University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Piotr Dittwald

Computational methods for large-scale data in medical diagnostics

PhD dissertation

Supervisors

dr hab. Anna Gambin
Institute of Informatics, University of Warsaw

dr hab. Paweł Stankiewicz
Baylor College of Medicine, Houston

May 2014

Author's declaration:
aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

May 7, 2014 . . . .
date Piotr Dittwald

Supervisors' declaration:
the dissertation is ready to be reviewed.

May 7, 2014 . . . .
date dr hab. Anna Gambin dr hab. Paweł Stankiewicz

Computational methods for large-scale data in medical diagnostics

Abstract

This thesis covers a topic of fast and reliable processing of the high-throughput biomedical data, that is currently needed in genetics and proteomics. We therefore concentrate on these two rapidly developing research areas in life sciences.

First, we perform a systematic analyses of human reference genome build in the context of its potential local instability caused by recurrent genomic rearrangements, e.g. deletions, duplications, and inversions. Our approach enables also to analyze large and unique clinical database.

Secondly, we present various analyses of mass spectrometry data. In particular, we propose isotopic distribution at many levels of accuracy; more precisely we consider aggregated and fine isotopic structures. We also show some case application studies involving high-throughput processing, potentially applicable in proteomics and lipidomics.

Of note, this thesis is also an exemplification of interdisciplinary approach for basic science, where a deeper and complex understanding of both biomedical and computational aspects can be mutually beneficial.

Keywords: computational methods, bioinformatics, mass spectrometry, recurrent genomic rearrangements

ACM Classification: J.3

Metody obliczeniowe dla wielkoskalowych danych w diagnostyce medycznej

Streszczenie

Niniejsza rozprawa opisuje efektywne metody przetwarzania wielkoskalowych danych w biologii molekularnej, co jest szczególnie istotne w genetyce i proteomice. Właśnie te dwie dynamicznie rozwijające się gałęzie nauk o życiu stanowią obszar naszych zainteresowań.

Na początku przeprowadzamy systematyczną analizę referencyjnego genomu człowieka. Nasze badania dotyczą jego potencjalnej lokalnej niestabilności spowodowanej przez nawracające rearanżacje, takie jak delecje, duplikacje oraz inwersje. Przedstawione podejście pozwala również, w przypadku delecji i duplikacji, przeanalizować dużą i unikalną bazę danych przypadków klinicznych.

W drugiej części rozprawy prezentujemy modele wykorzystywane w analizie danych spektrometrycznych. W szczególności zajmujemy się wpływem wariantów izotopowych na wyniki uzyskiwane w eksperymentach. Nasze badania prowadzimy wykorzystując różne stopnie dokładności przy reprezentowaniu rozkładów izotopowych – podejście zagregowane oraz dokładne. Ponadto przedstawiamy przykłady analizy wieloskalowych danych w proteomice.

Pragniemy podkreślić, że niniejsza rozprawa prezentuje interdyscyplinarne podejście do badań podstawowych. Ponadto, nasze badania są przykładem kompleksowego wykorzystania w nauce o życiu metod obliczeniowych popartych teorią nauk matematycznych.

Słowa kluczowe: metody obliczeniowe, bioinformatyka, spektrometria mas, nawracające rearanżacje genomowe

Klasyfikacja tematyczna ACM: J.3

Contents

Rodzicom i Siostrze

Acknowledgments

Na początku dziękuję moim promotorom – dr hab. Annie Gambin oraz dr. hab. Pawłowi Stankiewiczowi – za wszystko, czego się od nich nauczyłem podczas minionych lat, a co niekoniecznie zawarte jest w niniejszej rozprawie. Oraz za wsparcie.

Dziękuję moim koleżankom i kolegom ze studiów (na MISDoMPie oraz na MIMie, ze szczególnym uwzględnieniem pokoju 5810), pracownikom Uniwersytetu, którzy pomagali mi na najrozmaitsze sposoby (nierzadko odznaczając się przy tym dużym poczuciem humoru), a także studentom!

I would like to thank all people I met during my PhD studies, especially my coauthors – that was a pleasure and unforgoteable experience to work with you. Especially, I would like to thank Dr. Dirk Valkenborg for huge number of inspirations, not only scientific ones – Dank U!

Dziękuję Wszystkim za życzliwość, wyrozumiałość, ciekawe rozmowy i wspólne chwile, które bardzo mi pomogły i mam nadzieję, że pomagać będą nadal – na co bardzo liczę!

Science can purify religion from error and superstition; religion can purify science from idolatry and false absolutes. Each can draw the other into a wider world, a world in which both can flourish.

Saint John Paul II

(Letter to the Rev. George V. Coyne, S.J., Director of the Vatican Observatory, 1 June 1988)

1 Introduction

The bottleneck of the large-scale data processing has made bioinformatic analyses a crucial component in the life sciences workflows. The two large fields in biomedical studies, whose rapid development in the recent years has depended on computational methods, are genetics and proteomics. They both are strictly connected to each other, e.g. structural organization of the genome affects the variety of proteins in the organism; on the other hand, proteins are the crucial functional molecules that participate in the process of extracting the information encoded in the genome. In this thesis, we present selected bioinformatic methods used and discuss their application in basic research as well as in clinical diagnostics.

1.1 Methods for genome stability analysis

Human genome organization

A wide range of the human organism functions are encoded in a deoxyribonucleic acid (DNA). The structure of this molecule was discovered in 1953 by Watson and Crick, who once stated: ”It has not escaped our notice that the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material” (Watson and Crick, 1953).

The DNA double helix is composed of two strands of nucleotides oriented in opposite directions. Each nucleotide is built of a sugar-phosphate backbone and one of four nucleobases: adenine (A), guanine (G), thymine (T), or cytosine (C); A and G are classified as purines and T and C as pyrimidines. Two DNA strands are connected by hydrogen bonds, two between

A and T and three between C and G. The nucleobases in the DNA strand are connected by the bonds between the third and fifth carbon of the sugar molecules. Thus, each DNA strand has two ends termed 5' and 3'. The complementary nucleobase is referred to as base pair (bp) and is considered as a standard unit of DNA length.

Human DNA is compacted as chromatin and divided into 23 pairs of chromosomes[1]: 22 autosomes (numbered from 1 to 22), and one pair of sex chromosomes (X and Y). Males have one chromosome X and one chromosome Y and females have two chromosomes X. A complete set of chromosomes in a somatic cell is referred to as a karyotype. Human genome is diploid, i.e. all autosomes have the homologous copies; each chromosome in a pair is inherited from one parent in the process of meiosis. During fertilization, male and female gametes fuse, forming a single cell zygote that further divides in a process of mitoses, replicating the initial double helix DNA.

Each metaphase chromosome in human has been represented as an X-shaped structure, with two short p arms and two long q arms that are connected by a centromere[2]. The regions near ends of chromosomes are called telomeres composed of thousands of repeated TTAGGG sequences and stabilized by an enzyme, telomerase. This simple classification has been further subdivided based on G-banding chromosome staining, a technique, in which separate regions of chromosomes dyed by Giemsa stain show different banding pattern visible in a light microscope. These bands have been classified in a standard cytogenetic nomenclature, e.g. 1q21.1 designates chromosome 1, arm q, region 2, band 1, and sub-band 1. Relative location on a chromosome arm is referred to as proximal or distal when closer to or farther away from a centromere, respectively. A basic functional unit of DNA sequence is a gene. In humans, genes consist of exons (protein-coding intervals) and introns, which are removed in a process of splicing[3]. In addition, genes are usually accompanied by regulatory sequences such as promoters and/or enhancers.

Humans share the vast majority of nucleotides on the analogous (allelic) chromosomal loci, and determination of these base pair sequences was a primary goal of the international scientific research endeavor called Human Genome Project (HGP) initiated in 1990. HGP announced almost complete human DNA reference sequence (IHGSC, 2004) as hg17/NCBI Human Build 35 (May 2004). This genome assembly has been continually updated to the current version hg38/NCBI Human Build 38 (December 2013) by the Genome Reference Consortium (GRC). It should be noted that there are regions in the human reference genome, for which the exact nucleotide sequence is still not well determined. These sequences are

[1] In addition to this linear nuclear genome, humans have a circular-shaped mitochondrial DNA, however, we do not consider the mitochondrial genome here.

[2] For simplicity we use this nomenclature for chromosome coordinates regardless the phases of chromosome life cycle.

[3] We do not consider here genes that do not code proteins.

described as gaps mostly located in the telomeric and centromeric regions.

Each human individual DNA sequence is defined as a genotype. The differences between genotypes are caused, among others, by gene variants (alleles). Among many other factors, such as environmental interactions or post-translational modifications, these variants can contribute to the observable traits, i.e. a phenotype. Phenotypic trait expressed by these genes can be inherited as either autosomal recessive (AR; manifested when two alleles associated with this trait are mutated) or autosomal dominant (AD; one mutated allele is sufficient to manifest the phenotype).

In males, chromosome X is inherited from the mother and Y from the father. For the vast majority of X-linked or Y-linked genes, males have only one copy; disruption of the genes on chromosome X in males usually has phenotypic consequences as, opposed to disruption of genes on autosomes. X-linked genes are responsible for X-linked recessive traits when only one allele is mutated and not manifesting in female carriers of the allele. In case of X-linked dominant disease, both males and females with the mutated allele are affected.

Mutations can be lethal, cause non-lethal diseases of various severity levels, or might be not associated with pathogenic consequences. A set of genomic nonpathogenic mutations can be inherited as haplotypes that further segregate in a population. This variability should be taken into account when considering the human reference genome build as a golden standard for an individual genotype.

Mutation mechanisms

Mutation of a DNA sequence can be caused by errors in DNA replication, recombination, or repair. Mutations are classified based on their inheritance pattern, as they can occur *de novo* or can be inherited from a parent. Moreover, we can distinguish meiotic mutations (originating in germ cells) being present in 100% of child's cells[4], as constitutional, and mitotic somatic mutations that are acquired and propagated only to some cells (somatic mosaicism).

A change of a single base pair is referred to as Single Nucleotide Variants (SNV). Depending on SNV location, SNV may affect the coding region by changing the encoded amino acid, e.g. missense mutations, cause premature stop codon, e.g. nonsense mutations (nonsynonymous mutations), or alter the codon without changing the transcribed products (synonymous mutations, i.e. silent). Moreover, insertions or deletions of small portion of nucleotides (i.e. indels) might also cause a shift of the transcription reading frame (frameshift mutations).

Deviations from the 46 number of chromosomes (i.e. numerical chromosomal aberrations) often result from an abnormal chromosome segregation and manifests with pathogenic phenotypes. This is usually caused by nondisjunction during meiosis, when the chromosome pair is not properly separated, causing an imbalanced chromosome complement (i.e. mono-

[4] If not altered by other mutation.

somy or trisomy) in the daughter cells. The best known examples of numerical aberrations are trisomy of chromosome 21 (Down syndrome), trisomy of chromosome 18 (Edwards syndrome), or three sex chromosomes XXY in males (Klinefelter syndrome).

In addition to single base pair changes, also structural aberrations, e.g. deletions, duplications, translocations, insertions, or inversions of the chromosomal fragments, are observed. A portion of the abnormal number of copies of one or more DNA fragments resulting in an imbalance of DNA is referred to as a Copy-Number Variant (CNV). CNV size can vary from a few to thousands (i.e. kb) or millions of base pairs (i.e. Mb). The term genomic disorders has been coined for both the rearrangements themselves as well as the resulting pathogenic features (Lupski, 1998), caused e.g. by gene disruption or change in gene copy-number.

Mechanisms for structural aberrations origin

In this thesis, we focus on recurrent genomic rearrangements, i.e. rearrangements occurring *de novo* in the same genomic loci in different individuals. The main mechanism responsible for recurrent rearrangements is nonallelic homologous recombination (NAHR), wherein recombination breakpoints are located within highly similar DNA sequences, e.g. low-copy repeats (LCRs).

LCRs or segmental duplications (SDs) (Bailey et al., 2002) are defined as pairs of DNA fragments with fraction matching (homology score) over 90% and longer than 1 kb. It has been shown (Stankiewicz and Lupski, 2002) that for long LCR elements with high homology (originally the parameters were suggested to be 10-400 kb and 97%), the NAHR events might occur within LCRs causing inversions (for inversely oriented LCRs), deletions or reciprocal duplications (for directly oriented LCRs)[5], or reciprocal translocations.

Alternative mechanisms for CNVs origin such as microhomology-mediated break-induced replication (MMBIR) (Hastings et al., 2009) or fork stalling and template switching (FoSTeS) (Lee et al., 2007) have been also described. These are the two major DNA replication error mechanisms leading to nonrecurrent genomic rearrangements, for example complex duplication and deletion events (Lee et al., 2007).

Molecular experiments for genome analysis

A DNA sequence for a specific region can be determined using a Sanger sequencing reaction[6]. In this technique, the investigated DNA fragment is typically amplified using polymerase chain reaction (PCR). In a first phase, the short but unique DNA primers flanking the analyzed fragments are designed. The chain reaction, based on thermal cycles, enables

[5] These rearrangements have often prefix micro referring to their sub-microscopic size.

[6] Currently, next generation sequencing (NGS) is a broadly used alternative to Sanger sequencing and is also useful for CNV detection; however, NGS technology will not be used in research covered in this thesis.

replication of DNA material, growing exponentially with time. Then, the amplified DNA fragment is analyzed in a chain-termination reaction using A, C, G, and T deoxynucleotides and four radioactively or fluorescently labeled dideoxynucleotides (substituting one of the original nucleotides and terminating the nucleotide chain). Finally, all possible prefixes of the analyzed DNA sequence are obtained, each terminated by a tag easy to recognize. Using gel electrophoresis, these prefixes can be sorted according to their length to identify the DNA sequence.

Genome analysis of CNVs longer than 5 Mb can be visualized in the light microscope after chromosome staining, using e.g. G-banding. For smaller DNA changes, molecular biology techniques are used, e.g. fluorescent in situ hybridization (FISH). FISH is based on the concept of fluorescently-labeled probes binding (hybridizing) to a specific target DNA locus that can be analyzed in a fluorescent microscope (O'Connor, 2008). FISH technique is fast and easy for visual interpretation of single CNVs.

Microarray-based Comparative Genomic Hybridization (aCGH) is a method allowing for high-throughput genome-wide data processing in one experiment (Chial, 2008). Thousands or millions of DNA fragments (e.g. oligonuceleotide probes) can now be placed on a single glass slide (array). By analyzing control and patients DNA samples, and labeling them differently with fluorescent dyes (e.g. green for control, red for patient), it is possible to compare intensities of the fluorescent signals referring to copy-number ratios. For example, in a case of equal copy-numbers the yellow signal is observed, whereas more red/green signal is associated to duplication/deletion in the patient's genome, respectively. The aCGH method allows for detection of CNVs as small as tens of kilobases.

Genome (in)stability analyses

The (in)stability of the human genome is directly related to its structure. Both the size of the genome (over 3 billion of base pairs) and its complexity make it unfeasible to be systematically analyzed without application of automated algorithms. Moreover, fast and reliable processing of the outputs (e.g. of cytogenetic testing by chromosomal microarray analysis, CMA) should be integrated with phenotypic data provided by physicians and current knowledge about the genotype/phenotype correlation. Finally, the wet-bench experiments serve as an indispensable way to confirm the molecular bases of the identified rearrangements. In this interdisciplinary approach, using multiple data resources, the crucial steps of analyses involve understanding the biological background, designing the computational workflow and discussion of the results in a medical context.

Genome-wide analyses of known and potential novel recurrent deletions and reciprocal duplications

In this study, based on the literature data, we systematically analyzed the genomic regions of genetic diseases and syndromes associated with NAHR-mediated recurrent deletions and reciprocal duplications. Moreover, we queried and cross-referenced large database of high-resolution genomic analyses performed at Baylor College of Medicine on patients referred for CMA. The applied algorithms using custom scripts allowed us to filter out CNVs that correspond to NAHR-syndrome regions flanked by directly oriented paralogous LCRs (DP-LCRs). The causative association of the patients' rearrangements with the known genetic syndromes involved manual specification of the selected parameters to tackle the issue of different sensitivity of CMA arrays. As a result, we were able to determine the prevalence of the known recurrent genomic disorders in the clinical CMA database. We also determined the frequencies of the novel rearrangements. To this aim, we narrowed the study to the in silico cases with genomic breakpoints of the investigated CNVs mapped with a sufficient resolution. A statistic model based on a quasi-Poisson regression, suitable for count data with missing values, has been used to report genomic features that correlate with the frequencies of *de novo* recurrent rearrangements. Moreover, several architectural features of the LCR clusters flanking the interrogated regions have been investigated.

Furthermore, we constructed a new genome-wide map of the DP-LCR-flanked regions in the human genome (build hg19), i.e. the genomic regions where recurrent deletions or reciprocal duplications might occur via LCR-mediated NAHR. We also introduced a concept of computationally determined LCR cluster using a hierarchical clustering algorithm (Figure 1.1) and investigated the multiple parameters to propose the cut-off height of the clustering tree. The clustering approach enabled us to systematically distinguish between overlapping and adjacent regions, and to combine very similar regions. For example, we identified four novel recurrent NAHR-mediated deletions involving chromosome 2q12.2q13, which were previously referred to as a single region. Selected breakpoints of these novel rearrangements were sequenced using wet-bench experiments, and further clinically characterized. Using annotation of gene location and the OMIM database (http://www.omim.org/), we not only identified potentially disrupted genes, but also those of them that might cause known disease via NAHR, and might be useful in diagnostics.

Finally, the homology between the LCR clusters flanking the newly defined NAHR-prone regions has been visualized using Miropeats program (Parsons, 1995). This information might be useful for researchers to better understand the complexity of genomic regions where recombination hot spots occur.

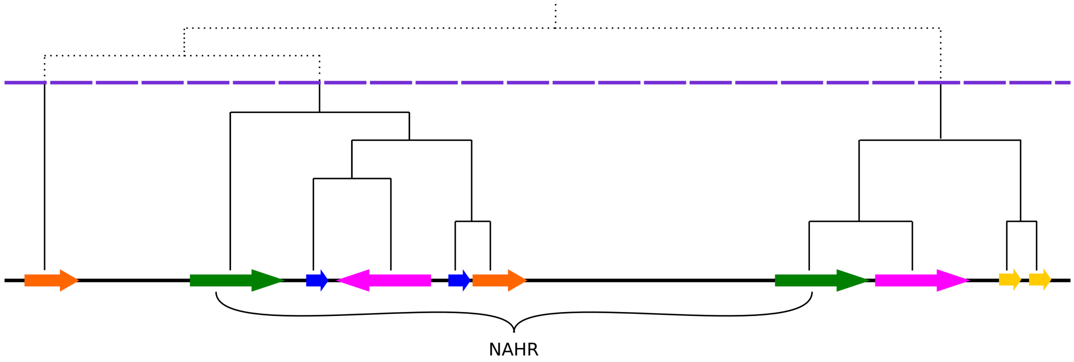

Figure 1.1: Schematic visualization of the concept of LCR clusters. LCR elements in which NAHR breakpoints occur (green) are accompanied by other elements that can be grouped into LCR clusters. The hierarchical clustering algorithm constructs a clustering tree that can be then pruned at a given height (violet dotted line). Source: Dittwald et al. (2013c)

Genome-wide analyses of potential recurrent inversions

It should be noted that balanced genomic rearrangements (e.g. paricentric or paracentric inversions) are not detectable by CMA assays. Our genome-wide computational approach aimed to investigate human genome instability potentially caused by balanced genomic inversions. We identified a set of inversely oriented, paralogous LCRs (IP-LCRs) that can potentially mediate recurrent inversions via NAHR, by integrating the latest version of human genome build (hg19), and the criteria from the literature applied for directly oriented LCRs that can potentially mediate deletions and duplications. Similarly to the previous section, our algorithms utilized efficient operations on intervals to efficiently analyze the genome. The set of IP-LCRs allowed us to estimate the fraction of the human genome where inversion breakpoints might be located, as well as the fraction of genome potentially unstable due to NAHR mediated by IP-LCRs.

The balanced rearrangements may disrupt the genes harboring the recombination site. Therefore, we reported a set of genes, for which at least one inversion breakpoint is located within such a gene, and identified genes that are dosage-sensitive and/or associated with diseases. We also analyzed the X-linked genes, as they have relatively high likelihood of clinically manifesting the recurrent inversions. Further, we focused on the known disease genes, i.e. those for which NAHR-mediated inversion might cause the already known disease. We also processed the genomic inversions from the Database of Genomic Variants (DGV) (Zhang et al., 2006) that could be associated with NAHR and estimated the statistical significance of such events.

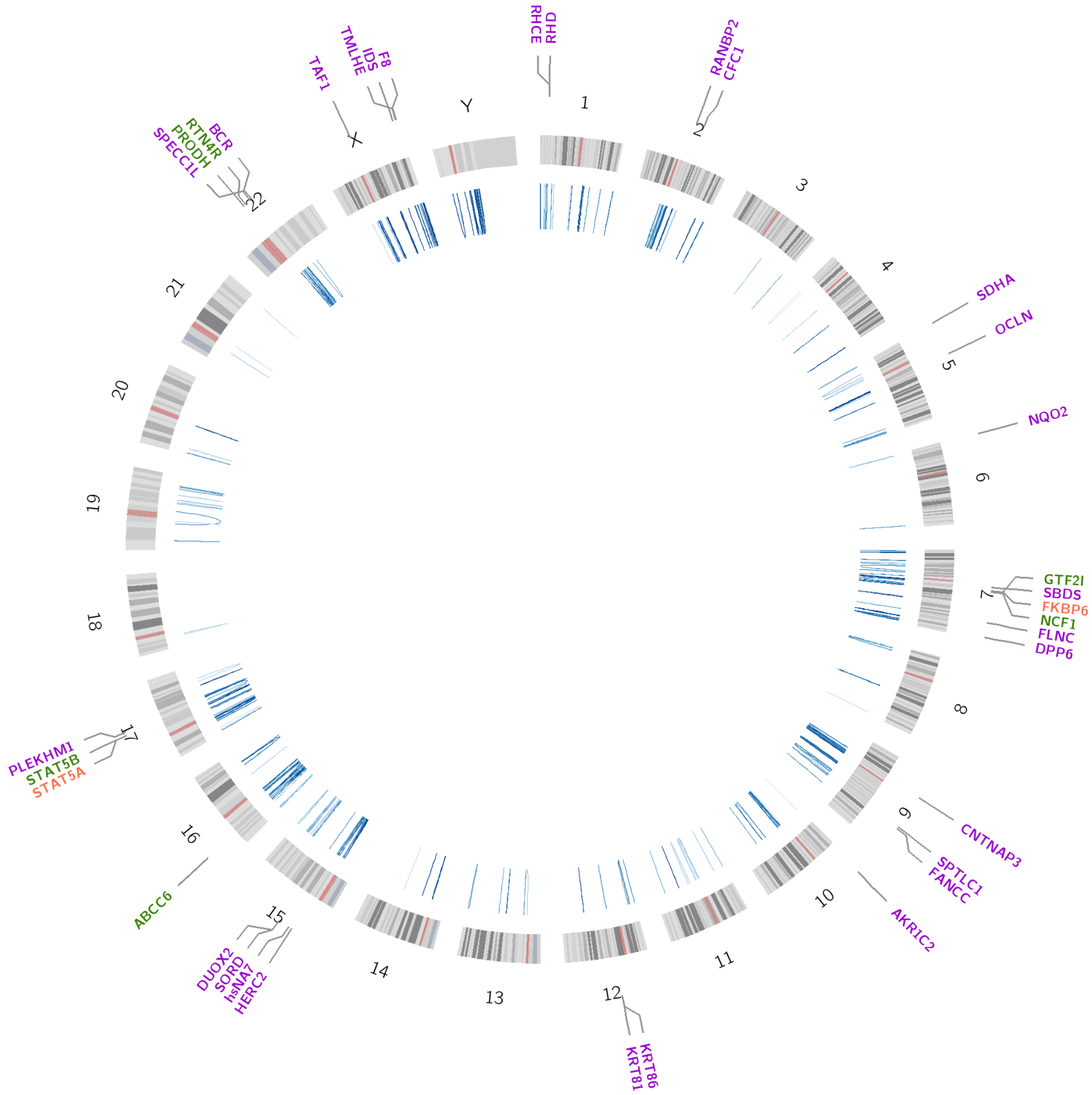

Figure 1.2: The Circos plot (Krzywinski et al., 2009) that depicts the identified genes potentially disrupted by NAHR-mediated inversions genome-widely. We highlighted the genes that are associated with diseases (violet), dosage sensitive (red), and those from both previous groups (green). Figure source: Dittwald et al. (2013b)

1.2 Methods for proteome analysis

Human proteome organization

The Central Dogma of molecular biology describes the flow of information from genes to proteins (Crick, 1970). First, the sequence of nucleotides is transcribed into mRNA molecule, which is further translated to amino acid sequence, composing a protein molecule. The structure of proteins can be considered at different levels: the primary structure describes a sequence of amino acids, while the secondary structure covers the hydrogen-bonds-driven substructures, e.g. α-helices or β-sheets. The tertiary and the quaternary structures refer to three-dimensional folding and cristal forming of proteins, respectively. Tertiary structure, also called conformation, is highly linked to the protein function. The information about the whole set of proteins expressed in the organism (i.e. proteome) – their amounts, functions, and interactions – is crucial for describing biological systems.

The 20 naturally occurring amino acids are built from five chemical elements: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S)[7]. The structure of amino acid can be divided into: amino group[8], a carboxyl group, the central carbon atom (C_α), and a side chain. Peptide (polypeptide chain[9]) is a short sequence of the amino acids linked by peptide bonds. As a product of forming single peptide bond, a water molecule (H_2O) is released. A polypeptide chain can be created e.g. as a product of enzymatic digesting of a protein. By convention, a polypeptide chain is described from its N-terminus (the end with free $-NH_2$ or $-NH_3^+$ group) to its C-terminus (the end with free $-COOH$ or $-COO^-$ group).

