COMPUTATION AND ANALYSIS OF LARGE CHEMISTRY DATA SETS

S. STANLEY YOUNG AND CHRIS E. KEEFER

Glaxo Wellcome Inc., Research Triangle Park, NC 27709, USA.

E-mail: ssy0487@glaxowellcome.com ; cek43215@glaxowellcome.com

Received: 25th May 2000 / Published 11th May 2001

ABSTRACT

Very large screening data sets are becoming available; hundreds of thousands of compounds are screened against panels of biological assays. There is a need to make sense out of the data; screeners need to know which compounds to screen next and medicinal chemists need to know which series of compounds are active and what features are associated with activity. We use the statistical technique recursive partitioning and simple molecular descriptors, atom pairs and topological torsions, to analyze these data sets based upon the 2D representation of the compounds. We use more general features and a special 3D representation of the compounds for pharmacophore identification. The benefit of this work is that we can rapidly evaluate screening data and make sound recommendations for additional screening work or how to proceed with lead optimization.

INTRODUCTION

Enormous numbers of compounds are now available for screening. Large companies will have over five hundred thousand compounds in inventory; over one million compounds are available commercially; library synthesis offers many millions of possible compounds. It is not feasible to screen all available compounds in all screens. Indeed, with the ongoing genetics efforts there will be an explosion of drug targets over the next several years, increasing the number of available screens.

There is a need to be able to examine screening data and make recommendations on how to proceed.

Which compounds should be screened next? Which compounds acquired for screening? When to stop screening and move to lead optimization? For lead compounds, what are the important features? Statistical analysis of large screening sets can help with all of these questions. In this paper we describe the use of recursive partitioning for the analysis of large chemistry data sets.

METHODS

We use rather simple compound descriptors. See Figure 1 for examples of atom pairs, [1] atom triples, and topological torsions. [2] For pharmacophore identification we use standard pharmacophore features. [3]

Recursively splitting a data set into homogeneous subsets was first proposed by Morgan, and Sonquest. [4] Statistical methods for univariate recursive partitioning are described by Hawkins and Kass, [5] Hawkins *et al*. [6] and Rusinko *et al*. [7] Basically, all potential variables are examined and the single variable that will best split the entire data set into two daughter data sets is selected and the split made; those compounds with the feature go to the right daughter node and those without the feature go to the left. See Figure 2.

Atom pair
N(3,0) - 7 - S(2,0)

Atom triple
N(3,0) - 7 - S(2,0)
S(2,0) - 6 - N(2,0)
N(3,0) - 12 - N(2,0)

Topological torsion
N(3,0)C(2,0)-1-
C(3,1)O(2,0)

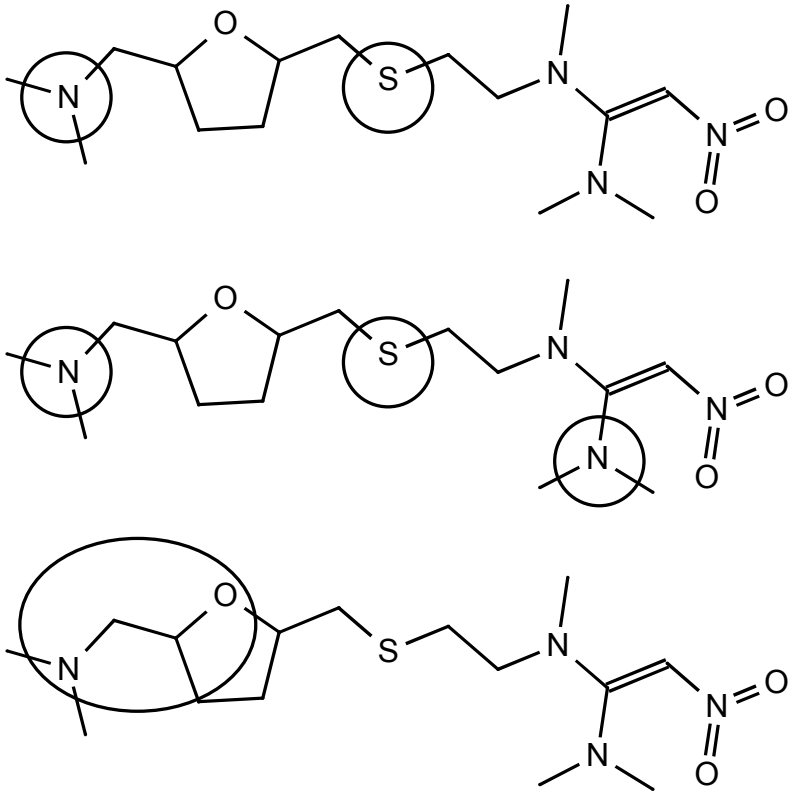


Figure 1. Atom pair, atom triple and topological torsion molecular descriptors.

