BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data

Christoph Bartenhagen[1*], Hans-Ulrich Klein[1], Christian Ruckert[1], Xiaoyi Jiang[2], Martin Dugas[1]

## Abstract

**Background:** Visualization of DNA microarray data in two or three dimensional spaces is an important exploratory analysis step in order to detect quality issues or to generate new hypotheses. Principal Component Analysis (PCA) is a widely used linear method to define the mapping between the high-dimensional data and its low-dimensional representation. During the last decade, many new nonlinear methods for dimension reduction have been proposed, but it is still unclear how well these methods capture the underlying structure of microarray gene expression data. In this study, we assessed the performance of the PCA approach and of six nonlinear dimension reduction methods, namely Kernel PCA, Locally Linear Embedding, Isomap, Diffusion Maps, Laplacian Eigenmaps and Maximum Variance Unfolding, in terms of visualization of microarray data.

**Results:** A systematic benchmark, consisting of Support Vector Machine classification, cluster validation and noise evaluations was applied to ten microarray and several simulated datasets. Significant differences between PCA and most of the nonlinear methods were observed in two and three dimensional target spaces. With an increasing number of dimensions and an increasing number of differentially expressed genes, all methods showed similar performance. PCA and Diffusion Maps responded less sensitive to noise than the other nonlinear methods.

**Conclusions:** Locally Linear Embedding and Isomap showed a superior performance on all datasets. In very low-dimensional representations and with few differentially expressed genes, these two methods preserve more of the underlying structure of the data than PCA, and thus are favorable alternatives for the visualization of microarray data.

## Background

DNA microarrays allow the measurement of transcript abundances for thousands of genes in parallel. Applications in quality assessment and interpretation of such high dimensional data by clustering [1,2] and visualization [3,4] make use of algorithms that reduce its dimension. Two and three dimensional visualizations are often a good way to get a first impression of properties or the quality of a dataset or of special patterns within the data by showing clusters such as diseased and healthy patients, revealing outliers, a high level of noise or to generate hypotheses for further experimentation [5-8]. In general, there are two different approaches to reduce a datasets' dimension. Feature selection methods [9-11] compute a ranking on all genes by means of some given score and pick a gene subset based on this ranking. Feature extraction methods define a mapping between the high-dimensional input space and a low-dimensional target space of a given dimension. Both methods are used in machine learning concepts. Most classification algorithms use many or all features in a complex (nonlinear) manner whereas approaches like [12,13] are based on the relative expression of only two or three genes to overcome the "black box" character of the other classifiers. So they allow an easy traceability of the genes leading to the classification result. On the other hand, applications like the visualization of high-dimensional data may profit from extracting information from all features. This results in feature extraction methods usually being more suited for low-dimensional representations of the whole data. In the

* Correspondence: Christoph.Bartenhagen@ukmuenster.de
[1]Department of Medical Informatics and Biomathematics, University of Münster, Domagkstraße 9, 48149 Münster, Germany
Full list of author information is available at the end of the article

following, we refer to feature extraction methods when speaking of dimension reduction techniques.

Considering visualization, these kind of mappings are often unsupervised, because they don't use further information of the data like class labels and allow an unbiased view of the structure within the data. Supervised methods are more applicable to improve classification or regression procedures, assuming that less non-differential or noisy features are reduced after the mapping.

All features, that are related to special properties of the data or a separation into classes or clusters, often lie in a subspace of a lower (intrinsic) dimension within the original data. A 'good' dimension reduction technique should preserve most of these features and generate data with similar characteristics like the high-dimensional original. For example, classifications should work at least as well on the low-dimensional representation and clusters within the reduced data should also be found, preferably more distinct. Principal Component Analysis (PCA) is a widely used unsupervised method to define this mapping from high-to low-dimensional space. Availability of large datasets with high-dimensional data, especially in biological research (e.g. microarrays), led to many new approaches in the last years.

Other studies, that deal with the assessment of dimension reduction techniques, either compare them against the background of classification [14-18], and hence mainly discuss supervised methods like Partial Least Squares [19,20], Sliced Inverse Regression [21] or other Regression models [22], or come from Computer Vision and deal with text, image, video or artificial data like the Swiss Roll [23-28]. This study instead, focuses on microarray data and its two and three dimensional visualization. We compare PCA to six recent unsupervised methods to find out if and under which conditions they are able to outperform PCA. In the following sections, we describe a benchmark, consisting of classifications and cluster validations, to compare the visualization performance of seven dimension reduction techniques on ten real microarray and several simulated datasets. After some technical details in the methods section, we present and discuss all results, based on one representative dataset. Further details of the other nine datasets are available in the supplement.

## Methods

### Dimension Reduction

Seven unsupervised dimension reduction techniques were compared within this study: Principal Component Analysis (PCA), Kernel PCA (KPCA), Isomap (IM), Maximum Variance Unfolding (MVU), Diffusion Maps (DM), Locally Linear Embedding (LLE) and Laplacian Eigenmaps (LEM). These dimension reduction techniques can be divided into two groups: linear and nonlinear methods. While PCA belongs to the former, due to a linear combination of the input data, the other six methods were designed with respect to data lying on or near a nonlinear submanifold in the higher dimensional input space and perform a nonlinear mapping.

Given an input space $\mathbb{R}^D$ and target space $\mathbb{R}^d$ (with $d << D$) let $X \in \mathbb{R}^{N \times D}$ be an input dataset of $N$ samples and $D$ features (gene expression values) and $Y \in \mathbb{R}^{N \times d}$ its low-dimensional representation. A dimension reduction technique is a mapping $\Phi$: $\mathbb{R}^D \to \mathbb{R}^d$ that optimizes a cost function $\in : R^d \to \mathbb{R}$ on the target space. This problem can often be reduced to an eigenvalue problem, whose eigenvectors will define the embedding $Y$.

### Principal Component Analysis

Principal Component Analysis (PCA) [29,30] builds a new coordinate system by selecting those $d$ axes $w_1, \ldots, w_d \in \mathbb{R}^D$, which maximize the variance in the data:

$$w_1 = \arg\max_{\|w\|=1} \operatorname{var}(Xw) = \arg\max_{\|w\|=1} w'Cw,$$

$w_2, \ldots, w_d$ are chosen in the same way, but orthogonal (independent) to each other (here, $C \in \mathbb{R}^{D \times D}$ denotes the covariance matrix of the data $X$). So, the principal components $p_i = Xw_i$ explain most of the variance in the data. Before mapping the data, the samples in $X$ were centered by subtracting their mean. Since PCA only considers the variance among samples, it works best if those features, that are relevant for class labeling, account for a large part of the variance. Sometimes, the first two or three principal components are not sufficient for a good representation of the data [26]. This can lead to a high target dimensionality and prevent a well suited visualization. Furthermore, the covariance matrix grows rapidly for high-dimensional input data. To overcome this issue, we substituted the covariance matrix by the matrix of squared Euclidean distances $D_E = \frac{1}{N} XX' \quad \left( D_E \in \mathbb{R}^{N \times N} \right)$ [14,31].

