ELSEVIER

A retrospective on how the application of data integration and visualisation has been used to connect disparate information sources into a drug-discovery focused, decision making environment.

Visualizing the drug target landscape☆

Stephen J. Campbell[1], Anna Gaulton[1], Jason Marshall[1], Dmitri Bichko[2], Sid Martin[1,3], Cory Brouwer[1] and Lee Harland[1,4]

[1] Computational Sciences Centre of Emphasis, Pfizer Global Research & Development, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK
[2] Computational Sciences Centre of Emphasis, Pfizer Research Technology Centre, 620 Memorial Drive, Cambridge, MA 02139, USA
[3] Research Informatics, Pfizer Global Research & Development, Ramsgate Road, Sandwich, Kent CT13 9NJ, UK

Generating new therapeutic hypotheses for human disease requires the analysis and interpretation of many different experimental datasets. Assembling a holistic picture of the current landscape of drug discovery activity remains a challenge, however, because of the lack of integration between biological, chemical and clinical resources. Although tools designed to tackle the interpretation of individual data types are abundant, systems that bring together multiple elements to directly enable decision making within drug discovery programmes are rare. In this article, we review the path that led to the development of a knowledge system to tackle this problem within our organization and highlight the influences of existing technologies on its development. Central to our approach is the use of visualization to better convey the overall meaning of an integrated set of data including disease association, druggability, competitor intelligence, genomics and text mining. Organizing such data along lines of therapeutic precedence creates clearly distinct 'zones' of pharmaceutical opportunity, ranging from small-molecule repurposing to biotherapeutic prospects and gene family exploitation. Mapping content in this way also provides a visual alerting mechanism that evaluates new evidence in the context of old, reducing information overload by filtering redundant information. In addition, we argue the need for more tools in this space and highlight the role that data standards, new technologies and increased collaboration might have in achieving this aim.

Stephen Campbell is a principal scientist in the Computational Sciences group at Pfizer where he has specialised in data mining and visualization techniques related to drug target and biomarker identification. With a background in Biochemistry and Bioinformatics, he holds a Ph.D. in molecular modelling carried out at University College London and Leeds University (UK).

Anna Gaulton holds a B.Sc. in biochemistry from UMIST, Manchester and a Ph.D. in bioinformatics from the University of Manchester, for work on analysis of G protein-coupled receptors with Professor Terri Attwood. Following this, she moved to Pfizer, where she worked in the Computational Biology group, building drug target databases and developing tools for target analysis and assessment of druggability. In 2009 she moved to the EMBL-European Bioinformatics Institute to join the newly established ChEMBL group, led by John Overington, where she is working to develop public domain chemogenomics resources.

Lee Harland leads computational biology research at Pfizer Regenerative Medicine, Cambridge, UK, supporting novel stem cell–based therapies. He holds a B.Sc. in biochemistry from the University of Manchester, UK, and a Ph.D. from Kings College London, UK, for research in gene therapy. His primary interests concern how informatics can support drug discovery through the assembly and interpretation of integrated data.

☆ This article is a reprint of a previously published article. For citation purposes, please use the original publication details; Drug Discovery Today 15/1–2(2010), pp. 3–15.
DOI of original article: 10.1016/j.drudis.2009.09.011
Corresponding author:. Harland, L. (lee.harland@pfizer.com)
[4] Current address: Pfizer Regenerative Medicine, UCB Building, Granta Park, Cambridge CB21 6GS, UK.

GLOSSARY

Biological rationale the scientific thought process and collective evidence behind treating a gene product as a potential therapeutic target on the basis of its connectivity to a disease or phenotype.
Druggable whether a protein might be a potential target because it exhibits properties indicating that it might be amenable to modulation by a small molecule or biological therapeutic.
Mash-up a single web application that integrates information or functionality from more than one source.
Natural language processing (NLP) the development and application of text-mining software for the 'automated reading' of documents. The aim of NLP is to recognize human language constructs so that key facts can be extracted and represented in more formal ways.
Ontology a representation of concepts within an information domain and the relationship between those concepts. Often, ontologies are used to standardize descriptions of particular areas of science with the aim of knowledge sharing and reuse.
Really Simple Syndication (RSS) formatted web feeds that are used to rapidly publish information from frequently changing information resources. Typically, users select feeds that are of interest to them and read those feeds through desktop or mobile software.
Semantic web a technical approach that uses the Resource Description Framework (RDF) to describe, integrate and share data both on the World Wide Web and within corporate enterprise systems.
Web 2.0 the second generation of the World Wide Web, where information is exchanged and presented in a two-way manner (read and write) rather than one-way manner (read). The phrase is commonly used in the context of collaboration. Examples of web 2.0 technologies include blogs and social networking sites.

Visualization adding value

Technological advances in the past few decades have dramatically increased our ability to generate primary data in the study of human disease. While one laboratory generates a dataset to test a particular hypothesis, the accumulation of multiple experimental results in online databanks can facilitate new knowledge discovery by others [1,2]. Yet there remain several challenges that prevent these resources being fully exploited by drug discovery scientists, particularly the size, complexity and poor integration of the data [3]. Visual interfaces are a crucial element in tackling this problem by enabling data exploration and engaging the human ability to synthesize complex visual inputs into meaningful understanding [4]. Here, we review the development and benefits of a visual environment designed to aid the discovery of drug target opportunities through the integration of disparate genomic and chemogenomic data. Although several innovative visualization approaches already exist in this area, the majority consider only a limited range of the data and do not provide a broad coverage of this extensive landscape. Nevertheless, these systems provide important insight into how key design elements can be used to increase the utility or usability of the software, as illustrated by the following examples.

Beyond predetermined content

The major genome viewers (University of California Santa Cruz, or UCSC [5]; Ensembl [6]; and the Generic Genome Browser, or GBrowse [7]) illustrate the benefits of systems designed to grow in parallel with experimental science. Each of these platforms can be uniquely customized to a particular experiment by allowing researchers to upload and view their own genome-based data in custom 'data tracks'. As a consequence, the systems convey much more focused and contextually relevant information, while also eliminating the need for complex data import and export steps. Furthermore, the development of standard formats and protocols for these data tracks promotes the creation of numerous tools and extensions, increasing the longevity of the software. An excellent example of this is the Galaxy system [8], which provides a streamlined environment for the manipulation of large genomic datasets, simplifying basic data manipulation operations and enabling biologists to spend more time interpreting the data. With a few simple operations, experimental results are visible as custom tracks within the UCSC browser, providing a complete workflow for the analysis of next-generation sequencing data. Without the foresight of these custom tracks, this entire workflow would have been much more complex and perhaps required the development of additional software components duplicative of much of the genome browser functionality.

Expanding dimensions

The very nature of a biological system lends itself to representation as a mathematical graph – nodes representing entities (genes, proteins and drugs) and edges representing action or interaction. To this end, a vast number of visualization programs for network analysis have been developed, offering an extensive range of features [9,10]. A common limitation of such software is the rapidly diminishing ability to visually decode the network as the number of nodes and edges increases. To combat this, tools such as ProViz [11] and FORG3D [12] provide pseudo-three-dimensional network views that generate separation between entities by increasing the apparent visual area. The 3D-SE system [13] extends this further by mapping relationships across an interactive virtual sphere, enabling users to explore many more connections than a two-dimensional view. However, three-dimensional displays might not always succeed in making the information clearer, particularly in situations in which there are multiple diverse aspects to the data. This led the designers of Arena3D [14] to propose an alternative methodology that separates multi-factorial datasets into distinct two-dimensional subgraphs. Different types of information, such as protein–protein interactions, protein–disease connections and protein–structural domain relationships, are drawn on their own layers that are subsequently stacked upon each other within three-dimensional space. In this way, Arena3D achieves the main aim of related systems (the greater use of spatial organization) while retaining the benefits of two-dimensional representation (clarity in specific relation types). Finally, any consideration of three-dimensional space within biology cannot ignore that much of the data is generated from the context of living cells and tissues. One of the most prominent tools here is the Allen Brain Atlas [15], which provides an unprecedented view of gene expression information by placing the data within an accurate representation of brain physiology. However, a system also

worthy of note is Illoura [16], which combines novel visualization with the use of established standards for data exchange between biological resources. Through this mechanism, Illoura was used to probe insulin granule distribution patterns in pancreatic beta cells by combining a three-dimensional cell model with protein subcellular localization patterns drawn dynamically from an online database.

