

DOCTORAL DISSERTATION

Statistical methods for the analysis of high-throughput proteomic and genomic data

Doctoral dissertation submitted to obtain the degree of Doctor of Science: Statistics, to be defended by

Fatemeh Zamanzad Ghavidel

Promoters: Prof. Dr Tomasz Burzykowski | UHasselt / tUL





and statistical Bioinformatics

Acknowledgements

This work would not have been possible without the encouragement and support of a number of people who were directly and indirectly involved in the completion of my PhD. This is the only way to express my sincere gratitude towards them.

First and foremost, I would like to express my special thanks to my promoter, Prof. dr. Tomasz Burzykowski. Dear Tomasz, thank you for your constant support, invaluable suggestions, and inspiring guidance throughout the course of my PhD. I have benefited a lot from your keen supervision. I extend my gratitude to my nonofficial co-promoter, Dr. Dirk Valkenborg. Dear Dirk, thank you for your continuous encouragement, helpful discussions, valuable suggestions and never-ending patience for my uncertainties.

I would like to acknowledge my co-authors for providing valuable comments and having useful discussions during my research. I am also grateful for the constructive feedback and suggestions I received from the jury members to an earlier version of my thesis.

I would like to thank all my colleagues in CenStat and my special thanks go to my officemates in offices C107b and E129. It has been a pleasure to share the office with you. Trishanta and Yimer, thank you for creating a pleasant and enjoyable working atmosphere. Thanks Jürgen for all your remarks during my PhD and helping me with the Dutch Samenvatting.

I owe my sincere thanks to two of my best friends, Reza and Amin. The friendship and company of you made my stay in Belgium especially unforgettable.

I would like to express special gratitude to my beloved parents, Parvaneh and Javad. Thank you for you lifelong and heartening support and infinite inspiration.

Finally, my deep appreciation is extended to my brother, Amir, without whom I

would have never gone so far. Thank you Amir for always being beside me and your spiritual support. I cannot imagine being the person I am today without such a great brother like you.

Thank you all! Fatemeh Zamanzad 12 December, 2014 Diepenbeek

Contents

| Ta | able | of Contents | iii | |
|---------------|--|---|---|--|
| Li | st of | Publications and Reports | vi | |
| \mathbf{Li} | st of | Abbreviations | $\mathbf{i}\mathbf{x}$ | |
| 1 | The | The focus and content of the dissertation | | |
| 2 | Cor | nsidered datasets | 3 | |
| | 2.1 | Bovine Cytochrome C mass spectra | 3 | |
| | 2.2 | Whole-cell lysate of $Caenorhabditis \ elegans$ tandem mass spectra | 5 | |
| | | 2.2.1 LC-MS setup | 5 | |
| | | 2.2.2 Database searching | 6 | |
| | 2.3 | Whole genome sequencing data of ethanol-tolerant $S.$ cerevisiae strains | 6 | |
| | | | | |
| | | | | |
| Ι | Sta | atistical Methods for Proteomics | 9 | |
| I 3 | Sta | atistical Methods for Proteomics | 9 11 | |
| I 3 | Sta Intr 3.1 | atistical Methods for Proteomics roduction Amino acids | 9 11 12 | |
| I 3 | Sta Intr 3.1 3.2 | atistical Methods for Proteomics roduction Amino acids . Proteins . | 9 11 12 12 | |
| I 3 | Sta Intr 3.1 3.2 3.3 | atistical Methods for Proteomics roduction Amino acids Proteins Protein structure | 9 11 12 12 13 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids | 9 11 12 12 13 15 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids | 9 11 12 12 13 15 15 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids Proteins Protein structure Principle of mass spectrometry 3.4.1 Ionization source 3.4.1.1 | 9 11 12 12 13 15 15 16 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids | 9 11 12 12 13 15 15 16 17 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids Proteins Protein structure Principle of mass spectrometry 3.4.1 Ionization source 3.4.2 Mass analyzer 3.4.3 | 9 11 12 12 13 15 16 17 19 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 | atistical Methods for Proteomics roduction Amino acids Proteins Proteins Protein structure Protein structure Protein structure 3.4.1 Ionization source 3.4.1.1 Matrix-assisted laser-desorption ionization (MALDI) 3.4.2 Mass analyzer 3.4.3 Ion detector Tandem mass spectrometry Ion | 9 11 12 12 13 15 15 16 17 19 19 | |
| I 3 | St : Intr 3.1 3.2 3.3 3.4 3.5 3.6 | atistical Methods for Proteomics roduction Amino acids Proteins Protein structure Principle of mass spectrometry 3.4.1 Ionization source 3.4.1.1 Matrix-assisted laser-desorption ionization (MALDI) 3.4.2 Mass analyzer 3.4.3 Ion detector Tandem mass spectrometry Collision-induced dissociation (CID) | 9 11 12 13 15 16 17 19 19 19 | |

| | | 3.7.1 | Peptide identification | 20 |
|----------|----------------|----------|---|-----------|
| | | | 3.7.1.1 De novo sequencing algorithms | 21 |
| | | | 3.7.1.2 Database search algorithms | 21 |
| | | 3.7.2 | Search engines | 21 |
| | | | 3.7.2.1 SEQUEST | 22 |
| | | | 3.7.2.2 MASCOT | 22 |
| | | 3.7.3 | Error sources of peptide identification | 22 |
| | | 3.7.4 | Post-processors for improving identification | 23 |
| | 3.8 | Isotopi | c distribution | 23 |
| 4 | Cor | npariso | on of the Mahalanobis distance and Pearson's χ^2 statistic | |
| | as r | neasure | es of similarity of isotope patterns | 27 |
| | 4.1 | Introdu | uction | 27 |
| | 4.2 | Experi | mental data | 29 |
| | 4.3 | Metho | dology | 30 |
| | | 4.3.1 | Prediction of the isotopic distribution | 31 |
| | | 4.3.2 | Similarity metric | 31 |
| | 4.4 | Results | 5 | 33 |
| | 4.5 | Conclu | usions | 37 |
| 5 | The | e use of | f the isotopic distribution as a complementary quality | |
| | met | ric to a | assess tandem mass spectra results | 39 |
| | 5.1 | Introdu | uction | 39 |
| | 5.2 | Materi | als and methods | 40 |
| | 5.3 | Implen | nentation | 40 |
| | 5.4 | Results | 5 | 43 |
| | | 5.4.1 | Region I | 45 |
| | | 5.4.2 | Region II | 46 |
| | | 5.4.3 | Region III | 49 |
| | | 5.4.4 | Region IV | 49 |
| | 5.5 | Conclu | usions | 52 |
| 6 | \mathbf{Ass} | essing t | the agreement between peptide assignments for different | |
| | sear | rch eng | ines | 55 |
| | 6.1 | Introdu | uction | 55 |
| | 6.2 | Implen | nentation | 56 |
| | 6.3 | Results | 5 | 58 |
| | 6.4 | Conclu | usions | 61 |

| Π | St | atistical Methods for Gene Mapping | 63 |
|----|------|---|-----|
| 7 | Intr | oduction | 65 |
| | 7.1 | Chromosomes, DNA, and genes | 65 |
| | 7.2 | Genetic variation and meiotic recombination | 66 |
| | 7.3 | Genetic linkage | 68 |
| | 7.4 | Molecular markers | 70 |
| | | 7.4.1 Single nucleotide polymorphisms | 70 |
| | 7.5 | Construction of genetic maps and QTL analysis | 70 |
| | 7.6 | DNA sequencing | 71 |
| | | 7.6.1 Illumina | 72 |
| | 7.7 | NGS analysis pipeline | 74 |
| 8 | A h | idden Markov-model for QTL-mapping based | |
| | on v | vhole-genome sequencing data | 75 |
| | 8.1 | Hidden Markov-models | 78 |
| | | 8.1.1 Forward-backward algorithm and parameter estimates \ldots . | 80 |
| | 8.2 | A hidden Markov-model for BSA experiments | 81 |
| | | 8.2.1 The hidden states $\ldots \ldots \ldots$ | 81 |
| | | 8.2.2 State-dependent emission probabilities | 81 |
| | 8.3 | Conclusions | 83 |
| 9 | A b | i-directional (dual) hidden Markov-model for QTL-mapping | 85 |
| | 9.1 | A bi-directional (dual) hidden Markov-model | 86 |
| | 9.2 | A dual hidden Markov-model for BSA experiments | 88 |
| | 9.3 | Results | 88 |
| | | 9.3.1 Uni-diectional HMMs and filtering | 88 |
| | | 9.3.2 Chromosome XIV, pool 1 | 90 |
| | | 9.3.3 Chromosome XIV, pool 2 | 92 |
| | | 9.3.4 Chromosome IX, pool 1 | 95 |
| | | 9.3.5 Chromosome IX, pool 2 | 98 |
| | | 9.3.6 Comparison of the DHMM with the basic HMM $\ldots \ldots \ldots$ | 102 |
| | 9.4 | Conclusions | 103 |
| 10 | A n | on-homogeneous hidden Markov-model for QTL-mapping | 105 |
| | 10.1 | An introduction to non-homogeneous Markov-models | 106 |
| | | 10.1.1 Forward-backward algorithm and parameter estimates | 106 |
| | 10.2 | A non-homogeneous hidden Markov-Model for BSA experiments $\ . \ . \ .$ | 107 |

v

| 10.3 Results | 107 | | | |
|---|-----|--|--|--|
| 10.3.1 Chromosome XIV, pool 1 | 107 | | | |
| 10.3.2 Chromosome XIV, pool 2 | 112 | | | |
| 10.3.3 Chromosome IX, pool 1 | 114 | | | |
| 10.3.4 Chromosome IX, pool 2 | 116 | | | |
| 10.4 Conclusions | 118 | | | |
| 11 A joint hidden Markov-model for QTL-mapping | 121 | | | |
| 11.1 Methodology | 122 | | | |
| 11.2 A joint hidden Markov-Model for BSA experiments | 123 | | | |
| 11.3 Results | 123 | | | |
| 11.3.1 Chromosome XIV | 123 | | | |
| 11.3.2 Chromosome IX | 126 | | | |
| 11.3.3 Chromosome II | 129 | | | |
| 11.4 Conclusions | 131 | | | |
| 12 Concluding remarks and future work | 133 | | | |
| 12.1 Concluding remarks | 133 | | | |
| 12.2 Topics for future work | 135 | | | |
| 12.2.1 Assessing the agreement between peptide assignments for dif- | | | | |
| ferent search engines | 135 | | | |
| 12.2.2 The hidden Markov-model | 136 | | | |
| Appendices 1 | 149 | | | |
| A.1 Tables | 151 | | | |
| A.2 Figures | 154 | | | |
| A.3 Nielsen <i>et al.</i> | 160 | | | |
| Samenvatting 161 | | | | |

List of Publications and Reports

The material presented in Part I has been based on the following publications and reports:

- Ghavidel, F.Z., Claesen, J., Burzykowski, T., and Valkenborg, D. (2014) 'Comparison of the Mahalanobis distance and Pearson's χ² statistic as mea- sures of similarity of isotope patterns'. Journal of the American Society for Mass Spectrometry, 25(2), 293-296.
- Ghavidel, F.Z., Mertens, I., Baggerman, G., Laukens, K., Burzykowski, T., and Valkenborg, D. (2014) 'The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results'. *Journal of Proteomics*, 98, 150-158.
- Ghavidel, F.Z., Laukens, K., Bittremieux, W., Burzykowski, T., and Valkenborg, D. (2014) 'Cohen's kappa: a concordance measure for peptide assignments between different search engines'. *Under revision*.

The material covered in Part II has been based on the following reports:

- Ghavidel, F.Z., Claesen, J., and Burzykowski, T. (2014) 'Bi-directional hidden Markov-model for gene-mapping based on whole-genome sequencing data'. *Submitted to BMC Genetics.*
- Ghavidel, F.Z., Claesen, J., and Burzykowski, T. (2014) 'Non-homogeneous hidden Markov-model for gene-mapping based on whole-genome sequencing data'. *Submitted to Journal of Computational Biology, accepted.*

• Ghavidel, F.Z., Claesen, J., and Burzykowski, T. (2014) 'Joint hidden Markovmodel for gene-mapping based on whole-genome sequencing data'. *Technical report*.

List of Abbreviations

| bp | base pair |
|-------|---|
| BRAIN | Baffling recursive algorithm for isotopic distribution calculations |
| CID | collision-induced dissociation |
| Da | Dalton |
| DDA | data-dependent acquisition |
| DHMM | dual hidden Markov-model |
| DNA | deoxyribonucleic acid |
| EM | expectation-maximization |
| EI | electron impact |
| ESI | electrospray ionization |
| FAB | fast atom bombardment |
| FD | field desorption |
| FDR | false discovery rate |
| FWHM | Full width at half maximum |
| HMM | hidden Markov-model |
| IUPAC | International Union of Pure and Applied Chemistry |
| LC-MC | Liquid-chromatography mass spectrometry |
| LDA | linear discriminant analysis |
| MALDI | matrix-assisted laser-desorption ionization |
| m/z | mass-to-charge ratio |
| mRNA | messenger RNA |

| MS | mass spectrometry |
|----------------|-------------------------------------|
| MS/MS | tandem mass spectrometry |
| NGS | next generation sequencing |
| NH-HMM | non-homogeneous hidden Markov-model |
| PEP | posterior error probability |
| \mathbf{PSM} | peptide-spectrum match |
| QTL | quantitative trait locus |
| SNP | single nucleotide polymorphism |
| SVM | support vector machine |
| TOF | time-of-flight |
| | |

Chapter 1

The focus and content of the dissertation

Over the last two decades, a number of technologies have emerged to profoundly advance the efficiency of biological and cellular investigation. The data explosion caused by these techniques has given rise to the range of scientific -omics subfields such as genomics, transcriptomics, proteomics, and metabolomics, representing a focus on the use of large-scale information of the subject under the study. For instance, proteomics, a more recent -omics term was first proposed by Marc Wilkins in 1995 [1], denoting the large-scale study of the structure and function of proteins. Highthroughput 'omics' technologies including micro-arrays, next-generation sequencing technology, mass spectrometry and many more methods, have been broadly applied. Mass spectrometry-based proteomics has become established as the method of choice for protein identification and quantification. Rapid advances in next-generation sequencing technology have led to an increase in the amount of genomic information. The advent of high-throughput technologies in genomics and proteomics promoted to the development of novel statistical methods for handling and analyzing enormous amounts of complex data being produced to extract important information regarding biological processes.

In this dissertation, we propose several methodologies for handling problems that come up when analyzing proteomics and genomics data. The focus of the first part is on statistical methods for proteomics.

In Chapter 3, we give a short introduction to proteomics, mass spectrometry, tandem mass for protein/peptide identification, and database searching methods. In

the following two chapters two applications of the isotopic distribution, an important characteristic of mass spectrometry-based proteomics, are introduced. In Chapter 4, we illustrate how different similarity measures perform to discern noise from a true signal by applying features of the isotopic distribution. Consequently, the best performed similarity measure is chosen and used in Chapter 5, in which the accuracy of the protein/peptide identification by search engines in shotgun proteomics setting is validated with the isotopic distribution patterns. Chapter 6 compares the statistical agreement between the two most widely used search engines, MASCOT and SEQUEST, in terms of peptide assignment/identification.

In the second part, the focus shifts to genomics, more explicitly the genetic dissection of phenotypic trait based upon next-generation sequencing data. Chapter 7 gives an introduction to genomics, linkage analysis, genetic map, and sequencing methods. In Chapter 8, we present the overview of the hidden Markov-model and describe a hidden Markov-model for gene-mapping. The methods presented in Chapter 8 are extended in the subsequent three chapters. In particular, we investigate the application of the bi-directional hidden Markov-model and non-homogeneous hidden Markov-model for genetic dissection of the phenotypic traits in Chapter 9 and Chapter 10, respectively. In Chapter 11, an extension of the hidden Markov-model for identification of causal genes is proposed, which uses multiple related experiments.

To conclude, in Chapter 12, we discuss the proposed methods and introduce topics for further research.

Chapter 2

Considered datasets

The methods for the analysis of proteomic and genomic data, considered in this dissertation, are applied to a number of data sets. These data sets will be briefly described in this chapter.

2.1 Bovine Cytochrome C mass spectra

Bovine Cytochrome C is a relatively small protein related to mitochondria in a cell. It is a chain of 105 amino acids:

MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFSYT DANKNKGITWGEETLMEYLENPKKYIPGTKMIFAGIKKKGEREDLIAYLKKA TNE.

A peptide mixture of tryptic-digested bovine Cytochrome C was purchased from LC Packings and mixed with five internal standards from Laser BioLabs used for the calibration of the mass spectrometer. According to the data sheets of the suppliers, the bovine Cytochrome C tryptic-digest and internal standard mixture contains 17 protein fragments. The amino acid sequences and theoretical monoisotopic masses (m) of these fragments are known and are presented in Table 2.1. The tryptic-digested bovine Cytochrome C and internal standards were mixed with the matrix molecules and automatically spotted 384 times on one stainless-steel plate by a robot. The plate was processed on a 4800 MALDI-TOF/TOF analyzer (Applied Biosystems) mass spectrometer, which resulted in 384 mass spectra. Figure 2.1 illustrates a full scan of mass spectrum of a tryptic digest of bovine Cytochrome C.

These mass spectra are primarily used for the comparison of the similarity measures described in Chapter 4.

| Bovine Cytochrome C (CC) | | | | | |
|----------------------------|--------------------------------|-------------------------|--|--|--|
| nr | Sequence | Monoisotopic mass (m) | | | |
| CC1 | IFVQK | 633.38445 | | | |
| CC2 | YIPGTK | 677.37428 | | | |
| CC3 | MIFAGIK | 778.44059 | | | |
| CC4 | KYIPGTK | 805.46924 | | | |
| CC5 | EDLIAYLK | 963.52715 | | | |
| CC6 | TGPNLHGLFGR | 1167.61434 | | | |
| $\rm CC7$ | GEREDLIAYLKK | 1433.78728 | | | |
| CC8 | TGQAPGFSYTDANK | 1455.66248 | | | |
| CC9 | KTGQAPGFSYTDANK | 1583.75744 | | | |
| CC10 IFVQKCAQCHTVEK | | 1632.81107 | | | |
| $\rm CC11$ | GITWGEETLMEYLENPK | 2008.94465 | | | |
| $\rm CC12$ | GITWGEETLMEYLENPKK | 2137.03961 | | | |
| Internal standards (IS) | | | | | |
| $\mathbf{n}\mathbf{r}$ | Sequence | Monoisotopic mass (m) | | | |
| IS1 | RPPGF | 572.30653 | | | |
| IS2 | DRVYIHPF | 1045.53397 | | | |
| IS3 | ZLYENKPRRPYIL | 1671.90508 | | | |
| IS4 RPVKVYPNGAEDESAEAFPLEF | | 2464.19051 | | | |
| IS5 | FVNQHLCGSHLVEALYLVCGERGFFYTPKA | 3493.67346 | | | |

Table 2.1: Peptides from a bovine Cytochrome C tryptic digest and internal standards.



Figure 2.1: Full scan mass spectrum of bovine Cytochrome C peptides.

2.2 Whole-cell lysate of *Caenorhabditis elegans* tandem mass spectra

The nematode *Caenorhabditis elegans* is a model organism that has seen extensive use over the last four decades in multiple areas of investigate. The free-living nematode C. elegans is one of the best-studied multicellular model organisms. C. elegans is used in Chapter 5 and 6 for peptide identification. For this purpose, C. elegans N2 (CGC, University of Minnesota) nematodes were grown on NGM plates seeded with OP50 strains of Escherichia coli (CGC, University of Minnesota). Mixed staged worms were collected in M9 buffer and sucrose-floated to remove all bacteria. The worms were then lysed in 50 mM ammonium bicarbonate (pH 7.8) using a Branson Cell Disrupter (Branson Ultrasonics Corp., USA) for 5 times 20 s followed by 60 s ice incubation. The lysate was centrifuged 10 min at 4000 rpm at 4° C to remove cell debris. Subsequently, the sample was centrifuged 10 min at 14,000 rpm at $4^{\circ}C$ to separate soluble lysate from insoluble lysate. The sample was then denaturated using 0.1% Rapigest (Waters Corp.) and boiled at 100° C for 5 min. Protein concentration was measured using the Pierce BCA Protein assay (Thermo Scientific). Proteins were reduced using 5 mM DTT for 45 min at 56° C and alkylated using 15 mM IAA for 15 min in the dark. Samples were digested using trypsin gold (Promega, USA) at 37°C overnight at a substrate to enzyme ratio 50:1. To remove Rapigest, samples were treated with 200 mM HCl.

2.2.1 LC-MS setup

The samples were analyzed on a Thermo Scientific LTQ Velos Orbitrap mass spectrometer. Proteolytic digests were separated on an Eksigent nanoLC system using a C18 reverse phase column (Dionex Acclaim pepmap 100, 3 μ m particles, 75 μ m i.d.× 150 mm). A gradient length of 150 min was used (350 nl/min of 2-35% acetonitrile in 0.1% formic acid). For data dependent acquisition, the method was set to fragment the top 10 most intense ions observed in the MS scan using CID. The nanoLC was interfaced to the Velos LTQ orbitrap (Thermo scientific) by means of an advion nanomate with LC-coupler. Spraying voltage was set to 1.8 kV.

The mass spectrometer was operated in the data-dependent mode, switching automatically between orbitrap MS and LTQ Velos MS/MS. Survey full scan spectra were acquired from m/z 300 to 2000 in the orbitrap with resolution of 60,000 at m/z 400. One million charges were accumulated in the linear iontrap for analysis in the orbitrap. Most intense ions, up to a maximum of 10 per MS1, were sequentially isolated in the iontrap for collision induced dissociation. Fragment ions were analyzed in the iontrap; up to 100,000 charges were accumulated. To obtain sub 3 *ppm* mass accuracy in MS1, the lock mass option was activated and the polydimethylcyclosiloxane (PCM) ions, generated in the electrospray process from ambient air (protonated (Si(CH₃)₂O))6; m/z-445.120025), were used for internal recalibration in real time.

2.2.2 Database searching

The spectra were interpreted by Proteome Discoverer 1.3 as a workflow manager. The 20,581 tandem MS data were searched using both SEQUEST and MASCOT against a *C. elegans* protein database (wormpep229.fasta) downloaded at ftp://ftp.sanger.ac. uk/pub/databases/wormpep/wormpep229/. All tandem mass spectra in the range of 300 Da to 10,000 Da were interpreted. Monoisotopic peak assignment, charge state determination, co-isolation interference, and mass difference between the measured and theoretical monoisotopic masses were determined by Proteome Discoverer. Precursor mass tolerance was set at 5 *ppm*, while fragment mass tolerance was set to 0.8 Da. A maximum of five missed cleavages by trypsine was allowed for. A static modification of 57.021 Da on cysteine was defined to allow for carbamidomethylation. Further, a dynamic modification of 15.9955 Da was introduced to account for possible oxidation of methionine. The use of average precursor masses and average fragment masses was prohibited. It is worth noting that only first ranked PSMs were considered for further analysis, i.e., only one sequence annotation for a fragment ion mass spectrum.

The charge state and monoisotopic mass of the precursor ion for each tandem MS spectrum were included in a target list and were exported from Proteome Discoverer to a comma-separated-value file format. The full scan MS data were transformed into ASCII MS1-file format by using RawXtract 1.9.9.2 downloaded from the Yates-lab [2].

2.3 Whole genome sequencing data of ethanoltolerant *S. cerevisiae* strains

Saccharomyces cerevisiae, a single-celled eukaryote yeast, is an intensively studied model organisms in molecular and cell biology. It is the most traditionally used yeast strain in food processing and industry, specifically as a fermenter of alcoholic beverages. High ethanol tolerance is one of the most prominent characteristic of this organism. S. cerevisiae is used in Chapter 9, 10, and 11 for the mapping of quantitative trait loci involved in tolerance to high ethanol levels (16% and 17%). For this intention, a highly ethanol-tolerant yeast strain was crossed with a laboratory strain (without the trait) of a moderate ethanol tolerance resulting in 5,974 viable haploid yeast cells. Haploid offspring was screened for high ethanol tolerance, first in a medium containing 16% of ethanol producing 136 ethanol-tolerant segregants and subsequently with 17% ethanol giving rise to 31 segregants.

The genomic DNA from both pools, 16% (pool 1) and 17% ethanol (pool 2), and the parent strains with high ethanol tolerance were submitted to a pooled-segregant genome-wide sequencing analysis by means of high-throughput next generation sequencing (NGS) using the Illumina/Solexa NGS technique. The technique produced DNA sequences with a length of 40 to 100 basepairs. These reads are subsequently aligned to a DNA sequence of the parental laboratory yeast strain (without the trait) and single nucleotide polymorphism (SNPs) as genetic markers are identified. In this experiment, the bulk segregant analysis (BSA, [3]) is combined with NGS to allow simultaneous identification of markers.

For each identified SNP, the chromosomal position, the number of sequencing events (reads), and the number of times nucleotides A, C, G, and T were present in the offspring were recorded. The larger the proportion of differences in terms of the nucleotides (mis-match/SNP frequency) between the offspring and the parental strain, the higher the chance of a presence of a trait-related gene in the vicinity of the chromosomal location.



Figure 2.2: The mis-match frequency for SNPs on chromosome XIV, pool 1.

Part I

Statistical Methods for Proteomics

Chapter 3

Introduction

The term proteomics refers to the study of the proteome [1, 4], the complete set of proteins including their (post-translational) modifications, produced by an organism or a cellular system. A single protein can be synthesized by a cell in different forms and with different modifications, consequently, thousands of genes can produce up to millions of proteins [5]. The large increase in protein diversity makes proteomics to be a complex study. The complexity of the proteomics research requires new technologies and new analytical protocols for sample preparation, protein detection and subsequent data analysis.

Over the last decade, high-throughput proteomics technologies have evolved rapidly. This event has resulted in the broadening of applications and potential uses of proteomics, most importantly in identification and quantification of proteins. Nowadays, proteomics has numerous applications, including: (i) study of post-translational modifications; (ii) protein-protein interactions; (iii) structural proteomics; (iv) functional proteomics; (v) computational proteomics, and many more.

3.1 Amino acids

Amino acids are the basic building blocks of peptides and proteins. They are carbon compounds that contain two functional groups: an amino group (N-terminus) and a carboxylic acid group (C-terminus). A side chain (usually denoted as R) attached to the compound gives each amino acid a unique set of characteristics. Figure 3.1 illustrates a structure of a typical amino acid. The key chemical elements of an amino acid are carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S). Each possible set of three nucleotides (codon) in the DNA encodes for one of the twenty amino acids. However, a specific amino acid can also be coded by several codons. The twenty possible amino acids are summarized in Table 3.1.



Figure 3.1: Basic chemical structure of an amino acid.

The sequence of amino acids in a protein can be defined by the sequence of nucleotides on the DNA strain. DNA is first transcribed to a messenger-RNA, which is further translated to a protein.

3.2 Proteins

Proteins (also known as polypeptides) are the most diverse class of biomolecules. They have evolved to accomplish many tasks in living systems. They can serve as enzymes, structural materials (e.g., keratin), specific binding (e.g., antibodies), substance carrier such as hemoglobin, and many more.

| Name | TLC | OLC | Formula | Mass (Dalton) |
|---------------|----------------------|--------------|----------------------|---------------|
| Glycine | Gly | G | $C_2H_5NO_2$ | 75.07 |
| Alanine | Ala | Α | $C_3H_7NO_2$ | 89.09 |
| Valine | Val | V | $C_5H_{11}NO_2$ | 117.15 |
| Leucine | Leu | \mathbf{L} | $C_6H_{13}NO_2$ | 131.18 |
| Isoleucine | Ile | Ι | $C_6H_{13}NO_2$ | 131.18 |
| Serine | \mathbf{Ser} | \mathbf{S} | $C_3H_7NO_3$ | 105.09 |
| Threonine | Thr | Т | $C_4H_9NO_3$ | 119.12 |
| Cysteine | \mathbf{Cys} | \mathbf{C} | $C_3H_7NO_{2S}$ | 121.15 |
| Methionine | Met | Μ | $C_5H_{11}NO_{2S}$ | 149.21 |
| Phenylalanine | Phe | \mathbf{F} | $C_9H_{11}NO_2$ | 165.19 |
| Tyrosine | Tyr | Υ | $C_9H_{11}NO_3$ | 181.19 |
| Tryptophan | Trp | W | $C_{11}H_{12}N_2O_2$ | 204.23 |
| Proline | Pro | Р | $C_5H_9NO_2$ | 115.13 |
| Asparagine | Asn | Ν | $C_4H_8N_2O_3$ | 132.12 |
| Glutamine | Gln | \mathbf{Q} | $C_5 H_{10} N_2 O_3$ | 146.15 |
| Aspartic acid | Asp | D | $C_4H_7NO_4$ | 133.1 |
| Glutamic acid | Glu | Е | $C_5H_9NO_4$ | 147.13 |
| Lysine | Lys | Κ | $C_6 H_{14} N_2 O_2$ | 146.19 |
| Histidine | His | Η | $C_6H_9N_3O_2$ | 155.16 |
| Arginine | Arg | R | $C_6H_{14}N_4O_2$ | 174.2 |

Table 3.1: 20 amino acids found in nature. The abbreviation TLC and OLC stand for Three Letter Code and One Letter Code, respectively.

3.3 Protein structure

Proteins are polymers of 20 different amino acids that are covalently joined together by peptide bonds. The sequence of the different amino acids in a protein, is its primary structure (Figure 3.2a). The primary structure determines how the protein folds into higher level structures. The secondary structure of the polypeptide chain refers to the spatial arrangement of amino acid residues. The most common types of secondary structure are the α -helix and the β -pleated sheet. The secondary structure formed through regular hydrogen-bonding interactions between NH and C=O groups of the protein back bone (Figure 3.2b). Protein tertiary structure refers to a protein's geometric shape (Figure 3.2c). Several proteins are formed by association of the folded chains of more than one polypeptide; this results in the quaternary structure of a protein (Figure 3.2d). Hemoglobin, the oxygen carrying component of blood, is an example of a protein in a quaternary structure.



Figure 3.2: Levels of protein structure. (a) primary structure, (b) secondary structure, (c) tertiary structure, and (d) quaternary structure. Taken from [6].

3.4 Principle of mass spectrometry

Mass spectrometers are devices, that measure the mass-to-charge ratio m/z (in Dalton (Da)) and intensity of ions. It is used for determining masses of particles and the elemental composition of molecules. The data generated by a mass spectrometer can be represented by a mass spectrum: a two-dimensional representation of signal intensity (ordinate) versus m/z (abscissa), see Figure 3.3. When applied in proteomics, an MS is an important method to identify and characterize proteins [7–11]. It allows separating peptide/protein molecules by their different masses. In its simplest form, a mass spectrometer consists of three parts, which are essential for their function; the ion source, the mass analyzer, and the detector. In the next subsections, we briefly discuss the different parts of the mass spectrometer used in the first part of this dissertation.



Figure 3.3: Graphical representation of the MS data.

