BIOINFORMATICS

Vol. 19 no. 7 2003, pages 842–850
DOI: 10.1093/bioinformatics/btg081

# Reconstructing the temporal ordering of biological samples using microarray data

Paul M. Magwene[1], Paul Lizardi[2] and Junhyong Kim[1, 3, 4,*]

[1]Department of Ecology and Evolutionary Biology, [2]Department of Pathology, Yale University School of Medicine, [3]Department of Molecular, Cellular, and Development Biology and [4]Department of Statistics, Yale University, New Haven, CT, USA

Received on January 29, 2002; revised on August 26, 2002; accepted on October 21, 2002

## ABSTRACT

**Motivation:** Accurate time series for biological processes are difficult to estimate due to problems of synchronization, temporal sampling and rate heterogeneity. Methods are needed that can utilize multi-dimensional data, such as those resulting from DNA microarray experiments, in order to reconstruct time series from unordered or poorly ordered sets of observations.

**Results:** We present a set of algorithms for estimating temporal orderings from unordered sets of sample elements. The techniques we describe are based on modifications of a minimum-spanning tree calculated from a weighted, undirected graph. We demonstrate the efficacy of our approach by applying these techniques to an artificial data set as well as several gene expression data sets derived from DNA microarray experiments. In addition to estimating orderings, the techniques we describe also provide useful heuristics for assessing relevant properties of sample datasets such as noise and sampling intensity, and we show how a data structure called a PQ-tree can be used to represent uncertainty in a reconstructed ordering.

**Availability:** Academic implementations of the ordering algorithms are available as source code (in the programming language Python) on our web site, along with documentation on their use. The artificial 'jelly roll' data set upon which the algorithm was tested is also available from this web site. The publicly available gene expression data may be found at http://genome-www.stanford.edu/cellcycle/ and http://caulobacter.stanford.edu/CellCycle/.

**Contact:** junhyong@sas.upenn.edu

## INTRODUCTION

Biological systems are inherently dynamical. A common approach to studying the behavior of dynamic biological phenomena is to sample individuals, tissues, or other relevant units at intervals throughout the temporal progression of the system under study. An ordered collection of such samples is called a time series. Well-characterized time series can be used as benchmarks (e.g. cell and developmental stages) for measuring the temporal progress of biological systems and as diagnostic tools for assessing and treating disease.

Several problems associated with sampling from dynamic biological systems (synchronization, temporal sampling, and rate heterogeneity) often make it difficult to obtain accurate time-series. Time series data are usually drawn from a population and without synchronization the samples will contain mixtures of the temporal process. Rate heterogeneity among the members of the population further complicates the temporal samples and can lead to situations where sample orderings are correct with respect to absolute time but are incorrect with respect to the dynamics of the biological process (Rice, 1997; Kim *et al.*, 1999). These two problems can be addressed through mixture modeling and dynamic transforms (e.g. Rice, 1997; Aach and Church, 2001).

When there are no easily identifiable phenotypic or genotypic markers, it may be impossible to explicitly time index samples. This is a fundamental problem in that without time-indices, we do not have even an approximate time-series. In this paper, we address the problem of *ab initio* reconstructing time ordering from data without explicit time information. We use microarray data as a set of high-dimensional measurements whose values can be used to track temporal ordering by assuming that the temporal changes in the transcriptome are relatively smooth and continuous (see Rifkin and Kim, 2002).

An example of how such data and algorithms for reconstructing temporal orderings might be profitably employed comes from the field of cancer research. Model organisms such as mice provide experimental access to cancer development within periods encompassing a few weeks where sampling all developmental stages is a realistic goal. Two key technology advances now make it possible to sample cancer development in a meaningful way: (a) laser capture microdissection (LCM, Bonner *et al.*, 1997) permits the isolation of homogeneous cell

*To whom correspondence should be addressed at present address. Department of Biology, Uhiversity of Pennsylvania, Philadelphia, PA, USA

populations comprising a few hundred epithelial cells, which represent a specific stage of the cancer process; (b) techniques for linear RNA amplification (Phillips and Eberwine, 1996) now enable the successful analysis of a few hundred laser capture microdissected cells by mRNA profiling experiments based on microarrays (e.g. Ohyama *et al.*, 2000; Luzzi *et al.*, 2001; Rubin, 2001; Sugiyama *et al.*, 2002; Mori *et al.*, 2002). However, the samples obtained from a mouse model of cancer do not represent a time series with a well-defined developmental order, because early cancer foci are: (1) very small with no external ontogenetic markers, (2) number in the hundreds within any tissue but are not synchronized. Thus, the only sensible approach to study temporal patterns of gene expression during cancer development is to perform separate microarray analysis observations in a large number of samples, followed by *ab initio* time indexing of multiple experiments. Other possible applications include cell lineage tracing in cell biology and developmental genetics (see Stern and Fraser, 2001 for review). It should be noted that biologists in practice often attempt to overcome the problems of temporal placement through experimental approaches (e.g. using tracer molecules) as the possibility of a computational solution is not widely appreciated.

