer (see Fig. 4), a set of physiochemical molecular descriptors were calculated including Molecular Weight, Hydrogen Bond Donors, Hydrogen Bond Acceptors, Topological Polar Surface Area and Octanol - Water Partition coefficient (cLogP).

At the Processing Layer, the following tools were used:

- *GRANATUM Rule of Five (Ro5) filter* (see Fig. 4 Processing Layer): Molecular Weight between 160 and 700, Hydrogen Bond Donors less or equal to 5, Hydrogen Bond Acceptors less or equal to 10, Topological Polar Surface Area less than 140, and Octanol - Water Partition coefficient (cLogP) between -0.4 and 5.6.

Table 2. Mutagenicity QSAR models comparison, LiSIs versus reference.

Mutagenicity QSAR Model Properties		
Description	A model for predicting mutagenicity of compounds	
Dataset Details	6512 compounds, 3503 mutagenic, 3009 non-mutagenic	
Classes	Mutagenic (positive), Non-mutagenic (negative)	
	LiSIs	Reference QSAR Model by Hansen <i>et al.</i> A
Chemical Descriptors	Molecular Weight, Hydrogen Bond Acceptors, Hydrogen Bond Donors, cLogP, Topological Surface Area, Molecular Complexity, Number of Rings, Molecular Flexibility	Molecular descriptors were selected from blocks 1, 2, 6, 9, 12, 15, 16, 17, 18, and 20 of DragonX version 1.2 B based on a 3D structure generated by Corina version 3.4 C
Fingerprint Descriptors	Morgan (circular) Fingerprints Size: 512 bits Format: Bit-vector Radius: 3 Includes chemical features	Not Used
Algorithms Used	Support Vector Machines	Support Vector Machines
	Decision Tree	Gaussian Process
	Random Forests	Random Forests
	k-Nearest Neighbours	k-Nearest Neighbours
Algorithm with best performance	Random Forest	Support Vector Machines
Performance	Sensitivity = 0.82	Sensitivity = 0.88
	Specificity = 0.80	Specificity = 0.64

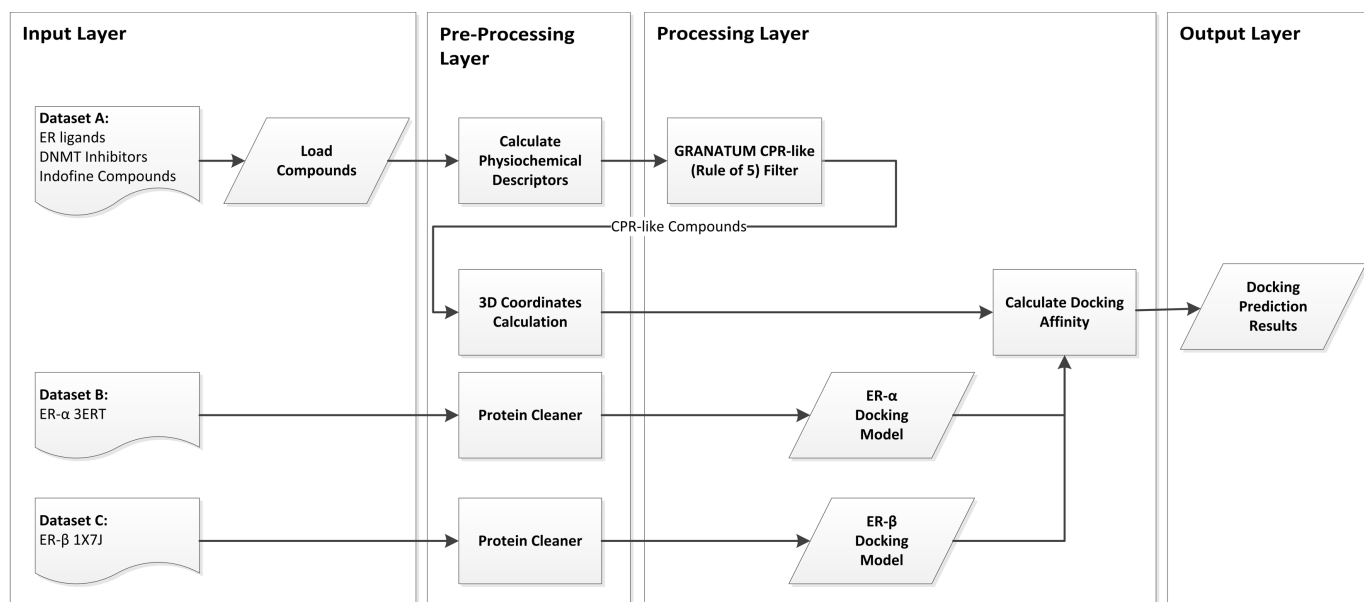
A. [53].

