TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 692 • Publication 692

Miika Ahdesmäki

Robust Signal Processing Methods for Genomic Time Series and Protein Accessibility Data

Tampere 2007

Tampereen teknillinen yliopisto. Julkaisu 692
Tampere University of Technology. Publication 692

Miika Ahdesmäki

Robust Signal Processing Methods for Genomic Time Series and Protein Accessibility Data

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB222, at Tampere University of Technology, on the 28th of November 2007, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2007

Custodian: Professor Olli Yli-Harja
Tampere University of Technology
Tampere, Finland

Opponent: Dr. Jaakko Hollmén
Helsinki University of Technology
Espoo, Finland

Reviewers: Dr. Andreas Beyer
Technische Universität Dresden
Dresden, Germany

Professor Samuel Kaski
Helsinki University of Technology
Espoo, Finland

ISBN 978-952-15-1872-0 (printed)
ISBN 978-952-15-1911-6 (PDF)
ISSN 1459-2045

Abstract

The aim of systems biology is to study living beings at the system level. This means that instead of studying just single molecules, we also try to understand the dynamics of larger systems such as biochemical and gene regulatory networks. By entering the genome and proteome wide level we are faced with great opportunities but also challenges.

The introduction of high-throughput measurement technologies for cellular level studies during the last decade has made it necessary to use advanced signal processing methods in computational systems biology and bioinformatics. The gene activity and protein level measurement technologies available today produce huge amounts of data that cannot be processed manually. Thus, advanced computational methods for analysing the data and making conclusions are essential.

The aim of this thesis is to introduce efficient signal processing methods that can be used in making relevant decisions based on systems biological measurement data. The thesis has been divided into three logical parts.

In the first part of the thesis, gene expression microarray measurements are studied. These measurements provide one of the main type of data used in the analyses later in the thesis. A simulation model is then introduced for the generation of microarray data with realistic statistical and biological properties. This data can be used *e.g.* in the generation of ground truth data for simulation studies.

In the second part, time series signals measured from genes with microarrays are studied. Periodicity detection analysis of gene microarray data is especially difficult due to short time series length, the vast number of measured genes and unknown type of noise in the measurements. We introduce different robust methods for both uniformly and nonuniformly sampled time series. The introduced methods are shown to be insensitive to changes in the assumed statistical model for the data and thus improve on robustness if compared to classical methods.

Finally, in the third part we move from genomic data to the actual end products of genes, proteins. A method is presented that can discern locations in the protein sequence that are more prone to pathogenic mutations on average than other locations in the sequence. The data we use is measured from clinical patients and depict the hydropathy of different parts of the

sequence. Changes in the hydropathy of a protein have been shown to relate to structural and functional changes and thus provide an interesting field of study.

Preface

I would first like to thank Professor Olli Yli-Harja and Dr. Harri Lähdesmäki for the support and guidance they have provided me during my research work. As the supervisors of my post-graduate studies, they have given me the possibility to advance in my academic career and provided me the support and encouragement to successfully finish my studies. In addition, I also want to thank all my colleagues in the Computational Systems Biology group for all the insightful discussions and collaboration. I am especially indebted to Dr. Matti Nykter for his invaluable comments, suggestions and remarks regarding my thesis.

This work has been mainly carried out in Institute of Signal Processing (ISP), Tampere University of Technology. The faculty and administration of ISP are gratefully acknowledged. I have also gained invaluable experience while working in Institute for Systems Biology (ISB, Seattle, WA, USA), Spring and Summer 2007. I want to express my sincerest gratitude to Professor Ilya Shmulevich for providing me the chance of working in the international environment at ISB. The financial support of Tampere Graduate School in Information Science and Engineering (TISE), Jenny and Antti Wihuri Foundation, Tekniikan edistämissäätiö and Academy of Finland (Finnish Centre of Excellence program 2006-2011) is gratefully acknowledged.

Finally, I am grateful to my wife Outi, mother Anneli and brother Mikko for all the love and support.

Tampere, November 2007

Miika Ahdesmäki

Markku Ahdesmäki in memoriam 25.08.1951 - 05.03.2006

Contents

List of publications

The contents of this thesis serve as an introduction to the following publications. The publications are referred to as Publication-I, Publication-II and so on in the text.

I Nykter, M., Aho, T., Ahdesmäki, M., Ruusuvuori, P., Lehmussola, A. and Yli-Harja, O. (2006) Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, **7**:349.

II Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H. and Yli-Harja, O. (2005) Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, **6**:117.

III Ahdesmäki, M., Lähdesmäki, H., Gracey, A., Shmulevich I. and Yli-Harja, O. (2007) Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. *BMC Bioinformatics*, **8**:233.

IV Ahdesmäki, M., Lähdesmäki, H. and Yli-Harja, O. (2007) Robust Fisher's test for periodicity detection in noisy biological time series. In *Proceedings of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics (Gensips'07)*, Tuusula, Finland, June 10-12, 2007.

V Ahdesmäki, M., Thusberg, J., Huttunen, H., Vihinen, M. and Yli-Harja, O. (2007) Detection of pathogenic mutation prone locations from protein sequences using solvent accessibility measurements. In *Proceedings of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics (Gensips'07)*, Tuusula, Finland, June 10-12, 2007.

The author's contribution to the publications is as follows.

M. Nykter and M. Ahdesmäki designed and implemented the microarray simulation model in Publication-I. M. Nykter drafted the manuscript. T. Aho was responsible for simulating the biological ground truth data. P. Ruusuvuori and A. Lehmussola performed the image processing experiments. Publication-I has also been included in the thesis of Dr. Nykter.

M. Ahdesmäki and H. Lähdesmäki were equal contributors to Publication-II. M. Ahdesmäki carried out an implementation of the methods, performed most of the extensive simulations and co-drafted the manuscript. H. Lähdesmäki developed the statistical methods, assisted in performing the simulations and mainly drafted the manuscript. Publication-II has also been included in the thesis of Dr. Lähdesmäki.

M. Ahdesmäki mainly drafted the manuscripts of Publication-III and Publication-IV. H. Lähdesmäki co-drafted these manuscripts. M. Ahdesmäki also performed all the simulations and analyses of measurement data, with the exception of A. Gracey performing the gene set enrichment analysis.

M. Ahdesmäki carried out the implementation of the methods, performed the computations and partly drafted the manuscript of Publication-V. J. Thusberg co-drafted the manuscript and performed the biological analyses. H. Huttunen helped in the computations and statistical analyses.

Chapter 1

Introduction

All living organisms are composed of cells. Simple light elements such as carbon, hydrogen, oxygen, nitrogen and phosphorus provide the building blocks for biomolecules such as carbohydrates, amino acids and deoxyribonucleic and ribonucleic acids (DNA, RNA) [22]. Together, these building blocks eventually form cells that are the basic units of life on earth. In all cells, be it the simple one cell prokaryotes or more complex eukaryotes, chromosomal DNA is the component that carries the hereditary information (genome) that eventually leads to protein synthesis and other important functions in the cell.

The ability to find the actual sequences of genes in DNA has been mediated by the different genome sequencing projects. Although all humans have different genomes, the differences between individuals are so small that it has been seen profitable to discover a reference human genome sequence [63, 124]. Although the human genome is huge with approximately 3.2 billion nucleotide base pairs (the bases adenine (A), guanine (G), cytosine (C) and thymine (T)), only some 2 percent of the genome is comprised of genes. The current estimate is that there are approximately 20000 to 25000 protein-coding genes in humans [26]. The rest of the genome consists of so called non-coding regions that are not coding proteins but instead have other possible functions like providing structural stability or regulation of genetic expression. The genomes of other organisms have been studied extensively as well. For example, *Escherichia coli* [13], a species of bacteria (prokaryote) living in the intestines of mammals, *Saccharomyces cerevisiae* [46], the quick growing budding yeast (eukaryote), and *Drosophila melanogaster* [23], the fruit fly, have been studied as model organisms. Based on an assumption that all living beings are descended from a common ancestor, different species also share similar properties. Thus, studying these model organisms is hoped to help also in studying humans (see *e.g.* [8]).

The function of genes is to produce meaningful proteins for the cell. Proteins serve both functionally and structurally in cells and are composed of

chemical compounds called amino acids. There are a total of 20 different amino acids and each amino acid is coded by three nucleotides of the corresponding gene. Proteins compose large sets called proteomes, which are much more complex than genomes, since single genes can give instructions for tens or hundreds of proteins. The function of the genome is to virtually stay the same whereas the proteins coded by the same gene can vary under different conditions. Thus, for example the human proteome project [49] is a much more demanding task than the genome project.

The synthesis of proteins, translation, is initiated in the cell nucleus by the transcription of messenger RNA (mRNA) from the part of the DNA that codes the corresponding gene [20]. The mRNA is then transported out of the nucleus to the cytoplasm (see Figure 1.1). In the cytoplasm the four-letter alphabet of nucleic acids is then translated on ribosomes into the twenty-letter alphabet amino acid sequences, *i.e.* proteins. The ribosomes themselves consist of three large RNA molecules and a collection of proteins. Post-translational processing of proteins includes *e.g.* folding of the protein into correct structure, cleavage of different parts of the protein, chemical modification by for example attachment of new chemical groups and removal of unneccessary parts. The finished protein can have different purposes; some act as structural components and others have functional purposes like signalling. Proteins can also initiate or inhibit the expression of genes. A special class is self-regulating proteins that activate their own transcription so that once the gene has been turned on it is expressed continuously. These changes in the protein sructures and regulation loops cannot be accounted for by the knowledge of the genome sequence alone and thus systematic more wider approaches are needed to understand the functional properties of the genome and the cell as a system.

As pointed out in [59], systems biology is a new field in biology that aims at systems theoretic analysis of biological systems. Further, to understand biological systems in their entirety, it is necessary to first investigate the structure, *e.g.* genes, metabolism, physical structures etc., of such systems. The next step is to understand the dynamics of the systems and how to control them. Finally, the goal is to be able to both design new and modify existing systems for desired properties. The approach of systems biology is to cyclically carry out scientific experiments and perform computational simulations and modelling so that the experiments are analysed computationally and the simulations are validated by further experiments.

According to [60], computational biology has two distinct branches. The first branch is data mining and includes the extraction of hidden patterns from experimental data. Applications of this approach include *e.g.* the inference of gene regulatory networks from gene expression profiles and detection of cell cycle regulated genes based on microarray time series data [115]. The main part of this thesis is devoted to developing novel analysis methods for these types of studies. The second branch is termed simulation-based anal-

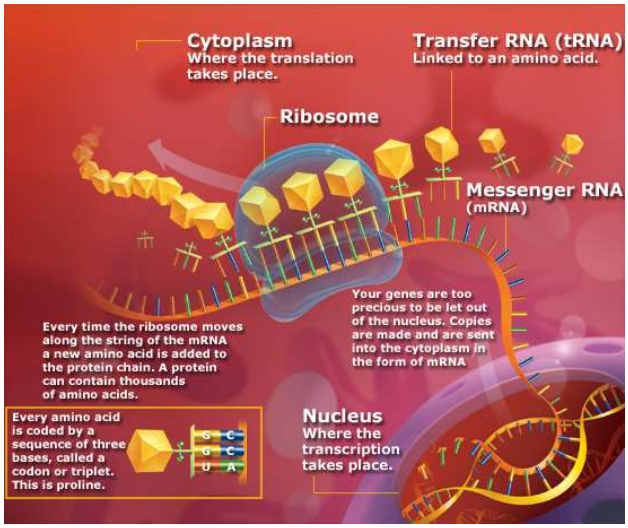


Figure 1.1: Visualisation of protein synthesis in a cell. Illustration reproduced courtesy of the Canadian Museum of Nature, Ottawa, Canada.

ysis, which tests hypotheses on computers (relatively cheap) and provides predictions to be tested by real experimental studies. The objective is to predict the dynamics of the underlying system so that the assumptions made about the operation of the system can be tested. The computational models are compared with experimental observations and inconsistencies mean that our knowledge of the system is at best incomplete. Models surviving the validation are used to make predictions that can be tested by experimenters.

Although the simulation based approach has received less attention in the past, the current experimental molecular biology is now supplying the high-throughput data necessary for also supporting the simulation approach. With the progress of the genome and proteome projects and fast increase in computational resources, there is tremendous potential in the systems approach and computational modelling and analysis are hoped to provide biological insights and predictions for targets like metabolic analysis, cell cycle oscillations and so on. The attempt to understand biological systems as systems, targeting the indentification of structure and dynamics, to control cellular behaviour by external stimuli and designing genetic circuits can only be achieved by combining computation, system analysis, comprehensive quantitative measurements and biological high-throughput experimental data.

There are many different components in a living cell that interact with each other, including *e.g.* genes and their products. These components give rise to the execution of normal cellular functions, seemingly complex (although coherent according to [60]) behaviour and interplay with the surrounding environment (including other cells). Cells are representative of

systems where the "whole" is seemingly more than the "sum of parts". A systems-wide analysis approach of such systems can be useful in gaining insight into their behaviour, requiring the introduction of a quantitative model of the components and their interactions. Further, mathematical, statistical and simulation tools are needed to understand the behaviour and how it relates to experimental data.

Cellular level high-throughput measurement technologies for studying biological organisms are the main tool of systems biology. The measurements have to be first comprehensive, secondly quantitatively accurate and thirdly systematic. For example, the ability to measure the expression of genes of a whole genome in a parallel way was made possible by the introduction of novel microarray technologies [113]. Although it is not yet possible to measure the gene expression in single cells on a large scale but instead in cell populations, this technology has truly changed the way genes can be studied. With DNA gene expression microarrays we can inspect genome-wide patterns of gene expression in any cell type, at any time and under any specified set of conditions [6] thus enabling us to study also the interplay between different genes. In Chapters 2 and 3 a more in depth discussion is given of the measurement technology, microarray data and how to simulate microarray measurements in a statistically sound way.

As pointed out, statistics and its applications in systems biology play an important role in processing the biological high throughput data produced by the modern measurement technologies. The high dimensional data we are dealing with consists usually of only few samples (large p small n) and the characteristics of interfering noise are unknown and non-Gaussian. Thus, the classical analysis methods are simply not guaranteed to be optimal when making conclusions based on the measurement data and it is imperative to use robust analysis methods. One of the main aspects of this thesis is to improve on robustness of computational analysis methods so that if our initial assumptions about the nature of the stray signals do not exactly hold, the results are still fairly reliable [84].

Cell cycle is an excellent example of how feedback loops work in a cell. The cell cycle is depicted in Figure 1.2 where we can see the growth of the cell and the eventual cell division after which the parent and descendant cell begin the growth process again. Cell cycle can be studied with help of microarray data by measuring gene expression over time. If the cells under study are in synchrony, the expression of genes that are mostly involved in the cell cycle should also oscillate at the period of the cycle. Chapter 4 deals with the problems of periodicity detection, including the detection of cell cycle related genes, in time series measurements with applications in microarray time series.

The effects of gene mutations are eventually observed in the proteins that genes produce. Some nucleotide polymorphisms (variation of a nucleotide in the DNA) do not change the corresponding amino acid alphabets at all

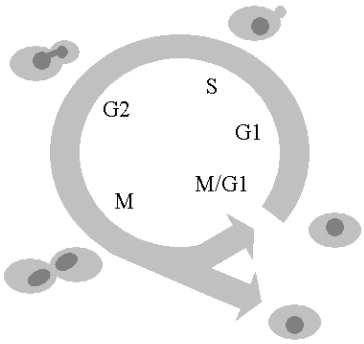


Figure 1.2: Illustration of the different phases of cell cycle, eventually leading to cell division.

and some amino acid mutations do not change the operation of the protein but the most interesting changes in proteins are the ones that alter the operation. The change of operation of a protein can be either beneficial or unfavourable. The detection of unfavourable changes that lead to malignant cancer is of utmost importance since knowing what causes the cancer can help finding specifically targeted cures. A disease phenotype can arise if amino acid substitutions result in structural alterations or loss of critical function in the amino acid sequence. However, it is also known that protein folds are rather robust and allow insertions to numerous sites without loss of function [100]. Protein folding refers to the process of amino acid sequences taking the three dimensional shape that depends on the surrounding environment. Hydrogen bonding and what is called hydrophobic effect are large contributors to the stability of the three dimensional shape of proteins [94]. In general, hydrophobic side chains of amino acids repel water and minimising their contact with water is an important factor contributing to the three dimensional shape. A backbone hydrogen-bonding based theory of folding is given in [106].

Predicting pathogenic mutations based on amino acid strings is difficult. Protein amino acid strings are seemingly random in the way that random permutations of real protein amino acid strings are difficult to discern from the original sequence by statistical means [137]. On the other hand, protein hydropathy sequences, which also relate to protein solvent accessibility, have been shown to deviate from randomness and are closely related to protein folding. Since amino acid changes modify the corresponding hydropathy values and cause misfolding, further leading to incorrect function, the study of hydropathy values is of great interest. A nonlinear analysis method known as recurrence quantification analysis (RQA) [126] has been successfully used in discriminating between different mutation classes based on hydropathy data [99]. A modification of RQA was used by us in Publication-V in pre-

dicting pathogenic mutation locations based on protein solvent accessibility data. Therefore, Chapter 5 is dedicated to the analysis of protein solvent accessibility sequences and how, with this information, we can predict pathogenic mutations in protein sequences.

Finally, concluding remarks, review of the results and future plans are discussed in Chapter 6

Chapter 2

Overview of microarray technologies

Modern high throughput methods for measuring gene expression, the activity of the genome to produce proteins in cells, have been in brisk development in the past decade. Protein production is mediated by the gene specific ribonucleic acids (RNA), so measuring the RNA content of a cell is indicative of the activity of its genes. Modern gene microarray technologies provide an elegant high throughput method of measuring this activity at the systems level.

In general, microarray technologies include deoxyribonucleic acid (DNA) hybridisation arrays, such as spotted two-channel arrays [113] and single channel arrays [74, 77], but also different protein arrays, tissue arrays and combinatorial chemistry arrays. The technological development in the area of different microarray technologies is likely to continue at a fast pace due to their huge success [6].

Gene expression microarrays make it possible to study genome-wide patterns of gene expression in any cell type at given times and set of conditions [6]. In gene expression microarray experiments, the total RNA of a cell population under study is reverse-transcribed into complementary DNA (cDNA, radioactively or fluorescently labelled), which is then hybridised on a glass or membrane support holding target DNA at known fixed positions. The cDNA that has attached to their counterparts on the support is then read by a laser excited scanner or other imaging techniques to produce gene expression measurements for thousands of target genes under various experimental conditions. The amount of data made available by microarray experiments is enormous and is hoped to provide fundamental insights into biological processes from gene function to development and cancer, among others. The amount of data also calls for efficient and statistically sound computational methods for making decisions based on the data.

It should be noted that gene microarrays can also be used to other ends

besides studying gene expression. For example, the ability of a group of proteins called transcription factors (TF) to bind to a promoter sequence of a gene can be studied with what is termed ChIP-chip microarrays (see [128] for a review).

We first review here the technological background of gene microarray technology and then present a model for simulating the microarray gene expression measurement process, based on Publication-I and [1].

2.1 Fabrication of gene expression microarrays

In practice, microarrays can be for example microscope slides that contain individual ordered samples (RNA, DNA, protein, tissue) where the type of the sample defines whether the assay is *e.g.* a DNA microarray or a tissue microarray. Since we know exactly where each sample or sequence is, the data that is obtained from the experiment can be traced back to any of the samples. Therefore each gene is addressable.

DNA microarrays are the most commonly used microarray type. There are more than one way of preparing a microarray, depending on how the array is fabricated. In two channel microarrays [113] the DNA that is printed on the microscope slide is enzymatically generated by polymerase chain reaction (PCR) from cellular messenger RNA (mRNA), using available DNA libraries. The DNA samples, or probes that are to be used in the array are fixed on the slide either by covalent bonding or with the help of electrostatic interactions. In these arrays, one spot on the slide usually corresponds to one gene and two samples with different markers are usually hybridised, a reference and a test, to obtain a differential measurement. Ratios of the reference and test are often used in studies.

A whole different approach is to synthesise the DNA directly on the slide itself by a photolithographic process using oligonucleotides (see Figure 2.1) [77]. In this context we usually speak of single-channel intensity-based oligonucleotide arrays or Affymetrix Inc. Genechips. With these kind of oligonucleotide arrays, the amount of target probes on the array is more exactly known. There is however a limit to how long sequences can be photolithographically synthesised and one gene is usually represented by more than one probe. These microarrays are usually single-channel and give good estimates of the absolute values of gene expression due to the well known amounts of targets on the array. The comparison of two conditions with single-channel microarrays requires usually the fabrication of two separate arrays.

Agilent Technologies provide also two-channel non-contact inkjet spotted oligonucleotide arrays [66]. These arrays combine some of the favorable properties of single-channel oligonucleotide arrays (specificity) and two-channel spotted arrays (length of the nucleotide chains) [28]. In addition, they also

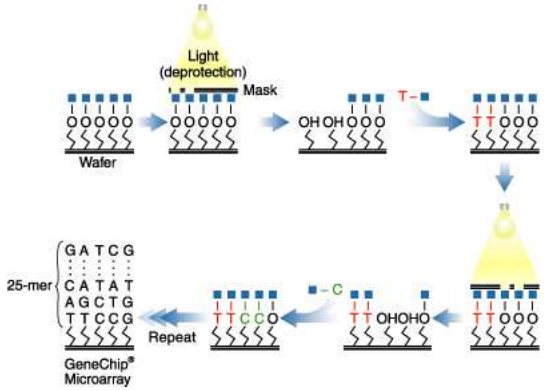


Figure 2.1: Fabrication of oligonucleotides on a Genechip (Affymetrix, Inc.)

offer simpler single-channel spotted oligonucleotide arrays.

DNA microarrays can be used to determine both the expression levels of genes in a sample (expression profiling) and the sequence of genes in a sample (minisequencing, mutation or single nucleotide polymorphism (SNP) analysis) [62]. To perform microarray experiments, it is not required to build a laboratory from scratch. Microarrays can be obtained from a variety of sources and commercial microarrays are of high quality, density and available for the most commonly studied organisms, including human, mouse, rat and yeast ([62]). A typical scheme of a two-channel microarray experiment can be seen in Figure 2.2 where samples from two cell populations are hybridised on the probes attached to the slide.

There are several steps left to perform after attaching the probes to the slide to obtain the gene expression values. These steps are reviewed next (as in [62]).

2.2 Extraction and labeling of the RNA samples

The measured quantity in DNA microarray experiments is actually the amount of mRNA present in cells. mRNA indirectly indicates what proteins are being synthesised. The three steps to labelling the mRNA are isolation of the mRNA from the cell population, labelling the mRNA by a reverse transcription procedure with fluorescent markers (most commonly Cy3 or Cy5) and purification of the labelled products. In the labelling process the labeled molecules are actually cDNA molecules which are produced from the mRNA by using the reverse transcriptase enzyme. If two samples are hybridised on the same array, each population is given a corresponding different label. If the amount of produced cDNA is small, PCR amplification can also be used.

