Quantitative Analysis of Live-Cell Growth at the Shoot Apex of Arabidopsis thaliana: Algorithms for Feature Measurement and Temporal alignment.

Oben M. Tataw, G. Venugopala Reddy, Eamonn J. Keogh, Amit K. Roy-Chowdhury

Abstract—Study of the molecular control of organ growth requires establishment of the causal relationship between gene expression and cell behaviors. We seek to understand this relationship at the shoot apical meristem (SAM) of model plant Arabidopsis *thaliana*. This requires the spatial mapping and temporal alignment of different functional domains into a single template. Live cell imaging techniques allow us to observe real time organ primordia growth and gene expression dynamics at cellular resolution. In this paper, we propose a framework for measurement of growth features at the 3D reconstructed surface of organ primordia, as well as algorithms for robust time alignment of primordia. We computed areas and deformation values from reconstructed 3D surfaces of individual primordia, from live cell imaging data. Based on these growth measurements, we applied a multiple features landscape matching algorithm (LAM-M), to ensure a reliable temporal alignment of multiple primordia. Although the original landscape matching algorithm (LAM) motivated our alignment approach, it sometimes fails to properly align growth curves in the presence of high noise/distortion. To overcome this shortcoming, we modified the cost function to consider the landscape of the corresponding growth features. We also present an alternate parameter free growth alignment algorithm which performs as well as LAM-M for high quality data, but is more robust to the presence of outliers or noise. Results on primordia and guppy evolutionary growth data show that the proposed alignment framework performs at least as well as the LAM algorithm in the general case, and significantly better in the case of increased noise.

Index Terms— Growth alignment, Algorithms, Growth Alignment, Growth Feature Measurement.

1 Introduction

ADVANCES in live cell imaging techniques such as those developed in [20] present opportunities to tackle important plant development questions. Live imaging allows us to investigate the causal relationship between cell behavior, organ growth and genes that work in networks. This is possible because with live imaging, we are able to observe in real time, organ primordia growth and gene expression dynamics at cellular resolution. In this paper, we present a framework for quantitative study of primordia growth at the shoot apical meristem of Arabidopsis, (shown in Fig. 1.), based on live imaging data. In doing so, we present algorithms for temporal alignment of primordia growth data and any other developmental growth data for that matter. Time synchronization is often a necessary step in any meaningful analysis of time lapse data. Time lapse data generated from live imaging experiments is very noisy and often requires robust time synchronization for reliable deduction of these relationships.

In addition to plant biology, a similar problem is faced in other areas of science including biological and medical sciences, where experimental data is often very large and usually unsynchronized.

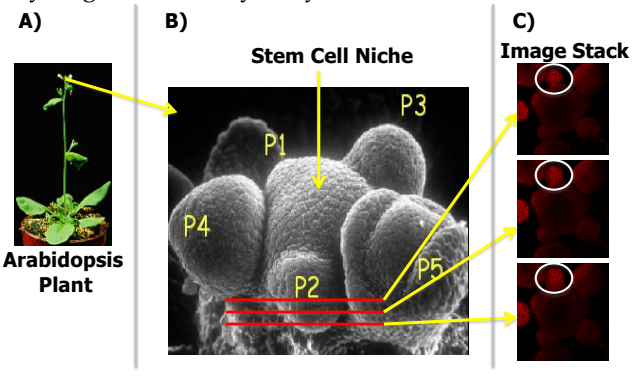

Fig. 1. Location of the Shoot Apical Meristem in model plant Arabidopsis *thaliana*. (A) Location of SAM. (B) Detailed view of SAM showing multiple primordia (P1-P5) at different developmental stages. (C) Illustrative sample of 3 out of 23 slices from an imaging session. White circles indicate location primordia per slice.

Quantitative time series data generated from these ex-

Manuscript received (1/18/2013). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

periments is sometimes collected through the measurement of interesting variables over varying timeframes. Such measurements and subsequent analysis are essential in efforts to gain greater insight into biological systems.

In the medical community, the study of multiple characteristics of human respiration can lead to the discovery of important dynamics in patient outcome studies [9],[28]. In order to achieve an acceptable level of reliability, such data needs to be properly time synchronized across multiple patients. In proteomics studies, there is a need for quantitative comparison of multi-class Liquid chromatography-Mass spectrometry (LC-MS) data [3]. The very nature of data generation in LC-MS studies is itself a source of large variability. Given the degree of noise and variation, a robust alignment of generated time series is often a necessary step towards reliable multi-class comparison.
One common and necessary requirement in these different domains and many others is the need to collect a large amount of unsynchronized data, properly align them and then perform further analysis.

Although there has been a fair amount of progress in this problem domain, to the best of our knowledge, no one has looked at the applicability of similar techniques in the study of plant development. Specifically, we do not know of any similar study applied to primordia growth dynamics at the shoot apex of Arabidopsis *thaliana*. Given high resolution data, many existing techniques have proven to be reliable and robust. However our problem presents many challenges in the form of low resolution time series, with high variability. Specifically we are challenged by the fact that we have about 5 to 13 time points per observed event. With such low resolution, the effect of noise and variability can be greatly magnified.

1.1 Main Contribution

This paper makes four specific contributions in the study of biological growth at the level of the organ (e.g. primordia):

- We show how to measure growth features from reconstructed 3D structures of organ primordia, based on live imaging data. (Section 3.2).
- We show the application of deformation field morphometry (DBM) for quantification of deformation at the surface of organs. (Section 3.2.2).
- We present a modified landscape matching algorithm for the alignment of low resolution, noisy time series using multiple features. (Section 3.3.2).
- We present a parameter free growth alignment algorithm. (Section 3.3.3).

1.2 Background

Although a full biological description of the primary system under study is beyond the scope of this paper, we present a brief description and motivation for our work. The Shoot Apical Meristem (SAM) is made up of stem-cells that provide cells for the development of all above ground plant structures. The organ primordia are regions of the SAM that develop into different plant organs. At each point in time, the SAM contains multiple primordia at different developmental stages, as shown in Fig. 1.B. The SAM is subdivided into different functional domains, with unique and overlapping gene expression patterns. Differential expression analyses have led to some understanding of the expression patterns of many SAM genes. In [26], analysis of just 3 cell types revealed more than 2000 genes with distinct and overlapping expression patterns. However, such an analysis is static and does not provide the dynamic context of observed patterns. It also does not provide the relationship between observed expression pattern, cell-cell communications, cell expansion/growth rates and organ growth.

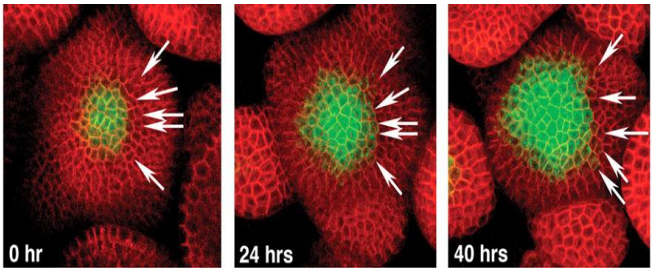

Fig. 2. Live imaging data of clavata-3 gene expression showing expression rates at 0hr, 24hrs and 40hrs. Live-imaging techniques allow for real time observation of growing SAMs. Proper alignment allows for dynamic gene expression analysis that is invariant to spatiotemporal changes. Data from [19]

Data for our project was collected using the live imaging technique developed in [20]. Live imaging allows for real time observation of primordia growth at the resolution of the cell. It also allows for real time observation of gene expression and cell division over a period of time, while the plant is alive and growing. These techniques use laser scanning confocal microscopy to obtain 3D volumes of SAM structure at different time instances throughout an observation period (usually several days), without damaging the growing plant. With live-imaging, we are able to capture the dynamic context of discovered genes, as illustrated in the images in Fig. 2., which were published in [19]. Live imaging also allows for computational tracking of cell division patterns as in [13]. We are

proposing a framework that will lead to the possibility of performing dynamic analysis of live imaging and other growth data. Such an analysis will be invariant to spatiotemporal changes in plant structure over time.
Data collected at each time instance of live imaging is in the form of a 3D stack of slices representing the view of the SAM at different focal planes from top to bottom.

