Global analysis of large-scale chemical and biological experiments

David E Root, Brian P Kelley & Brent R Stockwell*

Address
Whitehead Institute for Biomedical Research
Nine Cambridge Center
Cambridge
MA 02142
USA
Email: stockwell@wi.mit.edu

*To whom correspondence should be addressed

Current Opinion in Drug Discovery & Development 2002 **5**(3):355-360
 ISSN 1367-6733

Research in the life sciences is increasingly dominated by high-throughput data collection methods that benefit from a global approach to data analysis. Recent innovations that facilitate such comprehensive analyses are highlighted. Several developments enable the study of the relationships between newly derived experimental information, such as biological activity in chemical screens or gene expression studies, and prior information, such as physical descriptors for small molecules or functional annotation for genes. The way in which global analyses can be applied to both chemical screens and transcription profiling experiments using a set of common machine learning tools is discussed.

Keywords Chemical genetics, chemical genomics, descriptors, high throughput, informatics, transcription profiling

Introduction

Research in the life sciences has become dominated by high-throughput data collection methods. It is now common to screen many thousands or millions of small molecules in miniaturized biological tests, such as protein-targeted assays or cell-based assays [1•]. In addition, it is common to perform microarray-based transcription profiling, which involves the simultaneous hybridization of thousands of DNA sequences to spatially arrayed targets [2]. An emerging challenge is the analysis and integration of the large datasets generated by these disparate high-throughput techniques.

Until recently, only a few genes or compounds postulated in advance to be modulators of a phenotype or to have activity of interest were selected for study. High-throughput methods now permit use of a hypothesis-generating strategy in which large libraries of genes or chemicals are tested for biological effects of interest. One relies on the large size and diversity of the initial collection to yield active genes or compounds rather than prior knowledge of the screening candidates or the biological processes being studied. This strategy uncovers a large and varied set of active compounds or genes that can then be studied with a targeted, hypothesis-driven approach.

Ideally, the dataset from each new high-throughput experiment is interpreted in the context of all previous results. It then becomes part of the context in which all future screens are analyzed. Building on previous results is not new, but doing so takes on a new level of importance and complexity when datasets are vast and involve extremely inter-related information, and the relevant prior experimental data cannot be stored and organized in the mind of one scientist. We use the term 'global analysis' to refer to an emphasis on greater integration and analysis of data from all sources.

Challenges involved in the global analysis of experimental data are illustrated by the new fields of chemical genetics and chemical genomics [1•]. By analogy to classical genetics, chemical genetics uses small molecules in place of mutations as modifiers of protein function. Small molecules that modulate a process or phenotype of interest are identified through large-scale screening and serve as probes of the mechanisms underlying the biological process. Chemical genetics, like other large-scale screening approaches, integrates information from several large datasets. The activity profile of a library of compounds in a particular assay is measured and correlated with structural and chemical properties of the compounds, as well as previously documented biological activities. Chemical genomics involves the integration of chemical and genomic information and technologies. One example of the challenges of a chemical genomic approach is the integration and analysis of both transcription profiling and chemical screening data.

We will review work reported primarily within the last year that is applicable to global analyses of the properties of both small molecules and genes, focusing on: (i) selection and evaluation of physical descriptors for small molecules; (ii) new applications of machine learning algorithms; and (iii) novel approaches for analyzing microarray-based transcription profiling data.

Selecting chemical entities to screen

We restrict our discussion of chemical screens to low molecular weight organic molecules as these compounds are of particular interest in drug discovery efforts and in biological research. Small molecule screens are preferred for drug discovery because the resulting lead compounds can be more easily developed into orally available pharmaceuticals. Many of the tools for global analyses that we describe can also be applied to screens involving peptide, RNA, DNA or protein reagents.

The problem of selecting compounds to screen is a difficult one. The total number of possible organic compounds increases with molecular weight, thus, without a defined molecular weight cut-off there is an infinite number of possible compounds. Published estimates of the number of theoretical small molecule drugs range as high as 10^{66}, which is close to the number of atoms in the universe [3].

One strategy for selecting compounds for screening is to purchase or make a representative set of molecules based on physical properties or functional groups. This approach amounts to an attempt to select an optimally diverse subset of the obtainable compounds for an initial screen. Jorgensen *et al*, for example, developed a method for evaluating the

diversity of a compound collection using common subgraphs or substructural elements [4]. Xu *et al*, on the other hand, developed a drug-like index to aid the selection of compounds for screening. The index was trained on 4836 compounds from the Comprehensive Medicinal Chemistry database [5]. Reynolds *et al* evaluated two stochastic sampling algorithms for their ability to select both diverse and representative subsets of a chemical library space [6].