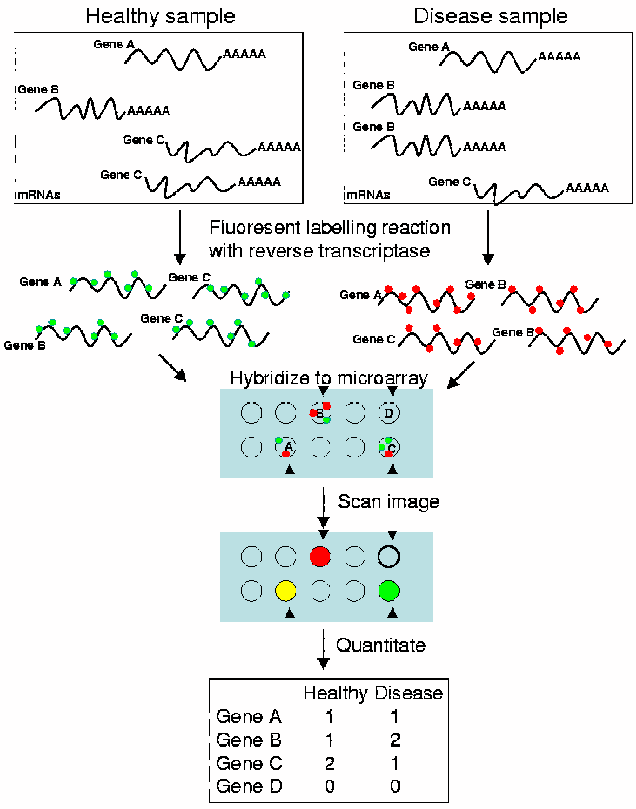


Figure 2.2: Layout of a two-channel gene expression microarray experiment [62].

2.3 Hybridisation of the sample on the slide

The fluorescently labelled samples are next hybridised onto the microarray slide where binding of the sample occurs with matching probes on the array. To keep the environment favourable, specific hybridisation chambers are used where temperature is kept constant and humidity can be controlled. The sample liquid is introduced onto the slide and incubated for several hours (overnight) for binding to occur before removing the part of the sample that has not bound to anything (*i.e.* no matching probe).

2.4 Scanning of the slide

After hybridisation, the fluorescently labeled bound molecules (probably not all cDNA samples had a target to bind to) can be read with a scanner. Usually microarray readers are scanning confocal microscopes with laser exciting at wavelengths proper for the (Cy3 or Cy5) dyes. The light emitted by the bound molecules is captured in a photomultiplier tube and the amount of radiation emitted is directly proportional to the amount of bound molecules. A two-channel microarray image is shown in Figure 2.4, where the channels are given their respective colours (Cy3 corresponding to green and Cy5 corresponding to red) to visually distinguish them. A single-channel array image is seen in Figure 2.3 [36]. The spots in the single-channel Genechip arrays are usually rectangular instead of round. Furthermore, the the error caused by non-specific binding for each gene perfect match (PM) probe is corrected with help of a mismatch (MM), which is a modified probe that should not correspond to any gene in the system under study.

2.5 Image processing of the data

The scanned microarray image must be further processed to obtain values corresponding to levels of gene expression. Since the Genechip type Affymetrix microarray images are usually stored in a closed format and are often preprocessed by their proprietary software, the following text is mostly focused on two-channel microarrays. At first, the spots on the array are separated by gridding the image. The aim of gridding is to detect the location of each spot so that the genes they correspond to can be addressed. Due to the high quality of the present day microarray technologies, automatic gridding is no longer a huge problem, even if there were some artefacts present on the slides. Such artefacts could be caused by for example slight scratches on the slide, misalignments of the spots or spots of varying size and shape. One approach to automatic gridding is shown in Figure 2.4, where horizontal and vertical projections of the image intensities are used to find approximate locations of columns and rows. Should the rows and columns be badly aligned

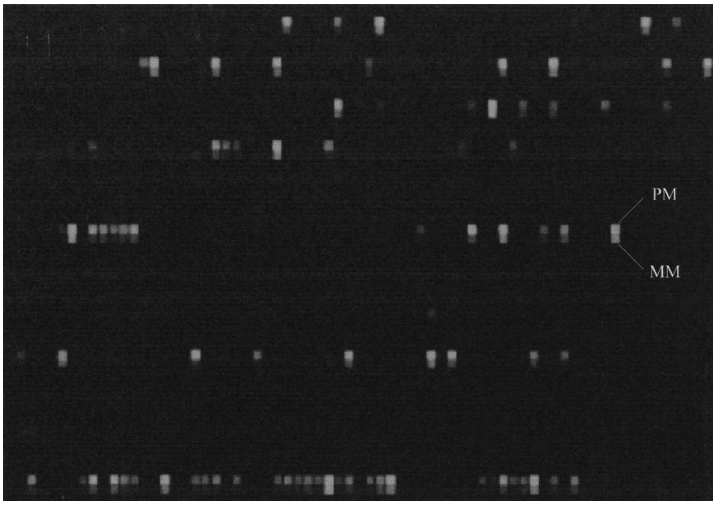


Figure 2.3: An example of a single-channel Genechip image [36]. The spots are rectangular rather than round, which is usually the case in printed two-channel microarrays. PM corresponds to perfect match and MM to mismatch.

or significant amounts of noise be present, the automatic gridding results can vary. Usually, the gridding algorithms in microarray analysis software allow manual adjustment of the gridding.

After gridding, segmentation of the image follows. The purpose of segmentation is to divide the image into a fore- and background so that the foreground is composed of the actual spots. If the spots were ideal, the segmentation would be trivial. However, since microarray images tend to be noisy and the spots less than ideal, robust approaches are needed to automatically separate the foreground from the background. One of the advanced segmentation algorithms is the watershed algorithm that does not assume a strict circular shape for the spots, see for example [67, 105]. After the spots are located, the intensity of each spot is estimated. A (trimmed) mean or median of the pixels inside the spot is computed and background corrected to yield the gene expression level estimate. The estimation of the background is based on either the whole image background or the local neighbourhood of the spot.

2.6 Preprocessing and normalisation of the data

As explained in [62], preprocessing and normalisation of the data are needed before further analyses. Preprocessing includes several steps. First, handling missing values and their possible imputation is important since missing values can seriously interfere with statistical testing and clustering. Compar-

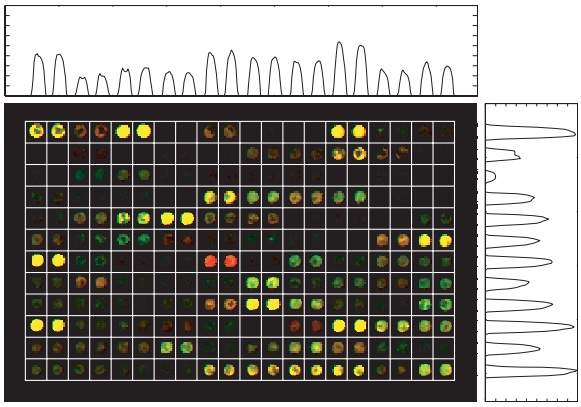


Figure 2.4: Automatic gridding of a two-channel microarray image. The two channels are given their respective colours corresponding to the wavelength of the laser used in reading the spots - red and green in this case [62].

ison between two conditions (channels or arrays) is usually conducted by considering the ratio of the intensities. Since a low expression of the test sample is confined in ratio in the interval between zero and one and high expression from one to infinity, the ratio is usually log-transformed to make the distribution more symmetric. In addition, different biological and technical replicates can be used to assess nonbiological variation in the data. Sometimes there can be higly inconsistent values in the data that can lead to biased statistical analyses. The detection and handling of these gross errors (outliers) is also necessary. For more details on data preprocessing, see [62] and references therein.

Normalisation, which can be thought of as part of preprocessing, is an important process of removing systematic variation not corresponding to gene expression levels in data [114]. Normalisation is necessary to be able to compare different channels or arrays in a meaningful way. The aim of normalisation is to remove systematic biases in the data but the problem is how to tell what is biological and what is not. The systematic biases can be caused by differences in labelling between the used dyes, differing powers of the lasers in the scanners, uneven hybridisation and so on. As an example, it is possible that in a two-channel microarray experiment, one of the dyes labels the sample more efficiently than the other, by a factor greater than one. Without normalisation, the gene expression in the first sample could be falsely considered higher even if they were equal. Different algorithms and methods (mean centering, median centering, standardisation, lowess smoothing to linearise channels) have been proposed for normalisation, but the basic idea is to scale the mean intensity ratio between the two channels or chips to equal to one. Then, the logarithm of the ratio for nondifferentially expressed genes should be zero. These methods are usually applied with-

in slide, but if several arrays have been prepared for the same experiment, between-slide normalisation can also be applied. It should be noted that normalisation makes assumptions about the data but is necessary for making comparisons.

2.7 Uses of microarray data

Currently microarrays are used mainly to monitor the expression levels of genes in comparison between two conditions, whether they are environmental, nutritional, chemical or related to temperature changes. This kind of study, namely gene expression profiling, is used to assess the function of specific genes under changes in the above-mentioned conditions [62].

In the past years there has been an explosion of available microarray gene expression data. Time series experiments, where the changes in gene expression are observed over time, have also become abundant. Because the data is inherently noisy and usually sparsely sampled, better algorithms are being developed to extract as much information as possible.

Chapter 3

Simulation of gene expression microarrays

In Publication-I and [1] we introduced a modular microarray simulation platform that combines all the steps taken in *e.g.* real gene microarray experiments, from gene regulatory networks to the scanned microarray slide. A big issue in *e.g.* validating different computational data analysis algorithms on simulated data is the mathematical model chosen for the data. The analysis methods that assume a similar model that is used in generating the data are favoured over others. Our modular approach allows the use of many different models for the data so that the analysis methods can be compared over different classes of models.

The introduced simulation platform has many possible uses in simulating the images of different microarray platforms, simulating the effects of noise on different biological systems and validating different computational algorithms, such as image processing, background noise removal, normalisation, clustering, classification and regulatory network inference. To validate different data analysis methods on actual measurement data requires a lot of measurements and knowledge on the biological ground truth. Since it can be very expensive to perform microarray experiments just for computational method development and the ground truth can be hard to discern (discussion in [88]), a simulation based approach is always welcome. For meaningful results, the simulated data must have similar characteristics to those in real biological data. In addition, if a biologically relevant model can be simulated accurately, it can be used in testing hypotheses *in silico*, in a fast and cost effective way. For Affymetrix type data, there exists real measurement data (known as spike-in and dilution data, see [56]) where the ground truth is approximately known for the measurements (using controlled samples). This type of data is especially good for evaluating the performance of normalisation algorithms [15] and assessing differentially expressed genes. This does not invalidate our simulation platform in the evaluation of the performance

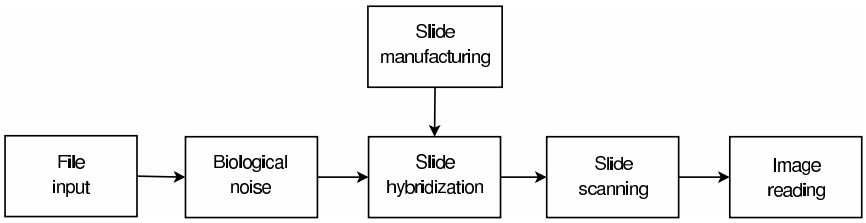


Figure 3.1: Layout of the microarray simulator. The different blocks of the simulator are modular in the way that only their inputs and outputs are specified, whereas the implementation can be changed.

of computational methods, since the platform can be used for generating all types of different data for different purposes, provided that a reasonable model exists for the simulated target.

We first discuss different ground truth models and then review the microarray simulation platform. The layout of the platform can be seen in Figure 3.1, where the modularity of the platform is also illustrated.

3.1 Ground truth models

Generating meaningful ground truth data depends on the application where the data is to be used. Since a microarray experiment may include comparisons of different classes of samples, measuring response to perturbations or measuring time series behaviour among other things, different ground truth data generation schemes must be considered.

The simplest approach of generating ground truth data is to take random samples from a specified distribution. The parameters of the distribution can be chosen ad hoc or estimated [31] from real data sets, if such data exists. Should it be necessary to compare for example how different analysis methods perform in detecting differentially expressed genes, the ground truth data could be sampled from two distributions, which have differing location parameters.

Real measurement data can also be used as ground truth data, to verify for example how data analysis algorithms can handle different amounts of added noise in data. By using real intensity data it is also possible to simulate the scanned images.

For a list of other ground truth models, see Publication-I.

3.2 File input

The file input block converts the input data from a file to the internal format of the simulator. Supported formats include for example simulated expres-

sion values as intensities or ratios. Different parameters include for example the number of subarrays and number of columns and rows in each subarray.

3.3 Biological stray signals and error models

The aim of the biological noise module is to take into account the fact that whatever the true expression level of a gene is in a cell, it cannot be directly exactly measured with the currently available technology. The characteristics of simulated data should thus include biological and measurement technology related errors. Biological errors are usually considered to include phenomena such as the stochastic noise of the cells and sample preparation errors [12, 39]. Measurement technology related limitations and errors are typical of the chosen platform; one- or two-channel microarray-based [121]. In [121] the authors state that for oligonucleotide-based microarray experiments the sample preparation noise is small if compared to the hybridisation related noise.

Since actual microarray measurements are performed for cell populations, the measured expressions are averages over many cells. To take this population effect into account, smoothing of the ideal ground truth data can be executed for example by using an averaging kernel [81]. After taking the population effect into account, noise with different characteristics can be added to the simulated data. Different noise models and noise parameter estimation methods have gained a lot of attention lately [25,28,30,31,48,51–53,69,70,76,87,93,104,116,121,129,132]. The approaches usually also produce lists of differentially expressed genes. Many of the models are technology specific but some can also be applied to both one- and two-channel array data. Frequentist models are usually based on maximum likelihood or least squares estimation, see *e.g.* [28,51,53,70,104] ([116] for an extended quasi-likelihood approach), whereas Bayesian approaches rely on the use of priors and posterior probability density estimates of the parameters, see [25, 30, 31, 48, 52, 69, 76, 87]. An attempt to unify a number of error models is presented in [132]. The models for the gene expression level are all captured by the following model

$$y = f(x) + e, \tag{3.1}$$

where y is the observed expression value, x is the true gene expression value, f is a nonlinear function dependent on the gene expression level and e is an error term independent of the expression level. More specific models are usually given for f and e to allow estimating the error in real data. The models typically include separate terms for gene specific noise, measurement specific noise, array specific noise, biological sample specific noise and other possible noise sources [25, 31].

As mentioned, some of the error sources must be implemented as technology specific. For example, in Genechip oligonucleotide arrays several probes correspond to one gene so they should have some level of dependence. Further, the perfect match and mismatch probes need to be handled independently [52]. These effects have been implemented, among others, in our simulation model and new features can be added easily.

A simple error model for single-channel arrays was introduced in [51] that makes the assumption of multiplicative noise only, transforming to a log-additive model. The model for the log-transformed data y_{ij}, where i indexes the spiked controls and j indexes the separate chips, is

$$y_{ij} = \mu_i + \rho_j + \epsilon_{ij}. \tag{3.2}$$

The authors assume that ϵ_{ij} is randomly distributed and drawn from the central normal distribution with variance σ_i^2, ρ_j corresponds to the chip specific error and μ_i corresponds to the true log-transformed gene expression value. Thus, the log-reported expression levels are distributed

$$y_{ij} \sim N(\mu_i + \rho_j, \sigma_i^2). \tag{3.3}$$

The authors also derive maximum-likelihood and maximum a posteriori estimates for the parameters and apply them to *Saccharomyces cerevisiae* data obtained from Affymetrix GeneChips.

A more complicated model suiting both one- and two-channel arrays was introduced in [104]. The model is

$$y_{ij} = \alpha_j + \mu_i e^{\eta} + \epsilon j, \tag{3.4}$$

where y_{ij} is the intensity measurement, α_j is the mean intensity of unexpressed genes and μ_i is the expression level in arbitrary units. The error terms are assumed $\epsilon j \sim N(0, \sigma_\epsilon^2)$ and $\eta \sim N(0, \sigma_\eta^2)$, which is a proportional error present in all measurements but noticeable mainly for highly expressed genes. The authors give instructions on how to estimate the background using negative controls, replicate measurements and what to do if no replicates are available. The estimation of σ_η^2 from high level expressed genes is also discussed. After estimating the parameters, the authors give instructions on how comparison of the expressions can be evaluated. Besides these two simpler models, many more elegant models exist. Future plans include the verification of the performance of the most promising error models.

These errors will affect the ideal ground truth data but not the physical appearence of the array. Physical effects are introduced in the slide manufacturing, hybridisation and slide scanning modules. The current implementation of the simulator includes several error models proposed in the literature [25, 31, 51, 52, 93, 104]. The methods proposed in [31, 51, 52] can also be used to estimate the error parameters from the expression data.

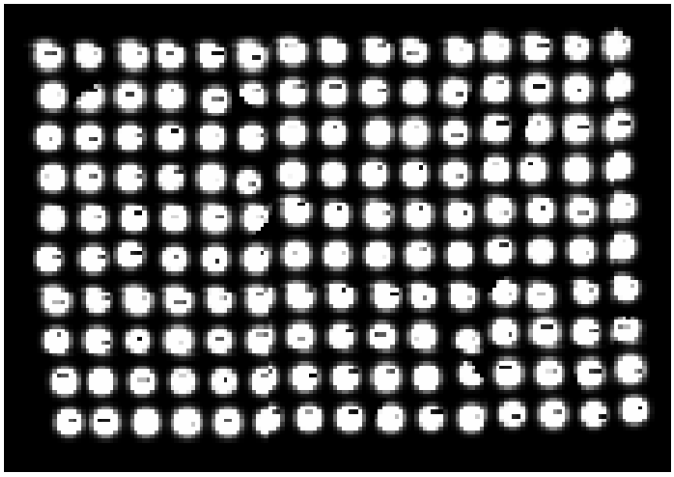


Figure 3.2: Visualisation of different physical artefacts related to slide manufacturing.

3.4 Slide manufacturing

Since the spots on a microarray slide that are composed of the probes are not always the same size and quality, artefacts related to slide manufacturing were also considered in Publication-I. Error sources related to array fabrication that are usually visible include variation in spot sizes (not so much in Genechip oligonucleotide arrays), marks done by the printing tip and deformations in the spot shapes, like chords cut away from the spots. These artefacts are shown for a two-channel type microarray template in Figure 3.2 where printing tip holes, cuts in spots and varying spot sizes are visible. The effect of misaligning the spots is also visible. For more information, see Publication-I.

3.5 Hybridisation

Hybridisation, the phase where the sample populations of labelled cDNA are introduced on the array, is simulated in the model by modifying the spots generated in the slide manufacturing phase. Since different array technologies produce different types of spots, different spot shapes are included in the model.

A simple Gaussian distribution has been shown to fit spotted two-channel microarray spots [21]. Other more complicated models were considered in Publication-I as well, including polynomial hyperbolic spot shapes [35] and rectangular shapes for Genechip type arrays. Hybridisation effects are included in the simplest form by applying multiplicative Gaussian noise, with user tunable parameters, to the ideal spot pixels. The hybridised spot is

then obtained by multiplying the spot of high brightness with the ground truth value (from zero to one). Other error sources, like background noise, spot bleeding, scratches and artificial air bubbles are included and whose parameters are controllable.

A visualisation of ideal and noisy Gaussian shaped two-channel array spots is shown in Publication-I Figure 4 (a-b). A noisy simulated single-channel array spot is visualised in Publication-I Figure 4 (c). Some of the different simulated hybridisation errors are also visible in the simulated image in Figure 3.3.

3.6 Slide scanning and image reading

In real microarray experiments, hybridisation is followed by scanning of the slide and reading the intensity values from the image. Since all real scanners have a limited dynamic range, saturation effects are included in the model. Different channels in multi-channel microarrays may also get misaligned and the slide may not be scanned straight, so these artefacts are also considered. Dye effects that can give different scanner readings for same true expression levels of two samples are also controllable. Since the simulated images are comparable to real microarray images, any microarray image reading software can be used for gridding and reading the intensities. However, an automatic grid alignment and image segmentation algorithm is included in the simulator package but can be replaced by any other software just as well.

3.7 Simulated images and results

Microarray images simulated with the introduced platform are shown in this section. The results obtained in Publication-I are also briefly reviewed. To verify that the simulator produces meaningful results, several ground truth data and noise models are considered and compared to real microarrays.

First, ground truth reference data was generated using random network topology with kinetic rate laws for gene mRNA amounts [89]. A random gene knockout was then simulated to obtain a test sample (second channel). Both models were simulated for a hypothetical time of 200 minutes and a hierarchial error model [25] was applied to add noise to the data. The resulting simulated microarray images can be seen if Figure 3.3. In the leftmost array (10 minutes) most spots are yellow, indicating no differential expression. On the right (200 minutes), the effects of gene knockout are visible so that some spots are green and some are red, indicating differential expression. This visualises the ability of the platform to imitate real experiments.

Secondly, a self versus self experiment was simulated by assigning same ground truth values for the two samples. The values were drawn from an ex-

Figure 3.3: Simulated two-channel microarrays corresponding to 20 minutes (a) and 200 minutes (b) after the beginnning of the hypothetical experiment. More differentially expressed spots can be seen in (b) and several physical artefacts like scratches and misalignments are visible.

ponential distribution and centered normal noise was added independently to both channels to make the channels differ. A transformation of the data was then performed as suggested in [5] to generate the true intensity values. Two experiments were simulated, *i.e.* an array for the generated data without applying further noise models and an array where a hierarchial error model [25] was applied. Scatter plots of the read intensities of these arrays were then compared with the original noise free data. Considerable similarities to scatter plots of real data are visible (plots not shown here). See Publication-I for further details.

Next, real Affymetrix Genechip readings were passed through the simulator. Part of the simulated microarray image can be seen with the original Affymetrix image in Publication-I Figure 10. The proprietary text in the upper part is clearly visible in both the original and the simulated image.

Finally, the capability of the simulator in making comparisons of different data analysis algorithms (spot segmentation in this case) was illustrated in Publication-I in a consistent way.

These examples illustrate the usefulness of the microarray simulation platform and give only a few suggestions on how to use it. The modularity of the platform gives the user the freedom to suit the needs of any imaginable microarray related validation or verification purpose.

Chapter 4

Detection of periodicity in gene expression time series

Data obtained from microarray experiments can be measured in consecutive time points, thus allowing access to time series. If the cell populations under study are in synchrony according to the periodicity of interest (see [110,115]), we can assess the periodically expressed genes from the measured time series by using traditional time series analysis methods.

Periodic phenomena on the cell and gene level include for example cAMP oscillations, cell cycle and circadian rhythms [122]. The technologies that measure these phenomena are usually indirect and consist of multiple phases that each add stray signals to the actual signal of interest.

