Sorting Points Into Neighborhoods (SPIN): Data Analysis and Visualization by Ordering Distance Matrices

D. Tsafrir[1], I. Tsafrir[1], L. Ein-Dor[1], O. Zuk[1], D.A. Notterman[2] and E. Domany[1]

[1] Department of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel;

[2] Departments of Pediatrics and Molecular Genetics, UMDNJ-Robert Wood Johnson Medical School.

January 25, 2005

Abstract

We introduce a novel unsupervised approach for the organization and visualization of multi-dimensional data. At the heart of the method is a presentation of the full pairwise distance matrix of the data points, viewed in pseudo-color. The ordering of points is iteratively permuted in search of a linear ordering, which can be used to study embedded shapes. Several examples indicate how the shapes of certain structures in the data (elongated, circular and compact) manifest themselves visually in our permuted distance matrix. It is important to identify elongated objects since they are often associated with a set of hidden variables, underlying continuous variation in the data. The problem of determining an optimal linear ordering is shown to be NP complete, and therefore an iterative search algorithm with $O(n^3)$ step-complexity is suggested. By using SPIN to analyze colon cancer expression data we were able to address the serious problem of sample heterogeneity, which hinders identification of metastasis related genes in our data. Our methodology brings to light the continuous variation of heterogeneity - starting with homogeneous tumor samples and gradually increasing the amount of another tissue. Ordering the samples according to their degree of contamination by unrelated tissue allows separation of genes associated with irrelevant contamination from those related to cancer progression.

Availability: Software package will be available for academic users upon request

Contact: fedafna@wisemail.weizmann.ac.il

Exploratory data analysis is critical in a broad range of research areas, where large collections of data need to be meaningfully arranged and presented. It is especially relevant in biology, where the past decade has witnessed an explosion in data production, largely attributed to the wide spread use of high-throughput technologies, such as gene expression arrays. One major challenge in the analysis of large-scale expression data is effective data organization and visualization [Eisen et al., 1998], where typical goals include class discovery and feature extraction [Ramaswamy

et al., 2001]. However, in many cases the data is characterized by inherently gradual progression rather than by clear, abrupt changes between discrete states. For such cases clustering algorithms, whose aim is to partition the data into several distinct groups, fail to capture the gradual nature of the phenomena. For example, when studying the evolution of a certain disease one expects the existence of continuous variables associated with disease progression. Therefore, any attempt to place a sharp division on such an inherently continuous phenomenon is doomed to be somewhat arbitrary.

The problem of uncovering and presenting continuous trajectories and variables led us to develop SPIN (Sorting Points Into Neighborhoods), an unsupervised *sorting method*, very different in spirit, philosophy and implementation from clustering. SPIN uses an iterative process to find an informative permutation of the data points. This is challenging since permutation space is factorially large, containing a very small measure of meaningful orderings (needle in a haystack problem). SPIN's intuitively color coded image of the reordered distance matrix uncovers elongated structures that reveal the existence of continuous variables that govern the variation in the data. Our ordering approach is especially appropriate for studying scenarios characterized by accumulation of gradual changes, since it excels at tracking progression. There exist other methods that search efficiently for special order-preserving sub-matrices in expression data [Ben-Dor et al., 2003, Getz et al., 2000, Lepre et al., 2004]. One of these [Ben-Dor et al., 2003] is limited to the case when the expression levels of the selected genes vary monotonously over the ordered samples, whereas SPIN uncovers with equal ease a multi-dimensional subspace of genes in which the samples trace a complicated trajectory, along which not a single gene varies monotonously. Another method [Lepre et al., 2004] captures efficiently a group of samples that form a tight sphere in the special subspace of genes, but was not designed to capture continuous variation.

The purpose of identifying shapes is to gain insight about the underlying process, such as the continuous nature of cell differentiation or the closed loop formed by cells along different stages in the yeast cell-cycle. In our main application we demonstrate a possible resolution of a well-known problem in micro-array measurements, namely that of sample heterogeneity. Previous expression-array based studies of cancer [Alon et al., 1999] recognized the issue of variability in tissue-composition of samples, and showed that some of the variation in expression between normal and cancer tissue can be attributed to such causes. This may result in identification of differentially expressed genes that are unrelated to the biological agenda, wrongly implicating them with cancer. We demonstrate how SPIN can be used to distinguish disease related genes from irrelevant ones.

We start with the concept and intuition of SPIN, explaining how to infer shape characteristics from SPIN-permuted distance matrices. This is done by going through a set of examples of increasing complexity, including toy models as well as real biological data. Next, we explain how the algorithm works, and the rationale behind a specific implementation. We prove that the problem the algorithm attempts to solve is NP-hard, and also show that our

heuristic quickly converges ($O(n^2) - O(n^3)$) to useful solutions. We stress the general applicability of SPIN to any data set where a dissimilarity metric between points can be defined. Finally, we describe an application to analysis of large-scale gene-array data, specifically addressing to issue of disease progression versus sample contamination in colon cancer.

Demonstrating the concept

Ferreting elongation

Our methodology assumes that a distance matrix, D, can be defined, whose element D_{ij} represents the dissimilarity between points i and j (Euclidean distance was used throughout this article). Starting with a random ordering of data points, the corresponding initial unordered distance matrix is impossible to interpret. However, the permuted image, obtained after reordering the data by $SPIN$, is highly informative. For our first example consider points uniformly distributed within a cylinder, as presented in fig. 1a1. SPIN orders the points from one end of the cylinder to the other, so that the correspondingly permuted distance matrix has a characteristic pattern, as seen in fig. 1a3: the elements near the main diagonal stand for short distances (colored blue), with a clear gradient of increasing distances (colors vary from blues to reds) as one moves away from the main diagonal. Although both the ordered (fig. 1a3) and unordered (fig. 1a2) matrices contain exactly the same elements, only the permuted matrix allows an observer to deduce structural information.

