Reconstructing the Temporal Progression of Biological Data Using Cluster Spanning Trees

Ryan Eshleman and Rahul Singh*

Abstract—Identifying the temporal progression of a set of biological samples is crucial for comprehending the dynamics of the underlying molecular interactions. It is often also a basic step in data denoising and synchronization. Finally, identifying the progression order is crucial for problems like cell lineage identification, disease progression, tumor classification, and epidemiology and thus impacts the spectrum of disciplines spanning basic biology, drug discovery, and public health. Current methods that attempt solving this problem, face difficulty when it is necessary to factor-in complex relationships within the data, such as grouping, partial ordering or bifurcating or multifurcating progressions. We propose the notion of cluster spanning trees (CST) that can model both linear as well as the aforementioned complex progression relationships in temporally evolving data. Through a number of experimental investigations involving synthetic data sets as well as data sets from the cell cycle, cellular differentiation, phenotypic screening, and genetic variation, we show that the proposed CST approach outperforms existing methods in reconstructing the temporal progression of the data.

Index Terms—Time series reconstruction, spanning trees, cell cycle modeling, phenotypic screening.

I. INTRODUCTION

BIOCHEMICAL processes can be thought of as spatial-temporal progressions. A small class of generating processes may characterize such progressions. For example, linear or polynomial processes (cell growth [1]), cyclical functions (cell cycle [2]), and branching (bifurcating or multifurcating) processes (cancer progression [3], and epidemic spread [4]). If the system under study can be sufficiently synchronized, as is the case with cell synchrony methods [5], characterizing the underlying progression becomes relatively straightforward. Often however, this is not possible and the temporal order has to be reconstructed from a sampling of the process. We focus on this latter case and note that its complexity arises due to both epistemic and intrinsic factors such as the unknown nature of the mechanisms of action, their (putative) non-linearity and anisotropy, phase shifts, and rate heterogeneity, as well as extrinsic factors such as undersampling, measurement errors and noise.

If we think of a biological process as a series of states evolving with respect to time, the problem of constructing the temporal ordering requires specifying the multidimensional generating function \( f(t) = [x_1(t), x_2(t), \ldots, x_d(t)] \), where \( x_i(t) \) denotes the value of dimension \( i \) at time \( t \) and the output \( f(t) \) is a point in \( d \) dimensions representing the state of the process at time \( t \). For our problem, this function has to be reconstructed from a sampling of the data \( S = \{s_1, \ldots, s_n\} \), where \( s_i = f(i) + \epsilon \) with \( \epsilon \) denoting the noise.

Graph-theoretic representation of the biological data provides a powerful formalism for this problem where the data is represented by a graph \( G_c = (V, E) \) with each datum corresponding to a vertex in \( V \) and the edges in \( E \) connecting the vertices based on some criterion. Within this framework, Minimum Spanning Trees (MST) constitute a powerful representation for progression reconstruction [3], [6], [7]. However, a MST cannot account for groupings in the data representing sub-processes or describe relations between such groups. Furthermore, the topology of a tree can be quite sensitive to the edge selection methodology. Consider for example, three methods to reconstruct the progression of gene expression during the cell cycle (Figure 1): a MST-based method [6], the Sample Progression Discovery (SPD) method [7] and the proposed Cluster Spanning Tree (CST) approach [8]. In this example 20 proteins associated with different phases of the cell cycle are chosen from the cell cycle cDNA expression dataset [2]. Each phase in the cycle can be understood as a subprocess and a crucial test of a method’s efficacy lies in retaining these groupings in the reconstruction. The MST-based method and SPD are found to accurately group proteins from the G1/S, S, and G2/M phases but introduce errors for the G2 phase. Only the proposed CST method correctly groups the G2 phase proteins as well as proteins from the other phases. Moreover, the CST is the only method that arranges the proteins in the proper order that reflects the stages of the cell cycle, namely: G1/S, S, G2, G2/M. A detailed analysis is presented in the results section.

