High Throughput Screening Informatics

Xuefeng Bruce Ling*

Biotechnology Core, Lucile Packard Children's Hospital, Stanford Medical Center, Stanford University, CA, USA

Abstract: High throughput screening (HTS), an industrial effort to leverage developments in the areas of modern robotics, data analysis and control software, liquid handling devices, and sensitive detectors, has played a pivotal role in the drug discovery process, allowing researchers to efficiently screen millions of compounds to identify tractable small molecule modulators of a given biological process or disease state and advance them into high quality leads. As HTS throughput has significantly increased the volume, complexity, and information content of datasets, lead discovery research demands a clear corporate strategy for scientific computing and subsequent establishment of robust enterprise-wide (usually global) informatics platforms, which enable complicated HTS work flows, facilitate HTS data mining, and drive effective decision-making. The purpose of this review is, from the data analysis and handling perspective, to examine key elements in HTS operations and some essential data-related activities supporting or interfacing the screening process, and outline properties that various enabling software should have. Additionally, some general advice for corporate managers with system procurement responsibilities is offered.

Keywords: HTS, data mining, sample bank, compound, drug target, lead discovery.

INTRODUCTION

The completion of the Human Genome Project has significantly advanced our understanding of human biology and the nature of many diseases, unraveling a plethora of novel therapeutic targets. Modern high throughput methodologies and robotics of remarkable efficiency and precision have significantly accelerated the industrialization of the drug discovery process, where millions of compounds are routinely screened to identify tractable chemical series (HTS hits) as starting points for drug design. Subsequently, HTS hits are transformed into lead series with the requisite biological activity against the target of choice, and can serve as tool compounds to understand the role of a particular biochemical process *in vivo*.

As a brute-force approach and a complicated automated industrial process, HTS "manufactures" unprecedented amounts of experimental data -- usually observations about how some biological entity, either proteins or cells, reacts to the exposure of various chemical compounds in a relatively short time. Due to the high cost, technical specialization and operational sophistication, small molecule lead discovery and the development of chemical research tools has been limited to pharmaceutical or large biotech companies. However, there is a recent trend towards academic pursuit of chemical screening due to the increasing availability of screening facilities [1]. Given the ever-increasing complexity of HTS work flows and emergence of high content screening technologies, innovative strategies of how to acquire and streamline new functionalities, evolve currently available software solutions, and ultimately deliver cost-effective, scalable and maintainable HTS computing platforms with limited budget, are operationally essential. The purpose of this review is, from the informatics perspective, to examine key elements and essential data-related activities supporting or interfacing the screening process, outline requirements and properties that various enabling scientific computing systems should have, and offer some general advice for managers with system procurement responsibilities. The breadth of the discipline of HTS has made it necessarily to constrain the scope and detail of this review in some areas. However, more detailed updates and additional supplementary information will be available online at http://hts.stanford.edu.

HTS KEY ELEMENTS IN BOTH PROCESS AND INFORMATICS

High-throughput screening is a critical link in the industrialized lead discovery chain. Fig. **1** diagrams the key operational elements and database systems necessary to support HTS. Typically, once a druggable target has been selected and prioritized for HTS by biologists with therapeutic area expertise, biology and HTS groups collaborate to develop an assay and transfer it to the HTS organization. HTS scientists need to optimize assay protocols for high throughput automation, ensuring the following key requirements are met: I) simplicity and robustness to allow unattended runs over extended periods, with constant quality, II) potential for miniaturization which saves compound and reagents by allowing screening in high density plates, III) production of a fluorescent, luminescent, absorbance, radioactivity or other signal amenable to detection with available plate reading systems. The primary and confirmatory screens against the compound collection, which routinely comprises a million or more of compounds of diversified structures, ideally lead to the identification of reproducible HTS hits. After compounds with undesirable chemical properties have been removed, potency, efficacy, and target selectivity are determined from dose-response experiments and counter screening. Additional data pertaining to compound purity and PKDM properties are typically compiled into a lead information package which is finally handed off to medicinal chemistry and the

*Address correspondence to this author at the Biotechnology Core, Lucile Packard Children's Hospital, Stanford, Stanford Medical Center, Stanford University, CA 94305-5164, USA; Tel: (408) 667-9454;
E-mail: xuefeng_ling@yahoo.com

