Screening informatics: adding value with meta-data structures and visualization tools

Bryn R. Roberts

Dramatic increases in the quantity and complexity of HTS data require effective informatics to convert this data into knowledge, and support decision-making within drug discovery. Enabling full exploitation of HTS data requires meta-data structures, which describe underlying data, providing context and the ability to associate screens according to common attributes. Visualization tools can be used to great effect during HTS for quality control, database navigation and data-mining.

Bryn R. Roberts
AstraZeneca, Enabling Science and Technology – Lead Informatics
Alderley Park, Macclesfield
UK SK10 4TG
tel: +44 1625 514537
fax: +44 1625 513441
e-mail: bryn.roberts@astrazeneca.com

▼ The objective of screening informatics is to enable full access to, and exploitation of, all screening data, thereby helping to improve decision-making. This will have an impact on many areas, including target selection, compound library design, assay design and lead compound identification.

The challenges of a changing drug discovery environment

The advent of genomics, combinatorial chemistry and HTS has revolutionized the drug discovery process. The impact of these new technologies requires a stepwise shift in the way data and, ultimately, the knowledge that they generate are managed[1]. The volumes and complexity of the data produced mean that traditional methods and tools for accessing, analysing and distributing the data and knowledge are no longer viable[2].

With ultra-HTS (uHTS), data volumes of 100,000 points-per-day-per-screen are becoming common[3–5]. A typical HTS operation in a large pharmaceutical company will be generating approximately 50 million data points-per-year, covering approximately 50 targets[6]. Compared with the situation ten years ago, when 200,000 data points-per-year was the usual rate, this represents an increase of two orders of magnitude in the data volume.

In addition to the increasing volume and rate of data generation, the diversity and complexity of HTS data is increasing with the introduction of high-content or high-information screening systems such as the ArrayScan™ by Cellomics[7] (Pittsburgh, PA, USA), Acumen by The Technology Partnership [Royston, Hertfordshire, UK; for reference, see Blenkinsop, P. High information screening (HIS). The way ahead… *5th Annual Conference of the Society for Biomolecular Screening*, Edinburgh, UK, 13–16 September 1999 abstract book p. 77], and technologies such as fluorescence correlation spectroscopy (FCS)[8]. A traditional HTS assay generates a single data point-per-compound, whereas the Acumen instrument, for example, will output a detailed image file as well as more than ten derived parameters for each compound and concentration.

From data to knowledge

Storage and retrieval of these large volumes of complex data are not the primary issues with current storage media and databases. Exploitation of the data for knowledge generation and decision-support is a much greater challenge to the drug discovery business, particularly if the derived knowledge is to be exploited fully by a global drug discovery enterprise. For example, during the development of a screen, many decisions are made, regarding such properties as the source of the target protein(s) and the assay technology. Each screen brings knowledge and learning that could be reused across the business. This knowledge is costly to rediscover, and potentially even more costly if repeatedly overlooked.

The inclusion of all data and associated knowledge in central corporate repositories and the

1359-6446/00/$ – see front matter PII: S1359-6446(00)01480-X

integration of these stores are the foundations for the development of informatics applications for decision-support.

What is involved in screening informatics?

For screening informatics to be effective, several factors must be considered:

- The underlying data model (such as which data are stored, their relationships, partitioning of pre-production and validated data, the database type)
- Meta-data structures (see later)
- Data entry tools (must be fast, easy to use and should control the vocabulary to maintain data integrity)
- Query tools (must be fast and intuitive, even for non-experts)
- Data navigation and presentation (such as summary Web views of data for projects, meta-data navigation tools)
- Data analysis and visualization tools (must address the questions asked).

The remainder of this article will focus on two of these aspects in more detail: meta-data and visualization.

Meta-data structures

In the context of screening informatics, meta-data are data that describe other data[9]. They can be used as a standard method of describing, for example, relationships, organization and sources of data, thus making it possible to search, cluster, navigate and add context to the data. For example, the assay technology and target classification descriptors can be considered meta-data, describing certain aspects of a screen and therefore the data generated by that screen.

What can be gained from meta-data?