Analyzing primordia morphogenesis based on measurable growth features is an important problem to plant biologists. It is important precisely because genes drive primordia growth. As a result of this, biologists need to understand the spatiotemporal dynamics of the interaction between gene expression and primordia growth. We seek to deduce the principles underlying the relationship between gene expression, cell division, cell-cell communication and overall primordia growth. Such an understanding will move us a step closer towards developing a dynamic gene expression atlas for the SAM of model plant Arabidopsis thaliana. Working towards this goal (see big picture in Fig. 3), we solve two important problems:

- Measurement of growth features at the surface of primordia
- Alignment of time series of growth features

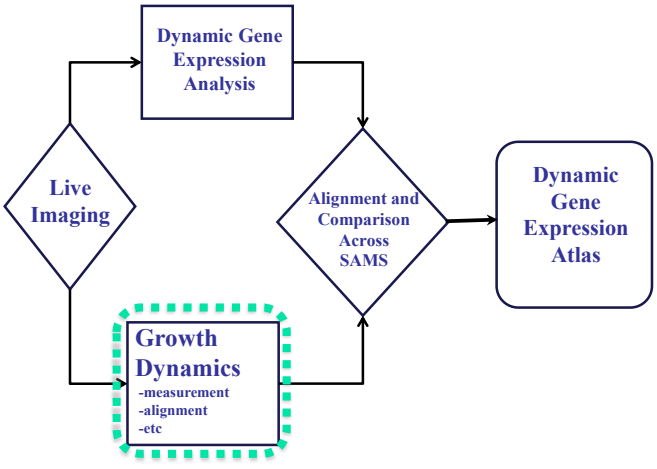

Fig. 3. Motivation for our study of growth at the SAM of Arabidopsis *thaliana*. Proper temporal alignment of growth data will allow for reliable development of dynamic models that integrate gene expression and quantitative data, both at the global (SAM) level and at the resolution of the cell.

Given the advantages of live imaging, one might imagine that we could just label all discovered genes in a single plant and perform our analysis using live-imaging. We are however limited in live imaging, because technical limitations mean we can only use 2 to 4 gene markers at a time. As a direct result of this restriction, we cannot study more than a few genes at a time using a single plant. Given that there are thousands of genes, and there is variation in SAM size and shape, we must perform analysis of different genes in different plants, and then integrate the results into a common template. Given that the SAM consists of multiple primordia at different developmental stages, time lapse data for a set of primordia have varying start and end times. Such data needs to be time synchronized in order to properly perform reliable analysis. This also requires complex alignments and comparisons across plants which calls for computational tools for SAM growth analysis. For this reason, our study of automatic growth alignment is a necessary first step in this process.

2 RELATION TO PREVIOUS WORK

The nature of the problem in our work is such that we are analyzing multiple agents per subject. These agents are undergoing similar biological processes (growth), but at any given moment, they are at different stages of that process. The challenge does not only come from the variability across SAMs, there is also great variability within the same subject (SAM).

In terms of measuring growth features, there has been some work within the past decade [5],[7],[11]. Our work differs from these studies in many ways. First of all our goal is to study the spatiotemporal dynamics of growth and use the output of our studies in the development of a dynamic model. Secondly, our dataset is significantly different. In [5], they used replicas taken from the surface of individual apex as input into their shape tracking system. In contrast, we use live imaging data that allows us to track gene expression at the resolution of the cell, while they used replicas taken from the surface of each individual apex to track shape changes. In [7], similar imaging equipment was used for data collection. However they used a different technique in their live imaging experiment. Their technique differed in the sense that they employed a multi-angle image acquisition approach, and then applied an image reconstruction technique that integrated these images into a single image with better resolution. They had a 24 hour time interval between observations, a constraint that was imposed by their approach. Such a long time window can present the risk of missing significant growth dynamics. In our case, we allowed 6 hour intervals between observations and took images from a single angle (horizontal). With a 6 hour interval, attempting the imaging approach employed in [7] will be very

traumatic to the plant, which will most likely not survive the adverse effect of laser intensity.

Our proposed system takes 3D surface points as input. These points represent the reconstructed 3D shape of primordia. Since we are focused on the study of global patterns in primordia growth, any contour based 3D reconstruction scheme is suitable to use with the proposed framework. Many image reconstruction techniques have been proposed in the past decades [1],[6],[14]. Methods like the classical Marching Cubes approach [14] or the more recent technique based on the Multi-level Partition of Unity (MPU) models [1] are examples of techniques that have been shown to produce acceptable results in many application domains.

In this work, we further show how to compute local surface deformation from 3D reconstructed live imaging data. Our approach to computing deformation was inspired by the phenomenon of deformation based morphometry (DBM). Although DBM originated from solid mechanics, it has been widely embraced in the medical imaging community to study a variety medical conditions. For example DBM has been applied to study the effects of alcoholism in [21], changes in human brain [17],[22] , as well as analysis of gray matter deformation in [4]. The validity of many variations of DBM has also been the focus of previous studies [8][23]. Although the adoption of DBM in the bio-medical imaging community has seen wide success, the approach has never before been used to study primordia development in the way we are proposing here.

Another major problem we address in this paper is that of time series alignment. In fact, we present two solutions for the temporal alignment of growth data, a task that is at the heart of this paper. Time series alignment is a problem that has attracted significant interest in the past. A widely adopted solution to this problem is dynamic time warping (DTW), an algorithm that has attracted interest from the data mining community [10],[27], although it came about as a result of work in the speech recognition community. Over the years, many variants of the DTW algorithm have been developed [16]. However, the key to most of these is the utilization of a robust cost function, in cases of high variability in amplitude and time. In addition to DTW, there are other approaches that have yielded reasonable results [2],[12]. In [2], a landscape based cost function was applied to develop what is referred to as the Landscape matching algorithm (LAM). This method was used to align data from respiration patterns of multiple patients. A statistics based approach was employed in [12], where multiple time series are simultaneously aligned and eventually a latent trace was inferred from the aligned set. This approach is very robust. However, because our plant growth data has very low temporal resolution, we cannot guarantee the correctness of such an approach. Our approach to align varying length and noisy sequences is detailed in Section 3.3.

3 Detailed Methodology

In this section, we give a detailed presentation of the entire proposed framework, from growth measurement, to eventual growth alignment of time lapse data. Although the primary goal of this paper is to present solutions for aligning growth data, we think for completion, it is important to give details on how we measure some of the features being aligned.

3.1 Overview of Proposed Framework

At any moment in time, the SAM is made up of multi- ple growing primordia at different stages of growth. The proposed framework is designed to handle the analysis of primordia from different plants. As shown in Fig. 4, the system is made up of two distinct modules: the growth computation module (GCM) and the growth alignment module (GAM). The GCM includes sub-modules to measure growth. For example in the study of primordia growth, the GCM sub-modules will play the role of measuring growth features at the surface of reconstructed primordia. The output of this module is a time series of primordia growth measurement, for all plants under study. Once we have a collection of time series, the growth alignment module performs a multiple series alignment on all the primordia. Deformation values computed by the GCM are used to improve alignment results.

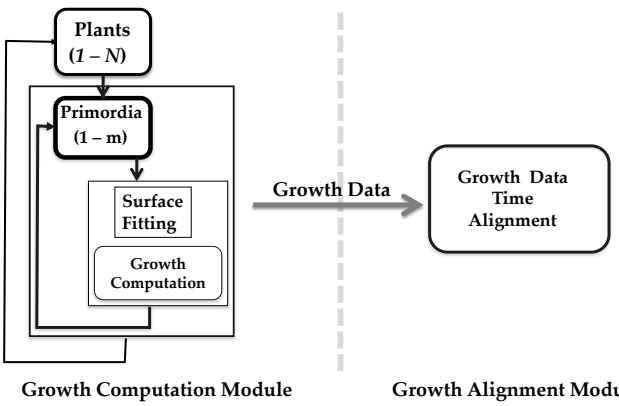

Fig. 4. The proposed system is made up of the growth computation module (GCM) which performs growth measurements, and the growth alignment module (GAM).

The next sections will give details on the methodology used for growth computation, and the alignment of

measured features.

