Progression Reconstruction from Unsynchronized Biological Data using Cluster Spanning Trees

Ryan Eshleman and Rahul Singh(✉)

Department of Computer Science, San Francisco State University,
San Francisco, CA 94132, USA
rahul@sfsu.edu

Abstract. Identifying the progression-order of an unsynchronized set of biological samples is crucial for comprehending the dynamics of the underlying molecular interactions. It is also valuable in many applied problems such as data denoising and synchronization, tumor classification and cell lineage identification. Current methods that attempt solving this problem are ultimately based either on polynomial and piece-wise approximation of the unknown generating function or its reconstruction through the use of spanning trees. Such approaches face difficulty when it is necessary to factor-in complex relationships within the data such as 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 data. Through a number of experiments on synthetic data sets as well as datasets from the cell cycle, cellular differentiation, and phenotypic screening, we show that the proposed CST approach outperforms the previous approaches in reconstructing the temporal progression of the data.

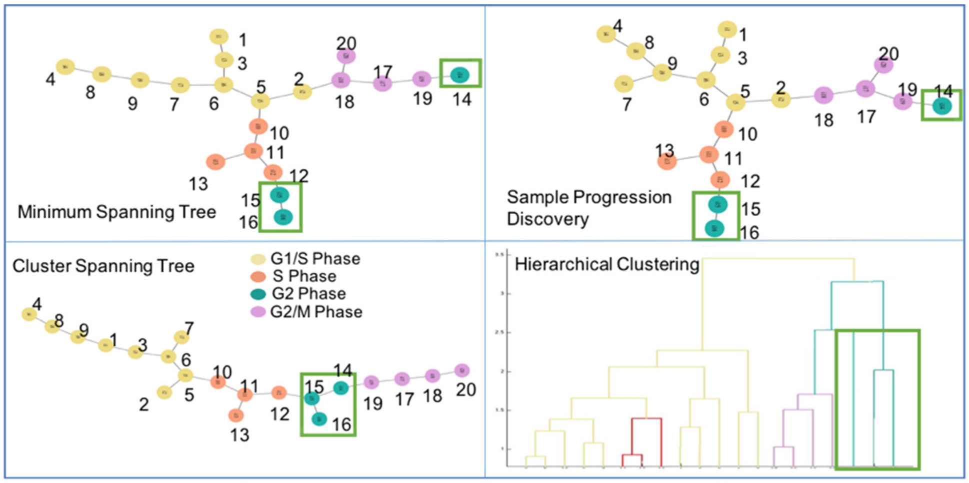
1 Introduction

Biochemical processes are dynamic processes expressed over time (and space). In terms of characterizing their temporal progression, a small set of generating functions can characterize such processes. For example, linear or polynomial functions (cell growth [11]), cyclical functions (cell cycle [12]), and branching (bifurcating or multifurcating) functions (cancer progression [13]). If the system under study can be sufficiently synchronized, as with cell synchrony methods [22], characterizing the underlying progression is 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 it is complicated due to epistemic and intrinsic factors such as the unknown nature of the molecular mechanisms of action, their (putative) non-linearity, phase shifts, and rate heterogeneity, as well as extrinsic factors such as undersampling, and noise.

Formally, if we think of a biological process as a series of states evolving with respect to time, the problem of constructing the temporal ordering for a set of samples requires specifying the function $f(t) = [x_1(t), x_2(t), \ldots, x_d(t)]$, where $x_i(t)$ is the value of dimension i at time t, so the output $f(t)$ is a point in d dimensions representing the state of the

A. Bourgeois et al. (Eds.): ISBRA 2016, LNBI 9683, pp. 1–12, 2016.
DOI: 10.1007/978-3-319-38782-6_12

process at time t. This function has to be reconstructed from the samples $S = \{s_1, \dots s_n\}$, where $s_i = f(i) + \varepsilon$ with ε denoting the noise. Noise modeling is often simplified by using well characterized distributions, such as a Gaussian. Graph-theoretic representation of the biological data provides a powerful formalism, especially for representing non-linear progressions. In such a representation, the complete data set is represented by a graph $G_c = (V, E)$ with each data point 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 [9, 10, 13]. MST-based methods assume that the tree with the minimum total edge weight best represents the underlying process. This does not account for relationships present in the data, such as groupings corresponding to subprocesses. Furthermore, the connectivity of a tree can be sensitive to how edges are selected and a poor choice may misrepresent relationships in the data. To illustrate this point, we use three different methods to reconstruct the progression of gene expression during the cell cycle. In this example 20 proteins associated with different phases of the cell cycle are chosen from the cell cycle cDNA expression micro array dataset [12]. Figure 1 shows the temporal ordering reconstructed by the MST-based method [9], the Sample Progression Discovery (SPD) method [10] and the proposed Cluster Spanning Tree (CST) approach. All three methods accurately group proteins from the G1/S, S, and G2/M phases, however only CST correctly groups the G2 phase proteins. Moreover, the CST is the only method that arranges the proteins in the proper order that reflects the stages of the cell cycle: G1/S, S, G2, G2/M. A more complete evaluation on this dataset is presented in the results section.