Chemical atoms are built of protons (positively charged), neutrons (not charged), and electrons (negatively charged). Protons and neutrons, also called nucleons, form the nucleus, where the vast majority of the atomic mass is concentrated (therefore the electron mass will be omitted further in this thesis). Many chemical elements have isotopes[10], i.e. the variants that differ by the amount of neutrons. Here, we will consider only stable isotopes of the five chemical elements building peptides, namely C, H, N, O, and S. The lightest isotope variant is called monoisotopic (in our case these are ^{12}C, ^{1}H, ^{14}N, ^{16}O, ^{32}S). A mass unit commonly used for chemical molecules is dalton (Da), defined as $\frac{1}{12}$ the mass of carbon ^{12}C, and approximately equal to 1.66×10^{-27} kg. The nominal mass of the element is the mass of its isotopic variant rounded to the integer value. The five considered elements have two (car-

[7] Amino acids can also contain other elements, like phosphorus, as a result of post-translational modifications (PTMs).

[8] Imino group in case of prolyne.

[9] Basically, peptides are short sequence of amino acids, while polypeptide are longer, however, we will not distinguish in this thesis between the two classes.

[10] We will consider only stable isotopes, and ignore the radioactive forms which spontaneously undergo the radioactive decay.

bons: ^{12}C, ^{13}C; hydrogens: ^{1}H, ^{2}H; nitrogens: ^{14}N, ^{15}N), three (oxygens: ^{16}O, ^{17}O, ^{18}O), or four (sulphurs: ^{32}S, ^{33}S, ^{34}S, ^{36}S) isotopic variants. Each of these isotopes has a certain exact mass, denoted as $M_{C_{12}}, \ldots, M_{S_{36}}$, and appears in the nature with a certain probability, denoted as $P_{C_{12}}, \ldots, P_{S_{36}}$. The average mass of the element is a weighted sum of its isotopes.

Mass spectrometry and its applications in proteomics

According to Eidhammer et al. (2007), the main tasks for analytical methods in proteomics are:

1. to identify the protein in the sample;
2. to characterize the various features of the protein (regardless its identification);
3. to quantify the amount of the protein in the sample;
4. to compare the occurrence/abundance/modifications of the proteins between the samples.

Mass spectrometry (MS) is one of the most popular analytical method used in proteomics to investigate the content of the chemical mixture, which has already brought a huge insights into the role of biological systems (Cravatt et al., 2007; Chandramouli and Qian, 2009). The instrumentation used in this method, i.e. mass spectrometer, is composed of the three main parts:

1. the ionization source – the molecules are charged (i.e. ions are created) and brought to a gas phase;
2. the mass analyzer – ions are separated by their mass-to-charge (m/z) ratio;
3. the detector – the spectrum of signals or peaks is produced, it assigns abundance, i.e. number of ions, for a given m/z.

Of note, MS was invented more than a hundred years ago by Thompson, however, its rapid growth is dated in the last decades of the XX century, when soft ionization techniques (producing almost no fragmentation of the analyzed molecules) was proposed by John Fenn and colleagues. This technology was called in a vivid manner as ”Electrospray Wings for Molecular Elephants” in Fenn’s Nobel Prize lecture (Fenn, 2002). In addition, before the sample is analyzed by the mass spectrometer, it is often fractionated in order to increase the detection rate, e.g. by gel electrophoresis of liquid chromatography (LC). It should be also noted that in the existing instruments used in proteomics many types of the described components occur (cf. Table 1.1) (Aebersold and Mann, 2003).

Table 1.1: Selected types of mass spectrometry instruments used in proteomics. The comparison of the Orbitrap with FT-ICR and TOF MS is presented in (Zubarev and Makarov, 2013).

name	type	description	reference
matrix-assisted laser desorption ionization (MALDI)	ionization source	the sample is mixed with a matrix, and further released e.g. by an ultraviolet (UV) beam, usually an ion is singly protonated;	(Peter-Katalinic, 2007)
electrospray (ESI)	ionization source	the sample is ionized within a very thin needle using high voltage; then, the droplets are injected into the atmosphere, where the solvent evaporates, producing the multicharged ions of the analyzed molecule; this method is especially useful in proteomics thanks to the ease of combining with liquid based separation of sample;	(Cole, 1997; Gross et al., 2002)
time-of-flight (TOF)	mass analyzer	ions are separated using the time they reach the detector after being accelerated in the electric field – the square of velocity of the accelerated ion is proportional to the m/z ratio;	(Cotter, 1994)
ion trap	mass analyzer	three-dimensional (quadrupole ion trap) or rectangular (linear ion trap) construction produces an electric or magnetic field within a high vacuum system; the field (its frequency and potential) is manipulated in a such way that only the molecules with selected m/z ratio reach the detector;	(Brancia, 2006)
Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR MS or FTMS)	mass analyzer	the m/z ratio can be calculated using the frequency of rotation of the investigated ions in the spatially uniform, cyclic magnetic field;	(Marshall et al., 1998)
Orbitrap	mass analyzer	ions are orbiting around the electrode within the electrostatic field, and the frequencies of their harmonic oscillations are proportional to $(m/z)^{-1}$.	(Makarov, 2000)

The ability to discriminate between the neighboring peaks is described by the resolution coefficient $R = \frac{M}{\Delta M}$. The Full Width at Half Maximum (FWHM) approach defines ΔM as the peak width at half of its height, and M is the mass at top of the peak. The resolution can be expressed in parts-per-milion (ppm), i.e. multiplied by 10^6 factor.

Finally, it should be noted that two basic approaches for MS proteomics are of use. The top-down analysis investigates the intact molecules, while the bottom-up analysis investigates at once the mixture of proteins digested by the proteolytic enzymes called peptidases (e.g. trypsine) (Yates and Kelleher, 2013).

MS data preprocessing

We distinguish two types of noise in the mass spectra associated with its origin: chemical (producing unexpected peaks, e.g. from contaminants) and electronic (fluctuations of the measurements). There are many preprocessing steps that try to remove the noise, and the standard workflow include:

1. baseline correction (removing errors with systematic dependencies);
2. smoothing (removing random fluctuations);
3. peak detection/peak picking (distinguish between signals and background).

The other processing step would transform m/z ratio to mass domain. For the molecule M with z additional protons of mass p, we can have $\frac{m}{z} = \frac{M+zp}{z} = \frac{M}{z} + p$, and therefore $M = z(\frac{m}{z} - p)$. The non-trivial problem is then to predict the charge value of the molecule. In practice, quite accurate prediction can be made using the Fourier and Patterson transform (Senko et al., 1995b).

This thesis will not cover the algorithms for the preprocessing phase. However, the appropriate methods at this step are crucial to accurately retrieve the signal measured by the spectrometers. Some computational approaches to this problems are presented e.g. in Eidhammer et al. (2007) and Yang et al. (2009). It should be also noted that very often (if not always) the raw data returned by the instruments are already the output of the built-in algorithms.

Isotopic distributions of the molecules

Let us consider the molecule[11] $\xi(v, w, x, y, z)$ of a chemical formula $C_vH_wN_xO_yS_z$, i.e. composed of v carbon, w hydrogen, x nitrogen, y oxygen, and z sulphur atoms. For simplification, we will further omit the parameters v, w, x, y, z, where their presence is obvious from the context.

[11] We will not distinguish between molecules and ions.

Analogously to the elements, we can also consider isotopic variants of the molecule. Each isotopic variant has its exact mass and a probability, being a sumaric mass and a product of probabilities of occurrence of its atoms, respectively.

The lightest isotopic variant (the one composed purely from the monoisotopic atoms) of the molecule is called a monoisotopic variant. The monoisotopic variant of ξ has an exact mass:

$$M_{mono} = vM_{C_{12}} + wM_{H_1} + xM_{N_{14}} + yM_{O_{16}} + zM_{S_{32}}, \tag{1.1}$$

which is also called a monoisotopic mass of ξ, and a probability:

$$P_{mono} = P^v_{C_{12}} \times P^w_{H_1} \times P^x_{N_{14}} \times P^y_{O_{16}} \times P^z_{S_{32}}. \tag{1.2}$$

One can look at the molecule with a different level of accuracy. In a very precise approach, we can consider isotopic fine structure of ξ, where we distinguish between any two isotopic variants as long as they are composed of different number of particular isotopes [12]. For example for $\xi(1, 0, 0, 2, 0)$, a carbon dioxide CO_2, we will consider 12 fine isotopic variants, namely $^{12}C^{16}O^{16}O$ (monoisotopic variant), $^{13}C^{16}O^{16}O$, $^{12}C^{16}O^{17}O$, $^{13}C^{16}O^{17}O$, $^{12}C^{16}O^{18}O$, $^{13}C^{16}O^{18}O$, $^{12}C^{17}O^{17}O$, $^{13}C^{17}O^{17}O$, $^{12}C^{17}O^{18}O$, $^{13}C^{17}O^{18}O$, $^{12}C^{18}O^{18}O$, and $^{13}C^{18}O^{18}O$. The approaches to the problem of effective isotopic variants representation involved symbolic polynomial expansion (Yamamoto and McCloskey, 1977; Brownawell and Fillippo, 1982), and the multinomial expansion (Yergey, 1983); see also Valkenborg et al. (2012) for the review of the models. However, even for a very small molecules, the number of fine variants is quite large, and while increasing the number of atoms we can quickly fall into the problem of huge number of configurations that cannot be easily handled.

The simplification of the fine approach is to look at the aggregated isotopic variants, where we group together variants with the same number of additional neutrons[13]. For example, for $\xi(1, 0, 0, 2, 0)$, we have only 6 aggregated variants with $0, 1, \ldots, 5$ additional neutrons. Of note, the aggregated variant with 0 additional neutrons is always composed of a single fine variant, i.e. the monoisotopic one. The center-mass of aggregated variant is the average mass of all its fine variants.

The most coarse approach considers the average mass of the molecule ξ, namely:

$$\bar{M} = v\bar{M}_C + w\bar{M}_H + x\bar{M}_N + y\bar{M}_O + z\bar{M}_S, \tag{1.3}$$

where $\bar{M}_C$, $\bar{M}_H$, $\bar{M}_N$, $\bar{M}_O$, and $\bar{M}_S$ are the average masses of the corresponding elements, e.g. $\bar{M}_C = P_{C_{12}}M_{C_{12}} + P_{C_{13}}M_{C_{13}}$. However, this approximation looses a lot of characteristic information for the isotopic structure of the molecule.

[12] We do not distinguish between isoforms, where the order of isotopes matters.

[13] Additional neutrons in comparison to the monoisotopic variant of considered element or molecule.

As the isotopic variants are analyzed here in the context of mass spectra, we will also refer to them as to peaks. However, it should be noted here that the peak, when taken from the data or appropriately modeled, is a signal associated with the variant, not the variant itself.

1.3 Results for proteome analysis

Aggregated isotopic variants

Our aim in this part of the analysis is to effectively model and process isotopic distribution using the concept of aggregated variant. We also wanted to investigate the usefulness of this approach to isotopic distribution for the purpose of molecule identification. As a result, we presented the algorithm called BRAIN (Baffling Recursive Algorithm for Isotopic distributioN calculations) that is able to compute the aggregated isotope distribution for the molecule $C_vH_wN_xO_yS_z$. The algorithms makes use of two polynomial generating functions. First of these functions, Q, is defined as:

$$\begin{aligned} Q(I;v,w,x,y,z) = \left(P_{C_{12}}I^0 + P_{C_{13}}I^1\right)^v &\quad \times \\ \left(P_{H_1}I^0 + P_{H_2}I^1\right)^w &\quad \times \\ \left(P_{N_{14}}I^0 + P_{N_{15}}I^1\right)^x &\quad \times \\ \left(P_{O_{16}}I^0 + P_{O_{17}}I^1 + P_{O_{18}}I^2\right)^y &\quad \times \\ \left(P_{S_{32}}I^0 + P_{S_{33}}I^1 + P_{S_{34}}I^2 + P_{S_{36}}I^4\right)^z &\quad , \end{aligned}$$

and computes the probabilities of the variants with the same number of additional neutrons. The second function, U, is used to calculate the corresponding center-masses, and is defined with the usage of the function Q:

$$\begin{aligned} U(I;v,w,x,y,z) = \\ vQ(I;v-1,w,x,y,z)\left(P_{C_{12}}M_{C_{12}} + P_{C_{13}}M_{C_{13}}I^1\right) \\ +wQ(I;v,w-1,x,y,z)\left(P_{H_1}M_{H_1} + P_{H_2}M_{H_2}I^1\right) \\ +xQ(I;v,w,x-1,y,z)\left(P_{N_{14}}M_{N_{14}} + P_{N_{15}}M_{N_{15}}I^1\right) \\ +yQ(I;v,w,x,y-1,z)\left(P_{O_{16}}M_{O_{16}} + P_{O_{17}}M_{O_{17}}I^1 + P_{O_{18}}M_{O_{18}}I^2\right) \\ +zQ(I;v,w,x,y,z-1) \times \\ \left(P_{S_{32}}M_{S_{32}} + P_{S_{33}}M_{S_{33}}I^1 + P_{S_{34}}M_{S_{34}}I^2 + P_{S_{36}}M_{S_{36}}I^4\right) \ . \end{aligned}$$

The algorithm calculate iteratively the coefficients of both generating functions using the theory of Newton-Girard and Viète's formulas (Séroul, 2000; Vinberg, 2003).

Moreover, we implemented BRAIN as a part of R Bioconductor repository together with the stopping criteria to calculate the substantial part of the isotopic distribution, and applied it in the case study involving batch processing of a large protein dataset extracted from the Uniprot database. Namely, we build the linear model predicting the monoisotopic mass based on the corresponding most abundant center-mass. This kind of approach might be potentially useful for experimentalists, who are not able to observe monoisotopic mass for heavy ions, but would like to use it for molecule identification.

Furthermore, we introduced BRAIN 2.0., involving two improvements to decrease both time and memory complexity in obtaining the aggregated isotope distribution, and a concept to represent the element isotope distribution in a more generic manner than in original BRAIN.

Finally, we proposed an automatic procedure for discrimination between lipid and peptide signals. The bunch of random forest classifiers is able to distinguish between lipids and peptides based on the features derived from the aggregated isotopic distribution. Moreover, we propose to extend the classification for discrimination between the different lipid classes.

Fine isotopic structure

In the next step of the analyses we tried to characterize the fine structure of aggregated isotopic variants (in practice, we especially looked at the most abundant peaks). We presented a generating function based approach to calculate the variance and the information theory entropy of mass for the aggregated isotopic variants. After processing the Uniprot database, we built the linear model for the variance of the most abundant aggregated peak based on its center-mass . Further, we also estimated the spread of mass distribution and number of configurations for the aggregated variants.

Organization of the thesis, articles and co-authors

Chapter 1 is an Introduction. Its first part covers the analysis of the human genome stability. More precisely, it presents the biological background, biological and bioinformatic methods, and subsequently summarizes the results, which are described in more details in Chapters 2 and 3. The second part of Chapter 1 covers the analysis of proteome using mass spectrometry. It describes the organization of the proteome in organisms, mass spectrometry as an analytical method for proteomics study, introduces the isotopic distribution, and summarizes results from Chapters 4-6

Chapter 2 describes the potential human genome instability that can be caused by recurrent genomic inversions mediated via NAHR and its content is mostly taken from the article (Dittwald et al., 2013c). This analysis was made by PD, partially during his visit at Baylor College of Medicine (BCM) in Houston, and Dr. Tomasz Gambin from Warsaw University of Technology. PD analyzed the frequencies of large clinical CMA database (maintained by Medical Genetics Laboratories at Baylor College of Medicine in Houston and preprocessed by the group of Dr. Chad A. Shaw), retrieved the cases of known deletions/duplication associated with NAHR syndromes, and prepared data used by Dr. Tomasz Gambin in quasi-Poisson modeling. PD and Dr. Paweł Stankiewicz identified four novel recurrent NAHR-mediated deletions involving 2q12.2q13. For further classification, we designed and performed wet-lab experiments to identify breakpoint hotspots (Dr. Przemysław Szafrański), and contacted the referring physicians to obtain the clinical characteristics of the studied patients. We also used the CMA data from Signature Genomic Laboratories in Spokane, USA. Moreover, Drs. Anna Gambin, Paweł Stankiewicz, and PD developed the concept of LCR clusters, and Dr. Tomasz Gambin prepared the Miropeats diagrams showing the homology of LCR clusters flanking NAHR-prone regions. This study was also performed in a collaboration with Dr. James R. Lupski in the Department of Molecular and Human Genetics at BCM, who helped to shape the final version of the manuscript.

Chapter 3 describes the potential human genome instability that can be caused by recurrent inversions mediated via NAHR mechanism. As there are very limited numbers of clinical cases associated with these rearrangements, this study covers the automated processing of the human genome database integrated with several biological annotations (genes, phenotypic characteristics, CNVs in normal patients cohort) based on the parameters from the literature about NAHR events responsible for deletions and reciprocal duplications. The content of this chapter is mostly taken from the article (Dittwald et al., 2013b). The computational analyses of genome instability potentially mediated by IP-LCRs was done together by PD and Dr. Tomasz Gambin and supervised by Drs. Anna Gambin and Paweł Stankiewicz. The clinical context these results was mainly analyzed by Dr. Pawel Stankiewicz. In this article, we also presented the complex genomic rearrangements with a duplication-inverted triplication-duplication (DUP-TRP/INV-DUP) structures performed by the group from BCM (computational analysis made by Dr. Claudia Gonzaga-Jauregui), not included in this thesis.

Chapter 4 presents the algorithm BRAIN (Baffling Recursive Algorithm for Isotopic distributioN calculations) for calculating both the aggregated isotope distribution and corresponding center-masses. Furthermore, BRAIN is evaluated in terms of speed and precision, and compared with existing alternatives. This part is mostly taken from the article (Claesen et al., 2012). The algorithm was developed by PD and Dr. Dirk Valkenborg. The evaluation of the BRAIN (using MATLAB implementation) was performed by Dr. Jürgen Cleasen. Finally, we present also the compiled (C++) implementation (useBRAIN), which is based on Hu et al. (2013) – the C++ implementation called useBRAIN was written by Han Hu, while the performance analysis was done by PD.

Chapter 5 introduces improvements in the original BRAIN. This part was mostly taken from (Dittwald and Valkenborg, 2014), the improvements were developed by PD and Dr.

Dirk Valkenborg, and the tests were implemented by PD.

Chapter 6 presents several applications of the BRAIN algorithm: Bioconductor BRAIN package together with use case of high-throughput data processing (both implemented mainly by PD; the results are taken from the BRAIN package online documentation and the article (Dittwald et al., 2013a)), and the lipid/peptide classifier (implemented and tested by PD; real MS data provided by Vanderbildt University and preprocessed by Vu Trung Nghia) described in the prepared manuscript (not yet accepted for publication). The preliminary results of this study were also presented as a poster at ASMS conference in Vancouver in 2012.

Chapter 7 is based on results obtained mainly by PD and described in the manuscript in preparation. The preliminary results of this study were also presented as an oral presentation and poster at Polish Bioinformatics Society annual meeting in Wrocław in 2013.

Finally, Chapter 8 contains concluding remarks and further works.

The majority of Figures and Tables in this dissertation is taken from the corresponding manuscripts. Moreover, both supervisors helped in correcting and editing the thesis.

2

Genome-wide analyses of recurrent deletions and duplications

Copy-Number Variants (CNVs) involve large portion of the human genome and is responsible for various genomic disorders (Stankiewicz and Lupski, 2010; Girirajan et al., 2011). We can distinguish recurrent CNVs that re-appear in the same genomic loci, as independent events. This phenomenon can be explained by a specific structure (architecture) of the particular genomic region that predisposes some loci to *de novo* rearrangements via Nonallelic Homologous Recombination (NAHR). It has been shown (Stankiewicz and Lupski, 2002) that the NAHR-mediated rearrangement breakpoints fall within the flanking highly homologous pairs of LCRs. In this Chapter, we focus on recurrent deletions and reciprocal duplications. In particular, we consider large unbalanced events as they can be detected by aCGH technology. Moreover, they are likely causing phenotypic manifestation in patients because they usually involve larger number of genes than small CNVs. In this study we:

1. present a novel approach (based on LCR clusters) to systematically analyze the genomic regions prone to NAHR events;

2. analyze large and unique clinical CMA database and report both prevalence and *de novo* frequencies of known NAHR syndromes as well as the evidence of novel potential syndromes;

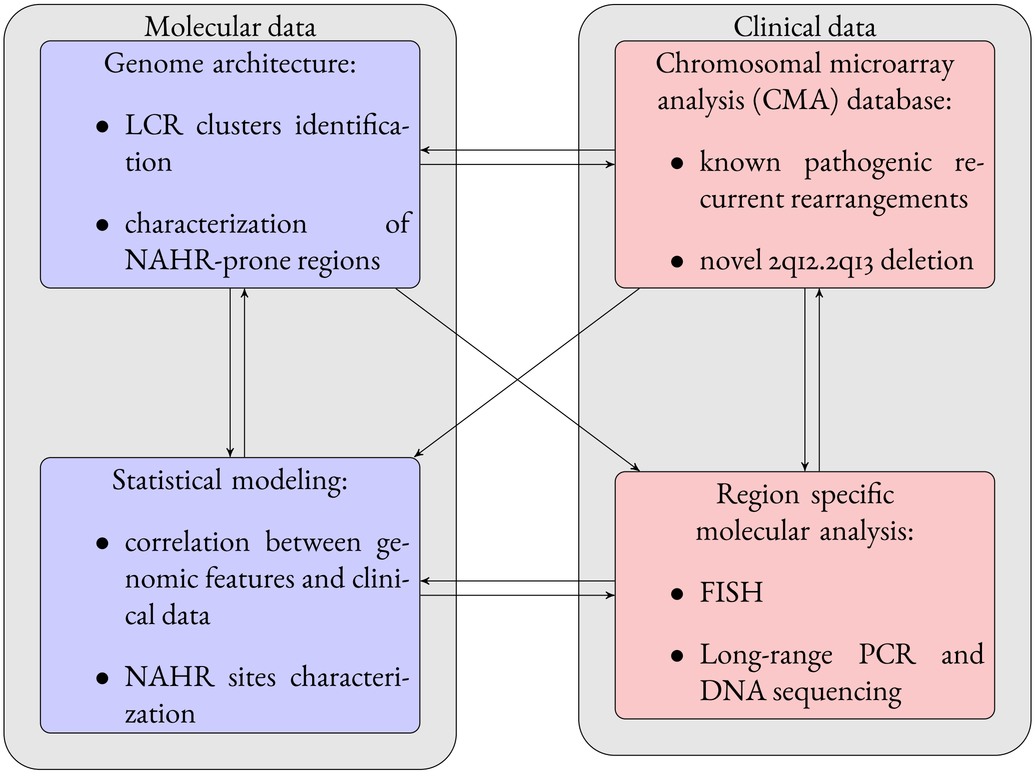
Molecular data
Genome architecture:
LCR clusters identification
characterization of NAHR-prone regions
Statistical modeling:
correlation between genomic features and clinical data
NAHR sites characterization
Clinical data
Chromosomal microarray analysis (CMA) database:
known pathogenic recurrent rearrangements
novel 2q12.2q13 deletion
Region specific molecular analysis:
FISH
Long-range PCR and DNA sequencing

Figure 2.1: A schematic workflow of the study. The violet and pink colors mark molecular and clinical data, respectively. The arrows indicate the data transfer, which was usually done using automated or semi-automated procedures. Figure courtesy: Dr Anna Gambin.

3. correlate *de novo* frequencies of different syndromes with selected genomic features to get some insights into molecular basis of the NAHR mechanism;
4. analyze statistically the genomic features related to the NAHR breakpoint regions.

Of note, these tasks are suited for intensive computational processing of large-scale data: CMA assays and human reference genome. To give a better idea about the complexity of this interdisciplinary study, we present schematically its workflow in Figure 2.1.

2.1 Previous studies

The analysis of directly oriented, highly homologous LCRs is a common approach to develop genome-wide map of NAHR-prone regions. There are two main studies that previously generated such map, and therefore serve as a good reference point to our results:

1. Sharp et al. (2006) analyzed older version of human genome build, i.e. hg16 (July 2003), and predicted 130 intervals flanked by directly oriented LCRs longer than 10 kb, with sequence identity above 95% and separated by $0.05 - 10$ Mb of intervening sequence; cf. also Sharp et al. (2005).

2. Liu et al. (2012) applied the same LCRs parameters as the previous study, but on genome build hg19, and found 608 intervals that collapsed to 89 regions.

Of note, LCR identified in hg16 and hg19 reveal several differences, which resulted in some discrepancies between the two studies. Moreover, Sharp et al. (2006) and Liu et al. (2012) used different methods to collapse the overlapping regions.

2.2 Direct Paralogous LCRs (DP-LCRs)

In our study, we decided to analyze the Segmental Duplication track (Bailey et al., 2002) available via UCSC Genome Browser for hg19. This track provides a set of LCR pairs, for elements longer than 1 kb, and homology measure (called fraction matching) between the corresponding elements above 90%. For further analyses, we chose the following subset of the directly oriented LCR pairs located on the same chromosome:

- elements longer than 8 kb – this parameter takes into account that elements shorter than those considered in (Sharp et al., 2006; Liu et al., 2012) can mediate known syndromes on Xp22.31 (STS syndrome (Hernandez-Martin et al., 1999)) and Xq28 (El-Hattab et al., 2011); on the other hand, we did not want to relax the length parameter too much, to avoid too large set of LCRs pairs;

- pairs separated by 50 kb - 10 Mb (plus length of a smaller copy) – this bounds the length of the deletions/duplications that can be caused by the considered elements and corresponds to the known recurrent NAHR syndromes sizes;

- excluding pairs that flank centromeres – restriction which eliminates CNVs that are expected to be lethal;

- fraction matching $> 95\%$ – a parameter corresponding to those used in Sharp et al. (2006) and Liu et al. (2012); this homology measure is provided by the Segmental Duplication UCSC track.

The above-defined subset of LCRs will be further referred to in this Chapter as Direct Paralogous LCRs (DP-LCRs). In total, we identified 653 DP-LCRs.

2.3 LCR clusters

Of note, LCRs that flank NAHR are often accompanied by other LCR elements. Therefore, we decided to systematically introduce a concept of LCR clusters, that can be computationally identified and adapted to the whole-genome analysis. First, we defined LCR seeds as interval on chromosome composed purely of either LCR elements or Gaps (i.e. unsequenced regions). Then, we calculated a distance between any pair of LCR seeds (denoted as `LCRseed1` and `LCRseed2`) according to the following rule:

Pseudocode chunk 2.1

```
if (getChromosome(LCRseed1) != getChromosome(LCRseed1))
return MAXVALUE ## big constant
else{
if startRegion(LCRseed1) < startRegion(LCRseed2)
return (startRegion(LCRseed2) - endRegion(LCRseed1))
else
return (startRegion(LCRseed1) - endRegion(LCRseed2))
}
```

In other words, the ”quasi distance”[1] for two seeds on the same chromosomes is defined as the distance between the closest ends of these seeds. While the seeds are located on distinct chromosomes, they are not comparable (the constant `MAXVALUE` is used).

