Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 37 (2004) 366–379
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Mapping high-dimensional data onto a relative distance plane—an exact method for visualizing and characterizing high-dimensional patterns

R.L. Somorjai*, B. Dolenko, A. Demko, M. Mandelzweig, A.E. Nikulin, R. Baumgartner, N.J. Pizzi

Institute for Biodiagnostics, National Research Council Canada, 435 Ellice Avenue, Winnipeg MB, Canada R3B 1Y6

Received 22 April 2004
Available online 15 September 2004

Abstract

We introduce a distance (similarity)—based mapping for the visualization of high-dimensional patterns and their relative relationships. The mapping preserves *exactly* the original distances between points with respect to any two reference patterns in a special two-dimensional coordinate system, the relative distance plane (RDP). As only a single calculation of a distance matrix is required, this method is computationally efficient, an essential requirement for any exploratory data analysis. The data visualization afforded by this representation permits a rapid assessment of class pattern distributions. In particular, we can determine with a simple statistical test whether both training and validation sets of a 2-class, high-dimensional dataset derive from the same class distributions. We can explore any dataset in detail by identifying the subset of reference pairs whose members belong to different classes, cycling through this subset, and for each pair, mapping the remaining patterns. These multiple viewpoints facilitate the identification and confirmation of outliers. We demonstrate the effectiveness of this method on several complex biomedical datasets. Because of its efficiency, effectiveness, and versatility, one may use the RDP representation as an initial, data mining exploration that precedes classification by some classifier. Once final enhancements to the RDP mapping software are completed, we plan to make it freely available to researchers.

1. Introduction

Biomedical [magnetic resonance (MR), infrared (IR) or Raman] spectra of biofluids and tissues [1–5], mass spectra, e.g., from proteomics [6–9] and microarray expression profiles [10–21] are being acquired at an ever-increasing rate. The principal goal is to discriminate non-invasively diseases and disease states. Biomedical data are characterized by relatively few *patterns* (N = O(10)–O(100)) that are initially presented in a very high-dimensional feature space (L = O(1000)–O(10000)). A major problem is the difficulty of reliably visualizing such high-dimensional patterns and their relative relationships. An accurate visualization would help the assessment of the characteristics, peculiarities, failings, etc., of the high-dimensional dataset, prior to the desired processing (e.g., classification).

There are several techniques that map L-dimensional patterns to a lower, M-dimensional space, $M \ll L$, typically $M = 2$ or 3. We can group them into two categories: linear and nonlinear. An excellent early review is in [22], with experimental comparisons in [23]. All mapping methods aim to preserve *all* interpoint distances as accurately as possible.

Linear methods generally involve no extensive optimization. High-dimensional data are frequently mapped onto a plane, whose coordinate axes are formed by the

* Corresponding author. Fax: +1 204 984 5472.
E-mail address: ray.somorjai@nrc-cnrc.gc.ca (R.L. Somorjai).

1532-0464/$ - see front matter.
doi:10.1016/j.jbi.2004.07.005

first two principal components (PCs) obtained from principal component analysis (PCA) [24]. However, these PCs rarely discriminate between patterns from known classes, because PCA finds orthogonal directions that sequentially explain the variance in the data, independently of class information. (Good class separation is accidental.) Multidimensional scaling (MDS) [25–28] and Kohonen's self organizing map (SOM) [29] are nonlinear variants, requiring the minimization of some objective (cost) function of the distances. None can preserve all relative distances exactly.

In general, MDS-based mapping helps visualize *proximity* relations of samples, represented in the *original* L-dimensional feature space by (Euclidean) distances D^{o}_{ij} between points i and j, with $D^{\mathrm{o}}_{ij} = (\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{t}}(\mathbf{X}_i - \mathbf{X}_j)$ is the (squared) distance between pattern vectors $\mathbf{X}_i$ and $\mathbf{X}_j$. The corresponding relations in the *reduced*, transformed M-dimensional space are $D^{\mathrm{R}}_{ij} = (\mathbf{Y}_i - \mathbf{Y}_j)^{\mathrm{t}}(\mathbf{Y}_i - \mathbf{Y}_j)$ is the (squared) distance between mapped versions Y of these same vectors [25]. Frequently, the $\mathbf{X} \to \mathbf{Y}$ mapping does not exist and only some *proximity* information is available. This is typically represented in terms of *dissimilarity* values. The (symmetric) dissimilarity between objects i and j is denoted by $\delta_{ij} = \delta_{ji}$. Quite often a monotonic nonlinear transformation $G(\delta_{ij}) = D^{\mathrm{R}}_{ij}$ is applied to the raw dissimilarity values. A general choice for the objective function J to be minimized is

$$J = \sum_{i<j} w_{ij}(D^{\mathrm{o}}_{ij} - D^{\mathrm{R}}_{ij})^2$$

with weights w_{ij} controlling whether short, medium or long distances are matched best. As an example, a general, one-parameter mapping proposed in [30] minimizes the objective function J_q, where

$$J_q = (1/D_q)\sum_{i<j}[D^{\mathrm{o}}_{ij}]^{q-1}(D^{\mathrm{o}}_{ij} - D^{\mathrm{R}}_{ij})^{q+2},$$

$$D_q = \left(\sum_{i<j} D^{\mathrm{o}}_{ij}\right)^{q+1},$$

with q a user-defined parameter. The choice of q determines whether the shorter or the longer distances are reproduced more faithfully in the reduced, low-dimensional space. For $q = 0$ we recover Sammon's mapping [31]. Whatever the value of q, the optimal mapping is obtained by some multivariate minimization process, with the minimization space typically multimodal. One serious difficulty with these approaches is that unless an explicit form is available for the $\mathbf{X} \to \mathbf{Y}$ mapping, the mapping/optimization has to be repeated for new patterns.

Kohonen's self-organizing maps (SOMs) are based on neural network ideas [29,32,33]. SOMs have recently been applied to microarray data [34,35] and to data visualization in particular [36]. SOMs are subject to the same general constraints as the other mapping methods. An additional restriction is that the mapping is to a 2- or 3-dimensional grid, typically rectangular or hexagonal. Furthermore, only topological relations are preserved, and a colouring scheme, such as the U-matrix method [37] is needed to visualize the actual distance interrelations. SOM tends to preserve the shorter distances. A recent variant, ViSOM [38], is claimed to be computationally simpler than SOM, with the additional advantage that not only the topological relations, but also the actual map distances are preserved (approximately). Furthermore, no retraining of the map is needed when new data points are presented.

Projection Pursuit (PP) [39,40] approaches the problem differently. Its 1-dimensional version attempts to find "interesting directions," by projecting (mapping) the data onto lines traversing the input space. "Interesting" is usually defined as "least Gaussian." As an extreme example, if the projection is multimodal, it is definitely "interesting" because it indicates the presence of structure, i.e., clusters, groupings in L-space. To find these "interesting directions," PP uses nonlinear optimization, with all its attendant, known difficulties. Furthermore, PP requires the setting of several parameters and the selection of a monotonically decreasing function. Note that when the projection is directly from the original L-space, PP is a linear method.

Ideally, we would like a visualization method that does not require optimization, yet displays the high-dimensional data in some low (3-, 2- or even 1-) dimensional manifold, without distorting any of the original L-space distances. Although this cannot be done exactly for *all* distances, we propose a simple approach that achieves certain important aspects of this goal. It can be used either as an exploratory tool, or as a confirmatory one, if class labels are available for members of the different classes comprising the dataset.

Our approach is a distance (similarity)-based, intrinsically nonlinear projection. It only requires a *single* computation of a distance matrix $\mathbf{D} = \{D_{jk}\}$, $j < k = 1, \ldots, N$ in some user-selected metric. This is computationally quite feasible for the number of samples ($N = \mathrm{O}(100)$) we generally encounter in biomedical applications. Our method is based on the simple but essential fact that the three distances between any three points in the original L-space are *exactly* preserved when displayed in a 2-dimensional coordinate system (S, T). Although this appears trivial and was recognized as the basis of a mapping approach called the *triangulation method* [41], it has not been exploited, especially in the context we propose. This is likely because the emphasis for all mapping methods always seems to have been on the best possible (hence necessarily approximate) preservation of *all distances*. In particular, the triangulation approach, although recognizing that $2N - 3$ of the $N(N-1)/2$ interpattern distances can be preserved exactly, only focuses on the display and maintenance of the minimum spanning tree (MST) of the

graph whose nodes are patterns and whose edge weights are the interpattern distances. However, MST still has some arbitrariness in its planar representation and does not appear to provide as good a discriminatory visual display as does our mapping.

