THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Bayesian inference for detection problems in biology

Ronny Hedell

Division of Applied Matematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Gothenburg, Sweden, 2017

Bayesian inference for detection problems in biology
Ronny Hedell
ISBN 978-91-7597-652-5

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4333
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
SE–412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 10 00

Printed in Gothenburg, Sweden, 2017.

Bayesian inference for detection problems in biology

Ronny Hedell

Swedish National Forensic Centre,
Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis is about different kinds of detection problems in biology: detection of DNA sequences in crime scene samples, detection of harmful bacteria in feed and food stuff and detection of epidemical diseases in animal populations. In each case, biological data is produced or collected in order to determine which DNA sequences, bacteria types or diseases are present, if any. However, the state of nature will often remain uncertain due to limited amounts of samples, low quality samples and imperfect methods for detection and classification. For correct and efficient interpretation of such data it is therefore often necessary to use statistical methods, taking the different sources of uncertainty into account. Several Bayesian models for analysis of such data, for determining the performance of detection methods, and for deciding on the optimal analysis procedure are developed and implemented.

In paper I of this thesis it is investigated how the quality in forensic DNA profiles, such as allele dropout rates, changes with different analysis settings, and how the results depend on features in the DNA sample, such as the DNA concentration and marker type. Regression models are developed and the better analysis setting is determined. In paper II Bayesian decision theory is used to determine the optimal forensic DNA analysis procedure, after the DNA concentration and level of degradation in the sample have been estimated. It is assumed the alternatives for DNA analysis are 1) using a standard assay, 2) using the standard assay and a complementary assay, or 3) the analysis is cancelled. In paper III detection models for bacteria are developed. It is shown how heterogeneous experimental data can be used to learn about the sensitivity of detection methods for specific bacteria types, such as *Bacillus anthracis*. As exemplified in the paper, such results are useful e.g. when evaluating negative analysis results. Finally, in paper IV a Bayesian method for early detection of disease outbreaks in animal populations is developed and implemented. Based on reported neurological syndromes in horses, connected e.g. with the West Nile Virus, the probability of an outbreak is computed using a Gibbs sampling procedure.

Keywords: Bayesian inference, Forensic DNA analysis, PCR, Allele dropout, Bacillus anthracis, Syndromic surveillance, Markov chain Monte Carlo

Acknowledgements

First I would like thank my main supervisor Petter Mostad for all support, inspiration and nice discussions during these years, and for giving me so many insights about statistical reasoning. Thanks to my co-supervisor and colleague Anders Nordgaard for all support, encouragement and good advices, and for giving me the opportunity to work within the field of forensic statistics. Many thanks to my co-supervisor Gunnar Andersson at SVA for all commitment, support and fruitful ideas. Thanks also for providing contact with many researchers within Sweden and abroad. Thanks to my co-supervisor and colleague Johannes Hedman for all support and encourgement during these years, and for giving me a better insight into the field of forensic DNA analysis.

Thanks to my colleague and former boss Birgitta Rasmusson for all encouragement and for making this PhD project possible. Many thanks to my colleague Ricky Ansell for giving me the opportunity to work within the field of forensic DNA analysis, for all good advice, and for the many insights I have been given into the field. Thanks to my examiner Olle Nerman for encouragement and for showing interest in my work. Thanks to my other colleagues at NFC and the Biology unit for support and friendship. Additional thanks to Charlotte Dufva, Lina Boiso and Malin Sanga for great co-authorship and help with data management. Thanks to my co-authors Olga Stephansson, Céline Faverjon and Agnès Leblond. It has been a pleasure working together with you. Thanks to Ivar Simonsson, Maja Sidstedt and my other PhD fellows for nice discussions and for showing interest in my work.

I am eternally grateful to my mother Tina, father Torbjörn, sister Marie, and the rest of my family for always being there. Thanks to Marianne and Lennart for all your kindness. Finally, to my wife and love Margje, I am forever thankful for all your support, for always being there and for sharing your life with me.

List of Papers

This thesis includes the following papers:

I. **Hedell R.**, Dufva C., Ansell R., Mostad P., Hedman J., (2015), *Enhanced low-template DNA analysis conditions and investigation of allele dropout patterns*, Forensic Science International: Genetics **14**: 61-75.
DOI: http://dx.doi.org/10.1016/j.fsigen.2014.09.008

II. **Hedell R.**, Hedman J., Mostad P., (2017), *Determining the optimal forensic DNA analysis procedure following investigation of sample quality*, International Journal of Legal Medicine.
DOI: http://dx.doi.org/10.1007/s00414-017-1635-1

III. **Hedell R.**, Stephansson O., Mostad P., Andersson G., (2017), *Detection probability models for bacteria, and how to obtain them from heterogeneous spiking data. An application to Bacillus anthracis*, International Journal of Food Microbiology **241**: 78-88.
DOI: http://dx.doi.org/10.1016/j.ijfoodmicro.2016.10.005

IV. **Hedell R.**, Andersson M.G., Faverjon C., Marcillaud-Pitel C., Leblond A., Mostad P., (2017), *Surveillance of equine diseases through implementation of a Bayesian spatio-temporal model: an example with neurological syndromes and West Nile Virus*, (Work in progress).

Additional papers not included in this thesis:

Albinsson L., Norén L., **Hedell R.**, Ansell R., (2011), *Swedish population data and concordance for the kits PowerPlex ESX 16 System, PowerPlex ESI 16 System, AmpFlSTR NGM, AmpFlSTR SGM Plus and Investigator ESSplex*, Forensic Science International: Genetics **5**: e89–e92.

Hedell R., Nordgaard A., Ansell R., (2011), *Discrepancies between forensic DNA databases*, Forensic Science International: Genetics Supplement Series **3**: e135-e136.