1386-2073/08 $55.00+.00

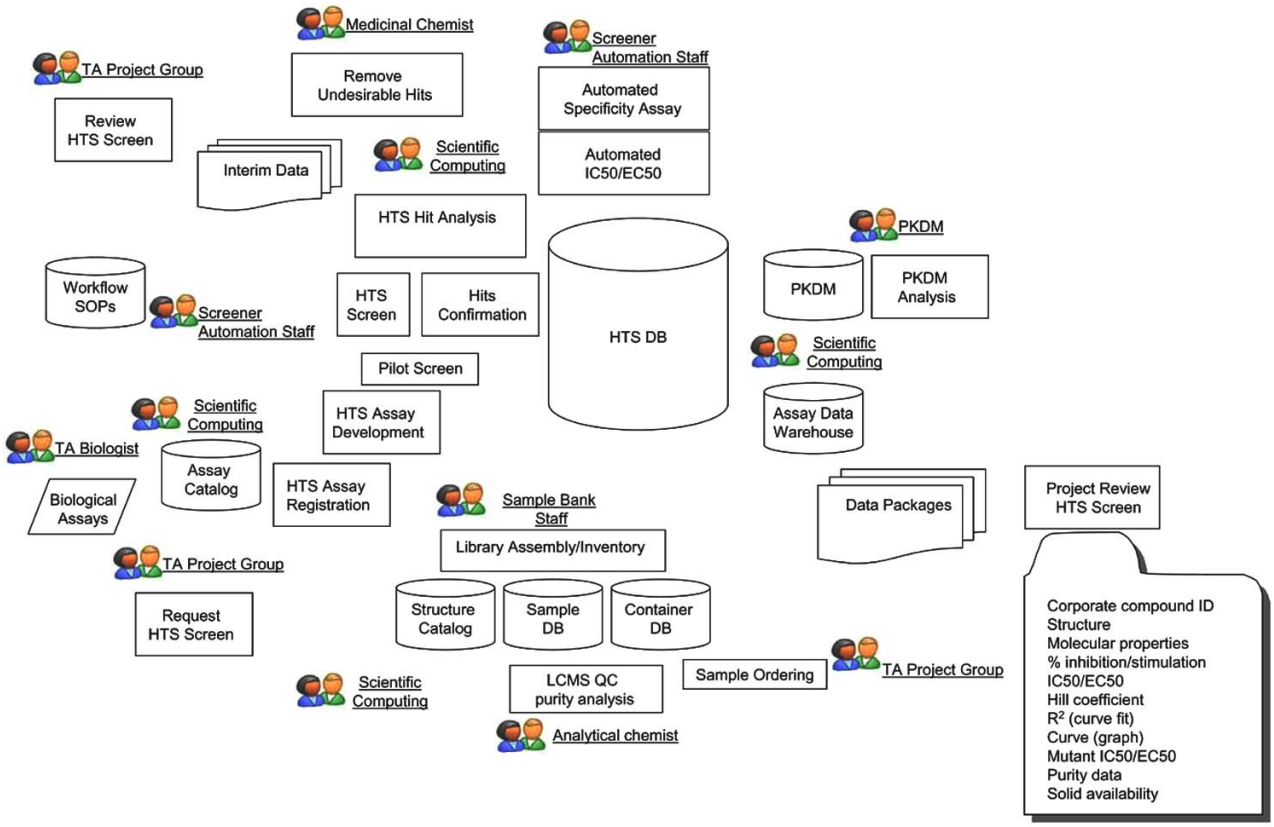

Fig. (1). Flow chart of HTS key elements.

therapeutic area where it provides the start point for future lead optimization.

Obviously, the development, deployment and maintenance of enterprise data handling systems are essential to enable these complex HTS workflows and facilitate optimal decision making under the various scenarios that may arise in the HTS arena. Data management requirements include real time tracking of information transactions, efficient interfaces to other operational and scientific databases, and robust storage of vast and diverse data sets including: I) compound structures, II) biological samples, III) various types of containers, IV) assay definitions, V) results from single-concentration and dose-response assays, and VI) workflow SOPs for assay/automation/inventory management. The integration of various information resources and interoperability between different software applications and robotic instruments are not only important for HTS operations, but also for multidisciplinary teamwork with other corporate functions including therapeutic areas, laboratory automation group, sample bank, analytical/computational/medicinal chemistry, scientific computing, and project teams. In general, scientific computing should consider the informatics elements outlined in Fig. **2**, [2]: I) a robust underlying Oracle data model, II) a well defined and structured meta data layer organized as ontologies and implemented in XML or Oracle to provide sufficient contexts for HTS data sets, III) fast data entry tools that are capable of processing large data sets and utilize controlled vocabularies to maintain data integrity, IV) fast and intuitive query tools, V) data navigation and reporting interfaces, and VI) intuitive visualization tools capable of drilling multi-dimensional data sets for quality control, database navigation and data mining. Similarly to the aftermath of genome sequencing where special ontological algorithms were essential to decode the genome [3], full exploitation of HTS data require novel data mining methods to convert information into knowledge, and support effective decision-making.

"BUY *VS* BUILD"

"Buy *vs* Build" has been a constant dilemma haunting managers responsible for HTS data analysis [4]. There are obvious advantages to building HTS applications in house. Through the software life cycle of business modeling, analysis and design, implementation, testing and final deployment, the company ultimately owns a system tailored specifically to its unique HTS processes. Additionally, the company retains an internal software development team with domain expertise to provide real time support and flexible customization, achieving substantial long term return on investment. However, given the complexity and dynamics of HTS operations, and budget/headcount constrains, development of a complete custom solution through internal resources is often impossible, even for large pharmaceutical companies. Thanks to the HTS tech-boom and intensive vendor software development, off-the-shelf software became available that is designed to support HTS "out of the box". Vendor lists (Table **1**), although not exhaustive, have been compiled and tabulated against HTS workflow or integration needs. Albeit at significant monetary cost, such software packages can rapidly jumpstart HTS operations in a new institutional setup, introduce best practices and standardized common vocabularies for HTS business process management, remove