Information maps

Cartographic techniques have also shown great value in representing large, multi-factorial datasets such as gene expression arrays and other life science data. The 'springScape' approach [17] follows this philosophy, representing the entire Gene Ontology [18] (see Glossary) as a series of distributed points across a landscape. These are subsequently interconnected by software 'springs', the tension of which increases the more two entities are related, drawing similar functions closer within the network. Intensities from gene expression analyses can be overlaid and rendered as coloured peaks rising from the surface map to directly indicate functional changes in the data. In the case of the *Caenorhabditis elegans* transcriptome, an analogous surface map was constructed by clustering genes based on co-expression patterns across more than 500 microarray experiments [19]. Surface peaks or 'gene mountains' were shown to represent high-density regions rich in genes with similar functions, aiding the classification of this genome. The authors of GeneTerrain [20] used protein–protein interaction data to create a surface map of signalling networks involved in Alzheimer's disease. As with previous examples, the overlay of disease-relevant gene expression profiles onto this landscape created three-dimensional contour projections that draw the eye to regions of interest. Importantly, GeneTerrain demonstrates how a very different view of protein interaction networks can provide a powerful way to navigate large genomic datasets. Finally, sample classification problems such as patient stratification can also be addressed through this type of approach. FreeViz [21] plots individual samples as points on a two-dimensional surface, with the distance between any two samples determined by their overall similarity. The background of the visualization is organized into differently coloured regions, each displayed as a gradient of colour intensity. The region in which any sample point is located dictates its cluster designation, and the underlying intensity provides confidence in membership of that group, helping investigators to more readily understand patient stratification.

Zooming in

An alternative mechanism for reducing visual overload is to view data at multiple levels of resolution. Although a zooming function is ubiquitous across modern software, genome browsers such as X:Map [22] and Genome Projector [23] show how groundbreaking technologies from other domains, such as Google Maps (http://maps.google.com), can be employed within scientific analysis. The benefit in this case is a greatly improved ability for dynamic genome exploration, enabling users to pan across and probe regions of interest through an intuitive and responsive interface. It should be noted, however, that the ability to adjust the resolution within a particular view is distinct from the concept of 'semantic zooming' [24]. An example of this is the display of different levels of granularity of cell signalling through the use of 'metagraphs' – networks in which each node is itself a more specialized regulatory circuit. Software such as VisANT [25] has been specifically designed to facilitate this form of browsing, enabling users to switch between views of high-level biological connectivity and low-level processing. Ultimately, this type of approach might enable navigation from tissue physiology, through cell–cell communication and signalling pathways and into genome-level events, connecting previously disjointed areas of biological information. Although this is some distance away, software such as BrainSnail [26] embodies this concept by enabling knowledge capture and navigation within distinct 'planes' of different information types, providing another angle on the concept of filtering the view of biological data.

The human element

Although functionality and analytical power are primary factors in choosing scientific software, good visualizations coupled with high usability can deliver applications that actively engage research scientists. For example, according to one survey (http://evolution.genetics.washington.edu/phylip/software.html), there are a staggering 437 different software options for phylogeny analysis. Yet the compelling graphics of the Interactive Tree Of Life system [27] or the innovative display of very large taxa implemented in Dendroscope [28], matched with their ease of use, present excellent starting points for many users. The same could be said of tools such as Circos [29], which redefines the presentation of complex tables and chromosomal ideograms through an innovative circular representation. Similarly, the Utopia system [30] enables common scientific tasks (such as sequence analysis, reading literature and pathway navigation) through software that both provides a compelling user experience and adheres to a strong standard-based architecture. These tools demonstrate how an empathy with the user community and understanding of gaps and frustrations is important to help to create applications that scientists want to – rather than have to – use.

Drug target hypotheses

Building biological rationale

Visualization approaches are employed across many facets of drug discovery and play an important part in understanding of chemical space [31], high-throughput screening results [32], compound toxicity [33], pharmacological relatedness [34], inter-disease relationships [35] and drug repurposing [36]. However, many of these examples focus on a particular analysis or data type and form only part of an overall picture of target-disease rationale. Scientists wishing to gather a complete picture of the current therapeutic landscape for an indication of interest are often forced to employ *ad hoc* knowledge assembly processes. These are not only laborious to perform and update but also make it difficult to share the resulting output between different groups. A major factor driving this situation is the wide range of disconnected sources required to collate information around biological and chemogenomic rationale. For example, databases tracking clinical programmes form the primary source of target–indication combinations that have moved into the development phase [37,38]. Sources such as Prous (http://www.prous.com), the Investigational Drugs Database (http://www.iddb.com), Adis Insight (http://www.adisinsight.com) and TrialTrove (http://www.citeline.com) provide large

amounts of information gathered from patents, conferences, websites and other materials. Evidence of earlier stage programmes can also be found in these databases and can serve as a useful indicator of pharmaceutical interest, although many of these will be some time away from clinical validation. For more novel targets, one can draw upon data from genetic association studies and phenotype observations from *in vivo* gene manipulation experiments. There are a range of these sources, including the genome-wide association database [39], the genetic association database [40], the mouse knock-out database [41] and Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim). The GeneRif [42] system is also of use here, providing access to community-based gene-centric annotation that includes highly accurate disease connections [43]. These relationships can be supplemented with further gene–disease associations derived from text-mining analyses across the biomedical literature. Commonly, co-occurrence of terms for genes, diseases and phenotypes within text is performed with subsequent ranking of these associations by simple document counts (perhaps favouring co-occurrence in title or co-occurrence in the same sentence) or statistical methods, such as *t*-tests [44] and *Z*-scores [45]. Although co-occurrence suffers from both false positives and false negatives, it remains a powerful technique for the triage of existing knowledge [46,47]. In addition, natural language processing-based analysis can be used to extract more specific relationships from the literature by looking for specific language constructs. Examples include searches for patterns of interaction-type verbs (such as 'activates' or 'inhibits') near to protein names, suggesting physical or at least functional interactions [48].

In addition to direct gene–disease connections, resources such as the Gene Ontology [18] can be used to identify genes involved in biological processes of relevance to a disease. Although these indirect connections might not always prove as meaningful as their direct counterparts, they nevertheless provide an important mechanism for identifying disease-relevant genes [47]. Finally, gene–disease associations can also be inferred from high-throughput genomic experiments, such as DNA microarrays. Often, these can be derived from integrated meta-analyses across multiple profiles within a given disease system, either from the literature [49–51] or generated internally. It should also be noted that even 'normal' expression profiles, such as those held in systems such as the GNF SymAtlas [52], provide important information regarding genes expressed in a relevant tissue.

The chemogenomics element

In addition to disease biology, a further crucial judgement in any target assessment is the likelihood of obtaining a safe, efficacious therapeutic agent such as a small molecule or biotherapeutic (i.e. 'druggability'). For small molecules, it is important to have access to the full spectrum of any existing chemical matter and associated data. If the target under study is the subject of clinical or later-stage discovery activity, competitor intelligence databases described above might provide key information. The competitor and clinical trial-orientated databases are supplemented by resources that contain detailed pharmacology information, such as MDDR (http://www.symyx.com/products/databases/bioactivity), Wombat (http://www.sunsetmolecular.com), GVKBIO (http://www.gvkbio.com/informatics.html), BioPrint (http://www.cerep.fr), PubChem (http://pubchem.ncbi.nlm.nih.gov), BindingDB [53] and ChEMBL [54], as well as internal screening results and the patent literature [55]. The physiochemical properties of compounds found in these sources can help build a profile of the chemical space and opportunities for further development and differentiation for a particular target [56]. Consequently, the generation of a complete picture requires integration across as many of the public and commercial resources as possible. Indeed, a recent analysis demonstrated that despite considerable overlap, most systems contain molecules not found in other databases [57]. Although it is reassuring to know each of these is contributing some unique content, it also illustrates the challenge facing any scientists wishing to generate a complete view of the chemogenomic landscape.

