ELSEVIER

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 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