### Kernel PCA

To make PCA more suitable for nonlinear data, Kernel PCA (KPCA) maps the data into a higher dimensional feature space before applying the the same optimization as PCA. [32,33]. The mapping can be done implicitly by using a kernel function. The Gaussian kernel $K(x_i, x_j) = \exp\left( \frac{-\|x_i - x_j\|^2}{\sigma^2} \right)$ was applied in our study.

### Isomap

Isomap (IM) [27,28], a nonlinear modification of Multidimensional Scaling [34], preserves the global structure of the input data in its low-dimensional representation. This is done by constructing a neighborhood graph $G$, weighted by shortest geodesic distances $D_G \in \mathbb{R}^{N\times N}$ between all $k$ nearest neighbors. This way, Isomap captures paths along a nonlinear manifold instead of the direct Euclidean distance. The embedding into the low-dimensional space is done by selecting $y_1, \ldots, y_d \in \mathbb{R}^N$ such that

$$\epsilon = \| D_G - D_Y \|_{L^2}$$

,is minimized, with $D_Y(i,j) = \|y_i - y_j\|^2$ being the pairwise distance matrix of neighbors $y_i$ , $y_j$ in the target space.

Previous work in [23] addressed problems in visualizing datasets consisting of several well separated clusters. Since Isomap is known to suffer from holes in the underlying manifold [14], it is suggested to modify the method by selecting $\frac{k}{2}$ nearest and $\frac{k}{2}$ farthest neighbors when constructing the graph, instead of the $k$ nearest neighbors. Both, IM and IM(mod), will be discussed in the results section.

### Maximum Variance Unfolding

Similar to Isomap, Maximum Variance Unfolding (MVU) [25,26] preserves the distances among $k$ nearest neighbors by means of a neighborhood graph $G$. But it varies in considering squared Euclidean distances between two neighbored samples, instead of geodesic distances and in maximizing the Euclidean distance between all points $y_i$ , $y_j$ in the target space (to 'unfold' the data) while preserving the distances in the neighborhood graph. This leads to the optimization problem.

$$\max \sum_{ij} \|y_i - y_j\|^2 \text{ subject to } D_G = D_Y$$

Based on the same concept, MVU shares some weaknesses with Isomap like suffering from erroneous connections in the graph.

### Diffusion Maps

Diffusion Maps (DM) [35,36] start with building a graph $G$ as well, but differ in weighting the edges by the Gaussian kernel function: $W(i,j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right)$. With the rows being normalized by $\hat{W} \in \mathbb{R}^{N\times N}$, the weights $\hat{W} \in \mathbb{R}^{N\times N}$ can be seen as a Markov Matrix that defines the probability to move from one sample to another in one time step. The transition probability for $t$ time steps, denoted $\hat{W}^{(t)}$, is given by $\hat{W}^t$. It can be used to control the local connections among neighbored samples. Here, we set it to $t = 1$. Diffusion Maps retain a weighted $L^2$ distance, the 'diffusion distance'

$$D^{(t)}(x_i, x_j) = \sqrt{\sum_{l=1}^{N} \frac{\left(\hat{W}^{(t)}(i,l) - \hat{W}^{(t)}(j,l)\right)^2}{\psi(x_l)}}$$

The term $\Psi(x_i) = \frac{\Sigma_j \hat{W}(i,j)}{\Sigma_{jl} \hat{W}(j,l)}$ leads to stronger weighting of samples from dense areas in the graph. Since the diffusion distance between two points is computed over all possible paths in the graph, Diffusion Maps are more robust to noise.

### Locally Linear Embedding

Unlike Isomap and MVU, Locally Linear Embedding (LLE) [24,37] attempts to preserve local properties of the data. Each sample $x_i$ is represented by a linear combination of its $k$ nearest neighbors:

$x_i = \sum_{j=1}^{k} W(i,j) x_j$. The weights $W \in \mathbb{R}^{N\times N}$ are estimated by minimizing the reconstruction error

$$\sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} W(i,j) x_j \right\|^2$$

subject to $W(i,j) = 0$, if $x_i$ is not a neighbor of $x_j$, and $\sum_{j=1}^{N} W(i,j) = 1$. The last constraint ensures an invariance to translation next to rotation and rescaling. By minimizing

$$\epsilon(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{k} W(i,j) y_j \right\|^2$$

the low-dimensional representation that best preserves the weights in the target space is chosen.

### Laplacian Eigenmaps

As well as LLE, Laplacian Eigenmaps (LEM) [38,39] are a local technique. Similar to Diffusion Maps, this method first constructs a neighborhood graph, weighted

