BIOINFORMATICS

Vol. 20 no. 6 2004, pages 874–880
DOI: 10.1093/bioinformatics/btg496

Approximate geodesic distances reveal biologically relevant structures in microarray data

Jens Nilsson[1,], Thoas Fioretos[2], Mattias Höglund[2] and Magnus Fontes[1]*

[1]Centre for Mathematical Sciences, Lund University, Box 118, SE-221 00 Lund, Sweden and [2]Department of Clinical Genetics, Lund University Hospital, SE-221 85 Lund, Sweden

Received on February 19, 2003; revised on September 12, 2003; accepted on September 15, 2003
Advance Access publication January 29, 2004

ABSTRACT

Motivation: Genome-wide gene expression measurements, as currently determined by the microarray technology, can be represented mathematically as points in a high-dimensional gene expression space. Genes interact with each other in regulatory networks, restricting the cellular gene expression profiles to a certain manifold, or surface, in gene expression space. To obtain knowledge about this manifold, various dimensionality reduction methods and distance metrics are used. For data points distributed on curved manifolds, a sensible distance measure would be the geodesic distance along the manifold. In this work, we examine whether an approximate geodesic distance measure captures biological similarities better than the traditionally used Euclidean distance.

Results: We computed approximate geodesic distances, determined by the Isomap algorithm, for one set of lymphoma and one set of lung cancer microarray samples. Compared with the ordinary Euclidean distance metric, this distance measure produced more instructive, biologically relevant, visualizations when applying multidimensional scaling. This suggests the Isomap algorithm as a promising tool for the interpretation of microarray data. Furthermore, the results demonstrate the benefit and importance of taking nonlinearities in gene expression data into account.

Contact: jensn@maths.lth.se

INTRODUCTION

The study of gene expression data has been greatly facilitated by the development of the microarray technology. High density oligonucleotide arrays (Lockhart *et al*., 1996) and cDNA microarrays (Schena *et al*., 1995a, b) measure the expression of thousands of genes simultaneously. Comparing the transcription profiles of different types of tissue specimens permits the identification of genes that best distinguish the samples. When samples correspond to different pathological states of the same tissue, or subtypes of the same malignancy, transcription profiling holds promise as a method for classifying cancers from a molecular rather than from a morphological perspective (Ramaswamy *et al*., 2001; Pollack *et al*., 2002; Khan *et al*., 2002). Furthermore, complex biological processes, such as the onset of the cell cycle (Iyer *et al*., 1999) or cellular responses elicited by various growth factors (Fambrough *et al*., 1999), are now open for a detailed analysis by the study of dense time series.

A main problem in microarray data analysis is how to extract the central features of the vast amount of information generated. Mathematically, the expression profile of a sample can be represented as a point in a gene expression space with coordinates given by its expression levels. Put in another way, the location of a cell sample in gene expression space is determined by its transcriptional state. Genes interact with each other in regulatory networks and as a consequence, the functional relations between genes restrict the distribution of possible gene expression states of the cell to some manifold, or surface, in gene expression space. Typically, the number of genes measured is very large and, consequently, so is the dimension of the studied gene expression space. A variety of mathematical methods have been described that reduce the dimensionality of the datasets so as to find the principal features of the data (Quackenbush, 2001). Two established and commonly used unsupervised methods are multidimensional scaling (MDS) and principal component analysis (PCA) [see e.g. Alter *et al*. (2000) and Bittner *et al*. (2000) for applications to expression data]. These methods work best when data are linearly distributed in data space. For the more general case of nonlinearly distributed data, there are several dimensionality reduction methods like, e.g. Principal Curves (Hastie and Stuetzle, 1988) or Kernel PCA (Schölkopf *et al*., 1996), but so far methods like these have been sparsely applied to gene expression data.

A natural way to handle nonlinearities is to adopt a different distance metric in data space. In most of the applied methods, Euclidean metrics or correlation is applied when estimating similarities/differences between biological samples. In

*To whom correspondence should be addressed.