B. [54].

C. [55].

Table 3. Known ER ligands used as positive controls for the validation of the *in silico* results.

A/A	Estrogen Ligand	Docking Score ER- α	Docking Score ER- β
1	Raloxifene	-11.70	-8.72
2	Lilly-117018	-11.53	-3.80
3	3-HydroxyTamoxifen	-11.02	N/A
4	Nafoxidine	-10.88	N/A
5	ICI-182780	-10.73	N/A
6	Pyrolidine	-10.04	N/A
7	Clomiphene A	-10.01	N/A
8	Nitrofinene Citrate	-9.87	N/A
9	ICI-164384	-9.82	-9.13
10	Moxestrol	-9.38	-9.77
11	Naringenine	-8.55	-7.80
12	Triphenylethylene	-8.50	N/A
13	Afema	-8.15	-7.78
14	Danazol	-6.99	N/A
15	Ethamoxytripheto	-6.67	N/A
16	4-HydroxyTamoxifen	-6.60	N/A
17	Dioxin	-6.22	N/A
18	Estralutin	-5.86	-3.80
19	Cyclopentanone	-4.88	N/A
20	Miproxifene Phosphate	-4.48	N/A
21	EM-800	N/A	N/A

**Fig. (4).** LiSis workflow for case study: “Identify natural compounds able to bind to Estrogen Receptor- α (ER- α) and/or Estrogen Receptor- β (ER- β)”.

This filter was defined by CPR experts participating to the GRANATUM project [8].

The filtering resulted in 1834 compounds with CPR-like features and 654 compounds without CPR-like features. The compounds with CPR-like features were pushed for docking experiments.

- *Docking experiment against ER- α and ER- β (see Fig. 4 Processing Layer):*

LiSIs uses AutoDock Vina [36, 37] and has been setup to provide us with the maximum docking affinity score. The current key aim of the GRANATUM project was to identify ER- α antagonists and ER- β agonists. Docking experiments on the filtered combined dataset have been performed by employing receptors ER- α 3ERT [PDB:3ERT] and ER- β 1X7J [PDB:1X7J]. The appropriate Docking Models were created using protein structures obtained from the PDB database [59] and related LiSIs tools for automated Protein Cleaning (see Fig. 4 Pre-Processing Layer) and Docking Model Preparation.

Fig. (5A) is a graphical representation of the docking affinity score predicted by LiSIs docking experiment tool for ER- α , and Fig. (5B) is a graphical representation of the docking affinity score predicted by LiSIs docking experiment tool for ER- β . The predicted binding affinity scores of the known ER inhibitors (see Table 3), depicted with red colour in Fig. (5), indicate the validity of the docking models prepared and the ability of these models to assign a lower score to inhibitors and reproduce ground truth. The cyan dots represent DNMT inhibitors characterized in [57]. The lower (most negative) the value of the docking score is, the higher the binding affinity. Consequently, the models are applicable in a VS context, i.e. for the prioritization of unknown compounds based on their predicted binding affinity to estrogen receptors.

Finally a selection of molecules highly ranked was hand-picked; a small sample of those is shown in Table 4. These molecules have undergone *in vitro* investigation to provide feedback for the calibration of the tools used by LiSIs platform and also to select a small set for further research.