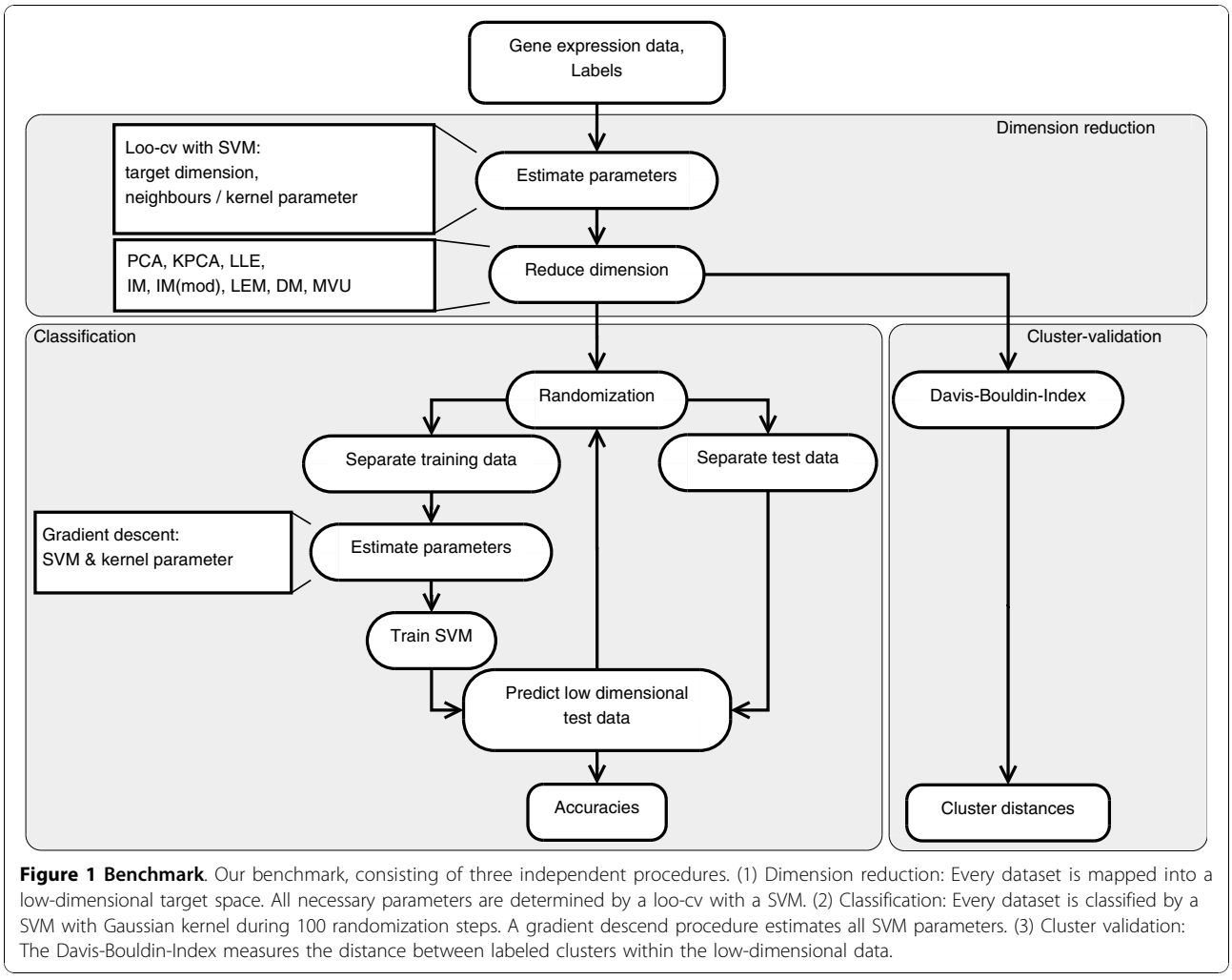


**Figure 1 Benchmark**. Our benchmark, consisting of three independent procedures. (1) Dimension reduction: Every dataset is mapped into a low-dimensional target space. All necessary parameters are determined by a loo-cv with a SVM. (2) Classification: Every dataset is classified by a SVM with Gaussian kernel during 100 randomization steps. A gradient descend procedure estimates all SVM parameters. (3) Cluster validation: The Davis-Bouldin-Index measures the distance between labeled clusters within the low-dimensional data.

with values $W(i, j)$ from the Gaussian kernel function. By minimizing a cost function

$$\epsilon(Y) = \sum_{ij} \left\| y_i - y_j \right\|^2 W(i,j)$$

for neighbored $y_i$, $y_j$ ($W(i, j) = 0$ otherwise), the distances between the low-dimensional representations are minimized and nearby samples $x_i$, $x_j$ are highly weighted, and thus brought closer together. This way, Laplacian Eigenmaps implicitly enforces natural clusters in the data.

## Methods of Assessment

### Benchmark

Our benchmark (Figure 1) is divided into three parts. First, the studied dimension reduction methods were applied to the complete dataset. The low-dimensional datasets were then assessed by two different approaches, namely classification and cluster validation. To evaluate and compare the performance of each method, the classification accuracies of Support Vector Machines [40] (with Gaussian kernel) and the compactness and distance of clusters within the low-dimensional representations were used. In the following, each step of our benchmark is described in detail.

#### *Datasets*

The methods were tested on ten published microarray datasets as well as on simulated data. Each published dataset was divided into two classes according to a binary variable corresponding to the samples' disease status, the presence of certain molecular mutations or other sample characteristics as shown in Table 1. Since microarray data is technically provided with a more or less high level of noise, we reran the benchmark on the microarray datasets combined with normally distributed noise with zero mean and an increasing variance between 0 and 0.1. Before adding noise, all data was scaled to values between 0 and 1 to overcome the varying means and standard deviations of the datasets.

**Table 1 Microarray datasets**

| Dataset | samples | features | class 1(#samples) | class 2(#samples) |
|---|---|---|---|---|
| 1 Wang et al. - Breast cancer [50] | 286 | 22.283 | ER+(209) | ER-(77) |
| 2 Verhaak et al. - Leukemia [51] | 461 | 54.675 | NPM1 pos.(140) | NPM1 neg.(321) |
| 3 Haferlach et al. - Leukemia [52] | 251 | 54.675 | NPM1 pos.(138) | NPM1 neg.(113) |
| 4 Haferlach et al. - Leukemia [52] | 77 | 54.675 | AML with t(8;21)(40) | AML with t(15;17)(37) |
| 5 Golub et al. - Leukemia [53] | 72 | 7.129 | ALL(47) | AML(25) |
| 6 Chiaretti et al. - Leukemia [54] | 22 | 12.625 | CLL stable(8) | CLL progressive(14) |
| 7 Alizadeh et al. - Lymphoma [55] | 38 | 18.432 | Activated B-like DLBCL(17) | GC B-like DLBCL(21) |
| 8 Nutt et al. - High-grade glioma [56] | 50 | 12.625 | Glioblastoma(28) | Anaplastic oligodendroglioma(22) |
| 9 Alon et al. - Colon cancer [57] | 62 | 2.000 | Tumor(42) | Normal(20) |
| 10 Singh et al. - Prostate cancer [58] | 102 | 12.600 | Tumor(52) | Normal(50) |

Summary of all ten microarray gene expression datasets we used for testing the dimension reduction techniques. Here, we focus on the data by Wang et al., which represents best the results of the whole benchmark. Datasets 2-10 are shortly discussed in the supplement to this work. All datasets were separated into two classes according to two characteristics or the diagnosis of a disease.

The simulated data is based on a 50 sample dataset whose 10.000 gene expression values are normally distributed with zero mean and standard deviation one. The covariances of all genes are given by a block diagonal matrix with coefficients $\rho = 0.2$ within and $\rho = 0$ outside the blocks of size $50 \times 50$. To separate the data into two classes, between 10 and 500 genes were randomly chosen to be differentially expressed by adding a constant of 0.6 to the expression values of the first 25 samples. We generated 100 datasets for testing.

In the same manner as for the ten microarray datasets before, normally distributed noise with zero mean and an increasing variance between 0 and 0.2 was added to the simulated data. We repeated the benchmark on 50 of these noisy artificial datasets. The number of differential features was fixed to 300.