Periodicity detection applied to gene expression data has been previously considered in [3, 24, 29, 45, 58, 72, 75, 78, 79, 131, 139]. These methods vary on a wide scale of approaches from using the frequentist approach in the detection of periodic genes at *a priori* unknown frequency (Fisher's test, applied in [131]) and *a priori* known (or hypothesised) frequency [29, 58] to Bayesian periodicity detection [3]. Periodicity in gene expression has been shown to be indicative of for example cell-cycle regulation [17] and circadian rhythms [27]. The recent increasing interest in detecting cell-cycle regulated genes has been mediated by the verification of connection between cell-cycle and cancer [130].

Non-uniform time series sampling is often encountered in systems biological studies and microarray experiments. Most previously published periodicity detection methods intrinsically assume uniform sampling and are therefore not directly suitable for detection in non-uniformly sampled data. Exception to this is presented in [45] where the authors use a modification of the periodogram (Lomb-Scargle periodogram) to detect periodic patterns in gene expression time series with no prior knowledge on the cycle period. A refined approach using B-splines is considered in [72]. In [29, 58, 75, 78, 79] non-uniform sampling is taken into account but the methods need accurate

prior information about the cycle period.

The Bayesian approach into spectrum estimation and periodicity detection has been covered in [3, 18, 102, 140]. The Bayesian methods make use of prior knowledge such as the approximate frequency of the oscillation and prior distributions for the estimated variables. In [3] the authors show that their approach is also robust in the sense that it can handle Laplacian and uniform noise in the data besides Gaussian. The method can also handle non-uniform sampling but no discussion is given about robustness against outlying data (a definition of an outlier is given below).

In this chapter we introduce robust nonparametric methods for periodicity detection that are dependendent on distributional assumptions only in the approximate sense. Therefore the introduced methods do not necessarily produce the best results in the case of normality assumption but produce better results on a wider scale of different distributions than the classical methods [84]. With classical we mean the basic methods, which assume that the distributional assumptions (usually Gaussian) hold exactly [84]. A basic example is the sample mean, which is the minimum variance unbiased (MVU) estimator for location of the normal distribution. The word robustness in this context implies insensitivity to changes in the distributional assumptions. A method that can provide reliable results even if the assumed distribution for the data does not exactly hold is called robust [84].

Outliers, which are usually defined as points that are inconsistent with the majority of the data [97], are closely related to the notion of robustness. Outliers are gross errors which can be the result of *e.g.* decimal point shifts or scratches in microarray slides that cause erroneous scanner readings. Non-robust methods usually give false results if the data is contaminated by outliers. The least squares based methods square-weight the residuals that are to be minimised so outliers have a tendancy to bias the estimates in a degrading manner.

In microarray gene expression studies the measured signals are usually covered by noise whose characteristics are not well known. The distribution of the samples can be strongly non-Gaussian [91] and a lot of samples can be missing [14] or outlying. These are the main reasons why classical time series analysis methods should be used with caution and why we have developed robust methods for periodicity detection. Most of the previously published periodicity detection methods are not robust when outlying data are present in the time series.

In the following we first review the background of the classical time series analysis methods by Hilbert space formulation and the connection of the periodogram spectral estimator to the projection theorem [19]. The periodogram spectral estimator, although non-robust, is a natural classical choice for periodicity detection in time series. Fisher's test for the detection of hidden periodicities, a test based on the periodogram, is then reviewed before introducing the robust estimators. Of the robust estimators we first

consider a robust rank based periodicity detection method (Publication-II) that performs well both with simulated and measured microarray time series. The rank based method is however not designed to handle non-uniform sampling. Therefore we next consider the problem of non-uniform sampling and introduce a robust regression based framework for periodicity detection (Publication-III). Finally, we consider the regression framework in the special case of uniform sampling in Publication-IV. In this approach, we replace least squares fitting in Fisher's test with robust M-estimate (maximum likelihood type) regression, for which we show that the original analytic null hypothesis distribution of Fisher's test approximately holds. This yields a robust test for periodicity that is fast to compute.

4.1 Stochastic processes, stationarity and Hilbert spaces

As according to [19], the analysis of time series is initiated by selecting a proper mathematical model or class of models for the data. Due to the unpredictable nature of measured variables it is reasonable to assume that our observed variables are realisations of certain random variables. We therefore define the time series $\{x_t, t \in T_0\}$ as a realisation of the family of random variables $\{X_t, t \in T_0\}$. This suggests modelling the data as a realisation of a stochastic process $\{X_t, t \in T\}$ where the index set $T \supseteq T_0$ (usually $\{0, \pm 1, \pm 2, \ldots\}$ or $\{1, 2, 3, \ldots\}$). Further, a stochastic process is a family of random variables $\{X_t, t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$ with Ω as the sample space, $\mathcal{F}$ the so-called sigma-algebra of the subsets of Ω and P the probability measure (restricted on the interval $[0, 1]$). It is noted that for a fixed $t \in T$, $X_t(\cdot)$ is a function on the set Ω and, on the other hand, for each fixed $\omega \in \Omega$, $X_\cdot(\omega)$ is a function on T.

Kolmogorov's theorem asserts the connection between a stochastic process and its distribution function, *i.e.* the probability distribution functions $\{F_{\mathbf{t}}(\cdot), \mathbf{t} \in \mathcal{T}\}$ ($\mathcal{T}$ is the set of vectors $\{\mathbf{t} = (t_1, \ldots, t_n)' \in T^n : t_1 < t_2 < \ldots < t_n, n = 1, 2, \ldots\}$) are the distribution functions of some stochastic process if and only if for any $n \in \{1, 2, \ldots\}, \mathbf{t} = (t_1, \ldots, t_n)' \in \mathcal{T}$ and $1 \leq j \leq n$,

$$\lim_{x_j \to \infty} F_{\mathbf{t}}(\mathbf{x}) = F_{\mathbf{t}(j)}(\mathbf{x}(j)), \tag{4.1}$$

where $F_{\mathbf{t}}(\mathbf{x}) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n)$. The $(n-1)$-component vectors $\mathbf{t}(j)$ and $\mathbf{x}(j)$ are obtained by deleting the j^{th} components of $\mathbf{t}$ and $\mathbf{x}$, respectively. This simply states that each function $F_{\mathbf{t}}(\cdot)$ should have marginal distributions that coincide with the specified lower dimensional distribution functions.

Before defining (weak) stationarity we first recall the definition of the autocovariance function. If $\{X_t, t \in T\}$ is a process such that $\forall t \in T$:

$\mathrm{Var}(X_t) < \infty$, then the autocovariance function $\gamma_X(\cdot,\cdot)$ of $\{X_t\}$ is defined by

$$\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s) = E[(X_r - E(X_r))(X_s - E(X_s))],\ r,s \in T. \quad (4.2)$$

Stationarity is then defined for a time series $\{X_t, t \in \mathbb{Z}\}$ with index set $T = \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ as

$$\begin{array}{lll} \text{(i)} & E|X_t|^2 < \infty, & \forall t \in \mathbb{Z}, \\ \text{(ii)} & E[X_t] = m, & \forall t \in \mathbb{Z} \\ \text{(iii)} & \gamma_X(r,s) = \gamma_X(r+t, s+t), & \forall r,s,t \in \mathbb{Z}. \end{array} \quad (4.3)$$

which means that the expected values of the squared variables must be limited, the time series variables have the same (limited) expected value and the autocovariance function is time shift-invariant.

4.1.1 Hilbert spaces

There are some important properties of Hilbert spaces (of which the Euclidean space is a special case) that are extremely useful in spectral analysis and periodicity detection in time series [19]. The most important properties of Hilbert spaces needed in this thesis are related to the norm induced by the inner product and the projection theorem, both of which are briefly reviewed in Appendix A. It should be noted that the periodogram spectral estimator, which will be introduced in the next section, is obtained by projecting the measured time series on a set of orthonormal sinusoidals. A more in-depth presentation of Hilbert spaces is given in [19].

4.2 Periodicity detection in stationary processes

In this section we consider the periodogram spectral estimator in the context of statistical inference for time series in the frequency domain. We first define the periodogram and then present tests for the presence of hidden periodicities in data (as given in [19]). Different models and hypotheses for the time series are discussed.

4.2.1 The periodogram

Let $x_1, \ldots, x_n$ represent n complex valued data observed at integer times $1, \ldots, n$. The vector $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{C}^n$ can be represented as a linear combination

$$\mathbf{x} = \sum_{j \in F_n} a_j \mathbf{e}_j, \quad (4.4)$$

where (with i the imaginary unit) $\mathbf{e}_j = n^{-1/2}(e^{i\omega_j}, e^{i2\omega_j}, \ldots, e^{in\omega_j}),\ j \in F_n$ and $F_n = \{j \in \mathbb{Z} : -\pi < \omega_j \equiv 2\pi j/n \leq \pi\} = \{-[(n-1)/2], \ldots, [n/2]\}$ and

$[x]$ denotes the integer part of x [19]. This holds since the vectors $\mathbf{e}_j,\ j \in F_n$ can be shown to constitute an orthonormal basis for the space $\mathbb{C}^n$. Thus

$$a_j = \langle \mathbf{x}, \mathbf{e}_j \rangle = n^{-1/2} \sum_{t=1}^{n} x_t e^{-it\omega_j}. \tag{4.5}$$

The sequence $\{a_j,\ j \in F_n\}$ is also called the discrete Fourier transform of $\mathbf{x} \in \mathbb{C}^n$.

The periodogram of $\mathbf{x} \in \mathbb{C}^n$ is defined by

$$I(\omega_j) = |a_j|^2 = |\langle \mathbf{x}, \mathbf{e}_j \rangle|^2 = n^{-1} \left| \sum_{t=1}^{n} x_t e^{-it\omega_j} \right|^2,\ j \in F_n \tag{4.6}$$

and further we can show that the periodogram decomposes $\|\mathbf{x}\|^2$ in the following way

$$\|\mathbf{x}\|^2 = \sum_{j \in F_n} I(\omega_j), \tag{4.7}$$

which allows us to view the frequency components in the analysis of variance sense. It should be noted that harmonics outside the interval $(-\pi, \pi]$ cannot be distinguished by the periodogram based on observations at integer times only.

For real valued signals $\mathbf{x} \in \mathbb{R}^n$ if $\omega_j, -\omega_j \in (-\pi, \pi]$ ($\omega_j = 2\pi j/n$, also known as the Fourier frequencies), then $a_j = \overline{a}_{-j}$ and $I(\omega_j) = I(-\omega_j)$. In this case Equation (4.4) reduces to

$$\begin{array}{rcl} \mathbf{x} & = & a_0\mathbf{e}_0 + \sum_{j=1}^{[(n-1)/2]} (a_j \mathbf{e}_j + \overline{a}_j \mathbf{e}_{-j}) + a_{n/2}\mathbf{e}_{n/2} \\ & = & a_0\mathbf{e}_0 + \sum_{j=1}^{[(n-1)/2]} \sqrt{2} r_j (\mathbf{c}_j \cos\theta_j + \mathbf{s}_j \sin\theta_j) + a_{n/2}\mathbf{e}_{n/2}, \end{array} \tag{4.8}$$

where we have expressed $a_j = r_j \exp(i\theta_j)$ in the polar form and $a_{n/2} = 0$ in case n is odd. Vectors $\mathbf{c}_j$ and $\mathbf{s}_j$ are defined as

$$\begin{array}{l} \mathbf{c}_j = \sqrt{(2/n)}(\cos\omega_j, \cos 2\omega_j, \ldots, \cos n\omega_j)', \\ \mathbf{s}_j = \sqrt{(2/n)}(\sin\omega_j, \sin 2\omega_j, \ldots, \sin n\omega_j)' \end{array} \tag{4.9}$$

and $\{\mathbf{e}_0, \mathbf{c}_1, \mathbf{s}_1, \ldots, \mathbf{c}_{[(n-1)/2]}, \mathbf{s}_{[(n-1)/2]}, \mathbf{e}_{n/2}\}$ (excluding $\mathbf{e}_{n/2}$ if n is odd) constitute an orthonormal basis for $\mathbb{R}^n$. As in the case of complex valued signals, we can now decompose $\sum_{i=1}^{n} x_i^2$ into components corresponding to the vectors that form the basis thus allowing the analysis of variance (Table 4.2.1). Usually for frequencies $\omega_j,\ 1 \le j \le [(n-1)/2]$ a single component per frequency is given, corresponding to the length of the projection of $\mathbf{x}$ onto the two-dimensional subspace $\text{span}\{\mathbf{c}_j, \mathbf{s}_j\}$, as in Table 4.2.1.

For convenience, we redefine the autocorrelation function for a stationary process (for which $\gamma(r,s) = \gamma(r-s, 0)\ \forall r, s \in \mathbb{Z}$ [19]) as

$$\gamma_{_X}(k) \equiv \gamma_{_X}(k, 0) = \text{Cov}(X_{t+k}, X_t),\ \forall k, t \in \mathbb{Z}.$$

A stationary time series $\{X_t\}$ with mean μ and absolutely summable autocovariance function $\gamma_X(\cdot)$ has a continuous spectral density

$$f(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_X(k) e^{-ik\omega}, \ \omega \in [-\pi, \pi], \tag{4.10}$$

The correlogram spectral estimator, which is a scaled estimate of the above spectral density, is given at the Fourier frequencies by

$$\begin{cases} I(\omega_j) = \sum_{|k|<n} \hat{\gamma}_X(k) e^{-ik\omega_j}, & \omega_j \neq 0 \\ I(0) = n|\hat{\mu}_x|^2, \end{cases} , \tag{4.11}$$

where $\hat{\mu}_x = n^{-1} \sum_{t=1}^{n} x_t$. If the autocovariance function estimate is chosen as $\hat{\gamma}_X(k) = n^{-1} \sum_{t=1}^{n-|k|} (x_t - \hat{\mu}_x)(x_{t+|k|} - \hat{\mu}_x)$, which is actually a biased estimate, we can show that the correlogram spectral estimator coincides with the periodogram spectral estimator. This shows the potential value of the periodogram for spectral density estimation and also opens up possibilities for modifying the periodicity detection tests (introduced in the next subsection) that are based on the periodogram ordinates. This is because we can replace the periodogram spectral estimator with the correlogram in the tests and also consider robust alternatives for the correlogram (as in Publication-II).

4.2.2 Tests for the presence of hidden periodicities

Based on the periodogram, several statistical tests can be formed to test the null hypothesis H_0 that the data $\{X_1, \ldots, X_n\}$ are generated by a Gaussian white noise process, against the alternative hypothesis (denoted H_1) that the data have a superimposed deterministic periodic component in addition to white Gaussian noise [19].

Table 4.1: Analysis of variance table for the harmonic decomposition of a real valued signal (as in [19]).

Source	Degrees of freedom	Sum of squares
ω_0 (mean)	1	$a_0^2 = n^{-1}(\sum_{t=1}^{n} x_t)^2 = I(0)$
ω_1	2	$2r_1^2 = 2\|a_1\|^2 = 2I(\omega_1)$
$\vdots$	$\vdots$	$\vdots$
ω_k	2	$2r_k^2 = 2\|a_k\|^2$
$\vdots$	$\vdots$	$\vdots$
$\omega_{n/2}$	1	$a_{n/2}^2 = I(\pi)$
Total	n	$\sum_{j=1}^{n} x_j^2$

The model for the data is chosen as

$$X_t = \mu + A\cos\omega t + B\sin\omega t + Z_t, \tag{4.12}$$

where Z_t is zero-mean Gaussian white noise with variance σ^2, A and B are deterministic constants and ω is the frequency of periodicity. As a side note, any sinusoidal $\tau\sin(\omega t + \theta)$, $\tau \in \mathbb{R}$ can be expressed by the above sum of cosine and sine terms. To formally define the null and alternative hypotheses, we choose

$$\begin{aligned} &H_0 : A = B = 0, \\ &H_1 : A \text{ and } B \text{ are not both zero.} \end{aligned} \tag{4.13}$$

In case we have prior knowledge on the frequency of periodicity, and ω is one of the Fourier frequencies $\omega_k = 2\pi k/n \in (0, \pi)$, the analysis of variance (Table 4.2.1) provides an easy test for periodicity. We can write the model in Equation (4.12) in a similar form as in Equation (4.8), *i.e.*

$$\mathbf{X} = \sqrt{n}\mu\mathbf{e}_0 + \sqrt{(n/2)}A\mathbf{c}_k + \sqrt{(n/2)}B\mathbf{s}_k + \mathbf{Z},\ \mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I_n). \tag{4.14}$$

In this case we reject H_0 in favour of H_1 if $2I(\omega_k)$ in Table 4.2.1 is sufficiently large. Under H_0

$$2I(\omega_k) = \|P_{\mathrm{span}\{\mathbf{c}_k,\mathbf{s}_k\}}\mathbf{X}\|^2 = \|P_{\mathrm{span}\{\mathbf{c}_k,\mathbf{s}_k\}}\mathbf{Z}\|^2 \sim \sigma^2\chi^2(2), \tag{4.15}$$

which is independent of

$$\|\mathbf{X} - P_{\mathrm{span}\{\mathbf{e}_0,\mathbf{c}_k,\mathbf{s}_k\}}\mathbf{X}\|^2 = \sum_{i=1}^{n} X_i^2 - I(0) - 2I(\omega_k) \sim \sigma^2\chi^2(n-3). \tag{4.16}$$

This leaves us to reject H_0 at level α if

$$(n-3)I(\omega_k)/\left[\sum_{i=1}^{n} X_i^2 - I(0) - 2I(\omega_k)\right] > F_{1-\alpha}(2, n-3). \tag{4.17}$$

If ω is not one of the Fourier frequencies, the testing is a bit more complicated but follows the same directions. Using the model of Equation (4.14) but replacing $\mathbf{c}_k, \mathbf{s}_k$ with the non-orthogonal vectors

$$\begin{aligned} \mathbf{c} &= \sqrt{(2/n)}(\cos\omega, \cos 2\omega, \ldots, \cos n\omega)', \\ \mathbf{s} &= \sqrt{(2/n)}(\sin\omega, \sin 2\omega, \ldots, \sin n\omega)', \end{aligned} \tag{4.18}$$

yields a modified test. In this case we reject H_0 in favour of H_1 if

$$2I^*(\omega) := \|P_{\mathrm{span}\{\mathbf{e}_0,\mathbf{c},\mathbf{s}\}}\mathbf{X} - P_{\mathrm{span}\{\mathbf{e}_0\}}\mathbf{X}\|^2 \tag{4.19}$$

is sufficiently large. Under H_0, $2I^*(\omega) \sim \sigma^2\chi^2(2)$ and $I^*(\omega)$ is independent of $\|\mathbf{X} - P_{\mathrm{span}\{\mathbf{e}_0,\mathbf{c},\mathbf{s}\}}\mathbf{X}\|^2 \sim \sigma^2\chi^2(N-3)$. Therefore we reject H_0 at level α if

$$(n-3)I^*(\omega)/\|\mathbf{X} - P_{\mathrm{span}\{\mathbf{e}_0,\mathbf{c},\mathbf{s}\}}\mathbf{X}\|^2 > F_{1-\alpha}(2, n-3). \tag{4.20}$$

To obtain the estimates we compute

$$\begin{aligned} P_{\text{span}\{\mathbf{e}_0\}}\mathbf{X} &= \sqrt{n}\sum_{i=1}^{n} X_i \mathbf{e}_0, \\ P_{\text{span}\{\mathbf{e}_0,\mathbf{c},\mathbf{s}\}}\mathbf{X} &= \sqrt{n}\hat{\mu}\mathbf{e}_0 + \sqrt{(n/2)}\hat{A}\mathbf{c} + \sqrt{(n/2)}\hat{B}\mathbf{s}, \end{aligned} \tag{4.21}$$

where the parameters $\beta = (\hat{\mu}, \hat{A}, \hat{B})'$ are obtained by solving $W'W\beta = W'\mathbf{X}$ for β with $W = [\sqrt{n}\mathbf{e}_0, \sqrt{(n/2)}\mathbf{c}, \sqrt{(n/2)}\mathbf{s}]$.

In case the frequency of the periodic signal is unknown we must use a different approach. Fisher's test for hidden periodicities ([38]) was constructed to test the null hypothesis that $\{X_t\}$ is Gaussian white noise against the alternative that $\{X_t\}$ contains an added deterministic periodic component of unspecified frequency. In Fisher's test, the null hypothesis is rejected if the maximum of the periodogram ordinates is substantially large when compared to the sum of the ordinates. Letting $q = [(n-1)/2]$, the test statistic is defined as

$$g = \frac{\max_{1\le k\le q} I(\omega_k)}{\sum_{k=1}^{q} I(\omega_k)}. \tag{4.22}$$

Note that we do not consider the frequencies 0 and π. The p-value for a realisation of the g statistic (g^*) is given (under the strict assumption of normality) by

$$P(g \ge g^*) = 1 - \sum_{j=0}^{q} (-1)^j \binom{q}{j} (1 - jg^*)_+^{q-1}, \tag{4.23}$$

where $x_+ = \max(x, 0)$. If this probability is less than α we can reject the null hypothesis at level α [19].

Before moving to robust periodicity detection, recall that the periodogram is an unbiased but not a consistent estimator, *i.e.* its variance does not approach zero as the sample size approaches infinity. The consistency can be corrected in many ways but the approaches rely on smoothing of the estimates. Unfortunately, the time series present in gene microarray experiments are usually of length 20 to 30 the most, so these smoothing approaches are not feasible here. We therefore modify and robustify the periodogram and Fisher's test using other approaches, as introduced in the next section.

4.3 Robust spectrum estimation and periodicity detection

We now consider robust modifications of the periodicity detection methods presented in the previous section. We consider both the detection of unknown and fixed frequency periodic signals. In the case of uniform sampling, there are several modifications of the periodogram that can be considered in periodicity detection, two of which are introduced. Methods that deal with non-uniform sampling are also introduced.

4.3.1 Rank based approach

A robust rank based spectral estimator was introduced in [98] that can, to a good degree, reject outliers in data and handle missing samples. The estimator uses ranks of the samples instead of the actual time series values in estimating the autocorrelation function, which relates closely to the autocovariance function in Equation (4.11). The robust autocorrelation function can thus be used in the estimation of the spectra of signals. Rank order filters and statistics have been previously used for example in image processing [133], multirate signal processing [4] and statistical testing [141] and are known to possess robust characteristics. Samples that have inconsistent values when compared to other data points usually bias the approaches based on minimising square sums, but this bias is much less pronounced if ranks are used instead.