Much effort has also focused on exploring and quantitating the notion of molecular complexity and determining the appropriate level of complexity for small molecules used in high-throughput screens. Barone and Chanon refined a quantitative index of complexity that uses the number and size of the rings in the smallest set of smallest rings and the connectivity of each atom [7•]. Alternatively, complexity can be defined as the number of interactive domains contained in a molecule. A molecule with low complexity has fewer sites of interaction with a target than a molecule with greater complexity. Hann *et al* devised a simple model in which complex molecules are more selective than simple compounds and, therefore, yield fewer hits in primary screens [8•]. This model predicts an optimal level of complexity for compounds used in primary screens as the result of a trade-off between sufficient affinity for detection versus sufficient promiscuity to yield a reasonable number of hits. This model is consistent with recent analyses affirming that successful lead compounds are generally less complex than the resulting drugs [8•,9•].

Given the virtually unlimited sources of small molecules, there has been interest in identifying characteristics of small molecules that are useful for drugs and for creating models that predict the probability that a given compound will be able to function as a drug (*vide infra*). It is difficult to evaluate the performance of these predictive models because of the great variability in crucial factors, such as the choice of the training sets of compounds and the choice of descriptors that define the actual criteria for discrimination. Furthermore, all empirically derived predictive models are essentially interpolative and extrapolative. Models that are better at assigning close structural analogs to members of the training set (interpolation) may be worse at generalizing more abstract properties to novel structures (extrapolation) and vice versa. Thus, one must beware of inferring the overall performance of a predictive model from a too limited set of test compounds.

Nonetheless, several efforts at discriminating drugs and non-drugs have been reported recently. Ertl *et al* used polar atom surface area to predict the extent to which small molecules exhibit a single property of drug transport (ie, bioavailability) [10]. Anzali *et al* used chemical descriptors consisting of multilevel neighborhoods of atoms to discriminate between drugs and non-drugs with some success. Their training and testing sets consisted of 5000 compounds from the World Drug Index and 5000 compounds from the Available Chemicals Directory (ACD) [11]. Muegge *et al* developed a simple functional group filter to discriminate between drugs and non-drugs using both the Comprehensive Medicinal Chemistry and MACCS-II Drug Data Report (MDDR) databases for drugs and the ACD for non-drugs [12]. Frimurer *et al* used a feed-forward neural network with two-dimensional (2D) descriptors based on atom types to classify compounds from the MDDR and ACD as drug-like or non-drug-like, respectively. They reported 88% correct assignment of a subset of each library that had been excluded from the training set. They also tested their model with a different library and claimed generalizability to compounds structurally dissimilar to those in the training set [13].

Drug versus non-drug comparisons emphasize characteristics common to all drugs over those characteristics specific to a particular receptor. Drugs share a number of general characteristics, such as target-binding affinity and the ability to permeate into cells, and they must also have favorable absorption, distribution, metabolism and excretion (ADME) properties. Models that discriminate drugs from non-drugs tend to select for ADME properties rather than properties that correlate with cellular biological activity. If one is interested simply in cellular biological activity rather than the full complement of required drug characteristics, a correspondingly appropriate compound training set must be selected. For example, in chemical genetic approaches, compound libraries with enriched protein-binding affinity are valuable, whereas compounds with favorable ADME properties have little added value.

Finally, it has been noted that many natural products do not conform to the canonical rules for selecting drug-like compounds. Moreover, many natural products have been directly developed as drugs without the need for significant (or any) analog synthesis. This observation has inspired a new strategy of synthesizing natural-product like compounds using combinatorial, diversity-oriented syntheses [14•,15•].

Descriptors

For comparisons that involve molecular properties, the structural, physicochemical, and/or biological properties of the molecules need to be represented in a consistent form to permit direct comparison. A standardized representation of a molecular feature is referred to as a 'descriptor'. The choice of descriptors plays a crucial role in the analysis of chemical screening data. A major challenge in descriptor analyses is the identification of the smallest, most easily and reproducibly calculated set of descriptors that retains all the information required to make the distinctions and comparisons of interest. Here, we discuss some general considerations concerning descriptor choice, and highlight some recent developments.