***Dimension reduction***

All dimension reduction techniques discussed here have one or two free parameters, that influence the embedding and the target dimension. Their determination was done by minimizing the error rate of a Support Vector Machine (SVM) within a leave-one-out cross-validation (loo-cv) schema: For $N$ samples, the dataset was divided $N$ times into a training and a test set. One sample was excluded for testing while the rest was taken for training. The average over all prediction accuracies gives an estimate of the SVMs' generalization error.

This procedure was repeated for every set of parameters within the following ranges:

Target dimensionality: $2 \le d \le 15$
Neighbors: $4 \le k \le 16$
Gaussian kernel: $1e-1 \le \sigma \le 5e5$

If the same loo-cv accuracies were achieved by using different parameter values for the target dimension, the lowest value was taken for reasons of a most simple representation. The same applies to the neighbor/kernel parameters.

After the loo-cv, the whole dataset was reduced in its dimension in an unsupervised manner, i.e. without consideration of class labels.

***Classification***

The first evidence for the quality of the different dimension reduction methods are the accuracies of a Support Vector Machine with Gaussian kernel.

The data was classified repeatedly during several randomization steps:

We randomly split the dataset a hundred times into a set to train the SVM and a test set for classification, and selected the median accuracy of all runs. Within the training set, a loo-cv was performed to determine the SVM parameters. For reasons of performance, a gradient descent procedure as proposed in [41] was used to minimize the loo-cv error. Every time during randomization, the training set consisted of two thirds of the original data and the test set of the remaining samples. The only constraint was to keep the balance between the number of samples in each class. Since SVMs do not restrict the dimension of the input data, the randomization results of the low-dimensional data can be compared to the high-dimensional original data, to see if more or less significant features got lost after the embedding.

***Cluster validation***

To measure the distances between the labeled clusters, we used the Davis-Bouldin-Index (DB-Index) [42]: Given $M$ clusters $C_i$ ($i = 1, \ldots, M$) and their centers

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x,$$

$$d_i = \frac{1}{|C_i|} \sum_{x \in C_i} \| x - \mu_i \|$$

is the average distance of the samples in cluster $C_i$ to its center. While $R_{ij} = \frac{d_i + d_j}{\|\mu_i - \mu_j\|}$ reports the compactness of clusters $C_i$, $C_j$ related to their distance, the DB-Index

$$DB = \frac{1}{M}\sum_{i=1}^{M} \max\left\{ R_{ij} \mid 1 \le j \le M, i \ne j \right\}$$

averages the worst cases of the clusters' separations. One might expect well separated clusters to have smaller values close to one. In our case, the DB-Index was computed for fixed target space dimensions 2,3,5, and 10.

### Implementation details

The presented benchmark was implemented in Matlab 7.8.0 (R2009a). Furthermore, libsvm (version 2.89) [43] served as Support Vector Machine implementation, in conjunction with Automatic Model Selection for Kernel Methods (Apr 2005) [44]. The Dimensionality Reduction Toolbox (version 0.7 - Nov 2008) [45], Isomap package (Release 1 - Dec 2000) [46], LLE routine [45] and MVU implementation (version 1.3) [47] were used for dimension reduction. Because the Isomap and LLE routines performed best in our benchmark, we converted their Matlab implementations for the statistical programming language R [48]. The R-package 'RDRToolbox', also including a routine to compute the Davis-Bouldin-Index and our microarray gene expression data simulator, can be downloaded from [49] (see also Additional file 1).

## Results and Discussion

The following sections present the results for the Wang et al. Breast Cancer dataset, which represents best the results of the whole procedure. For the sake of simplicity, the visualization example in Figure 2 refers to the Haferlach et al. Leukemia dataset, which consists of fewer samples. Further detailed analysis of all other datasets is available in the supplement (see Additional file 2).

A linear approach like PCA is known to recover the true structure of data lying on or near a linear subspace of the high-dimensional input space. The following results show that the structure of microarray data is often too complex to be captured well in very low dimensional target spaces in a linear manner. Nonlinear methods, in particular LLE and Isomap, preserve more information in the data than the first few principle components of a PCA are able to cover.

### Classification

The results of the randomization procedure are shown in Figure 3. In case of two and three dimensions, PCA performs worst, while all nonlinear methods, except Diffusion Maps, tend to retain the underlying structure of the data better in such low-dimensional target spaces.

Table 2 shows the parameters having the best loo-cv accuracies. The estimated target dimension was higher than ten in most cases. PCA and Kernel PCA result in the highest dimensions (14 and 15), while other methods like Laplacian Eigenmaps, MVU and Isomap worked best with less than ten dimensions. But classifications in two or three target dimensions often yield only slightly different accuracies. The classification accuracies on data with and without dimension reduction were often similar, even in two and three target dimensions.

While all methods perform nearly even in higher dimensions, Isomap, LLE and Laplacian Eigenmaps performed best in two and three dimensions. Only on two of ten datasets (Alizadeh et. al and Singh et. al), PCA performed as well as other nonlinear methods like Isomap in two or three dimensional target spaces (see Supplemental Figures S18/S19, S27/S28). On all ten datasets considered together (see supplement), Diffusion Maps and Laplacian Eigenmaps produce more varying results

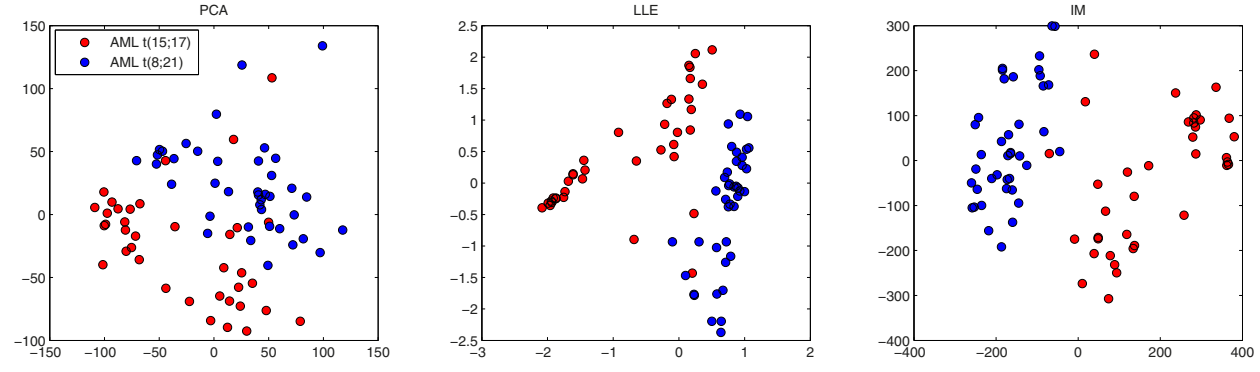


**Figure 2 Visualization**. Two dimensional visualization example of the Haferlach et al. Leukemia dataset. The LLE and Isomap embedding show more distinct clusters than the first two principal components of a PCA.