Bioinformatics 20(6) © Oxford University Press 2004; all rights reserved.

the present investigation we have applied geodesic distances as an alternative measure for similarity. As opposed to the straight-line Euclidean distance, geodesic distances are measured along the surface of the manifold on which data is assumed to lie. Approximations of the geodesic distances are calculated using the Isomap algorithm, originally described by (Tenenbaum *et al*., 2000) and developed as a tool for analysis of complex data, such as, e.g. digital images. Isomap tries to approximate the data manifold by a graph, constructed by locally connecting nearest neighbors. Approximate geodesic distances are then calculated as the distance of the shortest paths between samples in the graph. In the present study, we have applied the approximate geodesic distance measure on two previously analyzed datasets—one set of lymphomas (Alizadeh *et al*., 2000) and one set of lung cancer tumors (Garber *et al*., 2001), and shown that this approach reveals biologically relevant structures in the data not easily detected with a standard MDS analysis of the same data using Euclidean metrics.

SYSTEMS AND METHODS

Datasets

Two previously described microarray datasets were analyzed—one set of 96 lymphoma samples (Alizadeh *et al*., 2000) and one set of 73 lung cancer samples (Garber *et al*., 2001). The selection of genes was, in both cases, unsupervised. The lymphoma data were filtered so that the fluorescent intensity in each channel was greater than 1.4 times the local background for a gene to be included in the analysis, resulting in a total of 854 genes. The samples were divided into the nine diagnostic classes defined by Alizadeh *et al*. (2000). The lung cancer data were centered by sample mean and filtered so that the raw intensity in both channels was greater than or equal to 1.5 times the background, resulting in 831 genes. Samples were divided into five diagnostic classes as described by Garber *et al*. (2001).

Multidimensional scaling

MDS is a mathematical procedure that creates a lower-dimensional configuration of points $\{\bar{x}_i'\}$ so as to approximate optimally given distances between points $\{\bar{x}_i\}$ in a higher-dimensional space. MDS was performed using an implementation of non-metric MDS (Schiffman *et al*., 1981) available in the STATISTICA 6.0 software (Statsoft, Tulsa, OH). In short, the algorithm minimizes the raw stress defined as

$$\phi = \sum_{ij} \left[d(\bar{x}_i', \bar{x}_j') - f(d(\bar{x}_i, \bar{x}_j)) \right]^2$$

for different functions f belonging to a set M of monotone functions. The effect of the transformations f is such that the order relation between distances is preserved rather than the absolute values. The optimization procedure alternates between minimizing ϕ over M and the set of lower-dimensional configurations. The initial configuration in the optimization is found through PCA, i.e. by setting f to the identity.

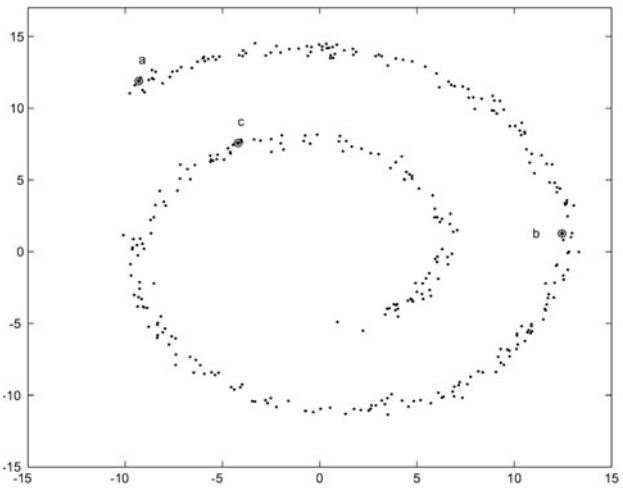

Fig. 1. Data distributed along a spiral. The geodesic distance along the spiral is presumably more reasonable than the Euclidean distance. Thus the distance between a and b should be considered shorter than that between a and c.

Isomap

Generally, MDS techniques work with distance data as it is given, possibly letting them undergo some monotone transformation as described above. The Isomap algorithm (Tenenbaum *et al*., 2000) differs in this respect since distances are transformed so that nonlinear dependencies in data are taken into consideration. Assume, e.g. that data are sampled from a spiral-shaped configuration (Fig. 1). Then the preferable distance measure between points is perhaps not the Euclidean distance, but the geodesic distance along the spiral. Consequently, in Figure 1, the distance between a and b should be considered shorter than that between a and c.

To handle this, Isomap constructs a graph G locally by connecting each data point to its nearest neighbors. The set of nearest neighbors of a point $\bar{x}_0$ is defined either as all points $\bar{x}_i$ within a distance $d(\bar{x}_0, \bar{x}_i) < \varepsilon$, for some chosen $\varepsilon > 0$, or as the K closest points, for some chosen integer $K > 0$. After the graph construction, approximations $d_G(\bar{x}_i, \bar{x}_j)$ to the geodesic distances between points $\bar{x}_i, \bar{x}_j$ are calculated by finding the shortest path in the graph between $\bar{x}_i$ and $\bar{x}_j$. MDS is then applied to these approximate geodesic distances instead of the original distances.