We call our mapping plane the relative distance plane (RDP). Our purpose and emphasis is different; in particular, we avoid any nonlinear optimization since we do not need to insist that in the projected space all original distances be preserved as closely as possible. Although we expound the RDP mapping in the context of supervised pattern recognition (known class labels), it still should be viewed as an exploratory data analysis method. As the examples demonstrate, its role is to help assess the properties/quality of the high-dimensional data via easy 2- and 1-dimensional visualization.

2. The RDP mapping method

Denote the members ("patterns") of the N-sample dataset in the L-dimensional feature space by $\mathbf{X}_j = \{X_{j1}, \ldots, X_{jL}\}$, $j = 1, \ldots, N$. This produces an $N \times L$ data matrix.

Our proposed procedure consists of the following steps:

1. Choose some distance (or similarity) measure and compute the corresponding $N \times N$ distance matrix.
2. Select any two points $\mathbf{R}_1 (\equiv \mathbf{X}_j)$ and $\mathbf{R_2} (\equiv \mathbf{X}_k)$ in the original L-dimensional space as *reference points* ("patterns"); the distance $D(\mathbf{R_1}, \mathbf{R_2}) \equiv D_{12}$ has already been computed in Step 1. The line through $\mathbf{R}_1$ and $\mathbf{R}_2$ defines a *reference axis* onto which one can further project the data.
3. For each pattern $\mathbf{X}_m$, $m \neq j, k$, of the dataset, denote its distances to the reference patterns by $D_{1m} \equiv D(\mathbf{X}_m, \mathbf{R}_1)$ and $D_{2m} \equiv D(\mathbf{X}_m, \mathbf{R}_2)$. The Euclidean (S, T) coordinates in the RDP for all points $\mathbf{X}_m$, $m = 1, 2, \ldots, N-2$, $m \neq j, k$, are:

$$S[\mathbf{X}_m] = (D_{12}^2 + D_{1m}^2 - D_{2m}^2)/2D_{12} - 1, \tag{1a}$$

$$T[\mathbf{X}_m] = (D_{1m}^2 - S^2)^{1/2}. \tag{1b}$$

Note that from the three given distances, it is equally easy to compute the (S, T) coordinates in another distance metric. Naturally, the RDP display, because it is distance-based, is independent of the metric the (S, T) coordinates are eventually expressed in. Only the (S, T) coordinates will be different. Eqs. (1a) and (1b) assume Euclidean distances, but in the L_1 norm the result is even simpler:

$$S[\mathbf{X}_m] = (D_{1m} - D_{2m} + D_{12})/2,$$
$$T[\mathbf{X}_m] = (D_{1m} + D_{2m} - D_{12})/2.$$

In implementing the mapping, for display and comparison purposes we place $\mathbf{R}_1$ at $(-1, 0)$, and $\mathbf{R}_2$ at $(1, 0)$ of the (S, T) coordinate system of the RDP. Thus, D_{12} is scaled (to 2) and the consequent scaling of D_{1m} to D_{1m}^* and D_{2m} to D_{2m}^* follows. In this scaled coordinate system, Eqs. (1a) and (1b) become:

$$S^*[\mathbf{X}_m] = (D_{2m}^{*2} - D_{1m}^{*2})/4, \tag{1a*}$$

$$T^*[\mathbf{X}_m] = (D_{1m}^{*2} - S^{*2})^{1/2}. \tag{1b*}$$

(For notational convenience, in the rest of the paper we remove the * from these scaled equations.)

In Step 1, the distance measure used in the input feature space can be the standard Euclidean distance, or its extension by some "kernel trick," any of the Minkowski distances, the Mahalanobis distance, the Anderson–Bahadur distance [42], some user-defined or data-driven weighted distance, or even some arbitrarily defined *dissimilarity measure*, (e.g., correlation, converted into a distance). The $N \times N$ distance matrix can be augmented by the 2 class centroids $\mathbf{C}_i$, $i = 1, 2$, creating the final $Q \times Q$ distance matrix, $Q = N + 2$.

In the original L-space, the distance of point $\mathbf{X}_m$ from any reference axis is $T[\mathbf{X}_m]$. $T_{\max}(\mathbf{R}_1, \mathbf{R}_2)$ is the radius of the hyper-cylinder that contains all Q points, with axis $\mathbf{R}_1$–$\mathbf{R}_2$. One may create a histogram by projecting all points in the RDP onto this axis. If the histogram is multimodal, then the $\mathbf{R}_1$–$\mathbf{R}_2$ axis provides a "potentially interesting" line traversing the original L-space. The points in the RDP will then display more-or-less distinct clusters. The current $\mathbf{R}_1$–$\mathbf{R}_2$ axis may not represent the most "interesting" line in L-space. However, all $N_{\text{pair}} = Q(Q-1)/2$ pairs of points $(\mathbf{R}_j, \mathbf{R}_k)$ are available for testing and inspection. When projecting the patterns onto the reference axis, one can view our approach as a discretized version of Projection Pursuit, consisting only of N_{pair} lines through existing pattern pairs, or some other readily computable directions.

Thus we can display all points of the dataset, without distorting their original distances *to the two reference patterns*. (We emphasize again that the original L-space distance between any two arbitrary patterns is not generally preserved in the RDP. However, for our purpose this is not relevant.) For any reference pair the RDP mapping preserves exactly $2N - 3$ distances.

An easy 3-dimensional visualization and appreciation of what mapping to the RDP means is to view the $\mathbf{R}_1$–$\mathbf{R_2}$ reference axis as the spine of a Rolodex, and the $N - 2$ points in L-space as $N - 2$ transparent cards in the Rolodex, with the relative position of any given point marked on its own card. Then the RDP mapping corresponds to aligning the $N - 2$ cards in a common plane. (This analogy is exact if the mapping is from 3 to 2 dimensions.)

In Fig. 1, we show these aspects of the RDP map, using a real-life, 2-class example (this particular

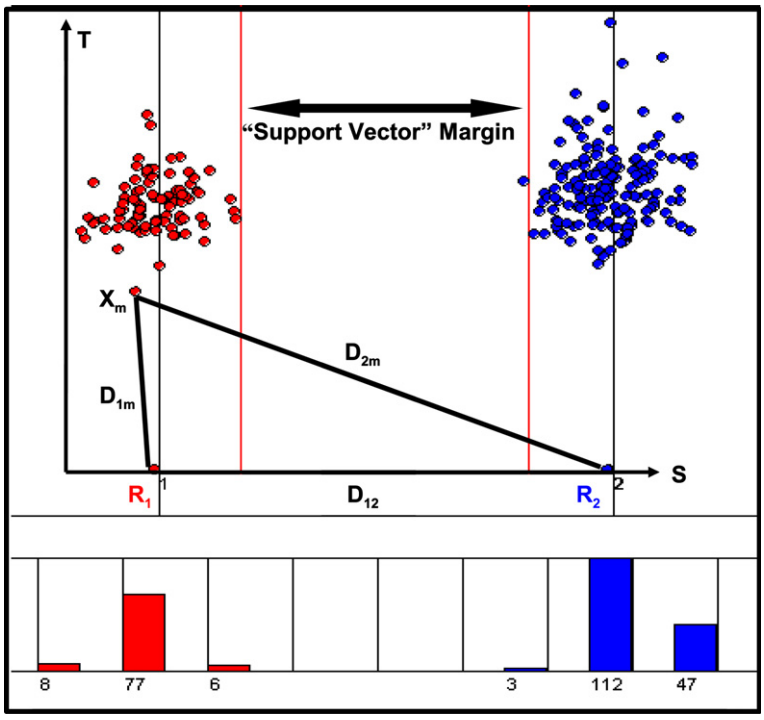

Fig. 1. Two-class RDP mapping from a 10-dimensional feature space. Class 1 (red) and class 2 (blue disks). Black double arrow: "support vector" margin (i.e., the smallest distance along the RDP coordinate S between the two classes).

