Data reduction and representation in drug discovery

Trevor J. Howe, Guy Mahieu, Patrick Marichal, Tom Tabruyn and Pieter Vugts

Johnson & Johnson Pharmaceutical Research & Development, Janssen Pharmaceutica N.V., Turnhoutseweg 30, B-2340 Beerse, Belgium

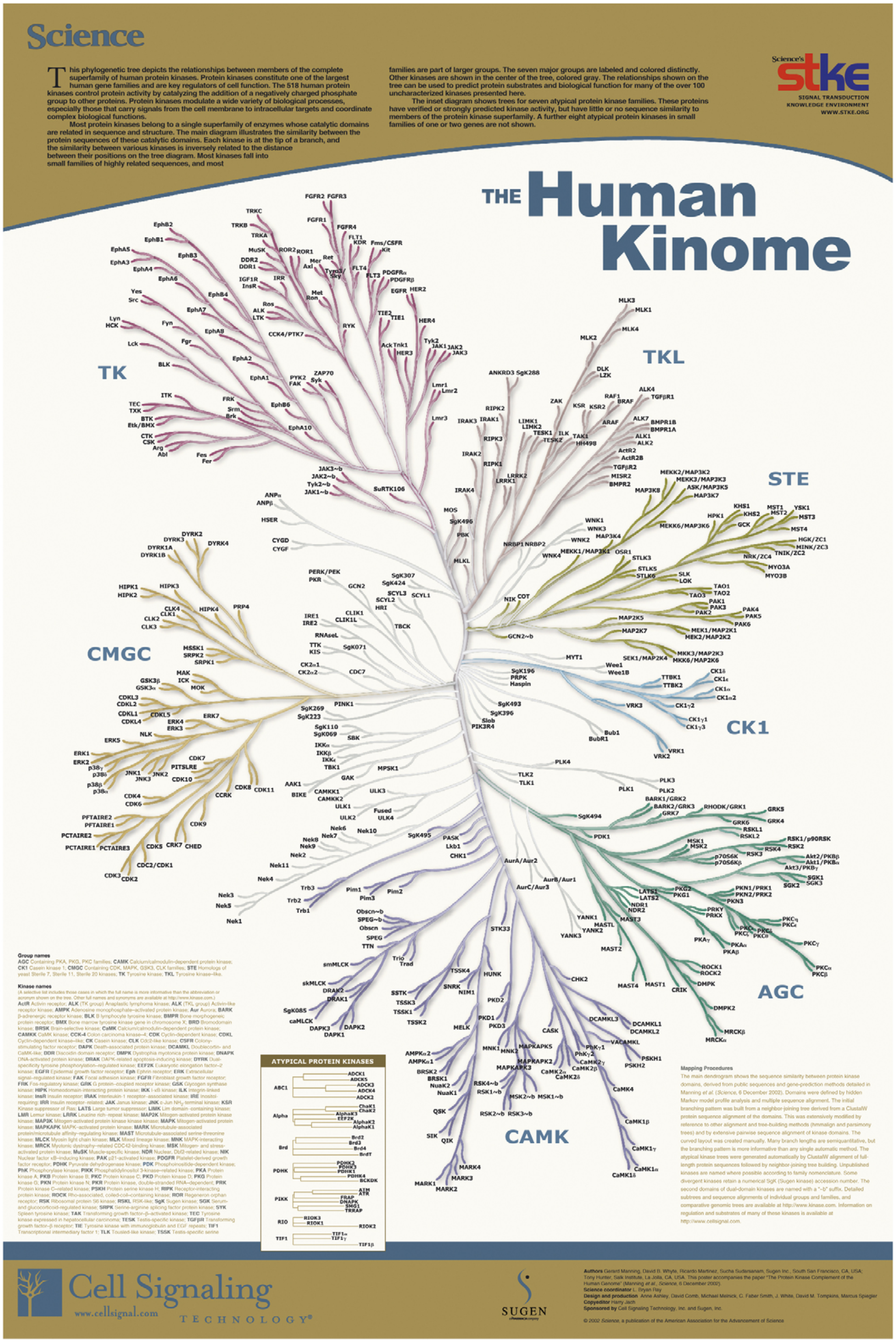
Pre-clinical drug discovery relies increasingly on huge volumes of inter-related multivariate data. To make sense of these data and enable quality decision-making based on this plethora of information they must be presented in an interpretable form. Reducing the dimensionality of the data often leaves a data set that is too complex to interpret readily, so intuitive visualization methods are needed. Bioinformatics has provided much of the impetus for visualizing complex data, the cheminformatics community has been aggressive with the data-reduction problem. The increasing appreciation of the inter-related multifactorial nature of pre-clinical drug discovery makes visualization a burgeoning and active field that spans biosciences, mathematics and visual psychology.

During the 1960s, drug discovery was driven largely by relatively restricted data sets. Single molecules were tested on low-throughput, *ex-vivo* biological assays and structure–activity relationships (SARs) often explained with a simple x–y plot. Since then, drug discovery has gone through many phases and been shaped by a variety of new technologies and external forces. On the technology side, molecular biology has contributed to a greater understanding of receptor pharmacology and, thus, the pursuit of highly target-specific drugs. In addition, high-throughput screening and the routine use of combinatorial libraries has greatly increased the volume and complexity of data in hit-identification stages. On the business side, there have many developments that attempt to compress development times. This has focused on improving the quality of pipelines and diminishing attrition rates through front-loading of drug-like characteristics. Consequently, multiple *in vitro* and *in silico* absorption, distribution, metabolism, excretion and toxicology (ADMET) assays are used. In addition, the increased regulatory demand for ultra-safe cardiovascular profiles has added further data volume and complexity through numerous human ether-a go-go related gene (hERG) assays. The drug discovery process itself has now gone full circle with so-called ‘phenotypic screening’ (screening compounds against either a ‘whole’ cell or organism to illicit a response) regaining popularity, complemented by functional genomics techniques to tease out likely targets for the compounds.

