# Visualization of Large Chemical Databases with Application to Drug Design

Dexuan Xie[*]

January 16, 2005

**Abstract**

This paper reviews our recent studies on the visualization of chemical database and its application in drug design at the molecular and cellular levels. It sketches the background of computer-aided drug design, and gives the mathematical formulations of fundamental database computational problems. Solutions of these problems rely on multivariate nonlinear optimization, combinatorial optimization, numerical linear algebra, and multivariate statistical analysis. This paper presents a fast algorithm for visualizing large chemical databases as a first step in solving the chemical database analysis problems. The algorithm is a novel combination of the singular value decomposition and principal component analysis techniques with the truncated-Newton minimization method (TNPACK) based on the classic distance geometry approach. Numerical experiments to real chemical databases show that our visualization algorithm has great potential in aiding the generation of drug candidates or the optimization of bioactive compounds.

## 1 Computer aided Drug Design

Before the 1970s, new drug candidates mostly were proposed from laboratory syntheses or extractions from Nature. Later, the methodology of "rational drug design" was introduced as the understanding of biochemical processes is increased, as the technology of computer is improved, and as the field of molecular modeling is recognized widely. Molecular modeling was thought to lead to dramatic progress in the design of drugs. However, the limited reliability of modeling molecular interactions between drugs and target molecules restricted its success. Indeed, the design of compounds with the correct binding properties is only a first step in the complex process of drug design; many other considerations and tests must be made to determine the drug's bioactivity and its various effects on the human body.

Since 1980s, new high throughput screening (HTS) synthesis techniques, such as robotic systems that can run hundreds of concurrent synthetic reactions, have emerged, enhancing synthesis productivity enormously [5, 6]. The rapid development of the HTS synthesis techniques makes many people to believe that drug design need not be "rational" if potential lead drugs can be searched by an exhaustive approach. However, as the HTS synthesis techniques are becoming cheaper and faster, chemical databases are becoming larger and larger. The amount of compounds to be synthesized becomes so vast that no a pharmaceutical company can afford the money and time required by an exhaustive approach even with the most advantaged HTS synthesis techniques.

In order to sharply reduce the time and money for identification and optimization of lead drugs [15], virtual HTS computational techniques have been developed, with which a huge database can be sorted through into a small focused dataset (e.g., a dataset in which compounds have the most potential to the activity with respect to a given target) or a small "most diverse" dataset that can be a

[*] Department of Mathematics and Ph.D. Program in Scientific Computing, University of Southern Mississippi, Hattiesburg, MS 39406, Dexuan.Xie@usm.edu.

representative of the huge database. Such small datasets eliminate most of guesswork in drug design for scientists so that the discovery process of a new drug can be significantly speeded up compared to traditional discovery methodologies. Combining the virtual HTS techniques with molecular modeling and database analysis techniques has become a very promise approach to computer-aided drug design. It has been applied to propose candidate molecules that resemble antibiotics, to find novel catalysts for certain reactions, and to design inhibitors for the HIV protease.

## 2 Chemical Databases and Chemical Descriptors

A chemical database consists of compounds with potential and/or demonstrated therapeutic activities. Most libraries are the properties of pharmaceutical companies, but public sources also exist. Examples of public databases include the MDL Drug Data Report (MDDR) database [1] and the Comprehensive Medicinal Chemistry (CMC) [2]. MDDR contains various information on known drugs and those under development, including descriptions of therapeutic action and cross-reference search utilities. Produced by MDL and Prous Science, MDDR contains over 100,000 biologically relevant compounds and well-defined derivatives, with updates adding about 10,000 a year to the database. CMC database provides 3D models and various biochemical properties of drugs, including drug class, logP, and pKa values for over 7,500 pharmaceutical compounds (1900-present). MDL updates CMC annually.

To analyze a chemical database by a virtual HTS technique, a corresponding virtual database needs to be established. That is, all compounds of the database are characterized as vectors by a list of "chemical descriptors". A basic assumption on such a virtual database is that the Euclidean distance of two compound vectors indicates the structure diversity (or similarity) of the two corresponding chemical compounds. Virtual databases will be simply referred to as databases in the remainder of this paper since all databases to be studied in our research work have vector expressions for compounds.

How to define chemical descriptors to satisfy the above basic assumption is an area of research on its own, and has been widely studied for more than twenty years. Generally, chemical descriptors can be a wide range of indices that indicate molecular connectivity, Kappa shape, topological state, electrotopological state, and other properties. There are many ways to formulate chemical descriptors, but the search for the most appropriate descriptors is an ongoing enterprise.

Examples of different classes of descriptors include: (1) molecular connectivity or topological indices that reflect molecular connectivity and other topological invariants; (2) binary encoded representations that indicate the presence or absence of a property, such as at least three nitrogen atoms, doubly-bonded nitrogens, or alcohol functional groups; (3) 3D structural indices that reflect geometric structural factors like van der Waals volume and surface area; and (4) electronic indices that characterize the ionization potential, partial atomic charges, or electron densities. See also [4] for further examples.