As shown in Table 4, three novel flavones, 3',4'-dihydroxy-a-naphthoflavone (Compound 2), 3,5,7,3',4'-pentahydroxyflavanone (Compound 5), and 4'-hydroxy-a-naphthoflavone (Compound 6) were among those with high binding scores for ER- α and ER- β as indicated from the *in silico* docking score. Flavones, a class of flavonoids, have previously been demonstrated to possess estrogenic activity in a number of hormonally responsive systems. Their estrogenic and antiestrogenic activities appear to correlate directly with their capacity to displace Estradiol from ER [60]. Our *in vitro* results showed that Compound 2 had the highest affinity for both receptors while Compound 5 also displayed similar affinity for both ER- α and ER- β . However Compound 6 was found to bind only weakly to ER according to the binding affinity assay. Furthermore, results from the *in silico* experiments showed that three previously not investigated coumarins, 3(2'-chlorophenyl)-7-hydroxy-4-phenylcoumarin (Compound 3), 3(3'-chlorophenyl)-7-hydroxy-4-phenylcoumarin (Compound 4) and 4-benzyl-7-hydroxy-3-phenylcoumarin (Compound 7) can potentially bind ER- α and ER- β based in their docking scores. Coumarins are natural or synthetic benzopyranic derivatives that form a family of active compounds with a wide range of pharmacological properties, including estrogen-like effects [61]. *In vitro* results showed that Compound 3 has greater affinity for ER- α while Compound 4 can bind with high affinity to both receptors. However, Compound 7 was not able to bind to either receptor as determined by the ER binding affinity assay.

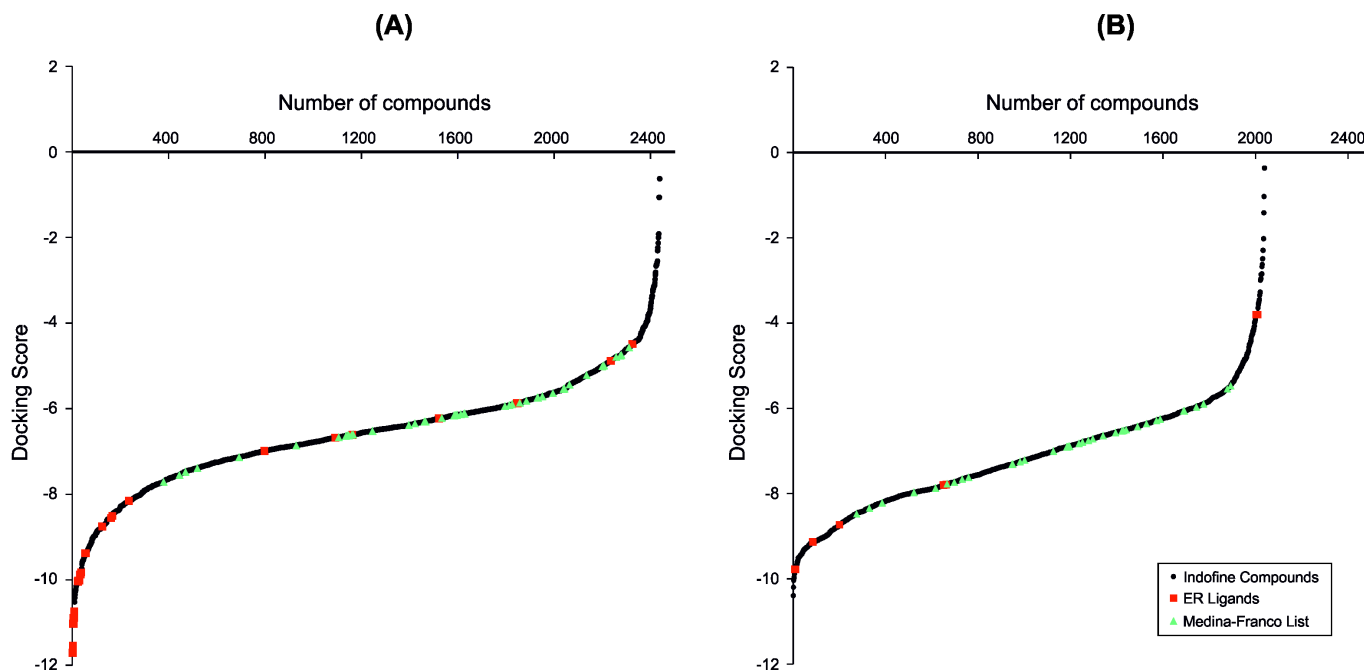
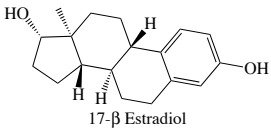
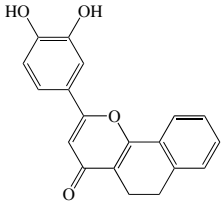
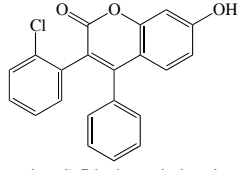
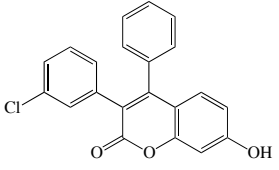
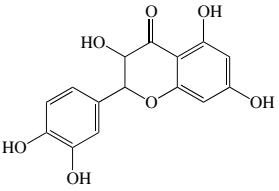
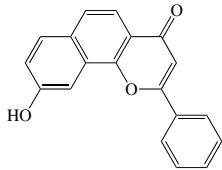
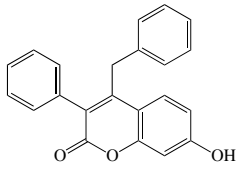