#	Phase	Protein
1	G1/S	CCNE1
2	G1/S	CDC25A
3	G1/S	CDC6
4	G1/S	E2F1
5	G1/S	MCM2
6	G1/S	MCM6
7	G1/S	NPAT
8	G1/S	PCNA
9	G1/S	SLBP
10	S phase	DHFR
11	S phase	RRM1
12	S phase	RRM2
13	S phase	TYMS
14	G2	CCNA2
15	G2	CCNF
16	G2	TOP2A
17	G2/M	BIRC5
18	G2/M	CCNB1
19	G2/M	CCNB2
20	G2/M	CDC20

Fig. 1. Progressions reconstructions from applying three reconstruction methods (MST, SPD, and CST) to a subset of the cell cycle micro array dataset in [12].

2 Background

Given a sampling S of size n of f, one way of reconstructing the underlying generating function is through polygonal approximation. Polygonal reconstruction [1] builds a connected graph $G = (V, E)$, where the vertices V are points from S and edges E connect the vertices such that each vertex has degree of 1 or 2 and for each set of adjacent vertices $[v_i\ v_j]$ corresponding to points $[f(i)\ f(j)]$, there does not exist a v_k: $f(i) \leq f(k) \leq f(j)$. This can be achieved by determining a traveling salesman path. The notion of principal curves can also be used to order data points when the manifold on which they lie has a curvature. Principal curves were introduced in [2] and constitute a non-linear generalization of principal components. For the set S, a principal curve is defined as a smooth function f_c that passes through the center of mass of the sample set S and is self-consistent, as defined by Eq. (1):

$$f_c(t) = E[S|t_f(S) = t] \tag{1}$$

In Eq. (1), $t_f(S)$ denotes points in S that are projected to point t. That is, each point on the principal curve coincides with the expectation of the data points that are mapped to it. As noted in [9], principle curves may require sampling at a denser rate than is provided in many biological contexts.

Neither polygonal reconstruction nor principal curves can be used to model branching processes. In such cases the system at time t has more than one possible state at time $t+1$. To address such issues, piece-wise representations, such as a spanning trees, have been employed that create tessellated representations of the data and reconstruct temporal ordering in each tessellate. A spanning tree of a complete graph $G_c = (V,E)$ is the connected graph $G_s = (V, E')$ where $E^{'} \subseteq E$and$\exists u \in V$:$(u, v) \in E' \vee (v, u) \in E' \forall v \in V$. Plainly, the subset $E^{'}$ contains edges that span all vertices in V. Because of the limited number of edges, a spanning tree enforces a unique path between vertices. Per Cayley's formula [3] there are n^{n-2} spanning trees on any complete graph. Therefore we must add constraints to find those trees which are biologically meaningful. An MST on G_c is a spanning tree with the additional constraint that $\sum_{e \in E'} e$ is the minimum across all spanning trees on G_c. MSTs can be constructed with one of many greedy algorithms, such as Kruskal's [4] or Boruvka's [5] that iteratively collect edges with the least weight to build the tree. The methods described in [9, 10] employ variations on the MST approach. In [9], the diameter path through the MST (or multiple candidate diameters with a PQ tree in the presence of noise) are used to determine the progression. In [10], an automated feature selection step is incorporated where MSTs are constructed on subspaces of the original feature space. The subspaces that generate the most similar MST topologies are merged to form the final putative MST progression.

As discussed earlier, the MST formulation cannot represent interrelationships such as natural sub-processes or groupings in the data. However, hierarchical clustering methods (like UPGMA [6]) may be used to identify data clusters which should be maintained in the resulting temporal reconstruction. Indeed, a method like UPGMA may be used directly for reconstructing temporal progression as in phylogenetics. A generic application of phylogenetic methods to this problem is however precluded, since such methods always impose a bifurcating structure on the data.