## BACKGROUND

### Ordering and curve reconstruction

If one assumes that the biological process under study can be treated as a continuous function with respect to time, then the problem of ordering samples can be solved by reconstructing vector-valued functions of the form $\vec{f}(t) = [x_1(t), x_2(t), \ldots, x_d(t)]$ where each $d$-dimensional point on the function represents the state of a system at a particular point in time. A sample from this curve is a random vector $\vec{\psi}(s) = \vec{f}(s) + \vec{\delta}(s)$, of the curve $\vec{f}(t)$ sampled at time $s$ with some noise vector $\vec{\delta}(s)$ from some distribution (say, multivariate normal). More commonly, we may expect the noise vector to be time homogeneous, $\vec{\delta}(s) = \vec{\delta}$. It may also be appropriate to consider the sampled time, $s$, as a random variable. An observed data set consists of a finite number of sampled points from the curve $V = \{\vec{\psi}(s_0), \vec{\psi}(s_1), \ldots, \vec{\psi}(s_n)\}$ where we assume that the sampled time points, $s_i$, are unknown except possibly the start, $s_0$, and the end, $s_n$. The characteristics of the data are determined by the size of the time interval between successive sampling points (which we call time intensity), the number of replicate observations for each time point or a small time interval (multiplicity), and the size of the noise vector relative to the length of the arc joining successive time samples (relative noise).

The problem of estimating the geometry of $\vec{f}(t)$ from finite points has been referred to as the curve reconstruction problem. The techniques that have been applied in such contexts can be roughly categorized into two classes—polygonal reconstruction approaches and principal curves techniques.

*Polygonal reconstruction and principal curves* If $\vec{f}$ is a smooth, twice differentiable curve in $\boldsymbol{R}^d$, and $V$ is a finite sample of points on $\vec{f}$, then Amenta *et al.* (1998) defined a polygonal reconstruction of $\vec{f}$ from $V$ as a graph that connects every pair of samples that are adjacent on $\vec{f}$, and no others. Two points, $\vec{f}(t_0)$ and $\vec{f}(t_1)$ are defined as being adjacent if no point in $V$ exists that is a point on the arc $\{\vec{f}(s) : t_0 \leqslant s \leqslant t_1\}$. That is, adjacency, here and below, refers to adjacency in terms of the time parameter.

Common assumptions for polygonal reconstruction are that the samples are drawn from a smooth curve embedded in $\boldsymbol{R}^d$ dimensions, that the points are sampled without error, and that sampling intensity is sufficient to achieve reconstruction (Amenta *et al.*, 1998; Dey and Kumar, 1999; Figueiredo and Gomes, 1995). Recent work by Giesen (1999, 2000) and Althaus and Mehlhorn (2000) has demonstrated that the requirement of smoothness can be relaxed somewhat; requiring only that curves be 'benign semi-regular'. Giesen showed that, given sufficiently dense sampling, the traveling salesman path (TSP) provides the correct reconstruction for curves in $\boldsymbol{R}^d$. A similar result holds for the minimum spanning tree (MST; Figueiredo and Gomes, 1995).

While the TSP or MST methods work well when applied to sample data sets that are observed without error, the presence of noise introduces undesirable results to the polygonal reconstruction algorithms. For example, even when noise is distributed uniformly around $\vec{f}$ then polygonal reconstructions tend to zigzag back and forth through the data cloud.

The second major class of algorithms that have been applied to the curve reconstruction problem is comprised of variants of the principal curve algorithm (Hastie and Stuetzle, 1989). Hastie and Stuetzle defined a principal curve as 'self-consistent' smooth curve that passes through the middle of a $d$-dimensional data cloud. Principal curves and related techniques explicitly assume that the observations in the sample represent points sampled with noise (Hastie and Stuetzle, 1989; Kégl *et al.*, 2000) with the idea that if sufficient points are available a kind of a 'mean' curve may be estimated.

Both polygonal reconstruction methods and principal curve methods attempt to recover the full curve $\vec{f}(t)$ within the sampled interval. Thus, both classes of algorithms require sufficiently dense temporal sampling (high time intensity) such that each segment of the curve can be linearly approximated (see below). In addition, principal curve type algorithms require high multiplicity

to reliably estimate the self-consistent mean trajectory. Microarray expression data is typically sparsely sampled with relatively high noise making it difficult to hope to recover the details of a trajectory. However, as noted above, just the temporal sequencing of the unknown time order can provide important biological insights. In this paper, we concentrate on this simpler problem.

### The ordering problem in a time series

We assume that there is a time-parameterized vector valued function, $\vec{f}(t)$, which is sampled at a finite number of *unknown* time points with error. We also allow for multiplicity in the sampling such that the collection may contain multiple samples of the same time point. Let $V = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n\}$ be the collection of observed samples. For each observation, $\vec{f}_i$, let $s_i$ be the unknown time index of the $i$th observation. Then we have:

DEFINITION 1. A permutation $\pi$ of the index set $\{1, 2, \ldots, n\}$ is a temporal ordering of the points, $V = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n\}$, if $\pi(i) \leqslant \pi(j) \Rightarrow s_i \leqslant s_j$ for all $i, j$ in the index set.

The ordering problem is to find the temporal ordering permutation, $\pi$, given the data $V = \{\vec{f}_1, \vec{f}_2, \ldots, \vec{f}_n\}$. The ordering problem arises in other contexts, for example when trying to build archeological chronologies from collections of artifacts (Kendall *et al.*, 1970; Buck and Sahu, 2000).

