Leveraging process integration in early drug discovery

Nicolas Fay and Dirk Ullmann

Recent advances in new analysis and prediction concepts in informatics, statistics and computational chemistry have drawn attention to mining the enormous flood of information generated from ultra-high-throughput screening (uHTS) and early drug discovery more effectively. This review analyses current infrastructure and process concepts in data analysis, storage and mining, with a particular focus on high-throughput technologies. It also provides examples of how these techniques have been applied successfully together with underlying reasons for these developments.

***Nicolas Fay and Dirk Ullmann**
Evotec OAI
Schnackenburgallee 114
D-22525 Hamburg
Germany
*tel: +49 40 5608 1253
fax: +49 40 5608 1222
e-mail: nicolas.fay@evotecoai.com

To obtain a sufficient number of high-quality leads for drug development from ultra-HTS (uHTS), it was often desirable to test as many molecules as possible during a screen. As a consequence, compound collections used in uHTS have expanded and diversified dramatically over recent years. Miniaturized assay system toolboxes that are capable of evaluating more than 300,000 compounds a day have been developed [1,2]. At the same time, the enormous quantity of hits obtained from uHTS presents new challenges, such as: (1) how to mine the data to select the correct compounds for subsequent screening campaigns, (2) how to 'weed out' hits that appear positive but that are actually screening artefacts, and (3) how to select compounds contained in screening libraries to make screening more efficient in the first instance.

Statistics, informatics and computational chemistry have the potential to play a tremendously important and diverse role during all steps of drug discovery [3,4]. Using their strengths and implementing them into an efficient technical infrastructure can transform uHTS into a networking key element (simultaneously interacting with other elements) within the drug discovery process. This review discusses current trends in technology and process elements for these data integration techniques.

Data inventory

There have been staggering advances in recent years in both life science technologies and in computing and storage technologies. These developments promise to be able to cope with an increase of 100% in data density over 18 months and to build a solid basis for extracting the knowledge that is potentially buried in the hundreds of terabytes of data that will be accumulated in the near future. With the advent of high-throughput technologies, such as combinatorial chemistry and HTS, the challenge of efficient data management has become more important than ever. Integrated and highly automated screening systems are delivering particularly high throughputs, thus requiring intelligent analyses and boosting the demands for computational and storage capabilities to the limit.

Designs for data inventories focus increasingly on support of all processes, ranging from follow-up analysis and data refinement to information extraction and decision support. Following a current trend, this is often realized as a three-tier structure, comprising a data warehouse, a middle-layer and an application layer, which promises to provide efficiency as well as flexibility. An interesting example of such an implementation is the collaboration of Aventis (http://www.aventis.com) and IBM (http://www.ibm.com), where DiscoveryLink (IBM) was deployed to integrate Aventis's heterogeneous information sources on a global scale. The application layer (application portal) serves the graphical user interface (GUI) software, such as visualizers, chemical structure database GUIs or spreadsheets, whereas IBM's middle-layer links genomics-, HTS-, chemistry-, biology- and textual data from across all Aventis sites with the application portal, resulting in significant time savings in regular operation.