In Publication-II we introduced a modification of Fisher's test for the detection of hidden periodicities that uses the rank based spectral estimator described in [98]. We first here review the spectral estimator and the modification of Fisher's test after which we show the key results for the detector. From this point on $\{X_t\}$ is assumed to be real. In addition, note that the autocorrelation function $\rho_X(k)$ is related to the autocovariance function through $\rho_X(k) = \frac{\gamma_X(k)}{\sigma_X^2}$ where $\sigma_X^2 = E[X_t - \mu_X]$ is the variance of X_t and $\{X_t\}$ is stationary. We can use the autocorrelation function in the correlogram, since the autocorrelation function is just a shifted and scaled version of the autocovariance function. We therefore consider spectral estimators that are of the form

$$\tilde{S}(\omega) = \sum_{k=-L}^{L} \tilde{\rho}_X(k) \exp(-i\omega k), \tag{4.24}$$

where $\tilde{\rho}_X(k)$ is the estimate of the autocorrelation function between $\{X_t\}$ and $\{X_{t+k}\}$ and L is the maximum lag to be considered (for the rank based method $L \leq n-2$ as in Publication-II). The rank based autocorrelation function estimate is defined as

$$\tilde{\rho}_X(k) = \frac{1}{C\tilde{\sigma}^2} \sum_{t \in I_k} \left(R_x(t) - \tilde{\mu}_X\right) \left(R'_x(t) - \tilde{\mu}_X\right), \tag{4.25}$$

where I_k is the set of indices t for which both x_t and x_{t+k} exist (in case there are missing values), $K_k = |I_k|$ (and assuming $K_k \geq 2$), C is a normalisation factor, $\tilde{\sigma}^2 = \frac{K_k^2 - 1}{12}$ can be shown to be the variance of the rank sequence, $\tilde{\mu} = \frac{K_k+1}{2}$ is the mean of the rank sequence, $R_x(t)$ denotes the rank of x_t in the set $S = \{x_t : t \in I_k\}$ and $R'_x(t)$ the rank of x_{t+k} in $S' = \{x_{t+k} : t \in I_k\}$. C can be chosen in multiple ways. If we choose $C = K_k$ we get the unbiased estimator. However, by choosing $C = n$ yields a spectral estimate analogous to the periodogram and $C = n$ is therefore used.

It should be noted that in Publication-II we have used the term correlation coefficient for the autocorrelation function (as defined here) and the term autocorrelation function for the autocovariance function that has not been mean subtracted, as sometimes defined in signal processing applications. The definitions are equivalent for mean subtracted processes.

Since the robust rank based autocorrelation estimator is analogous to the classical autocorrelation estimator, it can readily be used in robustly estimating spectra and detecting periodic time series, as will be illustrated shortly. Other good properties include straightforward calculation (no iterations needed) and ease of implementation. Additional discussion and illustrations on how the rank based approach performs as a spectral estimator are given in [98] and Publication-II.

To combine the Fisher's test and the rank based spectral estimator, we proposed the following test statistic in Publication-II ($q = [(n-1)/2]$)

$$g = \frac{\max_{1 \le k \le q} |\tilde{S}(\omega_k)|}{\Sigma_{k=1}^{q} |\tilde{S}(\omega_k)|}, \tag{4.26}$$

assuming a similar (without the location term μ) model as in Equation (4.12) and the null hypothesis (4.13). The absolute value of $\tilde{S}$ is used, since, unlike the periodogram, $\tilde{S}$ is not guaranteed to be non-negative. For this modified test statistic the null hypothesis distribution is not readily available in an analytical form, so two different p-value evaluation methods were considered in Publication-II, namely Monte Carlo simulation and permutation tests, to be explained shortly. Due to not using an analytical distribution, different interpolation schemes were also considered in Publication-II. That is, in the stage where the robust autocorrelation function estimate is Fourier transformed, frequencies were sampled more densely than just at the harmonic frequencies. It was found that interpolation of the spectrum for example to twice the density of the original yielded somewhat better results, although interpolation does not actually create any new information. This could be due to some frequencies of periodicity being between two harmonic frequencies, in which case interpolation could help bring about these cases.

Before elobarating on the p-value computation, an important property of the rank based periodicity detection method is emphasised: due to the g statistic, as given in (4.26), being dependent on $\{X_t\}$ only through the ranks, it is distribution free (see Publication-II and [103] for further details). It follows that for each n, regardless of the type of noise, the g statistic has exactly the same null distribution as long as the noise term is continuous and i.i.d. This important feature is utilised in the Monte Carlo simulation type p-value computation in the way that different noise types do not need to be considered separately; it suffices to simulate the null hypothesis distribution for one i.i.d. noise type only, for example Gaussian white noise. This null hypothesis distribution can then be used regardless of the noise type present in the data.

The Monte Carlo approach of estimating p-values to assess periodicity in time series is performed by simulating a large set of time series from any null distribution whose samples are i.i.d. and computing the g statistic for each of these series. If we denote the set of g statistics evaluated this way as G, a p-value for a time series of interest can in the simplest form be approximated by

$$p^* = P(g \geq g^*) \approx \frac{|\{g \geq g^* : g \in G\}|}{|G|}, \tag{4.27}$$

where g^* is the realisation of the g statistic for the time series under study. Other more novel approaches like kernel density estimation methods can also be used to form the null hypothesis distribution based on which p-values can be estimated using numerical integration.

Permutation tests, which exist for any test statistic (regardless of whether or not the distribution of the test statistic is known) can also be used in evaluating the p-values [47]. The idea is relatively simple:

1. Evaluate the g statistic on the original time series.

2. Randomly permute the time series P times and for each permutation $\pi_j,\ j = 1, \ldots, P$ evaluate the g statistic to obtain g^j. Usually $P \ll n!$ and in the range of hundreds or thousands. For small n it may also be feasible to consider all the permutations.

3. Based on the original g statistic and the sample generated in point 2, estimate the the p-value similarly as in the Monte Carlo case, see Equation (4.27).

The procedure is then repeated for all the time series in the set. To be able to apply permutation tests, it suffices that the time series samples are exchangeable under the null hypothesis [47]. A sequence of random variables $\{X_t,\ t = 1, \ldots, n\}$ is exchangeable, if the joint distribution of $X_{\pi_1}, \ldots, X_{\pi_n}$ is the same as that of the original sequence for all permutations π. Under the null hypothesis, the elements of the time series are assumed to be i.i.d. and therefore exchangeable.

Although permutation tests are theoretically exact and nonparametric and can thus be used without knowledge of the exact distribution of the data at hand, the Monte Carlo simulation approach is more feasible here. This is because the Monte Carlo-approach involves less computation here and is also exact in case we are using the rank based detector.

After obtaining p-values (depicting distrust in the null hypothesis) for all the genes in the dataset, the problem of choosing a proper cut off point for periodic genes arises. A strict significance level, *e.g.* $\alpha = 0.05$, means that under the null hypothesis there is a 5% chance of accepting a false positive (when considering one time series). To control the number of false positives

in the case of multiple time series, several multiple testing correction methods have been proposed (see [32]). The Benjamini-Hochberg method [10], which controls the false discovery rate (FDR), is an easy-to-apply procedure. The procedure can be used on a population of p-values (see Publication-II for details) and gives a cutoff point that is adaptive to the data.

We now briefly also consider the case where knowledge on the frequency of periodicity is available. A modified g statistic is defined by

$$g' = \frac{|\tilde{S}(\omega')|}{\sum_{k=1}^{q} |\tilde{S}(\omega_k)|}, \tag{4.28}$$

where ω' is a chosen frequency but $S(\omega')$ is not necessarily the maximum of the spectral estimate. This modified test is also distribution free and both methods for p-value estimation for the time series can be used, the Monte Carlo simulations or permutation tests.

Comparison of the performance of the rank based method to other periodicity detection methods is given in the next subsection, after introducing the regression based approaches. The rank based method was also applied to several publicly available microarray data sets in Publication-II.

4.3.2 Regression based approach

What is usually the case in cell level high throughput studies, like gene microarray experiments, is that the measurements are conducted non-uniformly in time. The reasons for this are many; the optimal experimental design may be based on non-uniform sampling or it might be too expensive or otherwise unfeasible to conduct the experiments at constant time intervals. In addition, the biologists conducting the experiment might not realise the good computational properties of uniform sampling. Although the rank based detector, which was intoduced in the previous subsection, is insensitive to heavy contamination of outliers, missing values, short time series length and nonlinear distortions, it cannot be easily modified to handle non-uniform sampling. Thus, alternative approaches are needed.

In this subsection, we consider a regression based formulation of the periodogram and its use in periodicity detection, as presented in Publication-III and Publication-IV. We first note that a scaled version of the model in (4.8) can be given in matrix form as

$$\mathbf{x} = \begin{bmatrix} 1 & & & 1 \\ 1 & & & -1 \\ \vdots & A_1 & A_2 & \vdots \\ 1 & & & (-1)^{n-1} \end{bmatrix} \begin{bmatrix} a_0 \\ \mathbf{a_1} \\ \mathbf{a_2} \\ a_{n/2} \end{bmatrix}, \tag{4.29}$$

where $\mathbf{x}$ is the measured time series and omitting $a_{n/2}$ (and the last column

in the matrix) if n is odd. Matrices A_1 and A_2 are then

$$A_1 = \begin{bmatrix} \cos(\omega_1 t_0) & \cdots & \cos(\omega_q t_0) \\ \cos(\omega_1 t_1) & \cdots & \cos(\omega_q t_1) \\ \vdots & & \vdots \\ \cos(\omega_1 t_{n-1}) & \cdots & \cos(\omega_q t_{n-1}) \end{bmatrix} \tag{4.30}$$

$$A_2 = \begin{bmatrix} \sin(\omega_1 t_0) & \cdots & \sin(\omega_q t_0) \\ \sin(\omega_1 t_1) & \cdots & \sin(\omega_q t_1) \\ \vdots & & \vdots \\ \sin(\omega_1 t_{n-1}) & \cdots & \sin(\omega_q t_{n-1}) \end{bmatrix}. \tag{4.31}$$

By scaling $t.$ to integers (still assuming uniform sampling) and using ordinary least squares regression (inverse of the model matrix in this case) to solve for $\begin{bmatrix} a_0 & \mathbf{a_1}^T & \mathbf{a_2}^T & a_{n/2} \end{bmatrix}^T$, we get the periodogram ordinates [101] by

$$\begin{aligned} I(\omega_0) &= n(\hat{a}_0)^2, \\ I(\omega_k) &= \frac{n}{4}\hat{a}_{1k}^2 + \frac{n}{4}\hat{a}_{2k}^2, k = 1, \ldots, q, \\ I(\omega_{n/2}) &= n(\hat{a}_{n/2})^2. \end{aligned} \tag{4.32}$$

This regression based formulation provides a convenient way of introducing non-uniform time indices and also replacing least squares minimisation with robust alternatives.

When considering non-uniform sampling, the harmonic Fourier frequencies are not well defined anymore. Therefore, to imitate uniform sampling, we scale the measurement times so that the first time point scales to zero and the last time point scales to $n-1$, where n is the number of samples in the time series. These time indices are then used in the model matrices in (4.30-4.31). We also estimate what the the sampling time would be on average, as if the sampling was performed uniformly. If we denote the original measurement times as a vector $\boldsymbol{\tau}$, new scaled indices are computed in the following way

$$\mathbf{t} = \frac{(\boldsymbol{\tau} - \tau_0 \cdot \mathbf{1})(n-1)}{\tau_{n-1} - \tau_0}. \tag{4.33}$$

The average sampling time is

$$T_s = \frac{1}{n}(\tau_{n-1} - \tau_0), \tag{4.34}$$

which can be used to make the connection between a real frequency and the hypothetical Fourier frequencies. By letting the sampling be non-uniform, the column vectors in the model matrix are no longer orthogonal and therefore the model matrix may under severely bad conditions become singular. This is, however, highly unlikely in practice.

Replacing least squares estimation with robust alternatives, like maximum likelihood-type robust regression methods (M-estimation), the high dimensionality (non-convergence) in (4.29) becomes a problem. Therefore the dimension of the problem should be reduced and the frequencies fit one at a time. In [119] the authors point out that when fitting sinusoidals one frequency at a time, it is imperative to fit to the residuals of the last fit to avoid overfitting, *i.e.* first choose the order the frequencies are fit and then after fitting the sinusoidals at one frequency, use the residual signal in the next fit. The order in which the frequencies are fit can be chosen for example based on an initial spectral estimate where no residual fitting is performed, the strongest component getting highest priority. The reduced model is given as

$$\mathbf{x} = X(\omega)\mathbf{b} + \mathbf{e}, \tag{4.35}$$

where $\mathbf{e}$, the residual, is used as the $\mathbf{x}$ of the following fitting and $\mathbf{b}$ can be used in estimating the spectrum, corresponding to the amplitudes of the fitted sine and cosine terms (like $\hat{a}$ in Equation (4.32)). Matrix $X(\omega)$ is now

$$X(\omega) = \begin{bmatrix} 1 & \cos(\omega t_0) & \sin(\omega t_0) \\ 1 & \cos(\omega t_1) & \sin(\omega t_1) \\ \vdots & \vdots & \vdots \\ 1 & \cos(\omega t_{n-1}) & \sin(\omega t_{n-1}) \end{bmatrix}, \tag{4.36}$$

This spectral estimate can then be used in Fisher's test similarly as the rank based estimator was used in Equation (4.26). The biggest difference to the rank based estimator is that we should not use Monte Carlo simulations in p-value computation since the regression based approach is not guaranteed to be distribution free.

Should prior information on the frequency of periodicity exist, we can also test for periodicity at just one frequency. This is also computationally more feasible since we need to estimate only one spectral component. We define the test statistic as

$$g_m = \hat{b}_{1c}^2 + \hat{b}_{2c}^2, \tag{4.37}$$

where $\hat{b}_{1c}^2$ and $\hat{b}_{2c}^2$ are the estimates of the cosine and sine terms corresponding to the chosen frequency in Equation (4.35). Permutation tests are then applied to produce p-values depicting distrust in the null hypothesis. It should also be noted that we have discarded the normalising term (previously sum of all the spectral components) in Equation 4.37, because firstly for non-Fourier frequencies the normalising term does not make sense anymore, secondly the gains in computational time are significant (the sinusoidals are fitted for one frequency only) and lastly because the term is not necessarily needed (permutation tests are used).

Before considering the performance of the regression based periodicity detection methodology on simulated and real data, we briefly review the

idea of M-estimation. As was shown in Publication-III, M-estimators, more specifically the Tukey's bisquare, provide good robust characteristics on a wide range of distributions and were shown to be the most viable option among the group of considered regression based estimators. Since it is reasonable to assume that the predictors (time points) are fixed and nonrandom, M-estimators can reject outliers (in the measurement signal) to a good degree. In the following we follow the presentation in [84] and assume a linear data model

$$\mathbf{x} = A\mathbf{b} + \mathbf{e}, \tag{4.38}$$

where $\mathbf{x}$ is the vector of measured data, A is the $n \times p$ ($p \leq n$) predictor matrix (nonrandom and known in what follows), $\mathbf{b}$ is the vector of unknown parameters to be estimated and $\mathbf{e}$ is a vector of random variables (the error terms). The first column in A is usually a constant $\mathbf{1}$ to take location (intercept) into account. If e_k has density

$$\frac{1}{\sigma} f_0 \left(\frac{e_k}{\sigma} \right), \tag{4.39}$$

where σ is a scale parameter, then for independent x_k the density function is

$$\frac{1}{\sigma} f_0 \left(\frac{x_k - A_k' \mathbf{b}}{\sigma} \right), \tag{4.40}$$

where A_k' is the kth row of A. The likelihood function for $\mathbf{b}$ assuming a fixed σ is

$$L(\mathbf{b}) = \frac{1}{\sigma^n} \prod_{k=1}^{n} f_0 \left(\frac{x_k - A_k' \mathbf{b}}{\sigma} \right), \tag{4.41}$$

and computing the maximum likelihood estimate corresponds to maximising (4.41). Thus, we need to find $\widehat{\mathbf{b}}$ such that

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}} \frac{1}{n} \sum_{k=1}^{n} \rho_0 \left(\frac{r_k(\mathbf{b})}{\sigma} \right) + \log \sigma, \tag{4.42}$$

where $\rho_0 = -\log f_0$ and $r_k(\mathbf{b}) = x_k - A_k' \mathbf{b}$. Assuming that σ is known, differentiating with respect to $\mathbf{b}$ and solving for the stationary point can be useful in the search for the solution of Equation 4.42. Thus, we get the equation

$$\sum_{k=1}^{n} \psi_0 \left(\frac{r_k(\widehat{\mathbf{b}})}{\sigma} \right) A_k = \mathbf{0}, \tag{4.43}$$

where $\psi_0 = \rho_0' = -f_0'/f_0$. It should be noted that the zero point of the derivative does not necessarily guarantee a unique solution to Equation (4.42). In case f_0 is the standard normal density function, then there exists

a unique $\widehat{\mathbf{b}}$, which is also known as the least squares (LS) estimate. If f_0 is the Laplacian density, then $\widehat{\mathbf{b}}$ satisfies

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}} \sum_{k=1}^{n} |r_k(\mathbf{b})|\,, \tag{4.44}$$

also known as the L1 estimate. It is worth noting that the L1 estimate is the regression equivalent of the median location estimate and that the L1 estimate is independent of any scale. No explicit expression for the L1 estimate exists, unlike for the LS estimate, but fast computational algorithms implementing the L1 estimate do [84].

In general, regression M-estimates are solutions $\widehat{\mathbf{b}}$ to

$$\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}} \sum_{k=1}^{n} \rho\left(\frac{r_k(\mathbf{b})}{\widehat{\sigma}}\right), \tag{4.45}$$

where $\widehat{\sigma}$ is an error scale estimate. By differentiating (4.45) with respect to $\mathbf{b}$ we get the equation

$$\frac{1}{\widehat{\sigma}} \sum_{k=1}^{n} \psi\left(\frac{r_k(\widehat{\mathbf{b}})}{\widehat{\sigma}}\right) A_k = \mathbf{0}, \tag{4.46}$$

where $\psi = \rho'$ and (4.46) does not actually need to be the estimating equation of a MLE. In fact, if the underlying error distribution is unknown, as it usually is, our aim is to find a function ψ that produces good parameter estimates on a wide range of distributions, even if there is no real distribution corresponding to ψ. Henceforth solutions to (4.46) with monotone ψ are called monotone M-estimates whereas solutions corresponding to nonmonotone ψ are called redescending M-estimates. A favorable property of monotone estimates is that all solutions of (4.46) are also solutions of (4.45). On the other hand, the redescending estimates may have multiple solutions corresponding to local minima; this does not happen with monotone estimates. The reason why redescending estimates are used is that the redescending M-estimates can yield a better trade-off between robustness and efficiency and can reject outliers altogether [84]. Our main interest is thus in redescending ψ, especially in the Tukey's bisquare, defined by

$$\psi(x) = x\left[1 - \left(\frac{x}{k}\right)^2\right]^2 I(|x| \leq k), \tag{4.47}$$

where $I(x) = 1$ when x is true and zero otherwise. It can be shown that the selection $k = 4.68$ yields a 95% efficiency on the normal distribution [84]. The bisquare function is everywhere differentiable and vanishes outside $[-k, k]$, thus zero weighting outlying samples. It can also be shown

that M-estimates with vanishing ψ outside an interval are not MLEs of any distribution [84].

Computation of redescending M-estimates is highly dependent of choosing a good starting point, which is usually provided by a monotone estimate (the L1 for example). For smooth ψ we can solve Equation (4.46) using an iterative algorithm called iteratively reweighted least squares (IRWLS). Defining [84]

$$W(x) = \begin{cases} \psi(x)/x & \text{if} \quad x \neq 0 \\ \psi'(0) & \text{if} \quad x = 0 \end{cases}, \tag{4.48}$$

(where L'Hôpital's rule was used for $x = 0$) we can rewrite (4.46) as

$$\sum_{k=1}^{n} w_k r_k A_k = \sum_{k=1}^{n} w_k A_k (x_k - A_k' \widehat{\mathbf{b}}) = \mathbf{0}, \tag{4.49}$$

where $w_k = W(r_k/\widehat{\sigma})$ and omitting the multiplier $\widehat{\sigma}$, since it has no effect on finding the zero point. The equations can also be given in the matrix form as

$$(WA)'(\mathbf{x} - A\mathbf{b}) = \mathbf{0}, \tag{4.50}$$

where W is a diagonal matrix whose diagonal elements are the weights. These are called weighted normal equations and if w_ks were known, could be solved by applying LS to $\sqrt{w_k}x_k$ and $\sqrt{w_k}A_k$. This, however, is not the case, since w_ks depend on the data. The procedure for IRWLS with a tolerance parameter ε is

1. As an initial estimate of $\widehat{\mathbf{b}}$, compute the L1.

2. Estimate σ as the normalised median of residuals of the L1-fit: $\widehat{\sigma} = \frac{1}{0.675}\text{Med}_k(|r_k| \mid r_k \neq 0)$, considering only nonnull residuals to prevent underestimating σ.

3. For $j = 0, 1, 2, \ldots$:

   (a) Given $\widehat{\mathbf{b}}_j$, for $k = 1, \ldots, n$ compute $r_{k,j} = x_k - A_k'\widehat{\mathbf{b}}_j$ and $w_{k,j} = W(r_{k,j}/\widehat{\sigma})$.

   (b) Compute $\widehat{\mathbf{b}}_{j+1}$ by solving

   $$\sum_{k=1}^{n} w_{k,j} A_k (x_k - A_k'\widehat{\mathbf{b}}) = \mathbf{0}. \tag{4.51}$$

   for $\widehat{\mathbf{b}}$

4. Stop when $\max_i(|r_{k,j} - r_{k,j+1}|)/\widehat{\sigma} < \varepsilon$.

The algorithm converges if $W(x)$ is nonincreasing for $x > 0$ and the solution is unique for monotone ψ [84].

In Publication-III, several regression based estimators were considered (least squares, bisquare M-estimator, least trimmed squares [107], minimum covariance determinant [108]) and compared also to the rank based method (Publication-II) in a simulation study where first the frequency of periodicity was unknown and sampling was non-uniform with 20 samples. In these simulations, 300 time series were generated from the null hypothesis (noise) distributions and 300 were generated from the alternative hypothesis (periodic signal plus noise) distributions. The different tests were then applied to the data. With knowledge of the ground truth, it is possible to assess the amount of true and false positives and negatives and visualise the performances by plotting receiver operating characteristic (ROC) curves for the different methods (see Publication-III). Based on the simulation results and ROC curves, which plot sensitivity (true positive rate) versus 1-specificity (false positive rate), the M-estimator was chosen as the best representative of the regression based methods. The method was then compared to a recently published Bayesian detector [3], in addition to the rank based method, in a simulation where the frequency of periodicity is approximately known. By this we mean that we fixed the frequency of periodicity in the simulated signals but deliberately chose the frequency, at which periodicity is to be detected, approximately 10% off the true value. According to the authors, the Bayesian detector should be robust to distributional changes. In Figure 4.1 we can see that in case the noise is purely Gaussian, the Bayesian detector is superior since it uses a prior centered around the chosen frequency (which is 10% off the true value) to be detected and does not assume a strict value. The left hand column data in Figure 4.1 is sampled according to a real non-uniformly sampled microarray data set measured from *Mytilus californianus* (introduced shortly) and the right hand column corresponds to an artificially deteriorated version of the left hand column sampling. The M-estimator and rank based methods perform relatively well even though they assume a strict value for the periodicity (deliberately chosen wrongly). Going down Figure 4.1, we see the effect of added outliers. The performance of the Bayesian detector degrades in an alarming manner whereas for the robust M-estimator and the rank based method there is no huge degradation. Since microarray data is known to be noisy and the noise characteristics are not guaranteed to be well known, methods that cannot reject outliers should be used with caution. It can also be seen that the rank based detector performs relatively well even though it is not designed to handle non-uniform sampling.