Fig. (5). Compounds were tested against ER- α (A) and ER- β (B) using *in silico* docking tools.

Table 4. Selection of highly ranked compounds from the final virtual screening results.

A/A	Chemical Structure	Molecular Weight (g/mol)	Concentration (μ M)	ER- α LDB		ER- β LDB	
				Binding Affinity	Docking Score	Binding Affinity	Docking Score
1	 17- β Estradiol	272.38	10	1	-9.4	1	-10
2	 3',4'-dihydroxy-a-naphthoflavone	304.29	1 10	0.11 0.22	-7.59	0.05 0.34	-10.39
3	 3(2'-chlorophenyl)-7-hydroxy-4-phenylcoumarin	348.78	1 10	0.21 2.71	-9.73	N/A 0.34	-10.03
4	 3(3'-chlorophenyl)-7-hydroxy-4-phenylcoumarin	348.78	1 10	0.24 2.23	-10.34	0.13 2.75	-9.67
5	 3,5,7,3',4'-pentahydroxyflavanone	304.26	1 10	N/A 0.27	-8.81	0.06 0.18	-9.61
6	 4'-hydroxy-a-naphthoflavone	228.29	1 10	N/A N/A	-8.18	0.05 N/A	-9.88
7	 4-benzyl-7-hydroxy-3-phenylcoumarin	328.37	1 10	N/A N/A	-10.13	N/A N/A	-9.13

CONCLUDING REMARKS

In recent years, many high-throughput methods have been established in the effort to identify novel Estrogen Receptor binders with anticancer activity. However, *in vitro* assays often produce disappointing results due to the small percentage of novel active Estrogenic compounds discovered. To identify novel compounds that act as effective ER- α co-activator binding inhibitors (CBIs), Gunther *et al.* applied a time-resolved fluorescence resonance energy transfer (TR-FRET) assay developed in a 384 well format [62]. This assay measures the binding of a Cy5-labelled SRC-1 nuclear receptor interaction domain to the ligand binding domain (LBD) of labelled ER- α leading to TR-FRET signal generation. Compounds that interfere with the TR-FRET signal are identified as potential CBIs or conventional ligand antagonists. Based on this method, only 1.6% of the total compounds screened were identified as active as reported in (Pubchem ID 629).

In the present study, we used a VS workflow implemented using the LiSIs platform to screen the Indofine database of 2413 compounds. Based on their drug-like criteria and docking results we selected 18 potential ER ligands. These were further investigated *in vitro* with the ER binding assay described by Gurer-Orhan *et al.* [63] with minor modifications. In this manner it was found that five agents displayed strong affinity for ER- α , three showed selectivity for ER- β and seven were able to bind to both receptors with similar affinity. In total 15 out of 18 compounds (83.3%) were experimentally confirmed active. Therefore, the use of LiSIs system may allow researchers to execute complex biomedical studies and *in silico* experiments on largely available and high quality data repositories in order to facilitate the selection and prioritize the investigation of novel chemopreventive compounds *in vitro*.

Compounds with high binding affinity to the ERs based on the *in silico* results, display structural characteristics that are similar to Estradiol-17 β (E2). All contain a phenolic ring which is indispensable for binding to the estrogen receptor [64]. The phenolic ring of Compounds 2 - 7 contains at least one hydroxyl group which mimics the 3'-OH of E2. Furthermore, all compounds have low molecular weight comparable to that of E2 (Molecular Weight equal to 272). All agents are highly hydrophobic which is required for binding in the ER binding pocket [65]. The differences observed in the binding affinities of compounds may be attributed to differences in structural characteristics. The lower ER binding affinity of Compound 5 (when compared to Compound 2) may be attributed to the hydrophilic hydroxyl group at C-11 of Compound 5 which, due to steric hindrance, lowers its binding affinity for both receptors [65].