1359-6446/02/$ – see front matter PII: S1359-6446(02)02439-X

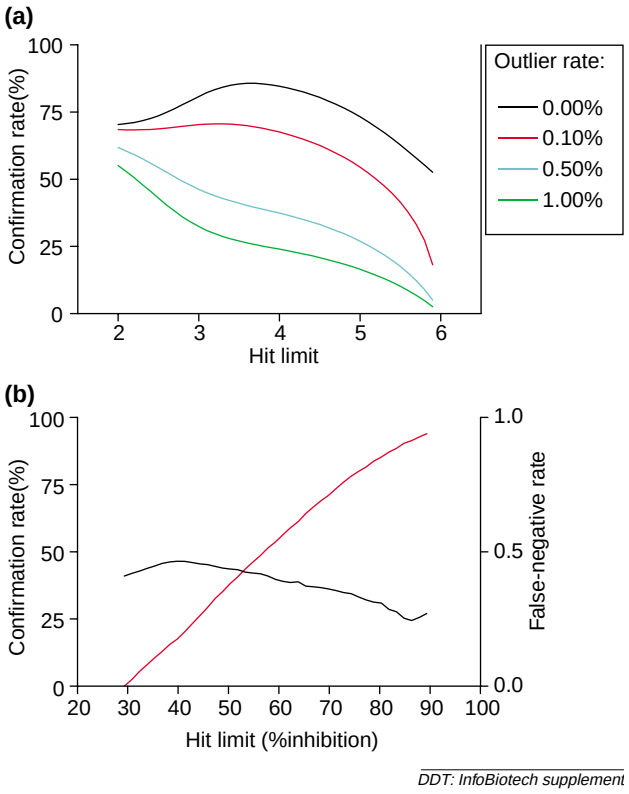


Figure 1. Numerical solution of a model (based on Ref. [7]) of the confirmation rate versus the hit selection threshold. (a) A normally distributed population of negative controls, a uniformly distributed population of hits and a uniformly distributed population of false positives are taken as inputs for the computations. The solution confirms the behavior of the confirmation rate of experimental data from HTS both in terms of curve shape and also numerically (see Fig. 1b). $\Delta\sigma$ is the distance above the mean of all negative controls in standard deviations of the negative controls population. (b) Screening result data taken from HCV-protease assay. The confirmation rate reaches a maximum of 45% at a hit selection threshold of 40% inhibition and then drops steadily. At the same time the rate of false negatives increases significantly. This illustrates that choosing a lower threshold resulting in a slightly lower confirmation rate helps to keep the number of false negatives as low as possible.

Successful implementation of these requirements in terms of performance, as well as flexibility or modularity, establishes a suitable source for the next step on the path of data refinement and knowledge extraction.

Online transaction and analysis processing

With rising throughputs, not all accumulated raw data are ready for transfer to a data warehouse or even use in follow-up analyses. Reduction of raw data and enhancement of transactional data is essential at this point to ensure only relevant information is passed for further analyses. These steps are usually termed 'online quality control', 'online transaction processing' and 'online analysis processing'. An illustration of selected process steps that are applied to enhance data quality is given in the next section.

Online quality control is especially important in HTS programs. Real-time quality control enables immediate retesting of failed measurements, which might occur because of, for example, pipetting errors. However, in addition, the biology itself might generate misleading patterns and the degradation of assay ingredients could bias the assay response. To control the quality, not only hardware logs but also statistical measures should be applied to control both the variability as well as the size of the screening window [5].

This pre-processing is a prerequisite for an accurate determination of whether a compound is a 'hit' or a 'miss'. Several approaches exist to select hits from the screening results. One of these approaches chooses a low theshold value at 3σ percent inhibition to separate hits from misses (σ is the standard deviation based on all inhibition values of the negative controls, assuming these values are normally distributed) and gives a confidence of ~99%. The advantage of this method over a basic ranking (that is, taking as many of the compounds with highest measured activity as can be pursued in follow-up assay studies) is that false positives that often show apparent high activity (e.g. because of pipetting errors) are not enriched in the final hit selection. Furthermore, the method enables the identification of weak-affinity compounds that might belong to a different structure class to other, highly active members of the hit population. These weak-affinity compounds could be a new starting point for the optimization of a novel, not yet patented chemical structure. At the same time, the number of false negatives is limited and this is considered advantageous because false negatives that have been excluded before 'cherry picking' are irrevocably lost, whereas false positives can be reliably identified and discarded before costly follow-up studies during hit confirmation.

Surprisingly, the confirmation rate, which is defined as the ratio of the number of confirmed hits to the number of hits selected from the primary screen, does not necessarily increase with higher thresholds. This is observed with real data (Fig. 1a), as well as with numerical solutions to simplified models ([6] and Fig. 1b). The reason for the decrease of the confirmation rate in this case is the presence of false positives above any chosen threshold (assuming a normally distributed population of non-hits). Adding a sparse real-hit population with activities evenly distributed between negative and positive controls lets the confirmation rate temporarily increase with rising threshold, whereas with even higher thresholds the influence from the false positives dominates again and the confirmation rate decreases. A significant proportion of false positives could occur through fluidics issues such as pipetting errors. Because the false positives from pipetting errors often result in very high activity values, a rising hit selection threshold favors these

false positives and enriches them in the final hit selection, resulting in a fully misleading value for the confirmation rate.