In gene expression data an elongated conformation may be associated with a gradual process, such as cells going through several successive stages of differentiation. In [Rozovskaia et al., 2003] the U95 affymetrix chip was used to determine expression profiles of acute lymphoblastic leukemias (ALLs), including tumors at various stages of differentiation (such as pre-B, pro-B, and T cell ALLs). One particular group of genes identified in [Rozovskaia et al., 2003] displayed expression profiles that are sensitive to the differences between the early and late differentiation stages (pro-B vs. pre-B and T cell tumors). Here the dissimilarity matrix between samples was calculated using only the differentiation-implicated genes as features. In fig. 1a4 the SPIN permuted matrix displays a clear pattern of elongation, with early cells (pro-B) placed at one side and more differentiated cells (from pre-B and T cell tumors) located at the other end of the trajectory.

If the elongated object closes upon itself, i.e. represents a cycle, then the corresponding fingerprint is also periodic, as shown in fig 1b1-3. The phasing in the colors as one progressively deviates from the main diagonal can be understood by considering the organization of a circle. Starting from any arbitrary point A and going around the ring, the distance of the current point to A increases monotonously (colors change from blue to red) until the diametrically opposing point is reached. At this stage the distances begin to decrease (colors go back to blue), as we approach the point of origin from the other side. The most obvious fingerprint of a circular object is the appearance

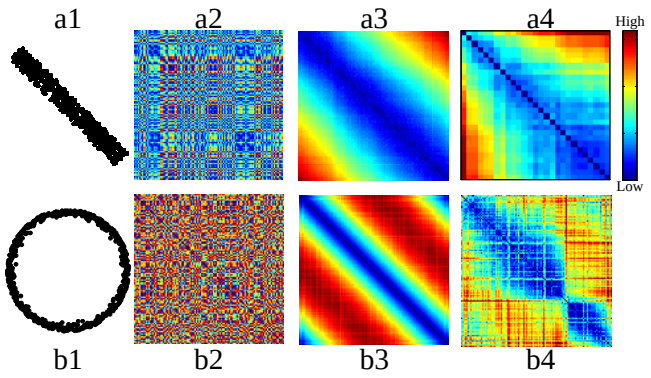

Figure 1: Shapes of simple objects, each consisting of 500 points (except for a4 where only 27 samples are available). (a) 1. Points uniformly distributed within a cylinder. 2. The corresponding distance matrix: the color of element D_{ij} reflects the relative distance between points i and j , where blue (red) denotes small (large) distances. This randomly ordered distance matrix is the input to $SPIN$. 3. The final permutation, in which points are ordered along the trajectory of the cylinder. Only the permutation of the rows/columns changes. 4. Leukemia samples: the permuted distance matrix for 27 blasts at various stages of differentiation, including pre-B, pro-B, and T cell ALLs. (b) Corresponding images for a cyclic object: A ring of points is characterized by a cyclic pattern, with small distances (blue) near the main diagonal and at the corners. 4. Yeast cell cycle: 500 genes with highest standard deviation across the samples were analyzed, using the raw expression data without any manipulation (except for thresholding at the 99th percentile to avoid spikes). The ordered image reveals the heterogeneous nature of the ring, corresponding to separation into different stages of the cell-cycle.

of blue at the corners far from the main diagonal of the distance matrix.

For a real-world application consider the yeast Elutriation-Synchronized cell-cycle expression data (taken from [Spellman et al., 1998]). [Spellman et al., 1998] employed a supervised *phasing* method to assign genes to five known classes, namely G1, S, S/G2, G2/M and M/G1, utilizing the expression profiles of genes that were previously known to participate in specific phases of the cell cycle. They then proceeded to perform unsupervised analysis, specifically hierarchical clustering, and found that most genes belonging to the same class were clustered together. Here we simply filtered out the most highly varying transcripts, as monitored during progression of cell-cycle, and proceeded to calculate their pairwise dissimilarity matrix. As seen in fig. 1b4, the permuted distance matrix contains the signature of a ring. Assigning such a cyclic nature to genes associated with cell-cycle is in accordance with known biological dynamics and functions [Alter et al., 2000]. This example highlights the ease of ordering gene expression data in SPIN, and the informative and intuitive nature of the color-enhanced output. Previous studies have recognized the inherent cyclic nature of this data set [Alter et al., 2000], but required several stages of data manipulation and normalization, followed by a manual ordering using the PCA projection to convey the results that are easily captured in SPIN.

Multiple clusters

The most common approach to analyzing a data set composed of multiple objects is clustering. However, by emphasizing partitioning of data, the clustering approach neglects the issue of elucidating the shapes of embedded

objects in multi-dimensional data. SPIN, on the other hand, focuses on meaningful ordering and presentation, thus gaining insight into local and global structures. This is demonstrated on artificial data (fig. 2a), where a presentation of the permuted distance matrix (fig. 2b) brings to light the separation into four groups, as can be seen by the sharp boundaries. Furthermore, one can infer the shape of each cluster, as well as the global conformation. The two tight spherical clusters (eyes) appear as dark blue squares on the main diagonal. From the light blue color of the squares between them we can deduce that the eyes are relatively close to each other, i.e. their relative placement. The next cluster (smile) has a gradient of colors, from dark blue on the main diagonal to light blue at the corners. As explained above, this indicates an elongated structure. The fourth cluster cycles through the entire spectrum, returning to dark blue at the corners, signifying a cyclic shape (see fig. 1b). The fact that the distance between opposing points on the ring is the largest (i.e. the darkest red in the matrix) indicates that the ring encompasses all other points.