3 Methods

We propose the idea of cluster spanning trees (CST) that can maintain temporal and hierarchical clustering structure of the data and investigate three algorithmic variations for CST construction. At the fundamental level, this method is a process of traversing a hierarchical tree which represents the relations in the data and iteratively adding edges between nodes or groupings thereof. A binary hierarchical tree $G_b = (V_b, E_b)$ in our formulation contains $2n\text{-}1$ vertices, n is the number of data points being clustered. The n leaf vertices represent the data points. Each of the n-1 internal vertices represent the union of its descendants. Accordingly, the root is a set of size n. Each internal vertex v_i has two children, c_{i1} and c_{i2} each containing disjoint sets where $v_i = \{c_{i1} \ \cup c_{i2}\}$. Figure 2 shows an example. The CST is constructed as follows: 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, e.g. data points in the root node of G_b and E_{CST} is the empty set. For each non leaf vertex v_i in V_b an edge is added to E_{CST} from the child vertices of v_i, c_{i1} and c_{i2}, that connects a point in c_{i1} to a point in c_{i2} and minimizes a distance function $d(c_{i1,}\ c_{i2})$. While the order in which the vertices are traversed is arbitrary and does not affect the resulting CST, if an in-order traversal is performed, this algorithm can be understood as the iterative merging of a set of trees into a single tree.

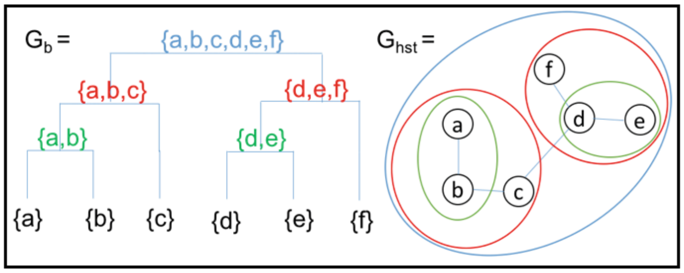


Fig. 2. **Left**, dendrogram and subsets assigned to the binary tree. Each internal node contains the union of the two child nodes. **Right**, Cluster Spanning Tree constructed from the hierarchical tree. For every internal vertex of G_b there is a connected subtree of G_{CST}.

When this operation has been performed over all internal nodes, we are guaranteed that for every internal node v_i in V_b there exists a connected sub-tree of $G_{CST,}$ $G'_{CST} = (V'_{CST}, E'_{CST})$ where $V'_{CST} \subseteq V_{CST}, E'_{CST} \subseteq E_{CST} \mathit{and} V'_{CST} = v_i$. Accordingly, the hierarchical clusters identified at the clustering stage are represented as sub-trees of the CST. As a general framework for the downward projection of a binary hierarchical tree of $2n-1$ vertices into a tree of $n-1$ vertices, there are two major algorithmic components to consider, namely, hierarchical data clustering and cluster merging.

3.1 Hierarchical Data Clustering

There are a number of established hierarchical clustering techniques that can be utilized to perform the initial data clustering. Methods we have investigated include the Unweighted Average (UPGMA) [6], Weighted Average (WPGMA) [6], Complete Linkage [6], Centroid [7], Median [7] and Incremental Sum of Squares (Ward) [8] methods. Details on these methods can be found in the references. All of these methods induce a hierarchical structure on the data that can be used to obtain a hierarchical clustering of the data. Non-hierarchical clustering techniques can also be employed for this problem. To limit the scope of this paper, they are not discussed.

3.2 Cluster Merging

The second algorithmic component is the strategy used to draw edges between points in the subsets at each bifurcation of the hierarchical binary tree. This consists primarily of choosing a distance function to minimize. The first vertex merging strategy is the nearest neighbor approach. An edge is drawn from the point in c_{i1} to the point in c_{i2} that are nearest in terms of some distance measure, for example Euclidean (used in the next three examples). Formally,

$$\operatorname{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a,b) = \sqrt{\sum\nolimits_{j=1}^{t} (a_j - b_j)^2} \tag{2}$$

This method is similar in principle to the traditional MST approach, except edges are constructed between the hierarchically derived subsets. This approach can be sensitive to outliers, for example if two outlying points in adjacent clusters happen to present the minimum distance. To minimize the influence of outliers, we employ the second merging method, called weighted centroids (defined in Eq. (3)) where we incorporate into the objective function, the distance from the centroid of the corresponding cluster point. This gives us the convex combination described in Eq. (3).

$$\begin{aligned}\operatorname{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a,b) = (1-\lambda)\sqrt{\sum\nolimits_{j=1}^{t} (a_j - b_j)^2} + \\ \lambda\left(\sqrt{\sum\nolimits_{j=1}^{t} (a_j - \overline{c_{i1}})^2} + \sqrt{\sum\nolimits_{j=1}^{t} (b_j - \overline{c_{i2}})^2}\right)\end{aligned} \tag{3}$$

Here, $\overline{c_{i1}}$ is the mean value of points in c_{i1} equivalent to the centroid of points in the set and λ is a mixing value between 0 and 1. At $\lambda = 0$ this 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}.

$$\operatorname{argmin}_{a \in c_{i1}, b \in c_{i2}} d(a,b) = \sqrt{\sum\nolimits_{j=1}^{t} (a_j - \overline{c_{i2}})^2} + \sqrt{\sum\nolimits_{j=1}^{t} (b_j - \overline{c_{i1}})^2} \tag{4}$$

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

4 Results

We evaluated our methods on two synthetic datasets including simulated state transitions and data generated through a noisy polynomial generating function. We also evaluated on biological datasets from cellular differentiation, the cell cycle, and phenotypic screening. That the method can be successfully employed on widely differing data sets, underscores its generic nature and broad applicability.

