Visual And Computational Analysis Of Structure-Activity Relationships In High-Throughput Screening Data

Peter Gedeck * and Peter Willett †

* Novartis Respiratory Research Centre, Novartis Pharmaceuticals UK Ltd., Wimblehurst Road, Horsham, West Sussex, RH12 5AB, United Kingdom

e-mail: peter.gedeck@pharma.novartis.com

phone: +44-1403-32 30 51

fax: +44-1403-32 33 07

† Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom

e-mail: p.willett@sheffield.ac.uk

phone: +44-114-22 22 633 / 630

fax: +44-114-27 80 300

Summary of Recent Advances

Novel analytic methods are required to assimilate the large volumes of structural and bioassay data generated by combinatorial chemistry and high-throughput screening programmes in the pharmaceutical and agrochemical industries. This paper reviews recent work in visualisation and data mining that can be used to develop structure-activity relationships from such chemical/biological datasets.

Keywords

Data analysis, data visualisation, data mining, drug-likeness, classification,

Introduction

The search for lead compounds in the pharmaceutical industry (and also in the agrochemical and related industries) has historically followed an inherently sequential process, in which individual compounds are synthesised and then tested for biological activity, with the results of such experiments being fed back to inform the selection of further molecules. Developments in combinatorial chemistry [1-5] and in high-throughput screening (HTS) [6-8] mean that such operations have been largely replaced by a massively parallel mode of processing, in which many thousands of molecules can be synthesised and tested at the same time. This has resulted in an explosion in the volume of data that is available for the identification of structure-activity relationships (SAR). Quantitative SARs, typically using physicochemical parameters or 3D molecular fields with statistical techniques such as multiple regression or principal components analysis, have been an important tool for medicinal chemists for many years. However, such methods are normally used for the detailed analysis of small numbers of structurally-related molecules, and are not applicable to the large, structurally heterogeneous datasets that characterise modern HTS systems. There is hence much current interest in novel soft-computing approaches that might be applicable to the analysis of such datasets, and we here review recent work in this area, focusing upon the use of visualisation and data mining techniques.

Visualisation Techniques

Information and data visualisation plays an important role in practically all areas of scientific research. Consequently, many visualisation techniques like scatter plots or histograms have been developed [9,10]. That these established, simple visualisation techniques can help to identify patterns also in large datasets is demonstrated nicely in an article by Hand *et al.* [11*]. However, very interesting developments have been made in information and data visualisation over the last years that are especially aimed at large datasets [12**,13]. A thorough introduction to this important field was recently published by Card *et al.* [12**], who provide a collection of important classic and cutting edge articles in this field.

With the increasing importance of high-throughput chemistry and screening, the consequent increase in data volume [14] requires more effective methods to visualise and structure data produced in research. In addition, the emerging consolidation of research data in chemical data warehouses [15] makes it now more feasible to mine these sources. However, only few applications in chemistry have appeared over the past years. It can however be expected that visualisation will have a major impact on drug discovery over the next years [16].

Two general purpose data visualisation programs will act as examples of what is currently possible. Spotfire [17,18] (Spotfire® is a product of Spotfire Inc., Cambridge, MA, USA) is probably one of the best-known data visualisation and mining programs. Although it only provides basic graphs like scatter-plots, histograms and pie charts, its special features like database connectivity and interactive query devices make it a powerful tool for interactive visualisation and information analysis (see Figure 1). Any change of the control elements is instantly executed and the user gets an immediate feedback. Another interesting visualisation program that has been applied to pharmaceutical research data is OmniViz Pro™ (OmniViz Pro is a trademark of OmniViz, Inc., Columbus, OH, USA). In contrast to Spotfire, this software is based on new methods for information visualisation developed by the *Information Visualisation* group of the *Pacific Northwest National Laboratory* (see http://multimedia.pnl.gov:2080/infoviz/). Figure 2 gives an example of the type of graphs that can be created with OmniViz Pro.

In many cases, the available data are multidimensional. This is especially true for chemical compounds that can numerically be represented either by fingerprints or a set of descriptor values. In order to explore multidimensional data, it is necessary to map the data points into a 2- or 3-dimensional space. This mapping is frequently called non-linear mapping. The aim of non-linear mapping is in most cases to preserve neighbouring properties, so that data points that are close together in the multidimensional space will be close together in the low-dimensional space. A variety of methods have been used over the last years in chemistry for the visualisation of databases [19], in diversity analysis [20,21] and for the analysis of structure-activity relationships [22,23].

An established method for non-linear mapping is multidimensional scaling [24]. Principal component analysis or singular value decomposition can be used to get an initial estimate for the low dimensional representation of data points. In a second step, this projection is improved by optimising the data point separation in the low-dimensional space so that they resemble the distances of the data points in the high-dimensional space better. The quality of the mapping is measured by using the Sammon's stress function or a variation thereof. Clark *et al.* [25*] proposed a particularly interesting modification of the original stress function. They observed that when mapping chemical structures based on fingerprints, the local similarity is not well preserved. This is due to the fact that the similarity measure used is an insufficient measure of dissimilarity. The suggested modification of the stress function ignores contributions of compounds with similarity smaller than a given value. The mapping obtained with the modified version clearly shows a more pronounced clustering of similar compounds.