. Nordgaard A., **Hedell R.**, Ansell R., (2012), *Assessment of forensic findings when alternative explanations have different likelihoods—"Blame-the-brother"-syndrome*, Science and Justice **52**: 226–236.

Norén L., **Hedell R.**, Ansell R., Hedman J., (2013), *Purification of crime scene DNA extracts using centrifugal filter devices*, Investigative Genetics **4**:8.

Forsberg C., Wallmark N., **Hedell R.**, Jansson L., Ansell R., Hedman J., (2015), *Reference material for comparison of different adhesive tapes for forensic DNA sampling*, Forensic Science International: Genetics Supplement Series **5**: 454-455.

Sidstedt M., Romsos E.L., **Hedell R.**, Ansell R., Steffen C.R., Vallone P.M., Rådström P., Hedman J., (2017), *Accurate digital PCR quantification of challenging samples applying inhibitor-tolerant DNA polymerases*, Analytical Chemistry **89**: 1642–1649.

Contents

Chapter 1

Introduction

In this thesis different kinds of detection problems in biology are addressed. This includes detection of DNA sequences in crime scene samples, detection of harmful bacteria in feed and food and detection of epidemical diseases in animal populations. All the problems stem from issues faced by expert authorities in Forensic Science, Veterinary Science and Food Control. As later described in detail, biological data are analysed by these authorities in order to determine which DNA sequences, bacteria types or diseases are present, if any. The detection problems are connected with different scientific hypotheses (here denoted by the letter H). For example:

H_1: The suspect is the source of the DNA
H_2: Someone else than the suspect is the source of the DNA

or

H_1: Genotype G is present in the sample
H_2: Genotype G is not present in the sample

or

H_1: The bacterium *Bacillus anthracis* is present in the material
H_2: The bacterium *Bacillus anthracis* is not present in the material

or

H_1: An outbreak of a neurological disease is ongoing
H_2: An outbreak of a neurological disease is not ongoing

As discussed later, many of the hypotheses above are composite and can thus be broken down into sub-hypotheses.

Due to limited amounts of samples, low quality samples or imperfect methods for detection and classification the raw data may not reveal the true hypothesis (state of nature) directly. For correct and efficient interpretation of biological detection data it is often necessary to use statistical methods, taking the different sources of uncertainty into account. Ideally, the finder of facts is able to assign a probability to each hypothesis, and doing so using formal statistical methods. Using Bayes' theorem the posterior probability $P(H_i|\boldsymbol{y})$ of each hypothesis H_i with data $\boldsymbol{y}$ can be computed exactly or approximately. Required for the computations using the theorem are the prior probabilities $P(H_i)$ and the value of the mass function or probability density function $f(\boldsymbol{y}|H_i)$ for data $\boldsymbol{y}$ for each hypothesis (i.e. the *likelihood* of each hypothesis). With unknown model parameters $\boldsymbol{\theta}$ with prior distribution $\pi(\boldsymbol{\theta}|H_i)$ the (marginal) likelihood $f(\boldsymbol{y}|H_i)$ can be expressed as $\int f(\boldsymbol{y}|\boldsymbol{\theta}, H_i)\pi(\boldsymbol{\theta}|H_i)d\boldsymbol{\theta}$, integrating over all values of $\boldsymbol{\theta}$.

Hence, to assign probabilities to hypotheses, such that the probabilities are aligned with data and prior knowledge, it is important to find appropriate statistical models for the components above, and to use proper computational methods. Ideally, the detection methods used for sample analysis provides data that, when combined with the statistical models, gives substantial support to the correct hypothesis.

Connected to each detection problem are also decision problems for decision makers: e.g. to decide on a verdict that a suspect is guilty of committing a crime, that a batch of food should be destroyed or that disease control actions in an animal population should be initiated. As discussed later, Bayesian decision theory can be used to determine what action is optimal and what detection procedure provides the most useful results for the decision maker, taking economic and social costs for different outcomes into consideration.

In papers I - IV several Bayesian models for analysis of biological detection data, for determining the performance of detection methods, and for deciding on the optimal analysis procedure are improved, developed and implemented.

A general background to the topics of this thesis is given in chapter 2. A background to forensic DNA analysis is given in section 2.1. An introduction to microbiological sampling is given in section 2.2, followed by a general background to syndromic surveillance in section 2.3. A short introduction to Bayesian inference and numerical tools for Bayesian computations is given in section 2.4. Finally, a summary with discussion about the four papers I - IV is given in section 3.1 - 3.4.

Chapter 2

Background

2.1 Forensic DNA analysis

Forensic science can be defined as science done in relation to investigation of crimes or in the evaluation of civil disputes [1, 2]. One such discipline discussed in this thesis is forensic DNA analysis [3]. Routinely, crime scene samples are collected and sent to forensic laboratories for DNA analysis. The results are compared to the DNA profiles of potential sources of the DNA (i.e. the suspects or victims), assessing how much the results speak in favour of or against them as being the source. A brief description of the basis and technology for DNA analysis follows.

The DNA molecule has a double stranded helix structure with the four nucleobases A (adenine), T (thymine), C (cytosine) and G (guanine) as the building blocks of the genetic code. At several locations in the genome there are repeating patterns with 2-7 basepairs of the nucleobases A, T, C, G as the repeating unit (short tandem repeats, STRs). The number of such repeats often differ between individuals and can therefore be used to discriminate one person's DNA from another. By investigating the number of repeats at several locations of the genome a *DNA profile* is obtained; a numerical vector with the number of repeats at the different locations investigated in the genome.