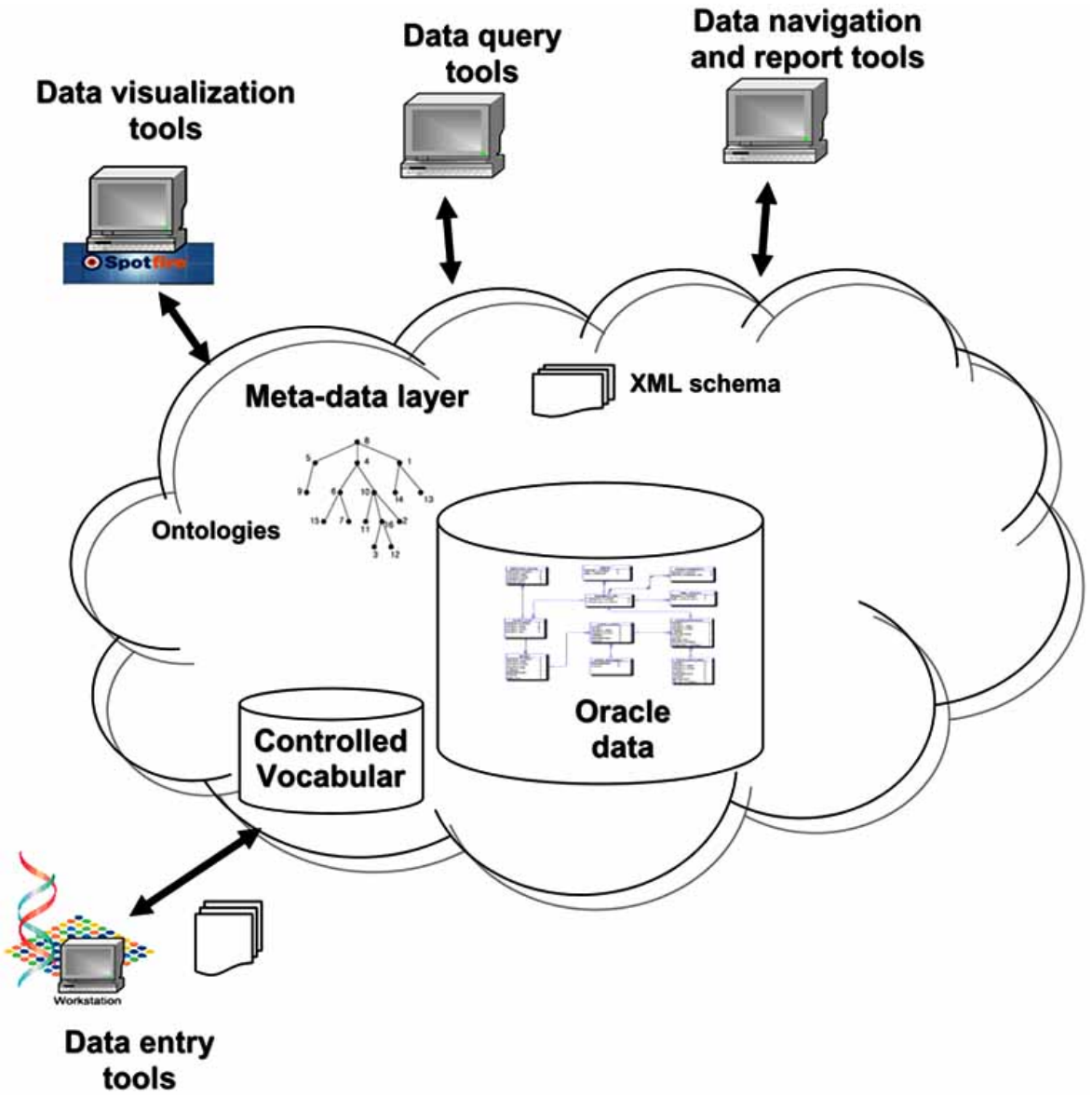

Fig. (2). HTS informatics key elements.

customization bottlenecks, or satisfy pressing needs resulting from newly acquired or developed HTS functionalities and capabilities. However, these software packages are mass-produced, each with some unique features, different strength and limitations, and consequently contain generic content not developed specifically for one particular organization or user group. A marketed product may serve parts of the HTS process well, but typically won't deliver a complete solution. Furthermore, special consideration must be given to downstream maintenance costs, which can be substantially larger than the initial purchase. Any feature enhancement, customization, or version upgrade requires vendor consultation, and product specific domain knowledge, which can lead to unexpected operational disruptions and unpredictable extra cost. In reality, effective lead discovery informatics has to balance both aspects of building proprietary *vs* licensing commercial systems, and it is important to avoid squandering internal programming resources by "reinventing the wheel" when a viable commercial solution is available at reasonable cost.

INTEGRATION OF ROBOTICS

The heavy use of robotic instrumentation characterizes modern industrialization of the drug discovery process, and large capital investments are defended not only based on labor cost savings, precision and higher throughput, but also by the financial gains associated with and shortening of drug discovery time lines. Many modern robotics systems come with some form of scheduling software in addition to instrument control software. Moreover, most, if not all, common robotic applications are designed for the Microsoft Windows platform. Nonetheless interoperability between different robotic systems can be extremely challenging simply due to the proprietary nature of vendor software, and the distribution, over time, of robotic applications on PC workstations with different Windows versions. Furthermore, due to the standalone nature of much robotic software, lack of proper integration with discovery Oracle databases, and platform incompatibility between Windows Desktop applications and UNIX based enterprise applications, it is often im-