It is usual, therefore, to have highly complex data sets at any stage of the pre-clinical drug discovery process, from hit identification to the late stages of lead optimization. Intelligent interpretation of this plethora of data is the challenge. Take, for example, a hit-identification scenario where the starting point is a high-throughput screen. In such screens it is not untypical to test 500,000 molecules, and it is likely that the same set of molecules will be tested on several similar targets. Presuming (and this is a conservative presumption) tests on four related targets, will add an additional 2,000,000 data points. Easily calculated properties, such as rule-of-five type definitions, can add at least 12,500,000 further data points, and existing *in vitro* ADME, toxicity and cardiovascular profiles for many compounds will add 125,000,000 data points. Therefore a data set of some 140,000,000 points is constructed easily. In our group, we add additional indicators of cross-reactivity using Bayesian models of pharmacological activity. It can be argued that much of this data set is redundant because only a few thousand molecules will be sent for dose–response level testing and, moreover, such data sets are rarely exhaustive; incomplete matrices are common. Nevertheless, the complexity of the data set that results from even this set should not be underestimated. The primary pharmacological activity, historical data from 20 related biological targets, and ADMET assays, can result in a matrix of 100,000 data points from

Corresponding author: Howe, T.J. (thowe@prdbe.jnj.com)

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.10.014



This phylogenetic tree depicts the relationships between members of the complete superfamily of human protein kinases. Protein kinases constitute one of the largest human gene families and are key regulators of cell function. The 518 human protein kinases control protein activity by catalyzing the addition of a negatively charged phosphate group to other proteins. Protein kinases modulate a wide variety of biological processes, especially those that carry signals from the cell membrane to intracellular targets and coordinate complex biological functions.

Most protein kinases belong to a single superfamily of enzymes whose catalytic domains are related in sequence and structure. The main diagram illustrates the similarity between the protein sequences of these catalytic domains. Each kinase is at the tip of a branch, and the similarity between various kinases is inversely related to the distance between their positions on the tree diagram. Most kinases fall into small families of highly related sequences, and most families are part of larger groups. The seven major groups are labeled and colored distinctly. Other kinases are shown in the center of the tree, colored gray. The relationships shown on the tree can be used to predict protein substrates and biological function for many of the over 100 uncharacterized kinases presented here.

The inset diagram shows trees for seven atypical protein kinase families. These proteins have verified or strongly predicted kinase activity, but have little or no sequence similarity to members of the protein kinase superfamily. A further eight atypical protein kinases in small families of one or two genes are not shown.

Group names

AGC Containing PKA, PKG, PKC families; **CAMK** Calcium/calmodulin-dependent protein kinase; **CK1** Casein kinase 1; **CMGC** Containing CDK, MAPK, GSK3, CLK families; **STE** Homologs of yeast Sterile 7, Sterile 11, Sterile 20 kinases; **TK** Tyrosine kinase; **TKL** Tyrosine kinase–like.

Kinase names

(A selective list includes those cases in which the full name is more informative than the abbreviation or acronym shown on the tree. Other full names and synonyms are available at http://www.kinase.com.)

ActR Activin receptor; **ALK** (TK group) Anaplastic lymphoma kinase; **ALK** (TKL group) Activin-like receptor kinase; **AMPK** Adenosine monophosphate–activated protein kinase; **Aur** Aurora; **BARK** β-adrenergic receptor kinase; **BLK** B lymphocyte tyrosine kinase; **BMPR** Bone morphogeneic protein receptor; **BMX** Bone marrow tyrosine kinase gene in chromosome X; **BRD** Bromodomain kinase; **BRSK** Brain-selective kinase; **CaMK** Calcium/calmodulin-dependent protein kinase; **CAMKK** CaMK kinase; **CCK-4** Colon carcinoma kinase–4; **CDK** Cyclin-dependent kinase; **CDKL** Cyclin-dependent kinase–like; **CK** Casein kinase; **CLK** Cdc2-like kinase; **CSFR** Colony-stimulating factor receptor; **DAPK** Death-associated protein kinase; **DCAMKL** Doublecortin- and CaMK-like; **DDR** Discoidin domain receptor; **DMPK** Dystrophia myotonica protein kinase; **DNAPK** DNA-activated protein kinase; **DRAK** DAPK-related apoptosis-inducing kinase; **DYRK** Dual-specificity tyrosine phosphorylation–regulated kinase; **EEF2K** Eukaryotic elongation factor–2 kinase; **EGFR** Epidermal growth factor receptor; **Eph** Ephrin receptor; **ERK** Extracellular signal–regulated kinase; **FAK** Focal adhesion kinase; **FGFR** Fibroblast growth factor receptor; **FRK** Fos-regulatory kinase; **GRK** G protein–coupled receptor kinase; **GSK** Glycogen synthase kinase; **HIPK** Homeodomain-interacting protein kinase; **IKK** I-κB kinase; **ILK** Integrin-linked kinase; **InsR** Insulin receptor; **IRAK** Interleukin-1 receptor–associated kinase; **IRE** Inositol-requiring; **IRR** Insulin receptor–related; **JAK** Janus kinase; **JNK** c-Jun NH_2-terminal kinase; **KSR** Kinase suppressor of Ras; **LATS** Large tumor suppressor; **LIMK** Lim domain–containing kinase; **LMR** Lemur kinase; **LRRK** Leucine rich–repeat kinase; **MAP2K** Mitogen-activated protein kinase kinase; **MAP3K** Mitogen-activated protein kinase kinase kinase; **MAPK** Mitogen-activated protein kinase; **MAPKAPK** MAPK–activated protein kinase; **MARK** Microtubule-associated protein/microtubule affinity–regulating kinase; **MAST** Microtubule-associated serine-threonine kinase; **MLCK** Myosin light chain kinase; **MLK** Mixed lineage kinase; **MNK** MAPK-interacting kinase; **MRCK** Myotonic dystrophy–related CDC42-binding kinase; **MSK** Mitogen- and stress-activated protein kinase; **MuSK** Muscle-specific kinase; **NDR** Nuclear, Dbf2-related kinase; **NIK** Nuclear factor κB–inducing kinase; **PAK** p21-activated kinase; **PDGFR** Platelet-derived growth factor receptor; **PDHK** Pyruvate dehydrogenase kinase; **PDK** Phosphoinositide-dependent kinase; **PhK** Phosphorylase kinase; **PIKK** Phosphatidylinositol 3-kinase–related kinase; **PKA** Protein kinase A; **PKB** Protein kinase B; **PKC** Protein kinase C; **PKD** Protein kinase D; **PKG** Protein kinase G; **PKN** Protein kinase N; **PKR** Protein kinase, double-stranded RNA–dependent; **PRK** Protein kinase C–related kinase; **PSKH** Protein serine kinase H; **RIPK** Receptor-interacting protein kinase; **ROCK** Rho-associated, coiled-coil–containing kinase; **ROR** Regeneron orphan receptor; **RSK** Ribosomal protein S6 kinase; **RSKL** RSK-like; **SgK** Sugen kinase; **SGK** Serum- and glucocorticoid-regulated kinase; **SRPK** Serine-arginine splicing factor protein kinase; **SYK** Spleen tyrosine kinase; **TAK** Transforming growth factor–β–activated kinase; **TEC** Tyrosine kinase expressed in hepatocellular carcinoma; **TESK** Testis-specific kinase; **TGFβR** Transforming growth factor–β receptor; **TIE** Tyrosine kinase with immunoglobulin and EGF repeats; **TIF1** Transcriptional intermediary factor 1; **TLK** Tousled-like kinase; **TSSK** Testis-specific serine