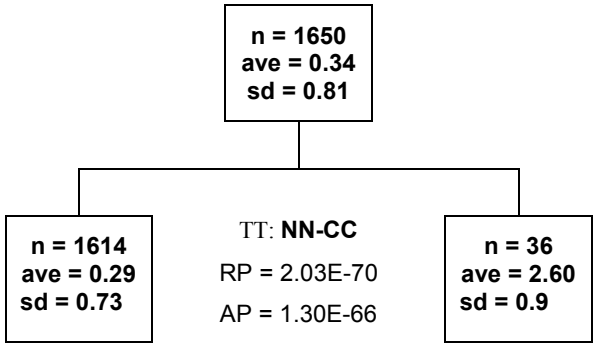


Figure 2: The data set is split using a t-test.

Each daughter node is split in turn. Splitting stops when there are no statistically significant splits remaining. For multivariate recursive partitioning we replace the Student t-test with the Hotelling T^2. [8]

RESULTS

Recursive partitioning is capable of identifying multiple chemical classes of compounds from a data set, and is thus a method for deconvoluting mixtures. [7] Figure 3 gives a skeleton of the recursive partitioning tree.

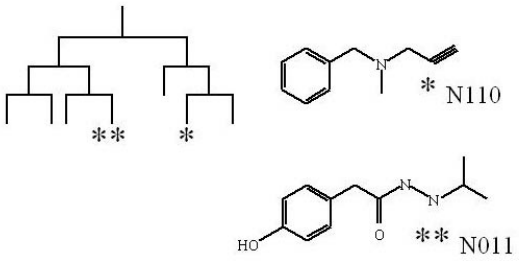


Figure 3: Tree and active compound classes identified.

Also given are representative compounds from two of the terminal nodes. These compounds act through different mechanisms to block the MAO enzyme, see references in Rusinko *et al*. [7]

A data set of 20989 compounds with 4 tumor responses was obtained from the NCI website.

Multivariate recursive partitioning was run. Figure 4 gives a skeleton tree with blowups of two of the terminal nodes. Terminal node N0101 has a

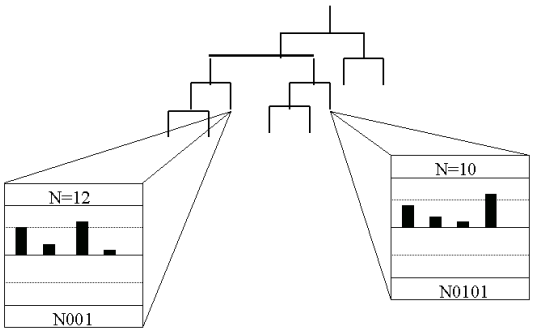


Figure 4: Multivariate recursive partitioning tree, NCI data.

relatively high incidence of the first and last tumor types, Lung and Melanoma, and a relatively low incidence of the second and third tumor types, Colon and Breast. Terminal node N001 has a high incidence of the first and third tumor types. The bits in the node names note the absence, 0, or presence, 1, of chemical features characteristic of compounds in the terminal nodes.

An internal data set of 1444 compounds with IC50 values for the kinase CDK2 was analyzed using typical pharmacophoric features, H-bond donor, H-bond acceptor, *etc*. [3] Multiple conformations were computed and distance between features were binned. After each split, constrained conformations were computed. A total of about 1.4M conformations were computed and the analysis took about 14 hr. CPU time. The resulting recursive partitioning tree is given in Figure 5. The resulting 3D pharmacophore was comparable to crystal structure results, Figure 6.