Table 1. Vendors and their Corresponding Software Packages Supporting HTS

Vendor	Bioassay Registration	Sample Management	Structure Management	Assay Data Capturing	Activity Potency (End Point)	EC50/IC50 Dose Response	Integrated Data Warehouse	Data Visualization	Workflow Integration	Statistical Packages	Electronic Notebook	Website
Acitivitybase (IDBS)	+	+	+	+	+	+	+	+	-	-	+	http://www.idbs.com/activitybase/
Accelrys	+	+	+	+	+	-	-	-	+	-	-	http://www.accelrys.com/products/accord/
Biosoft	-	-	-	-	+	+	-	-	-	-	-	http://www.biosoft.com/
CambridgeSoft	+	+	+	+	+	+	+	+	-	-	+	http://www.camsoft.com/
Daylight	-	-	+	-	-	-	-	-	-	-	-	http://www.daylight.com
ChemAxon	-	-	+	-	-	-	-	-	-	-	-	http://www.chemaxon.com
Elsevier MDL	+	+	+	+	+	+	+	-	-	-	-	http://www.mdli.com/
GraphPad Software	-	-	-	-	+	+	-	-	-	+	-	http://www.graphpad.com/prism/
Genedata	-	-	-	+	+	+	+	+	-	-	-	http://www.genedata.com/
Insightful (S plus)	-	-	-	-	-	-	-	-	-	+	-	http://www.insightful.com
Inforsense	-	-	-	-	-	-	-	-	+	-	-	http://www.inforsense.com
R project	-	-	-	-	-	-	-	-	-	+	-	http://www.r-project.org/
Rescentris	-	-	-	-	-	-	-	-	-	-	+	http://www.rescentris.com/products.html
Systat software	-	-	-	-	+	+	-	-	-	+	-	http://www.systat.com/
SAS	-	-	-	-	-	-	-	-	-	+	-	http://www.sas.com/
Symyx	-	-	-	-	-	-	-	-	-	-	+	http://www.symyx.com/
The Mathworks (Matlab)	-	-	-	-	-	-	-	-	-	+	-	http://www.mathworks.com/
TIBCO (Spotfire)	-	-	-	-	-	-	-	+	-	-	-	http://www.spotfire.com
Tripos	-	-	+	-	-	-	-	-	-	-	+	http://www.tripos.com/
Waters	-	-	-	-	-	-	-	-	-	-	+	http://www.waters.com

possible to retrieve real time information, track the transactions of various business objects, and subsequently update related records in corporate Oracle databases with the results of the robotic operation. Sometimes, vendor robotic software offers APIs (application programming interfaces), usually as Microsoft Visual Basic (VB) and ActiveX controls, which can be extended to communicate to external Oracle databases. Nevertheless, the launch of the vendor application is still an event driven effort, requiring manual input and causing operational bottlenecks if different robotic equipment and enterprise Oracle based applications are required to work interactively. Thanks to the advent of the Microsoft.NET framework, additional application development can be used to "publish" the capability of a particular robotic Windows application as a "web service" such that the business logic, data and processes of any given robotic application can be shared through a programmatic interface across the corporate network. Through the integration of the three main robotic specific components characteristic of automation platforms (a robotic specific application, the hardware modules of the mechanical system, and the module level software), the .NET integration application allows live data transactions and interoperability with server based enterprise applications, including Oracle databases, and any other equipment vendor applications.

SAMPLE BANK AND COMPOUND MANAGEMENT

It can be argued that the impact of screening and the integrity of the screening data are only as good as the quality of the compounds being tested and the accuracy of the managed compound information. Successful HTS campaigns critically depend on the integration of sample bank functionality, both at the level of physical samples, structures, and containers as well as the flow of the associated data and real time tracking of the transactions [5]. Subsystems for chemical information management, sample registration, sample inventory, and sample ordering/reordering are required to prepare sample bank to support HTS operations and follow up studies. It is essential that such systems and the underlining Oracle databases provide enterprise-wide access and seamless integration with all HTS related databases and

automation tools. Introduction of Oracle data cartridge technology through several vendors' implementations has ultimately relieved the bottleneck of the chemical information management, allowing for custom indexing and searching of ever increasing numbers of chemical structures in a similar fashion to text and numeric data. However, migration of legacy systems to the Oracle cartridge technology demands extreme attention to detail including: a careful inventory of current in house systems, identification of interdependencies, and special treatment of a number of specific molecular and chemical representation features such as stereochemistry, tautomers, etc. Compound collections for screening are usually purchased from commercial sources, produced through combinatorial chemistry processes, or internally synthesized by medicinal chemists. A sample registration application registers non-biological compound data for newly acquired or synthesized compound lots and seamlessly integrates with the inventory system. Sample storage and retrieval systems, in particular those in production within large pharmaceutical companies, have evolved into a fully automated process where the inventory tracking system and sample dispensing automation are directly linked. The inventory system tracks the quantity of every sample in its various sample formats and associated containers in the archive upon sample addition or depletion. To support plate-based screening and subsequent data analysis, the inventory system also creates and manages plate map information such as compound-well location, sample molecular weight, and sample concentration of the compounds. By linking to the inventory database, sample ordering systems can allow the requester to view real-time inventory information, such as available quantity and format, and enables easy designation of the format in which the requested sample is to be delivered. Through coordination of concurrent requests and online negotiation with sample bank staff, large numbers of compounds can be efficiently delivered to global and multidisciplinary teams. Ultimately, the time and effort needed to access samples from the sample archive can be greatly reduced.

ASSAY DATA CAPTURING AND ANALYSIS

The proprietary software shipping with many plate reading instruments commonly limits the data output from HTS assays to a restricted number of characteristic signal(s) per well. Both the deficient interoperability between HTS readers and high throughput data capturing systems, and the inadequacy of available solutions that effectively standardize the capture and reduction of data sets from complex kinetic biological responses [6], have become major bottlenecks in high throughput data capturing. Limited cooperation across HTS communities, including HTS equipment vendors, has hampered the formal introduction of rigorous data exchange standards that could resolve this interoperability bottleneck. As a result, it is not trivial to push the different reader output formats to downstream HTS data management systems. To accommodate this, instrument control software has to be manually configured to export screening datasets as ASCII text files that can be subsequently parsed by the data capturing system. The introduction of new instruments or changes to reader control software requires either hard-coding of special "parsers" or manual amendment of "reader templates" used by the data capturing system. With more-widespread implementation of kinetic detection techniques to track cell signaling events, innovative multi-parameter analytical algorithms are needed to standardize the reduction and extraction of kinetic assay data for downstream processing.