Chemical descriptors

The compounds in a database are normally identified by their 2D structural representations, which consist of a list of the constituent atoms, their interconnectivity and sometimes their relevant stereochemistry. Aside from experimental data, these 2D representations of the molecular structure typically contain all the available information distinguishing the compounds in the library. For each compound, a common set of structural/physical/chemical descriptors is generated from these 2D structures. Choosing this set of descriptors amounts to defining the 'chemical space' spanned by all possible descriptor representations. A correlation between regions in this chemical space and

bioactivity is assumed to arise from the binding of the chemical to specific biological targets. Here, we concentrate on the case in which there is no specific knowledge of the presumed binding sites and there is a purely empirical relationship between structure and activity.

There is a tremendous range in both the complexity and the reliability of descriptors. Simple descriptors, such as atom counts, may be obtained directly and reliably from the 2D structural representation. At the other extreme of both complexity and reliability are three-dimensional (3D) descriptors that involve 3D geometry-optimization and provide no assurance of producing a conformation with *in vivo* relevance. A widely varying number of descriptor dimensions have been employed to describe chemical libraries, but these have all involved a reduction in dimensions and, thus, a loss of information versus the original representation. Removing information that does not distinguish molecules by the properties of interest (eg, bioactivity) decreases the computational expense involved in computing and manipulating the descriptor representations and the 'noise' associated with the descriptors that do not contribute to the distinction of interest. One family of widely used descriptors consists of database hash keys, which were originally designed to filter compounds quickly in substructure searches. Although experience shows that these keys are unreliable when used alone to represent compounds, they have proven useful when used in conjunction with other descriptors [16•,17•,18].

Considerable effort has been devoted to determining the importance of 3D (conformational) information relative to more simply and reliably obtained 2D information, but the results seem to be highly dependent on the details of the analysis and the nature of the correlation being sought. 3D conformational analysis is generally avoided in the interest of computational speed and reproducibility. Estrada *et al* found a significant correlation between 2D topological indices and the dihedral angle in a series of alkylbiphenyls, demonstrating that 3D properties may be implicitly represented without resorting to geometry optimization [19]. In addition, Ertl found that 2D topological information was sufficient to calculate a molecular surface polar area descriptor that was essentially identical to the value obtained with the comparable 3D calculation [10]. One limitation of topological descriptors is that they cannot distinguish between stereoisomers. To help address this problem, Golbraikh *et al* [20] and Lukovits and Linert [21] have introduced interesting ways of combining chirality with 2D topological information.

The descriptors chosen to describe a compound library may be very different from one another with respect to their range and distribution. Godden and Bajorath used measures derived from Shannon entropy to quantify the information content of each descriptor within a compound library. They extended this method to compare the distributions of a descriptor between different libraries [22•].

Biological descriptors

There are a number of biologically relevant quantities that can be used as independent variables in a manner directly analogous to the chemical descriptors described above. Biological descriptors can be used in the global analyses of microarray-derived transcription profiling data or to interpret the results of a screen for biological activity in terms of previously known activities of compounds in the library. Chromosomal location can also serve as a descriptor. For example, Wyrick *et al* used chromatin immunoprecipitation and subsequent hybridization to genomic DNA microarrays to identify autonomously replicating sequences (ARS) in yeast cells. Using chromosomal location in the list of generated sequences, these authors determined that ARSs are overrepresented in subtelomeric and intergenic regions of chromosomes [23••].

Properties can be calculated directly from DNA sequence information in a manner analogous to the calculation of physical descriptors for small molecules. For example, enrichment of the fraction of guanine/cytosine base pairs (GC content) in promoter regions can be calculated directly from genomic DNA sequence. Konu *et al*, for example, found that gene expression levels were correlated with the GC content of the third nucleotide codon position of the message [24]. One can relate the presence of splice site sequences, promoter elements and transcription factor binding sites to gene expression level using similar strategies. For example, Bernstein *et al* determined that binding sites for the transcription factor Ume6p were enriched upstream of genes that are induced in *sin3* mutant yeast cells [25]. This type of global analysis correlates genomic sequence information with gene expression data.