The ability of the Isomap algorithm to produce good approximations of the geodesic distances on the underlying manifold depends on the density of data points and the choice of K (or ε) (Bernstein *et al*., 2000, http://isomap.stanford.edu). If the parameter is too small, a single connected graph is not achieved and distances cannot be calculated between all sample pairs. If, on the other hand, the parameter is too large,

shortcuts, not following the surface of the manifold, may appear in the graph. For the K-rule, the latter situation is likely to appear for large parameter values and low data density. It is reasonable to assume that the dimension of the underlying nonlinear data manifold is fairly large, thus the densities of the presently analyzed datasets are expected to be low. With this in mind and after trying different parameter values, we chose to construct the graph using the K-rule with $K = 2$.

Projection quality

The accuracy of an MDS approximation is quantified by the raw stress of the final point configuration. Lower stress values correspond to a better approximation of the original distances. To evaluate how well an individual sample $\bar{x}_i$ is represented in a projection one can calculate the raw stress over the distances between $\bar{x}_i$ and all other samples, i.e. $\phi_i = \sum_j \left[d(\bar{x}'_i, \bar{x}'_j) - f(d(\bar{x}_i, \bar{x}_j))\right]^2$. Samples with higher stress values are then less well approximated by the projection than samples with lower stress values.

Since the calculation of Isomap graph distances depends on the distribution of data, it is desirable to investigate how stable an acquired Isomap visualization is to changes in the data. This can be done by excluding one sample at a time, constructing Isomap graphs for each of the remaining data subsets and noting for which samples the Isomap graph structure changes drastically. Let G_0 be the graph that is constructed when the whole dataset is used and let G_i be the resulting graph when the i-th sample is left out. For each left-out sample we calculate δ_i, the Euclidean norm of changes in graph distance between points present in both G_0 and G_i divided by the Euclidean norm of graph distances in G_0 between points present in both G_0 and G_i as

$$\delta_i = \frac{\sqrt{\sum_{k,l \in J \times J} \left[d_{G_0}(\bar{x}_k, \bar{x}_l) - d_{G_i}(\bar{x}_k, \bar{x}_l)\right]^2}}{\sqrt{\sum_{k,l \in J \times J} d_{G_0}(\bar{x}_k, \bar{x}_l)^2}},$$

where $J = \{k; \bar{x}_k \in G_i\}$. Then δ_i is a measure of how deformed the Isomap graph is.