Most HTS assays produce large numbers of individual measurements with high inherent variability and errors. Therefore, statistical methods are indispensable for efficient analysis of such noisy data [7]. Although running assays in duplicate has been recommended and can significantly reduce false positives and false negatives (http://iccb.med.harvard.edu/screening/guidelines.htm [7]), time and cost considerations have caused the routine generation of true replicate measurements in primary screens to be relatively rare across the HTS community. Control readings are essential to a well-designed assay, and every assay should be equipped with both plate-based and assay wide controls for complete monitoring of assay quality, dynamic range, and subsequent statistical normalization of signals to facilitate identification of reproducible hits. Special attention should be given to the commonly used terminology of controls, including "positive", "negative", "blank", which can be highly ambiguous and context-dependent. This can easily lead to errors in computing normalized compound activities, miscommunication between multidisciplinary teams, and long term database storage confusion and contradictions. To resolve these potential issues, and facilitate consistent data processing, Fig. **3** proposes one set of terms to describe screening well types. The expected signal levels of various controls have been tabulated for better illustration of this controlled vocabulary.

With continuous advances in server capabilities, algorithm development, and statistics packages, HTS scientific computation can currently process vast amounts of HTS data to calculate compound activity end values, perform routine nonlinear regression for dose response and enzyme kinetic parameters, and gauge assay reproducibility, performance, and sensitivity. With the statistical analysis applications in place, extensive validation of HTS assays can facilitate decision making before committing to an HTS campaign. Quality control parameters can guide scientists to accept or reject certain data sets and to keep process errors within acceptable ranges. While not intended to be comprehensive, examples of formulas (Table **2**) commonly used in HTS, have been complied for easy reference. For brute force primary and confirmation screening, %Activity, %Activation for agonist or %Inhibition for antagonist screening respectively, are usually the final end values of normalized compound activity. Both of the Z and B score methods normalize compound signals without using controls under the assumption that most compounds are inactive [8]. While harder to compute, the B score method introduces row and column correction and should be preferred if row or column biases are suspected [7]. The signal to background (S:B) ratio describes the dynamic range while the signal to noise (S:N) ratio serves as a metric for the "signal strength" of an assay. The higher the variability of an assay, the larger the S:B ratio should be in order to qualify an assay for HTS. Both signal window and Z-Factor [9] describe dynamic range of an assay. However, the Z-Factor evaluates the assay quality considering both the dynamic range and data variation, which makes it a more rigorous statistical criterion when assessing the suitability of an assay for HTS. It is acceptable practice to exclude data outliers from statistical calculations, which