II. BACKGROUND

Given a sampling \( S \) of size \( n \) of \( f \), one way of reconstructing the underlying generating function is through polygonal approximation. Polygonal reconstruction [9] builds a
data point. Consequently, the resulting reconstruction is highly constrained of a TS-path is that it must pass through every point represented by a sequential traversal of the path. Since, distance. The reconstruction induces an explicit ordering of the a traveling salesman (TS)-path by minimizing the total edge distance. The reconstruction induces an explicit ordering of the

Unfortunately, this method was designed to obtain shape representations and its extension to our problem is non-trivial. Neither polygonal reconstruction nor principal curves can be used directly to model branching processes or processes which self-intersect. This problem was solved in the context of shape skeletonization in [13]. This method tessellated the curves, representing different parts of a shape were connected no assumptions are made about the probabilistic nature of the on the sampled data in a non-probabilistic setting. In particular, methods as well as the proposed approach, a tree structure is used to empirically reconstruct the underlying process based on the sampled data in a non-probabilistic setting. In particular, no assumptions are made about the probabilistic nature of the branching process, such as modeling it as a Markov process.

III. Method

Our method constructs and traverses a hierarchical tree which represents the empirical relations in the data and then iteratively adds edges between nodes or groupings thereof. A hierarchical binary tree $G_b = (V_b, E_b)$ in our formulation contains $2n$-1 vertices, where $n$ is the number of data points being clustered. In the tree, the $n$ leaf vertices represent the data and the $n$-1 internal vertices encode the possible (hierarchical) structures in it with each internal node representing the union of its descendants. The root represents a set of size $n$. Each internal vertex $v_i$ has two children, $c_{i1}$ and $c_{i2}$ each containing disjoint sets where $v_1 = \{c_{11} \cup c_{12}\}$. An example illustrating these concepts is presented in Figure 2. In the following, depending on the context, we shall use $v_i$ to refer to both vertices on the tree as well as the clustered data points at that vertex.

Beginning with $G_b$ and a graph of disconnected vertices $G_{CST} = (V_{CST}, E_{CST})$, where $V_{CST}$ is the set of original $n$ data points and $E_{CST}$ is the empty set, a CST is constructed as...
A. Hierarchical Data Clustering

There are a number of established hierarchical clustering techniques that can be utilized to perform the initial data clustering. Because the quality of the initial clustering is key to the reconstruction and data with different characteristics lend themselves to different clustering algorithms (for example, single linkage clustering is preferred over UPGMA for non-convex clusters) we examine several methods. Details on these methods can be found in Table 1. In addition to the well documented methods, we introduce a approach based on the Hausdorff Distance between sets [21], which we describe in the following. All of these methods induce a hierarchical structure on the data that can be used to obtain a hierarchical clustering of the data.

The first six methods in Table 1 describe the common similarity measures used to perform hierarchical clustering. Another approach for determining the similarity of two sets involves computing the extent to which each element of one set is similar to some element of the other set. This notion is captured by the Hausdorff distance [21]. Let \( A = \{a_1, a_2, \ldots , a_k\} \) and \( B = \{b_1, b_2, \ldots , b_l\} \). The directed Hausdorff distance between sets \( A \) and \( B \), \( h(A, B) \), is defined as [22]:

\[
h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||
\]

That is, to compute \( h(A, B) \), we first find the point \( a_i \in A \), which is farthest from any point in \( B \) and then identify the point \( b_j \in B \), that is the closest to it. The distance between the pair of points \( (a_i, b_j) \) defines the value of \( h(A, B) \). In general, the directed Hausdorff distance is non-symmetric. Given the directed Hausdorff distances \( h(A, B) \) and \( h(B,A) \), the Hausdorff distance between \( A \) and \( B \) is defined as: \( H(A, B) = \max(h(A, B), h(B, A)) \). For a given Hausdorff distance between sets \( A \) and \( B \), it follows from the definition that every point in \( A \) will be within that distance of some point in \( B \) and vice-versa. This property putatively allows temporally relating two groups of biological measurements using the relationship between a small number of data points, allowing us thereby to overcome incomplete or unavailable temporal relationship-information across the groups. At the same time, it should be noted that the Hausdorff distance between bounded sets depends on their suprema or infima and ignores the distribution of the other points in the set.