3.2 Growth Measurement

3.2.1 Surface Area Calculation

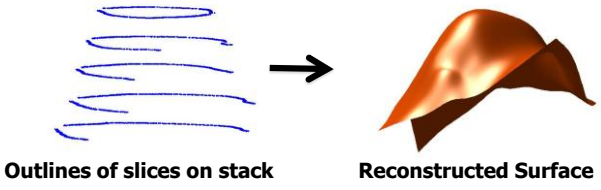

Fig. 5. Sample input into the system (left) is in the form of slice contours which are then used to reconstruct a smooth surface. Here we show a Matlab™ visualization of the reconstructed surface.

Given a stack of boundary points representing a primordia, we use existing tools to generate a smooth surface as shown in Fig. 5. The surface area of our reconstructed primordia surface is computed by performing a Delaunay triangulation, and then taking a summation on the areas of the triangles. Summation of the areas of all these surface triangles gives us an accurate estimate of the surface area of the primordia. Although there are many ways of computing the area of a triangle, we chose to use Heron's formula for computing the area of our triangle.

3.2.2 Deformation Computation

One of the fundamental patterns biologists seek to understand is the variational nature of deformation on the surface of the organ over time. We cannot begin to deduce the dynamic nature of organ deformation without first of all deriving a quantitative representation of such activity. In order to solve this, we used deformation based morphometry (DBM), an approach that has been widely adopted in the medical image analysis community [4],[8],[17],[21],[22]. Our DBM approach is composed of three major steps:

- Establish Point Correspondence between target surface T and source surface S, where T is the primordia surface at time t+1 and S is the same primordia surface at time t.
- Compute deformation field based on point correspondences.
- Quantify deformation by looking at the Jacobian at each point on the surface.

3.2.3 Finding Correspondences

Optimal correspondence can be achieved by solving the registration problem between surfaces S and T. This is done by addressing it as an optimization problem, where the goal is to minimize some error functional. Once we have our two point clouds properly registered, we chose as optimal, the correspondence that led to convergence in our optimization. The approach we employed is based on the well known Iterative Closest Point (ICP) algorithm for rigid registration of point sets [15]. Given the system we use for primordia growth data collection, surfaces generated from our data set are almost registered. This observation makes our data suited for use of the ICP algorithm, given its tendency to get stuck in local minima. Leveraging off the formalisms in [15], we now give a simple description of the basic ICP algorithm.

Given an initial state, our task is to minimize the sum of squared differences between the transformed target points and their correspondences on the source surface. This is done by minimizing the error function E_M, through an iterative process of computing correspondences and then optimizing the transformation parameters based on these correspondences.

$$E_M = min \sum_{i=1}^{n} [\|M(T_i) - S_i\|^2]$$

where E_M, is the error due to transformation M, and n is the number of correspondences and i is the current correspondence from a point in T to a point in S. Note that each point on the target has a correspondence to a point on the source surface.

The first step in the process of finding the optimal correspondence is to compute the initial transformation parameters between the two 3D point clouds representing our surfaces. Let the rotation parameter be r_c and the translation parameter be t_c. Given these initial or current state parameters, the next step is using the current state to find new correspondences $C_{(S_i,T_i)}$.

$$C_{(S_i,T_i)} = \arg\min[\|(r_c T_i + t_c) - s_i\|^2]$$

where $C_{(S_i,T_i)}$ is the i^{th} correspondence between T and S. Based on these new correspondences, update r_c and t_c, using the following equation

$$[r,t] = argmin[\left\|(r_c T_i + t_c) - S_{C_{(S_i,T_i)}}\right\|]$$

where $[r,t]$, is the updated rotation and tranlation parameters, based on previous state and correspondences computed from those previous values.

If the transformation error due to the updated transformation parameters is not less than a pre-set threshold, repeat the process over by computing new correspondences based on current transformation. The

correspondence that leads to convergence is selected as the optimal correspondence.

3.2.4 Quantifying Deformation

Once correspondence between points on surfaces at time t and t + 1 have been established, we use these correspondences to derive a displacement field. Displacement fields are represented in the form of displacement vectors. The structural difference between the two surfaces is encoded in this displacement field and this fact gives us a way to directly quantify deformation.

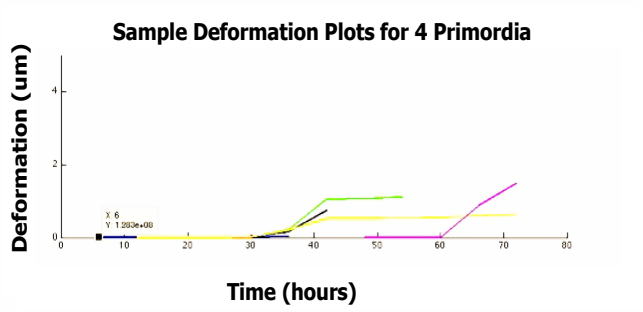

Fig. 6. Sample plots of total deformation computed from four different primordia. Total deformation is calculated by taking the sum of deformation values at points on the surface.

Given a displacement field from point correspondences, we quantify deformation by evaluating the local change at the resolution of each point on the surface. This is done by taking the determinant of the Jacobian matrix D at each surface point. Let U be the displacement vector at point $P(x, y, z)$ on the target surface. The deformation at P is defined as $\det(D)$, where D is defined as follows:

$$D_U = \begin{bmatrix} \frac{\delta U_x}{\delta x} & \frac{\delta U_x}{\delta y} & \frac{\delta U_x}{\delta z} \\ \frac{\delta U_y}{\delta x} & \frac{\delta U_y}{\delta y} & \frac{\delta U_y}{\delta z} \\ \frac{\delta U_z}{\delta x} & \frac{\delta U_z}{\delta y} & \frac{\delta U_z}{\delta z} \end{bmatrix}$$

where U is he displacement vector at the current point. A sample plot of total deformation over time for multiple primordia is shown in Fig. 6. As expected, primordia initially exhibit little or no deformation, but that changes at later stages of growth.

3.3 Growth Alignment

In our proposed system, output from the GCM is a time series of surface areas (or another growth measurement), computed for each primordia considered. Growth alignment as used in this paper refers to the temporal alignment of times series of growth measurements for multiple primordia. In all our experiments, we will be aligning growth data from multiple objects. However, the fundamental step in all these is the alignment of one growth time series against another. As such, our description of solutions for alignment of growth data will focus on alignment of two time series. After that, we will finally discuss how to apply these solutions as a sub-routine in the alignment of multiple time series.

Now we give a high level presentation of the alignment problem addressed here.

The Alignment Problem:

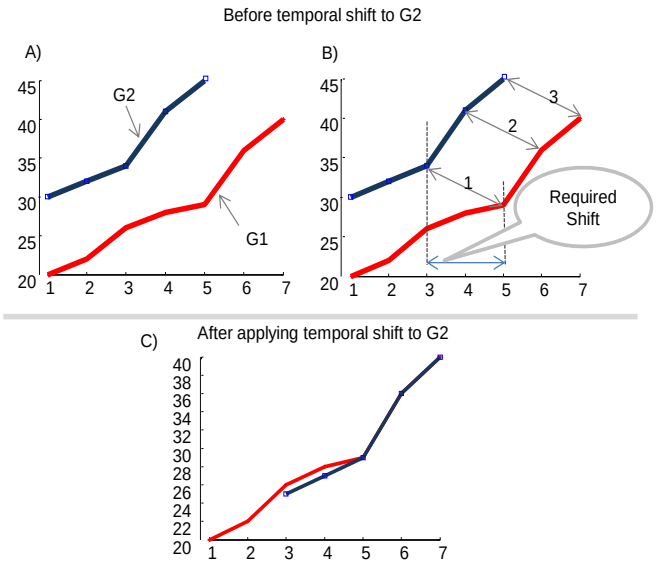

Fig. 7. The alignment problem we seek to solve involves finding the shift St necessary to bring G2 into an alignment with G1, such that some cost function is optimized.