Mapping Procedures

The main dendrogram shows the sequence similarity between protein kinase domains, derived from public sequences and gene-prediction methods detailed in Manning *et al.* (*Science*, 6 December 2002). Domains were defined by hidden Markov model profile analysis and multiple sequence alignment. The initial branching pattern was built from a neighbor-joining tree derived from a ClustalW protein sequence alignment of the domains. This was extensively modified by reference to other alignment and tree-building methods (hmmalign and parsimony trees) and by extensive pairwise sequence alignment of kinase domains. The curved layout was created manually. Many branch lengths are semiquantitative, but the branching pattern is more informative than any single automatic method. The atypical kinase trees were generated automatically by ClustalW alignment of full-length protein sequences followed by neighbor-joining tree building. Unpublished kinases are named where possible according to family nomenclature. Some divergent kinases retain a numerical SgK (Sugen kinase) accession number. The second domains of dual-domain kinases are named with a "~b" suffix. Detailed subtrees and sequence alignments of individual groups and families, and comparative genomic trees are available at http://www.kinase.com. Information on regulation and substrates of many of these kinases is available at http://www.cellsignal.com.

Authors Gerard Manning, David B. Whyte, Ricardo Martinez, Sucha Sudarsanam, Sugen Inc., South San Francisco, CA, USA; Tony Hunter, Salk Institute, La Jolla, CA, USA. This poster accompanies the paper "The Protein Kinase Complement of the Human Genome" (Manning *et al.*, *Science*, 6 December 2002).
Science coordinator L. Bryan Ray
Design and production Anne Ashley, David Comb, Michael Melnick, C. Faber Smith, J. White, David M. Tompkins, Marcus Spiegler
Copyeditor Harry Jach
Sponsored by Cell Signaling Technology, Inc. and Sugen, Inc.

© 2002 *Science*, a publication of the American Association for the Advancement of Science

FIGURE 1

Kinase dendogram. The methodology behind the image production of the kinase dendogram, which seeks to show the sequence and functional relationships of all human kinases, is contained in the captions. The dendrogram shows the sequence similarity between protein kinase domains that are derived from public sequences and gene-prediction methods detailed in Ref. [12]. Domains were defined by hidden Markov model profile analysis and multiple

5000 compounds. Add to this descriptors that characterize the compounds (in Johnson & Johnson PRD Beerse, we often use ~200 MolConnZ properties) and the scale of the data set increases to 20,000,000 data points. Again, this scale is deceptive as many of these properties will be cross-correlated and redundant; even so, the set will be sufficiently large to require data-reduction techniques to eliminate redundancy and define relevance to the SAR. It is clear that data reduction and data visualization are crucial to quality informed drug discovery.

A classic problem of data collection and interpretation on these scales can currently been seen in the US National Institutes of Health (NIH) drug-discovery initiative. This envisages screening compounds from an increasing combinatorial library against multiple screens in six screening centres [1]. The NIH compound library is expected to contain 3,000,000 compounds by 2009; it is easy to see that huge volumes of data will be generated and there is concern about what to do with these. Opinions stating that 'such activities will further add to the mountain of data' [2] and that the collection of data has 'outpaced the ability to interpret and apply' [3] are issues that data reduction and visualization seek to address.

Data reduction

Although there are methods to view high-density data and high-volume data, these are often meaningless unless some element of data reduction is applied. The human eye and mind cannot easily interpret more than three dimensions, four at best, and so some form of data reduction is required. The resulting information can then be visualized readily in various graphical forms that combine xy space, xyz space and combinations of these space dimensionalities with colour. Having mathematically and, perhaps, artificially reduced data to fewer dimensions for easy interpretation, mathematical treatments are then required to make meaningful comparisons of data structures and data types.

Data-reduction techniques and data-comparison techniques are, therefore, linked inextricably to visualization methods. Data in high dimensional space, which has to be reduced and reinterpreted in as few meaningful dimensions as possible, has been the province of the mathematical and statistical community for many hundreds of years. These techniques are modified frequently and applied to biosciences. Of particular note are the following methods: principle component analysis (PCA) [4]; projections to latent structures [5]; Bayesian analysis [6]; hierarchical clustering [7]; and similarity-measuring techniques such as Tanimoto [8].

Principle component analysis is a variation of factor analysis, the invention of which is credited largely to Charles Spearman. He was a British Army officer (more of the British Army later) who, late in life, studied to become an experimental psychologist. To help analyse his results, he developed factor analysis [9], publishing his seminal work in 1904. He was interested in formulating measures of intelligence for which he needed to explain observed random variables in terms of a smaller number of unobserved random variables or 'factors'. The observed variables are described by linear combinations of the factors. Thus, many variables can be described by very few factors. This idea was elaborated by Kari Karhunen and Michel Loèveto to become PCA in the form now used in biosciences [10]. A PCA is a mathematical transformation of a data set to new coordinate system where the greatest variance is described in the first axis (the first principal component), the second greatest variance on the second axis, and so on. Each principal component contains several factors and weightings that are not necessarily precluded in other principal components. Ideally, all of the variance is described in the first two or three principal components. This enables the familiar visualisation of a PCA in either two- or three-dimensional scatter plots, each axis representing a principal component. In deriving quantitative structure–activity relationship equations (QSARs) it is often a problem to identify the limited number of (relevant) descriptors needed to build a valid QSAR equation from the thousands of possible descriptors. Therefore, a PCA is often used to identify highly weighted descriptors that contribute to the top principal components that explain the maximum variance in the PCA. These desciptors are then used in the creation of QSAR equations.

As the statistical literature is used increasingly by biosciences to help reduce data size and dimensionality, so the methodology of the maths becomes linked inextricably to its graphical representation. One such technique and its representation is hierarchical clustering, shown in Figures 1 and 2.

Kinase dendogram and hierarchical cluster

Clustering is a means of classifying similar objects, for example, chemical structures [11], into groups that share some common features or traits. Usually, the clusters and the partitions between them are defined by some form of distance measure. A classic example in biological clustering is the well-known dendogram of kinase sequence similarity, produced by Cell Signalling Technology and Sugen [12]. A non-hierarchical variation of this method that is used often in chemistry is k-means clustering [13] in which all the points in the cluster for each dimension are averaged and the description of the variation is from the centroid of each cluster.

