Process-Driven Information Management System at a Biotech Company: Concept and Implementation

Alberto Gobbi,* Sandra Funeriu, John Ioannou, Jinyi Wang, and Man-Ling Lee*

Anadys Pharmaceuticals, Inc., San Diego, California 92121

Chris Palmer and Bob Bamford

Managed Ventures, LLC., Irvine, California 92618

Robin Hewitt

Hewitt Consulting, San Diego, California 92150

Received November 17, 2003

While established pharmaceutical companies have chemical information systems in place to manage their compounds and the associated data, new startup companies need to implement these systems from scratch. Decisions made early in the design phase usually have long lasting effects on the expandability, maintenance effort, and costs associated with the information management system. Careful analysis of work and data flows, both inter- and intradepartmental, and identification of existing dependencies between activities are important. This knowledge is required to implement an information management system, which enables the research community to work efficiently by avoiding redundant registration and processing of data and by timely provision of the data whenever needed. This paper first presents the workflows existing at Anadys, then ARISE, the research information management system developed in-house at Anadys. ARISE was designed to support the preclinical drug discovery process and covers compound registration, analytical quality control, inventory management, high-throughput screening, lower throughput screening, and data reporting.

1. INTRODUCTION

Today's drug discovery relies heavily on electronic information management, because in every stage of this process more and more data are produced. The data need to be stored, organized, and connected to other data to become an adequate source of information. Present-day drug discovery also involves an increasing number and diversity of groups that both create data and consume it. Therefore, information systems have evolved from being simply a database with the corresponding data retrieval application into complex applications with many interrelated modules.

Previously, papers discussing general concepts of drug discovery information systems[1−4] and specific modules[5−7] have been published. This paper describes processes and concepts as well as an implementation of a complete system. As such, it may be an example for other companies with similar needs.

Within the drug discovery process, multiple series of interactions between different groups of scientists take place sequentially or in parallel. These activities are diverse and so are the data needed by these groups as well as the data these groups generate. Results of the operations of a group will affect the decisions and activities of other groups. Therefore, timely distribution of information between the groups is essential. With the introduction of high-throughput (HT) operations, such as HT screening or combinatorial synthesis, the amount of data becomes more than can be managed by exchanging paper documents or electronic files.

Newly founded biotech companies have to design their Information Management System from the ground up. They have to decide whether to acquire a commercial off-the-shelf solution[8] or to integrate individual in-house or external applications.[9] In reality, the implementations are variations between these two extremes. That is because human and financial resources as well as development time have to be taken into account. The system of choice will vary hugely depending on available know-how, the needs of the company, and the company's financial situation. In contrast to established companies, new companies do not need to integrate with existing applications. This allows for more flexibility in the implementation.

In any case, it is essential to investigate the overall process and the activities involved prior to the design of the information management system. Thorough understanding of processes and data flows is essential to streamline the physical work and to avoid redundancy in data processing as much as possible. This understanding is also critical for defining appropriate program-to-program or user-to-program interfaces, a prerequisite for designing modular systems consisting of well-integrated programs and applications.

This paper presents the information management system designed and implemented at Anadys Pharmaceuticals. This system supports scientists throughout the discovery process

* Corresponding authors phone: +1 (858) 530 3657; e-mail: agobbi@anadyspharma.com (A.G.) and phone: +1 (858) 530 3658; e-mail: mlee@anadyspharma.com (M.-L.L.).

10.1021/ci034269o CCC: $27.50 © 2004 American Chemical Society
Published on Web 03/12/2004

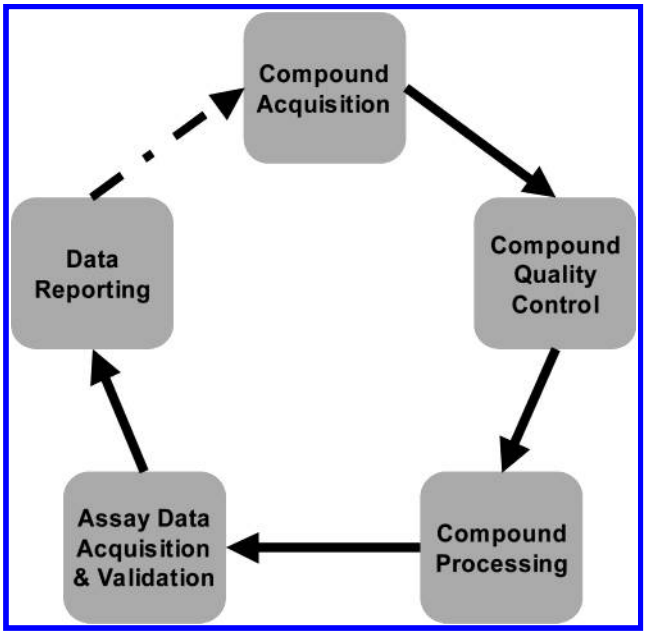


Figure 1. The interdepartmental process connections.

from HT screening, through lead confirmation and on through hit optimization. This includes reagent and compound library management as well as assay data upload and reporting.

A requirement for us was to have a system that is flexible, easy to maintain, and easy to extend, because drug discovery is an ever-changing process, and there will always be new areas of the discovery process that have to be supported. We did this by creating a highly modular system that allows us to change one piece without impacting others.

This paper summarizes the processes at Anadys and then gives an overview of the general implementation and of the individual modules. Finally, it presents an estimation of invested resources.

2. PROCESS ANALYSIS

Within Anadys, we identified five primary data-driven processes. These are

1. Compound Acquisition (and Registration)
2. Compound Quality Control
3. Compound Processing (Compound Library Management)
4. Assay Data Acquisition and Validation
5. Reporting

As shown in Figure 1, these processes are connected and, in the real world, form an optimization loop where results from one cycle feed into and improve the next cycle. The order within the cycle reflects dependencies between processes and data flow that the Anadys information system has to handle. The key issues of concern are collecting data and providing these data to subsequent processes.

Compound Acquisition activities fall within several main categories: (1) ordering, tracking, and registering reagents for medicinal chemistry, (2) registering and tracking single compounds synthesized by medicinal chemists for screening, (3) enumerating and registering libraries synthesized in house intended for screening, and (4) selecting, acquiring, and registering commercial screening compounds on plates. Not only are these activities diverse, moreover they are carried out by different people. Therefore, efficient support represents a challenge to the information management system, because it must bring together all relevant data for each task.

Compound Quality Control is considered a separate process. Supporting this process required the integration of analytical instruments with each other and with the database. In the case of in-house synthesized compounds, analytical data are required for registration to provide evidence of structure and purity. Commercial compounds, i.e., reagents and screening compounds, are registered without comprehensive analytical results because purity is guaranteed by the vendors. However, statistical quality control is done for commercial screening compounds, and the corresponding data are uploaded afterward. Because today's analytical laboratories are instrument parks generating a large amount of data, automated uploading of relevant data for reports and for archiving is mandatory.

Compound Processing activities are those activities carried out in the company's Compound Library department. They are manifold and diverse, e.g. creation of stock solutions and plates, compressing and replicating plates, and distribution of these compounds and plates upon request. The high number of items that have to be processed at once makes accurate tracking of these items crucial.

Once compounds are tested, assay data must be collected and made available to the research community. We separate Assay Data Acquisition from Data Reporting because there are various groups that need to see the results but are not involved in generating assay data. HT screening is a complex experiment on a huge number of samples. Support is needed both to upload the data into the database and for data validation.

The last of the five processes, Reporting is, in a sense, the most important of all. Reporting is the ultimate goal of any information system. It provides the user with access to the results of all the other processes and allows the company to extract value from the work done in the various groups. Individual scientists need reporting to keep track of their work, and project teams need data to drive the projects. The emphasis in Reporting is on flexibility, usability, and performance. Export functionality is important to allow further manipulation and dissemination.

A considerable amount of time was spent in close interaction with scientists to define and, wherever possible, standardize the workflow. However, in a dynamic research-driven company, flexibility will always be an important consideration. In the following sections, we describe key aspects of ARISE. These should be seen as examples and not as the only possible approach. This is especially true since the system's purpose is to optimally support Anadys' specific workflow and requirements.

