'Tools of the trade' for target identification, validation and biomarker discovery

As automation in the drug discovery arena increases then so does the amount and complexity of the data. We discuss an emerging solution that now integrates various technologies to provide a scientific data intelligence tool giving scientists increasing scope to identify, evaluate and assess potential targets with reduced 'lab-based' experimentation.

By Dr Paul Denny-Gouldson

The pharmaceutical industry is showing growing interest in the automation and simplification of many processes in drug discovery and development. One such area that has felt the effects of automation is the 'omics area of biology, such as genomics, proteomics and metabolomics. Here, the automation and miniaturisation of the process has brought significant improvements in the amount and type of 'screens' that can be achieved.

For example, a few years ago the typical chip array size for a genomics experiment was measured in thousands, but this number now reaches tens of thousands or even hundreds of thousands. The rapid changes in array technologies bear an analogy to advances in plate-based screening technologies, which went through a similar throughput and density increase in a relatively small space of time during the 90s.

The miniaturisation and throughput increase of chip-based assays delivers a problem – the amount and complexity of the data produced also increases. Typically this data is multidimensional, meaning that there are many different variables to the data that need to be correlated in order to understand the significance of a result.

So, how can we conquer this mountain of data to make sense of it?

Answer: one step at a time.

The advantages of automation

Even though the 'omics approaches have many hurdles that must be overcome both technically and ethically, we are beginning to see the value in data generated by automation. From the information obtained by these types of experiments, scientists are now developing the ability to understand different disease states at many levels previously immeasurable and apply that knowledge in drug development.

So in genomics, for example, experiments are now able to reveal the genes that are upregulated (their overall presence increased), those that are down-regulated (their overall presence decreased), and those that do not show any change. From this, scientists are able to take a closer look at each of

the 'hits' genes. By linking the gene to pathways using other experimental techniques, either in the lab or with bioinformatics or both, they can decipher how disease states show at the genetic level. This often manifests as the typical "New gene found that is related to disease X" statements that we see in the press.

The subsequent steps in this 'systematic' analysis of disease states, and in fact normal states, involves looking at the effects of disease on the proteome (proteomics). In the same way, researchers have more recently used metabolomics and transcriptomics to investigate how the metabolite and mRNA transcription changes in disease states.

'Systems biology' is the covering discipline that has emerged with the very complicated task of linking all the 'omics data together with all other types of biological data to get a 'systematic map of a disease state', encompassing how all the elements of a living organism interact at the molecular level.

When all these factors, data and observations are brought together scientists are able to understand the nature of disease and look for 'tell-tale signs' – commonly known in the industry as a biomarker. These biomarkers can be from any part of biology, not just genomics. The table below shows a number of example biomarkers:

DISEASE STATE	BIOMARKER	TYPE	REFERENCE
Allergy – asthma	ADAM33	Gene	Holgate et al PATS 2006
Allergy – asthma	NO (nitrous oxide)	Chemical	
Allergy – asthma	Difficulty breathing	Observation	Personal experience
Allergy	Ig(x)	Antibody (protein)	
Allergy	Histamine	Neurotransmitter	
Heart attack	Troponin (I or T)	Protein	Wikipedia

So we can see that a biomarker is not just something that is very complex and scientific – it can be as simple as an observation. The complexity comes when trying to find new biomarkers for diseases or when proving that an observation is related to a disease state. This is where the combinations of the 'omics technologies, good experimental design and systems biology really start to show promise.

Once we have all the data how can we use it to help drug discovery?

It is now common to see drug companies investing in both the 'omics technology areas and in systems biology. Streamlining the target identification, target validation and biomarker discovery areas of science continues to be of interest to the pharmaceutical industry because each area can have an impact on the discovery process.

While all of these research areas can benefit from the 'omics data combined with other sources of data, each raises different questions.

The scientist might ask:

Firstly, find me all the potential targets of a given disease. ***New target identification for a new drug.***

So how would 'omics help here? With systematic analysis of disease states across these technologies, a map of potential targets emerges. For example, if a disease state has an upregulation of a gene that in turn leads to a higher level of a certain receptor in the body, and this in turn leads to the disease manifestation in the body, one would hypothesise that a drug that acted at the receptor that is up-regulated would be beneficial.

This is a very simplistic view of how this data and knowledge can help find new targets of disease but it shows the process and how the data from these areas can be used. The 'omic data can provide the basis for a scientist to develop a hypothesis that can be tested further. This type of analysis also requires the collation of data from many systems since there is no single database that contains all the data about all diseases for all areas of 'omics.

Subsequently, the scientist asks:

Where might this already-developed drug act? ***New target for a known drug.***

This is less about 'omics and more about complex searching of existing data where systems biology begins to play a role. At first glance this seems simple but when you consider the complexity and distributed nature of the data, databases, documents, websites, etc, the problem becomes instantly more complex.