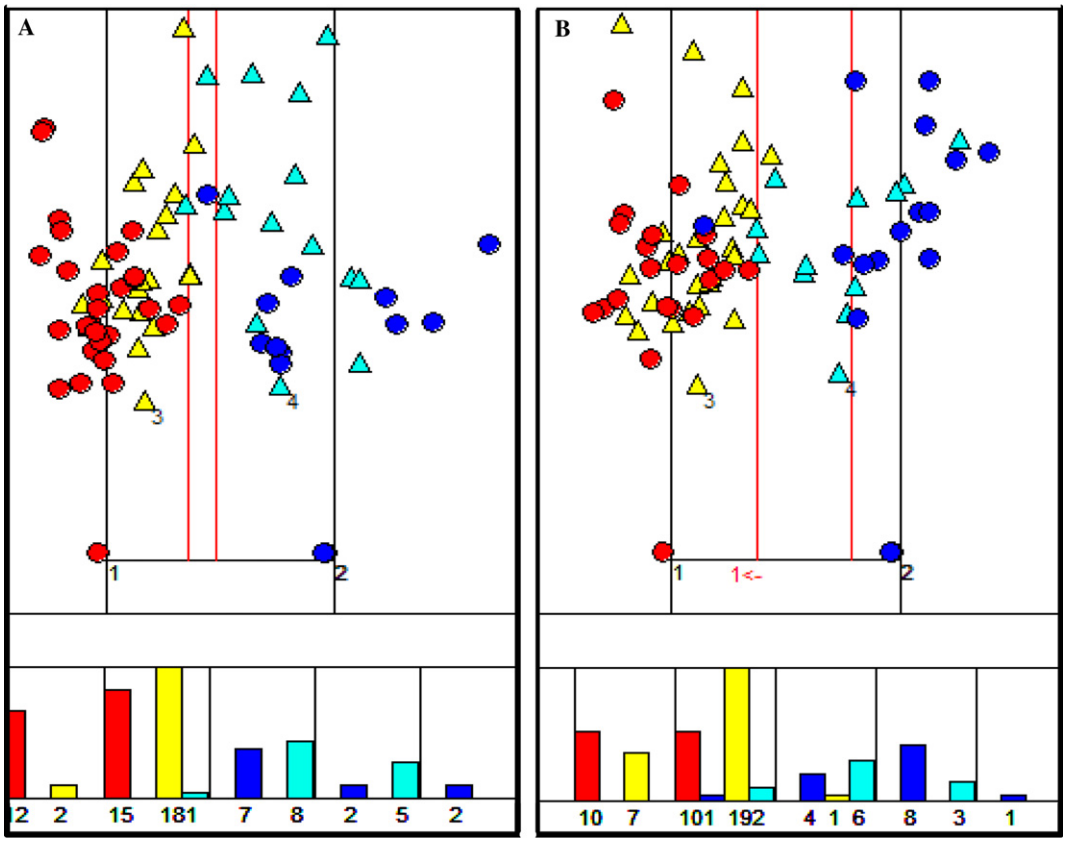

Fig. 2. ALL vs. AML. RDP maps, from 7129 dimensions; L_2 norm. (A) 38 TS, 34 VS (10/315 pairs). Reference points: 1, 2. (B) 34 TS, 38 VS (1 misclassified) Reference points: 1, 26.

mapping display is from 10 original features). The red and blue *disks* represent samples from class 1 and class 2, respectively. We depict the distances of an arbitrary data point $\mathbf{X}_m$ to $\mathbf{R}_1$ and $\mathbf{R}_2$ as D_{1m} and D_{2m}, respectively. When the mapping separates the data perfectly, the vertical red lines pass through the nearest patterns that separate the two classes in the RDP, their separation distance indicated as the "*Support Vector Margin*." In Fig. 1, all vertical lines between the margins represent the hyperplanes (a "hyper slab") in L-space that separate exactly the two classes. We also display the histograms formed when the data points of the RDP are binned and projected onto the reference axis. When no perfect separation of the two classes is possible, the vertical red line pairs are placed to minimize visual "misclassifications" in the data.

For any two classes, a particularly informative subset of the total number N_{pair} of possible reference pairs is the subset $\{\mathbf{R}_1 \equiv \mathbf{X}_i(1),\ \mathbf{R}_2 \equiv \mathbf{X}_j(2)\},\ i = 1,\ldots,N_1 + 1,\ j = 1,\ldots,N_2 + 1$, where $\mathbf{X}_i(1)$ denotes sample i from class 1, $\mathbf{X}_j(2)$ sample j from class 2. This is the subset of all $(N_1 + 1)$ $(N_2 + 1)$ pattern pairs (including the class centroids $\mathbf{C}_1$, $\mathbf{C}_2$) whose members belong to different classes. By cycling through this subset, we can explore the dataset in detail by designating any or all of these pairs as new reference points ($\mathbf{R}_1$, $\mathbf{R}_2$), and mapping the remaining patterns. In fact, this process transcends the conventional notion of classifiers that rely on using the class centroids. Thus, we may identify better separating hyperplanes than the one based on the $\mathbf{C}_1$–$\mathbf{C}_2$ axis. As a matter of fact, more frequently than not, there are lines in the original L-space that lead to better class discriminators than the one passing through the class centroids (such as produced by LDA).

3. What can one do with the RDP mapping—examples

We demonstrate various features of the RDP mapping on publicly available gene microarray data, and on mass spectra from proteomics. Both types of data are characterized by a small, O(10–100) number of patterns in the classes, and very large, O(1000–10,000) numbers of attributes. We focus on supervised pattern recognition, i.e., when class labels are known. We are in the process of developing an unsupervised version ("clustering") of the RDP mapping. *Direct* classification of the mapped patterns in the RDP will be discussed in another publication.

3.1. Direct mapping from a high-dimensional feature space

The first example is the well-studied two-class leukemia set, acute myeloid leukemia, AML (47 samples) vs. acute lymphoblastic leukemia, ALL (25 samples), based on microarray expression profiles [14]. The cDNA microarrays contain 7129 genes. The creators of the dataset partitioned it into training (38 samples, 27 AML + 11 ALL) and validation sets (34 samples, 20 AML + 14 ALL). As shown in Fig. 2, two of the many gene pairs, (2300, 4847) and (4211, 4847), give no errors for the training set (TS; red and blue disks for class 1 and 2, respectively) and 1 misclassification for the validation set (VS; yellow and turquoise triangles for class 1 and 2, respectively) [43]. However, as a **direct** mapping (without feature reduction) from 7129 dimensions to the RDP demonstrates in Fig. 2, this is an "easy" dataset to classify: identical classification results are obtainable, without using sophisticated classifiers or feature selection, with the simplest possible classifier, the Euclidean-distance-based Nearest Mean Classifier. In fact, 10 of the possible 315 reference pairs classify the TS perfectly with respect to the TS margin, and when the reference pair consists of the two class centroids, there is only one error with respect to the VS margin. When TS and VS are swapped, the best mapping from 7129-space to the RDP, with the reference pair the class centroids of the VS, produces a single misclassification in the new TS (original VS) and no error in the new VS (second panel). (This single misclassified sample is the same as found prior to the TS–VS swap.)

The next two datasets (nos. 2 and 3) were downloaded from the NIH/FDA Clinical Proteomics Program Databank (http://clinicalproteomics.steem.com). The second is a prostate and the third an ovarian cancer mass spectroscopy dataset. Both contain 15,154 "features" (mass/charge, M/Z values). With these datasets we demonstrate several useful features of the RDP mapping.

3.2. Consequences of different distance metrics and feature space reduction

Currently, we have implemented the following distance metrics in the original high-dimensional space:

$$\mathscr{M}_2 = \|\mathbf{X}_j - \mathbf{X}_k\|_2 = \sum_n (X_{jn} - X_{kn})^2$$

(Euclidean distance (L_2 norm)),

$$\mathscr{M}_1 = \|\mathbf{X}_j - \mathbf{X}_k\|_1 = \sum_n |X_{jn} - X_{kn}|$$

(Cityblock distance (L_1 norm)),

$$\mathscr{M}_\infty = \|\mathbf{X}_j - \mathbf{X}_k\|_\infty = \max_n |X_{jn} - X_{kn}|$$

(Max (L_∞) norm),

$$\mathscr{M}[\mathbf{W}(c)] = (\mathbf{X}_j - \mathbf{X}_k)^{\mathrm{t}}\mathbf{W}(c)^{-1}(\mathbf{X}_j - \mathbf{X}_k)$$

(Anderson–Bahadur (AB) distance).

$\mathbf{W}(c)$ is the pooled covariance matrix, $\mathbf{W}(c) = cp_1\mathbf{W}_1 + (1-c)p_2\mathbf{W}_2$, with $\mathbf{W}_k$ the covariance matrix of class k, $k = 1,2$, p_1,p_2, are the prior class probabilities, $p_1 + p_2 = 1$, and c is a parameter, $0 \leqslant c \leqslant 2$. $\mathbf{W}(1.0)$ gives the Mahalanobis distance, used in LDA. When $\mathbf{R}_1$, $\mathbf{R}_2$ are the two class centroids, we can optimize c to equalize the misclassification probabilities P_1, P_2 for the two classes, or, for a given P_1, minimize P_2 and vice versa [42].

Several other distance measures are possible, e.g., based on any other Minkowski metric, on the Bhattacharya distance, or on some similarity measure, such as correlation converted into a distance, or the L_2 norm extended by the kernel approach. Below, we shall demonstrate on the same datasets the outcomes of selecting the L_2, L_1, L_∞ norms, or the AB or Mahalanobis distance (when possible).