The algorithm for hierarchical clustering of the set of seeds Ω is as follows:

Pseudocode chunk 2.2

```
S := Omega #leaves of the tree
while (|S| > 1){ #until single cluster obtained
(a, b) := findAndRemoveTwoClosest(S, d)
c := merge(a, b)
addElement(S, c)
}
return S
```

We might represent this iterative clustering as a binary tree, where each internal node refers to merging its two sons into a single cluster. The edge lengths are proportional to the distance between clusters on its adjacent nodes.

[1] We call it ”quasi distance” as the triangle inequality is not satisfied.

Single linkage distance

Single linkage "quasi distance" between two clusters A and B of elements from Ω ($A, B \subseteq \Omega$; $dist : \Omega \times \Omega \rightarrow \mathbb{R}_{\geq 0}$ is a "quasi distance" function for each two elements of Ω) is defined as:

$$d_{SL}(A, B) := \min\{dist(a, b) | a \in A, b \in B\}$$

To obtain the clustering tree, we used the hierarchical clustering algorithm that starts from LCR seeds (as leaves) and merge clusters according to the single linkage distance (d_{SL}). Clusters were extracted by pruning this tree on the given threshold (schematically, this is represented in Figure 1.1). We decided to prune the clustering tree on a height corresponding to a set of $3,000$ LCR clusters. It should be emphasized, that the pruning threshold can be modified, or the clustering tree might be even pruned on various heights for different chromosome regions.

The NAHR-prone regions are defined as a set of genomic regions flanked by LCR clusters, considered as intervals C_α, C_β, and there exist a pair of DP-LCR elements (also processed as intervals) α, β such that $\alpha \subseteq C_\alpha$ and $\beta \subseteq C_\beta$. Of note, this definition allows for $C_\alpha = C_\beta$, i.e. NAHR-prone region can map within single LCR cluster containing a pair of DP-LCR elements. We observe such situation in case of the 12q14.2 region associated with Globozoospermia (MIM# 613958; (Elinati et al., 2012).

As a result of our analyses, we identified 198 NAHR-prone regions (full coordinates are available as a Supplemental Table S1 in Dittwald et al. (2013c)): 105 flanked by two distinct LCR clusters, and 93 composed of a single LCR cluster. These regions are graphically compared with previous approaches (in case of Sharp et al. (2006) we used only 92 out of 130 regions which successfully lifted over to the hg19 coordinates) in Figure 2.2. Figure 2.3 is a good example to appreciate that the LCR clustering approach allows for different configurations between the neighboring regions. Practically, we were able to distinguish between:

- regions that overlap (e.g. ids 61 and 65);
- one region that is a subset of another without sharing common LCR cluster (e.g. ids 69 and 66);
- one region that is a subset of another with sharing common LCR cluster (ids 65 and 66);
- two adjacent region sharing a common LCR cluster (ids 61 and 66).

To date, approximately 40 distinct (i.e. non-overlapping) loci on both autosomes and chromosome X associated with clinical syndromes have been classified (Lupski, 1998, 2009; Mef-

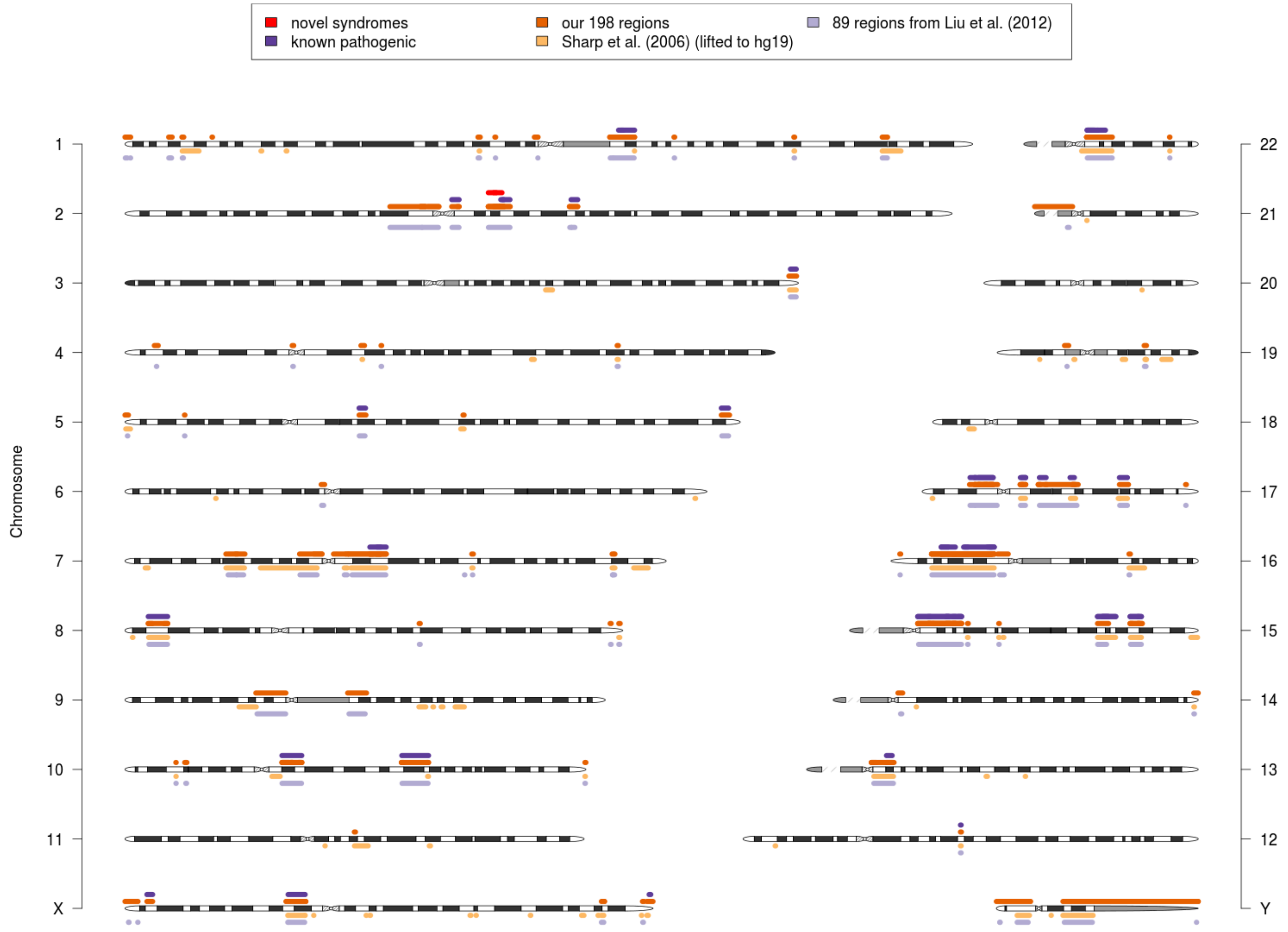

Figure 2.2: Ideogram for NAHR-prone regions according to Sharp et al. (2006) (yellow; only regions lifted to hg19), Liu et al. (2012) (light blue), and our method (orange). In addition we indicated the 2q12.3q13 region with candidates for novel syndrome (red), and known pathogenic genomic disorders caused by NAHR-mediated deletions and/or duplications (dark blue). Source: Dittwald et al. (2013c).

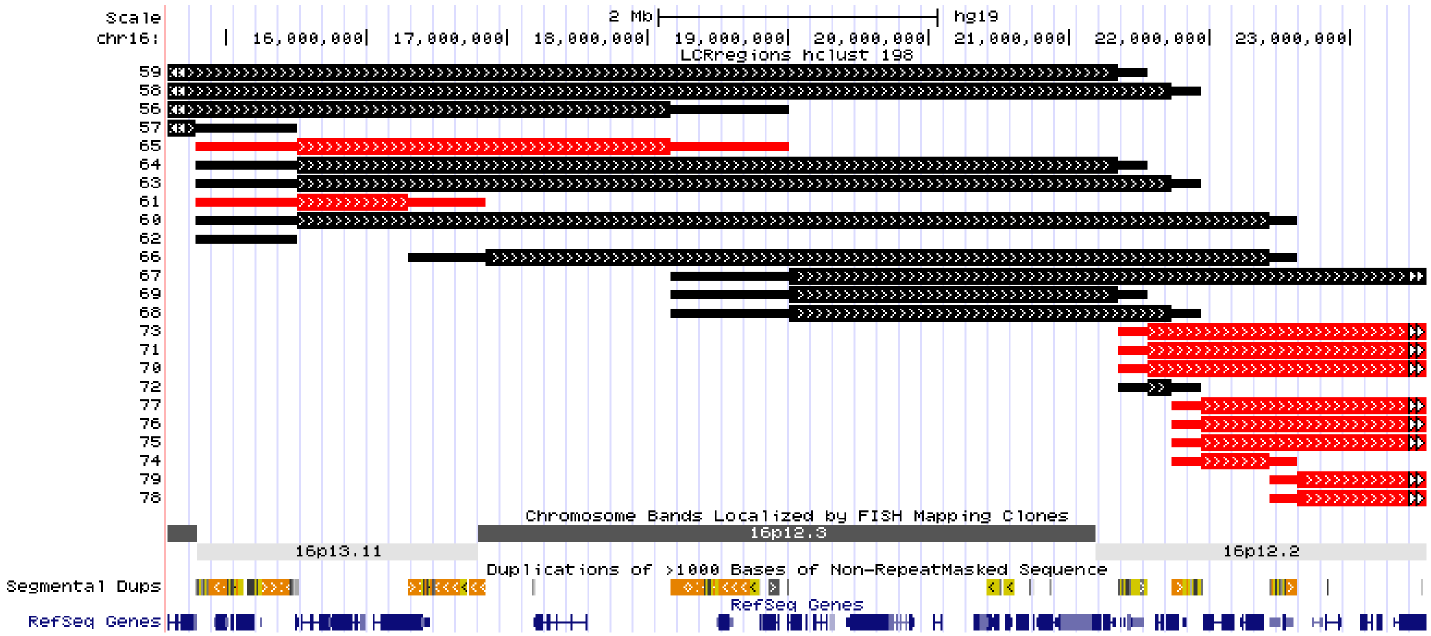

Figure 2.3: A subset of NAHR-prone regions identified by our method on $\sim$ 7 Mb long region spanning chromosome 16. The thin bars mark LCR clusters, connected by an intervening sequence (the wider bars). Red colors indicate already known NAHR-mediated syndromes. Source: Dittwald et al. (2013c).

ford et al., 2009; Vissers and Stankiewicz, 2012). Of note, these syndromes were associated with 53 NAHR-prone regions identified by our approach. Our method allows for a better classification of the selected similar regions that cause different phenotypes. For example, Thrombocytopenia-Absent Radius syndrome (TAR) region on 1q21 (Klopocki et al., 2007; Albers et al., 2012), and the 1q21.1 deletion/duplication syndrome region (Mefford et al., 2008; Brunetti-Pierri et al., 2008) were previously considered together, and now can be distinguished by our approach. On the other hand, we did not detect small *CHRNA7* deletion/duplication in 15q13.3 and 17q21.31 deletion/duplication region (Sharp et al., 2006) due to the fact that there were identified for the haplotypes that differ from the reference genome, and several variants 15q24 deletion syndromes for which the flanking LCRs reveal fraction matching smaller than 95%.

The convincing application of our method is to use it for detection of the new potential syndromes. Therefore, in the remaining (i.e. not associated with known syndromes) set of NAHR-prone regions, we analyzed the clinical cases. For the 2q12.2q13 locus (considered as a single NAHR-prone region by Liu et al. (2012)), we identified four adjacent and/or overlapping intervals, for which we found an evidence for large (between ∼0.6 and ∼1.9 Mb) deletions (Figure 2.4). In this step, we first queried the clinical CMA database. To identify more cases, we also used data provided by Signature Genomics. For two regions (2q12.2q12.3 and 2q12.3q13), the NAHR events were confirmed using long range PCR experiments. More details about the patients phenotypes (according to the reports sent by physicians) and long-range PCR/DNA sequencing experiments can be found in the online Supplementary Materials from Dittwald et al. (2013c). The characterized region reveals high homology between all LCR clusters, which is depicted using the Miropeats graphics (Figure 2.5).

2.4 Known deletion and duplication syndromes

In our study, we aimed to identify the known deletion/duplication syndrome regions in the BCM CMA database of over $25,000$ patients (diagnoses for these patients are depicted in Figure 2.6). To this aim, we manually collected the data about the LCR clusters flanking the syndromes (these data are provided as a Supplemental Table S2 in (Dittwald et al., 2013c)). Moreover, we designed the automatic workflow to filter out the rearrangements that overlap with the regions of interests (i.e. chromosome Y was not taken into consideration in further analyses). In particular, we used Bioconductor IRanges library efficiently operating on the intervals in order to detect the regions that are close to the syndromes loci. The CMA data were first preprocessed by the BCM bioinformatic laboratory (headed by Dr Chad A. Shaw), and provided as a region narrowed down to minimal/maximal start and stop coordinates. Then, we have assigned $2,129$ CNVs ($1,053$ deletions and $1,076$ duplications) to syndromes (we manually curated this step to get better association). This information, de-

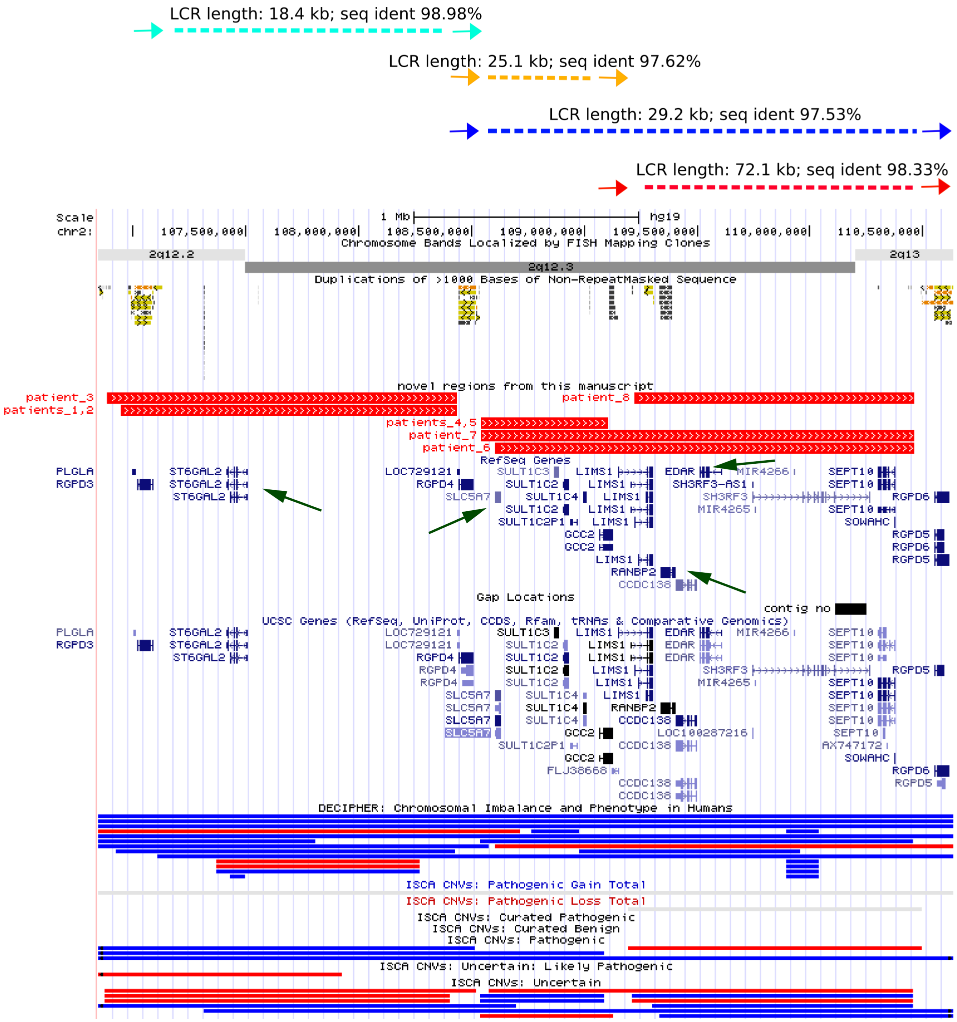
LCR length: 18.4 kb; seq ident 98.98%
LCR length: 25.1 kb; seq ident 97.62%
LCR length: 29.2 kb; seq ident 97.53%
LCR length: 72.1 kb; seq ident 98.33%
Scale
chr2:
1 Mb
hg19
107,500,000
108,000,000
108,500,000
109,000,000
109,500,000
110,000,000
110,500,000
Chromosome Bands Localized by FISH Mapping Clones
2q12.2
2q12.3
2q13
Duplications of >1000 Bases of Non-RepeatMasked Sequence
novel regions from this manuscript
patient_3
patients_1,2
patient_8
patients_4,5
patient_7
patient_6
RefSeq Genes
PLGLA
RGPD3
ST6GAL2
LOC729121
RGPD4
SLC5A7
SULT1C3
SULT1C2
SULT1C4
SULT1C2P1
GCC2
LIMS1
RANBP2
CCDC138
EDAR
SH3RF3-AS1
SH3RF3
MIR4266
MIR4265
SEPT10
SOWAHC
RGPD6
RGPD5
Gap Locations
contig no
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs &amp; Comparative Genomics)
LOC100287216
AX747172
FLJ38668
DECIPHER: Chromosomal Imbalance and Phenotype in Humans
ISCA CNVs: Pathogenic Gain Total
ISCA CNVs: Pathogenic Loss Total
ISCA CNVs: Curated Pathogenic
ISCA CNVs: Curated Benign
ISCA CNVs: Pathogenic
ISCA CNVs: Uncertain: Likely Pathogenic
ISCA CNVs: Uncertain

Figure 2.4: The 2q12.2q13 region where four potential novel rearrangements have been identified (the corresponding DP-LCR are represented as arrows on the top). The red bars in the middle indicate the events found in the clinical databases (BCM and Signature Genomics). Red and blue thin bars at the bottom indicates available entries from the DECIPHER and ISCA databases. The interesting genes harboring the regions of rearrangements (*ST6GAL2*, *SLC5A7*, *EDAR*, *RANBP2*) are indicated by green arrows. Source: Dittwald et al. (2013c).

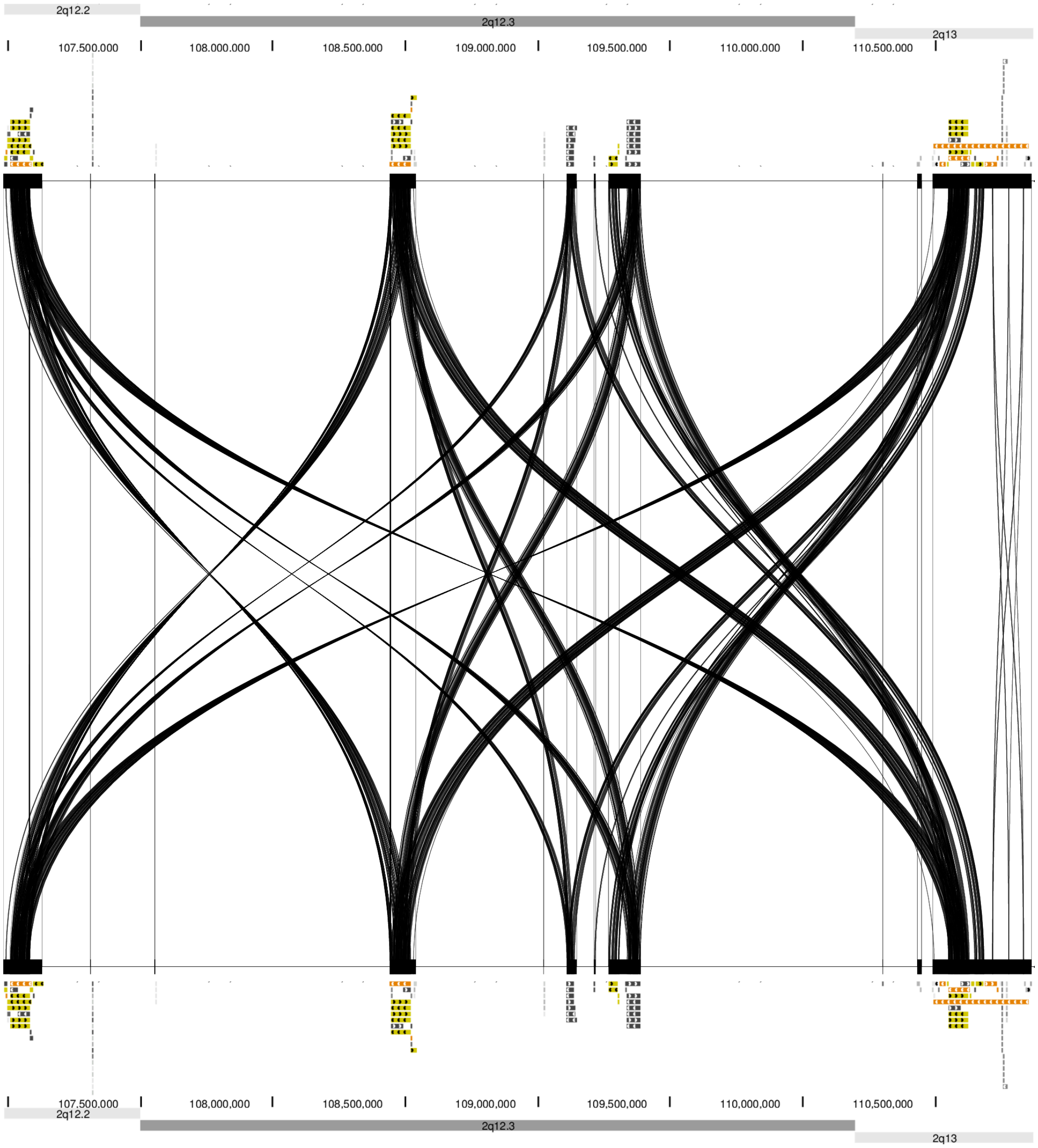

Figure 2.5: Schematic representation of the homology within the 2q12.2q13 region (same interval shown at the top and the bottom of the figure). Visualization is made utilizing ICAass algorithm (v.2.5) and the Miropeats program (v.2.01) (Parsons, 1995). Figure source: Dittwald et al. (2013c).

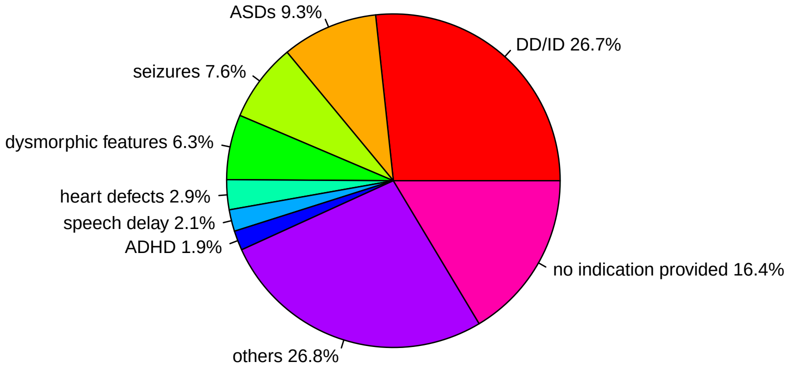

Figure 2.6: The pie chart with the main diagnoses on the patients from BCM CMA database. Abbreviations used: DD/ID – developmental delay/intellectual disability, ASDs – autism spectrum disorders, ADHD – attention deficit hyperactivity disorder. Data from Dittwald et al. (2013c) (initially provided by Ian M. Campbell).

picted as a histogram in Figure 2.7, describes the prevalence of the syndromes in our database (which is not the same as prevalence in the population as our database is biased towards abnormal phenotypic manifestation; cf. also Figure 2.6). The tree most common recurrent rearrangements observed were *NPHP1* duplications (233 cases), *CHRNA7* duplications (175 cases), and 22q11.21 deletions (DGS/VCFS common, 166 cases). We have extracted the inheritance information associated with samples (available for only $\sim 25\%$ of the analyzed CNVs), and each rearrangement was then characterized as *de novo* (190 CNVs), inherited (355 CNVs), or of unknown origin (1,584 CNVs). This information was obtained by additional FISH analyses in the parents (same method was used to confirm CNVs in patients). Then, we restricted our analysis to *de novo* vs. inherited cases in order to get more insights about the frequency of new events in our patients' cohort (cf. Figure 2.8). Of note, the most frequent *de novo* rearrangements were deletions: 22q11.21 (DGS/VCFS common), 16p11.2 (593 kb), and 7q11.23 Williams-Beuren syndrome (WBS).

2.5 Genes that are prone to cause abnormal phenotypes when deleted or duplicated

We analyzed the RefSeq genes extracted from the USCS Genome Browser – $2,145$ of them overlapped with the genomic regions flanked by DP-LCRs. We identified a subset of 39 genes reported to be dosage-sensitive, as increase or decrease of their expression may cause phenotypic manifestation (Huang et al., 2010). In addition, we queried OMIM database using OMIM API, and found 232 genes with associated diseases (all identified genes are presented as Supplementary Table S3 in Dittwald et al. (2013c)).

