The role of the informatics framework in early lead discovery

Nicolas Fay

Evotec AG, Schnackenburgallee 114, D-22525 Hamburg, Germany

Recent developments in screening technologies and data analysis have been driven by promises that the numbers of new lead compounds will increase. Although many of these promises have become reality, the success of this strategy also depends on the information framework that ties the individual components together. In particular, high-content technologies represent a new force in challenging established informatics frameworks; largely because of their data volume, variety of assay parameters and increased scientific complexity. A successful informatics framework design can be regarded as crucial for new technologies, both in terms of scientific content and information, and process integration across large corporate networks.

Looking at the evolution from simply collecting and querying local data to sophisticated approaches of data integration and algorithm implementation on an enterprise scale, it is safe to say that the benefit of high-throughput technologies in many areas – from genomics to HTS – often relies solely upon availability of the most recent computational infrastructure designs and technologies. Questioning the return from investment in these new technologies, the answer lies beyond a compendium of individual tools and algorithms. It involves their integration as a crucial step towards a more efficient process structure, affecting large-scale very-high-throughput or small-scale focused screening environments.

But integration not only means streamlining or pipelining data flow (i.e. automation on a purely technical level, although this is a reoccurring challenge in many companies with legacy hard- and soft-ware) it also particularly refers to the discovery process, with emphasis on process. The network of activities such as HTS, specificity testing or hit-compound series analysis with clustering methods (referred to as a process, each part individually contributing to discovery projects) has become increasingly complex. The 1D time-bar (i.e. the sequential succession of process steps within a project, not using information derived from other projects or projects run in parallel) of drug discovery going from target identification to clinical trials is being replaced by a network of processes. If the processes 'speak the same language', information- and knowledge-sharing in the network is facilitated, the flow of data and information improved and, thus, it is possible to extract and provide relevant information for different parts of the network quickly. This is the role of informatics frameworks – to provide an interconnection between individual processes and establish workflows and tools for collaborating researchers.

In this review, I discuss the reasons for the paradigm shift in life-science informatics from three different perspectives: the first key driver of the parallel and mutually fruitful development of information technologies and assay technologies is the portfolio of new signal-detection methods, high-content screening (HCS) and high-content analysis (HCA) technologies – how we deal with their output and how we generate additional value from it; second, is the role of statistics, which ensures quality and comparability of results across various stages of discovery projects; and, finally, we investigate how these and other drivers impact underlying informatics infrastructure designs and implementations, what the issues and core requirements are (both in terms of technology and strategies) and how they are going to be consolidated.

The impact of high-content technologies

A key driver of recent advances in assay and detection technology for early drug discovery is the promise to obtain better insight into biology with reduced investments in instruments and reagents on shorter time scales. Although traditional technologies such as homogeneous assays based upon absorbance, radioactivity and

Corresponding author: Fay, N. (nicolas.fay@evotec.com)

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.10.009

fluorescence light emission [1] are still in place, recent developments in the area of microscopy at subdiffraction resolution [2], two-photon excitation [3], cross correlation [4] and fluorescence lifetime [5] promise to deliver considerable value for screening in the near future. The fluorescence-based methods in particular combine flexibility, robustness and ease-of-use, lacking severe problems accompanied by the presence and disposal of radioactive material in the laboratory. The wide spectrum of methods applicable in screening comprises well-established technologies (e.g. fluorescence polarization) that are commercially available – usually in large-scale setups. Additionally, new readout technologies such as FCS+plus [6] are able to resolve assay species on a molecular level – rather than simply detecting the bulk fluorescence from all of the fluorescent particles. Thus, these new technologies deliver additional information to enhance the interpretation of test results. An example application is the removal of fluorescence artifacts that probably occur with all fluorescence-based methods; compounds contaminate the assay signal either through autofluorescence or by interfering with the molecular environment of the ligand. Traditional technologies are rather defenseless here, but high-content technologies have already proven successful in reducing the impact of such contamination on the assay results [7,8].

Although HCA can be used in a broader sense, its origins are actually in the area of a new generation of imaging readers (e.g. ArrayScan® by Cellomics; http://www.cellomics.com, IN Cell Analyzer by GE Healthcare; http://www.amershambiosciences.com, OPERA™ by Evotec Technologies; http://www.evotec-technologies.com, Pathway™ Bioimager by Becton Dickinson; http://www.bd.com, and ImageXpress® Ultra by Molecular Devices; http://www.moleculardevices.com. A selection of suppliers for HCA software can be found in Table 1). These new readers can resolve processes inside cells (i.e. spatially and timely) with precision and traceable fluctuations of several fluorescence markers.

HCS has already proven successful as a method to deliver more relevant information simultaneously in one experiment, rather than delivering a single readout in a series of sequential experiments [9–12]. A prototype scenario might be the series of simultaneously available readouts obtained from a cellular assay. One parameter identifies cells (i.e. membrane dye at first wavelength), another determines the stage of mitotic change (e.g. fragmented and condensed nuclei at a second wavelength) and a third parameter classifies the apoptotic stage using morphological criteria at a third wavelength. Certainly, these analyses can already be performed almost autonomously with very high throughput. But an appropriate software environment and a unifying informatics platform is required to take full advantage of this plethora of parameters obtained from HCS approaches; this prototype scenario is an example for this situation because of the abundance of parameters that have to be selected as relevant, combined and interpreted in a biological, chemical or mathematical context. This comparative and selective process is challenging enough based upon the high-content instrument and the assays themselves. If the data from the ongoing experiment are insufficient for interpretation of the experiment and additional data are required – maybe data captured in earlier experiments with different equipment and even under different conditions – the situation quickly demands new integrative framework technologies to ensure safe and fast access to the right scientific and technical information. After all, the sheer volume of raw data also demands new solutions (e.g. a fast integrated network for image transfer, storage and analysis).