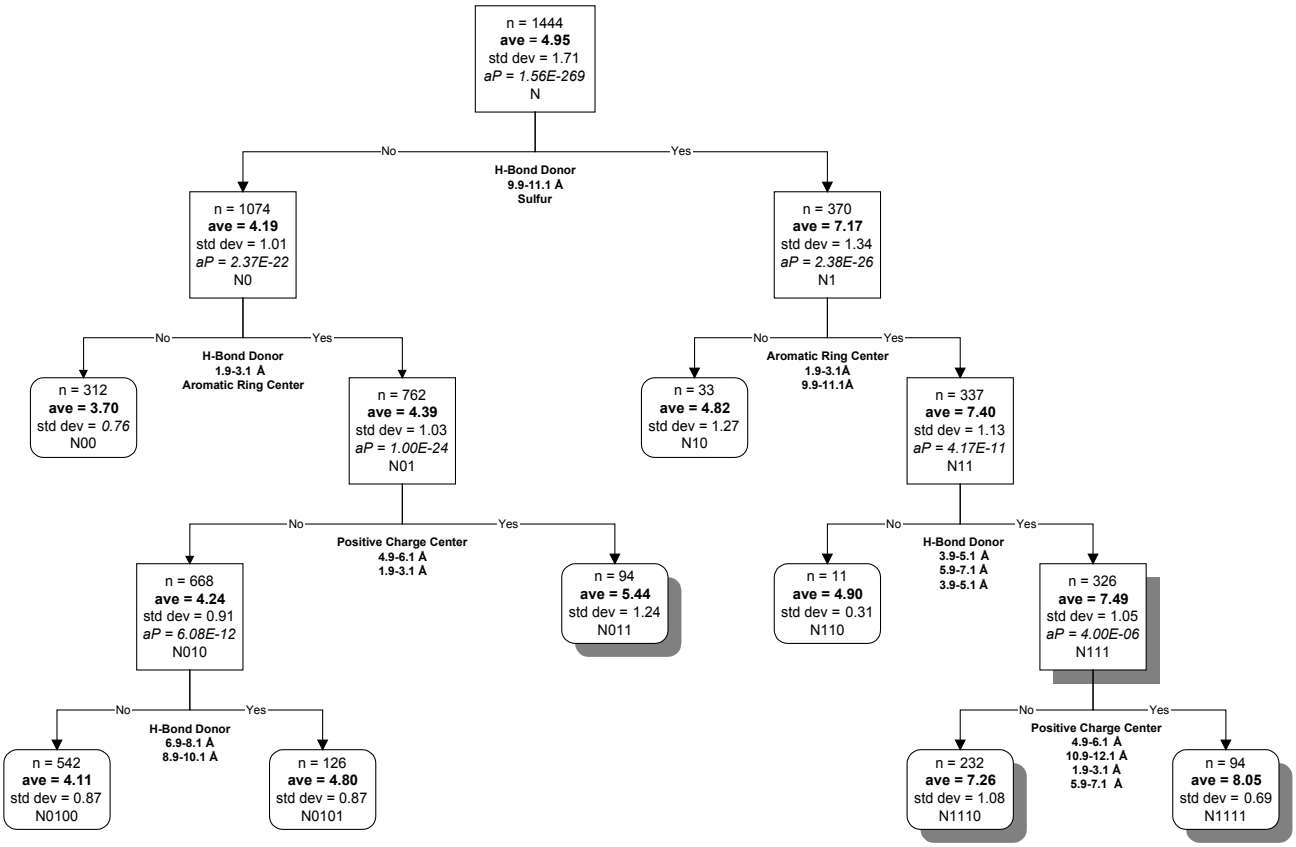
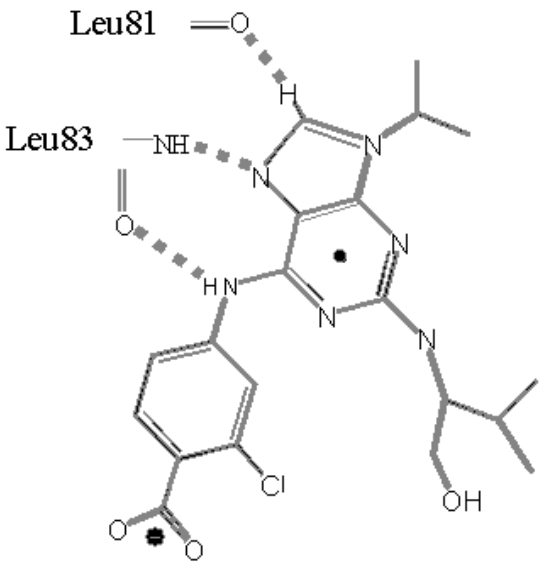
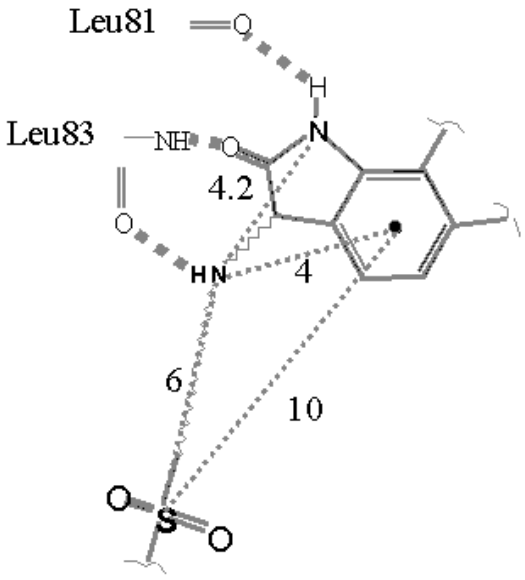


Figure 5: 3D recursive partitioning tree for CDK2 data set.