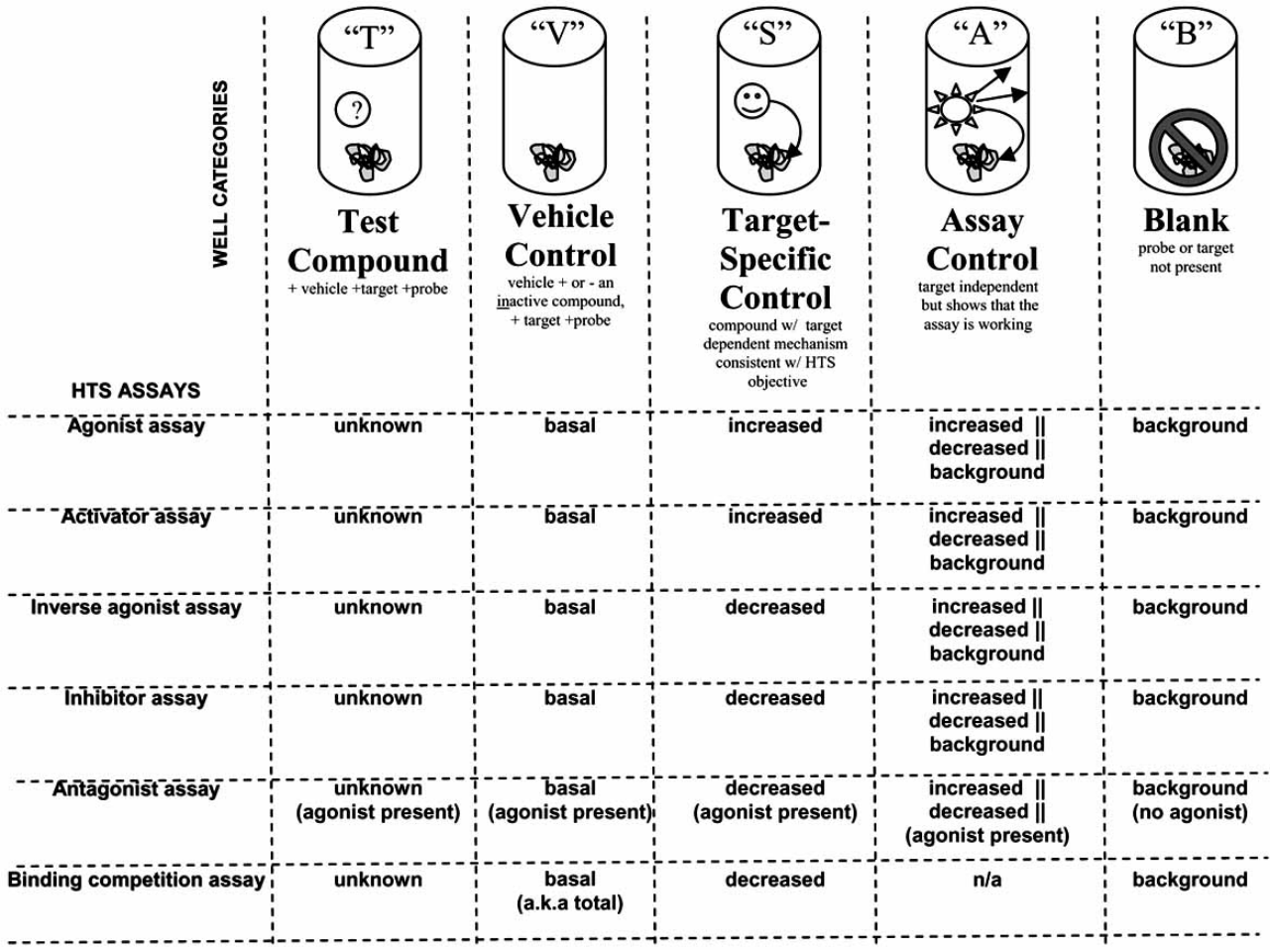

Fig. (3). Controlled vocabularies proposed to describe HTS screening well types. The expected signal levels of various controls are tabulated respectively.

increases the Z-factor value. If more than 2% of the control wells, and/or more than 5% of compound wells are discarded, a Z-factor should be considered artificially enhanced. In assay reproducibility tests, minimum significant ratio (MSR) and minimum significant difference (MSD) are important parameters for potency and efficacy evaluation respectively. The potency of compounds derived from dose-response experiments is typically represented by the EC50 or IC50 value which is defined as the compound concentration which produces 50% of the maximal response. The accuracy of EC/IC50 values varies widely depending on the methodologies used (i.e. replication, curve fitting method). The most common regression method for dose response curve fitting is the four parameter logistic model (4PL), also called the Hill-Slope model. When either the top or bottom asymptote is not available as a result of the compound potency falling outside the dosing range, 3PL model should be used to reduce fitting error and improve the curve fit. In situations where experiments cannot be repeated to yield better quality data, the data analysis platform should offer more interactive curve fitting and parameter options to report meaningful results. Because of its relative independence of assay conditions, K_i values derived by the Cheng-Prusoff equation for simple competitive assays can be more useful when comparing potency levels than IC50 values. The analysis of enzyme modulator kinetics can be challenging due to enzyme mechanism complications [10]. Non-linear fitting of the Michaelis-Menton equation derives enzyme kinetic parameters including V_{max} and K_m.

INFORMATICS ISSUES IN HIGH CONTENT SCREENING

High content screening, including high-throughput analysis of cellular and molecular images, has become an increasingly powerful tool in lead discovery, allowing acquisition of richer information around multiple biochemical or morphological pathways at the single-cell level at an early stage in the development of new drugs [11]. However, mature HCS data handling solutions for efficient acquisition and processing of large amounts of image data have yet to be developed [12]. Current HTS data centers may no longer be suitable for HCS data storage as a 1000-compound screen can easily consume 0.5 TB of disk space [11]. Familiar statistical measures of HTS like Z-factor and S:N ratios pre-suppose one measurement per well. Hence, they are no longer applicable to multi-dimensional HCS datasets. All of this demands a different data analysis strategy. A major unmet need in the HCS field is the development of robust tools for image processing in general and pattern recognition in particular, enabling researchers to quickly quantify phenomena on the

Table 2. Examples of Common HTS Computing Equations and Formulas

Name	Formula	HTS Accepted Range	Application	Definition of Terms
Compound signal	$\%Control = \left(1 - \frac{\mu_{max} - Y_{obs}}{\mu_{max} - \mu_{min}}\right) \times 100$		Normalize the readout to the controls	μmax: maximum response in the assay; μmin: minimum response in the assay; Yobs: observed response in the assay
Z score	$z = \frac{Y - \mu}{\sigma}$		Plate based compound signal normalization	б: s.d. of the plate raw measurement; μ: mean of the plate raw measurement. Y: test compound raw measurement
B score	$B = \frac{Rij}{MAD}$		Plate based compound signal normalization	Rij: the residual of the measurement for a particular plate row i and column j; MAD: median absolute deviation which is a robust estimate of spread of the residue values
Coefficient of variation	$\%CV = \frac{\sigma}{\mu} \times 100$	< 15%	Measure the dispersion of the measured signals	б: s.d. of the signal; μ: mean of the assay signal
Signal to noise	$S:N = \frac{\mu_{max} - \mu_{min}}{\sigma_{min}}$		Describe the "signal strength" of the assay	
Signal to background	$S:B = \frac{\mu_{max}}{\mu_{min}}$	> 2	Describe the dynamic range of the assay	
Signal window	$SW = \frac{\mu_{max} - \mu_{min} - 3(\sigma_{max} + \sigma_{min})}{\sigma_{max}}$	> 2	Describe the assay dynamic range	
Z-Factor	$Z = 1 - \frac{3\sigma_{sample} + 3\sigma_{control}}{\lvert\mu_{sample} - \mu_{control}\rvert}$	> 0.6	Describe the assay quality considering both dynamic range and data variation	
Z'-Factor	$Z' = 1 - \frac{3\sigma_{c+} + 3\sigma_{c-}}{\lvert\mu_{c+} - \mu_{c-}\rvert}$	> 0.6	Describe the assay quality considering both dynamic range and data variation	C+: positive control; C-: negative control
Minimum significant ratio	$MSR = 10^{2\sigma d}$	< 3.0	Use in reproducibility test of potency values	бd: standard deviation of the run difference in log-potency
Minimum significant difference	$MSD = 2\sigma d$	< 20	Use in reproducibility test of efficacy values	бd: standard deviation of the run difference in efficacy
Hill-Slope model (four parameter logistic model 4PL)	$y = bot + \frac{top - bot}{1 + (x / EC50)^{slope}}$	Fitting error < 40% of EC/IC50	Use to fit a dose-response curve to obtain the EC/IC50 when both asymptotes can be defined by the data	Relative EC/IC50: "top", "bot" are the fitted top and bottom of the curve. Absolute EC/IC50: "top" and "bot" are defined by the assay dynamic range where "top" is the max control level and "bot" is the min control level
Hill-Slope model (three parameter logistic model 3PL)	$y = bot + \frac{top - bot}{1 + (x / EC50)^{slope}}$	Fitting error < 40% of EC/IC50	Use to fit a dose-response curve to obtain the EC/IC50 when either asymptotes can not be defined by the data and a fixed value will be used instead	If the data do not define a top asymptote, then fixing the top at 100%. If the data do not define the bottom asymptote, then fixing the bottom at 0%
Michaelis and Menton	$v = \frac{[S]V_{max}}{[S] + K_m}$		Calculate Vmax and Km through the nonlinear regression analysis	V: reaction rate; Vmax: max reaction rate; [S]: substrate concentration; Km: Michaelis-Menton constant
Cheng-Prusoff	$K_i = \frac{IC_{50}}{(1 + [S]/K_m)}$		Calculate Ki under condition of competitive inhibition	[S]: substrate concentration
Cheng-Prusoff (ligand-binding)	$K_i = \frac{IC_{50}}{(1 + [L]/K_m)}$		Calculate Ki under condition of competitive inhibition	[L]: ligand concentration