All these new requirements drive the development of individual data management and analysis applications towards being mapped into workflows and decision processes. But, before I discuss how informatics frameworks deliver their value, successfully exploiting the new technologies depends on how reliable and comparable the primary data obtained from using these technologies are. This is the enduring role of quality assurance (QA).

The importance of QA for data normalization

In the past, placing bets on state-of-the-art technologies often paid off, the benefit delivered by the new approaches rectified previous sometime-risky investments; however, the resulting heterogeneity of methods and data sources and their wealth of useful information evoke consequences at many places. How can information systems leverage the value of the new technologies, which new parameters are available and how can they be exploited? To understand why these factors are important, we have to bear in mind that the core challenge for QA is not only to tame the statistical accuracy of the data; it is to ensure that data from different sources, at different points in time or under potentially varying environmental conditions, are still comparable later in another project – a process known as data normalization.

Again, QA is still an essential component of data processing, but this time enriched by potential that is buried in recent technology developments: detailed hardware monitoring (e.g. quantitative flow control in liquid-handling systems, as well as advanced detection technologies, as discussed earlier, provide plenty of parameters for online QA). Errors are handled either autonomously on a hardware level or they are passed immediately to the operator. Solved and unsolved problems or errors are flagged and passed into a second stage of QA, which is applicable to (and actually performed on) aggregated results, as soon as the problems occur.

Here, on-demand and interactive visualizations are an important element of a wide range of QA tools and strategies that are in place. The trellis view, a method to display multivariable data in an efficient way [13], which is suitable to show plenty of color-coded parameters obtained from the assay and hardware, is a well-known and useful example of a QA tool.

However, even fully transparent and automated procedures have been established as part of informatics frameworks for data analysis, data management and workflow control. Key to understanding the impact of these procedures on the design of such frameworks is the fact that each algorithm often requires a different type of implementation and always challenges compatibility with the framework design. Standard applications are plate-trend analyses and automated correction of plate-uniformity problems, which inevitably lead to decreased throughput and congesting follow-up work with junk. To cope with problems such as plate uniformity distortions, a variety of efficient algorithms have been developed that monitor the screen and detect and correct for quality issues inherent to the statistical nature of HTS [14–21]. The algorithms that utilize robust estimators of the model parameters have proven particularly successful – in this context robust means insensitive against outliers. A simple but efficient example

TABLE 1

Selection of software vendors and products relevant for data analysis and data management[a]

Product	Application	Comments
Visualization and visual data mining		
Spotfire DecisionSite® http://www.spotfire.com	Visual data mining	Visual and explorative data mining of large datasets. Very fast and intuitive user interface. Has connectivity to ISIS host and provides various structure-related analyses. Guided analytics (guides predefined analysis workflows). Integrates computational services for R-project and S-PLUS® and connects SAS files.
Batelle OmniViz® http://www.omniviz.com	Visual data mining	Mining in multidimensional space with very large sets of numerical, categorical, chemical and textual data. Plug-in interface for user scripts and tools. Feature extraction (relativity tool), dimensionality reduction for visualization, intuitive user interface. Cross-platform compatibility (runs under Solaris, Windows and Linux). Scriptable and plug-in interface for user scripts and tools. OmniViz® offers customization services.
Statistics		
R-project http://www.r-project.org	Statistical computing and graphics	Based upon S-language (Bell Laboratories). Used in statistical method-profiling and applications for generalized linear modeling, (nonparametric tests, nonlinear regression, classification, clustering, etc.) Open Source and arbitrarily customizable. Native interfaces to user-C, C++ and Fortran code and R-packages. Graphing capabilities.
Insightful® S-PLUS® http://www.insightful.com	Statistical computing and graphics	Value-added version of S-language (see R-project). Most features of the R-package. Extended features for robust and nonparametric regression, multivariate analyses, graphics, etc. Handles very large datasets. S-PLUS® provides modules and packages for specific applications (clinical trials, wavelets, optimization, etc.).
Umetrics Enterprise platform http://www.umetrics.com	Statistical computing and design of experiments	SIMCA (soft independent modeling of class analogies) products focus on multivariate analyses (21 CFR Part 11-compliant). Interactive and versions for batch modeling and analysis. MODDE variant used particularly in design of experiments (DoE). SMILES (simplified molecular input line entry specification) aware. Enterprise Platform product integrates SIMCA family.
General data mining		
MathWorks Matlab® and Simulink® http://www.mathworks.com	Numerical computation and visualization.	(MATrix LABORATORYoratory). High-level language and environment for algorithm development, visualization, analysis and numeric computation. Add-on toolboxes (collections of MATLAB® functions) provide specific services from signal and image processing to computational biology. Fully integrates with user C++, Fortran, Java, COM (SDK). Simulink® is a simulation platform with block editor that fully integrates with MATLAB®. Runs on various platforms.
Partek® Discovery Suite, Screener's Solution and QSAR Solution http://www.partek.com	Data mining, artificial intelligence	Integrated environment for visual and quantitative data analysis of very large datasets. Matrix-like data handling. Various distance and similarity metrices, PCA (principle component analysis), clustering, NLM (non Pattern recognition, classification and prediction capabilities with artificial neural networks and genetic algorithms (multi
High-content data analysis and management		
Genedata Screener® and Phylosopher® http://www.genedata.com	Screening data analysis, information management	Screener application supports quality control and analysis of interactively managed early-stage and large volume screening datasets. Provides exhaustive interactive visualizations based upon a broad range of statistical analyses to help prioritize compound sets for follow-up work. Phylosopher® integrates metadata from drug discovery projects ranging from genomics to pathway data and mode of action (MOA) studies.

