PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation

Kejun Liu,† Jun Feng,† and S. Stanley Young*

National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, North Carolina 27709

Received May 5, 2004

Ideally, a team of biologists, medicinal chemists and information specialists will evaluate the hits from high throughput screening. In practice, it often falls to nonmedicinal chemists to make the initial evaluation of HTS hits. Chemical genetics and high content screening both rely on screening in cells or animals where the biological target may not be known. There is a need to place active compounds into a context to suggest potential biological mechanisms. Our idea is to build an operating environment to help the biologist make the initial evaluation of HTS data. To this end the operating environment provides viewing of compound structure files, computation of basic biologically relevant chemical properties and searching against biologically annotated chemical structure databases. The benefit is to help the nonmedicinal chemist, biologist and statistician put compounds into a potentially informative biological context. Although there are several similar public and private programs used in the pharmaceutical industry to help evaluate hits, these programs are often built for computational chemists. Our program is designed for use by biologists and statisticians.

INTRODUCTION

The purpose of this software system is to create an operating environment for biologists and statisticians for viewing or browsing medium to large molecular SD files (Capacity is only limited by available memory; we have routinely worked with files of 50k compounds.), computing descriptors useful for judging if a compound is drug-like, and finding near-neighbor compounds from annotated chemical collections. In recent years the need has increased for the nonchemist scientist to study relatively large chemical data sets. Toxicologists are interested in commercial and private compounds; biologists now screen tens to hundreds of thousands of compounds and want to look at and evaluate molecular structures of the active compounds. Even at large companies, medicinal chemistry can be a limited resource. Biologists that work at small biotech companies may not have ready access to medicinal chemists. Statisticians are being asked to help in the high throughput screening process, designing screening sets and statistically evaluating HTS data; they typically are not trained in chemistry or computational chemistry. So there is a need for a range of useful techniques along with annotated data sets to be embedded in software to provide some of the functionality of medicinal and computational chemists to scientists of other disciplines.

In our particular case, we are working with biologists from across the world studying Huntington's disease. They are doing basic biology and setting up and running medium throughput screens, typically less than 50k compounds. They would like to see the compounds used in their screen and have ways to evaluate the hit compounds. In addition to being able to view the hit compounds, they can compute properties to help judge if the hit is drug-like (Reactive Group Present, Blood-Brain Penetration, Molecular Weight, logP, Number of Hydrogen Bond Acceptors, Number of Hydrogen Bond Donors, and Polar Surface Area). They can do similarity searching for compounds from their "hits" file against annotated databases. Here a "hit" compound is put into the context of the near neighbors in the annotated database. The annotated data sets serve as a surrogate for a medicinal chemist.

There are a number of molecular file display programs;[1−5] often these programs contain other useful functionality, clustering etc. Biologists are conducting more screening in cells or whole animals so there is a need to put the screening results of active compounds into possible biological context. Chemical genetics[5−10] has as a guiding strategy, the screening in intact biological systems so that biological targets might be simultaneously identified along with active compounds. For example versions of the gene that is mutant in Huntington's disease gene (Htt) have been inserted into a variety of mammalian cells as well as into model organisms: yeast, *Saccharomyces cerevisiae*, nematode, *Caenorhabditis elegans*, and fruit fly, *Drosophila melanogaster*. Under experimental conditions, expression of the mutant Htt genes generates toxicity, variously measured as cell death, cell dysfunction or as a decline in growth rate. The screen consists of treating the organism with a test compound and looking for a decline or reversal in these toxic effects. There is interest in sorting out possible mechanisms of active compounds; annotated, structural near-neighbors might suggest the down stream consequences of the inserted Huntington's gene. For example, what is the mechanism of a compounds increase the life span of a nematode with the Hunginton's gene?

One possible strategy for determining a biological mechanism is to screen only compounds of known mechanism. When an active compound is found, we hypothesize that the compound perturbs its known mechanism and that it is likely

† These authors contributed equally to this paper.

* Corresponding author e-mail: young@niss.org. Corresponding author address: National Institute of Statistical Sciences, P.O. Box 14006, RTP, NC 27709.

10.1021/ci049847v CCC: $30.25
Published on Web 01/28/2005

to be responsible for observed activity. Root et al. (2003) recommend this strategy, and they give an annotated database to use for this purpose.[11] They used the database to sort out possible mechanisms in a cell-based cancer screen: If a hit compound is similar to an annotated compound, then it might be operating through the same mechanism.

There are reasons to have additional strategies to help identify mechanisms. In most screening projects there is the capacity to screen a large number of compounds, and most of these compounds will not have any assigned mechanisms. Also cell-based or organism-based screens will have proteins that are analogous to, but not identical to, human proteins. It is useful to have a strategy to deal with compounds where the screened compounds do not have an assigned mechanism. We want the strategy to be sensible even if the cell or organism does not have an exact copy of a human protein.

Our strategy is to use existing knowledge contained in databases. We use the ACL and Tocris databases to examine possible mechanisms. We use other data sets to assess mutagenicity,[12] and three sets of compounds, P-glycoprotein (P-gp) substrates (201, 116+/85-), human intestinal absorption (HIA) (196, 131+/65-) and compounds that induce torsades de pointes (TdP) (361, 85+/276-), reported by Xue et al.[13] We are using a "guilt by association" strategy; we expect our hit compound to behave like compounds with similar structures. If the active compound is similar to an annotated compound we will hypothesize that the active compound is operating through the same mechanism. To follow this strategy, we numerically describe the hit compound and compute its similarity to the compounds in the annotated database. We read off the mechanisms of the compounds similar to the active compound and make a judgment on how similar the compounds are and how plausible the suggested mechanism is.