Some properties, such as gene function, may be linked to a DNA sequence through a strategy of annotation. Other possible annotations include chromosomal location, protein interactions and co-regulated expression groups. Each of these descriptors can serve as an independent variable for global analyses. Using functional annotation categories, Bernstein *et al* determined that the expression of carbon metabolite and carbohydrate utilization genes was greater in yeast cells with a *HDA1* deletion [25].

The construction of a descriptor vector for each gene used in a microarray experiment can be envisaged. Each sequence (eg, gene or chromosomal fragment) would have an associated value for GC content, the number of splice sites, the number and type of promoter elements, the number of binding sites for each of many transcription factors and a quantitative assignment (perhaps binary) for each functional annotation category. Once these vectors are constructed, they allow rapid analysis of the relationship between active and inactive genes for each of these descriptor categories. By applying computational strategies described in the next section, it is possible to extract the relationship between, for example, the number of AP-1 binding sites in a gene promoter and the level of induced expression in an experiment. Moreover, such methods would permit the detection of non-linear and combinatorial relationships among these descriptors, eg, 'stress-response genes with AP-1 binding sites and > 40% GC content in their promoter are enriched in response to stimulus X'. Finally, data from global analyses could be used to develop a predictive model to classify untested genes.

Data analysis

It is important to make a distinction between two fundamentally different applications of high-throughput screening data. Such methods may be used simply to identify compounds exceeding a certain activity threshold (hits) or to identify a more comprehensive correlation between the measured activity, molecular structure and/or previously determined biological activity or mechanism. This distinction is important because the acceptable false positive and false negative rates for the two approaches are substantially different. In a 'threshold' screen, high false negative and false positive rates are acceptable because secondary screening of the hits is used to distinguish between true positives and false positives. Since the identification of true positives is the ultimate goal in a 'threshold' screening approach, false negatives are not a concern as long as a sufficient number of true positives is found. In a global analysis, however, the false positive and false negative rates must be minimized because all results are used in a quantitative or semi-quantitative analysis. Global analyses can be quite powerful but are more expensive in terms of time and money to perform, and may require the use of sophisticated computational methods (*vide infra*).

Analysis of screening data

Screening results typically exhibit a continuous range of activities, usually with a Gaussian distribution. A cut-off value is chosen for the selection of hits and the active elements are normally confirmed in a secondary assay. The cut-off criteria for determining hits may be based on absolute activity (ie, 2-fold activity versus control), distribution (ie, three standard deviations or greater from the mean) or a desired number of compounds to be retested. Once confirmed actives have been identified, it may be desirable to search for additional active elements by testing or retesting candidates that are related in form or function. In transcription profiling screens, retesting entails performing a search of the original gene set for genes that are related to the active genes in terms of sequence or function. The screen comes to its natural conclusion with the selection of a set of actives that can be pursued in subsequent experiments.

Global analyses

Various learning techniques have been used to generate hypotheses and form models of relationships between descriptors and biological activity. These techniques may be divided into two main categories: classification and clustering. For simplicity, we assume that the data to be analyzed are compound descriptors and that the classes of compounds are active and inactive.

The goal of a classifier is to produce a model that can separate new, untested compounds into classes using a training set of already classified compounds. Classification routines attempt to discover those descriptors or sets of descriptors that distinguish the classes from each other. Neural networks, genetic algorithms and support vector machines attempt to discover regions in descriptor space that separate pre-defined classes. Unknown compounds that are subsequently placed in these regions can be classified as active or inactive [26-28]. These techniques optimize a learning function in order to fit the given number of classes while minimizing an error function based on the mismatch of the classifier in the assignment of compounds. One of the main issues of training is overfitting, in which the initial classes are learned so narrowly that no new members are allowed into a class. The learned model should be specific so that it seldom misclassifies compounds from the original training set but general enough to recognize new compounds that should belong to a class.

Recursive partitioning and decision trees first find the best single descriptor to split active and inactive populations into two groups and then successively find the next best descriptor to further divide the newly formed groups. These are known as greedy algorithms because they select the best solution at every step but do not necessarily find the global optimum [29].

Statistical methods can also be used to form probability models or estimate the likelihood of particular descriptors forming the known classes. These approaches generally involve the use of the training set to form a probability model that generates both a classification and a probability of being in a class. Simple statistical methods include k-nearest neighbors and the Naïve Bayes classifier. Support vector machines are also examples of statistical classifiers.