Unfortunately, multidimensional scaling is not well suited for large datasets, as the method scales quadratically with the number of data points. A significant improvement compared to conventional methods was achieved by Xie *et al.* [26], who applied the truncated-Newton optimisation method to improve the initial mapping obtained from singular value decomposition. They were able to demonstrate that the truncated-Newton optimisation can be up to 100 times faster than using the steepest descent method for optimisation. The approach is however not suitable for sets of several thousand data points.

A variety of different approaches were used to apply neural nets for non-linear mapping. The advantage of neural nets is that they can be used to predict positions of new data points in the low-dimensional space. A number of studies have used self-organising maps [27*] to visualise and analyse the diversity of databases [19,22,23]. It is also possible to use a multi-layer back-propagation neural net with n input and m output neurons (m=2,3). The output of the neural net for each output can be used as its m lower-dimensional co-ordinates [28]. Izrailev and Agrafiotis modified this method [29*]. Instead of using the full dataset, they suggest to train a feed-forward neural network to learn the projection obtained from conventional non-linear mapping of a subset of all data. The trained network can subsequently be used to project the whole compounds set. In an example of a combinatorial

library containing 57498 compounds, a subset of only 100 compounds (0.2 %) was already sufficient to generate a reasonable map of the whole dataset. Another approach uses a neural net with *n* neurons in the input and output layer and several middle layers. One of the hidden layers has *m* neurons. The net is trained to reproduce the input variables at the output neurons. The reduced dimensionality representation of a compound can then be read out from the *m* neurons of the middle layer [30].

While the described non-linear mapping techniques try to preserve the neighbouring relationship, the generated map might not necessarily be the best mapping if the aim is to visualise a classification. The classification mapping methods proposed by Su *et al.* [31*] aim to achieve this. The examples demonstrate that the techniques are able to give a qualitative or semi-quantitative picture.

Probably the most important problem during lead optimisation in drug discovery is to determine SAR information. While it is quite feasible to develop this SAR knowledge for small numbers of compounds manually, it is necessary to automate this process for large datasets. The aim is to identify sets of similar compounds that have a common structure and show a systematic variation in one part (e.g. a substituent, a spacer or a ring system). Therefore, it may be interesting to compare one particular compound to a variety of other compounds in the dataset, to visualise the common features, and hence the potential pharmacophore patterns, that are present.

Sheridan and Miller looked at recurrent topological substructures [32*]. They compare the structures of pairs of compounds and determine all common, possibly disconnected substructures. These substructures are scored and the highest scoring common substructure determined. This approach allows identifying 2D pharmacophores for a set of compounds. Another approach based on maximum common substructures is used by Distill (Distill is a trademark of Tripos Inc., St. Louis, MO, USA). This program develops a hierarchical organisation of compounds using maximum common substructures. The approach is however limited by the fact that there is only one classification tree for the structures created. This means that only one of all possible groupings of compounds can be explored, which will limit the SAR information that could be extracted from a set of compounds. Instead of using only a tree, the program LeadPharmer (LeadPharmer is a trademark of BioReason

Inc., Santa Fe, NM, USA) constructs ‘phylogenetic-like groupings’ of possible substructures [33-35]. The term ‘phylogenetic-like groupings’ was chosen to indicate that substructures are related. The determined substructures are used to assign compounds to different classes. It is possible that a compound can be assigned to more than one class. Using available activity information, interesting classes can be identified and the effect of structural variations on activity studied.

One drawback of the approaches mentioned so far is that the construction of the tree classification can be a quite time-consuming process. The program LeadScope (LeadScope is a trademark of LeadScope Inc., Columbus, OH, USA) tries a different approach [36*]. The program does not construct possible substructures for a number of given compounds, but uses a set of predefined structure fragments (large taxonomy of familiar structural features such as functional groups, aromatics, and heterocycles) to classify the compounds. Therefore, compounds can be assigned to more than one group depending on the structural fragments they contain. The compound classification can be used to explore structure activity relationships in a dataset and search other databases for related structures (see Figure 3).

Data Mining Techniques

Visualisation enables a chemist to interact directly with sets of compounds, but can prove difficult when very many data points need to be considered. Data mining methods, which seek to identify meaningful inter-variable relationships in large, multidimensional datasets, are now being used in a wide range of subject domains, and it is hardly surprising that several of these methods have been used to investigate SARs. Three good general sources on data mining methods are the KDNUGGETS Web site (see URL http://www.kdnuggets.com), and the classic texts by Mitchell [37*] and by Duda and Hart [38*]. The basic problem addressed by all of these methods is that of classification: given a set of molecules for which the activity (or inactivity) is known (the training set), derive a rule that will enable new molecules (the test set) to be classified into the predicted-active or predicted-inactive classes. Training data can be generated internally from ongoing lead-discovery programmes or from publicly available files such as the *MACCS Drug Data Report* (MDDR), *Available Chemicals Directory* (ACD) and *Standard Drug File* databases; the resulting classifications can then be used to

guide the selection of new molecules for synthesis and testing. Thus far, chemical applications have involved the following principal approaches: statistical criteria, decision trees and neural networks.