There are a great number of ways to numerically characterize a compound; see Chapter 3 of Leach and Gillet (2003)[14] or the encyclopedic book of Todeschini and Consonni (2000), hereafter TC.[15,16] There are two typical styles of descriptors, bit strings of 0's and 1's and vectors of continuous numbers. If a bit string is used, each bit typically notes the presence or absence of a particular molecular fragment. If a vector of continuous numbers is used, the elements of the vector typically note the values of molecular properties, e.g. molecular weight, polar surface area, etc.

Given molecular descriptors computed both on the target molecule and all the molecules in the annotated database, we compute the similarity between the target and each candidate molecule. For bit string similarity we use the Jaccard distance, called the Tanimoto distance in chemistry literature. For continuous variable vectors we use the Euclidian distance. An extensive coverage of molecular similarity is given in Johnson and Maggiora (1990).[17]

METHODS − CHEMICAL DESCRIPTORS

Six molecular descriptor sets are used. Four are bit string and two are continuous.

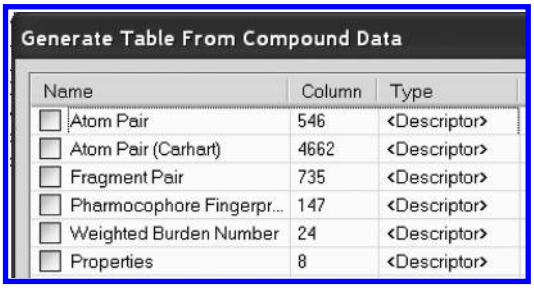
For the bit string descriptors, each bit is set to "1" when a certain feature is presented and "0" when it is not. We adopt the Carhart strategy where a feature refers to two chemical groups or atoms separated by a certain 2D path length;[18] we consider all path lengths up to seven bonds. For example, two phenyl rings, which are separated by two bonds, are expressed as Ph_02_Ph. We implemented four binary descriptors: atom-based, fragment-based, and pharmacophore-based. Atom-based feature pair descriptors are useful for finding very close analogues, while fragment- or pharmacophore-based descriptors are aimed at finding more diverse analogues. Note that we deviate from Carhart in noting presence/absence rather than using counts of features.

Table 1. List of the Carhart Atom Types Used in PowerMV

C(1,0)	C(2,0)	C(3,0)
C(4,0)	C(1,1)	C(2,1)
C(3,1)	C(1,2)	C(2,2)
O(1,0)	O(2,0)	O(1,1)
O(2,1)	N(1,0)	N(2,0)
N(3,0)	N(4,0)	N(1,1)
N(2,1)	N(3,1)	S(1,0)
S(2,0)	S(2,1)	S(1,1)
S(3,1)	S(4,2)	F
Cl	Br	I
P(4,1)	Si	B
Se	As	Y

Table 2. Fragment-Based Descriptors

5 or 6-member aromatic rings	primary, secondary, and tertiary amine, with positive charge or not
carboxylic acid, sulfinate, sulfone	trifluromethyl, nitrile, nitro, sulfonamide,
halogen	double bond center
triple bond center	disulfide, ester
hydroxyl, thiol, primary amine	aliphatic ring centers
amide, carboxylic ester	ketone, imine, thione
ethyl	methyl



For Carhart-style atom-based descriptors, 36 single atom based features are defined; see Table 1. Each atom is further defined based on number of its bonded non-hydrogen atom connections and number of pi electrons. For example, C(1,0) means the carbon atom bonded with 1 heavy atom and has 0 pi electrons, like the carbon atom in the methyl group. Halogen atoms only have one possibility, (1,0), in organic molecules, so their extended notation is ignored. All undefined atom features are assigned to feature Y.

For fragment-based descriptors, 14 classes are defined; see Table 2.

Fragment-based descriptors were built based on bioisosteric principles. For example, the disulfide (−S−) is often used to replace ester group (−O−), so we assign these two groups to the same type. This type of thinking leads to our pharmacophore-based descriptors, giving six classes; see Table 3.

These features are widely used in 3D pharmacophore identification so we include them for similarity-based searching.

In addition to binary descriptors, continuous descriptors can be used for nearest neighbor searching. For continuous

Table 3. Pharmacophore-Based Descriptors

An atom bearing a formal negative charge or groups such as carboxylic, sulfinic, tetrazole, and phosphinic acids
An atom bearing a formal positive charge or groups such as nitrogen in primary, secondary and tertiary amines
Hydrogen bond donor, oxygen or nitrogen atom with hydrogen attached
Hydrogen bond acceptor, oxygen or nitrogen atom with a lone pair electron
Aromatic center, any five- or six-member aromatic ring system
Hydrophobic center, a fragment in which most atoms are hydrophobic atoms, like aliphatic carbon ring systems or aliphatic carbon chains with few heteroatom substitutions

