Math. Program., Ser B 95: 161–185 (2003)

Digital Object Identifier (DOI) 10.1007/s10107-002-0345-7

Dexuan Xie · Suresh B. Singh · Eugene M. Fluder · Tamar Schlick

Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases

Received: December 21, 2000 / Accepted: August 19, 2002
Published online: September 27, 2002 –

Abstract. The similarity and diversity sampling problems are two challenging optimization tasks that arise in the analysis of chemical databases. As a first step to their solution, we propose an efficient projection/refinement protocol based on the principal component analysis (PCA) and the truncated-Newton minimization method implemented by our package TNPACK (PCA/TNPACK). We show that PCA can provide the same initial guess as the singular value decomposition (SVD) for the optimization task of solving the distance-geometry optimization problem if each column of a database matrix has a mean of zero. Hence, our PCA/TNPACK approach is analogous to the SVD/TNPACK projection/refinement protocol that we developed recently for visualizing large chemical databases. Using PCA/TNPACK and the Merck MDDR database (MDL Drug Data Report), we further investigate the projection/refinement procedure with regards to the preservation of the original clusters of chemical compounds, the accuracy of similarity and diversity sampling of chemical compounds, and the potential application in the study of structure activity relationships. We also explore by simple experiments accuracy and efficiency aspects of the PCA/TNPACK procedure compared to those of a global optimization algorithm (simulated annealing, as implemented by the program package SIMANN) in terms of producing the projection mapping of a database. Numerical results show that the 2D PCA/TNPACK mapping can preserve the distance relationships of the original database and is thus valuable as a first step in similarity and diversity applications. Of course, the generation of a global rather than local minimizer and its interpretation in terms of pharameceutical applications remains a challenge. Since all numerical tests are performed on the Merck MDDR database, results are representative of realistic cases encountered in the field of drug design, and may help analyze properties of medicinal compounds.

1. Introduction

Two fundamental research themes in chemical database analysis are similarity and diversity sampling [2, 15, 21, 29]. Their solution has great potential in benefiting computer-aided drug design [3, 7]. 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 pharmaco-

D. Xie: Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201-0413, e-mail: dxie@uwm.edu

S. B. Singh, E. M. Fluder: Merck Research Laboratories, Mail Stop RY50-SW100, Rahway, NJ 07065, e-mail: suresh_singh@merck.com, fluder@merck.com

T. Schlick: Departments of Chemistry and Mathematics, Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 251 Mercer Street, New York, NY 10012, e-mail: schlick@nyu.edu

Mathematics Subject Classification (2000): 65K10, 62H25, 92C50

logical agent is sought. The diversity problem is essentially a combinatorial optimization problem, which has a non-polynomial complexity. Hence, even for a small dataset (more than hundreds of compounds), solving the diversity problem represents a grand challenge. In contrast, the similarity problem is much simpler. Still, due to the large size of real databases (millions and billions of compounds), solving the similarity problem is a practical challenge, as exhaustive procedures are not feasible. Hence, all algorithms for addressing these problems are *heuristic,* and any systematic schemes to reduce the computing time involved can be valuable.

Recently, we developed an algorithm for visualizing large chemical databases in a low-dimensional space (2D or 3D) as a first step to the solution of the above two challenging problems [35, 36]. To illustrate, consider n compounds in the chemical database, described as n vectors in the m-dimensional Euclidean space R^m, with associated pairwise distances reflecting the similarity or diversity of the corresponding compounds. Our algorithm combines a distance-geometry approach with the singular value decomposition (SVD) [24]. The distance-geometry approach finds an optimal projection mapping of the database from the high-dimensional space R^m onto the low-dimensional space R^l with $l \ll m$ such that the intercompound distances in R^l approximate the corresponding distances in R^m as closely as possible. In our approach, we formulate the distance-geometry problem by a new "smooth" target function and then minimize this multivariate function by our truncated-Newton package TNPACK [33, 39] with a starting point generated by SVD. Since the SVD starting point is an optimal approximation of the original database in the sense of the Euclidean norm [24], it provides a reasonable starting point for a local minimization algorithm such as truncated Newton. Our application of TNPACK to refine the projection mapping generated by SVD (termed the SVD/TNPACK method) produces distance relationships in such chemical databases that are reasonable [35, 36].

Several distance-geometry algorithms, such as nonlinear mapping [32], multi-dimensional scaling [6], and the Sammon method [1], have been proposed to generate a 2D projection mapping of chemical databases. While they use different target functions of minimization, they rely on the simple steepest descent minimization algorithm with a randomly chosen initial guess. As such, they may suffer from slow convergence and variable quality of the 2D projection mapping. Our application of the truncated Newton method accelerates that local minimization aspect of the problem, but any other efficient large-scale method like the limited-memory BFGS method [11, 12] will likely perform just as well.

Principal component analysis (PCA) [25] and the related factor analysis [25] are two fundamental statistical techniques in multivariate data analysis. Since PCA can produce a compressed database in the low-dimensional space R^l that is optimal in the sense of maintaining the variation of the original database in R^m, PCA provides another promising approach for generating a good initial candidate for refinements which project the data from a high to a low-dimensional space.

In this paper, we formulate the PCA projection procedure and then show that PCA can produce the same projection mapping as SVD if each column of a database matrix has a mean of zero. The relationship we describe here may be of practical utility to researchers in the field, since PCA is widely used.

Similar to the SVD/TNPACK approach, we combine the PCA projection mapping with TNPACK for solving the distance-geometry problem. Numerical results show that while the PCA and SVD mappings are different for matrix databases that have columns with non-zero means, subsequent refinement by TNPACK can significantly reduce the differences in the coordinate scales. Hence, the PCA/TNPACK and SVD/TNPACK mappings generally produce similar accuracies in terms of retaining the distance relationships of the original database.

Furthermore, using PCA/TNPACK and Merck drug compounds (each represented by 300 topological atom-pair descriptors [9]) selected from a commercially available database from MDL, Inc. called MDDR (MDL Drug Data Report), we numerically investigate several important issues (as summarized below) related to the projection refinement approach.

First, we explore through simple tests the degree of improvement possible when the projection mapping of a database is defined by a global instead of a local (as found by PCA/TNPACK) minimum. Of course, the distance-geometry problem is a challenging global optimization task [5, 16, 17], but the computing cost involved is generally prohibitive. Thus, we solve the distance-geometry problem *approximately* by applying an efficient local minimization algorithm such as TNPACK together with a good starting point such as that generated from PCA or SVD.

To compare some accuracy and efficiency aspects of the PCA/TNPACK procedure to that of a global optimization algorithm, we construct a subset of the database with 15 compounds; the corresponding global minimizer of the distance-geometry problem can be easily located by the simulated annealing program SIMANN [10, 23, 22]. Numerical results show that the 2D projection mapping defined by the global minimizer provides minor improvements to the 2D PCA/TNPACK mapping in terms of the accuracy of retaining the database distance relationships. We also compare the performance of SIMANN with that of TNPACK for a dataset of 342 compounds. With a proper choice of starting temperature (obtained by experimentation), SIMANN identifies the same local minimizer as TNPACK does (with the same starting point generated by PCA) but costs more than 5000 times more CPU time. Clearly, global optimization approaches are important, but cost may be prohibitive for realistic applications, and the local minimization approach presents a valuable first step.

Second, we investigate how the 2D PCA/TNPACK mapping preserves distance relationships. Namely, from a dataset of 342 Merck compounds, we randomly select 12 spatially distant points and four nearby pairs of points based on the 2D PCA/TNPACK mapping and compare the similarity/diversity of the original chemical compounds in the database (before projections). Tests show that the chemical interpretation of the 2D mapping appears reasonable — distances between points in the projection roughly correspond to chemical similarity/diversity in terms of chemical structures and bioactivities. We also show that the projection accuracy of both the PCA and PCA/TNPACK mappings can improve significantly as the projection dimension l increases. For example, with $l = 20$ (rather than 2 or 3), almost all of the distance segments in the PCA/TNPACK mapping are within 10% of the original distance values. Though 2D and 3D projections are simpler for graphical illustrations, compressed datasets with higher dimensionality are valuable for compound similarity applications.

Third, we test how the 2D PCA/TNPACK mapping can retain certain clusters of chemical compounds. For this investigation, we construct a subset of MDDR that consists of four clusters of compounds in the 300-dimensional chemical space. We then project this dataset onto 2D by PCA/TNPACK. The four clusters are reasonably preserved in the 2D PCA/TNPACK mapping.