Then, the scientist asks:

Where does a biomarker exist that we can measure when developing a new assay? ***A biomarker for target disease.***

Here, the scientist needs to design a new experiment that will allow them to demonstrate that the drug candidate does show effect on the disease state. To do this they must have something to measure that is associated with a disease, ie, a biomarker.

On the other hand, the scientist typically asks a more complex question:
Show me all the data and information that we know about target X and how it is related to a given disease. Then show me related targets based on:

- Gene similarity and regulation in disease state
- Protein similarity and regulation in disease states
- Binding site similarity
- Assays available in house
- Patent information
- Compounds already tested and HTS profiles
- Similar compounds already in house
- Research information from lab books
- Safety profiles of compounds already tested
- …etc

Or the scientist might ask the 'simple' question:
Find me a possible biomarker for disease X, that has a differential of >24% between disease and non-disease states – and make it soluble.

Only when all the data has been collated can the scientist proceed and give an analysis as to the possible validity of a given target for a new intervention – without testing a single thing in the lab. This is a potential area for dramatic savings because the scientist can evaluate a target without doing any experimental work and without incurring expensive laboratory costs.

In order to make these types of analyses and hypotheses, the scientist typically needs to search many data types, as shown in **Figure 1**.

However, the search for information is then compounded by the fact that each data type or concept can come from many different sources, as shown in **Figure 2**.

With current systems, getting the data and information together is a very protracted and laborious task requiring many months of searching and report collation. Scientists prefer to do the work in the lab as it is easier than being a 'search engine jockey'. In addition, most scientists may tend to avoid the data collation because they feel they need expertise in many different querying techniques.

Tools for the job

There are a number of technical issues that need to be resolved before this process can be streamlined. The most apparent is the ability to search multiple data sources simultaneously – databases, flat file repositories, websites and web services – and deconvolute the results into a meaningful format that can then be analysed and comprehended by the scientist.

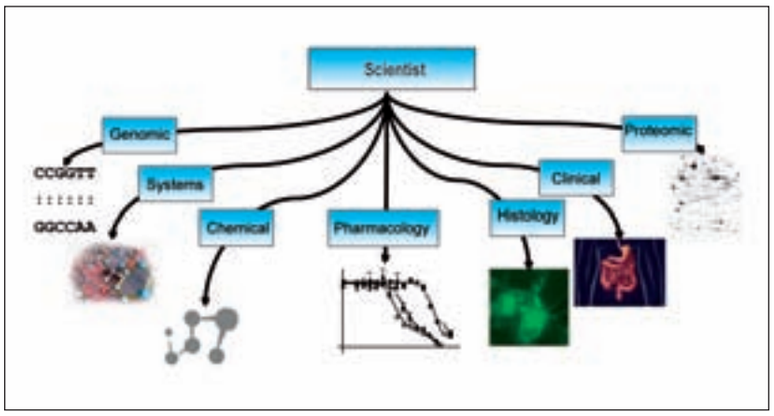
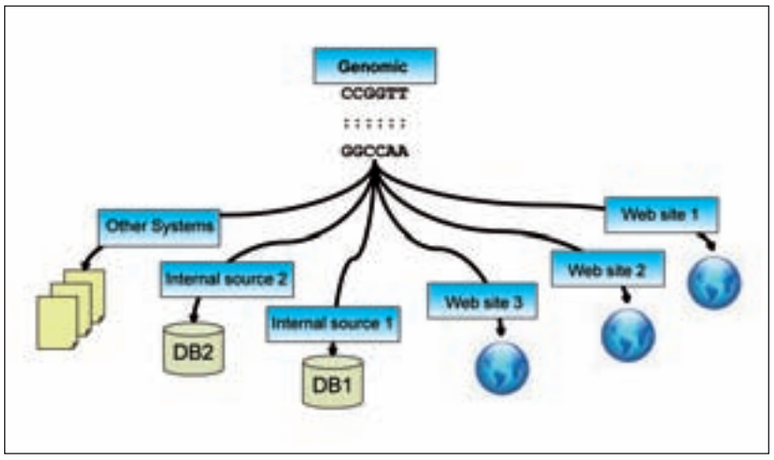


Figure 1
Evaluating a target typically involves searching many different data types

This is not a new problem – many companies have approached this problem by building data warehouses. This is where all the known data about an object is put into one large database so that at any point a user can search for data on any given object. The benefits of 'warehousing' particularly include the speed at which a search can be performed, as the warehouse design is usually optimised for query performance, often being 'star schema' in their nature. Such schema require data to be uploaded into the warehouse according to a given set of rules, involving the use of dictionaries that validate the data and its source. All these things are done to ensure that when data from multiple objects is compared, it is in fact 'correct' or scientifically valid for useful comparison to occur. These concepts are not new and they form the fundamental basis for good data management.