B. Cluster Merging

The second algorithmic component of our method is the strategy used to connect subsets at each bifurcation of the tree. This involves choosing a distance function to minimize, so as to relate two groups of data. Given an arbitrary norm \( ||.|| \), the distance between two sets \( A \) and \( B \), \( d(A, B) \) is defined as \( d(A, B) = \inf \{ ||a_i - b_j|| : a \in A, b \in B \} \). In this equation, if the sets \( A \) and \( B \) do not intersect (i.e. \( A \cap B = \{\} \)), then \( d(A, B) \) is positive and non-zero. Since we cluster the data hierarchically, as described in Section 3 A., the sets

<table>
<thead>
<tr>
<th>Method</th>
<th>Formulation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted Average (UPGMA) [18]</td>
<td>( d(A, B) = \frac{1}{</td>
<td>A</td>
</tr>
<tr>
<td>Weighted Average (WPGMA) [18]</td>
<td>( d(A, B) = \frac{(d(C, B) + d(D, B))}{2} )</td>
<td>Recursive definition where cluster ( A ) was created by combining ( C ) and ( D ).</td>
</tr>
<tr>
<td>Complete [18]</td>
<td>( \max(\text{dist}(A, B)) )</td>
<td>Maximum distance between two points in clusters</td>
</tr>
<tr>
<td>Centroid [19]</td>
<td>( d(A, B) = \text{dist}(\bar{A}, \bar{B}) )</td>
<td>( \bar{A} ) is the centroid of cluster ( A )</td>
</tr>
<tr>
<td>Median [19]</td>
<td>( d(\bar{A}, \bar{B}) = \frac{1}{2}(C + D) )</td>
<td>Where ( \bar{A} ) is the recursively defined centroid and cluster ( A ) was created by combining clusters ( C ) and ( D ).</td>
</tr>
<tr>
<td>Incremental sum of squares (Ward) [20]</td>
<td>( d(A, B) = \sqrt{\frac{2</td>
<td>A</td>
</tr>
<tr>
<td>Clustering with the Hausdorff Distance [21]</td>
<td>( H(A, B) = \max(h(A, B), h(B, A)) )</td>
<td>Measure of set proximity, see text.</td>
</tr>
</tbody>
</table>
being compared are finite and strictly non-intersecting. In the following, we describe a number of measures that can be used to facilitate cluster merging. A note on notation, for the remainder of the section, we use \( c_{ij} \) to refer to sets, because they are child nodes at a bifurcation, and \( d \) denotes the dimensionality of the data.

The first vertex merging strategy is the nearest neighbor approach. An edge is drawn from a point in \( c_{i1} \) to another point in \( c_{i2} \) that are closest in terms of some distance measure, for example the Euclidean distance (used in the next three examples). Formally,

\[
\text{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a, b) = \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2}
\]

This method is similar in principle to the traditional MST approach, except edges are constructed between the hierarchically derived subsets. It can be sensitive to outliers; for example, if two outlying points in adjacent clusters happen to be closest. To minimize the influence of outliers, a second merging method, called weighted centroids can be used. In it, the objective function incorporates the distance of the data point from the centroid of the corresponding cluster, yielding the convex combination described in Eq.(3).

\[
\text{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a, b) = (1 - \lambda) \sqrt{\sum_{j=1}^{d} (a_j - b_j)^2} + \lambda \left( \sqrt{\sum_{j=1}^{d} (a_j - c_{i1})^2} + \sqrt{\sum_{j=1}^{d} (b_j - c_{i2})^2} \right)
\]

In Eq.(3), \( c_{i1} \) is the centroid of \( c_{i1} \) and \( \lambda \) is a mixing value between 0 and 1. At \( \lambda = 0 \) Eq. (3), becomes the same as the nearest neighbor strategy. Our third method, centroid points, explicitly encourages the best alignment to cluster centroids by choosing a point in \( c_{i1} \) closest to the centroid of \( c_{i2} \).

\[
\text{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a, b) = \sqrt{\sum_{j=1}^{d} (a_j - c_{i2})^2} + \sqrt{\sum_{j=1}^{d} (b_j - c_{i1})^2}
\]

While the above methods do not necessarily guarantee the construction of a minimum spanning tree they do guarantee that higher groupings within the dataset are maintained.