Given two time series G_1 and G_2, the growth alignment problem is defined as the task of finding the temporal shift S_t, such that when applied to G_2, it leads to the best possible alignment of G_2 to G_1. As shown in Fig. 7.B, there are points in the two time series that are most similar and thus can be used as a basis for finding the required shift. Fig. 7.C shows the expected output, at least in this toy example, that is should come out of a good alignment. In this case, series G_2 has been shifted by a factor of two in the temporal direction. Note that in Fig. 7.A and Fig. 7.B, G_2 is plotted with a vertical shift of +5 to allow for easy display and illustration of the problem.

This definition of our alignment problem requires that the most similar subsequences between the two time series fall within the overlapping region that resulted from applying the shift S_t to G_2. What we have left now is to present our solution to this problem, given the application domain in consideration. We now present two solutions to the problem of aligning low temporal

resolution time series, with a focus on the alignment of plant growth data. The first approach is built on the Landscape Matching algorithm (LAM), while the second is a new parameter free alternative.

3.3.1 Landscape Based Alignment

We now give a brief review of the Landscape Matching (LAM) algorithm presented in [2]. This algorithm utilizes a landscape vector λ, calculated from the overlapping regions of the two series shown in Fig. 7.A, to calculate a matching score using the following cost function from normalized values of G_1 and G_2 :

$$M_{(G_1,G_2)} = argmax \ \frac{1}{L_1}\sum_{n=1}^{L^*}\frac{\left(1+\lambda^{\theta}{}_{(G_1^*,G_{2,s_t}^*)}(n)\right)}{1+|\ G_1^*(n)-G_{2,s_t}^*(n)\ |},$$

where L^* is the length of the overlapping region, λ is the landscape vector, L_1 is the length of G_1, s_t is the temporal shift of the current iteration, and G_1^* and G_2^* are the overlapping segments.

This algorithm considers all possible temporal shifts, with the goal of maximizing the landscape function $M_{(G_1,G_2)}$. As we shift G_2 across G_1 from left to right, starting with an initial overlap, we get different matching scores for resulting overlaps. The shift that leads to the highest matching score is selected as the required shift for the best possible alignment.

In Fig. 8, we show an example of calculating λ from the overlapping regions of two time series. Calculation of λ requires us to set a threshold parameter θ, within which we consider two corresponding points from the time series to be a match. The matching locations are then sorted in ascending order by time.

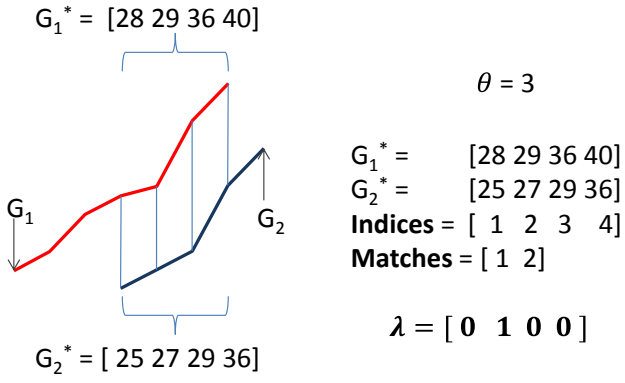

Fig. 8. Calculating the landscape vector. The threshold is set to 3 for this example.

The first position in λ and any other index location that is not in the list of matching positions is set to 0. For the rest of the index locations in the overlapping regions, the values are set to the absolute difference between the value at the index position in G_1, and the value of the prior matching index in G_1.

3.3.2 Multiple Features Landscape Based Alignment

The LAM algorithm requires that the threshold parameter be meaningful, since the result of the landscape function is sensitive to this parameter, and as such the algorithm can sometimes be sensitive to noise and outliers. With this motivation in mind, we decided to generalize the LAM function so that it can be used for alignment based on multiple growth features. The intuition of this generalization is that, provided other features have a property that significantly discriminates different developmental stages for example; we could leverage such property to improve the alignment due to the first feature. Consequently, a false positive due to noise from a single feature will be compensated by a strong match due to other features.

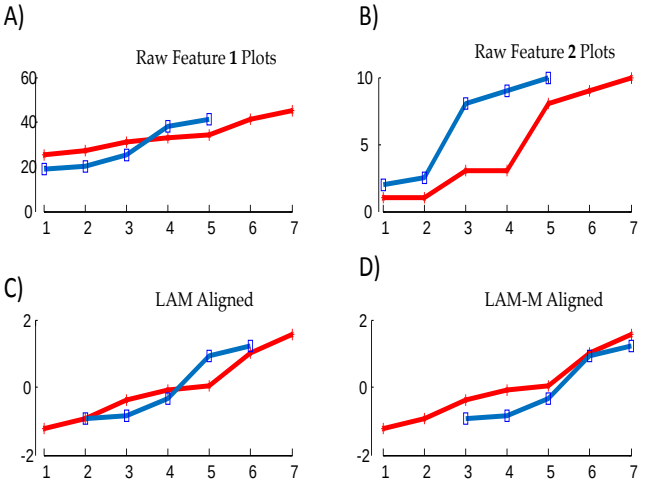

Fig. 9. Plots showing an example where single feature LAM algorithm fails to find the proper shift. By incorporating a second feature, LAM-M computed the correct shift.

For multiple features, we want to find the shift such that the cumulative landscape function of all features is maximized. For a particular feature f_i, we define the new landscape function for normalized values of the time series as follows:

$$m_{f_i} = \frac{1}{L_1}\sum_{n=1}^{L^*}\frac{\left(1+\lambda^{\theta}{}_{f_i,(G_1^*,G_{2,s_t}^*)}(n)\right)}{1+|\ G_1^*(n)-G_{2,s_t}^*(n)\ |}$$

The new cost function for multiple features is then defined as:

$$M_F = \prod_{i=1}^{k} m_{f_i}$$

This equation shows the new cost function for a multiple-feature LAM, which for ease of presentation, we refer to as LAM-M. The landscape vector for each fea-

ture is still calculated in the same way, and the goal now is to maximize the joint landscape of all the features.

In Fig. 9, we show an example where the basic LAM algorithm fails to properly align two fairly similar time series. In Fig. 9.B, we also show raw plots for a second feature, in this case deformation curves. In Fig. 9.C, we show alignment due to the original LAM algorithm, while in Fig. 9.D, we show how incorporation of multiple features leads to a correct alignment of the two time series.

3.3.3 *Parameter-Free Alignment*

Although the LAM algorithm works well, and our extension allows us to improve performance in the rare cases were LAM fails, we recognize two issues that could be problematic in some cases: first of all our data could be such that, the landscape function is too sensitive to the threshold parameter lambda. The second issue is that we might not have the luxury of a second meaningful corresponding feature. For these reasons, we now present a parameter free simple alternative to the LAM algorithm.

Our proposed parameter free algorithm looks to minimize the mean of Euclidean distances between subsequences in two time series G_1 and G_2. Unlike the LAM, it does not seek to analyze the entirety of the overlapping regions at any moment in time. It simply looks at the most similar subsequences by minimizing the mean of the Euclidean distances between corresponding points on the subsequences/motifs. Note that we use Euclidean distances here, but the distance metric could be substituted.

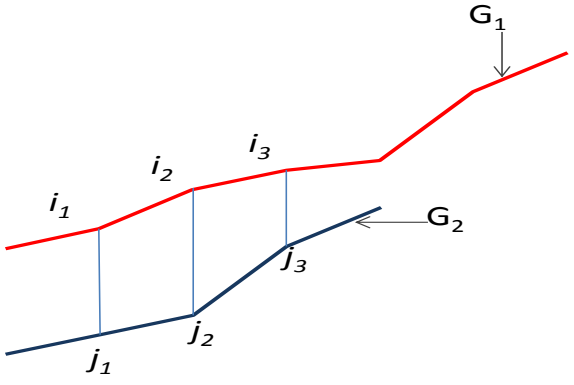

Fig. 10. Calculating the similarity score between two subsequences. Subsequences are labeled with i and j indices

Once we find the most similar subsequences, we calculate the shift necessary for the best possible alignment, based on the first matched indices of the subsequences. Because the most similar subsequences do not have to include the entirety of overlapping regions between two time series, this algorithm tends to be more robust to outliers that could throw off the landscape function in the LAM algorithm.