A specific region targeted for DNA analysis is referred to as a genetic *marker*. The variant forms at a genetic marker are referred to as *alleles* (e.g. for STR markers the alleles are determined by the number of repeats). The markers selected for human forensic DNA analysis are found in regions between the genes, i.e. in the regions not coding for any protein. As humans have two sets of chromosomes, one maternal and one paternal, it is possible to have two different alleles at each marker. In such a case the individual is *heterozygous* at that marker (e.g. alleles 11/13), otherwise *homozygous* (e.g. alleles 13/13).

A method known as Polymerase Chain Reaction (PCR) is used to analyse specific DNA sequences [3]. In this reaction the target DNA sequences are copied in a chain reaction process producing a large number of copies of the sequences enabling them to be detected. There are several important components in the PCR reaction: the extracted DNA sample; primers, which "find" the DNA markers; nucleobase building blocks for the A,C,T,G bases; DNA polymerase that adds the nucleobases to the correct positions; and fluorescence dyes, attached to the DNA molecules to enable detection. By altering the temperature in cyclic patterns more and more copies of the target DNA sequences are created from the mix of substances. First the temperature is increased to typically about 94°C where the DNA stands separate. Then the temperature is lowered to about 60°C where the primers bind to the DNA. Increasing the temperature to 72°C the building of new DNA copies takes place. This process is repeated a number of times (e.g. 30) to create a large number of copies of the target DNA sequences. If the process runs with maximum efficiency, the total number of DNA copies is doubled in each cycle.

As the non-DNA content of the sample may inhibit the reaction, those substances should ideally be removed prior to the PCR analysis. This step is known as *DNA extraction*. Different methods are available to extract the DNA and remove much of the non-DNA substances. Typically, with the purification of DNA there is also some loss of DNA molecules.

To detect the DNA fragments after the PCR run a method known as capillary electrophoresis (CE) is often used. A sub-sample of the PCR product is injected into the CE instrument and the molecules are detected using lasers. The end result is a diagram, known as an electropherogram, with peaks that represent how strong the fluorescence signals are from the different markers analysed. The fluorescence intensity is correlated with the number of DNA copies. The positions of the peaks in the electropherogram are used to determine what alleles are present. Typically, additional smaller peaks are also present that are due to different sources of noise. To avoid overlap between markers in the electropherogram different fluorescence dyes are used in the PCR analysis.

Due to e.g. very low initial DNA concentrations (*low-template DNA*), degradation of the DNA, inhibitory substances in the sample or due to sampling effects it is possible that some of the target DNA sequences that were originally present in the sample are not detected. Failure to detect one or several alleles in the sample is known as allele *dropout* [3, 4].

The crime scene sample may be affected by DNA degradation due to exposure to, e.g. oxygen, heat and ultraviolet radiation [3]. The larger the analysed DNA fragment, the more likely it is to be damaged by degradation, causing allele dropout. In order to enhance the analysis, the sample may be processed in different ways. For instance, if the DNA concentration is low, the sample may be analysed in replicate to increase the possibility of allele detection and

reduce the uncertainty about the DNA profile [4, 5]. For degraded samples, smaller STR allele fragments can be chosen for analysis [6].

Sometimes the DNA analysis is affected by minor, sporadic DNA contaminations causing one or perhaps a few extra alleles to be detected. Alleles that are due to such sporadic contaminations, unlikely to be detected in a repeated analysis, are known as *drop-in* alleles.

In addition to peak heights, dropout and drop-in rates there are several other quality measures used for assessing the PCR performance. Important examples are the heterozygote balance: the peak height ratio of the two alleles in a heterozygous marker; stutter ratios: the peak height ratio of a neighbouring artefact "false" allele and the true allele; fluorescence saturation and bleed-through: non-representative peak heights and false peaks caused by an exaggerated number of DNA copies in the CE.

Another method to detect the DNA, different from CE, is real time quantitative PCR (qPCR) [3]. In qPCR the fluorescence intensity is measured in every PCR cycle, in contrast to conventional PCR with CE where the fluorescence intensities are measured after the last cycle. An initial DNA analysis using qPCR can be used to determine the sample quality in terms of DNA concentration, DNA degradation and presence of inhibitory substances, and thereby guiding the scientist about the better analysis procedure.

Given now the DNA profile of a suspect and the electropherogram of the crime scene sample two competing hypotheses are considered:

H_1: The suspect is the source of the DNA

H_2: Someone else than the suspect is the source of the DNA

As genetic relatedness is an important factor in the statistical analysis several sub-hypotheses of H_2 are considered. Typical hypothesis pairs are:

H_1: The suspect is the source of the DNA

H_2: A person unrelated to the suspect is the source of the DNA

and

H_1: The suspect is the source of the DNA

H_2: A full sibling to the suspect is the source of the DNA

The forensic scientist assesses the relative likelihood of these hypotheses given the DNA profile of the suspect and the EPG data [7]. The general mathematical framework for evaluating the strength of evidence is described later. Several statistical models are available taking stochastic effects, such as the probability of dropout and drop-in, into account [7].

The whole DNA analysis process is complex and finding the optimal setting or analysis strategy for a sample is not trivial. The laboratory may choose to analyse the sample in different ways depending on the type and quality of the sample, the importance of the case and the cost and performance of the available analysis methods. Theoretically well-founded guidelines for the choice of analysis method are, however, lacking in most situations. As a starting point, the performance and limitations of PCR analysis for samples with different quality have to be understood. These issues are discussed in paper I and II of the thesis.

2.2 Microbiological sampling

There are many harmful bacteria that for different reasons might be present in food or feed or in the environment. Thus, investigating the presence of such pathogens can be very important for protecting humans and animals. Such cases may be related to feed and food quality regulations, tracing or confirmation/falsification of an outbreak, bioterrorism and forensic examinations.

Typically, only a subset of a batch of food or an area can be investigated, due to limited resources. Samples from the target location or batch may be collected in a heuristic manner, or by using some probabilistic sampling method such as random sampling, systematic sampling or stratified sampling [8].