### C. Method Refinement

The hierarchical clustering step provides us with aggregate representations of the data. Because each level has an associated cluster, we can use the clusters and their centroids as a higher level representation of the underlying dataset. To do so, we consider the cluster centroids to represent the central tendency of the cluster points. Replacing the original points with these centroids effectively performs a locally adaptive kernel smoothing where the kernel size adapts to the local data distribution. We’ll see in the results section that using centroids in place of the original data can lead to improved reconstructions.

### IV. Experiments and Results

#### A. Data

To assess the approach, we first generate two synthetic datasets to model distinct biological systems. The first set simulates a sequence of transitions between discrete states and the second simulates a smooth linear trajectory generated by a polynomial function. Varying degrees of Gaussian noise is added to the elements of those two sets. In the next experiment, three biological datasets are used. These include a stem cell differentiation dataset, a dataset consisting of cDNA microarray samples of the cell cycle, and screening data involving phenotypic changes occurring over time in the juvenile Schistosoma mansoni parasites due to drug exposure. Finally, we present a case study where we apply our method on a dataset of human single nucleotide polymorphisms.

#### B. Synthetic Datasets

We created a synthetic dataset by randomly generating points from six discrete states with an implicit ordering along the abscissa. Points were perturbed with varying amounts of Gaussian noise. The diameter of the CST was found to pass through each state in sequence while the MST took a much simpler path and failed to pass through all states. This result is due to the fact that by considering local structures, the CST method allows local centers of mass to (correctly) influence the reconstruction. We direct the reader to our preliminary report [8], where a visualization of these results is presented.

To simulate a process characterized by a continuous linear trajectory we sampled the polynomial \( y = x^2 + 3x^2 - 6x - 8 \) with additive Gaussian noise which allows us to directly measure the reconstruction error and observe the robustness of the methods to increasing noise. Figure 3.1 plots the reconstruction errors of the CST method as a function of the seven hierarchical clustering techniques described in Table 1. The reconstruction error of the MST method is also plotted for comparison. We see that all CST methods outperform the baseline MST method at all noise levels. In Figure 3.2, four cluster merging methods are compared, using the average (UPGMA) clustering technique. The weighted centroids method is examined with two values for the mixing coefficient \( \lambda \): .01 and .05. We see that the nearest neighbor and weighted centroid methods have comparable reconstruction error. The centroid points strategy does result in higher reconstruction error, it is, however, still better than the standard MST construction.

#### C. Reconstruction of Embryonic Stem Cell Differentiation

The polynomial reconstruction and simulated state changes are both examples of non-branching processes that we can represent as the diameter path through the tree. However, many biological progressions are characterized by branching processes where, at given branch points, the trajectory may follow one of several paths. Pluripotent embryonic stem cell (ESC) differentiation constitutes one such example. We use the data from [23] which contains 44 samples of mouse stem cells at different stages of differentiation. Interventions were performed on these samples to induce differentiation into trophoblasts, neural cells, endoderm lineages,
ordering should be able to reconstruct the sequence of phases in the cycle and as a preliminary task, group proteins by their role in the cycle. In the next experiment, we apply the CST, MST, and SPD methods to the expression levels of the 1099 genes in the human tumor cell cycle dataset provided in [2]. The CST reconstructed cycle is shown in Figure 5.1. In this figure, each gene is depicted as a vertex in the tree colored by the associated phase.

