algorithm for feature selection in Section IV. Section V describes related work on unsupervised feature selection. Experimental results on seven microarray datasets are presented and discussed in Section VI. Finally, we give concluding remarks in Section VII.

II. LDA AND MMC

Linear discriminant analysis (LDA) aims to find a set of projection vectors that maximize the between-class scatter and simultaneously minimize the within-class scatter, thereby achieving maximum discrimination [9].

Let $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ be a sample matrix containing $\boldsymbol{x}_i, i = 1,\ldots,n$ as columns, where n is the number of samples and p is the number of features. The between-class scatter matrix $\boldsymbol{S}_b$ and the within-class scatter matrix $\boldsymbol{S}_w$ are defined as:

$$\boldsymbol{S}_b = \frac{1}{n}\sum_{k=1}^{c} n_k(\boldsymbol{m}^{(k)} - \boldsymbol{m})(\boldsymbol{m}^{(k)} - \boldsymbol{m})^T,$$

$$\boldsymbol{S}_w = \frac{1}{n}\sum_{k=1}^{c}\sum_{j=1}^{n_k}(\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)})(\boldsymbol{x}_j^{(k)} - \boldsymbol{m}^{(k)})^T,$$

where c is the number of classes; n_k is the number of samples in class k; $\boldsymbol{x}_j^{(k)}$ is the jth sample in class k; $\boldsymbol{m}^{(k)}$ and $\boldsymbol{m}$ are the mean vector of class k and the total mean vector, respectively. Then, classical LDA finds the projection matrix $\boldsymbol{W}$ by maximizing the Fisher criterion

$$J_{LDA}(\boldsymbol{W}) = trace\big((\boldsymbol{W}^T\boldsymbol{S}_w\boldsymbol{W})^{-1}(\boldsymbol{W}^T\boldsymbol{S}_b\boldsymbol{W})\big). \quad (1)$$

By solving a generalized eigenvalue problem, $\boldsymbol{W}$ can be found as the eigenvectors of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ corresponding to the largest eigenvalues. However, when the dimensionality of samples is larger than the sample size, i.e. $p > n$, $\boldsymbol{S}_w$ becomes singular and we cannot compute $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$, which is a major drawback of classical LDA. This is known as the singularity problem or the small sample size problem.

To overcome this problem, Li *et al.* [18] recently proposed to use the maximum margin criterion (MMC) instead of (1) to find the projection vectors. The MMC is defined as:

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \boldsymbol{S}_w)\boldsymbol{W}\big). \quad (2)$$

In this case, the projection matrix $\boldsymbol{W}$ that maximize the criterion (2) can be found as the eigenvectors of $\boldsymbol{S}_b - \boldsymbol{S}_w$ corresponding to the largest eigenvalues. Li *et al.* proposed an efficient algorithm to compute the projection matrix of the MMC under the constraint that $\boldsymbol{W}^T\boldsymbol{S}_t\boldsymbol{W} = \boldsymbol{I}$, where $\boldsymbol{S}_t$ is the total scatter matrix. This is found to be the same as the uncorrelated LDA (ULDA) algorithm in [32]. Also, an efficient algorithm for the MMC subject to the orthogonality constraint on $\boldsymbol{W}$, i.e. $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$, was presented in [23]. In both cases, we need not compute the inverse of $\boldsymbol{S}_w$, hence the singularity problem can be easily avoided.

It should be noted that the MMC is not equivalent to the Fisher criterion. As Loog [21] disproved the theorem in [26], the discriminant vectors obtained by maximizing (2) are not generally the same as those obtained by maximizing (1). More precisely, although ULDA can be considered as an extension of classical LDA to small sample size cases [32], the MMC with the orthogonality constraint does not necessarily yield projection vectors that are optimal for discrimination. In practice, a better discrimination can be achieved by balancing the between-class and within-class scatters using the following criterion as in [20]:

$$J_{MMC}(\boldsymbol{W}) = trace\big(\boldsymbol{W}^T(\boldsymbol{S}_b - \mu\boldsymbol{S}_w)\boldsymbol{W}\big), \quad (3)$$

where μ is a non-negative constant. It is clear that (2) is a special case of (3). In the following sections, we focus on the MMC defined by (2) with the orthogonality constraint.

III. UNSUPERVISED LLDA

A. Extension of LLDA to unsupervised cases

We can write the total and within-class scatter matrices as follows:

$$\begin{aligned}\boldsymbol{S}_t &= \frac{1}{n}\boldsymbol{X}\big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T\big)\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_g)\boldsymbol{X}^T, \\ \boldsymbol{S}_w &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \sum_{k=1}^{c}\frac{1}{n_k}\boldsymbol{e}^{(k)}\boldsymbol{e}^{(k)T})\boldsymbol{X}^T \\ &= \frac{1}{n}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}_\ell)\boldsymbol{X}^T,\end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{e} = (1,1,\ldots,1)^T$ is an n-dimensional vector, and $\boldsymbol{e}^{(k)}$ is an n-dimensional vector with $e_i^{(k)} = 1$ if $\boldsymbol{x}_i$ belongs to class k, and 0 otherwise. In terms of graph Laplacians [6], $\boldsymbol{I} - \boldsymbol{W}_g$ can be viewed as the global Laplacian of a graph such that all vertices are connected with a constant weight of $1/n$, and $\boldsymbol{I} - \boldsymbol{W}_\ell$ as the local Laplacian of a graph such that vertices are connected with a constant weight of $1/n_k$ only when both belong to the kth class.

From the relationship [9]

$$\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_w,$$

it follows that

$$\begin{aligned}\boldsymbol{S}_b - \boldsymbol{S}_w &= \boldsymbol{S}_t - 2\boldsymbol{S}_w \\ &= \frac{1}{n}\boldsymbol{X}\big((\boldsymbol{I} - \boldsymbol{W}_g) - 2(\boldsymbol{I} - \boldsymbol{W}_\ell)\big)\boldsymbol{X}^T. \quad (4)\end{aligned}$$

The MMC represented in this form is referred to as Laplacian linear discriminant analysis (LLDA) in [26] and was applied to extract discriminant features in supervised scenarios.

In this study, we extend (4) to unsupervised cases. We first define the global similarity matrix $\boldsymbol{K}_g$ and the local similarity matrix $\boldsymbol{K}_\ell$ as:

$$[\boldsymbol{K}_g]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } i \neq j \\ 0, & \text{otherwise,}\end{cases}$$

$$[\boldsymbol{K}_\ell]_{ij} = \begin{cases} k(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{if } \boldsymbol{x}_i \text{ is among } k_\ell \text{ nearest neighbors of } \boldsymbol{x}_j \\ & \text{or } \boldsymbol{x}_j \text{ is among } k_\ell \text{ nearest neighbors of } \boldsymbol{x}_i \\ 0, & \text{otherwise.}\end{cases}$$

Here, $k(\cdot,\cdot)$ represents the similarity between each pair of samples, and the standard measures include heat kernel (Gaussian kernel), inner product and Euclidean distance. In supervised cases, prior class information can also be reflected to guide the graph construction [31]. Let $\boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$ be the normalized global and local Laplacian matrices, respectively as:

$$\boldsymbol{L}_g = \boldsymbol{I} - \boldsymbol{D}_g^{-\frac{1}{2}}\boldsymbol{K}_g\boldsymbol{D}_g^{-\frac{1}{2}},$$

$$\boldsymbol{L}_\ell = \boldsymbol{I} - \boldsymbol{D}_\ell^{-\frac{1}{2}}\boldsymbol{K}_\ell\boldsymbol{D}_\ell^{-\frac{1}{2}},$$

where D_g and D_ℓ are diagonal matrices such that $[D_g]_{ii} = \sum_j [K_g]_{ji}$ and $[D_\ell]_{ii} = \sum_j [K_\ell]_{ji}$. Then, we seek to find a set of projection vectors W that maximize the following criterion:

$$J_{LLDA}(W) = trace\big(W^T(S_g - 2S_\ell)W\big), \qquad (5)$$

where S_g and S_ℓ are the global and local scatter matrices defined as:

$$S_g = \frac{1}{n} X L_g X^T,$$
$$S_\ell = \frac{1}{n} X L_\ell X^T.$$

It is easy to check that, when we set $[K_g]_{ij} = 1/n$ for all i, j; and $[K_\ell]_{ij} = 1/n_k$ if x_i and x_j are both in the kth class, and 0 otherwise, L_g and L_ℓ respectively become $I - W_g$ and $I - W_\ell$, hence (5) includes the MMC as a special case.

In general, (5) does not require class information and can be used in an unsupervised manner. The construction of the local scatter matrix is based on the assumption that, if x_i and x_j are close, they are likely to belong to the same cluster. Under the condition that class labels are unavailable, we cannot explicitly consider the separability of different clusters, which is represented by the between-class scatter in classical LDA and the MMC. In the objective function (5), it is implicitly represented by the difference between the global scatter and the local scatter. Therefore, discriminative features can be extracted even in unsupervised scenarios. In this paper, we refer to unsupervised LLDA simply as LLDA.

Note that the reason for using the normalized graph Laplacians is that the criterion (5) without normalization may be affected by the scale of the similarity measure or by the choice of the number of nearest neighbors, since (5) is defined as the difference rather than the ratio of the global scatter to the local scatter. Also, the use of normalized graph Laplacian is known to be effective in spectral clustering (e.g. [24]).

B. *Efficient algorithm for LLDA*

Similarly to the case of (2), the projection matrix W that maximize the criterion (5) subject to the orthogonality constraint can be found as the eigenvectors of $S_g - 2S_\ell$ corresponding to the largest eigenvalues. When p, the number of features, is very large as in microarray data, however, it is computationally demanding to directly perform the eigenvalue decomposition (EVD) of $S_g - 2S_\ell$, which is of size $p \times p$. In [26], two approaches for computing LLDA have been presented. The first one directly computes the eigenvalues and eigenvectors, hence demands expensive computational costs. The other approach achieves this via the spectral decomposition of Laplacian matrix, but it still needs to compute the eigenvalues and eigenvectors of a $p \times p$ matrix. Even worse, the eigenvectors corresponding to the non-positive eigenvalues are discarded, thus it does not provide the exact solution to the maximization problem and results in losing discriminatory information.

Here, we propose a novel algorithm for computing W, which is particularly efficient when the feature size is much larger than the sample size, i.e. $p \gg n$, as is often the case with microarray data. The proposed algorithm is based on the following theorem (see the Appendix for the proof).