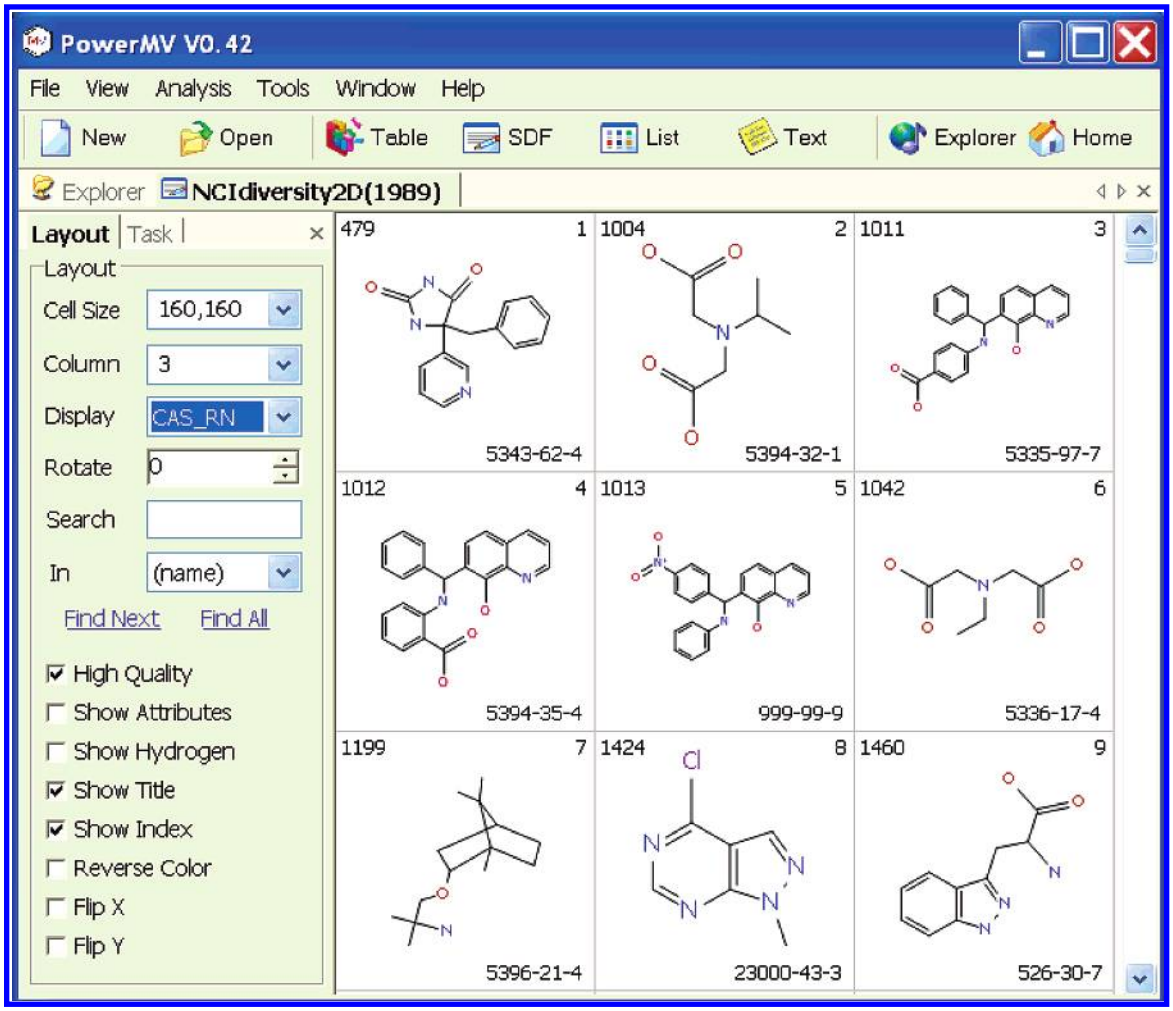


Figure 1. PowerMV: Molecular Viewer screen.

descriptors, Tanimoto distance is replaced with Euclidean distance to measure similarity. The continuous descriptors we implemented are a variation on the Burden number.[19] We place one of three properties on the diagonal of the Burden connectivity matrix:[15,19] electro negativity, Gasteiger partial charge or atomic lipophilicity, XLogP.[20] It is common to scale the off diagonal elements of the connectivity matrix[21] before computing eigen values. The off-diagonal elements were weighted by one of the following values: 2.5, 5.0, 7.5 or 10.0. We use the largest and smallest eigen values. This procedure gives us a total of 24 numerical descriptors. This procedure is similar to the method used by Dr. Pearlman calculating his BCUT descriptors. Dragon software, TC, also has Burden Number inspired eigen value descriptors. All three methods are computed somewhat differently, but all are inspired by Burden. Finally, we can compute eight descriptors useful for judging the drug-like nature of a molecule, XlogP, PSA,[22] number of rotatable bonds, H-bond donors, H-bond acceptors, molecular weight, blood-brain indicator and bad group indicator.

METHODS − NEAR NEIGHBOR SEARCHING

For bit string similarity we used the Tanimoto coefficient, TC, pages 395−400 and for continuous descriptors we used Euclidian distances, TC, Page 396. In both cases, the distance from the target compound to the near neighbors in the candidate database is computed.