For detection of bacteria in the samples a number of detection methods are available. A common method is cultivation by using so called agar plates and testing whether colonies of the bacterium appears on them (see below). Another possibility is to use PCR based techniques. The amounts of bacteria in a sample is often defined as the number of colony forming units (CFU). This is the number of viable bacteria that could potentially duplicate and create colonies or infect humans or animals. Prior to analysis with PCR or with plates the sample generally undergoes *pre-enrichment*, where the CFUs are allowed to multiply in a medium ("pre-enrichment broth") suitable for bacterial growth at about 37°C for a number of hours. Thereafter sub-samples are taken for analysis by DNA extraction and qPCR analysis or with agar plates. Using agar plates the sub-sample from the pre-enrichment broth is placed on a plastic plate containing a medium suitable for growth of the specific bacteria. Eventually colonies of the bacterium may appear on the plates, if they are present in the sample.

Even if the bacteria are present in the batch they might not be detected. They might not be present in the samples analysed due to sampling effects, or they are present in the samples but not detected by the subsequent analysis methods. In the latter case the reason can be due to competition from background flora, anti-bacterial substances or PCR inhibitors in the sample, loss of bacteria or DNA in the sub-sampling phases or in the DNA extraction phase. Thus, for correct and efficient evaluation of negative analysis results it

is important to take these different sources of uncertainty into account. For instance, incidences in the US 2001, where letters contaminated with anthrax were sent to several people, gave rise to a large scale sampling campaign that later highlighted the need for sound probabilistic methods for evaluating negative analysis results [9]. It follows that the statistical models used for sample analysis should take the detection probability of the detection method and the distribution of bacteria in the batch into account.

To assess the detection probabilities of the detection method an experiment can be performed where the target material[1] is spiked (i.e. artificially contaminated) with the target bacterium at different concentrations [10, 11, 12]. The number of detections and non-detections at the different concentrations used serve as basis for establishing the detection probabilities.

For safety reasons or due to limited capacity of safety labs the spiking experiments are sometimes performed using a surrogate bacterium instead of the target bacterium. Since the growth characteristics and detection probability can differ between bacterium strains [13] the difference might have to be accounted for in the evaluation of sampling results. As an example, the bacterium *Bacillus cereus* is sometimes used as surrogate for the extremely harmful bacterium *Bacillus anthracis*, the cause of anthrax [14].

Similarly, sometimes results are available from other spiking experiments with other matrices. Again, any important difference in detection probabilities between the matrix types should be accounted for if taking such experimental data into account.

Thus, for efficient use of heterogeneous experimental data, and for evaluation of sampling results, flexible statistical methods has to be developed and implemented. In paper III a modelling framework is presented.

2.3 Syndromic surveillance

Epidemic diseases in human and animal populations cause different kinds of syndromes in the affected individuals. Hence, the occurrence of such syndromes may be evidence of an outbreak of an epidemic disease. The same syndromes may however be explained by other causes or non-epidemic diseases. One important pathogen causing serious health issues in horses and humans is the West Nile virus [15], spread by infected birds via mosquitos. In horses a common syndrome is neurological problems.

In a *syndromic surveillance system* the count of syndromes are analysed statistically, together with other available and relevant information, such as clustering of syndromes in space and time, historical data on the occurrence of syndromes, seasonality of the disease, import risk assessment, etc, to determine

[1] "Materials" is sometimes referred to as "matrices" with "matrix" as the singular form

the probability of a disease outbreak [16, 17]. Alternatively, instead of the posterior probability of an outbreak being reported, an alarm is triggered based on some summary statistics of the syndromic data and a model for the number of syndromes during a non-epidemic period. For instance, an unexpectedly high number of reported syndromes might be indicative of an outbreak [18, 19]. The approach assumes that the false positive and false negative rates of the alarm system have been established as low [18].

An efficient syndromic surveillance system is thus able to detect both common and unusual diseases in an early stage based on the pre-diagnosis data. Regardless of the statistical method used, a decision maker may want to use the conclusions to take disease control actions: e.g. to initiate epidemiological investigations or to initiate vaccination programs, or wait for more data. If the utilities for correct and incorrect actions are available, e.g. based on the monetary cost for different outcomes, as well as the probability of an outbreak, the optimal action can be determined using a Bayesian decision theoretic framework [20].

Important for the enhancement of surveillance systems are more advanced disease spread models, improved computational algorithms and instructive examples based on realistic data. These issues are discussed in paper IV.

2.4 Bayesian inference

Throughout the thesis, statistical inference is performed mainly using Bayesian analysis. Many of the problems dealt with have a hierarchical structure, or have several sources of information and data that should be taken into account. In some cases probabilities of hypotheses (i.e. posterior probabilities) are requested. The flexibility and features of the Bayesian approach thus makes it an attractive choice. The flexible modelling and computational frameworks available, as well as the interpretational ease of results, are other arguments for adopting the approach. Finally, the approach offers an intuitive and relatively straightforward method for decision making.

Ideally, historical data, expert knowledge or logical and physical constraints are available to create informative prior distributions. Alternatively, the robustness of the posterior distributions may be checked by applying different vague prior distributions.

A brief discussion of Bayesian inference and numerical tools for Bayesian computations is given below, with topics relevant for the papers of the thesis. Throughout, it is assumed the probability distributions are proper.