Theorem 1: Let $P\Lambda Q^T$ be the reduced SVD [11] of $X \in \mathbb{R}^{p \times n}$, where $P \in \mathbb{R}^{p \times n}$ and $Q \in \mathbb{R}^{n \times n}$ are orthonormal matrices and $\Lambda \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Further, let $V\Delta V^T$ be the EVD of a symmetric matrix $\Lambda Q^T(L_g - 2L_\ell)Q\Lambda$, where $V \in \mathbb{R}^{n \times n}$ is an orthonormal matrix and $\Delta \in \mathbb{R}^{n \times n}$ is a diagonal matrix. Then, W is constituted by the eigenvectors in PV corresponding to the largest eigenvalues in Δ.

Note that the main computation of the algorithm consists of the SVD of a $p \times n$ matrix and the EVD of an $n \times n$ matrix. Thus, it is very efficient in the case of $p \gg n$. The previous study [23] first removed the null space of the total scatter matrix via the SVD, thereby reducing the dimensionality of the data to $n - 1$, and then applied the MMC in the reduced space, where the rank of the mean subtracted matrix of X was implicitly assumed to be $n - 1$. However, in a more general case where the samples show multi-colinearity, the rank degenerates to less than $n-1$ and needs to be estimated appropriately. In contrast, the proposed algorithm does not involve the dimension reduction before applying LLDA, allowing to deal with the degenerate case.

In this way, the graph Laplacian representation of the MMC enables both the extension to unsupervised LLDA and the efficiency of the algorithm.

IV. LLDA-RFE: Feature Selection Based on LLDA

The proposed algorithm for LLDA can be used in both supervised and unsupervised cases to extract discriminant features from high-dimensional data often encountered in e.g. face recognition [16], [18], [31], [32], text categorization [5], [32], and microarray cancer classification [18], [32], [33]. In the context of microarray data analysis, the features so extracted correspond to *metagenes*, a linear combination of multiple genes, but we are rather interested in identifying discriminative genes themselves.

To this end, the previous study [23] proposed to combine the MMC with recursive feature elimination (RFE). The MMC-RFE algorithm in [23] recursively removes features with the smallest absolute values of the discriminant vectors of the MMC. The RFE approach has recently proven to be effective with regression [19], [34] as well as with support vector machine (SVM) [13]. In the present study, we propose an unsupervised recursive feature selection method using the discriminant vectors of LLDA to identify features that potentially reveal clusters in the samples.

While the number of discriminant vectors extracted by classical LDA is limited to at most $c-1$, the MMC and LLDA are capable of extracting more than $c-1$ discriminant vectors. It can also be shown that the maximum value of $J_{LLDA}(W)$ with the obtained discriminant vectors is equal to the sum of the corresponding eigenvalues. Because the eigenvalues reflect the discrimination ability, we use only the discriminant vectors corresponding to the positive eigenvalues to calculate the weight of each feature. Let $\delta_1 \geq \delta_2 \geq \ldots \geq \delta_n$ be the eigenvalues in Δ. Then, we define the weight of feature j as the sum of the absolute values of d discriminant vectors in W, i.e. $\sum_{i=1}^{d} \sqrt{\delta_i}|[W]_{ji}|$, where d is the number of positive eigenvalues. Here, the discriminant vectors are weighted by the corresponding eigenvalues.

Our proposed algorithm, LLDA-RFE, can be summarized as follows:

Algorithm: LLDA-RFE

Input: sample matrix : $\boldsymbol{X} \in \mathbb{R}^{p \times n}$
k_ℓ : the number of nearest neighbors
Output: r top-ranked features
0. Set $q \leftarrow p$,
Repeat the following steps until $q = r$
1. Construct the complete and k_ℓ-nearest neighbor graphs on $\boldsymbol{X}$ and compute $\boldsymbol{K}_g, \boldsymbol{K}_\ell, \boldsymbol{L}_g$ and $\boldsymbol{L}_\ell$;
2. Perform the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{Q}^T$;
3. Compute $\boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{Q}^T(\boldsymbol{L}_g - 2\boldsymbol{L}_\ell)\boldsymbol{Q}\boldsymbol{\Lambda}$;
4. Perform the EVD of $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{V}^T$;
5. Set $\boldsymbol{W}$ to the eigenvectors in $\boldsymbol{PV}$ corresponding to the positive eigenvalues in $\boldsymbol{\Delta}$;
6. Remove the jth feature with the smallest weight of $\sum_{i=1}^{d} \sqrt{\delta_i}|[\boldsymbol{W}]_{ji}|$.
7. Set $q \leftarrow q - 1$, form $\boldsymbol{X}$ and go to step 1.

V. Related work on unsupervised feature selection

Data variance is one of the most common unsupervised feature selection criteria, and often used as a baseline method for comparison [15], [28]. Although variance ranking can be useful for selecting features that show large variation across all samples, it is not suited for selecting ones that contribute to characterize different clusters in the samples. Hastie *et al.* [14] developed gene shaving to select informative genes from microarray data. Gene shaving iteratively removes genes having lowest correlation with the leading principal component. Because the principal components are found so that they capture the directions of maximum variance in the data, gene shaving is also unsuitable for identifying genes that reveal different clusters. The assumption that discriminative genes exhibit large variance is not necessarily valid particularly for noisy microarray data, due to the large number of irrelevant genes. Indeed, recent studies [28], [30] have shown that variance ranking, principal component analysis and gene shaving are not effective for yielding distinctive patterns between different classes of samples.

The latest and probably more effective unsupervised methods include Laplacian score [15], the $Q-\alpha$ algorithm [30] and SVD-entropy [28]. Among these, Laplacian score and SVD-entropy are employed for comparison in this study. In the following, we give a brief overview of these two methods.

A. Laplacian score

The idea of Laplacian score is to evaluate each feature by its locality preserving power, which is similar in spirit to Locality Preserving Projection [16].

Let $\boldsymbol{f}_r = (f_{r1}, \ldots, f_{rn})^T$, $r = 1, \ldots, p$, denote the rth feature for n samples. First, we construct a nearest neighbor graph in the same way as for the LLDA-RFE algorithm. Then, we compute the weight matrix $\boldsymbol{K}$, the diagonal matrix $[\boldsymbol{D}]_{ii} = \sum_j [\boldsymbol{K}]_{ji}$, and the graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$. Finally, the Laplacian score L_r of the rth feature is computed as

$$L_r = \frac{\tilde{\boldsymbol{f}}_r^T \boldsymbol{L} \tilde{\boldsymbol{f}}_r}{\tilde{\boldsymbol{f}}_r^T \boldsymbol{D} \tilde{\boldsymbol{f}}_r},$$

where

$$\tilde{\boldsymbol{f}}_r = \boldsymbol{f}_r - \frac{\boldsymbol{f}_r^T \boldsymbol{D} \boldsymbol{e}}{\boldsymbol{e}^T \boldsymbol{D} \boldsymbol{e}} \boldsymbol{e}.$$

It is worth noting that Fisher score [8] can be related to Laplacian score as shown in [15].

TABLE I
Characteristics of the Datasets Used in This Study.

Dataset	# samples	# classes	# genes
Leukemia	38	2	7129
Colon cancer	62	2	2000
Medulloblastoma	60	2	7129
Breast cancer	76	2	4918
Lung adenocarcinoma	86	2	7129
MLL	57	3	12582
SRBCT	63	4	2308

B. SVD-entropy

Let us assume that $p > n$ for a given sample matrix $\boldsymbol{X} \in \mathbb{R}^{p \times n}$. Denoting by s_j the singular values of $\boldsymbol{X}$, an SVD-based entropy is defined as [2]:

$$E = -\frac{1}{\log(n_r)} \sum_{j=1}^{n_r} V_j \log(V_j),$$

where

$$V_j = s_j^2 \Big/ \sum_{k=1}^{n_r} s_k^2.$$

Here, $n_r \leq n$ is the number of positive singular values, which is equal to the rank of $\boldsymbol{X}$. Then, the contribution of the ith feature to the entropy is defined as [28]:

$$CE_i = E(\boldsymbol{X}_{[p \times n]}) - E(\boldsymbol{X}_{[(p-1) \times n]}),$$

where $\boldsymbol{X}_{[(p-1)\times n]}$ denotes the sample matrix with the ith feature being removed.

Varshavsky *et al.* [28] have proposed three feature selection strategies based on SVD-entropy: simple ranking (SR), forward selection (FS) and backward elimination (BE). SVD-entropy-based BE has high computational complexity in the case of a large number of features, hence impractical to apply to microarray datasets. This is due to the fact that CE_i is calculated on a leave-one-out basis. Indeed, they used only SR in their experiments on microarrays. Accordingly, we employ SR in this study; top-ranked features are those with the largest values of CE_i.

VI. Experimental results

A. Datasets and preprocessing

In the experiments, we used seven public datasets of cancer microarrays. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, the different methods were primarily compared using binary-class datasets: ALL versus AML for Leukemia [10], normal versus tumor for Colon cancer [1], outcome prediction on Medulloblastoma [25], Breast cancer [27], and Lung adenocarcinoma [4]. In addition, we used multi-class datasets on MLL [3] and SRBCT [17] to assess their performances. The characteristics of these datasets are summarized in Table I, and the details are given below:

- Leukemia [10]: This Affymetrix high-density oligonucleotide array dataset contains 38 samples from 2 classes of leukemia: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML). The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