The time series $\vec{f}(t)$ is a one-dimensional curve embedded in the space of our measurements; e.g. gene expression values. Let the geometry of the measurements be determined by some quadratic form determining the notion of a distance, say the standard Euclidean inner product. Given two points, $\vec{f}_a$ and $\vec{f}_b$, if we were able to measure the arc length (= geodesic distance), $l_{a,b} = \int_a^b \sqrt{1 + [f'(t)]^2}$, of the curve between $\vec{f}_a$ and $\vec{f}_b$ then the ordering could be found by simple sorting of the arc length distances between the points. However, the only information we have is the embedded geometry and if the curve has significant curvature it will be difficult to approximate the arc length from the embedded geometry except for short segments. For a complete reconstruction of the original curve, the segments need to be quite short to reliably approximate the geodesic path on the curve. However, for the temporal ordering problem, we do not need to know the geodesic path distance. We only need the embedding distance to be monotonically related to the geodesic distance which motivates the following 'rules':

- If two points, $\vec{f}_a$ and $\vec{f}_b$, are relatively similar (compared to other pairs of points), then compute the standard 'pairwise' dissimilarity measure of the embedded geometry.
- If $\vec{f}_a$ and $\vec{f}_b$ are relatively dissimilar, then compute a 'pathwise' dissimilarity measure – the sum of a series of short pairwise 'hops' between points intervening $\vec{f}_a$ and $\vec{f}_b$.

Tenenbaum *et al.* (2000) similary argued that the use of geodesic manifold distances between pairs of data points preserves the intrinsic geometry of the data. Figueiredo and Gomes (1995) suggested that when noise is present, rather than finding a path, one should find a tree graph and consider the diameter path (longest path) of the tree for the curve reconstruction. Unfortunately, this ignores the fact that when the curve has high curvature relative to sampling intensity, segments of the actual path may be present in the branches off the diameter path. Here we modify Figueiredo and Gomes's idea by using the structure of the tree graph to delineate the noise versus path components.

## ALGORITHM AND IMPLEMENTATION

### Ordering observations using minimum spanning trees

Our new algorithm can be briefly described as follows (see Fig. 1):

(1) Find the minimum spanning tree, $G_{\text{mst}} = \{V, E_{\text{mst}}\}$ of the weighted graph $G$, where $G$ is a complete graph whose vertices represent the sample observations and whose edge weights are pairwise dissimilarities (distances in the embedding geometry). If $G_{\text{mst}}$ is a path (i.e. it has no branches), take this to be the best estimate of the ordering.

(2) If $G_{\text{mst}}$ is not a path, assess the diameter path noise ratio, branch distribution, and sampling intensity as defined below. If the sampling appears to be relatively intense and the diameter path branch distribution appears to be relatively uniform, then the diameter path gives an estimate of the ordering. Samples contained in branches off the diameter path are assigned the same ordering index as the diameter path element to which they connect.

(3) If the diameter path sampling intensity ratio is large and the diameter path branch distribution appears to be non-uniform with a few long branches coming off the diameter path two additional steps are taken. First a data structure called PQ-tree is used to summarize all the uncertainties of path variations. Next, a secondary criterion of 'shortest path ordering' (motivated by the TSP algorithm) is applied to the variations of the paths. Each of the $X$ shortest paths that are consistent with the PQ-tree are reported, where $X$ is a user defined value.