An exhaustive list of statistical techniques that directly link their intrinsic method to the representation they produce is not possible here, but clustering methods illustrate beautifully the link between the underlying mathematics and the representation. Many methods are in use that contain, in some way, the essential elements of the data linked and presented in a graphically digestible form for ease of interpretation.

Visualization

Perhaps unexpectedly, one of the earliest pioneers of data visualization in medical research was Florence Nightingale. Although she is well known for her contribution to improving nursing in battlefield conditions, she is less known for how she conveyed this

sequence alignment. The initial branching pattern was built from a neighbour-joining tree derived from a ClustalW protein sequence alignment of the domains. This was modified extensively by reference to other alignment and tree-building methods (hmmalign and parsimony trees) and by extensive, pair-wise sequence alignment of kinase domains. The curved layout was created manually. Many branch lengths are semi-quantitative, but the branching pattern is more informative than any single automatic method. The atypical kinase trees were generated automatically by ClustalW alignment of full-length protein sequences followed by neighbour joining tree building. Unpublished kinases are named where possible according to family nomenclature. Detailed subtrees and sequence alignments of individual groups and families, and comparative genomic trees are available at kinase.com (http://www.kinase.com). Information on the regulation and substrates of many of these kinases is available at Cell Signalling Technology (http://www.cellsignal.com).

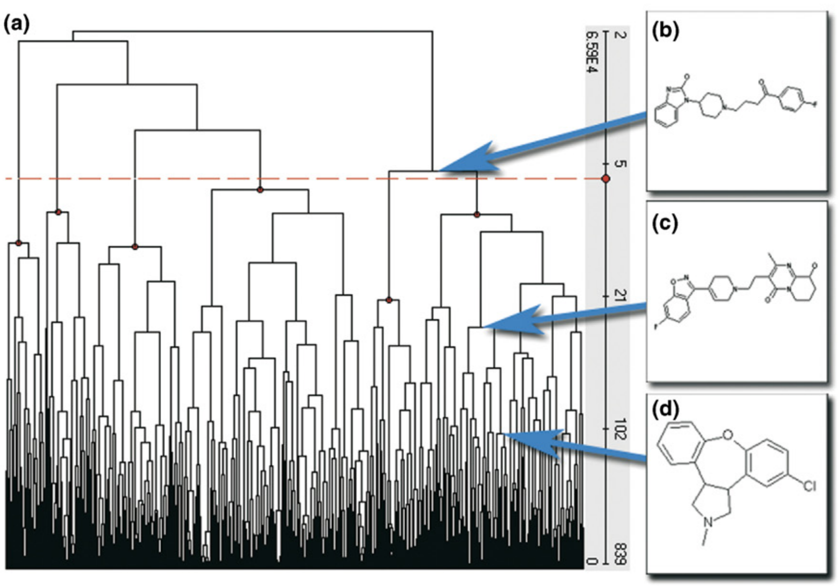
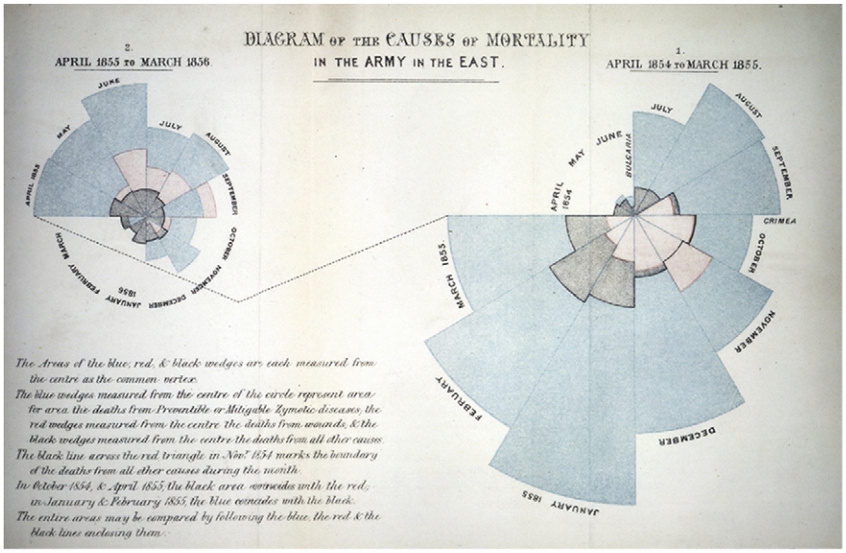


FIGURE 2

Chemical clustering of 839 dopamine D2 receptor antagonists using Ward's non-hierarchical clustering technique. The molecules are described by MDL keys and Kier-Hall indices, **(a)**. Chemical structures of the centroids at three levels on the dendogram are shown in **(b)**, **(c)** and **(d)**. As the clusters become deeper, more divergent molecules are described as the centroids of smaller clusters. This can be seen in (d), where a tricyclic structure is the centroid of a cluster. Nearer the top of the dendorgam, this compound would not be sufficiently diverse to represent all elements of a cluster such that it could be described as a common ancestor to all other structures.

message to the Army authorities. The dogma of the time was that soldiers were dying because they were soldiers; that is what happens to soldiers in battle. In fact, battlefield casualties were small compared to non-battlefield casualties. In her 1858 publication, *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army* [14], she invented a graphical illustration to show the difference between battlefield and non-battlefield deaths. Such was her success with these representations, which she



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

FIGURE 3

Florence Nightingale's diagram of the causes of mortality in the army in the east produced in 1858. To convince the British Army authorities that improvements in hygiene conditions in its own hospitals were beneficial, she illustrated the causes of sickness and death during the Crimean war in diagrams she called coxcombs. Each segment represents a month of data. The Coxcombs show that most of the British soldiers who died during the Crimean War died of sickness (blue), rather than of wounds or other causes (red or black). This novel presentation of data supported her case for better hygiene by using published Army figures to show that the death rate decreased after the Sanitary Commissioners arrived in March 1855 as they sought to improve conditions in the hospitals. Image reproduced by courtesy of the Florence Nightingale Museum Trust, London.

called coxcombs (Figure 3), that she was elected to the Statistical Society of England.