cellular and subcellular level and draw meaningful conclusions from the derived visual data. Additional challenges, including the cataloging of vast number of images, interoperating between HCS platforms and peripheral compound inventory databases, and integrating HCS with other HTS screening data, have yet to be resolved as well.

ELECTRONIC LABORATORY NOTEBOOKS

Today, lead discovery research activities, including experimental design, planning, and execution as well as data collection, processing, and reporting, rely heavily on computer systems. The initial drive for the laboratory notebook to go "electronic" came from the computerized chemical information management arena, and the use of HTS introduced biologists to enterprise computing platforms. Recent FDA encouragement of electronic submission and the relaxation of its interpretation of salient regulations (21 CFR part 11) has started to remove general concerns of whether electronic records provide sufficient evidence of invention, driving the widespread adoption and implementation of electronic laboratory notebook (ELN) systems across US pharmaceutical companies [13, 14]. Current vendor implementations provide for the capture and management of chemical structures and reactions, analytical data such as spectra, chromatograms and parameters, the text of authored reports and comments, spreadsheets and tables, images and drawings, scans and other multimedia files. Scientists can use ELNs to capture, process, save and search notebook data in a completely digital, networked environment, driving productivity and collaboration. However, in contrast to HTS and sample management systems, an ELN mainly manages unstructured data. While their data model is well structured for chemical structures and reactions, ELNs usually lack a rigorous software framework for other data types. Therefore, interoperability and integration of ELNs with enterprise HTS and sample management applications remains an unresolved issue if ELNs are to unleash their full potential. With the widespread use of ELNs and continuous developments from the software vendors, ELNs are expected to become fully integrated, or even eventually merge with HTS and/or sample bank applications, offering a "one-stop-shop" for scientists to fulfill their data analysis and management needs across the complete spectrum of drug discovery.

FROM DATA TO KNOWLEDGE

Data mining aims to identify new patterns and deeper insights once different types of datasets can interoperate with each other and data integrity can be managed rigorously. When HTS data continue to grow exponentially in size and complexity, a major challenge is to empower scientists, who may not necessarily be computationally savvy, to navigate diverse data in meaningful ways. In the current reality, only computationally skilled individuals can exploit a broad range of computational and statistical approaches to sort, drill and report data.

One effective means of exploring large datasets for expected or unexpected trends or outliers is to use visualization techniques. Several commercial packages such as Spotfire™ Pro (TIBCO, USA) offer advanced data visualization capabilities, which can be integrated for HTS quality control, database navigation and data mining. However, the effectiveness of rendering HTS data sets directly influences the outcome of the data visualization and subsequent mining process. For comprehensive and effective data drilling through visualization interfaces it is imperative to compile data sets from different Oracle tables or even different Oracle instances such that relevant information can be digested at many levels of detail and from different perspectives. Key technical issues, including the design of the visualization layout, efficient creation of integrated data views, indexing, and proper update of the underlying data sets, need to be addressed to aid HTS data mining.