Data Sets. ACL Compounds. Root et al. (2003)[11] assembled a collection of ∼2000 biologically active compounds from Sigma-Aldrich Co. by identifying all compounds that were either approved as drugs or were annotation reporting some biological activity. Three representative members of the compound class were chosen where there were many compounds in a class. Structurally distinct FDA-approved small molecule drugs from the electronic Orange

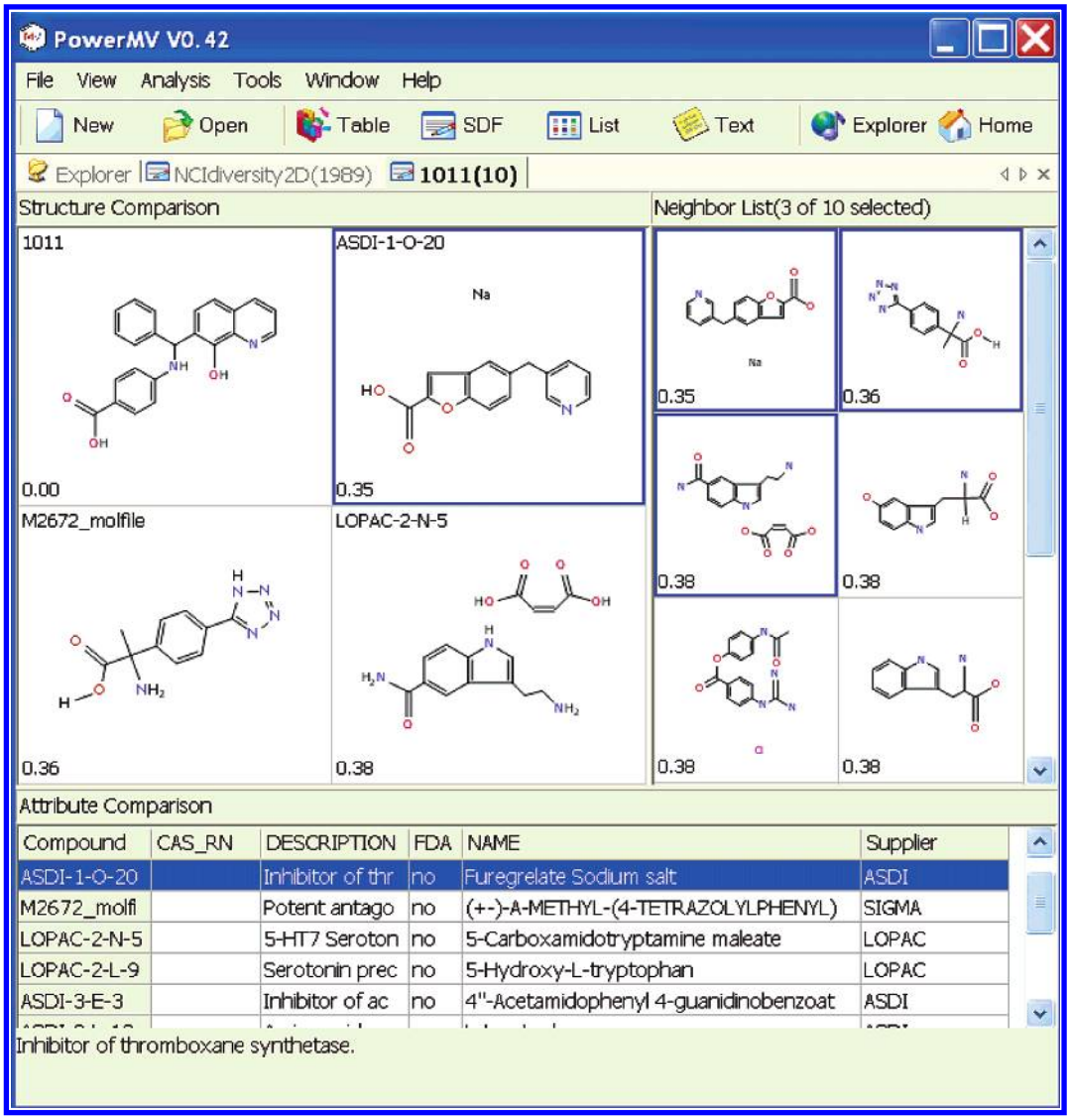


Figure 2. PowerMV: This screen contains three sections, Structure Comparison, Neighbor List and Compound Attributes.

Book were included. This annotation was largely computer generated, and it should be noted that there are duplicate compounds in the database.

Tocris Catalog Compounds. Tocris is a compound vendor, and they make available an electronic file of 939 biologically active compounds. Literature citations and biological activity class are given on the Tocris web site. See data set availability.

Mutagenicity. The mutagenicity data set[12] includes 1806 compounds where roughly half the compounds are mutagenic.

NCI Diversity Set. The NCI created a diversity set of 1990 compounds. They were selected based on 3D pharmacophore features from the 71,756 for which they had one gram of compound. See data set availability.

Other Literature Data Sets. Three different pharmacokinetic and toxicological properties are given in ref 13.

Programming and Features. The PowerMV package was written in Visual C# and C++; it runs under Microsoft .NET framework. PowerMV includes the following features.

1. Importing, viewing and sorting SD files (Capacity is only limited by available memory.). 2D compound viewing of multiple compounds in a r-rows by c-columns grid. The data set can be sorted by numerical or alphabetic attributes. The compound structures and compound attributes can be exported to Microsoft Excel (Office XP and above). Full screen compound viewing of single compounds is also supported. In the Compounds can be flipped and rotated to orient one compound to another to facilitate structural comparisons.

2. Users can generate their own annotated database for similarity searching and results viewing. Input for generating a database is a SD file. PowerMV will automatically generate descriptors for the compound set and save the chemical structures, attributes as well as the descriptors in a single database file. The saved database becomes an annotated database for similarity searching.