At Johnson and Johnson in Beerse, we have independently developed a similar idea implanted in a software tool called VlaaiVis (Box 1), a name derived from the Flemish for pie (Vlaai) and an abbreviation of visualization (Vis). The intention of this name is to distinguish the concept from classical pie charts. Data are represented in slices and normalized to simplify interpretation. This is used to display multifactorial data at all stages of drug discovery, from hit identification through to late-stage lead

BOX 1

Visualizing data using VlaaiVis

The colours in VlaaiVis are standardized for ease of interpretation. In the example shown in Figure I, the red wedges (toxicity and cardiovascular safety assays) are all near the edge, inside the grey, outer rim that defines the upper and lower bounds of the TPP. The first eight wedges (counting clockwise) of the green segment (ADME data) comply with the TPP although the third green wedge is hashed, which indicates a data error. The last three wedges are near the origin of the circle, which is far from the preferred range. They are either too high or too low and some ADME issues need to be addressed here. The grey segment (*in silico* parameters) complies with the TPP, with only one wedge out of the preferred range. The width of the wedges can be controlled so that the 'weight' of factor in the SAR can be indicated using a thin wedge for less important factors. The blue values (primary pharmacological activity) are more dispersed. In this case *in vivo* (dark blue) and *in vitro* (pale blue) data are included, showing a clear opportunity for optimization. White wedges indicate missing values.

Complete data set representation

When investigating an entire data set in VlaaiVis, as many circles/ compounds as desired can be displayed (Figure II). The user can also easily drop or add segments and wedges to focus interpretation, and ether relax or tighten the TPP if the focus of the project changes. A 'percentage filled' (comparable to a score) is calculated on-the-fly to give a crude idea of how well a compound adheres to the TPP. Other information is also provided in the interface. The compound structure is available as the mouse is dragged across a circle and the full assay data are also provided as a bar chart for more detailed analysis. By clicking on the segment colours at the bottom of the application, any or all of the segments can be easily dropped and added. Also, right-clicking on the mouse while over an individual wedge removes (or adds) these to ease interpretation.

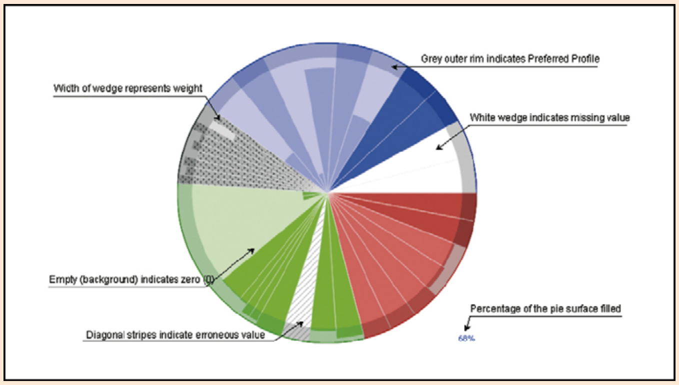
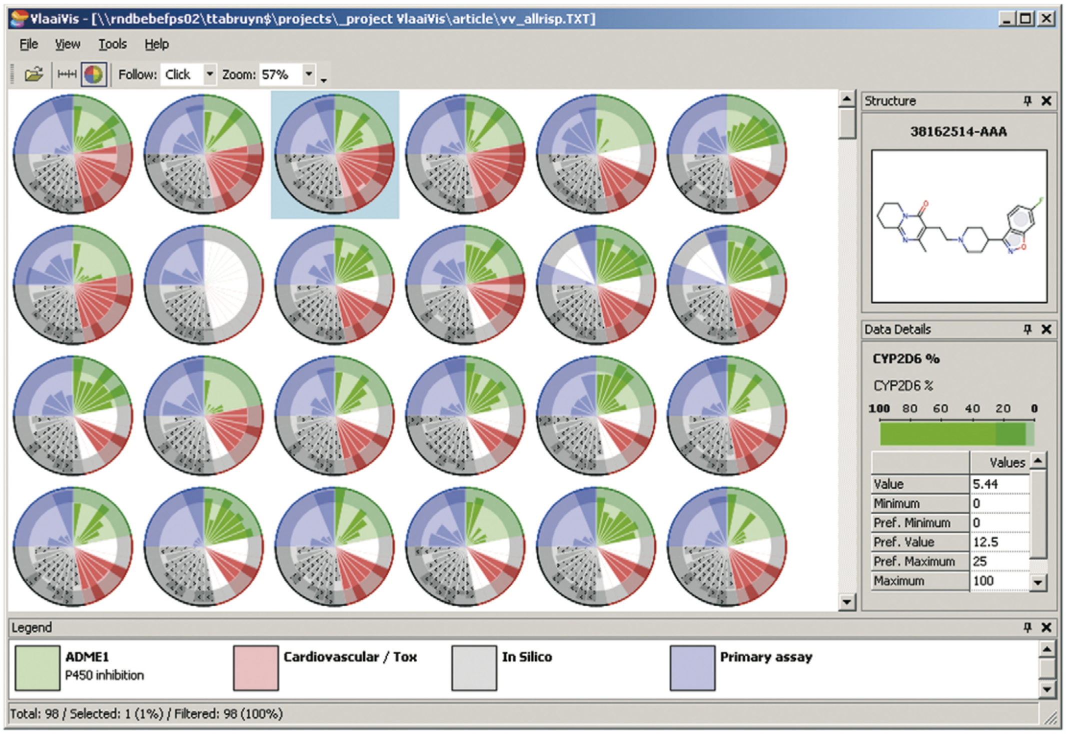


FIGURE I



VlaaiVis - [\\rndbebefps02\ttabruyn$\projects_project VlaaiVis\article\vv_allrisp.TXT]

38162514-AAA

CYP2D6 %

CYP2D6 %

100 80 60 40 20 0

	Values
Value	5.44
Minimum	0
Pref. Minimum	0
Pref. Value	12.5
Pref. Maximum	25
Maximum	100