RESULTS AND DISCUSSION

Analysis of the lymphoma dataset

To analyze the lymphoma samples, at first a distance matrix based on Euclidean metrics was produced from the data obtained by Alizadeh *et al*. (2000). A lower-dimensional representation of the data was obtained by performing non-metric MDS. Without previous knowledge of subclasses within the sample set no distinct clusters were seen. However, when the classification used by Alizadeh *et al*. (2000) was applied, it was seen that cases belonging to the same classes were mainly located in the same regions of the projection (Fig. 2a). In marked contrast, a similar MDS analysis of the calculated Isomap distances already produced distinct structures when projected into two dimensions. When samples were marked according to their classification a clear connection between classification and structure appeared (Fig. 2b). The two-dimensional Isomap visualization revealed three well-separated groups, all consisting of samples previously known to be of divergent origin. These groups were located at the periphery of the projection; one constituting the chronic lymphocytic leukemia (CLL) samples (yellow), one the activated blood-B samples (light blue), and a third group including resting/activated T cells (red) and transformed cell lines (pink). The other samples were positioned in the center of this structure. One interesting observation, already apparent in the two-dimensional representation, was the misclassification of one of the transformed cell lines (pink in Fig. 2b). This case, SUDHL-5, was grouped together with the other transformed cell lines by hierarchical cluster analysis (Alizadeh *et al*., 2000). In contrast, the Isomap algorithm placed this case at a distance from the transformed cell line class and between the activated blood-B and the diffuse large B-cell lymphoma (DLBCL) samples. Hence, this cell line seems to be more similar to the DLBCL and the activated blood-B class, than the other transformed cell lines. This is perhaps not surprising, given the fact that SUDHL-5 is a cell line established from a DLBCL tumor (Epstein and Kaplan, 1979), whereas at least three of the remaining cell lines are of T-cell origin (Tweeddale *et al*., 1987; Mehra *et al*., 2002). The third dimension revealed even further informative structures that could be linked to previous biological knowledge (Fig. 2c and d). For example, the central group of the samples in Figure 2b showed an extended distribution in the third dimension, revealing two arms extending upwards; one consisting of the follicular lymphoma group (FL, green) and the other of the DLBCL group (blue) interconnected by two cases of germinal center B-cells (GC B-cells; orange). When examining the FL cases (green in Fig. 2c, d and f) these could be separated into two groups; one located more closely to the GC B-cells and one more closely to the resting blood-B samples. The proximity of the latter group with the resting blood B-cells (violet) and CLL samples (yellow), could reflect the low proliferation rates of these samples, as also suggested by Alizadeh *et al*. (2000). The largest and most heterogeneous group of tumors was the DLBCL, which formed an extended central cluster (Fig. 2c and d). When labeling this group into those belonging to the 'germinal center B cell-like' (GCBL) or 'activated B cell-like' (ABL) DLBCL types as described by Alizadeh *et al*. (2000), the GCBL cases occupied the upper-half of the structure (red in Fig. 2e), whereas the ABL group preferentially occupied the lower half (green in Fig. 2e). As expected, the GCBL group extended towards the two GC B-cell samples (orange in Fig. 2c, d and f), whereas the ABL group was positioned in-between samples of normal lymph node/tonsil (gray in Fig. 2b–d) and the activated blood B samples (light blue in Fig. 2b–d). The proximity of the GCBL and ABL tumors to these normal cell samples was also seen