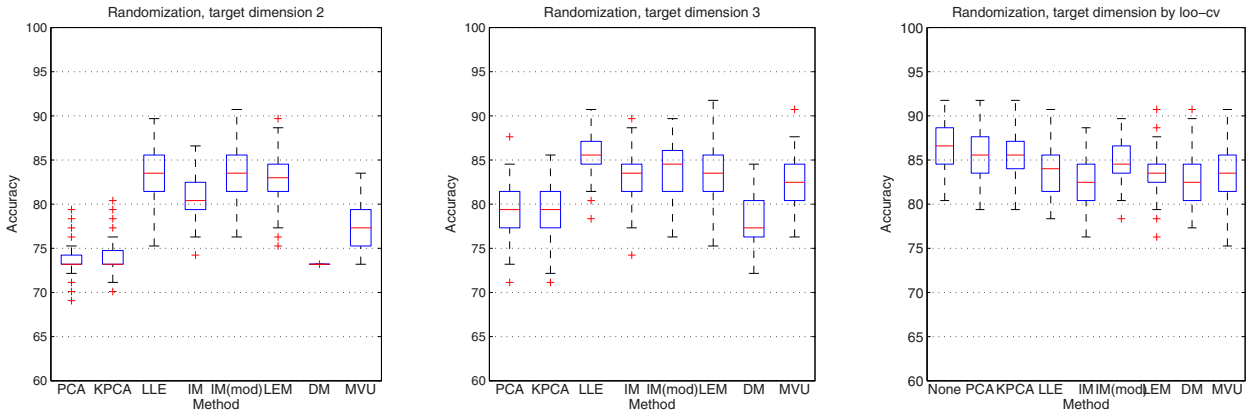


**Figure 3 Randomization accuracies**. Support Vector Machine classification accuracies of the Wang et al. breast cancer dataset. The data was randomized a hundred times for fixed target space dimensions two (left) and three (center) and for dimensions estimated by loo-cv (right). In the last case, the plot also shows the results for the original high-dimensional data without reduction. Especially in two and three dimensions, all nonlinear methods are superior to PCA.

and especially Diffusion Maps are very sensitive to the choice of the kernel parameter (see for example Figure 3, dimension two). But like Kernel PCA, they perform quite similar to PCA in most cases. MVU, which is based on Multidimensional Scaling like Isomap, is comparable to Isomap's good accuracies.

The initial publications on Isomap and MVU [25,27], covering text classification and face recognition, pointed out, that PCA might need higher dimensional target spaces than its nonlinear counterparts to lead to similar results. Since PCA only considers the variance in the data, it works best if those features, which are relevant for the class labeling, account most for the variance. Considering complex microarray data, the first two or three principal components were often not enough to cover the information necessary to sufficiently distinguish different classes within the data. This might prevent a well suited visualization, which is true to the original. LLE, Isomap and MVU, which classified best most of datasets, take advantage of overlapping local neighborhoods to create an image of the global geometry of the data. Although this approach may suffer from "holes" within the data (manifold), it proved more useful for accurate low-dimensional representations.

Well sampled datasets may overcome this issue of sparse data. But the Chiaretti et al. leukemia (22 samples), Alizadeh et al. lymphoma (38 samples) and Nutt et al. high-grade glioma dataset (50 samples) show that even with relatively few samples, a true to the original embedding is possible. The classification accuracies of most of the dimension reduction methods on these datasets (in ≥ 2 target dimensions) are comparable and sometimes even better than the accuracies on the high-

**Table 2 Parameter estimation**

| method | dim | neighbors/$\sigma$ | loo-cv accuracy |
|---|---|---|---|
| PCA | 14 | - | 87.4 |
| KPCA | 15 | 5e5 | 87.1 |
| LLE | 12 | 14 | 88.5 |
| IM | 8 | 10 | 85 |
| IM(mod) | 15 | 4 | 87.4 |
| LEM | 5 | 4 | 85.3 |
| DM | 13 | 5e5 | 84.3 |
| MVU | 5 | 14 | 85 |

All parameters of the dimension reduction techniques for the Wang et al. breast cancer dataset, estimated by leave-one-out cross-validation. PCA has no additional parameter, while KPCA and DM have a kernel parameter $\sigma$ and IM, LEM and MVU take the number of neighbors as argument.

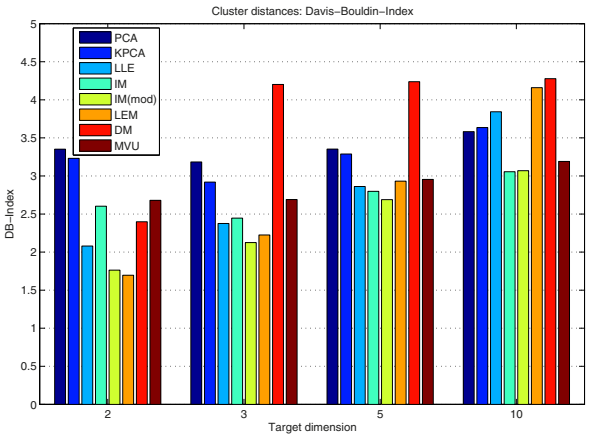


**Figure 4 Cluster validation**. Davis-Bouldin-Indices of the reduced Wang et al. breast cancer dataset for fixed target space dimensions 2,3,5 and 10. In most cases, the nonlinear methods produce more distinct clusters than PCA.

dimensional data (see Supplemental Figures S15, S18, S21).

### Cluster validation

The cluster distances, presented in Figure 4, confirm the above conclusions. In two and three target dimensions, PCA results in worse scores than most nonlinear methods. DM performs the worst for more than two dimensions. With increasing target space dimension all methods converge, while the DB-Index itself increases as well. Although Laplacian Eigenmaps implicitly enforce natural clusters in the data, they show only slight different scores than e.g. LLE, which clusters best on most of the datasets. Just in case of ten target dimensions, LLE's and Laplacian Eigenmaps score remarkably worser on four of our ten datasets, while the other methods, including PCA, hold steady (see Supplemental Figures S9, S12, S21, S24). While Isomap might map well separated clusters to very close points, the slight modification of regarding nearest and farthest neighbors seems to correct this behavior on three datasets (Supplemental Figures S3, S6, S24), but performs similar or (much) worse otherwise (see for example Supplemental Figures S15, S18, S27). MVU scores similar to Isomap, but fails on three other datasets in two dimensional target spaces (Supplemental Figures S6, S15, S18).