The goal in clustering a dataset is to group similar data together. Clustering forms groups of compounds that maximize internal class similarity while simultaneously minimizing external class similarity. Clustering can be accomplished by either a supervised method, where the number of classes is known, or through unsupervised learning, where the data are not grouped into a fixed set of classes.

In many cases, classes produced by clustering can be used for classification. Unknown compounds that group with predominately active compounds have a higher probability of also being active [30]. One drawback to this strategy is the fact that the higher hit rate only applies to the relatively small number of compounds that lie close to known hits. Furthermore, models of activity are not generated from clustering techniques and must be deduced by expert analysis. Indeed, descriptors that cluster compounds together may not be related to activity at all. As with classification, there are a variety of available clustering algorithms. These include hierarchical methods, such as Ward's clustering, and non-hierarchical methods, such as Jarvis-Patrick [31] and Self-Organizing Maps [32]. Examples of statistical-based clustering include the use of Bayesian neural network to cluster drugs and non-drugs [33] and the use of k-nearest neighbor analysis to cluster compounds at various stages of the screening process [34].

In a recent global analysis of both compound screening and gene expression data, Staunton *et al* used a statistical classifier to identify a correlation between gene expression and cell sensitivity to compounds. Sixty cancer cell lines were exposed to numerous compounds at the National Cancer Institute, and were determined to be either sensitive or resistant to each compound. Using a Bayesian statistical classifier, Staunton *et al* showed that for at least one third of the tested compounds, cell sensitivity can be predicted with the gene expression pattern of untreated cells [35••]. This

example demonstrates the power of global analyses to identify subtle but important relationships among variables in large-scale datasets.

Conclusion

Global analyses can be performed on data from compound screening and transcription profiling experiments using similar computational methods. The goal of such analyses is to discern sometimes-subtle relationships within these datasets and to make correlations between large sets of multidimensional data. Recent advances are making global analyses increasingly feasible and powerful.

There are numerous future challenges in this area. Firstly, it will be valuable to identify robust chemical descriptors that best define global chemical space, as well as the ligand-rich regions therein. Standardized tests for evaluating classification methods would enable more meaningful comparisons. Finally, methods for automatic incorporation of publicly accessible data into such analyses would be enormously powerful, as the range of testable relationships would expand dramatically.

Acknowledgments

Brent R Stockwell, PhD, is a Whitehead Fellow and is supported in part by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

References

•• of outstanding interest
• of special interest

1. Stockwell BR: **Chemical genetics: Ligand-based discovery of gene function.** *Nat Rev Genet* (2000) **1**:116-125.

• *A comprehensive review of the history, promise and current state of chemical genetics.*

2. Mills JC, Roth KA, Cagan RL, Gordon JI: **DNA microarrays and beyond: Completing the journey from tissue to cell.** *Nat Cell Biol* (2001) **3**:E175-E178.

3. Smit WA, Bochkov F, Caple R (Eds): *Organic Synthesis: The Science Behind the Art*. Royal Society of Chemistry, Cambridge, UK (1998).

4. Jorgensen AMM, Pedersen JT: **Structural diversity of small molecule libraries.** *J Chem Inf Comput Sci* (2001) **41**:338-345.

5. 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.

6. Reynolds CH, Tropsha A, Pfahler LB, Druker R, Chakravorty S, Ethiraj G, Zheng W: **Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms.** *J Chem Inf Comput Sci* (2001) **41**:1470-1477.

7. Barone R, Chanon M: **A new and simple approach to chemical complexity. Application to the synthesis of natural products.** *J Chem Inf Comput Sci* (2001) **41**:269-272.

• *Reviews quantitative indices of structural complexity, and introduces refinements.*

8. Hann MM, Leach AR, Harper G: **Molecular complexity and its impact on the probability of finding leads for drug discovery.** *J Chem Inf Comput Sci* (2001) **41**:856-864.

• *Presents an interesting model for the optimal complexity of small molecules to be used in high-throughput screens.*

9. Oprea TI, Davis AM, Teague SJ, Leeson PD: **Is there a difference between leads and drugs? A historical perspective.** *J Chem Inf Comput Sci* (2001) **41**:1308-1315.

• *Analyzes the differences in complexity between lead molecules and drugs.*

10. Ertl P, Rohde B, Selzer P: **Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties.** *J Med Chem* (2000) **43**:3714-3717.

11. Anzali S, Barnickel G, Cezanne B, Krug M, Filimonov D, Poroikov V: **Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS).** *J Med Chem* (2001) **44**:2432-2437.