Medicinal chemists have known for many years that certain types of molecule are unlikely to possess the characteristics necessary for a successful drug: they may be too large to pass the blood-brain barrier, they may be insoluble, they may contain toxic or highly reactive functionality, *etc.* Attempts to quantify such characteristics started with Lipinski's 'Rule of Five' and there have been several, more recent statistical analyses of sets of drug molecules (e.g., [39,40,41*]). A more sophisticated mode of analysis considers also sets of non-drug (or, more usually, presumed non-drug) molecules, this allowing the identification of rules that can be used to assess the 'drug-likeness' or 'drugability' of molecules. An obvious starting point is the distribution of global molecular properties in sets of drug and non-drug molecules. This approach was first studied by Gillet *et al.* [42**], using the distributions of molecular weight, numbers of rotatable bonds, numbers of aromatic rings and of hydrogen bond donors and acceptors, CLOGP and the ${}^{2}K_{\alpha}$ shape index. Here, the distributions for the value of some property in the drug and non-drug molecules is processed by a genetic algorithm (GA) [43] to produce a bioactivity profile, a set of weights that maximise the separation between the distributions for the two classes of molecule; the profiles, are then applied to the property values for test-set compounds so as to obtain a ranking of them in decreasing order of predicted drug-likeness. Gillet *et al.* subsequently described the use of the profiles in a GA for selecting combinatorial libraries of structurally diverse, drug-like molecules [44]; an analogous compound selection procedure has been reported by Sadowski [45] and there is now an extensive literature on the inclusion of drug-likeness in library design procedures [46-49]. A very similar set of global molecular properties has been studied by Oprea in a detailed analysis of several publicly-available datasets [50**], this analysis resulting in the specification of rules for compound-selection that are noticeably more precise than the original Rule of Five.

Statistical analyses of the presence of fragment substructures in active and inactive molecules provides a simple, and convenient alternative to the use of property information. Such approaches were first described almost three decades ago but current requirements for effective compound-selection

procedures has resulted in a surge of interest (se, e.g., [36*,51-53]). Similar approaches can be used to highlight substructures that are undesirable for drug activity (e.g., on grounds of toxicity or unwanted reactivity) [41*,54*].

Neural networks have been applied to a wide range of chemical problems [55**] and they were one of the first such techniques to be applied to drugability studies, the two papers by Ajay *et al*. [56**] and by Sadowski and Kubinyi [57**] appearing contemporaneously with the GA-based approach of Gillet *et al*. [42**]. Here, the network is trained using sets of drugs and non-drugs, and a scoring threshold derived that can maximally discriminate between the two classes; test molecules can then be classified by calculating the score when they are presented to the network. Work in this area is exemplified by the recent study of Frimurer *et al*. [58*]. These authors used sets of molecules from the MDDR and ACD databases to exemplify drugs and non-drugs, with each molecule represented by normalised counts of the numbers of CONCORD atom-types present. These representations were input to a multilayered feed-forward neural network which, after appropriate training, was able to achieve a success rate of 88% in classifying MDDR and ACD compounds that had not been involved in the training; importantly, when used in a predictive manner, the network was able to identify drug-like molecules noticeably different from those obtained from conventional 2D similarity searches. Sadowski discusses the use of a similar neural-network system to discriminate between crop-protecting and non crop-protecting compounds [45].

Decision trees provide an alternative classification tool. Here, the root of the tree represents an entire dataset, and this is subdivided into two (or more) subsets depending on the value of some splitting criterion. Various types of criteria can be used, such as the presence or absence of a particular substructural feature or a CLOGP value lying within a particular range. The potential splitting criteria are scored in some way, and the most advantageous chosen to split the dataset; the procedure is then repeated on the resulting sub-sets, and continued until some termination condition is satisfied. Several different splitting criteria and scoring schemes have been described [37*]. Decision trees were first used in drugability studies by Ajay *et al*. [56**]: these authors used the well-known C4.5 program (which employs an entropy-based scoring function) but who found that the resulting trees performed

less well than neural networks. More recently, however, Wagener and van Geerestein [59**] have used the successor program, C5.0, to distinguish between drug and non-drug compounds with a substantial measure of success on both public and corporate datasets. The most widely-used decision tree procedure for chemical applications has been the recursive partitioning approach, which uses a modified *t*-test for scoring potential splits. This approach has been popularised by Rusinko and co-workers, who have used it not only to analyse 2D fragment substructural data [60**] but also to suggest 3D pharmacophores [61,62]. Other recent examples of the use of recursive partitioning are provided by Cho *et al.* [63] and by Miller [64*]. Decision trees have the advantage over neural networks that they provide explicit, readily comprehensible sets of rules for discussion with medicinal chemists [59**], although Walters and Murcko believe that they are susceptible to over-training, producing classification rules with little predictive power [65]. Mello and Brown [66] have criticised them for assigning test data to just a single class, and have thus developed a hybrid approach that uses the feature-selection capabilities of recursive partitioning as the input to a Bayesian inference network, while Miller has combined recursive partitioning with *k*-nearest neighbour searching [64*]. Finally, Jones-Hertzog *et al.* [67*] describe the use of recursive partitioning to support a sequential HTS analysis of 14 G-protein-coupled receptor targets; other examples of data mining in sequential screening programmes are described by Stanton *et al.* [68] and Engels *et al.* [69*], using nearest neighbour and cluster analysis.

Drugability-based filtering is now common. That said, it must be emphasised that such schemes are still at a very early stage of development: they can often provide erroneous classifications if used without care [70], and they are arguably focused too much on known drugs rather than on the lead compounds that are the principal outputs of screening programmes [71*]. In addition, many of the reported studies thus far have focused on the difference between drugs and nondrugs; however, the same basic techniques can often be applied to the analysis of molecules from a particular therapeutic class if required [39,44,53,58*,72].