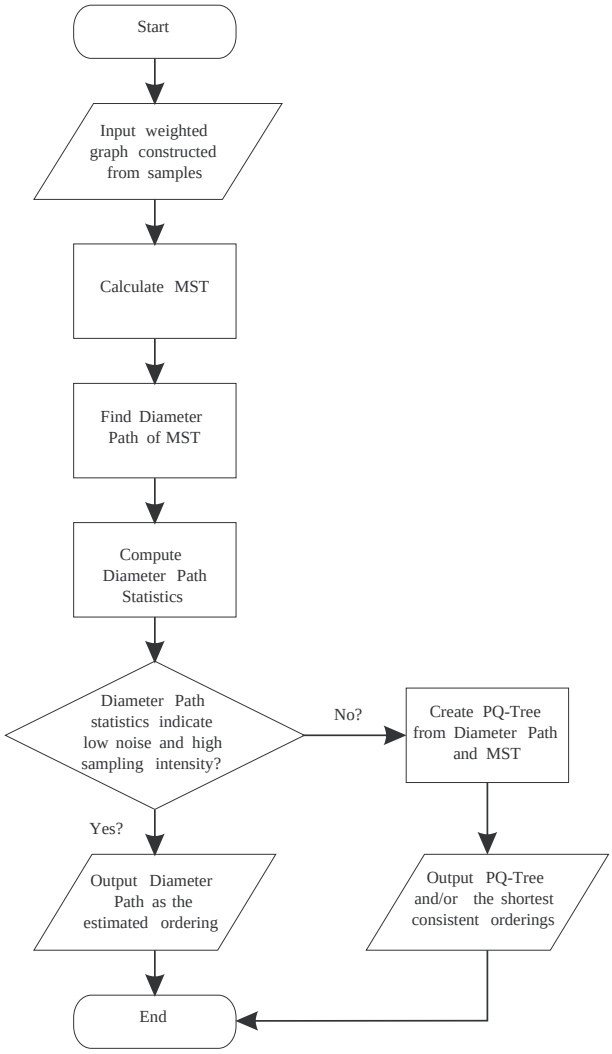


**Fig. 1.** Flowchart outlining the algorithm for ordering sample observations described in the text.

## Assessing noise and sampling intensity

The diameter path of a tree graph is the longest path in the tree (Fig. 2c). A noisy sample will have numerous edges dangling from it. We will refer to these dangling non-diameter path nodes and edges as the *branches* of the diameter path. One useful measure of the relative noise of the sample is the ratio of the number of points on branches to the total number of points. We call this the *diameter path noise ratio*. For example, a diameter path noise ratio of 0.05 indicates that only 5% of the sample points do not fall on the diameter path, which suggests that there is high signal-to-noise ratio in the data.

The distribution of points on branches can also be used as a heuristic tool to evaluate the quality of the diameter path as an estimator. If the branch points are truly noise, then the size of branches coming off the diameter path should be relatively uniform along the extent of the diameter path. However, if the distribution is non-uniform, relatively long branches may represent signal (part of the path) rather than noise. We refer to this distribution as the *diameter path branch distribution*.

Finally, the sampling intensity can be assessed by calculating the ratio of the average segment length to the total length of the diameter path. For a dense sample, this ratio should be small, while in a sparse sample this ratio will be large. We will use the term *diameter path sampling intensity ratio* to refer to this quantity. The critical cutoff of the diameter path sampling ratio necessary to achieve a suitable reconstruction will depend on the curvature of the underlying function $\vec{f}(t)$, but experimenting with a variety of artificial datasets suggests that a ratio less than 0.03 usually results in adequate reconstructions (data not shown).

## Representing temporal orderings using PQ-trees

The PQ-tree (Booth and Lueker, 1976) is a data structure that can represent reconstruction uncertainty, while at the same time retaining information about those parts of the reconstruction which seem to be well supported. The PQ-tree data structure has been exploited in a variety of applications from archeology (chronology reconstruction; Atkins *et al.*, 1998) to molecular biology (DNA mapping and sequence assembly; Greenberg and Istrail, 1995; Wilson *et al.*, 1997).

A PQ-tree of a set $U = \{u_1, u_2, \ldots, u_n\}$ is a rooted, ordered tree whose leaves are elements of $U$ and whose internal nodes are distinguished as either P-nodes or Q-nodes (Atkins *et al.*, 1998). P-nodes have at least two children; the children of a P-node can be arbitrarily reordered. Q-nodes have at least three children and permit only the rigid reversal of their children. One can think of P-nodes as having children on a string, while Q-nodes have children on a rigid ruler (see Fig. 3c). The leaves of the tree form an ordered set; for example, a tree with a single P-node represents the equivalence class of all permutations, while a tree with a single Q-node represents the left–right ordering of the leaves and their inverse. A tree with a mixed P- and Q-node structure represents the equivalence class of a constrained permutation where the exact structure of the tree determines the constraints (Atkins *et al.*, 1998).

An iterative algorithm for generating a PQ-tree whose equivalence class is consistent with the constraints imposed by a minimum spanning tree and its diameter path is:

(1) For a set of points, calculate the MST and diameter path.

(2) Designate an empty Q-node, $Q_{\text{main}}$.

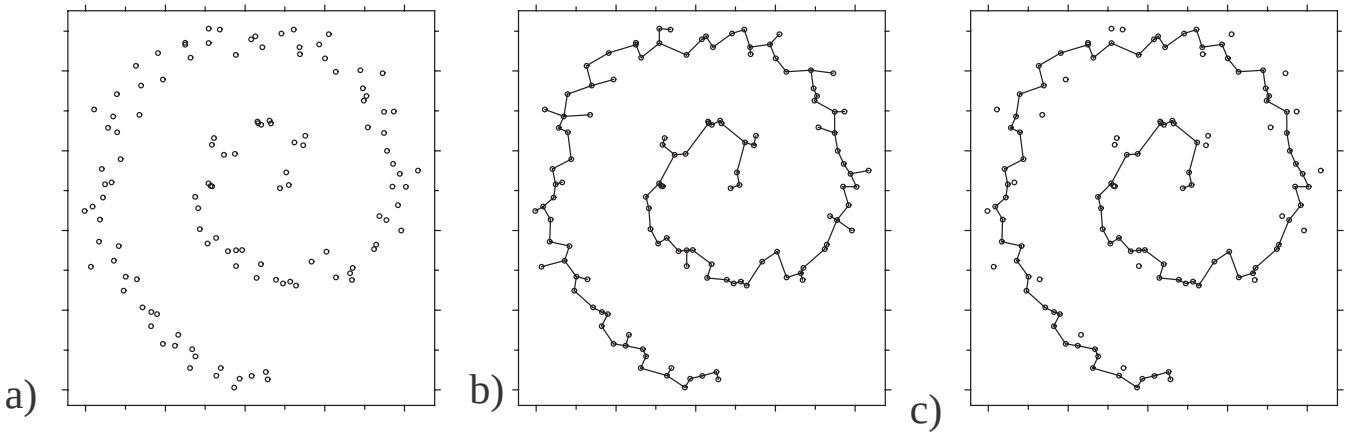


**Fig. 2.** Generated 'jelly roll' dataset. (**a**) Original data; (**b**) minimum spanning tree; (**c**) diameter path of the minimum spanning tree. We use the diameter path as the basis for estimating an ordering for this dataset. See text for further explanation.