The LiSIs platform aims to fill the current void in the application of advanced chemoinformatics and computational chemistry technology in determining efficacy and predicting possible mechanism of action or identifying a possible receptor for a chemopreventive agent in life sciences research. Its successful deployment may have a substantial impact on enabling biomedical 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. We have shown in this paper that by utilizing the LiSIs platform in conjunction to a widely used docking program we identified compounds that can bind to ER- α and/or ER- β with a high degree of success rate. This *in silico* approach is expected to facilitate the process of identification of lead compounds with estrogenic or anti-estrogenic activity and to enhance considerably the discovery process for new therapeutic agents.

CONCLUSION

In recent years, scientific workflow systems have been increasingly used by the chemoinformatics community. Several systems, both commercial and free, have been introduced with custom components catering to the needs of the drug discovery community and numerous applications have been described in the literature. In this paper we introduced LiSIs, a new SWMS platform designed and implemented to provide advanced computational chemistry and information technology tools in an online environment.

LiSIs enables the use of state of the art computational algorithms and techniques to design and implement solutions by reusing open source community tools such as RDKit [29], R [44] and AutoDock Vina [36, 37] to complement in-house code. Consequently, LiSIs users have access to numerous methods that enable operations such as molecular descriptor calculation, predictive model generation and use and docking experiments, among others. It is worth noting that, due to its online nature, LiSIs models, workflows and results can be easily shared with other platform users.

Future work will focus on current system limitations. Indicatively, LiSIs has limited visualization capabilities which need to be augmented and, certain components, such as the protein cleaner tool, need to be further expanded. System scalability is also a concern since the current infrastructure, which is limited to a single server with 12 processing cores and 16 GB of memory, will inevitably become a bottleneck as the system becomes more popular. We also intend to incorporate tools implementing multi-objective ranking and optimization methods to LiSIs in order to enable the consideration of multiple pharmaceutically important properties to VS and other library design experiments [66].

The potential of LiSIs has been highlighted through two case studies designed to illustrate the preparation of tools for the identification of promising chemical structures. The first case study involved the development of a QSAR model on a literature dataset while the second implemented a docking-based virtual screening experiment. Our results show that VS workflows utilizing docking, predictive models and other *in silico* tools as implemented in the LiSIs platform can identify compounds in line with expert expectations. For example, our experiments provided a chemical structure set that, once experimentally tested, was found to bind to ER- α and/or ER- β with a high degree of success as computationally predicted.

Moreover, the ability to readily share knowledge between researchers in the form of models, workflows, experimental data and results was found to be an additional beneficial feature that facilitated collaboration between distributed partners and, thus, the generation of new knowledge.

In conclusion, we have shown that LiSIs facilitates the discovery of cancer chemopreventive agents. We anticipate that the deployment of LiSIs, as currently implemented and available for use, can enable drug discovery researchers to more easily use state of the art computational techniques in their search for promising chemical compounds.

LiSIs is available as a web based application, accessed directly at [7] and also under the GRANATUM platform at [8].

ABBREVIATIONS

CBI	=	Co-activator Binding Inhibitors;
CPR	=	Chemoprevention research;
DDP	=	Drug Discovery Process;
DT	=	Decision Trees;
ER	=	Estrogen Receptor;
k-NN	=	k-Nearest Neighbours;
KNIME	=	Konstanz Information Miner;
LBD	=	Ligand Binding Domain;
LiSIs	=	Life Sciences Informatics;
PDB	=	Protein Data Bank;
PMML	=	Predictive Model Markup Language;
QSAR	=	Quantitative Structure Activity Relationship;
RF	=	Random Forests;
SDF	=	Structure Data File;
SMARTS	=	SMiles ARbitrary Target Specification;
SMILES	=	Simplified Molecular Input Line Entry Specification;
SW	=	Scientific Workflows;
SWMS	=	Scientific Workflow Management Systems;
TR-FRET	=	Time-Resolved Fluorescence Resonance Energy Transfer;
VS	=	Virtual Screening.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work has been partially supported through the EU-FP7 GRANATUM project, "A Social Collaborative Working Space Semantically Interlinking Biomedical Researchers, Knowledge and data for the design and execution of *In Silico* Models and Experiments in Cancer Chemoprevention", contract number ICT-2009.5.3.