3.4.1 Ionization source

The role of the ionization source is to generate atomic or molecular gas-phase ions. There are different techniques that can be used for this purpose, however, the type of ionization technique substantially depends on the disposition of the sample and the type of information required from the analysis. Ion sources can be broadly grouped into two classes. Hard ionization techniques, such as electron impact (EI), chemical ionization (CI), and field ionization (FI), result in breaking the molecule in a sample partly or completely into fragment ions during the ionization. On the contrary, soft ionization techniques, such as fast atom bombardment (FAB), field desorption (FD), matrix-assisted laser-desorption ionization (MALDI), and electrospray ionization (ESI), do not break up the molecule in a sample and provide m/z information on the intact molecule. The technique used in Chapter 4 of this dissertation, MALDI, will be briefly described below.

3.4.1.1 Matrix-assisted laser-desorption ionization (MALDI)

MALDI is suitable for the ionization and the analysis of large molecules $(> 1 \ kDa)$. It was developed by Tanaka [12] in the 1980s. For a typical MALDI analysis, the analyte substance is mixed with a high amount of matrix compound in a solution and spotted onto a stainless steel target plate in an array format. The matrix contains small organic molecules with an absorbance capacity at the laser wavelength. After evaporation and introduction of the target into the vacuum region of a mass spectrometer, the crystals formed on the target plate are illuminated with a pulsed laser beam. The energy from the laser pulse causes the matrix/analyte mixture to melt, evaporate, and eventually ionize resulting in the formation of charged ions.

The matrix, therefore, plays a key role by strongly absorbing the laser light energy and causing, indirectly, the analyte to vaporize. The ions generated by MALDI carry only a single charge, which makes the molecular weight determination straightforward. The basic principle of MALDI is depicted in Figure 3.4.



Figure 3.4: Basic principle of Matrix-Assisted Laster Desorption/Ionization (MALDI), re- produced from [13].

3.4.2 Mass analyzer

Once ions have been generated in the ion source, the role of the mass analyzer is to separate them according to their mass-to-charge-ratio (m/z). Mass analyzers use electric and magnetic fields to apply a force on charged ions. One of the mass analyzers used is the time-of-flight (TOF) mass analyzer, primarily interfaced with an MALDI ionization source. We briefly present the basic principle of the linear Time-Of-Flight Mass Spectrometer (TOF MS), which is schematically depicted in Figure 3.5.

Essentially, a TOF mass analyzer consists of an ion source, an acceleration region, a drift tube and a detector. After production of ions by MALDI a fixed potential difference (typically 20 to 30 kV) accelerates all the ions into a tube where TOF separation occurs. As all the ions are accelerated with the same potential, they all have the same kinetic energy. The linear TOF analyzer works by measuring the time required for ions generated in the source to fly through the tube and hit the detector at the other side. The principle is based on an ion of mass m leaving the ionization source with a charge z and accelerating potential V, thus having energy zV equal to the kinetic energy of the ion:

$$k = zV = \frac{mv^2}{2}. (3.1)$$

If the time taken, t, for the ion to fly the distance d of the flight tube at velocity V is given by

$$t = \frac{d}{V}.$$
(3.2)

Substituting (3.2) into (3.1) gives

$$t^{2} = \frac{m}{z} \left(\frac{d^{2}}{2V}\right). \tag{3.3}$$

The terms in parentheses (related to a fixed distance and accelerating potential) remain constant, thus m/z can be determined from t^2 . As all the ions are accelerated with the same potential then they all have the same kinetic energy. Because the ions have the same energy, yet a different mass, the ions reach the detector at different times. The smaller ions reach the detector first because of their greater velocity while the larger ions take longer due to their larger mass.

The key parameters of mass analyzers are sensitivity, mass resolution, and mass accuracy [14]. The sensitivity characterizes the ability of the mass analyzer to detect weak signals. Mass resolution and mass accuracy describe how well the analyzer is able to separate peaks with similar mass and how accurately it measures this mass, respectively. The mass resolution is a dimensionless unit and is expressed as the ratio of the mass of a signal peak in a mass spectrum and its Full-Width-At-Half-Maximum



(FWHM). The FWHM of a peak is illustrated in Figure 3.6.

Figure 3.5: Basic principle of a linear time-of-flight mass spectrometer. Taken from [15].



Figure 3.6: Illustration of the Full-Width-At-Half-Maximum (FWHM) of a peak. Taken from [13].

The mass accuracy is defined as the observed difference between the observed mass of an analyte and the expected mass:

$$mass\ accuracy = |\ mass_{observed} - mass_{expected} | . \tag{3.4}$$

The mass accuracy is often expressed in parts per million (ppm):

$$ppm = 10^6 \times mass \ accuracy/mass_{expected}.$$
(3.5)

To give an example, if we know that the expected mass of a compound is 1000,0 m/zand our mass spectrometer measures a signal for this compound at 999,99 m/z, then the accuracy of this measurement was $\simeq 10 \ ppm$.

3.4.3 Ion detector

The last component of the mass spectrometer instrument is the ion detector, which records the ions separated according to their mass to charge ratios. The role of the detector is to register the number of ions produced for each m/z, by detecting their impact on the detector surface. As a result, the mass spectrometer generates an output with two variables, the mass- to-charge ratio (m/z) and the corresponding intensity value.

3.5 Tandem mass spectrometry

Tandem MS (also called MS/MS or MS2) is used to determine the amino acid sequences of the peptide (identification or characterization) and provide detailed structural information. Tandem (MS/MS) mass spectrometers are instruments composed out of two (sometimes three) successive mass analyzers for which only the last mass analyzer has an ion detector. The first mass analyzer is used to select a particular m/z value (precursor ion). The selected ions pass through a region when they are activated and causes them to fall apart to produce fragments. The resulting MS/MS spectrum consists of product ions from the selected precursors.

3.6 Collision-induced dissociation (CID)

CID is the most common fragmentation method [16]. In CID, molecular ions are accelerated by an electrical potential and then allowed to collide with neutral atoms or molecules such as helium, nitrogen or argon. The collision converts some of the initial kinetic energy of the molecular ions into internal energy, causing chemical bonds to break. Two types of fragment ions, b- and y-ions, are commonly observed in MS/MS spectra obtained by CID fragmentation.

3.7 The shotgun proteomics workflow

An explicit goal of proteomics is to characterize all the proteins expressed in a cell or tissue. The improvements in MS instruments, protein and peptide separation techniques, and the availability of protein sequence databases for many species has facilitated the analysis of complex protein mixtures using shotgun proteomics. Shotgun proteomics is a powerful technology to study the protein population of a biological system. This approach generates high-throughput data in complex mixtures using a combination of LC with tandem MS. It is broadly used for the large-scale peptide and protein identification [14]. Shotgun proteomics is currently the dominant analytical approach in proteomics research. The typical workflow of a shotgun proteomics can be described as follows: The first step is to digest sample proteins into peptides. The digestion is typically done by using a specific protease (enzyme) that will cleave the protein sequence into peptides. The most common protease used for the protein digestion is trypsin which cleaves proteins on the carboxyl-terminal side of the arginine and lysine residues. The so-obtained peptide mixture is complex by nature and is therefore separated by LC to reduce the complexity. Subsequently, the separated peptides are injected into the mass spectrometer and are measured in a data-dependent acquisition (DDA) mode. In this acquisition mode, a predefined number of the most intense parent masses are selected from a full scan mass spectrum for a second interrogation by tandem mass spectrometry. During tandem MS, a selected parent ion, i.e., peptide, enters a collision cell and is fragmented in a pattern that is a characteristic for a particular amino acid sequence. The stream of data that is produced by this approach is interpreted by using computational tools.

3.7.1 Peptide identification

Numerous computational tools have been developed to support high-throughput peptide and protein identification by assigning sequences to tandem MS spectra [17]. In general, the peptide identification algorithms using tandem MS can be roughly categorized into two main paradigms : (i) *de novo* sequencing algorithms, and (ii) database search algorithms.

3.7.1.1 *De novo* sequencing algorithms

De novo sequencing algorithms obtain peptide sequences directly from the MS/MS spectrum by interpreting the mass differences between the generated MS/MS fragment ion sequence [18, 19]. These algorithms do not need a priori sequence information and hence can potentially identify protein sequences that are not available in a protein database. In recent years, many de novo sequencing algorithms and software packages were published. The most widely used de novo sequencing packages include PepNovo [20, 21] and PEAKS [22, 23]

3.7.1.2 Database search algorithms

In this approach, peptide identification is performed by correlating experimental tandem MS spectra with theoretical spectra predicted for each peptide contained in a protein sequence database. Many different algorithms have been developed for identifying tandem MS data using database search engines, including SEQUEST [24], MASCOT [25], X!Tandem [26], OMSSA [27], and ProbID [28]. All database search engines operate in a similar manner and follow the same general framework [29, 30]. Acquired tandem MS spectra are compared and correlated against theoretical spectra constructed for each database search algorithm that satisfies a certain set of database search parameters, i.e., mass tolerance, enzyme constraint, and types of posttranslational modifications, specified by the user. A scoring scheme is then used to measure the degree of similarity between the experimental tandem MS spectra and theoretical fragmentation patterns. Candidate peptides are ranked according to the computed score, and the highest scoring peptide sequence (best match) is selected for further analysis.

3.7.2 Search engines

Database searching algorithms remain the most efficient and widely used method for peptide identification. As a consequence, the computational analysis typically starts with database searching algorithms, and if needed, for example, *de novo* sequencing tools are applied to the remaining unassigned spectra. Our focus here is the database search algorithms, as they are the most relevant to the research presented in this dissertation. The main difference between different search algorithms is the scoring function used to quantify the degree of similarity between the acquired tandem mass spectrum and the candidate peptides retrieved from the database. As a result, they differ from one another in terms of speed, accuracy, sensitivity, and false positives (i.e., incorrectly identified peptides). The performance of database search algorithms have been compared in [17]. A detailed review of all different scoring schemes goes beyond the scope of this chapter; however the following discussion will focus on brief explanation of the database search tools that are used in this dissertation.

3.7.2.1 SEQUEST

SEQUEST [24] is one of the most widely used algorithms for database searching. It scores peptide sequences by the cross-correlation between the intensities of peaks on the observed and the theoretical spectra. The cross-correlation score (Xcorr) is computed as follows:

$$X corr = R_0 - \left(\sum_{t=-75}^{t=+75} R_t\right) / 151,$$
(3.6)

where $R_t = \sum_{i=1}^n x_i y_{i+t}$, x_i and y_i are the intensities of the peaks at location *i* (shifted along the m/z axis by *t* mass units) in the observed and theoretical spectra, respectively. For each experimental spectrum, the best scoring peptide assignment (highest Xcorr score) is kept for further analysis. In addition to Xcorr, the relative difference between the best and the second best Xcorr score, ΔC_n , is also computed and measures how different the top value is from the next best match. Generally, high values of ΔC_n are regarded as a correct assignment with respect to the top Xcorr value.

3.7.2.2 MASCOT

MASCOT [25], another commonly used database search algorithm, performs the probability-based searches of peptide database sequences by an extension of the MOWSE algorithm [31]. MASCOT estimates the probability of a match occurring by chance. An ion score is reported as $-10log_{10}(p)$, where p is the absolute probability. A higher score indicates a more confident match.

3.7.3 Error sources of peptide identification

All tandem MS database search tools return the best matching peptide found in the database for each acquired spectrum, what we call peptide-spectrum match (PSM), except when there are no candidate peptides in the searched database. However, the best match reported by the database search tool is not necessarily correct [32–34]. The main reasons why the database search tools fail to assign correct peptide sequences and a large fraction of the top ranked peptides are still wrong include:
- Deficiencies of the scoring scheme
- Low MS/MS spectrum quality
- Incorrectly determined charge state or peptide mass
- Restricted database search

Thus, the matches in the database search need to be further evaluated to limit false positives identifications [32, 33, 35].

3.7.4 Post-processors for improving identification

Different approaches have been developed to validate peptide assignments resulting from search algorithms. PeptideProphet and Percolator are the most commonly used post-processing computational tools, which attempt to improve the discrimination performance between correct and incorrect PSMs. PeptideProphet [33], originally developed to analyze SEQUEST search results, automatically validates peptide assignments to MS/MS spectra made by database search programs. PeptideProphet utilizes a machine learning algorithm called linear discriminant analysis (LDA) model to re-score PSMs sampled from a mixture distribution which represents the chance of a correct PSM and an incorrect PSM. The distribution of the correct and incorrect PSMs can be characterized. PeptideProphet applies an expectation maximization (EM) algorithm to generate a posterior error probability (PEP) for each PSM being a correct peptide identification.

Percolator [36] is an alternative post-processing software relying on target/decoy database search results to infer the q-value and PEPs. This system employs a semi-supervised machine learning method that iteratively trains a linear support vector machine (SVM) [37] classifier to discriminate between target and decoy PSMs.

3.8 Isotopic distribution

Most elements occur in nature as a mixture of isotopes. Isotopes are atom species of the same chemical element that have different masses, i.e., they have the same number of protons, but a different number of neutrons. The number of protons is referred to as the atomic number and determines the chemical element of an atom. Atoms with equal atomic numbers share the same chemical behavior and cannot be distinguished chemically.

As mentioned earlier, a peptide is composed out of amino acids, which are built from five elements: carbon (C, atomic number 6), hydrogen(H, atomic number 1), nitrogen (N, atomic number 7), oxygen (O, atomic number 8) and sulphur (S, atomic number 16). These elements are polyisotopic elements, i.e., they have naturally occurring variants with a different atomic mass (isotope). For example, carbon has two isotopes that occur in nature, ${}^{12}C$, which is comprised of six protons and six neutrons, and ${}^{13}C$, which carries six protons and seven neutrons. The most abundant natural isotope, such as ${}^{1}H$ or ${}^{12}C$, is also called the monoisotope. Note that in the case of these five atoms the monoisotope is also the lightest natural isotopic variant of the atom. Isotopes of each element are found in nature with certain abundance. For instance, the relative abundance of the monoisotopic carbon isotope ${}^{12}C$ is 98.93%, whereas the isotope ${}^{13}C$ has the relative abundance of 1.07%. Masses of atoms are measured in Dalton (Da), or equivalently in unified atomic weight units (u). According to International Union of Pure and Applied Chemistry (IUPAC), one Dalton is defined as 1/12 of the mass of one atom of the ${}^{12}C$ isotope. A list of the standard isotopes for elements specific to peptides is given in Table 3.2, together with their corresponding masses and probability of occurrence. When the atomic composition of a peptide is known, we can calculate the probability of occurrence of a particular isotopic variant using the probability of occurrence of the polyisotopic elements from Table 3.2. Therefore, the isotopic distribution is given by the probabilities of occurrence of all possible isotopic variants of a peptide. In a mass spectrometry, isotopic distributions appear as a series of peaks arising within mass spectra that are from compounds of the same molecular formula, but are different in their atomic isotope composition. An example of this phenomenon is shown in Figure 3.7. These peaks are grouped together with a mass-to-charge spacing of $\approx 1/z$, where z is the charges associated with the molecules during the ionization process.

To calculate the isotopic distribution, we need the information about the chemical composition of the peptide. Given the known chemical composition, the isotopic distribution can then be calculated, e.g., by using a Fourier transform as proposed by Rockwood [38], or the BRAIN algorithm [39–41]. However, the chemical composition of a peptide is often unattainable. As an alternative, the aggregated isotopic distribution can be predicted as a function of the mass. Several approaches [42–45] have been proposed to this aim. In this dissertation, we consider the polynomial approach to predict the distribution from the information about the monoisotopic mass of the peptide as suggested in [45].

Table 3.2: Isotopic variants of C, H, N, O and S as defined by the IUPAC1997 standard.

| Chemical element | Isotope | Atomic mass (ma/u) | Natural abundance (atom%) | | |
|------------------|----------|----------------------|---------------------------|--|--|
| Carbon | ^{12}C | 12.0000000000 | 0.9893 | | |
| | ^{13}C | 13.0033548378 | 0.0107 | | |
| Hydrogen | ^{1}H | 1.0078250321 | 0.999885 | | |
| | ^{2}H | 2.0141017780 | 0.000115 | | |
| Nitrogen | ^{14}N | 14.0030740052 | 0.99632 | | |
| | ^{15}N | 15.0001088984 | 0.00368 | | |
| Oxygen | ^{16}O | 15.9949146 | 0.99757 | | |
| | ^{17}O | 16.9991312 | 0.00038 | | |
| | ^{18}O | 17.9991603 | 0.00205 | | |
| Sulphur | ^{32}S | 31.97207070 | 0.9493 | | |
| | ^{33}S | 32.97145843 | 0.0076 | | |
| | ^{34}S | 33.96786665 | 0.0429 | | |
| | ^{36}S | 35.96708062 | 0.0002 | | |



Figure 3.7: Isotopic disribution of $C_{95}H_{159}N_{27}O_{36}S_1$ for the mass range [2286, 2292], calculated by the BRAIN algorithm [39–41].

The polynomial regression model proposed in [45] allows predicting the first three isotope ratios of an average peptide. The (x + 1)-th isotope ratio for a peptide with monoisotopic mass m is calculated as follows:

$$R(x+1,M) = \frac{H(x+1,M)}{H(x,M)},$$
(3.7)

where x (= 0, 1, 2, ...) indicates the particular isotope variant, with x = 0 denoting the monoisotopic variant, and H(x, m) is the probability of occurrence of the isotope variant x.

The method is based on the linear peptide model proposed in [42] and builds on the average amino acid model developed in [44], i.e., averagine. The linear peptide model spans a particular mass range by concatenating average amino acids. For each of the so obtained theoretical peptides, the isotopic distribution was calculated using the ICP isotope pattern calculator [19]. Subsequently, the resulting isotope ratios were modeled by a 4^{th} order polynomial model in function of the peptide mass in a similar fashion as proposed in [43]

$$R(x+1,m) = \beta_0 + \beta_1(\frac{m}{1000}) + \beta_2(\frac{m}{1000})^2 + \beta_3(\frac{m}{1000})^3 + \beta_4(\frac{m}{1000})^4.$$
(3.8)

Estimates of the parameters β_0 , β_1 , β_2 , β_3 , and β_4 for each isotope ratio were obtained by maximum likelihood estimation. In [45] was also suggested that models fitted to the consecutive ratios produce smaller errors than ratios with the monoisotopic peak as the common reference. This is because the monoisotopic peak is always among the most abundant peaks which would result in larger errors for the ratio estimation if it is taken as the common reference for these ratios. Chapter 4

Comparison of the Mahalanobis distance and Pearson's χ^2 statistic as measures of similarity of isotope patterns

4.1 Introduction

In high-resolution mass spectrometry, proteins and peptides appear in a mass spectrum as a series of locally correlated peaks. This specific characteristic is related to the isotopic distribution of a peptide (Section 3.8). The isotopic distributions contain potentially valuable information that can be used in a wide variety of applications. For example, it can be employed to discern genuine peptide peaks from noise [46, 47], to determine the monoisotopic peak [48, 49], or to study conformational dynamics of peptides and proteins using the hydrogen/deuterium exchange (HDX) [50], where isotopic distributions used to extract information on the isotopic states of exchanging hydrogens. The use of information about the isotopic distribution is not restricted to proteomics alone. In the field of metabolomics, isotope information is even more extensively used for metabolite identification [51] and low-level signal processing [52, 53].

As it has already been mentioned, information about the isotopic distribution can be employed to discern genuine peptide peaks from noise. This is because a series of peptide related peaks in a mass spectrum will express a particular pattern, corresponding to the underlying isotopic distribution for the peptide. On the other hand, noise peaks do not follow any particular shape or pattern [10]. In a spectrum, peptide peaks can be scrutinized by assessing the degree of similarity between the observed pattern of peaks and the isotopic distribution expected for a peptide with a similar mass [54-56]. The idea is illustrated in Figure 4.1. The left-hand-side panel of Figure 4.1 presents an observed series of equally distant peaks, which could be originating from a peptide. The-right-hand-side panel shows the expected isotopic distribution. The observed pattern is compared to the expected isotope distribution. If the computed value of a similarity measure is smaller than a pre-defined threshold, the selected peak is regarded as a genuine monoisotopic peptide peak. To this aim, a similarity measure is needed. Currently, the standard similarity measure is Pearson's χ^2 statistic and it has been rigorously investigated [50]. It is based on a weighted sum of the squared deviations between the expected and observed peaks [42, 43].

However, alternative similarity measures could be considered that can include information about possible correlation between the intensity peaks of an isotope distribution. For this reason we have evaluate the use of Pearson's χ^2 statistic and compare it to the Mahalanobis distance [57]. The latter similarity metric calculates the generalized distance and was described in a seminal paper by J. C. Mahalonobis. In mass spectrometry, the Mahalanobis distance is employed as a metric for outlier detection in the context of data quality assessment and it operates on a particular feature set [58–60]. Additionally, the metric is often included in the object function of machine learning methods as a global distance measure [61] to classify spectral data. Nevertheless, the Mahalonobis distance has never been proposed for the interpretation of the isotope patterns observed in mass spectra. For this purpose, we have conducted a controlled MALDI-TOF experiment on bovine Cytochrome C to evaluate its performance on resolved isotope peaks.



Figure 4.1: The panel on the left-hand side displays an observed isotope pattern. The panel on the right hand side shows a hypothetical isotope distribution. The lightest isotopic variant of a peptide, i.e., the variant which is composed out of ${}^{12}C$, ${}^{1}H$, ${}^{14}N$, ${}^{16}O$, and ${}^{32}S$ atoms, is called the monoisotopic variant. The peak corresponding to the monoisotopic variant is called the monoisotopic peak (indicated with an arrow in the left hand side panel). An isotope distribution can be calculated from the atomic composition and the elemental isotope distribution.

4.2 Experimental data

The method developed in this chapter are illustrated by using the data set described in Section 2.1. First, we focus on the series of four consecutive, 1 Da separated peaks, which are consistently found in more than 90% of the 384 spectra (Table A.1 in the Appendix). We call such series isotopic clusters. The reason for extracting the first four isotope peaks is that the mass of the peptides in the sample are predominantly in the range of 568.1 to 2465.2 Da. Consequently, it is reasonable to assume that the isotopic distributions of these peptides are sufficiently characterized by the first four isotope peaks. In total, 35 of such clusters are selected. For 12 clusters, the mass corresponds to the monoisotopic mass of one of the 17 protein fragments known to be present in the mixture. The additional 23 candidates were found to be related to peptides resulting from modifications or artifacts of the proteolytic background [62]. The heatmap in Figure A.1 in the Appendix indicates the mass location of the peptides in the 384 spectra. The color is an indication for the abundance of the corresponding peptide, with red indicating the highest intensity measurement. Chapter 4. Comparison of the Mahalanobis distance and Pearson's χ^2 statistic 30 as measures of similarity of isotope patterns

noise peaks. Figure 4.2 illustrates the selection of the peptide and noise clusters for a particular spectrum. The selected noise peaks are separated by 1 Da as well, but do not appear consistently across the 384 spectra. The noise peaks are located in mass regions in the neighborhood of the selected peptide isotope clusters. The data from the noise and peptide peak clusters are used as a benchmark to assess the ability of the similarity measures to discriminate noise peaks from peptide peaks.



Figure 4.2: An example for a selected noise cluster and a peptide isotopic cluster (m/z 2138.1).

4.3 Methodology

The comparison between an observed series of peaks with a hypothetical isotope distribution can be performed by considering isotope ratios (Section 3.8). The rationale for working with isotope ratios is that ratios are dimensionless and their use allows us to avoid scaling of the expected and observed intensity values. In addition, ratios are not sensitive to multiplicative noise.

Two components are required in order to compare an observed series of peaks with an hypothetical isotopic distribution. First, a model to predict the expected isotopic distribution and the corresponding expected isotope ratios of an average peptide is needed. Second, the measure to score the similarity between the observed and expected isotope ratios has to be defined. In the next two subsections, we discuss these issues.

4.3.1 Prediction of the isotopic distribution

In the proposed methodology, a model is required to predict the expected isotopic distribution. To this aim, the polynomial regression model described in [45] can be used. However, to define similarity measures that take into account the correlation, an estimate of the variance-covariance structure in the data is required. Alternately, an empirical estimate of the expected isotopic distribution and its corresponding variance-covariance structure can be applied. For this purpose, the Human HUPO database was digested *in-silico* by using trypsine as a protease. The digest led to 126,376 peptides with masses ranging from 400 to 4000 Da. The program BRAIN [39–41] was used to calculate the isotopic distribution and monoisotopic masses of the resulting peptides. For a given mass of m Da, a set of peptides with monoisotopic masses of the resulting to [m-5, m+5] Da was selected. Next, the mean value of isotope ratios was calculated for the peptides within the assumed mass interval and stored in vector R. Additionally, the variance-covariance matrix of the ratios Σ was estimated based on the selected data as well.

4.3.2 Similarity metric

By assessing the similarity between the observed isotope ratios of an isotope cluster in a spectrum and the corresponding expected values, we can decide whether the series of peaks might be generated by a peptide. To this aim, a similarity measure is needed. The standard measure (i.e., Pearson's χ^2 statistic) is defined as follows:

$$\chi^2 = \sum_{i=1} \left(O_i - R_i \right)^2 / R_i, \tag{4.1}$$

where O_i is the observed value of the *i*th consecutive isotopic ratio (i = 1, 2, 3) and R_i is the corresponding expected value. An alternative similarity measure could be the Mahalanobis distance [57]. The distance takes into account the variability and correlation of the ratios, and is defined as follows:

$$M = \{ (\mathbf{O} - \mathbf{R})' \Sigma^{-1} (\mathbf{O} - \mathbf{R}) \}^{1/2},$$
(4.2)

where O and R denote the vectors containing, respectively, the observed and expected consecutive isotope ratios, and Σ denotes the estimated variance-covariance matrix of the expected ratios. The expected ratios and corresponding variance-covariance matrix were calculated using the theoretical isotopic distributions from the Human HUPO database. Note that the expected values can also be computed by the polynomial model, which is more straightforward from a practical point of view. The Chapter 4. Comparison of the Mahalanobis distance and Pearson's χ^2 statistic 32 as measures of similarity of isotope patterns

use of the Mahalanobis distance is motivated by the fact that it takes into account the correlation between isotope ratios, which could allow for a better discriminatory performance. The motivation is illustrated in Figure 4.3. The figure presents the scatter plot of the first and second isotope ratio for peptides with a mass between 2000 Da and 2020 Da from the Human HUPO database. The grey diamonds above the histograms indicate the mean values of 1.1243 and 0.6082 for the first and second isotope ratios, respectively. The plot indicates that there is a substantial amount of correlation between the two ratios. Consider the two points, marked by the black circles. A similarity measure that does not take into account the correlation would more likely classify point 1 as a genuine peptide because its coordinates are close to the mean values. However, a measure taking into account the correlation, such as, e.g., the Mahalanobis distance, would most likely opt for point 2 because the coordinates of this point reflect the (linear) association resulting from the joint distribution of the two isotope ratios.



Figure 4.3: Scatter plot of the first isotope ratio (x-axis) and the second isotope ratio (y-axis) of peptides with a mass between 2000 and 2020 Da based on the Human HUPO database.

4.4 Results

For the selected peptide and noise peak-clusters, we calculate Pearson's χ^2 statistic and the Mahalanobis distance based on the data from the HUPO database. The top panel of Figure 4.4 presents the distribution of the computed Pearsons χ^2 values for the noise and peptide clusters using the data from randomly selected four spectra. The overlap between the distribution of the peptide (green) and the noise peak clusters (red) is small. It suggests that the statistic can reliably distinguish between the noise and peptide clusters. The bottom panel of Figure 4.4 presents the same information for the Mahalanobis distance. In this case, the overlap is much larger and suggests that using the distance would lead to more errors in classifying the peak clusters as peptide- or noise-related.



Figure 4.4: The overlap between the distributions of Pearson's χ^2 statistic and the Mahalanobis distance.

To check the variability in the overlap area in various spectra, Figures 4.5 and 4.6 present the overlap for the individual spectra. Note that, in both figures, the horizontal axis has been truncated to expose more details in the interesting region of the plot. From Figure 4.5 it can be clearly seen that, in agreement with the results

presented in Figure 4.4, the overlap area between the distribution of the peptide and the noise peak clusters was very small for all four individual spectra for Pearson's χ^2 statistic. On the other hand, Figure 4.6 shows that, for the Mahalanobis distance, the overlap is much larger in each of the four spectra. Thus, Pearson's χ^2 statistic is a better similarity measure to discriminate between peptide and noise signal observed in mass spectral data.



Figure 4.5: The overlap area between the distributions of the Pearson's χ^2 statistic values.



Figure 4.6: The overlap area between the distributions of the Mahalanobis distance values.

Figure 4.7 summarizes the performance of the two similarity measures for a set of randomly selected spectra. To this aim, the receiver operating characteristic (ROC) curve is used. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular cut-off point for the similarity measure. The area under the ROC curve (AUC) can be interpreted as the probability of the correct classification for a randomly selected subject (peak cluster in our case) from two populations (in our case, peptide- or noise-related clusters). In particular, a perfect discrimination is reflected by a curve passing through the (0, 1) point at the upper-left side, as seen in the plot for Spectrum 3 for Pearson's χ^2 . In that case, AUC is equal to 1. The ROC curves presented in Figure 4.7 indicate that the discriminative properties of Pearson's χ^2 statistic are better than the properties of the Mahalanobis distance. It is clear that AUC for the ROC curves for the Pearson's χ^2 statistic (blue, solid line) is larger than for the Mahalanobis distance (black, dashed line) for all four spectra. Figure A.2 in the Appendix illustrates that for the combined data from all four spectra.



Figure 4.7: ROC curves for each of the four analyzed spectra.

We also investigated the use of Pearson's χ^2 statistic calculated by using the polynomial model. Note that this was not possible for the Mahalanobis distance, as

Chapter 4. Comparison of the Mahalanobis distance and Pearson's χ^2 statistic 36 as measures of similarity of isotope patterns

the variance-covariance matrix of the expected ratios was not available. Figures 4.8 and 4.9 present for the selected spectra, the scatterplots of the values of Pearson's χ^2 statistic computed by using the *in silico* tryptic-digest database and by the polynomial model. The plots indicate that, for both sets of observed isotopic clusters, the obtained values were very close to those computed by using the *in silico* tryptic-digest database. Thus, they also performed better than the Mahalanobis distance in distinguishing between the peptide- and noise-related isotopic clusters. This is an important practical conclusion, as the calculation of the isotopic ratios from the polynomial models is much simpler than the use of the *in silico* tryptic-digest database.



Figure 4.8: Scatterplots of Pearson's χ^2 statistic computed by using the in silico tryptic-digest database and by the polynomial model for the putative-peptide clusters.



Figure 4.9: Scatterplots of Pearson's χ^2 statistic computed by using the in silico tryptic-digest database and by the polynomial model for the noise-peak clusters.