There is not yet a community standard that spans industry and academia to analyze and describe HTS information. However, the recent academic pursuit of screening has led to proposed guidelines for reporting small molecule HTS data [15]. From the HTS informatics perspective, this reporting standard is one type of metadata, trying to satisfy data interoperability after the completion of the screening process. For many years, metadata, technically "information about information", have been recognized as a significant component of the digital information environment for effective knowledge management. Indeed, it is has been long speculated that the comprehensive creation and effective management of metadata enable Google to seize search leadership from Yahoo. To enable intelligent knowledge management solutions for lead discovery, the enterprise or ideally the entire screening community needs to promote the widespread adoption of metadata standards and the development of specialized metadata vocabularies for standardizing multi-dimensional HTS data contents. The types of metadata to be stored and decisions as to how they should be structured must necessarily have a basis in the need to address currently unanswered and upcoming scientific questions. These metadata structures will serve as computing guidance to describe underlying data, provide data integration contexts, and cluster and join related data types according to common HTS attributes. The ultimate utility of HTS metadata sets depends solely on the quality and comprehensiveness of the HTS domain knowledge encapsulated. Therefore, special attention and resources should be allocated for this purpose. Rigorous evaluation and prototyping with the participation of multidisciplinary teams is strongly recommended. Given the dynamics of HTS business process, metadata should evolve accordingly. eXtensible Markup Language (XML) is ideal for metadata implementation as XML schemas are designed to be constantly evolving to address changes to business requirements and can be readily exploited by an analysis application, database or report generator. The construction of a well structured HTS metadata set and subsequent application development is expected to trigger a move towards effective business standardization, encouraging scientists to globalize research methodology. It is anticipated that innovative scientific computing capabilities, including a "Google" like search engine, will help to consolidate various lead discovery information under a knowledge-driven user interface for HTS data mining. Additionally, integration and interoperability of current applications, either acquired or built in house, will drive effective and efficient business integration.

OPPORTUNITIES AND FUTURE TRENDS

The ability of lead discovery to deliver high quality leads depends critically on the effectiveness and efficiency of sci-

entific computing. With the ever increasing complexity and sophistication of HTS operations, this review will end with a quick foray into some important future trends impacting HTS informatics.

First and foremost is the *emergence* of community based HTS data standards. Recently proposed guidelines as how to report screening results [15] represent one public effort in this direction. With the rapid expansion of the academic pursuit of chemical screening, it is fair to assume that this trend will persist, and ultimately lead to fully open, heterogeneous, and standardized HTS solutions. One dimension, both academia and pharmaceutical companies will benefit from and leverage, is interoperability.

Second is the *emergence* of trends to merge or consolidate various HTS related enterprise applications, including ELNs (electronic laboratory notebook), EDMs (enterprise data management system) and LIMS (laboratory information management systems), into one software framework.

Third is the *emergence* of a totally different data analysis strategy for high throughput content screening. Such an implementation would greatly influence HTS informatics in general.

Fourth is the *emergence* of an open, centralized repository, PubChem, for high throughput screening data. This effect could be similar to that observed in the course of human genome project. Once a critical mass is reached, the aggregated heterogeneous data contents may not only revolutionize academic research but also have a long term impact on industrial drug discovery.

Finally, "the increasing availability of data related to genes, proteins and their modulation by small molecules has provided a vast amount of biological information leading to the *emergence* of system biology and the broad use of simulation tools for data analysis [16]".

In summary, the aggressive pursuit of lead discovery, both in academia and industry, continuously drives the evolution of HTS scientific computing to deliver solutions effectively and efficiently support discovery decisions. In this regard, informatics and lead discovery are gradually engaging each other as partners in the discovery of new medicines or academic research tools.

ACKNOWLEDGEMENTS

The author is thankful to colleague scientists Drs. Tim Hoey (Oncomed pharmaceuticals), Marc Labelle (Lundbeck research USA, Inc.), Robert Bukar (Kalypsys Inc.), Steve W. Young (Amgen Inc.), Josh Xiao (Amgen Inc.), James Schilling (Stanford University), Yannick Pouliot (Stanford University), Jane Liu (USC Medical Center) for critical discussions.

ABBREVIATIONS

HTS	=	High throughput screening
HCS	=	High content screening
ELN	=	Electronic laboratory note book
FDA	=	Food and Drug Administration
PKDM	=	Pharmacokinetics and drug metabolism

REFERENCES

[1] *Nat. Chem. Biol.*, **2007**, *3*, 433.
[2] Roberts, B.R. *Drug Discov. Today*, **2000**, *1*, 10-14.
[3] Pouliot, Y.; Gao, J.; Su, Q.J.; Liu, G.G.; Ling, X.B. *Genome Res.*, **2001**, *11*, 1766-79.
[4] Weaver, D.C. *Pharm. Discov.*, **2005**, *1*, 42-47.
[5] Sofia, M.J.; Stevenson, J.M.; Houston, J. *Pharma DD*, **2005**, *1*,
[6] Simpson, P.B.; Wafford, K.A. *Drug Discov. Today*, **2006**, *11*, 237-44.
[7] Malo, N.; Hanley, J.A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. *Nat. Biotechnol.*, **2006**, *24*, 167-75.
[8] Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A. *J. Biomol. Screen*, **2003**, *8*, 634-47.
[9] Zhang, J.H.; Chung, T.D.; Oldenburg, K.R. *J. Biomol. Screen.*, **1999**, *4*, 67-73.
[10] Copeland, R.A. *Enzymes - A proactical introduction to structure, mechanism, and data analysis.* A John Wiley & sons, Inc.: 2000.
[11] Carpenter, A.E. *Nat. Chem. Biol.*, **2007**, *3*, 461-5.
[12] Zhou, X.; Wong, S.T.C. *IEEE Sign. Proc. Mag.*, **2006**, *23*, 63-72.
[13] Du, P.; Kofman, J. *J.A.L.A. Charlottesv. VA*, **2007**, *12*, 157-165.
[14] Taylor, K.T. *Curr. Opin. Drug Discov. Devel.*, **2006**, *9*, 348-353.
[15] Inglese, J.; Shamu, C.E.; Guy, R.K. *Nat. Chem. Biol.*, **2007**, *3*, 438-41.
[16] Oprea, T.I.; Tropsha, A.; Faulon, J.L.; Rintoul, M.D. *Nat. Chem. Biol.*, **2007**, *3*, 447-50.

Received: August 8, 2007 **Revised: November 28, 2007** **Accepted: December 3, 2007**