2.6 Statistical modeling involving genomic features potentially responsible for the NAHR rearrangements

As we collected the information about *de novo* NAHR frequencies, we used these data to link them with various genomic features that can predispose to genome instability. The previous study (Liu et al., 2012), limited to deletions in 17p11.2 region (Smith-Magenis syndrome; SMS), suggested that there is a correlation between rearrangement frequencies and the high percent of sequence identity between flanking direct paralogous LCRs. Here, we aimed to perform more systematic, genome-wide analyses. As our study introduces the concept of automatically derived LCR clusters, we considered separately genomic features characterizing DP-LCR and structural features associated with LCR clusters. Thus, for a subset of deletions that were likely to be caused by *de novo* NAHR events associated with known syndromes, two classes of NAHR hot spots were considered:

- ”active” – with identified *de novo* events;

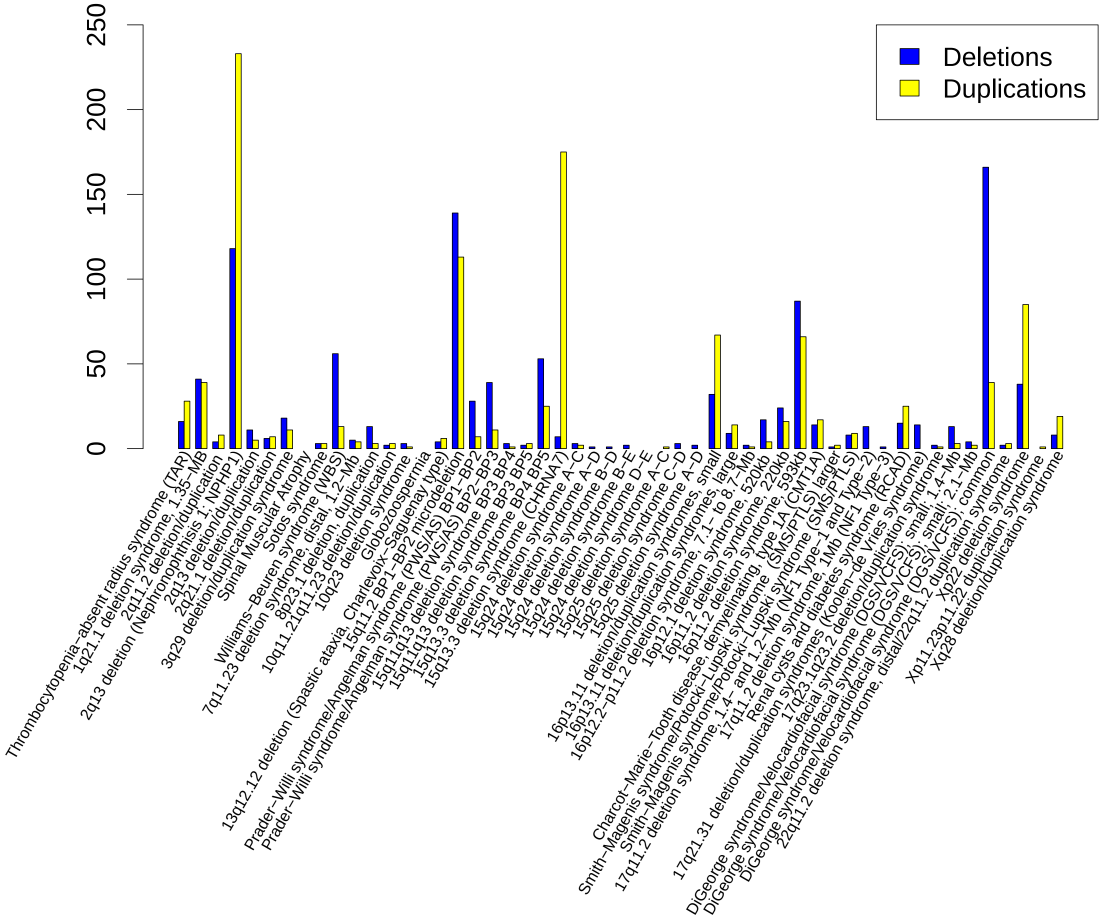

Figure 2.7: The prevalence (i.e. both inherited and *de novo* events are considered) of known syndromes associated with NAHR-mediated deletions and duplications among patients in BCM CMA database. Source: (Dittwald et al., 2013c).

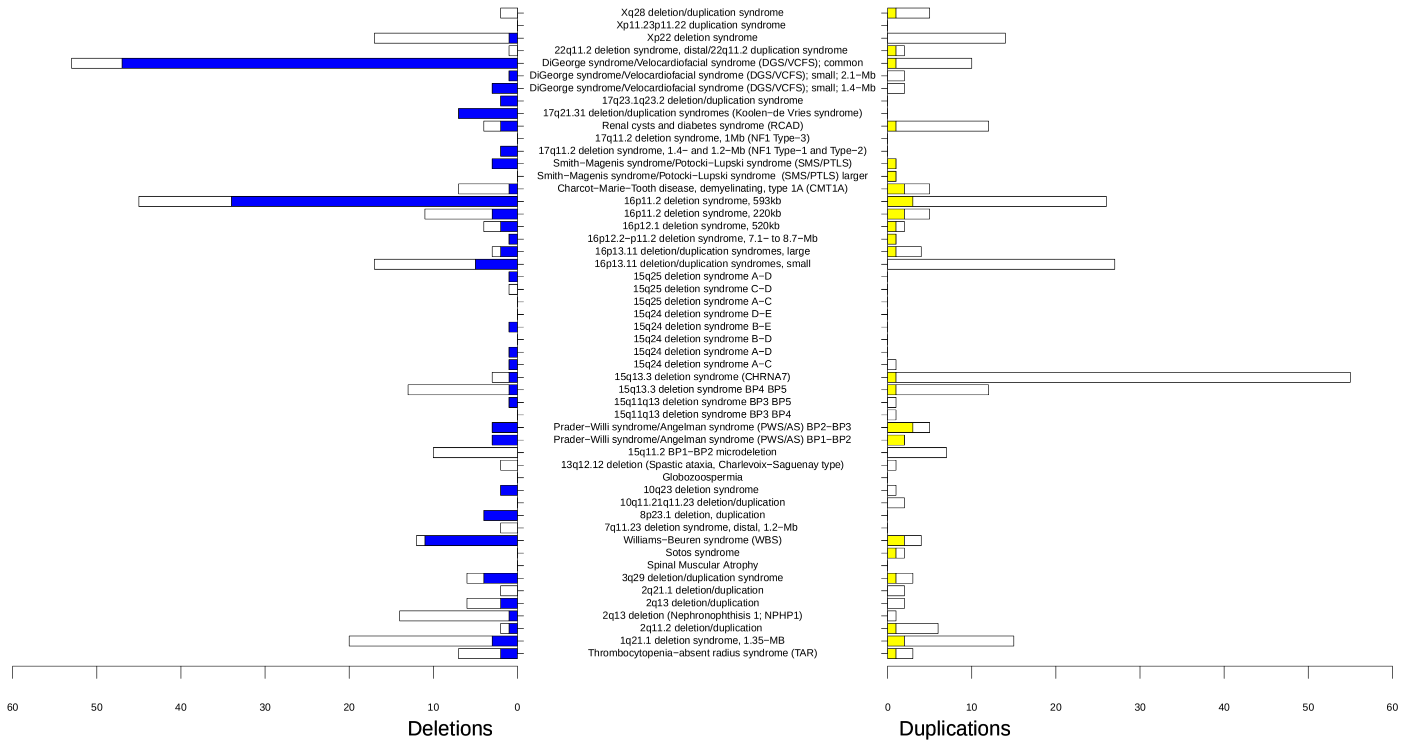

Figure 2.8: Frequencies of *de novo* events for known syndromes associated with NAHR-mediated deletions and duplications among patients in BCM CMA database. Source: (Dittwald et al., 2013c).

- ”inactive” – for the remaining set of regions.

A set of nonparametric Mann-Whitney-Wilcoxon tests has been made to explore the differences in genomic features between ”active” and ”inactive” hot spots.

Mann-Whitney-Wilcoxon test

For two samples A and B, from distributions X and Y, respectively, Mann-Whitney-Wilcoxon test validates if X and Y are statistically equal. More precisely, the null hypothesis H_0 says that $P(X > Y) = P(X < Y)$. The test uses the statistic U that measures the number of pairs (x, y), $x \in X, y \in Y$ such that $x \geq y$.

As a result, the statistically significant differences were noted for several genomic features, reported in Tables 2.1-2.2 (columns 2 and 3). In particular, ”active” hot spots reveal increased GC content and the increased density of the 13-mer motif (5’-CCNCCNTNNCCNC-3’) associated in (Myers et al., 2008) with recombination hot spots. Then, for DP-LCRs that flank more than two recurrent NAHR events, the Spearman rank correlation has been calculated.

Spearman rank correlation

Spearman rank correlation ρ between two variables X and Y of size n is a nonparametric measure of their dependence. For ranked (in case of ties, the mean rank value is used) values $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ obtained from original values drawn from X and Y, respectively, the following formula is used:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ states for mean value of vector x. The coefficient ρ has values from the interval $[-1; 1]$; for $\rho = \pm 1$ the perfect dependence is obtained, while for $\rho = 0$ two variables are assumed to be independent. The approximation of p-values can be done e.g. using the Student’s t distribution.

The strongest correlation tested to be statistically significant (p-value < 0.05) for DP-LCRs features was detected for the distance between DP-LCR elements (the negative correlation) and sequence identity. In addition, the homology length reveals strong correlation, which is, however, characterized by a low statistical significance (p-value ≈ 0.168). The more detailed results are provided in Tables 2.1-2.2 (column 4). For the following features of DP-LCRs and LCR clusters, the strongest correlation has been detected: maximal LCR homology, GC content and a maximal number of the 5’-CCNCCNTNNCCNC-3’ motifs.

Table 2.1: Analysis of the correlation between NAHR-mediated *de novo* deletions and various genomic features associated with DP-LCRs. Source: Dittwald et al. (2013c).

Feature of DP-LCRs	Comparison of DP-LCRs flanking active NAHR hot spots vs. DP-LCRs flanking inactive cold spots (P-values from Mann-Whitney-Wilcoxon test)		Correlation/regression of DP-LCRs feature and frequency of *de novo* deletions; DP-LCRs flanking reliable recurrent changes, i.e., genomic regions for which we detected at least three recurrent *de novo* deletions, were considered	
	Feature is greater in DP-LCRs flanking active NAHR hot spots, i.e., regions for which we detected at least one *de novo* deletion	Feature is greater in DP-LCRs flanking inactive NAHR cold spots, i.e., regions for which we did not detect any *de novo* deletion	Spearman rank correlation coefficients and P-values	Poisson regression coefficients and P-values
Length of homology of paralogous DP-LCRs	$(P = 1.86 \times 10^{-1})$		$0.29\ (P = 1.68 \times 10^{-1})$	
Distance between paralogous DP-LCRs	$(P = 1.18 \times 10^{-3})$		$-0.69\ (P = 2.19 \times 10^{-4})$	
Length of homology divided by distance between paralogous DP-LCRs		$(P = 7.64 \times 10^{-3})$	$0.6\ (P = 2.3 \times 10^{-3})$	$43.7\ (P = 1.08 \times 10^{-3})$
Fraction matching (percent identity) of paralogous DP-LCRs	$(P = 2.69 \times 10^{-1})$		$0.73\ (P = 8.18 \times 10^{-5})$	$29.72\ (P = 1.51 \times 10^{-2})$
Mean GC content of paralogous DP-LCRs	$(P = 7.53 \times 10^{-6})$		$-0.02\ (P = 9.05 \times 10^{-1})$	
Number of occurrences of the 13-mer recombination motif in the pair of DP-LCRs combined	$(P = 7.06 \times 10^{-5})$		$0.33\ (P = 1.17 \times 10^{-1})$	
Average density of the 13-mer recombination motif in the pair of DP-LCRs combined	$(P = 2.57 \times 10^{-6})$		$0.04\ (P = 8.55 \times 10^{-1})$	

Table 2.2: Analysis of the correlation between NAHR-mediated *de novo* deletions and various genomic features associated with LCR clusters. Source: Dittwald et al. (2013c).

Feature of DP-LCRs	Comparison of DP-LCRs flanking active NAHR hot spots vs. LCR clusters flanking inactive cold spots (P-values from Mann-Whitney-Wilcoxon test)		Correlation/regression of LCR cluster's feature and frequency of de novo deletions; clusters flanking reliable recurrent changes, i.e., genomic regions for which we detected at least three recurrent *de novo* deletions, were considered	
	Feature is greater in LCR clusters flanking active NAHR hot spots	Feature is greater in LCR clusters flanking inactive NAHR cold spots	Spearman rank correlation coefficients and P-values	Poisson regression coefficients and P-values
GC content within the cluster	(P = 1.11×10^{-4})		0.54 (P = 7.04×10^{-3})	27.1 (P = 1.34×10^{-25})
Maximum length of homology among LCRs within the cluster	(P = 1.41×10^{-1})		0.41 (P = 4.62×10^{-2})	1.4×10^{-5} (P = 5.43×10^{-11})
Total number of occurrences of the 13-mer recombination hot spot motif in the cluster		(P = 2.7×10^{-1})	0.51 (P = 1.17×10^{-2})	
Median number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster	(P = 7.1×10^{-2})		0.48 (P = 2.01×10^{-2})	0.45 (P = 3.5×10^{-3})
Third quartile of the number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster	(P = 1.07×10^{-1})		0.42 (P = 4.42×10^{-2})	-0.46 (P = 3.71×10^{-5})
Maximum number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster	(P = 3.81×10^{-5})		0.54 (P = 6.79×10^{-3})	
Total density of the 13-mer recombination hot spot motif in the cluster	(P = 1.96×10^{-3})		0.38 (P = 6.66×10^{-2})	
Third quartile of the density of the 13-mer recombination hot spot motif among LCRs within the cluster	(P = 2.72×10^{-2})		-0.05 (P = 7.9×10^{-1})	
Maximum density of the 13-mer recombination hot spot motif among LCRs within the cluster	(P = 7.85×10^{-5})		0.19 (P = 3.73×10^{-1})	

The next step was to build the regression model to explain the frequency information based on a set of genomic features. According to McElduff et al. (2010) the proper approach for count data analysis is a Poisson regression model.

Poisson regression

In Poisson regression, the response variable Y is assumed to have Poisson distribution, i.e.

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!},$$

where μ is a parameter. The second assumption is that the logarithm of expected value of Y (in our case $E(Y) = \mu$) can be modeled by a linear combination of parameters from vector X. Namely, for $Z = \log(E(Y|X))$, $Z = \beta X + \beta_0$.

The features that occurred to be statistically significant were presented in column 5 of Tables 2.1-2.2.

For the next analysis, we gathered the literature data about known NAHR recombination sites (presented as Supplementary Table S5 in (Dittwald et al., 2013c)). Finally, the closer investigation of the distribution of the 5'-CCNCCNTNNCCNC-3' motif revealed its enrichment in the distance up to 2 kb from breakpoints in contrast to other randomly selected 13-mer motifs.

3
Genome-wide analyses of NAHR-mediated inversions

In contrast to the several phenotypic NAHR-mediated deletions and duplications, only two recurrent inversions have been associated with clinical syndromes:

- hemophilia A (factor VIII deficiency; MIM #306700), where over 45% cases are associated with inversion disrupting the *F8* gene (Lakich et al., 1993; Naylor et al., 1992, 1993, 1995).
- mucopolysaccharidosis type II (Hunter syndrome; MIM #309900) – in this case a balanced inversion harbors the IDS gene (Bondeson et al., 1995).

On note, both aforementioned diseases are X-linked and map to Xq28 region. The low number of syndromes examples does not necessarily mean that NAHR-mediated inversions are rare, and can be explained by other reasons:

1. The balanced rearrangements are much more difficult to be detected, e.g. cannot be identified by aCGH assays.
2. Genes are mostly disrupted by inversion's breakpoints, in contrast to e.g. deletions and duplications, where the whole region between breakpoints is imbalanced. This causes that even long rearrangement usually causes not so severe phenotypic manifestation as in cases of deletions and duplications of the same size.

3.1 Genome instability via recurrent inversions

The project described in this Chapter aimed to investigate the potential genome instability caused by NAHR-mediated inversions. First, we identified the set of inversely oriented, paralogous LCRs (IP-LCRs) that are likely to mediate such events. Namely, we considered the UCSC Genome Browser Segmental Duplications track (Bailey et al., 2002) (genome build hg19) for the following parameters:

- minimal length of the LCR element over 1 kb (which was a limitation of the analyzed track);

- LCR elements separated by less than 10 Mb to exclude too long rearrangements;

- fraction matching above 95% – according to the parameter used for directly oriented LCRs (see Chapter 2).

As a result, we found $1,337$ pairs of such IP-LCRs (Figure 3.1). We also analyzed other fraction matching limitations: relaxed ($> 90\%$), and more stringent ($> 97\%$), which revealed $2,805$ and 915 pairs of opposite orientation LCRs, respectively (the term IP-LCRs is used in this Chapter only for dataset generated using fraction matching $> 95\%$). The set of IP-LCRs harbors in total 372.6 Mb, i.e. approximately 12% of the human genome, and in particular involves 43% of chromosome 17. This is a portion of the human genome that can be potentially altered by NAHR-mediated inversions utilizing IP-LCRs. The DNA covered by IP-LCRs elements is much shorter, i.e. 59 Mb (1.93% of the human genome, including over 11% of chromosome Y) – cf. Figure 3.2.

3.2 Analysis of the Database of Genomic Variants

As stated above, there are only two known syndromes associated with recurrent pathogenic inversions. Therefore, we analyzed the inversion from the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/)) – repository with data from healthy individuals (Zhang et al., 2006). In DGV, we found 587 inversions > 10 kb. In this set we were searching for events with two breakpoints within the corresponding IP-LCR elements. Specifically, we queried for inversions characterized by intervals R_α, R_β and a pair of IP-LCRs elements (also treated as intervals) α, β such that $\alpha \cap R_\alpha \neq \emptyset$ and $\beta \cap R_\beta \neq \emptyset$. As a result, we identified 47 such inversions (Figure 3.1), and tested their statistical significance by applying the following procedure.

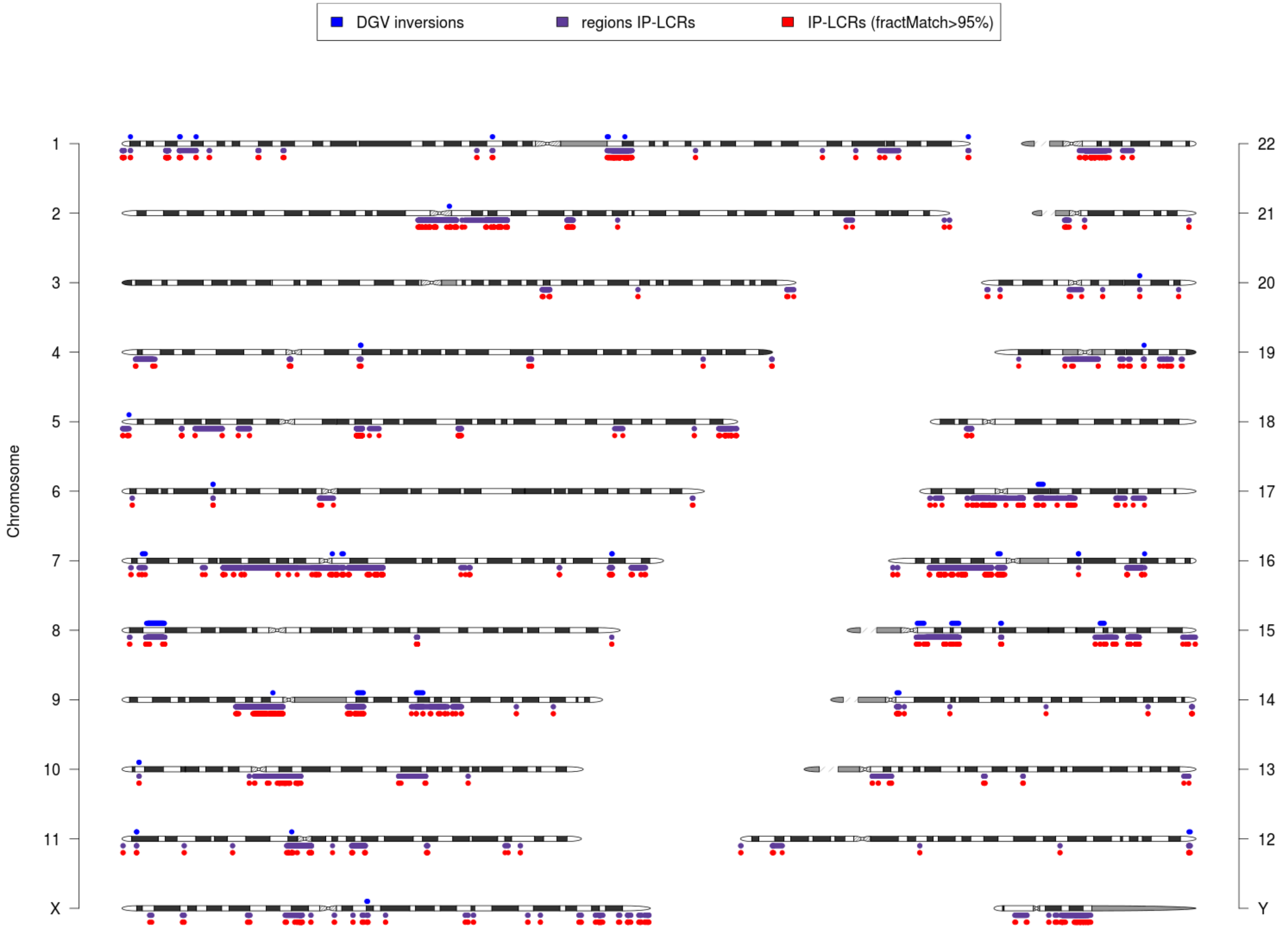

Figure 3.1: Ideogram of IP-LCR elements (red), and regions flanked by IP-LCRs (purple). In addition 47 inversions with breakpoints within IP-LCRs are marked (blue, above chromosomes). Source: Dittwald et al. (2013b).

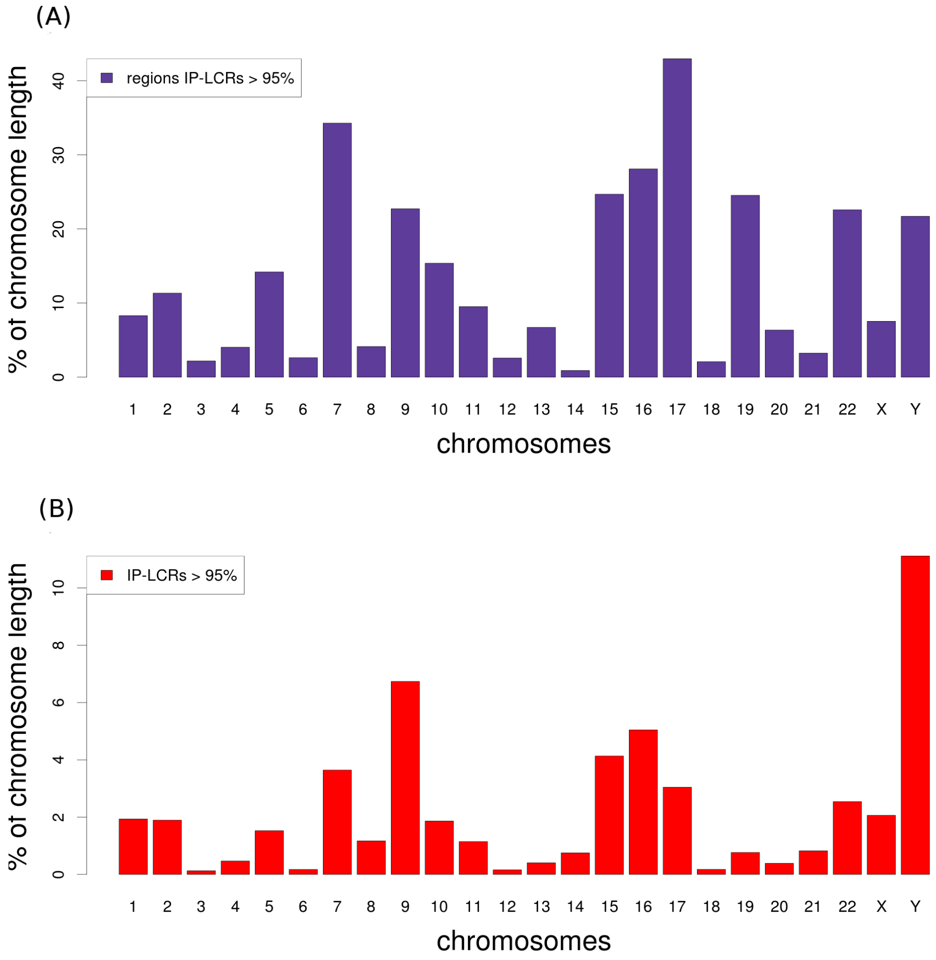

Figure 3.2: Portions of chromosomes covered by regions flanked by IP-LCRs (A), and IP-LCR elements themselves (B). Source: Dittwald et al. (2013b).

Pseudocode chunk 3.1

```
for inv in inversions{
  l <- length(inv)
  chr <- chromosomeOf(inv)
  seqs <- drawRandomSequences(l, hm, chr)
        ##draws hm sequences with two breakpoints outside gaps
  count <- 0
  for seq in seqs
        if (bothEndsInIPLCRs(seq))
            count <- count + 1
  pValue[inv] <- count/hm
}
return pValue
```

Namely, we estimated the probability (p-values) that the randomly generated sequence of a given length lying on the same chromosome as investigated inversion has both breakpoints within the corresponding IP-LCR elements. All estimated p-values are below 0.05 (each p-value was estimated independently, no correction has been applied).

3.3 Genes potentially disrupted by recurrent inversions

Here, we investigated genes where NAHR-mediated inversion breakpoints can be located. Using RefSeq genes from the track in UCSC Genome Browser, we identified 942 (99 X-linked) genes that intersect with at least one IP-LCR element; eight of them *ABCC6*, *FKBP6*, *GTF2I*, v*NCF1*, *PRODH*, *RTN4R*, *STAT5A*, and *STAT5B* are known as dosage-sensitive genes (Huang et al., 2010). Moreover, our research provided also the detailed characteristics of 31 genes that are known to be associated with diseases (cf. Table 3.1), and can serve clinicians for the diagnostic purposes. We also investigated the phenotypes associated with the found genes. For this purpose, we queried the Genetic Association Database (GAD) (Zhang et al., 2010), which is ”an archive of human genetic association studies of complex diseases and disorders” (according to the official webpage of the project `http://geneticassociationdb.nih.gov/`), and identified various disease classes as presented in Figure 3.3 (no class is dominating).