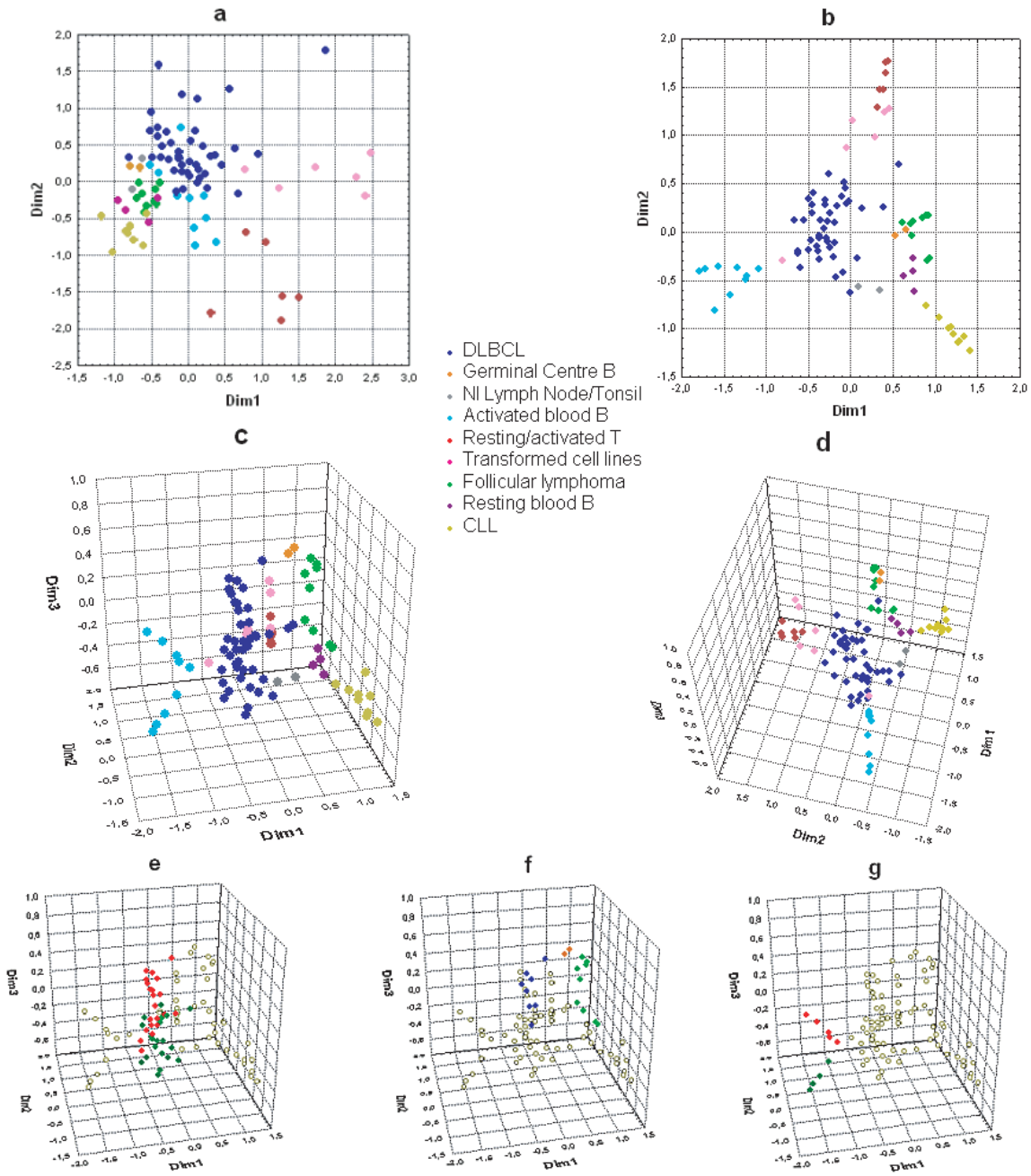

Fig. 2. Visualization of lymphoma microarray data. (**a**) A two-dimensional MDS representation of the Euclidean distances. (**b**) A two-dimensional MDS representation of the approximate geodesic distances. (**c**) A three-dimensional MDS representation of the approximate geodesic distances. (**d**) As in Figure 2c but from a different angle. Color codes in Figure 2a–d are as given in the figure. (**e**) DLBCL–GCBL, red; DLBCL–ABL, green. (**f**) FLs, green; GC B-cells, orange; DLBCLs with *t*(14;18), blue. (**g**) Blood-B cells activated for 6 h, green; blood-B cells activated for 24 h, red. For details, see text.

by Alizadeh *et al*. (2000) using hierarchical cluster analysis. Interestingly, a close inspection of the DLBCL cases (blue in Fig. 2f) extending upwards towards the GC B-cell samples (orange in Fig. 2f), revealed that these in fact were *t*(14;18)-positive as recently reported by Huang *et al*. (2002). Thus, Isomap placed tumors with similar primary genetic changes, i.e. DLBCL with a *t*(14;18) and FLs, which are known to be characterized by the same translocation, in close proximity and in a continuum, extending out from the normal GC B-cells. Hence, the data suggest that the latter two tumor types both initially develop from GC B-cells as suggested previously (Alizadeh *et al*., 2000; Küppers *et al*., 1999). In addition, as