Any mapping from higher to lower dimensions necessarily removes details. Whether this is beneficial or not depends on the data. In particular, the RDP mapping eliminates $L-2$ of the L original feature space dimensions. For microarray or spectral data in particular, many of the "feature" dimensions are experimental "noise" and/or are strongly correlated, hence not independent. Eliminating these is generally not detrimental to visualization or eventual classification.

To illustrate the importance and consequences of feature space reduction, we used the prostate cancer dataset ("JNCI 7-3-02"), with 42 class 1 and 42 class 2 samples in the TS and 21 and 27 samples in the VS.

In Fig. 3, we show mapping (L_2 norm) from the original 15,154-dimensional feature space to the RDP. This produced eight misclassifications in the TS and nine in the VS (Fig. 3A). After feature space reduction, two different 5-dimensional feature sets, with M/Z value positions 7-12-17-37-53 and 9-22-26-43-54, had no misclassification error for either TS or VS, when using LDA. The RDP maps for these two 5-dimensional classifiers show how the separability of the two classes improves over the original 15,154-dimensional dataset. Equally as importantly, the maps aid us in selecting one of these seemingly "perfect" classifiers over the other. Judging from the RDP maps, classifier "7-12-17-37-53" is more likely to generalize better than classifier "9-22-26-43-54," since, when using the L_2 norm, "9-22-26-43-54" had seven misclassifications in the TS and two in the VS (Fig. 3C), whereas "7-12-17-37-53" produced eight reference pairs with zero misclassification for the TS, three of which (one displayed in Fig. 3B) were also without misclassification in the VS. If the distance measure is the Anderson–Bahadur distance, then there are more reference pairs with perfect

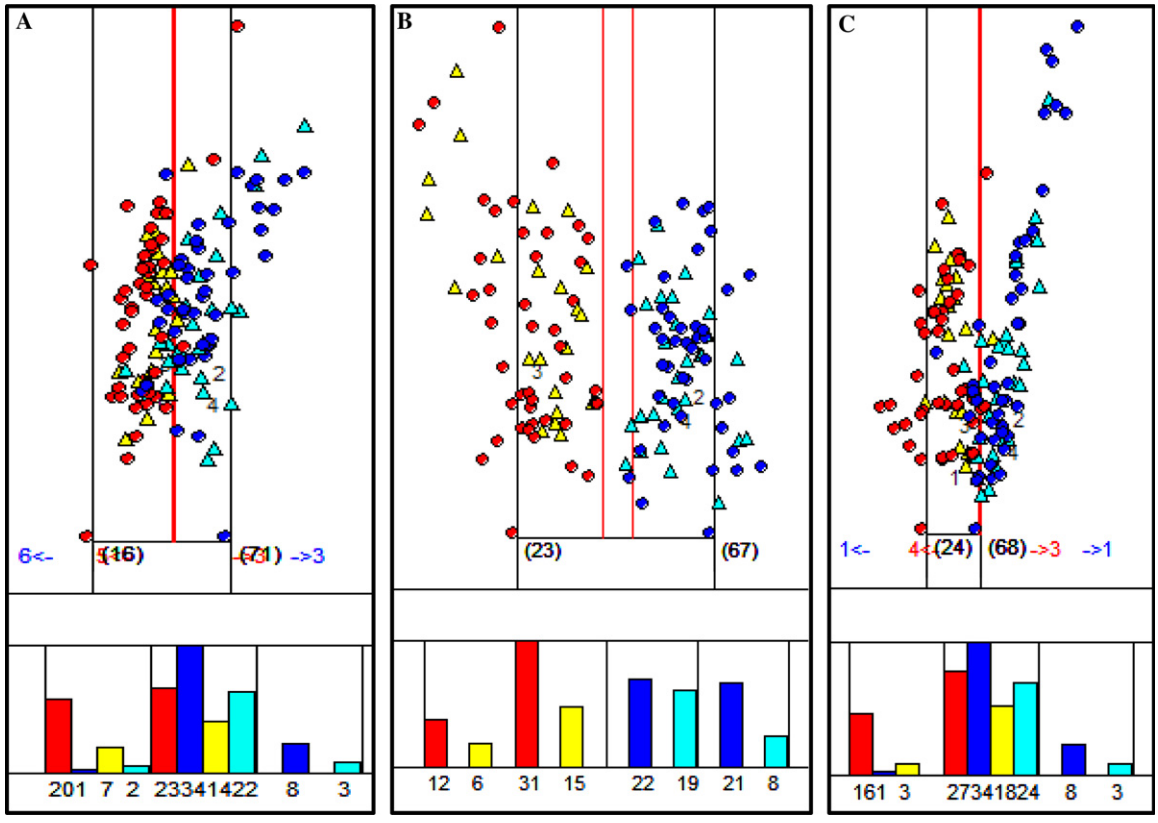

Fig. 3. Prostate cancer (JNCI-7-3-02). RDP mapping (L_2 norm). (A) From 15,154 dimensions: 8 TS + 9 VS errors. From 5 dimensions (in which there are two solutions with 0 TS + 0 VS errors): (B) Solution 1: Features (M/Z values): 7-12-17-37-53; 0 TS and 0 VS errors. (C) Solution 2: Features (M/Z values): 9-22-26-43-54; 7 TS and 2 VS errors.

TS + VS classification on RDP mapping. Superior class separation in the RDP is to be expected with both 5-dimensional feature sets, since an LDA-wrapper-based feature reduction from the original high-dimensional feature space was used.

The superiority of the "7-12-17-37-53" classifier is indeed preserved: 726 of the 1849 reference pairs give perfect results (Fig. 4 displays the pair with the largest TS margin), whereas only 394 of the 1849 pairs are perfect for classifier "9-22-26-43-54" (Fig. 5).

In the second column of Tables 1 and 2 we display the fraction of total reference pairs giving zero errors for both TS and VS, using the five different distance measures implemented (Mahalanobis ($c = 1.0$), c-optimized Anderson–Bahadur, L_2, L_1, and L_∞), for the "7-12-17-37-53" and "9-22-26-43-54" classifiers, respectively. In these tables we also show the best reference pairs, the rank of the reference pair comprising the two class centroids, and TS + VS misclassification errors with respect to the TS margins. The reference pairs are ranked in increasing order of misclassification errors. For equal TS misclassification errors, the pairs are first ranked in the order of decreasing VS errors, then according to decreasing TS margins. (For breaking ties, other ranking options currently implemented are the absolute $\mathbf{R}_1$–$\mathbf{R}_2$ distance, and the absolute and relative *spreads*. The absolute spread is the sum of the average of the deviations of the S-coordinates of class 1 from μ and the average of the deviations of the S-coordinates of class 2 from μ, where $\mu = (m_1 - m_2)/2$ and m_1, m_2 are the S-coordinates of the two margin samples. The relative spread is the absolute spread divided by the $\mathbf{R}_1$–$\mathbf{R}_2$ distance.)

An additional example is provided by the ovarian cancer dataset ("6-19-02"), which we randomly partitioned into TSs and VSs. The TSs contained 61 samples from both cancer and healthy classes, the VSs 30 cancer + 101 healthy samples. Mapping directly (L_2 norm) to the RDP from the original, 15,154-dimensional space (not shown) misclassified 3 TS + 8 VS samples with respect to the best reference pair, $\{1, 76\}$. (Note that when the reference points are the class centroids, the misclassifications increase to 12 TS + 15 VS samples.)

Using our wrapper-based feature selection method [43], only three features (2193, 2241, and 2349) were required to produce perfect separation for both TS and VS. When mapping from this 3-dimensional feature space into the RDP, perfect separation is retained with respect to the reference point pair $\{56, 112\}$. In fact, 54 other reference pairs out of the total possible 3844 pairs achieve zero misclassification (L_2 norm). In Table 3 we collect the mapping results for all five distance measures implemented. The best mapping is obtained with the AB distance, $c = 2.0$, with 3352 of 3834 possible reference pairs producing perfect TS classification. The best reference pair is $\{52, 102\}$. With the Mahalanobis distance