Conclusions

This brief review has highlighted some of the soft computing methods that are now being applied to the analysis of the structure-activity relationships present in HTS datasets. However, there are many other methods that have already been, or could be, applied to such problems: examples include ant-based computation [73,74], evolutionary Kohonen networks [27*], fuzzy clustering [23], support vector machines [75] and Bayesian learning [66]. We believe that methods such as these will prove invaluable in the analysis of the huge volumes of data that characterise modern pharmaceutical research, particularly when used in combination [76**.77**].

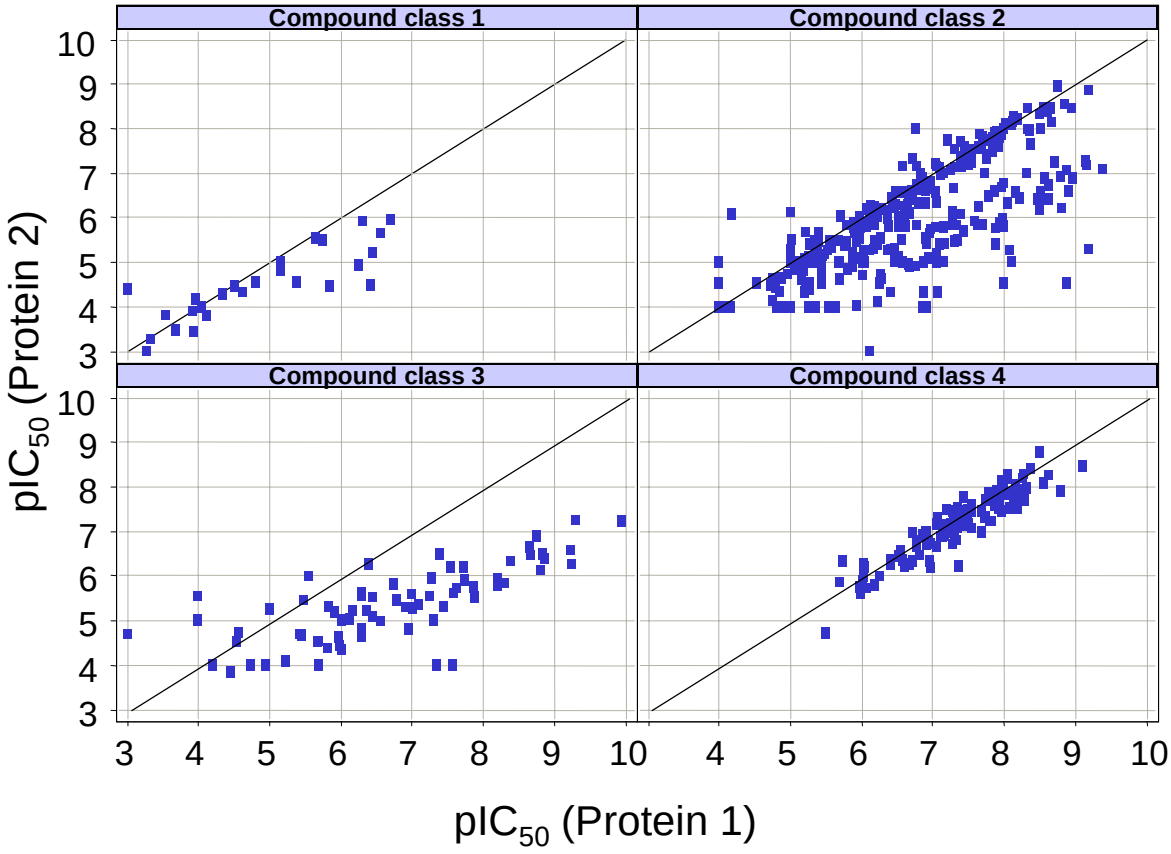
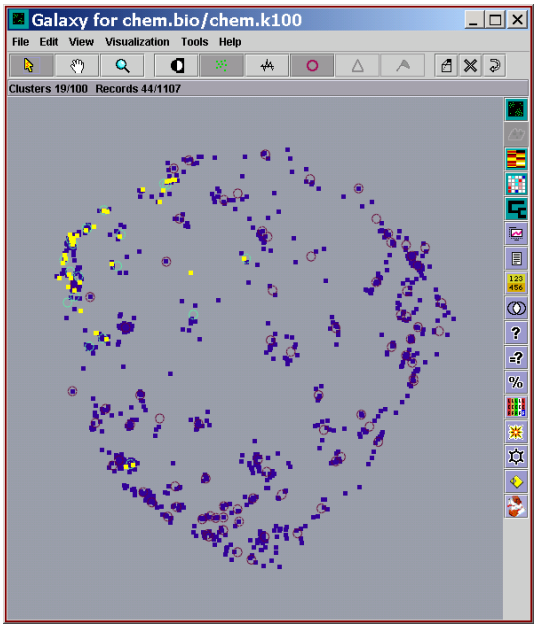