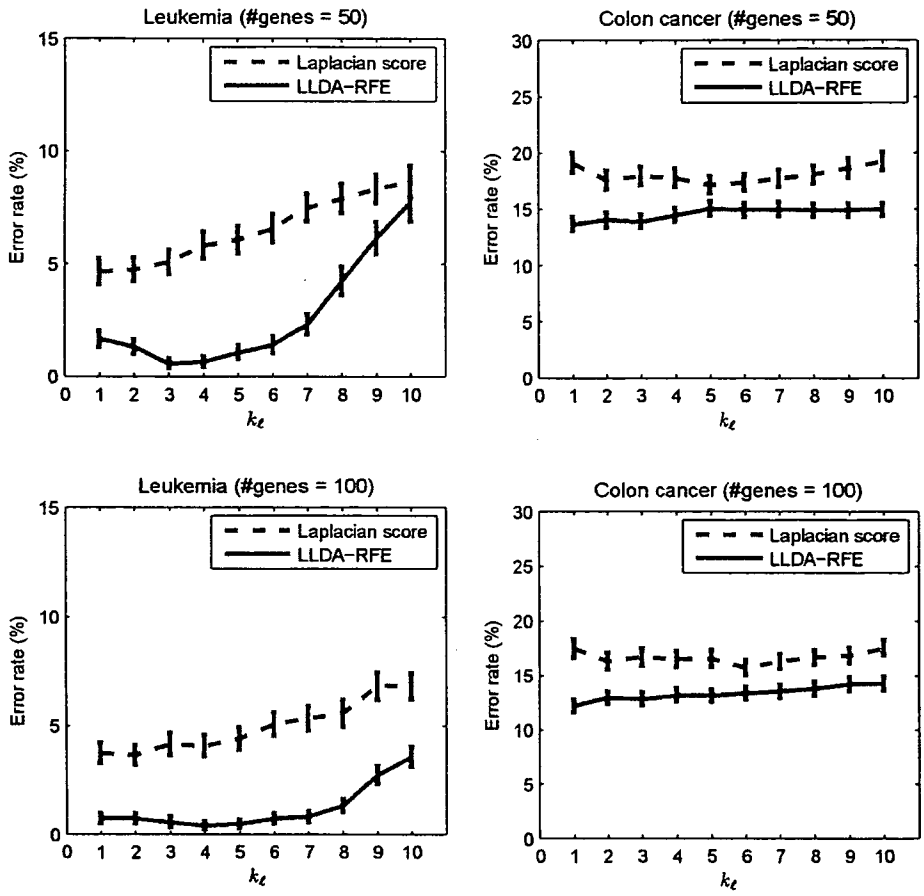

Fig. 1. Comparison between Laplacian score and LLDA-RFE using varying values of k_ℓ for Leukemia and Colon cancer

- Colon cancer [1]: This Affymetrix high-density oligonucleotide array dataset contains 62 samples from 2 classes of colon-cancer patients: 40 normal healthy samples and 22 tumor samples. The dataset is publicly available at http://microarray.princeton.edu/oncology/affydata/index.html.
- Medulloblastoma dataset [25]: This Affymetrix high-density oligonucleotide array dataset contains 60 samples from 2 classes on patient survival with medulloblastoma: 21 treatment failures and 39 survivors. The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- Breast cancer [27]: This cDNA microarray dataset contains 76 samples from 2 classes on five-year metastasis-free survival: 33 poor prognosis and 43 good prognosis. The dataset is publicly available at http://www.rii.com/publications/2002/vantveer.html.
- Lung adenocarcinoma [4]: This Affymetrix high-density oligonucleotide array dataset contains 86 samples from 2 classes on survival: an event of death for 34 and alive for 52. The dataset is publicly available at http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html.
- MLL [3]: This Affymetrix high-density oligonucleotide array dataset contains 57 samples from 3 classes of leukemia: 20 acute lymphoblastic leukemia (ALL), 17 mixed-lineage leukemia (MLL), 20 acute myelogenous leukemia (AML). The dataset is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
- SRBCT [17]: This cDNA microarray dataset contains 63 samples from 4 classes of small round blue-cell tumors of childhood (SRBCT): 23 Ewing family of tumors, 20 rhabdomyosarcoma, 12 neuroblastoma, and 8 non-Hodgkin lymphoma. The dataset is publicly available at http://research.nhgri.nih.gov/microarray/Supplement/.

For the Leukemia, Medulloblastoma, Lung adenocarcinoma and MLL datasets, expression values were first thresholded with a floor of 100 and a ceiling of 16000, followed by a base 10 logarithmic transform. Then, each sample was standardized to zero mean and unit variance across genes. For the Colon cancer dataset, after a base 10 logarithmic transform, each sample was standardized. For the Breast cancer dataset, after the filtering of genes following [27], each sample was standardized. For the SRBCT dataset, the expression profiles already preprocessed following [17] were used.

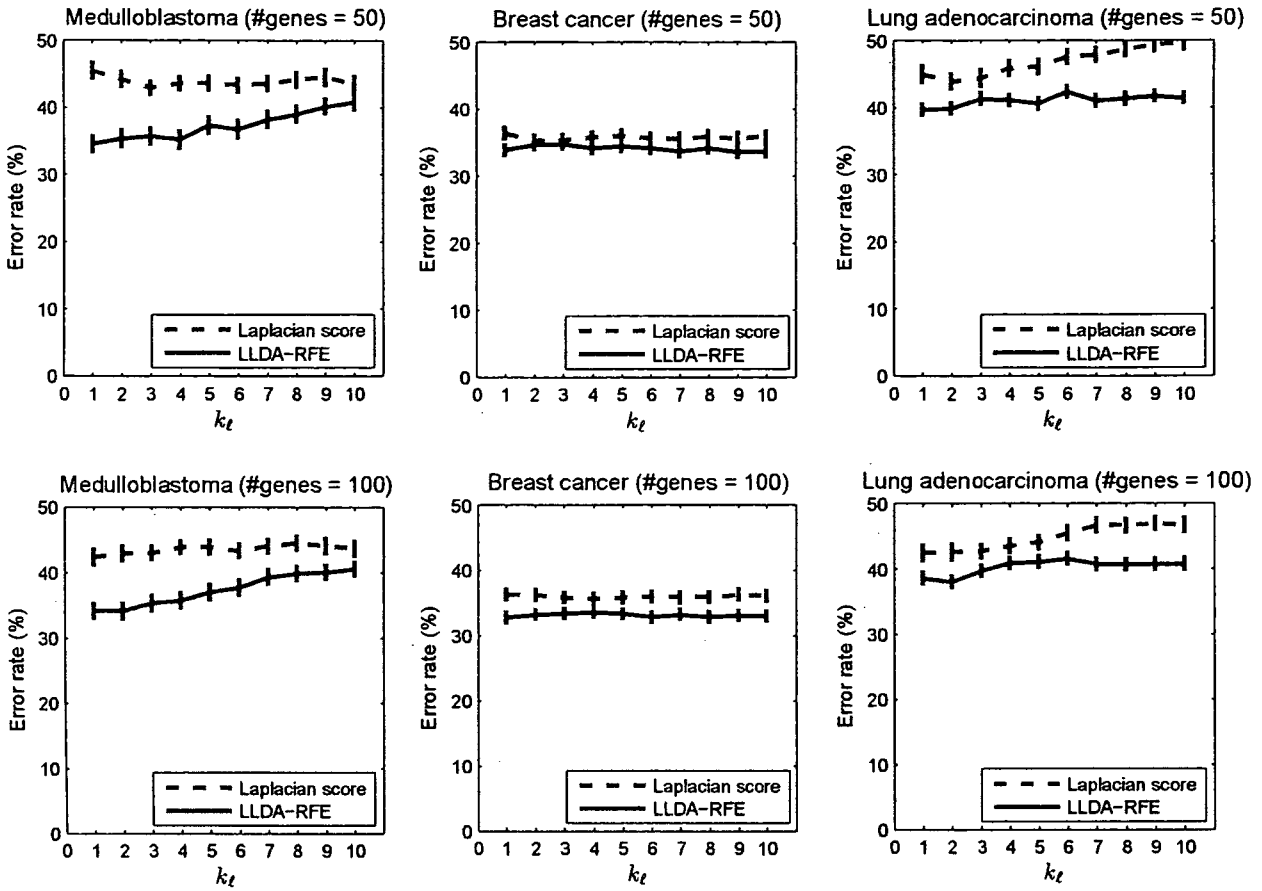

Fig. 2. Comparison between Laplacian score and LLDA-RFE using varying values of k_ℓ for Medulloblastoma, Breast cancer and Lung adenocarcinoma

B. Performance evaluation and experimental settings

We compare the performance of LLDA-RFE with those of state-of-the-art unsupervised feature selection methods, Laplacian score and SVD-entropy. The performances of the unsupervised methods are evaluated by their capability of identifying discriminative genes without using class information. Varshavsky *et al.* [28] employed the Jaccard score of clustering algorithms such as K-means, showing how clusters can be discovered by using a smaller number of genes selected from several thousand or more genes in the same samples. Wolf and Shashua [30] measured the performances by the classification accuracy of a linear SVM classifier using leave-one-out cross-validation; gene selection was performed in an unsupervised setting, but classification in a supervised setting using selected genes only. Because we also compare the performance between LLDA-RFE and a supervised gene selection method, Fisher score [8], we employed the nearest mean classifier (NMC) and measured the performances by its classification accuracy. It is known that NMC is highly effective for cancer classification despite its simplicity [29]. Note that since Fisher score is a supervised filter, it is generally expected to perform better than unsupervised methods.

We assessed the performance of each gene selection method with NMC by repeated random splitting as in [22]; the samples were partitioned randomly in a class proportional manner into a training set consisting of two-thirds of the whole samples and a test set consisting of the held-out one-third of the samples. To avoid selection bias, gene selection was performed using only the training set, and the classification error rate of the learnt classifier was obtained using the test set. This splitting was repeated 100 times. The error rates averaged over the 100 trials and the corresponding standard error rates are reported here.

To save computational time of RFE, we removed half of the genes until less than 500, and then a single gene at a time. For the computation of graph Laplacians, we used the Euclidean distance for nearest neighbor search and a simple 0-1 weighting as the similarity of the graph, i.e. $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are connected, and 0 otherwise.

C. Results and discussion

1) Effect of k_ℓ: We first compare the performance between LLDA-RFE and Laplacian score by varying the number of nearest neighbors, on the binary-class datasets: Leukemia, Colon cancer, Medulloblastoma, Breast cancer and Lung adenocarcinoma. Figs. 1 and 2 show the average error and standard error rates for $k_\ell = 1, \ldots, 10$. For LLDA-RFE, k_ℓ was fixed to the same value

during elimination. The number of genes selected and used for classification is 50 and 100.

It is clear that LLDA-RFE consistently achieves better performance than Laplacian score. This can be attributed to the difference that while Laplacian score is univariate, LLDA-RFE is multivariate and gene subsets are refined by the recursive elimination.

It may be difficult to set an appropriate value of k_ℓ in fully unsupervised settings, because we cannot rely on cross-validation unless class labels are provided and the value can also be largely dependent on the sample size of each dataset and on the potential number of clusters therein. Although an adaptive setting of the value might be preferable during elimination, our results suggest that $k_\ell = 1-3$ is a reasonable choice when applying LLDA-RFE to microarray datasets with small sample size.