4.5 Conclusions

Our analysis indicates that Pearson's χ^2 statistic offers a better discriminative power for detecting the peptide clusters than the Mahalanobis distance. This result is most likely due to the fact that the Mahalanobis distance is very much based on the assumed form of the variance-covariance matrix Σ . The matrix derived from the *in silico* tryptic digest database may not be adequate for the isotope ratios observed in a spectrum. Moreover, the definition of the Mahalanobis distance is very much based on multivariate normality, which may not necessarily apply to the values of the consecutive ratios observed in a mass spectrum. We checked the multivariate normality of the observed consecutive ratios by using two multivariate normality tests, including Royston's H test [63], and Henze-Zirkler's [64] multivariate normality tests. The small *p*-values for the Royston's test and Henze-Zirkler's test were equal to 1.776357e-15 and 6.844373e-05, respectively. They do indicate that the multivariate normality assumption of the observed consecutive ratios might not be tenable, indeed. Thus, Pearson's χ^2 is the preferred statistic for evaluating the isotope distribution in mass spectrometry data.

Considering another similarity measure, e.g., Euclidean distance (4.3), is also possible.

$$\sqrt{\sum_{i=1}^{N} (O_i - R_i)^2}.$$
(4.3)

Chapter 4. Comparison of the Mahalanobis distance and Pearson's χ^2 statistic 38 as measures of similarity of isotope patterns

Figure A.3 in the Appendix summarizes the performance of the three similarity measures for the noise and peptide clusters using the data from randomly selected four spectra. The figure indicates that the discriminative properties of the Euclidean distance are minimally better than the properties of the Pearson's χ^2 statistic. For noise clusters having similar pattern to peptide clusters, the Euclidean distance offers a better discriminative power than the Pearson's χ^2 statistic. This result could be related to the weights corresponding to expected ratios assigned to the Pearson's χ^2 statistic. However, a comprehensive comparison of the performance of the Euclidean distance and the Pearson's χ^2 statistic can be subjected to future research.

An important practical point related to the use of Pearson's χ^2 statistic is the choice of the threshold for deciding whether the observed isotope cluster is similar enough to the expected isotopic distribution. Based on our experiment, a threshold value of, e.g., 0.2 would be suitable. It is difficult to propose any concrete value for the threshold in general, though, as it most likely depends on the technological platform used to generate spectra. To empirically obtain a value of the threshold, an experiment and analyses similar to the ones presented in this chapter could be performed. The Pearson's χ^2 statistic can also be extended to account for the correlation between isotope ratios which can be a subject for future research.

Chapter 5

The use of the isotopic distribution as a complementary quality metric to assess tandem mass spectra results

5.1 Introduction

Shotgun proteomics employs tandem mass spectrometry for amino acid sequencing (Section 3.7). Fragmented ion masses that are produced by this approach can be used in correlative database-searching to identify proteins and peptides from complex mixtures. As it has already been mentioned in Subsection 3.7.1.2, database-search methods depend on a score function that evaluates the "match" between the predicted ion fragment masses and the ones observed in the tandem mass spectrum. The better the agreement/match between the observed and expected ion fragment spectrum, the more likely that the PSM confidently identifies the peptide sequence [24, 25]. The existing computational methods, e.g., Percolator and Peptide Prophet to improve the quality and confidence of peptide identifications are useful, but still ignore potentially relevant information present in the data. In particular, peptide

identification based on tandem MS and database-search algorithms operate solely on fragment information and ignore the information available in the full MS scans about the isotopic distributions of the precursor ions.

In this chapter, we present a workflow that provides a new perspective on the quality of PSM from database-searching strategies for peptide identification. Additional views on a dataset can facilitate a more hypothesis-driven interpretation of the mass spectrometry signals. The similarity metric on the PSM scores contemplates the isotopic profile and results in a measure that conveys a degree of biomolecular similarity observed from the precursor of the selected tandem MS spectra. A close agreement between the PSM score and the similarity metric will result in a higher confidence for the identification of the selected precursor ion.

5.2 Materials and methods

The methods developed in this chapter are illustrated by using the experimental workflow starting from the whole-cell lysate of *Caenorhabditis elegans*, shotgun proteomics set-up, and the data interpretation step described in Section 2.2.

5.3 Implementation

The main question that our implementation aims at is whether the mass region chosen for further fragmentation in the DDA approach does indeed have a biomolecular origin. For this purpose, we reevaluate the full scan MS and trace the precursor ions using the target list exported from the Proteome Discoverer suite of Thermo Scientific. It should be noted that this list contains only the first-ranked PSMs and that no thresholds for reporting PSMs are applied in the case of SEQUEST peptide spectrum matching. In other words, every tandem mass spectrum corresponds to one PSM identity. Next, an in-house algorithm extracts the related isotope peaks from the full scan MS data taking into account the charge state and the mass of the molecule and compares the observed peaks with a theoretical "expected" isotope distribution. Pearson's χ^2 metric is considered as a goodness-of-fit statistic to indicate the similarity between the observed isotope peaks and the theoretical expected isotope distribution of a peptide.

The "expected" isotope distributions were obtained by using a model similar to the one presented in [45]. Note, however, that in our case the model was built on data obtained from a complete peptide database, instead of using the averagine peptide proposed in [44]. In particular, an *in silico* digest of the IPI human protein database version 3.65 was performed with trypsin as a protease. This yielded 3,802,880 peptides in a mass range from 75 to 230,000 Da. Only unique amino acid sequences were considered. After removing the redundant peptides, we obtained 517,572 peptides in the aforementioned mass range. The set of unique peptides was further used to compute the atomic composition. As a result, 258,813 non-redundant atomic compositions were found in the peptide database. The number of peptides with 0, 1, 2, 3, 4, 5, and 5+ sulfur atoms was equal to 102,986, 81,957, 43,977, 17,879, 6672, 2635, and 2707, respectively. It should be observed that the set of peptides containing 1 and 2 sulfur atoms is substantial, constituting 31.6% and 16.9% of the total number of peptides, respectively. For this reason the model described in (5.1) accounts for these two subsets.

For each atomic composition, the aggregated isotopic distribution was calculated using the BRAIN algorithm [39–41]. The obtained isotopic distributions were modeled as a function of the monoisotopic mass to arrive at expected distributions. In particular, the probabilities of occurrence of the aggregated isotope variants were translated into isotope ratios (Section 3.8). Subsequently, a fourth-order polynomial model of the form

$$R(x+1,m) = \beta_0 + \beta_1 \left(\frac{m}{1000}\right) + \beta_2 \left(\frac{m}{1000}\right)^2 + \beta_3 \left(\frac{m}{1000}\right)^3 + \beta_4 \left(\frac{m}{1000}\right)^4$$
(5.1)

was applied to data for each of the first three isotope ratios obtained for the isotopic distributions resulting from the *in silico* digest of the IPI human protein dataset. Note that the model was applied separately to subsets of data differing by the number of sulfur atoms contained in the peptide, given the influence of sulfur on the isotope distribution [54]. Estimates of the parameters β_0 , β_1 , β_2 , β_3 , and β_4 for each isotope ratio were obtained by maximum likelihood estimation (the estimates can be found in the Appendix in Table A.2). The residuals of the polynomial model are further investigated. Figure A.4 in the Appendix presents the histogram of the residuals based on the parameters estimated from the polynomial model from the IPI human protein dataset. The plots clearly indicate that we can assume the normal distribution for the residuals of the polynomial model. In other words, a polynomial normal regression could be assumed. As a result, a predictive model was obtained, which provided the expected values of the isotope ratios for peptides with the monoisotopic mass m in the mass range of 300 to 8000 Da. The expected values for the first three isotope ratios, computed in the aforementioned manner, are shown in Figure 5.1 as solid lines.



Figure 5.1: Fit of the fourth-order polynomial model to the in silico digest of the IPI human protein database with the estimated parameters β_0 , β_1 , β_2 , β_3 , and β_4 .

The advantage of this approach, as opposed to the work described in [45], is that the regression is performed on the entire peptide database instead of only a few data points related to the average peptide model [44]. Such an approach leads to a more realistic model and provides additional insight in the variation we might expect in the isotope distribution of peptides. This variation is noticeable in Figure 5.1 from the spread of the data points around the mean value for a given mass and should be considered when applying a cut-off to the goodness-of-fit measure.

The peaks extracted from the full scan MS spectra are also transformed to the consecutive ratios to make them compatible with the model defined in (5.1). The observed ratios are then compared with the corresponding expected values, computed from the polynomial model (see Figure 5.1), by using Pearson's χ^2 statistic:

$$\chi^{2} = \sum_{x=0}^{2} \frac{\left[R_{E}(x+1,m) - R_{O}(x+1,m)\right]^{2}}{R_{E}(x+1,m)},$$
(5.2)

where $R_O(x + 1, m)$ is the ratio of the (x + 1)-th peak observed in a mass spectrum to the x-th peak, while $R_E(x + 1, m)$ denotes the corresponding expected ratio obtained from the polynomial model. Unless specified otherwise, the calculation is performed by using the values for the first three ratios ($x \in \{0, 1, 2\}$), representing the first four isotope peaks, including the monoisotopic one. The smaller the χ^2 score, the better the agreement between the observed and expected ratios, and the more likely that

5.4 Results

the series of peaks is genuinely generated by a peptide.

The proposed method of scoring the similarity between the observed series of peaks and the pattern expected for a peptide was applied to the *C. elegans* N2 dataset. Note that the dataset contained 20,581 tandem MS data, for which SEQUEST reported 19,881 PSM scores and MASCOT reported 8576 PSM scores above their internal threshold. The scores were not filtered with respect to a FDR threshold to allow false positive and false negative identifications in the data.

For the SEQUEST dataset, the mass and charge information in the target list of the 19,881 precursor ions were used to extract the first four isotope peaks from the full scan MS data. An error tolerance of 10 ppm was allowed on the mass location of the isotope peaks. The rationale for extracting the first four isotopic peaks is that, in the peptide-centric setting, peptides are predominantly in the range of 700 Da to 4000 Da. Consequently, it is reasonable to assume that the isotopic distributions of these peptides are sufficiently characterized by the first four isotopic peaks.

Out of the 19,881 precursor ions, there were 10,015 cases for which four peaks could be extracted from the spectrum. For these cases, the first three expected isotope ratios were computed from the polynomial model and compared to the three observed ratios. For 5961 ions, the last (fourth) isotopic peak was missing, most often because it fell below the limit of detection, especially in the case of low-mass ions. For these ions, the first two expected isotope ratios were computed and compared to the observed ones. There were also 2801 precursor ions, for which the second and/or third isotope peak could not be retrieved from the spectrum even with an increased error tolerance of 20 ppm. These precursor ions were omitted from considerations, leaving 17,080 PSMs available for the study. Surprisingly, when peak extraction was performed with an error tolerance of 3 ppm, which is within the precision of the Orbitrap class instrument, there were 112 precursor ions for which we could not retrieve the monoisotopic peak. This number decreased to 33 for an increased error tolerance of 20 ppm. These 33 precursor ions were also disregarded from the dataset.

In case of precursor ions for which multiple peaks occurred inside the tolerance windows (1071 PSMs), all possible configurations of peaks leading to a four-peak cluster were considered. Pearson's χ^2 was calculated for all of the configurations and

the minimum value, i.e., the best fit, was retained.

Both Pearson's χ^2 statistic and the PSM score aim at quantifying the agreement between experimental data and theoretical *in silico*-generated information. For this reason, we would expect to see a negative correlation between the two scores, indicating that high PSM scores (i.e., a good agreement with fragmentation for tandem MS) would correspond to low Pearson's χ^2 values (i.e., a good agreement with genuine peptide feature in the precursor scan), even though they both represent different layers of experimental information, i.e., MS1 and MS2 levels, respectively.

Figure 5.2 presents a scatter plot of the values of the SEQUEST Xcorr score and Pearson's χ^2 statistic for the 17,047 precursor ions from the *C. elegans* N2 dataset. The abscissa of the plot represents Pearson's χ^2 statistic and the ordinate represents the SEQUEST Xcorr score. The inserts show the regions with small values of Pearson's χ^2 statistic that indicate a good agreement between the observed and expected isotope patterns.

Figur 5.2 shows only a limited correlation between the SEQUEST Xcorr scores and Pearson's χ^2 statistic. The plot illustrates that there are cases when a low Xcorr score, e.g., below 2.5, is associated with a low value of Pearson's χ^2 s statistic, e.g., below 0.1. These are cases when an apparently valid peptide-related precursor ion was not identified. On the other hand, there are cases when a large value of Xcorr score is associated with a high value of Pearson's χ^2 statistic, implying a positive identification of a non-peptide-related ion.



Figure 5.2: A scatter plot of the SEQUEST Xcorr score and Pearson's χ^2 statistic.

44

These remarks indicate that, by systematically studying a scatter plot similar to the one shown in Figure 5.2, it could be possible to identify cases when, e.g., the identification is reliable or when it would require further scrutiny. To this end, the plot in Figure 5.2 can be conceptually subdivided in four regions, corresponding to the four possible combinations of low/high values of the SEQUEST Xcorr scores with low/high values of Pearson's χ^2 statistic. This principle is presented in Figure 5.3. In the next subsections, we will discuss characteristics of each of the four regions. In particular, we will discuss arbitrarily selected tandem MS data to illustrate the characteristics.



Figure 5.3: Four regions resulting from the scatter plot between the PSM scores and Pearson's χ^2 statistic.

5.4.1 Region I

In Region I, high SEQUEST Xcorr scores are accompanied with low values of Pearson's χ^2 statistic. A high SEQUEST Xcorr score indicates that the observed fragment masses match well with the theoretical fragment masses and hence lead to a positive sequence identification. On the other hand, a low value of Pearson's χ^2 statistic indicates that the observed series of MS1 peaks, linked to the precursor ion, is likely to be genuinely generated by a peptide. This score pair can be taken as a sign of a reliable identification. An example is the identification of the scan number 25,320 in Figure 5.4A). The figure displays the annotated product ion spectra for sequence 'YLGAYLLATLGGNASPSAQDVLK' with an Xcorr score of 6.5295 and $\Delta M = -0.06$ *ppm*. Figure 5.4B presents the observed series of peaks corresponding to the precursor ion. The close-up in Figure 5.4B shows the resulting observed (blue) isotope ratios, and the corresponding expected values (red). It can be seen that the observed values of the ratios correspond remarkably well to the expected values, which is confirmed by the low value of Pearson's χ^2 statistic equal to 0.03525.

5.4.2 Region II

In Region II, high SEQUEST Xcorr scores are accompanied by high values of Pearson's χ^2 statistic. A high SEQUEST X corr score suggests a good PSM. However, a high value of Pearson's χ^2 statistic indicates that the observed series of MS1 peaks, linked to the precursor ion, may not be genuinely generated by a peptide. There are three possible explanations for this discrepancy. First, the high value of Pearson's χ^2 statistic could be due to overlapping peptides. This overlap does not mean that the identification, indicated by the high PSM score, is wrong. A wide isolation window could yield enough ions to be fragmented and identified correctly by the score algorithm. This event could be assessed by calculating the extent of co-isolation based on the full MS1 scan. Second, the high value of Pearson's χ^2 statistic can be due to a problem with the monoisotopic mass determination [49]. This case is illustrated by the data for the scan number 25,651. Figure 5.5A displays the annotated product ion spectra for sequence 'LcYVALDFEQEMATAASSSSLEK' with an Xcorr score of 5.9267 and $\Delta M = -0.89 \ ppm$ ('c' means carbamidomethyl modification in the sequence). Figure 5.5B presents the observed series of four peaks corresponding to the precursor ion, together with the observed and expected isotope ratios (inset), producing the high value of Pearson's χ^2 statistic equal to 2.53364. The value suggests that the algorithm for monoisotopic mass determination might have selected the wrong peak. In such cases we may be wary about the fragment annotation, because the peptide database may have been filtered according to the assigned monoisotopic mass with an error tolerance of 5 ppm. The one Da discrepancy will most likely lead to a false positive identification. Running the mass spectrum with a corrected monoisotopic mass, similarly as is done in Bullseye [49], would resolve the issue. Finally, the third possible explanation is simply a false positive identification on a nonsense tandem mass spectrum, as might be the case when inspecting the corresponding fragment spectrum in Figure 5.5A. In conclusion, PSMs in region II should be treated as questionable and further investigated.



Figure 5.4: Region I: Panel A) displays the fragmentation spectrum with scan id 25320 of the precursor ion in Panel B) eluted at a retention time of 122.33 min. The isolation window is indicated in yellow and centered on 1162.13 m/z. The monoisotopic mass, indicated by the blue bar, is equal to 1161.621 m/z with a charge state of z = 2. The close-up displays the observed and expected isotope ratios (Panel B).



Chapter 5. The use of the isotopic distribution as a complementary quality

Figure 5.5: Region II: Panel A) displays the fragmentation spectrum with scan id 256510f the precursor ion in Panel B) eluted at a retention time of 123.90 min. The isolation window is indicated in yellow and centered on 1276.09 m/z. The monoisotopic mass, indicated by the blue bar, is equal to 1275.589 m/z with a charge state of z = 2.

5.4.3 Region III

In Region III, low SEQUEST Xcorr scores are accompanied by high values of Pearson's χ^2 statistic. This region suggests a bad identification, most likely based on a nonpeptide molecule. An example is provided by the data for the scan number 24,141 in Figure 5.6. A Pearson's χ^2 statistic of 2.65587, which implies a bad agreement between the theoretical and observed isotope patterns, is accompanied by the product ion spectra for sequence ' DVFFccNmcPYKAPTmNRcQR' with a weak Xcorr score equal to 0.5519 and ΔM = -2.97 ppm ('m' means methionine oxidation modification in the sequence). There can be several explanations for this case: overlapping peptides, co-isolation, atypical fragmentation, an error in the monoisotopic mass determination, etc. Alternatively, the DDA-approach may have selected a noisy peak for the tandem MS interpretation. In conclusion, PSMs in region III should be treated as unreliable and discarded. Alternatively, in the case of a defect monoisotopic mass determination, the database-search could be repeated after correction of the monoisotopic mass.

5.4.4 Region IV

Finally, in Region IV, low SEQUEST Xcorr scores are accompanied with low values of Pearson's χ^2 statistic. This indicates a bad identification, but one that is most likely based on a peptide molecule. An example is provided by the data for the scan number 17,496 in Figure 5.7 with a value of Pearson's χ^2 statistic equal to 0.00227 and the annotated product ion spectra for sequence 'EEPTDFSEENLVKK' with a low Xcorr score equal to 0.8904 and ΔM = -0.374 ppm. Note that the intensity of the precursor ion is large and it is well discernable from the noise. There can be several reasons why the PSM score fails to identify the tandem mass spectrum in this case. A simple explanation is that the correct peptide sequence is not present in the database, or that a particular post-translational modification is not accounted for in the search. Another possibility is that the peptide has an atypical fragmentation pattern, which results in fragment ions not predicted by the score algorithm, as could be the case in the fragment spectrum in Figure 5.7A. Manual or *de novo* interpretation of this set of peptides may address this problem. An enhanced fragmentation model that predicts more realistic fragment masses might solve the issue as well [65, 66]. Identifications situated in Region IV would be suited for manual or de novo reevaluation.



Chapter 5. The use of the isotopic distribution as a complementary quality

Figure 5.6: Region III: Panel A) displays the fragmentation spectrum with scan id 24141 of the precursor ion in Panel B) eluted at a retention time of 117 min. The isolation window is indicated in yellow and centered on 930.38 m/z. The monoisotopic mass, indicated by the blue bar, is equal to 929.715 m/z with a charge state of z = 3.



Figure 5.7: Region IV: Panel A) displays the fragmentation spectrum with scan id 17496 of the precursor ion in Panel B) eluted at a retention time of 91.78 min. The isolation window is indicated in yellow and centered on 832.91 m/z. The monoisotopic mass, indicated by the blue bar, is equal to 832.904 m/z with a charge state of z = 2.

51

5.5 Conclusions

It is generally accepted that scoring algorithms are essential to support the interpretation of the tremendous amount of tandem MS data generated by current mass spectrometers. On the other hand, it still occurs that the score algorithm is not able to discern peptides from noise based on the observed fragment spectra or that it simply fails to identify a good quality tandem mass spectrum. To address this problem, we propose to compare the isotope patterns of observed parent ions in the full MS scans with the expected isotopic distribution. Pearson's χ^2 statistic is well suited to indicate the agreement between the observed isotope pattern [67] and the one that is expected for a peptide with a similar monoisotopic mass. By merging this information with the database match score, additional insight in the reliability of the identification can be obtained.

The proposed approach will not resolve issues concerning FDR or deep-mining of tandem MS annotations for enhanced recovery of PSMs below the detection threshold. The approach solely aims at providing the users with an additional and different perspective on the data, extracted from a separate data layer, i.e., full-scan MS spectra. The definition of the four regions, depicted in Figure 5.3, can support the user to make decisions about the quality of the PSM. Tandem MS data resulting in scores situated in Regions I and III are clear-cut cases of go and no-go, respectively. On the other hand, scores from Regions II and IV should prompt further investigation. For example, PSMs in Region II could be critically assessed because of the bad resemblance with a peptide pattern observed in the full scan MS data for potential positive identification. The latter set is mainly situated in Region IV given the presence of peptidic patterns in full scan MS data and could be selected for further interpretation (manual, *de novo*, or with different modifications).

For the practical purpose of implementing the outlined strategy, one should define the thresholds defining the four regions. It is difficult to provide any general solution for this aspect. The threshold for Pearson's χ^2 statistic will depend on several factors.

An important factor is the MS platform used, as different instruments generate a different amount of measurement error for the observed peak intensities. For example, trap instrument can distort the isotope profile due to space-charge effects, as pointed out in [68]. Ion statistics also play a significant role when comparing isotopic distributions. Complex peptide mixtures combined with limited ion storage capacity in, e.g., ion traps, can lead to poor ion statistics of the isotopic representations for some peptides, while more abundant ones will be nicely presented. Spectral accuracy

52

is an important requirement for the procedure proposed in this manuscript. On top of the instrument variability, the natural variability in the atomic composition of molecules with a similar mass will be reflected in the isotopic distribution. Another aspect, related to the model used to predict the expected isotope pattern, is that the model does not account for post-translational modifications. Indeed, some modifications can alter the isotope profile of molecules severely, as is the case for, e.g., bromine or chlorine. As the proposed method adopts a peptide model that is based on the relation between mass and the atomic composition, it can be assumed that PTMs that are composed of C, H, N, O, and S are partially covered by the model. A possible extension of the model could incorporate modifications that are used by the search strategy to further scrutinize at the MS1-level whether the observed isotope profile is likely to be generated by modified peptides.

Taking into account the aforementioned sources of variation and factors, to obtain a concrete value of a threshold, one might need to run a designed experiment with a known mixture of peptides and to compare the distributions of Pearson's χ^2 statistic values obtained for the peptides and for randomly-selected peak clusters. Based on such a comparison, a value could be selected that offers a trade-off between the falsepositive and false-negative assignments of peak clusters as peptide or non-peptide generated. A threshold value of, e.g., 0.1, could be suitable for our study.

A concept related to the method considered in this chapter is used by the software program Bullseye [49]. Bullseye also performs a post-acquisition investigation of the targeted mass region in a full scan MS. However, the additional information is used to improve the efficiency of a database-search by accurately determining the monoisotopic mass of the selected parent ions for tandem MS, instead of providing an additional parameter for the evaluation of the database-searches. For this purpose, it uses persistent isotopic distributions observed in the full scan LC-MS data. A warrant monoisotopic mass relieves the need for a wide search space, i.e., less candidate peptides, as such, the efficiency of the database-search alleviates. On the other hand, the confidence in the identification increases because the accurate monoisotopic mass information can be used to filter-down the obtained PSM identifications. The method proposed in this chapter uses the MS data in a different way as compared to Bullseye. It applies the concept of the isotopic distribution to interpret the targeted region of full scan MS. However, instead of focusing on accurately determining the monoisotopic mass, as in Bullseye, the proposed method uses Pearson's χ^2 statistic to evaluate whether the observed peptidic features are legitimate.

Another relevant idea is the concept of a peptide window in which the accurate monoisotopic mass of a peptide should reside, [69]. We have performed a related analysis (see Appendix A.3) starting from the 19,881 tandem MS scans that were identified by SEQUEST. Regions I and II (see Figure 5.3) were found to be enriched for molecules outside the peptide window, while region III was found to be depleted. Finally, it is worth mentioning that the analyses presented in this chapter were also conducted using MASCOT E-values. The results are presented in the Appandix (see Figure A.5). The obtained conclusions are similar to those presented for the SEQUEST Xcorr.

54

Chapter 6

Assessing the agreement between peptide assignments for different search engines

6.1 Introduction

The data stream generated by shotgun proteomics approach (Section 3.7) is tremendous and automated annotation strategies have been devised to plough through the data. The variety of correlative search algorithms for peptide identification using MS/MS spectra is quite large [24, 70]. The variation of principles used in the search engines causes variation in peptide identifications. Thus, it seems that the choice of the search algorithm plays a role in the identification process of peptides and their corresponding proteins [71]. To improve the robustness and confidence of peptide and protein identification, recent studies suggest the use of consensus-based approaches that combine the results from two or more search engines. Most of these studies have shown that a consensus approach can greatly enhance mass spectral coverage and specificity, compared to the use of a single search algorithm [72-79]. For example, in [80], seven database search methods were studied using a composite score approach based on a calibrated expected value. It was concluded that, because of the possible weak correlation between different methods, accuracy and confidence of peptide identification can improve if different search engines are combined. Critical factors for the evaluation of peptide identification are which individual method (e.g., MASCOT, SEQUEST, X!Tandem) should contribute to a consensus method and how the method should be defined. A recent overview of the methods combining multiple search engines is provided in [76].

Besides improving the confidence, sensitivity, and specificity of peptide identification, it is of importance to measure the degree of agreement between the search algorithms regarding their peptide assignments. In this chapter, we employ Cohen's kappa coefficient to measure the agreement between the identification of peptide sequences obtained by SEQUEST and MASCOT. The coefficient has been used for decades by, e.g., pathologists to score concordance of grading of tissue slides [81]. To illustrate the concept of inter-rater reliability, we used a dataset of our benchmark organism described in Section 2.2.

6.2 Implementation

The main question that we tried to address was whether there was an agreement in terms of the PSM assignments between different search engines. The analysis of interrater agreement often provides a useful means of assessing the reliability of a rating system as is the case in many clinical examples, e.g., histological grading of cancer tissues. Various approaches have been proposed to study inter rater agreement. Percentage agreement [82] have traditionally been used to summarize observer agreement. It is equivalent to computing the proportion of individuals, in our case spectra, that received the same rating by both raters (in our case search engines). Although percentage agreement has some advantages that include computational simplicity and ease of interpretation [83], this statistic does not allow for the fact that a certain amount of agreement can be expected on the basis of chance alone. Cohen [84] proposed a measure of agreement that corrects for chance agreement. Specifically, Cohen's kappa coefficient measures the chance-corrected agreement between two raters (e.g., search engines) who independently classified the same n objects (e.g., spectra) into one of the k non-overlapping categories (e.g., non-redundant peptide sequences or protein accession numbers as defined by the protein database).

Table 6.1 presents the set up of a $k \times k$ table for the classifications. let p_{ij} , be the proportion of objects (spectra) that were placed in the *i*-th category by the first rater and the *j*-th category by the second rater (i, j = 1, ..., K). Also, let $p_{i.} = \sum_{j} p_{ij}$ denote the proportion of objects (spectra) placed in the *i*-th row/category (non-redundant peptide sequence) by the first rater (search engine), and let $p_{.j} = \sum_{i} p_{ij}$ denote the proportion of spectra placed in the *j*-th column/category by the second rater (search

engine). Then, the kappa coefficient is defined as follows:

$$\kappa = (p_o - p_e) / (1 - p_e) \tag{6.1}$$

where $p_o = \sum_{i} p_{ii}$ denotes the observed proportion of agreement between two raters (the proportion of spectra that received the same sequence annotation by the two search engines ,i.e., diagonal of the table), while $p_e = \sum_{i} \sum_{j} p_{i.} p_{.j}$ is the proportion of agreement expected by chance.

Table 6.1: Classification of spectra into k non-redundant peptide categories by two search engines.

| MASCOT | | | | | | | |
|---------|----------|----------|-------|----------|----------|--|--|
| SEQUEST | 1 | 2 | | k | Total | | |
| 1 | p_{11} | p_{12} | ••• | p_{1k} | $p_{1.}$ | | |
| 2 | p_{21} | p_{22} | ••• | p_{2k} | $p_{2.}$ | | |
| | ÷ | ÷ | | ÷ | : | | |
| k | p_{k1} | p_{k2} | • • • | p_{kk} | $p_{k.}$ | | |
| Total | $p_{.1}$ | $p_{.2}$ | • • • | $p_{.k}$ | 1 | | |

Possible values of κ range from -1 to 1, with the value of one indicating a perfect agreement. The value of zero indicates that the observed agreement is exactly what would be expected by chance, i.e., if the raters (search engines) had randomly assigned the ratings (sequence annotations). A value of Cohen's κ smaller than zero indicates an agreement worse than chance agreement as in the case when class labels would have been swapped. The strength of agreement is interpreted using the categories proposed in [81] as follows: 0.00 - 0.20 = slight agreement, 0.21 - 0.40 = fair agreement, 0.41 - 0.60 = moderate agreement, 0.61 - 0.80 = substantial agreement, 0.81 - 1.00 = almost perfect agreement.

In our study of SEQUEST and MASCOT, the resulting peptide matches, i.e., unique modified peptide sequences proposed by any (or both) search programs, define the rows and columns as in Table 6.1. The cells of a Table 6.1 contain the number of times a particular combination of sequences was observed for the SEQUEST and MASCOT identification. Based on the table, percentage agreement and Cohen's κ coefficient are calculated and interpreted as a degree of agreement between the SEQUEST and MASCOT search algorithms.

6.3 Results

In general, database-searching engines identify only a proportion of the MS/MS spectra of digested proteins. Figure 6.1 presents this proportion of the experimental MS/MS spectra identified by SEQUEST and MASCOT for the *C. elegans* dataset when FDR filtering is disabled. Several causes contribute to this phenomena, such as unexpected modifications, aberrant fragmentation patterns, sequence database in-accuracies, etc. Another nuisance factor are differences in the implementation of the search methods.

Figure 6.1 illustrates that SEQUEST and MASCOT are very different algorithms when it comes to reporting the search results. Due to differences in the implementation, SEQUEST reports a score for every tandem mass spectrum, even if this score is very low. In contrast, MASCOT adopts an internal reporting threshold to immediately eliminate low quality results. The scheme is explicitly coded into the MASCOT algorithm and cannot be adjusted by the user. In total, there were 20,233 (99%) PSMs by SEQUEST out of the 20,482 MS/MS scans. Among them, 12,128 (59%) were also identified by MASCOT (Figure 6.2). The histogram of the scores for the 20,233 peptides identified by SEQUEST is presented in Figure 6.3a. The histogram in Figure 6.3b displays the SEQUEST score for the spectra which did not receive a MASCOT annotation. When comparing Figure 6.3b to Figure 6.3a, it can be seen that the sequence identification omitted by MASCOT receive mainly low SEQUEST XCorr scores. This observation justifies the internal reporting threshold of MASCOT, but complicates the analysis of agreement.