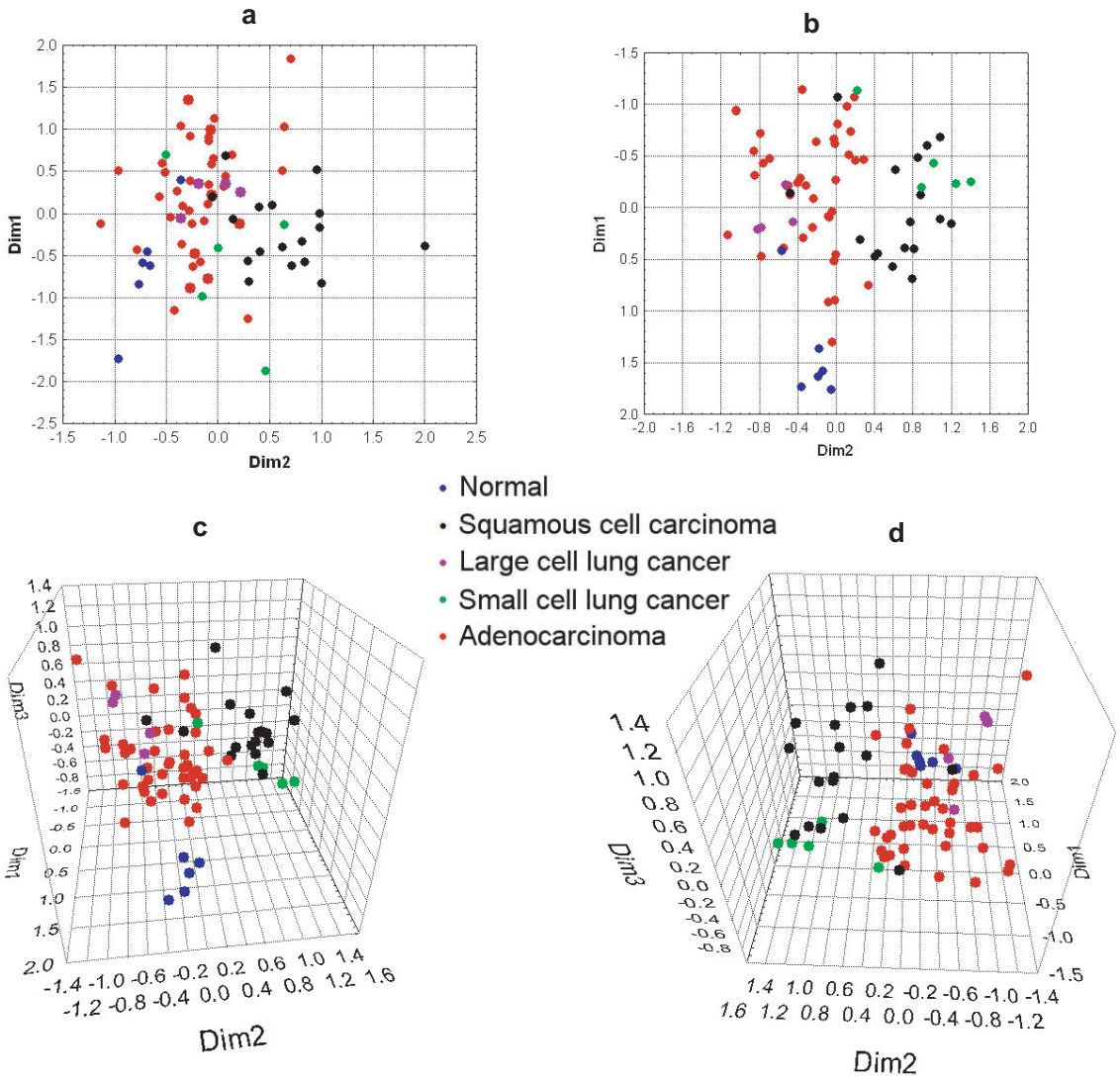

Fig. 3. Visualization of lung cancer microarray data. (**a**) A two-dimensional MDS representation of the Euclidean distances. (**b**) A two-dimensional MDS representation of the approximate geodesic distances. (**c**) A three-dimensional MDS representation of the approximate geodesic distances. (**d**) As in Figure 3c but from a different angle.

the tumor samples are organized in a linear order, originating from the GC B-cells, the observed order could possibly reflect gene expression alterations related to tumor progression. Finally, when identifying the individual samples within the activated blood-B samples, it was found that the upper arm (red in Fig. 2g) corresponded to cells stimulated for more than 24 h, whereas the lower arm (green in Fig. 2g) included the samples stimulated for 6 h. Hence, this observation further underscores the ability of the Isomap algorithm to differentiate between biologically similar samples.

Analysis of the lung cancer dataset

The same analysis was applied to the lung cancer dataset (Garber *et al*., 2001). First, a two-dimensional MDS analysis was performed based on Euclidean distances, displaying an unstructured cluster of tumor cases (Fig. 3a). When the classification used by Garber *et al*. (2001) was applied, it became evident that one-half of the structure was dominated by adenocarcinoma (AC; red) cases and the other half by squamous cell carcinoma (SCC; black) cases. In contrast, the approximate geodesic distances revealed further substructures when performing the corresponding MDS analysis (Fig. 3b). More specifically, the SCCs were separated further from the ACs. In addition, all but one of the small cell lung cancer (SCLC; green) cases were located within or adjacent to the SCC cluster. The remaining case (207-97-SCLC) was placed together with the ACs, suggesting a larger similarity of this tumor to that group. Further, five out of the six normal cases