2) Comparison on binary-class datasets: Table II shows the average error and standard error rates of NMC with four gene selection methods on the binary-class datasets. Fig. 3 plots the average error rates as a function of the number of genes from 1 to 100. The number of nearest neighbors for Laplacian score was set as follows: $k_\ell = 2$ for Leukemia, $k_\ell = 6$ for Colon cancer, $k_\ell = 3$ for Medulloblastoma and Breast cancer, and $k_\ell = 1$ for Lung adenocarcinoma. For LLDA-RFE, $k_\ell = 3$ was used for Leukemia and $k_\ell = 1$ for the other datasets.

We can observe that LLDA-RFE outperforms Laplacian score for a wide range of gene sizes. In comparison with SVD-entropy, LLDA-RFE yields lower error rates for Leukemia, Medulloblastoma and Breast cancer. Although SVD-entropy performs better for Colon cancer and Lung adenocarcinoma, LLDA-RFE consistently shows satisfactory performances for all the datasets. Also, note that LLDA-RFE performs better than Fisher score for Leukemia, Medulloblastoma and Breast cancer, despite the fact that LLDA-RFE is fully unsupervised. However, this does not imply that unsupervised gene selection is preferred to supervised one for these datasets. In fact, Fisher score, which can be viewed as a supervised version of Laplacian score, improves the performance of Laplacian score by using class information. Likewise, we can expect further improvement when using LLDA-RFE in a supervised manner.

3) Comparison on multi-class datasets: Table III shows the average error and standard error rates for the MLL and SRBCT datasets. Fig. 4 plots the average error rates as a function of the number of genes from 1 to 100. For Laplacian score, $k_\ell = 1$ was used for both datasets, and for LLDA-RFE, $k_\ell = 4$ and 3 were used for MLL and SRBCT, respectively.

It can be seen that LLDA-RFE reaches smaller error rates with a smaller number of genes, showing superior performance to Laplacian score and SVD-entropy. Notably, LLDA-RFE achieves even better performance than Fisher score. These results indicate that LLDA-RFE can also be useful for filtering genes from microarray samples potentially comprising multiple clusters.

In summary, our comparison using several microarray datasets has demonstrated that LLDA-RFE is effective for identifying genes that contribute to characterize different clusters in the samples. Although we used 0-1 weighting as the similarity measure, the performance could be improved by using other data-dependent similarity measures. Also, more discriminative features can be found by balancing the global and local scatters as in (3).

TABLE II
COMPARISON ON BINARY-CLASS DATASETS. BEST RESULTS IN BOLD FACE.

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
Leukemia				
20	4.9 ± 0.6	**2.6 ± 0.4**	9.6 ± 1.0	3.6 ± 0.5
50	3.9 ± 0.4	1.6 ± 0.4	4.8 ± 0.5	**0.6 ± 0.2**
100	2.9 ± 0.4	1.6 ± 0.4	3.7 ± 0.4	**0.6 ± 0.2**
Colon cancer				
20	**12.4 ± 0.6**	14.9 ± 0.6	18.2 ± 0.8	15.9 ± 0.6
50	13.0 ± 0.6	**11.5 ± 0.6**	17.4 ± 0.7	13.7 ± 0.6
100	12.8 ± 0.5	**11.7 ± 0.6**	15.7 ± 0.7	12.2 ± 0.6
Medulloblastoma				
20	38.8 ± 0.9	36.8 ± 1.1	43.7 ± 1.0	**34.1 ± 1.1**
50	38.7 ± 1.0	38.4 ± 1.0	43.0 ± 0.9	**34.6 ± 1.1**
100	38.5 ± 1.0	37.1 ± 1.0	43.1 ± 1.0	**34.2 ± 1.1**
Breast cancer				
20	35.2 ± 0.9	42.6 ± 0.9	35.1 ± 0.8	**33.3 ± 0.8**
50	36.0 ± 0.8	42.0 ± 0.8	35.4 ± 0.8	**33.8 ± 0.7**
100	36.3 ± 0.8	42.4 ± 0.8	35.9 ± 0.7	**32.8 ± 0.7**
Lung adenocarcinoma				
20	**37.8 ± 0.8**	40.5 ± 0.8	45.7 ± 1.2	42.6 ± 0.7
50	**36.3 ± 0.8**	40.0 ± 0.7	45.0 ± 1.1	39.7 ± 0.8
100	**35.1 ± 0.8**	38.3 ± 0.8	42.4 ± 1.1	38.6 ± 0.8

TABLE III
COMPARISON ON MULTI-CLASS DATASETS. BEST RESULTS IN BOLD FACE.

# genes	Fisher score	SVD-entropy	Laplacian score	LLDA-RFE
MLL				
20	7.2 ± 0.5	26.9 ± 0.9	10.2 ± 0.8	**6.1 ± 0.5**
50	6.6 ± 0.5	8.1 ± 0.6	9.4 ± 0.6	**5.1 ± 0.5**
100	5.9 ± 0.5	4.8 ± 0.4	9.1 ± 0.6	**3.8 ± 0.4**
SRBCT				
20	**3.6 ± 0.5**	22.6 ± 1.0	17.4 ± 1.2	16.2 ± 1.0
50	**2.6 ± 0.4**	17.4 ± 0.8	13.4 ± 1.0	12.0 ± 0.7
100	**4.6 ± 0.4**	11.8 ± 0.7	11.3 ± 0.9	11.4 ± 0.7

VII. CONCLUSIONS

In this paper, we have proposed a new unsupervised feature selection method based on Laplacian linear discriminant analysis (LLDA). In particular, we have extended LLDA to unsupervised cases and proposed an efficient algorithm for computing the discriminant vectors of LLDA. The LLDA-based Recursive Feature Elimination (LLDA-RFE) algorithm was applied to several microarray datasets to identify discriminative genes without using class labels.

Our comparison with other state-of-the-art unsupervised feature selection methods and with a supervised filter method has demonstrated the feasibility and effectiveness of the proposed algorithm. LLDA-RFE is capable of identifying discriminative features that contribute to reveal underlying class structures, providing a useful tool for the exploratory analysis of biological data.

A possible application of interest is the use of LLDA-RFE in semi-supervised scenarios; when labels are partially given, we

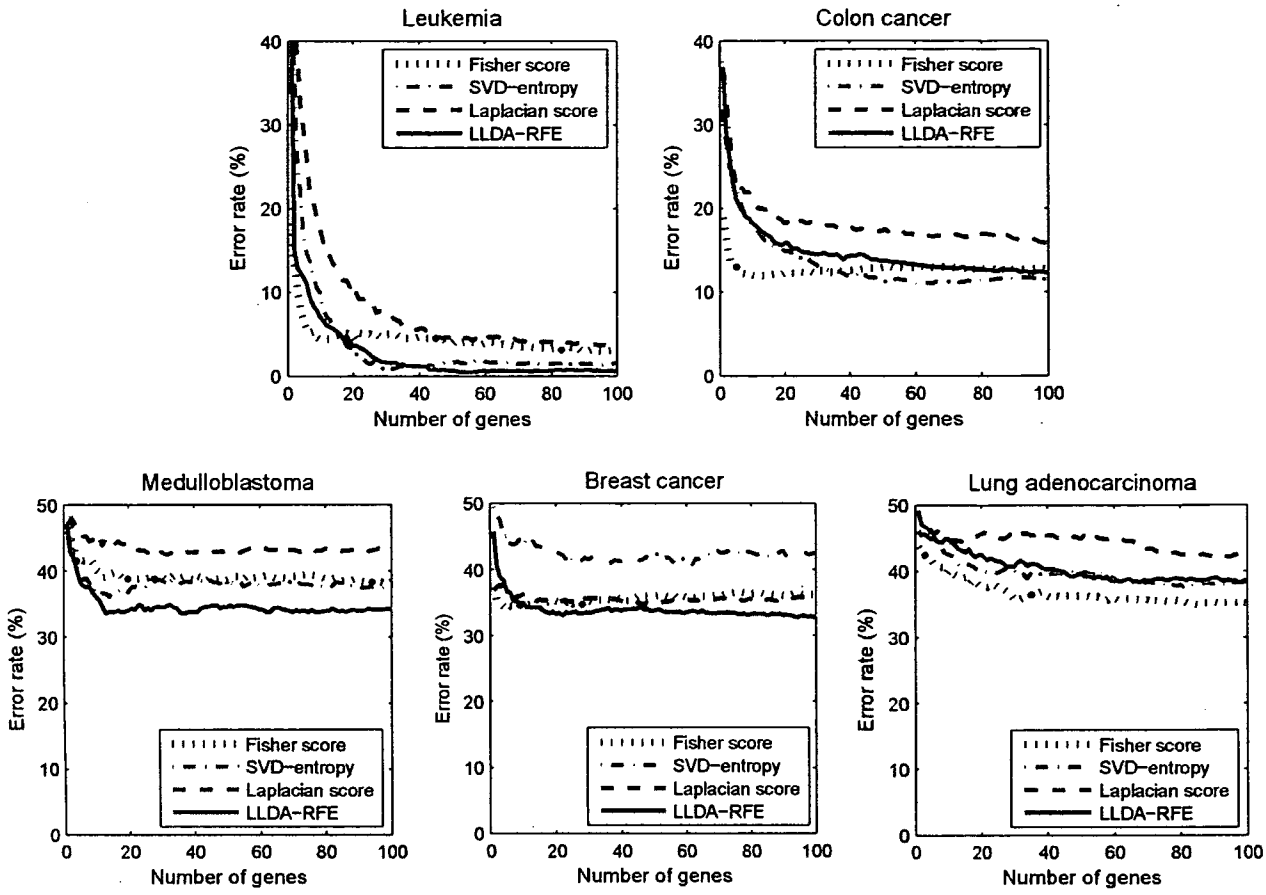

Fig. 3. Average error rates for Fisher score, SVD-entropy, Laplacian score and LLDA-RFE on binary-class datasets.