Figure 6.1: The proportion of the unknown spectra and identified spectra by SEQUEST and MASCOT.
In particular, when analyzing the agreement between the two search engines, we have to account for the cases when missing observations occur, i.e., a certain spectrum receives a (confident) sequence annotation for only one of the search engines. The fact that there are tandem mass spectra not identified by MASCOT should be taken into account when computing the value of Cohen's κ .



Figure 6.2: Venn diagram of defined spectra made by SEQUEST and MASCOT.



Figure 6.3: (a) Histogram of scores for uniquely identified peptides by SEQUEST with Xcorr scores. (b) Histogram of SEQUEST Xcorr for non-identified peptides by MASCOT.

Toward this aim, one can try to assess the possible range of agreement given various scenarios for the non-identified spectra. In particular, the agreement is accounted for by means of a sensitivity analysis that considers the concordance between the search engines under two extreme case scenarios. In the "conservative" scenario, we assume that the unannotated spectra by MASCOT would completely disagree with the SEQUEST findings. Therefore, we distribute the spectra with missing annotations randomly among the possible categories in the joint table. Note that in this procedure we prevent the possibility of a chance agreement. In the "liberal" scenario, we assume that the spectra with missing annotations do agree with the SEQUEST findings. By considering the two scenarios, we place an upper and a lower limit on the value of Cohen's κ .

Table 6.2: Percentage agreement and Cohen's κ coefficient calculated for agreement.

| | Percentage Agreement | Chance Agreement | Cohen's κ coefficient |
|--------------------|----------------------|------------------|------------------------------|
| Observed agreement | 0.7827 | 0.00012 | 0.7826 |
| Minimum agreement | 0.4691 | 0.00006 | 0.4691 |
| Maximum agreement | 0.8697 | 0.00012 | 0.8696 |

Table 6.2 shows the percentage agreement, chance agreement and Cohen's κ coefficient for three scenarios, i.e., ignoring the non-identified spectra ("Observed agreement"), and assuming that the spectral are discordantly ("Minimum agreement") or concordantly identified ("Maximum agreement"). Table 6.2 indicates that the probability of a chance agreement is very low and does not contribute much to the kappa statistic in this case. This is understandable, as the number of potential sequences that can be selected for an identification is very large. The "observed" agreement, calculated without taking into account the non-identified Mascot tandem spectra, could be classified as "substantial" agreement. However, the value of the κ coefficient underlying this conclusion should be interpreted with caution, as it is obtained by disregarding the missing data. In fact, the observed agreement in Table 6.2 assumes that the agreement for the missing data is equivalent to the agreement of the observed portion. However, if we assume that, for the non identified spectra, the two search engines would be more likely to give discordant identifications, then the obtained value of the κ coefficient suggests only a "moderate" agreement before the identifications for the two search engines. On the other hand, if we assume that, for the non identified spectra, the two search engines would be more likely to give concordant identifications, then the obtained value of the κ coefficient suggests an "almost perfect" agreement. Unfortunately, given that we have not got MASCOT identifications for the non identified spectra, we cannot make any more definitive statements other than those provided above.

6.4 Conclusions

In this chapter, we propose to apply Cohen's κ to analyze rater reliability in the context of database search engines. Assessing the agreement between SEQUEST and MASCOT is obtained by considering a sensitivity analysis that is connected to different interpretations: minimum agreement and the maximum agreement. The results of our study indicate that, in general, there is at least moderate agreement between the peptide identification results obtained for SEQUEST and MASCOT. Another observation, is that the percentage of chance findings is small, which is due to the high number of categories (non-redundant peptide sequence) to which a spectrum can be assigned. In fact, the chance agreement seems ignorable as it does not influence the kappa score substantially. However, this conclusion refers to the level of peptide identification. When assessing the rater reliability at the level of proteins, the number of categories would decrease and the chance findings would become more likely.

The scores obtained for SEQUEST and MASCOT were not filtered with respect to a FDR threshold. However, we can compare the two search engines on the set of confident peptide spectrum identifications, i.e., we filter out peptide identification results that do not comply with a FDR of 5% according to the target decoy approach. We could suggest that a combined SEQUEST and MASCOT search would not yield more peptide identifications when an alternative measure for peptide confidence could be developed. Moreover, the addition of another search engine will definitely be informative for the comparison of the agreement of the peptide assignments which can be a subject for future research.

Part II

Statistical Methods for Gene Mapping

Chapter 7

Introduction

Genomics is the study of the genome of an organism - its entire genetic material and hereditary information in the form of DNA, RNA, genes and chromosomes. It concentrates on understanding the structure and function of an organism's genetic material from the molecular level upwards, including interactions between genes, interactions between genes and the proteins they produce, and interactions between genes and environmental factors. It has obvious links to proteomics, which focuses on understanding the structure and function of the proteins produced by the genome. In the second part of this dissertation, we focus on the genetic dissection of phenotypic traits, also known as gene-mapping. This chapter provides a brief overview of genomics and principles of gene-mapping.

7.1 Chromosomes, DNA, and genes

The genetic information transmitted from parent to offspring is stored on chromosomes in the nucleus of eukaryotic cell. Chromosomes are threadlike structures that contain the genetic information. The number of chromosomes varies from species to species. In humans, for example, each cell normally contains 23 pairs of chromosomes. There are two different types of cells in eukaryotic organisms - haploid cells and diploid cells. The difference between haploid and diploid cells is related to the number of chromosomes that the cell contains. Diploid cells have two homologous copies of each chromosome, while haploid cells such as gametes only have one copy of each chromosome.

Double-stranded deoxyribonucleic acid (DNA), the basic biological material of inheritance, is a double-helix molecule and has two strands running in opposite directions to each other. (There are some examples of viral DNA which are singlestranded). Each strand is a polymer of complementary subunits called base pairs: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Each strand has a backbone made up of (deoxy-ribose) sugar molecules linked together by phosphate groups. Each sugar molecule is covalently linked to one of 4 possible bases. Complementary base pairing occurring in double strands of DNA means that the bases pair up in a specific way, i.e., Adenine binds to Thymine and Cytosine binds to Guanine through weak hydrogen bonds (Figure 7.1).

Each protein is encoded by a gene (a particular sequence of DNA nucleotides that specify how a single protein is to be made). Specifically, the order of nucleotides within a gene specifies the order and types of amino acids that must be put together to make a protein. In the process called transcription the gene is copied to messenger RNA (mRNA), which contains the required information of a particular protein. After the transcription is completed, the mRNA-molecules are translated into a polypeptide during polypeptide chain synthesis. The DNA sequence in a gene consists of coding and non-coding regions, i.e., exons and introns, respectively. Exons code for proteins, whereas introns do not code for proteins.

7.2 Genetic variation and meiotic recombination

Variation in the order of nucleic acids in a DNA molecule allow genes to encode enough information to synthesize the huge diversity of different proteins and enzymes needed for life. In addition to differences between genes, the arrangement of nucleic acids can differ between copies of the same gene. This results in different forms of individual genes. Variation in nucleic acid sequences can arise from mutations. The most common mutation is a base pair substitution, for instance, a single base pair T is replaced by A. Mutations may also involve the insertion or deletion of genetic materials, at the level of a few base pairs or even whole chromosomes.

Distinctive forms of a gene are called alleles, which are located at the same position, or genetic locus, on a chromosome but have an altered function. One gene might have many allelic variants. At any given locus, each diploid individual has two alleles (one allele for each of the two homologous chromosomes). If an individual has the same allele for both homologous chromosomes, the individual is homozygous for that allele. If an individual carries two different alleles, the individual is said to be heterozygous at that locus. The pair of alleles at a locus is referred to as the genotype.



Figure 7.1: Two strands of DNA are aligned anti-parallel to each other (panel a). Complementary primary nucleotide structures for each strand (panel b). Taken from [85].

Genetic variation together with environmental variation contribute to the organism's phenotype, the physical appearance of the genotype as it relates to a certain trait. The phenotypic traits of the different organisms can be of two kinds: qualitative and quantitative. The qualitative characteristics have distinct (separate) phenotypic classes. Classic examples are the Mendelian traits observed for pea-seed shape (wrinkled form *versus* smooth round) and blood group in humans. Usually, a single gene or small group of genes with little or no environmental modifications control qualitative traits. Quantitative traits, however, occur as a continuous range of variation. Some examples of quantitative traits include height and weight. Quantitative traits are influenced by genetic and environmental factors. A larger group of genes control quantitative traits as compared to qualitative ones. When multiple genes influence a trait, we can also describe it as a "polygenic trait". A genomic region that influences a quantitative trait is referred to as a quantitative trait locus (QTL).

The resulting offspring of sexually reproducing organisms has sometimes allele combinations which are not genetically identical to either parent or their siblings. This novel set of genetic information can be the result of recombination (crossing-over) during "meiosis". Meiosis is a process of eukaryotic cell division that produces haploid sex cells or gametes from diploid cells. During meiosis, homologous chromosomes pair up and crossing-over occurs. The location where crossing-over occurs is referred to the chiasma. At this location, part of the chromosomes is exchanged and recombinant chromosomes with new allele combinations are created. A schematic diagram of this process is presented in Figure 7.2.

7.3 Genetic linkage

In the mid-1800's, Mendel suggested that alleles of each gene separate independently from the other genes. However, in the early 1900s, Bateson and Punnett [86] realized their results did not conform to Mendel's law of independent assortment. Based on these findings, they proposed that certain alleles must somehow be linked with one another. The genes that are located on the same chromosome, and do not inherit independently. The understanding of genetic linkage was expanded by the work of Thomas Hunt Morgan on *Drosophila melanogaster*. In studying within-chromosome recombination, Morgan [87] proposed that the farther apart two genes were located on a chromosome, the more likely that crossing-over occurred. Alfred Sturtevant [88] took this argument one step further and constructed the first genetic map (also known as linkage map) of a chromosome in 1913.



Figure 7.2: Schematic diagram of meiosis. Taken from [89].

He proposed that the frequency of recombination between two genes can be used as measure of the chromosomal distances separating them. The (genetic) distance is expressed in units called genetic map unit (m.u.), or a centimorgan (cM). The recombination frequency of 1 percent corresponds to one centiMorgan. Sturtevant suggested that genes can range from being perfectly linked (recombination frequency = 0) to being perfectly unlinked (recombination frequency = 0.5) when genes are on different chromosomes or genes are separated very far apart on the same chromosome.

A map function is a mathematical relationship that converts map distance (d, measured in Morgans) to a recombination frequency (θ) . The most widely used is the Haldane map function [90]. The Haldane function is based on the assumption that cross-overs occur at random independently of one another. The resulting function is:

$$\theta = \frac{1}{2}(1 - e^{-2d}), \tag{7.1}$$

which is derived under the assumption that the locations of cross-overs follow a Poisson process.

7.4 Molecular markers

Specific regions on the DNA, both in the coding as well as non-coding regions, can be identified as markers. A molecular marker (genetic marker) is a fragment of DNA, ranging from 1 to 60 base pairs (bp), that is associated with a certain location within the genome. In the context of gene-mapping to infer the position of a gene that contributes to a specific trait, known chromosomal locations such as genetic markers play an important rule.

7.4.1 Single nucleotide polymorphisms

A SNP is the most common type of genetic marker. It is defined as a single base pair change in a DNA sequence. SNPs occur in non-coding regions more frequently than in coding regions. SNPs which occur in these non-coding regions does not have a direct genetic effects on disease or trait, but within a coding region, they can be disastrous.

7.5 Construction of genetic maps and QTL analysis

One of the main uses of molecular markers has been to construct genetic or linkage maps. The position of an unknown gene is inferred by detecting linkage between that gene and genetic markers. Linkage maps indicate the position and relative genetic distances between markers along chromosomes. One important use of linkage maps is to identify chromosomal locations containing major genes and QTL associated with traits of interest. QTL-mapping is a statistical analysis linking phenotypic information (the trait of interest) with genotypic data (segregation of molecular marker over the individuals) to provide specific genomic regions linked with the studied trait. A QTL-mapping study needs a population with as much variation as possible for the trait of interest. The mapping population is a cross arranged between two inbred lines which differ substantially in the quantitative trait of concern. Assume there are two parents having alleles QQ and qq at a certain locus on the chromosome. The offspring of these parents is called the F1 generation and have the allele Qq.

The purpose of mapping method is to find the QTL allele Q or q with unknown locus. All we know are markers along the chromosome. Recombination occurs when alleles cross over to another chromosome and recombination frequency can be used as a measure of closeness between two genes or between a gene and a marker. If the recombination frequency between marker and a QTL is small, then that marker is closely linked to a QTL. If the recombination frequency is large there may be no linkage between a QTL and the marker. The more markers, the more precise the genetic map is and the more accurate the mapping can be.

7.6 DNA sequencing

DNA sequencing is the act of determining the nucleotide sequence of given DNA molecules.

In 1977, two methods for sequencing DNA were introduced. One method is referred to as Maxam-Gilbert sequencing [91] and the other approach, developed by Frederick Sanger [92], is called the chain termination method (also called dideoxy sequencing). These approaches are used to generate DNA fragments subjected to polyacrylamide gel electrophoresis to separate strands of DNA that differ in size by as little as one base pair. The Maxam-Gilbert method is based on nucleotidespecific cleavage by chemicals and is best used to sequence short nucleotide polymers (usually smaller than 50 base-pairs in length). This method is rarely used as it is time-consuming and requires handling of toxic chemicals. Sanger *et al.*, [92] developed an enzymatic method based on the use of chain-terminating dideoxy nucleotides (ddNTPs). This method, on the other hand, offered overall higher efficiency after a series of optimizations, in particular switching from radioactive to dye labelling of nucleotides and using capillary gel electrophoresis instead of slab gels. This technique dominated DNA sequencing for decades.

However, Sanger method had some disadvantages. It was rather labor, reagent, time-consuming, and expensive. The demand for cheaper and faster sequencing methods has increased greatly. Consequently, the second-generation sequencing methods, or next-generation sequencing (NGS) methods has been developed. NGS performs massively parallel sequencing and decrease the time and cost of sequencing profoundly. Numerous NGS platforms have been launched [93, 94]. The first three platforms, which currently are still the most prevalent ones, are: 454 [95], an array-based pyrosequencing approach, Illumina [96], and SOLiD [97]. Each of these platforms have their own methods and ways of sequencing [98]. However, some fundamental features they share (Figure 7.3) are as follows:

- cell-free template amplification using emulsion PCR or solid phase amplification;
- immobilization of templates to some solid structure, which allows massive parallel processing ;
- imaging of nucleotides incorporated into synthesized molecules (sequencing-bysynthesis) or probe-hybridisation to templates (sequencing-by-ligation).

In the next subsection, a brief overview of the Illumina platform is given, as it is the most relevant to this dissertation.

7.6.1 Illumina

Illumina sequencing begins with the attachment of a specific adapter sequence to the DNA fragments. The fragment library is poured onto a solid surface, flow cell. Fragments get attached to the flow cell surface through binding to complementary adapter sequence and clusters of copies of the same DNA fragment are built by bridge amplification. The procedure then continues in a cyclic fashion, incorporating one nucleotide per cycle in each fragment cluster. All four nucleotides are added simultaneously, and the appropriate nucleotide is added to each fragment. Nucleotides carry reversible terminators. Furthermore, each nucleotide is fluorescently labeled. After fluorescence imaging, reactants are washed away, terminators are chemically removed, and another sequencing cycle can take place. At the end, all reads have the same length, as the number of sequencing cycles is the same for each cluster



Figure 7.3: Work flow of next-generation sequencing. Taken from [99].

7.7 NGS analysis pipeline

NGS technology has become a prominent tool in biological and biomedical research. However, NGS data analysis and the sequencing error-rate remain a major challenge. Different sequencing platforms generate various types of error. For example, 454 system has high error rate in insertion and deletion (indel) calls, while for the Illumina platform, indels are rare and the major sequencing errors come from miscall (with a typical rate of ~1%) [100]. In addition to identifying nucleotides, base-calling algorithms [100, 101] are also provided by NGS sequencers. This algorithm produces a quality score for each base indicating the probability of an incorrectly called base.

After the NGS reads are generated, the next step is aligning the reads to a reference genome or doing *de novo* assembly to reconstruct the original sequenced genome. This step is required as incorrectly aligned reads may lead to errors in SNP calling.

SNP or variant calling in the context of the NGS data analysis can be defined as the process of finding bases in the NGS data that differ from the reference genome. SNP calling is one of the most important applications of the NGS, with the challenge of separating the real variants from sequencing errors. SNP calling in early NGS studies is usually based on the simple filtering of quality scores in which only high-confidence bases would be kept. A commonly used cutoff is a Phred-type quality score of 20. However, this filtering based methods will lead to false negatives for heterozygosity calling in low to intermediate coverage datasets. Due to this disadvantage, most SNP calling methods use Bayesian algorithms to estimate the probability of calling a variant at a specific position.

Chapter 8

A hidden Markov-model for QTL-mapping based on whole-genome sequencing data

The analysis of polygenetic characteristics for mapping QTL has received a substantial amount of consideration in molecular genetics over the last decades. QTL analysis (see Section 7.5) links two types of information - phenotypic data (trait measurements) and genotypic data (usually molecular markers) - in an attempt to interpret the genetic basis of variation in traits and link certain phenotypes to specific regions of chromosomes. It requires reliable scoring of many genetic markers covering the entire genome. The advent of high-throughput technologies such as NGS [99] provides a new way to score large numbers of genetic markers. Recently, bulk segregant analysis (BSA, [3]) has been coupled with a high-throughput sequencing method that allows for simultaneous identification of genetic loci that contribute to the particular trait or phenotype of interest [102, 103]. QTL-mapping relies on the principle of cosegregation, i.e., molecular markers that are in close proximity of a specific gene will have a higher probability to be inherited together than markers that are not in close proximity of this gene (see Section 7.5).

The use of suitable statistical methods is always pivotal to analyze the genetic basis of phenotypic traits. A limited number of methods suited for gene-mapping based on markers identified by NGS have been proposed [103–108]. The methods are applicable to various quantitative traits with different levels of genetic complexity.

As an example, in [105], a method for BSA using NGS data based on the experimental design with an F2 population was proposed. In this method, each individual was measured for the trait of concern and subsequently two pools, i.e., with and without the trait, were selected. For each identified SNP, a G-test-statistic (8.1) was calculated for each pool using the observed number of reads.

$$G = 2\sum_{i=1}^{4} n_i \ln \frac{n_i}{\hat{n}_i},$$
(8.1)

Where n_i is the observed count and \hat{n}_i is the expected count of a 2×2 contingency table under the null hypothesis of no QTL in the vicinity of the selected SNP. The test-statistic was further averaged across neighboring SNPs. This approach was based on the smoothing version of the G-statistic using a smoothing kernel within a predefined window .The obtained *p*-values from the test-statistic detect the regions with contiguous significant SNPs. These regions were defined as a potential gene loci.

MULTIPOOL, a probabilistic model was proposed in [106]. In this model, a dynamic Bayesian network was applied to map genetic elements from pooled sequencing studies. A chromosome was first partitioned into discrete block with equal size. To each block, a hidden state reflecting the pool allele frequency (unobserved allele frequency) was assigned. Each block, i, may emit the observed allele frequency y_i according to its hidden state x_i . It was assumed that there is at most one casual block per analyzed chromosome. The value of the unobserved allele frequency was determined by the number of individuals N in a pool and the population allele frequency p. For each block a value for p was estimated and all estimated values for pwere tested against the null hypothesis of no association, p = 50%. If the population allele frequency p of a chromosomal region was significantly different from 50%, these segments could be identified as a potential gene loci related to the trait of interest.

In [108], a semi-parametric approach to map simultaneously gene loci for highethanol tolerance of yeast *S. cerevisiae* (see Section 2.3) based upon SNPs as molecular markers was suggested. In this approach, for each identified SNP, observed mismatch frequencies between the reads of the offspring and the parental reference strain were modelled by a binomial distribution with the probability of the difference between the parental and offspring strain. Genomic loci associated with the QTL were detected by analyzing trends in the mismatch frequencies along the genome. More specifically, the scatterplot smoother was applied with a smoothing spline as basis and a fixed number of knots to smooth trends. This method was able to model single as well as multiple pools of segregants at the same time. However, the identified region for potential QTLs were relatively wide.

To decrease the potential QTLs regions identified by scatter plot smoother, a hidden Markov-model (HMM) was proposed in [109] to map QTLs using NGS-assisted BSA. The HMM provides a flexible approach to classify each identified SNP into one of several pre-defined states with their own specific biological interpretation. The identified states of the HMM allow to identify genomic regions that may be likely to contain trait-related genes. The identified genomic regions of the HMM model are subparts of the relatively wide regions by the smoothing approach.

In [109], the HMM model was comprehensively compared with two methods suggested in [105] and [106]. The comparison indicated that for chromosome IX in which no casual gene(s) were identified [103, 108, 109], both the LOD-scores of MULTI-POOL and the smoothed G-values suggested that almost the complete chromosome contained a QTL. This finding is not in accordance with [103]. For chromosome XIV containing a QTL [103, 108, 109], only after filtering out the unreliable counts, both methods were able to detect the region of the causal genes. For MULTIPOOL a small chromosomal subregion was identified [109].

The HMM model proposed in [109] for mapping of multiple gene loci can also be extended in several ways. This includes, for instance, modelling two Markov-chains to consider the bi-directional dependance present in the data, including covariates to allow non-homogeneous transition matrix, and modelling multiple pools of segregants at the same time. These extensions were not considered in [109]. Therefore, in the remainder of this dissertation, the focus is on possible extensions of the HMM proposed in [109]. In particular, in Chapter 9, we explain how HMMs will be generalized to the bi-directional (dual) HMM. Non-homogenous HMM is described in Chapter 10. In Chapter 11, we present the joint HMM of multiple pools of segregants. First, however, we discuss the HMM proposed in [109] to map QTLs based on marker information obtained with high-throughput screening methods.

This chapter is organized as follows. In Section 8.1, we give a general introduction to the basic HMM. The term basic is defined in the sense that an HMM is based on a homogeneous Markov-chain without any trend or seasonal variation. The observations may be either discrete- or continuous-valued and we ignore information that may be available on covariates. An HMM used for gene localization based on the ethanoltolerant yeast dataset (Section 2.3) is described in Section 8.2. Concluding remarks and topics for further research are given in Section 8.3. In this dissertation we have limited our focus to the model to involve HMMs with a discrete valued observations.

8.1 Hidden Markov-models

Hidden Markov-models [110, 111] are a particular kind of mixture models, which allow for serially dependent observations. An HMM consists of two components. The first component is a "parameter process", a sequence $\{C_1, C_2, ...\}$ of discrete random variables that can assume one of m possible values ("states") from a set $\Sigma = \{s_1, s_2, ..., s_m\}$. We assume that the sequence of states forms a first-order Markov-chain, i.e., $P(C_i|C_1, ..., C_{i-1}) = P(C_i|C_{i-1})$. Moreover, we assume that, for all $i, P(C_i = s_j | C_{i-1} = s_k) \equiv \gamma_{jk}$, where γ_{jk} is the probability of transition (independent of step i) from state s_j to state s_k . The transition probabilities form the state transition probability matrix Γ , with the element in j-th row and k-th column equal to γ_{jk} . Additionally, we may need to specify the "initial state" probability distribution, i.e., $P(C_1 = s_j) \equiv \delta_j$, say.

The second part of an HMM is a "state dependent" process $\{X_1, X_2, ...\}$. The distribution of random variable X_i depends on state C_i , i.e., $P(X_i|X_1, ..., X_{i-1}, C_1, ..., C_i) = P(X_i|C_i)$. In particular, we assume that $P(X_i = x|C_i = s_j) \equiv p_j(x)$ is the "emission probability" of x for state s_j that depends on a (row)vector of parameters $\boldsymbol{\theta}_j$. We define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m)$.

In practice, we only observe a sequence of values $\{X_1 = x_1, X_2 = x_2, ..., X_N = x_N\}$, without the corresponding sequence of the generating states $\{C_1 = c_1, C_2 = c_2, ..., C_N = c_N\}$. In other words, the Markovian "parameter process" is hidden from the observer; hence the name of the model.

Figure 8.1 schematically presents the HMM described above. Note that transitions between the states are indicated by the arrows directed from left to right ("LtoR"). In applications, for finite sequences, one could also envisaged an HMM with transitions directed from right to left ("RtoL").

If the state sequence $\{C_1, C_2, ...\}$ is known exactly (i.e., $C_1 = c_1, C_2 = c_2, ...$), and given an HMM, the likelihood function based on the the observed sequence $\{X_1 = x_1, X_2 = x_2, ..., X_N = x_N\}$ and the state sequence can be represented as follows:

$$P(x_1, x_2, \dots, x_N, c_1, c_2, \dots, c_N) = \delta_{c_1} \gamma_{c_1, c_2} \dots \gamma_{c_{N-1}, c_N} p_{c_1}(x_1) \dots p_{c_N}(x_N).$$
(8.2)

The likelihood function, given in (8.2), is referred to as the complete-data likelihood, because it makes an assumption that both the observed sequence and the states are known.

Define $u_j(i)$ to be an indicator variable taking the value 1 if $C_i = s_j$ and 0 otherwise. Moreover, consider $v_{jk}(i)$ taking the value 1 if $C_{i-1} = s_j$ and $C_i = s_k$ and

0 otherwise. Sticking to the notation of an HMM, we can express the logarithm of (8.2) as follows:

$$\log(\mathbf{P}(x_1, ..., x_N, c_1, ..., c_N)) = \sum_{j=1}^m \{u_j(1) \log(\delta_j)\} + \sum_{j=1}^m \sum_{k=1}^m \{\sum_{i=2}^N v_{jk}(i) \log(\gamma_{jk})\} + \sum_{j=1}^m \sum_{i=1}^N \{u_j(i) \log p_j(x_i)\}.$$
 (8.3)

Note that the log-likelihood (8.3) is composed of three components, each depending on a different set of parameters.

However, in an HMM, the states are not observed. Thus, neither (8.2) nor (8.3) can be used to estimate the parameters of the model. Toward this aim, the observed-data likelihood has to be used, given by

$$L_N = P(x_1, x_2, \dots, x_N) = \sum_{j_1=1}^m \dots \sum_{j_N=1}^m \left\{ \delta_{j_1} \gamma_{j_1, j_2} \cdots \gamma_{j_{N-1}, j_N} p_{j_1}(x_1) \cdots p_{j_N}(x_N) \right\}.$$
(8.4)



Figure 8.1: A hidden Markov-model underlying the sequence of data values $(X_1, X_{i-1}, X_i, X_{i+1}, X_N)$. C_i is the hidden state for the observation X_i , $P(X_i|C_i)$ is the probability of emission of X_i in state C_i , and $P(C_i|C_{i-1})$ is the probability of transition from state C_{i-1} to state C_i .

In essence, the observed-data likelihood results from treating the sequence of states as missing data.

It is worth mentioning that the direct use of the likelihood (8.4) to estimate the parameters of an HMM is difficult and involve an enormous number of calculations even for a moderate N. However, this process can be simplified by using the forward-backward algorithm (also known as the Baum-Welch algorithm [112]), which is a

form of the Expectation-Maximization algorithm (EM, [113]). The algorithm is an iterative method for performing maximum likelihood estimation, while considering the sequence of the hidden states as missing.

After estimating the parameters of the HMM, the sequence of the hidden states, which could have generated the observed sequence of symbols, can be predicted. For instance, the most likely sequence ("global decoding") can be found with the help of the Viterbi algorithm [114]. Alternately, for each X_i , the most likely state can be assigned based on the conditional state-distribution given X_i ("local decoding"). Note that so-determined most likely state may differ from the one assigned by the Viterbi algorithm (though in particular applications the differences may be minor).

We discuss the forward-backward algorithm and estimation of the parameters of the HMM model in the following subsection.

8.1.1 Forward-backward algorithm and parameter estimates

The forward probabilities $\alpha_i(j)$ are defined as the joint probability of observing the sequence $\{x_1, ..., x_i\}$ when the state-chain ends in state s_j :

$$\alpha_i(j) = \Pr(X_1 = x_1, \dots, X_i = x_i, C_i = s_j).$$
(8.5)

The backward probabilities are the conditional probabilities that we observe the sequence $\{x_{i+1}, ..., x_N\}$ given that the underlying state-chain starts from state s_j :

$$\beta_i(j) = \Pr(X_{i+1} = x_{i+1}, ..., X_N = x_N | C_i = s_j).$$
(8.6)

We state the probabilities again below as a product of (row) vectors:

$$\boldsymbol{\alpha}_{i} = \boldsymbol{\delta} \boldsymbol{P}(x_{1}) \boldsymbol{\Gamma} \boldsymbol{P}(x_{2}) \dots \boldsymbol{\Gamma} \boldsymbol{P}(x_{i}) \quad for \ i = 1, \dots N,$$
(8.7)

$$\boldsymbol{\beta}_{i} = \boldsymbol{\Gamma} \boldsymbol{P}(x_{i+1}) \boldsymbol{\Gamma} \boldsymbol{P}(x_{i+2}) \dots \boldsymbol{\Gamma} \boldsymbol{P}(x_{N}) \boldsymbol{1}^{T}, \qquad (8.8)$$

where $\boldsymbol{\delta} \equiv (\delta_1, \delta_2, ..., \delta_m)$, $\boldsymbol{\Gamma}$ is the $m \times m$ matrix of transition probabilities and $\mathbf{P}_j(x)$ is the $m \times m$ diagonal matrix with the emission probability $p_j(x)$ as the *j*-th diagonal element. From (8.7) and (8.8), we can conclude that

$$L_N = \{ \boldsymbol{\delta} \mathbf{P}(x_1) \Gamma \mathbf{P}(x_2) ... \Gamma \mathbf{P}(x_i) \} \{ \Gamma \mathbf{P}(x_{i+1}) ... \Gamma \mathbf{P}(x_N) \mathbf{1}^T \}$$

= $\boldsymbol{\alpha}_i \boldsymbol{\beta}_i^T.$ (8.9)

The forward-backward algorithm iteratively computes the forward and backward probabilities. Each iteration of the algorithm consists of two steps: the expectation step (E-step), and the maximization step (M-step). In the E-step, $u_j(i)$ and $v_{jk}(i)$ are estimated given the observed data and current estimates of the model parameters. In the M-step, the likelihood function (8.2) is maximized with respect to δ , Γ , and θ , given the estimates of $u_j(i)$ and $v_{jk}(i)$. Since the parameters appear in separate components, we can seek the maximum of each of the three sums in (8.3) separately. These two steps are repeated until a convergence criterion is met. The convergence criterion used in this dissertation necessitates that the difference between the estimated parameters of two consecutive iterations should be smaller than the pre-defined tolerance. Standard errors of the parameter estimates are obtained by the method of Louis [115].