However, there are drawbacks to this approach. The data model that accepts the data is obviously fixed and so if a new data type is encountered, or a question is asked of the warehouse that cannot

Figure 2
Data types involved in a search may originate from disparate data sources



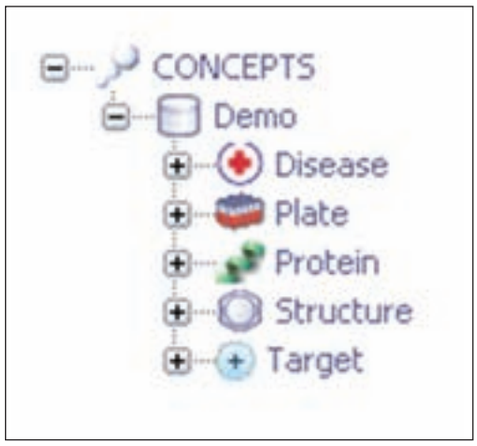


Figure 3
A concept model displaying data structured intuitively

be answered due to a 'missing link' in the data model, then the model has to be changed, resulting in all the tools that load the data into the warehouse not working.

So, warehouses are used when the data structure is very well defined and the types of questions that will be asked are well defined – because this allows relatively easy design of the data model.

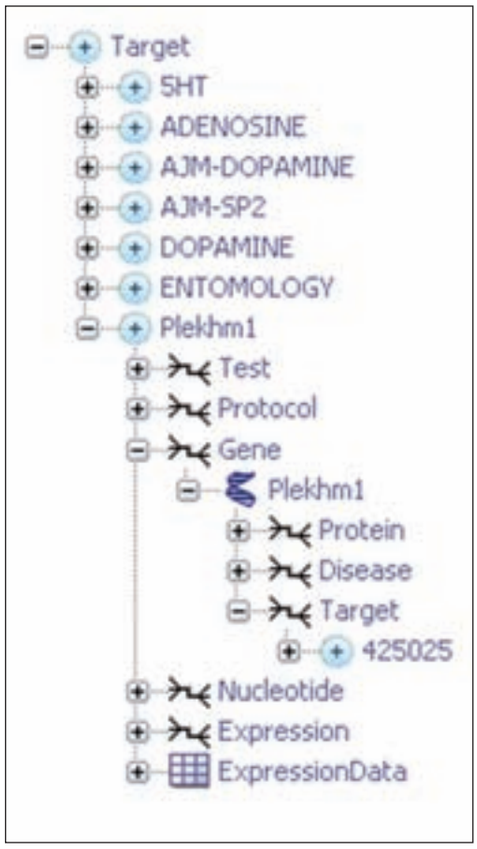


Figure 4
Linked concepts give easy access to different scientific areas

Unfortunately in some areas of research a rigid data model and full list of types of queries is not feasible. The problem often comes when many data sources are required to be searched at once and the nature of the query requires immediate responses. One further drawback of warehouses is that there is often a time lag between the data being produced compared to its availability in the database. When there are large quantities of data, this becomes even more problematic.

But this is not the only issue. The scientist must be able to navigate and search data without having to understand the underlying data structures and so requires a conceptualised data landscape that allows searching and presents data in a meaningful way. Scientists deal in compounds, assays, diseases, genes, proteins and targets and when searching for information, they want to use these familiar terms and have them linked in a rational manner. Many scientists work in a specific area of research and have a 'language' that they use on a day-to-day basis – an ontology. If this recognisable language is 'stitched' into their data navigation and search tools, scientists benefit greatly from this inherent 'intelligence' based on specific scientific disciplines. There are many ontologies already available as open resources, for example: Gene Ontology (GO) and Protein Ontology (PO) (see www.geneontology.org/ and http://proteinontology.info/).

For example, **Figure 3** shows how a typical concept model for a scientist might look.

If the concepts are related, a scientist can use familiar terminology to 'hop' across to different areas of interest, as shown in **Figure 4**.

A final important issue is making this data readily available to the rest of the organisation in a searchable manner, thereby 'sharing the captured knowledge'. Scientists want to share and analyse each other's data, so the manner in which the data is collated and stored is crucial as it must facilitate both contextual and factual searching – something that requires very specific and adept data management.

Search federation technologies provide the flexibility that a typical warehouse cannot. Federation involves a search being 'split' across the different data sources, and results concatenated when returned. Dividing the search across data sources removes the need to gather all the data into one database or warehouse. Federative technologies also allow new data sources to be added more quickly, because there is no need to change a database schema. However, the responsiveness of a search is lower than a warehouse approach since the searching can include any data type, for example,

documents, websites, databases, etc, all of which can increase the duration of a search.