2.4.1 Hierarchical models

Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the mass function or probability density function for data $\boldsymbol{y}$ given parameters $\boldsymbol{\theta}$. The prior distribution for $\boldsymbol{\theta}$ is denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{\mu})$ with the hyper-prior parameters $\boldsymbol{\mu}$ having prior distribution $\pi(\boldsymbol{\mu})$. The full set of variables and parameters together with their distributions and their multi-level dependencies among each other depicts a *Bayesian hierarchical model* [21, 22]. In a regression model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is modelled as a function of some predictor variables $\boldsymbol{x}$. For a concrete example, assume:

$$y_{i,j} \sim N(\alpha + \beta \cdot x_{i,j}, \sigma^2) \tag{2.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The parameters α, β and σ may e.g. be assumed independent with prior distributions $\pi(\alpha)$, $\pi(\beta)$ and $\pi(\sigma)$ respectively. Differences between groups $j = 1, \ldots, m$ could be modelled using *random effect parameters*, depicting a hierarchical model:

$$y_{i,j} \sim N(\alpha + \beta \cdot x_{i,j} + \epsilon_j, \sigma^2) \tag{2.2}$$

with random effects

$$\epsilon_j \sim N(0, \tau^2) \tag{2.3}$$

where τ has prior distribution $\pi(\tau)$. Hence, ϵ_j depicts the deviation from the mean intercept value α for each group $j = 1, \ldots, m$. Note that if the model is updated using Bayes' theorem and observed data $y_{i,j}$ for groups $j = 1, \ldots, m-1$ (i.e. not for group m) the posterior distribution for ϵ_m will typically also be affected due to the dependence on τ. This is a typical feature of a Bayesian hierarchical model. In contrast, if independent intercept parameters are used for all groups a similar "borrowing" of information between groups about the intercepts is not obtained.

2.4.2 Markov chain Monte Carlo

In some cases the posterior distributions of the model can be derived analytically, such as for conjugate models, but in general methods providing approximate results are required. Numerical integration, discretization of the parameter state space or rejection sampling are common methods [22]. Another common and general strategy, useful even for complex problems, is via Markov chain Monte Carlo (MCMC) simulation [21, 22, 23]. Assuming the model have unknown parameters $\boldsymbol{\theta}$ an initial vector of starting values $\boldsymbol{\theta}^{(0)}$ is chosen (e.g. arbitrarily) and a sequence of draws $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, ..., $\boldsymbol{\theta}^{(N)}$ are simulated via some of the available MCMC algorithms. Typically the draws are not independent. A feature of the simulation algorithms is that the empirical distribution of samples

converges (in distribution) to the true posterior distribution. In each iteration a new vector is sampled based on a transition distribution and the set of parameter values from the previous iteration, forming a Markov chain. The chain should satisfy several mathematical conditions in order to ensure convergence to the target posterior distribution:

- irreducible: for any state of the chain it should be possible to reach any other state.
- aperiodic: the chain must not get trapped in cycles.
- non-transient: it should always be possible to return to a state.
- the chain should have the target distribution as its stationary distribution.

As the number of iterations increases the draws behave more and more as (dependent) draws from the target posterior distribution. Due to sampling variation, auto-correlation and correlation between parameters a large number of iterations may be required to explore the parameter state space satisfactory. It is customary to discard some of the initial simulated values, assuming they are not from the stationary distribution.

In general there is no way of knowing the required number of iterations for when convergence can safely be assumed. However, a number of informal methods for checking non-convergence have been developed. A common recommendation is to run several MCMC simulations in parallel with substantial difference in their starting values $\boldsymbol{\theta}^{(0)}$. The different chains should eventually converge to the same distribution. Informally this can be checked in different ways. Often it is checked that some monitoring statistics are stable and similar between chains. Graphical inspections via time series plots of the sampled values (i.e. the *traceplots*) or via histograms of the sampled values are important tools. In both cases, the chains should be similar in terms of their distribution, assuming convergence. Another approach is via comparison of the between- and within variance of the different chains. Such methods are discussed e.g. by Gelman et. al. [22]. Another possibility is via standard hypothesis testing of equality between chains.

The more iterations and chains used and the more methods used for checking convergence, the more the scientist may become convinced that non-convergence is unlikely and that the chain has indeed converged.

One of the MCMC algorithms available is the *Gibbs sampler*. Denote the parameter vector as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, where k is the number of parameters and let $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$. Hence, $\boldsymbol{\theta}_{-i}$ is the parameter vector with all parameters except θ_i (or more general, all parameters except a subset of them). In each iteration $t = 1, 2, \ldots, N$ the Gibbs sampler cycles through the conditional density functions $\pi(\theta_i|\boldsymbol{\theta}_{-i}^{(t-1)}, \mathbf{y})$; $i = 1, \ldots, k$, and samples a value

from each of them using e.g. conjugacy, if possible, or some other sampling method, such as rejection sampling. The software OpenBUGS [24] implements algorithms for Gibbs sampling.

Another software for MCMC simulations is Stan [25], implementing algorithms for *Hamiltonian Monte Carlo* sampling [22]. The method utilizes gradient information about the parameter space. In many cases this allows for more efficient sampling compared to Gibbs sampling, with less sampling issues caused by correlated variables.

2.4.3 Model comparison and model check

Some common methods for model comparison in a Bayesian setting are briefly mentioned here.

A measure sometimes useful for comparing two models is the *Bayes factor* [22, 26, 27]. Assuming the two models are labeled as H_1 and H_2, and that their prior probabilities are positive, the Bayes factor is defined as:

$$\mathrm{BF}_{1,2} = \frac{P(H_1|\mathbf{y})/P(H_2|\mathbf{y})}{P(H_1)/P(H_2)} = \frac{f(\mathbf{y}|H_1)}{f(\mathbf{y}|H_2)} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta}, H_1)\pi(\boldsymbol{\theta}|H_1)d\boldsymbol{\theta}}{\int f(\boldsymbol{y}|\boldsymbol{\theta}, H_2)\pi(\boldsymbol{\theta}|H_2)d\boldsymbol{\theta}} \tag{2.4}$$

Hence, the Bayes factor expresses how the relative probability of H_1 to H_2 changes with the observed data. The ratio $f(\mathbf{y}|H_1)/f(\mathbf{y}|H_2)$ is known as the *likelihood ratio*. Sometimes the expression *likelihood ratio* is used only when there are no unknown parameters $\boldsymbol{\theta}$. If $P(H_1) + P(H_2) = 1$ the Bayes factor is the change in *odds* of H_1 to H_2. This framework is often used e.g. in Forensic statistics [7, 28]. The two hypotheses then depict the prosecutor's and the defence's position respectively, as mentioned earlier. Ideally, after the forensic scientist have computed the Bayes factor, or the likelihood ratio, the court or the police combines it with their prior odds to find the posterior odds. The Bayes factor or likelihood ratio may be reported together with an ordinal *scale of conclusions*, assigning verbal expressions to the numerical results [28, 29].