(3) Find the indecisive backbone of the diameter path defined as:
   (a) A vertex on the diameter path is called indecisive if its degree is greater than 2, otherwise it is called decisive.
   (b) The indecisive backbone is the longest continuous subset of the diameter path for which both the first and last vertex is indecisive. For example, if the diameter path has the following structure $\{i_1, d_2, d_3, i_4, d_5\}$, where the $i$'s are indecisive and the $d$'s are decisive, then the indecisive backbone is defined as vertices 1 to 4.

(4) Moving in order (from left to right) along the indecisive backbone, attach each decisive vertex, $V_i$, as a leaf-node to $Q_{\text{main}}$. For each indecisive vertex, $V_j$, attach a P-node, $P_j$ to $Q_{\text{main}}$ and make $V_j$ a leaf-node of $P_j$.

(5) For the points along the branch associated with each indecisive vertex, $V_j$ (i.e. the points along the twig off the main diameter path of the whole MST), find the diameter path of that branch and repeat the algorithm from step 1 and attach the resulting PQ-tree as a subtree of the P-node $P_j$. Recursively apply the algorithm to each branch emanating from indecisive vertices.

The resulting PQ-tree reflects the structure of the MST and diameter path (see Fig. 3b,c). The leaf nodes along the primary Q-node (the 'backbone') are the decisive vertices of the diameter path. The P-nodes, and their subtrees, represent the indecisive vertices and their associated diameter path branches. The orderings generated by the equivalence class of the PQ-tree are those permutations that are consistent with the diameter path.

## RESULTS

### Artificial dataset—the jelly roll

An artificial two-dimensional dataset of 150 points was generated with a common spiraling curve (jelly-roll; Fig. 2a). This dataset is fairly noisy and has relatively high time intensity and multiplicity. This data represents a difficult problem in that the local curvature is high and the global structure of the curve (the spiral) interferes with the estimation (many estimation procedures produce a curve that jumps back and forth between the different levels of the spiral).

The MST for this dataset is shown in Figure 2b, and the diameter path for the MST is shown in Figure 2c. The distribution of non-diameter path points (branches) is relatively evenly distributed along the extent of the diameter path and the diameter path noise ratio for this sample is $29/149 = 0.19$. About one fifth of the points in the sample are not included in the diameter path and the diameter path noise ratio confirms our visual assessment that the sample is noisy.

Finally we estimate sampling intensity from the diameter path. The diameter path has a total length of 95.08 units, and the average edge length in the diameter path is 1.07 units. Therefore the diameter path sampling intensity ratio is 0.011, confirming that sampling is relatively dense. Given the above considerations we use the diameter path of the MST (Fig. 2c) to estimate the ordering for this dataset. Each point off of the diameter path is assigned to the same ordering index as its corresponding root node on the diameter path.