Finally, we apply the 2D PCA/TNPACK mapping to the study of the relationship between structures and medicinal activity. For this purpose, we construct a dataset containing four different therapeutic groups of compounds and visualize these compound interrelationships in 2D by PCA/TNPACK. While the 2D PCA/TNPACK mapping produces reasonably clusters corresponding to the four therapeutic groups, we emphasize that the compounds belonging to the same therapeutic category can have very different chemical structures.

The remainder of the paper is organized as follows. Section 2 defines the database structure. Section 3 reviews the optimization problems arising from database analyses. Sections 4 and 5 describe the PCA and TNPACK mapping algorithms. Section 6 presents the numerical results, and Section 7 summarizes our conclusions. The relation between PCA to SVD (which we have not seen described in our context) is presented in the Appendix.

2. Database definition

We consider a database of n compounds generated from m highly correlated chemical descriptors $\{x_j\}_{j=1}^m$. Each compound X_i is represented as a vector

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

where x_{ij} denotes the value of descriptor j of compound X_i. The collective database $X = \{X_i\}_{i=1}^n$ is represented by the $n \times m$ matrix X:

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

Here the superscript T denotes the vector/matrix transpose.

Many different chemical descriptors have been proposed and used to quantify molecules [9, 14, 26]. For example, descriptors might characterize molecular connectivity, electrostatic interactions, or molecular geometry, as in the popular programs Molconn-X [18] or Molconn-Z [19]. Descriptors can also define the number of bonds along the shortest paths connecting an atom-pair ("topological atom-pair descriptors") [9], as in the large drug database MDDR. Atom-pair descriptors have been used successfully in structure-activity studies [2, 9]. They also have been recently expanded to include torsions of four consecutive atoms, charges of atom pairs, and hydrophobic characteristics, in order to reflect physicochemical and geometric features (see [2] for a short review).

A diversity score for two different compounds X_i and X_j can be measured by the *Euclidean distance* norm based on the compound descriptors:

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

To avoid the dominance of a few large descriptors on the diversity score, scaling becomes necessary if the chemical descriptors use different units and vary significantly in their magnitudes. It is customary to scale data entries x_{ij} by the following formula:

$$\hat{x}_{ij} = \frac{x_{ij} - \beta_j}{\alpha_j - \beta_j}, \tag{2}$$

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

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

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{3}$$

where μ_j and σ_j are the mean and the deviation of the jth 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.

3. Database optimization problems

The similarity and diversity sampling problems are two fundamental optimization problems that arise from database analyses. For a given target compound, *the similarity sampling problem* is to find a sampling set that contains a small number of most similar compounds to the target. Here compound X_{j0} is said to be the most similar to compound X_j in a dataset X if

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

Clearly, each search requires a total of $\mathcal{O}(nm)$ floating-point operations (flops) if X contains n compounds and each distance segment $\delta_{ij} = \|X_i - X_j\|$ requires $\mathcal{O}(m)$ flops to compute. An effective scheme is sought when n and m are large. The task of similarity sampling is often performed to find drugs with similar physiochemical and biological properties to a known drug.

In contrast to the above similarity problem, *the diversity search problem* is to find "the most diverse subset" (or the representative set of the database) $\mathcal{S}_0$ 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 is a given number satisfying $0 < n_0 \ll n$, and $\mathcal{S}$ denotes a subset of the database X with n_0 compounds. An exhaustive search of the most diverse subset requires a total of $\mathcal{O}(C_n^{n_0} n_0^2 m)$ flops because there are $C_n^{n_0}$ possible subsets of $\mathcal{S}$ of size n_0 and each subset takes $\mathcal{O}(n_0^2 m)$ flops. (Here $C_n^{n_0} = n(n-1)(n-2)\cdots(n-n_0+1)/n_0!$). Clearly, even for a small value of n (more than 100), it may take thousands of years to solve this problem. Hence, in practice, it is necessary to look for heuristics or for approximate solutions in some more restricted feasible sets.

Pharmaceutical companies often perform diversity sampling when they scan huge databases to 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.

Our approach to the above database analysis problems is to create a 2D view of the database so that similar and diverse compounds can be viewed simply. The most challenging part of this approach is the construction of a 2D mapping that retains the distance relationships of the database well. Based on the distance-geometry approach, we define the following objective function:

$$E(Y_1, Y_2, \ldots, Y_n) = \frac{1}{4} \sum_{i<j} \omega_{ij} \left(d_{ij}^2 - \delta_{ij}^2 \right)^2, \tag{4}$$

where Y_i is the projection vector of X_i into the l-D space, $d_{ij} = \|Y_i - Y_j\|$ and $\delta_{ij} = \|X_i - X_j\|$ are the Euclidean distances in the l-D and m-D spaces, respectively, ω_{ij} are the weights defined by $\omega_{ij} = 1/\delta_{ij}^4$ if $\delta_{ij}^4 \geq \eta_{min}$ and $\omega_{ij} = 1$ if $\delta_{ij}^4 < \eta_{min}$. The latter case (very small original distance δ_{ij}) may occur when two compounds in the datasets are very similar. We set the parameter η_{min} to a small positive number such as $\eta_{min} = 10^{-12}$.

Theoretically, we define the best projection mapping $(Y_1^*, Y_2^*, \ldots, Y_n^*)$ in the l-D space as the global minimum point of the objective function E. That is,

$$E(Y_1^*, Y_2^*, \ldots, Y_n^*) = \min_{\forall Y_i \in R^l} E(Y_1, Y_2, \ldots, Y_n). \tag{5}$$

Due to the difficulty of finding the global minimum point, especially for a large database, in practice, we often solve the distance-geometry problem (5) by an efficient local minimization algorithm, together with a good starting point close to the global minimizer. Note that the objective function E defined in (4) is smooth everywhere so that Newton algorithms such as TNPACK, which rely on second-derivative information, are good choices for solving (5).

The approximation of distance segments δ_{ij} by d_{ij} can be measured by the relative error expression

$$\begin{aligned} |d_{ij} - \delta_{ij}| \leq \eta\, \delta_{ij} \quad & \text{when } \delta_{ij} > d_{\min} \\ d_{ij} \leq \tilde{\epsilon} \quad & \text{when } \delta_{ij} \leq d_{\min}, \end{aligned} \tag{6}$$

where η, $\tilde{\epsilon}$, and $d_{\min}$ are small positive numbers less than one. Their selections are related to both machine precision and mapping accuracy. For example, we find that $\eta = 0.1$, which specifies a 10% relative accuracy, $d_{\min} = 10^{-12}$ and $\tilde{\epsilon} = 10^{-8}$, which indicate very similar compounds, are satisfactory for our database visualization application.

Let T_d be the total number of the distance segments $d(Y_i, Y_j)$ satisfying eq. (6). We introduce the percentage ρ of the distance segments satisfying eq. (6) to assess the degree of distance preservation of our mapping:

$$\rho = \frac{T_d}{n(n-1)/2} \times 100\%. \tag{7}$$

Compared with other measures, such as relative errors [35] and percentage of retained variation in a mapping as given in (19) [25], the value of ρ more directly indicates the accuracy of the projection mapping. The greater the ρ value (the maximum is 100%), the better the mapping and the more information can be inferred from the projected views of the database compounds (assuming reliable descriptors).

4. Projection by TNPACK

We minimize the objective error function E defined in (4) by our truncated Newton program package, TNPACK [33, 39]. In brief, TNPACK generates a sequence of iterates $\{Y^k\}$ expressed in the form

$$Y^{k+1} = Y^k + \lambda_k P^k, \quad k = 0, 1, 2, \ldots, \tag{8}$$

from an initial guess Y^0, and P^k is a descent direction generated by a "truncated" preconditioned conjugate gradient scheme for solving the classic Newton equation

$$H(Y^k)P = -g(Y^k).$$

The parameter λ_k in eq. (8) is the steplength generated by a line search scheme [30, 40], and g and H are the gradient vector and Hessian matrix of the objective function E, respectively.