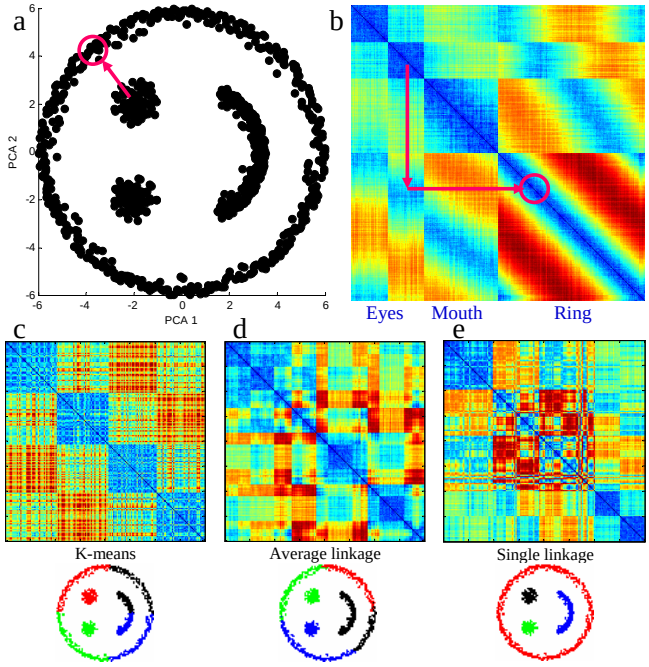

Figure 2: Relations and shapes of multiple clusters. SPIN's results for a toy data set of 800 points in 10-dimensions. The complex object was originally generated in 3-D, and then seven additional dimensions of noise (uniformly distributed between -1 and 1) were added. (a) The projection of the data points onto the first and second PCA plane. (b) The SPIN sorted distance matrix. From this organized matrix one can easily infer the shapes of the four clusters and their relative placement. For example, the position on the ring closest to the top eye is marked by a red circle. This can be inferred from the sorted matrix by locating the darkest blue elements in the rectangle corresponding to the distances between the eye and the ring, as shown by the arrows. For comparison, the results of three popular clustering methods were translated to permutations on the distance matrix: (c) K-means (with k=4) (d) Average linkage and (e) Single linkage. The division of the points into four clusters (red, blue, green and black) is presented below each matrix. K-means and average linkage fail to identify the correct clusters. Single linkage does identify the four clusters, but does not order the points within them in an informative manner.

Some clustering algorithms, such as average-linkage and k-means, fail to correctly cluster this data (see fig. 2c-d). Others, such as single-linkage, succeed in rightly separating the data (see fig. 2e), but are not able to

convey the different shape-characteristics of all four objects. A linkage algorithm can be supplemented by a leaf-ordering algorithm [Bar-Joseph et al., 2001], in order to provide a meaningful organization of points within clusters. However, even an ordered tree is lacking with respect to highlighting shapes. In SPIN, the inherent coupling between visualization and organization produces a powerful presentation tool. The permuted distance matrix captures the over all layout of compound structures, as well as the local conformation of its components.

Projection enhancement

In the exploration of gene expression data linear models were employed to describe the expression levels of genes as a linear function of common hidden variables. Singular Value Decomposition (SVD) [Alter et al., 2000] was used to decompose the gene profiles into linear combinations of *eigengenes*, i.e. the eigenvectors of the covariance matrix; Independent Component Analysis (ICA) [Liebermeister, 2002] produced a linear model based on hidden variables termed *expression modes*. In such approaches the projection of data to smaller subspaces reduces noise and allows useful visualization. In SPIN we suggest a different approach, in which the distance matrix is not subjected to any distortions, thus fully preserving the original structure of the data. One advantage of avoiding distortion is elimination of false positives, in the sense that the fingerprint of an elongated structure in the SPIN-permuted matrix invariably implies a genuine elongation in the data. In the supplementary material we further discuss the relationships between the SPIN permutation and projection according to PCA.

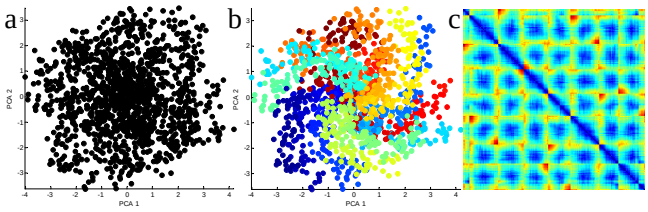

Figure 3: Intersecting rods. A set of seven orthogonal intersecting cylinders, comprised of 1400 points in seven dimensions. The rods were twisted by rotation with angles that increase linearly with the distance from the origin. (a) The points displayed in the first two PCA (b) The same projection with the points colored according to their placement in SPIN; the first point in the SPIN permutation is colored dark blue, going through blue, green, yellow and orange, with the last points colored dark red. In this example the coloring is crucial for making sense out of a complex image. (c) The correspondingly SPIN-permuted distance matrix. The region of the intersection creates blue patches in the off-diagonal regions of the distance matrix.

In the examples presented so far, by applying dimensionality reduction methods, the objects' shapes could be clearly discerned from a $3d$ projection of the data points. The next example illustrates SPIN's ability to deal with more complex objects embedded in a truly high dimensional space, objects whose structure is seriously distorted when projected onto three dimensions. In such cases even the most up-to-date dimensionality reduction methods are doomed to fail in finding three dimensions which properly capture the data structure. Fig. 3a shows a PCA projection of points constituting a set of seven intersecting twisted cylinders in $d = 7$ dimensions. Projecting such a

relatively complicated object onto the first principal components does not produce a clear image (fig. 3a). Coloring the points according to SPIN's linear ordering (see fig. 3b) produces a much more informative image. Furthermore, the distance matrix (fig. 3c) identifies each rod as an elongated structure (along the main diagonal). The relationships between the seven rods can be deduced from the patterns in the off-diagonal regions in the organized distance matrix. For example, the fact that the rods share a common nexus is reflected by a grid of blue patches. Combination of any dimensionality reduction technique with SPIN may serve to highlight the shape-characteristics of a high-dimensional object, that are not immediately made evident by projection onto a lower dimensional space.