Figure 1: A trellis display created using Spotfire. The four scatter plots compare activity data of compounds measured for two subtypes of a protein. In addition, the compounds were classified into four different structural classes. Each scatter plot shows the data points for one structural class. A comparison of the different scatter plots reveals interesting details. The compounds in class 3 are more selective for protein 1, whereas the compounds in class 4 are equipotent on both proteins. Compound class 2 shows no preference but two groups of compounds are clearly visible. A comparison of the structures in the different classes can reveal further information about the SAR.

a) 

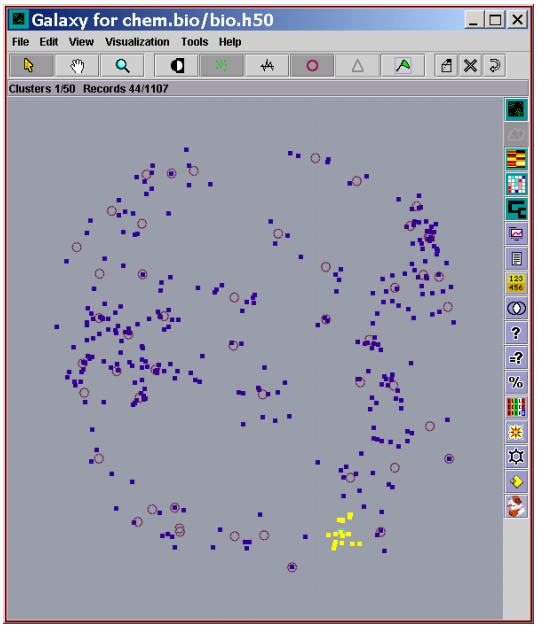
b) 

Figure 2: OmniViz Pro™, from OmniViz, Inc. (Columbus, OH, USA) provides integrated analysis of text, numeric, categorical, and genomic sequence data within a visual Cognitive Analytical Environment™. Here, the software was used in an analysis of 1107 compounds with defined fingerprints that had been screened in 19 biological assays.

In (a), the compounds are represented in a Galaxy™ view, a proximity map that shows how every record is related to every other record. In this map, the similarity is based on the fingerprints, providing a view of the chemical information space. Individual records (blue dots) represent each compound and the clusters of related compounds are evident (marked by circles). A separate Galaxy view in (b) shows the same compounds but with the similarity based on the biological activity profile (biological activity space). A cluster of compounds (lower right) was selected (highlighted yellow) in the biological activity space and the corresponding compounds were automatically highlighted in the structure-based Galaxy. The distribution of the compounds in the structure space indicates that most have similar fingerprint attributes (since they are located in close proximity), but a few are distant (e.g., two in the cluster at the bottom left). This suggests that an alternative set of structural attributes might create the same activity profile - suggesting a new class of structures to pursue.

Courtesy of OmniViz Inc.

Figure 3: The screenshot of the LeadScope user interface shows a comparison of two projects in the histogram view. The left panel shows the structural feature hierarchy open to reveal a portion of the Heterocycles:quinoline branch with quinoline, 2-phenyl selected. The central graphic panel shows parallel histograms comparing the contents of two projects relative to the structural features; each histogram bar gives the frequency of the feature class plotted on a log scale. The right panel contains a series of the property filters, which can be adjusted to select compounds with properties in specific ranges.

The database corresponding to the histogram on the left are compounds tested by the National Cancer Institute's (NCI) Developmental Therapeutics Program for growth inhibition and cytotoxicity against a panel of 60 human cancer cell lines. The comparison database – corresponding to the histogram on the right – are compounds available from Maybridge Chemical Company Ltd [Trevillett, Tintagel, Cornwall PL34 OHW UK].

Histogram bars are colour-coded based on the difference, expressed in number of standard deviations, between the mean activity of the subset of compounds containing a structural feature from the mean activity of the full set. In this example, IC50 data for the SF-295 cell line from the CNS panel is used for the NCI dataset. This technique can be used to locate subsets with unusually high mean activity and then identify new members of the structural class available from a commercial source.

Courtesy of LeadScope Inc.

References

1. Lee MS, Nakanishi H, Kahn M: **Enlistment of combinatorial techniques in drug development**. *Curr Opin Drug Disc Devel* 1999, **2**:332-341.

2. Edwards PJ, Gardner M, Klute W, Smith GF, Terrett NK: **Applications of combinatorial chemistry to drug design and development**. *Curr Opin Drug Disc Devel* 1999, **2**:321-331.

3. Calvert S, Stewart FP, Swarna K, Wiseman JS: **The use of informatics and automation to remove bottlenecks in drug discovery**. *Curr Opin Drug Disc Devel* 1999, **2**:234-238.

4. Thompson LA: **Recent applications of polymer-supported reagents and scavengers in combinatorial, parallel, or multistep synthesis**. *Curr Opin Chem Biol* 2000, **4**:324-337.

5. Kobayashi S: **Immobilized catalysts in combinatorial chemistry**. *Curr Opin Chem Biol* 2000, **4**:338-345.

6. Haupts U, Rüdiger M, Pope AJ: **Macroscopic versus microscopic fluorescence techniques in (ultra)-high-throughput screening**. *Drug Discov Today: HTS supplement* 2000, **1**:3-9.

7. Hertzberg RP, Pope AJ: **High-throughput screening: new technology for the 21st century**. *Curr Opin Chem Biol* 2000, **4**:445-451.

8. Roberts BR: **Screening informatics: adding value with meta-data structures and visualization tools**. *Drug Discov Today: HTS supplement* 2000, **1**:10-14.