CDK2-purvalanol B interactions determined by X-ray (*Science* **1998**, 281, 533)



Key interaction features of CDK2 inhibitors determined by 3DSCAM.

Figure 6: Node N111 in CDK2 Tree

DISCUSSION

The key problem to be overcome in the analysis of HTS data sets is that there are likely to be multiple, biological mechanisms. Some molecules may act through one mechanism and others by another. Some might bind in one orientation, others in a different orientation or even at a different location. In the case of the Abbott MAO data set, two mechanisms are known and compounds following each mechanism are found by recursive partitioning. For a large HTS data set there are likely to be multiple mechanisms and even for a single binding pocket, different compounds might bind in different orientations. Most statistical methods assume that there is one underlying model of a single process. If there are two processes, e.g. regular binding site and alosteric binding site, then the features important for one process are very unlikely to be important for the other. Most statistical methods, *e.g.* linear regression, will average the effect for each feature over the two processes. Results are likely to be bad and could be entirely misleading. Recursive partitioning is a simple statistical method that can deal with multiple mechanisms. A feature is identified and the data split based upon this feature. If the feature is important for a specific mechanism, then compounds with that feature (and binding by that mechanism) are separated out from the main body of the data. Following this set of compounds, the analysis is limited to just these compounds; other compounds in the data set have no affect on the subsequent analysis. In this manner, multiple mechanisms can be identified.

A second important problem with HTS data is that assay results for individual compounds are often only crudely determined. Speed and cost are important aspects of HTS. The main goal is to rapidly eliminate the vast majority of compounds

from further consideration. Recursive partitioning does not depend upon a single assay value. Recursive partitioning is driven by averages of compounds with a specific feature and averages are much more stable than single assay values. The node average is the average of all the compounds that have the features that lead compounds into that node. Because the recursive partitioning process is driven by averages, the derived structure-activity rules can have great statistical validity; p-values less than 10^{-100} are common even if the measured effects, increases in binding of less that five percent, are small.

A great deal of effort has been expended implementing these algorithms to make these codes fast. Univariate recursive partitioning runs in seconds for modest data sets, twenty five thousand compounds and ten thousand descriptors. Multivariate recursive partitioning is also fast. This speed has proven to be very useful. Obviously, time is money so completing an analysis quickly can help speed a drug to market. Just as important is that the speed can be used to explore alternative analyses. Medicinal chemists and biologists can interact with the data in real time increasing the likelihood that alternatives are considered and good decisions are made. The statistical methods are rigorous, *e.g*. p-values are adjusted for multiple testing, [9] and help keep the exploratory analysis soundly based.

Atom pairs and topological torsions could be criticized as too simple to be of use for structure activity determination. It is clear that binding into a protein is a three dimensional process; optical isomers often have very different effects. Knowledge of the binding conformation would seem to be essential for good SAR determination. It is clear both theoretically and empirically that these descriptors do capture some structural information. Our empirical results demonstrate that these simple descriptors, coupled with recursive partitioning, are effective in building simple, but useful, structure-activity models.

Building three dimensional pharmacophore models from large data sets is a challenge. We report here on modestly sized data sets, less than 2,000 compounds, where IC50 data is available. Computational speed for 3D recursive partitioning is good relative to commercial codes, but it would be helpful to increase speed. We are working on methods to increase speed with the goal of real-time analysis. In theory, 3D pharmacophore models should be better than 2D methods, but the superiority of 3D over 2D is largely undemonstrated. We plan benchmarking studies to address this question.

REFERENCES AND NOTES

[1] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, 25, 64.

[2] Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1987**, 27, 82.

[3] Chen, X.; Rusinko, A. III; Tropsha, A.; Young, S. S. *J. Chem. Inf. Comp Sci.* **1999**, 39, 887.

[4] Morgan, J. A.; Sonquest, J. N. *J. Amer. Stat. Assoc.* **1963**, 58, 415.

[5] Hawkins, D.M.; Kass, G.V. In *Topics in Applied Multivariate Analysis*; Hawkins, D. H., Ed.; Cambridge University Press, **1982,** 269.

[6] Hawkins, D.M.; Young, S.S.; Rusinko, A. III. *Quant. Struct.-Act. Relat.* **1997**, 16, 1.

[7] Rusinko, A, III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comp Sci.* **1999**, 39, 1017.

[8] Hotelling, H. *Ann. Math. Stat*. **1931** 2, 360.

[9] Miller, R. G. *Simultaneous Statistical Inference*. Springer-Verlag, **1981**