Because LLE and Isomap performed best on most of the datasets during classification and cluster validation, Figure 2 compares their two dimensional embedding of the Haferlach et al. Leukemia dataset to the first two principal components of a PCA. All three visualizations clearly show two clusters of AML patients with t(15;17) and t(8;21) respectively. But LLE and Isomap distinguish both classes best, while in the PCA embedding three more t(15;17) samples lie between samples of the other class. Since LLE and Isomap both map more samples correctly, there seems to be more information within the data, that the first two PCA components fail to preserve. On closer inspection, the common three t(15;17) outliers, that are in between or closest to t(8;21) samples in all three visualizations, are always the same samples #44 and #46 #57. Another visualization example of the Alon et al. Colon Cancer dataset with all eight dimension reduction techniques can be seen in Supplemental Figures S1 and S2.

### Noise evaluation

The tests on artificially noised microarray datasets reveal, that PCA, Kernel PCA and Diffusion Maps are most robust on noisy data (Figure 5). But the differences are less strong and the results more variable than for the classification and cluster validation without adding noise. The sensitivity to noise of all methods strongly depends on the given class labels and associated features, and thus leads to varying results between all ten datasets (see supplement). While Diffusion Maps are known to be robust to noise [14], all other nonlinear methods, especially Isomap and its modification, suffer most from unstructured data and lead to strongly varying cluster scores.

### Simulated data

Since LLE and Isomap performed best in the first two tests, the classifications on simulated data refer only to these methods. In all three cases, we fixed a two dimensional target space. Figure 6 shows that the results of the loo-cv on real microarray datasets can be reproduced on simulated data. With only few differential features, LLE and Isomap already capture more of the structure of the data than PCA. It takes more than 150 (of overall 10.000) differential features for PCA to perform nearly even. Furthermore, for less than 200 differential features, the accuracies of PCA are spreading much stronger, while LLE and Isomap give more stable

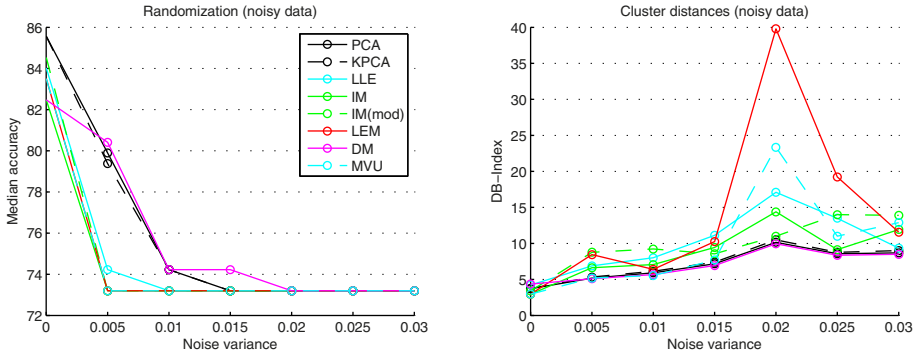


**Figure 5 Noise evaluation**. Randomization accuracies (left) and cluster-indices (right) for the Wang et al. breast cancer dataset combined with normally distributed noise with zero mean and different variances. PCA and DM react most stable on noise. All other methods lead to varying accuracies and cluster scores.

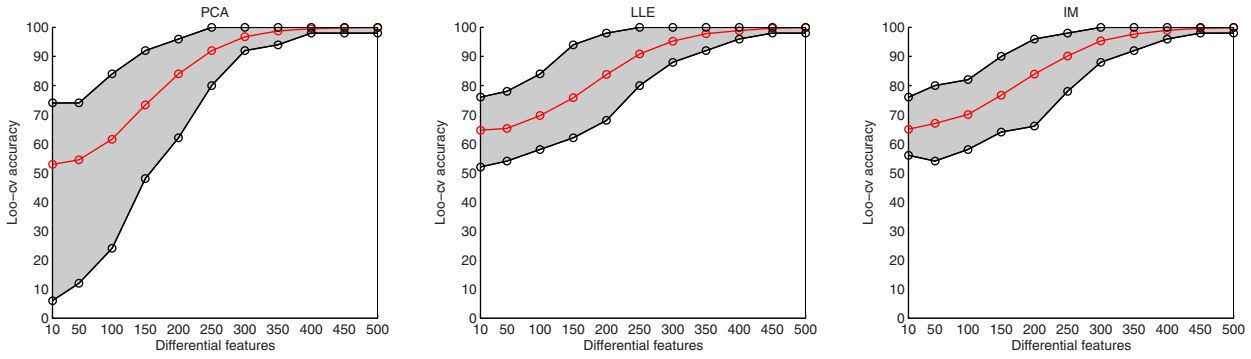


**Figure 6 Simulated data**. Leave-one-out cross-validation accuracies of the simulated datasets with increasing differential features. The red line marks the average accuracy of all 100 generated datasets within the 95% quantile. In two dimensional target spaces, LLE an IM capture more of the underlying structure of the data with much fewer significant features than PCA.

results. The findings for three target dimensions are similar to the two dimensional case and can be seen in the supplement (Supplemental Figure S30).

The benchmark with noisy simulated data, however, confirms the results of the noise evaluation for the ten microarray datasets. Supplemental Figures S31 and S32 show for two and three target dimensions, that PCA performs more robust than LLE and Isomap for both, classification and cluster validation, when noise within the data increases. These conclusions hold true for noisy data with a larger variance, since PCA, LLE and Isomap are invariant to multiplication of the data with a scalar.

### Statistical hypothesis test

In Table 3, we compare all results by applying the Wilcoxon signed-rank test on the accuracies and cluster scores for two dimensional data representations. We tested the null hypothesis, that the median of the differences between PCA and each of the nonlinear methods is equal to zero. This way, we computed the p-values of 14 paired samples. The p-values were not adjusted for multiple testing. Isomap and LLE show the most significant results in accuracy and clustering with p-values 0.0078 and 0.0273 respectively. Diffusion Maps led to results most similar to PCA.