TABLE 1 (*Continued*)

Product	Application	Comments
Definiens Cellenger http://www.definiens.com	Image analysis for high-content screening (HCS) and biomedical applications	Cellenger Developer Studio and Enterprise for automated (pre-defined work flows using Cellenger Server) object-oriented image analysis, uses structural and relational information in images (morphometric quantization) and realizes an image object hierarchy. Based upon 'Cognition Network Technology' aiming to mimick human perception of objects.
Evotec Acapella™ http://www.evotec-technologies.com	High-content data analysis	Interactive, fully scriptable and compatible with 3rd-party platform environments. Upgradeable with user libraries. Provides high-level language to reduce coding overhead for main applications in image analysis (HCS): object recognition, grouping and segmentation, morphologic filtering, image arithmetic. Libraries available also for Photon Statistics or specific applications like FLIPR kinetics analysis. SDK available.
Cellomics™ HCi™ http://www.cellomics.com	HCS – image management and analysis	Multi-tier integrated environment for large volumes of HCS data. Middle-layer manages data level (image store) and presentation (user) level with plug-in interfaces for additional functionalities like user data I/O, visualizations, workflow management and QA (vHCS Discovery Toolbox).
Molecular Devices Metamorph® and AcuityXpress http://www.moleculardevices.com	HCS – image management and analysis	Integrates with Molecular Devices HCS readers and MetaXpress. Image storage, analysis and mining software suite for cellular images with open image database API. Includes management tools for multi-user environments.
BioImagene (SciMagix) CellMine™ and SIMS™ http://www.bioimagene.com	HCS – image management	Main application is HCS. Multi-tier architecture for fast image-I/O of large volume HCS data from various instrument sources. Supports workflows for reorganization, aggregation and visualization of image and metadata for further analysis.
IDBS ActivityBase™ http://www.idbs.com	HTS data management and analysis	Biological assay data- and experiment-management platform. All data processing via a central relational database as the store and Microsoft Excel for data analysis (analysis workflows defined via Excel templates). Has chemistry cartridge and deals with drug metabolism and pharmacokinetics (DMPK) data specifics.
Evotec A+ http://www.evotec-technologies.com	HTS data management and data analysis	Scalable uHTS data management platform with a component architecture. A generic programmable data pre-processor (batch- and grid-enabled) captures and pre-analyzes instrument raw data in real-time. Automatically joins relevant data from external databases like compound management. Transparent data management from capturing to dose–response fitting and reporting. Access to result data through a web front-end and Spotfire® DecisionSite™, R-project and S-PLUS®, respectively, for high-level statistics and data post-processing and reporting.
SciTegic™ (now Accelrys) Pipeline Pilot™ http://www.scitegic.com	Data analysis and mining	Data analysis and workflow management based upon graphical programming (visual scripting): components are visually arranged to protocols. Publication of protocols for remote execution. Configurable components for chemistry, statistics, sequencing, text mining as well as integration of 3rd party applications.
NIH grant project Genepattern (GP) http://www.broad.mit.edu/genepattern/	Data analysis and mining	Workflow management system for data analysis and visualization. Provides graphical IDE and object browser. GP comes along with plenty of modules for statistics, visualization, machine learning, etc. to be arranged as a sequential or parallel pipelined workflow. GP modules are also accessible from within R-project, Java and MATLAB®.

TABLE 1 (*Continued*)

Product	Application	Comments
Agilent Synapsia Informatics Workbench http://www.chem.agilent.com	Knowledge management	Concurrent Synapsia provides the Discovery Manager desktop user interface: object hierarchies are mapped to a file- and directory-like structure whereby content and relationships can be displayed (e.g. with Spotfire®, as a SAR table or a phylogenetic tree). The open architecture and documented APIs enable integration of external tools for (e.g. BLAST searches). Together with Information Manager it represents a collaboration framework for cross-discipline R&D projects.
Data warehousing and document and content management		
SAS Institute http://www.sas.com	Enterprise data warehousing and mining	Statistics particularly in clinical trials. Integrates diverse capabilities in the areas of data warehousing and data I/O, analytics (statistics, mining, forecasting, QA monitoring, scheduling and simulation research), business intelligence and many others into a broad warehouse infrastructure including SDKs.
EMC^2 Documentum http://www.documentum.com	Content management	Managed collection of software tools to organize unstructured information originating from sources like documents, spreadsheets, web pages or e-mail databases according to defined business rules. Creates relationships, organizes metadata and provides tools for search, retrieval and presentation.
Dotmatics Suite http://www.dotmatics.com	Knowledge management	Knowledge management platform with integrated components for data and text mining, visualization and reporting. Gateway product as intuitive and fast visual gateway to projects with associated or related data and information (drill-down capabilities). Accompanying applications provide advanced querying features to external information sources (indexed), collaborative tools for cross-discipline work, data mining and decision management.
3rd Millennium® http://www.3rdmill.com	Knowledge management	Released as open source in 2003. Regarded as 'foundation system', open to be customized and extended using its APIs. Architecture is client-server with underlying database. Focus on sequence analysis.
Waters NuGenesis SDMS http://www.waters.com	Content management	Data and document management systems with interfaces to laboratory information management systems (LIMS) and analytics systems and applications.
LabVantage® Sapphire™ http://www.labvantage.com	LIMS	LIMS with an open architecture enabling free definition of workflows. Integrates external compound repository databases.
Oracle® Collaboration Suite http://www.oracle.com	Content management, groupware	Based upon Oracle® 10g Application Server – the collaboration suite provides a framework for management of unstructured disparate and heterogeneous information sources, real-time collaboration, unified messaging or project workspaces.
IBM® WebSphere® http://www.ibm.com	Information and process integration	Large software and consultancy portfolio aiming for the integration and application infrastructure.
Molecular modeling and cheminformatics		
Elsevier MDL® http://www.mdl.com	Cheminformatics, knowledge management	Product suites Discovery Framework (ISIS and Isentris plus supporting applications), Discovery Experiment Management (Assay Explorer, ACD-SC and more), Discovery Knowledge (reference work and literature links: Beilstein, Discovery Gate, Drug Data Report and many more) and Discovery Predictive Science (tools to predict compound properties featuring also 3rd party components). MDL® also offers consulting services.
Accelrys® Suites http://www.accelrys.com	Cheminformatics, computational chemistry	Broad offering of suites: Accord Enterprise Informatics for chemical information management, Discovery Studio built upon SciTegic™ Enterprise Server as an integrated environment for protein modeling, computational chemistry and crystallography. Also provides services for custom solutions.