Given two time series G_1 and G_2, we find subsequences of all lengths from G_2, such that the maximum subsequence length is at most the length of the shorter of the two time series. Within this space of subsequence lengths, we find the closest pairs between G_2 and G_1. Let the length of subsequence g_2 be l_2. We now define closeness score $S_{(g_1,g_2)}$ as follows:

$$S_{(g_1,g_2)} = \frac{1}{L} \sum_1^L d(g_{1,i}, g_{2,i}),$$

where $d(g_{1,i}, g_{2,i})$ is the distance between the corresponding points in the subsequence.

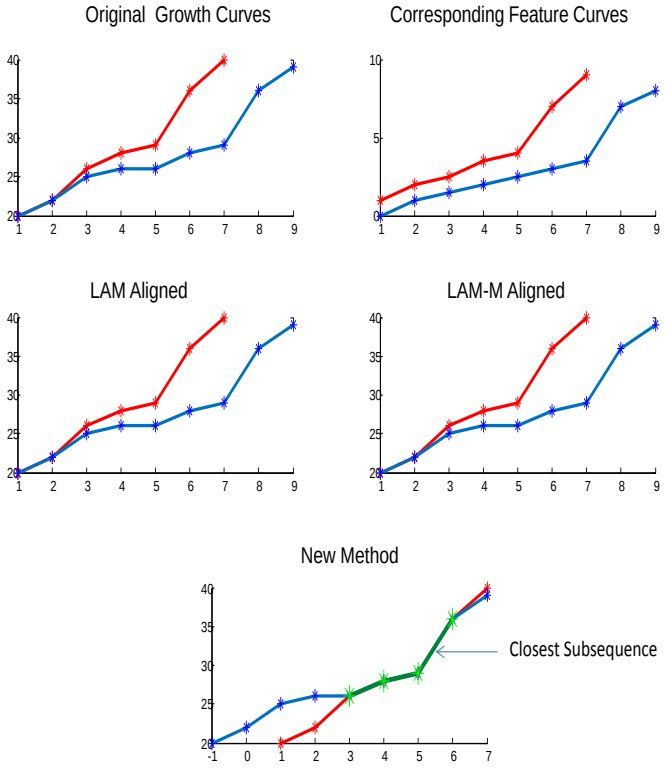

Fig. 11. In some cases, even the presence of a second feature might not be useful. As we see here, even LAM-M failed for this test data set. However, the new parameter free algorithm correctly aligned both time series.

In Fig. 10, we show an illustration of subsequence closeness scoring, for a subsequence of length 3. The process involves the summation of distances between corresponding points in subsequences being considered. In order to penalize matches that result from short sequences, we divide this sum by the length of the subsequence. We acknowledge that this penalty is

a somewhat naïve approach to favor longer subsequences. Others have presented intuitive approaches in the Minimal Description Length (MDL) space [18]. However, empirically our choice of simply taking the mean is extremely robust over many datasets. The matching pair that minizes the mean of the distances between corresponding points forms the basis for calculating the required shift S_t on G_2.

This parameter free algorithm sometimes outperforms LAM, even in the presence of multiple features. This is due to the fact that it relies on subsequences, as opposed to an entire overlapping region, and as such, it is less susceptible to outliers. Of course it is also true that there may exist a threshold at which LAM-M might do just as well.

3.4 Multiple Growth Alignment

Now that we have shown two alternatives to aligning two time series, we now show how we use them to align multiple time series. In the algorithm shown in Table 1, we assume the parameter free alignment algorithm. However, it can easily be changed to use the LAM-M optimization strategy instead.

This algorithm iteratively aligns the best possible pairs of time series, until there is no more time series to add to the aligned set. In Table 1.line 19, we make a call to our choice of alignment algorithm, in this case our parameter free algorithm. Lines 14 to 19 find the best match and shift between the current time series, and the rest of the time series that have yet to be aligned.

Note that this algorithm does not traverse the list of time series L, in order. For simplicity, we set the starting point to be the first time series in our list. This index location gets placed on the stack of touched time series in Table 1.line 7. In Table 1.line 15, we check to make sure we only compare against untouched time series. The time series that is best matched to the current source series becomes the source series in the next iteration and this is set in Table 1.line 26. In Table 1.line32 we consider the global context as we shift the current matches. The algorithm terminates once the list of touched indices is greater or equal to the length of input set of time series.

Table 1. Main Multiple Alignment Algorithm

Input:	*L* : list of time series
Output:	*S*: list of required shifts for time series in L.
1	Scores = {} // structure to hold similarity scores
2	S = [ ]
3	i = 1;
4	S(i) = 0;
5	C = [ ]; //empty set
6	**while** len(C) < len(L)
7	g = L(i)
8	C(len(C) + 1) = i;
9	**if** (~isset($Scores_i$.distance))
10	$Scores_i$.distance = inf
11	$Scores_i$.shift = 0
12	bestScore = inf
13	bestInd = i
14	bestShift = 0
15	**for** j from 1 => len(L)
16	**if** j ϵ C
17	continue
18	**end if**
19	h = L(j)
20	[shift score] = alignSeries(g,h)
21	**if**(score < bestScore)
22	bestScore = score
23	bestShift = shift
24	bestInd = j
25	**end if**
26	**end for**
27	i = bestInd
28	$Scores_i$.distance =bestScore
29	$Scores_i$.matchIndex = bestInd
30	$Scores_{bestInd}$.distance = Score
31	$Scores_{bestIndex}$.matchIndex = i
32	$Scores_{bestIndex}$.shift = $Scores_i$.shift + bestShift
33	S(i) = $Scores_i$.shift;
	end for

4 RESULTS AND ANALYSIS

We show alignment results on both synthetic and real data. We generated synthetic data to test the limits of the multi-feature landscape alignment (LAM-M). Synthetic data allows us to properly validate the alignment algorithm, since we are absolutely certain of the ground truth. In addition to aligning time series of primordial growth, we also show application of our algorithms on a longitudinal data set of individually marked guppies (*Poecilia reticulata*) in the wild.

4.1 Multiple-Features Landscape Alignment with Synthetic Data.

To generate synthetic data, we start with a single known time series, which we call the 'Main' pattern. From this pattern, we then generate multiple subpatterns of variable lengths by applying spatiotemporal random noise to different segments of the 'Main' pattern. This allowed us to create noisy samples that were sure to test the limits of our multiple features algorithm.

A plot of the unsynchronized version of the data is shown in Fig. 12. Note that the starting measurements of the generated growth curves are almost the same. This property can also be observed in the live imaging data. This is due in combination of the nature of primordia growth, as well as human error during imag-

ing, as well as the plant trying to adjust from the trauma caused by initial exposure to laser light. The fact that the generated data has this property makes it that more suitable for testing the multiple feature landscape alignment.

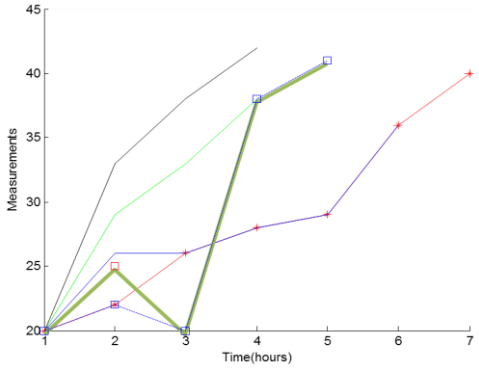

Fig. 12. Raw plot of the un-aligned synthetic data. This data has similar properties to those observed on real primordia growth data.

In Fig. 13., we show the results of aligning the raw time series using the LAM algorithm. Here we see that one series is seriously misaligned due to noise and outlying data points that cannot be overcome by the basic landscape function. To overcome this problem, we applied the LAM-M, to take advantage of the corresponding deformation values from the 'Main' trace and the incorrectly aligned curve.

As expected, the algorithm that incorporated the effects of multiple features was able to give a perfect alignment. Note that such results are only possible because this example shows the presence of a strong indicative second feature. As we showed before, there are cases where even the presence of a second feature is not helpful.

Fig. 13. Alignment of synthetic data using original LAM algorithm. Although LAM does fairly well, it failed to properly align one growth curve. We know this because we know the ground truth.