Some chemical descriptors can be computed on a computer. For example, more than 300 structural descriptors can be calculated by using Molconn-Z—a commercial program package of eduSoft. LC, which is originally developed by Lowell Hall [19]. Such descriptors have been found to be a convenient and reasonably successful approximation to quantify molecular structure and relate structure to biological activity (see review in [3]) and can be used in conjunction with other selectivity criteria based on activity data for a training set (e.g., [11, 16]).

Many commercial databases have contained compound vectors characterized by chemical descriptors. For example, the compounds in MDDR have vector expressions in terms of atom-pair descriptors, which are defined as the number of bonds along the shortest paths connecting an atom-pair [7], and several other descriptors that reflect torsions of four consecutive atoms, charges of atom pairs, and hydrophobic characteristics in order to reflect physicochemical and geometric features.

[1] http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=20

[2] http://www.mdli.com/cgi/dynamic/product.html?uid=$uid&key=$key&id=21

These compound vectors have been shown to be successful in structure-activity studies [3, 7].

# 3 Computational Problems in Database Analysis

In broad terms, two general problem categories can be defined in chemical database analysis and design: (1) Database systematics: analysis and compound grouping, compound classification, elimination of redundancy in compound representation (dimensionality reduction), data visualization, etc., and (2) Database applications: efficient formulation of quantitative links between compound properties and biological activity for compound selection and design optimization experiments. Both these general database problems are associated with several mathematical disciplines. Database systematic involves the tools of multivariate statistical analysis and numerical linear algebra. Similarity and diversity searches involve multivariate nonlinear optimization (for continuous formulations), combinatorial optimization (for discrete formulations), distance geometry techniques, and configuration sampling. The problem of combinatorial optimization, in particular, is known to have a non-polynomial time complexity.

Typically, the above combinatorial optimization problems can be solved by stochastic and heuristic approaches. These include genetic algorithms, simulated annealing, and tabu-search variants. See reference [2], for example, for a review. As in other applications, the efficiency of simulated annealing strongly depends on the choice of cooling schedule and other parameters. In recent years, several potentially valuable annealing algorithms such as deterministic annealing, multiscale annealing, and adaptive simulated annealing, as well as other variants, have been extensively studied.

In special cases, combinatorial optimization problems can be formulated as integer programming and mixed-integer programming problems. In this approach, linear programming techniques such as interior methods, can be applied to the solution of combinatorial optimization problems, leading to branch and bound algorithms, cutting plane algorithms, and dynamic programming algorithms. Parallel implementation of combinatorial optimization algorithms is also important in practice to improve the performance.

Other important research areas in combinatorial optimization include the study of various algebraic structures (such as matroids and greedoids) within which some combinatorial optimization problems can more easily be solved [8].

Our present studies focus on two fundamental research problems in database analyses: (1) the similarity sampling problem, and (2) the diversity sampling program [3, 12, 17, 21]. Their solution has great potential in benefitting computer-aided drug design [5, 6]. The *similarity problem* involves finding a drug from the database that is similar to another drug with known bioactive properties. The *diversity problem* involves defining a diverse subset of “representative” compounds so that researchers can scan only a subset of the huge database each time a specific pharmacological agent is sought. However, this formidable problem is a *combinatorial optimization.*

We used an innovative approach to solve the above two database problems. The approach consists of two steps. In the first step, a fast algorithm for dimensionality reduction and data visualization is developed so that the distance relationships of the database are displayed visually in the two-dimensional vector space. In the second step, algorithms for similarity sampling and diversity sampling are developed.

The main task of the first step has been completed recently. In the next section, a brief summary of the main results of this study is presented. It includes a mathematical framework of database analysis, an application of the principle component analysis (PCA) and *singular value decomposition* (SVD) methods in database reduction, and a fast database visualization algorithm defined by combining PCA and SVD with the truncated-Newton minimization method (TNPACK) based on the distance-geometry approach.

# 4 Mathematical Framework of Database Analysis

A mathematical framework of database analysis is essential to apply mathematics and computer sciences to the solution of database problems. It includes the definition of database, the measure of similarity and diversity, the scaling of database, and the mathematical formulation of basic database problems.

## 4.1 Expression of Chemical Database

Let $X$ be a chemical database of $n$ compounds. The $i$th compound $X_i$ is expressed as a vector of $m$-dimensional space $R^m$:

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T \quad \text{for } i = 1, 2, \ldots, n,$$

where $x_{ij}$ denotes the value of chemical descriptor $j$ of compound $X_i$. The database $X = \{X_1, X_2, \ldots, X_n\}$ is then represented as an $n \times m$ matrix

$$X = (X_1, X_2, \ldots, X_n)^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}. \tag{1}$$

Here the superscript $T$ denotes the transpose of vector and matrix. Usually, $n >> m$.

## 4.2 Diversity Measures

Formally, a diversity measure $\delta$ of database $X$ is a function from $X \times X$ into a real number space $R$ that is nonnegative ($\delta(X_i, X_j) \geq 0$), symmetric ($\delta(X_i, X_j) = \delta(X_j, X_i)$), and vanishes on the diagonal ($\delta(X_i, X_i) = 0$). It measures how different compounds $X_i$ and $X_j$ are in some sense.