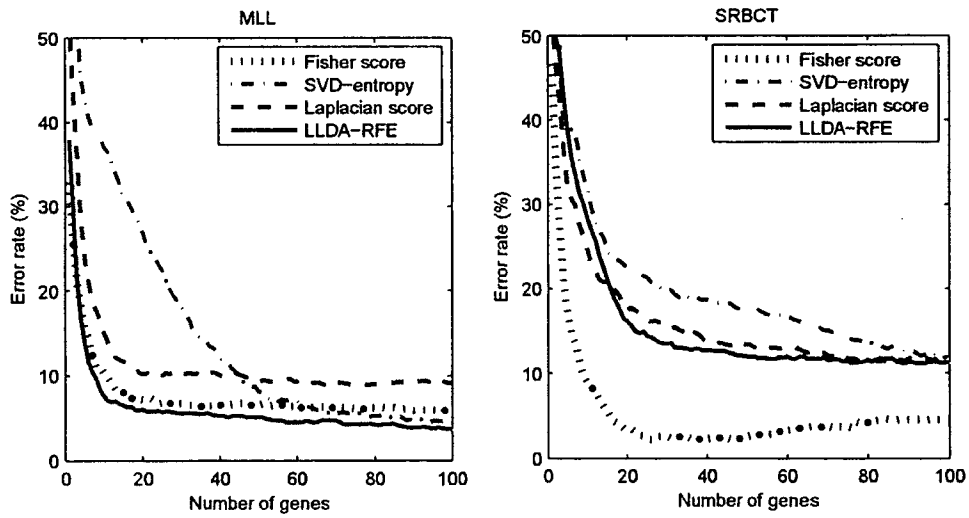

Fig. 4. Average error rates for Fisher score, SVD-entropy, Laplacian score and LLDA-RFE on multi-class datasets.

construct a graph such that samples with the same class are always connected, while those with different classes disconnected, and those with no class labels are adaptively connected or disconnected depending on the nearest neighbors.

APPENDIX
PROOF OF THEOREM 1.

It is straightforward to verify that $S_g - 2S_\ell$ can be decomposed as follows:

$$\begin{aligned} S_g - 2S_\ell &= \frac{1}{n} X(L_g - 2L_\ell)X^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell)(P\Lambda Q^T)^T \\ &= \frac{1}{n} P\Lambda Q^T (L_g - 2L_\ell) Q\Lambda P^T \\ &= \frac{1}{n} PV\Delta V^T P^T \\ &= \frac{1}{n} (PV)\Delta (PV)^T. \end{aligned}$$

Since both P and Q are orthonormal,

$$(PV)^T(PV) = V^T P^T PV = V^T V = I.$$

Hence, the theorem holds. □

ACKNOWLEDGMENTS

This work was supported by grant from the 21st Century COE program "Knowledge Information Infrastructure for Genome Science". A part of this work was done while S. Niijima was in the Graduate School of Systems Life Sciences at Kyushu University, and also supported in part by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to Prof. S. Kuhara).

REFERENCES

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA*, vol. 96, pp. 6745-6750, 1999.

[2] O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling," *Proc. Nat'l Academy of Sciences USA*, vol. 97, pp. 10101-10106, 2000.

[3] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia," *Nature Genetics*, vol. 30, pp. 41-47, 2002.

[4] D.G. Beer, S.L.R. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M.G. Taylor, M.D. Iannettoni, M.B. Orringer, and S. Hanash, "Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816-824, 2002.

[5] D. Cai, X. He, and J. Han, "Document Clustering Using Locality Preserving Indexing," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 12, pp. 1624-1637, 2005.

[6] F.R.K. Chung, *Spectral Graph Theory*, Regional Conference Series in Mathematics, no. 92, 1997.

[7] C.H.Q. Ding, "Unsupervised Feature Selection via Two-way Ordering in Gene Expression Analysis," *Bioinformatics*, vol. 19, no. 10, pp. 1259-1266, 2003.

[8] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Amer. Statist. Assoc.*, vol. 97, pp. 77-87, 2002.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, second ed. Boston, MA: Academic Press, 1990.

[10] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[11] G.H. Golub and C.F. Van Loan, *Matrix Computations*, third ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.

[14] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P.O. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology*, vol. 1, no. 2, research0003, 2000.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," In Y. Weiss, B. Schölkopf and J. Platt (eds.), *Advances in Neural Information Processing Systems 18*, pp. 507-514, Cambridge, MA: MIT Press, 2006.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H-J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328-340, 2005.

[17] J. Khan, J.S. Wei, M. Ringnér, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[18] H. Li, T. Jiang, and K. Zhang, "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 157-165, 2006.

[19] F. Li and Y. Yang, "Analysis of Recursive Gene Selection Approaches from Microarray Data," *Bioinformatics*, vol. 21, no. 19, pp. 3741-3747, 2005.

[20] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face Recognition Using Kernel Scatter-difference-based Discriminant Analysis," *IEEE Trans. Neural Networks*, vol. 17, no. 4, pp. 1081-1085, 2006.

[21] M. Loog, "On an Alternative Formulation of the Fisher Criterion That Overcomes the Small Sample Problem," *Pattern Recognition*, vol. 40, pp. 1753-1755, 2007.

[22] S. Michiels, S. Koscielny, and C. Hill, "Prediction of Cancer Outcome with Microarrays: a Multiple Random Validation Strategy," *Lancet*, vol. 365, pp. 488-492, 2005.

[23] S. Niijima and S. Kuhara, "Recursive Gene Selection Based on Maximum Margin Criterion: a Comparison with SVM-RFE," *BMC Bioinformatics*, vol. 7, 543, 2006.

[24] A.Y. Ng, M.I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," In T. Dietterich, S. Becker and Z. Ghahramani(eds.), *Advances in Neural Information Processing Systems 14*, pp. 849-856, Cambridge, MA: MIT Press, 2002.

[25] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub, "Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression," *Nature*, vol. 415, pp. 436-442, 2002.

[26] H. Tang, T. Fang, and P.-F. Shi, "Laplacian Linear Discriminant Analysis," *Pattern Recognition*, vol. 39, pp. 136-139, 2006.

[27] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[28] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *Bioinformatics*, vol. 22, no. 14, pp. e507-e513, 2006.

[29] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veenman, H. Dai, Y.D. He, L.J. van't Veer, "A Protocol for Building and Evaluating Predictors of Disease State Based on Microarray Data," *Bioinformatics*, vol. 21, no. 19, pp. 3755-3762, 2005.

[30] L. Wolf and A. Shashua, "Feature Selection for Unsupervised and Supervised Inference: the Emergence of Sparsity in a Weight-based Approach," *J. Machine Learning Research*, vol. 6, pp. 1855-1887, 2005.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: a General Framework for Dimensionality Reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.

[32] J. Ye, "Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems," *J. Machine Learning Research*, vol. 6, pp. 483-502, 2005.
[33] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, no. 4, pp. 181-190, 2004.
[34] J. Zhu and T. Hastie, "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics*, vol. 5, no. 3, pp. 427-443, 2004.

Published online 5 November 2007

Nucleic Acids Research, 2008, Vol. 36, Database issue ***D907–D912***
doi:10.1093/nar/gkm948

GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update

Yasushi Okuno[1,*], Akiko Tamon[2], Hiroaki Yabuuchi[1], Satoshi Niijima[1], Yohsuke Minowa[1], Koichiro Tonomura[1], Ryo Kunimoto[1] and Chunlai Feng[1]

[1]Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University and [2]Bio Science Group, IT Solution Div.1, Industry Solution Business Unit, Mitsui Knowledge Industry, Osaka city, Japan

Received September 6, 2007; Revised October 14, 2007; Accepted October 15, 2007

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GLIDA is a public GPCR-related Chemical Genomics database that is primarily focused on the integration of information between GPCRs and their ligands. It provides interaction data between GPCRs and their ligands, along with chemical information on the ligands, as well as biological information regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of Chemical Genomics research to easily retrieve such information from either biological or chemical starting points. GLIDA includes a variety of similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their conserved molecular recognition patterns and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. This article provides a summary of the GLIDA database and user facilities, and describes recent improvements to database design, data contents, ligand classification programs, similarity search options and graphical interfaces. GLIDA is publicly available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. We hope that it will prove very useful for Chemical Genomics research and GPCR-related drug discovery.

INTRODUCTION

The family of G-protein coupled receptors (GPCRs) represents one of the most important classes of pharmaceutical targets (1). Among the more than 1000 GPCRs encoded in the human genome, more than 400 are of potential therapeutic interest (2). Currently the drugs available on the market address only 30 GPCRs, which represent a small fraction of the GPCR target family. A large majority of human-derived GPCRs still remain promising drug targets, and thus a key goal of GPCR research related to drug design is to identify new ligands for such target GPCRs.

With the unprecedented accumulation of genomic information, databases and bioinformatics have become essential tools to guide GPCR research (3). The GPCRDB (2) and IUPHAR receptor database (IUPHAR-RD) (4) are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the GPCR gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in publicly collecting and applying chemical as well as biological information in the post-genome era (5–8). This new trend is called 'Chemical Genomics', and it aims to identify all possible chemical ligands and drugs for all targets families (9,10). There is a vast amount of information on the interactions between small molecules and proteins/genes. However, compound–protein interactions have not yet been analyzed on a large scale, and there are no effective methods to extract meaningful

*To whom correspondence should be addressed. Tel: +81 75 753 4559; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp

information from the data in a comprehensive manner. Therefore, we need to integrate chemoinformatics and bioinformatics into a common computational platform for mining of Chemical Genomics data (11).

GLIDA (GPCR-Ligand DAtabase) is a public GPCR-related Chemical Genomics database designed to simultaneously mine biological information on GPCRs and chemical information on their ligands. It provides various analytical data regarding GPCR–ligand correlations by incorporating bioinformatics and chemoinformatics techniques, and thus it should prove very useful for GPCR-related drug discovery from the viewpoint of Chemical Genomics research. There have been several major improvements to GLIDA since it was last described in Ref. (12): (i) there are more increments in the entries of the ligands and the corresponding ligand–GPCR pairs; (ii) the ligands are originally classified using a new strategy; (iii) additional options are available within the similarity search program for the GPCRs and ligands and (iv) the graphical interface to display the correlation maps between GPCRs and ligands has been enhanced.

DATA CONTENTS

GLIDA contains three types of primary data: biological information on GPCRs, chemical information on their ligands and information on binding of the GPCR–ligand pairs. The GPCR entries were acquired from human, mouse and rat entries deposited in the GPCRDB because these three species include sufficient information regarding ligands, and rats and mice are representative model animals used in drug discovery research. The ligand-binding information was manually collected and curated using various public web sites and commercial databases such as the IUPHAR-RD, PubMed (5), PubChem (5), DrugBank (13), Ki Database (14) and MDL ISIS/Base 2.5. Table 1 indicates the size and scope of the GLIDA database. In particular, we have dramatically expanded the entry number of ligands and the corresponding ligand–GPCR pairs. The latest GLIDA version includes 24077 ligand entries and 39140 GPCR–ligand pair entries, representing nearly 35-fold and 20-fold increases, respectively, since the last publication of GLIDA in 2006. The total number of GPCR entries remains unchanged, but entries with associated ligand information have increased slightly, suggesting that it is difficult to de-orphan the GPCRs whose ligands have not yet been identified (15).