The introduced regression framework can also be readily used in the case of uniform sampling, as considered in Publication-IV. As noted previously, by utilising bisquare M-estimation it is possible to obtain 95% efficiency on the normal distribution and also reject outliers in data. This further sug-

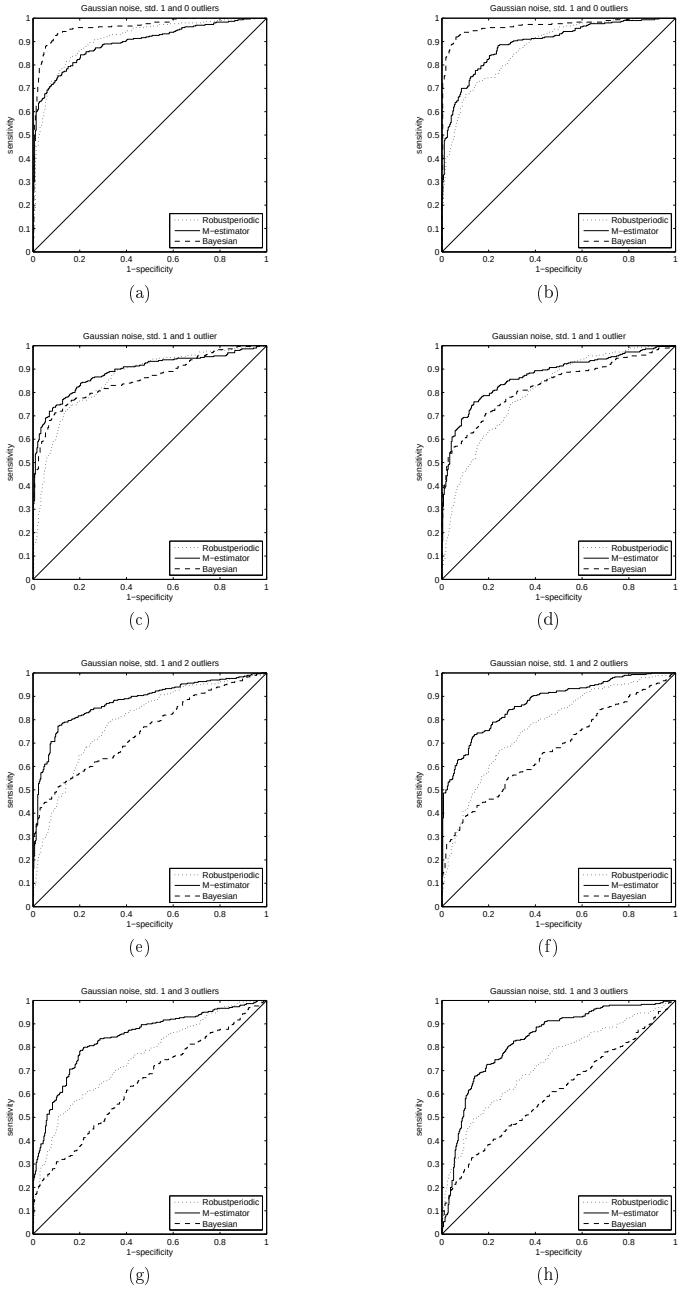


Figure 4.1: ROC curves for three methods, the rank based method (Robustperiodic), the bisquare M-estimate method (M-estimator) and the Bayesian method (Bayesian). The non-uniform sampling in the left hand column data is of a real microarray data set and the sampling in the right hand column data is a deteriorated version of the left hand column. On the first row the noise present in the signals is Gaussian white noise with standard deviation 1. On rows two to four every signal has one, two or three outliers, correspondingly, added to random locations in the signal.

gests that the analytical null hypothesis distribution of Fisher's test could be used in conjunction with the M-estimator in case the data is uniformly sampled. In Figure 4.3 we can see the theoretical null hypothesis distribution of Fisher's test (solid line), the estimate of the null hypothesis distribution when using periodogram based g statistic values (solid line with dots) and the estimates of the null hypothesis distribution when using M-estimation (dashed line for Gaussian data and dotted line for Gaussian data and outliers). The distribution estimates were obtained from the g statistic populations by using standard kernel density estimation methods in Matlab. The simulated null hypothesis time series (10000) were drawn from a centered unity variance Gaussian distribution and in the outlier case, 10% of the time series points were replaced by values in the interval $\pm[5, 6]$. The estimates are shown both for time series length 20 (a)) and 100 (b). As we can see, approximate p-values can be obtained for the robust test using the analytical null hypothesis distribution, yielding a very quick to implement test for detecting periodicity.

Figure (4.2) shows how the unmodified Fisher's test, the rank based approach and the M-estimator based test performed in a simulation study with uniform sampling. For the simulation, 300 time series were generated from the null hypothesis distribution and 300 from the alternative distribution. Gaussian distribution with standard deviation 0.75 was used as the null hypothesis distribution in the ROC curve of Figure (4.2) (a). In (b) to (d) outliers of amplitude $\pm(5 \dots 6)$ were randomly placed, one per time series in (b), two in (c) and three in (d). The alternative distribution was otherwise similar to the null hypothesis distribution but a sinusoidal of amplitude $\sqrt{2}$ and random phase and frequency was also added to represent a periodic signal. The time series length in the simulation was set to 20. The ROC curves illustrate the robust properties of both the rank based test and the M-estimator based test. The rank based test performs the best in the case of no outliers (a) but the M-estimator retains its performance better when outliers are added (b-c). In (d) where there are already 3 outliers (out of 20 samples), the ROCs of all the methods are rather close to the chance diagonal.

In addition to the simulation results, we also tested the M-estimate regression based method on gene microarray time series data measured from the mussel *Mytilus californianus* (available on the Internet in ArrayExpress, accession number E-TABM-287). The data was obtained by measuring gene expression over several days with non-uniform sampling. We hypothesised that the gene expression of the seaside mussel periodic at the circadian cycle and tidal rhythm would have a connection to the cell cycle or other known biological factors. However, the best ranked genes that were found periodic at the 24 hour cycle were not found, according to gene set enrichment analysis [117], to have a significant connection to 245 gene sets capturing biological prior knowledge, defined by their shared participation in a spe-

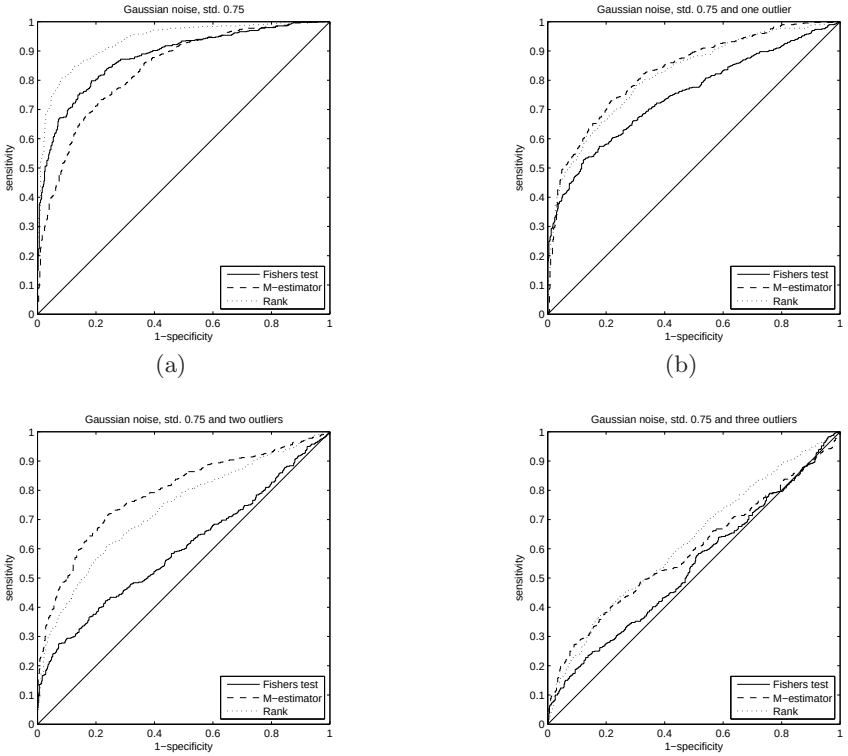
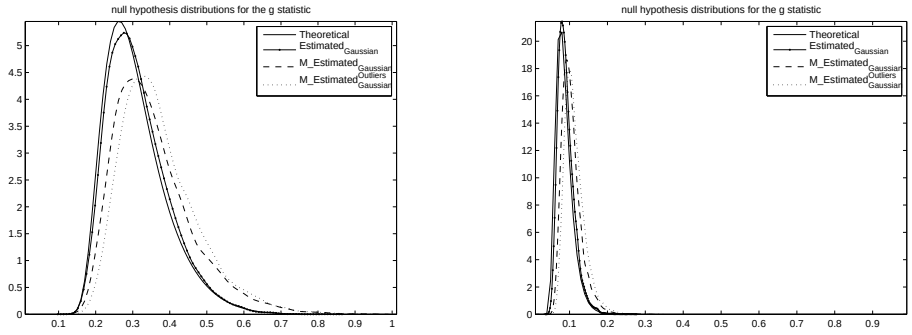


Figure 4.2: Receiver operating characteristic curves for Fisher's test (Fishers test) and two robust modifications of the test. The dashed line corresponds to the M-estimator based test (M-estimator) and the dotted line corresponds to the rank based test (Rank). The noise type is Gaussian with standard deviation 0.75. In addition, one outlier per time series is present in (b), two outliers in (c) and three outliers in (d).

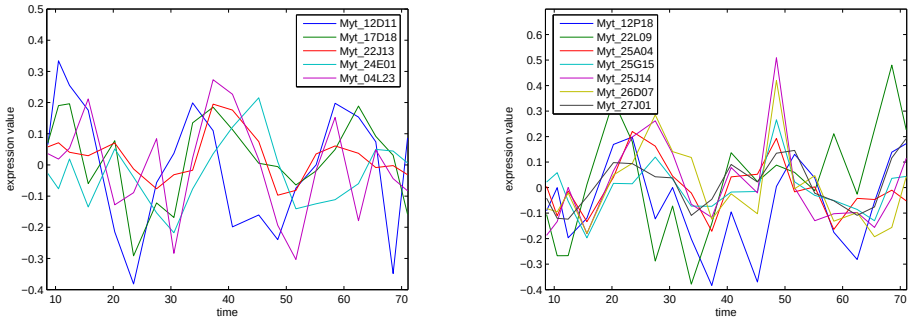
cific biological process in the Gene Ontology database. The 12 best ranked genes periodic in the 24 hour cycle can be seen in Figure 4.4. These genes were expressed at small amplitudes, which could be a reason they have not been studied much before and do not appear in data sets capturing prior biological relevance. Future studies include finding out the cycle frequencies at which the periodic genes that do have a biological relevance are expressed at.

The robust periodicity detection methods presented in this chapter provide a wide range of robust options to detect periodic events in sequence data that can be either uniformly or nonuniformly sampled and have noise characteristics of unknown distribution. As the results in Publication-II, Publication-III and Publication-IV show, detection at both priorly known and unkown frequencies have been implemented in an efficient way.



(a) (b)

Figure 4.3: Analytical (solid line), estimated (solid line with dots) and bisquare M-estimate estimated (dashed line for Gaussian data and dotted line for Gaussian data and outliers) null hypothesis distributions for Fisher's test. Time series length was set to 20 in (a) 100 in (b).



(a) (b)

Figure 4.4: The twelve best mussel gene expression data time series periodic at the 24 hour cycle.

Chapter 5

Detection of pathogenic mutation prone locations in protein sequences

The purpose of this chapter is to consider actual protein, *i.e.* gene end product, measurements and introduce methods for the detection of pathogenic mutation prone locations in the protein sequences.

Proteins play a crucial role in most biological processes [11,92]. Proteins have many different physiological functions, for example as catalysing enzymes; binding, storaging and transporting molecules; structure supporting molecules; antibodies; neurotransmitters; receptors and also as transcription factors for promoting or suppressing gene expression. Proteins are made up of building blocks known as amino acids. There are a total of 20 different amino acids that are in general used in proteins and each amino acid is coded by three nucleotides (many to one mapping actually). Some of the amino acids (valine, leucine and isoleucine) are termed hydrophobic, some hydrophilic (lysine, arginine and histidine are basic hydrophilic and aspartic acid and glutamic acid are acidic hydrophilic in nature) and others are more or less neutral in this sense. Hydrophobic amino acids tend to cluster in the inside region of a protein and away from the water surface, thus significantly stabilising the protein structure.

Proteins have a total of four levels of structure [11, 92], as illustrated in Figure 5.1. The primary structure refers simply to the amino acid sequence of the protein. The secondary structure refers to the simple three-dimensional structures of the amino acid chain. These arrangements can be for example helices or pleated sheets, as shown in Figure 5.1. Of the different structures, the tertiary structure is the most interesting in the context of this work. The tertiary structure refers to the overall three-dimensional arrangement of all atoms in a protein as opposed to the secondary structure, which only refers to the spatial arrangement of amino acid residues

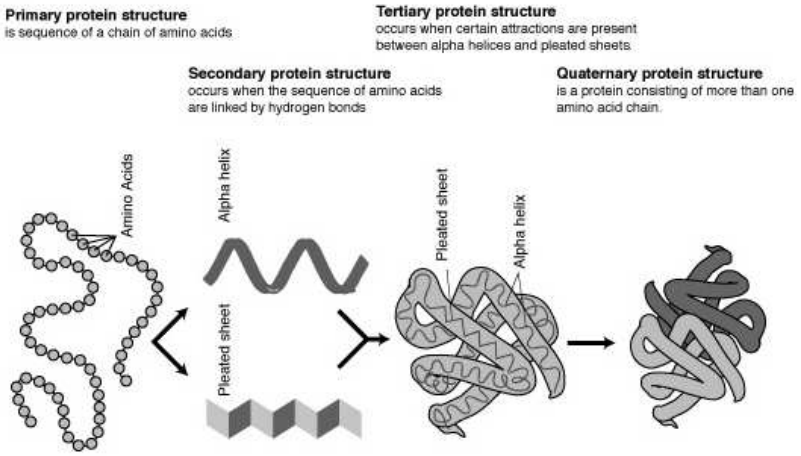


Figure 5.1: The four levels of protein structure.

that are adjacent in the primary structure. A protein spontaneously folds into a three-dimensional shape with a distinct inside and outside. Typically in an aqueous environment the interior of the protein is composed mainly of the hydrophobic amino acids present in the sequence that exclude water and the exterior mainly of hydrophilic amino acids, which increase the water solubility of the protein. Finally, the quaternary structure refers to how in some proteins that have separate amino acid chain subunits, the subunits arrange further into three-dimensional complexes (see Figure 5.1). The correct function of a protein is heavily dependent on the three-dimensional structure but the relation between *e.g.* the primary and tertiary structure is not extensively known. As according to [137], a major concern of biochemical research is to figure out the relationship between sequence embedded information and folding behaviour of proteins. This concern is especially pronounced in areas such as sequence and three-dimensional structure based functional predictions and folding mechanisms [34, 109].

The hydrophobicities and hydrophilicities of the different amino acids in the sequence play an important role especially in the formation of the tertiary (folding) structure of proteins. Later in this chapter we study mutations in protein amino acid sequences. We assume that the mutations that strongly change the hydropathy of certain critical parts (assessed by our algorithm) of the sequence will affect especially the tertiary structure of the protein and thus alter its operation to an unwanted direction. Previously, effects of pathogenic mutations have been predicted *e.g.* by observing the conservation of amino acid residue sequences in different protein families [80]. The methods proposed in this thesis and in [80] can complement each other in a useful manner and remain as a topic for future studies.

Only specific amino acid sequences are good for generating functional

proteins. However, statistically, there are only weak differences between random permutations of a protein sequence and the original sequence [137]. Especially, protein hydropathy sequences have been shown to differ from random sequences [55, 99]. According to [99], hydropathy is the only chemicophysical property of proteins that shows statistically significant nonrandomness. Further, prediction of 3D-structures of proteins based on sequence information alone may be impossible but hydrophobicity patterns have been shown to correlate with 3D-structures [42]. Hydrophobicity is believed to be linked to stabilising protein structures and points to the existence of specific constraints in the arrangement of hydrophobic and hydrophilic patterns along chains, leading to foldable structures.

It should be noted that hydropathy values are usually estimated from the amino acid sequence. A related measured quantity is the solvent accessibility of a protein. Solvent accessibility is closely linked to protein hydropathy, since hydrophilic parts of the sequence tend to be closer to the surface of the protein and are thus more accessible, whereas hydrophobic parts tend to be more inside of the protein or, for example, buried inside the cell membrane [2, 68]. Kyte-Doolittle method [61] is often used in the estimation of hydropathy values. In estimating hydropathy, each amino acid is first given a predetermined value between -4.6 and 4.6. The sequence is then passed through an averaging filter of length 9, a value suggested in [61], to take into account the interactions between the linked amino acids.

Protein hydropathy sequences have been studied extensively using a computational technique called recurrency quantification analysis (RQA), see for example [41, 42, 44, 99, 135–138]. RQA is based on quantifying the important characteristics of a plot known as recurrence plot. Recurrencies, on the other hand, are simply points that repeat itself [137]. In the context of chaotic systems, recurrencies can be indicative of unstable periodic orbits, which represent an element of order within chaos [16]. Recurrence plots and RQA are further introduced in the following subsections.

Recurrences are inherent in dynamical systems, whereas for random systems recurrences occur by chance alone [127]. This was visualised in [126], where a chaotic system known as Hénon strange attractor was simulated. The Hénon system is basically composed of two interconnected variables with nonlinear feedback. Plots of the simulated variables seem rather nondeterministic, but the recurrence plots of the variables show structures implying determinism that is not present in the recurrence plots of random permutations of the sequences. This gives a very good, although heuristic, motivation for trying to analyse protein hydropathy sequences with help of recurrence strategies.

The scale of applications where recurrence plots and RQA have been applied is wide. Applications include for example: Rhythmical physiological systems [126], surface electromyographic signals [37], characterising folding properties of chimeric sequences derived from two parental proteins [44],

discriminating between function retaining and nonfunctional mutants of β-lactamase with aid of principal component analysis [136] (principal component analysis has also been applied to protein p53 in [99]), prediction of thermophilic/mesophilic characteristics of rubredoxins [41], revealing hydrophobicity patterns in prions [138], predicting the presence of surface β-strands from amino acid sequence data [96], complexity studies [43, 85], studying chaotic systems [16, 57, 83, 111, 120], studying protein sequence-structure relationships [42], discovering hidden dynamics in epileptic electroencephalogram data [71] and numerous other studies. For more application areas and references, see [85], where the authors note that more than 1000 related references are found by the Scirus search engine.

We first review recurrence plots and related attempts to quantify the plots, including RQA. Much of the following is based on a recently published very formal 93-page study of recurrence plots and related metrics [85]. As an application, a modification of RQA is applied to real protein solvent accessibility data, as introduced in Publication-V. The method is successfully used to distinguish locations in the sequence that are more prone to pathogenic mutations than others.

5.1 Recurrence plots and statistics

Recurrence plots (RPs) were originally designed for studying dynamical systems, especially to detect hidden rhythms embedded within complex wave forms, independent of stationarity restrictions (possibly nonstationary signals) [33]. RPs provide a useful graphical representation of recurrent patterns in time or any ordered series [83].

Supposing we have a trajectory of a system, $\{\mathbf{x}_i\}_{i=1}^n$, the development of the system is then described by this series of vectors. The RP corresponding to this system is then based on the following recurrence matrix,

$$R_{i,j} = \begin{cases} 1: & \mathbf{x}_i \approx \mathbf{x}_j, \\ 0: & \mathbf{x}_i \not\approx \mathbf{x}_j, \end{cases} \quad i,j = 1,\ldots,n, \tag{5.1}$$

where $\mathbf{x}_i \approx \mathbf{x}_j$ means that the two vectors are separated by an error ε the most. This error term is essential, since most systems never recur to a formerly visited state exactly, just in the approximate sense. Recurrences are thus indicated as ones, usually black spots on white background in the plots, in the recurrence matrix and are indicative of where similar states in the underlying system occur [85].

Three examples of RPs are shown in Figure 5.2 [85], namely of (uniformly sampled) periodic motion on a circle (A), of the (uniformly sampled) chaotic Rössler system (B) [112], and of uniformly distributed I.I.D. noise (C). The

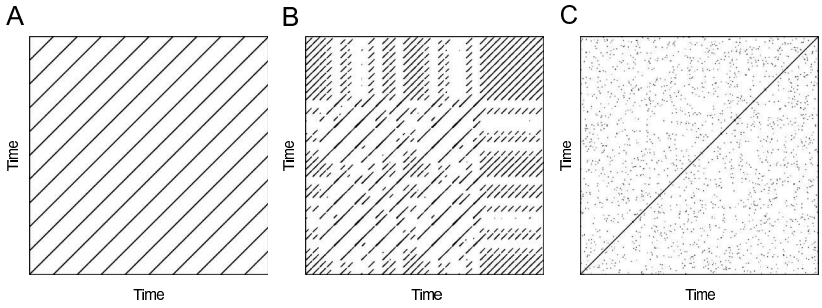


Figure 5.2: Recurrence plots of (A) a periodic signal with one frequency, (B) the chaotic Rössler system with parameters $a = b = 0.2,\ c = 5.7$ and (C) uniformly distributed noise [85].

Rössler systems is defined in terms of differential equations as

$$\begin{aligned} \frac{dx}{dt} &= -y - z \\ \frac{dy}{dt} &= x + ay \\ \frac{dz}{dt} &= b + z(x - c) \end{aligned} \quad , \tag{5.2}$$

where the parameters were chosen $a = b = 0.2$ and $c = 5.7$. A plot of the attractor can be seen in Figure 5.3. Recurrences can be observed in all the three systems, but the patterns are clearly different. In Figure 5.2 (A) the long diagonals reflect the periodic signal with the vertical distance between these lines corresponding to the period of the oscillation. In Figure 5.2 (B), for the Rössler system, the diagonals are shorter and there are vertical distances between the lines that are more irregular than for case (A). An exception is seen in the upper right corner in (B), where the rectangular patch looks like the RP of the periodic signal of (A). It is shown in [85] that this section corresponds to a nearly periodic structure on the attractor of the Rössler system, called an unstable periodic orbit (UPO). For the purely stochastic signal (C), the RP consists of mainly separate single recurrent points and next to none diagonal line structures. This leads us to the obvious conclusion that the shorter the diagonals in the RP, the less predictable the system is.