The advantage of truncated-Newton (TN) methods is their ability to incorporate available second-derivative information to accelerate convergence. We have found TN methods efficient in many molecular applications [37, 38]. As in other Newton methods appropriate for large-scale applications, the curvature information can be utilized efficiently so that the overall computational work is optimal. The TN outer/inner iteration design can be particularly robust and efficient for solving difficult optimization problems when a preconditioner is used in the PCG inner loop and care is taken in determining the steplength λ_k in eq. (8). In the present applications, we simply use TNPACK with

all default parameters and a numerical procedure to approximate matrix/vector products. That is, the product $H(Y^k)d$ is approximated by the first-order forward difference formula:

$$H(Y^k)d \approx \frac{g(Y^k + hd) - g(Y^k)}{h}, \tag{9}$$

where d is a vector, and h is a small number such as $h = 10^{-8}$. Since the Hessian matrix $H(Y^k)$ in TNPACK only appears in the calculation of the product $H(Y^k)d$, the approximation formula (9) leads to a linear dependency on the number of compounds for TNPACK's memory location.

We also use a simple preconditioner — a diagonal matrix consisting of the diagonal elements of the Hessian matrix for the present applications. We terminate the TNPACK iteration process when the solution iterate Y^k satisfies

$$||g(Y^k)|| < \epsilon_g(1 + |E(Y^k)|), \tag{10}$$

where ϵ_g is a small positive number (we used 10^{-6}). For details, see [33, 39]. We intend to further improve the performance of TNPACK through developing more efficient preconditioners and tailoring other components of TN to the database application in the future work. We emphasize that other local optimization methods like limited-memory BFGS algorithms should work equally well in our context.

5. Projection by principal component analysis (PCA)

The accuracy of the l-D projection mapping generated by TNPACK in approximating database X depends on the selection of the initial guess Y^0. In this section, we introduce PCA to generate a l-D projection mapping of database X. We then use it as the starting point Y^0 of TNPACK, resulting in the PCA/TNPACK protocol. Note that the PCA projection mapping is an "optimal" approximation to the 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 [25]. Hence, a PCA generated initial guess can be much better than a randomly selected one for an optimization algorithm for solving the application problem (5).

PCA is a classic tool for data reduction. In PCA, chemical descriptors are regarded as random variables, and the input database matrix X is considered their sampling matrix. In terms of the orthonormal eigenvectors of the covariance matrix C of the database X, PCA transforms the highly correlated descriptor variables into the uncorrelated variables called **principal components** (PCs). By using the first l PCs with $l \ll m$, the database matrix X can then be reduced to a smaller matrix with dimension $n \times l$.

The $m \times m$ covariance matrix C with elements $\{c_{ij}\}$ is defined by

$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \mu_i)\,(x_{kj} - \mu_j)\,, \tag{11}$$

where μ_j and μ_j are the means of the columns associated with descriptors i and j:

$$\mu_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki} \quad \text{and} \quad \mu_j = \frac{1}{n} \sum_{k=1}^{n} x_{kj}\,. \tag{12}$$

Since C is a symmetric positive semi-definite matrix, it has the spectral decomposition

$$C = V \Lambda V^T , \tag{13}$$

where $V = (v_1, v_2, \ldots, v_m)$, with $v_i \in R^m$, is the $m \times m$ orthogonal eigenvector matrix satisfying $VV^T = I_{m\times m}$, and Λ is a diagonal matrix of the m ordered eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0 .$$

The jth *principal component* vector Y_j of dimension $n \times 1$ is defined by the product of the original matrix X with the jth eigenvector v_j:

$$Y_j = X v_j \quad \text{for } j = 1, 2, \cdots, m . \tag{14}$$

The m PCs $Y_1, Y_2, \cdots, Y_m$ form the $n \times m$ matrix Y:

$$Y = (Y_1, Y_2, \cdots, Y_m).$$

Equation (14) can be written in the matrix form $Y = XV$. Since $VV^T = I$, we can express the dataset matrix X in terms of the PCs as:

$$X = YV^T . \tag{15}$$

Eq. (15) is a fundamental identity for the PCA data reduction. It can be used to reduce the dimensionality from m to l while optimally approximating the original variances of X. In fact, we write eq. (15) as a sum of m matrices:

$$X = \sum_{j=1}^{m} Y_j v_j^T , \tag{16}$$

where $Y_j v_j^T$ is an $n \times m$ matrix with rank of one. Note that the compound vector X_i is the transpose of the ith row vector of X, i.e.,

$$X_i = X^T e_i ,$$

where e_i is the $n \times 1$ unit vector with 1 in the ith component and 0 elsewhere. Therefore, from eq. (15) it follows that each chemical compound vector X_i has the following expression:

$$X_i = X^T e_i = \left(\sum_{j=1}^{m} Y_j v_j^T \right)^T e_i = \sum_{j=1}^{m} Y_j^T e_i \; v_j = \sum_{j=1}^{m} y_{ij} \; v_j , \tag{17}$$

where $y_{ij} = Y_j^T e_i$ is the ith component of the jth PC Y_j. This relationship implies that the compound vector X_i is transformed into the vector space spanned by orthonormal eigenvectors $\{v_j\}$, and can be simply denoted by

$$X_i = (y_{i1}, y_{i2}, \ldots, y_{im})^T .$$

According to expression (17), the l-th dimension projection Z_i of X_i can be defined by using the first l PCs as:

$$Z_i = (y_{i1}, y_{i2}, \ldots, y_{il})^T \quad \text{for } i = 1, 2, \ldots, n. \tag{18}$$

Note that the eigenvalues λ_j of the covariance matrix C represent the variances of the PCs. Hence, the ratio $\left(\sum_{j=1}^{l} \lambda_j\right) \Big/ \left(\sum_{j=1}^{m} \lambda_j\right)$ indicates the percentage of the variances of X that is retained in the l-D projected database $Z = (Z_1, Z_2, \ldots, Z_l)$.

In practice, the value of l is chosen according to the following criterion involving the threshold variance γ,

$$\left(\sum_{j=1}^{l} \lambda_j)/(\sum_{j=1}^{m} \lambda_j\right) \geq \gamma\,. \tag{19}$$

The higher the γ value, the better the approximation (that is, the better the original variances are maintained). The relation of PCA to SVD is given in the Appendix.

6. Numerical examples

In this section, we investigate the performance and potential application of the PCA/TNPACK combination for similarity and diversity sampling of chemical compounds as well as several issues related to the interpretation of the projection. We also compare accuracy and efficiency aspects of PCA/TNPACK to those of a global optimization algorithm in solving the distance-geometry minimization problem (5). All default values of TNPACK are used.

All numerical experiments are made on datasets selected from the large, commercially available database MDDR (http://www.mdli.com). Most computations are performed in double precision on a single R12000 300 MHz processor of an SGI Origin 2000 computer at the New York University. Only the simulated annealing applications (Section 6.8) are performed on a Dell Precision 610 workstation.

We use the package ARPACK [28] with all default parameters for computing the eigenvalue and eigenvectors required by the PCA and SVD procedures. ARPACK performs very well in our applications because it can compute the first l required PCs, singular values, and singular vectors for defining the PCA and SVD projection mappings in R^l, with an order of ln storage locations and an order of nm^2 floating point operations (the database matrix is $n \times m$).

6.1. Test datasets

All datasets of our tests are selected from the Merck drug database, which contains 7500 compounds retrieved from MDDR version 98.1 based on the assigned label in the "Source" field in the database. MDDR contains over 82,000 potential drug candidates. Most entries in the database have an MDDR registration number, chemical connectivity information, associated therapeutic category, mechanism of action, source/inventor, and literature citations. The compounds are derived based on patent literature, journal articles, professional conventions and symposia. There are about 25,000 unique atom-pair

descriptors and only a small number (several hundreds) are shared by most compounds in the MDDR.

For our tests, we select 300 atom-pair descriptors (i.e., $m = 300$) used by most compounds. We use atom-pair [9] and topological torsion [13] descriptors for all compounds. Atom pairs are substructure descriptors of a molecule defined as $AT_i - AT_j - r_{ij}$. The distance, r_{ij}, is the distance in bonds along the shortest path between an atom type AT_i and an atom type AT_j. The atom type reflects element type, number of non-hydrogen atom neighbors, and number of π electrons. Topological torsions are substructure descriptors of a molecule and are defined as $AT_i - AT_j - AT_k - AT_l$, where i, j, k, l are consecutively-bonded atoms. Topological torsion and atom-pair descriptors can effectively distinguish closely related compounds [2, 9] and are useful in similarity searching. Since the atom-pair descriptors in our test datasets are small (from zero to 25 only), no scaling is needed.

