An inquiry into the electro-physiology of bacteria

by

E. Prasanth Prahladan

M.Tech., Indian Institute of Technology Madras, 2013

B.Tech., Indian Institute of Technology Roorkee, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Electrical, Computer and Energy Engineering

2016

This thesis entitled:
An inquiry into the electro-physiology of bacteria
written by E. Prasanth Prahladan
has been approved for the Department of Electrical, Computer and Energy Engineering

Prof. Behrouz Touri

Prof. Joel M Kralj

Prof. Manuel Lladser

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Prasanth Prahladan, E. (M.S., Electrical Engineering)

An inquiry into the electro-physiology of bacteria

Thesis directed by Prof. Behrouz Touri

This thesis documents the exploration of tools and techniques which were used to determine the existence of patterns within and relationships between the voltage and calcium measurements from individual bacterial cells. The research lies in the broad domain of bacterial electro-physiology.

Acknowledgements

I would like to thank Prof. Joel Kralj, for giving me this oppurtunity to engage with a biological science problem, from the domain of a data-analysis. It was an exciting experience, to understand how fundamental biological questions are asked, and how measurements from newly engineered cell-structures are used to unravel truths about the biological world. I further extend my gratitude to my advisor, Prof. Behrouz Touri, for supporting my foray into an inter-disciplinary domain of research, which was not something he was familiar with.

Contents

Chapter

Figures

Figure

Chapter 1

Introduction

1.1 An analogy to the experimental process

Imagine the following scenario. There's a room in your house which is divided into two parts by a curtain. In one part of the room is an experimental apparatus with you instructor next to it. There is a curtain that separates that part of the room from the part in which you are standing.

The instructor sets up the experimental apparatus as follows. There's a flat roofed tent structure that has been set-up. We do not know the particulars of the material of the frame and the sheet that covers it. All we know is that it is strong enough to support the different structures that shall constitute the experiment. From different points on the ceiling of the tent, the instructor, hangs various simple-pendulums, with rods and balls made of similar materials. Further, upon each pendulum there are two measuring devices that are attached. One might be reading the angular displacement of the pendulum, the other might be measuring the velocity of the pendulum ball. The measuring devices are faulty and prone to very noisy measurements. Let's say that we obtain two measurements, 'signal-1' and 'signal-2' from each pendulum. Note that only the instructor has complete information about the experimental setup - the fact that each device is a pendulum, the fact that the measurement devices are noisy, the fact that the collection of pendulums are all hanging from the ceiling of a tent, etc.

So, what sort of information does the instructor give you across the curtain? Let's assume that there's an LCD screen set up on the curtain separating the two sections. The LCD screen displays the time-series measurements, 'signal-1' and 'signal-2' from each pendulum, and displays

it collectively on the screen. Further, the instructor also gives the coordinates of the locations of pendulum upon the ceiling-surface with a specific choice of coordinate system.

The instructor then asks you to determine if anything interesting can be said about the system from which those measurements are obtained, i.e. it would be fascinating if I could tell him that the signals are coming from a collection of pendulums hanging on a tent roof, as that would be information that he had not shared with me. . This thesis is thus a summary of my efforts at determining 'something interesting' about the experimental set-up based on the above information I was given by the instructor.

To understand the directions in which the research has proceeded, not particularly in a chronological order, it would be interesting to take a minute and consider how one would proceed with addressing this challenge.

Here are a set of questions that we can ask ourselves, while we consider the difficulty of the challenge.

(1) For every sub-unit of the system, there are two noisy-signals, 'signal-1' and 'signal-2'. The following questions arise about them:

   (a) Are there any patterns in the way either of the signals evolves over time? By 'patterns', we could mean either of the following

      - Can we describe it by a mathematical model that we are familiar with?
      - Can we detect motifs, subsequences within the measured signal-values, that seem to be repeated?
      - If the absolute values of the sub-sequences do not matter, but the shape of the curve traced by those values is important, we consider each of those traces to be a 'motif'. The question about repeating patterns, then becomes a question of identifying repeating shapes in the larger trace-structure of the entire time-series sequence that's recorded.

(b) Is there any relationship between 'signal-1' and 'signal-2'? Are they independent of each other?

(c) If we do assume the signals to be related to each other, such that one causes the other - What can you say about the rules that govern the behaviour of the sub-system from which the two signals are measured?

(d) Can we determine a mathematical model that relates the way the input-signal is converted into the output-signal?

(e) Is it possible that there is a feed-back loop in the sub-system that generates these signals i.e. can we reliably make the assumption that one signal actually causes the other signal? What if changes in one affects the other, and vice-versa?

(2) Let's just consider all 'signal-1' measurements from the different sub-units of the system. We could later repeat the analysis while considering all 'signal-2' measurements that we have. The following questions arise:

(a) We know that the different measurements are obtained from different points of a plane. Can we determine whether there is a spatial relationship between the signals? Do the sub-units that are closer to each other behave in a more-similar way than those that are far-away? Is it possible that each sub-unit disturbs the tent-material at which it is located, which then spreads across the tent surface and disturbs the behaviour of other sub-units?

(b) Do the different sub-units exhibit some form of synchronization behaviour?

(c) Are there groups or clusters of sub-units that behave in ways that are different from each other? Sub-units within a group behave similarly and display behaviours distinctly different from sub-units in other groups. If so, what can this be due to?

(3) Lastly, can we identify whether the instructor has followed a particular rule in spatially distributing the sub-units, based on the given information of the spatial locations of the

signal-sources? Or is it possible that the instructor has truly generated a random distribution of the sub-units over the plane.

Given the above set of questions as a guideline, we could proceed with an analysis of the observed signals and determine answers for the above questions. However, let's assume that the instructor tells you not to jump into your calculations directly. While you were pondering on those questions, he had been busy setting up another experimental set-up which is quite similar to the first set-up. It's possible that he changed the constituent materials of the tents, or he might have set up a group of spring-mass systems with dampers, instead of a simple-pendulum, etc. However, he makes sure that he records two signals from each sub-unit and presents the information to you on another LCD screen placed upon the curtain. Now, in addition to the questions that he had asked earlier, he further inquires,

(1) If he were to mix up the different pairs of signals from each experimental set-up would it be possible for you to determine which pair of signals came from which experimental set-up?

(2) Is there any parameter that can be measured from the signals in the two set-ups that helps differentiate between the two?

Compared to the scarcity of the information that we have, the list of questions that we have seem to be quite arduous. However, through the course of my thesis work, I have explored different fields of study that have experienced similar problems and have developed tools and techniques to address the above questions to a satisfactory level. We shall now proceed to understand the true scientific domain within which my research thesis aims to make a small contribution. Further, note that in the analogy above, the experimental set-up that was described is analogous to the biological system we are considering. The instructor represents the Biologist who sets up the experiments and obtains the recordings from the natural system and then provides it to me, the data analyst.

1.2 Introduction to the research problem

The domain of science within which this research is grounded, is known as Bacterial Electro-physiology, which is a subset of a parent field known as 'cellular electro-physiology'. This field of study concerns itself with understanding the structure and function of specific physiological structures and processes within living cells that are governed by the movement of electrically charged ions. Traditionally, the field of cellular electro-physiology has been predominated by the study of neurons, which are extremely differentiated cells found in higher-order multi-cellular organisms. When one talks about electrical behaviour in living organisms, one instantly thinks about the neuronal network in multi-cellular organisms. It would however be a very interesting endeavour to inquire into the evolutionary origins of neuronal behaviour.

Could the observed neuronal activity in higher-order organisms have evolved 'out of the blue'? Is it possible that the fundamental processes that governed their electrical behaviour had already developed and existed in much simpler life-forms like the uni-cellular organisms? Maybe those processes existed before the well-differentiated, super-specialized evolution of the neural network within the higher-order multi-cellular organisms? If so, what might be that link between the two worlds of different orders of biological complexity?

At a structural level, one may observe that the electrical behaviour of neurons is governed by the concerted work of specific cellular-structures called Ion-Channels and Ion-Pumps which govern the flow of ions across their cell-membranes. These structures dissipate energy while performing their function of moving ions either against their natural gradient or enabling their flow. They comprise of complex protein molecules that serve as gated-pores upon the cellular membrane, which might also have structures that filter and make them extremely selective of the type of ions they permit through them. These cellular structures interestingly are also present in unicellular organisms like Bacteria, Yeast, etc. However, their behaviour in these primitive organisms has been out of the scope of human scientific inquiry, due to the lack of specific technological advancements that help us probe into the functioning of these cells.

To get a more intuitive understanding of the above concepts, it would be good to have a comparison of the physical structures that are involved in the process:

- Size of biological structures:
  - Neuron Soma: $4 - 100 \times 10^{-6}$m; Nucleus diameter $3 - 18 \times 10^{-6}$m.
  - Ecoli: 2×10^{-6}m long, with diameter $0.25 - 1 \times 10^{-6}$m in diameter; Cell volume $0.6 - 0.7 \times 10^{-18} m^3$
- Size of physical structure/measuring devices:
  - Patch clamp pippette: $1 - 3 \times 10^{-6}$m diameter.
  - Visible light wavelength: $0.4 - 0.7 \times 10^{-6}$m.
  - Infrared light: $0.8 - 1000 \times 10^{-6}$m.

From the comparison of the above physical dimensions, we realize the importance of the probing technique that has been developed in measuring the intra-cellular ionic concentration and membrane potential. The technique by which the cells have been engineered to generate these signals is described below.

1.2.1 The biological system and the measurement process

The basic ideas behind cellular physiology can be explained as follows. The essential charge carrying particles in the biological realm are ions. The difference in the gradient of the ions across membrane structures leads to the formation of potential differences. The surface of a cell is made up of a permeable cell-membrane which comprises of various lipid and protein structures. Note that, given a permeable membrane, the existence of a difference in concentration of a particular ion, establishes a concentration gradient across each side of the membrane, which is measured as a potential difference across the membrane surfaces. By changing the conductivity of the membrane, the flows of the ions are regulated, which then leads a decrease in the membrane potential. Typically, a cell has low Sodium ion(Na^+) concentration inside it, in comparison to its exterior. Conversely,

we typically observe very high concentration of Potassium ion(K^+) inside the cell in comparison to its exterior. Both these ions carry unit positive charge ($\approx \pm 1.69 \times 10^{-19}$)C. Let's consider the situation, of a membrane with two ion channels, one each for Sodium and Potassium ions. The membrane potential is the difference in electric potential measured inside the cell to that outside it. We observe that when the conductivity of the Potassium channel is forced to zero, the membrane potential is positive. Similarly, when the conductivity of the Sodium channel is forced to zero, the membrane potential is negative. Thus, we observe that a cell need not directly cause a large flow of ions across its cell-membrane, to control its membrane-potential. All it needs to do, is to regulate the conductivity of the ion-channels that spread out across its surface.

Bio-chemical sensors Given the above description of the membrane potential, we now proceed to understand how the sensors work. The presence of a membrane potential, indicates the presence of an electric field across the membranes. Since, most life-forms have a good volume of water constituting the bulk of it body-mass, we observe hydrogen ions, H^+, obtained from the partial dissociation of the bipolar H_2O molecules, to move and accumulate near the surface of the positively charged membrane surface. The influence of the electric field is observed upto a few nano-meters in length from the surface, with the distance being called the 'bi-length'. The voltage established due to the concentration difference of a single ion, is called the "Nernstein Voltage" of a cell, given by

$$\delta V = (59mV)\log([H^+]) \tag{1.1}$$

Thus, a pH change of ± 1 (i.e a 10 fold change in H^+ concentration)leads to a change in voltage difference by $60mV$ across the cell-membrane.

1.2.1.1 PROPS: Voltage Indicators

The protein constituting the voltage-indicator ([16], [17]), can be imagined to be large lipid-protein structure embedded within the cell-membrane with a parts of it exposed to both the interior and exterior of the cell. Inside the protein structure,also contains a fluorescent component inside

it, which absorbs light within a particular range of frequencies and emits light in another range of frequencies. However, what's interesting about this chemical is that, the frequencies of absorbed and emitted radiation is regulated or controlled by its 'state of protonation' i.e the presence or absence of a H^+ ion inside its core.

1.2.1.2 CaGS: Calcium Indicators

The Calcium ion(Ca^{2+}) indicator is another bio-chemical sensor ([28]), that can be understood to consist of two functional segments. One part is the fluorescent part and the other is the part that binds to Calcium ions. The sensor molecule's behaviour is regulated by its structural conformation which changes in the presence or absence of Calcium ions. Note that, in a typical cell, the Calcium ion(carrying two unit positive charges each) concentration is higher outside the cell as compared to its interior.

In eukaryotic cells, like mammalian neurons, it is understood that the Calcium ions from the exterior of the cell, floods into its interior when the cell depolarizes. And hence, there is a relationship between the signals measured from the two types of sensors. It is also, understood that the direction of causality is from the voltage signals to the calcium signals. However, in the prokaryotic cells, like the bacterial cells, we do not yet know whether a similar relationship exists between the signals. Further, we are also unaware of the physical bio-chemical structures that are involved in the processes that lead to variations in the voltage differences and the calcium ion concentrations inside the cells.

1.2.2 On the exploration of what the signals mean

The dataset that served as the primary test template was a collection of voltage('signal-1') and calcium('signal-2') time-series signals for populations of cells present upon a petridish. The cells were nurtured in the same batch process and could be assumed to be sharing the same external environment. Henceforth, we shall illuminate the sequence of biological questions that were asked of the data-set and the techniques of inquiry adopted to determine an answer to the question.

(1) Assuming that the time-series are recordings of a zero-resource language, can we identify constructs or components of the language?

   (a) Are there any obvious visual patterns that can be identified in each class of the measured signals?

   (b) Are there any features that are repeatedly observed in all sets of measurements from either class of signals?

   (c) Given a mixed bag of signals can we differentiate between the signals of each type-class?

(2) Can we detect important events in the time-traces that indicate instances of important biological events?

   (a) Is it possible to determine the time-instances when biological change might have occurred?

   (b) If so, can we prove or verify that some significant biological event truly occurred at that time-instant, maybe through some other measurement process?

(3) Can we infer that the cells are communicating with each other?

   (a) Do we have measurements made from the medium in which the cells are lying, to determine chemical changes? If the cells do communicate with each other, we assume that it has to be through the changing the chemistry of the medium upon which they are placed.

   (b) Given only measurements from inside individual cells, can we determine relationship between these signals that indicate some sort of spatial/temporal influence of one upon the other? This influence can occur only via communication.

   (c) If the cells are communicating, do they exhibit any form of synchronization in their behaviours?

(4) Given any set of signal-measurements from cells of a given species, can we determine if there are intra-species differentiation, which might be attributable to genetic variations in the population that make the cells behave slightly differently?

(5) Can we extract features from the measured signals that may serve as unique signatures to differentiate between different species of cells?

(6) Do we observe similar types of intra-species differentiation in different species?

With the above questions of biological relevance highlighted, we shall now proceed to explain the tools and techniques that were adopted to determine answers to the above questions. In retrospection, the researchers have felt that it's more important to have the right questions. The search for the answers depends on the question being asked. It might seem quite intuitive, but the progress of the thesis depended on the questions that were asked and the answers that were obtained for each question. The results of the techniques used to answer the questions being probed led to an exploration of other questions and methods. The research methodology for the project was based on a broad search of techniques that could be applied, hoping to find some interesting patterns in the datasets. The presence or absence of obvious patterns further led to a reformulation of the problem statement. While we started with an exploration of unsupervised machine learning techniques to extract patterns in the dataset, we finally realized that it would be beneficial to adopt a more structured approach towards data-analysis, where classical techniques that are available for the study of time-series could be used to infer information from the dataset that might be more relevant than a blind trial-and-error approach towards the analysis of signals about which very little is known. This thesis shall thus comprise of the following main components:

(1) Spatio-temporal analysis of measured signals

(2) Unsupervised machine learning

(3) Development of data-visualization tool

An architectural view of the ways in which the different tools and techniques were used and put together, to obtain a pipeline that processes raw data to extract visualizations is provided below(Fig. 1.1). Every data-processing pipeline consists of the first-initial stage called 'Feature Extraction'. By a feature, we imply any attribute that can be computed from the data, to explain a certain characteristic about the signal - its time or frequency domain characteristics, its wavelet coefficients, the extent of noise in the signal, a measure of total-variation of the signal, simple statistical metrics from the time-trace, etc. The essential features can be extracted from the data, by using the tools and techniques available in different domains of Signal Processing, Statistics, and Time-Series Analysis.

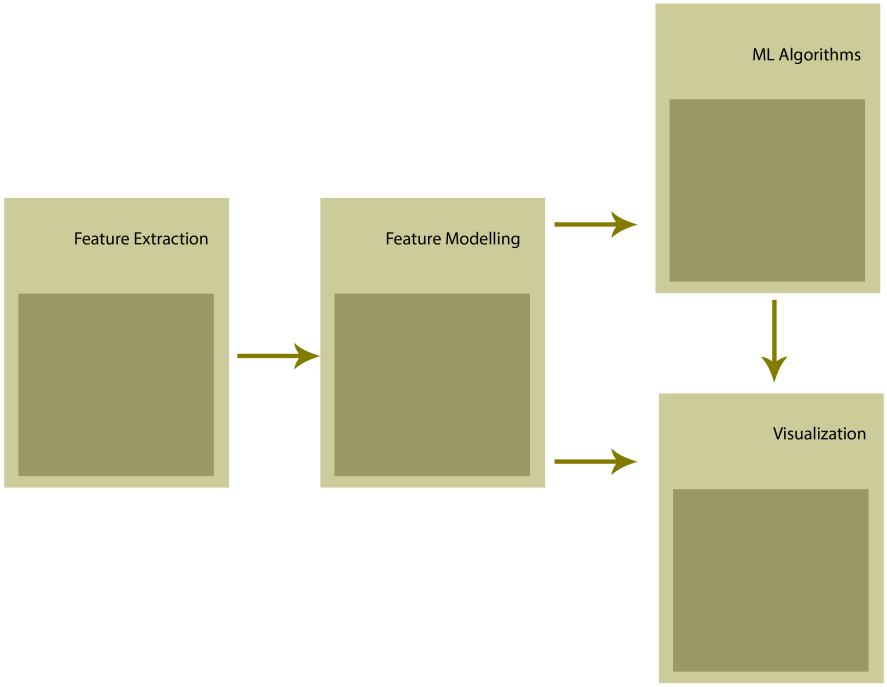


Figure 1.1: Architectural view of the different stages in the development of the processing pipeline that generates visualizations from the raw recorded data.

After the set of 'Features' have been determined, comes the stage of 'Feature modelling'. In this stage, different ways of modelling or aggregating the feature-information corresponding to each cell is considered. Different techniques like code-word histograms, probabilistic models, etc

can be used to approximate the mathematical object, from which the measured feature-vectors are assumed to be samples. After the information has been aggregated, if the data is low-dimensional, we can directly visualize the data; if the data is high-dimensional we may use linear or non-linear methods of dimension reduction to obtain the minimal dimensions into which these data points can be projected, to obtain a visualization. This basic pipeline is adopted to describe the methodology of using the different tools described in this thesis report.

One important question that arises, is the on the question of what exactly are we trying to model or describe using the mathematical formulation. Our dataset, consists of measurements from a dynamical system consisting of multiple-agents (bacterial cells) which may be interacting with each other. There are two particular modes in which we can pursue to address the problem, as illustrated in Fig.(1.2)

One mode tries to consider the observations as absolute, and thus use techniques that help us describe the observations - either in its entirety by some global statistical measure, or of the way in which the observations depends on its past values.

The other mode consists of trying to understand the dynamical system, that generates the observations. There is no absolute mapping of observations to a particular system model, since we can fit different mathematical models of the dynamical systems, to the same set of recorded observations. The value of this approach however, is that we have a body of knowledge regarding different mathematical models of dynamical systems, and hence once a dynamical model has been fit to the dataset, we can infer some properties of the model, and then formulate biological experiments to test for the conclusions that each model hypothesizes. The class of linear models(state-space models) and probabilistic models (hidden-markov models), etc fall into this category of mathematical models, that may be pursued to gain insight into the properties of the system.

In the present thesis, we have only explored the former mode of building mathematical models to describe the observations recorded from the system, based on itself. We do not try to infer any mathematical model to describe the dynamical system which generates these measurements. The following chapters shall describe different mathematical tools and techniques that were used during

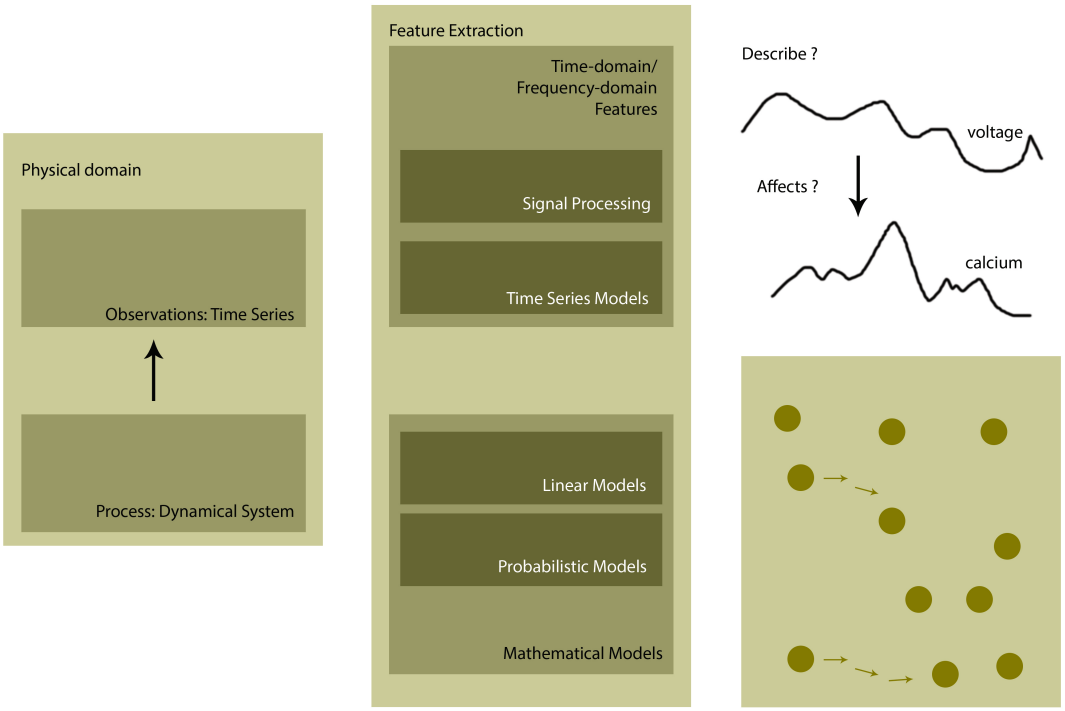


Figure 1.2: Two approaches to building mathematical models to describe the measured data.

the course of the research, to analyse the data and to gain insight into our biological system. To help summarize the inter-connections between the different tools and domains, kindly use the following info-graphic, which highlights an architectural view to the development of a data-processing pipeline for scientific inquiry.