8.2 A hidden Markov-model for BSA experiments

8.2.1 The hidden states

In [109], it was proposed to use three states (m = 3) to map QTLs. These states correspond to linkage or no linkage with genes responsible for the phenotype of interest. For a given SNP, at location *i*, it is assumed that there is no linkage when the number of offspring nucleotides concordant with the parental strain without the trait is equal to the number of discordant nucleotides. As a consequence, the resulting SNP frequency is 50%. In case of linkage, the number of concordant and discordant nucleotides are no longer equal to each other. If the number of discordant nucleotides is larger than the number of concordant counts, linkage with a locus of the parent with the trait is assumed. In case of the opposite, i.e, the number of concordant is larger than the number of discordant, it is assumed that there is a linkage with a gene loci of the parent without the trait. The HMM proposed in [109] is uni-directional, i.e., the direction of the transitions between the underlying states corresponding to the observed data is "LtoR" or "RtoL".

8.2.2 State-dependent emission probabilities

For each identified SNP, the number of times nucleotides A, C, G, and T were present in the offspring are observed. For a given SNP at a chromosomal location i, there are four possible nucleotides that can be presented in the reference (parental) strain. Let us denote by $n_{kl,i}$ the number of times the nucleotides k and l, $(k, l \in \{A, C, G, T\})$ are observed in the reference strain and in the offspring, respectively. The emission probabilities of a given SNP at location i can be modeled using the multinomial distribution for the observed nucleotide counts by considering an *m*-state (j = 1, 2, ..., m)HMM model [109]:

$$P_j(n_{AA,i},...,n_{AT,i},n_{CA,i},...,n_{TT,i}) = \left(\begin{array}{c}n_i\\n_{AA,i},...,n_{TT,i}\end{array}\right) \prod_{k,l \in \{A,C,G,T\}} (\mu_{kl,j})^{n_{kl,i}},$$
(8.10)

where $n_i = \sum_{k,l} n_{kl,i}$ and $\mu_{kl,j}$ is the probability of observing the pair of nucleotides for the *j*-th state. It is worth mentioning that, at a particular location *i*, only one nucleotide, say *r*, is presented in the reference strain. As a result, only four nonzero counts: $n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}$ can be observed. Conditioning on the reference nucleotide *r*, i.e., on the fact that all other counts are necessarily equal to zero, the emission probabilities derived from (8.10) can be expressed as

$$P_{j}(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \left(\begin{array}{c} n_{i} \\ n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i} \end{array} \right) \prod_{l \in \{A,C,G,T\}} \left(\frac{\mu_{rl,j}}{\sum\limits_{s \in \{A,C,G,T\}} \mu_{rs,j}} \right)^{n_{rl,i}}.$$
 (8.11)

Several simplifications of the (conditional) emission probabilities are possible. For example, it can be assumed that the probability of observing an offspring nucleotide discordant with the reference one is independent from that nucleotide, i.e., the discordance probabilities $\mu_{rl,j}$, for $r \neq l$, can be considered to be all equal to $\mu'_{r,j}$. In addition, it can be assumed that these discordance probabilities are also independent from the reference nucleotide, i.e., $\mu'_{r,j} \equiv \mu'_j$ and the concordance probabilities $\mu_{rr,j} \equiv \mu$. Then it follows that

$$\sum_{\{A,C,G,T\}} \mu_{rs,j} = \mu_j + 3\mu'_j \tag{8.12}$$

and, since $\sum_{k} \sum_{l} \mu_{kl,j} = 1$, we obtain that

$$\sum_{r \in \{A,C,G,T\}} \left\{ \mu_{rr,j} + \sum_{s \in \{A,C,G,T\}, s \neq r} \mu_{rs,j} \right\} = 4\mu_j + 12\mu'_j = 1$$
(8.13)

and

$$\mu_j + 3\mu'_j = \frac{1}{4}.\tag{8.14}$$

Consequently, the total probability of concordance (match) between the reference (parental) and offspring nucleotide for state j is equal to $4\mu_j$, whereas the total

probability of discordance (mis-match) is equal to $12\mu'_j$. Taking into account the equations (8.11), (8.12), and (8.14) the emission probabilities of each state can be expressed as

$$P_{j}(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \begin{pmatrix} n_{i} \\ n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i} \end{pmatrix} 4^{n_{i}} (\mu_{j})^{n_{rr,i}} \prod_{l \neq r} (\mu_{j}')^{n_{rl,i}}.$$
 (8.15)

Thus, the main emission probability parameters are μ_j , with $3\mu'_j = 1/4 - \mu_j$. The concordance (match) and discordance(mis-match) probabilities can be further reparameterized to an unconstrained scale. This can be achieved by considering the following transformation:

$$\theta_j = \log\left(\frac{\mu_j}{3\mu'_j}\right). \tag{8.16}$$

As a consequence, (8.15) assumes the following form:

$$P_{j}(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \begin{pmatrix} n_{i} \\ n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i} \end{pmatrix} \left(\frac{e^{\theta_{j}}}{1 + e^{\theta_{j}}}\right)^{n_{rr,i}} \prod_{l \neq r} \left(\frac{1}{3(1 + e^{\theta_{j}})}\right)^{n_{rl,i}}.$$
 (8.17)

In [109], the possibility of including sequencing errors in the model have been proposed. For a particular chromosomal location, sequencing errors can appear in the reference sequence and/or in the offspring sequence. In [109], it was assumed that there are no sequencing errors in the reference sequence. In addition, the probability of a sequencing error in the offspring sequence was assumed not to depend on the sequenced nucleotide. After incorporation of the sequencing errors, the equation (8.17) takes the following form:

$$P_{j}(n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i}) = \begin{pmatrix} n_{i} \\ n_{rA,i}, n_{rC,i}, n_{rG,i}, n_{rT,i} \end{pmatrix} \left(\frac{e^{\theta_{j}}(1 - 3\varepsilon) + \varepsilon}{1 + e^{\theta_{j}}}\right)^{n_{rr,i}} \prod_{l \neq r} \left(\frac{1 + \varepsilon(3e^{\theta_{j}} - 1)}{3(1 + e^{\theta_{j}})}\right), \quad (8.18)$$

where $\varepsilon_{tot}/3 = \varepsilon$ is the sequencing error probability for the offspring. In [109], $\varepsilon_{tot} = 5\%$ was assumed [116].

8.3 Conclusions

The HMM model proposed in [109] is flexible. It allows to fix one or more parameters of the model. The number of the hidden states can also be changed. Several extensions of the HMM proposed in [109] are possible. For example, including covariates to allow a non-homogeneous HMM, considering multiple experiments at the same time, and modelling two Markov chains to consider the bi-directional dependence present in the data. These extensions were not considered in [109].

Chapter 9

A bi-directional (dual) hidden Markov-model for QTL-mapping

The basic HMM proposed in [109] deals with the state transition information in one direction at the time (left-to-right or right-to-left) across the chromosome. This means that the state of the *i*-th SNP depends on the state of the (i - 1)-th SNP or, in the case of the right-to-left direction, the state of the i-th SNP depends on the state of the (i + 1)-th SNP. We will refer in the remainder of this dissertation to such models as uni-directional HMMs. Clearly, both the preceding state (i-1)and following state (i + 1) carry useful information about a current state of the *i*th SNP. Thus, there is an understandable reason to expect that a model explicitly conditioning on both uni-directions at each SNP position could be advantageous. This model was not considered in [109]. Toward this aim, we present a bi-directional ("dual") HMM (DHMM, [117]) for QTL-mapping that accounts for the bi-directional dependence present in the data. This model consists of two uni-directional HMMs. One of the HMMs deals with the left-to-right ("LtoR") state transition, while the other considers the right-to-left ("RtoL") state transition. DHMMs consider the full information on both uni-directions in order to improve the prediction accuracy of HMMs for parameter estimation.

This chapter is organized as follows. In Section 9.1, we give a general introduction to DHMMs. We describe the DHMM to map gene loci based upon NGS in Section 9.2. Section 9.3 shows the application of the proposed DHMM on the ethanol-tolerant yeast dataset (Section 2.3). Concluding remarks are given in Section 9.4.

9.1 A bi-directional (dual) hidden Markov-model

A DHMM was proposed in [117]. It consists of two uni-directional HMMs. Each of these uni-directional HMMs deals with a particular direction of dependence between the hidden states, i.e., "LtoR" or "RtoL". Figure 9.1 illustrates, schematically, the DHMM construction. The transitions between the states are indicated by the arrows showing the particular direction of dependence among the hidden states. Denote, symbolically, the models for the "LtoR" or "RtoL" directions as $\lambda^{(LtoR)}$ and $\lambda^{(RtoL)}$, respectively. Then, the DHMM can be expressed as $\lambda = (\lambda^{(LtoR)}, \lambda^{(RtoL)})$ where $\lambda^{(LtoR)}$ and $\lambda^{(RtoL)}$ are defined by the triplets $(\delta^{(LtoR)}, \Gamma^{(LtoR)}, \theta)$ and $(\delta^{(RtoL)}, \Gamma^{(RtoL)}, \theta)$, respectively, with $\delta^{(LtoR)} = (\delta_1^{(LtoR)}, ..., \delta_m^{(LtoR)})$ and $\Gamma^{(LtoR)}$ denoting, respectively, the initial state distribution and transition probability matrix for the "LtoR" HMM and $\delta^{(RtoL)}$ and $\Gamma^{(RtoL)}$ are the corresponding quantities for the "RtoL" HMM. Note that the DHMM involves the emission-probability distribution parameters θ that are assumed to be the same for both uni-directional models.



Figure 9.1: Bi-directional hidden Markov-model underlying the sequence of data values $(X_1, X_{i-1}, X_i, X_{i+1}, X_N)$, $P(C_i|C_{i-1})$ is the probability of transition from state C_{i-1} to state C_i indicating the "LtoR" transition and $P(C_i|C_{i+1})$ is the probability of transition from state C_{i+1} to state C_i illustrating the "RtoL" transition.

Consider a random state sequence $\{C_1 = c_1, C_2 = c_2, ..., C_N = c_N\}$. The probability of observing the sequence under the "LtoR" and the "RtoL" directions can be expressed as follows:

$$P(C_1 = c_1, C_2 = c_2, ..., C_N = c_N) = P\left(C_1 = c_1 | \lambda^{(LtoR)}\right) \times ... \times P\left(C_N = c_N | C_{N-1} = c_{N-1}, \lambda^{(LtoR)}\right).$$
(9.1)

$$P(C_{1} = c_{1}, C_{2} = c_{2}, ..., C_{N} = c_{N}) = P(C_{N} = c_{N} | \lambda^{(RtoL)}) \times ... \times P(C_{1} = c_{1} | C_{2} = c_{2}, \lambda^{(RtoL)}).$$
(9.2)

In [117], it was proposed to compute the total probability of $\{C_1 = c_1, C_2 = c_2, ..., C_N = c_N\}$ under the DHMM as follows:

$$P(C_{1} = c_{1}, ..., C_{N} = c_{N}) = \left\{ P\left(C_{1} = c_{1}, ..., C_{N} = c_{N} | \lambda^{(LtoR)}\right) + P\left(C_{1} = c_{1}, ..., C_{N} = c_{N} | \lambda^{(RtoL)}\right) \right\} / 2.$$
(9.3)

Consequently, given a particular DHMM, information is used from both directions. Using a weighting method by applying a mixture density [118], we combine information from both directions. In particular, the probability of an observed sequence $\{X_1 = x_1, X_2 = x_2, ..., X_N = x_N\}$ can be expresses as

$$L_N = \left\{ \pi_{(LtoR)} L_N^{(LtoR)} + \pi_{(RtoL)} L_N^{(RtoL)} \right\},$$
(9.4)

in which $\pi_{(LtoR)}$ and $\pi_{(RtoL)}$ are two nonnegative parameters that have to be estimated such that $\pi_{(LtoR)} + \pi_{(RtoL)} = 1$ and $L_N^{(LtoR)}$ and $L_N^{(RtoL)}$ are defined according to (8.4). The EM algorithms are very well studied for computing the maximum likelihood estimates of mixture model parameter [118]. In this model, we weight both directions equally, i.e., $\pi_{(LtoR)} = \pi_{(RtoL)} = 1/2$.

The forward-backward probabilities of the "LtoR" direction can be computed as explained in Subsection 8.1.1. The forward-backward component probabilities of the "RtoL" direction can be computed similarly for the "LtoR" direction:

$$\alpha_i^{(RtoL)}(j) = \Pr(X_T = x_T, ..., X_i = x_i, C_i = s_j),$$
(9.5)

$$\beta_i^{(RtoL)}(j) = \Pr(X_{i-1} = x_{i-1}, \dots, X_1 = x_1 | C_i = s_j).$$
(9.6)

Define $u_j^{(LtoR)}(i)$ and $v_{jk}^{(LtoR)}(i)$ to be indicator variables of the "LtoR" direction as explained in Section 8.1. The indicator variables of the "RtoL" direction, i.e., $u_j^{(RtoL)}(i)$ and $v_{jk}^{(RtoL)}(i)$, can be considered similarly for the "LtoR" direction. The parameters involved in (9.4) can be estimated by using the joint forward-backward algorithm. In the E-step, $u_j^{(LtoR)}(i)$, $v_{jk}^{(LtoR)}(i)$, $u_j^{(RtoL)}(i)$, and $v_{jk}^{(RtoL)}(i)$ are estimated given the observed data and current estimates of the model parameters. Then, in the M-step, the logarithm of the likelihood function (9.4) is maximized with respect to $\boldsymbol{\delta}^{LtoR}$, $\boldsymbol{\Gamma}^{LtoR}$, $\boldsymbol{\delta}^{RtoL}$, $\boldsymbol{\Gamma}^{RtoL}$, and $\boldsymbol{\theta}$, given the estimates of $u_j^{(LtoR)}(i)$, $v_{jk}^{(LtoR)}(i)$, $u_j^{(RtoL)}(i)$, and $v_{jk}^{(RtoL)}(i)$.

After estimating the parameters of the DHMM, the sequence of the hidden states, which could have generated the observed sequence of symbols, can be predicted. In particular, the most likely sequence of states for the DHMM is retrieved by taking the average of the Viterbi path [114] assigned to each uni-directions.

9.2 A dual hidden Markov-model for BSA experiments

The structure of the proposed DHMM is as follows. We consider three states, i.e., m = 3. The hidden states and the state-dependent emission probabilities are defined as in Subsection 8.2.1 and 8.2.2, respectively.

9.3 Results

We applied the DHMM, outlined in Section 9.2, to the data for chromosome XIV and chromosome IX of *S. cerevisiae* that were obtained in the experiment described in Section 2.3. On chromosome XIV, three genes responsible for high ethanol tolerance, i.e., APJ1, MKT1 and SWS2, have been identified with the help of artificial markers and a scatterplot smoother [103, 108]. All three QTLs, were found at approximately 470,000 *bp*. The same approach did not identify causal genes for ethanol tolerance on chromosome IX.

9.3.1 Uni-diectional HMMs and filtering

There are several issues about the data produced by BSA-NGS that need to be taken into account prior to the analysis. As it has already mentioned in Subsection 8.2.2, NGS process is error-prone, thus the possibility of including sequencing errors in the model was proposed in [109]. After including the sequencing-error correction a large number of SNPs with a low mismatch frequency, probably correspond to sequencing error, are assigned to the first state 1 (top panel of Figure 9.2). To remove the SNPs corresponding to the sequencing error, we can apply the filtering approach as proposed in [108] to the dataset, and subsequently use (8.18) (with error sequence equal to 5%). Filtering approach considers two selection criteria: the nucleotide should be sequenced at least 20 times [116] and have a SNP frequency of at least 80%. The scatterplot of mismatch frequencies along chromosome XIV, shown in bottom panel of Figure 9.2, illustrates the substantial increase of the reliable SNPs after including filtering approach. Table 9.1 presents the parameter estimates of the HMMs for the "LtoR" and "RtoL" directions.



Figure 9.2: The mismatch frequency of SNPs on chromosome XIV, pool 1. Top panel: including (8.18); bottom panel: including filtering approach.

Table 9.1: Parameter estimation of the three-state uni-directional HMM model for pool 1 segregants of chromosome XIV with sequencing-error correction and filtering approach. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2269 \ [0.220, \ 0.230]$ | 0.2268 [0.220, 0.230] |
| μ_2 | $0.1284 \ [0.127, \ 0.129]$ | $0.1284 \ [0.127, \ 0.129]$ |
| μ_3 | $0.0556\ [0.053,\ 0.057]$ | $0.0559\ [0.054,\ 0.056]$ |
| δ_1 | 0+ | 0+ |
| δ_2 | 1 ⁺ | 1 ⁺ |
| δ_3 | 0^{+} | 0+ |
| γ_{11} | $0.4511 \ [0.4415, \ 0.4611]$ | $0.4512 \ [0.4415, \ 0.4611]$ |
| γ_{12} | $0.4333\ [0.4242,\ 0.4426]$ | $0.4348\ [0.4257,\ 0.4441]$ |
| γ_{13} | $0.1151\ [0.1083,\ 0.1219]$ | 0.1138[0.1071,0.1205] |
| γ_{21} | $0.0213\ [0.0204,\ 0.0223]$ | $0.0214\ [0.0203,\ 0.0222]$ |
| γ_{22} | $0.9222\ [0.9217,\ 0.9227]$ | $0.9226\ [0.9222,\ 0.9230]$ |
| γ_{23} | $0.0561\ [0.0546,\ 0.0575]$ | $0.0560\ [0.0546,\ 0.0574]$ |
| γ_{31} | $0.0073\ [0.0036,\ 0.0110]$ | $0.0079\ [0.0043,\ 0.0115]$ |
| γ_{32} | $0.0739\ [0.0723,\ 0.0755]$ | $0.0737\ [0.0723,\ 0.0751]$ |
| γ_{33} | $0.9187 \ [0.9183, \ 0.9193]$ | $0.9189 \ [0.9184, \ 0.9194]$ |

9.3.2 Chromosome XIV, pool 1

Table 9.2 presents the estimates of the parameters of the DHMM. Note that the estimates of the concordance probabilities μ_j are assumed to be the same for the "LtoR" and "RtoL" directions. The total probabilities of discordance between the no-trait reference parent and the offspring can be estimated to be equal to $1-4 \times 0.2268=0.0928$, $1-4 \times 0.1284=0.4864$, and $1-4 \times 0.059=0.7764$ for the first, second, and third state, respectively. According to the interpretation of the hidden states (Subsection 8.2.1), the SNPs admitted to the second state are considered not to be linked with any gene(s) responsible for ethanol tolerance, whereas SNPs in the third state are linked to the gene(s) responsible for high-ethanol tolerance. The SNPs in the first state could be linked to gene(s) from the parent without the trait. However, due to the low number of SNPs in the first state (left-hand-side panel of Figure 9.3), and due to the fact that they are spread across chromosome XIV, one should consider these SNPs to be

Table 9.2: Parameter estimation of the three-state DHHM model for pool 1 segregants of chromosome XIV with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2268 \ [0.220, \ 0.230]$ | 0.2268 [0.220, 0.230] |
| μ_2 | $0.1284 \ [0.127, \ 0.129]$ | $0.1284 \ [0.127, \ 0.129]$ |
| μ_3 | $0.0559\ [0.054,\ 0.056]$ | $0.0559\ [0.054,\ 0.056]$ |
| δ_1 | 0+ | 0+ |
| δ_2 | 1 ⁺ | 1 ⁺ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.4512\ [0.4414,\ 0.4610]$ | $0.4512 \ [0.4415, \ 0.4611]$ |
| γ_{12} | $0.4333\ [0.4243,\ 0.4424]$ | $0.4348 \ [0.4257, \ 0.4441]$ |
| γ_{13} | $0.1153\ [0.1087,\ 0.1218]$ | $0.1138\ [0.1072,\ 0.1203]$ |
| γ_{21} | $0.0214\ [0.0204,\ 0.0223]$ | $0.0213\ [0.0203,\ 0.0222]$ |
| γ_{22} | $0.9223\ [0.9218,\ 0.9228]$ | $0.9225 \ [0.9220, \ 0.9230]$ |
| γ_{23} | $0.0561\ [0.0546,\ 0.0575]$ | $0.0561 \ [0.0547, \ 0.0576]$ |
| γ_{31} | $0.0073\ [0.0038,\ 0.0108]$ | $0.0074 \ [0.0038, \ 0.0110]$ |
| γ_{32} | $0.0737\ [0.0722,\ 0.0751]$ | $0.0736\ [0.0721,\ 0.0750]$ |
| γ_{33} | $0.9188 \ [0.9183, \ 0.9193]$ | $0.9188 \ [0.9183, \ 0.9193]$ |

the result of sequencing errors. The estimated transition probabilities indicate that the most likely transitions are from a given state to itself, as can be seen from the estimated values of $\gamma_{11} = 0.45$, $\gamma_{22} = 0.92$, and $\gamma_{33} = 0.92$. The estimated initial state probabilities indicate that the observed SNP frequency for the first SNP is most likely generated by the second state. The left-hand side panel of Figure 9.3 illustrates the states predicted for each SNP based on the most likely state sequence (global decoding) obtained from the fitted DHMM.

The plot clearly illustrates that state 3 (blue) is associated with a high SNP frequency, state 2 (green) with an intermediate frequency, and state 1 (red) with a low frequency. This panel shows a large number of SNPs to be linked to potential QTLs (blue). The SNPs shown in the right-hand-side panel of Figure 9.3 are in a chromosomal region which contains three genes responsible for high ethanol-tolerance [103]. In the plot, consecutive state-3 SNPs were joined by intervals if there was

no state-2 SNP between them (we ignore state-1 SNPs). These state-3-SNP-series illustrate the potential region where a QTL might be present.



Figure 9.3: Chromosome XIV, pool 1, predicted SNP-specific states for the DHMM model (left panel). Red stands for state 1, green for state 2, and blue for state 3 and zoom-in into the region (right panel) of three identified genes. The lines connect consecutive state-3 SNPs. The SNPs that are part of one of the three identified genes are located between orange lines.

9.3.3 Chromosome XIV, pool 2

The segregants of the second pool were screened for a higher tolerance as compared to pool 1, i.e., 17% versus 16%. Table 9.3 presents the estimates of the parameters of the DHMM. Also for this pool the estimates of the concordance probabilities μ_j are assumed to be the same for the "LtoR" and "RtoL" direction. The estimated total discordance probabilities are equal to 0.322 for the first state, 0.608 for the second state, and 0.87 for the third state. The transition probability estimates are almost identical for both directions. The interpretation of the second hidden state is no longer compatible with the one given in the analysis of pool-1 segregants. In contrast to what was assumed, the total discordance frequency of approximately 60% does not suggest that the second state is a state where the SNPs are not linked to a QTL. The second state could represent the SNPs which exhibit a weak linkage with QTL from the parent with the trait (Figure 9.4). The main reason for finding a second state

Table 9.3: Parameter estimation of the of the three-state DHHM model for the pool 2 segregants of chromosome XIV with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.1696\ [0.168,\ 0.171]$ | $0.1696 \ [0.168, \ 0.171]$ |
| μ_2 | $0.0979 \ [0.096, \ 0.099]$ | $0.0979 \ [0.096, \ 0.099]$ |
| μ_3 | $0.0309 \ [0.029, \ 0.032]$ | $0.0309 \ [0.029, \ 0.032]$ |
| δ_1 | 1+ | 0+ |
| δ_2 | 0^{+} | 1+ |
| δ_1 | 0^{+} | 0^{+} |
| γ_{11} | $0.8582 \ [0.7606, \ 0.9250]$ | $0.8593 \ [0.7617, \ 0.9267]$ |
| γ_{12} | $0.1249\;[0.1176,0.1325]$ | $0.1288 \ [0.1215, \ 0.1364]$ |
| γ_{13} | $0.0167\ [0.0112,\ 0.0256]$ | $0.0118 \ [0.0108, \ 0.0127]$ |
| γ_{21} | $0.1054\ [0.085,\ 0.12770]$ | $0.1027 \ [0.0823, \ 0.1250]$ |
| γ_{22} | $0.8163\ [0.7214,\ 0.8827]$ | $0.8151\ [0.7202,\ 0.8815]$ |
| γ_{23} | $0.0781 \ [0.0775, \ 0.0786]$ | $0.0821 \ [0.0815, \ 0.0826]$ |
| γ_{31} | $0.0079 \ [0.0044, \ 0.0140]$ | $0.0113 \ [0.0055, \ 0.0209]$ |
| γ_{32} | $0.0674 \ [0.0668, \ 0.0679]$ | $0.0640 \ [0.0634, \ 0.0645]$ |
| γ_{33} | $0.9246 \ [0.9073, \ 0.9384]$ | $0.9246 \ [0.9073, \ 0.9384]$ |

with discordance probabilities larger than expected is related to the assumption of a three-state DHMM model. As a result, regardless of the fact that there could be more than three states representing the data, the DHMM will classify the SNPs into three states.

The proposed DHMM with three states can be conveniently modified. For instance, the number of hidden states can be changed or one or more parameters of the DHMM can be fixed [109]. However, in this dissertation we do not modify the number of hidden states of the DHMM. Instead, we illustrate the effect of fixing the concordance probabilities. The initial state probabilities and transition probabilities can also be fixed, but we do not consider such a modification in this dissertation. In particular, we fix μ_2 to be equal to 0.125, forcing the total discordance probability to be equal to 0.50 for the second state. Table 9.4 illustrates the detailed results for the DHMM with the fixed value of μ_2 . Fixing the concordance probability influences the other estimated parameters (Table 9.4). In particular, the estimated transition probabilities γ_{jk} , initial state probability of the "LtoR" direction, and μ_1 change as compared to the model where no probabilities have been fixed (Table 9.3). These changes have an effect on the allocation of the SNPs to three states, i.e., the larger number of SNPs are classified to state 2 (Figure 9.5). This indicates that, in fact, for most of the state-2 SNPs for the latter model, the discordance probability of 0.5 (i.e., no linkage) could be assumed.



Figure 9.4: Chromosome XIV, predicted SNP-specific states for pool 2. The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3.
Table 9.4: Parameter estimation of the three-state DHHM model for pool 2 segregants of chromosome XIV with sequencing-error correction, when μ_2 is fixed. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with ⁺ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-----------------------------|-------------------------------|
| μ_1 | $0.1821 \ [0.180, \ 0.183]$ | $0.1821 \ [0.180, \ 0.183]$ |
| μ_2 | 0.125^{+} | 0.125^{+} |
| μ_3 | $0.0356\ [0.033,\ 0.037]$ | 0.0356[0.033, 0.037] |
| δ_1 | 0+ | 0+ |
| δ_2 | 1 ⁺ | 1 ⁺ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.8343\ [0.7367,\ 0.9011]$ | $0.8368 \ [0.7392, \ 0.9036]$ |
| γ_{12} | $0.1453\ [0.1380,\ 0.1529]$ | $0.1482 \ [0.1409, \ 0.1558]$ |
| γ_{13} | $0.0202\ [0.0147,\ 0.0291]$ | $0.0149 \ [0.0094, \ 0.0238]$ |
| γ_{21} | $0.0704\ [0.0698,\ 0.0710]$ | $0.0677 \ [0.0671, \ 0.0683]$ |
| γ_{22} | $0.8555\ [0.7606,\ 0.9219]$ | $0.8555 \ [0.7606, \ 0.9219]$ |
| γ_{23} | $0.0739\ [0.0733,\ 0.0745]$ | $0.0767 \ [0.0761, \ 0.0773]$ |
| γ_{31} | $0.0056\ [0.0021,\ 0.0117]$ | $0.0076 \ [0.0041, \ 0.0137]$ |
| γ_{32} | $0.0613\ [0.0607,\ 0.0619]$ | $0.0594 \ [0.0588, \ 0.0600]$ |
| γ_{33} | $0.9329\ [0.9156,\ 0.9467]$ | $0.9328 \ [0.9155, \ 0.9466]$ |

9.3.4 Chromosome IX, pool 1

Table 9.5 presents the parameter estimates of the DHHM obtained for the pool-1 data for chromosome IX. The total discordance probability is estimated to be equal to 0.08 for the first state, 0.38 for the second state, and 0.532 for the third state. The total discordance probability of 0.532 for state 3 indicates that SNPs in this state are not linked to loci from the parent with the trait. This finding is accordance with [103].

Figure 9.6 suggests that the second state does not correspond to a state where no linkage is present. It suggests a state related to chromosomal regions linked with gene loci from parent without the trait. The main reason for finding a second state with lower discordance probability than expected, is related to the assumed number of states (m = 3). Even if, in reality, there are only two states explaining the different regions, a three-state DHMM classifies the SNPS into three groups (Figure 9.6).

96 Chapter 9. A bi-directional (dual) hidden Markov-model for QTL-mapping



Figure 9.5: Chromosome XIV pool 2, predicted SNP-specific states for the DHMM model with fixed μ_2 . The colors indicate the state of the SNP. Red for state 1, green for state 2 , and blue for state 3.



Chromosome IX Bidirectional

Figure 9.6: Chromosome IX pool 1, predicted SNP-specific states for the DHMM model (left panel). Red stands for state 1, green for state 2, and blue for state 3.

Table 9.5: Parameter estimation of the of the three-state DHHM model for the pool 1 segregants of chromosome IX with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2366 \ [0.231, \ 0.247]$ | $0.2366 \ [0.231, \ 0.247]$ |
| μ_2 | $0.1553 \ [0.153, \ 0.156]$ | $0.1553 \ [0.153, \ 0.156]$ |
| μ_3 | $0.1172 \ [0.115, \ 0.118]$ | $0.1172 \ [0.115, \ 0.118]$ |
| δ_1 | 1+ | 0+ |
| δ_2 | 0^{+} | 1 ⁺ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.5273 \ [0.5176, \ 0.5372]$ | $0.5277 \ [0.5180, \ 0.5376]$ |
| γ_{12} | $0.3738\ [0.3697,\ 0.3779]$ | $0.3365\ [0.3324,\ 0.3406]$ |
| γ_{13} | $0.0987\ [0.0981,\ 0.0993]$ | $0.1356\ [0.1152,\ 0.1560]$ |
| γ_{21} | $0.0423\ [0.0408,\ 0.0438]$ | $0.0473\ [0.0458,\ 0.0488]$ |
| γ_{22} | $0.8092\ [0.8085,\ 0.8099]$ | $0.8068\ [0.8061,\ 0.8075]$ |
| γ_{23} | $0.1483\ [0.1417,\ 0.1549]$ | $0.1457\ [0.1401,\ 0.1513]$ |
| γ_{31} | $0.0197\ [0.0187,\ 0.0207]$ | $0.0146 \ [0.0132, \ 0.0160]$ |
| γ_{32} | $0.1689\ [0.1485,\ 0.1893]$ | $0.1723 \ [0.1503, \ 0.1943]$ |
| γ_{33} | $0.8113\ [0.7164,\ 0.9062]$ | 0.8130[0.7181,0.9079] |

The proposed DHMM with three states can be modified by fixing μ_2 to be equal to 0.125, forcing the total discordance probability to be equal to 0.50 for the second state. The estimated transition probabilities γ_{jk} and μ_1 change when μ_2 is fixed (see Table 9.6). A very small number of SNPs is classified to state 3 (see Figure 9.7), which do not form any contiguous region. Thus, we can indeed conclude that chromosome IX does not contain QTLs responsible for high-ethanol tolerance. Though the number of SNPs assigned to state 1 is larger than for the model with unconstrained μ_2 (Figure 9.6), there also does not seem to be any contiguous region where state 1 SNPs would be most prevalent. This can be taken as an argument against the suggestion that state-1 SNPs could be located in chromosomal regions linked with gene loci from the parent without the trait.