12. Muegge I, Heald SL, Brittelli D: **Simple selection criteria for drug-like chemical matter.** *J Med Chem* (2001) **44**:1841-1846.

13. 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.

14. Schreiber SL: **Target-oriented and diversity-oriented organic synthesis in drug discovery.** *Science* (2000) **287**:1964-1969.

• *Comprehensive review of the new field of diversity-oriented organic synthesis*.

15. Blackwell HE, Perez L, Stavenger RA, Tallarico JA, Cope Eatough E, Foley MA, Schreiber SL: **A one-bead, one-stock solution approach to chemical genetics: Part 1.** *Chem Biol* (2001) **8**:1167-1182.

• *Describes an industrialized process for synthesizing complex molecules on solid phase using split-pool synthesis.*

16. Bajorath J: **Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening.** *J Chem Inf Comput Sci* (2001) **41**:233-245.

• *An excellent overview of compound classification and molecular descriptor analysis.*

17. MacCuish J, Nicolaou C, MacCuish NE: **Ties in proximity and clustering compounds.** *J Chem Inf Comput Sci* (2001) **41**:134-146.

• *First demonstration that database hash keys have poor information content and that their use can compromise clustering results.*

18. Flower DR: **On the properties of bit string-based measures of chemical similarity.** *J Chem Inf Comput Sci* (1998) **38**:379-386.

19. Estrada E, Molina E, Perdomo-Lopez I: **Can 3D structural parameters be predicted from 2D (topological) molecular descriptors?** *J Chem Inf Comput Sci* (2001) **41**:1015-1021.

20. Golbraikh A, Bonchev D, Tropsha A: **Novel chirality descriptors derived from molecular topology.** *J Chem Inf Comput Sci* (2001) **41**:147-158.

21. Lukovits I, Linert W: **A topological account of chirality.** *J Chem Inf Comput Sci* (2001) **41**:1517-1520.

22. Godden J, Bajorath J: **Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis.** *J Chem Inf Comput Sci* (2002) **42**:87-93.

• *Develops a measure of the usefulness of various molecular descriptors.*

23. Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG, Young RA, Bell SP, Aparicio OM: **Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: High-resolution mapping of replication origins.** *Science* (2001) **294**:2357-2360.

•• *An excellent example of the use of chromosomal location information in a global analysis to identify the distribution of origins of replication throughput in the genome of yeast cells.*

24. Konu OO, Li MD: **Correlations Between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents.** *J Mol Evol* (2002) **54**:35-41.

25. Bernstein BE, Tong JK, Schreiber SL: **Genomewide studies of histone deacetylase function in yeast.** *Proc Natl Acad Sci USA* (2000) **97**:13708-13713.

26. Haykin S (Ed): *Neural Networks: A Comprehensive Foundation*, 2nd Edition. Prentice Hall, Upper Saddle River, NJ, USA (1999).

27. Vapnik VN (Ed): *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, Germany (1995).

28. Koza J, Banzhaf W, Chellapilla K, Deb K, Dorigo M, Fogel D, Garzon M, Goldberg D, Iba H, Riolo R (Eds): *Genetic Programming 1998*. Morgan Kaufmann, San Francisco, CA, USA (1998).

29. Rusinko R III, Farmen MW, Lambert CG, Brown PL, Young SS (Eds): **Analysis of a large structure/biological activity data set using recursive partitioning.** *J Chem Inf Comput Sci* (1999) **39**:1017-1026.

30. Engels MFM, Venkatarangan P: **Smart screening: Approaches to efficient HTS.** *Curr Opin Drug Discovery Dev* (2001) **4**:275-283.

31. Brown RD, Martin YC: **The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding.** *J Chem Inf Comput Sci* (1997) **37**:1-9.

32. Kohonen T (Ed): *Self-Organizing Maps. Springer Series in Information Sciences*, *Volume 30*. Springer, Berlin, Heidelberg, Germany (2001).

33. Ajay A, Walters WP, Murcko MA: **Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?** *J Med Chem* (1998) **41**:3314-3324.

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

35. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci USA* (2001) **98**:10787-10792.

•• *An excellent example of a global analysis that uses chemical screening data, transcription profiling data and a statistical classifier. The results demonstrate that gene expression profiles of untreated tumor cells can be used to predict their sensitivity to many chemical agents.*