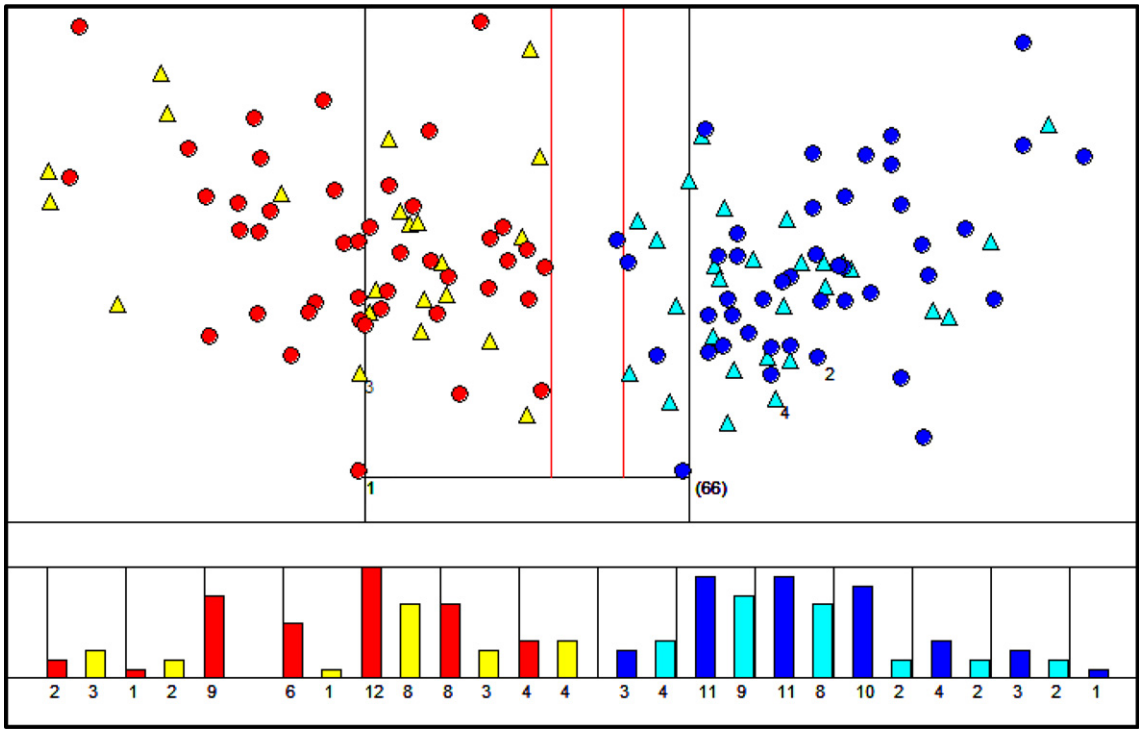

Fig. 4. Prostate cancer (JNCI-7-3-02). RDP mapping (Anderson–Bahadur distance, $c = 1.44$). From 5 dimensions (7-12-17-37-53); Best of the 726 reference pairs (1849 total) with 0 TS and VS errors.

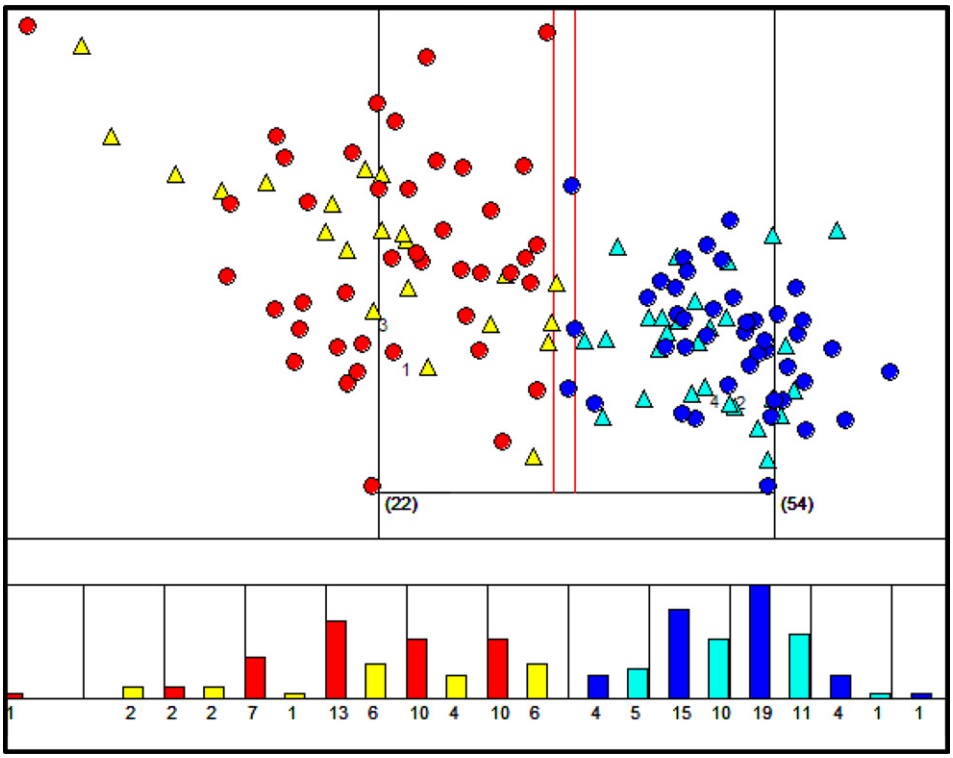

Fig. 5. Prostate cancer (JNCI-7-3-02). RDP Mapping (Anderson–Bahadur distance, $c = 0.9$). From 5 dimensions (9-22-26-43-54); Best of the 394 reference pairs (1849 total) with 0 TS and VS errors.

Table 1
Prostate cancer vs. healthy: 5 *M*/*Z* values (7, 12, 17, 37, 53)

Distance metric	Fraction of reference pairs error-free	R_1	R_2	Rank of reference pair (R_1, R_2)	TS margin-based errors (TS + VS)
Mahalanobis	627/1849	1	69	2	0 + 0
		1	2	48	0 + 0
Anderson–Bahadur	726/1849	1	66	1	0 + 0
$C = 1.44$		1	2	27	0 + 0
L_2	8/1849	23	67	1	0 + 0
		1	2	33	2 + 3
L_1	0/1849	22	70	1	2 + 1
		1	2	51	6 + 1
L_∞	1/1849	43	48	1	0 + 0
		1	2	284	6 + 6

Comparison of the mapping results for five distance metrics (AB, Mahalanobis, L_2, L_1, L_∞). The fraction of error-free reference pairs, the best pair and the rank of the class centroid pair, and the TS and VS errors with respect to the TS margin are shown.

($c = 1.0$), 3121 of 3844 reference pairs classified without error. It is evident from the RDP maps that the three features found are sufficient to produce classifiers with good generalization potential.

4. Detection of potential outliers

Potential outliers are readily identifiable visually via the RDP mapping. In particular, if the two reference patterns are the class centroids, any new point "well outside" the boundaries encompassing the current data points is a possible outlier. Furthermore, if certain data points appear to be outliers with respect to many reference pairs (corresponding to viewing the dataset from different perspectives), then the likelihood that these points are truly outliers is increased because of this consensus.

We demonstrate the outlier detection capability of the RDP mapping on the ovarian cancer dataset ("6-19-02") discussed above. As apparent in Fig. 6, with

Table 2
Prostate cancer vs. healthy: 5 M/Z values (9, 22, 26, 43, 54)

Distance metric	Fraction of reference pairs error-free	R_1	R_2	Rank of reference pair (R_1, R_2)	TS margin-based errors (TS + VS)
Mahalanobis	382/1849	37	52	5	0 + 0
		1	2	18	0 + 0
Anderson–Bahadur	394/1849	37	52	5	0 + 0
$C = 0.90$		1	2	17	0 + 0
L_2	0/1849	24	68	1	7 + 2
		1	2	505	25 + 13
L_1	0/1849	14	49	1	8 + 4
		1	2	287	19 + 7
L_∞	0/1849	18	59	1	13 + 8
		1	2	783	26 + 16

Comparison of the mapping results for five distance metrics (AB, Mahalanobis, L_2, L_1, L_∞). The fraction of error-free reference pairs, the best pair and the rank of the class centroid pair, and the TS and VS errors with respect to the TS margin are shown.

Table 3
Ovarian cancer vs. healthy: 3 M/Z values (2193, 2241, 2349)

Distance metric	Fraction of reference pairs error-free	R_1	R_2	Rank of reference pair (R_1, R_2)	TS margin-based errors (TS + VS)
Mahalanobis	3121/3844	49	78	1	0 + 0
		1	2	525	0 + 0
Anderson–Bahadur	3352/3844	52	102	1	0 + 0
$C = 2.0$		1	2	1332	0 + 0
L_2	55/3844	56	112	1	0 + 0
		1	2	1357	7 + 18
L_1	28/3844	12	84	1	0 + 0
		1	2	838	5 + 5
L_∞	1/3844	63	70	1	0 + 1
		1	2	1530	9 + 17

Comparison of the mapping results for five distance metrics (AB, Mahalanobis, L_2, L_1, L_∞). The fraction of error-free reference pairs, the best pair and the rank of the class centroid pair, and the TS and VS errors with respect to the TS margin are shown.

Mahalanobis distances, there are three likely outliers in the TS (samples 11, 15, and 22), and three in the VS (samples 185, 191, and 216). The likelihood of these six being genuine outliers is confirmed by viewing the RDP mappings with respect to the 10 best ranking reference pairs: For all 10, the triplet of $\{11, 15, 22\}$ appeared to be outliers, and for 9 of these 10, so did the triplet $\{185, 191, 216\}$.