There are many possible diversity measures, but all of them are equivalent since $R^m$ is of finite dimension. For simplicity, the *Euclidean distance* is often used as a diversity measure of two different compounds $X_i$ and $X_j$:

$$\delta_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

Here $\|\cdot\|$ is the Euclidean norm. There are $n(n-1)/2$ distance segments $\{\delta_{ij}\}$ in database $X = \{X_i\}_{i=1}^n$ for pairs $i < j$.

## 4.3 Database Scaling

To avoid the dominance of a few large descriptors on the diversity measure, scaling becomes necessary if the chemical descriptors use different units and vary drastically in their magnitudes. Generally, scaled descriptors $\{\hat{x}_{ik}\}$ can be defined by the following formula: For $k = 1, 2, \ldots, m$,

$$\hat{x}_{ik} = a_k(x_{ik} - b_k), \qquad 1 \leq i \leq n, \tag{2}$$

where $a_k$ and $b_k$ are real numbers, and $a_k > 0$. They are called the scaling and displacement factors, respectively. In practice, however, it is very difficult to determine the appropriate scaling and displacement factors for the specific application problem [25]. Given no chemical/physical guidance, it is customary to scale data entries $x_{ij}$ by the following formula:

$$\hat{x}_{ij} = \frac{x_{ij} - \beta_j}{\alpha_j - \beta_j}, \tag{3}$$

where $\alpha_j = \max_{1\leq i\leq n} x_{ij}$ and $\beta_j = \min_{1\leq i\leq n} x_{ij}$. This definition makes each column in the range $[0,1]$, and is also termed "standardization of descriptors".

Another scaling procedure arises from the multivariate statistical analysis. It sets the scaled component $\hat{x}_{ij}$ as follows:

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{4}$$

where $\mu_j$ and $\sigma_j$ are the mean and the deviation of the $j$th column of $X$, respectively, defined by

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \text{and} \quad \sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}.$$

This scaling effectively makes each column have a mean of zero and a standard deviation of one.

## 4.4 The Formulation of Database Computational Problems

The fundamental computational problems in database analysis are formulated mathematically as below.

**(1) The Similarity Search Problem.**

Suppose that compound $X_j$ is a given target. The similarity search problem for a database $X$ is to find the most similar compound $X_{j0}$ such that

$$\|X_{j0} - X_j\| = \min_{\substack{\forall X_i \in X \\ i \neq j}} \|X_i - X_j\|.$$

Since each distance segment $\delta_{ij} = \|X_i - X_j\|$ requires $\mathrm{O}(m)$ floating-point operations (flops) to compute, an exhaustive calculation over all $n$ candidates requires a total of $\mathrm{O}(nm)$ flops. An effective scheme is sought when $n$ and $m$ are large.

**(2) The Similarity Sampling Problem**

In practice it is often to find a sampling set that contains a small number of most similar compounds to the target $X_j$. Such a task is called *the similarity sampling problem.* Note that the vector expression $X_i$ of compound reflects the structure features of compound. Hence, according to the principle that similarity in structure may lead to similarity in bioactivity, it can be expected that performing this task can find these drugs with similar physiochemical and biological properties to a known drug.

**(3) The Diversity Search Problem**

The diversity search problem is to find a "representative subset" $\mathcal{S}_0$ that contains $n_0$ representative compounds such that

$$\sum_{\substack{X_i, X_j \in \mathcal{S}_0 \\ i<j}} \|X_i - X_j\| = \max_{\forall \mathcal{S} \subset X} \sum_{\substack{X_i, X_j \in \mathcal{S} \\ i<j}} \|X_i - X_j\|,$$

where $n_0 \ll n$, a fixed positive integer number, and $\mathcal{S}$ denotes a subset of the database $X$ with $n_0$ compounds. Diversity search naturally arises since pharmaceutical companies must scan huge databases each time they search for a specific pharmacological activity. Reducing the dataset of $n$ compounds to the small subset of $n_0$ representative elements is likely to accelerate such searches. The "representative subset" $\mathcal{S}_0$ might also be used to prioritize the choice of compounds to be purchased or/and synthesized, resulting in an accelerated discovery process.

An exhaustive search of the most diverse subset $\mathcal{S}_0$ requires a total of $\mathrm{O}(C_n^{n_0} n_0^2 m)$ flops because there are $C_n^{n_0}$ possible subsets of $\mathcal{S}$ of size $n_0$ and each subset takes $\mathrm{O}(n_0^2 m)$ flops. Here $C_n^{n_0} = n(n-1)(n-2)\cdots(n-n_0+1)/(1 \cdot 2 \cdot 3 \cdot \cdots \cdot n_0)$.

**(4) The Clustering Problem**

Clustering is a process of organizing objects into groups whose members are similar in some way. That is, for a given integer $n_0$, the database $X$ is partitioned into $n_0$ subsets $S_1, S_2, \ldots, S_k$ such that the sum of their costs, $\sum_{j=1}^{n_0} c(S_j)$, is minimized. Here each $S_j$ is referred to as a cluster of database, and $c(S_j)$ is a cost function of $S_j$. The cost function can be defined by

$$c(S_j) = \sum_{\substack{X_i, X_j \in S_j \\ i<j}} \|X_i - X_j\|.$$

Obviously, a representative set generated from a diversity searching algorithm can be used to produce the clusters of the database. In fact, if $\bar{X}^j$ is the $j$th representative compound, then the above cost function can be approximated as

$$c(S_j) = \sum_{X_i \in S_j} \|X_i - \bar{X}^j\|.$$

Here $\bar{X}^j$ is often called a centroid of cluster $S_j$. On the other hand, if a cluster partition is given, a representative compound $\bar{X}^j$ of cluster $S_j$ can be selected by performing the following task: Find $\bar{X}^j \in S_j$ such that

$$\sum_{X_i \in S_j} \|\bar{X}^j - X_i\| = \min_{x \in S_j} \sum_{X_i \in S_j} \|x - X_i\|.$$

In the above sense, both the diversity search problem and the clustering problem are equivalent.