Table 3.1: A table with 31 genes potentially disrupted by recurrent inversions, which are already known to be associated with diseases. Source: Dittwald et al. (2013b).

Gene	Gene description	Location	Intersection with LCR Size (Size of entire LCR) kb	LCR identity %	Disease	Inheritance	OMIM
ABCC6	ATP-binding cassette, sub-family C (CFTR/MRP), member 6	16p13.11	25 (128)	99.36	Pseudoxanthoma elasticum	AR	264800
AKR1C2	Aldo-keto reductase family 1, member C2	10p15.1	28 (47.5)	95.15	46,XY sex reversal 8	AR	614279
BCR	Breakpoint cluster region	22q11.23	7(10.5), 4(7)	95.98; 96.17	Chronic myeloid leukemia (CML)	-	608232
CFC1	Cripto, FRL-1, cryptic family 1	2q21.1	7 (227); 7 (229)	99.27; 99.27	Visceral heterotaxy-2 (HTX2); (a congenital heart disease; identified in patients with transposition of the great arteries and double-outlet right ventricle)	AD	605376
CHRNA7	Cholinergic receptor, nicotinic, alpha 7 (neuronal)	15q13.3	17 (307)	99.62	Chromosome 15q13.3 deletion syndrome	AD	612001
CNTNAP3	Contactin associated protein-like 3	9p13.1	5(208); 55(115); 130(155); 64(64); 22(49)	99.29; 98.72; 98.49; 98.3; 98.2	Candidate gene for bipolar disorder and bladder exstrophy	?	N/A
DPP6	Dipeptidyl-peptidase 6	7q36.2	105(105); 110(110)	98.4; 98.42	Paroxysmal familial ventricular fibrillation 2 (VF2)	AD	612956
DUOX2	Dual oxidase 2	15q21.1	1 (1)	97.46	Congenital hypothyroidism, Thyroid Dyshormonogenesis 6 (TDH6)	AR	607200
FANCC	Fanconi anemia, complementation group C	9q22.32	3(3)	95.98	Fanconi anemia, complementation group C	AR	227645
FLNC	Filamin C	7q32.1	3(3)	96.4	Myofibrillar myopathy, Distal myopathy 4	AD	609524; 614065
GTF2I	General transcription factor IIi	7q11.23	33 (144)	99.67	Williams-Beuren syndrome critical region, responsible for autism spectrum disorders	AD	194050
HERC2	HECT and RLD domain containing E3 ubiquitin protein ligase 2	15q13.1	4(4); 47(47); 1(1); 34(34); 6(103)	95.04; 97.31; 96.12; 97.07; 99.61	Juvenile development and fertility 2 (jdf2), skin/hair/eye pigmentation	AD?	227220
KRT81 and KRT86	Keratin 81 and keratin 86	12q13.13	4(4)	97.72	Monilethrix	AD	158000
NCF1	Neutrophil cytosolic factor 1	7q11.23	15.3(144)	99.67	Chronic granulomatous disease	AR	233700
NQO2	NRH:quinone oxidoreductase-2	6p25.2	2(2)	96.95	Breast cancer	-	114480
OCLN	Occludin	5q13.2	24(79)	99.67	Band-like calcification with simplified gyration and polymicrogyria (BLCPMG)	AR	251290
PLEKHM1	Pleckstrin homology domain-containing protein, family M, member 1	17q21.31	3(3)	95.79	Osteopetrosis, autosomal recessive 6	AR	611497
PRODH	Proline dehydrogenase	22q11.21	12(23); 2(2)	95.83; 96.37	Hyperprolinemia type 1; Schizophrenia	AD	239500; 600850
RANBP2	RAN binding protein 2	2q12.3	52(52); 52(52); 52(52)	97.59; 97.62; 97.67	Acute necrotizing encephalopathy (ANE1)	AD	608033
RHCE and RHD	Rh blood group, CcEe antigens	1p36.11	58(63) and 57(61)	98.07	RH-null disease	AD	268150
RTN4R	Reticulon 4 receptor	22q11.21	6(28)	95.84	Susceptibility to schizophrenia	AD	181500
SBDS	Shwachman-Bodian-Diamond syndrome	7q11.21	8(46)	96.77	Shwachman-Bodian-Diamond syndrome; Paragangliomas 5	AR	260400
SDHA	Succinate dehydrogenase,	5p15.33	0.5(5); 21(24)	96.18; 95.65	Leigh syndrome	AR	256000
SORD	Sorbitol dehydrogenase	15q21.1	4(4); 18(18); 17(25)	95.07; 97.31; 97.86	Deficiency in a family with congenital cataracts	N/A	N/A
SPECC1L	Sperm antigen with calponin homology and coiled-coil domains 1-like	22q11.23	5(5); 5(5)	97.06; 96.91	Oblique facial clefting-1 (OBLFC1)	AD	600251
SPTLC1	Serine palmitoyltransferase, long-chain base unit 1	9q22.31	11(11)	96.68	Neuropathy, hereditary sensory and autonomic, type 1, severe	AD	162400
STAT5B	Signal transducer and activator of transcription 5B	17q21.2	4(4)	97.4	Growth hormone insensitivity with immunodeficiency	AR	245590
TAF1	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa)	Xq13.1	3(3)	99.84	Dystonia 3, Torsion, X-linked (DYT3)	X-linked	314250
TMLHE	Trimethyllysine hydroxylase, epsilon	Xq28	16(51); 1(1)	99.92; 96.06	New error of carnitine metabolism	X-linked	N/A

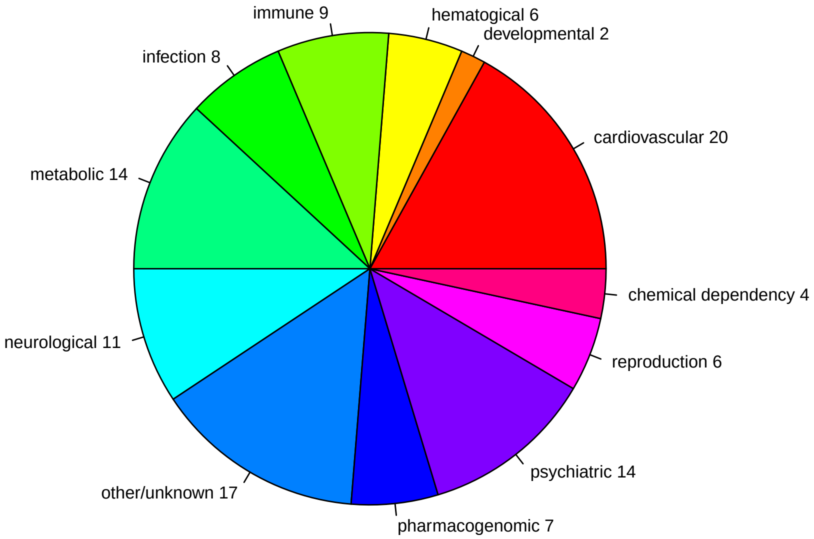

Figure 3.3: Disease classes among the genes intersecting with the IP-LCRs found in Genetic Association Database among entries associated with the genes (field Association(Y/N) equal to 'Y'), and harboured by potential NAHR-mediated inversions. Source: Dittwald et al. (2013b).

4

BRAIN – an algorithm for effective calculation of aggregated isotopic distribution

4.1 Original BRAIN algorithm

In Chapter 1, we already introduced several representations of the molecular isotopic distributions, as well as the corresponding notation. Recall, that by $\xi(v, w, x, y, z)$, we would represent the molecule of a chemical formula $C_vH_wN_xO_yS_z$, i.e. composed of v carbon, w hydrogen, x nitrogen, y oxygen, and z sulphur atoms. The number of stable isotopes for these five atoms equals two (C, H, N), three (O), or four (S). The elemental distribution in normal terrestial matter according to the IUPAC 1997 standard (Rosman and Taylor, 1997) is shown in Table 4.1[1]. Moreover, to calculate the mass (M_{mono}) and probability (P_{mono}) for monoisotopic variant of ξ, one can use Equation (1.1) and Equation (1.2), respectively. The average mass of ξ ($\bar{M}$) can be calculated by the closed formula from Equation (1.3).

The aggregated isotopic distribution merges variants with the same number of neutrons. The aggregated variants are indexed from 0, and the j-th aggregated variant refers to the variant with j additional neutrons in comparison to monoisotopic variant.

[1] In this thesis the calculations are performed for the elemental distribution that are equal or similar to the values from the Table 4.1.

Table 4.1: List of the stable isotopes for carbon, hydrogen, nitrogen, oxygen, and sulphur based on IUPAC 1997 standard (Rosman and Taylor, 1997). Source: (Claesen et al., 2012).

Isotope	Mass (ma/u)	Abundance (%)	Isotope	Mass (ma/u)	Abundance (%)
^{12}C	12.0000000000	98.93	^{16}O	15.9949146	99.757
^{13}C	13.0033548378	1.07	^{17}O	16.9991312	0.038
^{1}H	1.0078250321	99.9885	^{18}O	17.9991603	0.205
^{2}H	2.0141017780	0.0115	^{32}S	31.97207070	94.93
^{14}N	14.0030740052	99.632	^{33}S	32.97145843	0.76
^{15}N	15.0001088984	0.368	^{34}S	33.96786665	4.29
			^{36}S	35.96708062	0.02

By q_j we will denote a probability of j-th aggregated isotopic variant of the molecule ξ, which can be calculated as:

$$q_j = \sum_k p_{jk} \tag{4.1}$$

and the center-mass (i.e. expected value) for j-th isotopic variant is defined as:

$$E(m_j) = \bar{m_j} = \frac{\sum_k m_{jk} p_{jk}}{\sum_k p_{jk}}. \tag{4.2}$$

The m_{jk} and p_{jk} are, respectively, masses and probabilities of the fine variants (indexed by k) with j additional neutrons on comparison to the monoisotopic variant. Of note, the denominator – from Equation (4.1) – is equal to q_j.

In case of carbon dioxide, (CO_2; $\xi(1, 0, 2, 0, 0)$), we consider six aggregated variants with $0 - 5$ additional neutrons. These are (we distinguish between any two isotopic fine variants as long as they are composed of different number of particular isotopes, i.e. we do not differentiate between isoforms):

- monoisotopic variant composed of $^{12}C^{16}O^{16}O$;
- variant with 1 additional neutron composed of $^{13}C^{16}O^{16}O$, and $^{12}C^{16}O^{17}O$;
- variant with 2 additional neutrons composed of $^{13}C^{16}O^{17}O$, $^{12}C^{16}O^{18}O$, and $^{12}C^{17}O^{17}O$;
- variant with 3 additional neutrons composed of $^{13}C^{16}O^{18}O$, $^{13}C^{17}O^{17}O$, and $^{12}C^{17}O^{18}O$;
- variant with 4 additional neutrons composed of $^{13}C^{17}O^{18}O$, and $^{12}C^{18}O^{18}O$;
- variant with 5 additional neutrons composed of $^{13}C^{18}O^{18}O$.

Recall, in Chapter 1 we introduced the notation for isotopic masses ($M_{C_{12}}, \dots, M_{S_{36}}$), and corresponding probabilities, denoted as $P_{C_{12}}, \dots, P_{S_{36}}$. Using these elemental isotopic distributions for the monoisotopic variant, we have that $q_0 = p_{01} = P_{C_{12}} P^2_{O_{16}}$. For q_1, we should consider the sumaric probabilities of two fine peaks, $^{13}C^{16}O^{16}O$, and $^{12}C^{16}O^{17}O$. For the first of them ($k = 1$) $^{13}C^{16}O^{16}O$, $p_{11} = P_{C_{12}} P^2_{O_{16}}$. For the second fine peak ($k = 2$) $^{12}C^{16}O^{17}O$, $p_{12} = 2P_{C_{12}} P_{O_{16}} P_{O_{17}}$. As we ignore the atoms order, this fine variant is equivalent to $^{12}C^{17}O^{16}O$, and that is why multiplication factor 2 is included in the formula. Finally, for all aggregated variants we obtain the following equations:

$$\begin{aligned}
q_0 &= P_{C_{12}} P^2_{O_{16}}, \\
q_1 &= P_{C_{12}} P^2_{O_{16}} + 2P_{C_{12}} P_{O_{16}} P_{O_{17}}, \\
q_2 &= P_{C_{12}} P^2_{O_{17}} + 2P_{C_{12}} P_{O_{16}} P_{O_{18}} + 2P_{C_{13}} P_{O_{16}} P_{O_{17}}, \\
q_3 &= P_{C_{13}} P^2_{O_{17}} + 2P_{C_{13}} P_{O_{16}} P_{O_{18}} + 2P_{C_{12}} P_{O_{17}} P_{O_{18}}, \\
q_4 &= P_{C_{12}} P^2_{O_{18}} + 2P_{C_{13}} P_{O_{17}} P_{O_{18}}, \\
q_5 &= P_{C_{13}} P^2_{O_{18}}
\end{aligned} \tag{4.3}$$

In case of the center-masses, we have $m_{01} = M_{C_{12}} + 2M_{O_{16}}$ for a monoisotopic peak. For one additional neutron $m_{11} = M_{O_{17}} + 2M_{O_{16}}$, and $m_{12} = M_{C_{12}} + M_{O_{16}} + M_{O_{17}}$. Thus, $\bar{m}_1 = \frac{p_{11} m_{11} + p_{12} m_{12}}{p_{11} + p_{12}}$. This example can be continued further for the higher neutron numbers.

Our aim is to effectively calculate both q_j and $\bar{m}_j$. Here, we used the polynomial expansion method from (Rockwood, 1995). Let us first consider the following generating function:

$$\begin{aligned}
Q(I; v, w, x, y, z) = \left(P_{C_{12}} I^0 + P_{C_{13}} I^1\right)^v &\times \\
\left(P_{H_1} I^0 + P_{H_2} I^1\right)^w &\times \\
\left(P_{N_{14}} I^0 + P_{N_{15}} I^1\right)^x &\times \\
\left(P_{O_{16}} I^0 + P_{O_{17}} I^1 + P_{O_{18}} I^2\right)^y &\times \\
\left(P_{S_{32}} I^0 + P_{S_{33}} I^1 + P_{S_{34}} I^2 + P_{S_{36}} I^4\right)^z & \\
= \{Q_C(I)\}^v \times \{Q_H(I)\}^w \times \{Q_N(I)\}^x \times \{Q_O(I)\}^y \times \{Q_S(I)\}^z &
\end{aligned} \tag{4.4}$$

Furthermore, we will use also the shorter form of the previous equation:

$$Q(I; v, w, x, y, z) = \{Q_C(I)\}^v \times \{Q_H(I)\}^w \times \{Q_N(I)\}^x \times \{Q_O(I)\}^y \times \{Q_S(I)\}^z, \tag{4.5}$$

with $Q_C(I) = (P_{C_{12}} I^0 + P_{C_{13}} I^1)$ being an elemental polynomial for carbon, and $Q_H(I), \dots,$ $Q_S(I)$ defined analogously for other elements. The polynomial $Q(I; v, w, x, y, z)$ can be also written in its standard form:

$$Q(I; v, w, x, y, z) \equiv \sum_{j=0}^{n} q_j I^j \ , \tag{4.6}$$

where $n = v + w + x + 2y + 4z$ is a maximal number of additional neutrons. The coefficients $q_0, q_1, q_2, \ldots$ in Equation (4.6) correspond to the probabilities of aggregated isotopic variants, denoted by the same symbols in Equation (4.1). Indeed, in case of carbon dioxide, we obtain:

$$\begin{aligned}
Q(I; 1, 0, 0, 2, 0) &= \left(P_{C_{12}} I^0 + P_{C_{13}} I^1\right)^1 \times \left(P_{O_{16}} I^0 + P_{O_{17}} I^1 + P_{O_{18}} I^2\right)^2 \\
&= (P_{C_{12}} P_{O_{16}}^2) I^0 + \left(P_{C_{12}} P_{O_{16}}^2 + 2P_{C_{12}} P_{O_{16}} P_{O_{17}}\right) I^1 + \\
&\quad \left(P_{C_{12}} P_{O_{17}}^2 + 2P_{C_{12}} P_{O_{16}} P_{O_{18}} + 2P_{C_{13}} P_{O_{16}} P_{O_{17}}\right) I^2 + \\
&\quad \left(P_{C_{13}} P_{O_{17}}^2 + 2P_{C_{13}} P_{O_{16}} P_{O_{18}} + 2P_{C_{12}} P_{O_{17}} P_{O_{18}}\right) I^3 + \\
&\quad \left(P_{C_{12}} P_{O_{18}}^2 + 2P_{C_{13}} P_{O_{17}} P_{O_{18}}\right) I^4 + \left(P_{C_{13}} P_{O_{18}}^2\right) I^5 \\
&= \sum_{j=0}^{5} q_j I^j \ ,
\end{aligned} \tag{4.7}$$

which is consistent with Equation (4.3). Therefore, to obtain q_j, it is sufficient to effectively calculate coefficient near I^j in the polynomial $Q(I; v, w, x, y, z)$.

As already mentioned, this polynomial expansion method was applied by Alan L. Rockwood (Rockwood, 1995). His approach to evaluation of this function involved Fast Fourier Transform (Rockwood, 1995; Rockwood et al., 1995, 1996; Rockwood and Van Orden, 1996) (see also (Valkenborg et al., 2012)). The main advantage of this method is that after FFT, the convolution of the two vectors (which is conventionally done to perform polynomials multiplication) is replaced by multiplication of the coordinates. Finally, the inverse FFT is calculated and normalized. Here, we will show the alternative, algebraic approach to calculate the coefficients in $Q(I; v, w, x, y, z)$, that is easy to be implemented. In addition, in Chapter 5, we would also present some additional improvements that profit from its iterative nature.

Algebraic method to calculate aggregated isotopic probabilities

Let us denote a multiset of roots of polynomial P by $roots(P)$. From Viète's formulas, we know the relationship between the coefficients of polynomial $P(x) = \sum_{j=0}^{n} q_j x^j$ and the symmetric polynomials over $roots(P) = \{x_1, \ldots, x_n\} \subseteq \mathbb{Z}$.

Symmetric polynomials

An j-th symmetric polynomial over the set of variables $x_1, \dots, x_n$, denoted as $e_j(x_1, \dots, x_n)$, or e_j, when the variables $x_1, \dots, x_n$ are known from the context, is defined as a sum of all products of length j of the subsets of $x_1, \dots, x_n$. Namely,

$$\begin{aligned}
e_0 &= e_0(x_1, \dots, x_n) = 1 \\
e_1 &= e_1(x_1, \dots, x_n) = \sum_{k=1}^{n} x_k \\
\dots & \\
e_n &= e_n(x_1, \dots, x_n) = \prod_{k=1}^{n} x_k
\end{aligned}$$

Viète's formulas

The Viète's formulas for polynomial P (that result from its product form $P(x) = q_n(x - x_1) \dots (x - x_n)$) are then as follows:

$$\begin{aligned}
q_0 &= (-1)^n q_n e_n \qquad (4.8) \\
\dots & \\
q_k &= (-1)^{n-k} q_n e_{n-k} \\
\dots & \\
q_{n-1} &= -q_n e_1.
\end{aligned}$$

Unfortunately, it is not trivial to calculate the symmetric polynomials. However, there exist also another set of symmetric polynomials – denoted as $\phi_j(x_1, \dots, x_n)$, or ϕ_j when the variables are known from the context – where j-th polynomial is defined as a sum of j-th powers of $x_1, \dots, x_n$, i.e.:

$$\phi_j = \phi_j(x_1, \dots, x_n) = \sum_{k=1}^{n} x_k^j.$$

If we know $x_1, \dots, x_n$, it is then very easy to quickly calculate ϕ_j for any j. The Newton-

Girard identities (Séroul, 2000) provide the transformation from $e_1, \ldots, e_n$ to $\phi_1, \ldots, \phi_n$:

$$\begin{aligned} e_1 &= \phi_1 \\ e_2 &= \frac{1}{2}(e_1\phi_1 - \phi_2) \\ e_3 &= \frac{1}{3}(e_2\phi_1 - e_1\phi_2 + \phi_3) \\ \ldots \end{aligned} \tag{4.9}$$

Alternatively, from Equation (4.8) we have:

$$\begin{aligned} e_n &= (-1)^n \frac{q_0}{q_n} \\ \ldots \\ e_{n-k} &= (-1)^{n-k} \frac{q_k}{q_n} \\ \ldots \\ e_1 &= -\frac{q_{n-1}}{q_n}. \end{aligned} \tag{4.10}$$

By combining Equation (4.9) and Equation (4.10), we obtain:

$$\begin{aligned} \frac{q_{n-1}}{q_n} &= -\phi_1 \\ \frac{q_{n-2}}{q_n} &= -\frac{1}{2}(\frac{q_{n-1}}{q_n}\phi_1 + \phi_2) \\ \frac{q_{n-3}}{q_n} &= -\frac{1}{3}(\frac{q_{n-2}}{q_n}\phi_1 + \frac{q_{n-1}}{q_n}\phi_2 + \phi_3) \\ \ldots \end{aligned} \tag{4.11}$$

and multiplying all above equations by q_n gives us:

$$\begin{aligned} q_{n-1} &= -q_n\phi_1 \\ q_{n-2} &= -\frac{1}{2}(q_{n-1}\phi_1 + q_n\phi_2) \\ q_{n-3} &= -\frac{1}{3}(q_{n-2}\phi_1 + q_{n-1}\phi_2 + q_n\phi_3) \\ \ldots \end{aligned} \tag{4.12}$$

Of note, Equation (4.12) enables the iterative calculation of the coefficients q_j starting from the heaviest isotopic aggregated variants, while for a practical application it is much more useful to start from the lightest ones.

Let us now replace polynomial P by the mirror polynomial $\bar{P} = \sum_{j=0}^{n} q_{n-j}x^j$, and use the simple algebraic fact that $roots(\bar{P}) = \{x_1^{-1}, \ldots, x_n^{-1}\}$ (where $x_1, \ldots, x_n$ are roots of P). First, we notice that

$$\phi_j(x_1^{-1}, \ldots, x_n^{-1}) = \phi_{-j}(x_1, \ldots, x_n) = \phi_{-j}. \tag{4.13}$$

For simplification of notation, we introduce $\psi_l = \phi_{-j}$. Then, by applying Equation (4.12) for polynomial $\bar{P}$, we obtain:

$$\begin{aligned} q_1 &= -q_0\psi_1 \\ q_2 &= -\frac{1}{2}(q_1\psi_1 + q_0\psi_2) \\ q_3 &= -\frac{1}{3}(q_2\psi_1 + q_1\psi_2 + q_0\psi_3) \\ \ldots \end{aligned} \tag{4.14}$$

or, in a more compact form:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l. \tag{4.15}$$

Recall, that ψ_l is a sum of $(-l)$-powers of elements of $roots(Q(I; v, w, x, y, z))$.

Calculating coefficients ψ_j

We observe that Equation (4.15) uses the coefficients ψ_l for $j = 0, 1, 2, \ldots$. We will show on below that ψ_j can be easily calculated for any j. First, notice that for any $b \in \mathbb{R}$, the value b^{-j} can be calculated by two simple ways:

a) from the closed formula $b^{-j} = \exp(\log(b^{-j})) = \exp(-j \times \log(b))$,

b) iteratively, using previously calculated value $b^{-(j-1)}$, namely $b^{-j} = b^{-1}b^{-(j-1)}$.

The method from a) allows to calculate b^{-j} in a constant time, while the method from b) allows to calculate b^{-j} in a linear time $\Theta(j)$. However, when all coefficients from the range $b, b^2, \ldots, b^j$ have to be calculated anyway, the linear time cannot be beaten. From Equation (4.5) we see that the multiset $roots(Q(I; v, w, x, y, z))$ is equal to the sum of multisets $roots(\{Q_C(I)\}^v)$, $roots(\{Q_H(I)\}^w)$, $roots(\{Q_N(I)\}^x)$, $roots(\{Q_O(I)\}^y)$, and $roots(\{Q_S(I)\}^z)$. Of note, the polynomials $Q_C(I)$, $Q_H(I)$, and $Q_N(I)$ are linear, and their roots (denoted by r_C, r_H, and r_N, respectively) can be obtained from the equations:

$$r_C = -\frac{P_{C_{12}}}{P_{C_{13}}}, \quad r_H = -\frac{P_{H_1}}{P_{H_2}}, \quad \text{and} \quad r_N = -\frac{P_{N_{14}}}{P_{N_{15}}}\ .$$

The polynomial $Q_O(I)$ is quadratic, and its two complex conjugates roots r_O, $\bar{r}_O$ can be obtained as:

$$r_O, \bar{r}_O = \frac{-P_{O_{17}} \pm \sqrt{P_{O_{17}}^2 - 4P_{O_{16}}P_{O_{18}}}}{2P_{O_{18}}} .$$

Finally, the roots of $Q_S(I)$ can be calculated either by closed quartic formulas for the roots of fourth order polynomial ((Shmakov, 2011)) or by numerical approximations. We would obtain two pairs of complex conjugates: $(r_{S,1}, \bar{r}_{S,1})$ and $(r_{S,2}, \bar{r}_{S,2})$. Let us define $r_{O,all,j} = (r_O)^{-j} + (\bar{r}_O)^{-j}$ and $r_{S,all,j} = (r_{S,1})^{-j} + (\bar{r}_{S,1})^{-j} + (r_{S,2})^{-j} + (\bar{r}_{S,2})^{-j}$. It should be noted that

$$roots(\{Q_C(I)\}^v) = \{\overbrace{r_C, \ldots, r_C}^{v}\}.$$

In general, raising the polynomial P to the power j causes that in $roots(\{P\}^j)$ each element of $roots(P)$ is repeated j times. Finally, from definition of ψ_l we obtain:

$$\psi_l = v(r_C)^{-l} + w(r_H)^{-l} + x(r_N)^{-l} + y(r_{O,all,l}) + z(r_{S,all,l}),$$

which gives us the formula for ψ_j using only the roots of elemental polynomials $Q_C(I), \ldots, Q_S(I)$. Let us consider complex conjugates $z = a + ib = |z|(\cos\varphi(z) + i\sin\varphi(z))$ and $\bar{z} = a - ib = |z|(\cos\varphi(z) - i\sin\varphi(z))$, where $|z|$ and $\varphi(z)$ are the modulus and argument of z, respectively. For $n \in \mathbb{Z}$, we can apply de Moivre's formula $z^j = (a + ib)^n = |z|^n(\cos n\varphi(z) + i\sin n\varphi(z))$. As a result we obtain:

$$z^n + \bar{z}^n = 2|z|^n \cos n\varphi(z). \tag{4.16}$$

Therefore, for oxygen and sulphur we do not have to raise complex numbers to the power to obtain the values of $r_{O,all,l}$ and $r_{S,all,l}$.