Expression data analysis

In the context of gene expression data we implemented SPIN in an interactive GUI that accepts an expression matrix as input, and supports the following actions:

1. Ordering of samples using genes as features.
2. Ordering of genes using samples as features.
3. Zooming in on subsets of the original expression matrix to order objects in a reduced subspace.

A coupling between samples and genes is produced by the ability to identify a group of genes (samples) that fluctuate in a synchronized manner. Similar in spirit to the Coupled Two-Way clustering approach [Getz et al., 2000] we proceed by using the *zoom in* operation to order the samples (genes) in a selected reduced subspace. One can redefine the working space in a recursive manner.

Formal statement of the Problem

The input to SPIN is a distance matrix $D \in \mathbb{R}^{n \times n}$ calculated for data composed of n points, and its output is a reordered distance matrix, obtained by permuting the n objects according to a particular permutation $P \in S_n$ (the permutation group of n points). We denote by P also the permutation matrix associated with P. In search for criteria for an informative permutation, we observed that well-ordered distance matrices exhibit two distinct and sometimes competing properties. First, in many cases the values in the upper rows tend to increase with the column index (and in the bottom rows - decrease), as in fig. 1a3. This type of ordering demands that large distances are assigned to corners, far from the diagonal. The second, alternative aim is to ensure that the elements near the main diagonal tend to have smaller dissimilarity values, i.e. the linear ordering is such that if two points are positioned near each other, their distance in the full high-dimensional space is also small (see fig. 1b3). We term the two properties 'Side-to-Side' (denoted STS) and 'Neighborhood', respectively.

These attributes can be mathematically formulated by introducing an energy (or cost) function $\mathcal{F} \equiv \mathcal{F}_D : S_n \to \mathbb{R}$ quantifying the quality of a permutation. Thus, the ordering problem becomes finding the permutation P that minimizes $\mathcal{F}$. We concentrate on the following family of functions : $\mathcal{F}(P) = tr(PDP^TW) = \sum_{i,j=1}^{n} W_{ij} D_{P(i)P(j)}$, where tr denotes matrix trace, and $W \in \mathbb{R}^{N \times N}$ is some weight matrix. For this family, the optimization problem is known as the Quadratic Assignment Problem (QAP), introduced by [Koopmans and Beckmann, 1957]. The general QAP is considered an extremely difficult optimization problem. It is known to be *NP-Hard* even to approximate, and in practice, usually untractable for n more than 30. (See [Burkard et al., 1998] for a comprehensive survey of the problem).

The *STS* property is captured by setting $W = XX^T$, for some strictly increasing (column) vector X (in our implementation we worked with $X_i = i - (n+1)/2$). *Neighborhood* is reflected by choosing W to be symmetric and concentrated in a region, determined by a parameter σ, around its main diagonal (our choice of W is defined below). We show below that finding a global minimum for our particular choices of $\mathcal{F}$ is NP-hard, and we propose two iterative heuristic algorithms to search for minima. We prove, for both algorithms, that the energy is non-increasing on every iteration. Both algorithms were used in the examples presented in this article, but the displayed images are from *Neighborhood*.

The *STS* algorithm

We have shown that the *STS* problem is NP-Complete by reducing it to the well known *k-clique* problem in graph theory (see supplementary material).

The *STS* algorithm is given by :

Side-to-Side

Input : D and X.

1. Set $X^0 = X$, $t = 0$, define $P^{-1} = I_{n \times n}$.
2. Calculate $S^t = DX^t$.
3. Find P^t which sorts S^t in a descending order.
4. If $P^t S^t \neq P^{t-1} S^t$, set $X^{t+1} = {P^t}^T X^0$,
   set $t = t + 1$ and go to 2.
5. Output $P^t D {P^t}^T$.

We call each pass through steps $2-4$ a *STS* iteration, whose complexity is $O(n^2)$. Each *STS* iteration can be viewed as a mapping from the permutation group S_n to itself, $G_D : S_n \to S_n$. Thus P is a possible output of *STS* if and only if it is a fixed point of G_D.

In the supplementary material we prove that when the input matrix, D, is a distance matrix, convergence of *STS* to a fixed point is guaranteed after a finite number of steps. The proof is based on showing that every *STS* iteration reduces the cost function, $\mathcal{F}$, guaranteing convergence to a local minimum. Note that the *STS* procedure may converge to a P which does not correspond to the global minimum of $\mathcal{F}$; for different initial permutations the algorithm may terminate at different fixed points, with different values of $\mathcal{F}$. A known strategy to cope with this problem is to start the algorithm from many randomly generated initial permutations, and choose the best fixed point obtained. Moreover, it is also possible to have multiple global minima. For example, define for every permutation P its 'reverse' $\bar{P}$ by $\bar{P}(i) = P(n+1-i)$, $(i = 1, \ldots, n)$. If X is anti-symmetric we get for *STS*: $\mathcal{F}(P) = \mathcal{F}(\bar{P})$, leading to at least two global minima. Some data sets may contain further degeneracies due to inherent symmetries. In practice it is not essential to reach the global minimum since the fixed points to which the algorithm converges are often just as informative.

The *Neighborhood* algorithm

Claim : The *Neighborhood* problem is NP-Hard

Proof : The two ingredients of the problem are the distance matrix D and the weight matrix W. Setting $W_{ij} = 1_{|i-j|=1}$ gives $tr(PDP^TW) = \sum_{i=1}^{n-1} D_{P(i+1),P(i)} + \sum_{i=2}^{n} D_{P(i-1),P(i)} = 2\sum_{i=1}^{n-1} D_{P(i+1),P(i)}$. This is the cost function for the *Travelling Salesman Problem*, which is known to be NP-Hard, even in the Euclidian case [Papadimitriou, 1977]. ■

The following algorithm attempts to relocate a point A to a local neighborhood that *best fits* it, i.e. none of the points in the neighborhood of A are at a large distance from it.