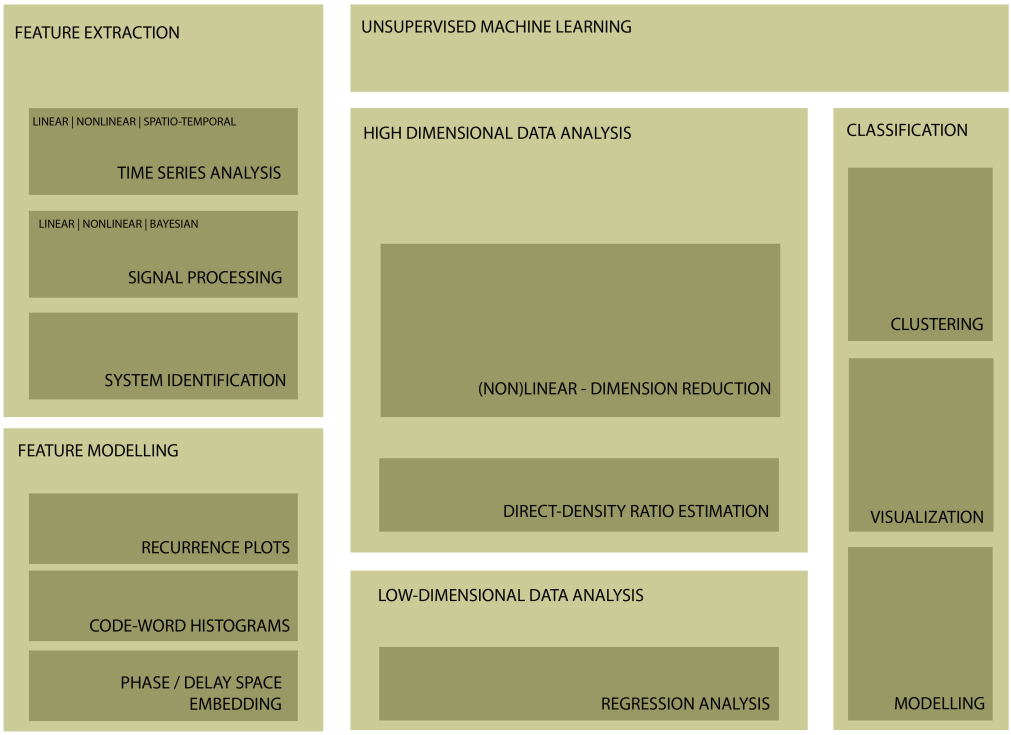


Figure 1.3: Architectural view of the process of developing data-analysis pipelines for gaining insight into the biological dataset.

Chapter 2

Spatio-Temporal Analysis

The dataset that we have is a collection of finite-time samples of two variables - the membrane potential, and the Ca^{2+} ion concentration. However, the two sets of data are not available in all experiments, as it depends on whether the batch of cells were prepared to express particular genes that generate the protein sensors in the cell-body. We are also provided with the location of the centroid of the bacterial cell-body with respect to the frame of view of the microscope. Thus, we are provided with two different types of information, the temporal signals, where the order of the signal sequence is important, and the spatial measurements, where the order is irrelevant. This dataset can thus be analysed individually in the spatial domain, to determine the distribution of the cells, and thus make inferences of the same. We may also, analyse the signals in the time-domain alone, thus inferring patterns in the time-sequences. The spatio-temporal analysis of the dataset would help us understand the contribution of the spatial distribution on the temporal progression of the measured time-sequences.

2.1 Temporal analysis of measured signals

Depending on the availability of only one or two sets of time-sequences corresponding to the two physical parameters - voltage and calcium ion concentration, we may have to consider the following two sub-domains of temporal analysis:

(1) Univariate Signal Processing

The set of tools in this domain are concerned with studying characteristics of signals of a

single type. The methods involve summarizing information regarding the signal sequences in the time-domain and the frequency domain, in addition to fitting models to describe the present behaviours based on past observations.

(2) Bi-variate Signal Processing

The set of tools in this domain are concerned with studying characteristics of signals of a two types. The methods involve the identification of input-output relationships between the two types of signals.

2.1.1 Univariate signal analysis

The first step in the analysis of univariate analysis, corresponds to removing the bias in the signal. This involves extracting the mean of a given signal-sequence, and subtracting it from the time-series. We undertake this step, as we do not expect to assign any real meaning to the magnitudes of the signals.

The time-series representation of the bacterial electrical signals, can be studied from the perspective of building linear or non-linear mathematical models of dynamical systems in high-dimensional spaces, who's trajectories in phase-space when projected onto a 1-dimensional manifold, provides us the traces measures. Else, we might pursue the analysis from a data-mining perspective, where we implement data-mining algorithms to detect motifs and patterns in the dataset.

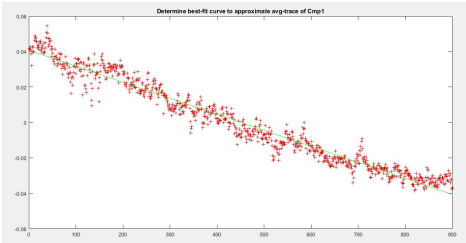
2.1.1.1 Analysis of synchronization between cells

To analyse the synchronization between the cells, we first determine the average-trace representative of the behaviour of all the cells constituting a particular component. Further, we attempt to obtain a best-fit line that approximates the average-trace of each component. By computing the difference of each original trace from the average-signal and the best-fit curve, we can obtain the distribution of the noise in the recorded signals. It is interesting to observe that the noise-distribution has a uni-modal property 2.2.

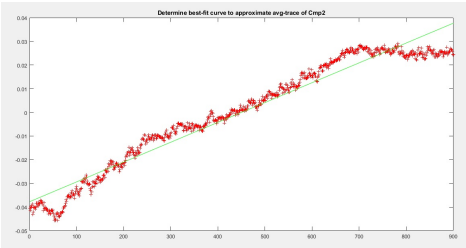
These observations raise the following directions of research:

(1) Is it possible to model the behaviour of the cells, as comprising of ramps and step-signals, with added noise.

(2) Does the noise-statistics correspond to a Gaussian Distribution?

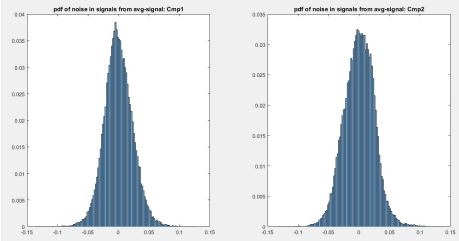


(a) avg-trace-1

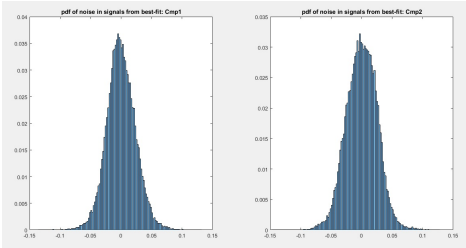


(b) avg-trace-2

Figure 2.1: (a)-(b) Average-Trace and Best-Fit trace, of cells in different components



(a) noise-from-avg



(b) noise-from-bestFit

Figure 2.2: Noise distribution in traces when compared to (a) Average-trace (b) Best-Fit-trace

The observations/measurements of a variable of interest that is obtained over a sequence of time-steps, is said to constitute a time-series. The time-series obtained, can be analyzed from two different approaches:

(1) Determine a dynamical model based on observations(measurements) themselves.

(2) Determine a dynamical model based on internal-states of a dynamical system, following specific deterministic rules of evolution. The measurements are certain reduced dimensional projections of the internal-state of the system.

The two approaches that we have, help to abstract away the idea of the physical world, by connecting the observed behaviour of a physical system, to the dynamical behaviour of an analytical mathematical model in the abstract mathematical universe. Each approach or assumption made about the recorded measurements or the process that generates these measurements, give us access to a set of tools and techniques compatible to that domain of study, which shall help us gather some insight into the real physical process. The similarities we observe, or the information we gather by the application of these tools to the dataset, can only help us identify patterns in the behaviour of the system. It cannot tell us why the particular system behaves in a particular way. The set of tools we use, shall help us in building up hypothesis about the nature of the biological system, the processes that generate the signals, and the possible dynamical relationships between the variables being measured. The hypothesis can be proved/rejected only by the design of suitable biological experiments.

2.1.1.2 Deterministic Processes

Time-series signals can be expected to be generated by deterministic processes, in which given the initial state of the system and a description of the rules of state-update, the future state at any given time can be exactly determined. It is true that the measurements we make of the biological system, seem to be corrupted with noise from the inadequacy of the measurement techniques adopted. However, one of the approximations we can make in studying the signals

obtained, could be to assume that the observations were made from a deterministic system that was evolving over time. Further, even though the signals contain noise and sometimes seem to be almost random, it's possible that the signals were generated by a chaotic deterministic process rather than a truly stochastic process.

Given this assumption, with regards to the generative process, from which these measurements are obtained, there are different questions we can ask about the system. One such question concerns the presence or absence of synchronization between the signals generated by the different cells.

Phase Synchronization The term synchronization refers to a process of mutual adjustment of dynamical evolution of two or more distinct but coupled units, in general assumed to be oscillators, which leads to a collective behaviour that is coherent - which might be periodic, quasi-periodic or chaotic. The phenomena of synchronization, can be studied under different physical attributes of the system. One such physical phenomena is called phase synchronization, i.e. a coherent motion with a fixed ratio of average frequencies.

At a very intuitive level, we would assume two coupled systems to be synchronized, if they were displayed spectral coherence, i.e. existence of oscillations with the same frequencies. However, this approach requires stationary conditions, with the presumed frequency contributing at all times with equal strength. In real world, however since the frequencies in signals are generally time-dependent, the assumption of stationary conditions is violated. This leads us to consider notions of frequency-time localized coherence, like 'wavelet-coherence'. However, phase-coherence analysis, deals with a notion of a time-varying 'instantaneous frequency of oscillations', the time-derivative of a suitably defined phase variable.

The question then arises, what do we understand by the 'phase' of a time-series? Well it depends on what we would like to consider as the 'phase' of the system. First, are there any particular properties that the a particular derived variable must follow, for it to be considered the phase of the system? Some of the methods adopted to define the 'phase' of the system are

(1) Classically, for a two dimensional dynamical system, we consider the plot of the velocity-vs-displacement of the state-variable, as phase-plot of the system. By extending this notion, we can consider embedding the obtained time-series in a finite-dimensional state-space. The Poincar'e sections of the time-series may be used to define points in time that correspond to fixed phase values of $2k\pi, k \in \mathbb{N}$.

(2) It is possible to extend an observed real-valued time-series into the complex plane, to obtain the complex-valued time-series $Z(t) = X(t) + iY(t)$, where $Z(t) \in \mathbb{C}, \{X(t), Y(t)\} \in \mathbb{R}$ and $Y(t)$ is the Hilbert-transform of $X(t)$. The phase-variable is then defined to be, the phase-angle at any given time, in the complex plane, ie.

$$\phi(t) = \tan^{-1}\left(\frac{Y(t)}{X(t)}\right)$$

(3) When $X(t)$ is noisy, its possible to define the phase of the signal as follows,

$$\phi(t) = \tan^{-1}\left(\frac{\dot{Y(t)}}{\dot{X(t)}}\right)$$

where $\dot{X(t)}$ is the local derivative of $X(t)$, and $Y(t)$ is the Hilbert-Transform of $X(t)$.

Once the phase-trajectories of the time-series has been determine, the relationship between them might be determined by either analysing the trajectory of their phase-difference, or even considering the scatter-plots of one with respect to the other.

Measures of phase coherence may be defined based on either the statistical properties of the phase differences, or the joint evolution of the phases.

Topological Phase Coherence One another approach to studying the phase relationships between signals, is interestingly, by not focussing on the 'definition' of phase in these signals. One can adopt a topological approach to determining phase coherence, based on the concept of 'recurrence plots'(RP)([20], [21], [26]). RPs were originally designed for visualizing the correlation pattern within a single time-series comparing the values at all times t_i with all observations at all other times t_j, as illustrated by the Recurrence Matrix,

$$R_{ij} = \Theta(\epsilon - ||X_i - X_j||)$$

where, $\Theta(\cdot)$ is the Heaviside function, $X_i = X(t_i)$, $||\cdot||$ is a suitable norm, and ϵ is a parameter defining the resolution of observation. The graphical visualization of R in terms of a monochrome structure is called a recurrence plot. Statistics on the distributions of continuous vertical and diagonal structures allow to estimate a variety of dynamic invariants, of the system generating the observations.

Interestingly, this methodology can be extended to measurements of dynamics in higher dimensional spaces. If the observed time-series, can be embedded in an N-dimensional space, then the state-vectors $X(t)$ for generating the recurrence plots correspond to the N-dimensional state-vector at time-t, obtained from the 1-dimensional time-series observation z_t.

The concept of recurrence plots, may be extended to study the joint evolution of two coupled variables $X(t)$ and $Y(t)$, the cross-recurrence plots are defined as

$$CR_{ij} = \Theta(\epsilon - ||X_i - Y_i||)$$

From a study of the signals generated by various dynamical systems - linear and non-linear systems, it is possible to characterize the distributions of the various diagonal and straight lines observed in the recurrence plots at a particular resolution. Further, the recurrence plots for the uni/bi-variate signals observed in our experiments can be generated, the statistics regarding the straight and diagonal lines can be computed, and thus compared with the information gathered regarding different mathematical models. A close fit to deterministic/stochastic systems exhibiting particular properties, might suggest that the current experimental observations are generated by a similar process.

2.1.1.3 Stochastic Processes

A stochastic process may be defined as a collection of random variables that are ordered in time and defined at a set of time-points, which may be continuous or discrete ([11]). Further, the random variables may take discrete or continuous values. We denote a random variable at time t, by $X(t)$ for continuous-time processes, and X_t for discrete-time processes. Every observed sequence

$z_t, t < \infty$ can be taken to be one-possible realization of a stochastic-process, which could have generated an infinite set of time-series given the same starting conditions but fixed update rules. Many models for stochastic processes are expressed by means of an algebraic formula relating the random variable at time t to past values of the process, together with an unobservable 'error' process. From this model, it may be possible to specify the joint-probability distribution of $X(t_1), \cdots X(t_k)$ for any set of times $\{t_1, \cdots, t_k\}$. A simpler more useful way of describing the process is to give moments of the process(obtained from the process ensemble(multiple realizations of the process)), in particular the mean and the auto-covariance function(acv.f) respectively.

$$\text{Mean Function, } \mu(t) = \mathbb{E}[X(t)] \tag{2.1}$$

$$\text{Variance Function, } \sigma^2(t) = \mathbb{E}[(X(t) - \mu(t))^2] \tag{2.2}$$

$$\text{Autocovariance Function, } \gamma(t_1, t_2) = \mathbb{E}[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))] \tag{2.3}$$

2.1.1.4 Models of stochastic processes

Moving Average Proceses Suppose that Z_t is a purely random process with zero mean and variance σ_z^2. Then X_t is said to be a moving average process of order q, MA(q) if

$$X_t = Z_t + \sum_{k=1}^{q} \beta_k Z_{t-k} \tag{2.4}$$

where $\{\beta_i\}$ are constants.

Auto-regressive Process Suppose that $\{Z_t\}$ is a purely random process with mean zero and variance σ_z^2. Then a $\{X_t\}$ is said to be an autoregressive process of order (p,q) if

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + Z_t \tag{2.5}$$

Mixed ARMA Process Suppose that $\{Z_t\}$ is a purely random process with mean zero and variance σ_z^2. Then a $\{X_t\}$ is said to be a mixed auto-regressive moving-average process of order (p,q) if

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + Z_t + \sum_{j=1}^{q} \beta_j Z_{t-j} \tag{2.6}$$

The importance of ARMA processes lies in the fact that a stationary time-series may often be adequately modelled by an ARMA model involving fewer parameters than a pure MA or pure AR model.

2.1.1.5 Stationary Signal Processing

An important class of stochastic processes are those that are stationary([5],). A time-series is said to be strictly stationary if the joint probability distribution of $X(t_1), \cdots, X(t_k)$ is the same as $X(t_1+\tau), \cdots, X(t_k+\tau)$ for all $t_1, \cdots, t_k, \tau$. In particular, for $k=1$, it implies the distribution of $X(t)$ is the same for all t, provided the first two moments are finite, ie. $\mu(t)=\mu, \sigma^2(t)=\sigma^2$. For $k=2$, we have the joint-distribution between any two time-instants, t_1 and $t_1+\tau$, depends only on the lag(τ). Thus, the auto-covariance function $\gamma(t_1,t_2)=\gamma(t_2-t_1)=\gamma(\tau)=cov(X(t),X(t+\tau))$. Similar constraints can be determined for larger set of random-variables.

In practice, a less restrictive definition of 'stationary property' is defined. A process is called second-order stationary(or weakly stationary) if its mean is constant and its acv.f depends only on the lag.

$$\mathbb{E}[X(t)] = \mu \tag{2.7}$$

$$\sigma^2(t) = \sigma^2 \tag{2.8}$$

$$cov[X(t), X(t+\tau)] = \gamma(\tau) \tag{2.9}$$

In other words, we deem a time series z_t stationary, if $\mathbb{E}[z_t]=\mu$ is independent of t and the auto-covariances $Cov(z_t, z_{t+k})$ depends only on k for all t. Let $\mathbb{E}[z_t]=\mu_z$. However, since we have only a finite number of time-samples, the sample estimate of the descriptive statistics are:

$$\bar{z_t} = \frac{1}{T-1}\sum_{t=1}^{T} z_t$$

$$\bar{\sigma_z}^2 = \frac{1}{T}\sum_{t=1}^{T}(z_t-\bar{z})^2$$

Assuming the time-series were realizations of random-variables, we use the following concepts from probability theory. For any two random variables, X and Y, we have

$$\text{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$
$$\text{Corr}(X,Y) = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}}$$

For a given time-series z_t, we can assume $X = z_t | 1 \leq t \leq T-k$ and $Y = z_t | k \leq t \leq T$. Assuming stationarity, i.e. $\mu_X = \mu_Y$, we define the auto-covariance and auto-correlation functions which are a function of the lag, k as

$$\text{auto-covariance}, \gamma(k) = \mathbb{E}[(z_{t+k}-\bar{z})(z_t-\bar{z})]$$
$$\text{auto-correlation}, \rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

However, note that the above definitions are valid only if the process is stationary.

Test for Independence between samples

The Auto-correlation Function,$\rho(k)$ plays an extremely crucial role in the identification of time-series models as it summarizes as a function of k, how correlated the observations that are k lags apart are. As a rule of thumb, estimating the above parameters from the sample-time series requires for a good approximation that the total number of observations $T > 50$ and choose $k < T/4$ in estimation of $\gamma(k)$ and $\rho(k)$.

It is known that if we have a time series, z_t $N(0,\sigma)$, then $\rho(k) = 0 \forall k > 0$. However, the sample estimates of the same will not be exactly zero. For large T, it is known that $\rho(k)$ $N(0, \frac{1}{T})$. Therefore, $\pm\frac{2}{T}$ is often used as a 95%-confidence interval. If $\hat{\rho}(k)$ plots beyond these confidence limits, we suspect that the assumption of independent time-series z_t is violated.

Identification of stationary signals

A useful tool to determine whether a given time-series is stationary or not, is the Variogram. This tool is well defined and applicable to stationary and non-stationary processes. The variogram G_k measures the variance of differences k time units apart relative to the variance of the differences

one unit apart. In particular, we have the variogram defined as

$$G_k = \frac{Var[z_{t+k} - z_t]}{Var[z_{t+1} - z_t]} \tag{2.10}$$

where G_k is plotted as a function of lags k. For a stationary process, we have

$$G_k = \frac{1 - \rho(k)}{1 - \rho(1)} \tag{2.11}$$

$$\lim_{k\to\infty} G_k = \frac{1}{1 - \rho(1)}, \text{ since } \lim_{k\to\infty} \rho(k) \to 0 \tag{2.12}$$

For a stationary process, with positive autocorrelation, the variogram could be interpreted as follows. Positive autocorrelation implies that consecutive observations are similar to each other. The variance of the first differences will therefore be less than the variance of the difference two lags apart, three lags apart, etc. However, as k gets larger the variance of the differences k-lags apart will be similar to that (k+1)-lags apart. Therefore, the variogram for small k gets larger as k increases, eventually it reaches an asymptotic value. The sample variogram can be estimated as follows:

$$\hat{Var}[z_{t+k} - z_t] = s_k^2 = \frac{1}{n-k-1} \sum_{t=1}^{n-k} (d_t|_k - \bar{d}|_k) \tag{2.13}$$

$$\hat{G}_k = \frac{s_k^2}{s_1^2} \tag{2.14}$$

where $d_t|_k = (z_{t+k} - z_t)$ and $\bar{d}|_k = \frac{\sum_t d_t|_k}{n-k}$. Once the sample-variogram has been plotted for the signals, we can determine whether the signal is stationary or not-stationary.