Constructing RPs

We now review the construction of recurrence plots in a more formal way. With focus on recurrences of states in dynamical systems, the recurrence plots measure recurrences of a trajectory $\mathbf{x}_i \in \mathbb{R}^d$ in d-space. If a scalar time series $y_i = y(i\Delta t)$, with $i = 1, \ldots, n$ and Δt as the sampling rate, has been measured, the phase space has to be reconstructed. Typically, the reconstruction is performed in the following way (for existing indices)

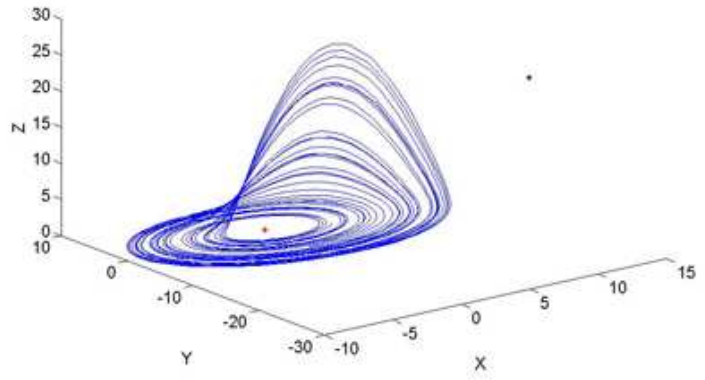


Figure 5.3: Illustration of Rössler attractor with $a = b = 0.2$ and $c = 5.7$.

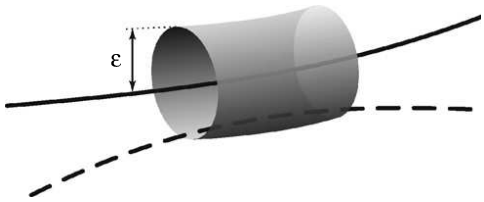


Figure 5.4: The ε-tube [85].

[85, 95, 118]

$$\widehat{\mathbf{x}}_i = \sum_{j=1}^{m} y_{i+(j-1)\tau} \mathbf{e}_j, \tag{5.3}$$

where m is the embedding dimension, τ is the integer time delay (which allows undersampling) and vectors $\mathbf{e}_j$ are unit vectors. Choosing the embedding parameters m and τ correctly is not trivial. Several rules of thumb have been proposed [42, 126] as well as methods to estimate the parameters [85]. The choice of these parameters is critical in the the reconstruction of the original d-dimensional space of trajectories. This is because only correct values yield embeddings that are guaranteed to be topologically equivalent to the original and unobserved dynamics [57,95,118]. The RP matrix is then defined as

$$R_{i,j}(\varepsilon) = u(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|), \ i, j = 1, \ldots, n, \tag{5.4}$$

where n is the number of measurement points, ε is a selected threshold, $u(\cdot)$ is the unit step function ($u(x) = 1$ for $x \geq 0$, zero otherwise) and $\|\cdot\|$ is a proper norm. The ε-neighbourhood is visualised in Figure 5.4. The recurrence plot is then obtained by plotting the recurrence matrix 5.4, plotting a black dot at (i, j) whenever $R_{i,j}$ equals to one (and white otherwise). Conventionally the point $(1, 1)$ is at the lower left corner of the plot. Since $R_{i,i}$ is trivially always one, the main diagonal is called the line of identity (LOI). If ε is

kept constant, the RP is also symmetric with respect to the main diagonal. Choosing the norm is usually a choice between the L_1-, L_2- (Euclidean norm) and L_∞-norms (maximum norm). The L_2-norm has been used a lot in the more biological publications (*e.g.* [41, 42, 44, 99, 135–138]), probably because of convention. However, it is actually possible to study some features in RPs analytically by choosing the L_∞-norm and it has therefore been used in the more theoretical works [16, 57, 83, 85].

Selection of the threshold ε is crucial in using RPs. Should ε be chosen too small, there are only few recurrent points in the RP and there is little to learn about the dynamics of the system. Of course, too big a ε leads to the situation that nearly every point is a neighbour of another and also to an effect called tangential motion. Tangential motion refers to including recurring points that are not actually recurring but just consecutive to real recurring points on the trajectory [85].

Several options for choosing the threshold have been presented in the literature. As pointed out in [57], the decision of the authors in [126] to prescribe a threshold corridor corresponding to the lower ten percent of the entire distance range is made without a comment. This cutoff is used by the authors of [126] in most of their later publications (which are many) as well. If ε is chosen based on a fixed percentage, the percentage should not exceed 10% of the mean or the maximum phase space diameter [85]. Another possibility is to choose ε so that the recurrence point density in the RPs is a constant. The advantage of this would be that comparing RPs of different systems would be easier without normalising the time series beforehand. Other possibilities for choosing ε are given in [85], including for example a noise-adaptive ε. No single correct method for choosing ε has been published so the decision remains application specific.

Patterns in RPs

RPs provide important insights into the time evolution of the trajectories of high dimensional systems [85]. Therefore, several important structures and patterns that are visible in RPs are next discussed. First, homogeneity of the patterns in RPs means that the process is fairly stationary. Second, fading to the upper left and right corners (away from the line of identity) means that the data is nonstationary and contains a trend or a drift. Third, white bands in the plot points to nonstationarities in the data and some states are either rare or far from the majority of data. Fourth, periodic and quasi-periodic patterns mean that there are cyclicities in the process. Distances between these line structures are indicative of the frequency of the period and differing distances point to quasi-periodic behaviour. Fifth, a RP with single isolated points means that the process is strongly fluctuating and is probably random. Sixth, diagonal lines parallel to the line of identity are indicative of states that evolve similarly at different locations. This can mean

that the process is deterministic or, if the diagonal lines occur beside single isolated points, chaotic. Periodic occurrance of diagonal lines can also point to unstable periodic orbits. Seventh, diagonal lines orthogonal to the line of identity means that the evolution of states is similar at different times but with reverse time. This could also be indicative of insufficient embedding. Eighth, vertical (or horizontal) lines and clusters mean that some states do not change or change only slowly. Lastly, long bowed line structures point to similar evolution of states at different points but with different velocity, meaning that the dynamics of the system could be changing.

It was discussed before that in case a scalar time series or sequence has been measured, the phase space must be reconstucted for example by the delay embedding technique. If the parameters m and τ are set to one, we get what is called an unembedded RP. Increasing the embedding dimension by one reduces always the length of all diagonal lines by one and removes isolated dots entirely [83]. It was shown in [85] that an increase in the embedding dimension cleans the RP from single recurrence points and emphasises the diagonal structures as diagonal lines. This affects any quantification of RPs that are based on diagonal lines. Therefore the embedding parameters must be chosen carefully or statistics that are invariant to the embedding dimension should be used. Higher dimensional embedding, even if advantageous, can however cause spurious correlations in the regarded system [85] and too large an embedding dimension can make random/stochastic systems display strong artificial patterns of recurrence. This can happen even though diagonal structures should be extremely rare for uncorrelated data. The bottom line here is that a stochastic signal that is embedded in a high dimensional space can give rise to diagonal lines in RPs and feign nonexisting determinism.

Modifications to RP construction

Several modifications to the presented RP evaluation have been published. The original definition of RPs in [33] used the L_2-norm and ε was chosen for each $\mathbf{x}_i$ separately so that the neighbourhood contained a fixed amount of nearest neighbours (FAN) $\mathbf{x}_j$. This leads to an asymmetric RP, since it is possible that $\mathbf{x}_i$ is one of the nearest neighbours of $\mathbf{x}_j$ but not necessarily vice versa. Further, all the columns in the RP have the same recurrence density. The neighbourhood with a FAN plays an important role in detection of generalised synchronisation and cross recurrence plots [85]. Another modification for visualisation purposes and studying phase space trajectories is to plot the actual distances between states without quantising to zero or one. This kind of a plot is called a global recurrence plot or unthresholded recurrence plot. For further information, see the review on different ways of choosing the neighbourhoods and figures of the corresponding RPs in [85].

Further extensions of recurrence plots include cross recurrence plots and

multivariate joint recurrence plots. Measures of complexity for recurrence plots (such as RQA) and how dynamical invariants for RPs can be derived are also considered in [85]. Furthermore, the potential of RPs for the analysis of spatial data, the detection of UPOs [16], detection and quantification of different kinds of synchronisation and effects of noise are considered. Comparisons with other methods (if applicable) that have been used in similar tasks are also given, see [85].

5.1.1 Recurrence quantification analysis

We now turn to quantifying the characteristics of RPs. One of the first attemps to quantify RPs was given in [126]. The approach is called recurrence quantification analysis and evaluates several important scores based on RPs. Five scores called recurrence rate (REC), percent determinism (DET), entropy of the diagonal line lengths ($ENTR$), trend ($TREND$) and ratio ($RATIO$) were initially introduced. Later, two more scores were introduced, the first quantifying vertical lines (laminarity LAM) and the second the average length of vertical line structures (trapping time TT) [86]. In addition, computation of these measures in small windows of the RP moving along the line of identity can be useful in detecting state transitions and the time dependent behaviour of these variables. Further scores and modifications have also been reviewed in [85]. For other reviews of RQA, see [127, 137].

A lot of criticism has erupted on the RQA metrics. In [16] the authors state that RQA cannot elucidate the spatiotemporal details of the dynamics of the underlying system. Further, RQA results on structurally dissimilar RPs can be virtually identical [57]. However, the authors in [126] claim that the goal of RQA is not to search for chaos or reconstruct attractors. Instead, the point is to use recurrence plot methodologies to reveal dynamical behaviour in sequences that is not so obvious and is not detected by standard linear techniques. Although having been subject to a lot of critique, RQA has been successfully used in many applications ranging from chemoinformatics [9] to economy [134].

We now focus on the application of RQA to RPs and consider the different RQA scores, their potentials and limits.

Recurrence density estimation

Recurrence rate (sometimes also called as correlation sum) of a system as measured from its RP is defined as

$$REC(\varepsilon) = \frac{1}{n^2} \sum_{i,j=1, i \neq j}^{n} R_{i,j}(\varepsilon). \tag{5.5}$$

Recurrence rate measures thus the density of recurrence points in the RP. In the limit $n \to \infty$, REC is the probability that a state recurs to its ε-neighbourhood in phase space [85]. The average number of neighbours each point on a trajectory has in its ε-neighbourhood is given by

$$N_n(\varepsilon) = \frac{1}{n} \sum_{i,j=1, i \neq j}^{n} R_{i,j}(\varepsilon). \tag{5.6}$$

Quantifying diagonal line structures

The histogram of diagonal line lengths is an important concept in quantifying diagonal line structures in RPs (see Figure 5.5 for examples). It is defined as

$$P(\varepsilon, l) = \sum_{i,j=1, i \neq j}^{n} (1 - R_{i-1,j-1}(\varepsilon))(1 - R_{i+l,j+l}(\varepsilon)) \prod_{k=0}^{l-1} R_{i+k,j+k}(\varepsilon), \tag{5.7}$$

and for simplicity of notation assuming that $R_{i,j} = 0$ outside the defined boundaries (*e.g.* at $R_{0,0}$).

Processes with uncorrelated or weakly correlated, stochastic or chaotic behaviour cause none or short diagonals and more deterministic processes have been shown to cause longer diagonals (and less isolated points) in RPs [85]. To measure the determinism and predictability of a system based on diagonal lines, the score DET is introduced as the ratio of recurrent points that form diagonal structures to all recurrent points, *i.e.*

$$DET = \frac{\sum_{l=l_{\min}}^{n} lP(l)}{\sum_{l=1}^{n} lP(l)}, \tag{5.8}$$

where $l_{\min}$ is the minimum length of diagonal lines considered (helping exclude diagonal lines formed by tangential motion) and omitting the symbol ε in $P(\varepsilon, l)$.

A diagonal line of length l indicates that a part of a trajectory is close to another segment during l τ-time steps. The lines are thus related to the divergence of trajectory segments. The average time that two segments of a trajectory are close to each other is quantified by the average diagonal line length

$$L = \frac{\sum_{l=l_{\min}}^{n} lP(l)}{\sum_{l=l_{\min}}^{n} P(l)}. \tag{5.9}$$

The complexity of RPs in respect of diagonal lines is reflected by the Shannon entropy of the probability $p(l) = P(l)/n_l$ to find a diagonal line of exactly length l in the RP,

$$ENTR = -\sum_{l=l_{\min}}^{n} p(l) \ln p(l), \tag{5.10}$$

where $n_l = \sum_{l \geq l_{min}} P(l)$ is the total number of diagonal lines. For uncorrelated stochastic signals the value of $ENTR$ is relatively small, since the RP is composed mainly of lines of length one, indicating low complexity.

The authors in [57] claimed that DET and $ENTR$ are independent of the embedding dimension. However, it was shown in [83] that this is only true for some low-dimensional chaotic processes whose recurrence rate scales approximately exponentially as $REC_m \approx Ae^{-K_2 m}$ for some K_2 (the Kolmogorov entropy rate) and embedding dimension m. For the Shannon entropy $ENTR$ this holds only in the case of perfect exponential scaling (data derived from an I.I.D. process).

It is useful in some instances, for example computing the variable $TREND$, to compute the variables REC and DET separately for each diagonal parallel to the line of identity. Therefore, RQA measures for a certain line parallel to and distance τ from the line of identity are denoted as τ-recurrence rate (REC_τ) and τ-determinism (DET_τ). Further, we denote $P_\tau(l)$ as the number of diagonal lines of length l on each diagonal $R_{i,i+\tau}$ parallel to the line of identity. The τ-recurrence rate for the diagonal lines at distance τ from the line of identity is

$$REC_\tau = \frac{1}{n-\tau} \sum_{i=1}^{n-\tau} R_{i,i+\tau} = \frac{1}{n-\tau} \sum_{l=1}^{n-\tau} l P_\tau(l). \tag{5.11}$$

This measure can be thought of as a generalised auto-correlation function [85]. The measure describes higher order correlations between points of trajectories depending on τ. An advantage over the linear auto-correlation function is that REC_τ can be determined for a trajectory in phase space and not only for a single observable of the trajectory of a system. It is also the probability that a state recurs to its ε-neighbourhood after τ time steps. The τ-determinism is defined as

$$DET_\tau = \frac{\sum_{l=l_{\min}}^{n-\tau} l P_\tau(l)}{\sum_{l=1}^{n-\tau} l P_\tau(l)}, \tag{5.12}$$

and describes the proportion of recurrence points forming diagonal lines longer than $l_{\min}$ to all recurrence points on the chosen diagonal. It is further noted in [85] that the τ-RQA measures are also important scores on their own. The measure REC_τ has been used in finding UPOs in low-dimensional chaotic systems [40, 64, 90]. The main motivation for using REC_τ in this is that periodic orbits are more closely related to the occurrance of longer diagonal structures. Other application areas are in studying nonstationarity in data [33] an analysing synchronisation between oscillators [85].

Yet another RQA measure is $TREND$, which provides information about nonstationarity in the process, for example whether a drift is present in the analysed trajectory. This variable is defined as the linear regression coefficient over the recurrence point density REC_τ of the diagonals parallel to

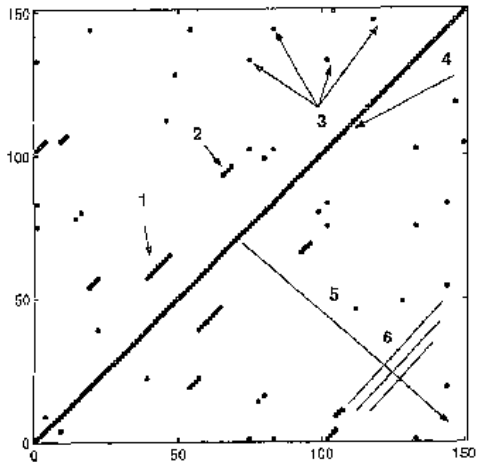


Figure 5.5: Typical diagonal line structures and points in RPs [137]. (1) a diagonal line composed of 8 recurrent points. (2) 4-point diagonal line. (3) Chance recurrent points. (4) The line of identity. (5) The line perpendicular to the line of identity, representing the x-axis in computing $TREND$. (6) Perpendicular lines along which recurrences may fall.

the line of identity, as a function of the time distance between the diagonals and the line of identity

$$TREND = \frac{\sum_{\tau=1}^{\tilde{n}}(\tau - \tilde{n}/2)(REC_\tau - 1/\tilde{n}\sum_{t=1}^{\tilde{n}} REC_t)}{\sum_{\tau=1}^{\tilde{n}}(\tau - \tilde{n}/2)^2}, \tag{5.13}$$

where $\tilde{n} < n$ is chosen so that the edges of the RP are excluded. This is because of the insufficient number of recurrence points near the corners of RPs. Another way to look at $TREND$ is to view it as the slope of the τ-recurrence rates when the line number 5 in Figure 5.5 is the x-axis and the τ-recurrence rates form the y-axis. For more discussion on choosing $\tilde{n}$, see [85].

Lastly, the measure $RATIO$ is defined as the ratio between DET and REC [126]. It is based on the number $P(l)$ of diagonal lines of length l as

$$RATIO = n^2 \frac{\sum_{l=l_{\min}}^{n} lP(l)}{(\sum_{l=1}^{n} lP(l))^2}. \tag{5.14}$$

The authors state in [126] that during certain types of qualitative physiological state transitions the number of recurrent points decreases and the proportion of points in line structures is less affected. Therefore, during physiological transitions $RATIO$ increases substantially but settles down again when a new state is achieved.

Quantifying vertical line structures

As mentioned in [85], large part of the vertical lines in RPs of continuous time systems, discretised with sufficiently high time resolution and using an appropriately large ε, are caused by tangential motion of the phase space trajectory. There are, however, some elements in the sets of vertical lines that do not correspond to tangential motion, for example in the presence of laminar states in intermittent regimes. Vertical lines are also present in the RPs of systems that have two or more different time scales. The total number of vertical lines of length v in the RP can be evaluated as

$$P_{\mathrm{vl}}(v) = \sum_{i,j=1}^{n} (1 - R_{i,j})(1 - R_{i,j+v}) \prod_{k=0}^{v-1} R_{i,j+k}. \tag{5.15}$$

The laminarity score is then evaluated analogously to DET as

$$LAM = \frac{\sum_{v=v_{\min}}^{n} v P_{\mathrm{vl}}(v)}{\sum_{v=v_1}^{n} P_{\mathrm{vl}}(v)}, \tag{5.16}$$

with $v_{\min}$ chosen large enough to decrease the influence of tangential motion (although $v_{\min} = 2$ is often used). The average length of vertical structures, known as trapping time, is given by

$$TT = \frac{\sum_{v=v_{\min}}^{n} v P_{\mathrm{vl}}(v)}{\sum_{v=v_{\min}}^{n} P_{\mathrm{vl}}(v)}, \tag{5.17}$$

and estimates the mean time the system will stay at a specific state.

The scores based on vertical line structures are able to find chaos-chaos transitions and allow for the investigation of intermittency in short and non-stationary data [86]. Chaos-order transitions can also be identified because for periodic dynamics the measures quantifying vertical structures are zero. It is further pointed out in [85] that the RQA variables quantifying vertical line structures can detect transitions between chaos and periodic windows based on just approximately 1000 data points, whereas some formerly proposed methods may need as many as 100000 data points.

5.1.2 Embedding independent properties of RPs

The previously introduced scores quantifying RPs are rather heuristic but are useful in finding various transitions in dynamical systems. The biggest problem of the RQA framework is that the measures are in general dependent on the embedding parameters used in reconstructing the phase space trajectory [85]. In some cases, if ground truth knowledge of the data at hand were available, resampling methods such as cross validation could be used in estimating the optimal embedding parameters. There are, however, some embedding invariant properties in RPs.

It is shown in [85] that the correlation dimension and correlation entropy are independent of the choice of the embedding dimension. The downside is that the accurate estimation of these variables may require tens or hundreds of thousands of data points; amounts that are not usually available for actual measurement data. On the other hand, correlation entropy can be used in estimating generalised mutual information, which quantifies the amount of information obtained from the measurement of one variable on another and has been applied to quantify dependencies within and between time series.

Recurrence plot statistics that are invariant to embedding dimension were also studied in [83]. It was noted in [83] that RPs of higher embedding dimensions can be obtained from the parental unembedded RP. Thus, embedding is not strictly necessary since all of the information is contained in the unembedded plot (*i.e.* $m = 1$). For further discussion on the effects of embedding, see [85].

5.2 Application of recurrence plots to solvent accessibility data

Since mutations in genes and deficiencies in post-translational processing of proteins can lead to serious illness and cancer, it would be very interesting to know which positions in the protein sequence are more susceptible to pathogenic mutations than others. In Publication-V, we studied protein sequences with help of recurrence plots and recurrence quantification analysis and distinguished locations where pathogenic mutations occur more frequently than elsewhere. This was verified by information on locations of pathogenic amino acid mutations in clinical patients. However, this additional information was not used in any kind of training of the algorithm; it was used only to validate the results.

Although protein solvent accessibility (or hydropathy) sequence is not a time series, it can be considered to be an ordered (spatial instead of temporal) sequence that RQA can equally well be applied to [43]. In [99], the authors show that the first principal component of the RQA variables (REC, DET, $ENTR$ and $TREND$) of different protein p53 mutants can discriminate between two known differently acting mutation types (binary classification). The authors used estimated hydropathy profiles for the wild type and mutated p53 proteins as their data set. In Publication-V we studied several protein domains with knowledge of pathogenic mutation locations in the sequences. Our first target, the human Bruton tyrosine kinase (BTK), is an extensively studied [7, 50, 54, 82, 123] protein that plays a crucial role in B cell development. Mutations in this protein result in X-linked agammaglobulinemia (XLA), which is an immunodeficiency characterised by failure to produce mature B lymphocyte cells [73]. We also analyse here the von Hipplel-Lindau (VHL) tumor suppressor protein [65]. Changes in this

protein can lead to failure in controlling a protein called hypoxia-inducible factor. Excess amounts of this factor stimulates cells to divide abnormally and can lead to development of cysts and tumors. The considered data sets include, besides the amino acid sequences of the protein domains, the solvent accessibility sequences and locations of pathogenic amino acid mutations discovered in clinical patients. For each pathogenic mutation location, several amino acid substitutions may have been found.