6.2. *Role of refinement*

We first study the role of TNPACK refinement in improving the accuracy of a 2D PCA mapping of a database in preserving the distance relationships of the database. We randomly selected 832 chemical compounds from the Merck MDDR to form a dataset called

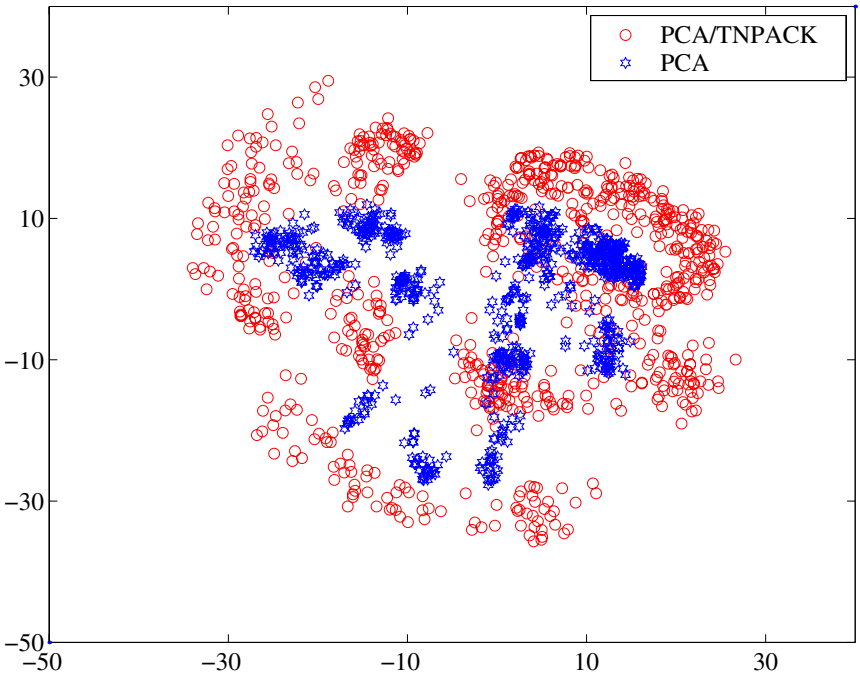

Fig. 1. The 2D PCA and PCA/TNPACK mappings for Merck set1 ($n = 832, m = 300$). See note at end on color figure

Table 1. The accuracy of the PCA and PCA/TNPACK mappings as a function of projection dimension l for Merck set1 ($n = 832$, $m = 300$), along with the performance of the PCA and PCA/TNPACK (TN) programs. Here the minimization objective function E is defined in (4), and the accuracy ρ is defined in (7) with $\eta = 0.1$

Dimension	Final E		Accuracy ρ		TN	CPU time (sec.)	
l	PCA	TN	PCA	TN	Itn.	PCA	Total
2	4.07×10^4	1.80×10^4	2.45	33.70	8	0.20	13.8
3	3.29×10^4	1.06×10^4	2.89	44.72	10	0.21	25.2
5	2.32×10^4	4.87×10^3	7.52	63.16	11	0.33	39.0
8	1.71×10^4	2.17×10^3	17.07	80.80	14	0.57	63.0
12	1.17×10^4	9.68×10^2	28.86	93.49	12	0.72	97.8
16	8.83×10^3	5.18×10^2	42.35	98.12	46	0.88	139.8
20	6.38×10^3	3.12×10^2	53.10	99.50	29	1.01	168.0
25	4.98×10^3	2.01×10^2	61.40	99.88	19	3.26	367.8
30	3.51×10^3	1.11×10^2	71.06	99.99	18	2.80	414.0
35	2.67×10^3	7.22×10	78.42	100	20	2.27	540.0
45	1.73×10^3	3.63×10	87.13	100	16	4.50	393.0
65	7.32×10^2	1.01×10	95.69	100	19	6.28	1141.2

Merck set1 (i.e., $n = 832$ and $m = 300$). Figure 1 compares the two 2D mappings for Merck set1 generated respectively by PCA alone and PCA/TNPACK. From the figure we see that the PCA mapping has been significantly changed following the TNPACK refinement procedure. Table 1 (first data row, for $l = 2$) shows that following TNPACK minimization, the accuracy ρ of the 2D mapping in approximating the distance relationships of Merck set1 has increased from about 3 to 34%. This minimization process only took about 14 seconds of CPU time.

Table 1 also shows that the projection accuracy of both PCA alone and PCA/TNPACK can be improved significantly as the projection dimension l increases (data shown for l ranging from 2 to 65). The accuracy of the PCA/TNPACK mapping in terms of ρ can be near maximal (100%) for $l = 16$, while PCA alone requires $l = 65$ to reach this maximal accuracy. From Table 1 we also see that the PCA program only took 0.2 to 6 seconds to compute the l-D PCA mappings for $l = 2$ to 65, which is very small compared to the total CPU time of 14 to 1141 seconds. Moreover, TNPACK only took 8 to 46 iterations to locate minimum values. Here the number of variables of the minimization target function E ranges from 1664 (2×832) to 54080 (65×832).

6.3. Similarity and diversity applications

As an example of similarity and diversity sampling, we selected four pairs of points that are close in the projection, as shown in Figure 2, and 12 distant points (letters A to L in Figure 3) from the 2D PCA/TNPACK mapping for Merck set1, respectively. 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 a general principle [21] that compounds with similar structures are likely to share physicochemical and biological properties, we

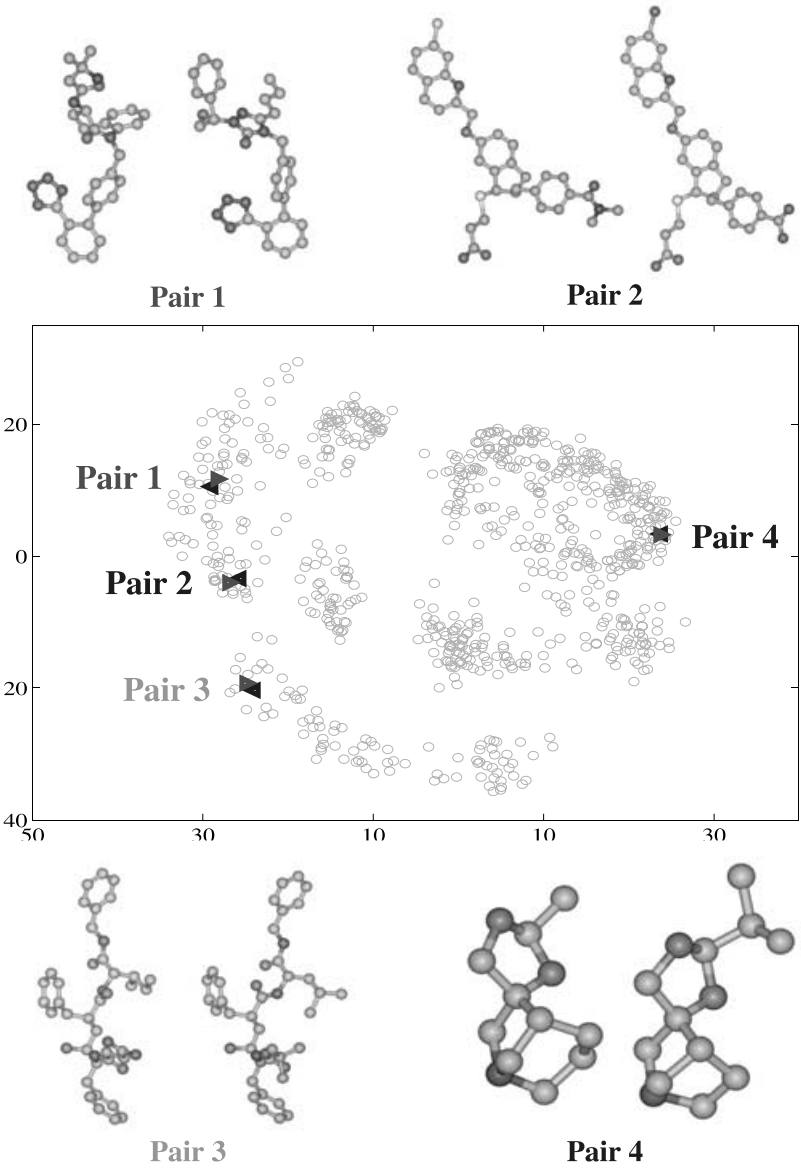

Fig. 2. Four similarity samples (Pair 1 to Pair 4) of chemical compounds based on the 2D PCA/TNPACK mapping and their chemical structures. See note at end on color figure

might predict that the 12 distant points belong to different therapeutic categories, while compounds corresponding to each pair are in the same therapeutic category. See Table 2 for the therapeutic categories corresponding to the sampling points. Still, as shown below (Figure 6), dissimilar structures can belong to the same therapeutic group. Clearly, visual inspection of structures is only one aspect of a compound's functionality.