There are many ways in which a time series can be non-stationary. One type of non-stationarity is where the level changes, but the process exhibits homogeneity in the variability, i.e. the first-difference of the time-series, $\nabla z_t = z_{t+1} - z_t$ is stationary. This is referred to as a (first order) homogenous non-stationary process. It might be possible that both the level and the slope of a time-series are non-stationary, but the variability otherwise exhibits homogeneity i.e the second difference $\nabla^2 z_t = \nabla(z_{t+1} - z_t) = (z_{t+1} - z_t) - (z_t - z_{t-1})$. Since the amplitudes of the time-signals measured are assumed to have no significant consequence, we shall continue to work with the first-difference or the second-difference of the signals.

Once the time-series has been determined to be stationary, we shall plot the auto-covaraince function(acv.f), the auto-correlation function(ac.f) and the partial auto-correlation function (pac.f) to identify the model family that might be fit to the data.

Once we determine a good model-fit, we expect the residuals to behave like uncorrelated white-noise. One way to determine the validity of the model, is to compute the ac.f and the pac.f of the residual time-series, and determine the presence of any significant coefficients. Further, since we assume that the errors are normally distributed during the modelling stage, we expect the residuals to be more or less normally distributed too. This can be determined by obtaining a normal plot of the residuals, which are then expected to lie along a straight upward sloping line.

Autocorrelation and the Correlogram (Auto-correlation function)

Another important tool to study the properties of the time-series is provided by a series of quantities called the sample auto-correlation coefficients. They measure the correlation, if any, between observations at different distances apart and provide useful descriptive information.

The ordinary correlation coefficient defined for N realizations of two random variables, X and Y, $\{(x_1, y_1), \cdots, (x_N, y_N)\}$, is given by

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2.15}$$

This quantity lies in the range [-1,1] and measures the strength of the linear assosciation between the two variables. The correlation is negative if 'high' values of one variable correspond with 'low' values of the other. If the two variables are uncorrelated or independent, then the true correlation is zero.

To determine the correlation of a signal, with a sub-sequence of itself obtained after a lag k units, we can use a modification of the definition of the random variables X and Y. Consider N observations of a time-series $\{x_t\}$. Then given a delay k units, we can obtain N-k pairs of observations, $\{(x_1, x_{1+k}), \cdots, (x_{N-k}, x_N)\}$, where each pair of observations is separated by k time-

units. Then we can compute the correlation of the observations as,

$$r_k = \frac{\sum(x_t - \bar{x}|_{k(1)})(x_{t+k} - \bar{x}|_{k(2)})}{\sum(x_t - \bar{x})^2} \tag{2.16}$$

2.1.2 Bivariate signal analysis

When considering bivariate signals, we can assume a causal relationship between the signals, in the form of an open-loop system where changes in input-signal X, causes changes in output signal Y, which is expressed as $Y = f(X)$. The relationship between the signals can be determined by regression. However, the relationship becomes complicated when the output response may be delayed and distributed over a period of time. In proceeding with this approach, however, one needs to be careful about the auto-correlation in the input-signal. The dynamic relationship between the input and output can be approximated by a linear transfer function

$$y_t = \eta_t + \sum_{k=0} \xi_k x_{t-k} \tag{2.17}$$

where the ξ_k are transfer function weights and η_t is generated by a white-noise process.

Some aspects of the observed cross-correlation between the signals, is an artifact of the auto-correlation of the input-signal with itself. One strategy that is adopted to filter out these artifacts, is called 'pre-whitening', which involves the following steps

(1) Identify and fit a time-series model to the input data, x_t.

(2) Extract the residuals for the time-series, which we expect to be generated by a white noise process.

(3) Pre-whiten the output series using the same model fitter to the input y_t. However, note that the residuals for the output-signal need not be uncorrelated.

(4) Compute the cross-correlation between the pre-whitened x_t and y_t.

(5) Compute the transfer function weights η_k

A rough estimate of the standard error of the cross correlations is given by $\frac{1}{\sqrt{n}}$, where n is the number of observations. For the above process, we consider the demeaned signals, $\tilde{x}_t = x_t - \mu_x$ and $\tilde{y}_t = y_t - \mu_y$. Let the ARMA model fit to the signals, generate the residuals α_t from $\tilde{x}_t$ and β_t from $\tilde{y}_t$. Thus, the transfer function model for the residual pre-whitened signals is

$$\beta_t = \eta_t + \sum_{i=1}^{k} \xi_i \alpha_{t-i} \tag{2.18}$$

where the rough estimates of the transfer function weights can be computed by the relationship

$$\hat{\xi}_j = r_{\alpha\beta}(j) \frac{s_\beta}{s_\alpha} \tag{2.19}$$

Once we determine the transfer function coefficients for the system, the coefficients can serve as feature vectors representative of the individual cell-system. Algorithms in machine learning and high-dimensional data analysis can then be used to identify patterns in these data-sets.

2.1.3 Change-Point Detection

Prior to the adoption of any methods of modelling the time-series data, there were a set of questions we asked. The methods we adopted to seek answers to those questions helped further determine the direction of the data-analysis we adopted. Given a set of time-series information, we hypothesized that, it would be interesting to determine any level of communication or coherence between the cells, based on the simultaneous occurrence of events in the cell-signals, or in comparing the time-intervals between the observed events in the cells.

The objective of change-point detection is to discover abrupt property changes lying behind time-series data. One of the question we were curious to explore, was to determine whether there were any observed patterns in the nature of events that occurred in the signals from different cells. We chose to follow the technique of change point detection in time-series data by relative density ratio estimation ([13]).

The method depends on the idea that one need not individually compute two probability densities, to compute their ratios. The direct-density ratio estimation methodology [24], is based

on the idea that knowledge of two densities implies knowledge of its ratio, but not vice-versa, since the decomposition of the density ratio into the two component densities is not unique. Therefore, the computation of the densities of two probability densities is substantially easier than density estimation. The KLIEP algorithm [15] was an algorithm that implemented the direct-density ratio methodology for detecting change-points in time series. It was reported to outperform other standard algorithms.

Detection of change-points in time-series depends on two specific steps:

(1) Determine a way to segment time-series to model it

(2) Determine the choice of algorithm for direct density ratio estimation

2.1.3.1 Change-point detection Problem Formulation

Let $y(t) \in \mathbb{R}^d$ be a d-dimensional time-series sample at time-t. Let,

$$\mathbf{Y}(t) := [y(t)^T, y(t+1)^T, \cdots, y(t+k-1)^T]^T \in \mathbb{R}^{dk}$$

be a subsequence of time-series t with length k. We thus consider, $\mathbf{Y}(t)$ to be the time-sample under study, instead of the observation $y(t)$. Define the set of samples $\mathcal{Y}(t)$ to contain n-representative subsequence samples starting at time-t:

$$\mathcal{Y}(t) := \{\mathbf{Y}(t), \mathbf{Y}(t+1), \cdots, \mathbf{Y}(t+n-1)\} \tag{2.20}$$

where $[\mathbf{Y}(t), \mathbf{Y}(t+1), \cdots, \mathbf{Y}(t+n-1)] \in \mathbb{R}^{dkxn}$ forms a Hankel matrix that is significant in change-point detection algorithms based on subspace learning methods. Note that, the sample of study, $\mathbf{Y}(t)$ at instant t, contains information observed from $\{y(t), \cdots, y(t+n+k-2)\}$.

For change-point detection, let's consider two consecutive segments, $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$. The algorithm tries to determine a dissimilarity measure between the two segments, and use it as a plausibility of change-points. The higher the dissimilarity measure, the more likely the point is a change-point.

2.1.3.2 Divergence based dissimilarity measure

For the two sets of time-samples, $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$, let P_t and P_{t+n} denote the probability distributions over the sample space from which the corresponding time-samples are realizations. Then the dissimilarity measure between the two sets can be defined as

$$D(P_t||P_{t+n}) + D(P_{t+n}||P_t) \tag{2.21}$$

where, $D(P||P')$ denotes the $f - divergence$ between the two distributions $\{P, P'\}$, given by

$$D(P||P') := \int p'(\mathbf{Y}) f\left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})}\right) d\mathbf{Y} \tag{2.22}$$

where $f(\cdot)$ is a convex function such that $f(1) = 0$ and $\{p(\mathbf{Y}), p'(\mathbf{Y})\}$ are probability density functions corresponding to $\{P, P'\}$ respectively.

The f-divergence includes the following:

(1) Kullback-Liebler Divergence, $f(t) = t * log(t)$

(2) Pearson Divergence, $f(t) = \frac{1}{2}(t-1)^2$

However, we note that in the description of the problem, P, P' are unknown and hence, the f-divergences cannot be computed directly.

A naive approach, that was traditionally adopted, was to compute estimates $\{\hat{P}, \hat{P}'\}$ of $\{P, P'\}$. However, density estimation is a hard problem. An alternative approach, based on direct-density ratio estimation can be used to compute the divergences between the two probability distributions. Thus, we can compute a measure of dissimilarity between two distributions, by having a direct method to compute the ratios between the two distributions.

A review of two methods for directly estimating the density ratio from samples contained in the sets $\mathcal{Y}(t)$ and $\mathcal{Y})(t+n)$ shall follow in the next subsections.

2.1.3.3 KLIEP

An application of direct-density ratio estimation, to the computation of KL-divergence between two probability distributions was reported in [24] and is abbreviated as KLIEP.

Density Ratio Model Let's model the density ratio $g(\mathbf{Y}) = \frac{P(\mathbf{Y})}{P'(\mathbf{Y})}$ by the following kernel model:

$$g(\mathbf{Y}|\theta) := \sum_{i=1}^{q} \theta_i K(\mathbf{Y}, \mathbf{Y}_i) \tag{2.23}$$

where $\theta := (\theta_1, \cdots, \theta_q)^T$ are parameters to be learned from the data samples, $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_q\}$ are p different sample points within a set and $K(\mathbf{Y}, \mathbf{Y}_i)$ is a kernel-basis function, like the Gaussian Kernel

$$K(X, X') = \exp\left(-\frac{||X - X'||^2}{2\sigma^2}\right). \tag{2.24}$$

where $\{X, X'\}$ are vectors of same dimensions, and $\sigma > 0$ is the kernel width. The kernel-width can be assumed, or determined by cross-validation. Note that the total number of parameters q can be chosen to equal to an arbitrarily chosen number, or equal to the total number of time-samples available.

Learning Algorithm From the above definition of the density ratio, we have a good estimation of the distribution $P(\mathbf{Y})$, $\hat{P}(\mathbf{Y}) = g(\mathbf{Y}|\theta)P'(\mathbf{Y})$. The learning algorithm thus seeks to determine the parameter vector θ that reduces the KL-divergence between $P(\cdot)$ and $\hat{P}(\cdot)$.

$$KL(P||\hat{P}) = \int p(\mathbf{Y}) log\left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})g(\mathbf{Y}|\theta)}\right) d\mathbf{Y} \tag{2.25}$$

$$= \int p(\mathbf{Y}) log\left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})}\right) d\mathbf{Y} - \int p(\mathbf{Y}) log(g(\mathbf{Y}|\theta)) d\mathbf{Y} \tag{2.26}$$

The KLIEP optimization algorithm thus becomes,

$$\max_{\theta} \int p(\mathbf{Y}) \log(g(\mathbf{Y}|\theta)) d\mathbf{Y} \tag{2.27}$$

$$\text{s.t} \int \hat{p}(\mathbf{Y}) d\mathbf{Y} = 1$$

$$\text{and } \theta \geq 0.$$

where the equality constraint comes from the requirement that $\hat{p}(\cdot)$ is a probability density function; and the inequality constraint asserts the non-negativity of the density ratio function. Since, the

unique global optimizer of the above problem , $\hat{\theta}$ can be computed. The optimization problem, in the discretised form becomes

$$\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{b=1}^{p} \theta_b K(Y_i, Y_b) \right)$$
$$\text{s.t } \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{b=1}^{q} \theta_b K(Y_i, Y_b) \right) = 1$$
$$\text{and } \theta \geq 0.$$

From solving the optimization problem, we obtain the optimal parameter $\hat{\theta}$ and the direct-density ratio estimator, $\hat{g}(\mathbf{Y}|\theta) = \sum_{b=1}^{q} \theta_b K(\mathbf{Y}, \mathbf{Y}_b)$, which shall be used for comparing the density ratios of the two sets of sample-points $\{\mathcal{Y}(t), \mathcal{Y}(t+n)\}$.

2.1.3.4 uLSIF

An application of direct-density ratio estimation, to the computation of PE-divergence between two probability distributions was reported in [15] and is abbreviated as uLSIF. It uses the same density-ratio model as KLIEP. However, the learning algorithm is different.

Learning Algorithm In the KLSIP algorithm, the objective was to reduce the KL-divergence, $\mathrm{KL}(P||\hat{P})$. However, in uLSIF, the objective function is chosen to be squared-loss between the true and the estimated probability distribution functions. The following squared loss $J(\mathbf{Y})$ is used

$$e(\mathbf{Y}) = \frac{p(\mathbf{Y})}{p'(\mathbf{Y})} - g(\mathbf{Y}|\theta) \tag{2.28}$$

$$J(\mathbf{Y}) = \frac{1}{2} \int p'(\mathbf{Y}) \Big(e(\mathbf{Y})^2 \Big) d\mathbf{Y} \tag{2.29}$$

$$= \frac{1}{2} \int p'(\mathbf{Y}) \left(\frac{p(\mathbf{Y})}{p'(\mathbf{Y})} \right)^2 d\mathbf{Y} + \frac{1}{2} \int p'(\mathbf{Y}) \Big(g(\mathbf{Y}|\theta) \Big)^2 d\mathbf{Y} - \frac{1}{2} \int p'(\mathbf{Y}) \Big(p(\mathbf{Y}) g(\mathbf{Y}|\theta) \Big) d\mathbf{Y} \tag{2.30}$$

The above optimization problem, when considering the variable terms and after discretizing the terms becomes,

$$\min_{\theta} \frac{1}{2}\theta^T \hat{H}\theta - \hat{h}^T\theta + \frac{\lambda}{2}\theta^T\theta, \tag{2.31}$$

$$\hat{H}_{ij} := \frac{1}{n}\sum_{k=1}^{n} K(Y'_k, Y_i)K(Y'_k, Y_j) \tag{2.32}$$

and the $\hat{\mathbf{h}}$ is the q-dimensional vector with the b-th element given by

$$\hat{h}_b := \frac{1}{n}\sum_{k=1}^{n} K(Y'_k, Y_b) \tag{2.33}$$

The optimal parameter, $\hat{\theta}$ can be determined, thus giving us the direct density-ratio estimator $\hat{g}(Y) = \sum_{b=1}^{q} \theta_b K(Y, Y_b)$.

Change-point detection by uLSIF The PE Divergence between the two distributions $\{P, P'\}$ is then numerically computed using

$$PE(P||P') := -\frac{1}{2n}\sum_{j=1}^{n} \hat{g}(\mathbf{Y}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\mathbf{Y}_i) - \frac{1}{2} \tag{2.34}$$

In essence, at every time-step, k, the Divergence value is computed for the extracted sets of sample-points, to extract the change-point score. A useful estimate of the parameters are $k = 10$ and $n = 50$ for the above two-algorithms.

2.2 Analysis of set of signals

2.2.1 Temporal analysis

The set of temporal sequences of a particular class of signals, if correlated with each other, would imply that they are coupled together by some form of communication - either explicit or implicit. By explicit communication, we consider intentional exchange of information via exchange of ions/chemical markers between the cells. By implicit communication, we consider each cell's independent interaction with its immediate environment, and the subsequent affect of the environment on the neighbouring cells.

While analysing the set of time-series of a particular class-type, we can make the following assumptions:

(1) The signals from each cell influences neighbouring cells through the exchange of chemicals with the environment. The mutual exchange of information creates a closed loop interaction.

(2) Each cell signal can be considered to be formed by the superposition of spatially dependent noise signals upon the independent time-series processes unique to each cell.

2.2.2 Spatial analysis

The spatial pattern observed in a given experiment, is a synthesis of dynamic processes operating at different spatial and temporal scales. Hence, the spatial structure at any given time can be understood to be one realization among several potential outcomes of the interactions among the processes. To make meaningful biological interpretations of the spatial pattern, we need to make some assumptions about the underlying processes.

One such assumption, is that the exogenous process affecting the distribution of the cells over the surface of the medium, is a stationary(homogeneous) process i.e the properties of the process are independent of the absolute location and direction in space.

The property of stationarity is required for making inferences from a model that characterizes the process of the spatial structure of data at locations that are not sampled i.e where the cells do not exist. Therefore, this assumption would help us treat the observed space as a continuous field upon which finitely many bacterial cells are laid, which then act as sensors of the local strength of the field. We do not have any rational basis for making this hypothesis, but its a good assumption to make, while we explore the results of applying statistical tools that correspond to this assumption.

Because stationarity is a property of the process, it cannot be tested directly. However, we need to note that only when stationary process prevails in all the study area, can the spatial statistics be used to characterize the study area with a single value([6]).

Significance testing of spatial data The significance of an observed measure is evaluated based on the assumption that the statistics computed using the observed data follows a reference distribution. When this reference is not known, boot-strap procedures can be used to generate the reference distribution from the data.

The basic procedure to generate the null reference distribution is:

(1) reallocate the sampled values of a variable over the sampling locations, with replacement.

(2) recompute the statistics for the new realization

(3) iterate a finitely large (10,000 or more) number of times

The probability at which the statistical decision of accepting or rejecting the null statistical hypothesis is made is proportional to the number of randomizations generated. Then this reference distribution is used to assess the probability of the observed data where the precision of the probability depends on the number of randomizations. However, the above methodology assumes that the realizations are independent realizations of a process. The presence of spatial dependence in determining the values of the measured variable, however can impair the use of randomization tests. Randomization tests assume that the sampled values are exchangeable, so that any arrangement that might arise by shuffling them is equally likely.

2.2.2.1 Spatial analysis of population data

In our experiments, we consider a case where there is a map of the locations of all 'events', which in our case consists of the location of the individual cells. The methods for analysing event-location data can be based on

(1) determining the neighbours of each event and making calculations based on the distances to them

(2) counting events in circles of given range of sizes centred on the events or randomly placed points.

(3) if the events have a quantitative variable(voltage, calcium reading) associated with them, then 'marked process' analysis can be used.

that depends on the pattern of points.

We now consider a set of methods to analyse maps of the positions of all points(events) of a particular type in a study plot. A test-statistic is calculated from the data and then compared to the expected value of the statistic under the null hypothesis of complete spatial randomness. It is important to make the distinction between patterns that are random and those that are under-dispersed(clumped/aggregated) or over-dispersed(spaced or regular) by comparison. The appearance and interpretation of pattern can change with the scale of study.

The basic ideas behind the available approaches are two-fold

(1) Nearest neighbour methods: How big can a circle centred on an event (or a random point) get before it encounters another event?

(2) Second-order methods: Given a circle of a given size, centred on an event, how many other events does it contain?

One of the most commonly used methods is called Ripley's K method. The approach is based on the concept that, if λ is the density of events per unit area, the expected number of points in a circle radius t centred on a randomly chosen point is some function of t, $\lambda K(t)$. If the points are over-dispersed, $K(t)$ will be close to 0 for small radii and increase for larger distances.

The calculation of the statistic for a given radius t, is based on counting all pairs of points separated by distance less than t. The sample estimate of the statistic $K(t)$ is computed as,

$$\hat{K}(t) = \frac{A}{n^2} \sum_{i \neq j}^{n} h_i(t) I_t(i,j) \tag{2.35}$$

$$h_i(t) = \begin{cases} 1 & c_i(t) \in A \\ \frac{a_{ij}}{A} & \text{asd} \end{cases} \tag{2.36}$$

where A is the area of the plot, with d_{ij} being the distance between the points i and j, $c_i(t)$ is the circle centred at node i with radius t, a_{ij} is proportion of the circle's area within the plot.

If the events follow complete spatial randomness, the number of points in a circle follows a Poisson distribution and the expected number of events in a circle of radius t is $\frac{n\pi t^2}{A}$. $\hat{K}(t)$ is compared with this expected value by subtracting the observed from the expected:

$$\hat{L}(t) = t - \sqrt{\frac{\hat{K}(t)}{\pi}} \tag{2.37}$$

Plots of $\hat{L}(t)$ can be interpreted with positive values indicating over-dispersion and negative values indicating clumping.

2.2.2.2 Spatial analysis of sampled data

The covariance and correlation of the values of a single variable with itself for all pairs of sampling units that are separated by a given spatial lag is called 'spatial autocorrelation'. The spatial auto-covariance, $C(d)$, of the variable x can therefore be estimated by computing the product of the deviation of the value of the variable x at the location i and at location i+d, from the expected value, $E(x_i)$ respectively.

$$C(d) = \mathbb{E}[(x_i - \mu_x)(x_{i+d} - \mu_x)] \tag{2.38}$$

where μ_x is the spatial mean. The spatial autocorrelation, $\rho(d)$, of the variable x at a distance class d, is the auto-covariance divided by the variance of the variable

$$\rho(d) = \frac{C(d)}{C(0)}, \tag{2.39}$$

$$C(0) = Var(x) = \sigma^2. \tag{2.40}$$

2.2.2.3 Dealing with spatial autocorrelation

When the assumption of independence of individual observations in the data under study, is violated, we need to understand the consequences of the spatial dependence on the description of the process. One source of the phenomena, is autocorrelation in the data due to causal interactions within the measured variable('inherent spatial dependence') or due to a functional dependence of the measured variable on an underlying variable which is itself auto-correlated ('induced spatial

dependence'). Autocorrelation is scale dependent, as we expect the physical processes based on movement of materials and energy cannot be independent of distances.