3. Fast searching. The descriptors for the candidate databases are pre-computed so for a search only descriptors for the target compound need to be calculated. An index-based file format is used to store the database, which leads to fast search speed. Similarity searching a two thousand compound ACL database takes less than one second on a Intel Xeon 2.4GHz workstation.

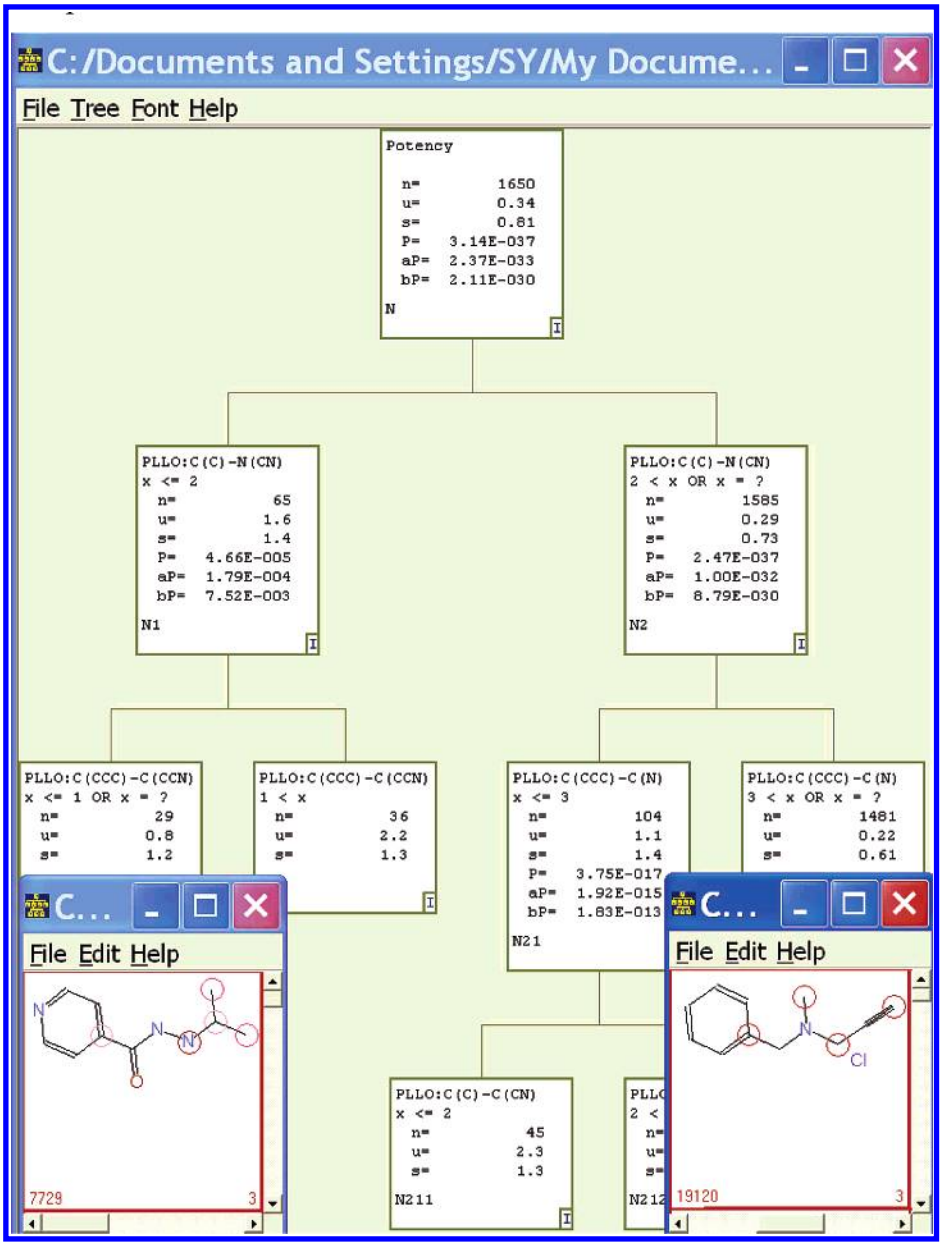


Figure 3. MAO Recursive partitioning tree from ChemTree including active compounds from two active nodes.

Examples. PowerMV is designed to view compound data sets and facilitate the comparison of a target compound to near neighbor compounds in an annotated compound set. The candidate data set is typically annotated with useful properties. Using near neighbors the scientist can judge the likely characteristics of the target compound. The program actually constitutes an operating environment for handling compound data sets, viewing their contents, and helping scientists put compounds into a useful context. We omit from this paper a description of the operating environment and concentrate our discussion on viewing and similarity searching. Figure 1 is a screen shot of the molecular viewer.

The user can control the cell size and number of columns of compounds viewed. The corners of the cell are used to display the compound Title, the Index number in the database, and one user defined characteristic. Under the Task tab, the user can export data sets in various forms, e.g. Excel spreadsheets, and do other compound and file manipulations. A compound in the target data set can be right-click selected and used as a target for a near neighbor search against any of the annotated databases with any of the similarity methods.

The results of a near neighbor search are given in Figure 2.