Meta-data can be used for data navigation (querying), where users want to ask questions such as 'show me all screens for kinase targets' or 'show me screens where a cytotoxic compound could produce a false hit'. Meta-data are also important for providing post-navigation context to data. For example, if a compound of interest was found to be active in several screens, the scientist might want to navigate more detailed information about the target classification, assay buffer constituents and incubation conditions for those screens. In addition, meta-data can be used by specialized applications to provide workflow representations for project teams or laboratory groups.

Why is meta-data important for screening informatics?

Without well-structured and accurate meta-data, even globally accessible data is difficult to navigate and even more difficult to interpret in a meaningful way. Meta-data enables clustering and joining of related data types to enable meaningful mining, analysis and visualization. It removes the necessity for expert knowledge of all the projects, targets, assays, data and their organization within the corporate database. However, the expert user also benefits greatly from meta-data entries, such as data reliability indicators (e.g. estimates of false-positive and false-negative rates in a screen)[10].

Examples of biological meta-data

As represented in Fig. 1, biological screen meta-data relate to specific screens, screen versions or screen occasions. They can be structured in several ways, two common examples being:

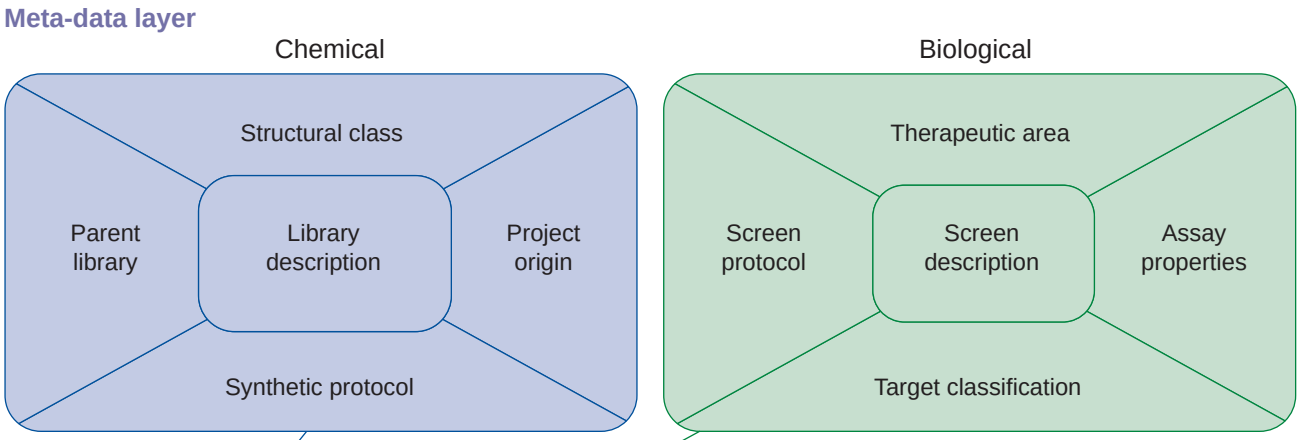
- Hierarchical – for example, target classification, where a receptor tyrosine kinase belongs to the class of tyrosine kinases, which belongs to the class of kinases, which belongs to the class of enzymes. One can query at any level in a hierarchy, depending on the level of detail known or required.
- Non-hierarchical – for example, assay properties such as assay volume or substrate concentration. These are useful for when there are no direct relationships represented between meta-data attributes.

Implementing the meta-data layer

Which meta-data are stored and how they are structured must be based on the questions and decisions that are to be supported. For the meta-data to be of maximum benefit, the data integrity must be 100%. Compliance by the scientific community can be encouraged if data entry is made easy and they perceive value for themselves in entering the data. However, a controlled vocabulary, which defines how meta-data values are named, is required if data integrity is to be high. For example, to a scientist, beta-adrenoceptor, β-adrenoceptor and beta-adrenergic receptor might mean the same thing – to a typical database they are quite different!