3. IMPLEMENTATION OVERVIEW

System Architecture. ARDS, the Anadys Research Data Store, was designed to be the central data repository for Anadys' Research department. All the compounds and related data are stored here and kept for reports or data mining. The centralized storage avoids redundancies. Data needs to be entered only once and is then available to all other members of the company. It also enables queries over multiple domains. For example, one can search for compounds in the

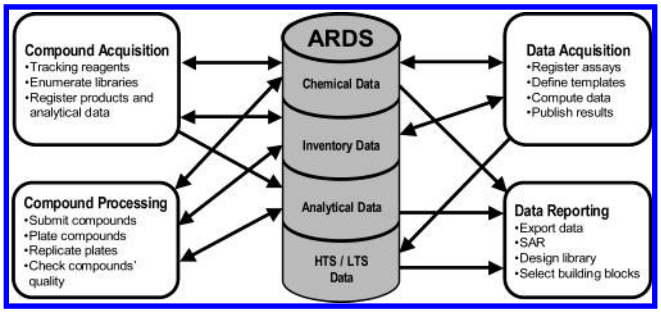


Figure 2. Architecture of the ARISE system. The arrows indicate the direction of the dataflow related to the activities (tasks) listed in the boxes.

Compound Library and the Chemistry Inventory, because the border between screening compounds and chemicals is not always strict.

Modules are in place to capture, process, and retrieve data. Users access them through the Web browser. Modules are usually grouped by processes and, if possible, organized according to the company's workflows. Direct database access is restricted to the development team. Figure 2 shows the architecture and data flow between the modules and the database.

Database Design. The data captured in the present system are chemical, analytical, and assay data. Chemical data includes not only structure and structure-related information about a compound but also "logistic" information about where the batches of the compound came from and where the existing samples are stored. The "chemistry" tables reflect these three tiers of compound data, i.e., structure, batch, and sample related data. These tables represent the core of the database. Analytical and screening data are linked to chemical data, preferably at the sample level. This ensures that the exact history of a sample including source or synthesis as well as distribution can be traced back in case of inconsistencies in experimental data.

High-throughput (HT) and low-throughput (LT) screening data are not stored in the same tables. This is due to the different number of data points created in each type of experiment and to the variety of assays. The high data volume produced in HT screens requires automation of data upload and quality control. Because the HTS workflow is well characterized and standardized, only a limited set of tables is necessary to store HTS data. Many more varieties of assays and detection methods are run, however, in LT mode.

To handle this variety there are two possible solutions: (1) create new tables whenever the data associated with a result differs from other experiments and (2) create a single generic vertical table where each parameter of a single result is stored in a separate row. These possibilities have been analyzed in depth by Miled et al.[10] We decided to use the first alternative because retrieval of one result is simpler, requiring only a single join. This simplicity also results in improved retrieval performance.[10]

Compound-related data, such as biological results, is usually generated for a given sample, and the exact sample history is important to be able to backtrack unexpected results. For Structure Activity Relationship (SAR) analysis however the relevant relationship is between experimental results and structure. The sample table is linked to the structure table via the batch table. To prevent queries for SAR involving many structures from being slow, the ARISE data model uses one denormalization: All tables storing experimental results include foreign keys not only to the sample but also to the batch and structure tables. This results in very quick retrieval for data queried at any level. To prevent data inconsistencies, triggers have been implemented to update the denormalized keys whenever there is a change.

Screening Data Collection. Screening data, in contrast to chemistry data, are very diverse, because of the huge number of assays and their variations. Additionally, data that are of interest have to be derived from raw data produced by the instruments. Therefore, decisions have to be made about which data should be captured.

In case of HT screening data, data down to the level of raw data are captured. This is essential for the automated computation of processed data and allows recomputation, if necessary. In case of LT screening data, the benefit of capturing low-level data are much smaller given the higher varieties of assays, instruments, and workflows. Therefore, we decided to collect only the computed final data, e.g. IC50 and percent inhibition.

Due to limited resources, data for those LT screens that produce data for many compounds are uploaded first. In vivo screens, which are done on a very small number of compounds, have a lower priority and have not been loaded yet. However, discussions about storage format were initiated to ensure that data stored locally is in an appropriate format for future uploads.

Software Components. The applications were designed around the Tomcat[11] application server by the Apache Jakarta Project.[12] Tomcat is open source software.[13] Further improvements are under active development and support is available by very active mailing lists.[14] Tomcat is the reference implementation for the Java Servlet[15] and Java Server Pages[16] technologies, which allow the easy development of fast and platform independent Web applications in Java. The ARISE applications also rely heavily on the huge pool of available open source Java modules:

- For graphing the PTPlot 2D data plotter package[17] from the UC Berkley is used.
- The batik[18] and FreeHEP[19] packages are very useful to generate graphics in PNG and EMF formats.
- For XML parsing and transformations the XERCES,[20] XALAN,[21] and JDOM[22] API's are used.
- For part of the user interface the Struts framework[23] is used.

On registration the chemical structures are normalized using struchk[24] by B. Rohde, and the canonical isomeric SMILES[25] is generated using the Daylight toolkit[26] to ensure uniqueness. Structure queries are entered using the ChemDraw plug-in.[27] Chemical structures in reports are displayed using smi2pict.[28] Display and analysis of NMR-spectra is done using NUTS.[29] The Daylight SMILES, SMARTS, and Fingerprint toolkits[26] were used along with in-house developed Java classes to provide chemical intelligence where needed.

Because most users are familiar with Microsoft Excel, and because its formatting, computing, and scripting features allow for a high degree of customization, Excel is extensively used for the preparation of the input data. For most of the data-upload modules, Excel templates are used. Some include

Table 1. List of Most Important Modules in ARISE

module	process	primary user	description
reagent ordering	compound acquisition	chemists, reagent manager	Supports reagent selection, including supplier and package size. Creates purchase orders.
reagent tracking	compound acquisition	chemists, reagent manager	Supports check-in and check-out of reagent bottles into reagent stockroom.
Jlab	compound acquisition	high-throughput chemists	Enumerates parallel synthesis products.
single compound registration	compound acquisition	chemists, reagent manager	Registers single screening compounds and reagents.
batch compound registration	compound acquisition	high-throughput chemists, compound library manager	Registers compound libraries.
DISE	compound acquisition	specialist	Selects a diverse set of screening compounds.
NMR spectra capturing	quality control	chemists	Automatically captures NMR spectra at spectrometer.
NMR spectra retrieval	quality control	chemists	Retrieves and analyses NMR spectra from compound report.
compound library management modules	compound processing	compound library manager	Supports plating, plate compression, plate replication and cherry-picking of compounds.
ATHENA	high-throughput screening	high-throughput screener	Supports data acquisition, quality control, data analysis and reporting of high throughput screening results.
low-throughput screening dataloader	low-throughput screening	biologist, biochemists	Supports data upload of low-throughput screening results.
SimpleSearch	reporting	all scientists	Retrieves compound-related data by ID or structure.
QueryBuilder	reporting	all scientists	Enables complex queries including constraints on structure, properties and biological results.
ListHandling	reporting	all scientists	Stores and retrieves hit-lists from QueryBuilder and SimpleSearch. Does set operations on hit-lists.

simple Excel functions for data reformatting. After the user has entered the results, a simple cut and paste into a Web page transfers the data to ARISE.

Although there are several high quality open source database systems,[30,31] we have chosen Oracle version 8.1.7, Standard Edition.[32] In addition to being widespread in the industry, Oracle provides the Extensible Indexing Interface,[33] which enabled the implementation of a chemical substructure and similarity search cartridge.[34] The cartridge is implemented in Java, which then is compiled into C code running inside the database process. This results in very high performance chemical search capabilities directly from within the SQL[35] query language.

Hardware Components. The production hardware for ARISE consists of just two servers, the database server and the application server. The database server has two CPUs running at 866 MHz, 4GB of RAM and a total of 400 GB of available disk space. The disks are configured as RAID5 and attached by a PERC3 hardware raid controller, which ensures failure-safety and performance. This server is running on the 6.2 enterprise version of the Red Hat Linux operating system since this is the only Red Hat Linux version on which Oracle 8.1.7.3 was certified. The application server has two CPUs at 2.8 GHZ, 2GB of memory, and 140 MB of available disk space through a PERC3 raid controller. For this server Red Hat Advanced Server 2.1 is used. Both versions of Linux provide a certified and stable operating system. Backups are done using an ADIC Scalar 220 and the excellent AMANDA open source backup application.[36] For testing and development a third server is used to run copies of the database so that test and development do not interfere with production data.