Compound 1101 was selected from the NCI Diversity data set, and the Tanimoto distance using atom pair descriptors was computed to all the ACL compounds. The user specified 10 near neighbors, and the 6 nearest neighbors are displayed on the right-hand side of the screen shot. Three of the nearest neighbors were Ctrl-click selected, and the program moved them to the Structure Comparison panel. The compounds can be flipped and rotated to aid in the visual comparison of the compounds. The Attribute Comparison panel can be used to display text and numerical descriptions of the compounds. The most similar compound is ASDI-I-O-20, an inhibitor of thromboxane synthetase. Selecting a compound in the Structure Comparison window will move text giving its biological annotation at the bottom of the Attribute Comparison panel.

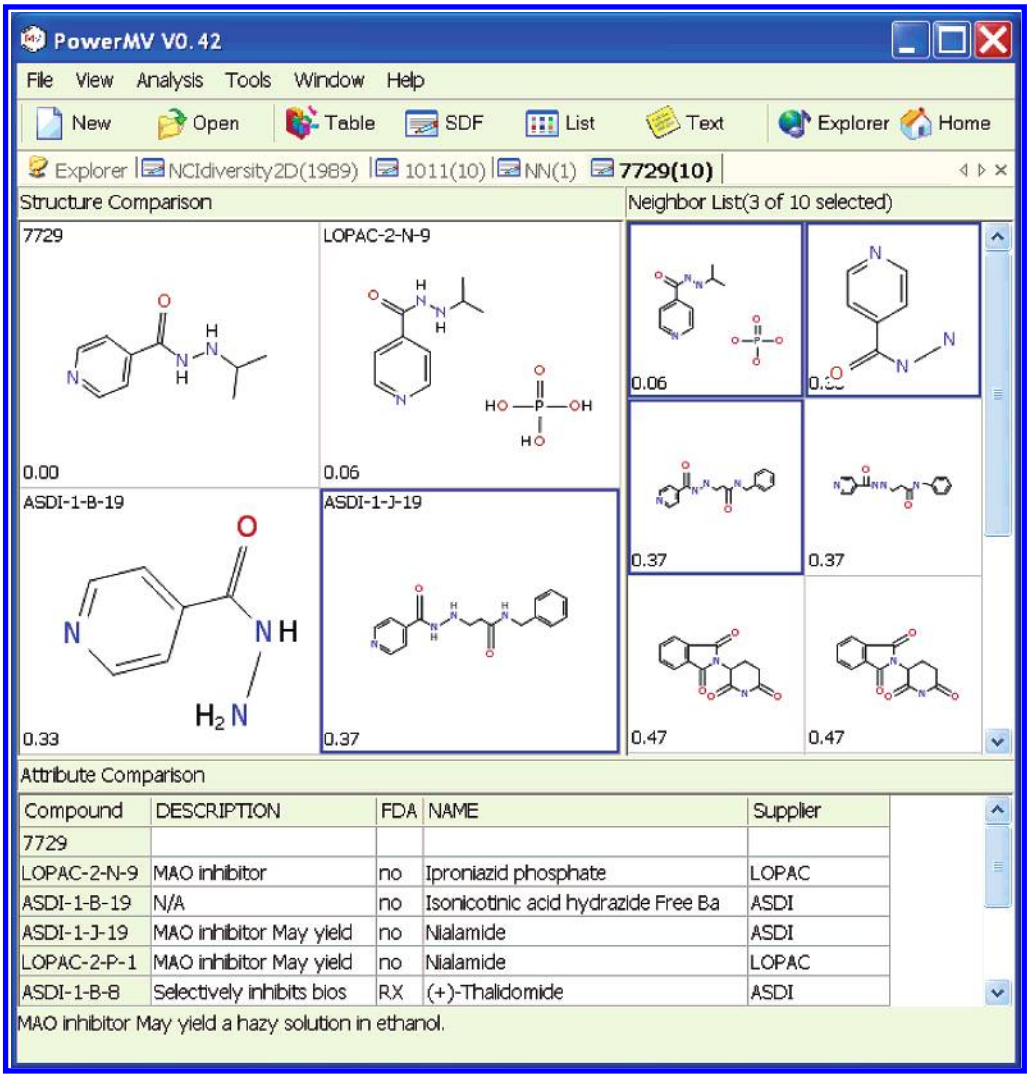


Figure 4. Results of search using an active MAO compound as the target and the ACL annotated compound data set as the candidate.

The results of a recursive partitioning analysis are given in Figure 3; see Rusinko et al. (1999)[23] and ChemTree-(2004),[24] of the well-studied Abbott MAO data set.

This data set is instructive because there are compounds active by each of two mechanisms. The data set is a mixture of inactive compounds and compounds active by these two mechanisms. One of the hallmarks of one mechanism is a CC triple bond, C#C. One hallmark of the other mechanism is two adjacent nitrogens, NN. In Figure 3 the left most terminal node contains NN compounds and node N211 contains C#C compounds. The atoms judged important by the ChemTree algorithm are marked in the compound displays, and these are in general agreement with MAO pharmacophore literature. Here we know the mechanism, but to demonstrate the approach we pretend that we do not and we conduct a near neighbor search of the ACL database to evaluate these two compounds, 7729 and 19120.

Figure 4 gives the results of the near neighbor search, using compound 7729 as the target and ACL as the candidate database.

We selected the three nearest neighbors from the Neighbor List, and they are displayed in the Structure Comparison panel. Note that LOPAC-2-N-9 is an exact match except for the salt form. (The distance would be 0.0 without the salt present.) There is no mechanism given for ASDI-1_B-19 (we might guess it is a MAO inhibitor). We do see that the next two compounds are identical and contain the NN feature; the compound is a MAO inhibitor. (Remember that the candidate database was computer generated and contains duplicate compounds.)