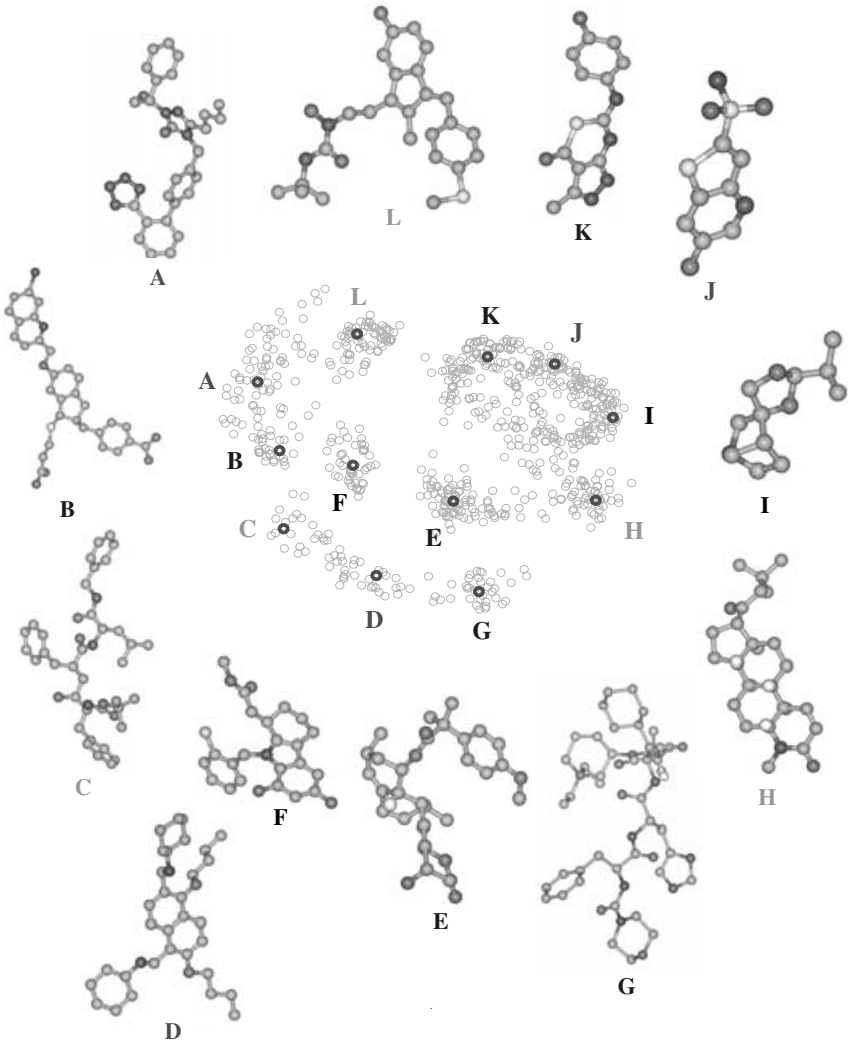

Fig. 3. A diversity sample (A–L) of chemical compounds based on the 2D PCA/TNPACK mapping and their chemical structures. See note at end on color figure

6.4. Cluster assessment

To test how the 2D PCA/TNPACK mapping can retain certain clusters of chemical compounds, we construct the dataset Merck set2 ($n = 342$ and $m = 300$) containing four clusters as shown below: select four distant target compounds K, E, H and I according to Figure 3 and then construct clusters for each from the Merck MDDR (a dataset of 7500 compounds) such that a compound j belongs to the cluster of a target compound i if the corresponding Euclidean distance d_{ij} is less than 15, and each compound belongs

Table 2. The therapeutic categories of the 20 sampling points selected in Figures 2 and 3

Sampling Points	MDDR Reg. No.	Therapeutic Category
Pair 1, blue	180311	Angiotensin II Blocker
Pair 1, red	190065	GH Growth Hormone, Somatotropin
Pair 2, blue	192077	Leukotriene Antagonist
Pair 2, red	192078	Leukotriene Antagonist
Pair 3, blue	157851	HIV Protease Inhibitor
Pair 3, red	160107	HIV Protease Inhibitor
Pair 4, blue	166899	Muscarinic M1 Receptor Agonist
Pair 4, red	166906	Muscarinic M1 Receptor Agonist
Point A	180311	Angiotensin II Blocker
Point B	192077	Leukotriene Antagonist
Point C	157851	HIV Protease Inhibitor
Point D	170280	Immunomodulator AIDS
Point E	141090	HMG-CoA Reductase Inhibitor
Point F	149705	Thromboxane Antagonist
Point G	152273	Renin Inhibitor
Point H	146649	Steroid 5alpha Reductase Inhibitor
Point I	166899	Muscarinic M1 Receptor Agonist
Point J	144259	Carbonic Anhydrase Inhibitor
Point K	190519	Elastase Inhibitor
Point L	188363	Lipoxygenase Inhibitor

to only one cluster. This produces clusters for E, H, I, and K containing 17, 77, 239, and 9 compounds, respectively.

We then project these four clusters of compounds from the 300-dimensional chemical space onto the 2D mapping space by PCA/TNPACK. The four clusters are displayed in the 2D PCA/TNPACK mapping by using four colors (blue for Cluster E, cyan for Cluster H, red for Cluster I, and green for Cluster K) in Figure 4. This 2D mapping has an accuracy of $\rho = 35\%$ for $\eta = 0.1$. The view shows that the four clusters are well retained in the 2D PCA/TNPACK mapping.

6.5. Structure and medicinal activity

To examine the relation between structural features (characterized by chemical descriptors) and medicinal activities, we construct the dataset Merck set3 ($n = 338$ and $m = 300$) that contains four different therapeutic groups. The names of the four groups and the number of compounds in each group are listed in Figure 5.

Figure 5 displays the distribution of the four categories based on the 2D PCA/TNPACK mapping. The accuracy ρ for $\eta = 0.1$ (defined in (6)) 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 Merck set3.

From Figure 5 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,

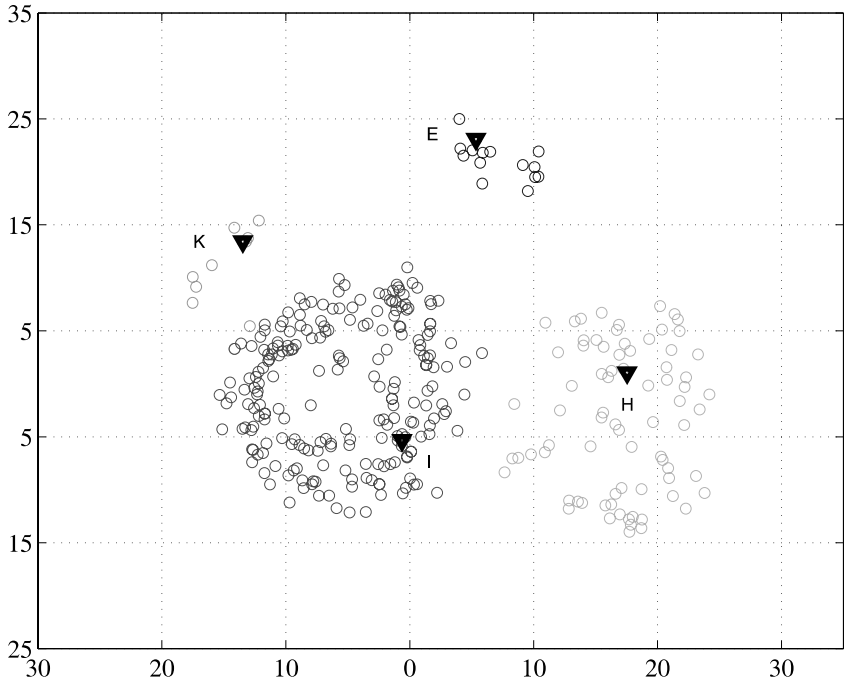

Fig. 4. The 2D PCA/TNPACK mapping for Merck set2 ($n = 342, m = 300$). Here set2 consists of four clusters of compounds: Clusters E, H, I, and K with 17, 77, 239, and 9 compounds, respectively. See note at end on color figure

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 5. This offers the possibility that corresponding drugs may have similar structures and similar medicinal activities.

Based on this mapping of Merck set3 in Figure 5, we select eight distant points (two points per group) as shown in Figure 6. Their chemical structures are also displayed in Figure 6. 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.