Although the distributions of the populations under consideration are heavily idealized, enabling compact mathematics, the actual results confirm exactly what is observed from uHTS campaigns. Such campaigns generating an increased proportion of these false positives do lower the confirmation rate significantly (Fig. 1b), whereas the confirmation rate does not tend to increase with rising threshold (Figs 1a and 1b), thus suggesting that a high threshold is rather disadvantageous.

Many of the reasons for false hits or misses are because of simple quality issues described previously. One of the serious obstacles specific to fluorescence techniques is autofluorescence, which is the term for an additional undesired fluorescence signal originating from the compound itself. Unfortunately, a bulk fluorescence measurement yields insufficient information to determine whether a compound shows real activity or whether autofluorescence mimics activity. This problem can be overcome through modeling the measured photon-histograms. The power of this method is the ability to resolve different bright species in the assay. It is therefore possible to resolve assay-signal and undesired autofluorescence even within single measurements of individual compounds (FCS+plus, Evotec OAI; http://www.evotecoai.com). Other established approaches rely on the analysis of the entirety of all measurements. They attempt to identify outliers in the dataset through modeling the relationship between bulk fluorescence intensity and polarization values. From this relationship, the proportion of autofluorescence can be estimated [1]. Others combine both of these approaches with principal component decomposition. The latter reduces many parameters obtained from the assay signal to those that contribute to the variability of the data most of all, by building linear combinations of the parameters, which are known as principal components. The methods try to predict compound behavior with regard to fluorescence artifacts with the partial least squares algorithm. Regular compounds, control inhibitors and certain dyes that mimic autofluorescence are used to calibrate the model [7]. Furthermore, prescreening of the compounds (i.e. screening without the bright ligand) is an approach to gather information for the compound's potential for autofluorescence. This can be used to flag the affected compounds.

Data mining

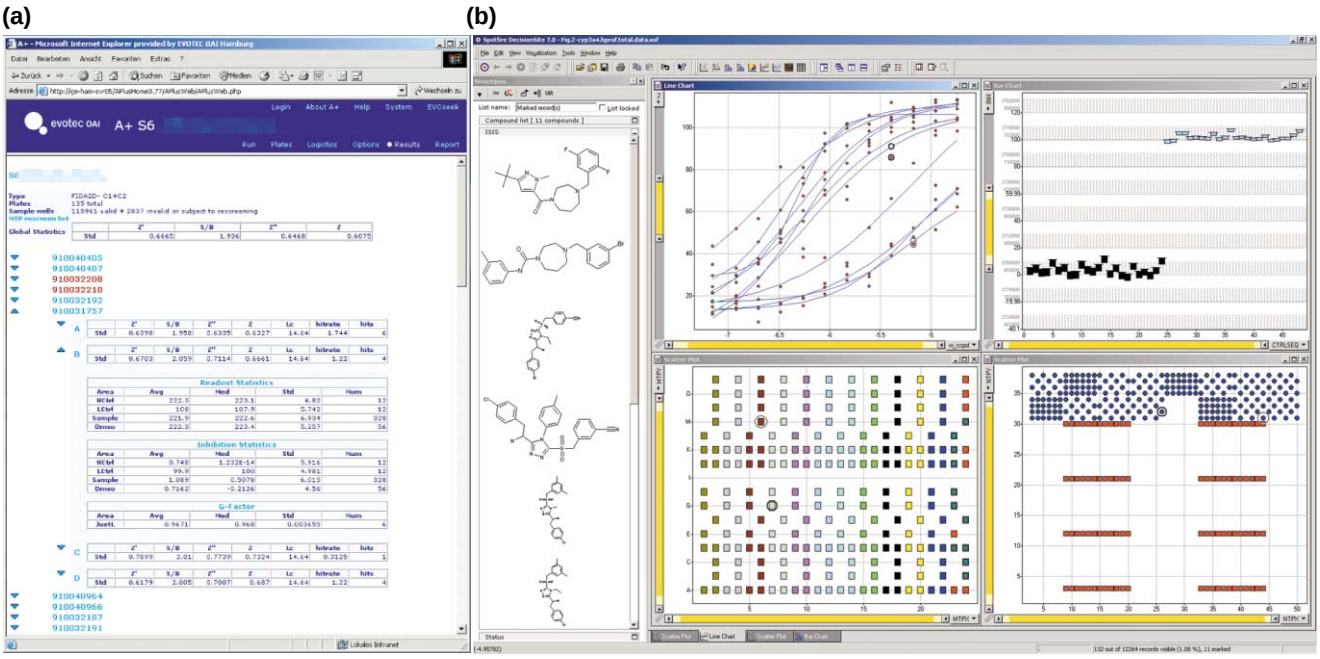
Once data are captured and refined, the next question to answer is what to do with the tremendous amounts of data that have finally been saved to the warehouse and, in addition, how to integrate the necessary tools for accessing and evaluating the data. A considerable number of software packages are now available for this process. However, none of them provides universal procedures that process data and create knowledge from it. Expertise is still required to choose proper algorithms and to apply them correctly. The goal is to detect patterns or relationships in the data that might lead to hidden information thereby enabling proactive, knowledge-driven decision-making. Because data mining is not a black box with simple inputs and useful outputs, applying data mining tools is an interactive and iterative process. Data mining tools are layered on top of the data warehouse infrastructure. Ideally they are fully integrated and can be used without additional data preparation efforts, such as transfer to a separate database or conversion to a propretary data format. They benefit particularly from advances in parallel computing, development of new algorithms and database technologies. The capability not only to handle data retrospectively but also to build powerful analytical models and derive probabilities for certain results (i.e. reliably predict trends and behavior) is facilitated through these developments. These include developments in statistics, artificial intelligence and rule-based approaches, as well as in clustering and classification methods [8].