For those targets for which no synthetic small molecules exist, estimates of the success of obtaining these can be generated through a variety of means. A primary question is whether a particular protein binds to an endogenous cellular small molecule and whether such molecules are 'drug like' according to guides such as Lipinski's Rule of Five [58]. Key data sources that provide this information are ChEBI [59], Human Metabolome Database [60], KEGG [61] and Stitch [62], and again, each provides overlapping but unique data. Alternatively, where protein structure information is available, this can also be used to identify potential binding surfaces for drug molecules through binding pocket identification algorithms [63,64]. Finally, more speculative sequence-based druggability predictions can be used to triage genome-scale datasets to identify potentially novel targets [65,66]. Such algorithms use statistical techniques to score entire proteomes according to the presence of certain key features (e.g. transmembrane helices, signal peptides and subcellular localization) that are enriched in known drug targets and have been successfully applied in several disease-relevant projects [67,68]. In addition, machine-learning algorithms can identify hidden traits shared by successful targets, which can then be applied across emerging target space to predict potential chances of success [69].

In addition to druggability analysis, large-scale pharmacology databases also provide the substrate for selectivity algorithms that identify non-specific target–ligand interactions. For example, affinity for the same small molecules can infer binding site similarities between proteins with no obvious relationship at the primary sequence level [70]. Alternatively, structural similarities between the ligands of individual targets might identify those proteins sharing similar areas of chemical space [71]. Because these types of algorithms are amenable to routine, automated computation, they represent a powerful way to exploit the data held in these repositories to address key questions in early drug discovery.

The target information landscape

At Pfizer, we have developed an internal data warehouse that holds data from many of the sources described above through integration via established strategies [3]. This includes more than 5 million data points from gene expression studies, more than 1 million gene–disease assertions mined from 22 million document abstracts, 400,000 clinical and competitor intelligence summaries and more than 500,000 active compounds and natural ligands. Even with this infrastructure, we learnt that scientists wishing to use this information to find drug target opportunities faced a complex process of manual data navigation and assembly. Most

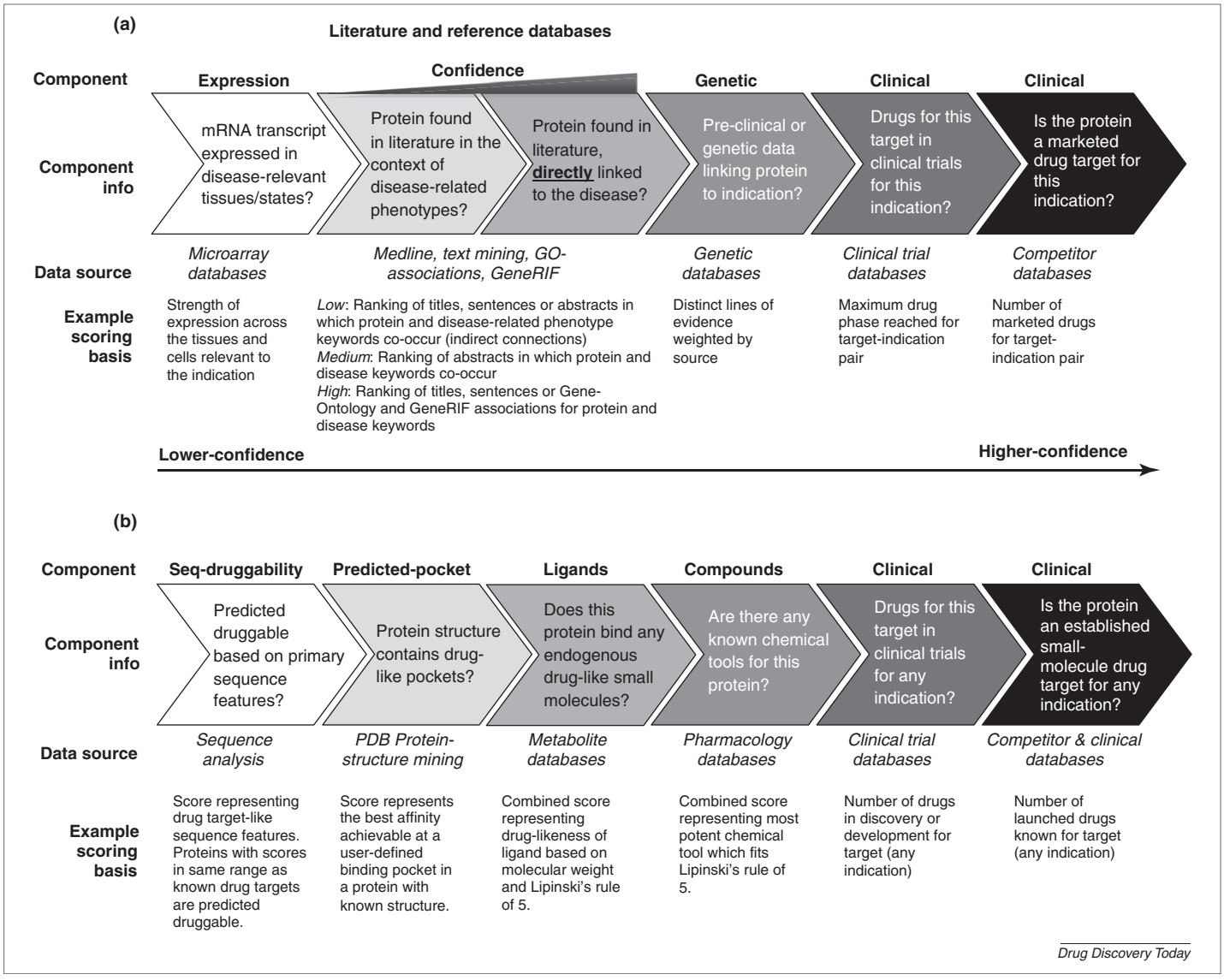


FIGURE 1

Data domains that contribute to the drug discovery landscape. **(a)** Ordered ranking of data categories that provide evidence connecting proteins to disease. **(b)** Ordered ranking of categories of evidence supporting small-molecule druggability for proteins.

often, this was performed through custom, *ad hoc* database or web queries and presented in the form of static Microsoft Excel datasheets. Crucially, scientists often added a layer of interpretation across the data, ranking the evidence along practical drug discovery perspectives to build confidence in any particular approach. Although this analysis was performed by many different groups, the overall questions were similar, namely: Which targets are most strongly connected to an indication or phenotype? Which are mostly likely to be druggable (and do any chemical or biotherapeutic starting points exist)? If a disease association is known, how far has this idea previously been taken by others? Thus, it became clear that what was needed was a mechanism to move this type of analysis forward, moving the emphasis from data gathering to interpretation and project decision making.

Rationale-based design

The use of scoring matrices to assess the rationale of potential therapeutic approaches is a common practice within the discipline and can be used to build a risk-balanced portfolio [72]. Different programmes can be assessed on the level of confidence individual pieces of evidence contribute to an overall likelihood that a project will be successful. This highly decision-focused view of information forms the basis for the design of our internal system. As Fig. 1 shows, the target opportunity landscape can be described as two series of ordered components that define increasing disease association and target druggability. These two series can also become the axes of a two-dimensional scatter plot representing increasing biological and chemogenomic rationale for proteins as related to a specific disease or phenotype. Within the plot, every human protein is assessed against the evidence represented by each component and then resolved as an individual point on this landscape, as shown in Fig. 2a. Of course, many proteins are expected to have evidence in multiple components along each axis, potentially resulting in a dense and confusing graphic. To combat this, we leverage the rank order of the data domains (Fig. 1) to simplify the display. Specifically, although the evidence for a protein is assessed