Neighborhood

Input : $D_{n\times n}$ and $W_{n\times n}$

1. Set $W^0 = W$, $P^{-1} = I_{n\times n}$, $t = 0$.
2. Compute $M^t = DW^t$.
3. Set $P^t = argmin_{Q\in S_n} tr(QM^t)$.
4. If $tr(P^tM^t) \neq tr(P^{t-1}M^{t-1})$, set $W^{t+1} = P^{t^T}W$, $t = t+1$ and go to 2.
5. Output $P^tDP^{t^T}$.

Each passage of steps $2-4$ constitutes one *Neighborhood* iteration. The size of the neighborhood is dictated by the choice of W, and in turn, affects the scale at which objects are distinguished. Step 3 can be accomplished by solving the Linear Assignment Problem. This solution reflects the best current guess for an improved location for all the data points. At every iteration, points are sent to their new location, based on the current ordering of the points.

That is, point A is sent to a new location $i(A)$ on the basis of the presently residing points near $i(A)$. However, since all the points are permuted simultaneously, there is no guarantee that this assignment remains optimal, since the points that were near $i(A)$ may have moved elsewhere. Hence the need to re-iterate. Since the Linear Assignment Problem is known to be solvable in time $O(n^3)$ [Dinic and Kronrod, 1969], the complexity of each iteration is $O(n^3)$.

We prove that the energy is improved on every iteration; thus convergence to a fixed point is guaranteed after a finite time

Claim : $tr(P^{t+1}DP^{t^T}W) \leq tr(P^tDP^{t-1^T}W)$

Proof : $tr(P^{t+1}DP^{t^T}W) = tr(P^{t+1}DW^{t+1}) \leq tr(QDW^{t+1})\ \forall Q \in S_n$

Using the symmetry of W and the property $tr(AB) = tr(BA)$ we get :

$$tr(QDW^{t+1}) = tr(QDP^{t^T}W) = tr((QDP^{t^T}W)^T) =$$

$$tr(WP^tDQ^T) = tr(P^tDQ^TW)$$

Taking $Q = P^{t-1}$ gives the desired result.

According to step 4, the algorithm terminates unless a strict inequality holds in the above claim. This prevents cycles of constant energy. Since the permutation space is finite, termination in a fixed point after a finite number of steps is guaranteed. ■

Our choice for the weight matrix is taken to be Gaussian, $W_{ij} = e^{-\frac{(i-j)^2}{n\sigma}}$, which is then normalized into a doubly stochastic matrix (i.e. sum of each row and column is equal to one). In this case the mismatch matrix $M = DW$ can be viewed as a Gaussian smoothing of variance σ^2 on each row of D. For a given data set, there exists a range of relevant length scales, where large scales reflect the over all layout of the data, while smaller values give a better local organization at the expense of possibly fragmenting larger structures. This is captured in SPIN by controlling the value of σ. One heuristic scheme that usually works well is starting with a very large σ, iterating several times, then lowering σ (e.g. by a factor of 2) and so forth, in the spirit of simulated annealing. Moreover, the solution of the linear assignment problem (step 3 in the algorithm) can be efficiently approximated by finding the minimum of each row of M, and then sorting the indices of the minima (ties are broken arbitrarily). This heuristic, though not guaranteed to reduce $\mathcal{F}$ at every iteration, generally yields a low energy solution, while considerably speeding up the calculations.

Application to colon cancer

The biological question addressed here is that of recognizing alterations in gene expression that may be linked with the progression of cancer. SPIN is especially appropriate for this analysis, since cancer evolution is an inherently continuous process, which arises from a gradual accumulation of genetic alterations that promote selection of cells

with increasingly aggressive behavior. Such continuity may be completely overlooked by traditional methods that emphasize clear separations. Colon cancer is a good model system since samples are readily available across several, well-defined, stages of the disease, enabling a study of the onset of the neoplastic transformation. Expression profiles were determined for seven types of samples using the Affymetrix U133A GeneChip [Tsafrir et al., 2004]: 47 primary carcinomas; 24 adenomas; 22 normal colon epithelium; 16 liver metastasis; 19 lung metastasis; 11 normal liver; and 5 normal lung. Standard pre-processing of the data included thresholding to 10 (i.e. all expression values smaller than 10 were set to 10) and $\log_2$ transformation. A variance filter was utilized to concentrate on the most relevant genes. We started with the 500 highest varying transcripts, then doubled the number; since there was a significant change in the results, the number of transcripts was doubled again, to 2000. Seeing as this did not alter the main conclusions to a noticeable degree we continued to work with the top 1000.

In the context of such complex data, the search for genes and pathways that are causally involved in cancer is complicated by the need to distinguish their signal from a large background of innocent bystander genes, whose expression levels appear altered due to secondary causes. An initial objective is to generate an overall impression of the data's structure, identifying major partitions and relationships. By filtering the highest variance genes and ordering the resulting expression matrix in SPIN (see fig. 4d) one can get a global view of the data. Two separate ordering operations were performed: one on the genes' distance matrix (rows; fig. 4c) and another on the samples' distance matrix (columns; fig. 4b). Thus, the two-way organized expression matrix allows one to study concurrently the structure of both samples and genes. In consecutive analysis stages, detailed in the following paragraphs, we proceeded to focus individually on sets of correlated genes that were identified in this initial step. SPIN is used to re-order the samples in the context of each gene-set separately, and the resulting permutation is shown to be informative of the underlying biology (see fig. 4e-g). This process of iteratively identifying and focusing on relevant subsets of the initial data matrix is reminiscent of the previously proposed Coupled Two-Way Clustering algorithm [Getz et al., 2000].