2.2.2.4 Determining Spatial Auto-Correlation: Moran's I and Mantel Test.

Moran's I coefficient

The Moran's I coefficient, is a coefficient of spatial autocorrelation, which is computed for distance class d as follows:

$$I(d) = \frac{\sum_{i \neq j} \sum_{j \neq i} w_{ij}(d)(x_i - \bar{x})(x_j - \bar{x})}{W(d)} \frac{n}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.41}$$

where $w_{ij}(d)$ is the distance class connectivity matrix. Positive auto-correlation is indicted by positive values, and the expected value for the absence of spatial autocorrelation is close to 0. The values generally lie in the range $(-1, 1)$. However, when there are too few pairs of sampling locations in distance class d, and the spatial layout of data looks non-stationary, the estimated value is unstable and can fall outside the expected bounded range.

Geary's C Coefficient

To avoid measures of spatial pattern based on deviations from the arithmetic spatial mean, the Geary's C Ceofficient was proposed that measures the difference between values of a variable at nearby locations so that the degree of spatial autocorrelation is based on differences of the variable values at a given distance class.

$$c(d) = \frac{\sum_{i \neq j} \sum_{j \neq i} w_{ij}(d)(x_i - x_j)^2}{2 * W(d)} \frac{n-1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.42}$$

This coefficient behaves like a distance measure and takes values from 0 to 2 or higher. A value of 0 indicates the highest value of positive autocorrelation and a value greater than or equal to 2, indicates strong negative autocorrelation. The expected value, $\mathbb{E}[c] = 1$, indicating the absence of spatial auto-correlation.

Variogram

In the field of 'geostatistics', the spatial structure is estimated from sampled data by computing

the spatial variance , which is then used to predict the values at unsampled locations by modelling the spatial structre using techniques known as krigging. The basic idea consists of assuming that the value of a variable z at a given location x is a particular realization of a random variable $Z(x)$. The observed value is composed of three components,

$$z(x) = m(x) + \epsilon(x) + \epsilon$$

where m(x) is the deterministic structural function of the variable at the location x; $\epsilon(x)$ is the spatially dependent residual from $m(x)$ and ϵ is the spatially independent normally distributed residual component. When we assume stationarity, $m(x) = \mu_x$, i.e $m(x)$ equals the spatial mean of the variable in the given study area. The spatial variance of the quantitative variable , z is estimated by the semi-variance function, $ga\hat{m}ma(h)$:

$$ga\hat{m}ma(h) = \frac{1}{2nh}\sum_{i=1}^{n(h)}(z(x_i) - z(x_i + h))^2 \tag{2.43}$$

where, the value of variable z is obtained at the i'th sampling location x_i and $n(h)$ is the number of pairs of sampling locations located at distance h from one another. Note that the semi-variance function is in the same units as the analysed data, and is not bounded in value.

The above tools were applied to determine the nature of the spatial-distribution of the data. However, the traditional approach to a rigorous use of the tools, involves determining the experimental variogram, having a mathematical model of the variogram, and estimating features from the same. We came across this domain of signal-analysis towards the very last stage of the research period, and hence we could not pursue this direction to the depth we intended.

2.2.2.5 Correlation between distance matrices

To measure the spatial relationship between spatially autocorrelated variables, we can constitute distance matrices corresponding to the different measured variables. The correlation between the distance matrices, then gives an indication of the similarity between the corresponding variables conditioned on the choice of the distance-metrics. This method of summarizing the spatial pattern

of data using a single number computed as a cross-product between two matrices is known as the 'Mantel Test'.

The null hypothesis of the Mantel test is that the distances in some matrix X are independent of the distances in another matrix Y sampled at the same locations. The Mantel statistics can be computed as follows,

$$Z_M = \sum_{i \neq j} \sum_{j \neq i} x_{ij} y_{ij} \tag{2.44}$$

$$= \mathbf{X} \cdot \mathbf{Y}, \text{ Hadamard Product} \tag{2.45}$$

and it is used to measure a linear relationship between the two symmetric matrices. Since, Z_M is unbounded, each matrix can be standardized before computing the Mantel correlation statistic to obtain a bounded r_M statistic that is bounded to lie within the range $(-1, 1)$.

$$r_M = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{x_{ij} - \bar{x}}{\sqrt{(x_{ij} - \bar{x})^2}} \frac{y_{ij} - \bar{y}}{\sqrt{(y_{ij} - \bar{y})^2}} \tag{2.46}$$

where, d is the number of elements in the lower triangular part of each matrix. Along with the computation of the statistic, one needs to determine the significance of the computed value, which can be obtained by a method of restricted randomization of one of the matrices, to ensure that the relationships between pairwise distances measured between sampling locations is not changed. This is accomplished by randomly shuffling the rows and columns of one of the matrices, while keeping the other constant.

Partial Mantel Test A partial correlation approach can be used to determine the contribution of a third variable, to the measured correlation between two variables. The procedure for the same, when applied to computation of the correlation between the distance matrices, corresponding to the two variables, as in the Mantel test, is called the 'partial Mantel three-matrix test'. The partial mantel statistic, $r_{XY.Z}$ is computed by detrending the linear effects of the values in matrix Z, on those in matrix X and Y, using a linear regression for the pairs $\{(X, Z), (Y, Z)\}$. Then, the Mantel test is performed on the derived residuals from both regressions, $Res_{X|Z}, Res_{Y|Z}$.

Alternatively, it may be computed as follows

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{1 - r_{XZ}^2}\sqrt{1 - r_{YZ}^2}} \tag{2.47}$$

Summary We have thus provided a broad description of the tools and techniques that were used to determine answers to the different biological questions we pursued. For more details regarding these tools, its advisable to refer to the citations indicated in the text, as they were the original primary or secondary sources from which these tools were studied.

Chapter 3

High-Dimensional Data Analysis

A time-series (one trace from the database) comprising of n-time samples can be considered to be an n-dimensional vector. The longer the time-duration, the larger the dimension of the space of representing a data-point. Further, our measurement system and observations can only permit us to sample a very small/finite number of data-points in that very large-dimensional data-space. Therefore, the analysis of this dataset would require we apply advanced algorithms, that are applicable to this domain of inquiry and research.

In this section, we intend to apply concepts from High-Dimensional data-analysis for the purpose of mining patterns in the dataset and hopefully predict and extract information about the bio-physical processes that constitute the lifetime of a Bacteria. We adopt the following architectural-framework for guiding our choice of methods, in the analysis of high-dimensional datasets.

(1) Embedding of Dataset into a Metric Space:

It is important to note that the representation of data-files in memory, needs to be chosen based on the application for which that data shall be utilized.For the purpose of classification, we need to design/develop a particular vector-representation that incorporates certain "descriptive features" of the traces. These "features" may be based on analytical notions of time-frequency domain representation. The representation of these features, as mathematical vectors embedded in a high-dimensional space, then permits us to use algorithms for processing vectorial data, for the purpose of automated classification/clustering. We intend to use the "spectrogram coefficients" and possibly other "high-level-feature" information

for each of the trace files.

(2) Visual Pattern Detection of Biophysical Trace-Database:

The preliminary step involves the use of a Data-visualization library, to determine whether there are any patterns like clusters in the data, which can then be exploited. It would be helpful to understand the logic/ideas behind the algorithms used to guide the visualization process, which could then be incorporated into the classification engine, that we need to develop.

(3) Determine Intrinsic Dimension of the Data:

If any patterns(clusters) have been detected in the data-visualization, it helps to assume that the different genres of high-dimensional data in fact is distributed along low-dimensional manifolds embedded within the high-dimensional space. We can then use algorithms([10]) for determining the intrinsic dimension of these manifolds (e.g "Correlation Dimension"). By determining the intrinsic-dimension of each of the genres, we can determine the minimum dimension of the space into which all the data can be embedded without losing much information.

(4) Metric Learning:

Ideally, when data are categorized into categories, we would like to believe that the traces/data-points that are "similar" to each other are "closer" to each other, and those that are "dissimilar" are "farther" apart. The closeness of two data-points is determined by the "distance metric" used to compute the distance between the two points. The reason such a "metric" would exist, can be deduced from the fact that the data points are distributed upon a Manifold. The measure of distance between any two points, thus becomes the distance travelled along the manifold surface between the points, rather than the shortest euclidean straight-line distance between them. The "metric-learning" problem deals with the problem of ascertaining the metric for each of the manifolds separately, or the notion of a "global-metric" for all the data-points.

(5) Dimension Reduction:

Once the minimal intrinsic dimension of the datasets has been determined, and we have a notion of "distance-metric" that helps compute the distances between the points, we then deal with the problem of reducing the dimension of the data. We note that this procedure, helps us determine how much of the information from the initial choice of vector representation of the dataset is redundant. The solution to this problem, is the design of a Map/function that projects every data-point from the high-dimensional space to the low-dimensional space. To ensure that the representation for each data-point is unique in the reduced dimension, we impose the constraint of an injective mapping, on this function.

(6) Clustering:

If we have identified patterns in the data-visualization stage, we would assume that the traces belonging to the same music-genre are distributed as clusters in the high-dimensional space. We expect that the dimension-reduction process does not hamper/destroy the clustering of the dataset, but rather reinforce/amplifies the extent of clustering - "similar" traces are brought closer together and "dissimilar" traces are pushed farther apart. It is obvious that the "distance-metric" used to measure the distances between traces in the reduced-dimensional space is very-different from that in the high-dimensional space. However, we can demand that the distance metric in the reduced-space be closer to the Euclidean space. The "metric" in the reduced-space shall then be used to identify clusters in the dataset.

(7) Query Processing (Cluster Assignment/Statistical Learning):

Once the abstract model - a spatial distribution/representation of the traces in the database has been determined, we have now developed the capacity to determine the genre/family that a particular trace-query might belong to. We can develop a method to determine different classes representing differentiation within species of bacteria, by determining the cluster that it belongs to, the proximity to its neighbours and the classification of its neighbours into the different classes.

3.0.2.6 Feature vector representation

To derive content-based features representative of each trace, we would have to implement digital signal-processing algorithms upon the .wav trace files, to compute quantitative information from each trace. Features extracted could be based on time-domain, frequency-domain or time-frequency domain. In this section, we focus on using the Spectral Information of the traces-represented by the spectrogram, to serve as the primary signature of a window-frame. Traces are thus analysed by using moving window of short-time duration. The window-length is arbitrarily chosen. To further, obtain a compact representation of the features, we proceed to develop a code-word histogram representation of each trace. (Dis)Similarity between the traces is then measured in the vector space of this final reduced feature-vector representation.

Code-word Histogram The code-word histogram [22] approach to feature-vector representation of a trace can be described intuitively as follows. All the traces in the database can be assumed to be 'sentences' that are made up of a fixed number of 'code-words' in the vocabulary of bacterial inter-cellular electrical signals. Each trace is represented by the number of times specific 'code-words' occur in it. The word-histogram is used to describe this method of counting the number of times a given code-word occurs in the trace. Thus, by assuming such a finite word vocabulary to describe all possible traces in the database, we enable a mechanism of drawing parallels between comparison of textual-documents via Natural-Language processing algorithms, and the analysis of biophysical signals.

Mathematically, the code-word histogram generation process can be described as follows. To obtain a compact summary of bacterial electrical traces, each trace is represented as a histogram over a dictionary of codewords. As a first step, a codebook is constructed by clustering a large collection of feature-descriptors. Once the codebook has been constructed, each trace is summarized by aggregating vector-quantization representations across all frames in the trace, resulting in codeword histograms. Finally, histograms are represented in a nonlinear kernel space to facilitate better learning of the distance metric between the traces.

Codebook training The primary spectral-feature descriptor used for training of the codebook, is the spectrogram coefficients of frames of a trace. The spectral-feature descriptors of all the traces in the database is aggregated into a single bag-of-features, which is then clustered to produce the codebook. In our implementation, the number of words in the codeword dictionary is arbitrarily chosen to be about 512.

For each trace x in the codebook training set X_C, we compute the spectrogram coefficients. These descriptors are then aggregated across all $x \in X_C$ to form an unordered bag of features Z, where each $z \in Z \subset \mathcal{R}^D$ is either an spectrogram feature-vector representation of the trace-frames.

To correct for changes in scale across different dimensions of $z \in Z$, each vector is normalized according to the sample mean $\mu \in \mathcal{R}^D$ and standard deviation $\sigma \in \mathcal{R}^D$ estimated from Z. The i'th coordinate is mapped by

$$z[i] \mapsto \frac{z[i] - \mu[i]}{\sigma[i]}. \tag{3.1}$$

The normalized feature vectors are then clustered into a set $\mathcal{V}$ of $|V| = 512$ codewords by k-means algorithm.

(Top τ) Vector Quantization Once the codebook $\mathcal{V}$ has been constructed, a trace $x \in X_C$, is represented as a histogram h_x over the codewords in the codebook. Each trace $x \in X_C$ is understood to be a time-sequence of feature-vectors, $z = z_i \in Z \subset \mathcal{R}^D$ where z_i is the feature-vector representation of a frame in the trace. Each $z_i \in z$ is normalized according to (3.1). The codeword histogram of trace $x \in X_C$ is constructed by counting the frequency with which each codeword $v \in \mathcal{V}$ quantizes the elements of z i.e

$$h_x[v] = \frac{1}{|z|} \sum_{z_i \in z} \frac{1}{\tau} \mathbb{1}\left\{ v = arg \min_{u \in \mathcal{V}}^{\tau} ||z_i - u|| \right\}. \tag{3.2}$$

where we have chosen to adopt multiple codeword quantizers of each vector z_i by defining the quantization set

$$arg \min_{u \in \mathcal{V}}^{\tau} = \left\{ u \text{ is a } \tau\text{-nearest neighbor of } z_i \right\}. \tag{3.3}$$

where $\tau \in 1, 2, \cdots, |V|$. Note that the codeword histograms are normalized by the number of frames $|z|$ in the trace in order to ensure comparability between traces of different lengths. Further, the normalization by $1/\tau$ ensures that $\sum_v h_x^\tau[v] = 1$, so that for $\tau > 1$, h_x^τ retains its interpretation as a multinomial distribution over $\mathcal{V}$.

3.0.3 Distance and Similarity in trace-space

The choice of the Distance-Metric depends on the chosen Feature-Vector-representation for each trace. The naive method to proceed would be to assume that all the traces are embedded in an Euclidean Space, and hence the distances can be computed using the L-2 Norm. However, for the specialized representations of the traces that we have adopted, we believe that the traces are distributed over a manifold, and hence the distance between the points need to be computed by using a metric intrinsic to that manifold. One approach would be to use metric-learning algorithms to learn the metric between the data points.

Conventional choices of the distance measure between the traces(high-dimensional dataset) are:

(1) Euclidean Distance

One can treat each n-sampled trace as a n-dimensional vector and then compute the euclidean distance of the same from another trace.

(2) Cosine Distance

One method to determine the distance between the traces, is to compute the Correlation Matrix between them. For any pair i,j of traces, we compute the cosine of the angle between them, by computing the inner-product of the normalized traces.

3.0.3.1 Similarity between traces in trace-space

The distance matrix computed above can further be processed by choosing a particular threshold value. We determine the adjacency matrix of the graph, representing the relation-

ship/relatedness between the traces. The graph-based path-distance between the traces, gives a measure of the similarity between the traces. Graph based methods for dimensional reduction can then be pursued for the given dataset.

When using only the distance information to describe similarity, we can choose the function to be any number of non-linear functions like:

(1) Thresholding $s_{ij} = \imath[d_{ij} > \theta]$, where θ is chosen threshold value.

(2) Gaussian Similarity $s_{ij} = exp(-d_{ij}^2/2\sigma)$,

(3) Inverse distance $s_{ij} = \frac{1}{1+d_{ij}^p}$, where $p \geq 1$,etc.

We use the Similarity Matrices computed above, as inputs to some of the nonlinear dimension reduction techniques like the Refined-Graph-Embedding.

3.0.4 Data Visualization

The t-SNE algorithm [19] is an award-winning algorithm, for visualizing high-dimensional datasets. We shall first implement this algorithm, to determine a possible visual representation of the database, that shall indicate visual patterns in the data. This visualization shall help us develop an intuition of what the data distributions might look like.

This algorithm provides us the opportunity to check the efficiency of the vector-space embedding of the selected features from the traces in our database. If the visualization does not separate our vector-space data set into a sufficient number of discrete clusters, then it is likely that our choice of features is inadequate to truly separate the different genres when implementing our own dimension reduction techniques. In this case we will augment the vector-space embedding by adding additional features from the traces. This visualization algorithm can also be used after each stage of the pipeline to check that the information in our dataset has not been adversely deformed. Further, this algorithm can also be used a dimension-reduction technique, and shall be described in the appropriate section below.

3.0.5 Dimension reduction

We shall create and compare multiple pipelines for the dimensional reduction process. Both linear and non-linear graph based methods shall be explored.

Non-linear, Graph-based Refined Embedding. This method builds a similarity graph, between all the datapoints, where the edge-weights are computed using a particular kernel function and distance-metric. From the similarity graph, we derive the Graph-Laplacian and follow the procedure of spectral embedding of this graph. The eigenvalues and the eigenvectors of the Graph Laplacian shall indicate the existence of the clusters.

3.0.5.1 Refined Graph Embedding

The Refined Graph Embedding relies on theoretical concepts covered in Spectral Graph theory. Let $G(V, E)$, with $|V| = n$ and $|E| = m$, be an undirected graph without self-loops. Let A be the adjacency matrix defined as $A_{ij} = \imath[(i, j) \in E]$ and D be the degree-matrix of the graph $D = diag(d_1, d_2, \cdots d_n)$ with $d_i = \sum_{i=1}^{n} A_{ij}$. We further define the following matrices

$$\begin{aligned} L &= D - A && \text{combinatorial unnormalized graph Laplacian} \\ P &= D^{-1}A && \text{markov probability transition matrix} \\ \mathcal{L} &= I - D^{-1/2} * A * D^{1/2} && \text{normlaized graph laplacian} \end{aligned} \tag{3.4}$$

The largest eigenvalues and the corresponding eigenvectors of $\mathcal{L}$ are computed. Next, a stationary probability distribution is computed for the Markov Chain transition matrix P. The coordinates of the point in the reduced dimensional-space are obtained from the eigenvectors and the stationary-probability distribution computed.

3.0.5.2 t-Distributed Stochastic Neighbour Embedding

The t-SNE algorithm [19], was introduced earlier as a technique for visualization of high-dimensional datasets, to built a visual intuition of the data-distribution. From the fact that local neighbourhood information of the data is preserved by the technique, it would serve as a use-

ful dimensional reduction technique, and hence used this method as an alternative to the Graph Embedding Technique.

tSNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. The similarity of data-point x_j to x_i is the conditional probability $p_{j|i}$ that x_i would pick x_j as a neighbour, if the neighbours were picked in proportion to their probability density under a Gaussian centred at x_i. Mathematically, the conditional probability is computed as

$$p_{j|i} = \frac{exp(-d_{ij}^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-d_{ik}^2/2\sigma_i^2)} \tag{3.5}$$

where σ_i is the variance of the Gaussian centred about x_i. Because, we are only interested in modelling pairwise similarities, we set the value of $p_{i|i} = 0$.

Now, we assume that there's a particular projection operation that projects each of the traces x_i to lower dimensional point y_i. For the low dimensional counterparts y_i and y_j of the high-dimensional points x_i and x_j, we can compute the conditional probability, $q_{j|i}$ as

$$q_{j|i} = \frac{(1 + d_{ij}^2)^{-1}}{\sum_{k\neq i}(1 + d_{ik}^2)^{-1}} \tag{3.6}$$

where, we use a t-Student distribution with a single degree of freedom, instead of a Gaussian distribution about point y_i. This is because, it has the particularly nice property that $(1 + d_{ij}^2)^{-1}$ approaches an inverse-square law for large pairwise distances in the low-dimensional map. This makes the maps representation of the joint probabilities almost invariant to changes in scale of the map for points that are far apart. Which also implies that large clusters of points that are far apart interact in just the same way as individual points, so the optimization operates in the same way at all but the finest scales.

If the projections correctly model the similarity of the points in the high dimensional space, then the conditional probabilities $p_{j|i} = q_{j|i}$. Motivated by this intuition, the tSNE algorithm aims to find a low-dimensional data representation that tries to minimize the mismatch between joint conditional probabilities P and Q. A natural measure of the distance between two distributions

is the Kulback Liebler-divergence. tSNE minimizes the KL-divergence between P and Q, via a gradient descent algorithm. The cost function is given by

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \left(\frac{p_{j|i}}{q_{j|i}} \right) \tag{3.7}$$

where it is assumed that $\sigma_i = \sigma$ for all data-points.

The gradient of the Kullback-Liebler divergence between P and the t-Student based joint probability distirbution Q is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + d_{ij}^2)^{-1} \tag{3.8}$$

where $d_{ij} = ||y_i - y_j||$

3.0.6 Results of Experiment: Codeword Histogram

The plots of the experiments using the spectrogram-codeword-histogram methodology gives a visual representation of the distribution of the traces in the 3 dimensional space. [Explain the procedure of constructing code-word histogram.]

