Information Surfaces in Systems Biology and Applications to Engineering Sustainable Agriculture

Hesam Dashti[1], Alireza Siahpirani[2], James Driver[1], and Amir H. Assadi[1]

[1] Department of Mathematics, University of Wisconsin, USA
[2] Department of Electerical and Computer Engineering, University of Wisconsin, USA
{Dashti,Fotuhisiahpi,Driver,Ahassadi}@wisc.edu

Abstract. Systems biology of plants offers myriad opportunities and many challenges in modeling. A number of technical challenges stem from paucity of computational methods for discovery of the fundamental properties of complex dynamical systems in biology. In systems engineering, eigen-mode analysis has proved to be a powerful approach to extract system parameters. Following this philosophy, we introduce a new theory that has the benefits of eigen-mode analysis, while it allows investigation of complex dynamics prior to estimation of optimal scales and resolutions. *Information Surfaces* organize the many intricate relationships among "eigen-modes" of gene networks at multiple scales. Via an adaptable multi-resolution analytic approach, one could find the appropriate scale and resolution for discovery of functions of genes in plants. This article pertains the model plant Arabidopsis; however, almost all methods can be applied to investigate development and growth of crops for research on sustainable agriculture.

Keywords: Dynamical Systems, Multiscale Analysis, Multiresolution Analysis, Eigen Analysis.

1 Introduction

The concept of dynamical systems has been proposed to investigate natural and synthetic time-dependent systems. Poincare first introduced dynamical systems to study the qualitative aspects of orbits in celestial mechanics [1]. The theory of dynamical systems has been extended to model broader classes of systems whose time-evolution may or may not have periodic orbits [1][2] [3] [4] [5].

The numerical study of dynamical systems is focused on modeling the current state of the system [3] for data mining purposes (i.e. supervised and unsupervised classification) [6]. On the other hand, one can argue the need for models that explain the potentially complex relationships among two or more systems [7]. In this direction, we introduce a measurement for quantifying the distance between two dynamical systems. We illustrate the utility and technical power by application of the theory to time-series of gene expression profiles. The data set is comprised of a set of genes stored in rows along time-steps corresponding to expression values in columns.. To analyze such arrays, we introduce the method of "*InfoSurf*"s in accordance to the

L.M. Camarinha-Matos et al. (Eds.): DoCEIS 2012, IFIP AICT 372, pp. 77–84, 2012.

three well-known mathematical theories, namely, multiscale analysis [8], multiresolution analysis [9], and Eigenanalysis [10]. The corresponding algorithms are implemented on a high-performance computing (HPC) platform. Briefly, the algorithm considers a two-dimensional array consisting of 'm' observations in the rows and 'n' time points in the columns. Clearly, when we regard the value of the i^{th} observation at the j^{th} time point as the height (z-coordinate) of a point ($x=i$, $y=j$) in a three-dimensional Euclidean space, then we would obtain a surface. In InfoSurf theory, one extracts the entries of the 2-dimensional array from eigenvalues of suitable operators, as outlined in Section 2.

In section 3, computational steps of the InfoSurf method are illustrated. In section 4 we apply InfoSurf's to a biological dynamical system associated to the *Arabidopsis Thaliana*.

2 Contributions to Value Creation

Sustainable agriculture is regarded as a domain that can greatly benefit from transformative innovations in molecular and cellular plant biology. Molecular methods in biotechnology and agricultural engineering promise rapid breeding of new lines of crops that would sustain stress from global warming and other harsh climatic events. Success of molecular methods depends on breakthroughs in molecular systems biology, and invention of new ways of understanding the complex dynamics formed by time-course data from genes, proteins and other biomolecules. The technical demand for development of new algorithms to surmount the present computational challenges requires re-examination of traditional methods that have proved successful in non-complex systems and their dynamics. In particular, researchers must address discovery of the necessary biological properties implicit in –omic data, and mine the abundance of dynamical features that could be observed only in appropriate scales and via optimal resolutions.

This research addresses some of the bottlenecks that are posed in providing effective applications of systems biology to sustainable agriculture. Thus, the applications of this research will contribute towards value creation and directly addressing critical scientific problems that face humankind today.