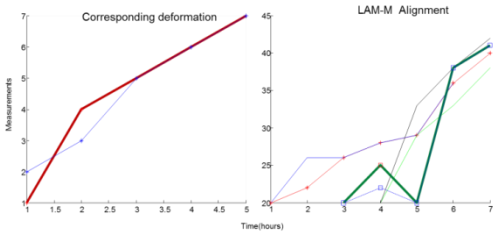

Fig. 14. Perfect alignment with LAM-M, using multiple features. (Left) Deformation plot for main trace and wrongly aligned series. (Right) Properly aligned set, due to strong second feature.

4.2 Results on Live Imaging Data.

Data used in this project was collected from a live imaging experiment that lasted up to 72 hours. Input into the system is a set of surface points for each primordia in a single plant over time. We use these surface points to compute the time series of surface areas for each primordia. For our experiment, we maintained a time interval of 6 hours between data points. An example of a single primordia input at a single time instance is shown in Fig. 5.right. Given a stack of images, primordia contours shown in Fig. 5.left can be automatically detected using available contour based feature detection techniques [24][25].

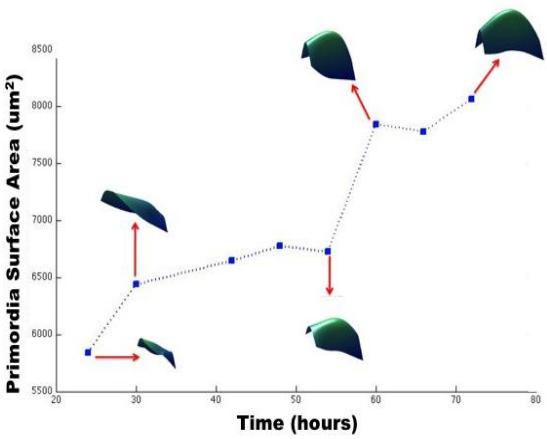

Fig. 15. Sample growth pattern, showing single primordia growth time series.

At the beginning of each live imaging experiment, developing primordia are at different developmental stages. At some point during the experiment, mature primordia that have been growing through out the observation period differentiate into other organs and exit the meristem. At that point we stop observing such primordia. Such exit is always accompanied by the emergence of new primordia. This behavior is the reason why in most of the results we show here, primordia will exhibit varying lengths of observation, a

situation that begs for robust time alignment. The rest of this section will present results of temporal alignment of time series of surface areas from a single plant, and then from multiple plants put together. In Fig. 15, we show an example of surface area growth pattern for one primordia. This figure shows the general pattern of primordia growth, which is multi-phasal. It is obvious from Fig. 15 that the sizes of the reconstructed surface area correlates with the calculated surface areas.

The SAM exhibits a great deal of variation within and across plants. As such, we tested our alignment algorithm on primordia from a single plant, as well as primordia from different plants put together. This experiment will test the stability of both the landscape based alignment and the parameter free alignment, which we call Minimum Mean of Distances alignment (MMD).

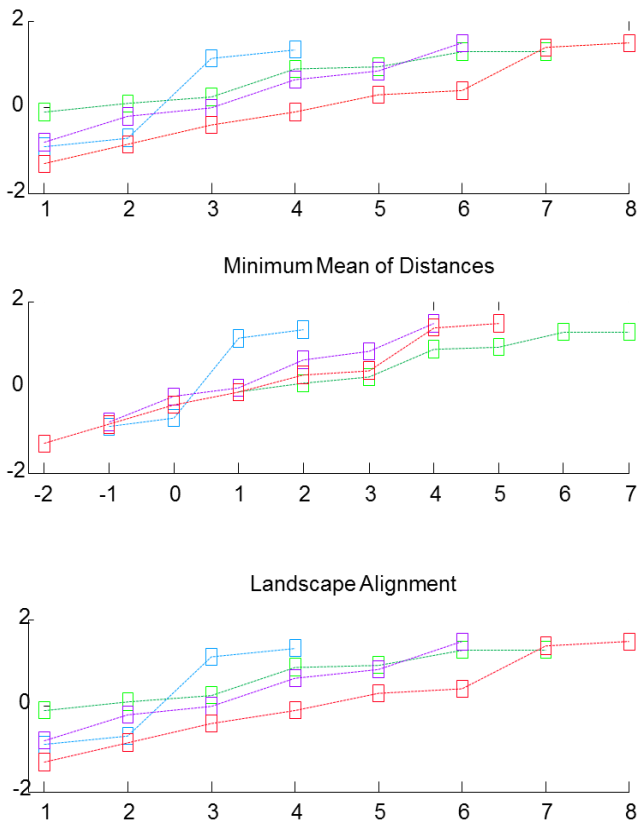

Fig. 16. Alignment of multiple primordia from a single plant. Raw growth curves (top). Alignment with parameter free algorithm labeled Minimum Mean of Distances (middle). Landscape based alignment (bottom). Even with a threshold parameter of 2, the landscape algorithm failed to find proper shifts.

Multiple Primordia from a Single Plant

The general nature of primordia growth and data collection introduces many opportunities for noise and outlier introduction. As such, the requirement for a threshold parameter is sometimes a limiting factor. As shown in Fig. 16.*bottom*, the alignment results from the landscape based alignment are no different from the original unsynchronized raw time series. This sensitive exhibited by the LAM based alignment occurred in multiple other plants. We show another example in Fig. 17.

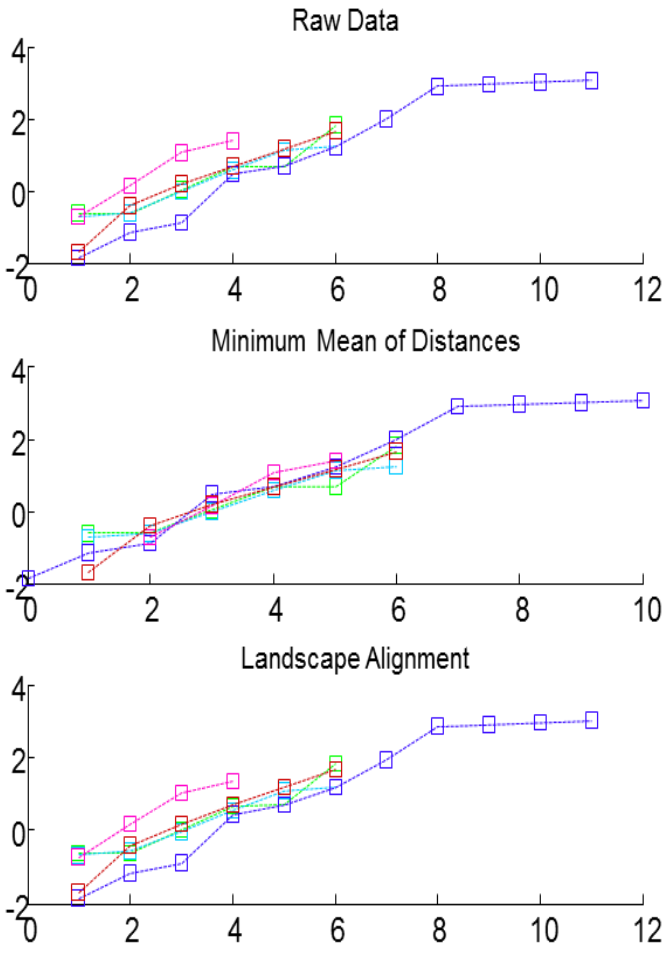

Fig. 17. Another example of parameter free Minimum Mean of Distances alignment being more robust than alignment based on landscape matching.

It is worth noting that LAM based alignment sometimes does yield reasonable results, especially in the presence of deformation. In Fig. 18, we show an example where both algorithms presented achieved reasonable alignments. Note that these results were obtained after computing the corresponding deformation of the primordia in question and applying the multiple feature version of LAM.