4. MODULE DESCRIPTION

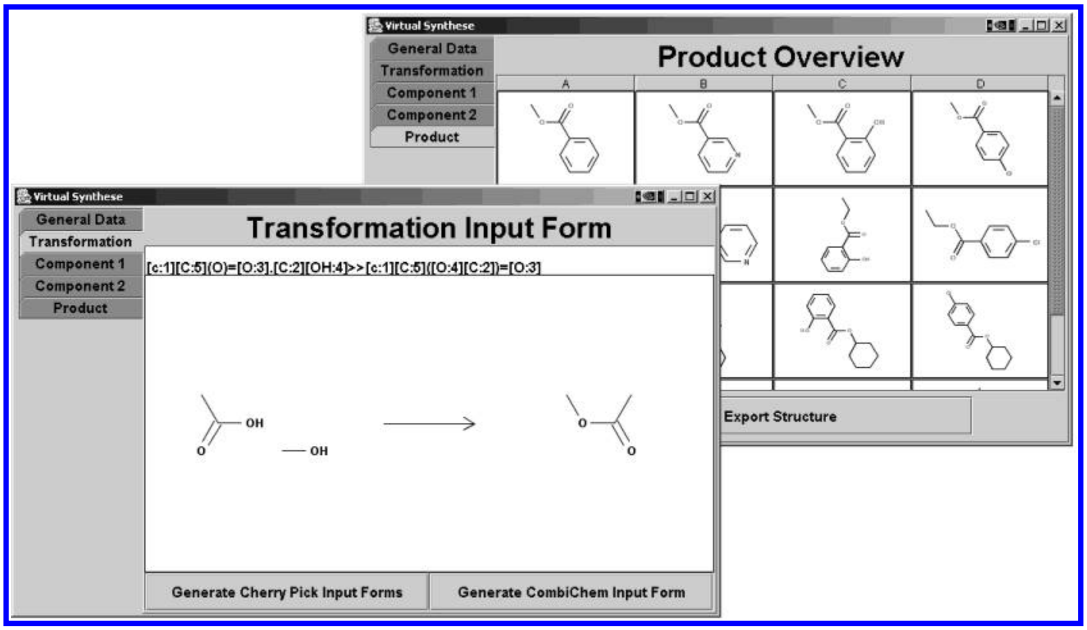
The following sections give descriptions of the most important modules that have been released to date. The modules are presented in the context of the processes for which they are used. The descriptions do not include details of the implementation. The goal is to point out the most important features and the interfaces between the modules. Table 1 gives an overview of all modules.

Compound Acquisition. The Compound Acquisition modules support reagent ordering and tracking, structure enumeration, synthesis, and compound registration. Depending on the requirements, these modules are implemented as Web applications, standalone programs, or Excel add-ins. Some applications produce temporary data, which are then used as input for follow-up modules. These temporary data are usually stored in Excel spreadsheets.

The decision to implement task-oriented modules and integrate them to a program suite was driven by the need to support three different workflows with similar tasks: screening-compound acquisition, classical compound synthesis, and high-throughput medicinal syntheses. The medicinal chemists' workflow varies with synthesis method. The system's modularity has helped us adapt quickly to workflow changes. Using reusable modules also reduced the initial implementation effort and has made routine maintenance easier.

Reagent Ordering. The Reagent Ordering application consists of two separate modules. The first one is Web-based. It requires a list of MFCD numbers, which are usually exported from ACD Finder[37] and the input of supplier, price, and package size preferences. After submission of this information, the program generates a customized Excel worksheet with available compounds and associated catalog data (package sizes and vendors). The chemists select the items to be ordered from this list.

A Visual Basic add-in generates the purchase order for the Purchasing department as well as an Excel worksheet with all the selected items for the person responsible for receiving and registering chemicals. Because the data in the worksheet are in the format required by the Compound Registration application, only the barcodes of the received chemicals have to be added for registration.



[c:1][C:5](O)=[O:3].[C:2][OH:4]>>[c:1][C:5]([O:4][C:2])=[O:3]

Figure 3. JLab, the structure enumeration application: Shown here are the tabs for generic reaction input and product structure display.

Reagent Tracking. A simple Web module is used for reagent tracking. As input, it takes a user ID, a list of barcodes, and a new location for the reagents. Then, the vials' current locations in the database are changed to the new location. The user ID of the person making the change is also recorded. The user interface allows the user to mark samples as depleted.

Structure Enumeration (JLab). JLab was implemented as a prototype of a Java application for reaction-based enumeration of structures. To guide the user through the multistep procedure, the input forms and reports are organized as tabs in a window. At first, only tabs to enter synthesis information and the generic reaction are available.

After submission, JLab will generate new tabs for each component in the generic reaction. The application can operate in one of two modes: combinatorial mode or cherry-pick mode. In combinatorial mode, the program expects lists of components for combinatorial enumeration of the products. In cherry-pick mode, a set of components is entered for each product. The user has complete control over the synthesis layout in this mode. Once finished with these inputs, the user presses the "Generate Product" button to initiate enumeration. When the enumeration finishes, another tab with the product structures appears including a button to export the data to Excel (Figure 3).

Since JLab is still a prototype, the accepted input format is limited. Generic reactions have to be entered as SMIRKS,[38] which can be obtained from the ChemDraw molecule editor.[27] The reagent inputs are SMILES with or without a corresponding compound ID. This input is easily obtained from the ARISE reports.

The exported Excel worksheet with products contains structure depictions, SMILES, and additional data: the exact mass, the molecular weight, the products' notebook references, and reagent IDs. The exact mass is needed for mass spectroscopy. The reagent IDs are used for tracking during synthesis with the IRORI[39] synthesizer. The other data are required for registration. The order of the columns is such that no rearrangement is needed for any of these three tasks.

The format of the exported Excel worksheet is the input format required by the batch registration program. JLab automatically creates empty columns for additional information that the chemists have to supply such as barcodes of the vials, amounts, purities, and other optional information for registration.

Compound Registration. We have implemented two different alternatives to register compounds. One is a Web-based form where compounds are registered one by one. Structures are entered in the ChemDraw Plug-in and the data input fields are standard HTML form elements. Since the master input form is a JSP page, customization for registration of both screening compounds and reagents is easy.

The second alternative is a command-line program, used for batch registration, which can handle large numbers of compounds.[40] An SD file and a configuration file are required as input. The latter is needed to map the fields in the SD file to the corresponding fields in the database. The separation of mapping and upload is especially important for registration of commercial compounds, because the format of the SD files differs from vendor to vendor.

Both registration programs call a set of PL/SQL stored procedures to upload the structure and compound-related data. These procedures must be used for compound registration, to ensure that structures are correct and unique. Ambiguous structure fragments, like the nitro group, are transformed according to the company's business rules. The other important function of the stored procedures is the correct assignment of the Anadys Number, the company's unique compound ID.

Compound Selection. Acquisition of compounds for high-throughput screening is an important task for a biotech company. The task is to select tens of thousands of compounds from millions of commercially available compounds, such that the compounds are diverse and optimally suited for screening. This is done using a command line application DISE. The DISE algorithm was described previously.[41]

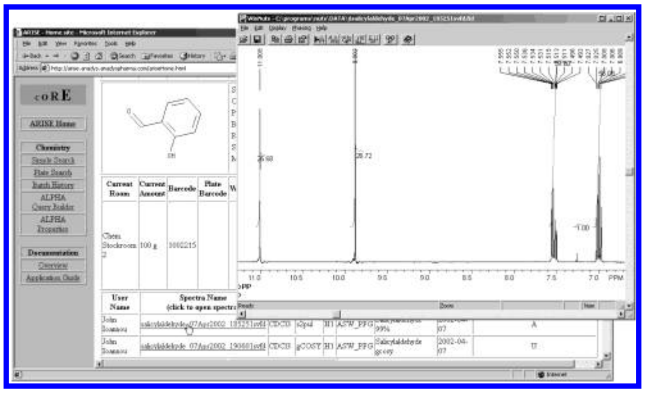


Figure 4. Integration of the NUTS software for NMR spectra analysis. Clicking on the link in the ARISE user interface opens up the NUTS software with the corresponding FID.

Analytical Data Integration. ARISE includes a module that integrates NMR spectra into the database. This allows users to access NMR spectra from their desktop within 30 min after they are taken.

The spectra acquisition is done on a Varian 400 MHz spectrometer. The samples are handled with a SMS500 Auto Sampler.[42] Chempack 2.1[43] was customized to guide users through the submission of a sample to the auto sampler. Macros in Varian's macro language query users for the lab-journal reference. The lab-journal reference, spectra parameters, and a link to the raw FID file are stored in ARISE. Because the lab-journal reference uniquely identifies a synthesized compound, it is used to link the FID to the compound. A mime type was defined on the users' workstations to associate the FID with the NUTS software. Thus, whenever a user would like to inspect a spectrum he/she can click on the HTML link in the compound report to open the NUTS software with the FID for further processing (Figure 4). A similar integration of the results from other analytical techniques such as HPLC and LC/MS is planned for a future release of ARISE.

Compound Library Management. The activities carried out by the Compound Library management are manifold. Fortunately, most of the workflows are standardized. Figure 5 shows a flowchart for processing plates. In-house synthesized compounds or commercial compounds purchased specifically for projects are submitted in vials to the Compound Library. They are directly transferred to 384-well plates (Plating). Compounds acquired as 96-well plates undergo plate compression to be in 384-well plates. Replications and further dilutions result in assay plates (APLs), which are distributed upon request to the HTS group. Compounds for hit confirmation are taken from the original vials or from the 96-well source plates (SPL) and are processed as shown in the flowchart.