Computational chemistry has received increased recognition as not only an alternative approach for hit identification, but also as a complementary method to 'wet'-screening. Although many of the tools developed and applied in chemoinformatics are still better applied by experts aware of potential pitfalls, attempts are being made to provide tools appropriate for medicinal chemists [9,10].

Fortunately, the sheer volume of uHTS data available for computational model building enables routine evaluation of these models. For this reason, early assessment of ADME/Tox (absorption, distribution, metabolism and excretion/toxicology) properties is receiving considerable attention and resources [11] Computational toxicology and chem-tox informatics are newly emerging fields that might change the way drug discovery and development interface in the future [12].

Integration and access

Cutting-edge infrastructure comprising both storage and algorithms, is beneficial only as long as it is realized as a network of transparent and smooth interacting components. Here, proper integration can guarantee significant efficiency in applying the individual algorithms. The future of these integration efforts probably belongs to protocols that manage communication among distributed applications such as Remote Procedure Call (RPC)-style web services. Ideally, they do not depend on a particular operating system, but aim for standards of communication that can be transported over the regular Internet protocol backbone, while not requiring platform-dependent runtime support at all. Web services work transparently and are thus different from traditional client-server implementations that provide GUIs. Instead, these GUIs are often web browsers. They are the top level of multi-tier components, and the success



DDT: InfoBiotech supplement

Figure 2. An Internet browser is used to connect to the main HTS data server A+. (a) Navigation, analysis, visualization and reporting are provided for simultaneous and independent user sessions through a convenient and highly integrated front end. Users can navigate through a combined raw data and result data tree. Analysis results are updated immediately upon changes in the data set. Detailed result visualization is provided through a direct mouse click on proper items in the tree, which pops up views such as the one shown in Fig. 2b. Every operational feature up to the creation of complete textual screening reports can be accessed from here. (b) Visual data exploration using DecisionSite for Lead Discovery (Spotfire Inc., Cambridge, MA, USA). The Figure shows selected IC50-curves obtained from a 12-point dose-response test of compounds whose chemical structures are simultaneously displayed on the left-side of the panel. Several control views on the ultra-high-throughput screening (uHTS) plate are enabled in this example. The full palette of DecisionSite features can be applied. This enables the user to quickly characterize assay performance and to identify visually interesting compounds (raw data, data pre-processing and refinement by Evotec OAI, Hamburg, Germany). Data is taken from a P450 isoenzyme screen.

of deploying and applying tools depends on efficient use of these user-interfaces. Not only is the integration of individual tools required for success, but also for intelligent guidance through a series of tools or steps, including seamless visualization of results and summaries (i.e. guiding the workflow).