(a) Info tab containing links to target summary and annotation view

Toolbar containing select, zoom and search functions

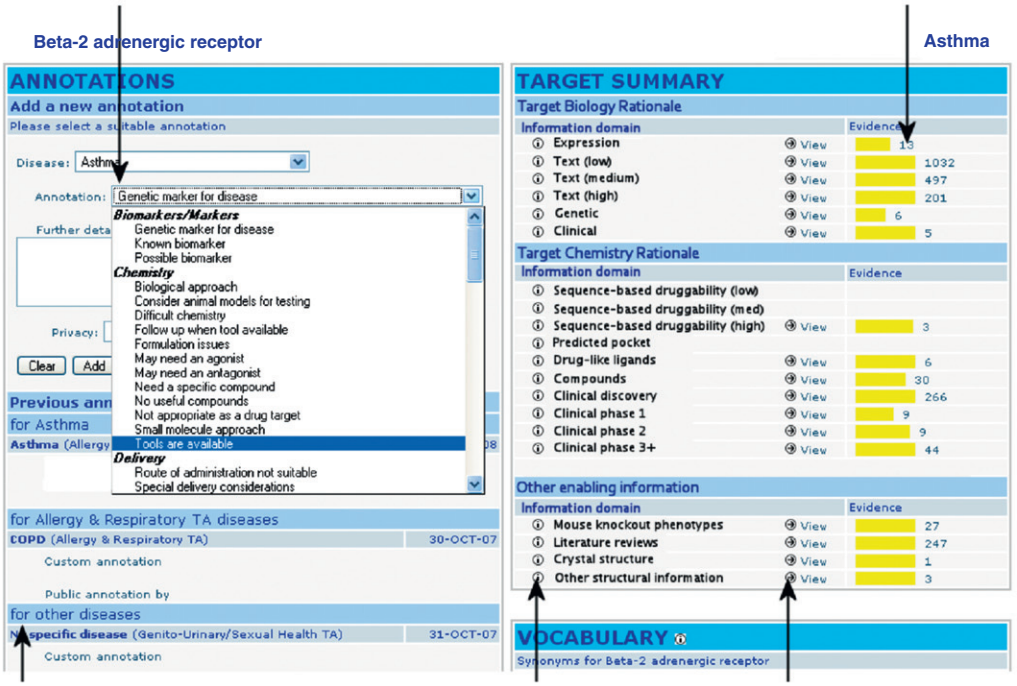
The filters tab has a hierarchical structure of gene/protein lists ranging from GO annotations to text-mined datasets filters can be combined using boolean logic.

Users can hover the mouse over each spot to view protein names and identifiers

A filter can be applied in three ways:
(i) Highlight genes/proteins resulting from filter combination
(ii) Only show genes/proteins resulting from filter combination
(iii) Hide genes/proteins resulting from filter combination

(b) Users select one or more target annotations from a hierarchy and add further comments

The raw score for the information domain (e.g. number of papers or compounds)



Previous annotations by users are displayed here

Hover to view a description of each information domain

Link to other system for viewing the supporting evidence such as literature or drug information.

Drug Discovery Today

within all of the components individually (and stored in an underlying database), the protein is drawn only once, plotted within the highest ranking *x*- and *y*-axis components for which there is evidence. For example, the position of a drug target used in current disease therapy would be within the clinical components at the right of the chart, regardless of how the protein is scored within other *x*-axis domains. Similarly, a protein for which there are known chemical tools would be located towards the middle of the *y*-axis, irrespective of the assessment of lower ranking sequence-based druggability evidence. Thus, the view specifically describes the degree to which each protein has progressed according to drug discovery precedence. Once a protein is assigned to the highest-ranking component, its exact position is fine-tuned with respect to evidence levels through quantitative scoring, as outlined in Fig. 1. We should stress that in the majority of cases, these calculations are either well described or generally obvious. Specific examples include grading targets based on the Rule-of-Five properties of their associated compounds [58], grading text-mining results on statistical scores [47] and grading targets in clinical space based on stage-gate progression times.

Pharmaceutical zones

Although the scatter plot is specifically designed to provide a simplified view of the available information, users could still be daunted by the number of proteins associated with a disease. Although standard zooming functions provide magnification to and focus on particular regions, the overall design of the landscape provides an additional mechanism to address this issue. As shown in Fig. 2a, specific regions of the plot are mapped to particular classes of drug discovery questions. For example, a disease team looking to identify a complete list of clinically precedented small-molecule targets would concentrate on region 'F' of the plot. Alternatively, key biotherapeutic opportunities will be enriched in region 'E' and/or 'H' (high disease rationale and low small-molecule druggability), whereas novel chemistry opportunities lie in region 'D' (potentially druggable but no clinical compounds). This subdivision of the visualization into readily identifiable, pharmaceutically distinct zones provides a highly contextual view of the information. The more directed mining that this enables is further illustrated by an application to compound repurposing, an important route to maximizing return on internal investment [73]. Region 'C' of the plot (Fig. 2a) is particularly important for this because it represents proteins with known chemical matter that are associated with the disease of interest but for which there is no evidence of clinical programmes within the competitor databases. This provides a potentially fruitful collection of druggable proteins that might provide a rapid entry into an exploratory drug discovery programme. Of course, many of these connections will not provide such an outcome, but the aim of the system is to enable more rapid discovery of these testable opportunities, reducing the need to trawl multiple unconnected databases. The simplicity with which users are able to perform this task now enables routine cross-referencing of the entire company target portfolio across all therapeutic divisions to rapidly identify value-added opportunities.

Ultimately, our aim is to grade the large number of potential opportunities according to the level of precedence suggested by the available evidence. Importantly, this does not mean that a protein towards the bottom left of the chart is any better or any worse a choice of target than one towards the top right. Such decisions are complex and are based on many scientific, business and human factors [74]. Rather, by gathering and partitioning the information in a more intuitive way, we hope to more rapidly generate subsets of potential target opportunities for deeper scientific scrutiny. Currently, we compute co-ordinates for more than 3600 different disease and phenotype landscapes, with monthly historical snapshots each represented by approximately 15 million data points. The entire system can be refreshed in less than 5 hours, satisfying the most demanding requirements for currency and responsiveness. The system is currently in use across all Pfizer research sites, covering an extensive range of disease and therapy areas.

Evolving functionality

Digging deeper

Development of the graphical interface to the landscape system was based upon a continuous conversation between research scientists, data analysts and informaticians. From the beginning, the tool was used in real-world projects, which directly highlighted key bottlenecks and limitations of the system. This not only allowed new feature development to be driven by immediate scientific need but also identified additional use cases that might not have been captured otherwise. Perhaps the most important feature request was the need to access the underlying evidence behind any assertion to understand exactly why a protein was located in a particular position. This functionality is realised as a report view (Fig. 2b) that provides full details of how the protein position is calculated, along with hyperlinks to source databases and original data within every axis component. A community annotation tool sits alongside the detailed information (shown on

FIGURE 2

Outline of the target landscape visualization. **(a)** Plots are constructed to represent a specific indication or phenotype, where linkage to disease and druggability is represented by the *x*- and *y*-axes, respectively. Axes are divided into subcomponents aligned to the qualitative ranking described in Fig. 1 and ordered in such a way as to provide increasing confidence towards the top right of the plot. A given protein (spot) is only shown on the plot in one location, appearing within the highest ranking component (on each axis) for which evidence is available. The size of each spot can be set by the user based on additional data: the plot above employs a scheme based on available research tools (mouse knock-outs, clones and crystal structures). The positions for all human proteins are computed for the landscape. The total known and potential target landscape for a disease can be divided into business-relevant areas, for example: (A) high-risk opportunities (owing to lack of disease-linking evidence) but testable hypotheses (owing to available chemical tools); (B) some novel chemistry or biotherapeutics opportunities but high false positives; (C) compound repurposing opportunities, in which there is evidence for disease linkage and available chemical matter; (D) novel chemistry opportunities, in which there is some evidence linking the protein to a disease but few chemical tools; (E) opportunities for biotherapeutics, in which small-molecule approaches might be difficult; (F) competitor activity based on information that has been disclosed by companies; and (G) and (H) competitor activity, in which small-molecule data are not unavailable or using biotherapeutics approach. **(b)** The target summary and annotation view for a target and disease pair. A user can select annotations from a controlled hierarchy and view previous annotations. Links are provided to source databases and other online resources, enabling the user to view all information that was used to place the target within the landscape.

the left of Fig. 2b), employing a controlled hierarchical vocabulary to aid the capture of assessment outcomes. These annotations include interpretation of the evidence, highlighting contradictions or overall decisions as to the suitability of a particular target. Importantly, this also enables scientists from different therapeutic areas to identify situations of mutual interest in a novel target, perhaps providing an additional incentive to initiate exploratory analysis.