In applying RQA to solvent accessibility data, the L_2-norm was used. The L_∞-norm was also tested but did not significantly change the results. In addition, embedding dimension 4 was chosen. The decision has no other than a heuristic reasoning; [44] claim that for hydrophobicity data, embedding dimension of 4 is dictated by a balance between the need to have a window large enough to keep track of between-residue interactions and on relying, at the same time, on a sufficient number of considered windows. Since a change of an amino acid in a protein can change the hydropathy and solvent accessibility around its location, we decided to simulate this effect in Publication-V by introducing values that are inconsistent with the rest of the data, *i.e.* outliers, to the wild type solvent accessibility sequence.

5.2.1 Outlier analysis of RQA

We demonstrate next how the proposed method works on actual measured protein solvent accessibility data. In previously shown RPs the origin has been in the bottom-left corner but in the following figures the origin is placed in the top-left corner, as in Publication-V. The data is obtained from three domains of the wild type human Bruton tyrosine kinase [125] (RPs in Figure 5.6 (a-c)) and from von Hippel-Lindau (VHL) tumor suppressor protein (Figure 5.6 (d), not considered in Publication-V). The proposed method was applied to other proteins as well and results were similar (not shown here).

In our approach, we subsequently change each value in the solvent accessibility sequence one location at a time by inserting a value (an outlier) that is at a far distance from all the other values. This effectively means that the affected windows can no longer be recurrent with any others. Although the effect of a real amino acid mutation to the solvent accessibility sequence is smoothened and spread to the adjacent positions, the outlier replacement coarsely approximates the worst case scenario of an amino acid change in the protein sequence. We next compute REC, by using ε corresponding to the 10% of the entire distance range (as in [126] and many other publications) in the wild type sequence, for all the modified sequences and plot REC as a function of the outlier location in the corresponding sequence. If the outlier hits an area of no recurrence or chance recurrence, REC will not change much. On the other hand, if the outlier hits an area contributing to line structures, REC will decrease more considerably. We call this procedure

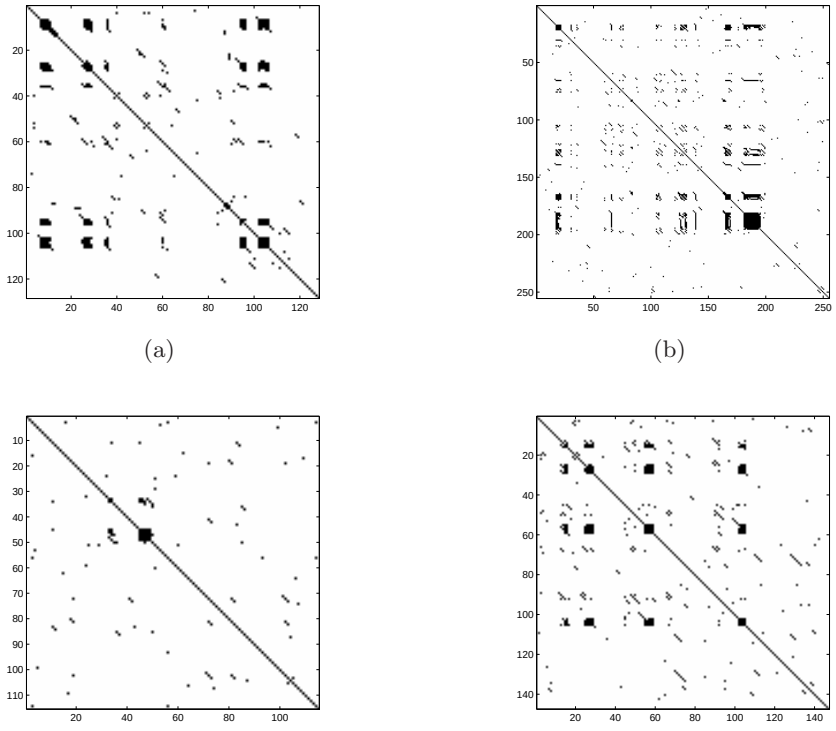


Figure 5.6: Recurrence plots for BTK PH domain (a), BTK kinase domain (b), BTK SH2 domain (c) and von Hippel-Lindau tumor suppressor protein (d).

as outlier analysis of RQA. If the data were generated by an independent noise process and the corresponding RP consisted mainly of line structures of length 1, the *REC* change plot would fluctuate slightly down (when the outlier affects one or a few recurrent points) and up (when the outlier has no effect on *REC*) and could maybe be modelled by a t-distribution. However, with longer line structures in the RP, the *REC* change plot will additionally make strong downward fluctuations that make the *REC*-change data asymmetrically distributed. Therefore, we opt to use a robust detection threshold to distinguish regions with a strong *REC* change. Least trimmed squares [107] is used in fitting a baseline to the *REC* change data and finding a robust estimate for the scale of the data. The threshold, below which points are treated as candidates for pathogenic mutation locations, is then chosen based on the scale estimate and t-distribution (one-sided 99% value). We also consider the receiver operating characteristic curve approach here that is independent of a single detection threshold. In other words, we use the recurrence percentage change values and knowledge of pathogenic mu-

tation locations (although this information is likely to be incomplete) to construct ROC curves that indicate how well our method performs on the different protein accessibility sequences.

We found that the change caused by the outliers in *REC* was more informative than the change in *DET*. This points to the possibility that the vertical lines in RPs, corresponding to laminar states, are also important when considering volatile locations in protein sequences. Since there is usually a considerable amount of measurement noise involved, some diagonal line structures may have also been split to single, although relatively close, recurrent points and thus make *REC* more interesting to observe.

The best results shown in Publication-V were obtained for the BTK PH domain whose corresponding RP is seen in Figure 5.6 (a), *REC*-change in Figure 5.7 (a) and ROC curve in Figure 5.8 (a). The detected locations cover 27.5% of the whole range but 62.5% of the known mutation locations reside in this area, indicating better than random detetection (15 true positives, 9 false negatives). The ROC curve, which is not dependent on the detection threshold, in Figure 5.8 (a) shows a clear deviation from the chance diagonal (area under the curve 0.74). Results for BTK kinase domain (area under the ROC curve 0.64) and VHL (area under the ROC curve 0.59) are shown in Figures 5.6-5.8 (b) and 5.6-5.8 (d), correspondingly. The results for BTK SH2 domain, whose ROC curve fluctuates around the chance diagonal (Figure 5.8 (c)) are not very good (area under the curve 0.5). This is possibly a result of the low initial recurrence percentage. Changing ε for this sequence did not yield any better results either. The low initial *REC* of the SH2 domain could be a result of shorter sequence length and/or measurement noise. It must be noted that in Publication-V, the value $\varepsilon = 10\%$ of the entire distance range was used only for the BTK kinase domain (Figure 5.6 (b)). The value $\varepsilon = 8\%$ of the distance range was accidentally used for the RPs of BTK PH and BTK SH2 domains. Figures 5.6 (a) and 5.6 (c) show the corrected plots, which differ from the ones in Publication-V. To see if the ad-hoc selection of $\varepsilon = 10\%$ of the entire distance range is reasonable, we observe the effect of varying ε next.

By varying ε in the analysis of BTK PH domain, the ratio of detected true mutation percentage to detected point percentage was highest at around 3.5 (detecting 37.5% of the mutations and covering 10.7% of the sequence) with $\varepsilon \approx 15\%$ of the maximum wild type Euclidian distance range. The ratio is shown for several values of ε in Figure 5.9. It can be observed that shortly after the value $\varepsilon = 13\%$, although the prediction ratio goes initially high, both the detected true mutations and detected locations approach zero. This is also evident with very low values of ε. The highest detected mutation percentage (62.5%) was obtained for $\varepsilon \approx 10\%$ of the range. It is also visible in Figure 5.9 that the ratio is nearly constant for a wide range of ε. Similar behaviour was also observed for BTK kinase domain (Figure 5.6 (b)), *i.e.* changing ε around 10% changed the results only slightly. The

choice of $\varepsilon \approx 10\%$ (based on literature [126]) of the range is therefore quite reasonable, at least here.

Based on these results, the conclusion can be drawn that the density of harmful mutations in the detected locations, mostly corresponding to the vertical and diagonal lines in the RPs, is generally higher than elsewhere.

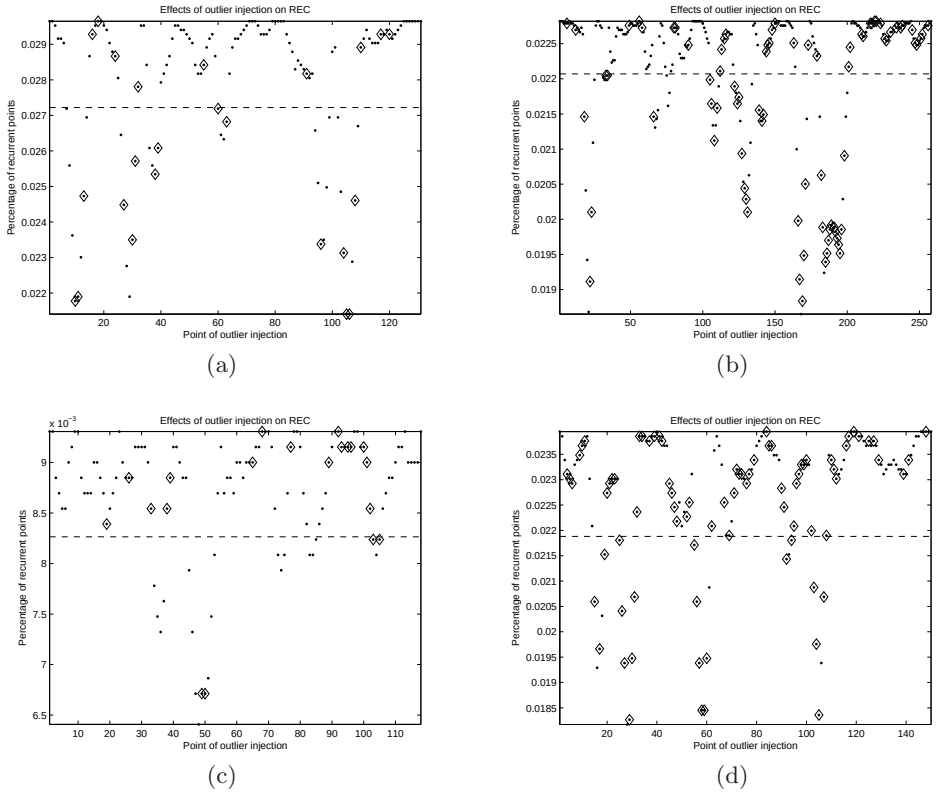


Figure 5.7: REC change plots for BTK PH domain (a), BTK kinase domain (b), BTK SH2 domain (c) and von Hippel-Lindau tumor suppressor protein (d). The plain spots correspond to REC percentages in locations where pathogenic mutations have not been observed in clinical patients. The spots surrounded by diamond shapes correspond to locations where pathogenic mutations have been observed in clinical patients. The dashed line shows the detection threshold, below which points are treated as susceptible to pathogenic mutations.

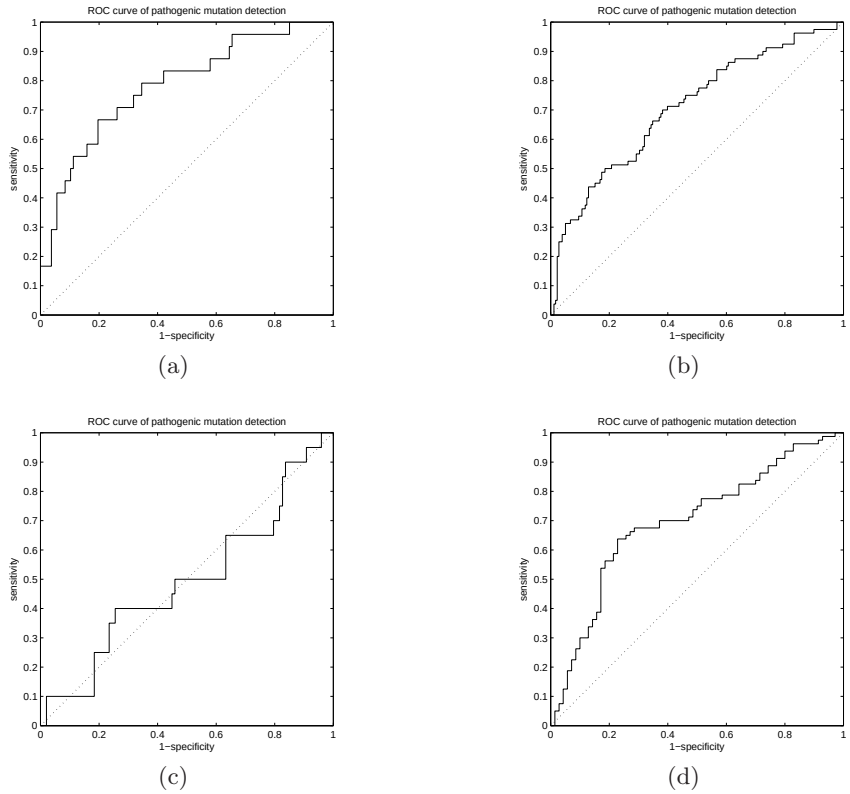


Figure 5.8: ROC curves of detection for BTK PH domain (a), BTK kinase domain (b), BTK SH2 domain (c) and von Hippel-Lindau tumor suppressor protein (d). The plots illustrate that for cases (a), (b) and (d) the detection deviates from random.

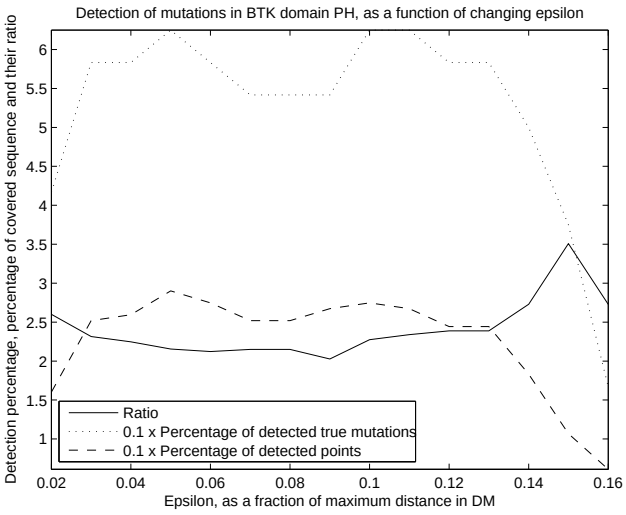


Figure 5.9: The effect of changing the value of ε on the value of detected locations in the sequence and detected true mutation locations.

Chapter 6

Conclusions

Several new robust statistical signal processing methods were introduced in this thesis for time series analysis in genomic and proteomic data. In addition, a general simulator framework for simulating gene expression microarrays was presented.

In Chapter 2 an overview of microarray technologies was given, of which gene microarrays represent the most advanced high throughput gene expression measurement technology to date. In Chapter 3, a highly modular gene expression microarray simulator was introduced that is able to simulate microarray measurements in a statistically sound way. One of the many uses of the simulator is to allow conducting microarray experiments in silico and help plan real microarrays. It can also be used in benchmarking different microarray preprocessing and analysis algorithms. Future plans include performing a comparison study of the different published error models for microarray data and implementing the simulation of other types of high throughput arrays. In addition, inference for the error terms and verifying the usefulness of the different models in processing microarray data should be considered.

Three robust periodicity detection methods were considered in Chapter 4. Robust spectrum estimation and periodicity detection have gained surprisingly little attention in the literature. Publication-II, Publication-III and Publication-IV are good exceptions to this and provide serious alternatives to classical methods in short length data with unkown noise characteristics. These methods can be used in a wide variety of applications where robustness is needed. To the author's knowledge, the algorithm introduced in Publication-II has already been applied to a wide range of measured microarray data as well as on nesting frequency data from leatherback turtles nesting in Gabon (personal communication based on algorithm implementation requests). There have been a total of 20 requests by E-Mail (as of October 18., 2007) for the implementation of the algorithm besides anonymous downloads on the companion website (for which no statistics exist). It

is also suprising that in the simulation studies considered in Publication-III the robust rank based algorithm has better receiver operating characteristics than the exact classical Fisher's test in the case of pure Gaussian noise. The results in Publication-III, when nonuniform sampling is considered, are equally surprising. It is evident from the simulation results that the algorithm presented in Publication-II operates outstandingly even when the sampling is not exactly uniform. In the case of more extreme nonuniform sampling schemes the regression based framework proposed in Publication-III performs robustly and with good receiver operating characteristics. Future work in this research direction includes considering wavelets in periodicity detection. As opposed to Fourier series based functions, some wavelet basis functions are localised and are known to capture the transient behaviour in signals. Wavelet analysis of short length time series and robustifying the wavelet transform are worth closer inspection.

The presentation given in Chapter 5 is the least developed aspect in this thesis. It was shown in Publication-V that the detected deterministic (even if weakly so) parts of human Bruton tyrosine kinase are volatile to pathogenic mutations and the density of pathogenic mutations is higher in the areas detected by our algorithm. This serves as a good motivation for continuing the work on recurrence based methods. Future work in this area should be devoted to formally quantifying the more or less ad-hoc measures and settings in the algorithms. It is important to quantify how much the paramaters in recurrence plots affect the results. Especially, the effects of the chosen norm and ε (although the choice of ε was shown to be quite liberal in Chapter 5) in constructing recurrence plots and the embedding parameters in recurrence quantification analysis should be quantified. Incorporating other biological (such as amino acid conservation) information and considering estimated hydrophobicities of the wild type and mutated sequences (and their comparison to the results obtained from solvent accessibilities) give also options for refining the algorithms.

Bibliography

[1] M. Ahdesmäki. Data dependent cdna-microarray simulator. In T. Aho, H. Lähdesmäki, and O. Yli-Harja, editors, *Proceedings of The 2nd TICSP Workshop on Computational Systems Biology, WCSB04*, 2004.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[3] C. Andersson, A. Isaksson, and M. Gustafsson. Bayesian detection of periodic mrna time profiles without use of training examples. *BMC Bioinformatics*, 7:63, 2006.

[4] G. Arce and T. Mu. Order statistic filter banks. *IEEE Transactions on Image Processing*, 5:827–837, 1996.

[5] Y. Balagurunathan, E. Dougherty, Y. Chen, M. Bittner, and J. Trent. Simulation of cdna microarrays via a parameterized random signal model. *Journal of Biomedical Optics*, 7(3):507–523, 2002.

[6] P. Baldi and S. Brunak. *Bioinformatics*. MIT Press, 2nd edition, 2001.

[7] E. Baraldi, K. Carugo, M. Hyvönen, P. Surdo, A. Riley, B. Potter, R. O'Brien, J. Ladbury, and M. Saraste. Structure of the ph domain from bruton's tyrosine kinase in complex with inositol 1,3,4,5-tetrakisphosphate. *Structure*, 7(4):449–460, 1999.

[8] K. Beckingham, J. Armstrong, M. Texada, R. Munjaal, and D. Baker. Drosophila melanogaster–the model organism of choice for the complex biology of multi-cellular organisms. *Gravitational and space biology bulletin*, 18:17–29, 2005.

[9] R. Benigni, A. Giuliani, J. Zbilut, S. Ellis, and D. Allorge. A signal analysis approach applied to the study of sequence, structure and function of the proteins. *Current Computer - Aided Drug Design*, 2:189–201, 2006.

[10] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the*

Royal Statistical Society: Series B (Statistical Methodology), 57:289–300, 1995.

[11] J. Berg, J. Tymoczko, and L. Stryer. *Biochemistry.* W. H. Freeman, 6th edition, 2006.

[12] W. Blake, M. KÆrn, C. Cantor, and J. Collins. Noise in eukaryotic gene expression. *Nature*, 422:633–637, 2003.

[13] F. Blattner, G. Plunkett III, C. Bloch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner, C. Rode, G. Mayhew, N. Gregor, J. Davis, H. Kirkpatrick, M. Goeden, D. Rose, B. Mau, and Y. Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277:1453–1474, 1997.

[14] T. Bø, B. Dysvik, and I. Jonassen. Lsimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32:e34, 2004.

[15] B. Bolstad, R. Irizarry, M. Åstrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

[16] E. Bradley and R. Mantilla. Recurrence plots and unstable periodic orbits. *CHAOS*, 12:596–600, 2002.

[17] L. Breeden. Periodic transcription: a cycle within a cycle. *Current Biology*, 13:R31–R38, 2003.

[18] G. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation.* Springer-Verlag, Berlin Heidelberg, 1st edition, 1988.

[19] P. Brockwell and R. Davis. *Time Series: Theory and Methods.* Springer-Verlag, New York, 2nd edition, 1991.

[20] T. Brown. *Genomes.* BIOS Scientific Publishers Ltd, 1st edition, 1999.

[21] N. Brändle, H. Bischof, and H. Lapp. Robust dna microarray image analysis. *Machine Vision and Applications*, 15:11–28, 2003.

[22] M. Campbell. *Biochemistry.* Saunders College Publishing, 2nd edition, 1995.

[23] S. Celniker and G. Rubin. The drosophila melanogaster genome. *Annual Reviews of Genomics and Human Genetics*, 4:89–117, 2003.

[24] J. Chen. Identification of significant genes in microarray gene expression data. *BMC Bioinformatics*, 6:286, 2005.

[25] H. Cho and J. Lee. Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics*, 20(13):2016–2025, 2004.

[26] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:915–916, 2004.

[27] A. Correa, Z. Lewis, A. Greene, I. March, R. Gomer, and D. Bell-Pedersen. Multiple oscillators regulate circadian gene expression in *neurospora*. *Proceedings of the National Academy of Sciences of the USA*, 100:13597–13602, 2003.

[28] A. Dabney and J. Storey. Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biology*, 8:R44, 2007.

[29] U. de Lichtenberg, L. Jensen, A. Fausbøll, T. Jensen, P. Bork, and S. Brunak. Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics*, 21:1164–1171, 2004.

[30] K.-A. Do, P. Müller, and F. Tang. A bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society*, 54:627–644, 2005.

[31] R. Dror, J. Murnick, N. Rinaldi, V. Marinescu, R. Rifkin, and R. Young. Bayesian estimation of transcript levels using a general model of array measurement noise. *Journal of Computational Biology*, 10(3-4):433–452, 2003.

[32] S. Dudoit, J. Shaffer, and J. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

[33] J. Eckmann, S. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4:973–977, 1987.

[34] D. Eisenberg, E. Marcotte, I. Xenarios, and T. Yeates. Protein function in the post-genomic era. *Nature*, 405:823–826, 2000.

[35] C. Ekstrom, S. Bak, C. Kristensen, and M. Rudemo. Spot shape modelling and data transformations for microarrays. *Bioinformatics*, 14:2270–2278, 2004.

[36] J.-B. Fan, X. Chen, M. Halushka, A. Berno, X. Huang, T. Ryder, R. Lipshutz, D. Lockhart, and A. Chakravarti. Parallel genotyping of human snps using generic high-density oligonucleotide tag arrays. *Genome Researh*, 10(6):853–860, 2000.