Typically, meta-data are stored in either a database (relational or object[11]) or in context-rich documents, such as word processor files. Databases have the advantage of data structure, relationships, integration, ability to query, as well as application development capabilities for the Web, using tools such as Oracle Application Server (Oracle Corp., Redwood Shores, CA, USA) and NetCharts™ (Visual Mining, Rockville, MD, USA). Database meta-data structures are highly appropriate for creating, for example, the target classification hierarchy (e.g. http://www.expasy.ch/enzyme for enzymes[12]).

By contrast, documents are familiar, flexible and can provide some data structure and relationships if well designed. They are more appropriate for meta-data such as screen protocol descriptions, which detail how a screen was performed and aspects associated with its operation (e.g. incubation conditions, safety issues). Data held in word processor documents is difficult to exploit (via searching or visualization), particularly



compound_ID	screen	result_type	result	result_date
compound1	ENZ1	%inhibition	54	01-01-2000
compound2	ENZ2	IC50	1.7E-8	01-01-2000
compound3	REC4	Kd	2.3E-9	01-01-2000
compound4	IMG1	image_jpg		01-01-2000

Drug Discovery Today: HTS supplement

Figure 1. A conceptual representation of example meta-data relating to typical drug discovery HTS data. Note that the distinction between data and meta-data layers can be 'fuzzy'; 'result_type' in this example could equally be considered as meta-data.

if no structure or controlled vocabulary have been applied. However, Internet technologies, such as Muscat's Linguistic Inference technology (The Dialog Corporation plc, London, UK), can be used to index, search and alert using document-based meta-data. One alternative to poorly structured word processor documents is the eXtensible Markup Language (XML).

eXtensible Markup Language

XML is generating much interest within the informatics community at present. The underlying technology (SGML; Standard Generalized Markup Language) is not particularly new, but full realization of what XML can offer is only just emerging[13]. XML is the emerging standard for data interchange and structured data document storage[14]. It enables the storage and transfer of data with structure (or meaning) rather than just format, and data held within XML documents can be readily exploited by an analysis application, database or report generator. In addition, XML files provide long-term accessibility options, because they are not dependent on specific file formats (see Box 1).

The advantage of moving to XML from unstructured documents for the storage of screen-related information such as the screen protocols or operating procedures, is obvious. Document construction will be relatively easy, especially as modern word processors will have native support for XML.

Well structured data and meta-data provide the foundation for effective data exploitation, a good example being data visualization.

Data visualization

Visualization tools, in their various forms, have many valuable applications within the HTS environment[2,10,15–17]. Some of these applications will now be discussed.

Quality control and data validation

Visualization provides a fast and robust method for checking the quality of HTS data, providing scientists with selected views of the experimental output[16]. Plate views, where wells are coloured or sized according to the data, enable edge-effects (where wells at the edge of an assay plate behave

differently from the centre wells because of variations in, for example, humidity or temperature gradients), blocked dispensing tips, or other artefacts (e.g. loss of enzyme activity during a screen) to be identified at very high throughputs. Aggregate value (e.g. average, maximum), frequency distribution and scatter plots enable meaningful comparisons to be made between screens, batches and plates, therefore enabling the rapid identification of problems[16].

Box 1. Example description of a receptor tyrosine kinase screen using XML

```
<?xml version="1.0"?>
        <SCREENDESCRIPTION>
                <SCREENAME>RTK1</SCREENAME>
                <TARGETCLASS>Kinase</TARGETCLASS>
                <ASSAYCLASS>TRF</ASSAYCLASS>
                <TOTALCMPS>200000</TOTALCMPS>
                <COMPLETEDATE>01-01-2000</COMPLETEDATE>
        </SCREENDESCRIPTION>
```

This is an example description of how a receptor tyrosine kinase (RTK) screen, using time-resolved fluorescence (TRF), which screened 200,000 compounds (CMPS), could be represented in XML. The tags (e.g. <SCREENAME>) provide structure (or meaning) to the contained text (e.g. RTK1), such that an application would recognize that RTK1 was the screen name and should be treated in a certain way.

Data-mining

One effective way to explore large data sets for expected or unexpected trends or outliers is to use visualization techniques[17]. Available visualization applications include Spotfire™ Pro (Spotfire, Boston, MA, USA), or custom applications can be developed using, for example, IDL™ from Research Systems (Boulder, CO, USA) or Toolmaster from Advanced Visual Systems (Waltham, MA, USA). To use an example to demonstrate data visualization, meta-data were used to provide a list of screens where cytotoxic compounds produced false positive responses. The visualization shown in Fig. 2 exemplifies how simple visualization can aid the identification of potential toxic compounds.