TABLE 1 (*Continued*)

Product	Application	Comments
Tripos^{TM} Suite	Cheminformatics, computational chemistry	Application server integrates tools and provides access to Tripos^{TM} and third-party discovery informatics software.
Foundation Server http://www.tripos.com		SYBYL as an optional environment provides tools for molecular modeling and cheminformatics.

[a] Discussed here in the context of informatics frameworks, roughly grouped by the main area of application, usually products cover several areas of applications at the same time.

is 'median polish' [22], which is used to quantify the spatial-trend structure of an assay plate, and enables predicting systematic deviations from the expected spatial or timely behavior of the experimental parameters (see Box 1).

Approaches that are more advanced than the simple median polish approach, which are shared by many software packages, have proven successful: methods such as global parametric models that model the experimental data with the assumption that one

BOX 1

Data normalization as a prerequisite for successful use in a broader project-spanning context: an example

Interpretation of experimental data is often improved when it can be compared with results from earlier experiments. A crucial prerequisite for this comparability is data normalization, which ensures that data can be compared 'out of the box', and details of the experiments are known so that they can be considered during a comparison. An element of this normalization process is shown in Figure Ia and Figure Ib: a common source of false-positives or false-negatives are plate patterns (i.e. systematic errors that shift the assay signal depending upon the position of the sample on the plate).

Figure Ia is a view along the columns of an original, uncorrected 384-well plate. Using a hit-selection threshold derived from the negative controls (red and light blue) false-positives are generated in the center of the plate (most of the dark-green dots in the center of the plate, which belong to the population of the light-green compound samples). 'Median polish' is a method that corrects for such plate patterns and can be performed 'on the fly' during raw-data import. The result is shown in Figure Ib: there are few 'real' hit compounds in the center of the plate (dark green) – most of the false-positives from the original dataset in Figure Ia are now correctly classified as inactive. The downside of the algorithm is that areas of the plate that are not affected by a systematic error can experience a slight artificial increase of variance. In Figure Ib, a slight over-correction is seen, which becomes manifest in an artificially increased noise level at the rim of the plate (left and right sides), whereas the center of the plate is not affected by the drift problem, although this is not relevant in terms of the overall advantage of median polish. Other correction approaches based upon different algorithms work in a similar way and each method has its own specific benefits and problems. Because of this, there is no general recommendation for one method only, and available methods need to be carefully assessed in their specific area of application.