3 Method

One of the novelties of InfoSurf theory is that it provides a new representation for "*global information contents*" in a dynamical system that could be localized in a heterogeneous manner. InfoSurf's allow such information contents (in the sense of Shannon) in a discretized dynamical system (M_{mxn}) to be considered as a surface in three-dimensional Cartesian coordinates, where appropriately defined estimates of (Shannon) information are assigned to the entries in rows and columns of the matrix constructed from the dynamics. In the case of gene expression time-series, the dynamical systems matrix consist of m rows (genes) with n columns (expression values sampled at time points), and typically $m \gg n$ for whole genome or a similar

High Throughput experimental assay due to that the high cost of performing experiments for each period of time[11][12]. On the other hand, in the time-series that we study, smooth interpolation of the few number of time points enable us to include a greater number of finer-scale and finer-resolution attributes for situations that the time-series implicitly encode such information about the dynamics [13].This method is a row-wise interpolation, and the choice of the algorithm is based on the regularity properties that are required from various real-valued or vector-valued functions. Further, regularity of the interpolation functions is important to ascertain the smoothness of the corresponding "information surfaces". Also, an InfoSurf requires regularity in how different columns are arranged in relative position (column-wise regularities). To achieve such regularity, an InfoSurf sorts the objects based on three features: the area underneath the curve for (a) the signal (a row), (b) its first derivative (speed of change), and (c) its second derivative (concavity). With these preliminary steps in mind, an InfoSurf is a transformation of the dynamical system onto a piecewise smooth surface (possibly without information loss, if so-desired, or according to estimates for lossy transformations) through multiscale and multiresolution analysis of singular value decompositions (SVD) of the numerous matrices that arise in the process.

3.1 Multiscale Analysis

Multiscale methods are used more commonly in recent years due to advances in computational speed that allow running parallel tasks for each scale simultaneously, as well as other hardware advances. In addition, an increasing number of biological modeling problems rely on disparate mathematics to describe phenomena at different spatial and temporal levels. Multiscale analysis [8] provides a bridge between these levels. Further, it allows one to analyze phenomena that are interdependent, to make their relationship explicit, and provide a synthesis of heterogeneous scales that might otherwise be impossible or too difficult to properly describe within the scope of a single model. Particularly in systems biology, biomolecular reactions occur at different rates (scales) and must be estimated at appropriate resolution that varies according to scale. In our setting,, multiscale analysis plays an important role to analyze data at different levels for biological realistic modeling, and as a result, requires us to identify new phenomena at different scales that may otherwise go undetected. This ability is especially important for the Arabidopsis systems biology, because the size of its genome is quite large (about 30,000 genes and other significant non-coding RNAs, or perhaps more.) To perform multiscale analysis on a dynamical system (M_{mxn}), InfoSurf theory considers a sliding window, a sub-matrix S, of size $k \times k$, $2 \leq k \leq \min(m, n)$, of M_{mxn}. The size of S varies between the construction of different surfaces but remains invariant for the entire surface under consideration and for the comparison of two surfaces as will be described later in this section. The sub-matrix slides in two directions; the first sub-matrix is defined by $S = \mathrm{M}(1{:}\,k, 1{:}\,k)$ (left-top), and slides to right and down by one in every iteration. The following pseudo-code illustrates the process:

```
for i=1:m-k
    for j=1:n-k
          S = M(i:i+k, j:j+k)
          //Performing analysis on S
    end
end
```

One finds that this process projects the matrix M_{mxn} to a super-matrix containing $(m\text{-}k)$x$(n\text{-}k)$ sub-matrices of dimension KxK. Overlaps in the sub-matrices reveals the continuous influence of objects on other groups of objects and allows for the method to proceed continuously and reveal information between data points that would otherwise be unaccounted for. Considering every point in different windows illustrates the effect of an object on other object/objects and is seen multiple times as the object remains in the sliding window. This amplifies the (probabilistic) effect(s) of the object and allows for it to be observed in different sliding windows. This allows for easier identification of an object and increases the accuracy of the algorithm when analyzing a dynamical system.

3.2 Multiresolution Analysis

Multiresolution analysis allows for larger features of a system to be reduced to the relationships of its fine features. For an example, in a gene expression time-series it allows for the detection of groups of genes that are potentially up or down regulated with respect to one another when verified through relevant biological data. Through use of surfaces, one can observe patterns of gene activity and reduce the macroscopic picture to the action of the individual genes responsible. Considering subsets of genes through different resolutions increases the accuracy of the InfoSurf algorithm. The different resolutions of InfoSurf are characteristic of the sliding window (S_{kxk}) described in the previous section. Through use of this window, InfoSurf's detection capabilities are increased and it allows for the extraction of specific attributes of genes and the construction of their interrelationships.