6.6. PCA/TNPACK vs SVD/TNPACK

As shown in the Appendix, PCA and SVD can generate the same projection mapping if the database matrix is scaled such that each column has a mean of zero. For an unscaled database, PCA and SVD may produce different projection mappings. Figure 7 compares the four 2D mappings for Merck set3, which is unscaled, generated by SVD, PCA, SVD/TNPACK, and PCA/TNPACK. Merck set3 contains four therapeutic groups and eight distant sampling points. The values of accuracy ρ with $\eta = 0.1$ for these four

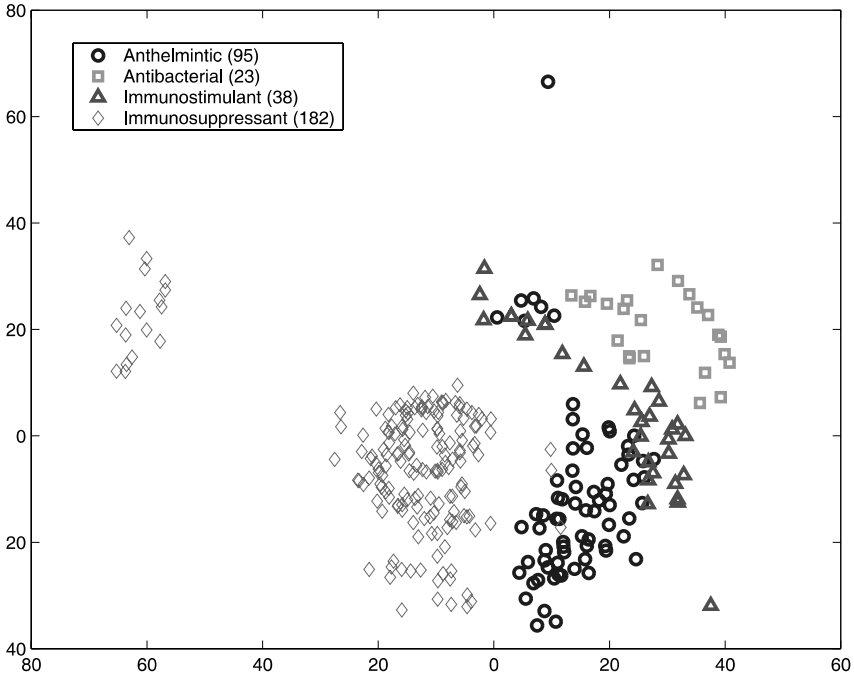

Fig. 5. The 2D PCA/TNPACK mapping for Merck set3 ($n = 338, m = 300$). Here set3 contains the four therapeutic groups as indicated in the figure. See note at end on color figure

projections are 10.03% (SVD), 10.04% (PCA), 36.68% (SVD/TNPACK), and 36.66% (PCA/TNPACK). Even though different projections are generated by SVD and PCA, from Figure 7 we see that the clusters corresponding to the four therapeutic groups are reasonably maintained.

6.7. PCA Starting Point vs. Randomly Selected Starting Point

To confirm that PCA can provide TNPACK with a better starting point than a randomly-generated point, we tested TNPACK for Merck Set 2 using the six random starting points defined by:

$$X^{0,k} = 5^{k-1}\Phi(k), \quad k = 1, 2, 3, 4, 5, 6,$$

where $\Phi(k)$ is a random vector with $2n$ components (as generated by a pseudo-random vector generator with seed k). Such starting points cover the search range from zero to 5^5 (3125) in each dimension direction. These numerical experiments are performed on a Dell Precision 610 workstation (a Pentium III Xeon 550 MHZ processor with 768MB RAM and 2MB cache) at the University of Southern Mississippi. The numerical results in Table 3 show that with the PCA starting point TNPACK located a lower minimum point in less CPU time than with the random starting points.

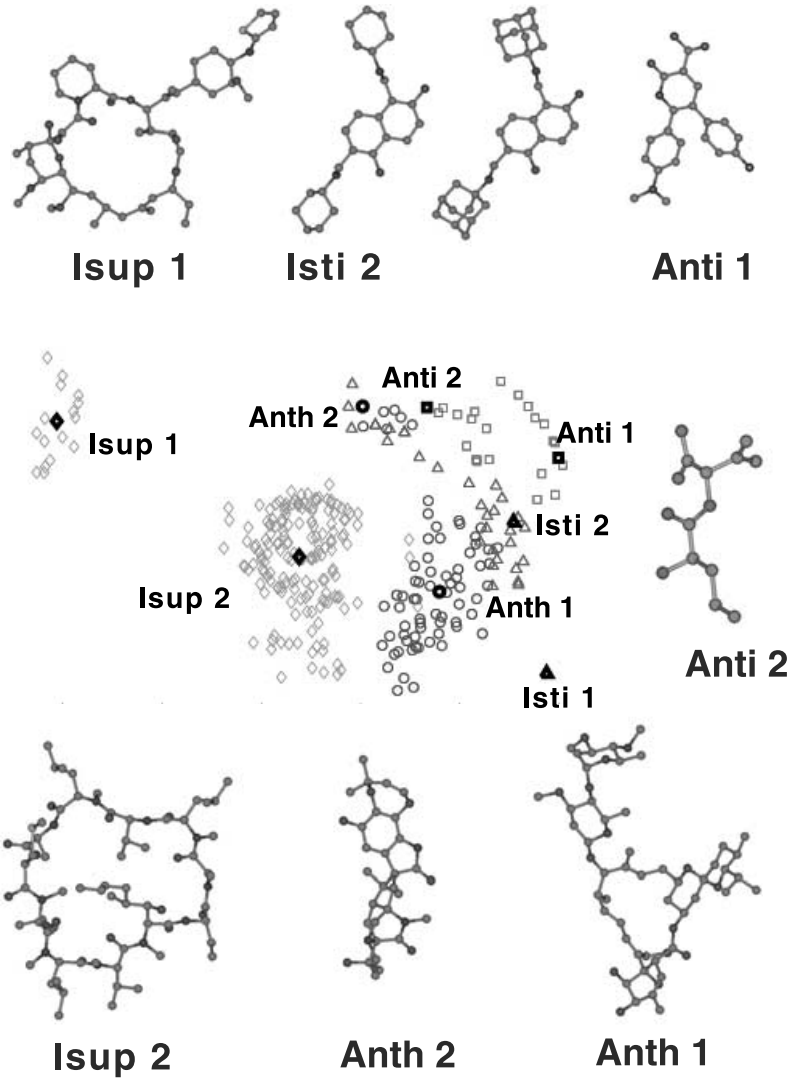

Fig. 6. The diversity of chemical compounds belonging to the same therapeutic group. See note at end on color figure

6.8. Comparison to a Global Optimization Scheme

To demonstrate that the local minimum point found by PCA/TNPACK may be a reasonable approximation to the global minimizer, we performed simple numerical tests for two datasets using simulated annealing implemented by the program SIMANN [22, 23] (available to the public).

We first compare the performance of PCA/TNPACK with that of SIMANN for solving the minimization problem (5) for Merck set2 ($n = 342$ and $m = 300$). Tests were

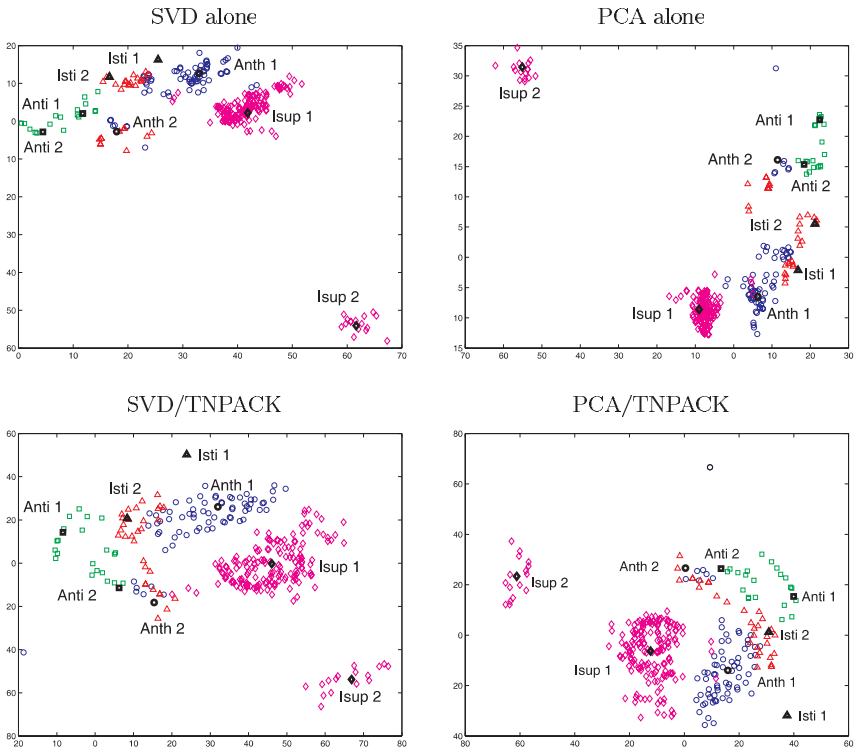

Fig. 7. Comparison of the four 2D mappings generated by SVD, PCA, SVD/TNPACK, and PCA/TNPACK for Merck set3 ($n = 338$, $m = 300$). See note at end on color figure