4.1 Synthetic Datasets

To show how CST captures the larger internal structures of a dataset, we generated synthetic data by sampling six discrete states that have an implicit ordering along the abscissa. Gaussian noise was introduced at varying intensities, as shown in Fig. 3. In this example, we see that the diameter of the CST correctly passes through each of the six states in order because it encourages the path to pass through local centers of mass. The MST takes a simpler path and does not pass through all states. The dendrogram, number 2 in Fig. 3, shows the hierarchical structure found by the UPGMA algorithm that was used to guide the tree construction.

The previous dataset allowed us to observe the reconstruction of state transitions. For a more rigorous evaluation we constructed a synthetic dataset by sampling the polynomial $y = x^3 + 3x^2 - 6x - 8$ with Gaussian noise. This allows us to measure the reconstruction error of our methods and quantify the effect of increasing noise on deviation from the ground truth polynomial as shown in Fig. 4. The CST method consistently outperforms the MST based approaches proposed in [9,10]. Interestingly, all trees, including MSTs, are rather robust to noise except for a significant initial spike. This phenomenon occurs because when the noise level is low enough, the diameter path will pass through every point. The reconstruction error will increase with noise as long as the diameter path passes through every point, however when noise increases and outlier points are no longer on the diameter path, the outlier error no longer contributes to the reconstruction error, and reconstruction error stabilizes.

4.2 Reconstruction of Embryonic Stem Cell Differentiation Data

The two previous examples showed the method’s ability to reconstruct processes that are non-branching by representing the progression as the diameter path in the tree. However, many biological progressions are characterized by branching processes. For example, the pluripotent embryonic stem cell (ESC) differentiation data set from [10] 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, and embryonic carcinoma. Each sample contains 25,164 gene expression measurements. After application of CST, all differentiation lineages are

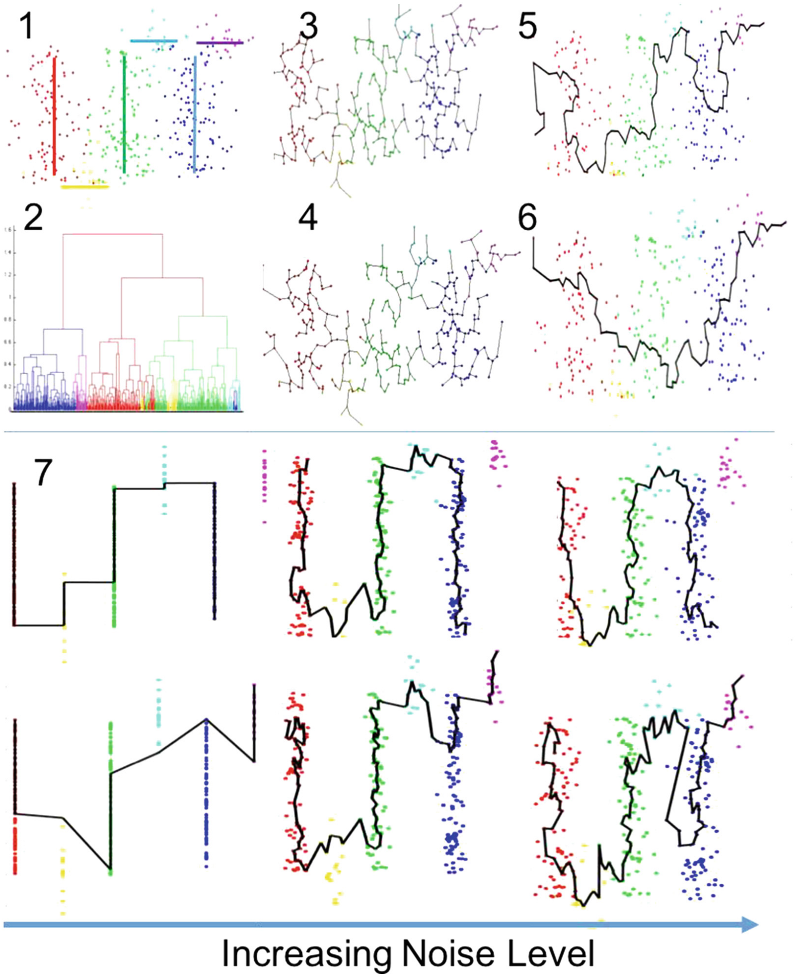
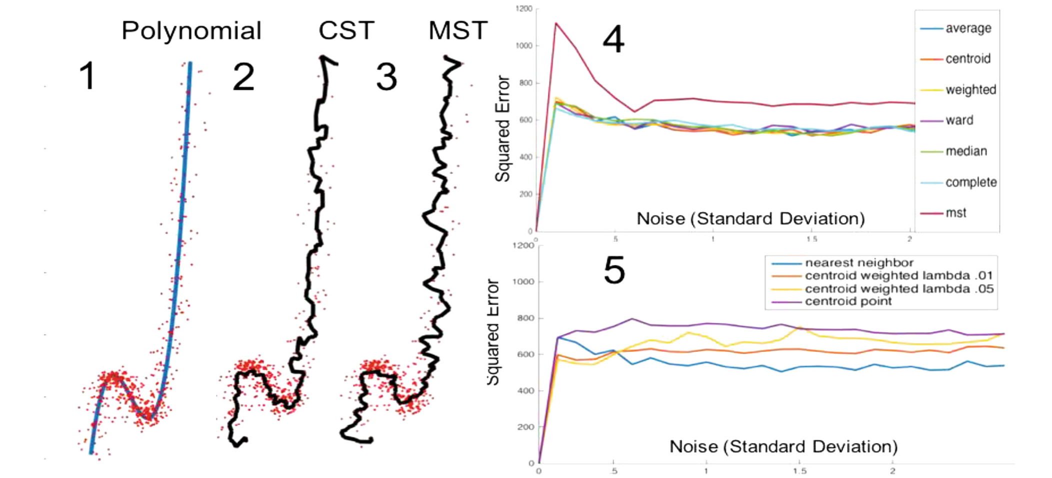