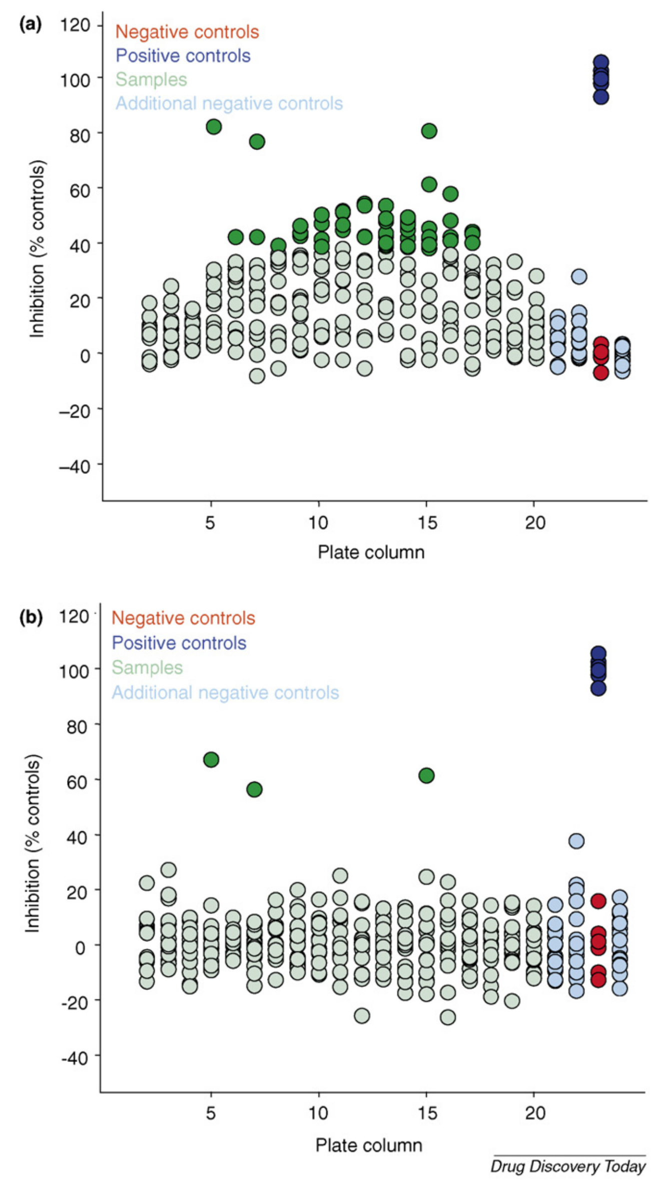

FIGURE I

(a) View of assay plate with uncorrected signal and (b) the same plate after 'median polish' correction.

model, one global analytic functional expression, is applicable and valid for the complete data set. These methods are well-suited for characterizing the general shape of signal drifts. However, for the purpose of correction and a more reliable prediction of compound activity, they are not as powerful as the local regression methods that require a local functional expression. A well-known and powerful software library for this purpose is LOCFIT (http://cm.bell-laboratories.com/stat/project/locfit), which is available from the R-project for statistical computing. Multistage strategies using several methods have also been applied [23].

Another category of methods that predict systematic errors is based upon machine learning (e.g. a group of pattern-recognition algorithms that are applicable in medical diagnosis or DNA sequencing). Many of these algorithms are also offered as part of various software solutions and software development kits (SDK) such as PRTools (http://www.prtools.org) for Matlab® or MLC++ (http://www.sgi.com/tech/mlc), and can be applied for recognition of QA-related patterns in spatial or timely signal behavior. Although some of these methods bear the risk of eroding clusters of true hits that might gather on certain plate areas, their benefits can be exploited safely if remote or metadata from, for example, a chemical structure database are included in the classification of true actives versus artifacts.

Multiparametric readouts delivered by modern screening technologies (along with mathematics and software engineering) play a crucial role in establishing generic countermeasures against reliability and comparability issues [24,25]. However, crucial for the success and efficiency of these countermeasures is the way in which they are finally used within the application, namely as part of an informatics framework. This emphasizes the particular demand for an integrative informatics environment – technically and logically.

The informatics framework

In parallel to all the technical developments discussed and their implication for informatics strategies for early drug discovery, the way projects are run or structured has also changed in a way that reflects the benefit that is delivered by the new technological opportunities. Hence, it is obviously expected that the underlying framework supports the seamless integration of technical components as well as logical components such as work- and information-flows.

Consider this basic, but unfortunately very common, situation: results from HTS need to be cleansed from frequent hitters, non-selective or nonspecific compounds. Although this sounds like a quick and simple 'query and put into a table for comparison' method, it can be tedious to struggle with technicalities that query remote and foreign databases from the desktop. Informatics frameworks aim to support this and other challenges such as the new screening strategy of iteratively testing smaller, focused sets of compounds, which represents a more rational approach than the simple random approach of mass-testing large compound collections. The success of this strategy in yielding more novel hit structures (together with considerable time and cost savings) has been shown [26–28]. Again, to what extent does this scenario affect informatics framework designs such as those depicted in Figure 1? Decisions in early drug discovery are always stimulated by the available pre-knowledge regarding the target and known binders because this directly affects the initial selection of compounds for the first round of testing [29]. In pursuing this strategy (i.e. of an initial selection of compounds, sometimes termed 'focused screening'), diverse information from logically and geographically disparate sources is required before a decision can be made as to what the next step in the project will be.

To gather this information, we are facing several hurdles such as actually performing the queries on the technical level. In addition, involving non-experts in this task bears the risk that retrieved data is taken to be valid and used to interpret experiments, although these data are actually incomparable, incompatible or cannot be combined for various reasons. This might be the case when accompanying metadata cannot be included in the query, is not considered at all or is simply unavailable – a serious problem if required metadata qualify and explain conditions where the requested data can be used within a given context. Metadata from sources with unstructured data such as patent databases or PubMed (http://www.pubmed.gov) also represent a specific semantic challenge, simply because of an incomplete data structure that usually ensures a defined and common meaning for the requested content [30] (for a semantic web prototype of a drug development dashboard see http://www.w3.org/2005/04/swls/BioDash/Demo). Instead, streamlined and certified workflows must provide more efficiency and reliability, both in terms of semantics and the technologies involved.

At this point, the informatics framework steps into the breach: it delivers a series of benefits such as guiding through predefined and certified workflows – each one providing potentially complex applications and algorithms – as well as a transparent and safe way to ensure standards of characteristic measures of the data that are being retrieved. Hence, requiring dedicated experts for standalone analyses or operations on remote, vast and disparate data sources is no longer required.

Although solutions for individual tasks in early drug discovery are available and efforts in making hardware more reliable succeed, the challenge is in tying together all this to meet the more rational screening paradigm, and in reducing friction between the various physical and logical components of the decision process, which ultimately lead to the decisions taken in hit-to-lead programs.