Customizable landscapes

For illustration purposes, we concentrate here on a generic set of biological and chemogenomic components. However, from the beginning, the system was designed to accommodate custom charts that omit certain elements or include new data sources, such as disease-specific databases. Even on the generic chart, there are several different algorithms that can be applied to rank proteins within a data component, which, again, can be customized as required by project teams. In addition, alternatives to small-molecule druggability, such as small interfering RNA (siRNA) [75] and monoclonal antibody [76] approaches, can also be assessed on the *y*-axis. Thus, the tool not only supports common questions but also can be tailored to the needs of individual disease research groups. All of the data within the system are also accessible via web services, enabling more computationally orientated scientists to incorporate the data into custom workflows and data pipelining tools [77].

Regardless of the landscape design, one of the most important features is the ability to limit the landscape to display only a specific subset of proteins through data filtering, providing extensive customization capabilities. We prebuilt several lists of proteins sourced from a range of public and commercial repositories that could be used to highlight, remove or specifically show matching entities. Important filters include those listing relationships to specific mouse knock-out phenotypes, Gene Ontology categories, ligand-derived selectivity indices and internal portfolios. Combining filters through Boolean logic enables the creation of highly focused subsets of the drug target landscape. An example is the use of Protein Data Bank and Gene Ontology filters to show only those

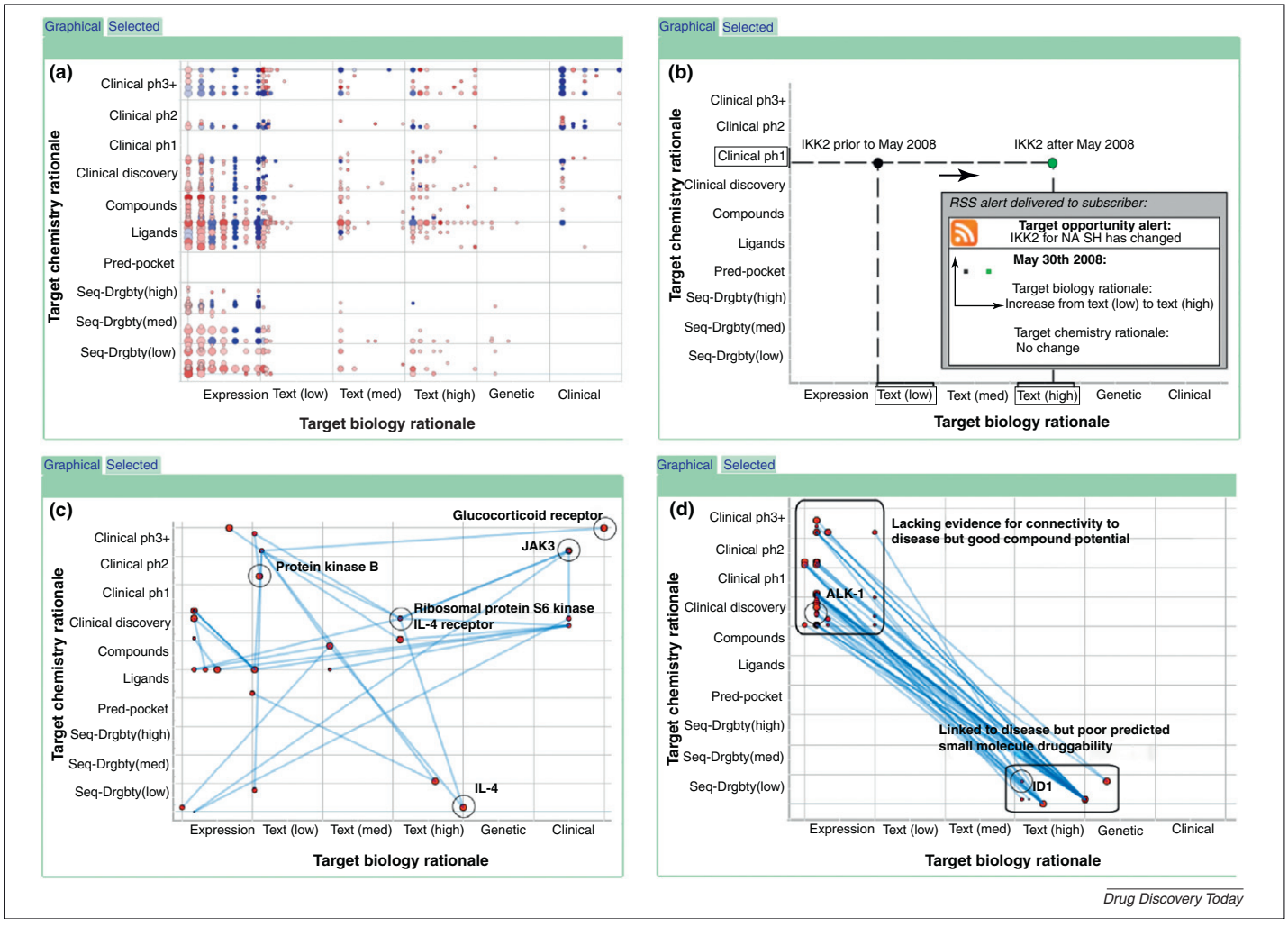


FIGURE 3

Exploring the target landscape in detail. **(a)** The heat map view, in which quantitative data such as gene expression results can be incorporated. **(b)** A new publication mentioning IKK2 causes the protein to jump across the NASH target landscape, triggering an RSS alert about this event, shown in the inset. The user can follow links to the evidence responsible for the jump. **(c)** The interleukin-4 signalling pathway is shown in the context of the target landscape for psoriasis. **(d)** Protein–protein interaction data used to filter the landscape scatter plot to identify chemically tractable proteins not obviously linked to a disease. The protein ID1, which has been linked to psoriasis, has low predicted druggability. However, ID1 connects to other proteins, such as ALK-1, which might provide alternative small-molecule opportunities.

proteins with solved crystal structures that also possess a particular biological function. In addition to pre-computed filters, user-specific gene and protein lists can be created, optionally associated with numerical values or scores. A common use for this function is to enable the incorporation of new experimental data, such as that from proteomic, microarray or siRNA experiments. The results are visualized using heat map style display (Fig. 3a), providing a mechanism to rapidly assimilate new results in the context of the existing information.

Assessing all the evidence

The summarization scheme initially masks data in lower ranking zones to represent a protein as a single point and simplify the presentation to the user. While all of the data are available through the summary view (Fig. 2b) on a per-target basis, users might also wish to use this information in triaging the full landscape. This can be achieved using the filter mechanism to display the scores calculated for any data component as a heat map across the proteome. For example, the values for the gene expression component can be used to highlight all proteins that exhibit this evidence, regardless of overall position. Alternatively, evidence from the druggability predictions might be applied; this can be useful to represent the likelihood of new chemical material even where some small molecules already exist. As with all filters, the process can be repeated for other domain scores, either individually or through Boolean combination. In this way, the system retains the power of the summary view while providing a mechanism to ensure decision making is made in light of all the available evidence and not just that in the highest component.