### Biological datasets—*Caulobacter* and yeast expression data

Caulobacter *microarray study* Laub *et al.* (2000) de-

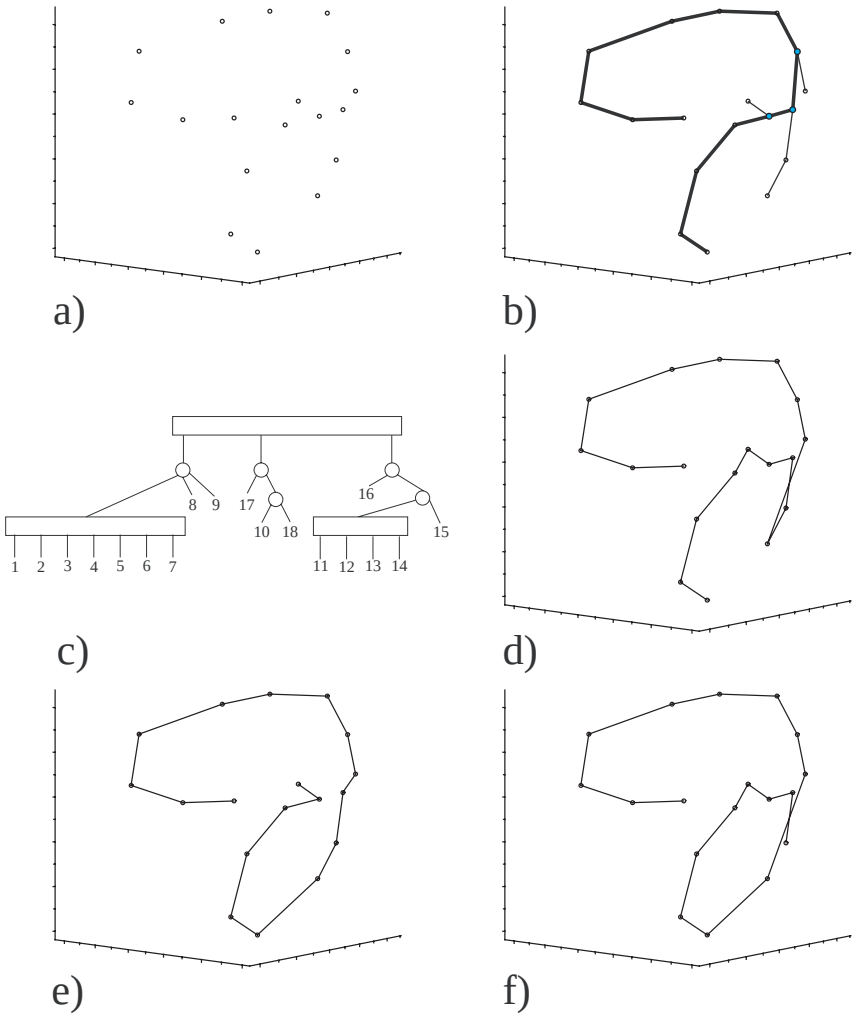


**Fig. 3.** Data from *Sacchoromyces* microarray dataset. (**a**) Ordination of sample points in the space of the three largest principal coordinates; (**b**) minimum spanning tree for this data; (**c**) PQ-tree based on the MST and diameterpath; (**d**) PQ-tree estimated ordering 1; (**e**) PQ-tree estimated ordering 2; (**f**) known ordering and path. See text for further explanation.

scribe a microarray-based analysis of gene transcription during a single cell cycle of the bacterium *Caulobacter crescentus* (11 samples at 15 minute intervals). We log transformed and mean-centered the published data on expression levels of 2966 predicted open reading frames (Laub *et al.*, 2000). We reduced the dataset to the 1594 ORFs for which expression measurements were available at all time points. We also obtained subsets containing the 500 most variable and 300 most variable genes. The estimated ordering was the same regardless of the sample used and results are described based on all 1594 ORFs.

With only 11 samples, the dataset is expected to be rather sparse. The diameter path of the MST for the sample elements has only a single branch with a single vertex. Though sparse, the diameter path appears to provide a relatively low-noise reconstruction and we take the points on the diameter path as the best estimate of the ordering. The known ordering of data points is given by the permutation [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. The reconstructed ordering gives the permutation [1, {2,3}, 4, 5, 6, 7, 8, 9, 10, 11] assuming either the start or end point is known. The reconstructed ordering therefore provides a near perfect estimate of the known temporal ordering.

*Yeast microarray study* Spellman *et al.* (1998) describe a time-series gene expression dataset (18 points at 7 minute intervals) for the yeast *Saccharomyces cerevisiae* synchronized by treatment with the mating pheromone $\alpha$-factor. The original dataset based on 6177 ORFs was reduced to a set of 5541 genes as described in Rifkin *et al.* (2000) and the data were log-transformed and normalized by the

**Table 1.** The ten shortest reconstructed orderings for the yeast microarray data set, based on the PQ-tree for the MST/diameter path. The known ordering of samples is [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] with a path length of 211.8

| Reconstructed ordering | Length of ordering |
|---|---|
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 18, 17, 16, 15, 14, 13, 12, 11] | 208.1 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18, 10, 11, 12, 13, 14, 16, 15] | 209.5 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18, 10, 11, 12, 13, 14, 15, 16] | 209.6 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 18, 17, 15, 16, 14, 13, 12, 11] | 209.8 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 18, 10, 17, 16, 15, 14, 13, 12, 11] | 210.3 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18, 10, 16, 15, 14, 13, 12, 11] | 210.9 |
| [1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 10, 18, 16, 15, 14, 13, 12, 11] | 210.9 |
| [1, 2, 3, 4, 5, 6, 7, 9, 8, 17, 18, 10, 11, 12, 13, 14, 16, 15] | 211.0 |
| [1, 2, 3, 4, 5, 6, 7, 9, 8, 17, 18, 10, 11, 12, 13, 14, 15, 16] | 211.1 |
| [1, 2, 3, 4, 5, 6, 7, 9, 8, 10, 18, 17, 16, 15, 14, 13, 12, 11] | 211.8 |

centroid. This ordering for this dataset is expected to be somewhat more difficult to reconstruct than the *Caulobacter* dataset because the data cover two cell-cycles and the sampling is sparser.

Of the approximately 5500 genes in the dataset, a great number do not exhibit appreciable variation in expression levels. The 500 genes exhibiting the most sample variation were used to construct a distance matrix among the 18 sample points. An ordination of these data in the space represented by the 3 largest principal coordinates of the distance matrix is depicted in Figure 3a. With only 18 samples, we expect that the sampling is rather sparse. The minimum spanning tree for this dataset, depicted in the space of a 3-D ordination of these points is shown in Figure 3b. The diameter path is highlighted in bold. The diameter path sampling ratio confirms that sampling is relatively sparse (intensity ratio of 0.077), and the distribution of branch lengths is non-uniform.

We constructed a PQ-tree based on the minimum spanning tree and diameter path; the PQ-tree is shown in Figure 3c. We then calculated the lengths of each of the orderings consistent with the PQ-tree. The ten shortest orderings and their path lengths are given in Table 1. The paths associated with two of these orderings are shown in Figures 3d and e. The known ordering is [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] (Figure 3f) and has a path length of 211.82 (slightly longer than the 10th shortest path). The orderings consistent with the PQ-tree share the major geometric features of the known true path. The main difficulty is where the two cycles meet and the path reconstruction inverts the true path. Denser sampling, especially near the conjunction of the two cycles, would lead to improved reconstructions.