To keep track of the plates and vials and to keep the database up-to-date, the Compound Library management has to register all new plates and update the amounts of parent samples. Despite the differences in activities, i.e., plating, compression, replication, cherry picking, and creation of titration plates, only two programs are needed to handle all plate registrations. The DataLoader module, which was initially written to upload assay data (see section on Low-Throughput Screening), is used to register plates produced in the plating, cherry pick, and titration activities. A new configuration file was written to enable DataLoader to perform the required task. A second program was written for the registration of replicated or compressed plates. This program treats all wells the same and requires only source and destination plate barcodes as input.

Both programs work with any plate size or sample concentration. Instead of entering the new absolute concentration, the user enters a dilution factor. This is more consistent with the workflow and allows for variations in

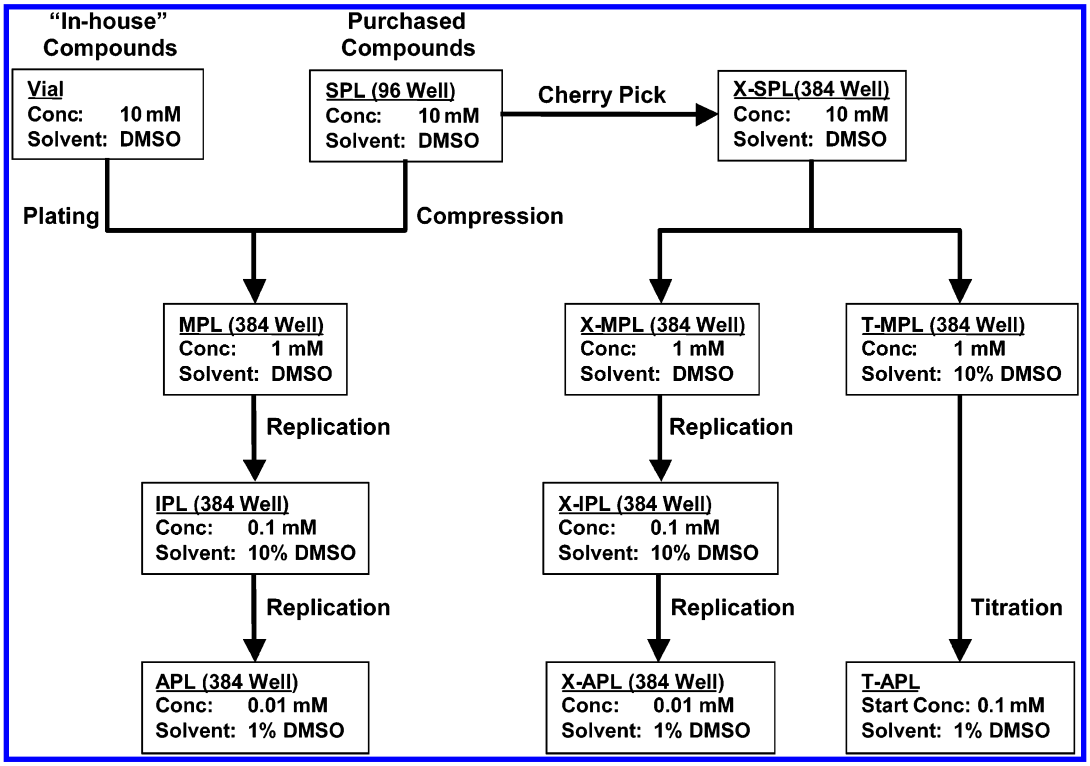
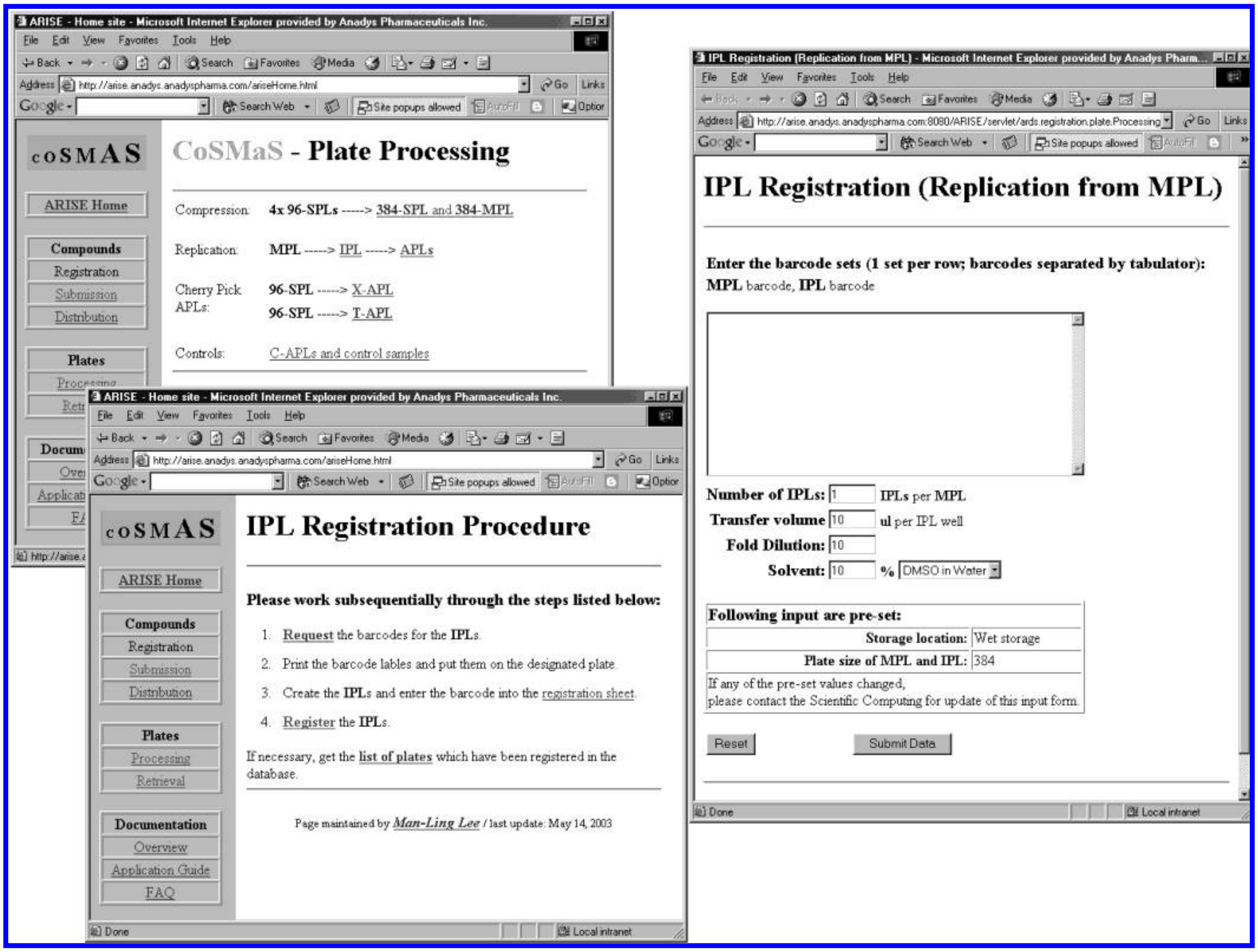


Figure 5. Flowchart of current Plate Processing at Anadys.


ARISE - Home site - Microsoft Internet Explorer provided by Anadys Pharmaceuticals Inc.
File Edit View Favorites Tools Help
Address http://arise.anadys.anadyspharma.com/ariseHome.html
coSMAS
ARISE Home
Compounds
Registration
Submission
Distribution
Plates
Processing
Retrieval
Documentation
Overview
Application Guide
FAQ
IPL Registration (Replication from MPL) - Microsoft Internet Explorer provided by Anadys Pharm...
Address http://arise.anadys.anadyspharma.com:8080/ARISE/servlet/ards.registration.plate.Processing
Reset
Submit Data
Done
Local intranet

CoSMaS - Plate Processing

Compression: 4x 96-SPLs -----> 384-SPL and 384-MPL

Replication: MPL -----> IPL -----> APLs

Cherry Pick APLs: 96-SPL -----> X-APL
96-SPL -----> T-APL

Controls: C-APLs and control samples

IPL Registration Procedure

Please work subsequentially through the steps listed below:

1. Request the barcodes for the IPLs.
2. Print the barcode lables and put them on the designated plate.
3. Create the IPLs and enter the barcode into the registration sheet.
4. Register the IPLs.

If necessary, get the list of plates which have been registered in the database.