9. Tufte ER: *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press; 1983.

10. Tufte ER: *Visual Explanations. Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press; 1997.

* 11. Hand DJ, Blunt G, Kelly MG, Adams NM: **Data Mining for Fun and Profit**. *Stat Sci* 2000, **15**:111-131.
A variety of examples illustrate the problems encountered with data mining of large datasets. This article provides an interesting, easy to understand introduction into the area.

** 12. Card SK, Mackinlay JD, Shneiderman B: *Readings in Information Visualization: Using vision to think*. San Francisco, CA: Morgan Kaufmann Publishers, Inc.; 1999.
As a collection of classic and cutting edge articles published over the last 15 years, this book provides

a thorough introduction into the area of information visualisation in one-, two-, and three-dimensional space.

13. Meyer RD, Cook D: **Visualization of data**. *Curr Opin Biotechnol* 2000, **11**:89-96.

14. Tropsha A: **Recent trends in computer-aided drug discovery**. *Curr Opin Drug Disc Devel* 2000, **3**:310-313.

15. Hayward J, Buchan I: **Benefits of data cartridge technology for handling chemical information**. *Curr Opin Drug Disc Devel* 2000, **3**:306-309.

16. Wedin R: **Visual data mining speeds drug discovery**. *Mod Drug Disc* 1999, **2**:39-47.

17. Ahlberg C: **Visual exploration of HTS databases: bridging the gap between chemistry and biology**. *Drug Discov Today* 1999, **4**:370-485.

18. Ladd B: **Intuitive data analysis: The next generation**. *Mol Divers* 2000, **3**:46-52.

19. Bernard P, Golbraikh A, Kireev D, Chrétien JR, Rozhkova N: **Comparison of chemical databases: analysis of molecular diversity with self organizing maps (SOM)**. *Analusis* 1998, **26**:333-341.

20. Agrafiotis DK: **Stochastic algorithms for maximizing molecular diversity**. *J Chem Inf Comput Sci* 1997, **37**:841-851.

21. Bayada DM, Hamersma H, van Geerestein VJ: **Molecular diversity and representativity in chemical databases**. *J Chem Inf Comput Sci* 1999, **39**:1-10.

22. Kirew DB, Chrétien JR, Bernard P, Ros F: **Application of Kohonen neural networks in classification of biologically active compounds**. *SAR QSAR Environ Res* 1998, **8**:93-107.

23. Ros F, Audouze K, Pintore M, Chrétien JR: **Hybrid systems for virtual screening: interest of fuzzy clustering applied to olfaction**. *SAR QSAR Environ Res* 2000, **11**:281-300.

24. Cox TF, Cox MAA: *Multidimensional Scaling*. London: Chapman & Hall; 1994.

* 25. Clark RD, Patterson DE, Soltanshahi F, Blake JF, Matthew JB: **Visualizing substructural fingerprints**. *J Mol Graph Model* 2000, **18**:404-411.
A modification of Sammon's stress function used in non-linear mapping is proposed. This modification improves the mapping of chemical structures represented by fingerprints considerably.

26. Xie D, Tropsha A, Schlick T: **An Efficient Projection Protocol for Chemical Databases: Singular value Decomposition Combined with Truncated-Newton Minimization**. *J Chem Inf Comput Sci* 2000, **40**:167-177.

* 27. Kohonen T: *Self-Organizing Maps.* 3rd Edition. New York: Springer-Verlag; 2000.
The latest edition of the standard text on self-organizing maps by the person who invented them.

28. Jiang J-H, Wang J-H, Liang Y-Z, Yu R-Q: **A non-linear mapping-based generalized backpropagation network for unsupervised learning**. *J Chemomet* 1996, **10**:241-252.

* 29. Agrafiotis DK, Lobanov VS: **Nonlinear mapping networks**. *J Chem Inf Comput Sci* 2000, **40**:1356-1362.
A combination of conventional non-linear mapping with feed-forward neural networks is proposed. The method allows mapping of datasets considerably larger than would be possible with conventional methods.

30. Garrido L, Goméz S, Roca J: **Improved multidimensional scaling analysis using neural networks with distance-error backpropagation**. *Neural Computation* 1999, **11**:595-600.

* 31. Su H, Che Z-H, Wu J-M, Li R: **Classification mapping and its application on chemical systems**. *J Chem Inf Comput Sci* 1999, **39**:718-727.
Classification maps preserve the classification structure while mapping multidimensional data into a two-dimensional space in a semiquantitative way. Different approaches for this type of non-linear mapping are presented and demonstrated with an example.

* 32. Sheridan RP, Miller MD: **A method for visualizing recurrent topological substructures in sets of active molecules**. *J Chem Inf Comput Sci* 1998, **38**:915-924.
The proposed clique-based subgraph detection method used to finding highest scoring substructures for a pair of molecules, offers an interesting approach to the analysis of structure activity relationships.