Figure 5 gives the results of the near neighbor search, using compound 19120 as the target.

We selected the three nearest neighbors from the Neighbor List, and they now appear in the Structure Comparison panel. LOPAC-2-H-8 is an exact match. LOPAC-2-F-16 is the next nearest neighbor and is a MAO-B inhibitor. The third compound in the Structure Comparison panel has the C#C but is not noted to be an MAO inhibitor.

DISCUSSION

This program and the annotated data sets are meant to be an aid in following up on active compounds when the assay is phenotypic. This program is designed for biologists (screeners) and statisticians. Computational chemists will have a number of programs that cover most of the functions of PowerMV. The program is simple to run and contains many of the functions useful for examination of initial screening hits. Our motivating example is Huntington's disease, and many of our screens involve cells or genetically

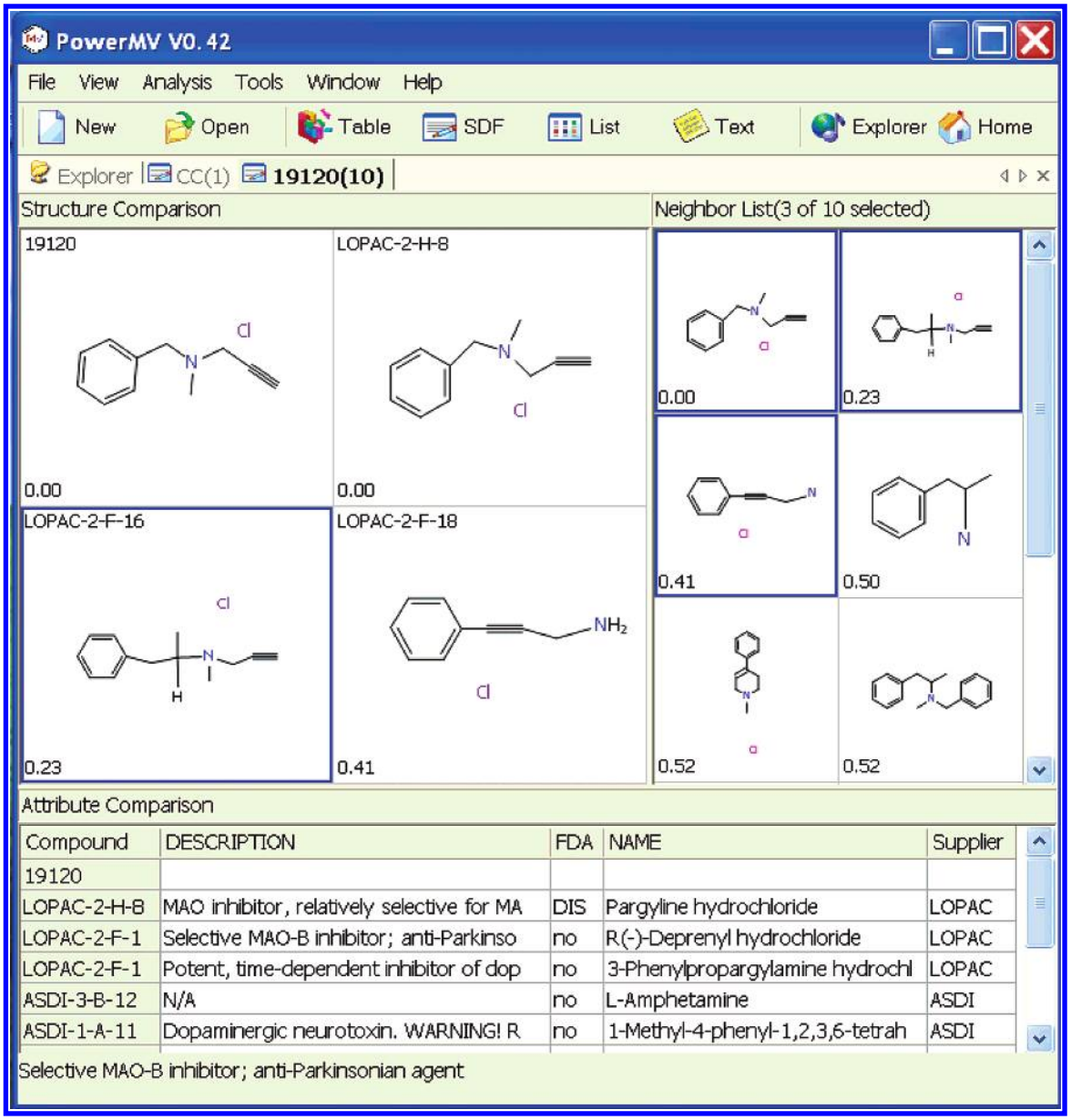


Figure 5. Results of search using an active MAO compound as the target and the ACL annotated compound data set as the candidate.

modified animals, fruit fly or the worm. The screened organisms are complex so one of the first questions with a hit is what mechanism might the compound be affecting. We would like for the compound to be affecting the introduced, knock in, genetic modification. There are several experimental ways to follow up a hit. Compounds similar to the active compound can be ordered and tested. We might find more active compounds. We might also begin to see chemical features that are consistent across the active compounds. The second experimental follow up is to order and test additional compounds that have similar biological action to the action suggested by the annotated database. Root et al. (2003) did this in their screening where tumor cells were sensitive to valinomycin, an ionophore. Other ionophores were ordered and tested and the killed tumor cells as well. So the annotated chemical database can point to a mechanism, and we can order biologically similar compounds in addition to structurally similar compounds.