**Table 3 Wilcoxon signed-rank test (p-values)**

| PCA compared to ... | Accuracies(dim 2) | DB-Index(dim 2) |
|---|---|---|
| KPCA | 0.1562 | 0.3223 |
| LLE | 0.0195 | **0.0273** |
| IM | **0.0078** | 0.1055 |
| IM(mod) | 0.0547 | 0.2324 |
| LEM | 0.1953 | 0.3750 |
| DM | 0.5000 | 0.8457 |
| MVU | 0.1953 | 0.7695 |

The p-values of the Wilcoxon signed-rank test based on the null hypothesis, that PCA yields results similar to each other nonlinear method. The median randomization accuracies and cluster scores (both for two target space dimensions) of all ten datasets served as input to the tests.

### Runtime

The computational complexity and memory requirements for all dimension reduction methods except MVU are equal, as shown in Supplemental Table S3. However, we observed differences in runtime between the methods due to different constant factors. Table 4 lists the runtime of all seven methods in seconds for the smallest dataset (Chiaretti et al. leukemia dataset, 22 samples) and the largest dataset (Verhaak et al. leukemia dataset, 461 samples). The target dimensionality was set to two. The embeddings were computed on an AMD Opteron processor with 2 GHz.

The runtime of all nonlinear methods (Kernel PCA, Isomap, LLE, LEM, DM, MVU) depends on the number of samples. Even for relatively large microarray datasets (461 samples in this case), runtimes between 9.4 and 21.9 seconds are acceptable for visualization purposes. Only the solution of a semidefinite program in the MVU algorithm takes two hours. For all methods, the computing time for datasets with more common sample sizes ($\leq$ 50) is less than a second.

## Conclusions

Classifications on high and low-dimensional data showed, that the most significant information within microarray data can be captured quite well in very few

**Table 4 Runtime**

| | PCA | KPCA | LLE | IM | LEM | DM | MVU |
|---|---|---|---|---|---|---|---|
| Chiaretti et al. | 0.09 s | 0.03 s | 0.14 s | 0.04 s | 0.04 s | 0.16 s | 0.25 s |
| Verhaak et al. | 9.4 s | 12.7 s | 21.9 s | 14 s | 15.2 s | 13.2 s | 2 hrs |

Runtime in seconds of all seven dimension reduction techniques on the Chiaretti et al. leukemia dataset ($N$ = 22, $D$ = 12.625, $d$ = 2, $k$ = 10, $\sigma$ = 10) and the Verhaak et al. leukemia dataset ($N$ = 461, $D$ = 54.675, $d$ = 2, $k$ = 10, $\sigma$ = 100). In the latter case, the runtime of MVU is given in hours.

dimensions compared to the thousands of features of the original data.

Our benchmark further revealed significant shortcomings of PCA in two and three dimensional target spaces and brought out two nonlinear methods, that distinguished most from PCA. Especially the performances of Locally Linear Embedding and Isomap in classification and cluster validation make them well suited alternatives to the classic, linear approach of PCA.

## Additional material

**Additional file 1: R-package**. RDRToolbox_1.0.0.tar.gz: A package for nonlinear dimension reduction using the Isomap and LLE algorithm. It also includes a routine for computing the Davis-Bouldin-Index for cluster validation, a plotting tool and a data generator for microarray gene expression data and for the Swiss Roll dataset.

**Additional file 2: Supplement**. Supplement.pdf: Contains information about preprocessing of the data, a discussion of the computational complexity of each dimension reduction method, further classification, cluster validation and noise evaluation results of nine other microarray datasets and further classification and randomization results for simulated datasets.

#### Acknowledgements
This study was supported by COST Action BM0801 Translating genomic and epigenetic studies of MDS and AML (EuGESMA) and by the European Leukemia Network of Excellence (ELN).

#### Author details
[1]Department of Medical Informatics and Biomathematics, University of Münster, Domagkstraße 9, 48149 Münster, Germany. [2]Department of Computer Science, University of Münster, Einsteinstrasse 62, 48149 Münster, Germany.

#### Authors' contributions
CB designed and implemented the benchmark and wrote the paper. HK and CR analyzed results and helped writing the manuscript. MD and XJ contributed to the benchmark design. All authors read and approved the final manuscript.

Received: 19 January 2010 Accepted: 18 November 2010
Published: 18 November 2010

#### References
1. Hibbs MA, Dirksen NC, Li K, Troyanskaya OG: **Visualization methods for statistical analysis of microarray clusters.** *BMC Bioinformatics* 2005, **6**:115.
2. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17(9)**:763-774.
3. Lim IS, Ciechomski PDH, Sarni S, Thalmann D: **Planar arrangement of high-dimensional biomedical data sets by Isomap coordinates.** *In Proceedings of the 16 th IEEE Symposium on Computer-Based Medical Systems* 2003, 50-55.
4. Baek J, McLachlan GJ, Flack LK: **Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010, **32**:1298-1309.
5. Butte A: **The use and analysis of microarray data.** *Nature Reviews Drug Discovery* 2002, **1(12)**:951-960.
6. Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G, Stephanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space.** *Genome research* 2002, **12(7)**:1112-1120.
7. Mramor M, Leban G, Demsar J, Zupan B: **Visualization-based cancer microarray data classification analysis.** *Bioinformatics (Oxford, England)* 2007, **23(16)**:2147-2154.
8. Dawson K, Rodriguez RL, Malyj W: **Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm.** *BMC Bioinformatics* 2005, **6**:195.
9. Umpai TJ, Aitken S: **Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes.** *BMC Bioinformatics* 2005, **6**:148.
10. Li T, Zhang C, Ogihara M: **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.** *Bioinformatics* 2004, **20(15)**:2429-2437.
11. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S: **RankGene: identification of diagnostic genes based on expression data.** *Bioinformatics* 2003, **19(12)**:1578-1579.
12. Geman D, d'Avignon C, Naiman DQ, Winslow RL: **Classifying gene expression profiles from pairwise mRNA comparisons.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3.**
13. Lin X, Afsari B, Marchionni L, Cope L, Parmigiani G, Naiman D, Geman D: **The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations.** *BMC Bioinformatics* 2009, **10**:256.
14. Van der Maaten LJP, Postma EO, van den Herik HJ: **Dimensionality reduction: a comparative review.** *Tech. rep., MICC, Maastricht University* 2008.
15. Chao S, Lihui C: **Feature dimension reduction for microarray data analysis using locally linear embedding.** *In APBC* 2004, 211-217.
16. Cho SB, Won HH: **Machine learning in DNA microarray analysis for cancer classification.** *APBC '03: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003* Australian Computer Society, Inc; 2003, 189-198.
17. Liu CCC, Hu J, Kalakrishnan M, Huang H, Zhou XJJ: **Integrative disease classification based on cross-platform microarray data.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:25.
18. Pochet N, De Smet F, Suykens JA, De Moor BL: **Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction.** *Bioinformatics* 2004, 3185-3195.
19. Nguyen DV, Rocke DM: **Tumor classification by partial least squares using microarray gene expression data.** *Bioinformatics* 2002, **18**:39-50.
20. Boulesteix AL: **PLS dimension reduction for classification with microarray data.** *Statistical Applications in Genetics and Molecular Biology* 2009, **3**:33.
21. Dai JJ, Lieu L, Rocke D: **Dimension reduction for classification with gene expression microarray data.** *Statistical applications in genetics and molecular biology* 2006, 5.
22. Antoniadis A, Lambert-Lacroix S, Leblanc F: **Effective dimension reduction methods for tumor classification using gene expression data.** *Bioinformatics* 2003, **19(5)**:563-570.
23. Vlachos M, Domeniconi C, Gunopulos D, Kollios G, Koudas N: **Non-linear dimensionality reduction techniques for classification and visualization.** *in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2002, 645-651.
24. Roweis ST, Saul LK: **Nonlinear dimensionality reduction by Locally Linear Embedding.** *Science* 2000, **290(5500)**:2323-2326.
25. Weinberger KQ, Saul LK: **Unsupervised learning of image manifolds by semidefinite programming.** *International Journal of Computer Vision* 2006, **70**:77-90.
26. Weinberger KQ, Saul LK: **An introduction to nonlinear dimensionality reduction by maximum variance unfolding.** *AAAI'06: proceedings of the 21st national conference on Artificial intelligence* AAAI Press; 2006, 1683-1686.
27. Tenenbaum JB, de Silva V, Langford JC: **A global geometric framework for nonlinear dimensionality reduction.** *Science* 2000, **290(5500)**:2319-2323.
28. Silva VD, Tenenbaum JB: **Global versus local methods in nonlinear dimensionality reduction.** *Advances in Neural Information Processing Systems 15* MIT Press; 2003, 705-712.
29. Hotelling H: **Analysis of a complex of statistical variables into principal components.** *Journal of Educational Psychology* 1933, **24**:417-441,498-520.
30. Jolliffe IT: **Principal Component Analysis.** *Springer* , 2 2002.
31. Chatfield C, Collins AJ: **Introduction to multivariate analysis.** *Chapman and Hall* 1980.
32. Schölkopf B, Smola A, Müller KR: **Nonlinear component analysis as a kernel eigenvalue problem.** *Neural Computation* 1998, **10(5)**:1299-1319.