Table 9.6: Parameter estimation of the three-state DHHM model for the pool 1 segregants of chromosome IX with sequencing-error correction, when μ_2 is fixed. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with ⁺ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-----------------------------|-------------------------------|
| μ_1 | $0.1916 \ [0.185, \ 0.197]$ | $0.1916 \ [0.185, \ 0.197]$ |
| μ_2 | 0.125^{+} | 0.125^{+} |
| μ_3 | $0.0922 \ [0.081, \ 0.103]$ | 0.0922[0.081, 0.103] |
| δ_1 | 1+ | 0+ |
| δ_2 | 0^{+} | 1 ⁺ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.6182\ [0.6085,\ 0.6279]$ | $0.6201 \ [0.6106, \ 0.6296]$ |
| γ_{12} | $0.3817\ [0.3773,\ 0.3861]$ | $0.3798\ [0.3760,\ 0.3837]$ |
| γ_{13} | 0^{+} | 0^{+} |
| γ_{21} | $0.1294\ [0.1281,\ 0.1307]$ | $0.1301\ [0.1286,\ 0.1316]$ |
| γ_{22} | $0.8363\ [0.8348,\ 0.8378]$ | $0.8234\ [0.8215,\ 0.8253]$ |
| γ_{23} | $0.0341\ [0.0270,\ 0.0412]$ | $0.0463\ [0.0386,\ 0.0540]$ |
| γ_{31} | $0.0276\ [0.0275,\ 0.0277]$ | $0.0224\ [0.0223,\ 0.0225]$ |
| γ_{32} | $0.3567\ [0.3363,\ 0.3771]$ | $0.3346\ [0.3046,\ 0.3646]$ |
| γ_{33} | 0.6156[0.5926, 0.6386] | $0.6429 \ [0.6219, \ 0.6639]$ |

9.3.5 Chromosome IX, pool 2

Table 9.7 presents the parameter estimates of the DHHM for the pool-2 data of chromosome IX. The total discordance probabilities are equal to 0.06, 0.48, and 0.624 for the first, second, and the third state, respectively. No QTLs were identified for this pool [109]. Note that the estimated SNP frequency of the third state is above 0.5. To check the sensitivity of the conclusions to the structure of the model, we fix μ_2 to be equal to 0.125, forcing the total discordance probability to be equal to 0.50 for the second state. The results are presented in Table 9.8. The estimated total discordance probabilities for state 1 and 3 are now equal to 0.088 and 0.656, respectively. Smaller number of SNPs are now assigned to state 3 (Figure 9.9) as compared to the model with unconstrained μ_2 (Figure 9.8), which are scattered widely across the whole chromosome and do not form any contiguous region.



Figure 9.7: Chromosome IX pool 1, predicted SNP-specific states for the DHMM model with fixed μ_2 . The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3.



Figure 9.8: Chromosome IX pool 2, predicted SNP-specific states for the DHMM model (left panel). Red stands for state 1, green for state 2, and blue for state 3.

100 Chapter 9. A bi-directional (dual) hidden Markov-model for QTL-mapping

Table 9.7: Parameter estimation of the of the three-state DHHM model for the pool 2 segregants of chromosome IX with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2356\ [0.229,\ 0.238]$ | $0.2356\ [0.229,\ 0.238]$ |
| μ_2 | $0.1380\ [0.136,\ 0.139]$ | $0.1380\ [0.136,\ 0.139]$ |
| μ_3 | $0.0944\ [0.093,\ 0.095]$ | $0.0944\ [0.093,\ 0.095]$ |
| δ_1 | 1+ | 0^{+} |
| δ_2 | 0^{+} | 1+ |
| δ_3 | 0^+ | 0^{+} |
| γ_{11} | $0.5325 \ [0.5234, \ 0.5418]$ | $0.5340 \ [0.5249, \ 0.5433]$ |
| γ_{12} | $0.3764\ [0.3723,\ 0.3805]$ | $0.3349 \ [0.3306, \ 0.3392]$ |
| γ_{13} | $0.0909 \ [0.0892, \ 0.0926]$ | $0.1309\ [0.1290,\ 0.1328]$ |
| γ_{21} | $0.0423 \ [0.0407, \ 0.0439]$ | $0.0481 \ [0.0463, \ 0.0499]$ |
| γ_{22} | $0.7261\ [0.6642,\ 0.7880]$ | $0.7235 \ [0.6533, \ 0.7937]$ |
| γ_{23} | $0.2314\ [0.2242,\ 0.2386]$ | $0.2283 \ [0.2210, \ 0.2356]$ |
| γ_{31} | $0.0146\ [0.0081,\ 0.0211]$ | $0.0104 \ [0.0042, \ 0.0166]$ |
| γ_{32} | $0.2045 \ [0.2002, \ 0.2088]$ | $0.2062 \ [0.2021, \ 0.2103]$ |
| γ_{33} | $0.7808 \ [0.7089, \ 0.8527]$ | $0.7833 \ [0.7103, \ 0.8563]$ |

Therefore, we can conclude that chromosome IX does not contain QTLs responsible for high ethanol-tolerance.

Table 9.8: Parameter estimation of the of the three-state DHHM model for the pool 2 segregants of chromosome IX with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2279 \ [0.2185, \ 0.2329]$ | $0.2279 \ [0.2185, \ 0.2329]$ |
| μ_2 | 0.125^{+} | 0.125^{+} |
| μ_3 | $0.0859\ [0.0849,\ 0.0869]$ | $0.0859\ [0.0849,\ 0.0869]$ |
| δ_1 | 1+ | 0+ |
| δ_2 | 0^{+} | 1+ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.5632 \ [0.5541, \ 0.5723]$ | $0.5647 \ [0.5555, \ 0.5739]$ |
| γ_{12} | $0.3556\ [0.3513,\ 0.3599]$ | $0.3549 \ [0.3507, \ 0.3591]$ |
| γ_{13} | $0.0802 \ [0.0783, \ 0.0821]$ | $0.0804 \ [0.0785, \ 0.0823]$ |
| γ_{21} | $0.0521 \ [0.0505, \ 0.0537]$ | $0.0491 \ [0.0473, \ 0.0509]$ |
| γ_{22} | $0.7261 \ [0.6642, \ 0.788]$ | $0.7251 \ [0.6549, \ 0.7953]$ |
| γ_{23} | $0.2218 \ [0.2149, \ 0.2287]$ | $0.2258 \ [0.2189, \ 0.2327]$ |
| γ_{31} | $0.0181 \ [0.0064, \ 0.0298]$ | $0.0144 \ [0.0084, \ 0.0204]$ |
| γ_{32} | $0.2045 \ [0.2006, \ 0.2084]$ | $0.2069 \ [0.2029, \ 0.2109]$ |
| γ_{33} | $0.7774 \ [0.7059, \ 0.8489]$ | $0.7732 \ [0.7012, \ 0.8452]$ |

102 Chapter 9. A bi-directional (dual) hidden Markov-model for QTL-mapping



Figure 9.9: Chromosome IX pool 2, predicted SNP-specific states for the DHMM model (left panel). Red stands for state 1, green for state 2, and blue for state 3.

9.3.6 Comparison of the DHMM with the basic HMM

To appreciate the differences between the DHMM and the uni-directional HMM proposed in [109], we compare the two models in terms of the parameter estimates and the SNPs allocation for the pool-1 data for chromosome XIV. Table 9.1 in Subsection 9.3.1 presents the parameter estimates of the HMMs for the "LtoR" and "RtoL" directions. It can be seen that the estimates of all the parameters for the two models are very similar. It is therefore not surprising that the results of the DHMM, shown in Table 9.2, are also very close to those obtained for the uni-directional HMMs, with a minimally higher precision for some of the DHMM estimated like, e.g., μ_3 . As a result, the state assignment is also similar for the two types of models (see Figure 9.10).



Figure 9.10: Chromosome XIV, pool 1, predicted SNP-specific states for the DHMM model with fixed μ_2 . The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3.

9.4 Conclusions

In this chapter, we have presented the application of the DHMM in QTL-mapping for high ethanol-tolerance. In the uni-directional HMM, the state of the *i*-th SNP depends on either the state of the (i-1)-th SNP in terms of the "LtoR" direction, or the state of the (i+1)-th SNP when considering the "RtoL" direction. Therefore, the uni-directional HMM ignores the information present in one of the directions. In the DHMM, information is used from both directions. The DHMM model was applied to data for chromosome XIV and chromosome IX of the case study. Based on the model, the potential regions responsible for high ethanol-tolerance on chromosome XIV could be identified, whereas no such regions were found on chromosome IX. These results are consistent with the previous findings [103, 108]. The comparison of the uni-directional HMM and the DHMM for chromosome XIV revealed only a slight difference in terms of the parameter estimates, with a minimal gain in precision of the estimation for the DHMM. As a result, the DHMM and the uni-directional HMMs assigned the SNPs to the same states. The main advantage of the DHMM is the fact that it produces a single set of estimates of the parameters of interest, i.e., emission (concordance) probabilities. However, this advantage comes at a price of increasing in the computational time and complexity of the model related to the joint forward-backward algorithm as compared to the basic HMM. As an example, for pool

1 segregants of chromosome XIV, fitting the model takes 3h 48m 22s and 9h 26m 01s for the basic HMM and the DHHM model respectively (on an HP Elitebook 8530p). The initial estimates of the joint forward-backward algorithm for the state-dependent emission probabilities are chosen in a way that the first state corresponds to linkage with a locus of the parent without the trait, the third state presents the linkage with a locus of the parent with the trait, and the second state does not show any linkage. Therefore, the total discordance probabilities of 0.2, 0.5, and 0.8 are considered for the first, second and the third state, respectively. In addition, the initial estimates of the transition matrix for all states are selected as the probabilities from a given state to itself is 0.8 and to another is 0.1. For pool 1 segregants of chromosome XIV, different initial estimates were chosen to check the sensitivity of the model to different starting points. The initial values and the estimates are presented in Table A.3 in the Appendix. The comparison of Table 9.2 and Table A.3 indicates that the choice of the initial values does not have a substantial influence on the parameter estimates. Although the filtering approach substantially increased the number of reliable SNPs, some SNPs with low mis-match frequency probably due to the sequencing error still remain (see red circles in the left-hand-side panel of Figure 9.3). Therefore we kept the same sequencing error of the basic HMM.

The simplification of the emission model was not assessed in this chapter, as the main intention was the comparison of the basic HMM proposed in [109] with its possible extensions while considering the similar emission model and the number of the hidden states. The simplification of the emission model towards the binomial distribution is possible and could reduce the complexity of the emission models in terms of the interpretation and notations. This could be a topic for future research.

Chapter 10

A non-homogeneous hidden Markov-model for QTL-mapping

The basic HMM proposed in [109] assumes that the identified SNPs are equally spaced across the whole genome. This assumption is not necessarily correct. In addition, the chance of co-segregation may depend on the distance between the SNPs. Hence, an extension of the HMM that accommodates the distance between SNPs is of interest. This extension was not considered in [109]. Toward this aim, the assumption that the latent Markov-chain is homogeneous, i.e., that the transition probabilities are constant, can be weakened. More specifically, the transition probabilities can be assumed to depend on the distance between SNPs. This results in a non-homogeneous hidden Markov-model (NH-HMM). NH-HMMs have been considered, e.g., in environmental studies [119–126].

In this chapter, we discuss an NH-HMM for QTL-mapping which was not evacuated in [109]. In particular, in Section 10.1, we give a general introduction to NH-HMMs. Section 10.2 shows an NH-HMM model to map QTL-loci based upon NGS data. The application of the proposed NH-HMM to the ethanol-tolerant yeast dataset is described in Section 10.3. Section 10.4 completes the chapter with conclusions and a discussion of topics for further research.

10.1 An introduction to non-homogeneous Markovmodels

In a homogeneous Markov-model, the transition probabilities are assumed to be constant. In a non-homogeneous HMM, the transition probabilities may depend on the position of the state sequence $\{C_1, C_2, ..., C_N\}$, i.e., the model is characterized by position-specific transition-probability matrices Γ_i . For instance, assume that, for each random variable X_i from the observable sequence $\{X_1, X_2, ..., X_N\}$, information is available about covariates \mathbf{Y}_i . The transition probability matrix can now depend on the covariates, i.e., $\Gamma_i = \Gamma(\mathbf{Y}_i)$. The form of the dependence of the transition probabilities on the covariates has to be specified in a way that is proper to the particular application. Several different parameterization for transition probabilities are possible [120]. A possible form of dependence of transition probabilities on covariates can be specified as follows:

$$\gamma_{jk,i} = P(C_{i+1} = s_k | C_i = s_j, y_i) = \frac{\exp(\alpha_{jk} + \beta_{jk} y_i)}{\sum_{k=1}^{m} \exp(\alpha_{jk} + \beta_{jk} y_i)},$$
(10.1)

in which the transition probabilities are associated to the covariate y_i through multinomial logit link functions. The unknown parameters of α_{jk} and β_{jk} are the coefficients of the link function that have to be estimated.

10.1.1 Forward-backward algorithm and parameter estimates

Assuming the non-homogeneous transition matrix as Γ_i , we state the forward and backward probabilities of the NH-HMM below as a product of (row)vectors:

$$\boldsymbol{\alpha}_{i} = \boldsymbol{\delta} \boldsymbol{P}(x_{1}) \boldsymbol{\Gamma}_{i} \boldsymbol{P}(x_{2}) \dots \boldsymbol{\Gamma}_{i} \boldsymbol{P}(x_{i}) \quad for \ i = 1, \dots N,$$
(10.2)

$$\boldsymbol{\beta}_{i} = \boldsymbol{\Gamma}_{i} \boldsymbol{P}(x_{i+1}) \boldsymbol{\Gamma}_{i} \boldsymbol{P}(x_{i+2}) \dots \boldsymbol{\Gamma}_{i}, \boldsymbol{P}(x_{N}) \boldsymbol{1}^{T},$$
(10.3)

where $\boldsymbol{\delta} \equiv (\delta_1, \delta_2, ..., \delta_m)$ and $\mathbf{P}_j(x)$ is the $m \times m$ diagonal matrix with the emission probability $p_j(x)$ as the *j*-th diagonal element. The non-homogeneous transition matrix $\boldsymbol{\Gamma}_i$ is described by the coefficients α_{jk} and β_{jk} . The forward and backward probabilities are used in the algorithm used to maximize the observed data likelihood (see Subsection 8.1.1).

After estimating the parameters of the NH-HMM, the sequence of the hidden states, which could have generated the observed sequence of symbols, can be predicted with the help of the Viterbi algorithm [114].

10.2 A non-homogeneous hidden Markov-Model for BSA experiments

The structure of the proposed NH-HMM is as follows. We consider three states, i.e., m = 3. The hidden states and the state-dependent emission probabilities are defined as in Subsection 8.2.1 and 8.2.2, respectively. In the proposed NH-HMM, the transition probabilities depend upon the distance between adjacent SNPs (in 10,000 base pairs). Denote the distance between the *i*-th and (i + 1)-th SNP by y_i . The transition probabilities can then be modelled as a function of y_i by the multinomial logistic link-function (10.1). To ensure estimability of the parameters α_{jk} and β_{jk} , we constrain α_{jj} and β_{jj} to be equal to zero.

10.3 Results

Figure 10.1 presents the distribution of the distance between neighboring SNPs (in 10K bp) for pool 1 of chromosome XIV and chromosome IX of the high ethanoltolerance dataset described in Section 2.3. As it can be seen from the histograms, the SNPs are not equally spaced. Therefore, we applied the proposed NH-HMM outlined in Section 10.2 to the data. We compare the results of the proposed NH-HMM with the outcome of the HMM with the homogeneous transition matrix in terms of the parameters estimates and the allocation of the SNPs to particular states. In all cases the sequencing-error corrections ($\varepsilon_{tot} = 0.05$) and filtering approach are used (see Subsection 9.3.1).

10.3.1 Chromosome XIV, pool 1

Table 10.1 shows the estimated concordance probabilities, initial probabilities, Akaike's information criterion (AIC), and Bayesian information criterion (BIC) for the H-HMM and NH-HMM.



Chromosome XIV, pool 1

Figure 10.1: The histograms of distance (10K bp) between neighboring SNPs in chromosome XIV, pool 1 (top panel) and chromosome IX, pool 1 (bottom panel).

Table 10.1: Chromosome XIV, pool 1: comparison of two forms of the HMM with three states and their corresponding parameter estimates (95% confidence intervals in brackets). For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Model | μ_1 | μ_2 | μ_3 | δ_1 | δ_2 | δ_3 | $-logL_N$ | AIC | BIC |
|--------|----------------------------|----------------|----------------|------------|------------|------------|-----------|----------|----------|
| H-HMM | 0.2268 | 0.1284 | 0.0559 | 0+ | 1+ | 0+ | 71662.85 | 143347 7 | 143400 7 |
| | [0.220, 0.230] | [0.127, 0.129] | [0.054, 0.056] | 0 | 1 | 0 | 11002.00 | 145547.7 | 143409.7 |
| NH-HMM | 0.2268 | 0.1283 | 0.0558 | 0^+ | 1+ | 0^+ | 71638.00 | 143208.2 | 143360-1 |
| | $\left[0.220, 0.230 ight]$ | [0.127, 0.129] | [0.054, 0.056] | 0 | 0.1. | | 71038.09 | 145296.2 | 140000.1 |

The comparison of the first and the second row of Table 10.1 reveals that on the basis of the information criteria, the NH-HMM is worth considering. Note that the estimates of the concordance probabilities μ_j are almost similar for both models. In particular, the total probabilities of discordance between the no-trait reference parent and the offspring can be estimated to be equal to $1-4\times0.2268=0.0928$, $1-4\times$ 0.128=0.488, and $1-4\times0.0559=0.776$ for the first, second, and third state, respectively. Thus, following the reasoning presented earlier (Subsection 8.2.1), the first state can be seen as corresponding to SNPs linked with a gene loci of the parent without the trait, the second state can be considered as corresponding to the SNPs not linked with any genes, whereas the third state can be seen as SNPs that are potentially linked to one or more genes responsible to high ethanol-tolerance. The estimated initial probabilities indicate that the most likely initial state for both fitted models is the second state ($\hat{\delta}_2 = 1$). This means that the first SNP is most likely generated by the second state for both models.

Table 10.2 shows the estimated coefficients of the logit link-function (10.1) describing the dependence of the transition probabilities on the distance between the adjacent SNPs (in 10K base pairs). The estimated transition probabilities of the Markov-chain, as a function of the distance between SNPs, are shown in Figure 10.2. The estimated coefficients indicate that, conditional on being in state 1, increasing the distance to the neighboring SNP increases the probability of transition to the second state. Similarly, for state 2, the larger the distance between the adjacent SNPs, the higher the chance of transition to state 3 as compared to state 1. For state 3, results similar to those for state 1 are obtained: increasing the distance between SNPs increases the probability of transition to the second state. In other words, for all three states, increasing the distance between the neighboring SNPs increases the chance of transition to another state. As a consequence, the transition probabilities of staying in the same state decrease as the distance between the neighboring SNPs

| Parameters | j = 1 | j = 2 | j = 3 |
|---------------|--------------------------------|---------------------------|----------------------------|
| α_{j1} | 0 | -3.591 [-4.055, -3.126] | -4.93 [-5.776, -4.084] |
| α_{j2} | -2.705 $[-4.088, -1.323]$ | 0 | -2.668 $[-2.950, -2.385]$ |
| α_{j3} | -3.926 $[-5.483, -2.369]$ | -2.866 $[-3.149, -2.583]$ | 0 |
| β_{j1} | 0 | -4.834 [-14.476, 4.806] | 3.167 [-2.939, 9.273] |
| β_{j2} | $438.909 \ [177.203,700.616]$ | 0 | $3.402 \ [1.184, \ 5.619]$ |
| β_{j3} | $437.078\ [175.235,\ 698.921]$ | 1.444 [-1.074, 3.964] | 0 |

Table 10.2: Parameter estimates (95% confidence intervals in brackets) for the non-homogeneous HMM with three states for pool 1 of chromosome XIV.

increases. For state 1 (Figure 10.2) this decline is especially large as compared to SNPs in state 2 and 3. The small number of SNPs, less densely spread out along the chromosome with SNPs frequency below 20% (Figure 10.3), is most likely responsible for this effect. On the other hand, the probability of co-segregation decreases as the distance between the adjacent SNPs increases. The presence of potential QTLs for high ethanol-tolerance in state 3, could be a possible reason for the high chance of transition from state 2 to 3 and from state 3 to 2 over large distances.

The left-hand-side panel of Figure 10.3 presents the states predicted for each SNP based on the most likely state sequence (Viterbi algorithm) obtained from the fitted NH-HMM (Table 10.2), while the right-hand-side panel presents the states predicted for each SNP obtained from the NH-HMM. Both plots clearly illustrate that state 3 (blue) is associated with a high frequency of mis-matched nucleotides, state 2 (green) with an intermediate frequency, and state 1 (red) with a low frequency. Except in few cases (10 SNPs), NH-HMM and H-HMM assign the SNPs to the same states.



Figure 10.2: Probabilities of transition from one state to another, as estimated by the three-state non-homogeneous hidden Markov-model for chromosome XIV, pool 1.



Figure 10.3: Chromosome XIV, pool 1, predicted SNP-specific states for the model with homogeneous HMM (left panel) and with non-homogeneous HMM (right panel). Red stands for state 1, green for state 2, and blue for state 3.

10.3.2 Chromosome XIV, pool 2

Table 10.3 shows the results for the H-HMM and NH-HMM for pool 2 of chromosome XIV. The corresponding total discordance probabilities are equal to 0.321, 0.609, and 0.876 for state 1, state 2, and state 3, respectively. Table 10.4 shows the estimated

Table 10.3: Chromosome XIV, pool 2: comparison of two forms of the HMM with three states and their corresponding parameter estimates (95% confidence intervals in brackets). For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Model | μ_1 | μ_2 | μ_3 | δ_1 | δ_2 | δ_3 | $-logL_N$ | AIC | BIC |
|--------|-------------------|----------------------------|----------------------------|------------|------------|------------|-----------|----------|----------|
| H-HMM | 0.1696 | 0.0979 | 0.0309 | 1+ | 0+ | 0+ | 74740.81 | 140521.6 | 140583.6 |
| | $[0.168,\!0.171]$ | $\left[0.096, 0.099 ight]$ | $\left[0.029, 0.032 ight]$ | 1 | 0 | 0 · | 14149.01 | 149521.0 | 149565.0 |
| NH-HMM | 0.1695 | 0.0979 | 0.0308 | 1+ | 0^+ | 0^+ | 74730 50 | 140501.9 | 140563-1 |
| | $[0.168,\!0.171]$ | $[0.168,\!0.171]$ | [0.168, 0.171] | 1 | 0 | 0 · | 14139.39 | 149501.2 | 149000.1 |

coefficients of the logit link-function describing the dependence of the transition probabilities on the distance between the adjacent SNPs (in 10K bp) for the NH-HMM. It can be observed that (see Figure 10.4), similarly to pool 1, the probabilities of staying in the same state decrease as the distance between the neighboring SNPs increases. However, for state 1 of pool 2, this decline is not as large as for spool 1 (see Figure 10.2). The larger number of SNPs around 20,000 bp, could be a possible reason for this behavior (Figure 10.5). If the distance between SNPs increases, the possibility of moving from state 2 to state 3 and for state 3 to 2 also increases. Except for 10 SNPs, NH-HMM and H-HMM assign the SNPs to the same states (see Figure 10.5).

Table 10.4: Parameter estimates (95% confidence intervals in brackets) for the non-homogeneous HMM with three states for pool 2 of chromosome XIV.

| Parameters | j = 1 | j = 2 | j = 3 |
|---------------|-----------------------------|-------------------------------------|-----------------------------|
| α_{j1} | 0 | -2.0530 [-2.324,-1.781] | -5.2343 [-6.163,-4.305] |
| α_{j2} | -2.0421 [$-2.323, -1.76$] | 0 | -2.8260 $[-3.130, -2.516]$ |
| α_{j3} | -4.1373 $[-4.868, -3.406]$ | -2.3521 $[-2.660, -2.043]$ | 0 |
| β_{j1} | 0 | 0.9012 [-1.529,3.331] | 8.0646 [4.057,12.071] |
| β_{j2} | $3.4682 \ [0.663, 6.273]$ | 0 | $5.646201 \ [3.095, 8.197]$ |
| β_{j3} | 4.5416 $[-0.671, 9.755]$ | $1.1887 \left[-1.414, 3.792\right]$ | 0 |



Figure 10.4: Probabilities of transition from one state to another, as estimated by the three-state non-homogeneous hidden Markov model.



Figure 10.5: Chromosome XIV pool 2, predicted SNP-specific states for the model with homogeneous HMM (left panel) and with non-homogeneous HMM (right panel). Red stands for state 1, green for state 2, and blue for state 3.

10.3.3 Chromosome IX, pool 1

Table 10.5 presented two models fitted for pool 1 of chromosome IX. Similar parameter estimates indicate that the results do not depend on the assumed non-homogeneous transition matrix. The total discordance probability can be estimated to be equal to 0.8, 0.38, and 0.532 for the first, second, and the third state, respectively. For both

Table 10.5: Chromosome IX, pool 1: comparison of two forms of the HMM with three states and their corresponding parameter estimates (95% confidence intervals in brackets). For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Model | μ_1 | μ_2 | μ_2 | δ_1 | δ_2 | δ_3 | $-logL_N$ | AIC | BIC |
|---------------|---|----------------|-------------------|------------|-------------|------------|---------------|----------|----------|
| н нмм | 0.236 | 0.155 | 0.117 | 1+ | $1^+ \ 0^+$ | - 0+ | 0^+ 64105.1 | 128232.3 | 128295.4 |
| Π-ΠΙΝΙΝΙ | $[0.231,\!0.247]$ | [0.153, 0.156] | $[0.115,\!0.118]$ | 1 | | | | | |
| NH HMM | 0.237 0.1556 0.1178 1 ⁺ 0 ⁺ | 0^+ | 0^+ | 0+ 640700 | 190101 0 | 199944.0 | | | |
| 1111-111/11/1 | [0.231, 0.247] | [0.153, 0.156] | $[0.115,\!0.118]$ | 1+ | 0 ' | 0, | 04079.9 | 120101.0 | 120244.9 |

models, the discordance probability of state 3 indicates that SNPs in this state are not linked to loci from the parent with the trait [103]. In addition, the discordance probability of state 2 shows that second state does not correspond to a state where no linkage is present. It rather suggests a state in which there are a linked chromosomal regions with gene loci from parent without the trait. To check the sensitivity of the conclusions to the structure of the model, we use a model in which μ_2 is fixed to be equal to 0.125, i.e., the total discordance probability for the second state is set to be equal to 0.50. The results are presented in Table 10.6. The estimated total discordance

Table 10.6: Chromosome IX, pool 1: comparison of two forms of the 3-state HMM with fixed mismatch probability $\mu_2 = 0.125$ and their corresponding parameter estimates (95% confidence intervals in brackets). For the parameters indicated with ⁺ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Model | μ_1 | μ_2 | μ_2 | δ_1 | δ_2 | δ_3 | $-logL_N$ | AIC | BIC |
|--------|--------------------------|-------------|---------------------------|------------|------------|------------|-----------|----------|----------|
| H-HMM | 0.1914 [0.172,0.203] | 0.125^{+} | 0.092 [0.091,0.093] | 1^{+} | 0+ | 0+ | 64654.8 | 129331.7 | 129394.8 |
| NH-HMM | 0.1905 [0.171, 0.202] | 0.125^{+} | 0.0913 [0.0903,0.0924] | 1^{+} | 0^{+} | 0^{+} | 64628.4 | 129279 | 129342.1 |

probabilities for state 1 and 3 are now equal to 0.234 and 0.632, respectively. The second state can now be considered as corresponding to the SNPs not linked with any genes. Table 10.7 shows the estimated parameters of the non-homogeneous transition matrix for each state after fixing the mismatch probability of the second state. The estimated transition probabilities are shown in Figure 10.6.

Table 10.7: Parameter estimates (95% confidence intervals in brackets) for the non-homogeneous HMM with three states for pool 1 of chromosome IX.

| Parameters | j = 1 | j = 2 | j = 3 |
|---------------|---|-----------------------------|---------------------------------|
| α_{j1} | 0 | -1.9192 [-2.082,-1.757] | -6.3019 [-7.925,-4.679] |
| α_{j2} | -0.5973 $[-0.790, -0.403]$ | 0 | -3.7624 $[-4.957, -2.567]$ |
| α_{j3} | -4.1373 [-4.868,-3.406] | -2.3521 $[-2.660, -2.043]$ | 0 |
| β_{j1} | 0 | 4.2321 [1.3436,7.1214] | 443.3612 [290.895,595.820] |
| β_{j2} | $6.50 \ [0.778, \ 12.230]$ | 0 | $433.4266\ [283.214,\!583.631]$ |
| β_{j3} | $10.468 \left[-270.120, 291.056\right]$ | $-0.6437 \ [-8.271, 6.983]$ | 0 |



Figure 10.6: Probabilities of transition from one state to another, as estimated by the three-state non-homogeneous hidden Markov model.

According to Figure 10.6, for all three states, the transition from each state to itself is decreasing if the distance between adjacent SNPs increases. In state 1, increasing the distance between the neighboring SNPs increases the chance of moving from state 1 to state 2. Surprisingly, the probability of transition from state 1 to state 3 is steady and equal to zero across the whole chromosome. According to Figure 10.7, SNPs assigned to state 1 and state 2, are located uniformly along the whole chromosome. In all states, at specific distance point between the adjacent SNPs (0.1 K bp, 0.4 K bp and 0.3 K bp for state 1, state 2 and state 3 respectively, Figure 10.6) transition probability between these two states is equal. This indicates that at that particular distance point, SNPs locating on state 2 or state 3 can either randomly stay in their own state or depart towards the other state. Increasing the distance after this point, moves these SNPs towards the other state. For the models with fixed concordance probability, a very small number of SNPs is classified to be in state 3 (see Figure10.7). Except in 23 cases, SNPs are classified similarly into states as shown by Viterbi path in Figure 10.7.



Figure 10.7: Chromosome IX pool 1, predicted SNP-specific states for the model with homogeneous HMM (left panel) and with non-homogeneous HMM (right panel). Red stands for state 1, green for state 2, and blue for state 3.

10.3.4 Chromosome IX, pool 2

The estimated parameters of the total discordance probability for pool 2 of chromosome IX (Table 10.8) is equal to 0.06, 0.4504 and 0.64 for the first, second, and the third state, respectively. The same trend of transition probabilities is observed in pool 2 of chromosome IX (Table 10.9 and Figure 10.8) as compared to pool 1. The SNPs assigned to state 3 (blue in Figure 10.9) are scattered widely across the whole chromosome and do not form any contiguous region. Therefore, we can conclude that chromosome IX does not contain QTLs responsible for high ethanol-tolerance.