**(5) The Diversity Sampling Problem**

However, it seems to be impossible to find the exact solution of the diversity search problem or the clustering problem because their combinatorial feature is an impossible obstacle. The number of possible subsets increases in a high rate with $n$, the search for an optimal solution is out of reach despite the enormous power of large scale parallel computers. Hence, in practice, it is always to look for heuristics or for approximate solution in some more restricted feasible sets. The diversity sampling problem is to generate a diversity sampling—an approximate solution of the diversity search problem with a satisfactory accuracy.

# 5 Our Recent Progresses

In our studies, a promise approach is used to solve the above database problems. The approach consists of two steps. The first step is to create a 2D mapping of the database with a satisfactory accuracy in retaining the distance relationships of the database. Based on the 2D mapping of the database, the second step is then to develop algorithms for solving the above database problems together with the applications in drug design. How to develop efficient visualization algorithms for large databases is a challenging research topic in this approach. This section presents the research progresses we made recently on this topic.

## 5.1 Database Visualization by PCA

To reduce the dimensionality of database $X$ by using PCA approach, the chemical descriptors are regarded as random variables, while the input database $X$ as their sampling. PCA transforms the highly correlated descriptor variables into the uncorrelated variables called **principal components** (PCs). By using the first $l$ PCs with $l << m$, the database matrix $X$ can then be reduced to a smaller matrix with dimension $n \times l$. It is well known that this small database matrix is an "optimal" approximation to the input database matrix $X$ in the sense that it retains the variation presented in $X$ in the $l$-dimensional ($l$-D) space as much as possible [18].

According to our recent paper [30], the $l$-dimension approximation $Y_i$ of $X_i$ can be defined by using the first $l$ PCs as:

$$Y_i = (y_{i1}, y_{i2}, \dots, y_{il})^T \quad \text{for } i = 1, 2, \dots, n. \tag{5}$$

Here $y_{ij}$ is the $i$th component of the $j$th PC. For the purpose of visualization, $l$ is set to be 2 or 3.

## 5.2 Database Visualization by SVD

The SVD technique [9] is a procedure for data compression used in many practical applications like image processing and cryptanalysis (code deciphering). Essentially, it is a factorization for rectangular matrices that is a generalization of the eigenvalue decomposition for square matrices. With SVD, the database rectangular matrix $X$ can be written as the sum of rank-1 matrices [14]:

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^T, \tag{6}$$

where $r$ is the rank of matrix $X$ ($r \leq m$), $u_k \in R^n$ and $v_k \in \mathcal{R}^m$, respectively, are left and right singular vectors, and $\sigma_k$ is the singular value. All singular values are arranged in decreasing order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \quad \text{and} \quad \sigma_{r+1} = \ldots = \sigma_m = 0.$$

In terms of the right singular vectors $\{v_k\}_{k=1}^m$, which is here an orthonormal basis of $R^m$, each compound vector $X_i$ has the following new vector expression

$$X_i = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, \dots, \sigma_r u_{ir}, 0, \dots, 0)^T, \tag{7}$$

where $u_{ij}$ is the $i$th component of the $j$th left singular vector $u_j$. Based on (7), the $l$ dimensional approximation vector $Y_i$ of $X_i$ is defined by

$$Y_i = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, \dots, \sigma_l u_l)^T, \quad i = 1, 2, \dots, n. \tag{8}$$

Here $l << m$. With $l = 2$ or 3, the database can be visualized.

## 5.3 Distance Refinement by TNPACK

In many applications of database analysis, it is important to study the distance relationships among the compounds as much as possible. Hence, it is essential for a visualization algorithm to retain the distance relationships well. However, the SVD/PCA mapping of database $X$ may be poor in retaining the distance relationships of the database. Hence, a step of distance refinement is required to improve the accuracy of the SVD/PCA mapping in retaining the distance relationships of the database.

The problem of distance refinement can be formulated as a classic distance-geometry problem: find $n$ points in the $l$-dimensional space $R^l$ ($2 \geq l << m$) so that their interpoint distances match the corresponding values from the $m$-dimensional space $R^m$ as closely as possible.

Specifically, with an objective error function $E$ to describe the discrepancy between the distance segments in $R^l$ and $R^m$, an optimization algorithm finds a minimum point $Y^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)$ with $Y_i^* \in R^l$ for $i = 1, 2, \ldots, n$ such that

$$E(Y_1^*, Y_2^*, \ldots, Y_n^*) = \min_{Y_i \in R^l,\ 1 \le i \le n} E(Y_1, Y_2, \ldots, Y_n), \tag{9}$$

where each $Y_i = (y_{i1}, y_{i2}, \ldots, y_l)^T$.