GPCR and ligand data

The database lists general information on GPCR and ligand data, respectively. The general information table listing GPCRs contains gene names, family names, protein sequences (in fasta format) and links to other biological databases, such as GPCRDB, UniProt (16), IUPHAR-RD, Entrez Gene (17) and KEGG (18). The ligand result page provides a general information table containing names, molecular structures, CAS registry numbers, formulas, molecular weights, structure files and links to PubChem, KEGG, ChEBI (8) and DrugBank that are in publicly available chemical databases.

Table 1. The current numbers of GLIDA ligands and GPCRs and their respective links.

Information item	Number of entries
GPCR entries	3738
Links to Entrez Gene	3073
Links to GPCRDB	3738
Links to UniProt	3738
Links to IUPHAR	446
Links to KEGG	595
Ligand entries	24077
Cas registry number	2425
Molecular structure	23216[a]
Links to PubChem	1821
Links to ChEBI	103
Links to KEGG	664
Links to DrugBank	479
Cluster number	300[b]
GPCR–ligand pair entries	39140
GPCR entries	410
Ligand entries	24077
Activity	
Agonist	8305
Full Agonist	2325
Partial Agonist	262
Antagonist	28132
Inverse Agonist	116

[a]Molecular structures consist of MDL MOL files and original files converted into KEGG atom types. The numbers of MDL MOL files and KEGG-type files are 23216 and 23214, respectively. PCA calculation was performed for 23214 KEGG-type files.
[b]This cluster number (300) is different from the number of the selected principal components (314). No compounds were assigned to 14 principal components.

Information on binding of GPCR–ligand pairs

The interaction information relating GPCRs to particular ligands, a key issue for GPCR-related drug discovery, is deposited in a relational database. GLIDA allows users to retrieve GPCR–ligand-binding information dynamically and continuously. When users retrieve a GPCR (or ligand) entry, its result page displays all entries showing the corresponding ligands (or GPCR) entries with their binding activity types, as well as references. The references are hyperlinked with the corresponding PubMed literature. The activity types include agonist, antagonist and full, partial or inverse agonist (Table 1). Here the detail annotations such as full, partial or inverse agonist are not finished yet. The ligands classified as agonists are possible full agonists or partial agonists. Inverse agonists can be also contained among the antagonists.

WEB INTERFACE AND APPLICATION

GLIDA is available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. The web interface of GLIDA includes a GPCR search page (Figure 1a) and a ligand search page (Figure 1b). Each page consists of a classification menu and a keyword search box. The users can search a GPCR (or ligand) manually using the classification tool,

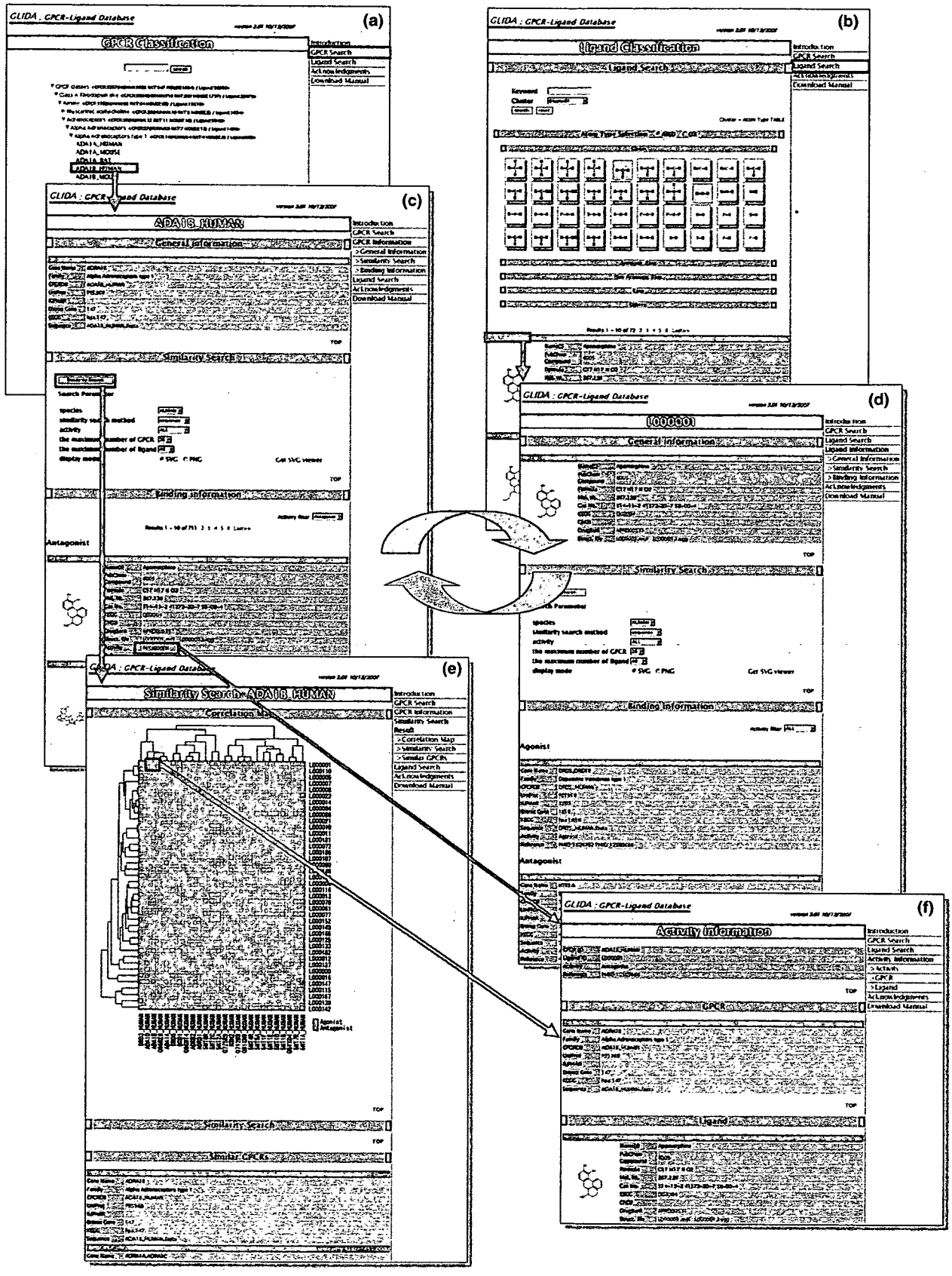

Figure 1. A screenshot of GLIDA showing linked relations among search pages (**a** and **b**), result pages (**c** and **d**), an analytical report page (**e**), and a binding information page (**f**). The analytical report page consists of a correlation map and a list resulting from a similarity search. Red and blue colors of the spots on the correlation map indicate the ligand activities of antagonists including inverse agonist and agonists including full/partial agonist, respectively.

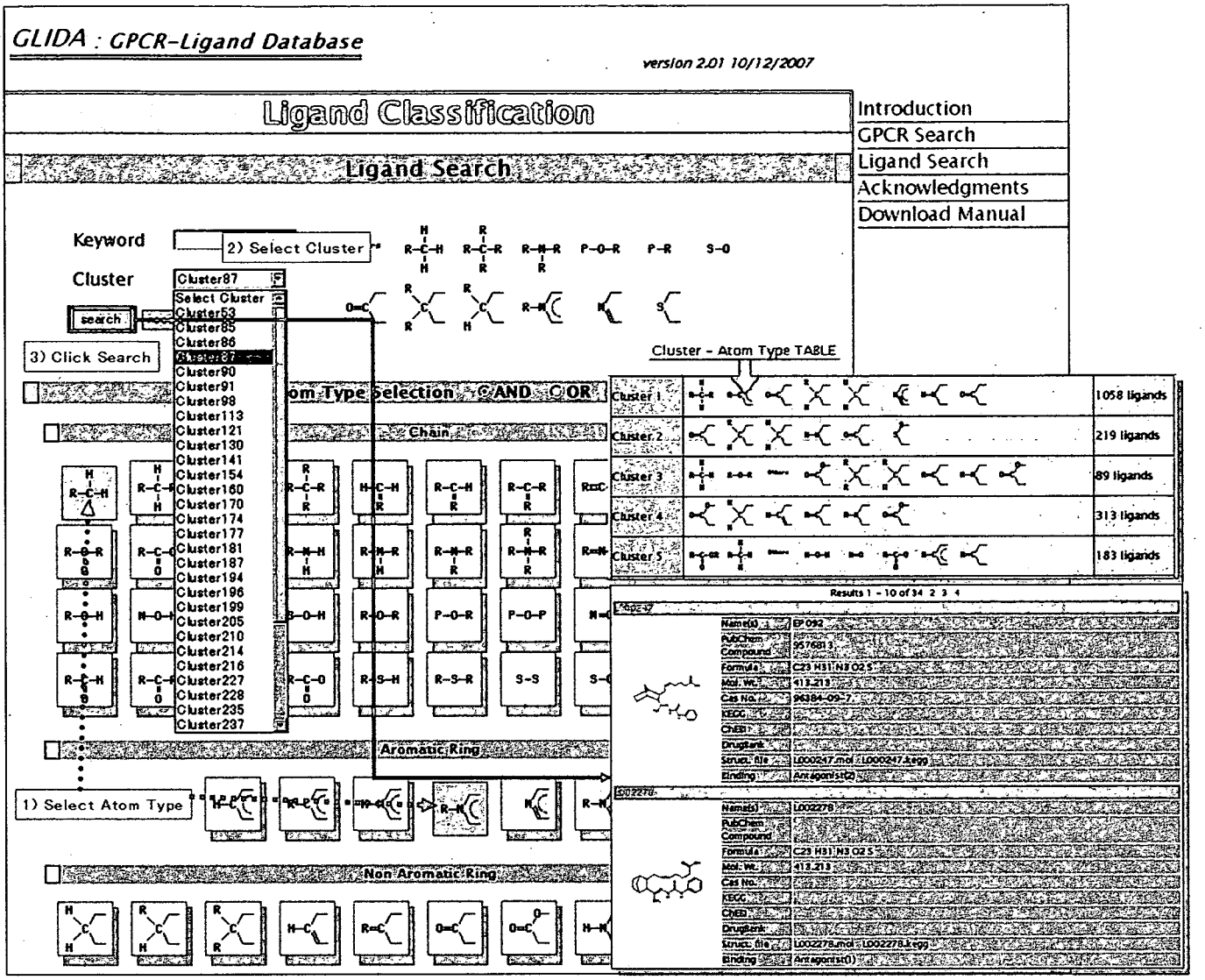

Figure 2. A screenshot of the ligand search process on the ligand classification page. Users can search the ligands from two starting points: keyword search and cluster selection. If they have a chemical structure of their query compound, the ligand search is performed using the cluster selection tool as follows. Selecting a set of atom types (step 1) that the query compound contains, the pull-down menu of cluster selection displays the list of the only clusters that include selected atom types as the principal components (step 2). By selecting a cluster from the list, users can check the principal component's atoms on the upper right section of the page. Finally, upon clicking the search button, GLIDA displays the list of all ligands classified in the selected cluster (step 3). The 'Atom Type TABLE' button links the user to the page showing the cluster size and representative atom types for each cluster.

or automatically by using the keyword search function. Every GPCR (or ligand) has its own results page (Figure 1c or d) containing a general information table regarding a GPCR (or ligand), a table of its correlated ligands (or GPCRs) and a menu button to carry out a similarity search and correlation analysis.