We have taken existing annotated chemistry data sets to demonstrate the features of PowerMV. These data sets are instructive of what should be possible, but they are far from ideal. Work is starting on annotated chemical databases, ChemBank of ICCB at Harvard, PubChem of ICBI, and ChEBI of European Bioinformatics Institute.[25−27] Annotation is difficult for a number of reasons. First, there is the problem of detail that is known about the compounds. For some compounds all we know is that they kill bacteria. For others we have much more detailed knowledge, e.g. the compound is a selective CB-1 inhibitor. It would be useful to have an ontology for representing chemical information. For example, we might organize our knowledge in a hierarchy: there are ∼500 drug targets.[28] For each target there are known active compounds. For each compound we can give a compound name, CAS number, and chemical structure, SD or smiles. There is an enormous range of possible explanations or level of biological detail of mechanism. A biology ontology would also be useful. Pointers to literature can help, and some of the databases we use do have pointers to literature, e.g. Tocris. Obviously, more complete annotation would be helpful. Second, there is an enormous amount of scientific

literature. Human searching through that literature is very difficult and time-consuming. The annotations in the ACL data set were generated using automated text mining methods. Creating and fitting existing knowledge into a useful ontology is labor intensive.

So in the whole scheme of things, PowerMV is modest. Biologists and statisticians can easily view compounds and compute basic numerical descriptors to judge if a compound is drug-like. Using near neighbor searches against annotated chemical databases, the scientist can judge the possible biological mechanism of the compound. With better-annotated databases, other properties could also be judged, e.g. expected toxic effects. Clearly, the near-neighbor searching utility of PowerMV depends on the quality of the annotated databases.

We have added robust singular value decomposition, a clustering method, k-means, and a statistical prediction method, Least Angle Regression, to PowerMV, but they will be discussed elsewhere.

Software Availability. PowerMV can be downloaded from www.niss.org/PowerMV.

Data Set Availability. The ACL set is described by Root et al. (2003) and can be downloaded from http://jura.wi.mit.edu/stockwell/StockwellLab/supplement/supplement_ACL.html. See their "Table S2. Compressed SD file for the compounds in the ACL." The Tocris compounds can be viewed at the Tocris web site http://www.tocris.com/shop/inital_page.php. The Mutagenicity data set can be obtained from the contact author. The P-glycoprotein, human intestinal absorption and torsades de pointes data sets are available from Xue et al. (2004). The Abbott MAO data set can be obtained from Yvonne Martin, yvonne.c.martin@abbott.com.

ACKNOWLEDGMENT

Jun Feng is supported by High Q Foundation. Jack Liu is supported by CIIT and NISS.

REFERENCES AND NOTES

(1) Wild, D. J.; Blankley, C. J. VisualiSAR: A Web-based application for clustering, structure browsing, and structure−activity relationship study. *J. Mol. Graphics Modeling* **1999** 17, 85−89.
(2) Shen, J. HAD: An Automated Database Tool for Analyzing Screening Hits in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2003** 43, 1668−1672.
(3) VIDA **2004** commercial, http://www.eyesopen.com/products/applications/vida.html.
(4) SDEditor **2004** freeware, http://www.issware.com/sdeditor.html.
(5) LeadNavigator **2004** commercial, http://www.lionbioscience.com/leadnavigator.html.
(6) Stockwell, B. R. Chemical genetics: ligand-based discovery of gene function. *Nat. Rev. Genet.* **2000** 1, 116−125.
(7) Schreiber, S. L. The small-molecule approach to biology. *C&E News* March 3, **2003** Pages 51−61.
(8) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004** 5, 262−275.
(9) Chemical genomics **2004** http://www.broad.mit.edu/broad/huicg.html.
(10) Schreiber, Chemical Genomics http://www-schreiber.chem.harvard.edu/home/pdffiles/8109genomics.pdf.
(11) Root, D. E.; Flaherty, S. P.; Kelley, B. P.; Stockwell, B. R. Biological mechanism profiling using an annotated compound library. *Chem. Biol.* **2003**, *10*, 881−892.
(12) Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S. S. Predictive toxicology: Benchmarking molecular descriptors and statistical methods. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 1463−1470.
(13) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen Y. Z. Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci.* (in press).
(14) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics;* Kluwer Academic Publishers: Dordrecht, 2003; Chapter 3.
(15) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany 2000.
(16) DRAGON is commercially available from http://www.dist.umimib.com.
(17) Johnson, M. A.; Maggiora, G. M. Eds. *Concepts and Applications of Molecular Similarity*; Wiley-Interscience: New York, 1990.
(18) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Defnition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(19) Burden, F. R. Molecular identifcation number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(20) Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom additive method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47−66.
(21) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(22) 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.
(23) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Statistical analysis of a large structure-biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *38*, 1017−1026.
(24) ChemTree **2004** www.goldenhelix.com.
(25) ChemBank **2004** http://chembank.med.harvard.edu/compounds/.
(26) PubChem **2004** http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pccompound.
(27) ChEBI, Chemical Entities of Biological Interest http://www.ebi.ac.uk/chebi/.
(28) Drews, J. Drug discovery: A historical perspective. *Science* **2000** 287, 1960−1964.

CI049847V