Fig. 3. Minimum spanning tree and UPGMA spanning tree path reconstruction for a noisy (additive Gaussian noise) synthetic data set composed of six states with an implicit horizontal ordering. **1.** The dataset showing mean values of the 6 states. **2.** The UPGMA dendrogram that shows the hierarchical clustering of the dataset used to enforce level-wise spanning tree construction. The clustering and class-color adjacencies in the dendrogram reveal how UPGMA spanning tree's constructed the correct path. **3.** Shows the CST built with the UPGMA and Centroid Point merging strategy. **4.** The MST built on this dataset. **5.** Is the diameter path of the CST which passes through all states in sequence. **6.** The diameter path of the MST which fails to pass through the light blue state. **7.** The diameter path of the data set with increasing noise with MST on top and CST below. The MSTs consistently fail to pass through the state coded in purple. (Color figure online)

reconstructed intact and in the proper temporal order; as shown in Fig. 5, the four cell lineages each branch off from the blue embryonic stem cells in the center of the tree. These results are comparable to those achieved by the Sample Progression Discovery method. The corresponding dendrogram confirms that the cell lineages are clustered in the clustering phase and the resulting reconstruction shows that temporal order is maintained within clusters.



	Standard	Squared Error										
	Deviation	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
Method	Average	0	669.8	617.5	578.4	539.4	534.6	549.5	538.8	548.6	563.7	543.3
	Centroid	0	669.3	579.4	584.2	540	520.8	539.5	515.3	539.3	562.4	550
	Weighted	0	650.9	573.2	576.7	552.5	532.5	530.3	523.4	539.5	561.7	591.6
	Ward	0	634.2	597.7	598.6	550.7	545.6	572	536.4	576.7	557.2	570.8
	Median	0	675.9	597.2	602.3	564.3	547.9	546.3	526.7	531.5	569.4	558.7
	Complete	0	622.3	584.8	590.2	580.6	574.6	553.1	551.9	542.5	566.9	542
	MST	0	989.6	720.8	706.5	716.4	696.6	675.5	686.3	694.6	696.8	683.3

Fig. 4. CST and MST performance on a synthetic dataset sampled from the polynomial $y = x^3 + 3x^2 - 6x - 8$ with Gaussian noise. **1.** The original curve over the sampled points. **2.** The CST construction, **3.** The MST reconstruction. **4**. Is the squared reconstruction error of the six clustering methods with nearest neighbor merging, and the MST. Noise increases left to right. Cluster trees show consistently lower reconstruction error. **5.** Reconstruction error of UPGMA clustering with the three merging strategies described in Sect. 3.2.