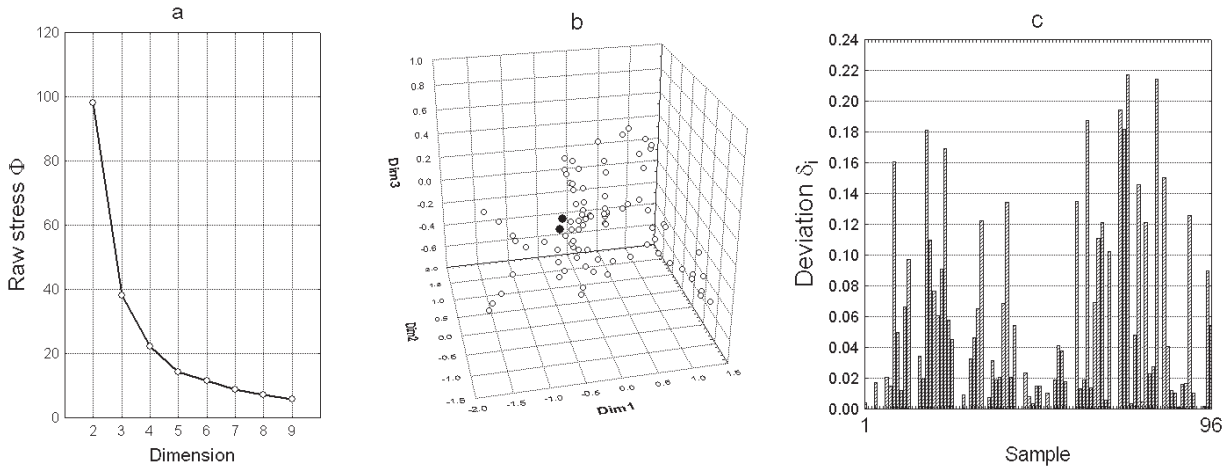

Fig. 4. Projection quality of lymphoma data. (**a**) Raw stress for MDS-projection of approximate geodesic distances relative to projection dimension. (**b**) Locations of the two samples (OCI Ly10 and DLCL-0011) showing high individual stress values in the three-dimensional MDS representation of the approximate geodesic distances. The representation is shown from the same angle as in Figure 2c. (**c**) Structure stability analysis. Deviation *i* in graph distance for each left-out sample.

(blue) formed a well-separated group at the periphery. These cases were derived from adult tissue, whereas the outlier, located among the AC tumors, was a sample obtained from fetal lung.

A three-dimensional MDS projection (Fig. 3c and d) based on approximate geodesic distances displayed an even better separation between the three major groups—ACs, SCCs and normal cases. Furthermore, the SCCs and the SCLCs, which clustered together in the two-dimensional visualization were now separated. Like in the two-dimensional projection, the ACs formed one heterogeneous group. Thus, we could not confirm the results of Garber *et al*. (2001) who, using hierarchical clustering, divided the ACs into three major subgroups and a fourth group of six samples, not included in any of these main clusters. It remains unclear whether this discrepancy stems from a shortcoming in the Isomap algorithm's capability to identify these suggested subgroups, or from the fact that hierarchical clustering always detects clusters in data regardless of whether any real underlying groups are present. Similarly, the large cell lung cancer (LCLC; violet) cases could not be separated from the ACs. These tumors are poorly differentiated and their expression similarities with ACs may suggest a common tumor origin.

Projection quality

Additional Isomap projections of the lymphoma data with dimensions from four up to nine were made. For each projection dimensionality, the overall raw stress was calculated and plotted in a scree plot (Fig. 4a). The scree plot indicated that the data would be well described by a three- or four-dimensional projection. Raw stress values were also calculated for individual samples, in order to evaluate the credibility of sample locations. Two samples, OCI Ly10 and DLCL-0011, had a substantially higher stress than the rest and these were marked in the visualization (Fig. 4b). However, none of these samples were crucial for the detailed biological interpretations made. In order to evaluate the robustness of the structure, 96 Isomap graphs were constructed, excluding one sample at a time. For each sample subset, the distance deviations in the Isomap graph were calculated (Fig. 4c). For the studied dataset and the used Isomap parameter settings, the samples with high graph distance deviations are apparently important in the calculation of graph distances and noise disturbances on these samples have a relatively large impact on the graph structure. Since there is no knowledge of the underlying data manifold, we cannot tell to what degree their positions in gene expression space are 'biologically correct' or if they have been dislocated by noise.

Raw stress analysis was performed also for the lung cancer data. A scree plot showed that a three- or four-dimensional projection was appropriate. To evaluate the goodness of fit for individual samples, individual raw stress values were calculated. The distribution of these values was more homogeneous than the corresponding distribution for the lymphoma data in that it did not contain any obvious outlier values.