C. C. Kannas, I. Kalvari, I. Kirmizoglou, Z. Antoniou and K. G. Achilleos developed LiSIs' tools and helped on the configuration of LiSIs server. G. Lambrinidis and E. Mikros provided valuable guidance in regards to ligand-protein docking. C. M. Neophytou, C. G. Savva, C. A. Pitta

and D. Scherf were responsible for the 2nd case study and also provided valuable feedback during the development phase. V. J. Promponas, C. A. Nicolaou, C. Gerhauser, A. I. Constantinou and C. S. Pattichis are the scientific experts. All authors contributed in the writing of this article. All authors read and approved the final manuscript.

REFERENCES

- [1] Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H. N.; Sastry, G. N. Virtual Screening in Drug Discovery -- a Computational Perspective. *Curr. Protein Pept. Sci.*, **2007**, *8*, 329-351.
- [2] Kar, S.; Roy, K. How Far Can Virtual Screening Take Us in Drug Discovery? *Expert Opin. Drug Discov.*, **2013**, *8*, 245-261.
- [3] *Virtual Screening: Principles, Challenges, and Practical Guidelines*; Sottriffer, C., Ed.; Methods and Principles in Medicinal Chemistry; 2011.
- [4] Barker, A.; Hemert, J. V. Scientific Workflow: A Survey and Research Directions. In: *Proceedings of the 7th international conference on Parallel processing and applied mathematics*; PPAM'07; Springer-Verlag: Berlin, Heidelberg, **2008**; pp. 746-753.
- [5] Yang, X.; Bruin, R. P.; Dove, M. T. Developing an end-to-end scientific workflow: a case study using a comprehensive workflow platform in e-science. *Comput. Sci. Eng.*, **2010**, *12*, 52-61.
- [6] Kannas, C. C.; Achilleos, K. G.; Antoniou, Z.; Nicolaou, C. A.; Pattichis, C. S.; Kalvari, I.; Kirmizoglou, I.; Promponas, V. J. A Workflow System for Virtual Screening in Cancer Chemoprevention. In: *2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE)*; **2012**; pp. 439-446.
- [7] LiSIs Life Sciences Informatics platform <http://lisis.cs.ucy.ac.cy/> (accessed Oct 24, **2014**).
- [8] GRANATUM - Project Vision <http://granatum.org/> (accessed Oct 24, **2014**).
- [9] Sporn, M. B. Approaches to prevention of epithelial cancer during the preneoplastic period. *Cancer Res.*, **1976**, *36*, 2699-2702.
- [10] Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Molecules*, **2002**, *7*, 566-600.
- [11] Hull, D.; Wolstencroft, K.; Stevens, R.; Goble, C. A.; Pocock, M. R.; Li, P.; Oinn, T. Taverna: A Tool for Building and Running Workflows of Services. *Nucleic Acids Res.*, **2006**, *34*, W729-W732.
- [12] Oinn, T.; Greenwood, M.; Addis, M.; Alpdemir, M. N.; Ferris, J.; Glover, K.; Goble, C.; Goderis, A.; Hull, D.; Marvin, D.; Li, P.; Lord, P.; Pocock, M. R.; Senger, M.; Stevens, R.; Wipat, A.; Wroe, C. Taverna: Lessons in Creating a Workflow Environment for the Life Sciences. *Concurr. Comput. - Pract. Exp.*, **2006**, *18*, 1067-1100.
- [13] Missier, P.; Soiland-Reyes, S.; Owen, S.; Tan, W.; Nenadic, A.; Dunlop, I.; Williams, A.; Oinn, T.; Goble, C. Taverna, Reloaded. In: *Scientific and Statistical Database Management*; Gertz, M.; Ludäscher, B., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, **2010**; Vol. 6187, pp. 471-481.
- [14] Berthold, M.; Cebon, N.; Dill, F.; Gabriel, T.; Kötter, T.; Meinel, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In: *Data Analysis, Machine Learning and Applications*; Springer Berlin Heidelberg, 2008; pp. 319-326.
- [15] Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz Information Miner: Version 2.0 and beyond. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 26-31.
- [16] Pipeline Pilot is Accelrys' scientific informatics platform <http://accelrys.com/products/pipeline-pilot/> (accessed Jul 16, **2012**).
- [17] IDBS - InforSense Suite - analytical data management <http://www.idbs.com/products-and-services/inforsense-suite/> (accessed Jul 16, **2012**).
- [18] Jorgensen, W. L. The many roles of computation in drug discovery. *Science*, **2004**, *303*, 1813-1818.
- [19] Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (qsar): a review. *Comb. Chem. High Throughput Screen.*, **2006**, *9*, 213-228.
- [20] Krovat, E. M.; Steindl, T.; Langer, T. Recent Advances in Docking and Scoring. *Curr. Comput. - Aided Drug Des.*, **2005**, *1*, 93-102.
- [21] Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. evaluation of different