The optimum solution would integrate both methods – the warehouse approach where the data structure is well defined, the data formats are fixed and the data is 'summary data' – and the federative approach, where the data is unstructured, dynamic and operational. Combining the two allows the scientist to explore the data landscape in a way that is not really possible at this time without huge time and effort cost.

Scientific intelligence tools

However, an emerging solution now integrates various technologies to provide a scientific intelligence tool – something akin to a business intelligence tool for scientists – that meets the complex requirements for target identification, target validation and biomarker discovery.

Below are real-world examples of how by using a combination of an electronic laboratory notebook (ELN) framework, data federation technology, industry knowledge and collaboration, this solution can solve the difficulties in accessing and extracting relevant data.

Example 1

Generating reports on potential new targets for drug intervention is excessively time consuming for pharmaceutical companies. Scientists at one major European pharmaceutical company were spending in excess of 15 days to complete such a task. Researchers and project teams were faced with the tedious and time-consuming task of manually collating data from up to 15 different sources, including data from structured databases, websites, and spreadsheets and documents in scattered network locations.

Using data federation technology, researchers defined a query based on a 'hit' from a genomics experiment (hit classified as either upregulated or down-regulated in disease state by a certain %) to search data sources ranging from Microsoft® Excel templates, PDFs, websites and web databases, to internal and external screening and chemical databases.

Concept mediation enables the scientists to 'search by navigating', whereby a simplified conceptual view of the available data types is presented (as shown above in **Figures 3** and **4**), masking the complexity of the data sources and data structures. The concepts are presented to the user in terms of a language and structure that fit their domain, for example, a medicinal chemist might think in terms of structures, plates, reactions, assays and protocols while a biologist could think in terms of genes, proteins,

diseases, subjects, sequences, etc. Scientists query data sources using familiar terminology as the search language, simplifying and so accelerating the search for information.

Researchers create either 'standard' report types that are pre-defended at an organisation level, by a simple drag and drop operation of the retrieved data into an ELN framework. Alternatively, the scientist can drag any result node into the ELN to create an ad hoc report, containing more data than the 'standard report'.

Capturing both structured and unstructured data in a compliant environment, the ELN framework stores knowledge in a manner that is searchable for both context, ie, conclusions, methods and factual data, ie, relative expression level <3.5.

Example 2

Combining unstructured data from separate programs in order to generate a lead profile report can present challenges to organisations that do not have a consolidated infrastructure and standardised data format.

A major biotechnology company wanted to integrate data from its screening program with its toxicology and ADMET programs, but this proved to be problematic due to data stored on different networks and in varying formats. Because this was an extended process, the data being sought often became out of date very quickly.

The scientists' aim was to easily generate a profile report on any lead from four separate repositories. This required deconvolution and reporting on *in vitro* and *in vivo* data while making sure that the report contained the most recent data prior to presenting at project meetings where lead selection would occur.

Again, using concept mediation, scientists are able to enter only two or three items of key information into a very simple search interface that is pre-defended at an organisation level. The results from all four sources that meet the search criteria are presented to the scientists, viewed and stored as a simple HTML table.

Summary

The combination of an ELN framework, federation technology and concept mediation provides a portal that allows scientists to search and mine data from any number of data sources using the language and concepts that are applicable to their specific discipline.

Data federation technology provides the flexibility to query data sources of different formats simultaneously, searching data in the native systems and consolidating inherently unstructured query results into structured data.

The concept mediation provides a simplified and intuitive search language that negates the need to understand all the different search interfaces and structures for the various data sources.

The ELN framework allows search results of interest to be easily imported and compiled into structured reports, eliminating manual data transcription. In addition, the ELN's compliant environment protects IP and allows scientists from all disciplines to search for and share information and knowledge.

Such domain-intelligent solutions give scientists increasing scope to identify, evaluate and assess potential targets with reduced 'lab-based' experimentation. By making it much easier to search more data sources, researchers can readily retrieve and analyse both contextual and factual data that can help in the formulation of a hypothesis – while bringing significant savings in both time and money due to the increased ease and speed of doing so. DDW

Dr Paul Denny-Gouldson *joined IDBS in May 2005 when his company Deffinity Solutions Ltd was purchased by IDBS. Paul set up Deffinity Solutions Ltd, a company specialising in Electronic Laboratory Note books (ELNs), in 2000. Paul's broad background and experience in pharmaceutical research and development meant that the system was designed to meet the scientist's need as well as corporate needs. Prior to this Paul spent four years at Sanofi-Aventis, where he worked first as a molecular and cell biologist, then a molecular pharmacologist and finally as a group leader and working group head for G-Protein coupled receptor (GPCR) research across the six research centres in France. Paul received his PhD from the University of Essex, Colchester, UK in computational chemistry and is a well established researcher with 14 peer reviewed articles and papers in leading journals across multiple disciplines.*