Liver contamination

Previous expression-data studies recognized the challenge posed by the heterogeneous composition of sampled tissues [Alon et al., 1999], which was not answered in the context of traditional analysis methods [Ghosh, 2004]. In the current data the clearest separation in the samples is according to their organ of origin - either colon, liver or lung - with the liver samples forming the most distinct group (see fig. 4b) . Even though the tissue samples were carefully dissected, the strongest expression signals are indeed related with the composition of the various samples. The most prominent gene-cluster, highlighted by the bottom black rectangle (Fig. 4c-d), is characterized by highest expression levels in the liver samples. The annotation of genes belonging to this cluster is related to liver functions

(including SERPINA3, CP, HP and APOC1), and therefore we refer to it as *liver-specific*. These liver-specific genes are totally irrelevant to the disease, and yet when performing a PCA projection of the samples (fig. 4a) the first principal direction (explaining 34 percent of variance) is dominated by the difference between normal liver and all other samples. The highly relevant aspect of this phenomenon is that some of the liver metastasis samples display elevated expression levels for the liver-specific genes, shifting their placement in the SPIN ordering towards the location of the normal liver samples. This hinders the ability of traditional statistical analysis methods to generate a list of genes associated with metastatic cancer; when searching for genes with high expression in liver metastasis versus carcinoma samples, liver-specific genes may be implicated. Indeed a supervised hypothesis test [Pan, 2002] generated a list of genes significantly over expressed in liver metastasis as compared to the primary tumor samples (387 transcripts out of the examined top 1000 passed the Wilcoxon ranksum test with FDR of $q = 0.05$ [Benjamini and Hochberg, 1995]). The vast majority of these (97 percent) are associated with liver functions and are in fact members of our liver-specific cluster (fig. 4e). The increased expression for these genes is probably a byproduct caused by contamination of the metastasis samples with normal liver tissue. Therefore, these genes could potentially serve as the basis for constructing a liver-metastasis classifier [Dudoit et al., 2002]; However, analysis based on SPIN clarifies that they do not play a role in the progression of cancer, but rather as a tissue-of-origin indicator.

Muscle and connective tissue contamination

As demonstrated in the previous section, the problem of tissue heterogeneity may be a major complication, and one that was mostly unresolved by traditional analysis methods. In some data sets an assessment by the pathologist of the percentage of relevant tissues in each sample is available [Notterman et al., 2001, Alon et al., 1999], and this information can be utilized to construct an appropriate statistical test [Ghosh, 2004]. In the current data no such knowledge is available, which prevents the proper employment of supervised methods, and necessitates the use of an unsupervised approach. For example, consider a group of genes that appear significantly under-expressed in the neoplastic samples as compared with normal tissue (434 transcripts out of the examined top 1000 passed the Wilcoxon ranksum test with FDR of $q = 0.05$). It has already been observed in colon cancer studies that tumor samples are more biased towards epithelium tissue then their normal counterparts, causing apparent under-expression of genes functioning in muscle and connective tissues [Alon et al., 1999]. In the SPIN-permuted data (fig. 4c-d) the transcripts that show reduced expression in diseased tissue clearly separate into two different gene-profiles. One of this gene-clusters (fig. 4f) exhibits extreme variation in expression in the context of the normal colon samples, which is visually manifested by a pattern of elongation in the relevant SPIN-sorted distance matrix (see fig. 4f)). The annotation of these genes associates them with smooth muscle and connective tissue. Therefore, a likely cause for the disparity in expression among the normal samples are the differences in tissue composition. The reduced

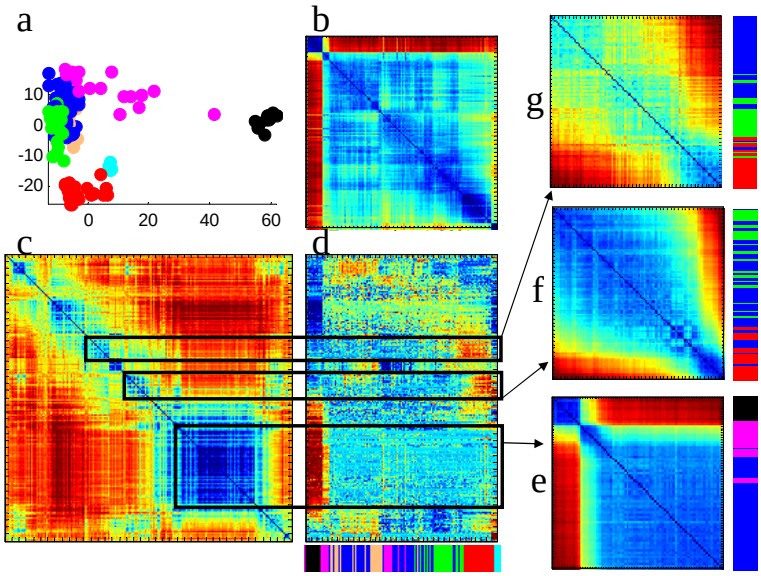

Figure 4: Colon cancer data; expression levels of the 1000 highest variance transcripts over all 144 samples. (a) Projection of the samples onto the first (x-axis) and second (y-axis) principal components, calculated in gene-space. The clinical identity of samples is indicated by a color: primary carcinomas (blue); adenomas (green); normal colon (red); liver metastasis (magenta); lung metastasis (orange); normal liver (black); and normal lung (cyan). This coloring scheme for tissues is kept in all sub-figures. The first PCA reflects 34 percent of variance, and is dominated by the differences between normal liver and all other samples. (b) SPIN-permuted distance matrix for the samples. Colors depict dissimilarity levels between samples, with red (blue) indicating large (small) distances. (c) Genes SPIN-permuted distance matrix. The genes display several distinct expression profiles. (d) Two-way sorted expression matrix. Here colors depict relative expression intensities, where red (blue) denotes relatively high (low) expression. The colored bar below the matrix provides the tissues' clinical identity. Some of the dominant gene-clusters and their expression levels are highlighted by dark rectangulars. Each gene-cluster is used to construct the distance matrix of a particular subset of the samples (e) The distance matrix of normal liver, liver metastasis and carcinoma samples, as calculated in the subspace of the liver-specific gene cluster. The normal liver and carcinoma samples form two distinctly separated, tight spherical clusters, while the metastasis form a connecting elongated cloud, with some of the samples displaying higher proximity (i.e. similarity) to the normal liver samples. The 6 metastasis samples that were placed farthest from the liver samples presumably contain lowest amounts of normal liver tissue, and are therefore referred to as *clean* metastasis. (f) Muscle and connective tissue associated genes. Expression profiles related to cell-mixtures can be distinguished in SPIN by the fact that affected samples tend to order into an elongated shape, due to the relatively high variation in samples' composition. Here the normal colon samples' ordering is indicative of levels of muscle and connective tissue contamination, lowest in the polyp samples. (g) Genes related with a gradual loss of differentiation. Note the placement of the polyp samples between normal and cancer tissue. In (e)-(g) the tissues' clinical identity is given by the colored bar to the right of each distance matrix.