For the chemical elements with corresponding elemental polynomial has an order higher than four there are no closed form solutions for roots (Abel-Ruffini theorem (Jacobson, 2007)). However, the roots can be approximated numerically (e.g. by the Newton-Raphson method (Press et al., 2007)). Moreover, as we will show in Chapter 5, there is no need to calculate the roots explicitly to obtain the values of ψ_j.

Calculating coefficients q_j

Using Equation (4.15), we can calculate coefficients $q_0, q_1, \ldots$ iteratively for each molecule ξ with a chemical formula $C_vH_wN_xO_yS_z$. Namely, we start from q_0 that is equal to the probability of the monoisotopic variant of ξ, for which we already shown the closed form solution, cf. Equation (1.2). Then we obtain from Equation (4.15) formulas for for $q_1, q_2, \ldots,$ namely:

$$\begin{aligned} q_1 &= -q_0 \times \psi_1 \\ q_2 &= -\frac{1}{2}(q_0 \times \psi_2 + q_1 \times \psi_1) \\ q_3 &= -\frac{1}{3}(q_0 \times \psi_3 + q_1 \times \psi_2 + q_2 \times \psi_3) \\ \ldots & \end{aligned} \tag{4.17}$$

Therefore, to calculate q_j, we need to know the values of $q_0, \ldots, q_{j-1}$ and $\psi_1, \ldots, \psi_j$ (which needs the memory of size $\Theta(j)$), and perform $\Theta(k)$ summations and multiplications for $k = 0, \ldots, j$. Therefore, if we know $\psi_{k_{k=1,\ldots,j}}$, then the computational time to obtain q_j is $\Theta(j^2)$. As all the coefficients $\psi_{k_{k=1,\ldots,j}}$ can be calculated in linear time (see the previous subsection), the total computation time to obtain q_j is $\Theta(j^2)$. The Equation (4.15) is based on algebraic identities, therefore the results are exact if the polynomial roots are known. However, the obtained results might involve numerical errors, e.g. if aggregated probabilities are very small. Anyway, we will show later that for many practical applications, the algorithm reveal convincing accuracy.

Algebraic method to calculate center-masses

Let us remind that the center-mass is defined by Equation (4.2). As already mentioned, the denominator in this equation is simply the aggregated isotopic probability, for which we have already shown the effective method of computation. Here, we concentrate on the task how to calculate numerator of Equation (4.2), namely $\sum_k m_{jk}p_{jk}$ for a given $j \in \mathbb{N}$.

Let us consider a polynomial:

$$U(I; v, w, x, y, z) = \sum_j \left(\sum_k m_{jk}p_{jk} \right) I^j \equiv \sum_j q^\star_j I^j \tag{4.18}$$

Obtaining $\sum_k m_{jk}p_{jk}$ is then an equivalent for calculating the coefficients $q^\star_j$ in $U(I; v, w, x, y, z)$.

For the sake of clarity, we introduce the new polynomial:

$$\begin{aligned} Q^*(I,K;v,w,x,y,z) &= \\ \left(P_{C_{12}}K^{M_{C_{12}}}I^0 + P_{C_{13}}K^{M_{C_{13}}}I^1\right)^v &\times \\ \left(P_{H_1}K^{M_{H_1}}I^0 + P_{H_2}K^{M_{H_2}}I^1\right)^w &\times \\ \left(P_{N_{14}}K^{M_{N_{14}}}I^0 + P_{N_{15}}K^{M_{N_{15}}}I^1\right)^x &\times \\ \left(P_{O_{16}}K^{M_{O_{16}}}I^0 + P_{O_{17}}K^{M_{O_{17}}}I^1 + P_{O_{18}}K^{M_{O_{18}}}I^2\right)^y &\times \\ \left(P_{S_{32}}K^{M_{S_{32}}}I^0 + P_{S_{33}}K^{M_{S_{33}}}I^1 + P_{S_{34}}K^{M_{S_{34}}}I^2 + P_{S_{36}}K^{M_{S_{36}}}I^4\right)^z, & \end{aligned} \tag{4.19}$$

which can be alternatively expressed by:

$$Q^*(I,K;v,w,x,y,z) \equiv \sum_j \left(\sum_k p_{jk}K^{m_{jk}}\right) I^j \tag{4.20}$$

We differentiate $Q^*(I,K;v,w,x,y,z)$, using Equation (4.20), with respect to K, and then set $K = 1$:

$$\begin{aligned} \frac{\partial}{\partial K}Q^*(I,K;v,w,x,y,z)\bigg|_{K=1} &= \sum_j \left(\sum_k m_{jk}p_{jk}K^{m_{jk}-1}\right) I^j\bigg|_{K=1} \qquad (4.21) \\ &= \sum_j \left(\sum_k m_{jk}p_{jk}\right) I^j = U(I;v,w,x,y,z) \end{aligned}$$

where the last equation follows from Equation (4.19). On the other hand, we might perform the same sequence of operations, using initially the Equation (4.19) and applying the formula of the differentiation the product – first equation follows from Equation (4.21):

$$\begin{aligned} U(I;v,w,x,y,z) &= \frac{\partial}{\partial K}Q^*(I,K;v,w,x,y,z)\bigg|_{K=1} \qquad (4.22) \\ &= vW_C(I)Q(I;v-1,w,x,y,z) + wW_H(I)Q(I;v,w-1,x,y,z) \\ &+ xW_N(I)Q(I;v,w,x-1,y,z) + yW_O(I)Q(I;v,w,x,y-1,z) \\ &+ zW_S(I)Q(I;v,w,x,y,z-1) \end{aligned}$$

where $W_C(I) = P_{C_{12}}M_{C_{12}} + P_{C_{13}}M_{C_{13}}I^1$, and $W_H(I)$, $W_N(I)$, $W_O(I)$, $W_S(I)$ are defined analogously.

Thus, $U(I;v,w,x,y,z)$ is a sum of five polynomials, each being a product of polynomials for which the roots can be obtained, and therefore can be calculated using formula analogous

Table 4.2: Ten biomolecules previously used in (Olson and Yergey, 2009), for which the performance of the selected algorithms for isotope distribution calculation has been tested. Table source: Claesen et al. (2012)

No.	Common Name	Molecular Formula	Mass (Da)	
			Monoisotopic	Average
(1)	Angiotensin II	$C_{50}H_{71}N_{13}O_{12}$	1045.534515	1046.181107
(2)	Bovine insulin	$C_{254}H_{377}N_{65}O_{75}S_{6}$	5729.600867	5733.510759
(3)	Human insulin	$C_{520}H_{817}N_{139}O_{147}S_{8}$	11616.849350	11624.448751
(4)	Human myoglobin	$C_{744}H_{1224}N_{210}O_{222}S_{5}$	16812.954775	16823.321352
(5)	Human intrinsic factor	$C_{2023}H_{3208}N_{524}O_{619}S_{20}$	45387.007033	45415.679370
(6)	Bovine serum albumin	$C_{2934}H_{4615}N_{781}O_{897}S_{39}$	66389.862474	66432.455561
(7)	Human Na/K ATPase Renal isoform, subunit	$C_{5047}H_{8014}N_{1338}O_{1495}S_{48}$	112823.879546	112895.125932
(8)	Human ATP binding cassette protein	$C_{8574}H_{13378}N_{2092}O_{2392}S_{77}$	186386.799265	186506.052594
(9)	Human intrinsic factor -hydroxocobalamin receptor	$C_{17600}H_{26474}N_{4752}O_{5486}S_{197}$	398470.366994	398722.972484
(10)	Human dynein heavy chain	$C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$	533403.475090	533735.214651

to Equation (4.15) (summing coefficients of polynomials can be easily done trough adding vectors by coordinates).

The method introduced here to calculate aggregated isotopic distribution is called BRAIN. It should be noted that although its core Equation (4.15) indeed presents the recursive relationship between coefficients, they are calculated iteratively, i.e. one after another.

4.2 Comparison with other packages

For the comparison, we use 10 biomolecules from Olson and Yergey (2009), that are presented in Table 4.2. Selected isotopic distribution are also plotted in Figure 4.1.

The performance was initially tested (Claesen et al., 2012) for the following algorithms:

- Emass – the probabilities and center-mass are calculated using super atoms (idea similar to exponentiation by squaring) which are systematically updated (convoluted) to obtain the investigated molecules; some pruning can be applied during this process (Rockwood and Haimi, 2006);

- Mercury – uses FFT approach to convolute peaks on a grid (Rockwood et al., 1995);

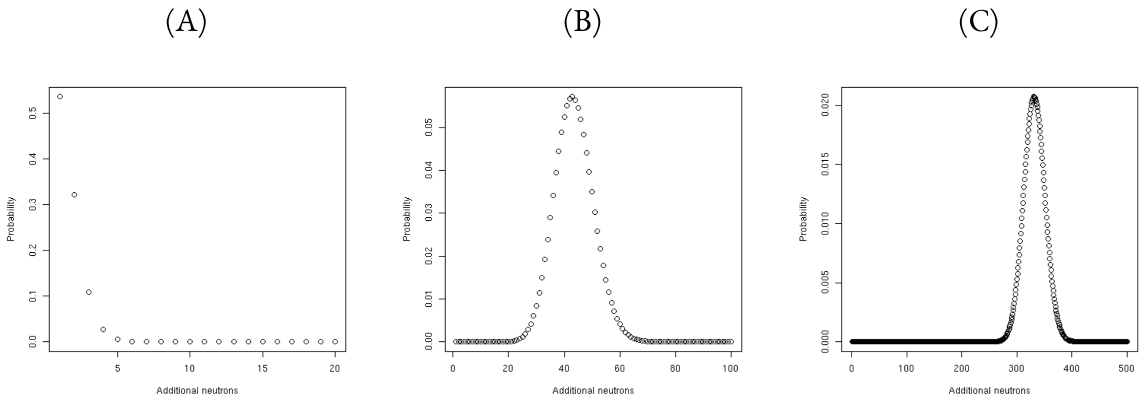

Figure 4.1: Isotopic distribution for three biomolecules (x-axes correspond to peak index, starting from monoisotopic variant, y-axes correspond to probabilities): (A) Angiotensin II ($C_{50}H_{71}N_{13}O_{12}$), (B) Bovine serum albumin ($C_{2934}H_{4615}N_{781}O_{897}S_{39}$), (C) Human dynein heavy chain ($C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$).

- NeutronCluster – uses a bunch of binomial formulas to calculate abundances, and a concept of simplified average mass of the additional neutrons to obtain center-masses (Olson and Yergey, 2009);

- IsoPro – uses multinomial expansion (Yergey, 1983);

- IsoDalton – calculates isotopic fine structure using dynamic programming, the pruning is reducing the list of the fine variants while the calculation is progressing (Snider, 2007);

- BRAIN (MATLAB implementation).

As the criterion to decide about the accuracy of the returned values, we used the average mass calculated by two methods and then compared. Namely, the theoretical average mass from the closed formula presented in Equation (1.3) was compared against the weighted mean $\sum_j q_j m_j$, which express exactly the same value (of course, when only a certain part of the distribution is computed, this is not exactly the same, but if the distribution is sufficiently covered, the difference is expected to be very tiny). When $\sum_j q_j m_j \approx \bar{M}$ for different molecules, the distribution is claimed to be accurately calculated.

In this assessment, both BRAIN and Emass returned the very accurate values of weighted mean, while the Mercury and NeutronCluster returned greater error. IsoPro and IsoDalton, beside low accuracy, were also very inefficient and only selected molecules from Table 4.2 (i.e. molecules 1-7, and 1-5, respectively) were tested. It should be mentioned at this point, that in an original BRAIN article (Claesen et al., 2012) we proposed a simple heuristic for estimating a number of peaks sufficient to cover the most informative part of the isotope distribution with the following formula:

$$n_{\text{stop}} = \max(2 \times \lceil \bar{M} - M_{mono} \rceil, 5), \tag{4.23}$$

The idea behind this equation is the assumption that the distribution has a bell-shape curve and taking this distribution symmetrically around its expected value should capture all informative peaks. The time performance – number of peaks set according to Equation (4.23) – for BRAIN gave around 0.04 sec for molecules (1-4) up to around 0.4 sec for the heavies molecule 10 (tests run on Intel Core 2 Duo processor with 2.26 GHz and 4 GB RAM). The full set of results conducted by Dr. Jürgen Claesen is presented in (Claesen et al., 2012). As a follow-up of this article, Dr. Sebastian Böcker presented an additional comparison with SIRIUS (a framework for *de novo* identification of metabolites, that employs the isotope pattern analysis (Böcker et al., 2009)) and BRAIN R Bioconductor package (see Chapter 6), where he pointed out that SIRIUS is as accurate as BRAIN and works even faster than BRAIN R implementation (however, the author admitted that this was not surprising due to the fact that SIRIUS was written in Java, i.e. a compiled language).

As correctly observed by Fernandez-de Cossio Diaz and Fernandez-de Cossio (2012), the stopping criterion from Equation (4.23) does not work well for the small molecules (to few peaks are computed). Therefore we suggested in Hu et al. (2013) to slightly modify this formula:

$$n_{\text{stop}} = \max(2 \times \lceil \bar{M} - M_{mono} \rceil, 50). \tag{4.24}$$

Simply, the only change is to increase the minimal number of computed peaks from 5 to 50. Of course, another stopping criteria can be also applied. For example, as discussed in Claesen et al. (2012) and Hu et al. (2013), the iterative procedure in BRAIN can be stopped when the cumulative distribution would reach the certain threshold, e.g. 99.9%.

Of note, Equation (4.23) is simply a sum of polynomial products, so can be also calculated via FFT. Indeed, Fernandez-de Cossio Diaz and Fernandez-de Cossio (2012) provided very efficient calculation of center-masses as a part of ISOTOPICA package implemented in C# (referred further as FTMC), that outperformed Bioconductor BRAIN implementation. However, as C# is an compiled language, and R is an interpreted language, we benchmarked also a C++ implementation of BRAIN written by Han Hu, called `useBRAIN` (available online at `https://code.google.com/p/brain-isotopic-distribution/`), that is also suitable for batch processing[2]. In Hu et al. (2013) we made a comparison between FTMC and `useBRAIN` and showed that the latter software can work faster for several tests (cf. Fig-

[2] This implementation works for all chemical elements, as it uses efficient way to calculate power root sums without calculating the roots explicitly. This method, called [RO] improvement, will be presented in the next chapter as a part of BRAIN 2.0 algorithm.

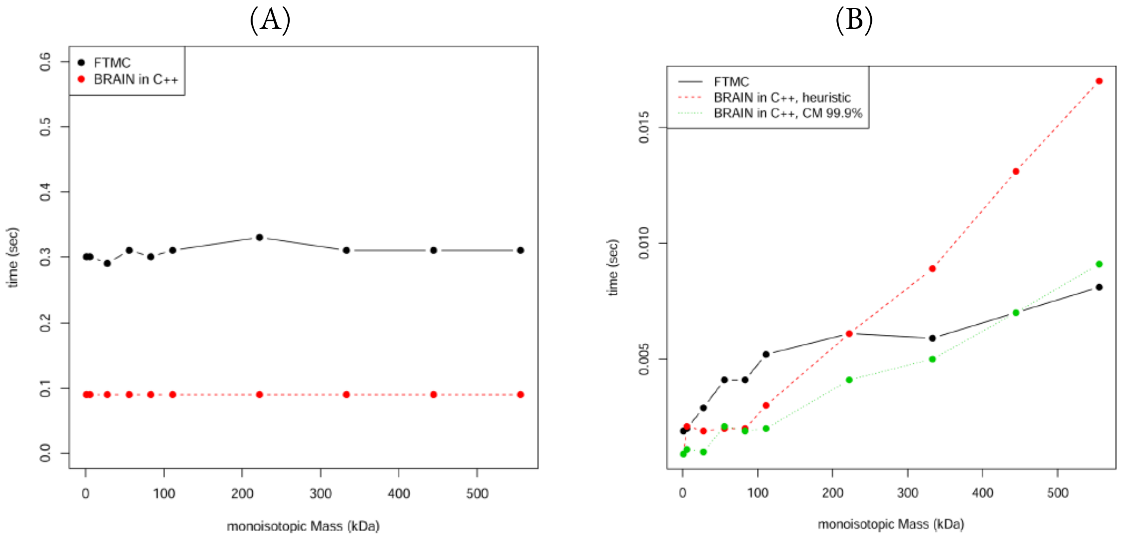

Figure 4.2: Comparison between FTMC (default number of computer peaks) and C++ implementation (the latter run for the number of peaks according to Equation (4.24) of the BRAIN for processing 10 molecules from averagine model with corresponding masses marked on x-axes. (A) Average (from 100 runs) time of processing a single molecule. (B) Elapsed time (divided by 100) from the batch-processing of the file with 100 the same molecules. In addition, we show the BRAIN in C++ for the peaks range that starts at the monoisotopic variants and ends when the coverage exceeds 99.9% (this number is precalculated using Bioconductor BRIAN package) – this heuristic is denoted as CM 99.9%. Source: Hu et al. (2013). The comparison code is available online at `http://www.mimuw.edu.pl/~pd219416/AnChemComment/`

ure4.2). It should be underlined that the algorithm performance depends also on different stop criteria, which usually result in different numbers of the calculated peaks.

5

BRAIN 2.0 – improvements to the original BRAIN

As already mentioned above, the original BRAIN calculates iteratively the $q_0, q_1, \ldots$ coefficients. In this section, we show that for the molecules composed of five chemical elements, C, H, N, O and S, the improvements in calculating probabilities of the aggregated isotopic variants, both involving speed and memory, can be applied. In addition, the extension of these methods for the other chemical elements is also discussed. Of note, the application of the improvements for calculating center-masses needs additional investigation.

5.1 Improvements presentation

Recurrence of constant length [RCL] improvement Equation (4.15) gives a formula for calculating q_j, namely the probability of the j-th aggregated isotopic variant. To this aim, we calculate the standard scalar product of vectors $(q_0, \ldots, q_{j-1})$ and $(\psi_j, \ldots, \psi_0)$. For large j, we would like to choose the natural index d such that $1 \leq d < j$, and trimming the sum in Equation (4.15) to the length d will give us the coefficient $\hat{q}_j$ that is approximating q_j

with a small error. First, we will split Equation (4.15) into two parts:

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l = -\frac{1}{j}(\sum_{l=1}^{d} q_{j-l}\psi_l + \sum_{l=d+1}^{j} q_{j-l}\psi_l) \tag{5.1}$$
$$= \underbrace{-\frac{1}{j}\sum_{l=1}^{d} q_{j-l}\psi_l}_{\hat{q}_j} \underbrace{-\frac{1}{j}\sum_{l=d+1}^{j} q_{j-l}\psi_l}_{error}$$

Then, we estimate $|q_j - \hat{q}_j|$:

$$\begin{aligned} |q_j - \hat{q}_j| &= |-\frac{1}{j}\sum_{l=d+1}^{j} q_{j-l}\psi_l| \\ &= \frac{1}{j}|\sum_{l=d+1}^{j} q_{j-l}\psi_l| \leq \frac{1}{j}\sum_{l=d+1}^{j} |q_{j-l}\psi_l| \\ &= \frac{1}{j}\sum_{l=d+1}^{j} |q_{j-l}||\psi_l| \overset{(\star)}{\leq} \frac{1}{j}\sum_{l=d+1}^{j} |\psi_l| \\ &\leq \frac{j-d-1}{j}\max_{l\in\{d+1,\ldots,j\}}|\psi_l| \leq \max_{l\in\{d+1,\ldots,j\}}|\psi_l|, \end{aligned} \tag{5.2}$$

where $(\star)$ results from $\forall_{k\in\{0,\ldots,j-1\}} 0 \leq q_k \leq 1$, as q_k are probabilities.

On the other hand, from Equation (4.16) we can estimate

$$\begin{aligned} |\psi_j| &= |v(r_C)^{-j} + w(r_H)^{-j} + x(r_N)^{-j} + y(r_{O,all,j}) + z(r_{S,all,j})| \\ &\leq v|(r_C)^{-j}| + w|(r_H)^{-j}| + x|(r_N)^{-j}| + y|(r_{O,all,j})| + z|(r_{S,all,j})|. \end{aligned} \tag{5.3}$$

Figure 5.1(A) shows the descending trends of $|(r_C)^{-j}|$, $|(r_H)^{-j}|$, $|(r_N)^{-j}|$, $|(r_{O,all,j})|$, and $|(r_{S,all,j})|$, while we increase j. Of course, the exact value of $|\psi_j|$ also depends on numbers of particular atoms, i.e. v, w, x, y, z, therefore for a given class of molecules, additional investigation should be performed.

Here, we will concentrate on proteins, where the proportions between these numbers are relatively constant. We will use four heaviest biomolecules from Table 4.2, i.e. molecules with labels $7-10$ (our improvement is aimed for the relatively big molecules, therefore for molecules 1-6 we suggest using the original BRAIN). Figure 5.1(B) presents that indeed, also the value of $|\psi_l|$ reveals the trend decreasing to 0, therefore if d is big enough, the value of

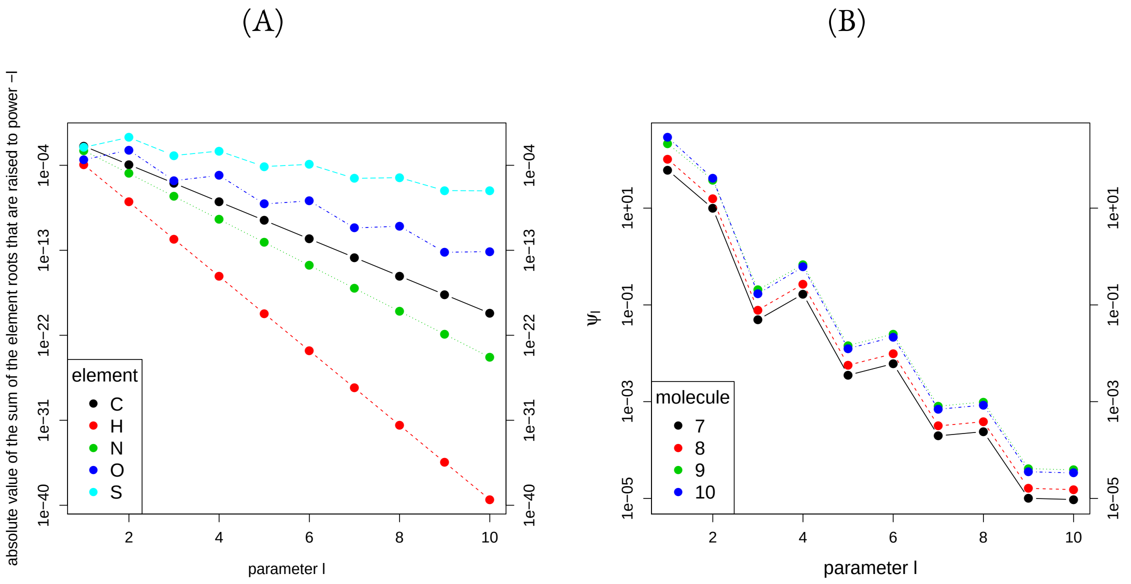

Figure 5.1: (A) The values of $|(r_C)^{-l}|, |(r_H)^{-l}|, |(r_N)^{-l}|, |(r_{O,all,l})|, |(r_{S,all,l})|$ decreases monotonically while we increase l. (B) The values of $|\psi_l|$ for 4 heavy biomolecules from Table 4.2. We observe the decreasing trend. Please, note the logarithmic scale on y-axes in both panels. Figure source: Dittwald and Valkenborg (2014).

$\max_{l \in \{d+1,\ldots,j\}} |\psi_l|$, and as a consequence the value of $|q_j - \hat{q}_j|$ should be sufficiently small. For example, for $d = 10$ and four analyzed biomolecules we have:

$$|q_j - \hat{q}_j| \leq \max_{l \in \{d+1,\ldots,j\}} |\psi_l| \leq |\psi_{10}| \leq 10^{-4}.$$

Late Starting Point [LSP] improvement The other characteristics of the BRAIN algorithm is that calculations start from the monoisotopic variant, i.e. q_0. However, this can be treated as a weakness of the method, as interesting peaks (i.e. big enough) might start much later. For example, the following method from Rockwood et al. (1995) can be applied to narrow the range of the investigated peaks:

1. Calculate σ, i.e. the standard deviation of the mass distribution, from the closed formula:

$$\sigma = \sigma_C + \sigma_H + \sigma_N + \sigma_O + \sigma_S, \tag{5.4}$$

where $\sigma_C = P_{C_{12}} M_{C_{12}}^2 + P_{C_{13}} M_{C_{13}}^2 - (P_{C_{12}} M_{C_{12}} + P_{C_{13}} M_{C_{13}})^2$, and $\sigma_H, \ldots, \sigma_S$ are calculated analogously.

2. Calculate N, i.e. number of investigated peaks, as:

$$N = \lceil \alpha \sqrt{(1+\sigma^2)} \rceil, \tag{5.5}$$

where α is a constant (typically $\alpha = 10$; in (Fernandez-de Cossio Diaz and Fernandez-de Cossio, 2012) $\alpha = 16$ is used). For simplicity, we assume further N is odd; in cases where N is even, the tiny adjustments are needed.

3. Calculate the middle point n_{middle} of the investigated distribution to be the closest one to the molecule average mass.

4. Calculate n_{start} and n_{stop} such that $n_{start} = n_{middle} - \lfloor \frac{N}{2} \rfloor$ and $n_{stop} = n_{middle} + \lfloor \frac{N}{2} \rfloor$.

An interesting observation is that if we use q_0 multiplied by the constant denoted γ and then calculate subsequent coefficients from Equation (4.15), then the subsequent coefficients will be also multiplied by γ. However, the ratios between consecutive coefficients remain unaffected, i.e. will be equal to $\frac{q_1}{q_0}, \frac{q_2}{q_1}, \ldots$. In other words, we can start our iterative formula from any arbitrary set number (e.g. from 1), and obtain true ratios of consecutive probabilities of aggregated variants. In practical applications, the peak heights (probability of aggregated variant can be also referred as the peak heights) are often normalized, e.g. by dividing by the maximal peak height. Therefore, there is not much loss of information if we consider only probabilities ratios (which we would alternatively call peak ratios) instead of the actual probabilities. Moreover, we can approximate the values of the probabilities from the normalized peak heights.