- docking/scoring combinations. *J. Med. Chem.*, **2000**, *43*, 4759-4767.
- [22] Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. Large-scale virtual screening for discovering leads in the postgenomic era. *IBM Syst. J.*, **2001**, *40*, 360-376.
- [23] Lyne, P. Structure-based virtual screening: an overview. *Drug Discov. Today*, **2002**, *7*, 1047-1055.
- [24] Nicolaou, C. A.; Brown, N. Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.*, **2013**, *10*, e427-e435.
- [25] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **1997**, *46*, 3-26.
- [26] Anderson, A. C.; Wright, D. L. The Design and docking of virtual compound libraries to structures of drug targets. *Curr. Comput. - Aided Drug Des.*, **2005**, *1*, 103-127.
- [27] Achilleos, K. G.; Kannas, C. C.; Nicolaou, C. A.; Pattichis, C. S.; Promponas, V. J. Open Source Workflow Systems in Life Sciences Informatics. In: *2012 IEEE 12th International Conference on Bioinformatics Bioengineering (BIBE)*; **2012**; pp. 552-558.
- [28] Simmhan, Y. L.; Plale, B.; Gannon, D. A survey of data provenance in e-science. *ACM SIGMOD Rec.*, **2005**, *34*, 31-36.
- [29] Landrum, G. RDKit: Open-source cheminformatics <http://www.rdkit.org/> (accessed Jul 17, **2012**).
- [30] Barone, R.; Chanan, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 269-272.
- [31] Morgan, H. L. The Generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Inf. Model.*, **1965**, *5*, 107-113.
- [32] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273-1280.
- [33] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64-73.
- [34] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, **1987**, *27*, 82-85.
- [35] Daylight Fingerprints <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html>.
- [36] Trott, O.; Olson, A. J. AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **2010**, *31*, 455-461.
- [37] AutoDock Vina - molecular docking and virtual screening program <http://vina.scripps.edu/> (accessed Jul 17, **2012**).
- [38] Kuhn, M. Building predictive models in r using the caret package. *J. Stat. Softw.*, **2008**, *28*, 1-26.
- [39] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics*, **2011**, *3*, 33.
- [40] Gardine, B.; Riemer, C.; Hardison, R. C.; Burhans, R.; Elnitski, L.; Shah, P.; Zhang, Y.; Blankenberg, D.; Albert, I.; Taylor, J.; Miller, W.; Kent, W. J.; Nekrutenko, A. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **2005**, *15*, 1451-1455.
- [41] Goecks, J.; Nekrutenko, A.; Taylor, J.; Galaxy Team, T. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **2010**, *11*, R86.
- [42] Blankenberg, D.; Kuster, G. V.; Coraor, N.; Ananda, G.; Lazarus, R.; Mangan, M.; Nekrutenko, A.; Taylor, J. Galaxy: a web-based genome analysis tool for experimentalists. In: *Current Protocols in Molecular Biology*; John Wiley & Sons, Inc., **2010**.
- [43] O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.*, **2008**, *2*, 5.
- [44] R. Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, **2008**.
- [45] De Roure, D.; Goble, C.; Bhagat, J.; Cruickshank, D.; Goderis, A.; Michaelides, D.; Newman, D. myExperiment: Defining the Social Virtual Research Environment. In *eScience, 2008. eScience '08. IEEE Fourth International Conference on*; IEEE, **2008**.
- [46] KNIME | Konstanz Information Miner <http://www.knime.org/> (accessed Jul 16, **2012**).