Visually, we identify the pipeline comprising of the following:

(1) Code-word Histograms of the spectrograms

(2) t-Distributed Stochastic Neighbourhood Embedding

(3) Clustering

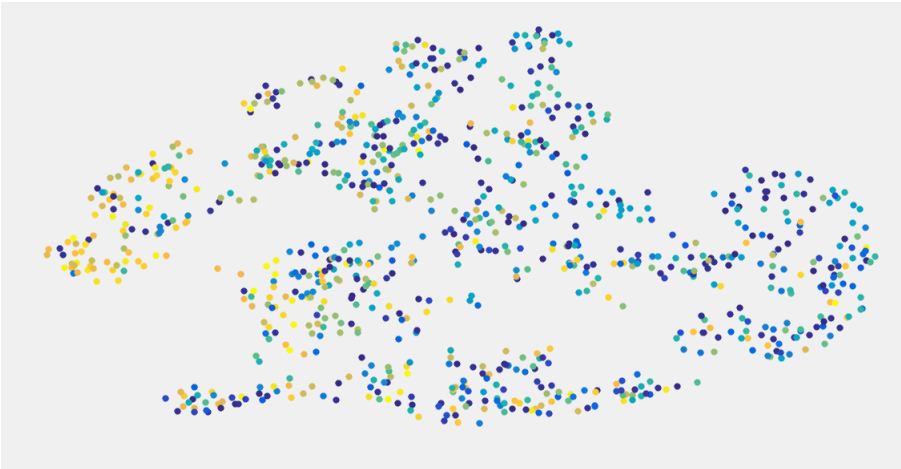
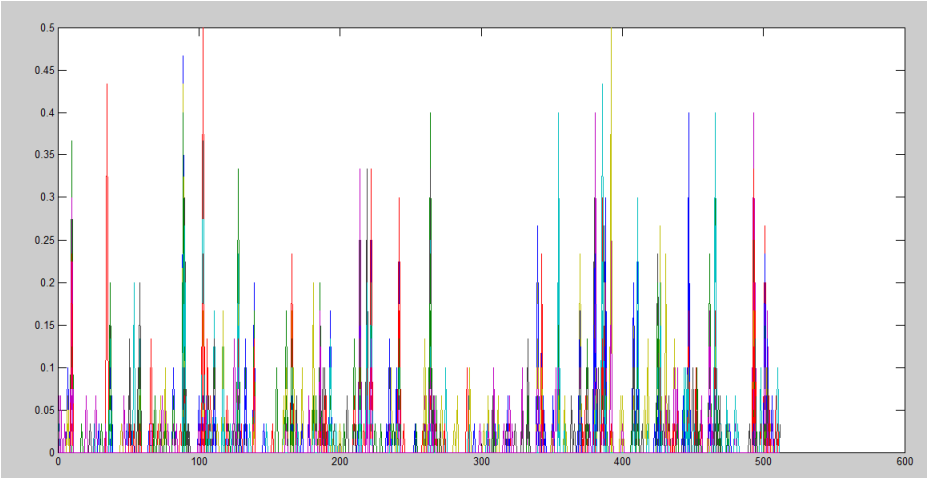


Figure 3.1: Plot of the traces categorized into classes based on graph components obtained by thresholding(0.69) the correlation matrix.



(a) cw-histogram

Figure 3.2: Plot of the traces when represented as (a)time-series dataset ;(b)code-word histograms based on choice of 512 codewords in "language".

Chapter 4

Data Visualization

During the course of the research project, it was understood that some of the tasks of data-visualiztion seemed to be quite repetitive, especially when it came to visualizing time-series information, and also, plotting interactions between the cells based on the correlations between the time-series. Further, the signal-analysis methods adopted, and the graphical plots, seemed to hide away the spatial information that was evident, while observing the video of cell activity.

This need for a better visualization tool, that would facilitate an interaction with the spatial cell distribution, while also depicting the temporal nature of the signals, led to the idea of developing an interactive data-exploration tool, as a deliverable of the research thesis.

While designing the data-visualizer, there were a few basic criteria we needed to adhere to:

(1) Since, we were recording univariate or bivariate signals from each cell, it was determined to build a tool designed for bivariate time-series signals.

(2) In the video recordings, the cell-activity was observed to be blinks of varying intensities. To make the activity more perceivable, it was decided to convert the intensities into a scaled version of cell expansion-contraction. Thus, the change in intensities of the light-signals measured from the cells was converted into pulsating expansion and reduction of the visual-elements indicative of cell-structure. This translation of light-intensity into cell-size, helped to visually determine whether there were any patterns in the behaviour of cells, as relative motion is more easily comparable than relative intensities for the human visual system.

(3) There was a possibility that the cells would be densely distributed in space, with high chances of overlap in many areas. To facilitate the representation of overlapping cells, it was decided to adopt a gradient-coloring scheme that was darkest at the center of each cell, while becoming lighter towards the edges of the cell. This would facilitate visual discrimination of over-lapping cell-structures.

(4) In reality, the cells being observed could be of various shapes and sizes. Ecoli cells are long and tubular. Salmonella cells are spherical. In order to adopt a uniform visualization scheme, the cell-centroids are computed from the video-recordings. Circular visual-elements are chosen to represent biological cells of all shapes. The cell-colour is used as discriminating factor to differentiate between different biological-cell species or different signal types.

(5) Two specific modes of interaction with the spatial cell-distribution were determined.

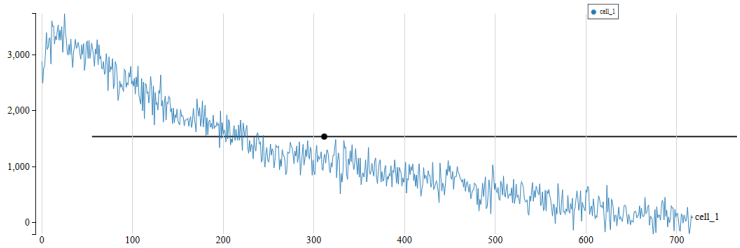
- Point and Click: Cells of interest could be clicked using the mouse-arrow. The time-series signals corresponding to the set of cells that were selected by this procedure, would be shown together in one plot. This method of interaction would help to compare cells that had large pair-wise distances between themselves.
- Drag and Drop: An entire spatial area of interest could be selected by clicking the mouse-arrow in the area representing the petridish, dragging the mouse and dropping the cursor, to define a rectangular sub-section of the petri-dish. The time-series signals corresponding to the set of cells lying within the selected region, would be shown together in one plot.

(6) The window-frame displaying the spatial-distribution of cells, can be treated to be a network of nodes and edges. The edges would be lines between the cells(nodes) that represent some property or relationship between the pair of cells. The edge color and thickness can be controlled to simultaneously represent multiple-features and properties indicative of the relationship between a pair of cells. The node color and radius may also be modified as per

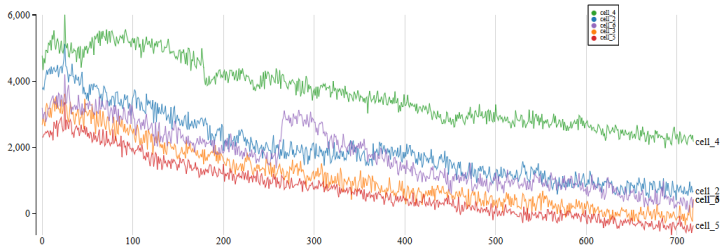
any individual property of the cells.

With the above criteria charted out, a novel data-visualization tool, customized for the biological dataset consisting of spatially distributed cell-units, with bivariate time-series measurements from each discrete spatial location was designed and implemented.

For the sake of ease of implementation, code portability and exchange-reuse of the research tools devleoped by the lab, we decided to adopt web-technologies like HTML, CSS and Javascript, along with W3-standards, to implement the project. The javascript data-visualization library, d3.js was extensively used as the back-bone of the visual-engine, with the visual representations being developed based entirely on the visual structures and attributes associated with the visual elements.



(a) far away cells



(b) local cell neighbourhood

Figure 4.1: Window frames depicting cell-traces for groups of cells in (a) far-away locations (b) local neighbourhood. The black line with a moving ball, is indicative of the position of the cursor and the current time-step in relation to the total time duration of the signal.

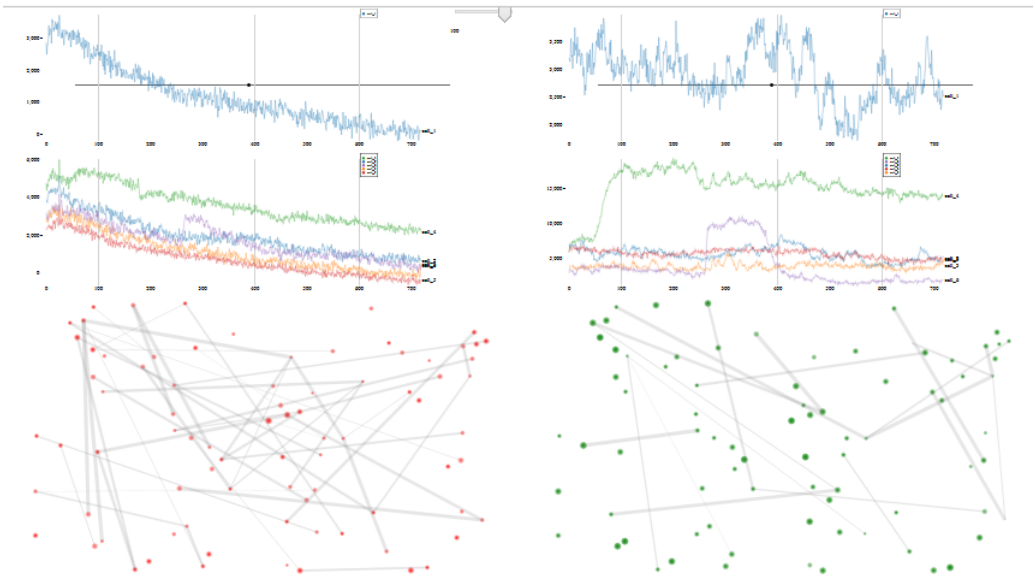


Figure 4.2: Overview of the window-frames depciting the spatial distribution of the cells, along with the voltage(left-pane) and calcium traces(right-pane). The diameter of the cells vary with time, according to the magnitude of the cell-signal being represented.

Chapter 5

Biological Experiments: Analysis of Data

We start with an initial dataset comprising of uni-variate or bi-variate signal measurements along with the spatial location of the centroid of bacterial cells from which the measurements are taken. For the first few months of the research, we were interested in identifying features from the time-series signals measured, which could be visually determined to be represent repeated patterns. The primary approaches were more focussed on time and frequency domain methods, without any resort to modelling of the entire time-series.

With regards to patterns in the signals, we were motivated by the following two approaches:

(1) Motif mining in the time-series

This is a time-domain approach towards identifying patterns in the signals, especially when the absolute magnitudes of the signal does not carry any significant information. We are interested in identifying the relative magnitudes in a subsequence which determines the shape of the subsequence. We term 'repeating motifs' to be significant and demonstrative of repeatable patterns in the signal. The algorithms like the motif enumerator ([23]), SAX([18]) and iSAX ([25]) were explored.

(2) Feature extraction in a zero-resource language

In this approach, we wished to treat every individual time-trace as an audio recording of a natural language, which we had not heard before, and of which we have not had any previous annotated content. The research attempt was directed towards reading up

literature on the same, and determining what sort of features can be extracted from the signals for identifying words and phrases in the different time-traces. The features being studied were primarily in the frequency-domain. One of the methods that we explored was the idea of Intrinsic Spectral Analysis([12]).

The above approaches were good experimental approaches, but the results of application of the algorithms didn't seem to help us extract any particularly significant conclusions. Further, the interpretation of the outcome of the algorithms, was especially difficult as we did not have any prior information from the biological system, which could indicate that the observed patterns were for real, or whether they were just artefacts of the algorithms used. This first stage of the exploratory data-analysis, was terminated after we realized that a blind pursuit of patterns in a generic dataset, might not be a fruitful direction of research.

From the above approach, we realized that we needed to create biological experiments, that would help us ascertain the occurrence of a biological event, and thus indicate the presence of patterns that ought to be modified as a consequence of the biological event. If we could identify features in the signals, that underwent changes as we changed the physical conditions of the bacterial cells, then it's possible that those patterns might be relevant and their significance as a signature to distinguish different states of existence can later be studied. My primary task was to identify a couple or so features that could help discriminate between the different states of existence induced by different physical/chemical conditions of existence.

The following biological experiments were designed to simulate scenarios where we expected biological-events to occur, or atleast to create a change in the bio-chemical environment of the cells:

(1) Study of different species

The same voltage and calcium measurements were recorded from Ecoli and Salmonella cell-species, under similar physio-chemical conditions. The cells were prepared to express the protein-sensors of both the membrane-potential and the calcium ion concentrations. Thus, we recorded a bivariate signal from each cell-line.

(2) Study of the effect of variation in population density

For a given species of cells, we modified the population density upon the petridish before measurements are recorded, by diluting the batch of cells with a sterile medium. Three different densities were prepared, corresponding to the following labels - 'LOW', 'MEDIUM' and 'HIGH'.

(3) Study of the effect of toxins upon the cell-signals

A given batch of cells were subjected to a significant dosage of different toxins to record cell-signal behaviour as the cell-environment is changed from a nutritions medium to a toxic medium. Three videos were prepared corresponding to the following states -

- 'BEFORE' : In this stage, the cells inhabit a nutritious environment, conducive for normal(healthy) cell-behaviour.
- 'DURING' : In this stage, while the measurements are being recorded, a specific concentration of toxin is uniformly introduced to the cell-culture, to ensure that all cells experience the similar stress in its physio-chemical environment.
- 'AFTER' : After the earlier stage of toxin infusion, we wait for the biological system to attain a state of equilibrium, after which the cell-signal measurements are again recorded. We expect that the cells would have attained a different mode of functioning in the presence of a toxic medium, and that this change would be indicated in the signals being measured.

For the same species, two sets of univariate signals - one set measuring the voltage and another set measuring the calcium ion concentration signals were recorded at different days, and analysed. The cells were not induced to express both the sensor proteins.

For all the above experiments, the experimental set-up consisted of the following

(1) Composition of Medium:

- 1% Agarose
- 50mM Sodium Phosphate
- 25mM Potassium Phosphate
- 25mM Ammonium Chloride
- 20mM Glucose

(2) pH of the medium was established at 7.5

(3) Concentration of the cell-culture:

The 'HIGH' density started at 10^6 cells per milliliter (1:1), which was then diluted further for different lower densities.

(4) Concentration of Toxins:

The toxins were applied at 20mM concentrations. The following toxins were considered: Carbonyl-Cyanide m-Chlorophenyl Hydrazone(CCCP, a chemical inhibitor of oxidative phosphorylation), 2-Deoxy-D-glucose(2DOG, a marker for tissue glucose uptake), Indole (a quorum sensing molecule) and Kanamycin(an antibiotic).

5.0.7 Experiment 0: The general case of univariate/bivariate signal measurements

5.0.7.1 Design of experiment

The experiment considered here isn't a single unique experiment. We consider different sets of data obtained from different batches of the same cell-line, which were observed on different days. This helps us establish a ground-line on the general behaviour of the cell-signals, which could be used to understand the nature of signals. Further, we adopted this approach, as we needed to determine what could be said, about the cell-signals, without any prior-knowledge about biological events or changes that the cells might be undergoing during the course of the measurement period. The cell-line consists of Escherichia Coli bacterial cells, which are gram-negative, anaerobic and

rod-shaped cells of the Enterobacteriaceae family. Most of the E.Coli strains are harmless and are often found as a constituent of human gut biota.

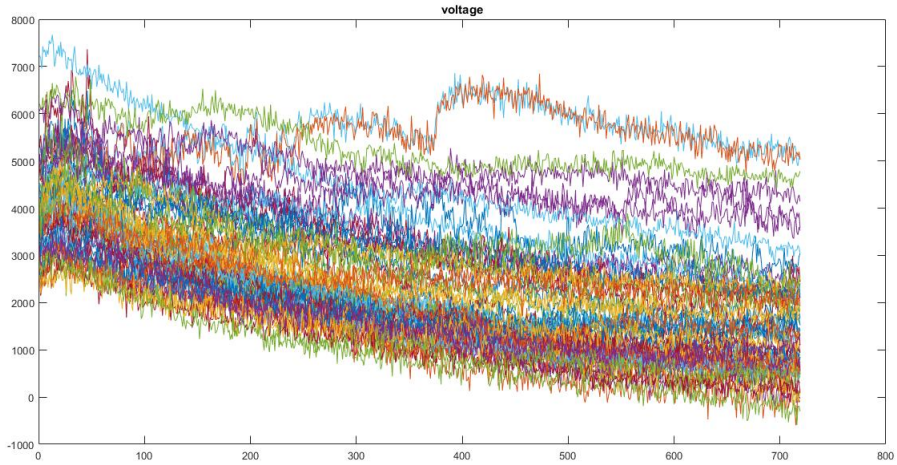
5.0.7.2 Analysis of measurements

We shall examine the results of the application of most of the tools that were described in the previous sections upon the available dataset. The results of the application of different tools were approached in a very qualitative manner, as our agenda was not to determine the perfect tool, but to have a broad experience with the use of different tools and ideas, and to learn from the outcome of these computational experiments, rather than to take a decision with regards to a single specific tool.

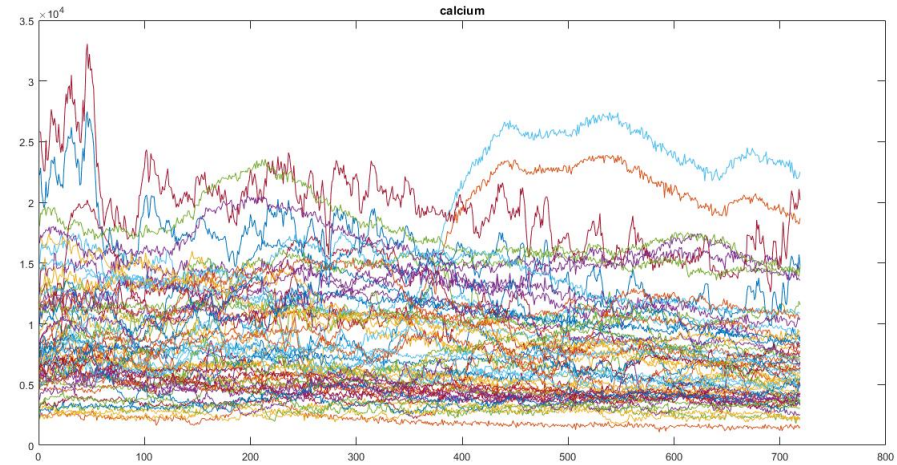
A typical set of measurements that we observe, is represented below(Fig.5.1), where we have plotted the population traces of the voltage and calcium signals. One of the first steps we undertake, is to determine the relationship between the present value of an observation, with the past value of the observation, which is called the delay-space representation of the signals. Further, we also consider the first difference of the signal, to correspond to the velocity of the observed variable, and plot it with the immediate value of the original signal. This plot is considered to be the phase-space representation of the signal. From the two plots(Fig.5.2), we realize that it is possible to relate the present measurements to the immediate past measurements by a approximately linear relationship.

Assuming spatial independence of the signals measured, its possible to treat the measurement of either the voltage or the calcium signal at any given time-instance, to be equally likely for each of the cell-locations. And thus, by using a statistical procedure called bootstrapping, we may obtain the spread of the values, in addition to the mean value at any given time instant. This is illustrated in the Fig.5.3. The decrease in the spread of the values at certain time-instances becomes indicative of some sort of coordination or interaction between the cells, due to which all the cells seem to attain values very close to each other.