Theoretically, the distance-geometry problem (9) is a global minimization problem, but in practice, only a local solution can usually be obtained. Hence, how to find an initial guess of solution that is near enough to the global optimal solution is an important and difficult objective.

**Objective Error Function** $E$**.** The objective error function $E$ can be formulated in many different ways [1, 22, 23]. In our algorithm, a particular objective error function $E$ is used:

$$\begin{aligned} E(Y_1, Y_2, \ldots, Y_n) &= \frac{1}{4} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} \left( d(Y_i, Y_j)^2 - \delta_{ij}^2 \right)^2, \qquad (10) \\ \omega_{ij} &= \begin{cases} 1/\delta_{ij}^4 & \text{if } \delta_{ij}^4 \ge \eta, \\ 1 & \text{if } \delta_{ij}^4 < \eta, \end{cases} \end{aligned}$$

where $d(Y_i, Y_j) = \|Y_i - Y_j\|$, $\delta_{ij} = \|X_i - X_j\|$, $\{\omega_{ij}\}$ denote weights, and the parameter $\eta$ is a small positive number such as $10^{-12}$. Note that the first and second derivatives of $E$ are well defined. Hence, an efficient second-derivative method like Newton-type algorithms [13] can be applied.

**The Quality Measure of Distance Approximation.** Various error measures can be used to assess the agreement between the original and projected pairwise distances. In our algorithm, the following percentage $\rho$ is used to measure the quality of the approximation of $d(Y_i, Y_j)$ to $\delta_{ij}$ for all pairs $i < j$:

$$\rho = \frac{T_d}{n(n-1)/2} \cdot 100. \tag{11}$$

The variable $T_d$ is the total number of the distance segments $d(Y_i, Y_j)$ satisfying

$$|d(Y_i, Y_j) - \delta_{ij}| \le \epsilon \delta_{ij} \quad \text{when } \delta_{ij} > d_{min}, \tag{12}$$

or

$$d(Y_i, Y_j) \le \tilde{\epsilon} \quad \text{when } \delta_{ij} \le d_{min}, \tag{13}$$

where $\epsilon, \tilde{\epsilon}$, and $d_{min}$ are given small positive numbers less than one. For example, we set $\epsilon = 0.1$ to specify a 10% accuracy ($d_{min} = 10^{-12}$ and $\tilde{\epsilon} = 10^{-8}$). The second case above (very small original distance) may occur when two compounds in the datasets are similar highly. The greater the $\rho$ values, the better the mapping and the more information can be inferred from the projected views of the complex data.

**Distance Refinement by TNPACK Minimization.** In our algorithm, a $l$-D mapping of database is first generated by SVD or PCA. When the accuracy of the SVD/PCA projection is not high enough (i.e., the value of the percentage $\rho$ defined in (11) is too small), the algorithm goes to the step of distance refinement: solve the distance-geometry problem (9) by our efficient truncated Newton program package, TNPACK [24, 27, 28]. Here, the SVD/PCA projection is selected as the initial iterate of TNPACK, and the TNPACK iterations are carried out until the $k$-th TNPACK iterate $Y^k$ satisfies

$$||g(Y^k)|| < \epsilon_g (1 + |E(Y^k)|), \tag{14}$$

where $\epsilon_g$ is a small positive number (we used $10^{-5}$), and $g$ is the gradient vector of $E$. Such an $Y^k$ defines the $l$-D projection of the database.

TNPACK was first published in 1992 [24] and updated recently [27, 28]. Our algorithm uses the new version of TNPACK [27]. One of the features of TNPACK is an application-tailored preconditioner matrix (that approximates the Hessian of the objective function) used to accelerate convergence [26]. This novel preconditioner makes TNPACK an efficient tool for the minimization of molecular potential functions in comparison to other available minimizers [10, 26]. For the present applications, we only used a simple preconditioner, namely, the diagonal part of the Hessian, or terms $\partial^2 E(Y_1, Y_2, \ldots, Y_n)/\partial y_{ik}^2$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, l$.

# 6 The Visualization Program Package

A visualization program package called SIEVER (SIngular Values and Error Refinement) has been developed. The core computing part of the package is a Fortran 77 program routine of our visualization algorithm described in the previous section. The minimization program package TNPACK [27] and the eigenvalue program package ARPACK [20] are adopted to SIEVER for computing distance refinements and the eigenvectors required by the PCA and SVD procedures. ARPACK allows users to compute only the first $l$ eigenvalues and eigenvectors at a complexity of the order $nm^2$ (with $n > m$) floating point operations per iteration; storage requirements are of order $nl$. Hence, it is effective for our application.

Recently, a web-based graphic interface system was established for SIEVER [32] so that SIEVER can be used on the Internet. This system has a user-friendly webpage for a user to input datasets and the initial parameters required by SIEVER, and to receive the graph of the 2D mapping of the dataset and the 3D structure graphs of compounds that are generated from the home website of SIEVER. The input dataset is assumed to be in the matrix form (1). The 3D molecular structure files and bioactivity files are required to be in the same ordering as that of the compounds in the database. With the 2D mapping graph of the dataset, the user can quickly make similarity sampling and diversity sampling virtually. By clicking a point of the 2D mapping graph, a compound of the dataset is selected for sampling, and a graphic window is generated to show the 3D molecular structure, chemical formula, bioactivities, and other important information of the compound.