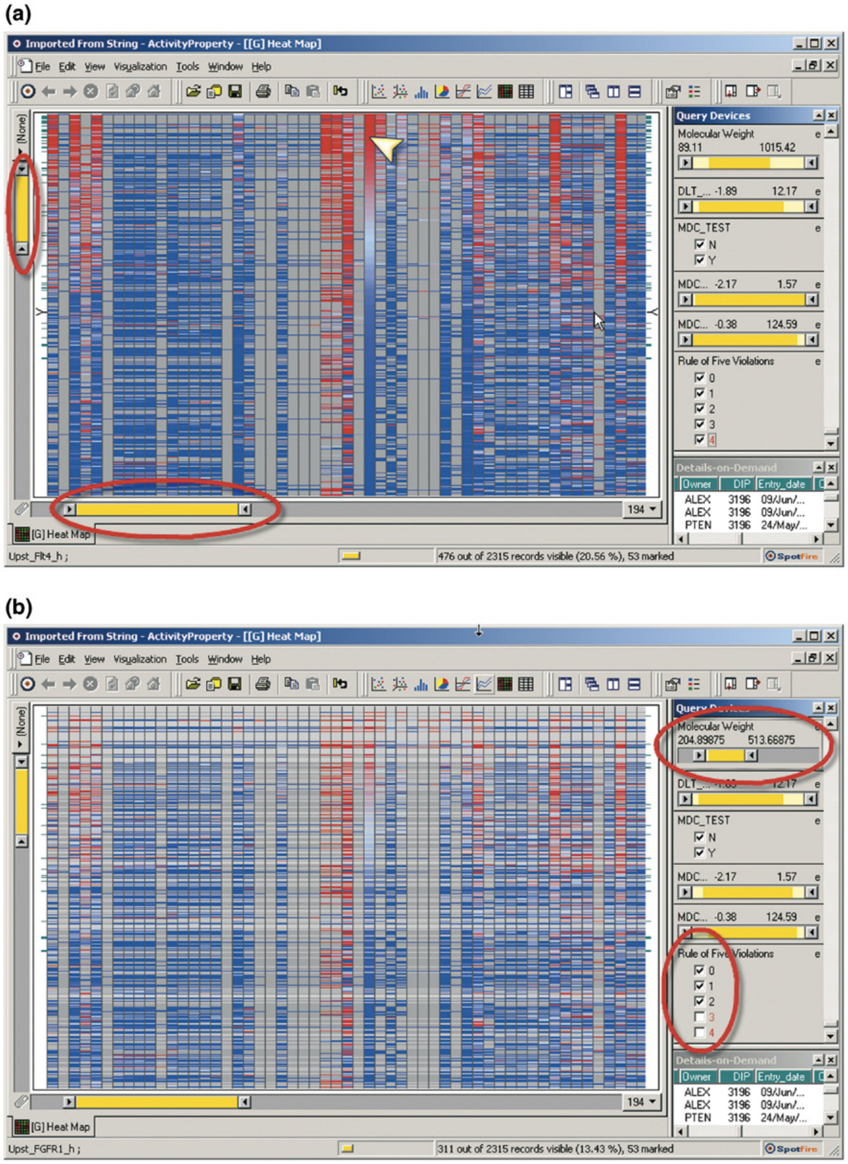
🟩 ADME1 P450 inhibition
🟥 Cardiovascular / Tox
⬜ In Silico
🟦 Primary assay

Total: 98 / Selected: 1 (1%) / Filtered: 98 (100%)

FIGURE II

optimization. The driver for this is that a drug-discovery team can often be blind to obvious trends in the data when interpreting large compound data sets that contain a variety of *in vivo*, *in vitro* and *in silico* tests. There might be several reasons for this, including, for example, personal bias toward a particular chemistry, and concentrating overly on a particular assay. These 'gut feelings' although often invaluable, can lead to unexplored SAR paths whereas others are either over-emphasized or ignored. Through VlaaiVis, the programme team is presented with a data visualization that stimulates thinking about the development of an SAR in the context of all of the data. Although not a substitute for detailed, multivariate, QSAR and SAR analyses the visualization is easily interpreted and provides a basis for data-driven analysis.

VlaaiVis is a simple concept, basically a compound-based normalized histogram of responses that are depicted in a circle. The initial development began using classic histograms with preferred



(a) Imported From String - ActivityProperty - [[G] Heat Map]

Query Devices

Molecular Weight 89.11 1015.42

DLT_... -1.89 12.17

MDC_TEST ☑ N ☑ Y

MDC... -2.17 1.57

MDC... -0.38 124.59

Rule of Five Violations ☑ 0 ☑ 1 ☑ 2 ☑ 3 ☑ 4

Details-on-Demand

Owner	DIP	Entry_date
ALEX	3196	09/Jun/...
ALEX	3196	09/Jun/...
PTEN	3196	24/May/...

Upst_Flt4_h ; 476 out of 2315 records visible (20.56 %), 53 marked

(b) Imported From String - ActivityProperty - [[G] Heat Map]

Query Devices

Molecular Weight 204.89875 513.66875

DLT_... 12.17

MDC_TEST ☑ N ☑ Y

MDC... -2.17 1.57

MDC... -0.38 124.59

Rule of Five Violations ☑ 0 ☑ 1 ☑ 2 □ 3 □ 4

Details-on-Demand

Owner	DIP	Entry_date
ALEX	3196	09/Jun/...
ALEX	3196	09/Jun/...
PTEN	3196	24/May/...

Upst_FGFR1_h ; 311 out of 2315 records visible (13.43 %), 53 marked

FIGURE 4

Heatmaps produced in Spotfire. Heatmaps are used increasingly commonly to show multidimensional data. **(a)** In this case, coarse-grained activity of compounds against a series of targets is shown whereby compounds are ordered by a hierarchical clustering technique on the y-axis. Various protein assay targets are on the x-axis. The probable activity of a compound, determined by a Bayesian model, is entered as a colour with data binned in a binary fashion, red for active and blue for inactive. Sorting the data set by the middle assay data (column indicated by the arrow) shows a correlation between this target protein and columns 1, 3 and 5 plus a few more near the right-hand side of the heatmap. The circled slider bars next to and underneath the heatmap indicate that only a fraction of the data are in the viewport. **(b)** Spotfire can be used to manipulate data through the use of slider bars. The same data set is shown using the sliders. The right-hand sliders are used to hide (greyed out) the rows that fall outside the molecular-weight range of 204–513, and have ≥three rule-of-five violations [26]. Few of the compounds are lost from the view, which indicates that this data set contains compounds that are all drug-like according to Lipinski's rules. The small panel at the bottom right contains the details-on-demand. These show the actual column values of the row selected in the heat map if the user wants to examine a compound in detail.

minimum and maximum values. Assay and descriptor responses were interpreted as either overshooting or undershooting the desired range. However, once the number of 'wedges' grew beyond three of four, the visualization became too complex for intuitive interpretation. To circumvent this problem, we normalized the data against a defined target product profile (TPP). A TPP is a series of maximum and minimum defined values for each of the assays under consideration, with the optimal value always ascribed to the edge of the circle, following the concept of the fuller the better. This makes interpretation simple: by looking for a completely filled circle, there is no necessity for detailed understanding of either the chemistry or even the exact assay data. The detailed data are always available underneath the VlaaiVis visualisation interface as a spreadsheet if this is needed, but, in essence, the application can be used for naïve interpretation and communication of strategic decisions simply through the normalised pie chart display. At any point in the cycle of the project, the TPP can be adjusted, and more assays included and some removed. The requirements of the TPP are likely to be more relaxed during the hit-identification stage, with fewer assays considered. As the project progress through hit-to-lead to lead optimization, the assay requirements might be more stringent and the number of assays (i.e. wedges) is also likely to increase.