Classification of GPCRs and ligands

The GPCR classification table on the search page was adapted from the phylogenetic tree within the GPCRDB information system (http://www.gpcr.org/7tm/phylo/phylo.html). The GPCR classification table displays the entries of the corresponding GPCRs at the tree branches, and these are hyperlinked with the corresponding result pages (Figure 1a). GLIDA also provides an original ligand classification (Figures 1b and 2). With the great increase in ligand entries, we have to improve our method of classifying all the ligands in GLIDA. Hierarchical clustering and its tree representation, which were used in the old version of GLIDA, are unsuitable for the data mining of huge chemical databases. We therefore have adopted principal component analysis (PCA) for clustering of 23 214 ligand structures in this new version, as follows. We generated frequency profiles of the atoms and the bonds converted into the KEGG atom types from MDL MOL files of ligand entries (19). The KEGG-type profile for each ligand is shown in 'Struct. file' item of

general information table of GLIDA. PCA was applied to the data matrix consisting of 700 KEGG-type features' columns and 23214 ligand entries' rows. The resulting principal components (PCs) constitute a new set of linearly independent, orthogonal axes that capture the directions of maximum variance in the data. The samples (chemical compounds) were then projected onto these PC axes. Herein, we used the top 314 PCs as seeds of clusters that account for >80% (cumulative proportion) of the total variance. Finally, each compound was assigned to the PC cluster having the maximum score among the 314 PCs. In order to annotate the features of each cluster (PC), we selected for each PC the atom types and their bonds corresponding to the top 10 loadings having the largest magnitude. The ligand classification page displays a table of all the atom types selected by PCA (Figure 2). By clicking on some of the atoms in this table, users can search clusters that include the selected atom types. Consequently, the ligands relevant to users' interests are included in the retrieved cluster.

Similarity search and correlation map for GPCRs and ligands

The fact that similar proteins bind similar ligands is the underlying principle of the Chemical Genomics approach to drug discovery (11). GLIDA has a variety of similarity search functions for GPCRs and ligands, respectively, on its result pages (Figure 1c or d). Alignment scores for protein sequences generated by the BLAST algorithm provide similarity measures for GPCRs. In addition to sequence similarity, gene expression patterns in tissue origins and developmental stages were used as similarity measures. The expression data for each GPCR was generated from the EST sequences in different libraries served from NCBI/Unigene (http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi). We can thereby retrieve the GPCRs that present tissue-/stage-specific distribution similar to a query GPCR. For example, co-expression information on specific GPCRs enables us to speculate about GPCR-heterodimerization that might have an effect on their activity (1). Ligand similarity is defined by the dissimilarity (distance) of frequency profile patterns generated from the constitutive atoms and bonds of the chemical structure, using the KEGG atom types (19,20). From the similarity search, the most similar GPCRs (or ligands) within the users' selected parameters are retrieved and listed with their similarity scores on an analytical report page (Figure 1e). In the latest GLIDA version, various parameters have been added as search options, such as selections of species, ligand activities, displayed number of GPCRs/ligands and map graphical mode. As another result of similarity search calculations, GLIDA illustrates the correlation map (Figure 1e) showing homologous GPCRs (or ligands) and their ligands (or GPCRs) that are retrieved. This map shows spots that match the GPCRs and their ligands in a 2D matrix. The ordering along the x-axis and the y-axis are calculated respectively by two-way clustering of the GPCRs and the ligands, based on their similarities. In particular, the ordering along the x- and y-axes allows users to evaluate the sequence similarities among GPCRs and the correlation coefficients among ligands simultaneously. By analyzing the correlation patterns between GPCRs and ligands that are illustrated by these maps, we can gain detailed knowledge about their interactions. We can then utilize this information to infer possible candidates for development of GPCR-specific drugs. Furthermore, we have enhanced a graphical interface to display the correlation map between GPCRs and ligands. Graphics are an important tool to aid visualization and interpretation of high-dimensional data. The old version of GLIDA used only the PNG (Portable Network Graphics) format to display a GPCR–ligand correlation map. Due to the great increase in entries, the latest GLIDA version introduces the SVG (Scalable Vector Graphics) format, which is adaptable to an enormous correlation map size. The SVG vector image can be scaled indefinitely without loss of image quality, while the PNG bitmap image cannot. Users must install the free plug-in software on their computer in advance to use the SVG format (http://www.adobe.com/svg/viewer/install/). In the case of uninstalled devices, PNG representation should be selected as a graphical mode. Figure 1 shows an example of the GPCR–ligand search and analysis process starting from a GPCR query using GLIDA.

DISCUSSION AND FUTURE DIRECTIONS

GLIDA provides a unique database useful for GPCR-related Chemical Genomics research and drug discovery. GLIDA is distinct from other public Chemical Genomics databases because it contains original, GPCR-specific chemical entries and offers a common mining platform of bioinformatics and chemoinformatics. GLIDA provides several advantages over other databases, in that a search can be started either from a GPCR or from a ligand. Thus, searches can be carried out in a dynamic and user-friendly way. GLIDA's coverage of chemical and biological information simultaneously also provides an advantage to users by saving them the time and labor required to search multiple databases. The ligand search page is another distinct characteristic of GLIDA, in that it displays the structural distribution of ligands. It thereby facilitates research on GPCR-related drugs by incorporating structural aspects of the ligand compounds into the search. The analytical report pages resulting from the calculated structural similarities of GPCRs and ligands can give the user deep insights into the GPCR–ligand relationships. The lists of neighboring ligands (or GPCRs) and the correlation maps are useful visualization tools for analyzing correlations among the structural features and the GPCR–ligand-binding properties. Because this database system can be applied to proteins other than the GPCR family, it may also be considered as a promising database for other types of Chemical Genomics research. One critical issue is how to define similarity metrics for proteins and ligands, because the underlying principle of GLIDA is that similar receptors bind similar ligands. For example, ligand similarity can be defined by manifold representations such as graph, fingerprint and descriptors.

Protein similarity can be also measured in different ways such as overall sequence homology (phylogenetic relationships), consensus motifs, common binding sites, 3D structures and reported functional annotations. Therefore we will add new menus incorporating these various similarity metrics for GPCRs and ligands. GLIDA will be updated continuously. In particular, we are now planning to add the drawing tool of chemical structures and to expand the ligand-searching function for an arbitrary chemical query.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, from the JSPS, KAKENHI, Grant-in-Aid for Publication of Scientific Research Results and from the Ministry of Health, Labour and Welfare of Japan. Financial support from the SUNTORY INSTITUTE FOR BIOORGANIC RESEARCH, the TATEISI SCIENCE AND TECHNOLOGY FOUNDATION and the Okawa Foundation for Information and Telecommunications is gratefully acknowledged. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. George,S.R., O'Dowd,B.F. and Lee,S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nature Rev. Drug Discov.*, **1**, 808–820.
2. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
3. Strachan,R., Ferrara,G. and Roth,BL. (2006) Screening the receptorome: an efficient approach for drug discovery and target validation. *Drug Discov. Today*, **11**, 708–716.
4. Foord,S.M., Bonner,T.I., Neubig,R.R., Rosser,E.M., Pin,J.P., Davenport,A.P., Spedding,M. and Harmar,A.J. (2005) International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. *Pharmacol. Rev.*, **57**, 279–288.
5. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, 12.
6. Schreiber,SL. (2004) Stuart Schreiber: biology from a chemist's perspective. Interview by Joanna Owens. *Drug Discov. Today*, **9**, 299–303.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, D402–D404.
8. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
9. Zerhouni,E. (2003) The NIH Roadmap. *Science*, **302**, 63–72.
10. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
11. Klabunde,T. (2007) Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.*, **152**, 5–7.
12. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, **34**, D673–D677.
13. Wishart,D.S., Knox,C., Guo,A.C., Shrivastava,S., Hassanali,M., Stothard,P., Chang,Z. and Woolsey,J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.
14. Roth,B.L., Lopez,E., Beischel,S., Westkaemper,R.B. and Evans,J.M. (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.*, **102**, 99–110.
15. Civelli,O. (2005) GPCR deorphanizations: the novel, the known and the unexpected transmitters. *Trends Pharmacol. Sci.*, **26**, 15–19.
16. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **35**, D193–D197.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
18. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
19. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
20. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Chem-Bio Informatics Journal, Vol. 7, No. 2, pp.24-34(2007)

Compound-Transporter Interaction Studies using Canonical Correlation Analysis

Masato Kitajima[1,2], Yohsuke Minowa[3], Hideo Matsuda[2], Yasushi Okuno[4*]

[1]*Current Address:Life Science Systems Dept. ,PLM Solutions Div. ,Fujitsu Kyushu System Engineering Limited, 2-2-1,Momochihama,Sawara-ku,Fukuoka,814-8589,Japan*
[2]*Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan,*
[3]*National Institute of Biomedical Innovation Toxicogenomics Informatics Project 7-6-8 Asagi Saito Ibaraki-City Osaka, 567-0085,Japan*
[4]*Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

**E-mail: okuno@pharm.kyoto-u.ac.jp*

(Received November 3, 2007; accepted November 15, 2007; published online December 5, 2007)

Abstract

The efficient screening of lead compounds or drug candidates for efficacy and safety is critically important during the early stage of drug development. Compounds are usually screened from a diverse 'chemical space' based only on its pharmacological effects, but this screening is not enough to guarantee drug safety. To solve this problem, we devised a chemical space that takes into account interaction information with proteins such as drug transporters. We also created and evaluated a mathematical model for predicting compound-transporter interactions. This was achieved by first generating an interaction correlation matrix based on drug transporters and their corresponding inhibitor compounds. To implement a screening scheme that takes into account interaction with drug transporters, we created a model using Canonical Correlation Analysis (CCA) that makes use of the known information on interaction between drug transporters and their corresponding inhibitors. Cross-validation of the model gave satisfactory test results and a physiologically relevant chemical space was constructed based on the model.