Clear evidence against a single model may sometimes be detected by comparing simulated data from the model to the observed data, or by comparing some summary statistics of the two kinds of data. The comparison is made either using the prior predictive distribution or the posterior predictive distribution. The data used for comparison is any observed data left out from the model fitting ("cross-validation") or the same data as used for model fitting [22, 23].

Another measure used for model comparison is the *log pseudomarginal likelihood* [21]. Assuming the data is measured using a common scale for all models and that the same sampling scheme is used it is defined as:

$$\text{LPML}_i = \sum_{j=1}^{n} \log(f(y_j|\mathbf{y}_{-j}, H_i)) \tag{2.5}$$

for model H_i, and with data vector $\mathbf{y}_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n)$. When comparing models the one with the higher LPML is preferred. A related measure is the Leave-One-Out Information Criterion (LOOIC), recently discussed and implemented by Vehtari et al. [30].

Another common statistic used in Bayesian analysis is the Deviance Information Criterion (DIC), a measure sometimes useful to determine what model has the better trade-off between model fit and model complexity [22, 23]. It is based on the deviance statistic

$$D_i(\boldsymbol{\theta}) = -2\log(f_i(\boldsymbol{y}|\boldsymbol{\theta})) + C \tag{2.6}$$

for model i and some constant C. Measuring the model fit by the expected value $E(D_i(\boldsymbol{\theta})|\boldsymbol{y})$ and model complexity by $p_{i,D} = E(D_i(\boldsymbol{\theta})|\boldsymbol{y}) - D_i(\hat{\boldsymbol{\theta}})$ for posterior mean $\hat{\boldsymbol{\theta}}$, DIC is defined as

$$\text{DIC}_i = E(D_i(\boldsymbol{\theta})|\boldsymbol{y}) + p_{i,D} \tag{2.7}$$

2.4.4 Decision theory

Assume the set $H_1, \ldots, H_N$ of hypotheses represent the possible states of nature in a scientific problem. A decision maker may have to act as if one of the hypotheses is true (c.f. the examples given in chapter 1). Denote the action of deciding on hypothesis $1, \ldots, N$ by $a_1, \ldots, a_N$ respectively. Assume that for $i = 1, \ldots, N$ and $j = 1, \ldots, N$ each combination (a_i, H_j) of actions and hypotheses is assigned a utililty $U(a_i, H_j)$. The utilities can have positive or negative values. Thus, if action a_i is taken when hypothesis H_j is true utility $U(a_i, H_j)$ is obtained (e.g. measured in monetary units).

In the Bayesian framework, given data $\mathbf{y}$ and probabilities $P(H_j|\mathbf{y})$, optimal decisions are obtained by choosing the action $a_1, \ldots, a_N$ that maximizes the *expected utility*[31]:

$$E(U(a_i|\mathbf{y})) = \sum_{j=1}^{N} U(a_i, H_j)P(H_j|\mathbf{y}) \tag{2.8}$$

Using the equation above as basis for optimal decisions, rules for determining optimal analysis strategies e.g. in forensic DNA analysis can be derived, as shown in paper II.

Chapter 3

Summary of papers

3.1 Paper I

Title: *Enhanced low-template DNA analysis conditions and investigation of allele dropout patterns.*

In this paper different measures of EPG quality are considered to find optimised settings for PCR cycle number and CE injection time for the ESX 16 analysis kit. 30–35 PCR cycles are applied to find the cycle number where allele detection is optimal and drop-in alleles and bleed-through peaks has a low impact. For each PCR product, three different CE injection times (5, 10 and 20 s) are applied. Mock crime scene DNA extracts of different quantities are prepared and analysed, including samples containing just a few DNA molecules and samples almost generating complete DNA profiles. In addition to dropout and drop-in rates, heterozygote balances, peak heights and stutter ratios are monitored. For dropout rates, heterozygote balances and peak heights regression models are created, using the DNA amount, PCR cycle number and CE injection time as predictor variables. The dye colour and marker type are also included as predictors for the dropout models. As the data is highly structured, with each of the generated PCR products being analysed using 5, 10 and 20 s in CE injection, Bayesian hierarchical models with random effects are used taking the dependencies between results into account. Thus, using the fitted models the scientist is able to predict the changes in EPG quality with altered analysis settings.

Irrespective of DNA amount, the dropout probability seems not to be affected by increasing the number of PCR cycles beyond 33 or by elevating the CE injection time. The results for 33, 34 and 35 PCR cycles indicate that heterozygote balance and stutter ratio are mainly affected by DNA amount, and

not by PCR cycle number and CE injection settings. 32 and 33 PCR cycles with 10 CE injection seconds were judged to be the optimal LT-DNA analysis conditions, maximising detection of true alleles while minimising the risk for problematic artefact peaks and bleedthrough. It is adviced using a similar systematic approach when optimising the detection limit for another STR kit.