We can pinpoint the core requirements for such a framework. They are: proper quality control criteria and procedures to ensure reliability; consistency and comparability; powerful ontologies for covering unstructured data; fast availability of recent and properly validated experimental data for analysis; compatibility of data formats and interlinks – interoperability within the framework as well as ergonomics with regard to user interfaces; and, finally, performance and maintenance efforts. However, from the various design requirements for such a framework, process safety (i.e. result reliability) often receives top priority. It has been clearly realized that quality is more relevant than the proverbial 'new world record' in throughput [31]. Thus, reliability of results is considered a primary goal, which is ensured by validated algorithms and certified workflows on the basis of data management solutions that guarantee data integrity. Further benefits are: low maintenance and administration; flexibility to cope quickly with individual demands; and being prepared for future developments of algorithms and applications.

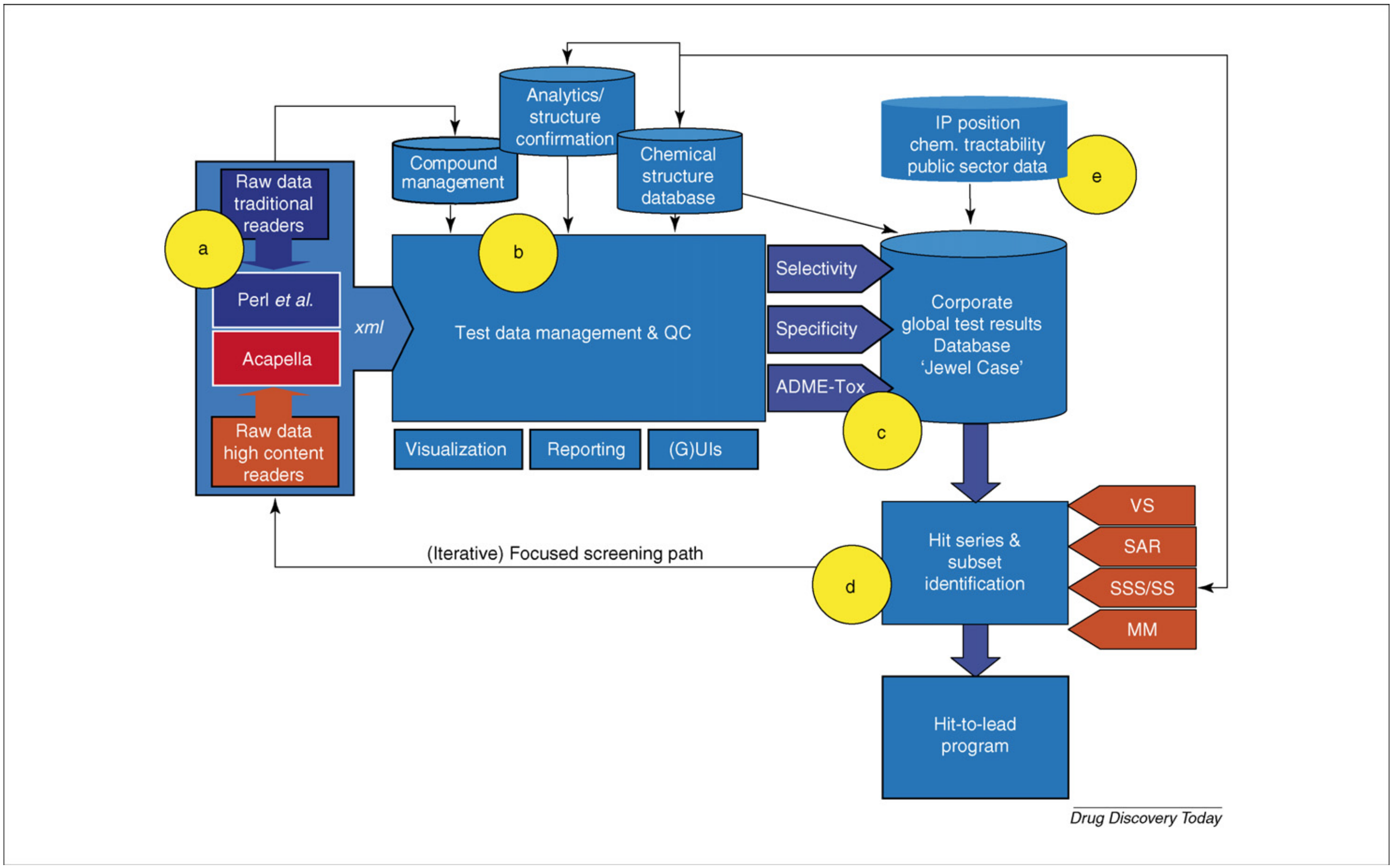

FIGURE 1

A model of an informatics framework reflecting the network of data and information flow. The network of drug discovery processes comprises various components of data management, data analysis, visualization, joining disparate and metadata from external sources, data mining, and prioritizing result sets for follow-up processes. Initially, raw data enter the network **(a)**. Two basic types of raw-data handlers cover raw-data capturing and pre-analysis. Passive importers convert different data formats that are generated by different instruments, whereas active importers are capable of complex operations 'on the fly' while importing. For active import almost every kind of preferable script language can be used, indicated on the left side of the Figure, by Perl and Acapella^{TM} (Evotec). In particular, active importers can perform various analyses ranging from identification and correction of errors (e.g. signal drifts before they are passed to the central test-data management system to meet basic data normalization requirements) to sophisticated image analysis operations. Once reliable primary data have been loaded from the instruments, disparate and metadata can be added from external sources **(b)** and data can be visualized or reported as needed. **(c)** Represents parallel results such as running potency tests on the primary target together with selectivity assays. They are then passed to a central repository that contains validated test results, which are available for exhaustive data mining. Again, external information that might be relevant at stage **(e)** is joined and finally passed to higher level analysis to prioritize results for either focused or refined retesting or even hit-to-lead studies **(d)**. Interfaces to virtual screening (VS), SAR, substructure searches or similarity searches (SSS/SS), or molecular modeling (MM) provide an integrated environment for easy access to the necessary tools.