### Simulation results—the effects of noise and sampling intensity

We carried out a computer simulation in which we generated samples using a simple generating function. The generating function was a three-dimensional spiral, where $x = \cos(t \cdot \theta)$, $y = \sin(t \cdot \theta)$, and $z = 4t/4\pi$. In our simulation the generating function spirals twice ($t : 0 \rightarrow 4\pi$) with a total translation along the $z$-axis of 4 units. To assess the effects of noise and sampling intensity we generated datasets in which the generating function was sampled over a range of step sizes ($\pi/64 \rightarrow \pi/3$ at intervals of $\pi/128$; small step sizes correspond to high sampling intensities) and subject to a range of multivariate random noise (standard deviation of 0 to 0.5 units in each of the $x$-, $y$-, and $z$-directions). For each pair of (step size, noise) parameters five replicate sample data sets were generated. For each sample data set we estimated the temporal ordering using the diameter path and calculated Spearman's rank-correlation between the estimated ordering and the known true ordering. A contour plot showing the average rank-correlation over the sampling parameter domain is shown in Figure 4. At low noise-levels the diameter path reconstruction performs well over a wide range of sampling intensities (step size). However as noise increases, the diameter path estimate performs best at high and intermediate sampling intensities (small to intermediate step sizes). Figure 4 also suggests that in the presence of moderate amounts of noise an intermediate sampling intensity may actually lead to better reconstructions than high sampling intensities. In the presence of noisy data, such as might be expected from biological phenomena, increased sampling effort may have little benefit for ordering purposes once a sufficient level of sampling intensity is achieved (particular parameter values are, of course, case dependent).

## DISCUSSION

The approach we describe for ordering samples is well suited to handle datasets that utilize a variety of dissimilarity measures and sampling conditions. The demonstration that our heuristic algorithm works well for both artificial and experimental datasets suggests there is significant practical value in utilizing this approach to study dynamic biological processes when it is not possible to time index samples.

Minimum spanning trees provide a natural geometric characterization for samples. This characterization is free of *a priori* distributional assumptions, and can be applied to datasets using a variety of dissimilarity measures. Modifications of the MST or the MST itself provide a useful basis for estimating orderings and reconstructing curves, and additionally can be used as a tool to evaluate

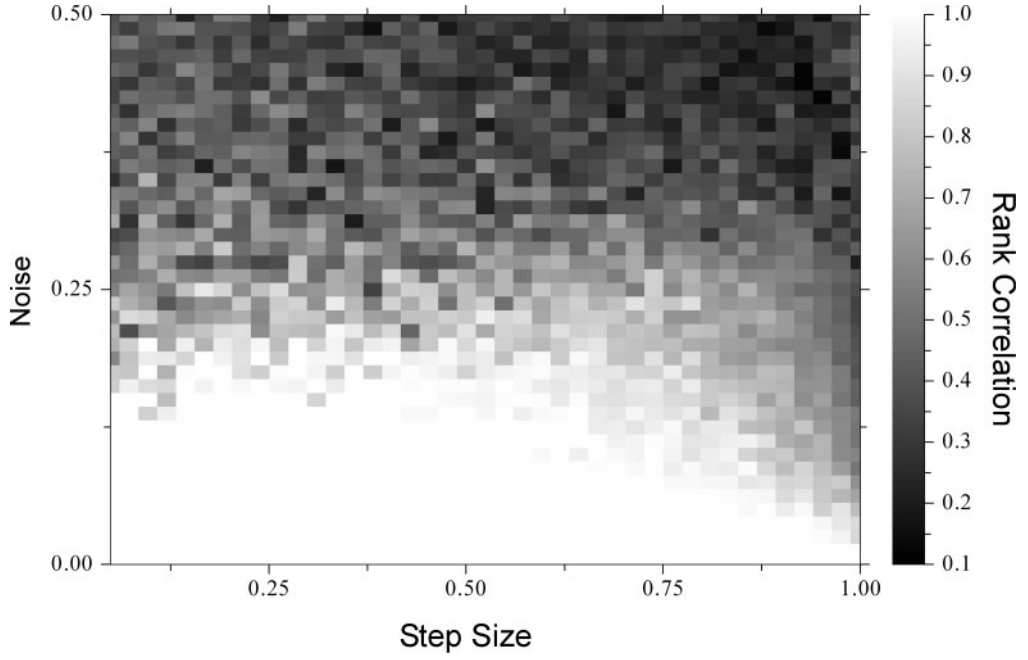


**Fig. 4.** Simulation to study the effects of sampling intensity and sample noise on order reconstruction. The gray-scale represents of rank-correlation between estimated diameter path ordering and the true ordering over the domain of simulation parameters. See text for further explanation.

distributional properties such as noise and sampling intensity.

Our approach can also be applied to data from multiple sources such as simultaneous information on phenotype, the transcriptome, and the proteome. Data such as these will facilitate the construction of well-characterized and informative time series. Constructing such time series is a necessary first step towards gaining a better understanding of the dynamical behavior of biological systems.

## ACKNOWLEDGEMENTS

This research was funded in part by an NIH training grant through the Yale Center for Medical Informatics and by an NSF Minority Postdoctoral Research Fellowship to P. Magwene. J. Kim acknowledges the support of an NIH pre-NPEBC to Yale University.