We gain additional confidence in identifying outliers when the same patterns appear as outliers for other distance metrics. For the ovarian cancer dataset this seems to be the case for $\{11, 15, 22\}$, both in the L_2 and L_1 norms (not shown), suggesting that they are truly outliers. Furthermore, RDP maps with respect to different reference pairs still indicate $\{11, 15, 22\}$ as likely outliers.

5. Distributional properties of training and validation sets

An important advantage of the RDP representation is that one can visually assess whether an "independent" validation set derives from the same probability distribution as the training set. This is essential to correctly assess the generalization power of any classifier based on the training set. We make use of histograms constructed for each class (as shown in the figures). By ordering the patterns of the two classes according to their distances to their own class means, these histograms show the class membership distributions (and the overlap between the two classes). A statistically more precise approach is to carry out the standard, one-dimensional Kolmogorov–Smirnov (KS) test (or some variant, e.g., Kuiper's test) [44] to assess the two distributions' similarity to each other, when mapped onto the reference axis.

The two-sample, two-sided KS test [44] compares the empirical cumulative distribution functions of the two groups to be assessed for distributional origin. It is a nonparametric, distribution-free test of the probability that the two groups are **not** from the same distribution. The null hypothesis, H_0, is that the two groups of samples come from the same distribution. The Δ-statistic is a measure of the distributional differences. It is the maximal absolute deviation between the two cumulative distributions. $\Delta = 1.0$ means that the origins of the two groups are unequivocally different; Δ small ($\sim$0 for sufficiently large sample sizes) suggests a common

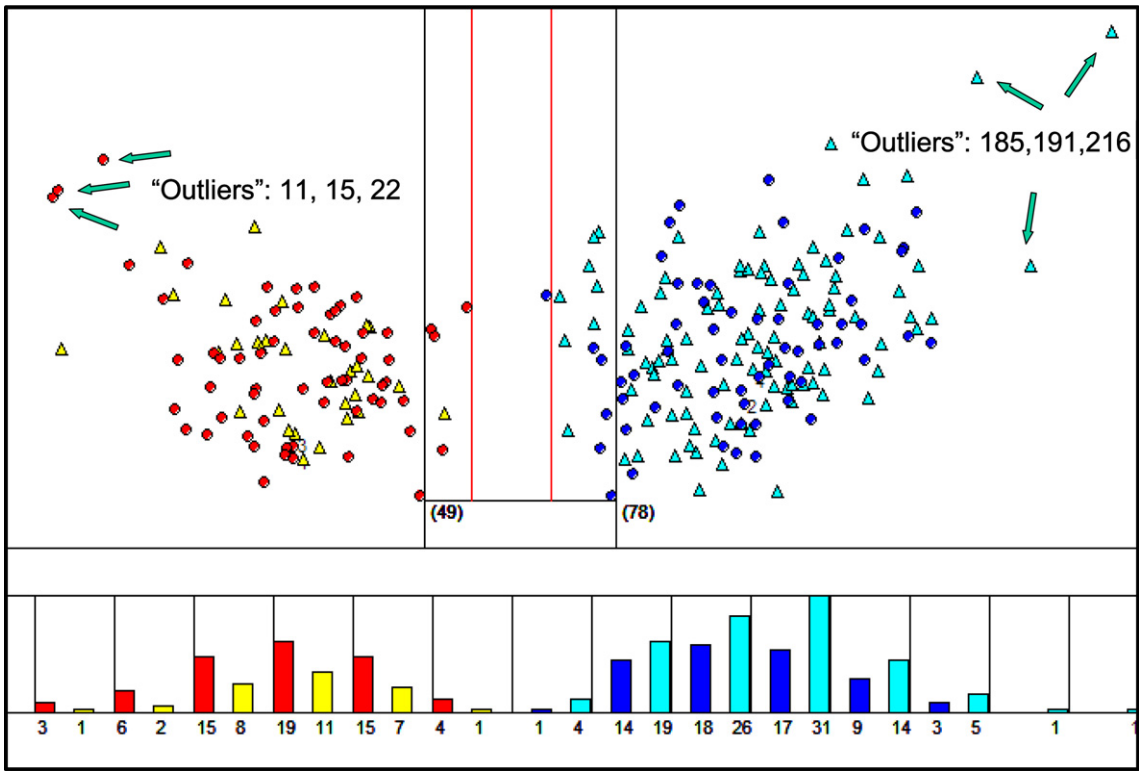

Fig. 6. Ovarian Cancer (6-19-02). RDP mapping (Mahalanobis distance, $c = 1.0$). RDP mapping from 3 dimensions; M/Z values: 2193, 2241, 2349; Best of the 3121 reference pairs (3844 total) with 0 TS + 0 VS errors; Potential outliers (identified by arrows): 11, 15, 22, 185, 191, 216.

distributional origin. We can assign statistical significance to the computed Δ values (e.g., at a significance level $\alpha = 0.05$) and compare it to sample-size-dependent, tabulated or easily computable critical values, $\Delta(N_1, N_2)$.

For a 2-class problem, let "**1**" denote the TS for class 1, "**2**" for class 2, "**3**" the VS for class 1, and "**4**" for class 2. Then if the classes separate perfectly for the TS, we expect the KS Δ-statistics to give $\Delta(\mathbf{1–2}) = 1.0$, and if the same applies to the VS, $\Delta(\mathbf{3–4}) = 1.0$. The more interesting tests are $\Delta(\mathbf{1–3})$ and $\Delta(\mathbf{2–4})$. The former assesses the similarity of the distributional origins of the class 1 TS and VS, the latter that of class 2. If the TS and VS derive from the same distribution, both should be less than $\Delta_\alpha(N_1, N_2)$. The mixed tests (e.g., $\Delta(\mathbf{1–4})$ and $\Delta(\mathbf{2–3})$) can be used for confirming consistency. They ought to match $\Delta(\mathbf{1–2})$ and $\Delta(\mathbf{3–4})$, respectively.

For the ovarian dataset, we display in Table 4 the Δ values obtained for the RDP mapping when projecting from the original 15,154-dimensional feature space (L_2 norm). In the six columns we display the six possible pairwise mappings of the four datasets. We also show the number of samples used in the mappings and the corresponding critical $\Delta(N_1, N_2)$ values for significance levels $\alpha = 0.05$ and 0.01. The six rows contain the six possible $\Delta(\mathbf{I–J})$ values, $\mathbf{I < J = 1, \ldots, 4}$. The values shown in bold match mapping pairs with their corresponding Δ values.

Table 4
Kolmogorov–Smirnov Δ statistics for the ovarian cancer vs. healthy

RDP Map	1 vs. 3	2 vs. 4	1 vs. 2	3 vs. 4	1 vs. 4	2 vs. 3
$N_1 + N_2$	61 + 30	61 + 101	61 + 61	30 + 101	61 + 101	61 + 30
$\Delta_{0.05}$	*0.303*	*0.221*	*0.246*	*0.283*	*0.221*	*0.303*
$\Delta_{0.01}$	*0.364*	*0.264*	*0.295*	*0.339*	*0.264*	*0.364*
$\Delta(1–3)$	**0.207**	0.255	0.142	0.177	0.095	0.112
$\Delta(2–4)$	0.120	**0.190**	0.100	0.175	0.118	0.113
$\Delta(1–2)$	0.115	0.311	**0.951**	0.934	0.967	0.885
$\Delta(3–4)$	0.168	0.383	0.957	**0.990**	0.967	0.937
$\Delta(1–4)$	0.137	0.273	0.915	0.925	**0.984**	0.908
$\Delta(2–3)$	0.240	0.452	0.967	1.000	0.967	**0.934**

RDP mappings (L_2 norm) from the original 15154 M/Z values. Δ values for all six possible mapping combinations are shown.

Columns **1** vs. **3** and **2** vs. **4** of the table reflect the RDP-mapping-based maximum separation achievable for the TS and VS of classes 1 and 2, respectively. Despite this maximization of the separations, in the case of both the **1** vs. **3** and the **2** vs. **4** mapping, the corresponding Δs are less than the critical Δ for both $\alpha = 0.05$ and 0.01, suggesting that the TS and VS for both classes derive from the **same** distribution. Of course, the relative dispositions of the four datasets in the RDP display depend on which two are used for the mapping. Thus, for the **1** vs. **3** mapping, the other five Δs are also less than the critical Δ for both $\alpha = 0.05$ and 0.01. In contrast, for the **2** vs. **4** mapping, all the other five Δs are greater than the critical Δs.