Table 3. Comparison of the PCA starting point with the randomly selected starting points for TNPACK. Here the objective function E is defined by (4) for Merck set2 ($n = 342$, $m = 300$)

Starting point	Initial E	Final E	Iterations	CPU time (sec.)
PCA	8.39×10^3	2.554×10^3	39	18
Seed 1	1.45×10^4	2.567×10^3	61	28
Seed 2	1.45×10^4	2.751×10^3	51	24
Seed 3	5.90×10^5	2.732×10^3	57	26
Seed 4	1.74×10^8	2.584×10^3	59	29
Seed 5	9.94×10^{10}	2.699×10^3	74	37
Seed 6	1.43×10^{14}	2.571×10^3	63	23

made using starting points generated either randomly or by PCA. Default values of SIMANN were used. For example, we set the temperature reduction factor to be 0.85, as suggested in [10], and the final temperature to be lower than 10^{-8}. The numerical results reported in Table 4 assess performance in terms of the number of function evaluations (E evals) and the CPU time. The total number of gradient evaluations (g evals) is also reported for TNPACK. Gradient evaluations in TNPACK are performed in the line search and in each inner loop of TNPACK, in which one matrix/vector product is evaluated by a finite difference of gradients (9).

Table 4. Comparison of TNPACK with the simulated annealing global optimization program SIMANN for minimizing objective function E defined in (4) for Merck set2 ($n = 342$, $m = 300$)

Methods	Final E	Final $\|g\|$	E evals.	g evals.	CPU time (min.)
Start point randomly generated					
SIMANN	3.22×10^3	1.29×10^{-4}	9,097,201	0	1409
TNPACK	2.56×10^3	4.21×10^{-8}	82	1970	0.76
Start point generated by PCA					
SIMANN	2.55×10^3	9.85×10^{-5}	9,028,801	0	1374
TNPACK	2.55×10^3	2.42×10^{-7}	39	613	0.26

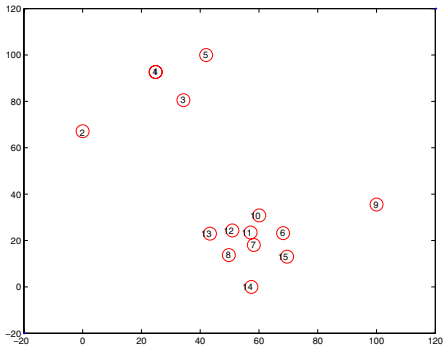

Fig. 8. The 2D mapping of the global/local minimizer located by SIMANN for the set containing the first 15 compounds of Merck set1 ($n = 15, m = 300$). See note at end on color figure

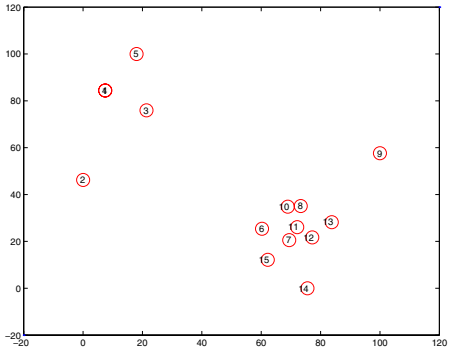

Fig. 9. The 2D mapping generated by the local minimum point located by PCA/TNPACK for the set containing the first 15 compounds of Merck set1 ($n = 15, m = 300$). See note at end on color figure

To compare computing time, we use the same starting point generated by PCA for both SIMANN and TNPACK. We then set the starting temperature of SIMANN to 100, a value obtained after many numerical tests, so that SIMANN locates the same local minimum as TNPACK. In this case, we see that TNPACK is about 5288 times faster than SIMANN. With a randomly selecting starting point, simulated annealing locates a higher minimum point than TNPACK.

Clearly, to allow SIMANN to find a lower minimum value than TNPACK, a higher starting temperature must be used. However, it is difficult to find such a starting temperature for a large-scale minimization problem, as no general rule is available. The starting temperature depends on the size of problem and may be higher than 10^{13} [34].

To save computing time for the SIMANN application, we construct a small dataset that contains the first 15 compounds of Merck Set2; this allows us to conduct many tests with different starting temperatures that range from 10 to 10^{17}. These tests produce the lowest value $E = 1.635$, which we believe to be the global minimum value of E, using the starting temperature of 10^{14}. The 2D mapping defined by this global minimizer, which has an accuracy of $\rho = 57.14$, is plotted in Figure 8.

For comparison, we also plot the 2D PCA/TNPACK mapping for this small dataset in Figure 9. Here, the minimum of E found by PCA/TNPACK is 1.731, and the 2D mapping has an accuracy of $\rho = 58.1$. The figure shows rather similar projections except regarding points 8 and 10, between which the original distance is 6.63: SIMANN approximates this distance well (value of 7.48) whereas the PCA/TNPACK approach yields 1.56. Of course, in general the global optimization approach is essential; yet from a practical point of view, a local solution can be helpful and can provide a first step for analysis.

7. Conclusions

We have described optimization problems that arise from database analyses and formulated the PCA/TNPACK procedure for generating a projection mapping of a database as a first step. Using PCA/TNPACK on Merck drug datasets, we performed numerical experiments to explore several relevant issues related to the projection and refinement protocol; these results demonstrate that the PCA/TNPACK protocol can have a reasonable accuracy in approximating the original distance relationships and retaining both the clusters of compounds and the distribution patterns of therapeutic groups. Hence, the PCA/TNPACK procedure is valuable for sampling similar and diverse compounds, and perhaps ultimately for aiding the generation of drug candidates or the optimization of bioactive compounds.

To compare the accuracy and efficiency of PCA/TNPACK with that of a global minimization algorithm for solving the distance-geometry optimization problem, we have experimented with simulated annealing as implemented in the program SIMANN [23, 22] for two small Merck datasets. Numerical results show that while the global minimizer can improve the accuracy of the 2D PCA/TNPACK mapping, the computing cost can be several thousand times greater. It is far from trivial to find the global minimizer by SIMANN for a large dataset. Thus, the local approach is a valuable first step.

To further improve the performance of TNPACK, we plan to develop more efficient preconditioners for TNPACK than the simple diagonal formulation used here. We will also investigate how to overcome the memory limitation that comes from the definition of the objective function E defined in (4), which requires an order of n^2 memory locations for the calculation of E efficiently (n is the number of compounds of a database). A possible solution to this problem has been suggested in [36], that is, to divide the huge dataset and apply the PCA/TNPACK procedure to each subset of database, followed by some suitable assembly technique. Another simple way to solve this problem is to define the objective function E by only using a banded distance matrix with a small band width so that the memory requirement is reduced to an optimal order of n. Global optimization approaches, of course, have greater potential for success in general for these kinds of problems, and might be worth the investment in algorithm development.

Acknowledgements. This work is supported by the National Science Foundation (ASC-9157582 and BIR 94-23827EQ), the National Institutes of Health (R01 GM55164) [to T.S.], the Oak Ridge Associated Universities (0221709103) and the University of Southern Mississippi (2221709006) [to D.X.]. T. Schlick is an investigator of the Howard Hughes Medical Institute.

Appendix

Relation of PCA to SVD

SVD is another classic technique used for data reduction in many practical applications. In [35, 36], we described its application to visualize chemical databases. In this section we show that PCA and SVD are closely related. In particular, they generate the same projection mapping if each column of a data matrix X has a mean of zero.