33. Bassett SI, Elling JW: **Automating data analysis for high throughput screening**. In *Book of Abstracts, 216th ACS National Meeting, Boston, August 23 – 27.* 1998

34. Elling JW, Hruska SI, Henne R: **Artificial intelligence-directed iterative clustering for lead discovery in high throughput screening (HTS) data**. In *Book of Abstracts, 215th ACS National Meeting, Dallas, March 29 – April 2.* 1998

35. Elling JW, Bassett SI, Nutt RF: **Pharmacophore model generation from high throughput screening (HTS) data sets**. In *Book of Abstracts, 217th ACS National Meeting, Anaheim, CA, March 29 – April 2.* 1999

* 36. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE, Jr.: **LeadScope: Software for Exploring Large Sets of Screening Data**. *J Chem Inf Comput Sci* 2000, **40**:1302-1314.
LeadScope is a software system for the analysis and visualisation of sets of molecules, including facilities for identifying significant differential occurrences of fragment substructures in sets of actives and inactives.

* 37. Mitchell TM: *Machine Learning*. New York: McGraw-Hill; 1997.
The standard text book for machine learning, providing in-depth treatments of a whole range of soft computing tools, several of which have already been studied in drugability studies.

* 38. Duda RO, Stork DG, Hart PE: *Pattern Classification and Scene Analysis. Part 1: Pattern Classification*. 2nd Edition. Chichester: Wiley; 2000.
The first edition of this much cited textbook appeared as long ago as 1973: this new edition is likely to have just as much an influence.

39. Ghose AK, Viswanadhan VN, Wendoloski JJ: **A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases**. *J Comb Chem* 1999, **1**:55-68.

40. Bemis GW, Murcko MA: **Properties of known drugs. 2. Side chains**. *J Med Chem* 1999, **42**:5095-5099.

* 41. Wang J, Ramnarayan K: **Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds**. *J Comb Chem* 1999, **1**:524-533.
Categorisation of a molecule as a drug (or not) based on the presence of connected fragment substructures that occur in a database of known drugs.

** 42. Gillet VJ, Willett P, Bradshaw J: **Identification of biological activity profiles using substructural analysis and genetic algorithms**. *J Chem Inf Comput Sci* 1998, **38**:165-179.
First description of the use of substructural analysis and genetic algorithms to discriminate between drugs and non-drugs, and also between specific therapeutic classes.

43. *Evolutionary Algorithms in Computer-Aided Molecular Design*. Edited by Clark DE. Weinheim: Wiley-VCH; 2000.

44. Gillet VJ, Nicolotti O: **Evaluation of reactant-based and product-based approaches to the design of combinatorial libraries**. *Perspect Drug Discov Design* 2000, **20**:265-287.

45. Sadowski J: **Optimization of drug-likeness of chemical libraries**. *Perspect Drug Discov Design* 2000, **20**:17-28.

46. Martin EJ, Critchlow RE: **Beyond mere diversity: tailoring combinatorial libraries for drug discovery**. *J Comb Chem* 1999, **1**:32-45.

47. Xu J, Stevenson J: **Drug-like index: a new approach to measure drug-like compounds and their diversity**. *J Chem Inf Comput Sci* 2000, **40**:1177-1187.

48. Brown RD, Hassan M, Waldman M: **Combinatorial library design for diversity, cost efficiency and drug-like character**. *J Mol Graph Model* 2000, **18**:427-437.

49. Mason JS, Beno BR: **Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity**. *J Mol Graph Model* 2000, **18**:438-451.

** 50. Oprea TI: **Property distribution of drug-related chemical databases**. *J Comput Aided Mol Des* 2000, **14**:251-264.
An extensive analysis of physicochemical properties in several public databases of chemical compounds, resulting in a considerably enhanced set of criteria for "Rule of Five"-like filters.

51. Gao H, Williams C, Labute P, Bajorath JW: **Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands**. *J Chem Inf Comput Sci* 1999, **39**:164-168.

52. Rhodes N, Willett P, Dunbar J, Humblet C: **Bit-string methods for selective compound acquisition**. *J Chem Inf Comput Sci* 2000, **40**:210-214.

53. Poroikov VV, Filimonov DA, Borodina YV, Lagunin AA, Kos A: **Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds**. *J Chem Inf Comput Sci* 2000, **40**:1349-1355.

* 54. Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N: **Strategic pooling of compounds for high throughput screening**. *J Chem Inf Comput Sci* 1999, **39**:897-902.
The Supporting Information for this paper (see URL http://pubs.acs.org) contains SMARTS definitions for an extensive set of substructures that are used at GlaxoSmithKline to filter compounds containing inappropriate functionality.

** 55. Zupan J, Gasteiger J: *Neural Networks in Chemistry and Drug Design*. 2 Edition. Weinheim: Wiley-VCH; 1999.
The standard text on chemical applications of neural networks.

** 56. Ajay W, Walters W, Murcko MA: **Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules?** *J Med Chem* 1998, **41**:3314-3324.

Joint-first description of the use of a neural network to discriminate between drug and non-drug compounds. Also the first use of decision trees for this purpose.

** 57. Sadowski J, Kubinyi H: **A scoring scheme for discriminating between drugs and nondrugs**. *J Med Chem* 1998, **41**:3325-3329.
Joint-first description of the use of a neural network to discriminate between drug and non-drug compounds.

* 58. Frimurer TM, Bywater R, Naerum L, Lauritsen LN, Brunak S: **Improving the odds in discriminating "drug-like" from "non drug-like" compounds**. *J Chem Inf Comput Sci* 2000, **40**:1315-1324.
Detailed study of the implementation of a neural network-based prediction system using not just public datasets but also a set of 136 proprietary GABA-uptake inhibitors, for which there was a noticeable relationship between activity and predicted drug-likeness.