Page maintained by Man-Ling Lee / last update: May 14, 2003

IPL Registration (Replication from MPL)

Enter the barcode sets (1 set per row; barcodes separated by tabulator):
MPL barcode, IPL barcode

Number of IPLs: 1 IPLs per MPL
Transfer volume 10 ul per IPL well
Fold Dilution: 10
Solvent: 10 % DMSO in Water

Following input are pre-set:

Storage location:	Wet storage
Plate size of MPL and IPL:	384

If any of the pre-set values changed,
please contact the Scientific Computing for update of this input form.

Figure 6. Some plate processing Web pages: (1) Workflow overview page with links to specific module sets, (2) IPL registration procedure with links to related modules, and (3) IPL registration input form.

the parent sample's concentration. The modules are accessible through easy to customize HTML or JSP pages shown in Figure 6. The entry page lists the steps for plate processing. The workflow-oriented menu makes it obvious when to use which module. Changes in the workflow mostly require only simple changes in the user interface.

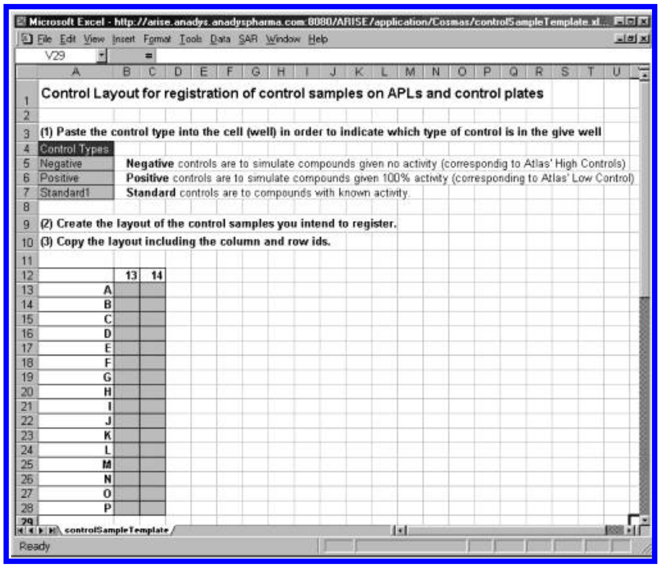
A small Java module was written to register the position of control wells on the assay plates. The module will accept any layout provided the given wells are not already associated with compounds. The layout is created in Excel (Figure 7) and then pasted into a textbox on a Web registration form.

Screening Data Acquisition and Analyses. As mentioned in the Introduction, the processes in high- and low-throughput screening are very different. Therefore, they are supported by different modules.

High-Throughput Screening and the ATHENA Application. The Anadys HTS system, ATHENA[44] consists of the following modules:

- Data Acquisition and Normalization
- Quality Control and Data Correlation
- Dose Response Computation
- Positive Identification

Data Acquisition and Normalization. The format of raw data may vary depending on the equipment used and the experiment performed. The first step in Data Acquisition is an equipment-dependent parser to put the data into a uniform format. The second step, if required, consists of any transformations needed to compute a raw value with linear dependence on inhibition. A small separate data upload application was implemented as a client-side Java application to accomplish these steps. A new piece of equipment or a new experiment type will usually require a modification or


Microsoft Excel - http://arise.anadys.anadyspharma.com:8080/ARISE/application/Cosmas/controlSampleTemplate.xl...
File Edit View Insert Format Tools Data SAR Window Help
V29
controlSampleTemplate
Ready

Control Layout for registration of control samples on APLs and control plates

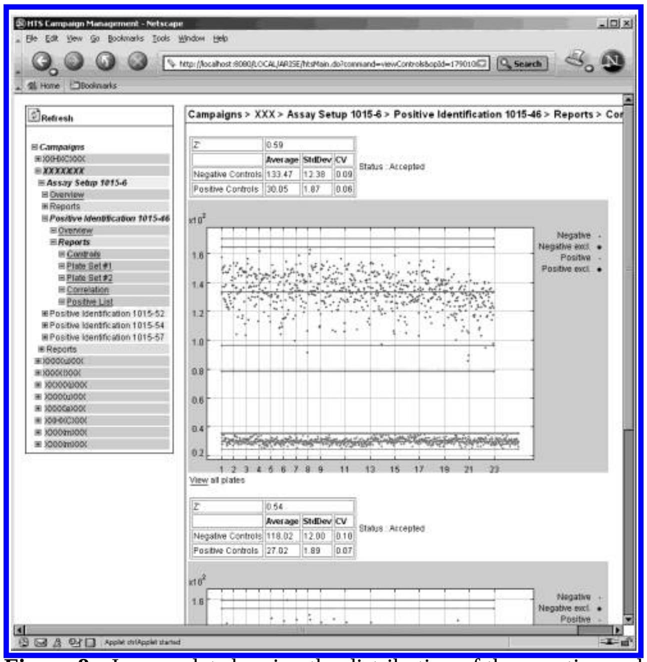
(1) Paste the control type into the cell (well) in order to indicate which type of control is in the give well

Control Types
Negative — **Negative** controls are to simulate compounds given no activity (correspondig to Atlas' High Controls)
Positive — **Positive** controls are to simulate compounds given 100% activity (corresponding to Atlas' Low Control)
Standard1 — **Standard** controls are to compounds with known activity.

(2) Create the layout of the control samples you intend to register.
(3) Copy the layout including the column and row ids.

	13	14
A		
B		
C		
D		
E		
F		
G		
H		
I		
J		
K		
L		
M		
N		
O		
P		

Figure 7. Control-registration template in Excel.


HTS Campaign Management - Netscape
File Edit View Go Bookmarks Tools Window Help
http://localhost:8080/LOCAL/ARISE/htsMain.do?command=viewControls&opId=179010
Search
Home
Bookmarks
Refresh
Campaigns
XXXXXXX
Assay Setup 1015-6
Overview
Reports
Positive Identification 1015-46
Overview
Reports
Controls
Plate Set #1
Plate Set #2
Correlation
Positive List
Positive Identification 1015-52
Positive Identification 1015-54
Positive Identification 1015-57
Reports
Campaigns > XXX > Assay Setup 1015-6 > Positive Identification 1015-46 > Reports > Co
x10²
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23
Negative
Negative excl.
Positive
Positive excl.
View all plates
Applet ctrlApplet started

Z'	0.59		
	Average	StdDev	CV
Negative Controls	133.47	12.38	0.09
Positive Controls	30.05	1.87	0.06

Status : Accepted

Z'	0.54		
	Average	StdDev	CV
Negative Controls	118.02	12.00	0.10
Positive Controls	27.02	1.89	0.07

Status : Accepted

Figure 8. Java applet showing the distribution of the negative and positive controls.

addition of a parser or transformer.

The only parser currently in use is for the Victor plate readers. Transformers are available for luminescence, fluorescence polarization, Time-Resolved Energy Transfer (TRET), and Fluorescence Resolved Energy Transfer (FRET). During the upload procedure scientists enter additional data such as the protocol used in the screen and a screen-dependent set of parameter values. The assay data are automatically linked to all compound-related data by plate barcode and well position.

Quality Control and Data Correlation. The next step is to check the quality of the controls, either plate-by-plate or for the whole set of plates. The average of the negative controls is used to define the 0% line and the average of the positive controls is used to define the 100% line for the computation of percent inhibition values. The difference of these two values and their respective standard deviations (σ) define Z', which is used to evaluate signal-to-noise ratio.[45] Z' should be larger than 0.5:

$$Z' = 1 - 3*\frac{\sigma_{negative} + \sigma_{positive}}{|avg_{negative} - avg_{positive}|}$$

The graph, shown in Figure 8, allows scientists to see the distribution of positive and negative controls. This provides a visual assessment of the signal-to-noise ratio. Outliers can be further analyzed by clicking on them in the graph, which then opens the Plate Viewer shown Figure 9.

The Plate Viewer shows detailed results on a per-plate basis and allows the user to exclude wells where the experiment failed, e.g. due to instrument failure. Excluding a well containing positive or negative controls is important because it affects the percent inhibition of all other wells. Once the exclusions have been completed, the final percent inhibition values are computed. To check for trends over the whole experiment, a Plate Set Viewer (Figure 9) is available. This shows all plates for an experiment so global trends can be identified.

Assay validation experiments are done in duplicate so that reproducibility can be checked. Graphs showing the correlation of results from duplicate experiments are inspected visually. The user also gets a preliminary overview of the expected number of hits by examining a bar graph showing the number of hits given various % inhibition thresholds.