An example that illustrates the use of such technologies in the surroundings of online quality control and online data analysis is the front-end that deals with the primary data from the screening system. Its purpose is to assess and refine the raw data from the measurement as early as possible to prepare validated secondary data for further processing on a smaller scale. To provide these capabilities network-wide, it has to meet high-load requirements with regard to input/output-performance, computing power and multi-user and simultaneous access capabilities.

Ideally, navigation as well as access to individual analysis tools is intuitive and self-explanatory. Figure 2a gives an example of an implementation that follows the guidelines of high integration and workflow support. The figure shows an html page that provides both access to the data to be analysed and calculation directives, including loading of screening runs, detailed plate-editing features or a link to initiate the creation of a full textual screening report. The currently selected html frame shows run summary information above a hierarchical tree of the barcodes of all screening plates. Statistical key values, such as accuracy and stability of the assay, along with signal statistics for each screening plate can be appraised in one view. Run names and barcodes are html-linked to open either the full run data or selected screening plate data in a Spotfire window (Spotfire Inc., Cambridge, MA, USA) (Fig. 2b). It shows how the whole environment is able to create a fast and convenient view of all dose-response measurements, which used to be performed in secondary process steps, rather than being fully automated on uHTS equipment. The open and modular design enables the user to implement new analysis components in a plug-in concept, which is important, especially in advanced software environments.

Close integration into an ensemble of data preparation, transforming, storing, exchanging or number-crunching components enables smooth operation, makes information

exchange as transparent as possible and minimizes the necessity for the operator's special knowledge of the underlying structures. This might culminate in a service that supports the whole process, beginning with raw data validation and ending with a complete textual or printable report on the full data set.

Concluding remarks

In recent years, efforts made by the pharmaceutical industry in assay miniaturization of disease targets for uHTS have undoubtedly been successful. As a result, there are significant savings in the use of reagents and hits are identified in larger numbers more quickly. Persistent concerns regarding miniaturized uHTS technologies have partly been because of the poor quality and insufficient reliability of the data. However, there have been improvements in all areas of uHTS, from fluidics and signal detection to the algorithms applied in the analysis of screening data and the embracing software architectures. This establishes the pre-eminent role of uHTS in early hit discovery, and makes it a fundamental pillar of the drug discovery process.

The amazing speed of co-developments in information technology paves the way for handling the flood of raw data more effectively, and should put an end to the approach of only optimizing throughput in the hope of exploiting the full potential of uHTS. The challenge is rather building an infrastructure that facilitates knowledge extraction at any proper time, but preferably as early as possible in drug discovery process [15–22].

This is one of the tenets that are enabled through new algorithms and the awareness of integration: to shift important decisions considerably toward earlier stages of the drug discovery pipeline. An important role here is provided by computational chemistry and computational biology, which have experienced significant progress and their current steep ascent seems to be only now beginning. Knowledge and methods are already sufficiently mature to work in a multi-tier environment of drug discovery programs. An illustrative example is the application of prediction tools for pharmacokinetic properties within uHTS because pertinent model systems such as liver microsome assays or P450 enzyme assays are well adapted to miniaturized uHTS systems. This bears the potential to significantly reduce the number of substances that are excluded from further expensive follow-up studies because of their side effects, no matter whether they have undesired ADME properties or whether they are toxic, which is one of the major reasons for their failure in trials.

A similar approach to benefit from the synergies of combined 'wet'-uHTS and *in silico* studies has been termed the 'sequential screening paradigm' [13,14]. The idea of sequential screening is that it uses knowledge obtained from *in silico* tests to direct further uHTS. A potential weakness of randomly testing substances is the size of the chemical space (i.e. the number of possible chemically different structures). Even if complete libraries comprising several million compounds were tested, it is a negligible fraction of what could be interesting for a particular target. Therefore selection of compound library subsets that augment certain desired substructures could be realized during directed uHTS, resulting in an iterative cycle of synthesis, testing and refinement.