An ever-changing landscape

One of the most crucial features to be identified was the need to provide an alerting mechanism to highlight changes to the landscape caused by new data. By storing bi-weekly snapshots of the evidence, it is possible to compare current and historical evidence for any protein and alert users to any new findings. Furthermore, the inherent design of the system allows for a more tailored alerting process that only triggers on potentially interesting increases in the disease rationale of potential targets. To achieve this, we set a customizable limit on the degree of movement required by a protein either within or across components to register a change. This enables scientists to keep abreast of major events without the need to scan large numbers of new publications and databases. For example, before 2008, the IkappaB kinase 2 (IKK2) protein would be found in the low-confidence text-mining domain of the non-alcoholic steatohepatitis (NASH) disease landscape as a result of an article describing limited connectivity between the gene and disease [78]. However, after the publication of a paper demonstrating that IKK2 inhibition directly blocks NASH initiation [79], the protein moved across the target landscape towards the far right of the high-confidence text-mining component because of this much stronger evidence. A user with a registered interest in NASH would receive an alert through a Really Simple Syndication (RSS) news feed (Fig. 3b). This follows the same visual theme as the landscape itself, designed to promote rapid interpretation of the impact of the new information. Thus, landscape-based alerting differs from other systems such as keyword alerts by leveraging what is already known about a protein–disease connection to filter out information that does not alter the relationship. Alerts are provided in a directly applicable context: for example, new disease-linkage information is provided within the context of druggability, and vice versa. Of course, this approach is not designed to replace more comprehensive monitoring of information required for targets under active research. However, because it is simply not possible for a scientist to monitor the entire proteome across multiple indications, the landscape alerting system provides a mechanism by which major new developments can be identified early.

Case studies

By developing the landscape system *in situ* with disease area programmes, we identified a wide range of scenarios in which the tool could be of use, which are summarized in Box 1. Below, we highlight some of these use cases in more detail.

Pathways in context

Cellular signalling pathways provide a deeper understanding of disease [80] and drive a more systems-orientated approach to drug discovery [53–56]. Researchers have a wide choice of pathway and protein–protein interaction systems such as Reactome [81], the NCI-Nature Pathway Interaction Database [82], HPRD [83] and many more (http://www.pathguide.org). However, subsequent cross-referencing of these pathways against internal portfolios, druggability information and clinical precedence often requires multiple import or export and data translation steps. Thus, the integration of this type of information with the target landscape system represents an important step towards addressing this. An exploration of the interleukin-4 (IL4) signalling pathway and its involvement in psoriatic disease [84] provides an illustrative case study. The IL4 canonical pathway can be obtained from sources described above and used as a 'network filter' to show only proteins involved in this system and their connections (Fig. 3c). This representation complements graph-based pathway diagrams using Cartesian positioning to place each protein node within a pharmaceutical context and highlights related nodes amenable to small-molecule modulation.

Extending target networks

In addition to canonical pathways, one can also incorporate high-throughput interaction data to identify additional druggable targets that interact with known disease-implicated proteins (Fig. 3d). In this example, several proteins with linkage to psoriasis have been used to seed an interaction network to identify binding or regulatory partners with lower rationale (i.e. more novel) but with higher druggability. One such protein, the DNA-binding protein inhibitor 1 (ID1) is visible as being strongly linked to psoriasis [85] but with limited expectation of being amenable to small-molecule modulation. Expansion with protein interaction data identifies the more druggable Activin-receptor-like kinase 1 (ALK-1) as a modulator of ID1 [86] with limited psoriasis rationale. However, ALK-1 is a participant in the TGF-beta system [87], which is known to participate in inflammatory skin disorders such as psoriasis [88] and might provide an interesting area for further research. Clearly, such correlations do not in themselves constitute robust evidence that a particular approach will work in the clinic or pre-clinical models. However, the integration of protein networks and drug

BOX 1

Applications of the drug target landscape

Rapid lookup of information

A simple mechanism to determine drug discovery precedence. Useful for new team members or for researchers looking at the role of a target across a range of diseases.

Targeting a protein family

Strong project enablers such as crystal structures, practical experience and available assays might lead to a protein family being favoured within a company. By computing landscapes for more than 1000 diseases, scientists are able to cross-reference each member of the family to all diseases of interest to the organization and, thus, maximize the effort.

Targeting a signalling pathway

Often, drug discovery programmes will investigate opportunities across a signalling pathway involved in disease. Scientists can view where each pathway member lies within the landscape, identifying both precedented and novel opportunities.

Augmenting a mechanism hypothesis

From a protein that is known to be involved in a disease, researchers can use in-house and external protein–protein interaction data to navigate to an alternative point of intervention. This might be more promising in terms of chemical tractability or previous experience with that pathway.

Early identification of new opportunities

Comparison of target landscape snapshots over time enables detection of new information that has impact or new opportunities requiring immediate evaluation. Because this greatly reduces the amount of information presented, scientists can also sign up to additional alerts on related phenotypes that they might not otherwise be able to monitor.

Visualizing high-throughput 'omics data

Filtering a target landscape with a list of genes from a high-throughput experiment enables a scientist to rapidly determine the level of precedence within the set. This could also include the identification of enriched membership in diseases other than that of immediate interest, suggesting possible connections in biology and cross-group opportunity.

Compound and mechanism repurposing

Targets with good chemical matter but no record of having ever been clinical targets for a specific disease of interest present rich opportunities for drug repurposing. Alternatively, knowledge of the existence of a chemical tool modulating a protein of interest might accelerate basic research into a biological system.

Target portfolio reloading

A team-based target triage exercise can use the landscape as a starting point, covering target space for any indication. Analysis might involve multiple data filter exercises, subsetting the data across important biological processes or related phenotypes. After an initial high-level triage, a team can then work through an iterative shortlisting process, leveraging the community annotation tool to record decisions.

Visualizing internal portfolios

By connecting the system to the internal drug target portfolio, scientists are able to view all targets of current interest to the disease area, against the background of precedence. This can be valuable when assessing the spread of risk across the portfolio. Alternatively, one can take the active targets within one therapy area and view them on a series of additional disease landscapes to obtain a view of potential cross-group collaborative opportunities.

Competitor assessment

Filtering to targets that are the subject of an organization's current discovery portfolio enables rapid determination of mechanisms that are under investigation by many companies and those that present key competitive opportunities.

discovery landscapes provides a much-needed methodology for systematic mining of available opportunities.

Continuous feedback

The landscape approach fits within a continuous cycle of target discovery and validation. A target triage exercise involving the system often begins by defining biological processes relevant to the disease and creating tailored charts through customized literature and database searches. A team-based *in silico* exercise is then performed, dividing the landscape into manageable regions and team members reviewing the underlying evidence for each protein. Results from previous analyses and newly commissioned studies (including microarray, proteomics and siRNA studies) are often included through the filtering mechanism. Throughout this exercise, hundreds of human annotations and decisions regarding protein–disease assessments are captured into the system, informing future cycles of the process. At times, annotations are made to suggest alternative uses for a gene or protein, such as a biological probe using siRNA study or as a potential biomarker for early research, providing additional aid to the programme. The dynamic nature of the target landscape system also enables continuous tracking of targets whose underlying evidence might be insufficiently compelling yet strong enough to maintain an active interest. Alerts can be initiated to monitor all such possibilities and notify scientists when there might be new evidence (say, the publication of a new small-molecule chemical series) that might

cause a re-evaluation of that target. Finally, the annotation capabilities also capture protein positions misplaced by the automated data gathering and analysis made by the system. This helps to increase the validity of the information to other team members and also highlights data quality and integration issues, improving aspects of the data processing steps.

Conclusions and future directions

The generation of hypotheses from large genomic and chemogenomic datasets is the subject of much research, generating many individual algorithms and web resources [89]. Yet as these mature, there is a need to ensure they move beyond the realm of computational scientists and are accessible to a wider population of drug discovery researchers. The target landscape represents such an attempt, specifically designed to connect several important databases to provide a more holistic picture of existing knowledge. Clearly, this is not designed to provide deep analysis into specific pathways or validate novel targets. Rather, the landscape system fits with a portfolio of analysis software and complements more quantitative experimentation and systems biology strategies. Indeed, landscape-based searches often form the first stage of target analysis, generating lists of proteins with known involvement in a disease as input to more complex modelling. Conversely, a large-scale genomic or pathway analysis might create a list of proteins that the user wishes to assess rapidly for disease novelty, as well as for connections with other disease areas under investigation by the organization. Thus, we see the landscape as one element of a multifaceted workflow, providing access to information in a more integrated and context-orientated manner than has previously been possible.

Although this retrospective has concentrated on the development of one specific system, the experience has provided more general insight into some of the considerations for future informatics tools in this space. Below, we highlight three key areas.