** 59. Wagener M, van Geerestein VJ: **Potential drugs and nondrugs: prediction and identification of important structural features**. *J Chem Inf Comput Sci* 2000, **40**:280-292.
Use of the C5.0 decision tree program to distinguish between drug and nondrug compounds in both public and corporate datasets. Their results suggest that surprisingly successful classifications can be achieved simply on the basis of testing for the presence of simple functional groups (e.g., hydroxyl, secondary or tertiary amino, phenol etc.).

** 60. Rusinko A, III, Farmen MW, Lambert CG, Brown PL, Young SS: **Analysis of a large structure/biological activity data set using recursive partitioning**. *J Chem Inf Comput Sci* 1999, **39**:1017-1026.
Description of the SCAM (Statistical Classification of the Activities of Molecules) program for carrying out recursive partitioning on very large chemical datasets (300K compounds with 2M fragment variables were analysed in less than one hour CPU time).

61. Chen X, Rusinko A, III, Young SS: **Recursive Partitioning Analysis of a Large Structure-Activity Data Set Using Three-Dimensional Descriptors**. *J Chem Inf Comput Sci* 1998, **38**:1054-1062.

62. Chen X, Rusinko A, III, Tropsha A, Young SS: **Automated pharmacophore identification for large chemical datasets**. *J Chem Inf Comput Sci* 1999, **39**:887-896.

63. Cho SJ, Shen CF, Hermsmeier MA: **Binary formal inference-based recursive modelling using multiple atom and physicochemical property class pair and torsion descriptors as decision criteria**. *J Chem Inf Comput Sci* 2000, **40**:668-680.

* 64. Miller DW: **Results of a new classification algorithm combining *k*-nearest neighbours and recursive partitioning**. *J Chem Inf Comput Sci* 2001, **41**:168-175.
A modification of the flexible metric nearest neighbour classification (FMNN) is presented and tested using HIV-1 infection data from the NCI database. In comparison with conventional *k*-nearest neighbours and recursive partitioning, the new method performs better.

65. Walters W, Murcko MA: **Library filtering systems and prediction of drug-like properties**. In *Virtual Screening for Bioactive Molecules.* Edited by Böhm H-J, Schneider G. Weinheim: Wiley-VCH; 2000:15-32.

66. Mello KL, Brown SD: **Novel 'hybrid' classification method using Bayesian networks**. *J Chemomet* 1999, **13**:579-590.

* 67. Jones-Hertzog DK, Mukhopadhyay P, Keefer CE, Young SS: **Use of recursive partitioning in the sequential screening of G-protein-coupled receptors**. *J Pharmacol Toxicol Methods* 1999, **42**:207-215.
Description of a sequential screening strategy based on decision trees. The authors show using experimental HTS data for 14 GPCR targets, that screening 10 to 20 % of a collection is sufficient to find 50 to 80 % of the active compounds.

68. Stanton DT, Morris TW, Roychoudhury S, Parker CN: **Application of nearest-neighbour and cluster analyses in pharmaceutical lead discovery**. *J Chem Inf Comput Sci* 1999, **39**:21-27.

* 69. Engels MFM, Thielemans T, Verbinnen D, Tollenaere JP, Verbeek R: **CerBeruS: A system supporting the sequential screening process**. *J Chem Inf Comput Sci* 2000, **40**:241-245.
Describes CerBeruS, a system for sequential screening using cluster-based selection. Use of the system on the NCI AIDS dataset (anti-HIV activity of 32110 compounds) increased the hit rate in subsequent screens.

70. Lepre CA: **Library design for NMR-based screening**. *Drug Discov Today* 2001, **6**:133-140.

* 71. Teague SJ, Davis AM, Leeson PD, Oprea TI: **The design of lead-like combinatorial libraries**. *Angew Chem Int Ed* 1999, **38**:3743-3747.
Analysis of successful lead compounds shows that they are often substantially less complex than fully-fledged drugs, with a consequent need to reflect this in computer-based filtering systems.

72. Ajay W, Bemis GW, Murcko MA: **Designing libraries with CNS activity**. *J Med Chem* 1999, **42**:4942-4951.

73. Bonabeau E, Dorigo M, Theraulaz G: **Inspiration for optimization from social insect behaviour**. *Nature* 1999, **406**:39-42.

74. Izrailev S, Agrafiotis DK: **A novel method for building regression tree models for QSAR based on artificial ant colony systems**. *J Chem Inf Comput Sci* 2001, **41**:176-180.

75. Christianini N, Shawe-Taylor J: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press; 2000.

** 76. Shi LM, Fan Y, Lee JK, Waltham M, Andrews DT, Scherf U, Paull KD, Weinstein JN: **Mining and visualizing large anticancer drug discovery databases**. *J Chem Inf Comput Sci* 2000, **40**:367-379.
Large-scale analysis of the NCI anticancer drug database using principal component analysis, cluster analysis, neural networks, multidimensional scaling and genetic function approximation.

** 77. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN: **A gene expression database for the molecular pharmacology of cancer**. *Nat Genet* 2000, **24**:236-244.
Gene expression profiles in 60 human cancer cell lines were used to correlate gene expression and drug activity. A clustered image map is used to effectively summarise the result. This work is an excellent demonstration how the gap between chemoinformatic and bioinformatic knowledge can be bridged.