variability detected in the tumor samples (most tumors form a tighter, less elongated shape in fig. 4f) is consistent with the observation made in earlier studies that those samples contain mostly epithelial tissue [Alon et al., 1999], and with the fact that in this experiment they were carefully dissected [Tsafrir et al., 2004]. The adenomas exhibit the lowest expression, perhaps associated with the fact that these benign precursors of cancer protrude into the lumen of the colon, making it easier to remove them surgically without inadvertently including some surrounding muscle or connective tissue. Therefore, using SPIN to study the profile of this gene-cluster clarified that even though the genes are significantly differentially expressed between normal and tumor they are not connected with the neoplastic transformation, but rather with tissue mixtures.

Gradual loss of differentiation

The analysis described in the previous section illustrates how an unsupervised visualization tool such as SPIN can serve to guide rigorous statistical analysis. Employing supervised statistical tests to compare our normal colon samples with the tumors resulted in a mixed list, which included some genes that the SPIN analysis revealed to be related with tissue-mixtures. It is further possible using SPIN to distinguish the desired set of disease-progression associated genes, and show that the reduction in their expression is correlated with the gradual onset of the cancer. Focusing on this subset of genes reveals that in this context the samples trace an elongated shape (fig. 4g), with the normal colon epithelium placed to one side, followed by the adenomas that show a somewhat reduced expression, which is even lower in the carcinoma samples. This set includes genes that were observed to be preferentially expressed in human epithelial cells and down-regulated in cancer, such as carbonic anhydrases [Notterman et al., 2001], Guanylate cyclase activators [Birkenkamp-Demtroder et al., 2002] and EPLIN [Maul and Chang, 1999]. A plausible hypothesis is that these genes are associated with colon functions, and that the SPIN-permutation highlights a gradual loss of differentiation in the transformed tissue. Perhaps the percentage of cells that still keep their colon functions is steadily reduced with the progression of the disease. To conclude, supervised tests were employed to answer a specific question - e.g. differential expression in sick versus healthy tissue, while the analysis in SPIN revealed that some of the implicated genes answer a very different question, i.e. which samples contain the highest proportion of muscle and connective tissue.

Metastasis associated signal

The analysis of the colon cancer data demonstrates a situation where SPIN can be used to assign new labels to samples, and employ this knowledge to improve the application of supervised methods. Metastasis samples, for example, can be marked according to the degree of surrounding normal tissue inadvertently included in the sample's preparation. One way of gaining this information is in the context of the *liver-specific* cluster, where the samples' expression profiles can be viewed as the result of a gradual mixing process, starting with samples extracted from the colon, that contain no liver tissue, and continuing with the metastasis samples that vary in the amount of liver contamination. The degree of liver mixture in each sample is reflected by the SPIN ordering, as can be seen in fig. 4e. The least contaminated metastasis samples can be distinguished by their placement next to the cluster of primary tumors, and labelled as *clean*. A clustering algorithm, such as average linkage, although clearly separating between normal liver and primary tumors, does not produce such meaningful ordering of the metastasis samples. Therefore, SPIN is especially useful in this situation since it can be used to perform a type of *electronic micro-dissection*, allowing identification of the cleanest metastasis samples. A similar procedure can be performed for the lung metastasis samples by using the normal lung samples. It is than possible to proceed by focusing on the *clean*

metastasis samples (from both liver and lung) to uncover genes relevant to the metastatic process. The resulting list included several known oncogenes (such as VEGF, CSE1L [Behrens et al., 2003],TGIF2 and UBE2C); in particular, some are located on chromosomal arm 20q, a region which has been previously shown to be amplified in metastatic colon cancer [Platzer et al., 2002]. In SPIN one can further observe that this group of genes exhibits a gradual elevation in expression which is coupled with the progression of the cancer - from normal tissue, through polyps, increasing in primary tumors and culminating in the *clean* metastasis samples.

Summary and discussion

The emphasis of SPIN is on providing an informative image of the data, one that facilitates extraction of meaningful characteristics. The distance matrix ordered by SPIN can reveal the finger-print of "objects" (e.g. regions with high density of data points) of various shapes that are embedded in a high dimensional representation of the data. We demonstrated this for objects of fairly complex general shapes, including an elongated rod, associated with a continuous variable and for a curve that closes upon itself indicating a cycle, as well as for gene-expression data on cell-cycle and differentiation [1]. Furthermore, the reordered distance matrix is also able to identify multiple objects, where the main linear variation of each entity is followed sequentially, and for which the inter-object relationships, such as their relative placement, can be also identified. The concept of presenting an organized distance matrix is not new, but the SPIN-permuted matrix is shown to be more informative than images produced by popular methods. SPIN was used to resolve the problem of tissue-mixtures in colon cancer expression data, and consequently to allow a clear identification of a set of genes implicated in the gradual loss of differentiation of the transformed tissue.