One of the initial steps to determining patterns in the signal, was to apply the change-point detection algorithm to the noisy and the denoised signal. We expected to identify patterns in the



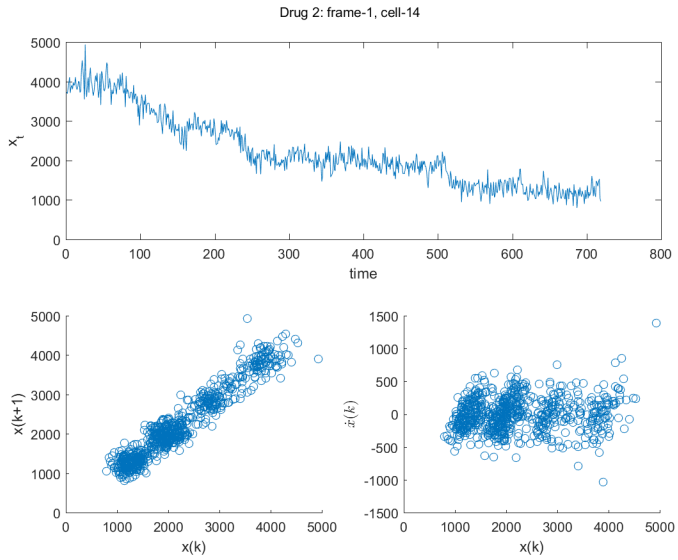
(a) voltage



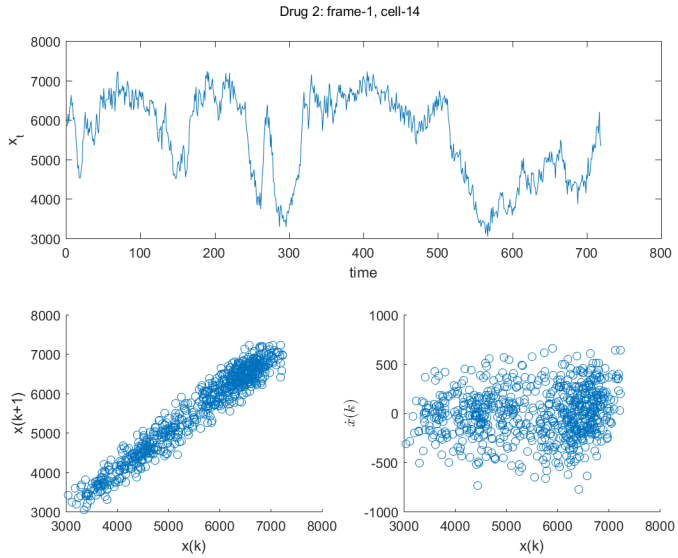
(b) calcium

Figure 5.1: Delay-space and phase-space representation of the voltage and calcium signals of a randomly chosen cell.

occurence of events, or the inter-event durations in the bi-variate time-series. A typical example for the consequences of applying the algorithm on the dataset, is represented in the plot below(Fig. 5.4). We note that under certain circumstances, the voltage-signals indicates change-points before the corresponding change-points in the calcium signal, and at other instances the converse is observed. This is suggestive of a feedback mechanism existing between the signals. Further, we attempted to study the histogram(Fig.**??**) of the times at which the maximum change-score events were detected,



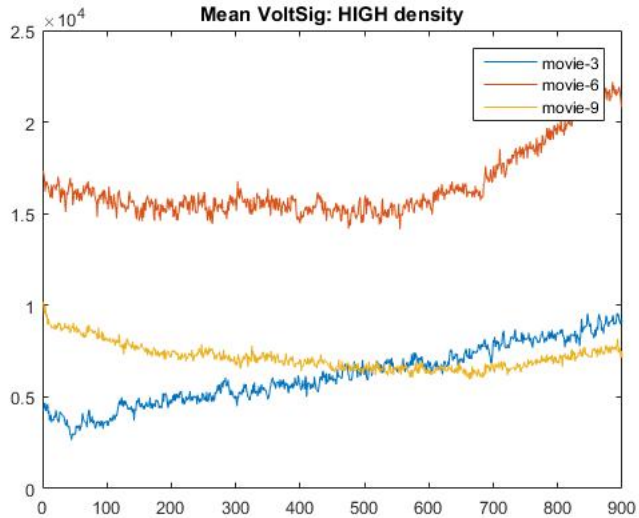
(a) voltage



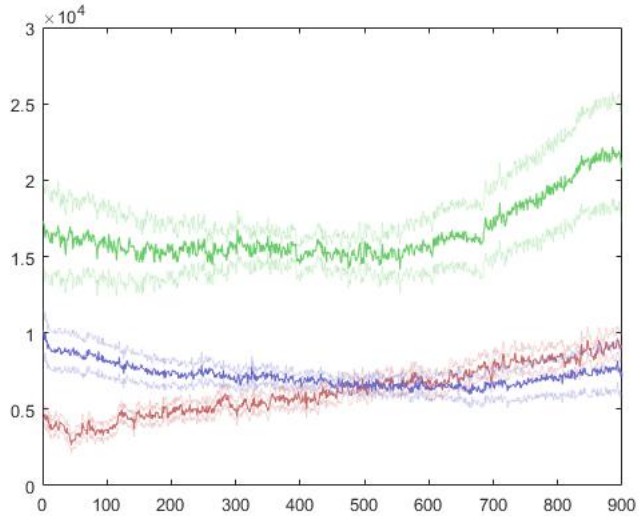
(b) calcium

Figure 5.2: Delay-space and phase-space representation of the voltage and calcium signals of a randomly chosen cell.

and also the time-intervals between consecutive change-point events.The motifs were extracted for windows of length ten time-samples in a time-series that stretched about 200 time-steps. We did not feel that this direction of research was helping us unravel anything interesting about the system, and hence abandoned further detailed exploration of the same.



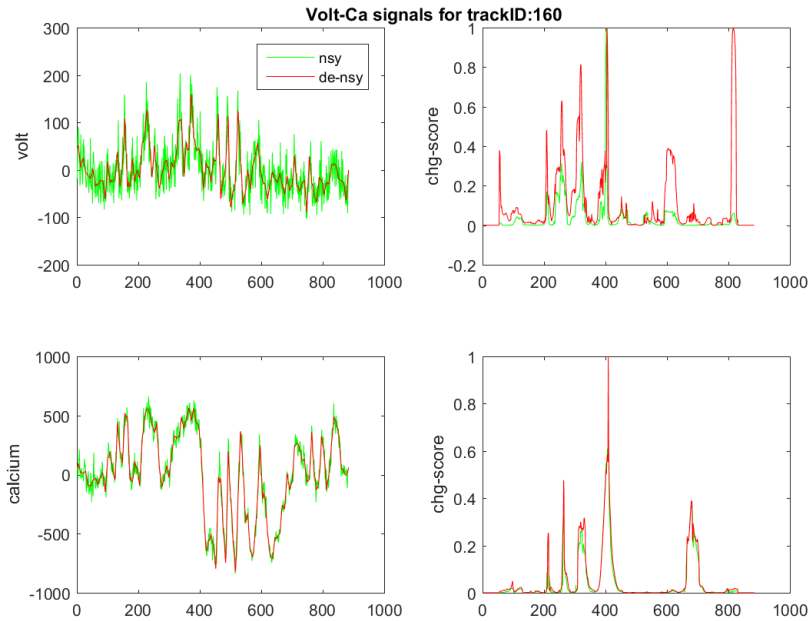
(a) population mean



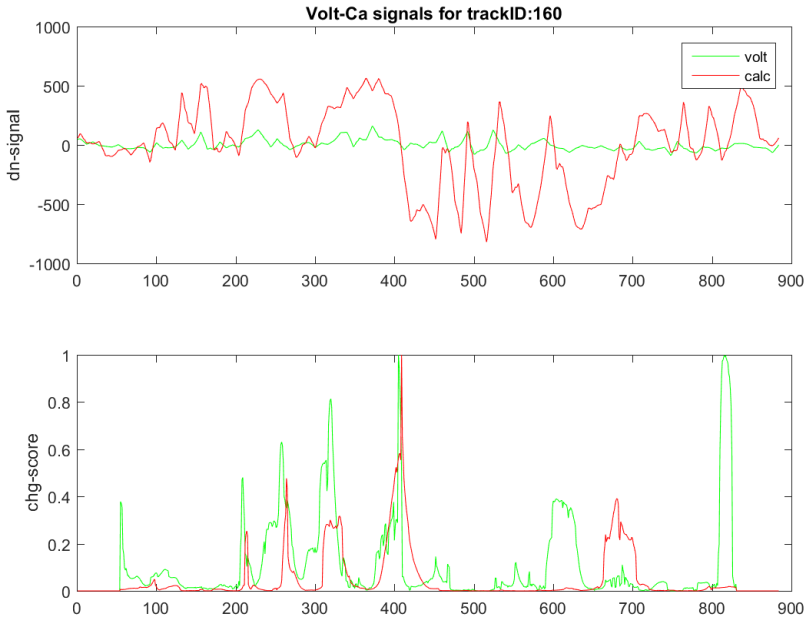
(b) spread about mean

Figure 5.3: Representation of the mean, 5 percentile and 95 percentile bounds on the spread of the signal values at different time instances.

The exploration of motifs in the time-series signals, and also a comparison of the motifs in Calcium signal concurrent with the presence of motifs in the voltage-traces were attempted by using some of the motif-mining algorithms that could be found in the literature. However, the results(Fig.5.6) we obtained were qualitative and did not help us deduce anything interesting about the patterns in the signals or about the cell-behaviour.



(a) voltage



(b) calcium

Figure 5.4: Change-point detection in voltage and calcium signals

Detection of sub-groups of cells behaving similarly

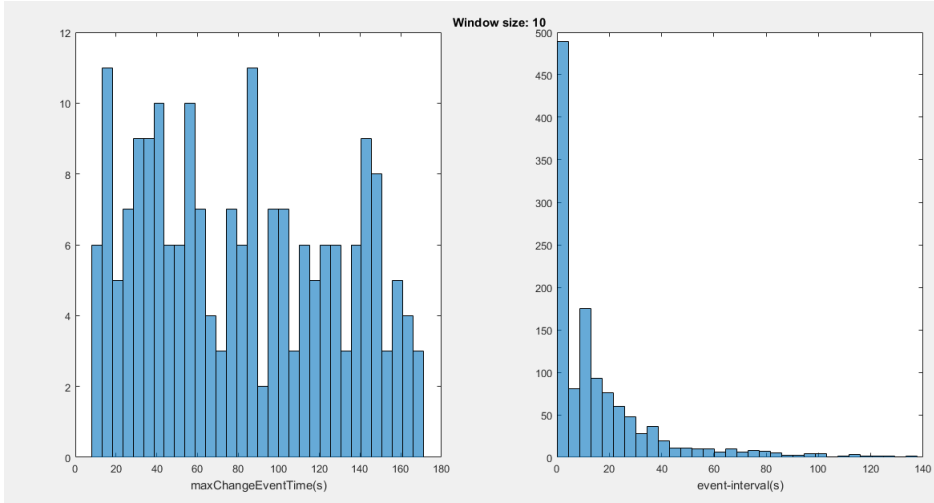
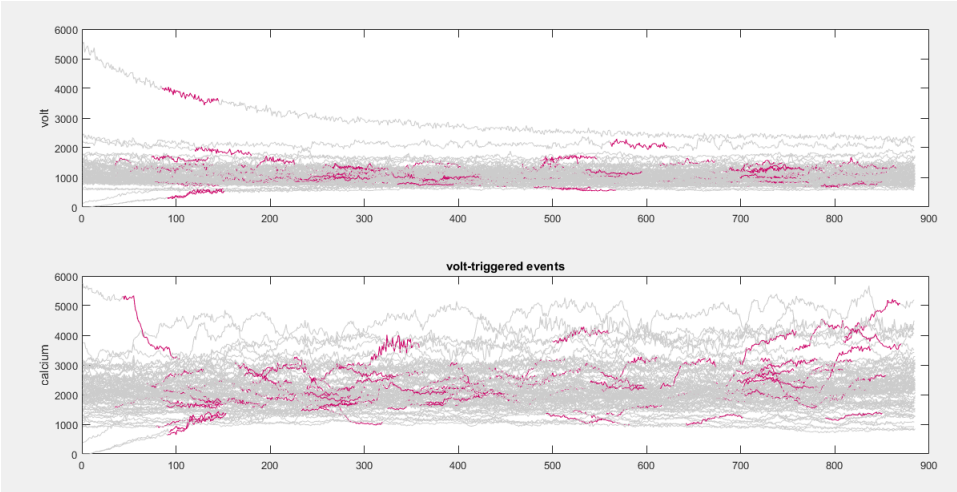


Figure 5.5: Histogram of time of maximum change-score events and the distribution of the time-intervals of consecutive events in a given set of time-traces.

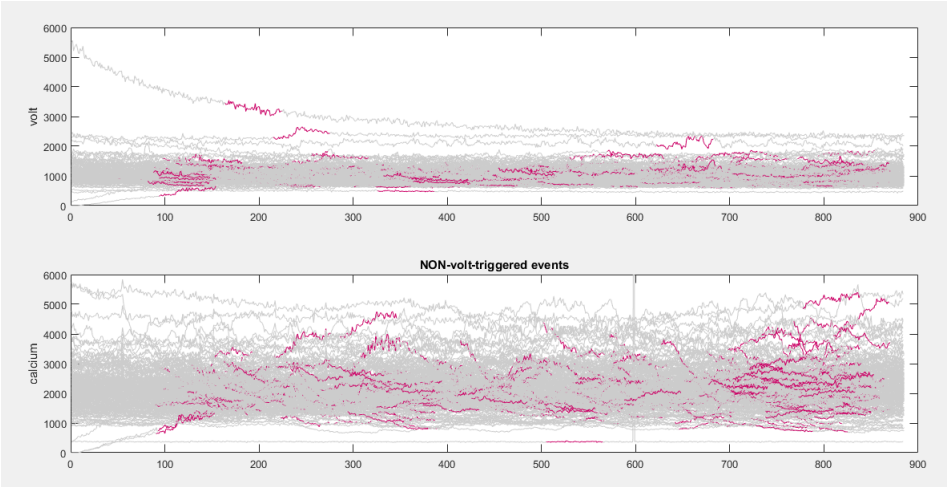
We can consider analysing each trace individually, to detect events and other "patterns", or we may look at all the traces as a multivariate time-series.

We adopted the following methodology of probing into the dataset. First, we assume that there is no time-lag between the trace-signals from the bacteria. It might be possible that the source bacteria are communicating with each other. And since, we believe that they may communicate with each other, only via a chemical process, we would ideally expect there to be a delay in the transmission of signals from one cell to another. However, we ignore these aspects for the initial tests on the data, and hope to see if there's any patterns that's worth exploring.

We compute the Correlation Matrix representing the cosine-distances between the traces by computing the inner-product of the normalized traces. A value of 1 indicates identical signals and -1 indicates opposing signals. To identify the population of cells that might be communicating with each other(almost in synch with each other) we choose a particular threshold value(θ) and extract an Adjacency Matrix, representing the connectedness between correlated cells. We then generate the correlation-graph, $G(V, E)$ representing with V being the set of cells and E being the set of edges $\{(i, j) \in E | CorrelationMatrix(i, j) \geq \theta\}$. A plot of the graph obtained, helps us identify the existence of a connected-components in the dataset. We extract the connected



(a) voltage-triggered motifs



(b) non-voltage triggered motifs

Figure 5.6: Motif detection in the calcium signals based on motif-mining algorithms

components from the graph by pursuing a depth-first search. Re-indexing the cell indices by using the ordering obtained from a depth-first search, then helps us obtain a Correlation Matrix, in which the correlated components become visually self-evident. The cluster membership as determined by the depth-first search, can then be used to guide any other visualization scheme that we might wish to do, using the given dataset.

Note that, at a threshold of 0.69 we obtain two large clusters, that are isolated from all other clusters. However, lower thresholds, have more inter-connectedness, and creates a Large-Connected-Component that links up a larger fraction of the population.

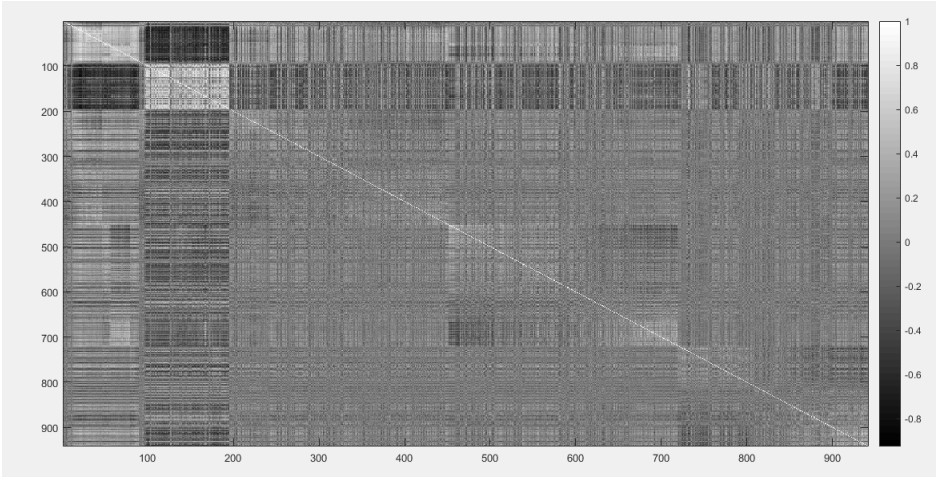


Figure 5.7: Determining connected components in correlation-matrix.

Note that the cluster-membership histogram helps identify the distribution of the cells into components connected by high correlation index.

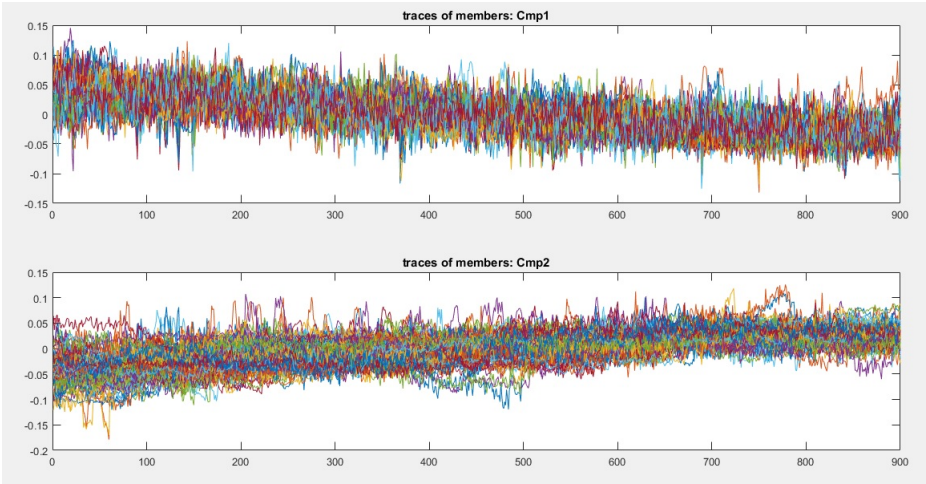
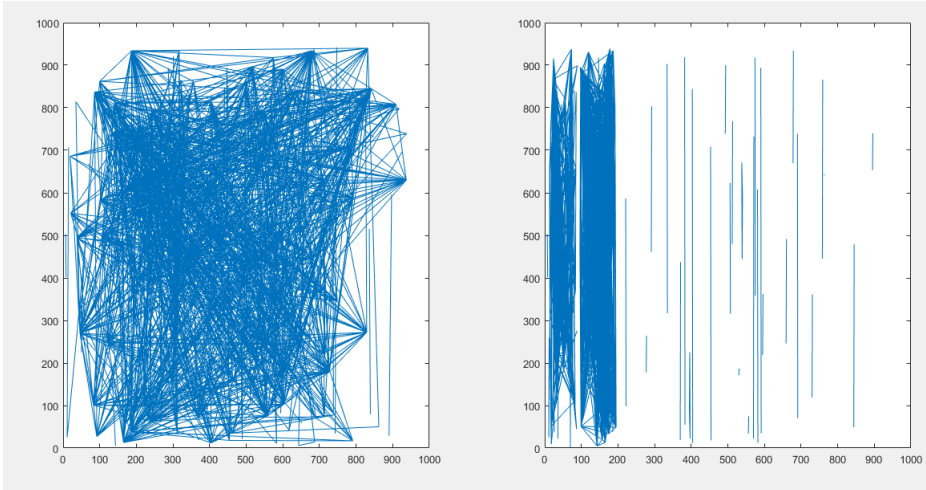


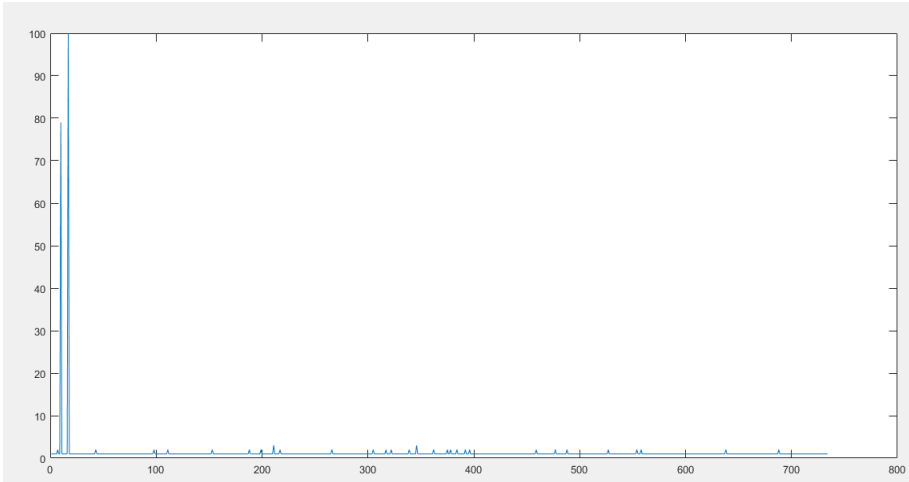
Figure 5.8: Traces of cells in different components

The presence of connected components with members that are synchronized between themselves, but anti-symmetric with the members of the other component, is an extremely surprising observation in the dataset, given that we have assumed an absence of time-lags in the communication between the cells.

This observation raises the following questions:



(a) graphical components



(b) correlation matrix

Figure 5.9: Determining connected components in correlation-graph using threshold= 0.69.

(1) The high correlation value between the cells, indicates the presence of Synchronization between cells. Can we determine whether these cells were spatially located in each other's vicinity, or are there cells that are far away, but synchronized with each other?

(2) Is it possible to determine the physical phenomena, by which synchronization at a distance might be realized?

(3) Can we characterize the behaviour of the cells in each component?

Fitting ARIMA(1,1,1) models to the signals

The suggestion of the presence of a linear model fit to the signal's present value based on its past values, indicates that its possible to fit a model to describe the data. We start with trying to identify a linear model, where the present value of the measured variable depends on its immediate past value and the value of its velocity. By determining the parameters that define the planes in the plots(Fig.5.10), we obtain the linear-model coefficients that describe the system.

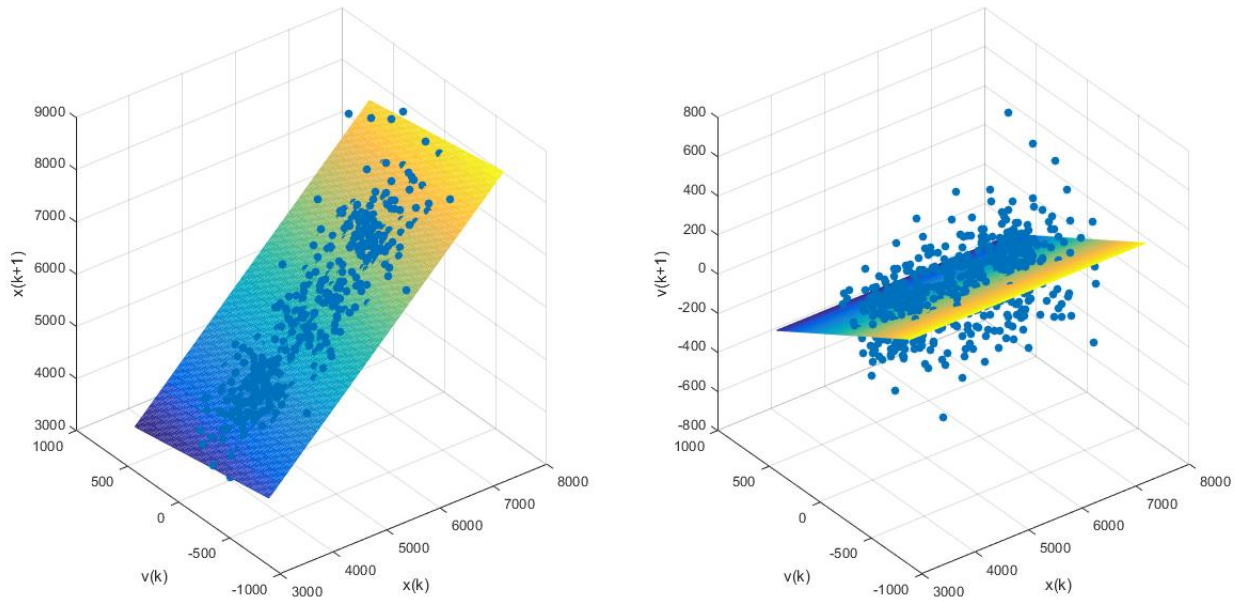


Figure 5.10: Linear model describing the relationship of the signal with its past values.