The degree of variation or 'noise' in each screen is immediately visible using visualizations, and can be considered in further manipulation or analysis. Compounds that have a high level of activity in all four screens can be clearly identified as those in the upper right corner of the plot with larger and redder data markers (one such point is indicated in Fig. 2 by the cursor arrow). Alternative views using multiple linked graphs can make exploring multidimensional data sets even more effective. For example, the 'brushing-and-linking' technique enables selected data records on one graph to be selected on all other graphs in the view. In addition, 'dynamic querying' enables the user to set various criteria (such as inhibition of cell growth IC_{50} <10 μM) to define which data are displayed on the graphs[2].

Database navigation and dynamic workflow representation

Graphical views on data within databases can be easily created with tools such as those from Visual Mining (e.g. NetCharts™ and Decision.Control™, their decision/information portal), and made available on the corporate intranet. This enables users to

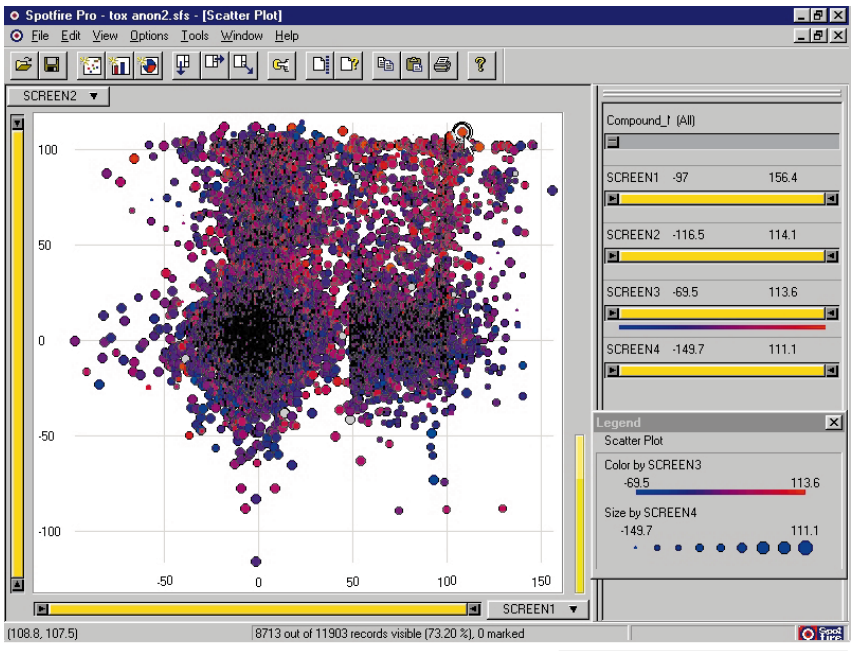


Figure 2. Visualization from Spotfire™ Pro (Spotfire, Boston, MA, USA) showing data from four high-throughput screens, identified from meta-data as being mammalian cell-based screens, susceptible to false-positives from cytotoxic compounds. The percentage effect in the four screens (1–4) are represented on one graph by: 1 – x-axis position; 2 – y-axis position; 3 – size of data marker; 4 – colour of data marker (blue to red = low to high). The record marked by the cursor arrow indicates a potential cytotoxic compound.

view summary or quality assurance data from screens, or to recreate concentration–response graphs from data in the database (negating the need to store static images). Composites of these visualization types can be constructed to represent a screening cascade for a project or to denote workflow in an HTS laboratory, by visually representing throughputs, bottlenecks and hit-rates in the portfolio of screens. These graphics on the intranet can also enable navigation of the underlying data and meta-data, for detailed evaluation or further analysis.