Differences in the risk for allele dropout are found between several STR markers. If amplification efficiencies were equal for all markers, no impact of marker or fluorescent label would be expected. The validity of the results also for casework samples is tested. The dropout rates are computed for several casework DNA extracts, low in DNA and analysed with 30 and 32 cycles in replicates. The results agree with the predictions from the statistical models.

3.2 Paper II

Title: *Determining the optimal forensic DNA analysis procedure following investigation of sample quality*

In this paper, it is shown how guidelines for the choice of analysis method of forensic samples can be created using Bayesian decision theory. The theory is applied to forensic DNA analysis, showing how the information from the initial qPCR analysis can be utilized. Clearly, a crucial factor when choosing method for analysis is that the results should be useful in court. Hence, a holistic approach is taken, connecting the results of the initial qPCR analysis of the crime scene samples with decisions about guilt or source. The expected value of different PCR analysis strategies are shown, using Bayesian decision theory and statistical modelling. It is assumed the alternatives for DNA analysis are 1) using a standard assay, 2) using the standard assay and a complementary assay, or 3) the analysis is cancelled. The costs for DNA analyses and costs for erroneous conclusions are two of the factors that have to be considered. Although the theory is applied to forensic DNA analysis, the framework is general and could be applied to other forensic disciplines as well.

One of the main features of the approach is how the DNA concentration and level of degradation, both estimated in the initial qPCR analysis, affect the probability of allele dropout. A regression model for the risk of dropout is derived, based on experimental data, and a simulation algorithm for the distribution of likelihood ratios is described. The likelihood ratios are simulated under each of the hypotheses of the case. Their distributions reveal how likely it is to obtain ”meaningful” evidence, i.e. with the potential to affect the court’s decision. Maps are produced showing the optimal DNA analysis strategy for different qPCR results. An important feature of the method is that the decision maps can be created prior to the DNA analysis, avoiding the need for real time simulations. Therefore, in terms of computational efforts, increasing the

complexity of the models is a viable option, e.g. by allowing for mixtures (DNA from more than one person).

Some ideas for estimation of the costs and utilities connected to the analysis are discussed. The importance of sensitivity analyses for different values of the model parameters are stressed. In the working example, the sensitivity analysis indicated no major differences in decisions when changing the parameter values moderately.

3.3 Paper III

Title: *Detection probability models for bacteria, and how to obtain them from heterogeneous spiking data. An application to Bacillus anthracis*

In this paper the evaluation of negative analysis results and the modelling of detection probability curves for bacteria are discussed. The latter is a crucial part in the evaluation of negative analysis results, together with a model for the distribution of bacteria in the target area or target batch. It is shown step-by-step how such detection models can be created and how different sources of data can be included, including results from both PCR analyses and analyses with agar plates, with different bacteria strains and matrix types. The theory is applied to evaluation of samples with *Bacillus anthracis* as the target bacterium, using *Bacillus cereus* as surrogate bacterium in the spiking experiments for the target matrix. Other available spiking data with *Bacillus anthracis* and *Bacillus cereus* is included in the analysis, to learn about the model parameters.

Two different modelling approaches, differing in whether the pre-enrichment step and the PCR detection step are modelled separately or together, are applied. There are some potential advantages with dividing the statistical modelling into a pre-enrichment step and a detection step. If the pre-enrichment protocol or the analytical detection protocol is changed in some way then parts of the model can be kept and do not have to be re-evaluated; e.g. if the PCR protocol is changed the posterior distributions of the pre-enrichment phase parameters could be kept as prior distributions for the new setup. If the knowledge in one part or the other of the model is good, a smaller experiment might be performed to assess the complete model rather than performing a more exhaustive experiment. It is also possible to use data from more types of experiments in order to further reduce the uncertainties in the detection probabilities, e.g. using data from quality assurance experiments specifically made to study the PCR performance (without pre-enrichment).

The different candidate models are compared via their log pseudo-marginal likelihood (LPML), and checked with observed data using the posterior predictive distributions. The statistical checks of the final models do not suggest any overall misfit.

The relative importance on the detection curves for various existing data sets are evaluated and illustrated. It is shown how extrapolating the information from all the experiments with the surrogate bacterium *Bacillus cereus* together with the data for *Bacillus anthracis* for all but one matrix gives similar results as using the complete set of data.

3.4 Paper IV

Title: *Surveillance of equine diseases through implementation of a Bayesian spatio-temporal model: an example with neurological syndromes and West Nile Virus*

In this paper a Bayesian method for early detection of disease outbreaks in animal populations is developed and implemented. Based on reported neurological syndromes in horses, connected e.g. with West Nile Virus (WNV), the probability for an outbreak is computed. The basic model structure and computational algorithms are general and could be applied also to other scenarios for animal or human disease surveillance as well. A discrete spatio-temporal model is developed, defining the spatial units by grid cells and counting the number of reported neurological syndromes in horses for each week in France. The magnitude of syndromes for a non-outbreak period is estimated using available syndromic data from 2006–2016. Based on known WNV outbreaks a disease spread model is derived and the expected number of syndromes for a WNV-like outbreak is estimated. Our model allows a cell to be unaffected by a disease causing neurological syndromes, being affected by a WNV-like disease, or being affected by a non-WNV like disease causing neurological syndromes. The probability for each of these states is estimated using a Gibbs sampling procedure. It is described how the simulations can be performed using known conditional distributions for all unknown parameters. For the disease status variables a Forward filtering Backward smoothing algorithm is described and implemented.

The use of the models for several outbreak scenarios are examplified. Important quality measures such as the sensitivity and specificity of the surveillance system are presented. Results are given for different threshold decision rules, based on the probability of an outbreak. The number of instances with high probability for an outbreak despite not being affected is relatively small. Many outbreaks are detected, in the sense of yielding a high probability of outbreak, within 2 weeks after the index case, and almost all within 4 weeks. A novel and relatively simple model for the spread of a disease between non-neighbours is described, enabling faster detection of outbreaks that have spread a long distance. In a limited example, the number of disease related syndromes for the coming week is predicted. The results can be used to concretely make decision makers aware about possible future events, if no disease control actions are taken.