Dose Response Computation. ATHENA has the capability to fit sigmoidal Dose Response curves in batch mode to establish IC50's. This module uses a simplex minimization algorithm[46] to find the optimal values for the four parameter of the sigmoidal curve (Min, Max, Hill, and IC50). The result is presented in an applet (Figure 10). Users can exclude measurement points or overwrite curve parameters.

Positive Identification. ATHENA offers an editable Positive List (Figure 11). After entering a threshold, the application presents all results that are within that input. The display includes the experimental data. The user can exclude single compounds based on overall performance or for any other reason. To do so, an exclusion reason has to be given to ensure that this decision remains comprehensible. The list can be exported to Excel anytime during and after the review process and used for additional analysis or for requesting compounds for further investigation.

Low-Throughput Screening. The variability of LTS experimental protocols can be large and includes experiments with many parameters such as dose, temperature, time, and additional substrates that may or may not be present. Although this variability complicates data acquisition and processing, the value of these data are much greater than the value of HTS data since they are usually higher quality and the compounds tested are in a more advanced stage of the discovery process.

DataLoader. To support LTS, the DataLoader module in ARISE provides an easy-to-use and highly configurable data loading capability. A scientist enters the results from her experiment together with the compound identifier (barcode) into a predefined template in Excel. The completed template is then transferred by copy and paste into a textbox in the DataLoader's Web interface. After data submission, the user gets a result sheet back that includes the result IDs needed for updates.

The DataLoader is configured by XML files located on the server. The template identifies which XML file to use. The XML contains information about column headings and data types. After checking the data types, SQL statements in the XML file are run to check the validity of the entries and to link the entries to other database tables (e.g. the location, user, or compound tables). Finally, the XML file contains a section that defines the insert and update statements that create records in the database. If a new type of experiment needs to be supported, all that needs to be done is to create the required tables, templates, and XML file for the DataLoader.

Reporting. Data acquisition is important, but the ultimate goal is to enable members of the company to retrieve the data they need. A small reporting application that allows users to search by structures, compound IDs, and synonyms (SimpleSearch) was released in conjunction with the compound registration application at the very beginning of the ARISE project. A small selection of predefined reports was provided including a compound data sheet and an Excel

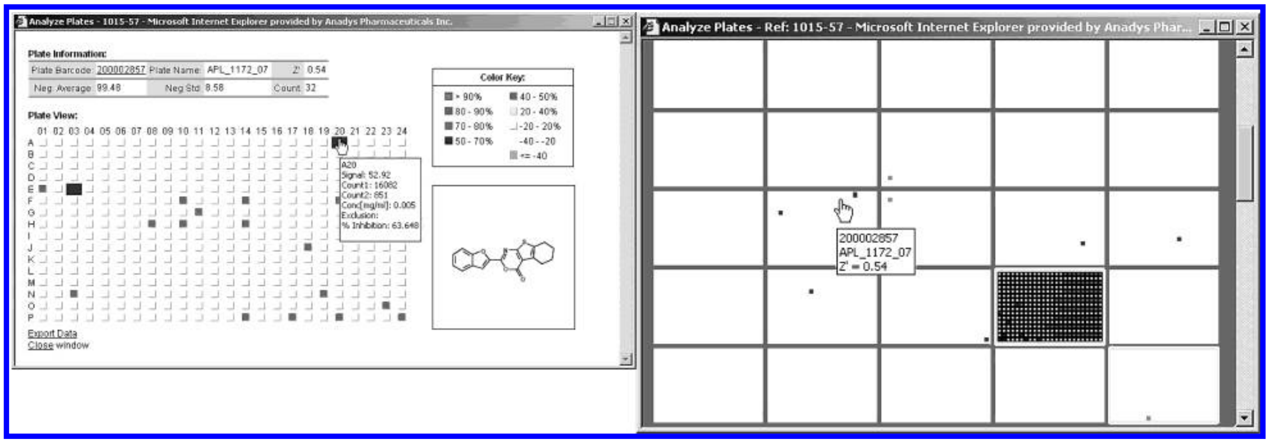

Analyze Plates - 1015-57 - Microsoft Internet Explorer provided by Anadys Pharmaceuticals Inc.
Color Key:
> 90%
80 - 90%
70 - 80%
50 - 70%
40 - 50%
20 - 40%
-20 - 20%
-40 - -20
<= -40
Plate View:
Export Data
Close window
Analyze Plates - Ref: 1015-57 - Microsoft Internet Explorer provided by Anadys Phar...

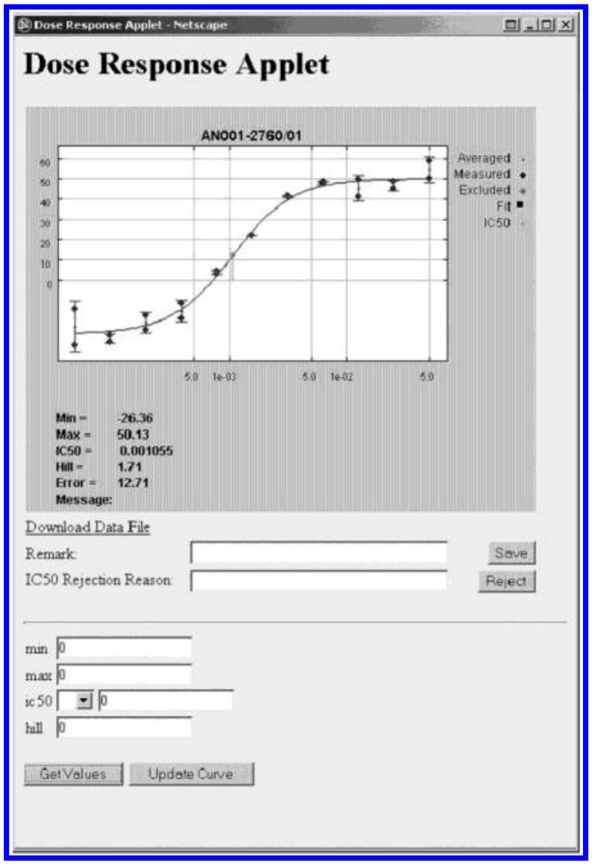
Plate Information:

Plate Barcode:	200002857	Plate Name:	APL_1172_07	Z':	0.54
Neg. Average:	99.48	Neg Std:	8.58	Count:	32

A20
Signal: 52.92
Count1: 16082
Count2: 851
Conc[mg/ml]: 0.005
Exclusion:
% Inhibition: 63.648

200002857
APL_1172_07
Z' = 0.54

Figure 9. The Plate Viewer displays details of the activity distribution on a single plate. The Plate Set Viewer shows all plates in a single view. In this and the following figures, the structures shown have been selected at random from a set of commercial compounds to protect internal discoveries.

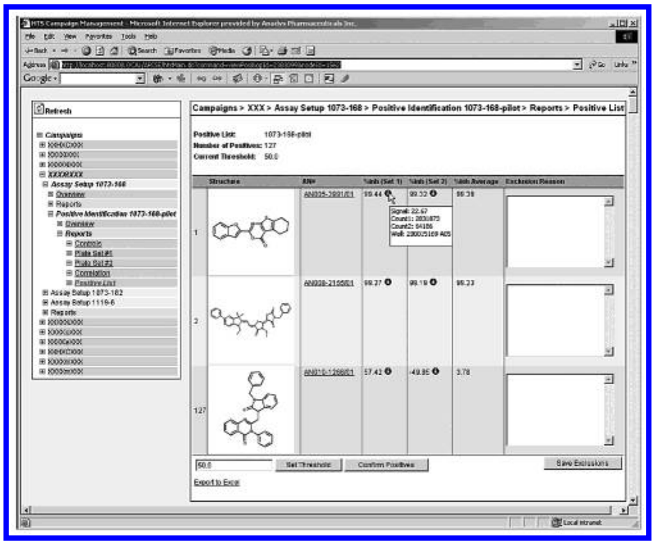


Min = -26.36
Max = 50.13
IC50 = 0.001055
Hill = 1.71
Error = 12.71
Message:

Remark:
IC50 Rejection Reason:

min 0
max 0
ic50 0
hill 0

Figure 10. Dose Response Applet.



Campaigns > XXX > Assay Setup 1073-168 > Positive Identification 1073-168-pilot > Reports > Positive List

Positive List: 1073-168-pilot
Number of Positives: 127
Current Threshold: 50.0

	Structure	AN#	%Inh (Set 1)	%Inh (Set 2)	%Inh Average	Exclusion Reason
1		AN005-3991/01	99.44	99.32	99.38	
2		AN008-2156/01	99.27	99.19	99.23	
127		AN010-1268/01	57.42	-49.85	3.78	

Signal: 22.67
Count1: 2830873
Count2: 54186
Well: 200005169 A05

50.0

Figure 11. Positive List. The picture was modified to show the top two and the last positives.

worksheet containing most of the captured data types for further processing.