The very interesting question is whether it is possible to start the iteration in Equation (4.15) later than from the monoisotopic variant. We assume there might be some burn-in period needed to retrieve the original values of the peaks ratios, and propose the following heuristic:

1. choose the first (n_{start}) and the last index (n_{stop});

2. depending on n_{start}, choose the appropriate value of the burn-in period $1 \geq b \geq n_{start}$;

3. using Equation (4.15), and setting $q_0, \ldots, q_{b-1}$ to 0 and q_b to 1, calculate coefficients $q_b, \ldots, q_{start}, \ldots, q_{stop}$;

4. calculate the ratios of the consecutive coefficients $\frac{q_{start+1}}{q_{start}}, \ldots, \frac{q_{stop-1}}{q_{stop}}$.

Root Omitting [RO] improvement In the original BRAIN, the sum of the roots' powers for each elemental polynomial was calculated using the explicitly obtained (from the closed formulae or numerically approximated) values of the roots. However, we would show here that using once again the Newton-Girard identities, it is possible to solve this problem easier.

Let us consider sulphur, a relatively complicated example. Of note, Equation (4.15) can be applied to elemental polynomial $Q_S(I)$ in the following manner:

$$\begin{aligned} P_{S_{33}} &= -P_{S_{32}} r_{S,all,1} \\ P_{S_{34}} &= -\frac{1}{2}(P_{S_{33}} r_{S,all,1} + P_{S_{32}} r_{S,all,2}) \\ 0 = P_{S_{35}} &= -\frac{1}{3}(P_{S_{34}} r_{S,all,1} + P_{S_{33}} r_{S,all,2} + P_{S_{32}} r_{S,all,3}) \\ P_{S_{36}} &= -\frac{1}{4}(P_{S_{34}} r_{S,all,2} + P_{S_{33}} r_{S,all,3} + P_{S_{32}} r_{S,all,4}) \\ 0 = P_{S_{37}} &= -\frac{1}{5}(P_{S_{36}} r_{S,all,1} + P_{S_{35}} r_{S,all,2} + P_{S_{34}} r_{S,all,3} \\ &+ P_{S_{33}} r_{S,all,4} + P_{S_{32}} r_{S,all,5}) \\ &\vdots \end{aligned} \tag{5.6}$$

For coefficients near $I^0, \ldots, I^4$, i.e. $P_{S_{33}}, \ldots, P_{S_{36}}$, the application of Equation (4.15) is straightforward, which allows to retrieve iteratively the values of $r_{S,all,1}, \ldots, r_{S,all,4}$. The only non-trivial step is to realize that the identities remain true for the $P_{S_{37}}, P_{S_{38}}, \ldots$ coefficients, which all are equal to 0. In general, we obtain:

$$0 = P_{S_{32+i}} = -\frac{1}{i}(P_{S_{36}} r_{S,all,(i-4)} + P_{S_{34}} r_{S,all,(i-2)} + P_{S_{33}} r_{S,all,(i-1)} + P_{S_{32}} r_{S,all,i})$$

Therefore, it is also possible to iteratively calculate values of $r_{S,all,5}, r_{S,all,6}, \ldots$. Finally, the iterative equation is as follows:

$$r_{S,all,i} = -(P_{S_{32}})^{-1}(P_{S_{36}} r_{S,all,(i-4)} + P_{S_{34}} r_{S,all,(i-2)} + P_{S_{33}} r_{S,all,(i-1)})$$

It should be pointed out that using the analogous argumentation, the proposed method can be applied to any chemical element.

5.2 Performance tests

To challenge the [RCL] and [LSP] improvements in practice (as already mentioned in the previous Chapter, the [RO] improvement was already implemented in C++ by Han Hu as a part of original BRAIN), we implemented them in the R programming language and compared with the original algorithm, in which we inactivated the center-masses calculations to have computation times comparable. We performed a set of tests, for which we have used again four heaviest biomolecules among those already used to benchmark the original BRAIN (molecules 7-10 from Table 4.2). To check if the improvements do not seriously affect the accuracy, we calculated the Pearson's χ^2 error statistic between isotopes ratios. More precisely, this statistic is defined as:

$$\chi^2 = \sum_{j=n_{start}}^{n_{stop}} \frac{(R_j^I - R_j^{II})^2}{R_i^I} \tag{5.7}$$

where R_j^I and R_j^{II} are the ratios between intensities of j-th and $(j+1)$-th aggregated peak obtained by original BRAIN and BRAIN 2.0 improvements being investigated, respectively.

Moreover, to set parameters d and b used in [RCL] and [LSP] improvements, we used the following rule of thumb (that in particular is compatible with the observations while exploring different values of these parameters for the molecules 7-10, cf. Figure 5.2)

$$b = d = \lceil \log_{10}(M_{mono}) + 5 \rceil. \tag{5.8}$$

The assessments are as follows:

1. We compared the original BRAIN with [RCL] improvement. We run both algorithms for the same number of N peaks (starting from the monoisotopic variant) as returned by heuristic from Equation (4.23). The results presented in Table 5.2 show that while the χ^2 remains very small (so both algorithms return pretty the same vectors of isotopic ratios), the [RCL] outperforms original BRAIN up to 2-fold speed up for the heaviest molecule.

2. In the second assessment, the original BRAIN is tested against the [LSP] improvement. In case of the first algorithm, the number of the computed peaks (recall, we start from monoisotopic variant) was set according to Equation (4.23). For [LSP] improvement we used Equation (5.5) with $\alpha = 10$ to specify the index of the first (n_{start}) and the last (n_{stop}) peak used to obtain the peak ratios (additional b peaks preceding this peak range needs to be computed according to the specificity of the [LSP] improvement), which indeed resulted in the better time performance (cf. Table 5.3) for

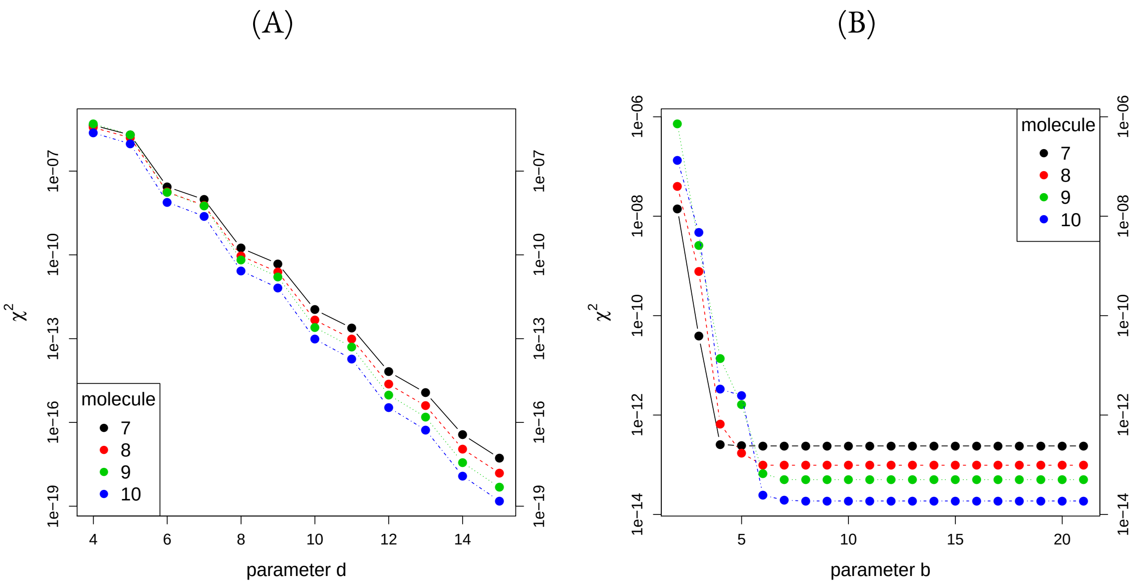

Figure 5.2: The change of χ_2 correlation between the isotopic ratio calculated according to BRAIN and BRIAN 2.0, when one parameter (length of recursion d and burn-in period b in panels (A) and (B), respectively) is set to 11 according to Equation (5.8), as a function of the other parameter. We observe that for $b = d = 11$ the value of χ^2 is smaller than 10^{-12}. Figure source: Dittwald and Valkenborg (2014).

Table 5.1: The speed-up evaluation when both [RCL] and [LSP] improvements are enabled. Speed is measured in seconds.

id	$monoMass(Da)$	b	d	χ^2	$speed_{BRAIN}$	$speed_{BRAIN2}$	$improvement$
7	112824	11	11	2.39e-13	0.00873	0.00473	1.85
8	186387	11	11	9.79e-14	0.0138	0.0054	2.56
9	398470	11	11	5.02e-14	0.0336	0.007	4.8
10	533403	11	11	1.87e-14	0.0493	0.00766	6.43

this method.

3. Finally, we combined [RCL] and [LSP] improvements in a single assessment. The number of peaks calculated for original BRAIN and improved versions were computed analogously to the previous point. While we do not observe serious loss of accuracy (in terms of the Pearson's χ^2) in comparison with the original BRAIN, the speed-up is larger than in any of the previous tests (cf. Table 5.1).

Table 5.2: The speed-up evaluation when only [RCL] improvement is enabled. Speed is measured in seconds.

id	$formula$	$monoMass(Da)$	d	n_{start}	n_{stop}	N	χ^2	$speed_{BRAIN}$	$speed_{BRAIN2}$	$improvement$
7	$C_{5047}H_{8014}N_{1338}O_{1495}S_{48}$	112824	11	1	143	143	1.07e-11	0.00863	0.00588	1.47
8	$C_{8574}H_{13378}N_{2092}O_{2392}S_{77}$	186387	11	1	239	239	2.45e-11	0.0137	0.00873	1.57
9	$C_{17600}H_{26474}N_{4752}O_{5486}S_{197}$	398470	11	1	506	506	7.65e-11	0.0336	0.0162	2.07
10	$C_{23832}H_{37816}N_{6528}O_{7031}S_{170}$	533403	11	1	664	664	6.2e-11	0.0484	0.0208	2.33

Table 5.3: The speed-up evaluation when only [LSP] improvement is enabled. Speed is measured in seconds.

			BRAIN			BRAIN 2.0						
id	$monoMass(Da)$	b	n_{start}	n_{stop}	N	n_{start}	n_{stop}	N	χ^2	$speed_{BRAIN}$	$speed_{BRAIN2}$	$improvement$
7	112824	11	1	143	143	26	117	92	1.75e-30	0.00859	0.00748	1.15
8	186387	11	1	239	239	61	178	118	8.63e-29	0.0136	0.00996	1.37
9	398470	11	1	506	506	167	338	172	2.08e-25	0.0336	0.0158	2.13
10	533403	11	1	664	664	235	429	195	8.68e-25	0.0484	0.0188	2.57

6

Applications of the BRAIN algorithm for large-scale data analyses

6.1 The Bioconductor BRAIN package

The BRAIN package is written in the R statistical language as a part of Bioconductor repository (Gentleman et al., 2004). The package provides the user a few functions, which as an input get the chemical formula (a list with numbers of C, H, N, O, S):

- `calculateMonoisotopicMass` – calculates monoisotopic mass of the molecule from Equation (1.1)
- `calculateAverageMass` – calculates average mass of the molecule from Equation (1.3)
- `calculateIsotopicProbabilities` – computes vector $(q_0, \ldots, q_{n_{stop}-1})$, of the aggregated isotopic variants probabilities;
- `useBRAIN` – the main functionality of the package; this computes all what the aforementioned functions offer plus a vector $(m_0, \ldots, m_{n_{stop}-1})$ with center-masses of the aggregated isotopic variants;

Recall that BRAIN computes isotopic distribution iteratively from the monoisotopic peak, therefore functions `calculateIsotopicProbabilities` and `useBRAIN` use the n_{stop} index, obtained from the input parameters. This index depends on one of the following stop criteria (represented by the values of the parameter *stopOption*):

1. *"nrPeaks"* – number of peaks explicitly provided by the user – by default it is a simple heuristic from Equation (4.23), computed by a function `calculateNrPeaks`, for the number of peaks, where the monoisotopic peak is the first one, that should cover the significant part of the isotopic distribution;

2. *"coverage"* – a fraction of coverage (value between 0 and 1) that should be covered by a cumulative distribution function; the computations stop when the defined value is reached;

3. *"abundantEstim"* – a number of consecutive peaks that are not higher than the current maximal peak; the computations stop when this criterion is satisfied.

In addition, the BRAIN offers a function `getAtomsFromSeq`, which takes as input the sequence of amino acids and returns a list with numbers of C, H, N, O, and S. This function can be useful in preprocessing step. The documentation of the package (a Reference Manual and a package vignette containing examples of usage) is available at Bioconductor online (`http://www.bioconductor.org/packages/release/bioc/html/BRAIN.html`), together with a source code.

6.2 High-throughput data processing

As already mentioned in the Introduction, the common masses assigned for the molecules are e.g. the monoisotopic mass and the average mass, and these values can be referred to in the databases when identification is done. For the small peptides, the monoisotopic peak is relatively high in the aggregated isotopic distribution (cf. Figure 4.1(A)). However, when the size of molecule increases, we observe a trend, where a bell-shaped distribution moves right (cf. Figure 4.1(B)-(C)). More precisely, the shift between the most abundant aggregated peak and the monoisotopic peek increases as well. As a result, for the large molecules, the monoisotopic peak is usually expected to be very tiny. This might cause biases in the molecule identification procedure. The average mass is estimated from the observed distribution, e.g. by a simple weighted mean calculation. The other approach might be to estimate the monoisotopic mass based on its dependence on the most abundant aggregated mass. To this aim, we processed the Uniprot database (Yamamoto and McCloskey, 2012) and calculated aggregated isotopic distributions for 52,589 cases with the monoisotopic masses smaller than 10^5 Da. We analyzed the relationship between the most abundant aggregated mass and corresponding monoisotopic mass, which occurred to be linear (cf. Figure 6.1(A)-(B)).

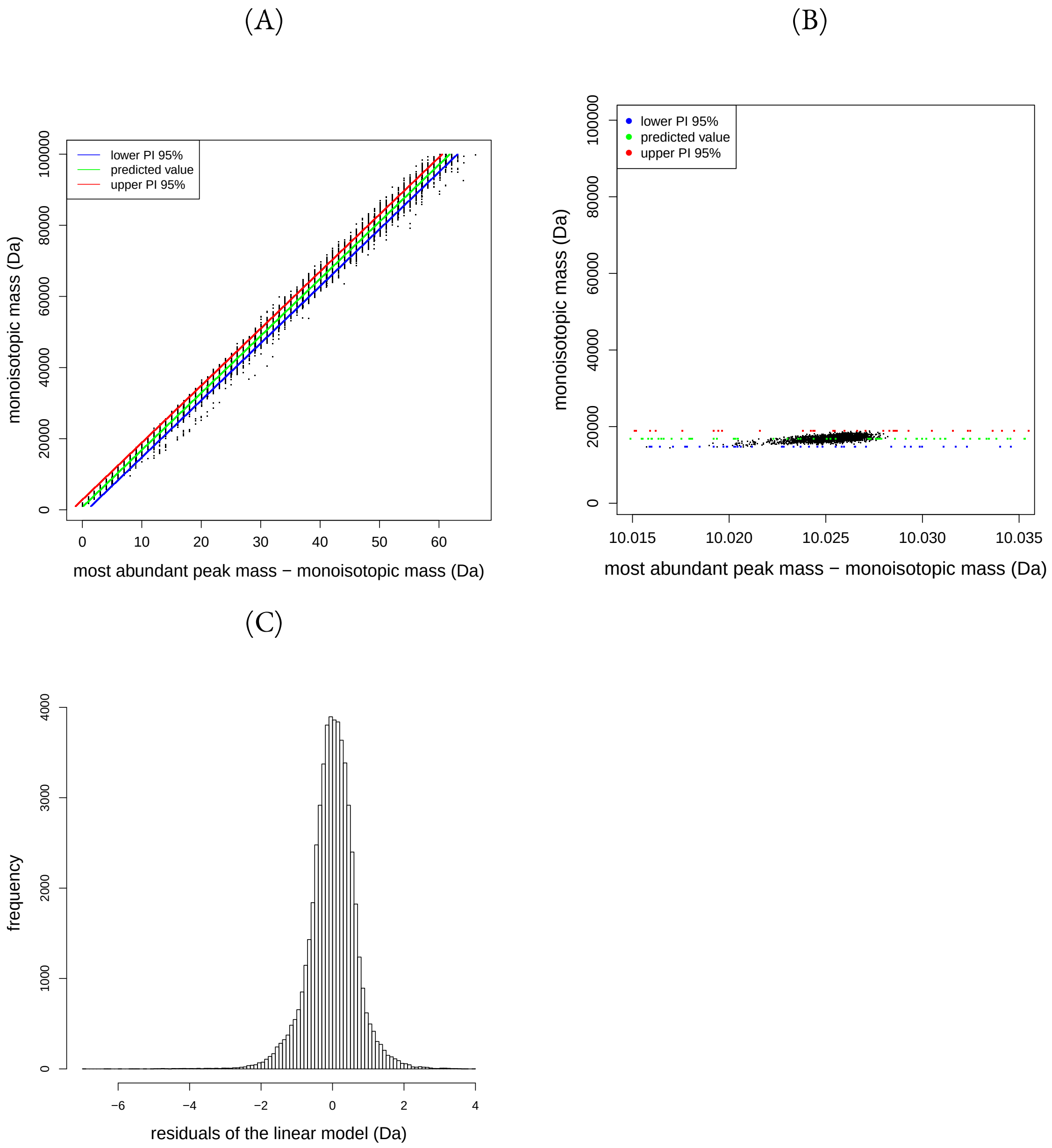

Figure 6.1: (A) The relationship between proteins' monoisotopic mass and the shift between the most abundant and monoisotopic mass, together with upper and lower 95% prediction interval (PI) – the zoom-in of this plot is presented in (B). (C) Histogram with the residuals of the model from Equation (6.1) – the vast majority of the values fits within the interval [-2Da; 2 Da]. Source: Dittwald et al. (2013a).

Therefore, the following linear model was constructed giving the following formula:

$$M_{mono} = 0.482 + 0.9994 \times m_{ma}, \tag{6.1}$$

(m_{ma} denotes most abundant peak mass) where both coefficients are statistically significant (p-values $< 2 \times 10^{-16}$). However, the residuals (for the same data as used to build the model) spanned a range of ± 2 Da (cf. Figure 6.1(C)). This suggests that the further study needs to be performed to improve the prediction (indeed, we are developing the hierarchical model in order to receive smaller residuals – see also Chapter 8). On the other hand, this simple case-study shows that the BRAIN package is suitable for a large-scale (high-throughput) data processing (package version 1.4.0. was used and run on PC with two Intel(R) Core(TM)2 2.40GHz CPUs; total processing of the 52,589 proteins took approximately 80 minutes).

6.3 Lipid Centrifuge

As already mentioned in Chapter 1, the mass spectrometry experiments process enormous amount of information, which cause that the accurate data processing constitutes a bottleneck in various assays. Therefore, automated procedures supporting the experimental workflows are highly desired by the community. In this study, we will propose a classifier that might help to distinguish lipids and peptides from a full scan mass spectra (i.e. when the full mass information within a predefined range is returned by an instrument).

Databases

To retrieve the actual chemical formulas of peptides, we have used the Human Uniprot protein database (Yamamoto and McCloskey, 2012). The proteins were then *in-silico* tryptically digested (no missed cleavages allowed). To this aim we have used the OrgMassSpecR package (function `Digest`) from the R CRAN repository (`http://cran.r-project.org/web/packages/OrgMassSpecR/index.html`). The motivation for digesting was to have peptides and lipids masses comparable (the intact proteins are in general much heavier). The lipids chemical formulas were extracted from the Lipid Maps gateway database (Fahy et al., 2009).

For the further study, we limited the data to molecules with monoisotopic masses below $2,800$ Da. This gave $263,897$ *in silico* tryptic digested peptides, and $6,313$ lipids. Of note, the latter set can be further subdivided into eight lipid classes (as defined by Lipid Maps consortium): fatty acyls (#913; FA), glycerolipids (#400; GL), glycerophospholipids (#1,415; GP), sphingolipids (#1,167; SP), sterol lipids (#604; ST), prenol lipids (#442; PR), saccharolipids (#76; SL), and polyketides (#1,296; PK); # tag indicates the number of items in each lipid class.

The general outline of the study is presented in Figure 6.2, and will be further explained in the following parts of this section.

Mass defect

The basic concept we used in this study is based on the well known *mass defect* phenomenon. More precisely, we wanted to use the tiny differences in fractional parts of masses between the chemical elements for the classification purposes. Of note, the mass defect based approach was used by (Kirchner et al., 2010) and (Bruce et al., 2006) when investigating a degree of phosphorylation in proteins. To give an intuitive explanation of our reasoning let us consider the molecules composed on C, H, N, O, and S, and take a look at its monoisotopic mass fractional part, which equals:

- 0 Da for carbon (by definition, as 1 Da = 1/12 of ^{12}C mass);
- 0.007 Da for hydrogen;
- 0.003 Da for nitrogen;
- 0.995 Da for oxygen;
- 0.972 Da for sulphur.

In a case of proteins and peptides, the specific structure of the molecule (chain of amino acids) results in relatively well defined constraints between numbers of chemical elements. The *averagine*, average peptide model (Senko et al., 1995a) based on the Protein Identification Resource database, gives the following formula to describe these proportions:

$$C : H : N : O : S = 4.9384 : 7.7583 : 1.3577 : 1.4773 : 0.0417. \tag{6.2}$$

From the *averagine* model we can see that for a given monoisotopic mass we can approximate the amount of each atom and, using the elemental mass defects, we can predict the overall mass defect of the monoisotopic mass. Of course, the proportions between amounts of atoms in real molecules differ from the model. However, even with some margins allowed, we can still conclude that for a given monoisotopic mass of the peptide, its fractional part can take only a certain range of values. In case of lipids, the analogous fractional parts of monoisotopic masses can potentially differ – mostly because of different structural characteristics of the molecule; in addition, lipids can also contain fluorine (F), bromine (Br), phosphorus (P), chlorine (Cl), sodium (Na), iodine (I), and potassium (K) atoms. We performed a simple computational experiment, and for a monoisotopic masses between 740 and 750 Da investigated the mass defect and placed values into appropriate bins. When the bin widths

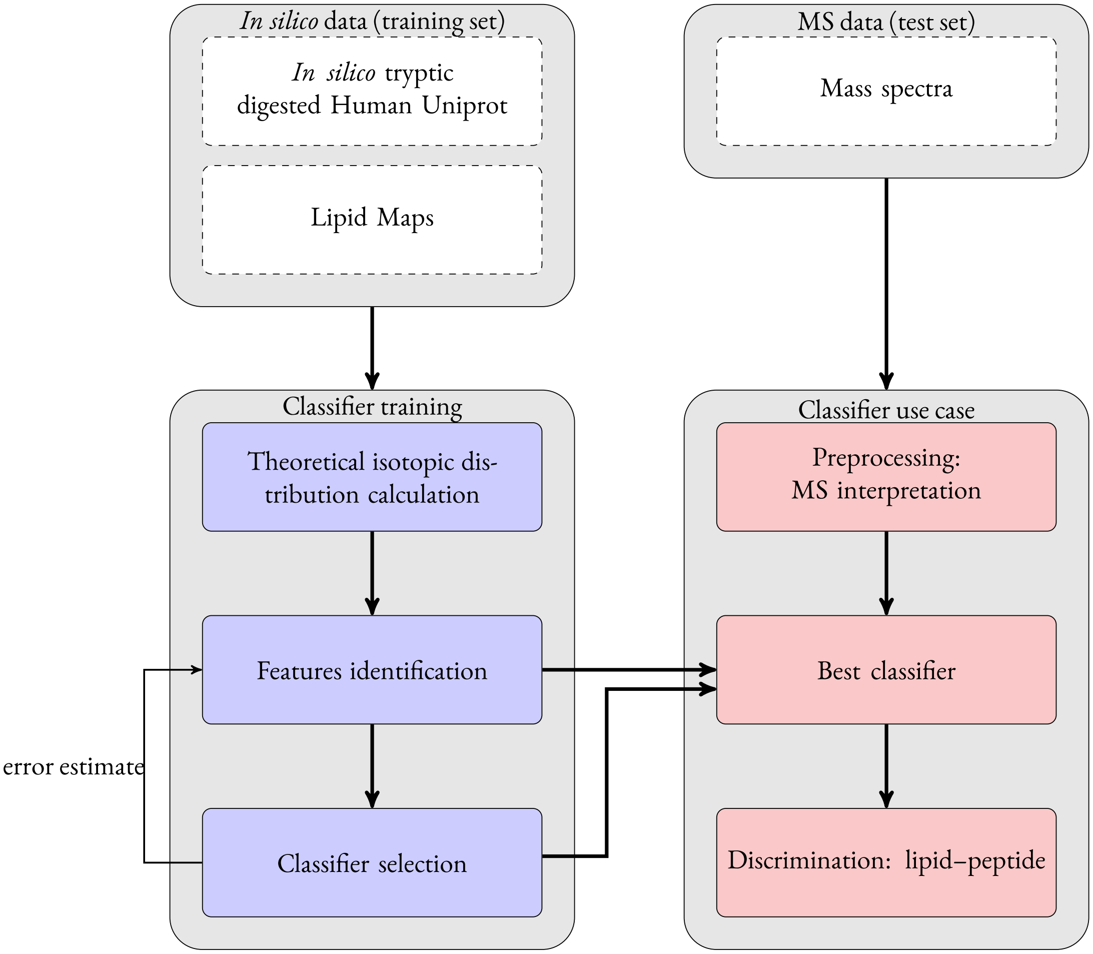
In silico data (training set)
In silico tryptic digested Human Uniprot
Lipid Maps
MS data (test set)
Mass spectra
Classifier training
Theoretical isotopic distribution calculation
Features identification
Classifier selection
error estimate
Classifier use case
Preprocessing: MS interpretation
Best classifier
Discrimination: lipid–peptide

Figure 6.2: Schematic workflow of the lipid-vs.-peptide classifier construction, validation and application. Figure courtesy: Dr. Anna Gambin

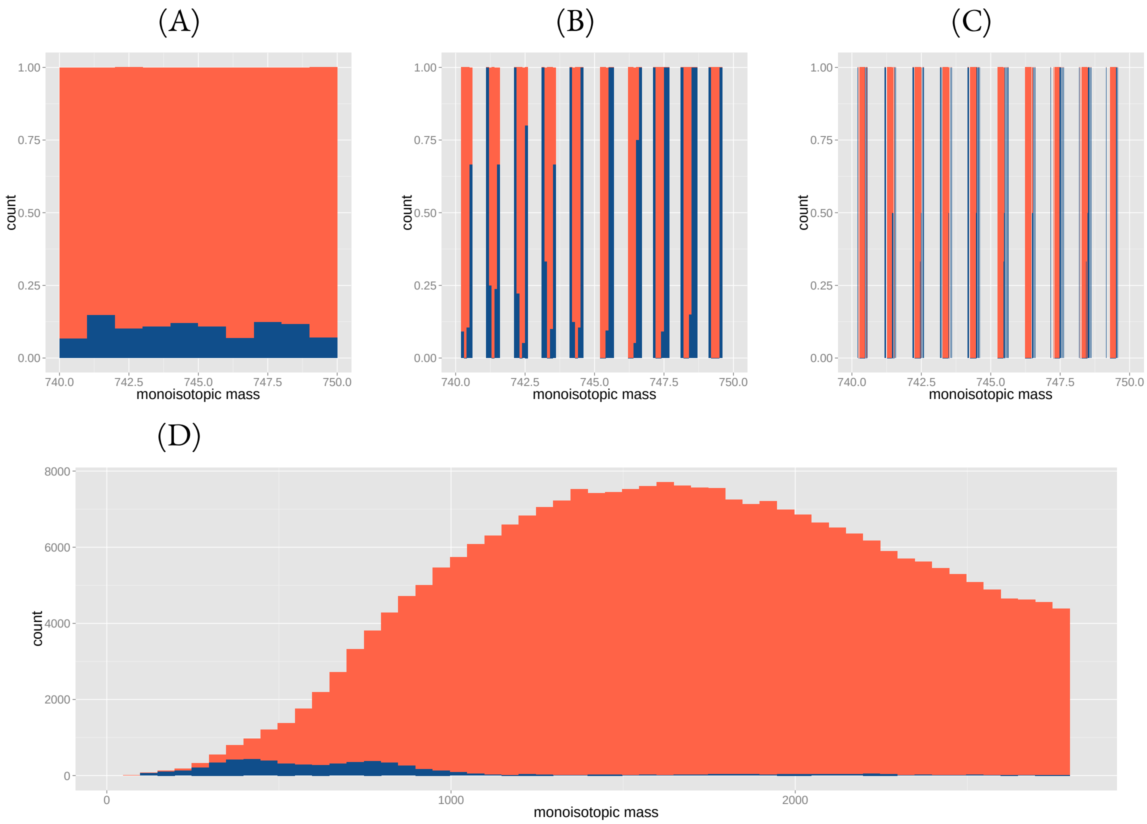

Figure 6.3: The normalized histograms showing the proportions between lipids (blue) and *in silico* digested peptides (red) for monoisotopic masses between 740 and 750 Da. The bin widths in panels (A), (B), (C) correspond to 1 Da, 0.1 Da, and 0.01 Da, respectively. (D) The monoisotopic mass distribution for the analyzed data set.

equal 1 Da, no discrimination is observed (cf. Figure 6.3(A)). However, when we decrease the widths to 0.1 Da (cf.Figure 6.3(B)) or even to 0.01 Da (cf. Figure 6.3(C)), the more and more visible trend is revealed – the fractional parts of the monoisotopic masses tend to occupy different mass ranges.