Starting multiscale analysis at a larger scale, larger k for the size of the sliding window, allows the algorithm to identify regions of differences of two dynamical systems. InfoSurf uses the multiresolution process to zoom into the regions with very fine sliding windows (smaller values of k) and identify the specific objects corresponding to the differences between the dynamical systems. It provides the ability to capture relationships between groups of objects (coarse scale) and tune it to identify relationships between the objects (fine scale) [14].

3.3 Eigen Analysis

Eigen analysis is a fundamental method of data analysis and the investigation of structural properties of datasets. The use of Eigen analysis in the InfoSurf algorithm was inspired by the kinematics of surface deformation as described in [15]. This analysis of InfoSurf is conceptually similar to what introduced in [16] in analysis of

neuronal activation data in experiments on rat Anterior Cingulate Cortex in pain research, and in [17] for MEG data of human brain for detection of activated brain regions by measuring the starting point and estimate on length of time of the magnetic fields generated by neuronal spiking and ion transport properties.

For every invariant sliding window (section 3.1), InfoSurf computes its eigenvalues and eigenvectors. The eigenvalues of a sliding window represents a) the heights of the surface. The distribution of eigenvalues is representative of the number of eigenvectors needed to reconstruct *S*. Since the number of necessary eigenvalues for reconstruction of a surface depends on the smoothness of the surface, b) eigenvalues can be used to represent the smoothness of the surface. For every dynamical system the InfoSurf computes the eigenvalues of every sliding window. Since the sliding window iterates in two-dimensions the eigenvalues are stored in entries of a matrix, E. For the r^{th} row and s^{th} column iterations, InfoSurf associates the absolute value of the sum of the eigenvalues to E(*r*, *s*). The matrix E is called an Eigensurface and represents the internal properties of data. After constructing the Eigensurface, the InfoSurf method calculates the first and second derivatives of the Eigensurface. These derivatives are useful for identifying circadian clock information of dynamical systems [18]. While the first derivative is characteristic of the slope of the change of the eigenvalues, relating the change of the information content of each window and the objects within it, the second derivative provides information on the concavity, or acceleration of changes, circadian clock, and shows whether a subset of objects within each window is having a larger or smaller effect as time progresses. After constructing the representative surfaces, the InfoSurf measures the dissimilarity between the dynamical systems. To compare two dynamical systems, the InfoSurf generates seven surfaces: a) distance of the Eigensurfaces, the surfaces of the first derivatives, and surfaces of the second derivatives. b) free-scale distance of the three representative surfaces, and c) the Jacobian matrix. The distance is the absolute value of the direct subtraction of two matrices (surfaces): $Dist(A,B) = abs(A-B)$, and the scale-free distance is defined $FreeDist(A,B) = \frac{abs(A-B)}{abs(A)+abs(B)}$. The distance surfaces show differences/similarities of the dynamical systems. Figure 1 shows different steps of the InfoSurf method.

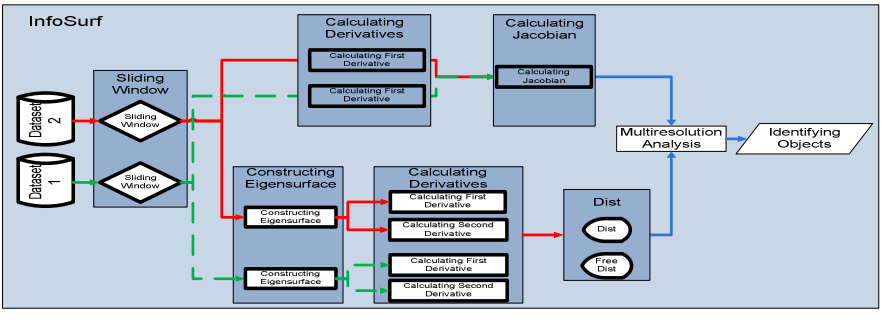

Fig. 1. The InfoSurf workflow diagram. This figure shows the flow of data in the algorithm.

After calculating the Distance, Free Distance surfaces, and finding the largest differences between the derivative surfaces, multiresolution analysis is applied to locate the biggest differences in behavior of Eigensurfaces, which in turn, focuses with a finer resolution and more accuracy towards an area in the original data set that caused these differences.