The architecture of this reporting application is such that the query module is clearly separated from the display module. The query module is responsible for performing searches in the database and retrieves a list of substance IDs of compounds that match the search criteria. This list is handed over to the display module, which retrieves the data for display and creates a report.

The separation of the query and display modules has several benefits:

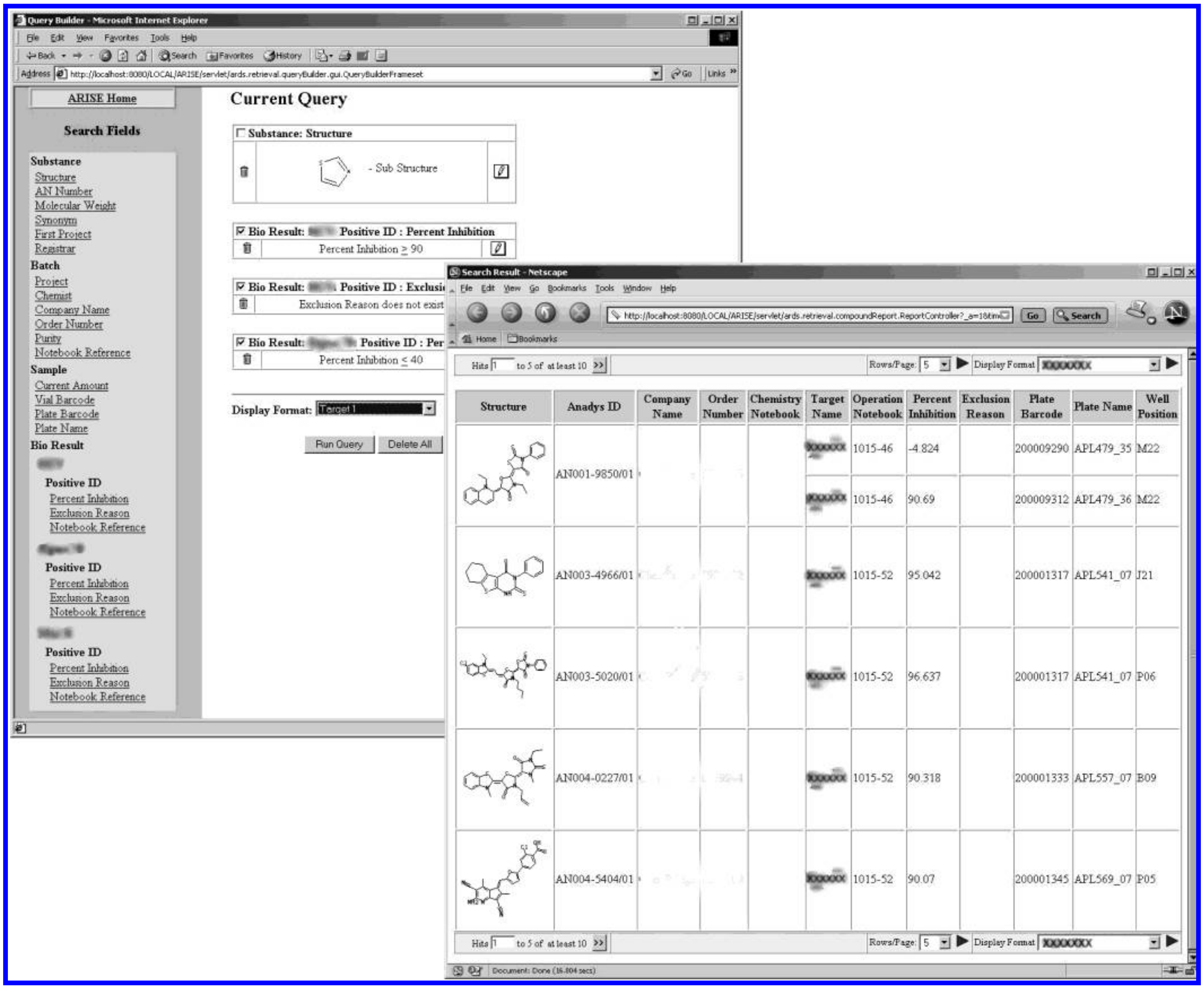
(1) Performance. The SQL query that retrieves the hit list can be optimized to complete very quickly, while the retrieval of report data is done on the fly and only for those compounds that are on the current display page.

(2) Flexibility. Hit lists can come from any source and can be displayed in various report formats. Currently, hit lists can come from three sources (QueryBuilder, SimpleSearch, and ListHandling).

(3) Ease of Implementation and Maintenance.

QueryBuilder[47] was released when the infrastructure for acquiring assay data was implemented and the first LTS data were uploaded. The timing was important, because as mentioned before, the variability of LTS data is large and therefore requires an easy-to-use application for users to formulate their questions. The interface is shown in Figure 12. On the left side a tree with the available search fields is displayed from which users can pick and set constrains. The query as whole is displayed on the right of the tree. Before running the query, the user has to choose the search domain and the display format. The search domain can be the entire database or any hit lists that the user may have saved from previous searches.

The current reporting applications interface to the ListHandling module.[47] This means that users can save results of searches and apply list logic, i.e., union, intersection, and



Query Builder - Microsoft Internet Explorer

Address: http://localhost:8080/LOCAL/ARISE/servlet/ards.retrieval.queryBuilder.gui.QueryBuilderFrameset

ARISE Home

Search Fields

Substance
Structure
AN Number
Molecular Weight
Synonym
First Project
Registrar
Batch
Project
Chemist
Company Name
Order Number
Purity
Notebook Reference
Sample
Current Amount
Vial Barcode
Plate Barcode
Plate Name
Bio Result
[redacted]
Positive ID
Percent Inhibition
Exclusion Reason
Notebook Reference
[redacted]
Positive ID
Percent Inhibition
Exclusion Reason
Notebook Reference
[redacted]
Positive ID
Percent Inhibition
Exclusion Reason
Notebook Reference

Current Query

□ Substance: Structure
- Sub Structure

☑ Bio Result: [redacted] Positive ID : Percent Inhibition
Percent Inhibition ≥ 90

☑ Bio Result: [redacted] Positive ID : Exclusi
Exclusion Reason does not exist

☑ Bio Result: [redacted] Positive ID : Per
Percent Inhibition ≤ 40

Display Format: Target 1

Run Query | Delete All

Search Result - Netscape

http://localhost:8080/LOCAL/ARISE/servlet/ards.retrieval.compoundReport.ReportController?_a=1&tim

Hits 1 to 5 of at least 10 >> Rows/Page: 5 Display Format XXXXXXX

Structure	Anadys ID	Company Name	Order Number	Chemistry Notebook	Target Name	Operation Notebook	Percent Inhibition	Exclusion Reason	Plate Barcode	Plate Name	Well Position
	AN001-9850/01	[redacted]	[redacted]		[redacted]	1015-46	-4.824		200009290	APL479_35	M22
					[redacted]	1015-46	90.69		200009312	APL479_36	M22
	AN003-4966/01	[redacted]	[redacted]		[redacted]	1015-52	95.042		200001317	APL541_07	J21
	AN003-5020/01	[redacted]	[redacted]		[redacted]	1015-52	96.637		200001317	APL541_07	P06
	AN004-0227/01	[redacted]	[redacted]		[redacted]	1015-52	90.318		200001333	APL557_07	B09
	AN004-5404/01	[redacted]	[redacted]		[redacted]	1015-52	90.07		200001345	APL569_07	P05

Hits 1 to 5 of at least 10 >> Rows/Page: 5 Display Format XXXXXXX

Document: Done (16.804 secs)

Figure 12. QueryBuilder and an example report view. Some fields have been blacked out.

subtraction. This functionality enables users to easily formulate queries involving Boolean logic including "OR" and "NOT". Like the query and display modules, the List-Handling module is encapsulated and interacts with the other modules through a well-defined interface, i.e., a list of compounds.

5. CONCLUSION

ARISE has proven itself to be a solid foundation as the data management system for Anadys. It is used by scientists from all groups on a daily basis. Starting from scratch it was possible to implement the required functionality in a reasonable time. The current database and data model is stable, flexible, and well able to cope with the increasing data volume. It now contains multiple hundreds of thousands of substances, many million samples, and HTS data points as well as multiple tens of thousands of LTS results.

Resources and Timetable. Undertaking a project like ARISE might seem like a big risk given the number of required modules. However as shown in the Introduction the various open source projects provided a solid foundation for much of the ARISE platform. This made it possible to implement ARISE in a reasonable amount of time and within budget constraints. It is important however to set priorities and to identify workflow requirements. Also, releasing simple implementations first and improving on those later helped to accommodate the evolving needs of the research community at Anadys. As mentioned previously, the reporting capabilities are what provide value to the users. Because of this the two major milestones to date in the ARISE project were as follows: (1) the release of a simple reporting application (SimpleSearch) together with the compound registration module and (2) the release of the QueryBuilder together with the registration module for biological data. The release of the other modules was mostly dictated by the needs of specific departments within Anadys.