Bibliography

[1] Eckert W.G., (ed), (1980), *Introduction to Forensic Sciences*, The C. V. Mosby Company, St. Louis, p. 9.

[2] Inman K., Rudin N., (2001), *Principles of Practice of Criminalistics - The Profession of Forensic Science*, CRC Press, Boca Raton, p. 15.

[3] Butler J.M., (2011), *Advanced topics in forensic DNA typing: methodology*, Elsevier Academic Press, San Diego.

[4] Gill P., Whitaker J., Flaxman C., Brown N., Buckleton J., (2000), *An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA*, Forensic Science International 112: 17–40.

[5] Kloosterman A.D., Kersbergen P., (2003), *Efficacy and limits of genotyping low copy number (LCN) DNA samples by multiplex PCR of STR loci*, Journal De La Societe De Biologie 197: 351–359.

[6] Butler J.M., Shen Y., McCord B.R., (2003), *The development of reduced size STR amplicons as tools for analysis of degraded DNA*, Journal of Forensic Sciences 48: 1054–1064.

[7] Buckleton J.S., Triggs C.M., Walsh S.J., (eds), (2004), *Forensic DNA evidence interpretation*, CRC Press, Boca Raton.

[8] Jongenburger I., Reij M.W., Boer E.P.J., Gorris L.G.M., Zwietering M.H., (2011), *Random or systematic sampling to detect a localised microbial contamination within a batch of food*, Food Control 22: 1448–1455.

[9] United States Government Accountability Office, (2006), *Anthrax detection: agencies need to validate sampling activities in order to increase confidence in negative results*, GAO-05-251.

[10] Koyuncu S., Andersson M.G., Häggblom P., (2010), *Accuracy and Sensitivity of Commercial PCR-Based Methods for Detection of Salmonella enterica in Feed*, Applied and Environmental Microbiology 76: 2815-2822.

[11] Krämer N., Löfström C., Vigre H., Hoorfar J., Bunge C., Malorny B., (2011), *A novel strategy to obtain quantitative data for modelling: combined enrichment and real-time PCR for enumeration of salmonellae from pig carcasses*, International Journal of Food Microbiology 145: 86-95.

[12] Koyuncu S., Häggblom P. (2009), *A comparative study of cultural methods for the detection of Salmonella in feed and feed ingredients*, BMC Veterinary Research 5: 6.

[13] De Siano T., Padhi S., Schaffner D.W., Montville T.J., (2006), emph-Growth characteristics of virulent Bacillus anthracis and potential surrogate strains, Journal of Food Protection 69: 1720–1723.

[14] Fricker M., Ågren J., Segerman B., Knutsson R., Ehling-Schulz M., (2011), *Evaluation of Bacillus strains as model systems for the work on Bacillus anthracis spores*, International Journal of Food Microbiology 145: 129–136.

[15] Leblond A., Hendrikx P., Sabatier P., (2007), *West Nile virus outbreak detection using syndromic monitoring in horses*, Vector-Borne and Zoonotic Diseases, 7: 403–410.

[16] Heaton M.J., Banks D.L., Zou J., Karr A.F., Datta G., Lynch J., Vera F., (2012), *A spatio-temporal absorbing state model for disease and syndromic surveillance*, Statistics in Medicine 31: 2123–2136.

[17] Dórea F.C., Vial F., (2016), *Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011–2016)*, Veterinary Medicine: Research and Reports 7: 157–170.

[18] Lawson A., Kleinman K., (Eds.), (2005), *Spatial and Syndromic Surveillance for Public Health*, Wiley & Sons, Ltd, Chichester.

[19] Faverjon C., Andersson M.G., Decors A., Tapprest J., Tritz P., Sandoz A., Kutasi O., Sala C., Leblond A., (2016), *Evaluation of a Multivariate Syndromic Surveillance System for West Nile Virus*, Vector-Borne and Zoonotic Diseases 16: 382–390.

[20] Andersson M.G., Faverjon C., Vial F., Legrand L., Leblond A., (2014), *Using Bayes' Rule to Define the Value of Evidence from Syndromic Surveillance*, PLOS ONE 9: e111335.

[21] Christensen R., Johnson W., Branscum A., Hanson T.E., (2011), *Bayesian Ideas and Data Analysis*, CRC Press, Boca Raton.

[22] Gelman A., Carlin J.B., Stern H.S., Rubin D.B., (2003), *Bayesian Data Analysis, second edition*, Chapman and Hall, London.

[23] Congdon P., (2010), *Applied Bayesian Hierarchical Methods*, CRC Press, Boca Raton.

[24] Lunn D., Spiegelhalter D., Thomas A., Best N., (2009), *The BUGS project: evolution, critique, and future directions*, Statistics in Medicine 28: 3049–3067.

[25] Stan Development Team, (2016), *RStan: the R interface to Stan.* http://mc-stan.org

[26] Berger J.O., (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.

[27] Kass E.R. Raftery A.E., (1995), *Bayes Factors*, Journal of the American Statistical Association 90: 773-795.

[28] Aitken C., Taroni F., (2004), *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester.

[29] Nordgaard A., Ansell R., Drotz W., Jaeger L., (2011), *Scale of conclusions for the value of evidence*, Law, Probability & Risk 11: 1-24.

[30] Vehtari A., Gelman A., Gabry J., (2016), *Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC*, Statistics and Computing 27: 1413–1432.

[31] Taroni F., Bozza S., Biedermann A., Garbolino P., Aitken C., (2010), *Data analysis in forensic science: A Bayesian decision perspective*, John Wiley & Sons, Chichester.