Table 10.8: Chromosome IX, pool 2: comparison of two forms of the 3-state HMM and their corresponding parameter estimates (95% confidence intervals in brackets). For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Model | μ_1 | μ_2 | μ_3 | δ_1 | δ_2 | δ_3 | $-logL_N$ | AIC | BIC |
|---------------|----------------|----------------|-------------------|------------|------------|------------|------------------------|----------|----------|
| н нмм | 0.2350 | 0.1377 | 0.0941 | 1+ | 0+ | 0+ | 60/11 3 | 138844.6 | 138007.8 |
| 11-111/11/1 | [0.228, 0.238] | [0.136, 0.138] | [0.093, 0.095] | 1.0. | | 0 | 09411.3 | 130044.0 | 138901.8 |
| NH HMM | 0.2349 | 0.1374 | 0.0936 | 1+ | 0^+ | 0^+ | 60387 65 | 128707 2 | 138860 5 |
| 1111-111/11/1 | [0.228, 0.238] | [0.136, 0.138] | $[0.092,\!0.094]$ | T 0. | 0 | 09301.05 | 9307.00 130797.3 13880 | 138800.5 | |



Figure 10.8: Probabilities of transition from one state to another, as estimated by the three-state non-homogeneous hidden Markov-model.

118 Chapter 10. A non-homogeneous hidden Markov-model for QTL-mapping

Table 10.9: Parameter estimates (95% confidence intervals in brackets) for the non-homogeneous HMM with three states of pool 2 of chromosome IX.

| Parameters | j = 1 | j = 2 | j = 3 |
|---------------|--|------------------------------|-----------------------------|
| α_{j1} | 0 | -3.0526 [-3.445,-2.658] | -3.6818 [-4.299,-3.064] |
| α_{j2} | $\textbf{-0.3960} \; [\textbf{-0.777,-0.013}]$ | 0 | -1.6406 $[-1.827, -1.4563]$ |
| α_{j3} | -1.8248 [-2.481,-1.166] | -1.5328 $[-1.724, -1.338]$ | 0 |
| β_{j1} | 0 | 411.2044 [-2.754,25.164] | -19.0388 [-58.362,20.284] |
| β_{j2} | 1.1466[-3.244, 5.538] | 0 | 20.12985 [14.476, 25.777] |
| eta_{j3} | -0.3409[-9.607, 8.925] | $23.0909\ [16.633,\!29.545]$ | 0 |



Figure 10.9: Chromosome IX pool 2, predicted SNP-specific states for the model with homogeneous HMM (left panel) and with non-homogeneous HMM (right panel). Red stands for state 1, green for state 2, and blue for state 3.

10.4 Conclusions

In this chapter we have presented the NH-HMM model for QTL-mapping. The approach adopted by the NH-HMM has a number of advantages over a basic HMM. Most importantly, by taking into account the distance between adjacent SNPs, an NH-HMM better models chromosomes where some regions are densely covered and others are covered at lower density. In addition, an NH-HMM can be extended to include other covariates except the distance between the adjacent SNPs. However,

further investigation should be devoted to the choice of the relevant covariates for the NH-HMM. One of the main concern in an NH-HMM model is choosing a suitable link function together with an appropriate distance scale. For instance, in our case, we choose the standard (multinomial) logistic link-function between covariates and the transition probabilities. Other link functions, such a probit are also possible, however, we did not investigate such a link function in our model. The distance scale of 10K bp between the adjacent SNPs was used. This scale is not the only option and different scaling of distance can also be applied, e.g., $\log(distance)$.

Another important issue for using the NH-HMM is the number of parameters in the model that have to be estimated. Applying the NH-HMM can increase the number of estimated parameters related to the transition probabilities. For larger number of states, the number of these parameters increases the computation time of the forward-backward algorithm. As an example, for pool 1 segregants of chromosome XIV, fitting the model takes 3h 48m 22s and 7h 14m 09s for the basic HMM and the NH-HMM model respectively (on an HP Elitebook 8530p). To decide which stochastic model fits the data best, AIC and BIC can be computed. In our case, there was a slight difference between the values of these information criteria for the basic HMM and the NH-HMM. Moreover, there was little difference in the estimated values of the parameters of interest, i.e., the concordance probabilities. Thus, in our application, the results did not depend on the assumption of (non-)homogeneity of the transition matrix.

Chapter 11

A joint hidden Markov-model for QTL-mapping

The basic HMM proposed in [109] analyzed data for only one pool of segregant, i.e., either pool 1 or pool 2 for a particular chromosome. However, if multiple segregated pools are available, one could consider modelling them simultaneously. The possibility to incorporate multiple segregant pools was considered in [108] with the application of a scatter plot smoother. The presented results in [108] indicated that differences between the observed trends of multiple segregated pools are advantageous for identifying potential loci associated with the trait of concern. The significant difference between the multiple pools of segregant could also be informative for identifying a minor QTL present in the reference strain. Therefore, the incorporation of multiple segregated pools could be an important feature for reducing the size of the identified chromosomal regions associated with the trait.

In analogy to the scatter plot smoother, we model multiple pools of segregant at the same time with the application of a joint HMM.

This chapter is organized as follows. In Section 11.1, we give a general introduction to the joint HMM. Section 11.2 presents a joint HMM model for BSA experiment. The application of the proposed joint HMM on the ethanol-tolerant yeast dataset (Section 2.3) is described in Section 11.3. Concluding remarks and topics for future research are given in Section 11.4.

11.1 Methodology

The basic HMM described in Section 8.1 dealt with only a single observation sequence to fit the model. In the joint HMM, we assume the use of multiple observation sequences that can be observed simultaneously.

Denote the set of k observation sequences as

$$\mathbf{X} = \left\{ \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(k)} \right\},$$
(11.1)

where $\boldsymbol{X}^{(k)} = (X_1^{(k)}, X_2^{(k)}, ..., X_N^{(k)})$ is the k-th observation sequence. In addition, we assume that all observation sequences are independent of each other. Denote the parameters of the HMM model for the k-th sequence as $(\boldsymbol{\delta}^{(k)}, \boldsymbol{\Gamma}, \boldsymbol{\theta}^{(k)})$.

The joint HMM model based on an *m*-state Markov-chain can be expressed symbolically as $\lambda = (\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\theta})$, with $\boldsymbol{\delta} = (\boldsymbol{\delta}^{(1)}, ..., \boldsymbol{\delta}^{(k)}) = (\delta_1^{(1)}, ..., \delta_m^{(1)}, ..., \delta_1^{(k)}, ..., \delta_m^{(k)})$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(k)}) = (\theta_1^{(1)}, ..., \theta_m^{(1)}, ..., \theta_1^{(k)}, ..., \theta_m^{(k)})$ denoting, respectively, the initial state distribution, transition probability matrix, and the emission-probability distribution parameters of the joint HMM model. Note that the same transition probability matrix is assumed for all the sequences.

Assuming independency among observation sequences, given a particular join HMM (δ, Γ, θ) , the log-likelihood of an observed sequence $\mathbf{X} = \mathbf{x}$ can be expressed as

$$log(L_N) = log\{L_N^{(1)}\} + \dots + log\{L_N^{(k)}\},$$
(11.2)

where $L_N^{(k)}$ is defined according to (8.4).

The forward-backward probabilities assigned to each observed sequence can be computed as follows:

$$\boldsymbol{\alpha}_{i}^{(k)} = \boldsymbol{\delta}^{(k)} \mathbf{P}(x_{1}^{(k)}) \boldsymbol{\Gamma} \mathbf{P}(x_{2}^{(k)}) \dots \boldsymbol{\Gamma} \mathbf{P}(x_{i}^{(k)}) \quad for \ i = 1, \dots N,$$
(11.3)

$$\boldsymbol{\beta}_{i}^{(k)} = \boldsymbol{\Gamma} \ \mathbf{P}(x_{i+1}^{(k)}) \boldsymbol{\Gamma} \mathbf{P}(x_{i+2}^{(k)}) \dots \boldsymbol{\Gamma} \mathbf{P}(x_{N}^{(k)}) \mathbf{1}^{T},$$
(11.4)

in which the transition matrix Γ is assumed to be the same for all observed sequences.

Define $u_j^{(k)}(i)$ and $v_{jk}^{(k)}(i)$ to be indicator variables of the k-th observation sequence as explained in Section 8.1. The parameters involved in (11.2) can be estimated by using a joint forward-backward algorithm. In the E-step, $u_j^{(1)}(i)$, $v_{jk}^{(1)}(i)$,..., $u_j^{(k)}(i)$, and $v_{jk}^{(k)}(i)$ are estimated given the observed data and current estimates of the model parameters. Then, in the M-step, the log-likelihood function (11.2) is maximized with respect to $(\boldsymbol{\delta}^{(1)},...,\boldsymbol{\delta}^{(k)})$, $(\boldsymbol{\theta}^{(1)},...,\boldsymbol{\theta}^{(k)})$, and $\boldsymbol{\Gamma}$, given the estimates of $u_j^{(1)}(i)$, $v_{jk}^{(1)}(i)$, and $v_{jk}^{(k)}(i)$. After estimating the parameters of the joint HMM, the sequence of the hidden states, which could have generated the observed sequence of symbols, can be predicted for each observed sequence separately. For instance, the most likely sequence ("global decoding") can be found with the help of the Viterbi algorithm [114].

11.2 A joint hidden Markov-Model for BSA experiments

The structure of the proposed joint HMM is assumed as follows. We consider three states, i.e., m = 3. The hidden states and the state-dependent emission probabilities are defined as in Subsection 8.2.1 and 8.2.2, respectively. Pool 1 and pool 2 of each chromosome are assumed to be two independent observation sequences. The common set of SNPs along the chromosome have been selected.

11.3 Results

We apply the joint HMM, outlined in Section 11.2, to chromosome XIV, chromosome IX, and chromosome II. These chromosomes are selected for validation of the method due to three possible scenarios, i.e., the presence of a major QTL (chromosome XIV), the absence of a QTL (chromosome IX), and the presence of a minor QTL (chromosome II). In all cases the sequencing-error correction ($\varepsilon_{tot} = 0.05$) and filtering approach are used (see Subsection 9.3.1).

11.3.1 Chromosome XIV

Table 11.1 presents the parameter estimates of the joint HMM. The total discordance probabilities of chromosome XIV are equal to $(1,1,1) - 4(\mu_1^{(1)},\mu_2^{(1)},\mu_3^{(1)}) =$ (0.1928, 0.5044, 0.7872), and $(1,1,1) - 4(\mu_1^{(2)},\mu_2^{(2)},\mu_3^{(2)}) = (0.2132, 0.502, 0.8492)$, for pool 1 and pool 2, respectively. Thus, following the argumentation presented in Subsection 8.2.1, for both pools, the first state can be seen as corresponding to SNPs linked with a gene loci of the parent without the trait, the second state can be treated as corresponding to the SNPs not linked with any gene(s) responsible for ethanol tolerance, while the third state can be seen as identifying SNPs linked to the gene(s).

The estimated transition probabilities indicate that the most likely transitions are from a given state to itself, as can be seen from the estimated values: $\gamma_{11} = 0.6882$, $\gamma_{22} = 0.8857$, and $\gamma_{33} = 0.9285$.

Figure 11.1 presents the states predicted for each SNP based on the most likely state sequence while considering the joint HMM model. The plot clearly illustrates that for both pools, state 3 (blue) is associated with a high SNP frequency, state 2 (green) with an intermediate frequency, and state 1 (red) with a low frequency.

We test whether there is a significant difference between the state-dependent emission-probabilities parameters of each pool. The significant difference could represent gene loci, therefore, might decrease the size of the identified region. For this

Table 11.1: Parameter estimation of the three-state model for the joint pool segregants of chromosome XIV with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space.

| Parameters | |
|------------------|-------------------------------|
| $\mu_1^{(1)}$ | $0.2018 \ [0.199, \ 0.205]$ |
| $\mu_2^{(1)}$ | $0.1239\ [0.123,\ 0.124]$ |
| $\mu_3^{(1)}$ | $0.0532 \ [0.053, \ 0.055]$ |
| $\mu_{1}^{(2)}$ | $0.1967 \ [0.194, \ 0.199]$ |
| $\mu_{2}^{(2)}$ | $0.1245 \ [0.124, \ 0.126]$ |
| $\mu_{3}^{(2)}$ | 0.0377[0.037, 0.039] |
| $\delta_1^{(1)}$ | 1+ |
| $\delta_2^{(1)}$ | 0^{+} |
| $\delta_3^{(1)}$ | 0^{+} |
| $\delta_1^{(2)}$ | 0^{+} |
| $\delta_2^{(2)}$ | 1^{+} |
| $\delta_3^{(2)}$ | 0^{+} |
| γ_{11} | $0.6882 \ [0.6623, \ 0.7132]$ |
| γ_{12} | $0.2769\ [0.2678,\ 0.2862]$ |
| γ_{13} | $0.0347\ [0.0333,\ 0.0361]$ |
| γ_{21} | $0.0526\ [0.0511,\ 0.0540]$ |
| γ_{22} | $0.8857\ [0.8634,\ 0.9060]$ |
| γ_{23} | $0.0615 \ [0.0612, \ 0.0617]$ |
| γ_{31} | $0.0062 \ [0.0030, \ 0.0118]$ |
| γ_{32} | $0.0652 \ [0.0637, \ 0.0666]$ |
| γ_{33} | $0.9285 \ [0.9277, \ 0.9292]$ |

purpose, the emission-probability distribution parameters of two pools are compared are compared with each other through the reparametrization $\theta_j = \log(\frac{\mu_j}{3\mu'_j})$ (equation (8.16)). The comparision of two pools is based on the value obtained for each of η_j :

$$\theta_j^{(2)} = \theta_j^{(1)} + \eta_j. \tag{11.5}$$

Particularly, we want to test the following hypothesis for each three states:

$$\begin{cases} H_0: \eta_j = 0\\ H_1: \eta_j \neq 0 \end{cases}$$
(11.6)

The Wald test $\frac{\eta_j - 0}{SE(\eta_j)} \sim N(0, 1)$ is used and the corresponding statistic and their *p*-values are presented in Table 11.2. According to the results, there is a significant difference between emission-probability distribution parameters in state 1 and state 3. The negative value of η_j for these two states indicates that the concordance probability for pool 1 is higher than the concordance probability for pool 2 for both states. The higher concordance probability for pool 1 in the first state can be referred to an additional effect around 200,000 *bp* for pool 2 (see Figure 11.1), where the SNP frequency drops to approximately 20%. This suggests the presence of a minor QTL in the reference strain, which was not present in the strain of the parent with a high ethanol-tolerance [108]. The higher concordance probability for the third state could be related to an enrichment effect in the area around the three QTLs for pool 2 [103], where the SNP frequency increases to 90%.

Table 11.2: Parameter estimation and the Wald test statistic (Z) of the difference in relative risk ratios for chromosome XIV. The values between the brackets are the corresponding 95% confidence intervals.

| Parameter | MLE | Z statistic | p-value |
|-----------|---|-------------|-------------|
| η_1 | -0.2286 $[-0.319, -0.137]$ | - 4.241 | < 0.0001 |
| η_2 | $0.0173 \ [-0.011, \ 0.045]$ | 1.198 | 0.2348 |
| η_3 | $\textbf{-}0.3560\;[\textbf{-}0.393,\textbf{-}0.318]$ | -18.651 | $<\!0.0001$ |



Figure 11.1: Chromosome XIV, predicted SNP-specific states for the joint HMM model. The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3. The vertical lines indicate the location of the three identified genes, i.e., MKT1, SWS2 and APJ1.

11.3.2 Chromosome IX

Table 11.3 presents the parameter estimations of the joint HMM for chromosome IX. The total discordance probabilities are equal to $(1, 1, 1) - 4(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}) = (0.2192, 0.474, 0.7436)$, and $(1, 1, 1) - 4(\mu_1^{(2)}, \mu_2^{(2)}, \mu_3^{(2)}) = (0.1608, 0.5352, 0.752)$ for pool 1 and pool 2, respectively. Thus, the first state can be seen as corresponding to SNPs linked with a gene loci of the parent without the trait, the second state can be treated as corresponding to the SNPs not linked with any gene(s) responsible for ethanol tolerance, while the third state can be seen as identifying SNPs linked to the gene(s). Figure 11.2 presents the states predicted for each SNP based on the most likely state sequence.

According to the values obtained for the η_j (Table 11.4), we can see that there is a significant deference between the emission-probabilities parameter for state 1 and state 2. The positive value of $\eta_1 = 0.2124$ indicates that the discordance probability in pool 1 is higher than the discordance probability in pool 2 for state 1 (Figure 11.2). The negative value of η_2 represents that concordance probability in pool 1 is higher than the concordance probability in pool 2 for state 2. The significant

Table 11.3: Parameter estimation of the three-state model for the joint pool segregants of chromosome IX with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals.

| Parameters | | | |
|------------------|-------------------------------|--|--|
| $\mu_1^{(1)}$ | $0.1952 \ [0.193, \ 0.197]$ | | |
| $\mu_2^{(1)}$ | $0.1315\ [0.130,\ 0.132]$ | | |
| $\mu_3^{(1)}$ | $0.0641\ [0.056,\ 0.071]$ | | |
| $\mu_1^{(2)}$ | $0.2098 \ [0.206, \ 0.212]$ | | |
| $\mu_{2}^{(2)}$ | $0.1162 \ [0.115, \ 0.117]$ | | |
| $\mu_3^{(2)}$ | $0.0620\ [0.059,\ 0.065]$ | | |
| $\delta_1^{(1)}$ | 1^{+} | | |
| $\delta_2^{(1)}$ | 0^{+} | | |
| $\delta_3^{(1)}$ | 0^{+} | | |
| $\delta_1^{(2)}$ | 1^{+} | | |
| $\delta_2^{(2)}$ | 0^{+} | | |
| $\delta_3^{(2)}$ | 0^{+} | | |
| γ_{11} | $0.6365\ [0.4963,\ 0.7298[$ | | |
| γ_{12} | $0.3595\ [0.2570,\ 0.4423]$ | | |
| γ_{13} | $0.0039 \; [0.0018, 0.0082]$ | | |
| γ_{21} | $0.0542 \ [0.0525, \ 0.0558]$ | | |
| γ_{22} | $0.9244\ [0.9021,\ 0.9467]$ | | |
| γ_{23} | $0.0214\ [0.0200,\ 0.0227]$ | | |
| γ_{31} | $0.0213\ [0.0138,\ 0.0261]$ | | |
| γ_{32} | $0.4637\]0.4386,\ 0.4896]$ | | |
| γ_{33} | $0.5148 \ [0.5104, \ 0.5192]$ | | |

difference between two pools in state 1 might suggest minor gene loci present in this chromosome. However, no potential QTLs were identified for chromosome IX [103, 108]. According to η_3 , there is not a significant difference between the emission parameters in state 3 between pool 1 and pool 2. The total discordance probabilities in state 3 for both pools indicate that this state can be seen as identifying SNPs linked to gene(s). Therefore, there is a problem with the model, as it suggests linkage, while no gene loci were found in biological analysis.

Table 11.4: Parameter estimation and the Wald test statistic (Z) of the difference in relative risk ratios for chromosome IX. The values between the brackets are the corresponding 95% confidence intervals.

| Parameter | MLE | Z statistic | p-value |
|-----------|-----------------------------|-------------|-------------|
| η_1 | $0.2124 \ [0.126, \ 0.298]$ | 4.863 | < 0.0001 |
| η_2 | -0.2298 $[-0.250, -0.209]$ | -22.034 | $<\!0.0001$ |
| η_3 | $0.0487 \ [-0.113, 0.211]$ | 0.588 | 0.5563 |



Figure 11.2: Chromosome IX, predicted SNP-specific states for the joint HMM model. The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3.

Chromosome IX

11.3.3 Chromosome II

The parameter estimates of the joint HMM for chromosome II is presented in Table 11.5. The total discordance probabilities are equal to $(1, 1, 1) - 4(\mu_1^{(1)}, \mu_2^{(1)}, \mu_3^{(1)}) = (0.2296, 0.5104, 0.7908)$, and $(1, 1, 1) - 4(\mu_1^{(2)}, \mu_2^{(2)}, \mu_3^{(2)}) = (0.2204, 0.546, 0.8128)$ for pool 1 and pool 2, respectively.

Therefore, for both pools, the first state can be seen as corresponding to SNPs linked with a gene loci of the parent without the trait, the second state can be treated as corresponding to the SNPs not linked with any gene(s) responsible for ethanol

Table 11.5: Parameter estimation of the three-state model for the joint pool segregants of chromosome II with sequencing-error correction. The values between the brackets are the corresponding 95% confidence intervals.

| Parameters | |
|------------------|-------------------------------|
| $\mu_1^{(1)}$ | $0.1926 \ [0.189, \ 0.195]$ |
| $\mu_2^{(1)}$ | $0.1224 \ [0.122, \ 0.124]$ |
| $\mu_3^{(1)}$ | $0.0523\ [0.051,\ 0.054]$ |
| $\mu_1^{(2)}$ | $0.1949\ [0.189,\ 0.199]$ |
| $\mu_2^{(2)}$ | $0.1135 \ [0.112, \ 0.114]$ |
| $\mu_{3}^{(2)}$ | $0.0468\ [0.046,\ 0.047]$ |
| $\delta_1^{(1)}$ | 0^{+} |
| $\delta_2^{(1)}$ | 1^{+} |
| $\delta_3^{(1)}$ | 0^{+} |
| $\delta_1^{(2)}$ | 0^{+} |
| $\delta_2^{(2)}$ | 0^{+} |
| $\delta_3^{(2)}$ | 1^{+} |
| γ_{11} | $0.5675 \ [0.4650, \ 0.6503]$ |
| γ_{12} | $0.3890\ [0.3798,\ 0.3983]$ |
| γ_{13} | $0.0433\ [0.0418,\ 0.0447]$ |
| γ_{21} | $0.0276\ [0.0260,\ 0.0291]$ |
| γ_{22} | $0.8879 \ [0.8656, \ 0.9082]$ |
| γ_{23} | $0.0843 \; [0.0828, 0.0857]$ |
| γ_{31} | $0.0077 \ [0.0042, \ 0.0148]$ |
| γ_{32} | $0.1819 \ [0.1615, \ 0.1798]$ |
| γ_{33} | $0.8102 \ [0.8046, \ 0.8157]$ |

tolerance, while the third state can be seen as identifying SNPs linked to the gene(s). According to the values obtained for η_j (Table 11.6), there is a significant difference between the emission-probabilities parameters for state 2 and state 3. The negative values of η_2 and η_3 represent that the concordance probability in pool 1 is higher than the concordance probabilities for state 2 indicate that this state can be considered as corresponding to the SNPs not linked with any gene(s). The significant difference between the emission-probabilities parameters in state 3 suggests the presents of a potential QTLs in this regions. The presence of one gene around 470.000 bp, i.e., LYS2, was confirmed in this chromosome [103].

Table 11.6: Parameter estimation and the Wald test statistic (Z) of the difference in relative risk ratios for chromosome II. The values between the brackets are the corresponding 95% confidence intervals.

| Parameter | MLE | Z statistic | P.value |
|-----------|-------------------------------------|-------------|----------|
| η_1 | $0.0561 \ [-0.037, \ 0.149]$ | 1.172 | 0.241197 |
| η_2 | -0.1431 [-0.163, -0.123] | -13.956 | < 0.0001 |
| η_3 | $-0.1396 \left[-0.113, 0.211 ight]$ | -6.521 | < 0.0001 |


Figure 11.3: Chromosome II, predicted SNP-specific states for the joint HMM model. The colors indicate the state of the SNP. Red for state 1, green for state 2, and blue for state 3. The vertical blue line indicates the location of the identified gene, i.e., LY S2.

11.4 Conclusions

In this chapter we present the application of a joint hidden Markov model in QTL mapping. The significant difference between the state-dependent emission-probabilities parameters can lead us to find a chromosomal location for the phenotype of interest. In case of the presence of minor QTLs, the significant difference between pool 1 and pool 2 could be informative as the minor QTLs are more distinctive in pool 2 (17%) as compared to pool 1 (16%) [103]. In chromosome XIV the estimated emission-probabilities assigned to state 1 and state 3 are significantly different from each other. This shows that the first and the third state could be the possible regions for the potential QTL and results were compatible with previous findings ([103, 108]). In chromosome IX, the application of the joint HMM led us to spurious regions as no gene loci have been found in this chromosome. The joint HMM could identified a

minor QTL presents in the pool-2 data of chromosome II. The initial estimates of the joint forward-backward algorithm for the state-dependent emission probabilities are chosen in a way that the first state corresponds to linkage with a locus of the parent without the trait, the third state presents the linkage with a locus of the parent with the trait, and the second state does not show any linkage for both pools. Therefore, for each pool the total discordance probabilities of 0.2, 0.5, and 0.8 are considered for the first, second and the third state, respectively. In addition, the initial estimates of the transition matrix for all states are selected as the probabilities from a given state to itself is 0.8 and to another is 0.1. The application of the joint hidden Markov model profoundly increased the computational time. As an example, for pool 1 segregants of chromosome XIV, fitting the model takes 3h 48m 22s and 13h 04m 42s for the basic HMM and the joint HMM model respectively (on an HP Elitebook 8530p). In general, a joint HMM can be used to detect potential regions responsible for ethanol tolerance. However, this method led us to a wider region, where no QTLs have been identified in a chromosome. Increasing the number of pools, e.g., three pools, could be result in more precise location of potential QTLs. This work can be subjected to future research.

Chapter 12

Concluding remarks and future work

12.1 Concluding remarks

In this dissertation, we proposed statistical methods for datasets from proteomics and genomics workflow.

Over the past decade, MS-based proteomics has emerged as a high-throughput method for the identification and quantification of proteins in complex samples. The high resolution MS data contains a large degree of noisy, redundant, and irrelevant information. Only a part of it includes the biologically meaningful signal, i.e., peptides and small proteins, making accurate classification between peptide/protein peaks and peaks generated by noise difficult. To overcome this obstacle, prior information related to the physical properties of the peptide/protein, i.e., isotopic distribution, is needed. However, a similarity measure is also required to distinguish between peptide and noise peaks clusters. In Chapter 4, we considered the use of Pearson's χ^2 statistic and the Mahalanobis distance for this purpose. We evaluated the performance of the two similarity measures by using a designed MALDI-TOF experiment. The results could extend to any high-resolution mass spectrum and indicated that Pearson's χ^2 statistic offered a better discriminative power for detecting the putative-peptide clusters than the Mahalanobis distance.

Protein identification is a key and essential step in the field of proteomics. For this purpose, shotgun proteomics is recognized as one of the main techniques for protein identification and quantification. In a standard computational pipeline, MS/MS spec-

tra from a mass spectrometer are searched against database search engines or de novo sequencing approaches. In database search algorithms, fragment ions derived from the unidentified protein are compared with theoretical data, and a score is assigned according to how well the two sets of data match together. The top score is expected to identify the unknown protein. The limiting factor in all database search tools is the tradeoff between false positives and false negatives. It is definitely essential to keep false positives to a minimum during protein identification. Principally, peptide identification based on tandem MS and database-search algorithms does not take into account information about isotope distributions of the precursor ions. To determine the effectiveness of these search algorithms in terms of their ability to distinguish between correct and incorrect peptide assignments, in Chapter 5, we proposed an additional metric that quantifies the similarity between the theoretical isotopic distributions for the precursor ions selected for tandem MS and the experimental mass spectra by using Pearson's χ^2 statistic. The observed association between Pearson's χ^2 statistic and the score function indicated that good scores can be obtained for molecules which exhibit atypical isotope profiles, while low scores can be obtained for fragment spectra which have a clear peptide-like isotope pattern. These results demonstrated that Pearson's χ^2 statistic can be used in conjunction with the score of database search algorithms to increase the sensitivity and specificity of peptide identification.

There are many search engines available for the analysis of proteomics data produced by MS/MS. These search algorithms vary in accuracy, sensitivity, and specificity due to the different principles in the underlying scoring mechanism. However, measuring the degree of agreement between different search engines in terms of peptide identification is always in our interest. For instance, how possible is the peptide identification obtained from SEQUEST can also be observed in MASCOT. In Chapter 6, we proposed Cohen's kappa coefficient (chance-corrected agreement) to determine the level of the agreement, between the MASCOT and SEQUEST. The results suggested that there is, in general, a good agreement between the peptide assignments for the two search engines.

The advent of high throughput sequencing methods, such as NGS has greatly accelerated biological and medical research and discovery. NGS has provided an effective approach to identify the large scale of DNA polymorphic loci used as molecular markers to distinguish gene loci responsible for the trait of concern. In Chapter 9, 10, and 11, we introduced different variants and generalizations of the basic HMM proposed in [109] used to map various QTLs responsible for high ethanol-tolerance in *S. cerevisiae* with NGS. One possible extension that can be dealt with the Marko-

vian model in the basic HMM is the direction of modelling. Both the preceding state of the (i - 1)-th SNP and following state of the (i + 1)-th SNP carry advantageous information about a current *i*-th SNP. Uni-directional HMMs ignore this influence, hence the motivation of applying the DHMM in Chapter 10. The comparison of the uni-directional HMM and the DHMM for chromosome XIV revealed only a slight difference in terms of the parameter estimates, with a minimal gain in precision of the estimation for the DHMM. As a result, the DHMM and the uni-directional HMMs assigned the SNPs to the same states. The main advantage of the DHMM is the fact that it produces a single set of estimates of the parameters of interest, i.e., emission (concordance) probabilities.

In chapter 10, we proposed the non-homogeneous HMM. The advantage of the NH-HMM is that it allows the transition probabilities of the basic HMM to vary in distance by exploiting covariate information. Our model assumed that taking into account the distance between the neighboring SNP can influence the state assignment to each SNP. The NH-HMM were able to detect gene loci responsible for high ethanol-tolerance in *S. cerevisiae*.

In Chapter 11, we considered joint HMM of two pools of segregants at the same time. The motivation was, the significant differences between the state-dependent probabilities between two pools might lead us to the potential regions of gene loci. Joint HMM was able to detect potential genomic regions for high ethanol-tolerance in chromosome XIV. However, the same approach was not able to work properly in chromosome IX.

12.2 Topics for future work

12.2.1 Assessing the agreement between peptide assignments for different search engines

The scores obtained for SEQUEST and MASCOT to assess the agreement in terms of the sequence annotations were not filtered with respect to a FDR threshold. However, we can compare the two search engines on the set of confident peptide spectrum identifications, i.e., we can filter out peptide identification results that do not comply with a FDR of 5% according to the target decoy approach. Applying the FDR threshold, could decrease the number of non-confident peptide identifications and consequently increase the peptide sequences that are both found by SEQUEST and MASCOT. However, such an extension would be a subject of further research.