Inspection of the two rows corresponding to Δ(**1–3**) and Δ(**2–4**) indicates that, independently of which of the six RDPs we consider, the TSs and VSs for both classes appear to belong to the same distribution. Similarly, the next four rows, Δ(**1–2**), Δ(**3–4**), Δ(**1–4**) and Δ(**2–3**), for dataset pairs comprising different classes, show unequivocally that these classes belong to different distributions (their Δ values are near to 1).

These conclusions carry over without any *qualitative* change when the RDP mapping is from the 3-dimensional reduced feature space. The *quantitative* difference (not shown) is that Δ(**1–3**) and Δ(**2-4**) are smaller and Δ(**1–2**), Δ(**3–4**), Δ(**1–4**), and Δ(**2–3**) are larger than their counterparts for the mapping from 15,154 dimensions.

6. Assessment of dataset sparsity

We cannot ignore the influences of dataset sparsity [43] on the TS and VS distributions. We demonstrate this on a microarray dataset [16], created for discriminating between the four classes of small, round blue-cell tumor (SRBCT). There are 2308 genes expressed on the chips. Of the four classes, we select two that have very few samples per class, Burkitt lymphoma (BL; TS: 8, VS: 3) and neuroblastoma (NB; TS: 12, VS: 6).

In Table 5, we repeat the RDP mapping and the KS test computations we have reported in Table 4 (there, for a dataset with larger number of samples). We now use the much sparser BL vs. NB class pair. The first important observation is that because of the limited number of samples, the critical Δ values are much larger than those computed for Table 4. Furthermore, most of the Δ(**I–J**) values are 1.0, suggesting either that no statistical decision can be made, or that because of dataset sparsity, TS and VS separate perfectly ($\Delta = 1.0$) for both classes, instead of showing a common distributional origin. In fact, of the 36 Δ(**I–J**) values, only eight are less than the maximum possible 1.0. These results (counterintuitive because we know what they ought to be) are simply illustrated and resolved in Fig. 7, where we display the RDP maps based on TS(BL) vs. VS(BL) (Fig. 7A) and on TS(NB) vs. VS(NB) (Fig. 7B). For instance, the Δ(**1–2**) value of $0.292 < 0.875$ ($\Delta_{0.05}$) would suggest that TS(BL) and TS(NB) come from the same distribution. Inspection of Fig. 7A reveals that this is simply an accident of that particular projection. We reach similar conclusions for the mapping shown in Fig. 7B. In contrast, whereas VS(BL) and VS(NB) separate perfectly visually (Fig. 7A), they separate only marginally at the $\Delta_{0.05}$ level, and no statistical decision is possible at the $\Delta_{0.01}$ level.

Thus, the RDP mapping, together with the final projection onto some reference axis and consequent execution of the 1-dimensional KS test, provides a visualizable yet more quantitative measure of the adequacy of the sample sizes for a statistically meaningful analysis. A 2-dimensional version of the KS test, or some variant, would be an even more appropriate statistical test to achieve this goal. We are in the process of implementing such a test.

7. Discussion

Note that by projecting exemplars belonging to a different class C_k, $k \neq 1$ or 2, onto the RDP patterns, one may immediately display their distance relation to the original two classes. For every new pattern, we need to calculate only two additional distances in the original

Table 5
Kolmogorov–Smirnov Δ statistics for SRBCT: Burkitt lymphoma vs. neuroblastoma

RDP Map	1 vs. 3	2 vs. 4	1 vs. 2	3 vs. 4	1 vs. 4	2 vs. 3
$N_1 + N_2$	8 + 3	12 + 6	8 + 12	3 + 6	8 + 6	12 + 3
$\Delta_{0.05}$	*0.875*	*0.667*	*0.625*	*1.000*	*0.667*	*0.833*
$\Delta_{0.01}$	*1.000*	*0.833*	*0.708*	[a]	*0.833*	*1.000*
Δ(1–3)	**1.000**	1.000	0.542	1.000	1.000	1.000
Δ(2–4)	1.000	**1.000**	0.833	0.667	1.000	1.000
Δ(1–2)	0.292	0.458	**1.000**	1.000	1.000	1.000
Δ(3–4)	1.000	0.833	1.000	**1.000**	1.000	1.000
Δ(1–4)	1.000	1.000	1.000	1.000	**1.000**	0.333
Δ(2–3)	1.000	1.000	1.000	1.000	0.500	**1.000**

RDP mappings (L_2 norm) from 2308 M/Z values. Δ values for all six possible mapping combinations are shown.
[a] Cannot reject H_0 regardless of observed Δ value.

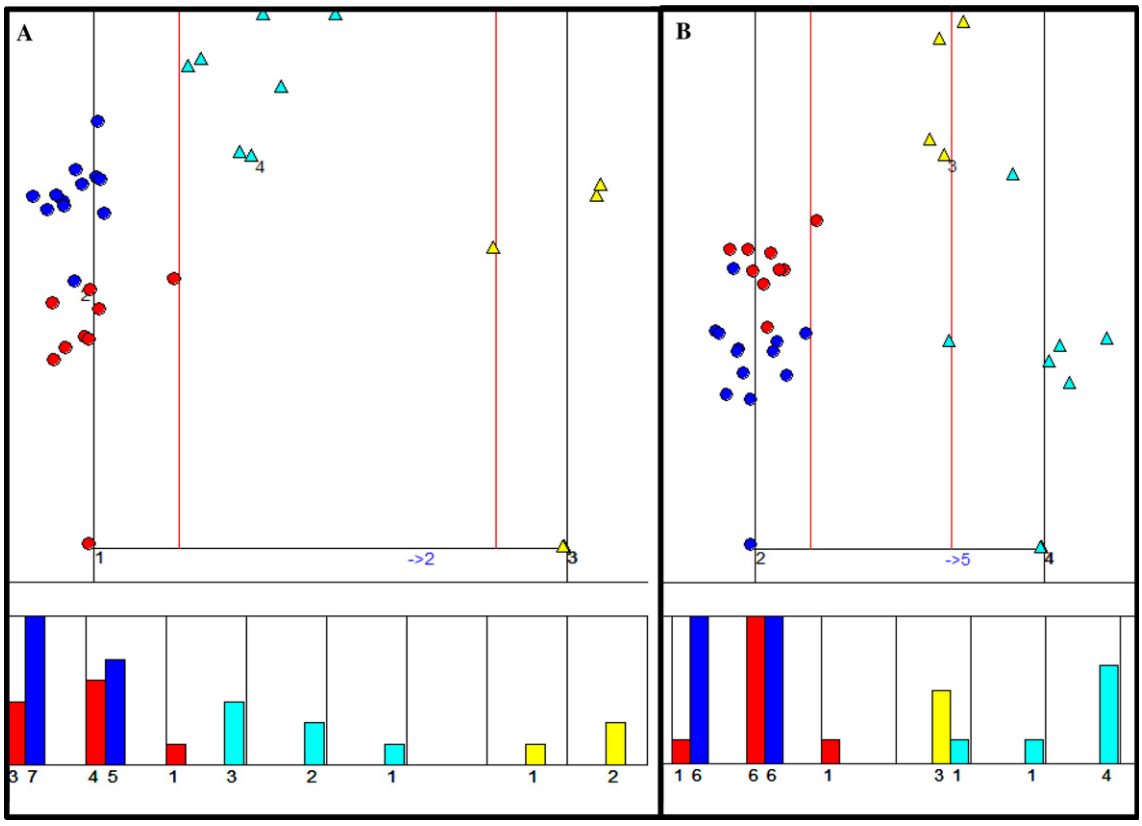

Fig. 7. Burkitt Lymphoma (BL) vs. Neuroblastoma (NB). RDP mappings (L_2 norm) from 2308 *M*/*Z* values. (A) TS(BL) vs. VS(BL). (B) TS(NB) vs. VS(NB).

L-space of attributes, making online classification possible.

Recall that with the RDP mapping, we can preserve only the distances to the two reference points (e.g., the two class centroids). Furthermore, in general, their RDP displays will not correspond to the best possible classification achievable in the original *L*-space. That is because any projection into a lower dimensional space leads to a loss of flexibility (degrees of freedom). Consequently, the number of misclassifications in the current RDP will likely be an upper bound to that in any dimension greater than two. Furthermore, we have the flexibility to choose any pair of other reference points (e.g., two putative or identified support vectors) to recompute a new RDP mapping and possibly improve classification accuracy. The examples presented clearly support this.

What can we do when the dimensionality is high, and the dataset size is large? The simplest approach is to first carry out some clustering, requesting O(100) clusters and use the cluster centroids in place of the individual samples. If class labels are available, clustering each class separately, and using the cluster centroids as representatives of the classes, will serve the same purpose.