Data standards

One of the biggest challenges in creating a system such as the target landscape is the difficult task of maintaining the currency of the underlying data, which, in turn, distracts from more important efforts to develop analytical aspects. A major contributor to this problem is the lack of standards across resources. This is a clear example in which the widespread adoption of core standards across industry, academic and commercial content providers would accelerate these efforts greatly. Of greatest need is the establishment of controlled vocabularies and taxonomies to describe common, core entities and processes within drug discovery. Although some of these exist through key public bodies, such as the Open Biomedical Ontologies Foundry [90], important concepts around mechanism of action, pharmacological data and screening assay terminologies (to name but a few) do not. We believe both industry and data providers must move away from *ad hoc*, individual and/or proprietary standards and move to those that directly facilitate data integration across resources. Success of such an initiative requires widespread adoption of the standards, something that might only be possible if they exist fully within the public domain and without usage restriction. Clearly, the funding of both the initial construction and the ongoing maintenance of these key assets will be a challenge. However, in the long term, the cost savings associated with the removal of laborious data transformation steps, as well as the potential for accelerated scientific discovery, should more than justify this effort.

Exploiting technology

In addition to data standards, it is also important to use electronic data formats that best facilitate analysis and hypothesis generation. It is here that the semantic web (SW) [91] and its associated descriptive language, known as Resource Description Framework (RDF), hold much promise [92]. RDF has the potential to encode information in a format amenable to interpretation by both humans (via presentation software) and, crucially, computers. The highly structured nature of RDF-encoded data facilitates the development of automatic reasoning tools that can scan a network of information to infer new causal linkages [93]. Yet the realization of the value of SW technology within drug discovery will not be through data integration or even through computational analysis but through the application of the integrated data to solving real problems in human health. To accomplish this, we must ensure that these tools are accessible to as many scientists as possible, leveraging the combination of new technology and human biological knowledge. Although visualization approaches are crucial to accomplishing this, only a handful of practical real-world examples currently exist (including Illoura [16] and Utopia [30]). Thus, as the SW matures, the development of better mechanisms to interrogate the resulting network of information will be crucial to realizing the potential of the advances made so far.

In addition to context-specific visualizations, we could also see greater benefit from more generic approaches. In particular, the concept of 'mashing up' different pieces of information from unconnected sources is gaining popularity across the Internet. Components of the so-called 'Web 2.0' generation of software such as Dapper (http://open.dapper.net), Yahoo Pipes (http://pipes.yahoo.com) and GreaseMonkey (http://www.greasespot.net) have emerged as powerful tools for extracting, connecting and manipulating web content for life sciences [94,95]. For example, the inventors of iHOPerator showed how GreaseMonkey can be used to dynamically augment protein function information on the Information Hyperlinked over Proteins (iHOP) website [95]. This represents a real paradigm shift in the use of the web, enabling individuals to customize and integrate content as they desire, no longer limited by the designers of the original website. In conjunction, several generic, high-quality data-charting applications have emerged on the web, including ManyEyes (http://manyeyes.alphaworks.ibm.com), Processing (http://processing.org), GapMinder (http://www.gapminder.org) and Axiis (http://www.axiis.org/). These enable users to rapidly transform numerical data held in databases and spreadsheets into dynamic and interactive graphs. The combination of these tools with mash-up technology and data standards might lead to a future in which ideas such as the target landscape can be created as *ad hoc* prototypes, providing timely support for individual projects.

The role of collaboration

The current surge in data coupled with decreasing revenues and increasing budget restraints creates a difficult environment for life science informatics within industry. Whether developed in-house or purchased from a commercial vendor, drug discovery software

and data carry costs that make assembling an optimal repertoire of tools and content very challenging. At the same time, there is an ever-increasing involvement of academic and non-profit organizations in drug discovery efforts, driving an increased availability of both practical drug discovery reagents [96] and pharmacological data [54] within the public domain. This creates a unique opportunity for partnership across the drug discovery community to develop the core resources required by all engaged in the discipline. We have argued that much of this work could be considered precompetitive and that the sharing of costs, dissemination of learning and pooling of ideas benefits both commercial and non-commercial organizations alike [97]. In addition, collaboration and discussion between companies might enable industry to speak with a common voice on key issues and could boost initiatives such as the adoption of data standards. A more collaborative approach might considerably aid the development of future hypothesis generation and information visualization tools for drug discovery within the public domain.

The rapidly decreasing cost of genomic technologies coupled with a glimpse of the vast complexity of genetic regulation [98] suggests a future in which the volume of data will dwarf those currently available. Major developments in informatics capabilities will be required in all areas, ranging from data storage and transfer to analysis algorithms and data integration. We would argue that an additional and essential development will be systems that place these data into disease-relevant contexts and facilitate interpretation and hypothesis generation by all members of the drug discovery team.

Acknowledgements

The authors acknowledge the guidance of Enoch Huang and substantial input from Robert Hernandez, Markella Skempri, Dave Burrows, Jerry Lanfear and Nigel Wilkinson into the target landscape visualization. In addition, we recognize the contributions from Pfizer Research Informatics, Knowledge Discovery, eBiology and TASIS departments and, specifically, Ian Harrow, Andrew Hopkins, William Loging, Ben Sidders, Anneli Sullivan, Sari Ward, Bryn Williams-Jones and Phil Verdemato. Finally, we thank the reviewers for suggestions on improving the original manuscript.

References

1 Blagosklonny, M.V. and Pardee, A.B. (2002) Conceptual biology: unearthing the gems. *Nature* 416, 373
2 Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 26, 99–105
3 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
4 Perini, L. (2005) Explanation in two dimensions: diagrams and biological explanation. *Biol. Philos.* 20, 257–269
5 Kuhn, R.M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* 37 (Database issue), D755–D761
6 Hubbard, T.J. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.* 37 (Database issue), D690–D697
7 Stein, L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.* 12, 1599–1610
8 Giardine, B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, 1451–1455
9 Pavlopoulos, G.A. *et al.* (2008) A survey of visualization tools for biological network analysis. *BioData Min.* 1, 12
10 Suderman, M. and Hallett, M. (2007) Tools for visually exploring biological networks. *Bioinformatics* 23, 2651–2659
11 Iragne, F. *et al.* (2005) ProViz: protein interaction visualization and exploration. *Bioinformatics* 21, 272–274
12 Paananen, J. and Wong, G. (2009) FORG3D: force-directed 3D graph editor for visualization of integrated genome scale data. *BMC Syst. Biol.* 3, 26
13 Naud, A. *et al.* (2007) Visualization of documents and concepts in neuroinformatics with the 3D-SE viewer. *Front. Neuroinform.* 1, 7
14 Pavlopoulos, G.A. *et al.* (2008) Arena3D: visualization of biological networks in 3D. *BMC Syst. Biol.* 2, 104
15 Lein, E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176
16 McComb, T. *et al.* (2009) Illoura: a software tool for analysis, visualization and semantic querying of cellular and other spatial biological data. *Bioinformatics* 25, 1208–1210
17 Ebbels, T.M. *et al.* (2006) springScape: visualisation of microarray and contextual bioinformatic data using spring embedding and an 'information landscape'. *Bioinformatics* 22, 99–107
18 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
19 Kim, S.K. *et al.* (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087–2092
20 Qian, Y. *et al.* (2008) GeneTerrain: visual exploration of differential gene expression profiles organized in native biomolecular interaction networks. *Inform. Vis.* 10.1057/palgrave.ivs.9500169
21 Demsar, J. *et al.* (2007) FreeViz – an intelligent multivariate visualization approach to explorative analysis of biomedical data. *J. Biomed. Inform.* 40, 661–671
22 Yates, T. *et al.* (2008) X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.* 36 (Database issue), D780–D786
23 Arakawa, K. *et al.* (2009) Genome Projector: zoomable genome map with multiple views. *BMC Bioinform.* 10, 31
24 Hu, Z. *et al.* (2007) Towards zoomable multidimensional maps of the cell. *Nat. Biotechnol.* 25, 547–554
25 Hu, Z. *et al.* (2008) VisANT: an integrative framework for networks in systems biology. *Brief. Bioinform.* 9, 317–325
26 Telefont, M. and Asaithambi, A. (2009) BrainSnail: a dynamic information display system for the sciences. *Bioinformation* 3, 289–290
27 Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128
28 Huson, D.H. *et al.* (2007) Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinform.* 8, 460
29 Krzywinski, M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645
30 Pettifer, S. *et al.* (2009) Visualising biological data: a semantic approach to tool and database integration. *BMC Bioinform.* 10 (Suppl. 6), S19
31 Villar, H.O. and Hansen, M.R. (2009) Mining and visualizing the chemical content of large databases. *Curr. Opin. Drug Discov. Dev.* 12, 367–375
32 Maniyar, D.M. *et al.* (2006) Data visualization during the early stages of drug discovery. *J. Chem. Inf. Model.* 46, 1806–1818
33 Xu, E.Y. *et al.* (2008) Integrated pathway analysis of rat urine metabolic profiles and kidney transcriptomic profiles to elucidate the systems toxicology of model nephrotoxicants. *Chem. Res. Toxicol.* 21, 1548–1561
34 Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690
35 Goh, K.I. *et al.* (2007) The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8685–8690
36 Ha, S. *et al.* (2008) IDMap: facilitating the detection of potential leads with therapeutic targets. *Bioinformatics* 24, 1413–1415
37 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* 5, 993–996
38 Zheng, C.J. *et al.* (2006) Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* 58, 259–279