A highly structured data hierarchy is necessary to enable logical and meaningful graphical representation of data, with the ability to navigate through the different layers of detail. Storing reduced data (such as IC_{50} values) alone is highly limiting for concentration–response screening. Ideally, sufficient data are stored to enable the recreation of the curves in a browser and to enable future reanalysis if necessary. For example, concentration–response data could be stored at the following levels:

- Concentration–response raw data with the controls
- Normalized and calculated data (e.g. percentage response)
- Derived or reduced data (e.g. IC_{50}, slope)
- Summary data (e.g. active, toxic, in the context of that screen occasion)
- Validation data (e.g. valid record, suspect record).

This would therefore enable the data to be viewed at many levels of detail. Meta-data might also be required to add further context to the data (e.g. for an enzyme screen, to describe the ratio of substrate concentration to its Km).

Conclusions

Screening informatics is a crucial resource in the modern HTS environment. Appropriate implementation and exploitation of meta-data, coupled with the use of various visualization approaches, will add considerable value to any drug discovery enterprise. Currently emerging technologies[18], such as Internet-platform databases (e.g. Oracle8i), new component-based development environments (e.g. Java Beans[19]), middleware layers enabling integration of disparate data sources and legacy systems (e.g. CORBA; Common Object Request Broker Architecture), new standards for data interchange and storage (e.g. XML), and the introduction of alerting tools and 'intelligent' agents[20], will have a significant impact on screening informatics over the next few years. This will provide an integrated information environment, where existing questions can be asked more easily and new questions will become possible.

Acknowledgements

Many of my colleagues have contributed to the thinking contained within this article, particularly: Stuart Bell, Rich Lawson, Ian Lloyd, John Major, Robert McNutt, Steve Peters and Jason Swift. My thanks also to Rich Lawson for his comments on this manuscript.

References

1 Gund, P. and Sigal, N.H. (1999) Applying informatics systems to high-throughput screening and analysis. *Trends Biotechnol.* (Suppl. S), 25–29

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

3 Oldenburg, K.R. (1999) Automation basics: robotics vs workstations. *J. Biomol. Screening* 4, 53–56

4 Wildey, M.J. *et al.* (1999) Allegro™: moving the bar upwards. *J. Biomol. Screening* 4, 57–60

5 Mere, L. *et al.* (1999) Miniaturized FRET assays and microfluidics: key components for ultra-high-throughput screening. *Drug Discovery Today* 4, 363–369

6 Drews, J. (1999) Informatics: coming to grips with complexity. *Pharmainformatics: A Trends Guide* 1–2

7 Conway, B. *et al.* (1999) Quantification of G-protein coupled receptor internalization using G-protein coupled receptor–green fluorescent protein conjugates with the ArrayScan™ high-content screening system. *J. Biomol. Screening* 4, 75–86

8 Moore, K. *et al.* (1999) Single-molecule detection technologies in miniaturized high-throughput screening: fluorescence correlation spectroscopy. *J. Biomol. Screening* 4, 335–353

9 Greene, J. (1997) *Oracle8 Server*, pp. 768–770, Sams Publishing

10 Spencer, R.W. (1998) High-throughput screening of historic collections: observations on file size, biological targets, and file diversity. *Biotechnol. Bioeng.* 61, 61–67

11 McCullough, C. (1997) *Oracle8 for Dummies*, pp. 89–103, IDG Books Worldwide

12 Appel, R.D. *et al.* (1994) A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server. *Trends Biochem. Sci.* 19, 258–260

13 Golfarb, C. (1998) *The XML Handbook*, Prentice-Hall

14 Johnston, H. (1999) Fresh approach helps interfacing. *Scientific Computing World* 49, 21

15 Somogyi, R. (1999) Making sense of gene-expression data. *Pharmainformatics: A Trends Guide* 17–24

16 Williams, G. (1999) ActivityBase 4. *Discovery* 1, 11–14

17 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349

18 Brocklehurst, S. *et al.* (1999) Creating integrated computer systems for target discovery and drug discovery. *Pharmainformatics: A Trends Guide* 12–15

19 Boyle, J. (1998) A visual environment for the manipulation and integration of JAVA Beans. *Bioinformatics* 14, 739–748

20 Ding, Z. and Liu, L. (1999) Machine learning and agent-based computing. *Dr Jobb's Journal* November, 88–96