Because the cell cycle is a non-branching process, we represent the underlying trajectory as the diameter path through the tree and connect the ends of the path to form a cycle. The reconstructed tree is not strictly linear and can be rather bushy, so, a diameter node is assigned the class through a majority vote of all its children in the off-diameter branches. To better represent phase regions of the diameter path, we perform neighbor smoothing whereby a vertex’s phase assignment is determined by the majority vote of its raw phase and that of each of its neighbors. The smoothed diameter paths are shown in Figure 5.2. In these diameter paths, we see that the CST method correctly reconstructs the phase sequence with the exception of two G1/M phase nodes in the G1/S phase region. This result can be explained by the implied overlap of G1 phase within the two regions. On the other hand, the MST method fails to represent the G1/M phase altogether while the SPD method combines M/G2, G1/M and G2 phase proteins. Thus the CST path not only provides a better representation, but its diameter path is also significantly longer which means that the reconstruction is at a higher resolution than the other two methods. We measured the reconstruction error by counting the number of nodes whose phase does not match the phase of its nearest diameter node. The CST method had the lowest reconstruction error of the three methods followed by MST and SPD respectively.

If we project the data in a two dimensional Principal Component Analysis (PCA) space, the cell cycle forms a circular point cloud with distinct phase regions. Plotting the diameter paths resulting from CST and MST in this space, we find that the paths traverse the exterior of the point cloud (see Figure 5.3 (left column)). This is expected because the diameter path is by definition, the maximal shortest path between any pair of vertices in the network, and the farthest pair-wise distances are found around the perimeter of the data set. An alternative view of the diameter path can be achieved by projecting all points in the dataset to the nearest point on the diameter in the original, non PCA space and then observing the phase densities along the path. Such a view is shown in Figure 5.3 (right). In so doing, we see that the G1/S, S and G2 phases have relatively distinctive regions (with some overlap) for both the MST and CST method. The problem for MST arises in distinguishing between G1/M and G2 phase proteins where the class densities are very similar. The CST method does not completely avoid overlaps. However, it does differentiate the region more clearly.

D. Cell Cycle Reconstruction

Gene expression measurements from the cell cycle let us evaluate our method on a well characterized phase sequence. During cellular reproduction, a cell passes through the G1 phase, S phase, G2 phase, and then M phase to complete one iteration of the cell cycle. Each phase can be characterized by a set of highly expressed genes that carry out the biological function. These expression levels can be measured using cDNA microarrays. As shown in [2], the expression profiles form natural clusters of genes associated with each phase of the cell cycle. A method for reconstructing the temporal

and embryonic carcinoma. Each sample contains 25,164 gene expression measurements. After application of CST, all differentiation lineages are reconstructed in the proper temporal order. As indicated in Figure 4, the four cell lineages each branch off from the embryonic stem cells in the center of the tree. These results are comparable to those achieved by the Sample Progression Discovery method in [7].

D. Cell Cycle Reconstruction

Gene expression measurements from the cell cycle let us evaluate our method on a well characterized phase sequence. During cellular reproduction, a cell passes through the G1 phase, S phase, G2 phase, and then M phase to complete one iteration of the cell cycle. Each phase can be characterized by a set of highly expressed genes that carry out the biological function. These expression levels can be measured using cDNA microarrays. As shown in [2], the expression profiles form natural clusters of genes associated with each phase of the cell cycle. A method for reconstructing the temporal
Fig. 5. 1. Cell cycle reconstruction. The CST (shown), MST and SPD methods were applied to the cell cycle gene expression microarray data. The cell cycle phases are: G1, S-phase, G2, and M. Each gene is represented by a node in the tree colored by its associated phase. CST properly separated the phases and reconstructed the sequence in the correct order. The phases could not however, be adequately separated with the MST and SPD methods. 2. The diameter paths of each tree (with 1-neighbor smoothing). The MST does not contain the G1/M phase. SPD mixes M/G2, G1/M and G2 proteins. Error is computed by summing the number of vertices that do not match the nearest diameter vertex’s phase. 3. Left: The diameter paths in a two dimensional principal component space. The non-smoothed paths are jagged and include many outlying data points. Centroid smoothing leads to a more consistent path. Right: (Phase) class densities projected onto the respective diameter paths. Both CST and Smoothed CST show relatively distinct class regions, however The MST diameter path combines the M/G2 and G1/M phases much more than the CST variants. 4. Centroids from hierarchical clustering. Circle diameter corresponds to cluster size and color indicates level of the hierarchy with the root in red and leaves in blue. We see a trend towards the center of mass with fingers showing outlying clusters being pulled towards the center. 5. Progression reconstruction of parasite phenotypic response. 1-4 days of exposure to 10\(\mu\)M concentration of the drug Mevastatin. CST successfully groups most of the early-stage and late-stage responses corresponding primarily (but not exclusively) to the first and fourth day samples respectively. A heterogeneous intermediate cluster is also created containing parasites with two and three days of exposure. Example parasite images from various points on the tree are shown as well as the progression reconstruction error. 6. Left. Path reconstruction of CST, MST and CST smoothed methods. CST and MST show similar paths. The path resulting from the smoothed dataset is more concise and closely follows the shape of the data distribution. Right: Class densities projected onto the diameter path 7. Feature distributions across the diameter. Top: parasite brightness fall with time. Middle: parasite image inconsistence increase. Bottom: parasite roundness and texture increase with time.