Although wet screening benefits, for example, from random testing by the possibility of finding new and non-patented structural classes of hit compounds 'by accident', it does not compete with computational or *in silico* techniques. Rather, the joint development of both *in silico* and wet methods derived from genomics, proteomics, chemoinformatics and medicinal chemistry bears the potential for the next breakthrough in drug discovery methodology.

Acknowledgments

We would like to thank Maciej Hoffman-Wecker and Pierre Ilouga for ongoing contributions to the development and implementation of the statistics, and Matthias Biel for programming our uHTS analysis framework A+. We also thank Joe Mernagh for valuable discussions and Tom Mander for reviewing the manuscript.

References

1 Turconi, S. *et al.* (2001) Real experiences of uHTS: a prototypic 1536-well fluorescence anisotropy-based uHTS screen and application of well-level quality control procedures. *J. Biomol. Screening* 5, 275–290

2 Wölcke, J. and Ullmann, D. (2001) Miniaturized HTS technologies – uHTS. *Drug Discov. Today* 6, 637–645

3 Bajorath, J. (2001) Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* 6, 989–995

4 Rusinko, A. *et al.* (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* 39, 1017–1026

5 Zhang, J-H. *et al.* (1999) A simple statistic parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screening* 4, 67–73

6 Zhang, J-H. *et al.* (2000) Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. *J. Comb. Chem.* 2, 258–265

7 Bingham, R. and Hutchinson, J. (2001) Characterization and Identification of Compound Interference Patterns. Presentation at the 7th Annual Conference and Exhibition of The Society for Biomolecular Screening (SBS), 10–13 September 2001, Baltimore, MD, USA

8 Goebel, M. and Gruenwald, L. (1999) A survey of data mining and knowledge discovery software tools. *Newsletter of the ACM special interest group on knowledge discovery and data mining, published twice a year, SIGKDD Explorations* 1, 20–32

9 Hann, M. and Green, R. (1999) Chemoinformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.* 3, 379–383

10 Olsson, T. and Oprea, T.I. (2001) Cheminformatics: a tool for decision-makers in drug discovery. *Curr. Opin. Drug Disc. Dev.* 4, 308–313

11 Manly, C.J. *et al.* (2001) The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today* 6, 1101–1110

12 Johnson, D.E. *et al.* (2001) Chem-tox informatics: data mining using a medicinal chemistry building block approach. *Curr. Opin. Drug Disc. Dev.* 4, 92–101

13 Engels, M. and Venkatarangan, P. (2001) Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Disc. Dev.* 4, 275–283

14 Young, S. St. *et al.* (2002) Initial compound selection for sequential screening. *Curr. Opin. Drug Disc. Dev.* 5, 422–427

15 Ladd, B., Kenner, S. (2000) Information visualization and analytical data mining in pharmaceutical R&D. *Curr. Opin. Drug Disc. Dev.* 3, 280–291

16 Walters, W.P. *et al.* (1998) Virtual Screening – an overview. *Drug Discov. Today* 3, 160–178

17 Kreusel, D. (2001) From raw data in the laboratory to information availability in the enterprise. *Drug Discovery World Winter 2001/2002*, 69–74

18 Schneider, G. and Böhm, H-J. (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today* 7, 64–70

19 Small, R.D. and Edelstein, H.A. (2001) Data mining in the pharmaceutical industry. *Drug Discovery World Fall 2001*, 39–48

20 Valler, M.J. and Green, D. (2000) Diversity screening versus focussed screening in drug discovery. *Drug Discov. Today* 5, 286–293

21 Entzeroth, M. *et al.* (2000) High throughput drug profiling. *J. Automated Methods & Management in Chemistry* 22, 171–173

22 Gund, P. and Sigal, N. (1999) Applying informatics systems to high-throughput screening and analysis. *Pharmainformatics: A Trends Guide (Trends supplement)*, 25–29