CONCLUSIONS

In this work, two alternative ways of measuring dissimilarities or distances between gene expression profiles were compared. Visualizations were created with both Euclidean and approximate geodesic distances as inputs in MDS. The results showed that the approximate geodesic distance measure gave rise to more informative visualizations on the investigated lymphoma and lung cancer data. Even without supervised filtering of the genes with respect to class differentiation, e.g. by creating a weighted gene list (Luo *et al*., 2001),

diagnostic classes appeared as discernible units. That the approximate geodesic distance measure seems more informative could be taken as an indication that tumor samples are distributed on a nonlinear manifold in gene expression space, which in turn would imply that functional relations between genes are nonlinear. Furthermore, the fact that the approximate geodesic distances correspond to the sum of incremental steps between slightly different tumor samples may open the possibility to capture aspects of tumor progression in the form of microarray data. More generally, the results demonstrate the benefit and importance of taking nonlinearities in gene expression data into account. To conclude, we anticipate that the conceptual framework of geodesic distances will prove useful in both practice and theory for the analysis of gene expression data.

ACKNOWLEDGEMENTS

We thank Kalle Åström for stimulating discussions and useful ideas. This work was supported in part by grants from the Swedish Cancer Society.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.R., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Bernstein,M., de Silva,V., Langford,J.C. and Tenenbaum,J.B. (2000) Graph approximations to geodesics on embedded manifolds. *Technical Report*, Stanford University.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E. Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al*. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Epstein,A.L. and Kaplan,H.S. (1979) Feeder layer and nutritional requirements for the establishment and cloning of human malignant lymphoma cell lines. *Cancer Res.*, **39**, 1748–1759.

Fambrough,D., McClure,K., Kazlauskas,A. and Lander,E.S. (1999) Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell*, **97**, 727–741.

Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Gengelbach,M., van de Rijn,M., Rosen,G.D., Perou,C., Whyte,R. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.

Hastie,T. and Stuetzle,W. (1988) Principal curves. *J. Amer. Stat. Assoc.*, **84**, 502–516.

Huang,J.Z., Sanger,W.G., Greiner,T.C., Staudt,L.M., Weisenburger,D.D., Pickering,D.L., Lynch,J.C., Armitage,J.O., Warnke,R.A., Alizadeh,A.A. *et al*. (2002) The t(14;18) defines a unique subset of diffuse large B-cell lymphoma with a germinal center B-cell gene expression profile. *Blood*, **99**, 2285–2290.

Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,J., Jr, Boguski,M.S. *et al*. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Khan,J., Saal,L.H., Bittner,M.L., Jiang,Y., Gooden,G.C., Glatfelter,A.A. and Meltzer,P.S. (2002) Gene expression profiling in cancer using cDNA microarrays. *Methods Mol. Med.*, **68**, 205–222.

Küppers,R., Klein,U., Hansmann,M.L., and Rajewsky,K. (1999) Mechanisms of disease: Cellular origin of human B-cell lymphomas. *N. Engl. J. Med.*, **341**, 1520–1529.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Luo,J., Duggan,D.J., Chen,Y., Sauvageot,J., Ewing,C.M., Bittner,M.L., Trent,J.M. and Isaacs,W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.*, **61**, 4683–4688.

Mehra,S., Messner,H., Minden,M., and Chaganti,R.S. (2002) Molecular cytogenetic characterization of non-Hodgkin lymphoma cell lines. *Genes Chromosomes Cancer*, **33**, 225–234.

Pollack,J., Van de Rijn,M. and Botstein,D. (2002) Challenges in developing a molecular characterization of cancer. *Semin. Oncol.*, **29**, 280–285.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nature*, **2**, 418–427.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Schena,M., Shalon,D., Davis,R.W. and Brown,P. (1995a). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P. and Davis,R.W. (1995b). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.

Schiffman,S., Reynolds,M. and Young,F. (1981) *Introduction to Multidimensional Scaling*. Academic Press, New York, pp. 362–371.

Schölkopf,B., Smola,A. and Müller,K.-R. (1996) Nonlinear component analysis as a kernel eigenvalue problem. *Technical Report 44*, Max-Planck-Institut für biologische Kybernetik.

Tenenbaum,J.B., de Silva,V. and Langford,J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2322.

Tweeddale,M.E., Lim,B., Jamal,N., Robinson,J., Zalcberg,J., Lockwood,G., Minden,M.D. and Messner,H.A. (1987) The presence of clonogenic cells in high-grade malignant lymphoma: a prognostic factor. *Blood*, **69**, 1307–1314.