[37] G. Filligoi and F. Felici. Detection of hidden rhythms in surface emg signals with a nonlinear time-series tool. *Medical Engineering & Physics*, 21:439–448, 1999.

[38] R. Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London*, 125:54–59, 1929.

[39] H. Fraser, A. Hirsh, G. Giaever, J. Kumm, and M. Eisen. Noise minimization in eukaryotic gene expression. *PLoS Biology*, 2(6):e137, 2004.

[40] R. Gilmore. Topological analysis of chaotic dynamical systems. *Reviews of Modern Physics*, 70:1455–1529, 1998.

[41] A. Giuliani, R. Benigni, P. Sirabella, J. Zbilut, and A. Colosimo. Nonlinear methods in the analysis of protein sequences: a case study in rubredoxins. *Biophys J*, 78:136–149, 2000.

[42] A. Giuliani, R. Benigni, J. Zbilut, C. Webber Jr., P. Sirabella, and A. Colosimo. Nonlinear signal analysis methods in the elucidation protein sequence-structure relationships. *Chem Rev*, 102:1471–1492, 2002.

[43] A. Giuliani, M. Colafranceschi, C. Webber Jr., and J. Zbilut. A complexity score derived from principal components analysis of nonlinear order measures. *Physica A*, 301:567–588, 2001.

[44] A. Giuliani, P. Sirabella, R. Benigni, and A. Colosimo. Mapping protein sequences by recurrence quantification analysis: a case study on chimeric structures. *Protein Engineering*, 13:671–678, 2000.

[45] E. Glynn, J. Chen, and A. Mushegian. Detecting periodic patterns in unevenly spaced gene expression time series using lomb-scargle periodograms. *Bioinformatics*, 2005.

[46] A. Goffeau, B. Barrell, H. Bussey, R. Davis, B. Dujon, H. Feldmann, F. Galibert, J. Hoheisel, C. Jacq, M. Johnston, E. Louis, H. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. Oliver. Life with 6000 genes. *Science*, 274:563–567, 1996.

[47] P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypothesis*. Springer, New York, 1st edition, 2000.

[48] R. Gottardo, A. Raftery, K. Yeung, and R. Bumgarner. Quality control and robust estimation for cdna microarrays with replicates. *Journal of the American Statistical Association*, 101(473):30–40, 2006.

[49] S. Hanash and J. Celis. The human proteome organization: a mission to advance proteome knowledge. *Molecular & Cellular Proteomics*, 1:413–414, 2002.

[50] H. Hansson, P. Mattson, P. Allard, P. Haapaniemi, M. Vihinen, C. Smith, and T. Härd. Solution structure of the sh3 domain from bruton's tyrosine kinase. *Biochemistry*, 37(9):2912–2924, 1998.

[51] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young. Maximum-likelihood estimation of optimal scaling factors for expression array normalization. *Proceedings of SPIE, Microarrays: Optical Technologies and Informatics*, 4266:132–140, 2001.

[52] A. Hein, S. Richardson, H. Causton, G. Ambler, and P. Green. Bgx: A fully bayesian integrated approach to the analysis of affymetrix genechip data. *Biostatistics*, 6(3):349–373, 2005.

[53] W. Huber, A. von Heydebreck, and M. Vingron. Error models for microarray intensities. *Bioconductor Project Working Papers*, 6, 2004.

[54] M. Hyvönen and M. Saraste. Structure of the ph domain and btk motif from bruton's tyrosine kinase: molecular explanations for x-linked agammaglobulinaemia. *EMBO J*, 16(12):3396–3404, 1997.

[55] A. Irbäck and E. Sandelin. On hydrophobicity correlations in protein chains. *Biophysical Journal*, 79:2252–2258, 2000.

[56] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[57] J. Iwanski and E. Bradley. Recurrence plots of experimental data: To embed or not to embed. *CHAOS*, 8:861–871, 1998.

[58] D. Johansson, P. Lindgren, and A. Berglund. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, 19:467–473, 2003.

[59] H. Kitano, editor. *Foundations of Systems Biology*. MIT Press, 1st edition, 2001.

[60] H. Kitano. Computational systems biology. *Nature*, 420:206–210, 2002.

[61] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

[62] M. Laine and J. Tuimala. *DNA Microarray Data Analysis*. CSC - Scientific Computing Ltd., 2nd edition, 2005.

[63] E. Lander and et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[64] D. Lathrop and E. Kostelich. Characterization of an experimental strange attractor by periodic orbits. *Physical Review A*, 40:4028–4031, 1989.

[65] F. Latif, L. Tory, J. Gnarra, M. Yao, F. Duh, M. Orcutt, T. Stackhouse, I. Kuzmin, W. Modi, L. Geil, and et al. Identification of the von hippel-lindau disease tumor suppressor gene. *Science*, 260:1317–1320, 1993.

[66] C. Lausted, T. Dahl, C. Warren, K. King, K. Smith, M. Johnson, R. Saleem, J. Aitchison, L. Hood, and S. Lasky. Posam: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biology*, 5:R58, 2004.

[67] A. Lehmussola, P. Ruusuvuori, and O. Yli-Harja. Evaluating the performance of microarray segmentation algorithms. *Bioinformatics*, 2006.

[68] A. Lesk and C. Chothia. Solvent accessibility, protein surfaces, and protein folding. *Biophysical Journal*, 32:35–44, 1980.

[69] A. Lewin, S. Richardson, C. Marshall, A. Glazier, and T. Aitman. Bayesian modeling of differential gene expression. *Biometrics*, 62(1):1–9, 2006.

[70] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the USA*, 98:31–36, 2000.

[71] X. Li, G. Ouyang, X. Yao, and X. Guan. Dynamical characteristics of pre-epileptic seizures in rats with recurrence quantification analysis. *Physics Letters A*, 333:164–171, 2004.

[72] A.-W. Liew, J. Xian, S. Wu, D. Smith, and H. Yan. Spectral estimation in unevenly sampled space of periodically expressed microarray time series data. *BMC Bioinformatics*, 8:137, 2007.

[73] J. Lindvall, K. Blomberg, A. Berglöf, and C. Smith. Distinct gene expression signature in btk-defective t1 b-cells. *Biochemical and Biophysical Research Communications*, 346(2):461–469, 2006.

[74] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(1 Suppl):20–24, 1999.

[75] D. Liu, D. Umbach, S. Peddada, L. Li, P. Crockett, and C. Weinberg. A random-periods model for expression of cell-cycle genes. *Proceedings of the National Academy of Sciences of the USA*, 101:7240–7245, 2004.

[76] K. Lo and R. Gottardo. Flexible empirical bayes models for differential gene expression. *Bioinformatics*, 23(3):328–335, 2007.

[77] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. a. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.

[78] X. Lu, W. Zhang, Z. Qin, K. Kwast, and J. Liu. Statistical resynchronization and bayesian detection of periodically expressed genes. *Nucleic Acids Research*, 32:447–455, 2004.

[79] Y. Luan and H. Li. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20:332–339, 2004.

[80] N. Luscombe and J. Thornton. Protein-dna interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*, 320:991–1009, 2002.

[81] H. Lähdesmäki, H. Huttunen, T. Aho, M.-L. Linne, J. Niemi, J. Kesseli, R. Pearson, and O. Yli-Harja. Estimation and inversion of the effects of cell population asynchrony in gene expression time-series. *Signal Processing*, 83:835–858, 2003.

[82] C. Mao, M. Zhou, and F. Uckun. Crystal structure of bruton's tyrosine kinase domain suggests a novel pathway for activation and provides insights into the molecular basis of x-linked agammaglobulinemia. *Journal of Biological Chemistry*, 276(44):41435–41443, 2001.

[83] T. March, S. Chapman, and R. Dendy. Recurrence plot statistics and the effect of embedding. *Physica D*, 200:171–184, 2005.

[84] R. Maronna, D. Martin, and V. Yohai. *Robust Statistics - Theory and Methods*. Wiley, 2006.

[85] N. Marwan, M. Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438:237–329, 2007.

[86] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths. Recurrence-plot-based measures of complexity and their application to heart rate variability data. *Physical Review E*, 66:026702.1–026702.8, 2002.

[87] G. McLachlan, R. Bean, and L.-T. Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22(13):1608–1615, 2006.

[88] T. Mehta, M. Tanik, and D. Allison. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nature Genetics*, 36:943–947, 2004.

[89] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl 12):122–129, 2003.

[90] G. Mindlin and R. Gilmore. Topological analysis and synthesis of chaotic time series. *Physica D*, 58:229–242, 1992.

[91] E. Motakis, G. Nason, P. Fryzlewicz, and G. Gutter. Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics*, 22:2547–2553, 2006.

[92] D. Nelson and M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 4th edition, 2004.

[93] M. Nykter, T. Aho, J. Kesseli, and O. Yli-Harja. On estimation of statistical characteristics of microarray data. In *Proc. Finnish Signal Processing symposium FINSIG 2003, Tampere, Finland*, 2003.

[94] C. Pace, B. Shirley, M. McNutt, and K. Gajiwala. Forces contributing to the conformational stability of proteins. *The Journal of the Federation of American Societies for Experimental Biology*, 10:75–83, 1996.

[95] N. Packard, J. Crutchfield, J. Farmer, and R. Shaw. Geometry from a time series. *Physical Review Letters*, 45:712–716, 1980.

[96] C. Palliser and D. Parry. Quantitative comparison of the ability of hydropathy scales to recognize surface β-strands in proteins. *PROTEINS: Structure, Function, and Genetics*, 42:243–255, 2001.

[97] R. Pearson. *Mining Imperfect Data: dealing with contamination and incomplete records*. Siam, 2005.

[98] R. Pearson, H. Lähdesmäki, H. Huttunen, and O. Yli-Harja. Detecting periodicity in nonideal datasets. *SIAM International Conference on Data Mining 2003, Cathedral Hill Hotel, San Francisco, CA, May 1-3*, 2003.

[99] A. Porrello, S. Soddu, J. Zbilut, M. Crescenzi, and A. Giuliani. Discrimination of single amino acid mutations of the p53 protein by means of deterministic singularities of recurrence quantification analysis. *PROTEINS: Structure, Function and Bioinformatics*, 55:743–755, 2004.

[100] E. Poussu, M. Vihinen, L. Paulin, and H. Savilahti. Probing the alpha*-complementing domain of e. coli beta*-galactosidase with use of an insertional pentapeptide mutagenesis strategy based on mu in vitro dna transcription. *Proteins*, 54:681–692, 2004.

[101] M. Priestley. *Spectral Analysis and Time Series*. Academic Press, London, 1981.

[102] Y. Qi, T. Minka, and R. Picard. Bayesian spectrum estimation of unevenly sampled nonstationary data. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:1473–1476, 2002.

[103] R. Randles and D. Wolfe. *Introduction to the Theory of Nonparametric Statistics*. Wiley, 1979.

[104] D. Rocke and B. Durbin. A model for measurement error for gene expression array. *Journal of Computational Biology*, 8(6):557–569, 2001.

[105] B. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41:187–228, 2001.

[106] G. Rose, P. Fleming, J. Banavar, and A. Maritan. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the USA*, 103:16623–16633, 2006.

[107] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

[108] P. Rousseeuw, S. Van Aelst, K. Van Driessen, and J. Gulló. Robust multivariate regression. *Technometrics*, 46:293–305, 2004.

[109] J. Rumbley, L. Hoang, L. Mayne, and S. Englander. An amino acid code for protein folding. *Proceedings of the National Academy of Sciences of the USA*, 98:105–112, 2000.

[110] G. Rustici, J. Mata, K. Kivinen, P. Lió, C. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bähler. Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, 36:809–817, 2004.

[111] M. Rustici, C. Caravati, E. Petretto, M. Branca, and N. Marchettini. Transition scenarios during the evolution of the belousov-zhabotinsky reaction in an unstirred batch reactor. *Journal of Physical Chemistry A*, 103:6564–6570, 1999.

[112] O. Rössler. An equation for continuous chaos. *Physics Letters A*, 57:397–398, 1976.

[113] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[114] G. Sherlock. Analysis of large-scale gene expression data. *Briefings in Bioinformatics*, 2:350–362, 2001.

[115] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[116] K. Strimmer. Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics*, 4:10, 2003.

[117] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the USA*, 102:15545–15550, 2005.

[118] F. Takens. *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, chapter Detecting strange attractors in turbulence, pages 366–381. Springer, Berlin, 1981.

[119] L. Tatum and C. Hurvich. High breakdown methods of time series analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 55:881–896, 1993.

[120] L. Trulla, A. Giuliani, J. Zbilut, and C. Webber Jr. Recurrence quantification analysis of the logistic equation with transients. *Physics Letters A*, 223:255–260, 1996.

[121] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences of the USA*, 99(22):14031–14036, 2002.

[122] J. Tyson. *Computational Cell Biology: An Introductory Text on Computer Modeling in Molecular and Cell Biology*, chapter Biochemical oscillations. Springer-Verlag, 2002.

[123] S. Tzeng, Y. Lou, M. Pai, C. Chen, S. Chen, and J. Cheng. Solution structure of the human btk sh3 domain complexed with a proline-rich peptide from p120cbl. *Journal of Biomolecular NMR*, 16(4):303–312, 2000.

[124] J. Venter *et al.* The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[125] J. Väliaho, C. Smith, and M. Vihinen. Btkbase: the mutation database for x-linked agammaglobulinemia. *Human Mutation*, 27(12):1209 – 1217, 2006.

[126] C. Webber Jr. and J. Zbilut. Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2):965–973, 1994.

[127] C. Webber Jr. and J. Zbilut. *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*, chapter Recurrence Quantification Analysis of Nonlinear Dynamical Systems. Retrieved May 24, 2007, from http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp, 2005.

[128] A. Weinmann. Novel chip-based strategies to uncover transcription factor target genes in the immune system. *Nature Reviews Immunology*, 4(5):381–386, 2004.

[129] L. Weng, H. Dai, Y. Zhan, Y. He, S. Stepaniants, and D. Bassett Jr. Rosetta error model for gene expression analysis. *Bioinformatics*, 22(9):1111–1121, 2006.

[130] M. Whitfield, G. Sherlock, A. Saldanha, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13:1977–2000, 2002.

[131] S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20:5–20, 2004.

[132] Z. Wu and R. Irizarry. A statistical framework for the analysis of microarray probe-level data. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, 73, 2005.

[133] X. Xiaoyin, C. Dongbin, and M. Sarhadi. Adaptive two-pass rank order filter to remove impulse noise in highly corrupted images. *IEEE Transactions on Image Processing*, 13:238–247, 2004.

[134] J. Zbilut. *Economics: Complex Windows*, chapter Use of Recurrence Quantification Analysis in Economic Time Series. Springer Milan, 2005.

[135] J. Zbilut, A. Colosimo, F. Conti, M. Colafranceschi, C. Manetti, M. Valerio, C. Webber Jr, and A. Giuliani. Protein aggregation/folding: The role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and a beta*(1-40). *Biophysical Journal*, 85:3544–3557, 2003.

[136] J. Zbilut, A. Giuliani, C. Webber Jr, and A. Colosimo. Recurrencee quantification analysis in structure-function relationships of proteins: an overview of a general methodology applied to the case of tem-1 β-lactamase. *Protein Engineering*, 11:87–93, 1998.

[137] J. Zbilut, P. Sirabella, A. Giuliani, C. Manetti, A. Colosimo, and C. Webber Jr. Review of nonlinear analysis of proteins through recurrence quantification. *Cell Biochemistry and Biophysics*, 98:67–87, 2002.

[138] J. Zbilut, C. Webber Jr, A. Colosimo, and A. Giuliani. The role of hydrophobicity patterns in prion folding as revealed by recurrence quantification analysis of primary structure. *Protein engineering*, 13:99–104, 2000.

[139] L. Zhao, R. Prentice, and L. Breeden. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Sciences of the USA*, 98:5631–5636, 2001.

[140] C. Zhou, J. Wakefield, and L. Breeden. Bayesian analysis of cell-cycle gene expression data. *UW Biostatistics Working Paper Series*, 276, 2005.

[141] B. Zumbo and D. Coulombe. Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, 51:139–149, 1997.

Appendix A

Hilbert space properties

An inner-product space ($\mathcal{H}$), whose completisation a Hilbert space is [19], is characterised for each pair x and y from the space by a complex number $\langle x, y\rangle$ (the inner product) for which

$$\begin{array}{lll} (a) & \langle x, y\rangle = \overline{\langle y, x\rangle}, & \\ (b) & \langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle, & \forall x, y, z \in \mathcal{H} \\ (c) & \langle \alpha x, y\rangle = \alpha\langle x, y\rangle, & \forall x, y \in \mathcal{H},\ \alpha \in \mathbb{C} \\ (d) & \langle x, x\rangle \geq 0, & \forall x \in \mathcal{H} \\ (e) & \langle x, x\rangle = 0 \text{ if and only if } x = 0. & \end{array} \tag{A.1}$$

In a complex finite-dimensional inner-product space the inner-product is defined as $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{j=1}^{k} x_j \overline{y}_j$ with $\mathbf{x} = (x_1, \ldots, x_k)' \in \mathbb{C}^k$ ($\mathbf{y}$ defined in a similar manner). In the Euclidean space the corresponding vectors are real valued and the complex conjugate of y reduces to y. The norm of an element x of an inner-product space is defined by

$$\|x\| = \sqrt{\langle x, x\rangle}. \tag{A.2}$$

As was mentioned, a complete inner-product space is called a Hilbert space. Completeness of an inner-product space is guaranteed if all Cauchy sequences in the space converge in norm to some element $x \in \mathcal{H}$, *i.e.* for every $\epsilon > 0$ there exist positive integers $n_{\epsilon_1}, n_{\epsilon_2}$ such that

$$\begin{array}{rl} \|x_n - x_m\| < \epsilon, & \forall m, n > n_{\epsilon_1}, \\ \Rightarrow \exists x : \|x_n - x\| < \epsilon, & \forall n > n_{\epsilon_2}, \end{array} \tag{A.3}$$

where $\forall n \in \mathbb{N} : x_n \in \mathcal{H}$. For example, $\mathbb{R}^k$, $\mathbb{C}^k$ and $L^2(\Omega, \mathcal{F}, P)$ (a complete linear norm space) are Hilbert spaces, but the inner-product space $C^r(A, B)$ that is the set of functions for which the rth derivative is continuous, in general is not.

In a Hilbert space, the orthogonal complement of $\mathcal{M} \subseteq \mathcal{H}$ is the set $\mathcal{M}^\perp \subseteq \mathcal{H}$ whose all elements are orthogonal to all the elements of $\mathcal{M}$, *i.e.*

$$x \in \mathcal{M}^\perp \Leftrightarrow \langle x, y\rangle = 0\ \forall y \in \mathcal{M}. \tag{A.4}$$

An orthonormal set $\{e_t, t \in T\}$ is a special case of orthogonal sets for which for every $s, t \in T$: $\langle e_s, e_t \rangle = 1$ if $s = t$ and 0 otherwise. The projection theorem states that for a closed subspace $\mathcal{M} \subseteq \mathcal{H}$ there is a (unique) $\hat{x} \in \mathcal{M}$ such that $\|x - \hat{x}\| = \inf_{y \in \mathcal{M}} \|x - y\|$ if and only if $\hat{x} \in \mathcal{M}$ and $(x - \hat{x}) \in \mathcal{M}^{\perp}$, where $\hat{x}$ is called the orthogonal projection of x onto $\mathcal{M}$.

In $\mathbb{R}^n$ every closed subspace $\mathcal{M}$ can be expressed as $\mathcal{M} = \text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ with $\{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ an orthonormal subset of $\mathcal{M}$ and $m \leq n$ (the dimension of $\mathcal{M}$). In case $m < n$, there exists an orthogonal complement of $\mathcal{M}$ such that $\mathcal{M}^{\perp} = \text{span}\{\mathbf{e}_{m+1}, \ldots, \mathbf{e}_n\}$. If we denote the projector onto $\mathcal{M}$ as $P_{\mathcal{M}}$ (the projector onto $\mathcal{M}^{\perp}$ is $I - P_{\mathcal{M}}$), then

$$P_{\mathcal{M}}\mathbf{x} = \sum_{j=1}^{m} \langle \mathbf{x}, \mathbf{e}_j \rangle \mathbf{e}_j. \tag{A.5}$$

We can also compute $P_{\mathcal{M}}\mathbf{x}$ directly from any set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ ($\mathbf{x}_i \in \mathbb{R}^n$) that are not necessarily orthogonal but span $\mathcal{M}$ (meaning that they are linearly independent). First denote

$$P_{\mathcal{M}}\mathbf{x} = \textstyle\sum_{j=1}^{m} \beta_j \mathbf{x}_j = X\beta, \tag{A.6}$$

for some m-vector β and where X is the matrix composed of the vectors spanning the subspace. Since $X\beta - \mathbf{x} \in \mathcal{M}^{\perp}$, it follows that

$$\langle \mathbf{x}_j, X\beta - \mathbf{x} \rangle = \mathbf{x}_j'(X\beta - \mathbf{x}) = 0, \; j = 1, \ldots, m \tag{A.7}$$

and $X'(X\beta - \mathbf{x}) = \mathbf{0}$ from which further follows that $X'X\beta = X'\mathbf{x}$. Matrix $X'X$ is non-singular (and thus has an inverse) since the column vectors of X are assumed to be linearly independent. Therefore we have a unique solution

$$P_{\mathcal{M}}\mathbf{x} = X(X'X)^{-1}X'\mathbf{x}. \tag{A.8}$$

The vector $\beta = (X'X)^{-1}X'\mathbf{x}$ can be viewed as the set of coordinates for the projection in the subspace $\mathcal{M}$.

Appendix B

Publications

Publication-I

Nykter, M., Aho, T., Ahdesmäki, M., Ruusuvuori, P., Lehmussola, A. and Yli-Harja, O. (2006) Simulation of microarray data with realistic characteristics. *BMC Bioinformatics*, **7**:349.

Publication-II

Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H. and Yli-Harja, O. (2005) Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, **6**:117.

Publication-III

Ahdesmäki, M., Lähdesmäki, H., Gracey, A., Shmulevich I. and Yli-Harja, O. (2007) Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. *BMC Bioinformatics*, **8**:233.

Publication-IV

Ahdesmäki, M., Lähdesmäki, H. and Yli-Harja, O. (2007) Robust Fisher's test for periodicity detection in noisy biological time series. In *Proceedings of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics (Gensips'07)*, Tuusula, Finland, June 10-12, 2007.

Publication-V

Ahdesmäki, M., Thusberg, J., Huttunen, H., Vihinen, M. and Yli-Harja, O. (2007) Detection of pathogenic mutation prone locations from protein sequences using solvent accessibility measurements. In *Proceedings of the Fifth IEEE International Workshop on Genomic Signal Processing and Statistics (Gensips'07)*, Tuusula, Finland, June 10-12, 2007.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O. Box 527
FIN-33101 Tampere, Finland