39 Johnson, A.D. and O'Donnell, C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.* 10, 6
40 Becker, K.G. *et al.* (2004) The genetic association database. *Nat. Genet.* 36, 431–432
41 Eppig, J.T. *et al.* (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.* 35 (Database issue), D630–D637
42 Maglott, D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 35 (Database issue), D26–D31
43 Osborne, J.D. *et al.* (2009) Annotating the human genome with Disease Ontology. *BMC Genomics* 10 (Suppl. 1), S6
44 Agarwal, P. and Searls, D.B. (2008) Literature mining in support of drug discovery. *Brief. Bioinform.* 9, 479–492
45 Yetisgen-Yildiz, M. and Pratt, W. (2006) Using statistical and knowledge-based approaches for literature-based discovery. *J. Biomed. Inform.* 39, 600–611
46 Hu, Y. *et al.* (2003) Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.* 2, 405–412
47 Weeber, M. *et al.* (2005) Online tools to support literature-based discovery in the life sciences. *Brief. Bioinform.* 6, 277–286
48 Wang, H. *et al.* (2009) Extract interaction detection methods from the biological literature. *BMC Bioinform.* 10 (Suppl. 1), S55
49 Hulbert, E.M. *et al.* (2007) T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res.* 35 (Database issue), D742–D746
50 Novershtern, N. *et al.* (2008) A functional and regulatory map of asthma. *Am. J. Respir. Cell Mol. Biol.* 38, 324–336
51 Rhodes, D.R. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166–180
52 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
53 Liu, T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35 (Database issue), D198–D201
54 Overington, J. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aid. Mol. Des.* 23, 195–198
55 Rhodes, J. *et al.* (2007) Mining patents using molecular similarity search. *Pac. Symp. Biocomput.* 12, 304–315
56 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815
57 Southan, C. *et al.* (2007) Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Curr. Top. Med. Chem.* 7, 1502–1508
58 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26
59 Degtyarenko, K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36 (Database issue), D344–D350
60 Wishart, D.S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35 (Database issue), D521–D526
61 Goto, S. *et al.* (1997) Organizing and computing metabolic pathway data in terms of binary relations. *Pac. Symp. Biocomput.* 175–186
62 Kuhn, M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 36 (Database issue), D684–D688
63 Cheng, A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75
64 Halgren, T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* 49, 377–389
65 Al-Lazikani, B. *et al.* (2008) The molecular basis of predicting druggability. In *Bioinformatics – From Genomes to Therapies*. 1315–1334
66 Bakheet, T.M. and Doig, A.J. (2009) Properties and identification of human protein drug targets. *Bioinformatics* 25, 451–457
67 Aguero, F. *et al.* (2008) Genomic-scale prioritization of drug targets: the TDR targets database. *Nat. Rev. Drug Discov.* 7, 900–907
68 Berriman, M. *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature* 460, 352–358
69 Zhu, F. *et al.* (2009) What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J. Pharmacol. Exp. Ther.* 330, 304–315
70 Frye, S.V. (1999) Structure–activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era. *Chem. Biol.* 6, R3–R7
71 Keiser, M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206
72 Wehling, M. (2009) Assessing the translatability of drug projects: what needs to be scored to predict success? *Nat. Rev. Drug Discov.* 8, 541–546
73 Chong, C.R. and Sullivan, D.J., Jr (2007) New uses for old drugs. *Nature* 448, 645–646
74 Korstanje, C. (2003) Integrated assessment of preclinical data: shifting high attrition rates to earlier phase drug development. *Curr. Opin. Investig. Drugs* 4, 519–521
75 Alvarez-Salas, L.M. (2008) Nucleic acids as therapeutic agents. *Curr. Top. Med. Chem.* 8, 1379–1404
76 Presta, L.G. (2005) Selection, design, and engineering of therapeutic antibodies. *J. Allergy Clin. Immunol.* 116, 731–736
77 Reidhaar-Olson, J. *et al.* (2002) Process biology: integrated genomics and bioinformatics tools for improved target assessment. *Targets* 1, 189–195
78 Zhao, C.Y. *et al.* (2007) An experimental study on the reverse mechanism of PPAR-gamma agonist rosiglitazone in rats with non-alcoholic steatohepatitis. *Zhonghua Gan Zang Bing Za Zhi* 15, 450–455
79 Beraza, N. *et al.* (2008) Pharmacological IKK2 inhibition blocks liver steatosis and initiation of non-alcoholic steatohepatitis. *Gut* 57, 655–663
80 Gandhi, T.K. *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* 38, 285–293
81 Vastrik, I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 8, R39
82 Schaefer, C.F. *et al.* (2008) PID: the Pathway Interaction Database. *Nucleic Acids Res.* 37 (Database issue), D674–D679
83 Mishra, G.R. *et al.* (2006) Human protein reference database – 2006 update. *Nucleic Acids Res.* 34 (Database issue), D411–D414
84 Weigert, C. *et al.* (2008) Interleukin 4 as a potential drug candidate for psoriasis. *Exp. Opin. Drug Discov.* 3, 357–368
85 Bjorntorp, E. *et al.* (2003) The helix–loop–helix transcription factor Id1 is highly expressed in psoriatic involved skin. *Acta Derm. Venereol.* 83, 403–409
86 Ota, T. *et al.* (2002) Targets of transcriptional regulation by two distinct type I receptors for transforming growth factor-beta in human umbilical vein endothelial cells. *J. Cell. Physiol.* 193, 299–318
87 Goumans, M.J. *et al.* (2003) Activin receptor-like kinase (ALK)1 is an antagonistic mediator of lateral TGFbeta/ALK5 signaling. *Mol. Cell* 12, 817–828
88 Li, A.G. *et al.* (2004) Latent TGFbeta1 overexpression in keratinocytes results in a severe psoriasis-like skin disorder. *EMBO J.* 23, 1770–1781
89 Yang, Y. *et al.* (2009) Target discovery from data mining approaches. *Drug Discov. Today* 14, 147–154
90 Smith, B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255
91 Berners-Lee, T. *et al.* (2001) The Semantic Web. *Sci. Am.* 284, 34–43
92 Ruttenberg, A. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinform.* 8 (Suppl. 3), S2
93 Slater, T. *et al.* (2008) Beyond data integration. *Drug Discov. Today* 13, 584–589
94 Cheung, K.H. *et al.* (2008) HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0. *J. Biomed. Inform.* 41, 694–705
95 Good, B.M. *et al.* (2006) iHOPerator: user-scripting a personalized bioinformatics web, starting with the iHOP website. *BMC Bioinform.* 7, 534
96 Edwards, A.M. *et al.* (2009) Open access chemical and clinical probes to support drug discovery. *Nat. Chem. Biol.* 5, 436–440
97 Barnes, M.R. *et al.* (2009) Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* 8, 701–708
98 Birney, E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816