The interface system contains the following programs:

- A HTML program to generate a webpage for users to input and receive data on the internet environment.
- A PERL-CGI program for processing the input data and passing the processed data to the core computing program. It also triggers the running of the core computing program. This program uses the CGI (Common Gateway Interface, see http://www.cgi-resources.com) technology, which is a standard interface for information servers.
- A PERL program for plotting the 2D mapping of dataset with a Linux plotting tool called Gnu-Plot. The graph of the 2D mapping can be zoomed to view the details. It also carries out the action of sending the 2D mapping graph to the user's website.
- A JAVA program for displaying 3D molecular structure and bioactivity information on compounds. The input 3D structure files are in the PDB format, which is a widely-used format for describing biomolecular structures. This program also sets up a cross link among the 2D mapping file, the 3D structure files, and the bioactivity files. Clicking a point of the 2D mapping graph gives a graphic window that shows the 3D structure of the corresponding compound as well as its bioactivities and other important chemical properties.

The website of SIEVER is still in construction. As soon as completed, it will be opened to the public.

# 7 Application Examples

Many experiments with over ten chemical datasets (of size 58 to 27255 compounds) have been made to demonstrate the effectiveness of our algorithm in visualizing databases and database analyses

Table 1: The accuracy of the PCA and PCA/TNPACK mappings as a function of projection dimension $l$ for DATASET1 ($n = 832$, $m = 300$), along with the performance of the PCA and PCA-TNPACK (TN) programs. Here the minimization objective function $E$ is defined in (9), and the accuracy $\rho$ is defined in (11) with $\eta = 0.1$.

| | Final $E$ | | Accuracy $\rho$ | | TN | CPU time | |
|---|---|---|---|---|---|---|---|
| Dimension $l$ | PCA | TN | PCA | TN | Itn. | PCA (sec.) | TN (min.) |
| 2 | $4.07 \times 10^4$ | $1.80 \times 10^4$ | 2.45 | 33.70 | 8 | 0.20 | 0.23 |
| 3 | $3.29 \times 10^4$ | $1.06 \times 10^4$ | 2.89 | 44.72 | 10 | 0.21 | 0.42 |
| 5 | $2.32 \times 10^4$ | $4.87 \times 10^3$ | 7.52 | 63.16 | 11 | 0.33 | 0.65 |
| 8 | $1.71 \times 10^4$ | $2.17 \times 10^3$ | 17.07 | 80.80 | 14 | 0.57 | 1.05 |
| 12 | $1.17 \times 10^4$ | $9.68 \times 10^2$ | 28.86 | 93.49 | 12 | 0.72 | 1.63 |
| 16 | $8.83 \times 10^3$ | $5.18 \times 10^2$ | 42.35 | 98.12 | 46 | 0.88 | 2.33 |
| 20 | $6.38 \times 10^3$ | $3.12 \times 10^2$ | 53.10 | 99.50 | 29 | 1.01 | 2.80 |
| 25 | $4.98 \times 10^3$ | $2.01 \times 10^2$ | 61.40 | 99.88 | 19 | 3.26 | 6.13 |
| 30 | $3.51 \times 10^3$ | $1.11 \times 10^2$ | 71.06 | 99.99 | 18 | 2.80 | 6.90 |
| 35 | $2.67 \times 10^3$ | $7.22 \times 10$ | 78.42 | 100 | 20 | 2.27 | 9.00 |
| 45 | $1.73 \times 10^3$ | $3.63 \times 10$ | 87.13 | 100 | 16 | 4.50 | 6.55 |
| 65 | $7.32 \times 10^2$ | $1.01 \times 10$ | 95.69 | 100 | 19 | 6.28 | 19.02 |

[31, 29, 30]. This section only shows three of them. All computations were performed in double precision on a single R12000 300 MHz processor of an SGI Origin 2000 computer at the New York University.

## 7.1 Accuracy as a Function of Projection Dimension

Table 1 shows that the projection accuracy of both PCA alone and the PCA-TNPACK method can be improved significantly as the projection dimension $l$ increases (data shown for $l$ ranging from 2 to 65). It also shows that the distance refinement step implemented by the TNPACK minimization algorithm can significantly and efficiently improve the accuracy of the 2D PCA mapping. For example, in the case of $l = 2$, the accuracy $\rho$ has increased from about 3 to 34% in about 14 seconds of CPU time. The accuracy $\rho$ of the PCA-TNPACK mapping is near maximal ($\rho = 100\%$) for $l = 20$, while PCA alone requires $l = 65$ to reach this maximal accuracy. The dataset in this experiment is called *DATASET1,*, which contains 832 chemical compounds from MDDR with a list of 300 chemical descriptors.