4 Discussion and Experimental Results

To evaluate the InfoSurf algorithm we used two dynamical systems from the Arabidopsis Diurnal Rhythms experiment [19]. The data represents gene expression levels of *Arabidopsis Thaliana* when stimulated with changes in temperature and light. Samples of 22,810 genes were taken in 4 hour intervals over a period of 48 hours. The first experiment consists of exposing the plants to constant light and 22^C in temperature for 12 hours, and then 12 hours of darkness with a 12^C temperature. The second experiment was from plants that were exposed to light during the entire experiment while the change in temperature was the same as the previous experiment. The first data set is called LDHC (Light, Dark, Hot and Cold) and the second data set is called LLHC (Light, Light, Hot and Cold). Each data set has genes in 22,810 rows (genes) and 12 columns (the four hour time steps in the experiments). We interpolated the data sets by the cubic SPLINE method row-wise, and then resampled uniformly to obtain 100 time points for each gene expression.

To acquire a smoother starting surface, we sorted the genes through row exchange based on the similarity of their time series (i.e. expression values). If we denote the time series of a gene by f(t), we consider the value of $g(x) = \int f(x) + \int f'(x) + \int f''(x)$ to be a good representation of the shape of the signal. The integrations are calculated by the trapezoidal approximation. We sort the LDHC data set and apply the algorithm to obtain the control surface; LDHC) and rearrange the second data set LLHC to impose the same order of genes in both data sets. The deformed surface corresponds to LLHC.

To find the genes that have different dynamic behavior in the two data sets, we considered the differences between the second derivatives of the two Eigensurfaces with the sliding window of size 40, and found the local extrema. These points represent a window of 40x40 in the original data sets (40 genes in 40 time steps) whose eigenvalues are different in the two data sets. To refine the selection of genes, we used a higher resolution sliding window (20x20) inside of the 40x40 matrix. Then the Eigensurface is constructed and the second derivative is calculated in order to elucidate a better understanding of the genes' dynamic behavior within their group and at a finer resolution. This leads to a 20x20 window in the original data that includes the local extrema. We further continued increasing the resolution by using the Eigensurfaces of the10x10 and 5x5 sliding windows which yield a 5x5 area (5 genes in 5 time steps). Algorithm 1 delineates these steps. We considered these 5 genes as potential candidates for being the culprit for the differences in the Eigensurfaces. We then looked up the phenotypic traits attributed to these genes using DAVID (the Database for Annotation, Visualization and Integrated Discovery[20]) to check their functionality, and found a gene whose functionality is related to the response of temperature or light stimulus. Due to the large amount of data we ran our program on high performance computing facilities of the Keeneland project [21]. The time required to run the MATLAB code that implemented the algorithm on our HPC

Cluster (64 nodes AMD Athlon 2.8 GHz and 32 GB RAM) exceeded 24 hours. This computing time was reduced to 3 hours once we implemented the algorithm for the Keeneland HPC platform. David listed "response to temperature stimulus" and "response to cold" as one of the functionalities of gene AT3G49910 (252235_at), and listed response to "light stimulus" and "response to light intensity" for AT2G06850 (266215_at). Output of analyzing these data is shown in a supplementary data at (http://vv811a.math.wisc.edu/InfoSurf).

Algorithms 1.

```
1- A ← interpolated LDHC; B ← interpolated LLHC.
2- Sort A according to similarity of signals; Rearrange B in the same
order.
3- eigA ← Eigensurface of A; eigB ← Eigensurface of B (window size 40).
4- D1A ← first derivative of eigA; D1B ← first derivative of eigB.
5- D2A ← second derivative of eigA; D2B ← second derivative of eigB.
6- Delta ← D2A-D2B.
7- E ← The local extrema of Delta.
8- for each point "e" in E, do the following:
    8.1- W2A, W2B ← 40x40 window from A and B that starts from
coordinates of e.
    8.2- Delta2 ← difference of second derivatives of Eigensurface of
            W2A and W2B (with sliding window of size 20).
    8.3- E2 ← The local extrema of Delta2.
    8.4- consider "e2" to be maximum of E2.
    8.5- W3A, W3B ← 20x20 window from A and B that starts from
coordinates of e2.
    8.6- Delta3 ← difference of second derivatives of Eigensurface of
            W3A and W3B (with sliding window of size 10).
    8.7- E3 ← The local extrema of Delta3.
    8.8- consider "e3" to be maximum of E3.
    8.9- W3A, W3B ← 10x10 window from A and B that starts from
coordinates of e3.
    8.10- Delta3 ← difference of second derivatives of Eigensurface of
            W3A and W3B (with sliding window of size 5).
    8.11- E3 ← The local extrema of Delta3.
    8.12- consider "e3" to be maximum of E3.
    8.13- select genes in the 5x5 window that starts from coordinates of
            e3, as possible candidates.
```