The biology driven sciences have, perhaps, made more progress in visualising massive data sets than the medicinal chemistry community have thus far demonstrated in their use of visual techniques in hit identification and hit to lead phases. Heat maps, for example, which are common across all fields of drug discovery (Figure 4) were popularized first in bioinformatics. The NCI-60 project has provided a particularly rich source for visualization of this type. Using cDNA microarrays [15], the variation in expression of 8000 genes in 60 cancer-cell lines (from brain, blood and bone marrow, breast, colon, kidney, lung, ovary, prostate and skin) has been explored and described by a series of heat maps. In addition, since the inception of these cell lines in 1992, there have been many chemical screening campaigns and the richness of the data generated by such studies has stimulated chemical visualization as illustrated by Covell's SAR study on *in vivo* and *in vitro* activities [16].

That the most novel and spectacular innovations in visualization have come from biology-driven science is perhaps not surprising given the impetus provided by the publication of the human genome [17]. The genomes of many other animals are now available at the European Bioinformatics Institute (http://www.ebi.ac.uk/genomes/), which archives and provides access to all publicly available genomes. Combined with various proteomic initiatives [18], this has stimulated interest in connecting networks of genes and proteins. A beautiful series of images across a range of sciences has been collected by Manuel Lima on the visualcomplexity site (http://www.visualcomplexity.com). The idea of visualcomplexity is to provide a unified resource for the visualization of complex networks in any technology sphere. As such, bioscientists can see the type of visualizations that are applied to other sciences and *vice versa*. A fine example comes from researchers at The Institute for Cellular and Molecular Biology at the University of Texas (http://bioinformatics.icmb.utexas.edu/lgl/#gallery) who have developed an application called Large Graph Layout [19]. This is a compendium of applications for making the visualization of large networks and trees tractable, and was motivated by the need to make the visualization and exploration of large biological networks more accessible. Figure 5 shows a protein homology network in which the results of ~92 billion pair-wise amino acid-sequence alignments between 289,069 proteins from 90 genomes are used.

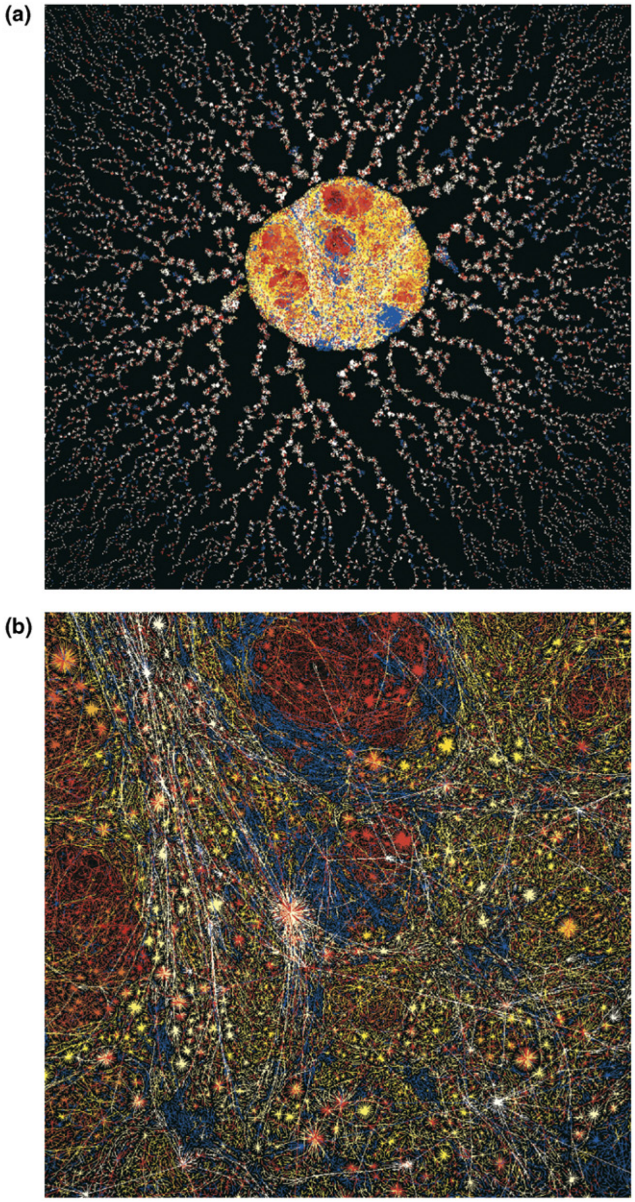


FIGURE 5

Protein homology network by Large Graph Layout (LGL) from the Bioinformatics, Proteomics and Functional Genomics Department of the University of Texas at Austin. The relationships between proteins and the homologies between proteins sequences are often key considerations in the validation of drug targets. Understanding these relationships across the proteome is a huge visual task. **(a)** This image shows an entire, massive, multi-species, protein homology network. **(b)** A detail from the centre of the network. The network summarizes the results of ~92 billion pair-wise, amino acid-sequence alignments between 289,069 proteins from 90 genomes. The final network is composed of 27,325 connected sets summing to 7,940,873 edges. An edge is coloured blue if it connects two proteins from the same species, and red if it connects two proteins from two different species. If that information is not available the edges are coloured based on layout hierarchy. Reproduced with permission from Alex Adai.

The vertices are proteins and an edge is defined between any two proteins that have a BLAST [20] e-value of $<10^{-12}$. An edge is coloured blue if the protein is close in homology to a protein from the same species and coloured red if it is close in homology to two different species. If no information is available, the edges are coloured based on layout hierarchy.

Perhaps the most common tool in the pharmaceutical industry for data visualization is Spotfire (Spotfire AB. SE-413 28 Göteborg, Sweden). Although a simple concept because few graphical solutions are provided, the sliding scales, colour categorization, picking capabilities and some chemical intelligence make it a powerful data manipulation and representation tool [21] Some pharma companies extend this use further, and complete chemical and biological databases are handled entirely within Spotfire's framework.

Spotfire's foundations [22,23] were laid during the early 1990s through collaboration between Christopher Ahlberg and Professor Ben Shneiderman of the Human–Computer Interaction Laboratory at the University of Maryland. The company website provides a history (http://www.spotfire.com/about/company_history.cfm) as does the Human Computer Interaction Laboratory (http://www.cs.umd.edu/hcil/spotfire/) in an excellent article 'Dynamic queries, starfield displays, and the path to Spotfire'. Going beyond pure representation and focusing on dynamic interfaces that enabled direct manipulation of the data, they introduced the now familiar concept of data sliders that enable rapid human interaction with the data. The first application of this idea enabled the user to change the rendering of a polynomial curve by moving sliders that represented its coefficients. FilmFinder, a later project by the same group, featured a scatter plot with dots representing films. The dots were colour- coded to match certain film genres (drama, action and comedy), the details of which pop up on-demand when a user clicks a data spot.