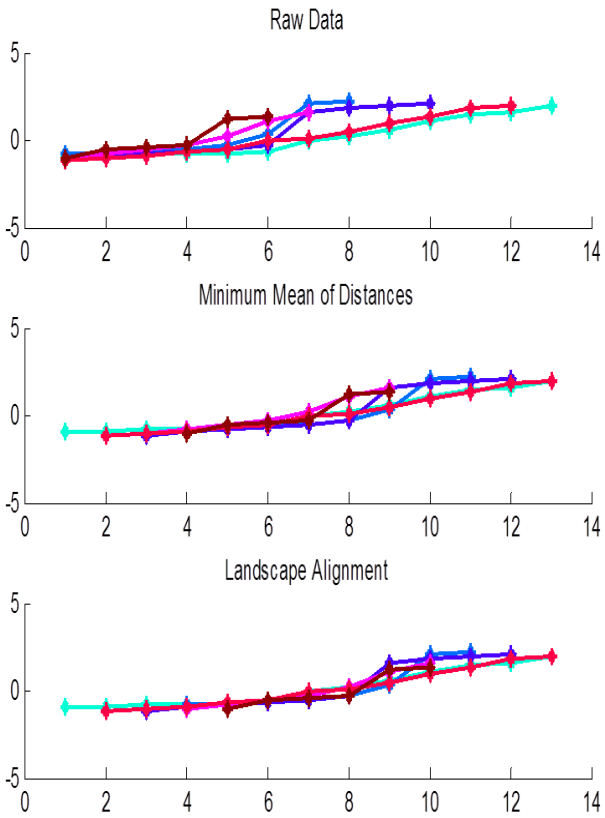

Fig. 18. Both algorithms yielding similar alignment results.

All these examples suggest that in the absence of a very strong second feature, our parameter free Minimum Mean of Distances (MMD) alignment algorithm is generally superior to the landscape based alignment. It is true that without a reasonable minimum subsequence length, MMD alignment could also fail, because it might elect very short sequences that give the minimum score, yet in the context of surrounding points, does not make for a good match. In this situation and in the presence of a strong second feature, the LAM-M multiple feature alignment will be the preferred option.

Alignment of Multiple primordia from multiple plants

In order to test the robustness of our algorithm, we show performance of our alignment algorithm when we mix primordia from different plants. For this experiment, we aligned a total of 24 primordia, using both algorithms presented Fig. 19. Even with expected variation across plants, we show that a robust alignment routine can achieve an acceptable alignment.

These results are very significant, especially in the study of the principles that govern meristem growth/maintenance. As mentioned earlier, live imaging limits the number of markers we can use at any given instance. As such it requires many plants to study a few genes. Proper alignment is the only way we can begin any accross plants analysis. Based on these results, LAM based alignment schemes seem to be very sensitive to growth type data like the multiphasal growth exhibited by primordia. LAM will work well for datasets like respiration data, since the patterns sort of periodic and variability between subjects is not as high as we see in meristem development. In such case it might be better to look at entire landscapes, rather than merely subsequences, even though one could argue that the space of overlapping landscapes is a subset of all possible subsequences.

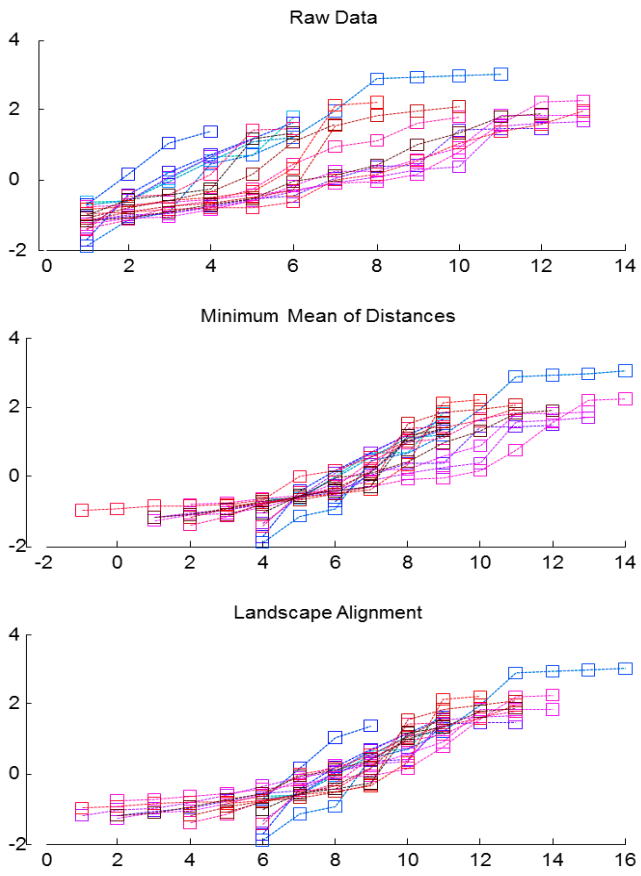

Fig. 19. Alignment of primordia collected from 5 plants. Total number aligned is equal to 24.

4.3 Results on Guppy evolutionary data.

The guppy data comes from a long-term project on rapid life-history evolution for which guppies were introduced in four isolated sections of streams in The Northern Range of the island of Trinidad. All guppies are marked with a unique combination of visible implant elastomer and are thus individually identifiable. The stream sections are censored monthly, when all individuals are caught, identified, and measured

(standard length and weight), and new recruits are given a new unique mark. Only individuals above 14mm of standard length are marked. Because birth is not observed, the age of individuals is unknown. A robust temporal alignment algorithm is therefore required for proper analysis. Data for this experiment was extracted from the time lapse average length measurements taken over a period of fifteen months. To show applicability, we randomly selected individuals, but made sure they were from the same cohorts or from cohorts that were not introduced into the ecosystem more than two months apart.

As we show in Fig. 20, the parameter free algorithm based on Minimum Mean of Distances achieved a very desirable alignment. Independent domain expert biologists who granted us the right to test on their data actually confirmed our alignment results as reasonable.

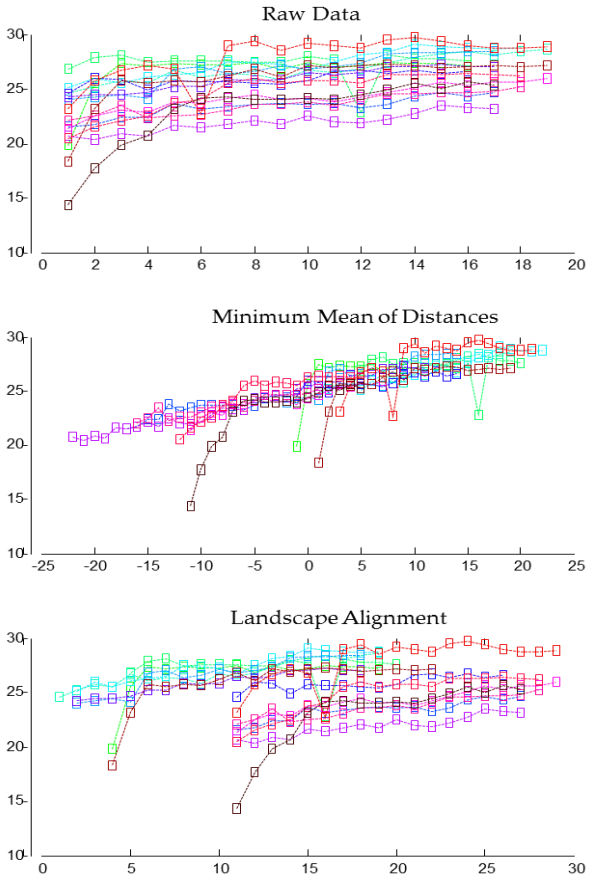

Fig. 20. Alignment of growth in length, of Trinidadian guppy. Each plot represents a single individual's growth pattern over time. The Y axis is the normalized lengths, while the x-axis represents time.

These results on guppy evolutionary data are significant, and present a new opportunity for biologist trying to study the effects of ecosystem conditions on adaptive evolution. Biologists now explore the possibility of estimating the age of individuals, or properly comparing individuals, in a way that is invariant to seasonal variation.

Given the variation in the data, it is no surprise that the parameter free MMD algorithm resulted in a more natural alignment. This is again another case where significant variation and noise poses a constraint on LAM based alignment.