4.3 Cell Cycle Reconstruction

Cellular reproduction is carried out in a well characterized and repeating sequence of biological phases. Specifically, a cell passes through the G1 phase, S phase, G2 phase, and then M phase to complete one iteration of the cell cycle, beginning again at G1 phase to repeat the process. Each phase has a number of genes that carry out the underlying biological function, these genes are often highly expressed during their associated phase. To capture the expression dynamics at each phase, cDNA microarray samples measure gene expression levels throughout the cycle. The gene expression profiles form natural clusters of genes that are associated with each phase [12].

To test our approach's ability to both capture the gene clusters and accurately reconstruct the sequence of phases in the process we applied the CST, MST, and SPD methods to the expression levels of the 1099 genes in the human tumor cell cycle dataset provided in [12]. Each gene is represented by a vertex in the tree with the color indicating its associated phase in Fig. 6. Visibly, the CST method performs better separation of the phases. Both the MST and SPD methods tend to merge the G2, M/G2 and G1/M gene

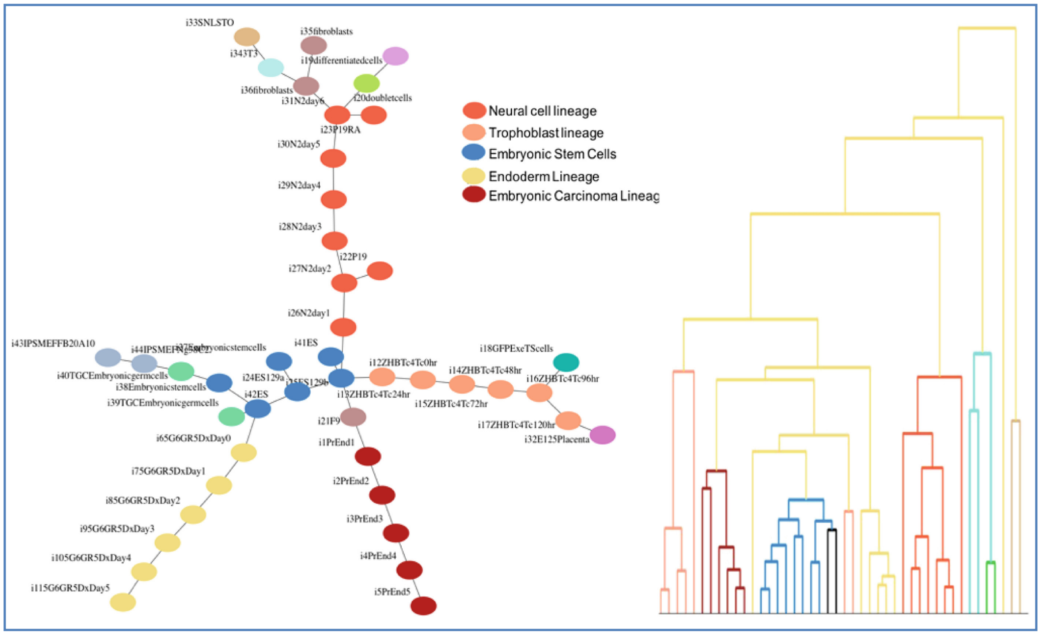


Fig. 5. Embryonic stem cell differentiation. Four cell differentiation lineages are reconstructed in order with sequential vertices representing increasing time. The dendrogram on the left shows the underlying hierarchical clustering that informed tree construction. The two images show that not only are the cell lineages generally clustered together, but their temporal order is maintained in the clusters.

groups. Because we know that the cell cycle is a repeating sequence with no branches, we observe the diameter path through the tree as a representation of the underlying biological sequence. To better represent phase regions of the diameter path, we performed 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 Fig. 6. The end points of the diameter are connected to show the cyclical nature of the process.

Observing these diameter paths, we see that the CST method correctly reconstructs the phase sequence with the minor exception of two G1/M phase nodes in the G1/S phase region, this can be explained by the implied overlap of G1 phase within the two regions. The MST method fails to represent the G1/M phase altogether while the SPD method combines M/G2, G1/M and G2 phase proteins.

Because most nodes do not appear on the diameter path and form clusters around the path, we measured the reconstruction error by counting the number of nodes whose phase assignment does not match the phase assignment of its nearest diameter node. The CST method had the lowest reconstruction error of the three methods followed by MST and SPD respectively.