## 7.2 Similarity and Diversity Sampling

From the 2D PCA/TNPACK mapping for DATASET1, we selected four pairs of points that are close in the projection, as shown in Figure 1, and 12 distant points (letters A to L in Figure 2). Their corresponding chemical structures are also displayed in these figures. From these figures we see that the chemical structures of the compounds corresponding to the four pairs of nearby points have some resemblance, while compounds corresponding to the 12 distant points are more dissimilar. According to the similar property principle [17] (i.e., the compounds with similar structures have similar physicochemical and biological properties), we might predict that the 12 distant points belong to different therapeutic categories, while the compounds corresponding to each pair are in the same therapeutic category. Except for pair 1, these predictions were confirmed in [30].

### 7.3 Structure and Medicinal Activity

To examine the relation between structural features (characterized by chemical descriptors) and medicinal activities, DATASET2, which has 338 compounds of MDDR with 300 chemical descriptors, was constructed that contained four different therapeutic groups. The names of the four groups and the number of compounds in each group are listed in Figure 3.

Figure 3 displays the distribution of the four categories based on the 2D PCA-TNPACK mapping of DATASET2. The accuracy $\rho$ for $\eta = 0.1$ (defined in (11)) of the 2D PCA-TNPACK mapping is 36%. With $\eta = 0.2$, we have $\rho = 90\%$. That is, 90% of the distance segments in the PCA/TNPACK mapping are within 20% of the original distance values. This shows that the PCA-TNPACK mapping has reasonably approximated the distance relationships presented in DATASET2.

From Figure 3 we note that the anthelmintic (circles, agents that destroy or cause the expulsion of parasitic intestinal worms) and immunosuppressant (diamonds, agents that depress the immune response of an organism) groups contain two separate clusters, while the antibacterial group and the immunostimulant (agents that stimulate an immune response) group are more spread out — perhaps having some similarities to other compounds. For example, some points from the immunostimulant group are close to some points from the anthelmintic group as shown in Figure 3. This suggest that the corresponding drugs may have similar structures and similar medicinal activities.

Based on the mapping figure in Figure 3, we selected eight distant points (two points per group) as shown in Figure 4. Their chemical structures are also displayed in Figure 4. From this figure we see that compounds belonging to the same therapeutic group may have very different chemical structures. Clearly, structural properties are only one factor in determining complex biological activity.

These numerical results show that the 2D PCA-TNPACK mapping can reasonably retain both the distance relationships of compounds and the distribution patterns of therapeutic groups. Hence, the 2D database mapping might be a valuable tool for similarity and diversity sampling of chemical compounds, and perhaps ultimately for aiding the generation of drug candidates or the optimization of bioactive compounds.

## 8 Future Works

Application of the SVD or PCA procedures to large datasets is straightforward, but the minimization refinement extension remains a challenge. A main difficulty comes from the limitation of computer memory because the objective function $E$ defined in (10) is defined in terms of two $n \times n$ distance matrices, which requires an order of $n^2$ memory locations for the calculation of $E$ efficiently. A possible solution to this problem is to divide a huge database as dictated by computer memory; the refinement procedure could then be applied to each subset of database; finally, a proper assembly technique is needed to combine appropriately these separate 2D mappings into a global representation. Another simple way to solve this problem is to define the objective function $E$ by only using a band distance matrix with a small band width so that the memory requirement is reduced to an optimal order of $n$. In addition to these memory issues, how to design an efficient preconditioner for TNPACK is an important research topic. In fact, the performance of TNPACK for large chemical datasets can be significantly improved by using a more efficient preconditioner than the simple one of the current algorithm. We will study these extensions in detail in our subsequent work.

## Acknowledgments

This work is supported by the University of Southern Mississippi (#2221709006 and #2221709007). The co-researchers of this project are Doctors T. Tamar (Courant Institute of Mathematical Sciences, New York University), A. Tropsha (School of Pharmacy, University of North Carolina), and S. B. Singh and E. M. Fluder (Merck Research Laboratories).

# References

[1] D. K. Agrafiotis. A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Science*, 6:287 – 293, 1997.

[2] D. K. Agrafiotis. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.*, 37:841–851, 1997.

[3] D. K. Agrafiotis. Diversity of chemical libraries. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 1, pages 742–761. John Wiley & Sons, West Sussex, UK, 1998.

[4] D. K. Agrafiotis, J. C. Myslik, and F. R. Salamme. Advances in diversity profiling and combinatorial series design. *Mol. Div.*, 4:1–22, 1999.

[5] L. M. Amzel. Structure-based drug design. *Curr. Opin. Biotech.*, 9:366–369, 1998.

[6] D. B. Boyd. Computer-aided molecular design. In A. Kent (Executive) and C. M. Hall (Administrative), editors, *Encyclopedia of Library and Information Science*, volume 59, pages 54–84. Marcel Dekker, New York, NY, 1997. Supplement 22.

[7] R. E. CArhart, D. H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, 25:64 – 73, 1985.

[8] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver. *Combinatorial Optimization.* John Wiley & Sons, New York, NY, 1998.

[9] J. W. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

[10] P. Derreumaux, G. Zhang, B. Brooks, and T. Schlick. A truncated-Newton method adapted for CHARMM and biomolecular applications. *J. Comp. Chem.*, 15:532–552, 1994.

[11] S. L. Dixon and H. O. Villar. Investigation of classification methods for the prediction of activity in diverse chemical libraries. *J. Comput.-Aided Mol. Design*, 13:533–545, 1999.