5 Conclusion

We have proposed a framework for measurement of growth features and alignment of unsynchronized primordia growth data from live imaging experiments. To move closer to the big picture goal of developing a dynamic gene expression atlas, a necessary first step is the development of robust temporal growth alignment algorithms. We presented an extension to an already reliable alignment algorithm (LAM), as well as presented a parameter free alternative to growth alignment. The extension to the LAM algorithm to allow for multiple feature optimizations allowed us to overcome unique challenges presented in data with increased noise. However, this extension is sensitive to outliers and high variation. We have shown cases where LAM-M outperforms LAM basic, and also cases where even LAM-M fails, yet our parameter free algorithm succeeds. To demonstrate the diversity of our algorithm, we also showed results on guppy evolutionary growth data. This tells us that in the presence of high variation and increased noise, our parameter free MMD alignment algorithm is superior to LAM based alignment algorithms. A robust alignment of time lapse data gives developmental biologist the ability to begin to deduce the causal relationship between gene expression, cell behaviors and organ growth. Such an analysis would not be reliable without an alignment framework such as the one we proposed.

Acknowledgment

The authors were partially supported by National Science Foundation ChemGen IGERT Grant No. DGE 0504249 and funded in part by grant IOS-0718046 to GVR.

The authors thank David Reznick and Andrés López-Sepulcre for access to the guppy data and help with its interpretation. The guppy work was funded by a NSF Frontiers in Integrative Biological Research grant (EF0623632) to D. Reznick

References

[1] Braude, I., Marker, J., Museth, K., Nissanov, J., and Breen, D. Contour-based surface reconstruction using mpu implicit models. Graphical Models (2007).

[2] Chen, L., McKenna, T., Gribok, A., and Reifman, J. Lam: A land-

scape matching algorithm for respiratory data alignment. SPIE The International Society for Optical Engineering (2006).

[3] Christin, C., Hoefsloot, H., Smilde, A., Suits, F., Bischoff, R., and Horvatovich, P. Time alignment algorithms based on selected mass traces for complex lc-ms data. Journal of proteome research (2010).

[4] Chung, M., Worsley, K., Robbins, S., Paus, T., Taylor, J., Giedd, J., Rapoport, J., and Evans, A. Deformation-based surface morphometry applied to gray matter deformation. Neuroimage (2003).

[5] Dumais, J., and Kwiatkowska, D. Analysis of surface growth in shoot apices. The Plant Journal (2002).

[6] Eid, A., and Farag, A. On the performance evaluation of 3d reconstruction techniques from a sequence of images. EURASIP J. Appl. Signal Process. (2005).

[7] Fernandez, R., Das, P., Mirabet, V., Moscardi, E., Traas, J., Verdeil, J., Malandain, G., and Godin, C. Imaging plant growth in 4d: robust tissue reconstruction and lineaging at cell resolution. Nature Methods (2010).

[8] Gaser, C., Nenadic, I., Buchsbaum, B., Hazlett, E., and Buchsbaum, M. Deformation-based morphometry and its relation to conventional volumetry of brain lateral ventricles in mri. NeuroImage (2001).

[9] Guilleminault, C., Poyares, D., Rosa, A., and Huang, Y. Heart rate variability, sympathetic and vagal balance and eeg arousals in upper airway resistance and mild obstructive sleep apnea syndromes. Sleep Med (2005).

[10] Keogh, E., and Pazzani, M. Derivative dynamic time warping. First SIAM International Conference on Data Mining (2001).

[11] Khairy, K., and Keller, P. Reconstructing embryonic development. Genesis (2011).

[12] Listgarten, J., Neal, R. M., Roweis, S. T., and Emili, A. Multiple alignment of continuous time series. In Advances in Neural Information Processing Systems (2005).

[13] Liu, M., Yadav, R., Roy-Chowdhury, A., and Reddy, G. Automated tracking of stem cell lineages of arabidopsis shoot apex using local graph matching. The Plant Journal (2010).

[14] [14] Lorensen, W., and Cline, H. Marching cubes: A high resolution 3d surface construction algorithm. SIGGRAPH Comput. Graph. (1987).

[15] P.J. Besl., and N.D. McKay. A method for registration of 3-d shapes. Trans. PAMI (1992).

[16] C. S. Myers and L. R. Rabiner. A Comparative study of several dynamic time-warping algorithms for connected-word recognition. The Bell system technical journal (1981).

[17] Pieperhoff, P., Su¨dmeyer, M., Ho¨mke, L., Zilles, K., Schnitzler, A., and Amunts, K. Detection of structural changes of the human brain in longitudinally acquired mr images by deformation field morphometry: Methodological analysis, validation and application. NeuroImage (2008).

[18] Rakthanmanon, T., Keogh, E., Lonardi, S., and Evans, S. Time Series Epenthesis: Clustering Time Series Streams Requires Ignoring Some Data. ICDM (2012).

[19] Reddy, G., and Meyerowitz, E. M. Stem-cell homeostasis and growth dynamics can be uncoupled in the arabidopsis shoot apex. Science (2005).

[20] Reddy, G. V. Real-time Lineage Analysis reveals Oriented Cell Divisions Associated with Morphogenesis at the Shoot Apex of Arabidopsis thaliana. Development 2004, 131: 4225-4237.

[21] Rohlfing, T., Sullivan, E. V., and Pfefferbaum, A. Deformation-based brain morphometry to track the course of alcoholism: Differences between intra-subject and inter-subject analysis. Psychiatry Res (2006).

[22] Soares, J., Kochunov, P., Monkul, E., Nicoletti, M., Brambilla, P., Sassi, R., Mallinger, A., Frank, E., Kupfer, D., Lancaster, J., and Fox, P. Structural brain changes in bipolar disorder using deformation field morphometry. Neuroreport (2005).

[23] Studholme, C., Cardenas, V., Blumenfeld, R., Schuff, N., Rosen, H., Miller, B., and Weiner, M. Deformation tensor morphometry of semantic dementia with quantitative validation. NeuroImage (2004).

[24] Tataw, O., Liu, M., Roy-Chowdhurry, A., Yadav, R., and Reddy, G. Pattern analysis of stem cell growth dynamics in the shoot apex of arabidopsis. ICIP, IEEE International Conference on Image Proccesing (2010).

[25] [Tsai, D.-M., Hou, H.-T., and Su, H. J. Boundary-based corner detection using eigenvalues of covariance matrices. Elsevier - Pattern Recognition Letters 20 (1999), 31–40.

[26] Yadav, R., Girke, T., Pasala, S., Xie, M., and Reddy, G. Gene expression map of the arabidopsis shoot apical meristem stem cell niche. PNAS (2009).

[27] Zhang, Y., and Edgar, T. F. A robust dynamic time warping algorithm for batch trajectory synchronization.

[28] Zwiener, U., Schelenz, C. h., Bramer, S., and Hoyer, D. Short-term dynamics of coherence between respiratory movements, heart rate, and arterial pressure fluctuations in severe acute brain disorders. Physiol Res (2003).

Oben M. Tataw is a PhD candidate in the department of computer science and engineering, at the University of California, Riverside. He is part of the data mining lab and the UCR Video Computing Group. He is also a Chem-Gen IGERT fellow.

G. Venugopala Reddy is an assistant professor of Cell Biology in the department of Botany and Plant Sciences. In 1999, he received the Jane Coffin Childs Memorial Fund for Medical Research Foundation post doctoral fellowship to pursue research at California Institute of Technology, Pasadena, USA, where he co-developed live-imaging method to study the dynamics in the shoot apex of Arabidopsis.

Eamonn J. Keogh is a full professor of computer science at the University of California Riverside. His research areas include data mining, machine learning and information retrieval, specializing in techniques for solving similarity and indexing problems in time-series datasets. He received best paper awards in IEEE ICDM 2007, SIGMOD 2001 and SIGKDD 2012. He has given over two dozen well received tutorials in the premier conferences in data mining and databases.

Amit K. Roy-Chowdhury is an Associate Professor of electrical engineering and a Cooperating Faculty in the Department of Computer Science, University of California Riverside. He leads the Video Computing Group at UCR. His group is involved in research projects related to camera networks, human behavior modeling, face recognition, and bioimage anaysis. He has been a principal investigator on a number of projects from the US National Science Foundation, Office of Naval Research, the National Endowment for the Humanities, as well as private companies, among others. He is the author of the book, Camera Networks, which is the first monograph providing an overview of research in this area.