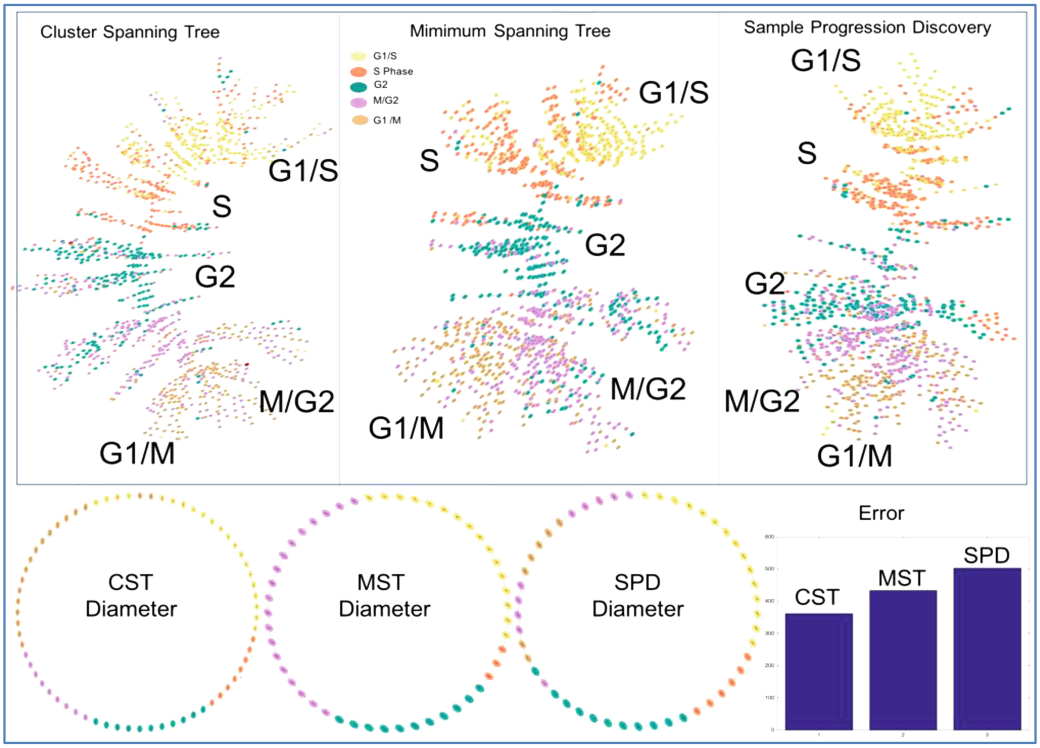


Fig. 6. Cell cycle gene reconstruction. The CST, MST and SPD methods were applied to the cell cycle gene expression microarray data. The cell cycle has a known sequence of phases: G1, S-phase, G2, M. Each gene is represented by a node in the tree colored by its associated phase in the cycle. The CST method properly separated the phases and reconstructed the sequence in the correct order. Phases were not sufficiently separated with the MST and SPD methods. The diameter paths of each tree with 1 neighbor smoothing are shown. 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.

4.4 Reconstruction of Macro-parasite Phenotypic Screening Data

We consider phenotypic screening against parasites that cause the disease Schistosomiasis. Our data set consists of images of 95 *S. mansoni* somules taken on the first, second, third, and fourth day of exposure to a 10μM solution of the HMG-CoA reductase inhibitor Mevastatin which has been studied for its potential anthelmintic effects [14]. Each parasite is represented by 43 quantitative image features that describe the parasite's shape and texture. Parasites tend to show increasingly apparent deleterious effects as exposure time increases.

Like with the cell cycle example, this dataset contains a known linear progression (exposure duration) and natural clustering (images of parasite groups taken on specific days), so we seek to reconstruct the time progression of the clusters from the dataset. Error is measured using the same metric from the cell cycle dataset, namely mismatches along the smoothed diameter path. Figure 7 shows the trees resulting from the three algorithms along with parasite images across the CST. The CST result shows strong grouping and correct ordering of parasites from days one and four. It is not surprising that the intermediate exposure days are rather heterogeneously grouped due to the varying rate of response that individual parasites show to the drug. While days two and three are merged, 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 underlying data, the three 'Day 3' parasites and one 'Day 1' parasite present in the 'Day 4' group all show significant effects and are properly placed, effect-wise, with the 'Day 4' parasites. Similarly, the two 'Day 4' parasites in the 'Day 1' group show idiosyncratic effects and are rightly not grouped with the other 'Day 4' parasites.