Table 2 shows a rough estimate of effort and the approximate time of the first release of each module. The numbers, however, do not reflect the total effort required for cheminformatics and data management. Not included is the time spent on tasks not related to software development such as process analysis, user training and support, project related data modeling, doing data uploads, compound selection, and general maintenance. A rough estimate would be that software developments represented only about 40% of

Table 2. Rough Estimates of Development Effort

module	FTE weeks	initial release
hardware setup	3	5/2001
chemical search engine	5	5/2001
database design	4	6/2001
diversity selection	3	6/2001
compound registration	6	9/2001
reagent tracking	2	9/2001
SimpleSearch	4	9/2001
compound library management modules	6	11/2001
NMR spectra integration	6	4/2002
HTS registration and analysis[44]	35	10/2002
compound enumeration	7	10/2002
QueryBuilder[47]	20	12/2002
LTS registration	6	1/2003
ListHandling[47]	5	5/2003
sum	112	

the effort involved in implementing ARISE. It is important to realize that data management in a biotech company is a complex process that requires a significant amount of planning and maintenance. To speed up the development of critical modules, we have found it to be particularly efficient to increase the workforce with proven consulting partners on large applications to shorten the implementation time.

Usage Monitoring. Monitoring module usage is very important for a project like this. We capture the number of logins and which kinds of tasks users are doing. These numbers help to evaluate over a longer period if things are going in the right direction. It also helps to identify power users who can give valuable input on problems and suggest improvements. After major releases, we consistently observed an increase in usage.

6. OUTLOOK

The current state of ARISE fulfills most informatics needs of the scientists at Anadys. However, a project like this will never be complete in a research driven and dynamic company. New requirements arise from new technologies or ideas developed or used throughout the company. The approach in developing ARISE was to quickly create simple-to-use modules that support the most needed functionalities. This means, however, that some modules will have to undergo multiple cycles until the optimal functionality is implemented. This has proven to be a very useful approach since the first implementation usually animates users to provide feedback and improvement ideas.

Currently, improvements to Reporting are under development. These will allow users to design reports, selecting which data to include and how to display it.

Another unmet need is additional integration of instruments with the database. This is especially important for instruments used in high-throughput operations. A streamlined data flow will allow faster access by scientists and ensure data consistency. A detailed analysis of the work and data flow for this project has begun. Additional future efforts will focus on data analysis and mining as well as on improving workflow support.

ACKNOWLEDGMENT

We would like to thank Dr. James Appleman for reviewing this paper and many helpful suggestions.

REFERENCES AND NOTES

(1) Ahlberg, C. Visual exploration of HTS databases: bridging the gap between chemistry and biology. *Drug Discov. Today* **1999**, *4*, 370−376.

(2) Fay, N.; Ullmann, D. Leveraging process integration in early drug discovery. *Drug Discov. Today (information biotechnology suppl.)* **2002**, *7*, S181−S186.

(3) Claus, B. L.; Underwood, D. J. Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* **2002**, *7*, 957−966.

(4) Peakman, T.; Franks, S.; White, C.; Beggs, M. Delivering the power of discovery in large pharmaceutical organizations. *Drug Discov. Today* **2003**, *8*, 203−211.

(5) Trepalin, S. V.; Yarkov, A. V. CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100−107.

(6) Ihlenfeldt, W.-D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. Enhanced CACTVS Browser of the Open NCI Database. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 46−57.

(7) Adams, N.; Schubert, U. S. From Data to Knowledge: Chemical Data Management, Data Mining, and Modeling in Polymer Science. *J. Comb. Chem.* **2004**, *6*, 12−23.

(8) Vendors of "off-the-shelf" chemical, analytical or assay data management system or components include: Advanced Chemistry Development Inc, Toronto, Ontario, Canada; CambridgeSoft, Cambridge, MA, U.S.A.; ID Business Solution Guildford, Surrey, UK.; MDL Information System Inc. San Leandro, CA, U.S.A.; NuGenesis Technologies Corporation, Westborough, MA.

(9) Examples of proprietary developments: Celltech's Registration System, http://www.daylight.com/products/casestudies/f_casestudy.html; Senomyx' HT Discovery Platform, http://www.daylight.com/meetings/mug03/Cohen/index.html; Actelion's OSIRIS System http://www.actelion.com/Apps/WebObjects/Actelion.woa/wa/dp?name=research_informatics.

(10) Miled, Z. B.; Liu, Y.; Powers D.; Bukhres O.; Bem M.; Jones R.; Oppelt R. J. An Efficient Implementation of a Drug Candidate Database. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 25−35.

(11) The Apache Jakarta Project, Apache Tomcat, http://jakarta.apache.org/tomcat/index.html.

(12) The Apache Jakarta Project, http://jakarta.apache.org.

(13) The Apache Software License, http://www.apache.org/licenses/LICENSE.

(14) The Apache Jakarta Project, Mailing Lists, http://jakarta.apache.org/site/mail.html.

(15) java.sun.com, Java Servlet Technology, http://java.sun.com/products/servlets.

(16) java.sun.com, JavaServer Pages, http://java.sun.com/products/jsp.

(17) Heterogeneous Modeling and Design UC Berkley, EECS, Ptolemy Project, Ptplot http://ptolemy.eecs.berkeley.edu/java/ptplot/.

(18) The Apache XML Project, Batik SVG Toolkit, http://xml.apache.org/batik/.

(19) The FreeHep Java Library, http://java.freehep.org/index.html.

(20) The Apache XML Project, http://xml.apache.org/.

(21) The Apache XML Project, Xalan-Java, http://xml.apache.org/xalan-j/index.html.

(22) JDOM, http://www.jdom.org/.

(23) The Apache Jakarta Project, Struts, http://jakarta.apache.org/struts/.

(24) Rohde B. private communication.

(25) Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

(26) James, C. A.; Weininger, D. Daylight Software Manual Version 4.42. Daylight Chemical Information Systems Inc., Irvine 1996. http://www.daylight.com.

(27) ChemDraw is a product of CambridgeSoft Corporation Cambridge, MA, http://www.camsoft.com/.

(28) Rohde B., private communication.

(29) NUTS NMR Data Procession Software is a product of AcornNMR, Inc. Livemore, CA, http://www.acornnmr.com/.

(30) PostgreSQL, http://www.postgresql.org/.

(31) MySQL is developed, supported and distributed by MySQL AB, Uppsala, Sweeden, http://www.mysql.com/.

(32) Oracle Corporation, Redwood Shores, CA, http://www.oracle.com/.

(33) Raphaely, D.; Rawles, J.; Murray, C.; Agarwal, N.; Al-Shaikh, R.; Banerjee, S.; Das, D.; Murthy, R.; Trezza-Miller, C. Oracle8i: Data Cartridge Developer's Guide, Oracle Corporation: Redwood Shores 1999.

(34) Gobbi, A.; Rohde B. unpublished work.

(35) Groff, J. R.; Weinberg, P. N. *SQL: The Complete Reference*, 2nd ed.; McGraw-Hill: U.S.A., 2002.

(36) Amanda, University of Maryland at College Park, 1997, http://www.amanda.org/. Further information may be found in the following: Preston, W. C. Unix Backup & Recovery; O'Reilly 1999.

(37) ACD Finder is a product from MDL Information System Inc. San Leandro, CA 94577, U.S.A., http://www.mdli.com/.

(38) SMIRKS is a language defined for generic reactions developed by Daylight Inc. (cf. ref 26), http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html.

(39) The IRORI synthesizer is a product of Discovery Partners International, San Diego, CA, http://www.discoverypartners.com/.

(40) The largest number of compounds registered with the batch registration program to date is around 41 500 compounds on plates. This took about 2 h.

(41) Gobbi, A.; Lee, M. DISE: Directed Sphere Exclusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 317−323.

(42) SMS500 Autosampler is available from Varian, Inc., Palo Alto, CA 94304, U.S.A., http://www.varianinc.com/.

(43) Chempack is an add-on to the Varian VNMR software.

(44) ATHENA was implemented together with of Managed Ventures LLC, Irvine, USA. OpenHTS a follow-up implementation of a very similar system is available from Managed Ventures.

(45) Zhang, J. H.; Chung, T. D.; Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screen.* **1999**, *4*, 67−73

(46) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipies in C*; Cambridge University Press: UK, 1992.

(47) QueryBuilder and the ListHandling module were designed and implemented together with Hewitt Consulting, San Diego, U.S.A.

CI034269O