Further, by using the variogram (Fig.5.11), we determine that the original signal-traces are non-stationary. We may obtain stationary signals from the same, by extracting the first-difference and second-difference signals. The stationary property of the signals is ascertained by the fact that their variogram plateaus out at a specific value, and also from observing the property of the auto-correlation function(ac.f) and the partial auto-correlation function(pac.f) of the signals. We note that the the ac.f and the pac.f helps us determine the order of the ARMA models that can be used to describe the first-difference signal.

The appropriate model choice of ARIMA(1,1,1) is determined to fit(Fig.5.12) the signal trace.

This is verified by determining if the distribution of the residuals computed after the model-fit, is normally distributed and uncorrelated with itself.

5.0.8 Experiment 1: The study of the signals from different species

5.0.8.1 Design of experiment

In this experiment, the above method of preparation was followed for a different species of cells. Salmonella, is a bacterial species, that is rod-shaped gram-negative of the Enterobacteriaceae family. Strains of this bacterial lineage is known to cause 'food poisoning' in humans. It's a constituent of the intestinal biota, similar to the E.Coli strains. However, the challenge here is to determine features in the measured signals to help differentiate between the two cell-species which otherwise cannot be separated visually.

5.0.8.2 Analysis of measurements

Time-Domain Features We adopt both time-domain and frequency-domain methods for the purpose of analysis of the measurements. In the time-domain, we attempt to fit an ARIMA(1,1,1) model to each of the time-series and try to extract the coefficients of the best-fit model. These features are then used as a signature of the time-series of each cell from the different species. These parameters are then visualized in a three-dimensional space, to determine if the cells belonging to different species form visually separable clusters.

From the above visualization(5.13), we observe that the model-parameters do form clusters rather than being uniformly spread. However, for the given set of features, the data aren't inseparable.

Frequency-Domain Features In the frequency domain, we extract the spectrogram of the subsequences of the signals, and model the features using a code-word histogram approach. The code-word histogram based vector representation of the frequency-domain information contained in the signals, is then subjected to non-linear dimensional reduction via the t-stochastic neighbourhood

embedding algorithm(tSNE [**?**]) to obtain a three-dimensional visualization of the cell-data. The visualization thus obtained has a distributional property of its points, that similar to the distribution of the feature-vectors in the original high-dimensional space.

From the plots indicated (Fig.5.14), we make the following observations:

(1) For cells from a given species, we observe that the high-dimensional feature vectors when reduced to the three-dimensional space, seem to occupy distinctive clusters which form a unique macro-pattern. This is interesting because, it is indicative of intra-species cell-differentiation. It might suggest that there are four-to-five different modes of functioning as observed in the different cell-behaviours. It would be interesting to determine, what particular characteristics each of the cells belonging to a specific cluster actually exhibit.

(2) The second observation that we make, is that the datasets of the two species, when combined, seem to overlap. The different clusters thus observed might not just be indicative of intra-species differentiation, but would rather be indicative of different modes of behaviour that's common to all cells from the Enterobacteriaceae family.

5.0.9 Experiment 2: The study of the effect of a change in population density on cell-signal behaviour

5.0.9.1 Design of experiment

A given batch of cells, which were prepared to express protein sensors of a previously determined signal-type, is diluted to different levels, before being placed upon a nutritive petridish for recording observations. Three levels of dilution are used, corresponding to the different levels - 'LOW', 'MEDIUM' and 'HIGH' densities.

5.0.9.2 Analysis of measurements

For the univariate signal traces obtained from cell-samples prepared at different population densities, we decided to use a simple metric to represent each cellular trace. This metric was called

the 'Total Variation'([1]) of the signal and was computed as follows - if z_t was the signal sequence measured over time, the total variation is given by

$$TV(z_t) := \sum_{k=2}^{N} (z_k - z_{k-1})^2 \tag{5.1}$$

Thus, its a measure of the square of the distance of the vector, obtained by a first-difference of the original time-series. The first-difference of the observed time-series is computed to remove the effect of any trends in the signal, and to extract the variations that happens at each time-step. Thus, it is closely related to the use of the ARIMA model that was earlier used to model the time-series obtained from the different species.

From the above figure(Fig.5.15), we observe that the normalized histogram of the total-variation computed for signal-traces from different datasets of same-density data are very similar. However, when the histogram for each density type are contrasted with each other(Fig.5.16), we observe that the mode of the distribution for the higher-density dataset is larger than the modes for the medium and low density dataset. From this observation, we can infer that the cells have an ability to detect the presence of others around it, and this sense-perception leads to increased agitation or signal variability in the measured voltage and calcium signals. The fact that the similar observation is noted in the calcium signals, which has a higher signal-to-noise ratio, than the voltage signals, helps ascertain this inference.

Further, from the histogram plots of the total-variation computed from the signal traces, it is possible to identify the cell-indices that have their total-variation greater than specified percentile thresholds. The probability of total-variation values lying beyond a particular threshold can be determined by identifying the cell-indices of the cells satisfying the threshold criteria. We may also compute the conditional probability of having a certain subset of cells, lying within the threshold limits for the calcium signal total-variation, given the information about the cells satisfying the thresholds for the voltage-signal total-variation. From the below plot(Fig.5.17) we note that the conditional probability of the occurrence of the event that a cell satisfies the threshold criteria for the calcium-signal, given that it satisfies the criteria for the voltage-signal is higher than if the two

events were independent. This observation, helps establish the fact that the two signals are not independent of each other.

5.0.10 Experiment 3: The study of the effect of toxins on the cell-behaviour

5.0.10.1 Design of experiment

In this experiment, we prepare the same batch of cells on multiple petri-dishes and take recordings from the specimens in the following stages. The healthy stage, where the cells are thriving in a nutritious medium, is called the 'BEFORE' stage. The measurements are taken to determine the ground-state behaviour of the cells. In the next stage, while the measurements are being taken, we introduce certain toxins onto the cell-culture without blocking the view of the microscope. The measurements recorded in this stage are considered to be the 'DURING' stage. After the required duration of measurements are obtained, we wait for the cells to settle down in their response to the chemical stress introduced to the environment. The measurements recorded in this state is considered to be the 'AFTER' stage.

5.0.10.2 Analysis of measurements

We computed the total-variation(5.1) for the cell-signals from the different sets of measurements made for each of the three stages - BEFORE, DURING and AFTER, corresponding to cell batches treated to a particular toxin type. The histograms were normalized and compared to determine changes in the probability distribution of the above measure of total-variation in the different stages.

From the plots above (Fig.5.19, 5.18), it was identified that mode of the distributions shifted to the left, as we transition from stage 'BEFORE' to the 'AFTER' stage. This is suggestive of the possibility that the presence of toxins inhibits cellular activity, as indicated by the total-variation of the time-series.

Since, the total-variation is a time-domain measure, it seems obvious that comparable differences should be observed in the time-series model-coefficients that would be extracted when fitting an ARIMA(1,1,1) model to the signals. The plots of the coefficients of the ARIMA(1,1,1) model and the linear-model(Fig.5.20) that is learned from the observations, also helps to visually identify different clusters corresponding to different behaviours in the two stages.

5.0.11 Discussion of results

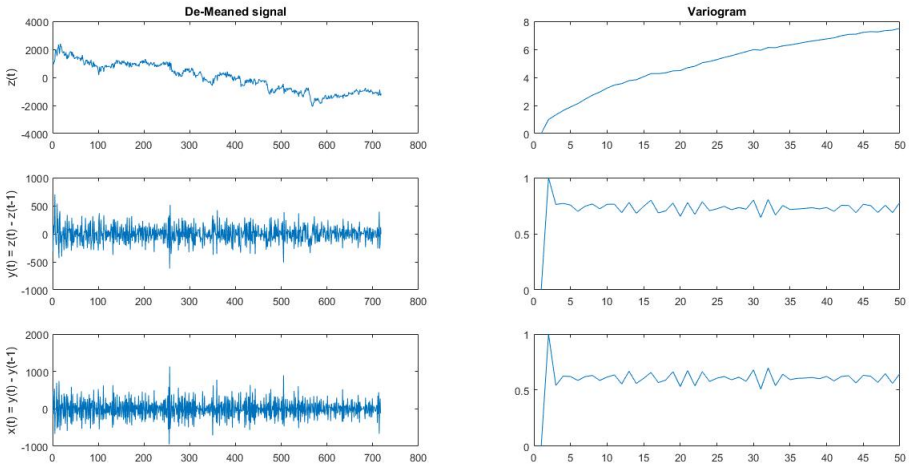
From the description of the patterns observed in the data corresponding to the above experiments, we make the following inferences

(1) The voltage and calcium signals are not independent of each other.

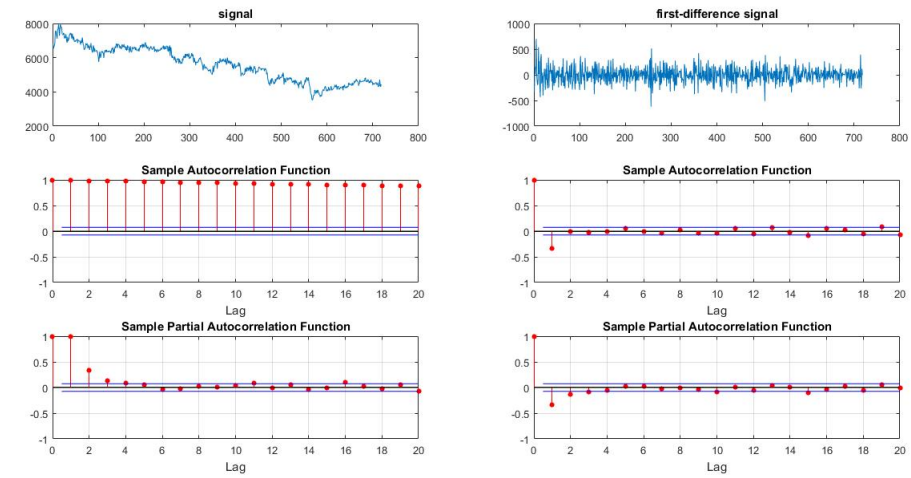
(2) It is possible to fit an ARIMA(1,1,1) model to the observed time-series data. Why this works, is still uncertain. However, this suggests that it would be an interesting research direction, to identify other methods to model the dynamical system from which these measurements are made, rather than just fit models to the observations obtained. Determining the complexity of the dynamical system represented by the collection of cells, shall help us better understand the macroscopic behaviour of the system components that relate the voltage to the calcium signals, which might guide the identification of the structural components participating in the process.

(3) It is observed that the constriction of cellular metabolic activity, by the presence of toxins, can be detected by the computation of total-variation in the voltage and calcium signals that are measured.

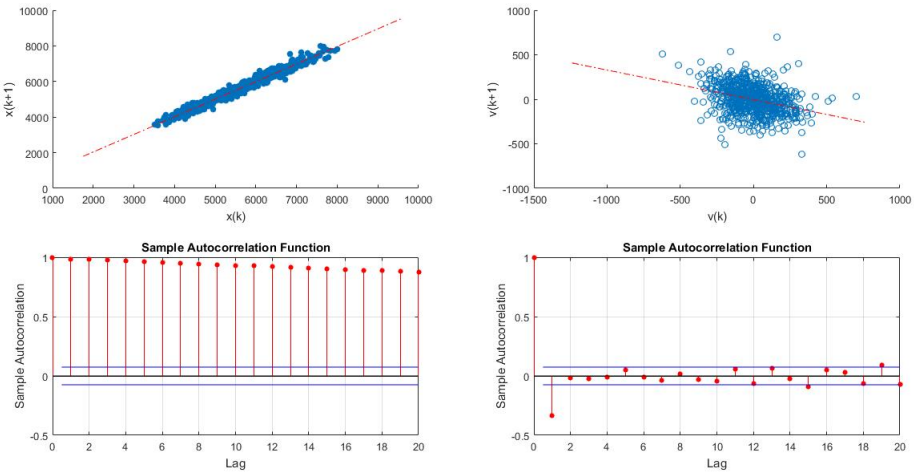
(4) By fitting linear models, to the voltage-calcium input-output system represented by each individual cell, we have detected clusters in the coefficient space, which suggests that there is intra-species differentiation occurring in the biological population. We need to determine biological experiments to verify the precise ways in which this differentiation occurs.



(a) variogram

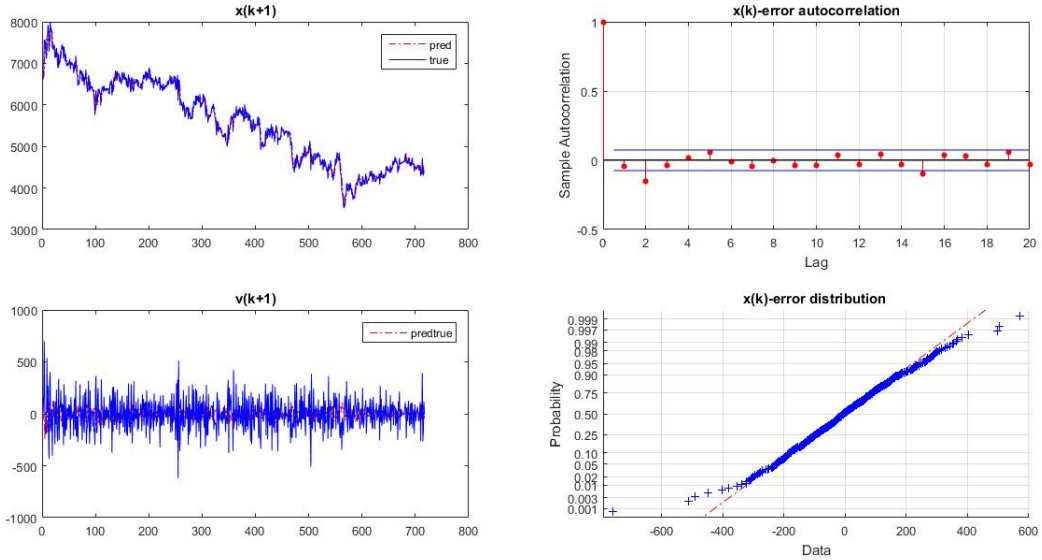


(b) first-difference

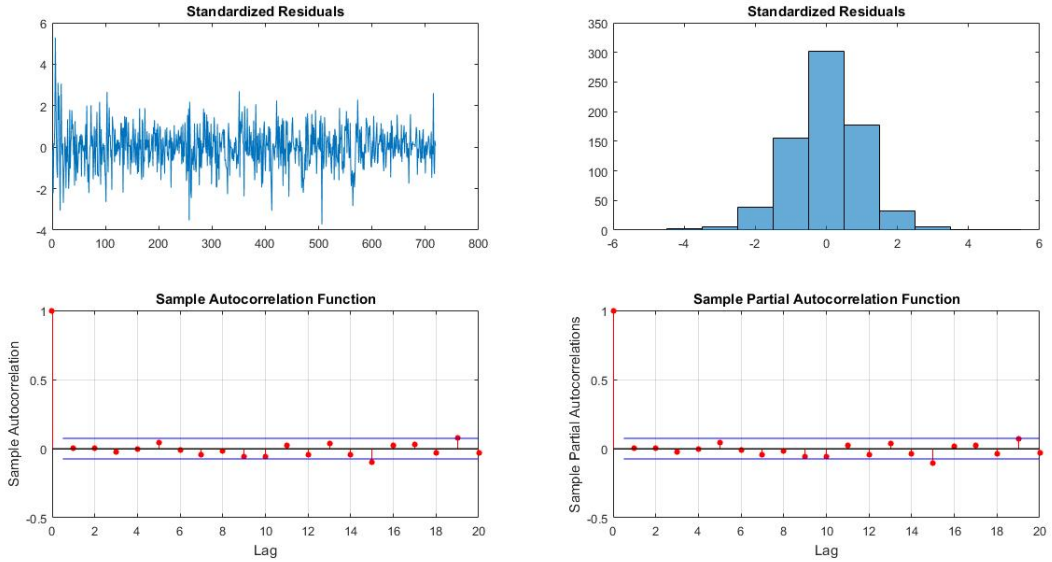


(c) linear relationship

Figure 5.11: Determination of model parameters for the signal traces



(a) model-fit



(b) residual histogram

Figure 5.12: Determination of properties satisfied by the residuals obtained after the ARIMA(1,1,1) model-fit to the signals.

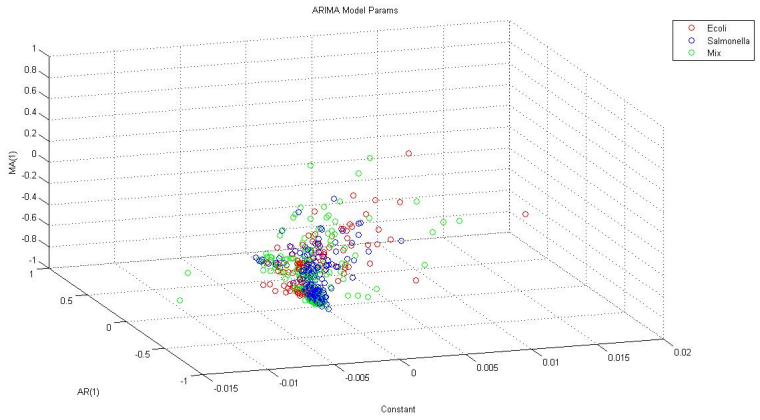
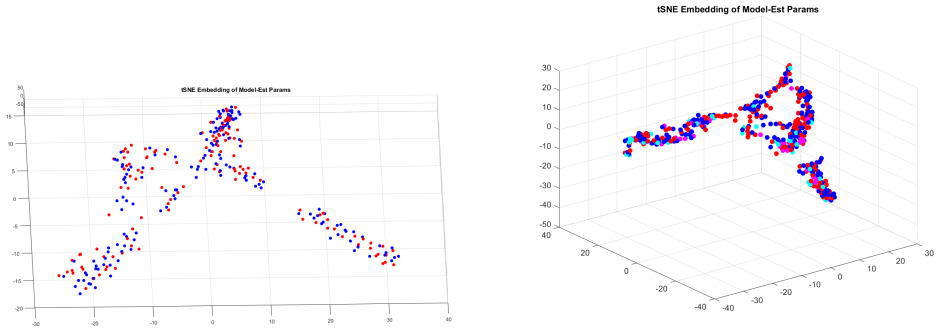
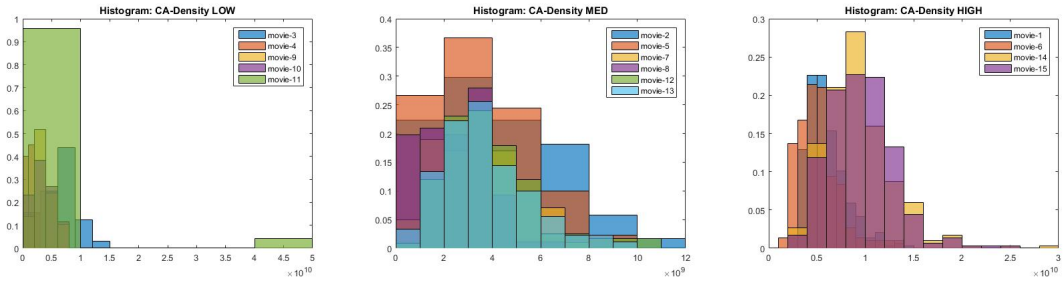


Figure 5.13: Visualization of the ARIMA(1,1,1) coefficients for the experimental data containing single-species and mixture of species datasets



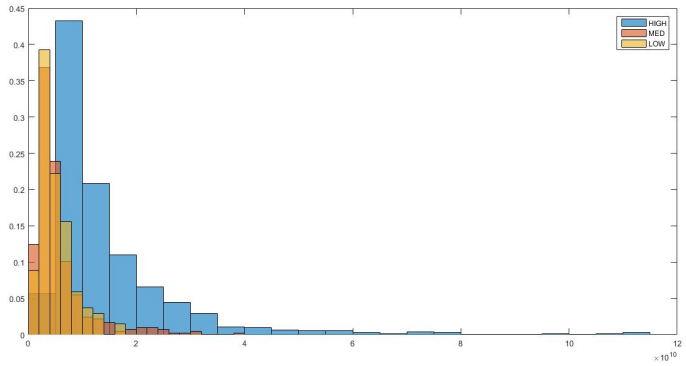
(a) tsne-species-one-day (b) tsne-species-two-day

Figure 5.14: Three dimensional projection of the code-word histogram representation of the frequency domain based feature-vectors for the two species. (a)Dataset for one experiment, with Ecoli(Red) and Salmonella(Blue). (b) Dataset from two experiments, with Ecoli(shades of Red) and Salmonella(shades of Blue).