33. Schölkopf B, Smola A, Müller KR: **Kernel principal component analysis.** *Advances in kernel methods: support vector learning* 1999, 327-352.
34. Cox TF, Cox MAA, Raton B: **Multidimensional Scaling.** *Technometrics* 2003, **45(2)**:182.
35. Nadler B, Lafon S, Coifman RR, Kevrekidis IG: **Diffusion maps, spectral clustering and reaction coordinates of dynamical systems.** *Applied and Computational Harmonic Analysis* 2006, **21**:113-127.
36. Lafon S, Lee AB: **Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization.** *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2006, **28(9)**:1393-1403.
37. Saul LK, Roweis ST: **Think globally, fit locally: unsupervised learning of low dimensional manifolds.** *Journal of Machine Learning Research* 2003, **4**:119-155.
38. Belkin M, Niyogi P: **Laplacian Eigenmaps for dimensionality reduction and data representation.** *Neural Comp* 2003, **15(6)**:1373-1396.
39. Belkin M, Niyogi P: **Laplacian Eigenmaps and spectral techniques for embedding and clustering.** *Advances in Neural Information Processing Systems 14* 2001, **14**:585-591.
40. Cristianini N, Shawe-Taylor J: **An introduction to Support Vector Machines and other kernel-based learning methods.** Cambridge University Press;, 1 2000.
41. Chapelle O, Vapnik V, Bousquet O, Mukherjee S: **Choosing multiple parameters for Support Vector Machines.** *Machine Learning* 2002, **46**:131-159.
42. Xu R, Wunsch D: **Clustering.** Wiley-IEEE Press;, illustrated 2008.
43. Chang CC, Lin CJ: **LIBSVM, a library for support vector machines.** 2001 [http://www.csie.ntu.edu.tw/~cjlin/libsvm], [last accessed at 29th of Oct 2010].
44. Chapelle O: **Automatic model selection for kernel methods.** *http://olivier.chapelle.cc/ams/*, [last accessed at 29th of Oct 2010].
45. van der Maaten LJP: **Matlab toolbox for dimensionality reduction.** [http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html], [last accessed at 29th of Oct 2010].
46. Tenenbaum JB: **Matlab Isomap package.** [http://isomap.stanford.edu/], [last accessed at 29th of Oct 2010].
47. Weinberger KQ: **Maximum Variance Unfolding.** [http://www.cse.wustl.edu/~kilian/code/code.html], [last accessed at 29th of Oct 2010].
48. **The R Project for statistical computing.** [http://www.r-project.org/], [last accessed at 29th of Oct 2010].
49. **RDRToolbox - A package for nonlinear dimension reduction with Isomap and LLE.** [http://www.bioconductor.org/help/bioc-views/release/bioc/html/RDRToolbox.html], [last accessed at 29th of Oct 2010].
50. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, van Gelder MEM, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *The Lancet* 2005, **365(9460)**:671-679.
51. Verhaak R, Wouters B, Erpelinck C, Abbas S, Beverloo H, Lugthart S, Löwenberg B, Delwel R, Valk P: **Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling.** *Haematologica* 2009, **94**:131-134.
52. Klein HU, Ruckert C, Kohlmann A, Bullinger L, Thiede C, Haferlach T, Dugas M: **Quantitative comparison of microarray experiments with published leukemia related gene expression signatures.** *BMC Bioinformatics* 2009, **10**:422.
53. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
54. Del Giudice I, Chiaretti S, Tavolaro S, De Propris MS, Maggio R, Mancini F, Peragine N, Santangelo S, Marinelli M, Mauro FR, Guarini A, Foa R: **Spontaneous regression of chronic lymphocytic leukemia: clinical and biologic features of 9 cases.** *Blood* 2009, **114(3)**:638-646.
55. Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan ea W C: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403(6769)**:503-511.
56. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN: **Gene expression-based classification of malignant gliomas correlates better with survival than histological classification.** *Cancer Research* 2003, **63(7)**:1602-1607.
57. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96(12)**:6745-6750.
58. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1(2)**:203-209.

doi:10.1186/1471-2105-11-567
**Cite this article as:** Bartenhagen *et al.*: Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics* 2010 **11**:567.