points in the cluster. Figure 5.4 shows how these centroids move across the PCA space at the different levels of the hierarchy. Increase in circle diameter corresponds to increase in cluster size, and circles are colored blue (leaf) to red (root) based on their level in the hierarchy. We see the centroids pulling towards the center of mass as the clusters grow and they ascend the tree. If the original data points are replaced with centroids to arrive at a locally smoothed representation, we find the resulting path to be more representative of the shape of the data distribution. Furthermore, the class densities along the diameter remain comparable to the raw CST data as shown in Figure 5.3.

E. Reconstruction of Phenotypic Screening Data

In this experiment, we consider phenotypic screening data of macroparasites that cause the disease schistosomiasis. The data consists of images of 95 \textit{S. mansoni} parasites taken on the first, second, third, and fourth day of exposure to a 10\(\mu\)M solution of the HMG-CoA reductase inhibitor
same metric as in the cell cycle dataset. Figure 5.5 shows the typic responses and measure the reconstruction error using the irregular shapes, become more pronounced. As the drug exposure time increases, the deleterious effects, such as bloating, darkening and forming of shape and texture. Each parasite is thus represented by 43 quantitative image features that describe the parasite’s shape and texture. As the drug exposure time increases, the deleterious effects, such as bloating, darkening and forming of irregular shapes, become more pronounced.

We seek to reconstruct the time progression of the pheno-typic responses and measure the reconstruction error using the same metric as in the cell cycle dataset. Figure 5.5 shows the tree resulting from the CST reconstruction which especially groups and orders the early and late stage phenotypes. The heterogeneous phenotypes from intermediate exposures (day 2 and day 3) are also clustered together. We can interpret the results as showing three intuitive groupings: initial response, intermediate response, and maximal response. It is worth noting that, upon visual inspection of the data, the four parasites that are grouped with the parasites which had suffered four days of drug exposure, all show significant degenerative effects. Similarly, the two parasites which had four days of exposure and are grouped with the single-day exposure group also have similar phenotypes.

Parasites at the fourth day of exposure are grouped by all three methods. However, only CST was able to group the parasites with one day of exposure, while both MST and SPD split them and placed them on opposite ends of the tree. By reviewing the spatial organization of the underlying data through a lower dimensional PCA projection we observe that, while the parasites from the first day of exposure are near to each other in feature space, the MST and SPD algorithms do not take into account the local organization and one misplaced edge has significant effects on the overall graph topology. The local constraints enforced by CST help to ameliorate this problem and improve the overall reconstruction.

Fig. 6. CSTs constructed from the 1000 genomes project single nucleotide polymorphism (SNP) dataset. SNPs from two genes, GUCY1A3, and KCNQ1 were used in tree construction. Distinct clustering of individuals with Africa origin are found in both trees.

Mevastatin which has been studied for its potential anthelmintic effects [24]. The parasites cannot be cloned and consequently are individuals with varying responses to the chemical insult. Thus, exposure duration does not exclusively determine phenotypic progression. We applied the image analysis pipeline described in [25] which consists of phase-congruency based parasite identification and feature extraction from individual parasites. Each parasite is thus represented by 43 quantitative image features that describe the parasite’s shape and texture. As the drug exposure time increases, the deleterious effects, such as bloating, darkening and forming of irregular shapes, become more pronounced.