A better framework simply provides a means to reduce the time required to access an even larger volume of information for a better-guided decision process.

Many powerful software suites (i.e. collections of applications marketed as a product) are available that cover individual topics among different areas of the drug discovery process. However, solutions are often difficult to join together into a homogeneous framework; valuable algorithms are accessible only through proprietary, mostly interactive, user interfaces, and data are internally processed using proprietary storage, which is often inaccessible from the environment without explicit human interaction. This type of software is difficult to integrate into a tailored and corporate-specific pipeline of data management and analysis steps. Sometimes, when features need to be re-implemented because they are not accessible off the shelf, it seems like reinventing the wheel. For example, the simple task of data retrieval and conversion from one proprietary data format to another can be a painful experience. Solving this collection of problems of interoperability requires widely expanded workflow support, and perhaps automation beyond high-throughput IC_{50} determination. Solutions must bridge the gap between the statistical process of sequential mass testing and the rational process of hit-series and lead-structure identification.

Many process steps, and the way they are supposed to interact with each other, are subject to corporate standards in numerous organizations. Obviously, even in the case of a binding assay, the binding affinity of a compound to the receptor in a biochemical assay is not the only parameter that renders the compound interesting. Counter-assays, selectivity, specificity and ADME–Tox assays are performed in parallel by HCS [32]. In addition, comparative filters automatically applied to the relevant selectivity assays, or *in silico* filters like Lipinski's Rule of Five and identification of reactive and toxic groups, help to rule out undesired structures quickly [33–35]. There are also other metadata sources that contribute to a decision or compound prioritization; results from virtual screening or docking can be taken into consideration as well as purity confirmation through LCMS, assessment of chemical tractability by a chemist and information from remote databases such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) or ZINC (http://blaster.docking.org/zinc/). Among these components are similarity searches that can analyze the compound library around an interesting hit compound, including structural clustering to help focus on a particular hit series with preferred properties or identification of structural clusters that can develop structure–activity models [36].

Having formulated the 'soft criteria' from the processes perspective on a technical level, the discussion is back where it started – how can all this be tied together? For user interfaces there is a trend towards platform-independent client applications that are based on web services [37] or Java Web Start. Few companies offer generic software platforms that promise smoother data exchange between the many heterogeneous elements of a corporate – sometimes global – software environment. There is still no standard for software applications, a software architecture or the involved protocols for information service infrastructure that would enable straightforward deployment of an interlink software system at all organizations, the larger the environment the less likely this is. It is important that the specification and remaining adaptation effort, and overall development and long-term maintenance and licensing costs for the integrating software alone, are still reasonable (i.e. do they pay off?) compared with tailored in-house solutions. However, although individual implementations of data-management platforms in the industry are not necessarily identical, they share some common requirements that pay-off in the long run if followed during system design. A high degree of modularity of individual applications, and open standards for the interfaces on a technical level and on a semantic level [38], enable easier implementation of the preferred algorithms, easier integration of new components, low redundancy and, thus, high performance in information management. In the end the preferred strategies, techniques and products must ensure faster and more-reliable access to the information buried within the data (i.e. to achieve a sufficient level of knowledge to determine the next step even earlier).

A diverse portfolio of solutions for different problems is available from commercial software vendors, from academia or from dedicated intra-company groups. Table 1 gives a clearly nonexhaustive overview of vendors and products relevant for the topics discussed here, and might serve as a starting point for more-specific investigations for dedicated informatics projects.

Concluding remarks

Lifescience informatics evolved into a close, or even symbiotic, relationship with drug discovery research, development and operations [39]. Traditional problems such as extreme data volumes or the contiguity of technically and geographically disparate data repositories have turned into a challenge, rather than being showstoppers because of a lack of technological ability. New assay and detection technologies are constantly triggering new standards in informatics and, at the same time, are triggered by the latest developments in informatics. This mutual fruitfulness seems to rectify increasing attention on informatics solutions, both in terms of expenditures and timely efforts – additional time spent on data analysis and mining is rarely wasted.

In medium and, in particular, large organizations the various individual contributions to the drug discovery process require a dedicated informatics infrastructure that can be realized by internal efforts or by partnering with external vendors that offer not so much prepacked software but collaboration tools and services that deal better with specific requirements. Deliberate designs of informatics frameworks help reducing costs at multiple frontiers: software and data maintenance; development; training and staffing; and, in particular, the time required and safety provided in gathering complex disparate information 'atoms' to develop the full picture, which actually supports a scientific, business decision.

Although this represents a huge challenge, the role of informatics networks will probably gain even more importance because of evolving systems biology [40,41], which does not simply focus on isolated or even artificial models and targets but assumes that observations can only be predicted and explained when understanding the biological network as a whole.

Although some of the developments mentioned might be a way off, metabolic pathways and phenotypic screening are further-advanced and some are already realities. Therefore, informatics has to provide a crucial component of the solution that exploits these exciting new opportunities.