## REFERENCES

Aach,J. and Church,G.M. (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Althaus,E. and Mehlhorn,K. (2000) TSP-based curve reconstruction in polynomial time. *Symposium on Discrete Algorithms*, 686–695.

Amenta,N., Bern,M. and Eppstein,D. (1998) The crust and the Beta-skeleton: combinatorial curve reconstruction. *Graphical Models and Image Processing*, **60**, 125–135.

Atkins,J.E., Boman,E.G. and Hendrickson,B. (1998) A spectral algorithm for seriation and the consecutive ones problem. *Siam J. Comput.*, **28**, 297–310.

Bonner,R.F., Emmert-Buck,M., Cole,K., Pohida,T., Chuaqui,R., Goldstein,S. and Liotta,L.A. (1997) Laser capture microdissection: molecular analysis of tissue. *Science*, **278**, 1481–1483.

Booth,K.S. and Lueker,G.S. (1976) Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. Sys. Sci.*, **13**, 335–379.

Buck,C.E. and Sahu,S.J. (2000) Bayesian models for relative archaeological chronology building. *Applied Statistics*, **49**, 423–440.

Cook,W. and Rohe,A. (1999) Computing minimum-weight perfect matchings. *INFORMS J. Comput.*, **11**.

Dey,T.K. and Kumar,P. (1999) A simple provable algorithm for curve reconstruction. In *Proceedings of the 10th Symposium on Discrete Algorithms*.

Figueiredo,L.H.d. and Gomes,J. (1995) Computational morphology of curves. *The Visual Computer*, **11**, 105–112.

Giesen,J. (1999) *Curve reconstruction in arbitrary dimension and the traveling salesman problem*, Discrete Geometry for Computer Imagery, Springer, pp. 164–176.

Giesen,J. (2000) Curve reconstruction, the traveling salesman problem, and Menger's theorem on length. *Discrete and Computational Geometry*, **24**, 577–603.

Greenberg,D.S. and Istrail,S.C. (1995) Physical mapping by STS hybridization: Algorithmic strategies and the challenge of software evaluation. *J. Comput. Biol.*, **2**, 219–274.

Hastie,T. and Stuetzle,W. (1989) Principal Curves. *J. Am. Stat. Assoc.*, **84**, 502–516.

Kendall,D.G. (1970) A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London*, **A269**, 125–135.

Kégl,B., Krzyzak,A., Linder,T. and Zeger,K. (2000) Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence*, **22**, 281–297.Kendalletal1970,BuckSahu2000 Bayesian models for relative archaeological chronology building

Kim,J., Kerr,J. and Min,G.S. (1999) Molecular heterochrony in the early development of *Drosophila*. *Proc. Natl Acad. Sci. USA*, **97**, 212–216.

Laub,M.T., McAdams,H.H., Feldblyum,T., Fraser,C.M. and Shapiro,L. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.

Luzzi,V., Holtschlag,V. and Watson,M.A. (2001) Expression profiling of ductal carcinoma in situ by laser capture microdissection and high-density oligonucleotide arrays. *Am. J. Pathol.*, **158**, 2005–2010.

Mori,M., Mimori,K., Yoshikawa,Y., Shibuta,K., Utsunomiya,T., Sadanaga,N., Tanaka,F., Matsuyama,A., Inoue,H. and Sugimachi,K. (2002) Analysis of the gene-expression profile regarding the progression of human gastric carcinoma. *Surgery*, **131**, S39–47.

Ohyama,H., Zhang,X., Kohno,Y., Alevizos,L., Posner,M., Wong,D.T. and Todd,R. (2000) Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization. *Biotechniques*, **29**, 530–536.

Phillips,J. and Eberwine,J.H. (1996) Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells. *Methods*, **10**, 283–288.

Rifkin,S.A., Atteson,K. and Kim,J. (2000) Structural analysis of microarray data using singular value decomposition. *Functional and Integrative Genomics*, **1**, 174–185.

Rifkin,S.A. and Kim,J. (2002) Geometry of gene expression dynamics. *Bioinformatics*, **18**, 1176–1183.

Rice,S.H. (1997) The analysis of ontogenetic trajectories: when a change in size or shape is not heterochrony. *Proc. Natl Acad. Sci. USA*, **94**, 907–912.

Rubin,M.A. (2001) Use of laser capture microdissection, cDNA microarrays, and tissue microarrays in advancing our understanding of prostate cancer. *J. Pathol.*, **195**, 80–86.

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

Stern,C.D. and Fraser,S.E. (2001) Tracing the lineage of tracing cell lineages. *Nat. Cell. Biol.*, **3**, E216–E218.

Sugiyama,Y., Sugiyama,K., Hirai,Y., Akiyama,F. and Hasumi,K. (2002) Microdissection is essential for gene expression profiling of clinically resected cancer tissues. *Am. J. Clin Pathol.*, **117**, 109–116.

Tenenbaum,J.B., de Silva,V. and Langford,J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Wilson,D.B., Greenberg,D.S. and Phillips,C.A. (1997) Beyond islands: runs in clone-probe matrices. *Proceedings of the First Annual International Conference on Computational Molecular Biology*. ACM Press, New York, pp. 320–329.