We seek to reconstruct the time progression of the pheno-typic responses and measure the reconstruction error using the same metric as in the cell cycle dataset. Figure 5.5 shows the tree resulting from the CST reconstruction which especially groups and orders the early and late stage phenotypes. The heterogeneous phenotypes from intermediate exposures (day 2 and day 3) are also clustered together. We can interpret the results as showing three intuitive groupings: initial response, intermediate response, and maximal response. It is worth noting that, upon visual inspection of the data, the four parasites that are grouped with the parasites which had suffered four days of drug exposure, all show significant degenerative effects. Similarly, the two parasites which had four days of exposure and are grouped with the single-day exposure group also have similar phenotypes.

Parasites at the fourth day of exposure are grouped by all three methods. However, only CST was able to group the parasites with one day of exposure, while both MST and SPD split them and placed them on opposite ends of the tree. By reviewing the spatial organization of the underlying data through a lower dimensional PCA projection we observe that, while the parasites from the first day of exposure are near to each other in feature space, the MST and SPD algorithms do not take into account the local organization and one misplaced edge has significant effects on the overall graph topology. The local constraints enforced by CST help to ameliorate this problem and improve the overall reconstruction.

Observing the class distributions across the diameter, we see that, as indicated in the tree representation, parasites exposed for one and four days tend to have significant regions, and the parasites with one day of exposure show a bimodal characteristic, which may be a product of noise or indicate distinct response states. Smoothing our dataset by replacing the original points with centroid values for clusters of size 2 (empirically derived) results in a diameter path that closely resembles the shape of the point cloud.

Using the smoothed diameter path, we examine the change in parasite feature values across the pseudo time represented by the diameter. Figure 5.7 shows the trends across time. The top plot shows the mean, median, and mode pixel intensity values which is a measure of parasite brightness. We see a downward trajectory which agrees with the general darkening effect associated with degenerate parasites. The middle plot shows the standard deviation of various pixel values within a parasite, and we see an interesting upward trend meaning that the parasite images become less internally uniform. The bottom plot shows two features, extent is a measure of roundness of the parasites and correlation is a measure the internal texture. Both show increasing trends across the diameter.

F. Case Study: Genetic Variation

The 1000 genomes project [26] has produced a dataset built from whole human genome sequencing that contains 84 million single nucleotide polymorphisms (SNPs) from individuals of African, American, European, South Asian, and East Asian ancestry. SNPs are indices of the genome that show high variance (>1%) across the population. We examined the SNPs from two genes from the genome of 1000 individuals: GUCY1A3 and KCNQ1. Both of these genes are either implicated or suspect to be implicated in a human pathology. Gene GUCY1A3 plays a role in the conversion of GTP to 3',5'-cyclic GMP and pyrophosphate. It has been studied for its ties to hypertension and myocardial infarction [27]. KCNQ1 is on the 11th chromosome and belongs to the family of genes that provide instructions for making potassium channels, it has been tied to susceptibility to type 2 diabetes mellitus [28].

The trees resulting from the CST method are shown in Figure 6 and show distinct internal structures. In both cases we see clustering of individuals of African origin. This cluster is centrally located in the GUCY1A3 tree and at an extremity of the KCNQ1 tree. Diabetes mellitus, which is linked to the GUCY1A3 gene has been shown to have a higher prominence in African americans [29] and common genetic factors in the African American sample may contribute to the clustering. Interestingly, the clustering and resulting tree topologies seem to support the ‘out-of-Africa’ theory of human migration [30] where the GUCY1A3 shows non-African individuals propagating out of a central African cluster and KCNQ1 has propagation potentially moving right to left.

V. Conclusions

Biological progressions contain internal structures such as groupings and partial orderings that can be leveraged to improve sample-order reconstruction. We have developed a
general framework that explicitly identifies these structures through hierarchical clustering and then uses them to guide the progression reconstruction. Application of this method was found to improve the order reconstruction for several datasets when compared with other methods at the state of the art.

ACKNOWLEDGEMENT

The authors thank Conor R. Caffrey for the S. mansoni phenotypic screening data.

REFERENCES