- [47] Data Mining Group (DMG) <http://www.dmg.org/>.
- [48] Abouelhoda, M.; Alaa, S.; Ghanem, M. Meta-Workflows: Pattern-Based Interoperability between Galaxy and Taverna. In *Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science*; Wands '10; ACM: New York, NY, USA, **2010**; pp. 2:1-2:8.
- [49] myExperiment <http://www.myexperiment.org/>.
- [50] Afgan, E.; Baker, D.; Coraor, N.; Chapman, B.; Nekrutenko, A.; Taylor, J. Galaxy Cloudman: Delivering Cloud Compute Clusters. *BMC Bioinformatics*, **2010**, *11*, S4.
- [51] Taverna - open source and domain independent Workflow Management System <http://www.taverna.org.uk/> (accessed Jul 16, **2012**).
- [52] Benchmark Data Set for In Silico Prediction of Ames Mutagenicity <http://doc.mli.tu-berlin.de/toxbenchmark/>.
- [53] Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.*, **2009**, *49*, 2077-2081.
- [54] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim; New York, **2008**.
- [55] Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1000-1008.
- [56] INDOFINE Chemical Company, Inc., NJ Chemicals, Los Angeles County Chemicals, NJ Molecules, Los Angeles County Molecules, Xian, Shijiazhuang, Shanghai, China, Call 908-359-6778 <http://www.indofinechemical.com/> (accessed Jul 18, **2012**).
- [57] Medina-Franco, J. L.; López-Vallejo, F.; Kuck, D.; Lyko, F. Natural Products as DNA Methyltransferase Inhibitors: A Computer-Aided Discovery Approach. *Mol. Divers.*, **2010**, *15*, 293-304.
- [58] The PubChem Project <http://pubchem.ncbi.nlm.nih.gov/> (accessed Jul 18, **2012**).
- [59] RCSB Protein Data Bank <http://www.rcsb.org/pdb/home/home.do>.
- [60] Collins-Burow, B. M.; Burow, M. E.; Duong, B. N.; McLachlan, J. A. Estrogenic and Antiestrogenic Activities of Flavonoid Phytochemicals Through Estrogen Receptor Binding-Dependent and -Independent Mechanisms. *Nutr. Cancer*, **2000**, *38*, 229-244.
- [61] Jacquot, Y.; Laños, I.; Cleeren, A.; Nonclercq, D.; Bermont, L.; Refouvelet, B.; Boubekeur, K.; Xicluna, A.; Leclercq, G.; Laurent, G. Synthesis, Structure, and Estrogenic Activity of 4-Amino-3-(2-Methylbenzyl)coumarins on Human Breast Carcinoma Cells. *Bioorg. Med. Chem.*, **2007**, *15*, 2269-2282.
- [62] Gunther, J. R.; Du, Y.; Rhoden, E.; Lewis, I.; Revennaugh, B.; Moore, T. W.; Kim, S. H.; Dingledine, R.; Fu, H.; Katzenellenbogen, J. A. A Set of Time-Resolved Fluorescence Resonance Energy Transfer Assays for the Discovery of Inhibitors of Estrogen Receptor-Coactivator Binding. *J. Biomol. Screen.*, **2009**, *14*, 181-193.
- [63] Gurer-Orhan, H.; Kool, J.; Vermeulen, N. P. E.; Meerman, J. H. N. A novel microplate reader-based high-throughput assay for estrogen receptor binding. *Int. J. Environ. Anal. Chem.*, **2005**, *85*, 149-161.
- [64] Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, J. A. The Estradiol Pharmacophore: Ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids*, **1997**, *62*, 268-303.
- [65] Agatonovic-Kustrin, S.; Turner, J. V. Molecular Structural Characteristics of Estrogen Receptor Modulators as Determinants of Estrogen Receptor Selectivity. *Mini-Rev. Med. Chem.*, **2008**, *8*, 943-951.
- [66] Nicolaou, C. A.; Kannas, C. C. Molecular library design using multi-objective optimization methods. In: *Chemical Library Design*; Zhou, J. Z., Ed.; Methods in Molecular Biology; Humana Press, **2011**; pp. 53-69.