Acknowledgments. The authors thank Professor Joanne Chory for providing the data sets and discussion about the biological problem. We thank personnel of the "Keeneland: National Institute for Experimental Computing" for their kind supports. This material is based upon work supported by the National Science Foundation under Grant No. 0923296. This projected is partially supported by the National Institute of Health under Grant No. EY21357.

References

1. Poincare, H., Magini, R.: No Title. Il Nuovo Cimento 10, 1895–1900 (1899)
2. Hannon, Bruce, Ruth, Matthias: Modeling Dynamic Biological Systems, http://www.springer.com/life+sciences/ecology/book/978-0-387-94850-8

3. Hari Rao, V.S.: Differential Equations and Dynamical Systems, http://www.springer.com/mathematics/journal/12591
4. Kaneko, K., Furusawa, C.: Consistency principle in biological dynamical systems. Theory in Biosciences = Theorie in Den Biowissenschaften 127, 195–204 (2008)
5. Alicki, R., Fannes, M.: Quantum Dynamical Systems. Oxford University Press, USA (2001)
6. Wingate, D., Singh, S.: Kernel Predictive Linear Gaussian models for nonlinear stochastic dynamical systems. In: Proceedings of the 23rd International Conference on Machine Learning - ICML 2006, pp. 1017–1024. ACM Press, New York (2006)
7. Mehta, P.G.: The Kullback–Leibler Rate Pseudo-Metric for Comparing Dynamical Systems. IEEE Transactions on Automatic Control 55, 1585–1598 (2010)
8. Gao, J., Cao, Y., Tung, W.-W., Hu, J.: Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond. Wiley-Interscience (2007)
9. Rohwer, C.: Nonlinear Smoothing and Multiresolution Analysis. International Series of Numerical Mathematics. Birkhäuser, Basel (2005)
10. Sehmi, N.S.: Large order structural eigenanalysis techniques: algorithms for finite element systems. John Wiley & Sons Inc. (1989)
11. Androulakis, I.P., Yang, E., Almon, R.R.: Analysis of time-series gene expression data: methods, challenges, and opportunities. Annual Review of Biomedical Engineering 9, 205–228 (2007)
12. Ernst, J., Bar-Joseph, Z.: STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics 7, 191 (2006)
13. Bar-Joseph, Z., Gerber, G.K., Gifford, D.K., Jaakkola, T.S., Simon, I.: Continuous representations of time-series gene expression data. Journal of Computational Biology: a Journal of Computational Molecular Cell Biology 10, 341–356 (2003)
14. Nicholson, H.: Modelling of Dynamical Systems. IEE control engineering series. Inspec/Iee (1980)
15. Lai, W.M., Rubin, D., Krempl, E.: Introduction to Continuum Mechanics. Elsevier (2009)
16. Fallahati, D.M., Backonja, M., Eghbalnia, H., Assadi, A.H.: Dynamic PCA for network feature extraction in multi-electrode recording of neurophysiological data in cortical substrate of pain. Neurocomputing 44-46, 401–405 (2002)
17. Wang, L., Baryshnikov, B., Eghbalnia, H., Assadi, A.H.: Extraction of nonlinear features in MEG and fMRI data of human brain. Neurocomputing 52-54, 683–690 (2003)
18. Aase, S.O., Ruoff, P.: Semi-algebraic optimization of temperature compensation in a general switch-type negative feedback model of circadian clocks. Journal of Mathematical Biology 56, 279–292 (2008)
19. Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., Givan, S.A., Yanovsky, M., Hong, F., Kay, S.A., Chory, J.: Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genetics 4, e14 (2008)
20. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4, P3 (2003)
21. Keeneland: National Institute for Experimental Computing, http://keeneland.gatech.edu/