We presented two different search heuristics for exploring permutation-space: *STS* generates a distance matrix that preferentially places red-colored elements (which denote large distances) near the top-right (and bottom-left) corners. Thus points that are placed far apart in the linear ordering are also distant in the full high-dimensional space. *Neighborhood*, on the other hand, tries to make sure that elements located near the main diagonal are blue-colored, i.e. neighboring points in the linear ordering are also close to each other in the high dimensional space. This subtle distinction in emphasis may lead to substantial difference in the results, as different energy functions reveal alternative aspects of the data, thus enabling the study of diverse properties. From a practical point of view, the *STS* algorithm is faster, while the *Neighborhood* algorithm produces better results for complex data, especially containing compound objects. Therefore, a user could start by applying *STS*, which would generate an image that visually manifests the major elongation in the data, and proceed by utilizing *Neighborhood* to study the more intricate objects.

The important advantages of SPIN are: (1) the simplicity of the underlying algorithm, which makes it easily

[1] In order to show that SPIN does not find structures that do not exist, we performed a random-permutation test on expression data. Results are presented in the supplementary material, section *Loss of structure in randomized expression data*

implementable, accessible and clear to a wide range of users. (2) Running time of $O(n^2)$-$O(n^3)$ gives almost instant feedback for data sets of reasonable size (see table 1 in supplementary material). (3) A fingerprint of an elongated structure in the sorted distance matrix invariably implies an elongation in the data. This follows from the fact that SPIN permutes the distance matrix with no distortions. (4) Synergy with other exploratory analysis techniques. For example, SPIN can be used to order within and between predefined clusters obtained with most standard clustering algorithms. It was also shown that SPIN can enhance dimensionality reduction analysis, as exemplified in fig. 3 where the color coded ordering significantly clarifies the PCA image. Furthermore, the colon cancer application demonstrates a biologically important scenario where the lack of sufficient labels prevents the exclusive employment of supervised statistical methods, while the continuous nature of the underlying biological process makes SPIN an especially appropriate exploration methodology.

Acknowledgments

This work was supported by the NIH under grant #5 $P01$ CA 65930-06. We thank P.B. Paty and W.L. Gerald for preparation of the colon cancer samples and acknowledge use of the Gene Expression Core Facility of the Cancer Institute of New Jersey. We acknowledge partial support by an EC Research Training Network (STIPCO), by the Ridgefield Foundation and by EC FP6 funding. This publication reflects the author's views and not necessarily those of the EC. The Community is not liable for any use that may be made of the information contained herein. We thank U. Feige, I. Kanter, A. Natanzon, Y. Pilpel and R. Raz for useful discussions and comments.

References

U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12):6745–6750, 1999.

O. Alter, P.O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, 97(18):10101–10106, 2000.

Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:S22–S29, 2001.

P. Behrens, U. Brinkmann, and A. Wellmann. Cse1l/cas: its role in proliferation and apoptosis. *Apoptosis*, 8(1): 39–44, 2003.

A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: The order-preserving submatrix problem. *J Comput Biol.*, 10(3-4):373–84, 2003.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.*, 57(1):289–300, 1995.

K. Birkenkamp-Demtroder, L. L. Christensen, S. H. Olesen nad C. M. Frederiksen, P. Laiho, L. A. Aaltonen, S. Laurberg, F. B. Sorensen, R. Hagemann, and T. F. Orntoft. Gene expression in colorectal cancer. *Cancer Research*, 62(15):4352–63, 2002.

R.E. Burkard, E. Cela, P. Pardalos, and S.L. Pitsoulis. *The Quadratic Assignment Problem*, volume 3, pages 241–339. Dordecht:Kluwer Academic Publishers, 1998.

E. A. Dinic and M. A. Kronrod. An algorithm for the solution of the assignment problem. *Soviet Math. Dokl.*, 10: 1324–1326, 1969.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.

G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, 97(22):1207912084, 2000.

D. Ghosh. Mixture models for assessing differential expression in complex tissues using microarray data. *Bioinformatics*, 20(11):1663–1669, 2004.

T. Koopmans and M. Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25: 53–76, 1957.

J. Lepre, J. J. Rice, Y. Tu, and G. Stolovitzky. Genes@work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data. *Bioinformatics*, 20(7):1033–1044, 2004.

W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

R. S. Maul and D. D. Chang. Eplin, epithelial protein lost in neoplasm. *Oncogene*, 18(54):7838–7841, 1999.

D.A. Notterman, U. Alon, A.J. Sierk, and A.J. Levine. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, 61(7):3124–3130, 2001.

W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554, 2002.

C. H. Papadimitriou. The euclidean traveling salesman problem is np-complete. *Theoretical Computer Science*, 4 (3):237–244, 1977.

P. Platzer, M. B. Upender, K. Wilson, J. Willis, J. Lutterbaugh, A. Nosrati, J. K. V. Willson, D. Mack, T. Ried, and S. Markowitz. Silence of chromosomal amplifications in colon cancer. *Cancer Research*, 62(4):1134–1138, 2002.

S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, 98(26):15149–15154, 2001.

T. Rozovskaia, O. Ravid-Amir, S. Tillib, G. Getz, E. Feinstein, H. Agrawal, A. Nagler, E. Rappeport, I. Issaeva, Y. Matsuo, U.R. Kees, T. Lapidot, F. Lo Coco, R. Foa, A. Mazo, T. Nakamura, C.M. Croce, G. Cimino, E. Domany, and E. Canaani. Expression profiles of acute lymphoblastic and myeloblastic leukemias with all-1 rearrangements. *Proc. Natl. Acad. Sci. USA*, 100(13):7853–7858, 2003.

P.T. Spellman, G. Sherlock, MQ. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

D. Tsafrir, W. Liu, Y. Yamaguchi, I. Tsafrir, Y. Wen, W. Gerald, R. Stengel, F. Barany, P. Paty, E. Domany, and D. Notterman. A novel mathematical approach to analyzing gene expression data: results from an international colon cancer consortium. In *proc. of AACR 2004*, 2004.