Key Words: Pharmacokinetics, Transporter, chemoinformatics, bioinformatics

Area of Interest: Molecular Recognition

1. Introduction

During the drug development process it is important to screen compounds for efficacy and safety at an early stage in order to prevent unnecessary and costly analysis later on. It is thought that the diverse chemical space may contain as much as 10^{60} chemical structures or more. Searching for a drug candidate with a good balance of efficacy and safety from this huge chemical space is obviously very difficult, as evidenced from the fall in the number of new drug applications in recent years [1][2].

The previously used screening approach involved sampling a diverse chemical library for those leads that display only promising pharmacological effects i.e. drug efficacy. It is important to select a compound with excellent pharmacokinetic properties (drug safety), not just pharmacological effects, when screening at the early stage of drug development. When analyzing the pharmacokinetic properties of drug candidates, ligand interactions with Phase I enzymes such as cytochromes P450, Phase II conjugation enzymes (e.g. GST and sulfotransferases), as well as transporter proteins that play a crucial role in Phase III, must be considered [3]. Of these, transporter proteins, which are important in facilitating absorption of compounds in the intestines as well as the degree of penetration across the blood-brain barrier, play a central role in determining the bioavailability of drugs.

This paper focuses on transporter proteins that play an important role in drug pharmacokinetics. The interaction studies were carried out based on information on the interactions of the compounds and the transporter proteins. To implement a screening scheme that takes into account interaction with drug transporters, we created a model using CCA that exploits the characterized interactions between drug transporters and their corresponding inhibitors. The model is evaluated and then used to create a physiologically relevant chemical space.

2. Method

2.1 Gathering of compound-transporter interaction data

The compound-transporter interaction data used in this study was extracted from the ADME Database (developed by Fujitsu Kyushu Systems Engineering Ltd., Fukuoka, Japan) [4][5][6]. The database is a collection of information on drug transporters as well as drug metabolizing enzymes found in the literature. Two kinds of transporter proteins were selected for this study; the ABC transporter family and the SLC transporter family. Compounds that interact with these transporter families were also extracted from the database. The compound-transporter interaction type available in the database includes substrates, inhibitors, inducers and activators. However, for the purpose of this study only the inhibitors were selected.

A total of 17 ABC transporter families and 110 different SLC transporter proteins were selected for this study. Data concerning the interaction of these transporter proteins with known compounds was extracted from the ADME Database. The database contains 5,860 compound-transporter interactions between the selected 117 transporter proteins and their interacting 3,275 compounds.

The selected transporter proteins are as follows.

Table 1. List of transporter proteins used in the study

ABCA1	SLC1A1	SLC5A6	SLC7A7	SLC19A1	SLC23A1
ABCA2	SLC1A2	SLC5A7	SLC7A8	SLC19A2	SLC23A2
ABCA9	SLC1A3	SLC5A8	SLC10A1	SLC21A11	SLC26A2
ABCA10	SLC1A4	SLC5A9	SLC10A2	SLC21A12	SLC26A3
ABCB1	SLC1A5	SLC6A1	SLC10A4	SLC21A14	SLC26A4
ABCB4	SLC1A6	SLC6A11	SLC13A1	SLC21A2	SLC26A6
ABCB5	SLC1A7	SLC6A12	SLC13A2	SLC21A20	SLC26A7
ABCB11	SLC2A1	SLC6A13	SLC13A3	SLC21A3	SLC26A8
ABCC1	SLC2A10	SLC6A14	SLC13A4	SLC21A6	SLC26A9
ABCC2	SLC2A11	SLC6A2	SLC13A5	SLC21A8	SLC27A4
ABCC3	SLC2A12	SLC6A3	SLC15A1	SLC21A9	SLC28A1
ABCC4	SLC2A13	SLC6A4	SLC15A2	SLC22A1	SLC28A2
ABCC5	SLC2A2	SLC6A5	SLC15A4	SLC22A11	SLC28A3
ABCC10	SLC2A3	SLC6A6	SLC16A1	SLC22A12	SLC29A1
ABCC11	SLC2A4	SLC6A9	SLC16A10	SLC22A16	SLC29A2
ABCG1	SLC2A6	SLC7A1	SLC16A3	SLC22A2	SLC29A4
ABCG2	SLC2A7	SLC7A10	SLC16A5	SLC22A3	SLC32A1
	SLC2A8	SLC7A11	SLC16A7	SLC22A4	SLC36A1
	SLC4A4	SLC7A2	SLC17A1	SLC22A5	SLC38A1
	SLC5A1	SLC7A3	SLC18A1	SLC22A6	SLC38A4
	SLC5A2	SLC7A5	SLC18A2	SLC22A7	SLC38A5
	SLC5A4	SLC7A6	SLC18A3	SLC22A8	SLC43A2

2.2 Organizing the collected data

A correlation matrix of compound-transporter interactions was constructed based on the collected compound-transporter interaction information. Here compounds that interact with a transporter protein are flagged '1', and those that do not interact are flagged '0' as shown in Table2.

The similarity between transporter proteins was calculated based on this interaction correlation matrix. The Tanimoto coefficient found below was used to evaluate similarity [7] .

$$\text{Tanimoto}(X,Y) = \frac{C}{A + B - C}$$

A: No. of bits in X that were flagged '1'
B: No. of bits in Y that were flagged '1'
C: No. of flagged bits common to both X and Y

The interaction similarity of the compounds was also defined as Tanimoto coefficient based on the correlation matrix. We used the distance measure transformed from the Tanimoto coefficient as shown below:

$$\text{Distance}(X,Y) = 1 - \text{Tanimoto}(X,Y)$$

Table 2. Excerpt from the correlation matrix of compound-transporter interactions.

Compound Name \ Transporter Name	ABCA1	ABCA10	ABCA2	ABCA9	ABCB1	ABCB11	ABCB4	ABCB5	ABCC1	ABCC10	ABCC11	ABCC2	ABCC3	ABCC4	ABCC5	ABCG1	ABCG2
Verapamil	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0
Cholyltaurine, Taurocholic acid, Taurocholate	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	1
4,4'-Diisothiocyanostilbene-2,2'-disulfonic acid, DIDS	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0
Phloretin	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
Bromosulfophthalein, Sulfobromophthalein, BSP	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
Cyclosporin A, Cyclosporine, Cyclosporin, Ciclosporin	0	0	0	0	1	1	1	0	1	1	0	1	1	1	0	1	1
Probenecid	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	0
Indomethacin	0	0	0	0	1	0	0	0	1	0	0	1	1	1	0	0	0
Progesterone	0	0	0	0	1	1	0	0	1	0	0	1	0	1	1	0	1
Quinidine	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
Cimetidine	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Phloridzin, Phlorizin	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Pravastatin, Pravastatin acid	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1
Rifampicin, Rifampin	1	0	0	0	1	1	0	0	1	0	0	1	1	0	0	0	1
1-Methyl-4-phenylpyridinium, MPP(+)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Estradiol 17beta-D-glucuronide, E(2)17betaG	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0
L-glutamine, Gln	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L-leucine, L-Leu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Methotrexate	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	0	1
MK571, MK-571	0	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	1
Chlorpromazine	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Desimipramine, Desipramine	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Diclofenac	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
Ketoprofen	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0

2.3 Canonical Correlation Analysis (CCA)

Next, CCA was performed using the collected compound-transporter interaction data. Dragon X was used to calculate the compound descriptors [8]. A total of 929 descriptors were calculated. Highly correlated descriptors were grouped together and then filtered to give a final total of 324 compound descriptors.

The transporter proteins were calculated as bigram (two amino acids) frequency in protein sequences, and were used to generate a total of 400 protein descriptors. CCA was then performed, which generated 324 components. CCA is a technique to extract common features from a pair of multivariate data (chemical and protein descriptors). CCA finds a linear transformation of the chemical and protein spaces such that the correlation coefficient is maximized. Therefore we can construct the chemical space with the higher correlation to the protein space by extraction of the some components with the higher correlation coefficients.

2. 4 Cross Validation

A 5-fold cross-validation test was performed using only the 44 CCA components with P value < 0.01 of correlation coefficient test. The whole training compound set was divided into five sets. The first set was left out for testing and the remaining four sets were used to train a model. The compounds from the first set were then used to evaluate the trained model. The model was evaluated by setting a Euclid distance threshold from a test compound and using the closest neighboring compounds within this threshold for prediction. The procedure was repeated for all five sets, each time leaving out one set for testing, until all compounds from all five sets had been evaluated.

3. Result

3.1 Analysis of compound-transporter interactions

A significant number of compounds were found to inhibit more than one transporter using the collected compound-transporter interaction data. Indeed, out of these compounds, 183 were identified as inhibiting five or more transporters. The list below shows the frequency for each "interaction count" of a compound.

Table 3. Frequency of Interaction count for a single compound

Interaction Count	Frequency
18	1
17	1
16	2
15	3
14	1
13	2
12	4
11	6
10	8
9	15
8	19
7	18
6	37
5	66
4	118
3	355
2	393
1	2226

Moreover, the similarity of each transporter protein was calculated from the interaction matrix profile by using the Tanimoto coefficient. The result of clustering based on this similarity measure is shown in Figure 1. As shown in the figure, ABCG1 and ABCB4 are similar by interaction pattern even though the sequence similarity between both proteins is very low. Analogous results were found for ABCG2, which was shown to be similar to ABCC1 and ABCC2.

Moreover, Cyclosporin A that interacts with 15 kinds of transporter proteins, and MK571 that interacts with 11 kinds of transporter proteins were examined and compared as shown in Figure 2 [9][10]. Even though the two compounds share a low degree of structural similarity, they were found to interact with the same 10 transporter proteins. Our results demonstrate that it is not possible to explain all the similarities in interaction by simply comparing the structure of the relevant compounds or by protein sequence alignments alone. We then performed CCA on the collected interaction data to construct a correlation model for building a chemical space that reflected the classification of the transporters.