12.2.2 The hidden Markov-model

The hidden Markov-models, proposed in part II of this dissertation, can be extended in several ways. The emission-state probabilities stated in (8.18), can be simplified towards a binomial distribution. This simplification will reduce the notational burden and provide an easier interpretation in terms of the "total" concordance and discordance probabilities. The first-order DHMM can be replaced by a higher-order chain. For example, a second-order Markov-chain can be characterized by the transition probabilities for each direction as follows:

$$\gamma_{jky}^{(LtoR)} = P(C_i = s_y | C_{i-1} = s_k, C_{i-2} = s_j)$$
(12.1)

$$\gamma_{jky}^{(RtoL)} = P(C_i = s_y | C_{i+1} = s_k, C_{i+2} = s_j)$$
(12.2)

The first-order Markov-chain allow to analyze for serial dependence between successive observations. Increasing the order of the Markov-chain could allow for the serial dependance beyond the recent neighboring SNPs.

In the NH-HMM model, one could include other covariates in the model. For example, besides the inclusion of the distance between the SNPs, the number of recombination events can also be included as a relevant covariate as the recombination rate is not uniform along the length of chromosomes.

Bibliography

- V.C. Wasinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K. Williams, and I. Humphery-Smith. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis*, 16(7):1090–1094, 1995.
- [2] http://fields.scripps.edu/.
- [3] K. Schneeberger, S. Ossowski, C. Lanz, T. Juul, A.H. Petersen, K.L. Nielsen, J.E. Jørgensen, D. Weigel, and S.U. Andersen. Shoremap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, 6:550– 551, 2009.
- [4] M.R. Wilkins, J.C. Sanchez, A.A. Gooley, R.D. Appel, I. Humphrey-Smith, D.F. Hochstrasser, and K.L. Williams. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology* and Genetic Engineering Reviews, 13:19–50, 1996.
- [5] C.S. Lane. Massspectrometry-based proteomics in the life sciences. *Cellular and Molecular Life Sciences*, 62:848–869, 2005.
- [6] http://sphweb.bumc.bu.edu/.
- [7] S. Hanash. Disease proteomics. Nature, 422:226-232, 2003.
- [8] E. Phizicky, P.I. Bastiaens, H. Zhu, M. Snyder, and S. Fields. Protein analysis on a proteomic scale. *Nature*, 422:208–215, 2003.
- [9] A. Sali, R. Glaeser, T. Earnest, and W. Baumeister. From words to literature in structural proteomicse. *Nature*, 422:216–225, 2003.

- [10] M. Tyers and M. Mann. From genomics to proteomics. Nature, 422:193–197, 2003.
- [11] H. Zhu, M. Bilgin, and M. Snyder. Proteomics. Annual Review of Biochemistry, 72:783–812, 2003.
- [12] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo. Protein and polymer analyses up to m/z 100,000 by laser ionization time-offlight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2:151– 153, 1988.
- [13] D. Valkenborg. Ph.d. dissertation: Statistical methods for the analysis of highresolution mass spectrometry data. *I-BIOSTAT*, Hasselt University, Belgium, 2008.
- [14] M. Mann and R. Aebersold. Mass spectrometry-based proteomics. Nature, 422:198–207, 2003.
- [15] http://alevelnotes.com/mass-spectrometry/124?tree=.
- [16] J.R. Yates, C.I. Ruse, and A. Nakorchevsky. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual Review of Biomedical Engineering*, 11:49–79, 2009.
- [17] A.I. Nesvizhskii. Protein identification by tandem mass spectrometry and sequence database searching. *Methods in Molecular Biology*, 367:87–119, 2007.
- [18] J.A. Taylor and R.S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11:1067–1075, 1997.
- [19] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. *De novo* peptide sequencing via tandem mass spectrometr. *Journal of Computational Biology*, 6:327–342, 1999.
- [20] A. Frank and P. Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. Analytical Chemistry, 77:964–973, 2005.
- [21] A. Frank, M.M. Savitski, M. L Nielsen, R.A. Zubarev, and P.A. Pevzner. De novo peptide sequencing and identification with precision mass spectrometry. Journal of Proteome Research, 6:114–123, 2007.

- [22] B. Ma, K.Z. Zhang, and C.Z. Liang. An effective algorithm for peptide de novo sequencing from MS/MS spectra. Journal of Computer and System Sciences, 70:418–430, 2005.
- [23] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. Peaks: powerful software for peptide *De novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17:2337–2342, 2003.
- [24] J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of The American Society for Mass Spectrometry*, 5:976–989, 1994.
- [25] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, and J.C. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.
- [26] R. Craig and R.C. Beavis. Matching proteins with tandem mass spectra. *Bioin-formatics*, 20:1466–1467, 2004.
- [27] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, Shi. W., and S.H. Bryant. Open mass spectrometry search algorithm. *Journal of Proteome Research*, 3:958–964, 2004.
- [28] N. Zhang, R. Aebersold, and B. Schwikowski. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2:406–1412, 2002.
- [29] R.G. Sadygov, D. Cociorva, and J.R. Yates. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods*, 1:195–202, 2004.
- [30] R.G. Sadygov, H. Liu, and J.R. Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Analytical Chemistry*, 76:1664–1671, 2004.
- [31] D.J. Pappin, P. Hojrup, and A.J. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3:327–332, 1993.
- [32] A.I. Nesvizhskii and R. Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today*, 9:173–181, 2004.

- [33] A. Keller, A.I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74:5383–5392, 2002.
- [34] S.D. Patterson. Data analysis: the Achilles heel of proteomics. Nature Biotechnology, 21:221–222, 2003.
- [35] J.D.A. Storey. A direct approach to false discovery rates. Journal of the Royal Statistical Society, 64:479–498, 2002.
- [36] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, and M.J. MacCoss. Semisupervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007.
- [37] A. Ben-Hur, C.S. Ong, S. Sonnenburg, B. Schlkopf, and G. Rtsch. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10):e1000173, 2008.
- [38] A.L. Rockwood. Relationship of fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry*, 9:103–105, 1995.
- [39] J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact centermasses. *Journal of The American Society for Mass Spectrometry*, 23(4):753–763, 2012.
- [40] J. Claesen, P. Dittwald, T. Burzykowski, and D. Valkenborg. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical Chemistry*, 85(4):1991–1994, 2013.
- [41] D. Valkenborg, I. Mertens, F. Lemière, F. Witters, and T. Burzykowski. The isotopic distribution conundrum. *Mass Spectrometry Reviews*, 2011:96–109, 2011.
- [42] E.J. Breen, F.G. Hopwood, K.L. Williams, and M.R. Wilkins. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21:2243–2251, 2000.
- [43] S. Gay, P.A. Binz, D.F. Hochstrasser, and R.D. Appel. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, 20:3527–3534, 1999.

- [44] M. Senko, S. Beu, and F. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry*, 6:229–233, 1995.
- [45] D. Valkenborg, I. Jansen, and T. Burzykowski. A model-based method for the prediction of the isotopic distribution of peptides. *Journal of The American Society for Mass Spectrometry.*, 19:703–712, 2008.
- [46] B.Y. Renard, M. Kirchner, H. Steen, J.A.J. Steen, and F.A. Hamprech. Nitpick: Peak identification for mass spectrometry data. *BMC Bioinformatics*, 9:355– 370, 2008.
- [47] S. Nicolardi, M. Palmblad, H. Dalebout, M. Bladergroen, R.A. Tollenaar, A.M. Deelder, and Y.E. Van der Burgt. Quality control based on isotopic distributions for high-throughput MALDI-TOF and MALDI-FTICR serum peptide profiling. *Journal of the American Society for Mass Spectrometry*, 9:1515–1525, 2010.
- [48] M.W. Senko, S.C. Beu, and F.W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry*, 6:229–233, 2005.
- [49] E.J. Hsieh, M.R. Hoopmann, B. Maclean, and M.J. MacCoss. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *journal of proteome research*, 9:1138–1143, 2010.
- [50] M. Palmblad, J. Buijs, and P. Hakanson. Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 12:1153–1162, 2001.
- [51] Y. Wang and M. Gu. The concept of spectral accuracy for MS. Analytical Chemistry, 82:7055–7062, 2010.
- [52] F.A. De Jong and C. Beecher. Addressing the current bottlenecks of metabolomics: Isotopic ratio outlier analysis TM, an isotopic-labeling technique for accurate biochemical profiling. *Bioanalysis*, 4:2303–2314, 2012.
- [53] R.A. Scheltema, S. Decuypere, J.C. Dujardin, D.G. Watson, R.C. Jansen, and R. Breitling. Simple data-reduction method for high-resolution LC-MS data in metabolomics. *Bioanalysis*, 1:1551–1557, 2009.

- [54] D. Valkenborg, P. Assam, G. Thomas, L. Krols, K. Kas, and T. Burzykowski. Using a poisson approximation to predict the isotopic distribution of sulphurcontaining peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry*, 21:3387–3391, 2007.
- [55] D. Valkenborg, G. Thomas, L. Krols, K. Kas, and T. Burzykowski. A strategy to analyse data from high performance liquid chromatography combined with high resolution mass spectrometry. *Journal of Mass Spectrometry*, 44:516–529, 2009.
- [56] M.W. Senko, S.C. Beu, and F.W. McLafferty. Automated assignment of charge states from resolved isotopic peaks for multiply-charged ions. *Journal of The American Society for Mass Spectrometry*, 6:52–56, 1995.
- [57] P.C. Mahalanobis. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India. Calcutta, 2:49–55, 1936.
- [58] M.M. Matzke, K.M. Waters, T.O. Metz, J.M. Jacobs, A.C. Sims, R.S. Baric, J.G. Pounds, and B.J. Webb-Robertson. Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics*, 27:2866–2872, 2011.
- [59] O. Schulz-Trieglaff, E. Machtejevas, K. Reinert, H. Schlter, J. Thiemann, and K. Unger. Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. *BioDataMining*, 2:Article 4, 2009.
- [60] D.A. Cairns, D.N. Perkins, A.J. Stanley, D. Thompson, J.H. Barrett, P.J. Selby, and R.E. Banks. Integrated multi-level quality control for proteomic profiling studies using mass spectrometry. *BMC Bioinforma*, 9:519, 2008.
- [61] Q. Liu, A.H. Sung, M. Qiao, Z. Chen, J.Y. Yang, M.Q. Yang, X. Huang, and Y. Deng. Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics*, 10(Suppl 1):S3, 2009.
- [62] P. Picotti, R. Aebersold, and B. Domon. The implications of proteolytic background for shotgun proteomics. *Molecular and Cellular Proteomics*, 6(9):1589– 1598, 2007.
- [63] J.P. Royston. Some techniques for assessing multivariate normality based on the shapiro wilk w. Applied Statistics, 32(2), 1983.

- [64] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. Communications in Statistics-Theory and Methods, 19(10):3595– 3617, 1992.
- [65] Z. Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. Analytical Chemistry, 76:3908–3922, 2004.
- [66] S. Degroeve and L. Martens. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–31203, 2013.
- [67] F.Z. Ghavidel, J. Claesen, T. Burzykowski, and D. Valkenborg. Comparison of the mahalanobis distance and pearson's χ^2 statistic as measures of similarity of isotope patterns. Journal of the American Society for Mass Spectrometry, 25:293–296, 2014.
- [68] M.V. Gorshkov, D.M. Good, Y. Lyutvinskiy, H. Yang, and R.A. Zubarev. Calibration function for the orbitrap FTMS accounting for the space charge effect. *Journal of the American Society for Mass Spectrometry*, 21(11):1846–1851, 2012.
- [69] M.L. Nielsen, M.M. Savitski, and R.A. Zubarev. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Journal of Proteome Research*, 4(6):835–845, 2005.
- [70] B. Paizs and S. Suhai. Fragmentation pathways of protonated peptides. Mass Spectrometry Reviews, 24:508–548, 2004.
- [71] R. G. Sadygov, D. Cociorva, and J. R. Yates. Largescale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods*, 1:195–202, 2004.
- [72] M. J MacCoss, C.C. Wu, and J. R. Yates. Probability-based validation of protein identifications using a modified sequest algorithm. *Analytical Chemistry*, 74:5593–5599, 2002.
- [73] D. C. Chamrad, G. Körting, K. Stühler, H. E. Meyer, and J. Klose. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics*, 4:619–628, 2004.
- [74] E. A. Kapp, F. Schutz, L. M. Connolly, J. A. Chakel, and J. E. Meza. An evaluation, comparison, and accurate benchmarking of several publicly available

MS/MS search algorithms: sensitivity and specificity of analysis. *Proteomics*, 5:3475–3490, 2005.

- [75] K. Resing, K. Meyer-Arendt, A.M. Mendoza, L.D. Aveline-wolf, and K.R. Jonscher. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical Chemistry*, 76:3556–3568, 2004.
- [76] D. Shteynberg, A.I. Nesvizhskii, R.L. Moritz, and E. W. Deutsch. Combining results of multiple search-engines in proteomics. *Molecular and Cellular Proteomics*, 12(9):2383–2393, 2013.
- [77] T. Kwon, H. Choi, C. Vogel, A.I. Nesvizhskii, and E.M. Marcotte. MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. *Journal of Proteome Research*, 10(7):2949–2958, 2011.
- [78] B.C. Searle, M. Turner, and A.I. Nesvizhskii. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research*, 7(1):245–253, 2008.
- [79] W. Yu, J.A. Taylor, M.T. Davis, L.E. Bonilla, K.A. Lee, P.L. Auger, C.C. Farnsworth, A.A. Welcher, and S.D. Patterson. Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. *Proteomics*, 10(6):1172–1189, 2011.
- [80] G. Alves, W.W. Wu, G. Wang, R.F. Shen, and Y.K. Yu. Enhancing peptide identification confidence by combining search methods. *Journal of Proteome Research*, 7:3102–3113, 2008.
- [81] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [82] R.A. Berk. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency*, 83:460–472, 1979.
- [83] D.M. Baer. Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis*, 10:117–119, 1977.
- [84] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement., 20:37–46, 1960.
- [85] http://www-cs-faculty.stanford.edu/ uno/abcde.html.

- [86] W. Bateson, E.R. Saunders, and R.C. Punnett. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society*, *Report II*, page 499, 1905.
- [87] T.H. Morgan. Random segregation versus coupling in mendelian inheritance. Science, 34:384, 1911.
- [88] A.H. Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43– 59, 1913.
- [89] http://www.accessexcellence.org/rc/vl/gg/comeiosis.php.
- [90] J.B.S. Haldane. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8:299–309, 1919.
- [91] A.M. Maxam and W. Gilbert. A new method for sequencing dna. Proceedings of the National Academy of Sciences of the United States of America, 74(2):560– 564, 1977.
- [92] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chainterminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12):5463–5467, 1977.
- [93] M.L. Metzker. Sequencing technologies the next generation. Nature Reviews Genetics, 11:31–46, 1977.
- [94] C.S. Pareek, R. Smoczynski, and A. Tretyn. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52:413–435, 2011.
- [95] M. Margulies and et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.
- [96] D.R. Bentley. Whole-genome re-sequencing. Current Opinion in Genetics and Development, 16:545–552, 2006.
- [97] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S.M. Johnson. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7):1051–1063, 2008.

- [98] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, and Law M. Lu, L. and. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012:251364, 2012.
- [99] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26:1135–1145, 2008.
- [100] H.C Bravo and R.A. Irizarry. Model-based quality assessment and base-calling for second generation sequencing data. *Biometrics*, 66(3):665–674, 2010.
- [101] j. Rougement, A. Amzalla, C. Iseli, L. Farinelli, I. Xenarios, and F. Naef. Probabilistic base calling of solexa sequencing data. *BMC Bioinformatics*, 9:431, 2008.
- [102] S. Swinnen, J.M. Thevelein, and E. Nevoigt. The statistics of bulk segregant analysis using next generation sequencing. *FEMS Yeast Research*, 12(2):215– 227, 2012.
- [103] S. Swinnen, K. Schaerlaekens, T. Pais, J. Claesen, G. Hubman, G. Yang, M. Demeke, M. Foulquie-Moreno, A. Goovaerts, K. Souvereyns, L. Clement, F. Dumortier, and J.M. Thevelein. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. *Genome Research*, 22:975–984, 2012.
- [104] I.M. Ehrenreich, N. Torabi, Y. Jia, J. Kent, K. Martis, J.A. Shapiro, D. Gresham, A.A. Caudy, and L. Kruglyak. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464:1039–1042, 2010.
- [105] P.M. Magwene, P.H. Willis, and J.K. Kelly. The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, 7:e1002255, 2011.
- [106] M. Edwards and D. Gifford. High-resolution genetic mapping with pooled sequencing. BMC Bioinformatics, 13(Suppl 6):S8, 2012.
- [107] L. Parts, J. Cubillos, J. Warringer, K. Jain, F. Salinas, S.J. Bumpstead, M. Molin, A. Zia, J.T. Simpson, M.A. Quail, A. Moses, E.J. Louis, R. Durbin, and G. Liti. The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, 7:e1002255, 2011.
- [108] J. Claesen, L. Clement, Z. Shkedy, and T. Burzykowski. Simultaneous mapping of multiple gene loci with pooled segregants. *PLoS ONE*, 8(2):e55133, 2013.

- [109] J. Claesen and T. Burzykowski. A hidden markov-model for gene mapping based on whole genome sequencing data. *Statistical Applications in Genetics* and Molecular Biology, doi:10.15.15/sagmb-2014-0007.
- [110] L.R. Rabiner and B.H. Juang. An introduction to hidden Markov models. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3:14–16, 1986.
- [111] W. Zucchini and I.L. MacDonald. Hidden markov models for time series. An introduction using R. CRC Press, 2009.
- [112] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.
- [113] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.
- [114] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.
- [115] T. A. Louis. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, 44:226–233, 1982.
- [116] J.C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequences. *Nucleic Acid Research*, 36:e105, 2008.
- [117] R. Plamondon and X. Li. Handling context dependence with dual hidden Markov model. *Vision interface*, '2000:95–100, 2000.
- [118] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.
- [119] W. Zucchini and P. Guttorp. A hidden Markov model for space-time precipitation. Water Resources Research, 27:1917–1923, 1991.
- [120] J.P. Hughes and P. Guttorp. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resources Research, 30:1535–1546, 1994.

- [121] J.P. Hudges, P. Guttorp, and S.P. Charles. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society*, Series C,48(1):15–30, 1999.
- [122] S.P. Charles, B.C. Bates, and J.P. Hughes. A spatiotemporal model for downscaling precipitation occurrence and amounts. *Journal of Geophysical Research Atmospheres*, 104:31657–31669, 1999.
- [123] E. Bellone, J.P. Hughes, and P. Guttorp. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climatology Research*, 15:1–12, 2000.
- [124] S.P. Charles, B.C. Bates, I.N. Smith, and J.P. Hughes. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, 18:1373–1394, 2004.
- [125] A.W. Robertson, S. Kirshner, and P. Smyth. Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. *Journal of Climate*, 17(22):4407–4424, 2004.
- [126] B. Betrò, A. Bodini, and Q.A. Cossu. Using a hidden Markov model to analyse extreme rainfall events in Central East Sardinia. *Environmetrics*, 19:702–713, 2008.
- [127] F. Yates. Tests of significance for 2 × 2 contingency tables. Journal of the Royal Statistical Society, 147:426–463, 1984.

Appendices

A.1 Tables

Table A.1: Peptides found in more than 90% of the 384 bovine cytochrome c tryptic-digest mass spectra.

| Mass | % | Mass | % |
|--------|-------|--------|------|
| 568.1 | 100 | 1584.8 | 100 |
| 779.4 | 100 | 1606.8 | 99.7 |
| 964.5 | 100 | 1633.6 | 100 |
| 1046.5 | 100 | 1649.6 | 100 |
| 1124.6 | 99.7 | 1672.9 | 100 |
| 1152.6 | 98.1 | 1820.7 | 91.7 |
| 1168.6 | 100 | 2010.0 | 100 |
| 1184.6 | 100 | 2026.0 | 100 |
| 1196.6 | 96.1 | 2032.0 | 100 |
| 1212.6 | 99.7 | 2042.0 | 98.4 |
| 1296.7 | 100 | 2058.0 | 97.1 |
| 1306.7 | 100 | 2138.1 | 100 |
| 1322.7 | 90.1 | 2154.1 | 100 |
| 1367.7 | 99.59 | 2160.1 | 97.4 |
| 1434.8 | 100 | 2170.1 | 98.4 |
| 1456.7 | 100 | 2186.1 | 95.3 |
| 1478.7 | 93.0 | 2465.2 | 100 |
| 1562.9 | 100 | | |

Table A.2: Estimated parameters of model (2) for ratios R(1,m), R(2,m), and R(3,m) for the indicated mass range.

| β | R(1) | R(2) | R(3) | | | |
|---|---------------------|----------------------|-------------------|--|--|--|
| No- sulfur-containing peptides, mass range :406-4000 | | | | | | |
| β_0 | -0.01824020003960 | 0.04836786201980 | 0.03182864110699 | | | |
| β_1 | 0.58774944141321 | 0.27156117306559 | 0.21751289233773 | | | |
| β_2 | -0.04374427526573 | -0.00231080382628 | -0.01989772300572 | | | |
| β_3 | 0.01569137227889 | 0.00128199330392 | 0.00517961166975 | | | |
| β_4 | -0.00188808146247 | -0.00019481372808 | -0.00051318105645 | | | |
| | One- sulfur-contain | ing peptides, mass r | ange : 406-4000 | | | |
| β_0 | -0.02967484608832 | 0.35639302719110 | 0.07467638437695 | | | |
| β_1 | 0.59144103400519 | -0.13157761023047 | 0.27251901416010 | | | |
| β_2 | -0.04590195614053 | 0.22773349494144 | -0.07667317606891 | | | |
| β_3 | 0.01707870774360 | -0.05797444396238 | 0.02329108779036 | | | |
| β_4 | -0.00208417838098 | 0.00545260364925 | -0.00243318653058 | | | |
| Two- sulfur-containing peptides, mass range :406-4000 | | | | | | |
| β_0 | -0.02384488784289 | 0.66680371115165 | 0.04919301885722 | | | |
| β_1 | 0.54435929270485 | -0.52043003662670 | 0.39221454011078 | | | |
| β_2 | -0.00744230605220 | 0.43849096392312 | -0.15353657635010 | | | |
| β_3 | 0.00573097136927 | -0.10969414060664 | 0.04302666609309 | | | |
| β_4 | -0.00093432040778 | 0.01016342015877 | -0.00425108044170 | | | |

Table A.3: Parameter estimation of the three-state DHHM model for pool 1 segregants of chromosome XIV with sequencing-error correction. The initial estimates of the total discordance probabilities of 0.1, 0.55, and 0.77 are considered for the first, second and the third state, respectively. The initial estimates of the transition matrix for all states are selected as the probabilities from a given state to itself is 0.7 and to another is 0.15. The values between the brackets are the corresponding 95% confidence intervals. For the parameters indicated with $^+$ a confidence interval could not be calculated as these parameters are at the boundary of the parameter space

| Parameters | "LtoR" Direction | "RtoL" Direction |
|---------------|-------------------------------|-------------------------------|
| μ_1 | $0.2255 \ [0.2187, \ 0.2320]$ | $0.2255 \ [0.2187, \ 0.2320]$ |
| μ_2 | $0.1283\ [0.1260,\ 0.1290]$ | $0.1283 \ [0.1260, \ 0.1290]$ |
| μ_3 | $0.0557\ [0.0530,\ 0.0580]$ | $0.0557 \ [0.0530, \ 0.0580]$ |
| δ_1 | 0^{+} | 0+ |
| δ_2 | 1 ⁺ | 1 ⁺ |
| δ_3 | 0^{+} | 0^{+} |
| γ_{11} | $0.4614\ [0.4519,\ 0.4709]$ | $0.4614 \ [0.4519, \ 0.4709]$ |
| γ_{12} | $0.4273\ [0.4181,\ 0.4365]$ | $0.4340\ [0.4250,\ 0.4430]$ |
| γ_{13} | $0.1111\ [0.1042,\ 0.1180]$ | $0.1046 \ [0.0979, \ 0.1113]$ |
| γ_{21} | $0.0210\ [0.0201,\ 0.0219]$ | $0.0210 \ [0.0201, \ 0.0219]$ |
| γ_{22} | $0.9226\ [0.9221,\ 0.9231]$ | $0.9227 \ [0.9222, \ 0.9232]$ |
| γ_{23} | $0.0563\ [0.0548,\ 0.0578]$ | $0.0563 \ [0.0548, \ 0.0578]$ |
| γ_{31} | $0.0074\ [0.0038,\ 0.0111]$ | $0.0075 \ [0.0039, \ 0.0111]$ |
| γ_{32} | $0.0740\ [0.0724,\ 0.0756]$ | $0.0739\ [0.0723,\ 0.0755]$ |
| γ_{33} | $0.9185\ [0.9181,\ 0.9189]$ | $0.9185 \ [0.9181, \ 0.9189]$ |

A.2 Figures





Figure A.2: ROC curves for all four analyzed spectra combined.



Figure A.3: ROC curves for all four analyzed spectra combined.



Figure A.4: Histograms of the residuals of the estimated polynomial model.



Figure A.5: Empirical scatter plot between the Mascot and Pearson's χ^2 statistic.

A.3 Nielsen *et al.*

Nielsen *et al.* [69], proposed the concept of a peptide window in which the accurate monoisotopic mass of a peptide should reside. We have performed such an analysis starting from the 19,881 tandem MS scans that were identified by SEQUEST. Interestingly, 17,838 spectra had a monoisotopic mass that fell within in the defined peptide window. However, 2,043 spectra fell outside this mass range. For 10,015 spectra, the four consecutive isotope peaks were found and a Pearson's χ^2 statistic was calculated. From this set 1,878 spectra fell outside the peptide window. The distribution of the data is presented in Table A.3 according to a XCorr threshold of 3.5 and a Pearson's χ^2 threshold of 0.1. For each region a two-sided hypergeometric test was performed according to Yates [127]. Region I and II were found to be enriched for molecules outside the peptide window, whilst region III was found to be depleted.

Table A.3: The distribution of the number of tandem MS scans having monoisotopic mass reside outside the peptide window for each region.

| | Outside peptide window | All |
|---------------------------|------------------------|-------|
| Region I | 190 | 691 |
| Region II | 232 | 724 |
| ${\rm Region}\ {\rm III}$ | 792 | 5192 |
| Region IV | 664 | 3408 |
| Total | 1878 | 10015 |

Samenvatting

In dit proefschrift stellen we statistische methodes voor waarmee gegevens over het proteoom en het genoom kunnen geanalyseerd worden. In het laatste decennium, wordt massa-spectrometrie-gebaseerde proteomica vaak gebruikt als high-throughput methode voor de identificatie en kwantificatie van eiwitten in complexe biologische stalen. Gegevens van dergelijke experimenten bevatten redundante en irrelevante informatie, en zijn vaak onderhevig aan ruis. Hierdoor is het moeilijk om de biologisch relevante signalen, i.e., de peptiden en eiwitten, te onderscheiden van ruis-signalen. Een oplossing voor dit probleem is het vergelijken van de gemeten signalen met theoretisch berekende, biologische signalen. In hoofdstuk 4 evalueerden we twee similarity measures, de Pearson χ^2 statistiek en de Mahalanobis-afstand, waarmee relevante biologische signalen, i.e., de isotopen-verdeling, gedetecteerd kunnen worden. Het onderscheidingsvermogen van de Pearson χ^2 statistiek was hoger dan het vermogen van de Mahalanobis-afstand in een MALDI-TOF experiment.

De identificatie van eiwitten speelt een belangrijke rol in proteomica. Een van de meest gebruikte technieken voor eiwit-identificatie en -kwantificatie is shotgun proteomics. Tandem MS spectra worden vergeleken met databanken met de hulp van gespecialiseerde zoekmachines. Deze zoekmachines vergelijken niet-gedentificeerde eiwitfragmenten met theoretische data, en kennen een score toe die uitdrukt hoe groot de gelijkenis is tussen het eiwitfragment en de theoretische data. Des te beter deze score is , des te waarschijnlijker de identificatie is. Het bepalen wanneer een score goed is, is niet eenvoudig en is voor vele zoekmachines een uitdaging. Zelden of nooit wordt door deze zoekmachines de isotopenverdeling van de precursor-eiwitten gebruikt. In hoofdstuk 5 stelden we de Pearson χ^2 statistiek als maatstaf voor om de gelijkenis tussen de geobserveerde en theoretische isotopenverdeling van de precursor-eiwitten

te bepalen. Dankzij de Pearson χ^2 statistiek konden we aantonen dat een goede score voor een bepaalde identificatie niet noodzakelijk overeenkomt met een grote gelijkenis tussen de geobserveerde en berekende isotopenverdeling van de precursor-eiwitten, en omgekeerd. Het combineren van de Pearson χ^2 statistiek en de zoekmachine-scores leidde tot een verhoogde sensitiviteit en specificiteit van de eiwit-identificatie.

Er bestaan vele zoekmachines voor de analyse van tandem MS data. De resultaten van deze zoekmachines zijn verschillend qua accuraatheid, sensitiviteit en specificiteit. De onderliggende reden hiervoor is de manier waarop de scores berekend worden. Desondanks deze verschillen, zijn we genteresseerd in de mate van overeenkomst tussen de resultaten van de zoekmachines. We vragen ons bijvoorbeeld af waarom een identificatie met SEQUEST gelijk of niet gelijk is aan een identificatie met MASCOT. In hoofdstuk 6, stelden we Cohen's kappa-cofficint voor om de mate van overeenkomst te bepalen tussen de MASCOT en SEQUEST identificatie-resultaten. Aan de hand van de Cohen's kappa cofficint vonden we dat er een goede overeenkomst is tussen de resultaten van MASCOT en SEQUEST.

De opkomst van high-throughput sequencing methodes, zoals NGS, heeft voor een omslag gezorgd in biologisch en biomedisch onderzoek. Dankzij deze techniek kan men in DNA efficint en op grote schaal polymorfe nucleotiden detecteren. Deze nucleotiden kunnen onder andere gebruikt worden als moleculaire merkers om de functie van bepaalde genen vast te stellen. In hoofdstuk 9, 10 en 11 introduceerden we een aantal aanpassingen aan een hidden Markov model [109] dat gebruikt werd om verschillende QTLs te identificeren die verantwoordelijk zijn voor abnormale ethanol tolerantie in S. cerevisiae. En van de mogelijke aanpassingen is gekoppeld aan de onderliggende afhankelijkheid tussen de moleculaire merkers. In hoofdstuk 10, stelden we een niet-homogeen HMM voor. In een iet-homogeen HMM zijn de overgangskansen een functie van n of meerdere covariaten. Op deze manier kunnen we rekening houden met de afstand tussen twee naburige merkers. Dit niet-homogene model kon eveneens verscheidene gekende genen identificeren die verantwoordelijk zijn voor een abnormale ethanol tolerantie. In hoofdstuk 11 breidden we het basis-HMM uit zodat het kan omgaan met merkers van twee verschillende groepen. Dankzij deze aanpassing kunnen significante verschillen tussen twee merker-groepen gevonden worden. Dit joint-HMM kon in chromosoom XIV potentiele chromosomale regio's identificeren die gerelateerd zijn aan ethanol tolerantie. In chromosoom IX werkte deze aanpak niet.