[12] G. M. Downs and P. Willett. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.*, 7:1 – 66, 1996.

[13] P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization.* Academic Press, London, England, 1983.

[14] G. H. Golub and C. F. van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, MD, second edition, 1986.

[15] M. Hann and R. Green. Cheminformatics – a new name for an old problem? *Curr. Opin. Chem. Biol.*, 3:379–383, 1999.

[16] P. A. Hunt. QSAR using 2D descriptors and TRIPOS' SIMCA. *J. Comput.-Aided Mol. Design*, 13:453–467, 1999.

[17] M. A. Johnson and G. M. Maggiora. *Concepts and Applications of Molecular Similarity.* Wiley, New York, 1990.

[18] I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[19] L. B. Kier and L. H. Hall. *Molecular Connectivity in Structure-Activity Analysis.* Research Studies Press, John Wiley and Sons, Letchworth, England, 1986.

[20] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Philadelphia, 1998.

[21] Y. C. Martin, R. C. Brown, and M. G. Bures. *Combinatorial Chemistry and Molecular Diversity.* Wiley, New York, 1997.

[22] T. Pinou, T. Schlick, B. Li, and H. G. Dowling. Addition of Darwin's third dimension to phylectic trees. *J. Theor. Biol.*, 182:505–512, 1996.

[23] D. D. Robinson, T. W. Barlow, and W. G. Richard. Reduced dimensional repressentations of molecular structure. *J. Chem. Inf. Comput. Sci.*, 37:939 – 942, 1997.

[24] T. Schlick and A. Fogelson. TNPACK — A truncated Newton minimization package for large-scale problems: I. Algorithm and usage. *ACM Trans. Math. Softw.*, 14:46–70, 1992.

[25] P. Willett. Structural similarity measures for database searching. In P. von Ragué Schleyer (Editor-in Chief), N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, III, editors, *Encyclopedia of Computational Chemistry*, volume 4, pages 2748–2756. John Wiley & Sons, West Sussex, UK, 1998.

[26] D. Xie and T. Schlick. Efficient implementation of the truncated Newton method for large-scale chemistry applications. *SIAM J. Opt.*, 10(1):132–154, 1999.

[27] D. Xie and T. Schlick. Remark on the updated truncated Newton minimization package, *algorithm 702. ACM. Trans. Math. Softw.*, 25(1):108–122, March 1999.

[28] D. Xie and T. Schlick. A more lenient stopping rule for line search algorithms. *Optimization Methods and Software*, 2000. In Press.

[29] D. Xie and T. Schlick. Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization. In C. A. Floudas and P. Pardalos, editors, *Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches*, pages 267–286. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[30] D. Xie, S. B. Singh, E. M. Fluder, and T. Schlick. Principal component analysis combined with truncated-newton minimization for dimensionality reduction of chemical databases. Submitted, 2000.

[31] D. Xie, A. Tropsha, and T. Schlick. A data projection approach using the singular value decomposition and energy refinement. *J. Chem. Inf. Comp. Sci.*, 40(1):167–177, 2000.

[32] J. Xie and D. Xie. A web-based biomolecular information query system. To be Submitted, 2001.

Figure 1: Four similarity samples (Pair 1 to Pair 4) of chemical compounds based on the 2D PCA-TNPACK mapping of DATASET1 and their chemical structures.

Figure 2: A diversity sample (A–L) of chemical compounds based on the 2D PCA-TNPACK mapping of DATASET1 and their chemical structures.

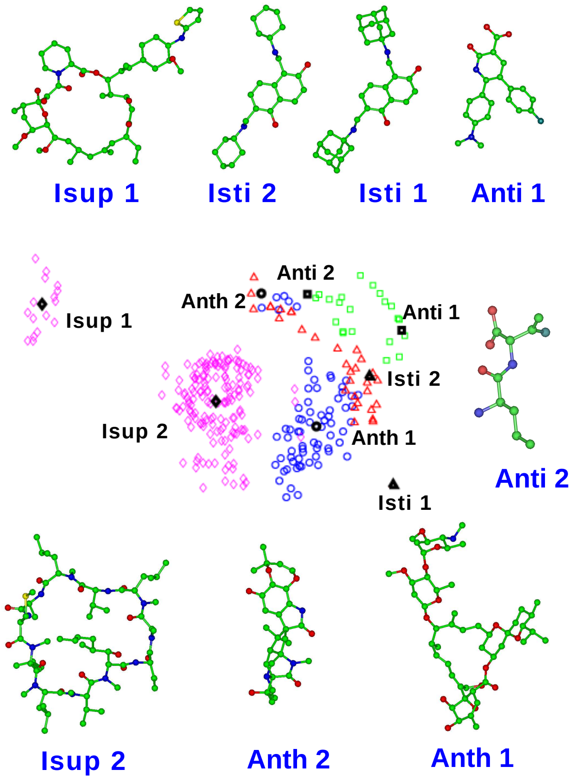

Figure 3: The 2D PCA-TNPACK mapping for DATASET2 ($n = 338, m = 300$). Here DATASET2 contains the four therapeutic groups as indicated in the figure.

Figure 4: The diversity of chemical compounds belonging to the same therapeutic group.