Ahlberg took these first steps further into his Information Visualization and Exploration Environment (IVEE), which gathered the explored ideas into a working solution for data interaction capable of handling flat files for import. In 1996, venture capital was raised, and IVEE turned into the commercial product now known as Spotfire. It quickly took hold in the discovery pharmaceutical field because of its visual data-mining capabilities. The concept of dynamic queries also made its way into other data-driven domains, such as website log visualizations, social studies [24] and query previews (http://www.cs.umd.edu/hcil/eosdis/) [25] for the 'Earth Observing System Data and Information System' of NASA. The sliders, filters, scatter plots, histograms and table views are all interlinked and provide different views of the same data. Changing the selection or filtering the data set in one view is reflected immediately in all other views. This opens up the concept of visual data exploration and stimulates thought processes that were unavailable previously from data in standard static graphical forms.

Spotfire facilitates the production of heat maps. A compound activity heat map as used in Johnson and Johnson PRD in Beerse is shown in Figure 4. Compounds are ordered by a hierarchical clustering technique on the y-axis. Various protein assay targets are on the x-axis. The probable activity of a compound, determined by a Bayesian model, is entered as a colour with data binned in a binary fashion (red for active and blue for inactive). These data sets can be filtered with the sliders, as also illustrated in Figure 4, where Lipinski 'rule of five' filters [26] are incorporated.

Conclusions

Drug discovery is not the only discipline in which interpretation and understanding is aided by simplification. Cartography learnt many of these lessons over 70 years ago during the 1930s. It took an electrical engineer, Harry Beck, to create the London Underground map [27]. Ignoring the cartographic convention of the day, which demanded accurate distance and exact geographic locations, he followed the wiring diagram rules he knew as an engineer and turned the most complex underground network in the world into a simple stylized image (http://www.tfl.gov.uk/tfl/pdfdocs/colourmap.pdf) that is still in use today and is the model for underground systems worldwide. In 1936, Phyllis Pearsall, frustrated at the difficulty in trying to locate a house in Mayfair where she was to attend a party, set about redrawing the London street map. Again, against cartography convention she drew the width of the streets far larger than their scale size, introduced an accurate index and designed a clear typeface to fit the streets. This map, which interprets and represents clearly the information needed to navigate the streets, was turned down by cartography companies as it flew against their usual representations. To publish her new map, the *A to Z of London* [28] she founded the Geographers' Map Company Ltd and was involved with the company until her death, aged 89, in 1996. The *A to Z of London* is still the defining road-navigation map of this highly complex city and the rules for mapping cities the world over were changed by a simplified representation. Drug discovery is beginning to wake up to these same opportunities and the variety of ways networks are described in visualcomplexity.com provides a 21st-century vision of what was achieved by chance in cartography during the 1930s. As we begin to ask questions of how we look at data, our interpretation of this data will change and we will be all the richer and more efficient through the new insights that data reduction, simplification and presentation will provide.

References

1 Couzin, J. (2003) Molecular medicine. NIH dives into drug discovery. *Science* 302, 218–221

2 Wiliams, Michael (2004) A return to the fundamentals of drug discovery? *Curr. Opin. Investig. Drugs* 5, 29–33

3 Lindpaintner, K. (2002) The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nat. Rev. Drug Discov.* 1, 463–469

4 Anton, H. (2000) *Elementary Linear Algebra. 8th Edition.* John Wiley, New York, U. S. A.

5 Eriksson, L. *et al.* (2001) *Multi- & Megavariate Data Analysis*. Umetrics Ab, Sweden

6 Berry, D.A. (1996) *Statistics: A Bayesian Perspective*. Duxbury Press, California, U. S. A.

7 Johnson, S.C. (1967) Hierarchical clustering schemes. *Psychometrika* 2, 241–254

8 Willett, P. *et al.* (1986) Implementation of nearest-neighbour searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* 26, 36–41

9 Spearman, C.E. (1904) Proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101

10 Jackson, J.E. (1991) *A Users Guide to Principal Components*. John Wiley & Sons

11 Willett, P. (1996) Molecular diversity techniques for chemical databases. *Inf. Res.* 2 (3), http://informationr.net/ir/2-3/paper19.html

12 Manning, G. *et al.* (2002) The protein kinase complement of the human genome. *Science* 298, 1912–1934

13 Willett, P. *et al.* (2004) Clustering files of chemical structures using the fuzzy k-means clustering method. *J. Chem. Inf. Comput. Sci.* 44, 894–902

14 Nightingale, F. (1858) *Notes on Matters Affecting the Health. Efficiency and Hospital Administration of the British Army.* Harrison and Sons

15 Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24, 227–234

16 Covell, D.G. *et al.* (2006) Assessment of *in vitro* and *in vivo* activities in the national cancer institute's anticancer screen with respect to chemical structure, target specificity, and mechanism of action. *J. Med. Chem.* 49, 1964–1979

17 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

18 Hancock, W.S. (2002) The challenges ahead. *J. Proteome Res.* 1, 9

19 Adai, A.T. *et al.* (2004) LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol.* 340, 179–190

20 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

21 Brodfuehrer, J. *et al.* (2004) The Implementation of an ADME enabling selection & visualisation tool for drug discovery. *J. Pharm. Sci.* 93, 1131–1141

22 Ahlberg, C. *et al.* (1992) Dynamic queries for information exploration: an implementation and evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Monterey, CA, USA* (Bauserfield, P. *et al.* eds), In pp. 619–626, ACM Press

23 Ahlberg, C. *et al.* (1994) Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Proceedings of the ACM CHI 94 Human Factors in Computing Systems Conference, Boston, MA, USA* (Adelson, B. *et al.* eds), In pp. 313–317, ACM Press

24 Ellis, J. *et al.* (1997) Putting visualization to work: ProgramFinder for youth placement. *Proc. of ACM CHI* 97, 502–509

25 Plaisant, C. *et al.* (1997) Query previews in networked information systems: the case of EOSDIS. In *Conference on Human Factors in Computing Systems; CHI '97 Extended Abstracts on Human factors in Computing Systems: Looking to the Future, Atlanta, GA, USA* (Edwards, A. and Pemberton, S., eds), pp. 202–203, ACM Press

26 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.* 23, 3–25

27 Garland, K. (1994) *Mr Beck's Underground Map.* Capital Transport Publishing

28 Pearsall, P. (1990) *From Bedsitter to Household Name.* Geographers A to Z Map Company

Reviews • INFORMATICS