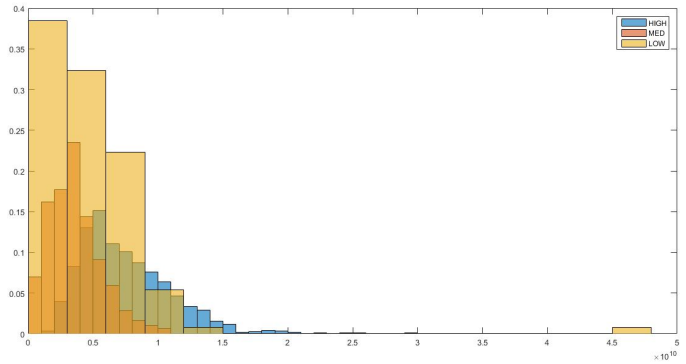


(a) LOW-density (b) MEDIUM-density (c) HIGH-density

Figure 5.15: Normalized histogram of total-variation computed for calcium-traces from datasets corresponding to different densities. (d)The traces from all videos corresponding to a given density-type are collected together to form a single histogram for that type.



(a) voltage



(b) calcium

Figure 5.16: Normalized histogram of total-variation computed for (a)voltage-traces and (b) calcium-traces from datasets corresponding to different densities.The traces from all videos corresponding to a given density-type are collected together to form a single histogram for that type.

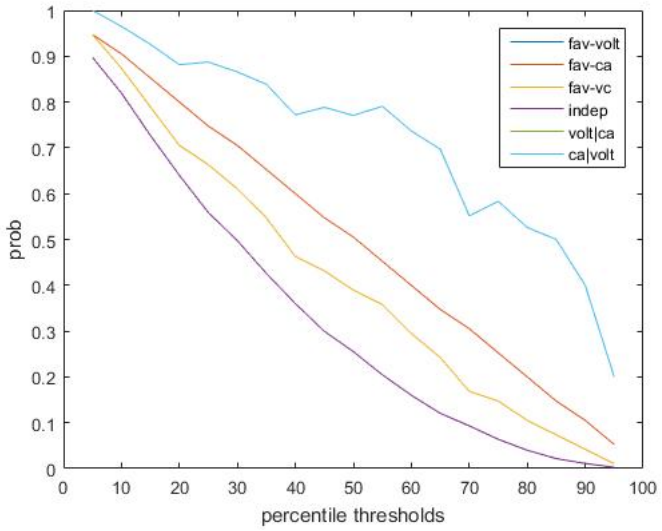
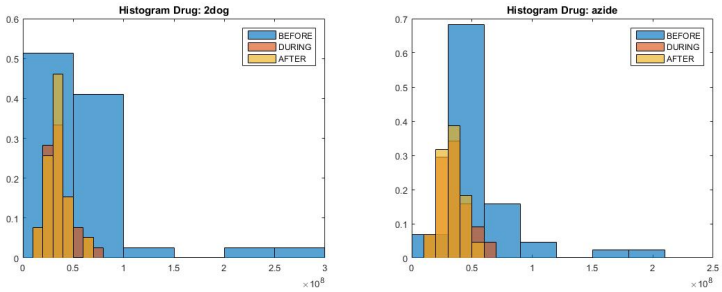
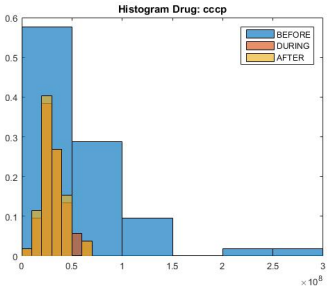


Figure 5.17: Computation of the probability of the events that the total-variation of the cell-signals satisfy percentile thresholds.

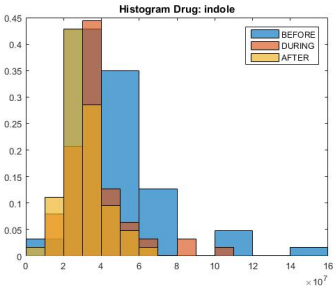


(a) 2DOG

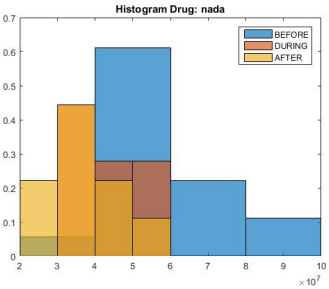
(b) Sodium Azide



(c) CCCP

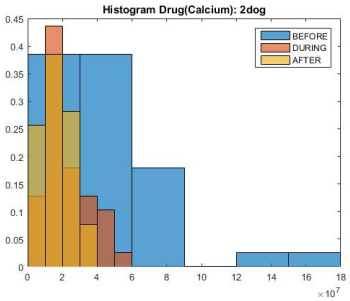


(d) Indole

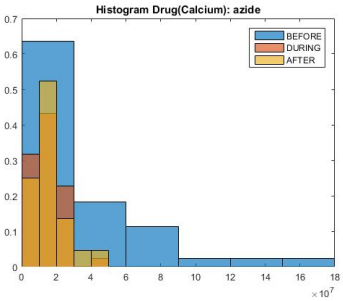


(e) NADA

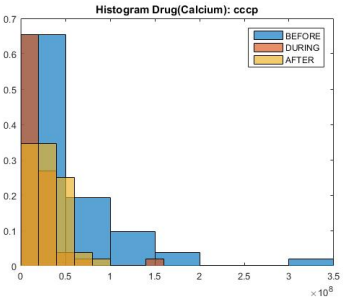
Figure 5.18: Normalized histogram of total-variation computed for voltage signal datasets of BEFORE, DURING and AFTER stages of introduction of different toxins upon the prepared cell-sample.



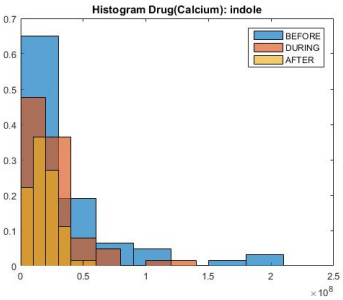
(a) 2DOG



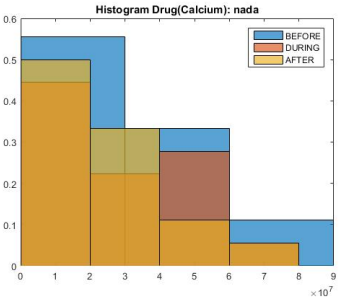
(b) Sodium Azide



(c) CCCP



(d) Indole



(e) NADA

Figure 5.19: Normalized histogram of total-variation computed for calcium signal datasets of BEFORE, DURING and AFTER stages of introduction of different toxins upon the prepared cell-sample.

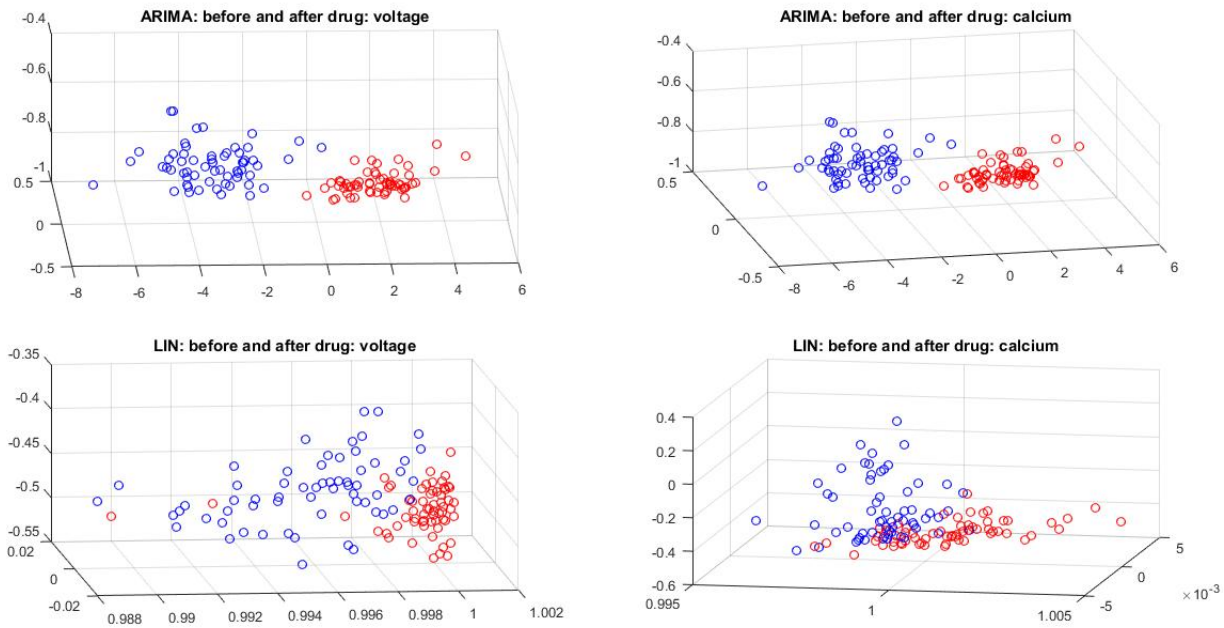


Figure 5.20: Plots of coefficients obtained from the ARIMA and linear modelling of the voltage and calcium signals in the two stages - BEFORE(blue) and AFTER(red).

Bibliography

[1] *Alvarez-Esteban, Pedro C., C. Eun, and J. Ortega.* "Time Series Clustering using the Total Variation Distance with Applications in Oceanography." arXiv preprint arXiv:1501.04050 (2015).

[2] *Ailon, Nir, and Bernard Chazelle.* "Faster dimension reduction." Communications of the ACM 53.2 (2010): 97-104.

[3] *Bisgaard, Sren, and Murat Kulahci.* Time series analysis and forecasting by example. John Wiley & Sons, 2011.

[4] *Box, George EP, et al.* Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[5] *Chatfield, Chris.* The analysis of time series: an introduction. CRC press, 2016.

[6] *Dale, Mark RT, and Marie-Jose Fortin.* Spatial analysis: a guide for ecologists. Cambridge University Press, 2014.

[7] *Davis, Jason V., et al.* "Information-theoretic metric learning." Proceedings of the 24th international conference on Machine learning. ACM, 2007.

[8] *Donner, Reik V., and Susana M. Barbosa.* "Nonlinear time series analysis in the geosciences." Lecture Notes in Earth Sciences 112 (2008).

[9] *Hanski, Ilkka, and Ian P. Woiwod.* "Spatial synchrony in the dynamics of moth and aphid populations." Journal of Animal Ecology (1993): 656-668.

[10] *Hein, Matthias, and Jean-Yves Audibert.* "Intrinsic dimensionality estimation of submanifolds in R d." Proceedings of the 22nd international conference on Machine learning. ACM, 2005.

[11] *Honerkamp, Josef.* Stochastic dynamical systems: concepts, numerical methods, data analysis. John Wiley & Sons, 1993.

[12] *Jansen, Aren, and Partha Niyogi.* "Intrinsic spectral analysis." IEEE transactions on signal processing 61.7 (2013): 1698-1710.

[13] *Kanamori, Takafumi, Taiji Suzuki, and Masashi Sugiyama.* "Statistical analysis of kernel-based least-squares density-ratio estimation." Machine Learning 86.3 (2012): 335-367.

[14] *Kantz, Holger, and Thomas Schreiber.* Nonlinear time series analysis. Vol. 7. Cambridge university press, 2004.

[15] *Kawahara, Yoshinobu, and Masashi Sugiyama.* "Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation." SDM. Vol. 9. 2009.

[16] *Kralj, Joel M., et al.* "Electrical spiking in Escherichia coli probed with a fluorescent voltage-indicating protein." Science 333.6040 (2011): 345-348.

[17] *Kralj, Joel M., et al.* "Optical recording of action potentials in mammalian neurons using a microbial rhodopsin." Nature methods 9.1 (2012): 90-95.

[18] *Lin, Jessica, et al.* "Experiencing SAX: a novel symbolic representation of time series." Data Mining and knowledge discovery 15.2 (2007): 107-144.

[19] *Maaten, Laurens van der, and Geoffrey Hinton.* "Visualizing data using t-SNE." Journal of Machine Learning Research 9.Nov (2008): 2579-2605.

[20] *Marwan, N., M. Thiel, and N. R. Nowaczyk.* "Cross recurrence plot based synchronization of time series." arXiv preprint physics/0201062 (2002).

[21] *Marwan, Norbert, ed.* Recurrence Quantification Analysis: Theory and Best Practices. Springer, 2015.

[22] *McFee, Brian, Luke Barrington, and Gert Lanckriet.* "Learning content similarity for music recommendation." IEEE Transactions on Audio, Speech, and Language Processing 20.8 (2012): 2207-2218.

[23] *Mueen, Abdullah, and Nikan Chavoshi.* "Enumeration of time series motifs of all lengths." Knowledge and Information Systems 45.1 (2015): 105-132.

[24] *Sugiyama, Masashi, et al.* "Direct importance estimation with model selection and its application to covariate shift adaptation." Advances in neural information processing systems. 2008.

[25] *Shieh, Jin, and Eamonn Keogh.* "i SAX: indexing and mining terabyte sized time series." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

[26] *Thiel, Marco.* "Recurrences: exploiting naturally occurring analogues." (2004).

[27] *Webster, Richard, and Margaret A. Oliver.* Geostatistics for environmental scientists. John Wiley & Sons, 2007.

[28] *Zhao, Yongxin, et al.* "An expanded palette of genetically encoded Ca2+ indicators." Science 333.6051 (2011): 1888-1891.

Appendix A

Motivation to the choice of the research problem

As an end note, I would like to elaborate on my educational background and the reason why I chose to work in this field for my Master's thesis. Science and Mathematics has been my favourite subjects ever since my high-school years. On reflecting about the years that have gone by, I've come to realize that the reason I loved these subjects is because they are reliable fields of study that can lead us to the truth. There's a sense of certainty, a security and a sense of belonging to the world, that comes from believing that a 'truth' exists, and that we have the capacity in us, as human beings to take steps to realize the truth. However, what's interesting is that my notion of truth has changed a lot over the years.

During my early-schooling, 'truth' referred to facts - descriptive facts and observational trivia about different aspects of the world. This was the easiest to comprehend and the easiest to indulge oneself in. One could come across the facts about the earth, about the physical processes, the living systems, by just curiously exploring an encyclopaedia. And at every stage, when you encountered a new object, or a new field of study - you realize that there's a whole different vocabulary, sets of words you need to learn. There are words in the language, that have been created to help describe, and classify the objects and processes that one shall come across in each field - be it geography, physics, chemistry or biology. The more you know about the objects('nouns') and the processes('verbs'), the more you can describe the static and dynamical aspects of that world.

During my undergraduate education in Electrical Engineering, I was introduced to different fields of study - Computer Programming, Digital Design, Power Electronics, Electrical Machines,

Signals and Systems, Digital Signal Processing, Control and Optimization, Data-Structures and Algorithms. My mathematical training was limited to Complex Analysis and Advanced Calculus. These fields of study helped me see a whole range of disciplines that existed, in which one could choose to major for one's research. At that time, I loved the idea of algorithms - the concept of solving an optimization problem by using techniques like genetic algorithms, swarm intelligence, etc that were ideas inspired by processes observed in nature. It felt amazing to realize that we could solve mathematical problems, by mimicking nature. My final year project, on the optimization of maximum power-point tracking of solar cells, essentially dealt with formulating an optimization problem with regards to a panel of solar-cells, and to determine the optimal operating point of voltage-current relationship, to ensure that we extract the maximum power from the device. The assumption here was that each power-generating device, which had its own physical characteristics determined by the physics of the materials and the interfaces connecting the materials that constituted the solar cell, had a specific power-generating characteristic curve, which looked like a concave function. The physical problem, could then be solved by building a mathematical model of the solar panel, extracting a mathematical model of the power-characteristic curve, determine the optimal operating point by solving the optimization problem, and then determining physical means by which the device could be forced to operate at the optimal-power point.

After graduation, I worked as a software developer and software test-analyst at an investment banking technology firm. I got exposed to the systems and processes within a corporate world, while also realizing that there were huge technological systems that were being used in the industry and world-wide, which formed the back-bone of the global economy. I was particularly involved with the development process of a Securities Core Processing platform, which was essentially a software-engine to handle the processing of the purchase and selling of securities in the stock-exchange markets. The software-engine was extremely large in scale, in terms of the number of transactions being processed, the number of servers and sub-systems involved in its architecture, and the number of humans working to design, develop and operate the system. In particular, I realized that though I was working on a system, I did not have the knowledge to be able to describe

and analyse the performance of these systems. I felt certain, that there were mathematical tools and technologies that exist in academia, that would enable one to describe the ways in which these machines behaved. And I was especially intrigued by the fact that, these systems though running on computers and hardware infrastructure built on the physical world, were essentially information systems - where it was the flow of bits and messages that constituted the dynamics. There were logical rules which governed the ways in which the information was transformed and stored, but these rules were designed and enforced by humans. These were not constraints or rules that were defined by nature. My realization of my own ignorance and inability to describe and explain these systems in a formal mathematical way, made me take a step towards educating myself in the topics of computer science and also, apply to graduate school.

I first got into a graduate program in Control and Instrumentation, which helped to introduce me to the topics of non-linear dynamical systems analysis and control, and the topics in computer science related to machine learning and reinforcement learning. My masters thesis was on an exploration of mathematical-techniques for developing a state-estimator(observer) for the Attitude Estimation and Control subsystem of a student-designed nano-satellite. The graduate school introduced me to a little more mathematical theory than what I had been exposed to during my early years. In essence, my education trained me to model, build and analyse engineering systems - devices that were being built to meet a particular purpose. However, I was not exposed to the area of probability theory until I graduated and worked as a researcher in a lab that was trying to understand methods to solve the urban traffic problem. Ever since I learned about probability and uncertainty, I realized that I wanted to study this further, atleast to an extent that I am familiar with the vocabulary, and the tools that are used so that I may explore and study it at my own leisure. I knew that this sort of an education can be received only within a University system, the outside world of industry cannot teach you new ways of thinking and describing the world, unless you find yourself in those particular situations. Thus, my initial graduate program while being concentrated towards implementation and simulation studies, rather than mathematical analysis, it exposed me to the mathematical topics that helped model uncertainty and learning in systems.

My intention during my graduate studies at CU, thus became concentrated on getting exposed to a lot more concepts in Applied probability and numerical methods. Linear Algebra, Markov Processes, Stochastic Processes, and Convex Optimization became the mathematical topics that I got introduced to over these years. Further, as an applied math project, I explored the subject of search engine and high-dimensional datasets, which helped me realize how abstract mathematical ideas were being used to classify information or structured data. In essence, all the course-work that I had done till now, made me realize the following - the fields of control theory and optimization, assume that we know the mathematical model of the system we wish to study. And once that is know, it provides a set of tools and techniques to analyse the behaviour of that model, and to help take decisions/control actions that determine how the system operates to the specifications we wish to ask of it. And further, most of the devices I was introduced to were engineered systems - human specified devices, of which we required a particular level of performance. It was good to develop a good mathematical maturity, to help you unravel truths about the mathematical models - to derive their limitations, or the constraints within which certain solutions could be obtained. My course in Game Theory, helped me realize that there's a mathematical tool-set that can be used to analyse and discuss the ways in which the complex system formed by the interaction of multiple-agents with their own autonomy of action choices could be analysed, and designed. This is one area that I could explore. The only limitation being the extent to which one has control over how much the knowledge gained through the analysis can actually be implemented in the world. The socio-economic, human-machine ecosystem is an extremely large and complex system, in which design and policy decisions are implemented at a scale of states, countries or corporations. It was science, it was a pursuit of truth - and I liked it for its beauty. However, based on my learning style, I realized that I would prefer to engage with systems at a more physical level, and understand them over multiple levels of abstraction.

When I came across, Prof. Kralj's research into bacterial electro-physiology, I was intrigued. It was an exercise of seeking the truth, about the biological world - about exploring the structures and processes that governed the functioning of the most primitive forms of life, from which we

could possibly gain deeper insight into the functioning of ourselves - the tissues and organs that constitute the body that we have. The challenge placed to me was as follows: We have signal recording from inside bacterial cells. In the past, we humans did not have the ability to measure physio-chemical changes occurring inside the cell-bodies that had an extremely small size - these were the most simplest of living forms. Its true, we have a community of physicists working with the smallest scales of matter, studying and analysing those objects using quantum theory and electrodynamic field theory. However, to understand the smallest scales of the biological world, is another interesting domain of research that one can pursue. And yes, there were research teams that had identified ways to make measurements, to look into certain aspects of the functioning of single-celled bacteria. The question then became, is there anything we can learn about these living beings from the measurements we are making.

The challenge here is multi-fold: 1) we do not know whether there is a relationship between the two different signals being measured from each cell. we assume they should be related. there might be biological processes that run inside the cell, that relate the voltage and the calcium dynamics in the cell - given the knowledge of the material scale of functioning of these livings systems; 2) we do not know whether there are ways in which the biological cells are communicating with each other, which influences the signal recordings that we observe; 3) we do not know the direction in which information is transmitted and the mechanisms by which that information is transmitted between the cells; 4) I personally, did not know how to go about handling such a problem. This was my first foray into feeling uncomfortable about my ignorance, about not knowing about tools and techniques that existed that could help me study something and discover something absolute or concrete, to be said about an unknown system. And this was science, as I had dreamed it to be. This was what I wanted to do all throughout my childhood. And finally this problem has made me realize, that my purpose as an engineer - is to be able to design systems and devices, to help make science - to discover truths about life and the living world.

Looking forward, I would like to pursue my PhD in learning about the field of probabilistic models and stochastic dynamical systems to describe. It would be interesting to learn how the

algorithms used to compute and analyse large systems can be programmed to run on distributed computer-architecture systems.