A database X can be modified into a database $\hat{X} = \{\hat{x}_{ij}\}$ with a mean of zero for its each column by defining each $\hat{x}_{ij}$ as

$$\hat{x}_{ij} = x_{ij} - \mu_j, \tag{20}$$

where μ_j is the mean of the jth column of X as defined in eq. (12). The SVD decomposition of $\hat{X}$ can be expressed as

$$\hat{X} = U \Sigma V^T, \tag{21}$$

where $U_{n\times n} = (u_1, u_2, \ldots, u_n)$ and $V_{m\times m} = (v_1, v_2, \ldots, v_m)$ with $u_i \in R^n$ and $v_i \in \mathcal{R}^m$ are orthogonal matrices, and $\Sigma = \text{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ is a diagonal matrix with the singular values arranged in decreasing order:

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

Here r indicates the rank of matrix $\hat{X}$.

According to eq. (21), we define the l-D projection mapping Y_i of each modified compound $\hat{X}_i$ (i.e., the transpose of the ith row vector of $\hat{X}$) as follows:

$$Y_i = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, \ldots, \sigma_l u_{il})^T, \tag{22}$$

where u_{ij} is the ith component of the jth left singular vector u_j (see [35] for details). It is well known that the SVD projection mapping $Y = (Y_1, Y_2, \ldots, Y_n)^T$ is an optimal approximation of $\hat{X}$ in the following sense

$$\|Y - \hat{X}\| = \min\{\|B - \hat{X}\| \mid \text{for all } n \times m \text{ matrices } B \text{ with rank } l\},$$

and that $\|Y - \hat{X}\| = \sigma_{l+1}$ [24].

Note that the covariance matrix of the database $\hat{X}$ is $C = \hat{X}^T \hat{X}$. Hence, combining the SVD factorization of $\hat{X}$ in eq. (21) with $U^T U = I$ and $V^T V = I$, we have

$$CV = \hat{X}^T \hat{X} V = (U \Sigma V^T)^T U \Sigma V^T V = V \Sigma U^T U \Sigma V^T V = V \Sigma^2,$$

or in vector form,

$$C v_j = \sigma_j^2 v_j \quad \text{for } j = 1, 2, \ldots, m.$$

This shows that the right singular vectors v_j are the eigenvectors of the covariance matrix C with σ_j^2 as the corresponding eigenvalues.

Therefore, the PC mapping Y of the database $\hat{X}$ is defined by $Y = \hat{X}V$, and can be written as

$$Y = \hat{X}V = U\Sigma V^T V = U\Sigma,$$

or in the vector form,

$$Y_j = \sigma_j u_j \quad \text{for } j = 1, 2, \ldots, m. \tag{23}$$

This shows that *PCs are the products of singular values and the corresponding left singular vectors.*

According to eq. (23), the components of PC Y_j can be expressed as $y_{ij} = \sigma_j u_{ij}$ for $i = 1, 2, \ldots, n$. Hence, the PCA mapping defined in (18) is identical to the SVD mapping defined in (22). This shows that PCA and SVD generate the same projection mapping for a database modified by eq. (20).

A database X whose columns have a mean of zero can be scaled by the scaling procedure (3) so that the range of variance in each column of X is the same. For a database scaled by (3), the total sum of all variances becomes m, and $\sum_{j=1}^{l} \sigma_j^2$ indicates the total variances retained in the projection mapping. Hence, based on SVD, rule (19) for selecting the dimension l can be simplified as

$$\frac{1}{m} \sum_{j=1}^{l} \sigma_j^2 \geq \gamma. \tag{24}$$

For a database X whose columns have a mean of nonzero, the PCA and SVD mappings are clearly different. In fact, PCA and SVD define two different orthonormal rotation transforms that map X into two different spaces (in terms of orthogonal vectors $\{v_j\}_{j=1}^{m}$). They also have different coordinate scales for the mapping points.

We use the efficient program package ARPACK [28] for computing spectral decomposition for SVD and PCA. 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.

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. Diversity of chemical libraries. In *Encyclopedia of Computational Chemistry*, P. von Ragué Schleyer, Editor-in-Chief, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and Schaefer, III, H. F., Eds, John Wiley & Sons, West Sussex, UK, 1:742 – 761, 1998.
3. L. M. Amzel. Structure-based drug design. *Curr. Opin. Biotech.*, 9:366 – 369, 1998.
4. 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.
5. M. Bakonyi and C. R. Johnson. The Euclidean distance matrix completion problem. *SIAM J. Matrix Anal. Appl.*, 16:646 – 654, 1995.
6. I. Borg and P. Groenen, Modern Multidimensional Scaling. *Springer-Verlag*, New York, 1997.
7. 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 Decker, New York, NY, 1997. Supplement 22.
8. R. D. Brown and Y. C. Martin. Information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sciences.*, 37:1 – 9, 1997.

9. 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.
10. A. Corana, M. Marchesi, C. Martini, and S. Ridella. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software*, 13:262 – 280, 1987.
11. D. C. Liu and J. Nocedal. On the limited memory BFGS method for large-scale optimization. *Math. Prog.*, 45:503 – 528, 1989.
12. S. G. Nash and J. Nocedal. A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization. *SIAM J. Optim.*, 1:358 – 372, 1991.
13. R. Nilakantan, N. Bauman, J. S. Dixon, R. Venkataraghavan. Topological torsions: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comp. Sci.*, 27:82 – 85, 1987.
14. J. Devillers and A. T. Balaban, Eds. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, the Netherlands, 1999.
15. G. M. Downs and P. Willett. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.*, 7:1 – 66, 1996.
16. W. Glunt, T. L. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matrix Anal. Appl.*, 11:589 – 600, 1990.
17. J. C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra Appl.*, 67:81 – 97, 1985.
18. L. L. Hall and L. B. Kier. *Molconn-X, version 2.0, a Program for Molecular Topology*. Hall Associates Consulting: Quincy, MD, 1995.
19. L. L. Hall and L. B. Kier. *Molconn-Z, version 3.1, a Program for Molecular Topology*. Hall Associates Consulting: Quincy, MD, 1998.
20. M. Hassan, J.P. Bielawski, J.C. Hempel, and M. Waldman. Optimization and visualization of molecular diversity and combinatorial libraries. *Molecular Diversity*, 2:64 – 74, 1996.
21. M. A. Johnson and G. M. Maggiora. *Concepts and Applications of Molecular Similarity*. Wiley, New York, 1990.
22. W. L. Goffe. SIMANN: A global optimization algorithm using simulated annealing. *Studies in Nonlinear Dynamics and Econometrics*, 1: 169 – 176, 1996.
23. W. L. Goffe, G. D. Ferrier, and J. Rogers. Global optimization of statistical functions with simulated annealing. *J. of Econometrics*, 60(1/2):65 – 99, 1994.
24. G. H. Golub and C. F. van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, MD, second edition, 1986.
25. I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.
26. L. B. Kier and L. H. Hall. *Molecular Structure Description: The Electrotopological State*. Academic Press, 1999.
27. T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, New York, 1997.
28. 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.
29. Y. C. Martin, R. C. Brown, and M. G. Bures. in J.F. Kerwin and E. M. Gordon, editors, *Combinatorial Chemistry and Molecular Diversity*, Wiley, New York, 1997.
30. J. J. Moré and D. J. Thuente. Line search algorithms with guaranteed sufficient decrease. *ACM Trans. Math. Software*, 20:286 – 307, 1994.
31. J. Pintér. Continuous global optimization software: a brief review. *Optima*, 52:1 – 8, 1996.
32. D. D. Robinson, T. W. Barlow, and W. G. Richard. Reduced dimensional representations of molecular structure. *J. Chem. Inf. Comput. Sci.*, 37:939 – 942, 1997.
33. 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.
34. S. R. White. Concepts of scale in simulated annealing. *Proceedings of the IEEE International Conference on Computer Design*, ICCD 84, New York 1984,646 – 651.
35. 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.
36. 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.
37. P. Derreumaux, G. Zhang, B. Brooks, and T. Schlick. A truncated-Newton method adapted for CHARMM and biomolecular applications *J. Comput. Chem.*, 15:532 – 552, 1994.
38. T. Schlick and M. L. Overton. A powerful truncated Newton method for potential energy functions. *J. Comp. Chem.*, 8:1025 – 1039, 1987.

39. 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.
40. D. Xie and T. Schlick. A more lenient stopping rule for line search algorithms. *Optim. Meth. Soft.*, In Press, 2002.

Note on Figures

(Color version of figures available at: monod.biomath.nyu.edu/index/papdir/pap_2_92.html)