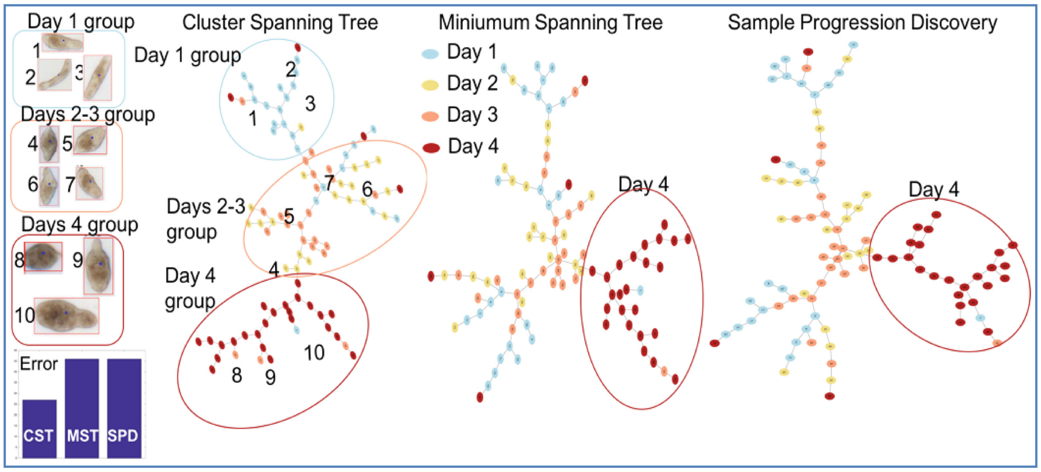


Fig. 7. Progression reconstruction of parasite phenotypic response after the first, second third and fourth day of exposure to 10 µM concentration of the drug Mevastatin. CST correctly groups the first and fourth day samples, while days two and three form a heterogeneous intermediate cluster. Example parasite images from various points on the tree are shown as well as the progression reconstruction error.

All three tree construction methods accurately grouped the 'Day 4' parasites, however only CST was able to group 'Day 1.' Both MST and SPD split the 'Day 1' group and placed them on opposite ends of the tree, significantly distorting the reconstruction. By reviewing the spatial organization of the underlying data through a lower dimensional projection (not shown) we observe that, while the parasites from Day 1 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.

Acknowledgements. This research was funded in part by the National Science Foundation grant IIS-0644418 and the National Institutes of Health grant 1R01A1089896.

References

1. Amenta, N., Bern, M., Eppstein, D.: The crust and the B-skeleton: combinatorial curve reconstruction. Graph. Models Image Process. **60**(2), 125–135 (1998)
2. Hastie, T., Stuetzle, W.: Principal curves. J. Am. Stat. Assoc. **84**(406), 502–516 (1989)
3. Aigner, M., Ziegler, G.M., Erdos, P.: Proofs from THE BOOK, vol. 274. Springer, Berlin (2010)
4. Kruskal, B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. **7**(1), 48–50 (1956)
5. Boruvka, O.: Contribution to the solution of a problem of economical construction of electrical networks. Elektronický Obzor **15**, 153–154 (1926)
6. Sokal, R.R.: A statistical method for evaluating systematic relationships. Univ Kans Sci Bull. **38**, 1409–1438 (1958)
7. Székely, G.J., Rizzo, M.L.: Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. J. Classif. **22**, 151–183 (2005)
8. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
9. Magwene, P.M., Lizardi, P., Kim, J.: Reconstructing the temporal ordering of biological samples using microarray data. Bioinformatics **19**(7), 842–850 (2003)
10. Qiu, P., Gentles, A.J., Plevritis, S.K.: Discovering biological progression underlying microarray samples. PLoS Comput. Biol. **7**, 4 (2011)
11. Bochner, B.R.: Global phenotypic characterization of bacteria. FEMS microbiology Rev. **33**(1), 191–205 (2009)
12. Whitfield, M.L., et al.: Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell **13**(6), 1977–2000 (2002)
13. Park, Y., Shackney, S., Schwartz, R.: Network-based inference of cancer progression from microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. **6**(2), 200–212 (2009)
14. Arreola, L.R., Long, T., Asarnow, D., Suzuki, B.M., Singh, R., Caffrey, C.: Chemical and genetic validation of the Statin drug target for the potential treatment of the Helminth disease. Schistosomiasis PLoS One **9**, 1 (2014)
15. Fitch, W.M.: Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**, 406–416 (1971)
16. 1000 Genomes Project Consortium.: A map of human genome variation from population-scale sequencing. Nature **467**(7319), 1061–1073 (2010)
17. Behrends, S., Vehse, K., Scholz, H., Bullerdiek, J., Kazmierczak, B.: Assignment of GUCY1A3, a candidate gene for hypertension, to human chromosome bands 4q31. 1 → q31. 2 by in situ hybridization. Cytogenet. Genome Res. **88**(3–4), 204–205 (2000)
18. Yasuda, K., et al.: Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat. Genet. **40**(9), 1092–1097 (2008)
19. Platt, O.S., et al.: Pain in sickle cell disease: rates and risk factors. N. Engl. J. Med. **325**(1), 11–16 (1991)
20. Allison, A.C.: Protection afforded by sickle-cell trait against subtertian malarial infection. Br. Med. J. **1**(4857), 290–294 (1954)
21. Ehret, G.B., et al.: Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature **478**(7367), 103–109 (2011)
22. Merrill, G.F.: Cell synchronization. Methods Cell Biol. **57**, 229–249 (1988)