References

1 Eggeling, C. *et al.* (2003) Highly sensitive fluorescence detection technology currently available for HTS. *Drug Discov. Today* 8, 632–641

2 Hofmann, M. *et al.* (2005) Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using reversibly photoswitchable proteins. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17565–17569

3 Tirri, M.E. *et al.* (2005) Two-photon excitation in fluorescence polarization receptor-ligand binding assay. *J. Biomol. Screen.* 10, 314–319

4 Bacia, K. *et al.* (2006) Fluorescence cross-correlation spectroscopy in living cells. *Nat. Methods* 3, 83–89

5 Hoefelschweiger, B.K. *et al.* (2005) Screening scheme based on measurement of fluorescence lifetime in the nanosecond domain. *J. Biomol. Screen.* 10, 687–694

6 Kask, P. *et al.* (2000) Two-dimensional fluorescence intensity distribution analysis. Theory and applications. *Biophys. J.* 78, 1703–1713

7 Turek-Etienne, T.C. *et al.* (2003) Evaluation of fluorescent compound interference in 4 fluorescence polarization assays: 2 Kinases, 1 Protease and 1 Phosphatase. *J. Biomol. Screen.* 8, 176–184

8 Fay, N. *et al.* (2005) Fluorescence artifacts in HTS. Poster at the 11th Annual Conference and Exhibition of the Scociety for Biomolecular Screening (SBS) September 11–15, Geneva, Switzerland.

9 Johnston, P.A. and Johnston, P.A. (2002) Cellular platforms for HTS: three case studies. *Drug Discov. Today* 7, 353–363

10 Giuliano, K.A. *et al.* (2003) Advances in high content screening for drug discovery. *Assay and Drug Development Technologies* 1, 565–577

11 Taylor, D.L. (2001) Real-time molecular and cellular analysis: the new frontier of drug discovery. *Curr. Opin. Biotechnol.* 12, 75–81

12 Monk, A.J. (2005) Faster, surer prediction. *The Biochemist* October, pp. 25–28

13 William, S. Cleveland (1993) *Visualizing Data*. AT&T Bell Laboratories.

14 Gribbon, P. *et al.* (2005) Evaluating real-life high-throughput screening data. *J. Biomol. Screen.* 10, 99–107

15 Huber, W. *et al.* (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology* 2, article 3

16 Gunter, B. *et al.* (2003) Statistical and graphical methods for quality control determination of high-throughput screening data. *J. Biomol. Screen.* 8, 624–633

17 Zhang, J.H. *et al.* (2005) Probing the primary screening efficiency by multiple replicate testing: a quantitative analysis of hit confirmation and false screening results of a biochemical assay. *J. Biomol. Screen.* 10, 695–704

18 Wu, X. *et al.* (2005) Further comparison of primary hit identification by different assay technologies and effects of assay measurement variability. *J. Biomol. Screen.* 10, 581–589

19 Tye, H. (2004) Application of statistical 'design of experiments' methods in drug discovery. *Drug Discov. Today* 9, 485–491

20 Brideau, C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* 8, 634–647

21 Martin, J.J. (2005) Coping with patterns in screening plates. Presentation at the 8th International Conference and Exhibition on Drug Discovery – MipTec, May 9–12, Basel, Switzerland 2005

22 Mosteller, F. and Parunak, A. (1985) Identifying extreme cells in a sizeable contingency table: probabilistic and exploratory approaches. In *Exploring data tables, trends and shapes* (Hoaglin, D.C. *et al.* eds), pp. 189–224, Wiley, N.Y

23 Kevorkov, D. and Makarenkov, V. (2005) Statistical analysis of systematic errors in high-throughput screening. *J. Biomol. Screen.* 10, 557–567

24 Sun, D. *et al.* (2005) Adopting a practical statistical approach for evaluating assay agreement in drug discovery. *J. Biomol. Screen.* 10, 508–516

25 Malo, N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 24, 167–175

26 Brideau, C. (2005) An efficient method for lead identification using iterative focused screening. Presentation at the 11th Annual Conference and Exhibition of The Society for Biomolecular Screening (SBS) September 11–15, Geneva, Switzerland 2005.

27 Karnachi, P.S. and Brown, F.K. (2004) Practical approaches to efficient screening: information-rich screening protocol. *J. Biomol. Screen.* 9, 678–686

28 Davies, J.W. *et al.* (2006) Streamlining lead discovery by aligning *in silico* and high-throughput screening. *Curr. Opin. Chem. Biol.* 10, 343–351

29 Lipinski, C. and Hopkins, A. (2004) Navigating the chemical space for biology and medicine. *Nature* 432, 855–861

30 Gardner, S.P. (2005) Ontologies in drug discovery. *Drug Discov. Today* 2, 235–240

31 Walters, W.P. and Namchuk, M. (2003) Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* 2, 259–266

32 Perlman, Z.E. *et al.* (2004) Multidimensional drug profiling by automated microscopy. *Science* 306, 1194–1198

33 Agrafiotis, D.K. *et al.* (2002) Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug Discov.* 1, 337–346

34 Oprea, T.I. (2005) Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J. Biomol. Screen.* 10, 419–426

35 Feng, B.Y. *et al.* (2005) High-throughput assay for promiscuous inhibitors. *Nat. Chem. Biol.* 1, 146–148

36 Rhee, A.M. *et al.* (2001) Retrospective analysis of an experimental high-throughput screening data set by recursive partitioning. *J. Comb. Chem.* 3, 267–277

37 Curcin, V. *et al.* (2005) Web services in the life sciences. *Drug Discov. Today* 10, 865–871

38 Gardner, S.P. (2005) Ontologies and semantic data integration. *Drug Discov. Today* 10, 1001–1007

39 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58

40 Butcher, E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.* 22, 1253–1259

41 Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? *Nat. Rev. Drug Discov.* 4, 461–467