An alternative approach is to pick randomly any two reference points (one from each class), and first compute only the distances to these two. By saving these distances, the distance matrix can be built up sequentially. Mapping only those points that have their distance **ratio** $R_m = D_{1m}/D_{2m}$ $(m \neq 1,2)$ within a pre-specified tolerance, e.g., $0.9 \leqslant R_m \leqslant 1.1$, will create an RDP map of only those points near the class boundary. This "on-line" version of the RDP mapping can be repeated, either for a fixed number of reference pairs, or exhaustively.

8. Extensions to higher dimensions and other generalizations

We can readily extend the mapping of multidimensional points into higher than two-dimensional spaces, without losing relative distance information. In particular, consider any two RDP reference points $\mathbf{R}_1$ and $\mathbf{R}_2$, and add to them any point already in the RDP, say $\mathbf{R}_3$. By construction, the original distances D_{km}, $m \neq k = 1,2,3$ are preserved. In the original *L*-space, an arbitrary point **P** has distances D_{kP} to the $\mathbf{R}_k$. The distances D_{kP} are preserved, but generally only in 3-space. The three coordinates (S_P, T_P, V_P) of any **P** in the 3-dimensional relative distance volume can be readily calculated analytically. We can then easily visualize the data with any 3-dimensional display software.

If there are three or more classes present, the three class centroids are the natural first choices for the $\mathbf{R}_k$s, although any three points can be selected. For a two-class problem, a reasonable choice for the triplet $\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$ is the pair of reference points that gave the best class separation in the RDP, and any one from either class that provides the best separation in 3 dimensions.

Such stepwise extension to higher (>3) dimensions is immediate, although the analytical expressions become increasingly more complex. Furthermore, anything beyond 3-space is unlikely to be useful for visual display. Of course, mapping from these progressively higher (e.g., 4 or 5) dimensions to 3 dimensions is possible and would distort less the other patterns' mutual distance relationships than mapping directly from the original L-dimensional feature space.

With N $(= N_1 + N_2)$ L-dimensional 2-class patterns, there are at least $N_P = N_1 N_2$ possible reference axes in the RDP if the reference pairs consist of two patterns not belonging to the same class. Thus, selecting any two reference axes $\mathbf{A}_j$ and $\mathbf{A}_k$, and using the corresponding $S_j(\mathbf{X}_m)$ and $S_k(\mathbf{X}_m)$ values as the two coordinates for sample $\mathbf{X}_m$, one can display, and classify all patterns in this (generally oblique) coordinate system. The choice of the pair $\mathbf{A}_j$ and $\mathbf{A}_k$ determines how good will be the classifier. Extension to higher dimensions is immediate. We shall report on these generalizations elsewhere.

If one found several, equally accurate classifiers in an M-dimensional $(2 < M < L)$ feature space via some feature reduction method, ties could be broken by comparing misclassification errors when going from M-space to the RDP. We gave an example, for the prostate cancer dataset, for which two error-free classifiers were found in a reduced, 5-dimensional feature space. However, mapping from these to the RDP clearly showed that one of these feature sets produces a classifier that generalizes considerably better than the other.

9. Conclusion

The RDP representation is a similarity-based mapping for the visualization of high-dimensional patterns and their relative relationships in which original distances between points are exactly preserved with respect to arbitrary reference pattern pairs in a special two-dimensional coordinate system, easily extendible to three or higher dimensions. As only a single calculation of a distance matrix is required, this method is computationally efficient, an essential characteristic for exploratory data analysis. The data visualization afforded by this representation permits a rapid assessment of class pattern distributions, an important consideration for any eventual classification strategy. We can explore any dataset in detail by identifying the subset of reference pairs whose members belong to different classes, cycling through this subset, and for each pair, mapping the remaining patterns. These multiple viewpoints help identify and confirm possible outliers. Furthermore, we can eliminate the curse of dimensionality and facilitate the construction of nonlinear decision boundaries between classes. We demonstrated the effectiveness of this method for the analysis of several complex biomedical datasets. Because of its efficiency, effectiveness, and versatility, one may use the RDP representation as an initial, data mining exploration that precedes eventual classification. We recommend using the RDP mapping in an interactive, feedback fashion during all stages of high-dimensional biomedical data analysis.

References

[1] Lean CL, Somorjai RL, Smith ICP, Russell P, Mountford CE. Accurate diagnosis and prognosis of human cancers by proton MRS and a three stage classification strategy. Annu Rep NMR Spectrosc 2002;48:71–111.

[2] Mountford CE, Somorjai RL, Malycha P, et al. Diagnosis and prognosis of breast cancer by magnetic resonance spectroscopy of fine-needle aspirates analysed using a statistical classification strategy. Br J Surg 2001;88:1234–40.

[3] Petrich W, Staib A, Otto M, Somorjai R. Correlation between the state of health of blood donors and the corresponding mid-infrared spectra of the serum. Vibr Spectrosc 2002;28:117–29.

[4] Somorjai RL, Dolenko B, Nikulin A, et al. Distinguishing normal allografts from biopsy-proven rejections: application of a three-stage classification strategy to urine MR and IR spectra. Vibr Spectrosc 2002;28:97–102.

[5] Staib A, Dolenko B, Fink DJ, et al. Disease pattern recognition testing for rheumatoid arthritis using infrared spectra of human sera. Clin Chim Acta 2001;308:79–89.

[6] Adam B-L, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. Cancer Res 2002;62:3609–14.

[7] Li J, Zhang Zh, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches of serum biomarkers to detect breast cancer. Clin Chem 2002;48:1296–304.

[8] Petricoin III EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.

[9] Vlahou A, Schellhammer PF, Mendrinos S, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158:1491–502.

[10] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745–50.

[11] Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–40.

[12] Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262–7.

[13] DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–60.

[14] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[15] Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc—Appl Case Studies 2002;97:88–99.
[16] Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7(6):1–10.
[17] Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415:436–42.
[18] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001;98:15149–54.
[19] Su AI, Welsh JB, Sapinoso LM, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001;61:7388–93.
[20] van't Veer L, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[21] Yeang Ch-H, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types. Bioinformatics 2001;17:S316–22.
[22] Siedlecki W, Siedlecka K, Sklansky J. An overview of mapping techniques for exploratory pattern analysis. Pattern Recogn 1988;21(5):411–29.
[23] Siedlecki W, Siedlecka K, Sklansky J. Experiments on mapping techniques for exploratory pattern analysis. Pattern Recogn 1988;21(5):431–8.
[24] Jackson JE. A user's guide to principal components. New York: Wiley; 1991.
[25] Borg I, Groenen P. Modern multi-dimensional scaling. Berlin: Springer Verlag; 1997.
[26] De Backer S, Naud A, Scheunders P. Non-linear dimensionality reduction techniques for unsupervised feature extraction. Pattern Recogn Lett 1998;19:711–20.
[27] Kim S-S, Kwon S, Cook D. Interactive visualization of hierarchical clusters using MDS and MST. Metrika 2000;51:39–51.
[28] Klock H, Buhmann JM. Data visualization by multidimensional scaling: a deterministic annealing approach. Pattern Recogn 2000;33:651–69.
[29] Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.
[30] Niemann H, Weiss J. A fast-converging algorithm for nonlinear mapping of high-dimensional data to a plane. IEEE Trans Comput 1979;C-28:142–7.
[31] Sammon JW. A non-linear mapping for data structure analysis. IEEE Trans Comput 1969;C-18:401–9.
[32] Haese K. Self-organizing feature maps with self-adjusting learning parameters. IEEE Trans Neural Networks 1998;9:1270–8.
[33] Muruzábal J, Munoz A. On the visualization of outliers via self-organizing maps. J Comput Graph Stat 1997;6:355–82.
[34] Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic dfferentiation. Proc Natl Acad Sci USA 1999;96:2907–12.
[35] Törönen P, Kolehmainen M, Wong G, Castrén E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999;451:142–6.
[36] Vesanto J. SOM-based data visualization methods. Intell Data Anal 1999;3:111–26.
[37] Ultsch A. 1993. Self-organizing neural networks for visualisation and classification. In: Opitz O, Lausen B, Klar R, editors. 16th annual conf. information and classification: concepts, methods and applications.
[38] Yin H. Multiclass classification of SRBCTs. ViSOM—A novel method for multivariate data projection and structure visualisation. IEEE Trans Neural Networks 2002;13:237–43.
[39] Friedman J, Tukey JW. A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput 1974;C-23:881–9.
[40] Huber PJ. Projection pursuit. Ann Stat 1985;13:435–75.
[41] Lee RCT, Slagle JR, Blum H. A triangulation method for the sequential mapping of points from *N*-space to two space. IEEE Trans Comput 1977;C-27:288–92.
[42] Anderson TW, Bahadur RR. Classification into two multivariate normal distributions with different covariance matrices. Ann Math Stat 1962;33:420–31.
[43] Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19(12):1484–91.
[44] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C++. The art of scientific computing. second ed. Cambridge: Cambridge University Press; 2002. pp. 628–633.