Feature sets

The natural extending of the mass defect idea is to use more information derived from aggregated isotopic distributions and check its usefulness for the classification purposes. Therefore, we used BRAIN to obtain the first three aggregated peaks, i.e. their intensities (alternatively called peak heights or the probabilities of the aggregated variants) and center-masses. Then, we calculated the following values that would be further investigated as the decision-making features:

- *mass.1*: center-mass of first isotope peak (i.e. monoisotopic mass);
- *mass.2*: center-mass of second isotope peak;
- *mass.3*: center-mass of third isotope peak;
- *mass.frac.1*: fractional part of center-mass of first isotope mass;
- *mass.frac.2*: fractional part of center-mass of second isotope peak;
- *mass.frac.3*: fractional part of center-mass of third isotope peak;
- *mass.diff.21*: difference between second and first isotope center-masses;
- *mass.diff.32*: difference between third and second isotope center-masses;
- *iso.ratio.21*: ratio of intensities of second and first isotope peaks;
- *iso.ratio.31*: ratio of intensities of third and first isotope peaks.

All but last two features depend on the center-masses. The last two features are isotopic ratio of heights of the consecutive peaks. As already mentioned in Chapter 5, in the real data processing the normalization procedure is commonly used. Therefore, the fact that we consider the peak height ratios instead of the peak heights does not affect the analysis strongly. On the contrary, considering the peak height ratios eliminates from the analysis the multiplicative noise associated to the isotopic abundances.

Classifiers – *in silico* studies

Our aim is to produce a classifier applicable for the experimental MS data, which takes into account both measurement noise and resolution limits. On the other hand, BRAIN models the theoretical, aggregated isotopic distribution, i.e. both probabilities and center-masses are modeled exactly in infinite resolution mode. Of note, lipid-vs.-peptide separation is straightforward for two dimensional plots – cf. Figure 6.4 for the ideal (no noise modeled) situation.

To mimic the experimental outputs, we artificially added the noise and resolution limitations to the modeled data. In case of the center-masses, the inaccuracy origins mostly from the resolution limits, which we simulated by assuming that the mass is measured only to a given decimal digit. More precisely, we rounded center-masses obtained from BRAIN to k-the decimal digits ($k = 1, \dots, 5$) corresponding approximately to FTICR, Orbitrap, TOF, ion trap and quadrupole instruments resolution, respectively. Of note, this simulates only

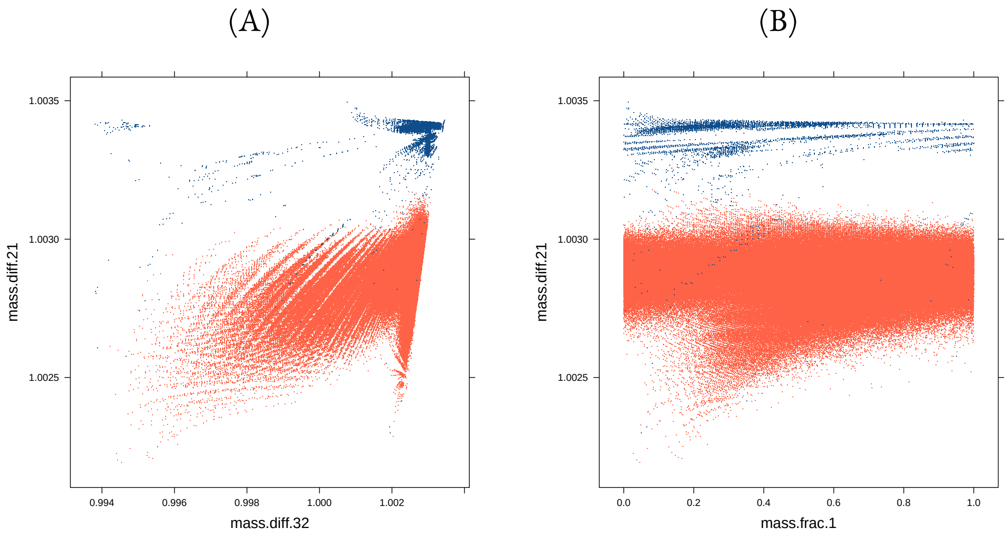

Figure 6.4: The simple two-dimensional visualization of the analyzed datasets with lipids (blue) and *in silico* digested peptides. The dimensions are (A) *mass.diff.21* vs. *mass.diff.32*, and (B) *mass.diff.21* vs. *mass.frac.1*

.

absolute mass errors, while for MS community the relative errors are often more informative. Therefore, we approximated the resolution by ppm ranges (according to masses of considered molecules) for each rounding (see Table 6.1 header).

For intensities, we multiplied the original probabilities (i.e. before calculating the peak ratios) by the Gaussian noise of mean 0 and standard deviation of 0.01, 0.1, 0.2, and 0.3. Theoretically, these normal distributions can take non-positive values. However, the probability of such situation is so small (e.g. for $\mathcal{N}(0, 0.3^2)$ the probability of non-negative value equals approximately 0.00043) that these cases were ignored in our experiments.

As a machine learning technique, operating on multidimensional data, we chose to use random forest (RF) classifier from (Breiman, 2001).

Random forest (RF)

Random forest (RF) is a classifier based on a bunch of decision trees that are constructed on a randomly sampled (with replacements) training sets. For each of these training sets, the tree is constructed, and RF makes a final decision by aggregating the single decision trees answers.

We run RF on each of the 30 data sets (6 levels of resolution for center-masses and 5 levels of noise; each combination possible). As a misclassification rate, we used out-of-bag measure

(OOB) (Narsky and Porter, 2013).

Out-of-bag measure (OOB)

Out-of-bag measure is constructed while creating a set of decision trees in RF construction. The single tree is constructed based on sampling with replacements, which gives around $1 - e^{-1} \approx \frac{2}{3}$ of the original data as training set. The remaining data (around $\frac{1}{3}$) are used as a test set to measure the misclassification error.

This measure is more informative when both classes are equally numerous (in other case we can image a situation when one class constitutes 99% of the data; then the blind classifier indicating always this class will have misclassification of 1% only), we sampled the subset of $6,313$ proteins (to check the stability of the classifier, we repeated this sampling procedure in selected places, which was then mentioned explicitly). We obtained the misclassification of 0.15% for ideal input (no rounding/noise added) up to almost 11% for least accurate data (cf. Table 6.1). In addition, we considered the reduced feature set, including only mass-derived features (i.e. all features but *iso.ratio.21* and *iso.ratio.31*). Of note, RF performs not much worse as in case of the full feature set with higher ($\sigma \geq 0.1$) normal noise modeled. On the other hand, in this case we do not have to worry about modeling inaccuracy of the intensity measurements.

Alternatively, to measure RF classifier performance we applied a 10-fold cross-validation scheme (Table 6.2), which includes the following steps:

1. dataset divided into 10 random (almost) equal parts; this is done for lipids and peptides sets independently;
2. repeat 10 times the following procedure (i.e. for $i = 1, \ldots, 10$);
   (a) build a classifier on test set with $9/10$ of dataset excluding i-th sets of lipids and peptides;
   (b) test a classifier on $1/10$ of dataset using i-th sets of lipids and peptides and return i-th misclassification rate;
3. at this point we obtain a vector of 10 misclassification rates – the result of 10-fold cross-validation is the mean of this vector.

The trends are similar to observed in Table 6.1, however, the standard deviation is larger.

Of note, a RF classifier also returns the ranking of feature importance. In other words, it provides a measure, called mean decrease in the Gini index, indicating which features revealed to be most useful in building the classification trees. We used this score to get more insights into the classifier performance, and run the following three tests.

approx. relative error for each column:	0 ppm	0.002-0.1 ppm	0.02-1.1 ppm	0.2-10.6 ppm	1.8-106.4 ppm	17.9-1063.5 ppm
sd of intensity noise	no mass rounding	mass rounding to 5th decimal digit	mass rounding to 4th decimal digit	mass rounding to 3rd decimal digit	mass rounding to 2nd decimal digit	mass rounding to 1st decimal digit
complete feature set						
0	0.15 (0.02)	0.14 (0.03)	0.18 (0.02)	2.81 (0.08)	*4.34* (0.18)	5.18 (0.23)
0.01	0.15 (0.03)	0.15 (0.02)	0.18 (0.03)	3.34 (0.14)	*5.45* (0.15)	6.24 (0.28)
0.1	0.19 (0.03)	0.17 (0.01)	0.20 (0.03)	5.44 (0.09)	*9.25* (0.36)	10.20 (0.34)
0.2	0.18 (0.03)	0.19 (0.01)	0.19 (0.03)	5.68 (0.11)	*9.74* (0.31)	10.69 (0.33)
0.3	0.19 (0.03)	0.17 (0.01)	0.19 (0.02)	5.66 (0.14)	*9.78* (0.29)	10.90 (0.31)
reduced feature set						
	0.16 (0.027)	0.14 (0.015)	0.19 (0.012)	5.57 (0.136)	10.27 (0.361)	10.78 (0.129)

Table 6.1: The misclassification out-of-bag measure (in percent) for the bunch of classifiers trained and tested on data with different precision, modeled bot for center masses (columns) and isotopic abundance (rows). In addition, each modeled precision for center-masses is estimated by corresponding relative error in ppm.

Table 6.2: The misclassification error (in percent) for the bunch of classifiers trained and tested on data with different precision, modeled bot for center masses (columns) and isotopic abundance (rows). The error is measured according to the 10-fold cross-validation scheme.

sd of intensity noise	no mass rounding	mass rounding to 5th decimal digit	mass rounding to 4th decimal digit	mass rounding to 3rd decimal digit	mass rounding to 2nd decimal digit	mass rounding to 1st decimal digit
complete feature set						
0	0.10 (0.10)	0.10 (0.11)	0.15 (0.14)	2.78 (0.28)	4.23 (0.59)	5.05 (0.67)
0.01	0.11 (0.09)	0.10 (0.11)	0.15 (0.14)	3.24 (0.54)	5.46 (0.61)	6.10 (0.59)
0.1	0.13 (0.11)	0.17 (0.13)	0.17 (0.15)	5.43 (0.58)	8.95 (0.85)	10.20 (0.59)
0.2	0.15 (0.09)	0.16 (0.12)	0.17 (0.16)	5.67 (0.53)	9.59 (0.60)	10.52 (0.79)
0.3	0.16 (0.14)	0.17 (0.13)	0.17 (0.18)	5.88 (0.51)	9.73 (0.66)	10.63 (0.78)
reduced feature set						
	0.063 (0.082)	0.063 (0.082)	0.063 (0.082)	5.591 (0.806)	9.093 (0.920)	6.986 (1.045)

1. We considered a full feature set. For each of the 6 considered resolution limits on center-masses and no noise modeled on isotopic intensities, we built a RF classifier and measured the feature importance (cf. Figure 6.5(A)). In this case, for high resolutions the most influential were *mass.diff.21* and *mass.diff.32*. However, when the mass is rounded to less than 3 decimal digits, these features become completely uninformative (which is not surprising, as center-mass differences presented in Figure 6.4 span a range of less than 0.01 Da), while *mass.1* and *mass.2* second center-masses become the most important features.

2. We considered a full feature set and the center-masses rounded to second decimal digit (to mimic MALDI measurement which will be further used as a validation for the real MS data). We multiplied the isotopic intensities by the normal noise of mean 0 and σ varying from 0 to 0.3. In general, the isotopic ratios and the exact center-masses were the most informative in a decision making process. However, the importance of the isotopic ratios decreased when the noise was higher.

3. We considered a reduced feature set (all features but *iso.ratio.21* and *iso.ratio.31*). We modeled various levels of resolutions and observed similar effects as described in point 1.

The next step in our *in silico* studies was to build a classifier aimed for distinguishing between eight lipids classes. The proof-of-concept visualization for two dimensional space (cf. Figure 6.6) suggests at least partial usefulness of this approach. E.g. we see that polyketides (PK) and glycolipids (GL) tend again to appear in different parts of the two dimensional plot. Of note, it is straightforward to use random forests, as they are easily applicable for multi-classification tasks. The train set consisted of $6,313$ lipids considered in previous step. The overall misclassification rate (OOB measure) of the classifier is $> 30\%$. However, when we analyze the confusion matrix for each of the eight classes separately (cf. Table 6.3), we observe that for the three most numerous classes (GP, PK, SP; over 1150 entries in each of these class; over 60% entries in total), the misclassification rate was smaller than 17%. Therefore, the RF classifier can potentially bring supporting information in a decision making process.

Classifiers – tests on real MS data

Finally, we run our classifier on the experimental MS data. To this aim, we utilized MALDI-TOF MS measurements performed on a lipid/peptides mixture. Using a reference list of a known substances within a mixture (cf. Table 6.5), we found in our data six molecules – four peptides and two lipids (cf. Figure 6.7 and Table 6.4). Then, we used the RF classifier trained on theoretical data, as described above. Namely, reduced feature set based on Lipid Maps and *in silico* digested Uniprot entries (training set consisted of $6,313$ lipids and the same number of the randomly drawn peptides) was produced and center-masses were rounded to the

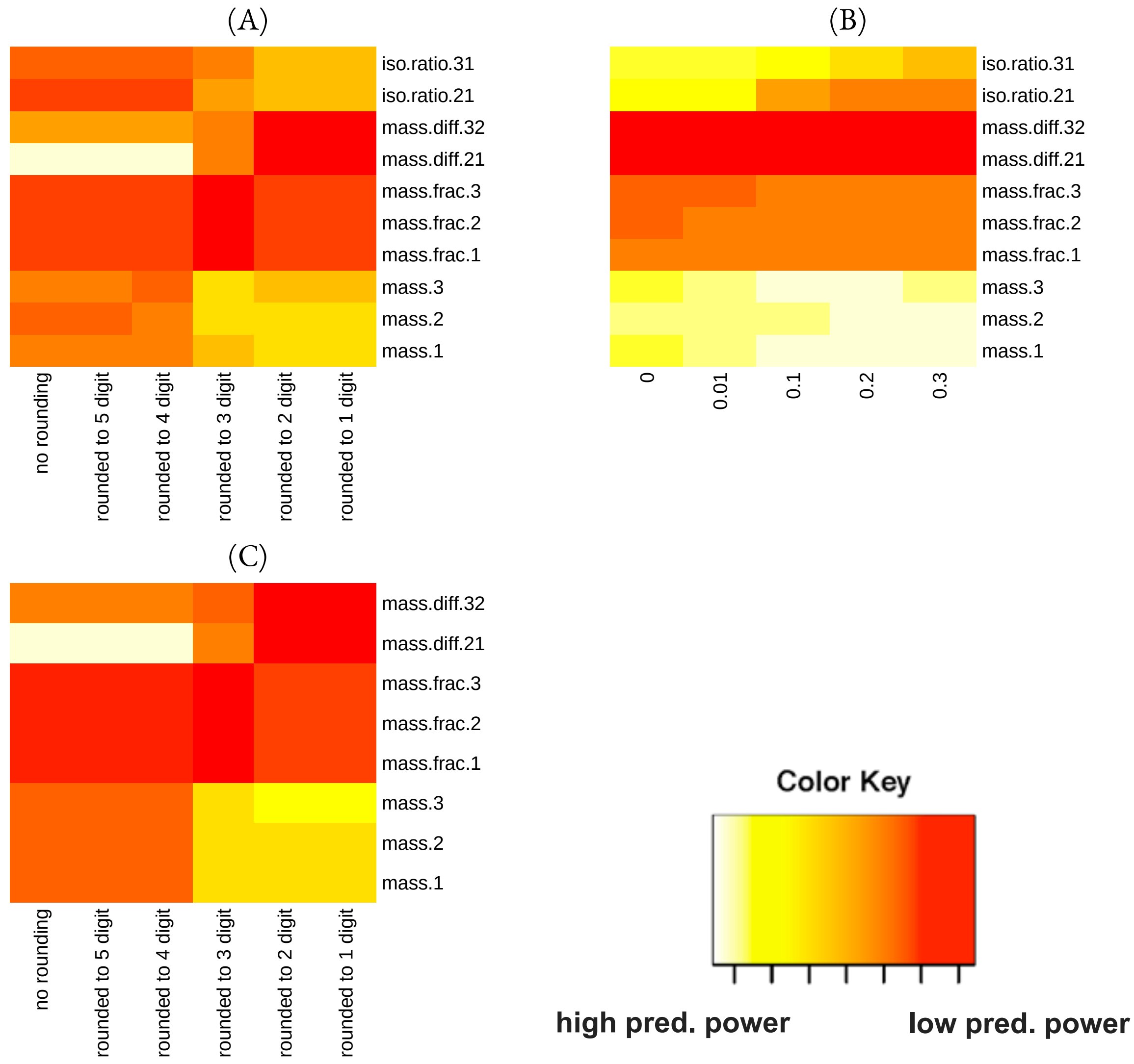

Figure 6.5: Heatmaps depicting the importance of given feature, according the mean decrease in Gini index. Each column should be considered independently. Full feature set for different precision modeled for (A) center-masses (intensities are exact) and (B) intensities (center-masses rounded to 2-nd decimal digit). (C) Reduced feature set for different precision modeled for center-masses.

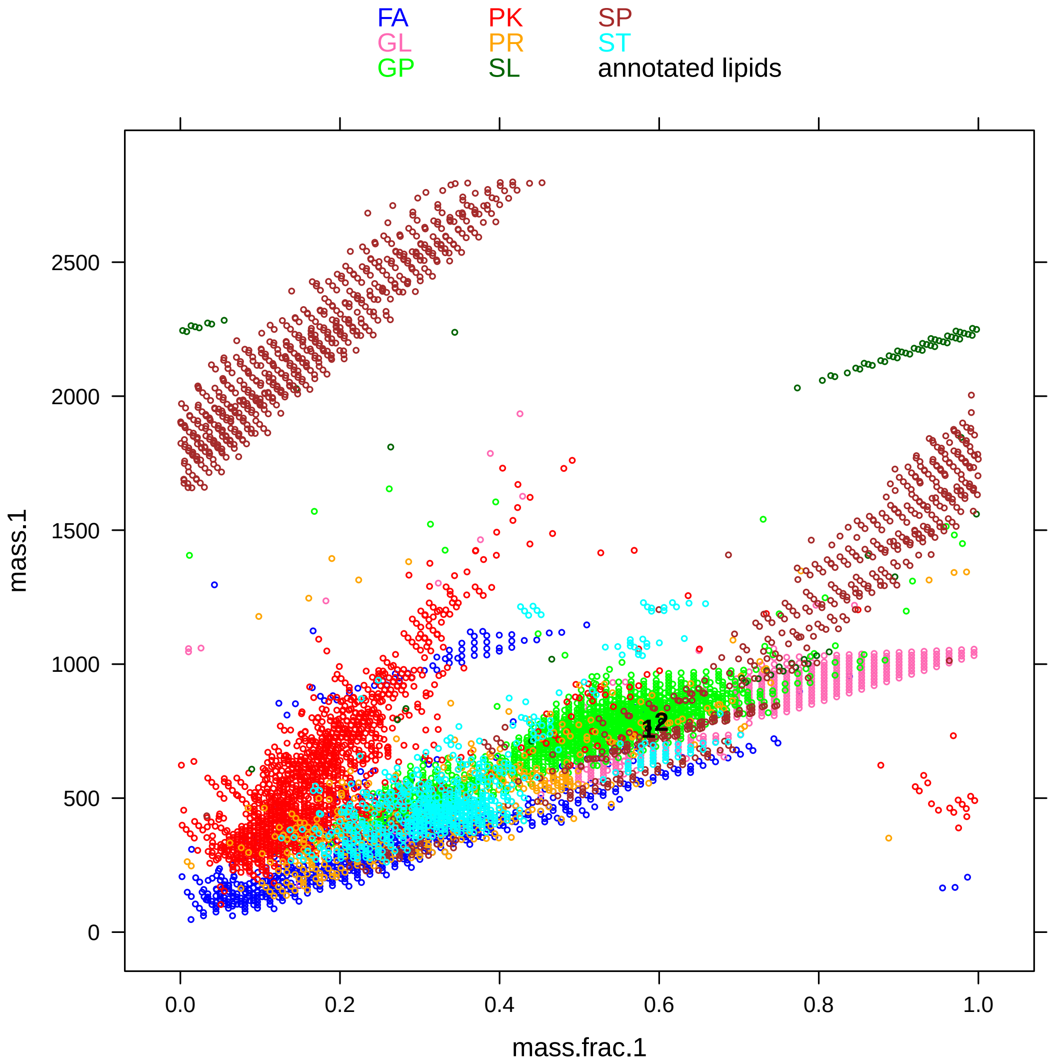

Figure 6.6: Two-dimensional plots with distribution of the analyzed data sets with eight classes of lipids, where dimensions are *mass.1* and *mass.frac.1*. The two analyzed molecules are denoted in black according to their labels from Table 6.4.

Table 6.3: Confusion matrix for within-lipid classification. The numbers indicate how many species from lipids class in row have been assigned to the class in column; last column indicates the misclassification (OOB) error.

	FA	GL	GP	PK	PR	SL	SP	ST	$\sum_{row}$	class.error (%)
FA	531	16	24	60	102	0	31	149	913	41.8
GL	14	255	57	6	8	1	30	29	400	36.2
GP	11	48	1178	36	24	2	47	69	1415	16.8
PK	38	1	51	1133	29	0	4	40	1296	12.6
PR	151	10	69	67	44	2	13	86	442	90
SL	0	7	3	3	1	49	12	1	76	35.5
SP	54	26	82	3	7	3	974	18	1167	16.5
ST	150	31	101	49	67	0	5	201	604	66.7
$\sum_{column}$	949	394	1565	1357	282	57	1116	593	6313	total class.err: 30.9

second decimal digits. The classifier was then run on the real data. It should be mentioned, that RF provides not only a label decision (”lipid” or ”peptide”), but also a probability score p_l that the given data belong to ”lipid” class (the corresponding probability of belonging to ”peptide” class is simply defined as $p_p = 1 - p_l$). This probability is based on the decisions made by decision trees used to build RF classifier. As a result, we obtained the p_l for lipids no. 1 and 2 of 0.9874 and 0.9996, respectively. For peptides no. $3 - 6$ we obtained p_p of $0.1576, 0.9738, 0.9996, 0.8476$, respectively (the presented scores are averaged over 5 runs of the classification based on different subsets of peptides used in training set; the corresponding standard deviation, σ, equals 0.189 for molecule 3 and $\sigma < 0.017$ for molecules $1 - 2, 4 - 6$). For the majority of the molecules the classification is correct, however, for peptide no. 3 the value of p_p is surprisingly smaller than 0.5. To get some insights into the origin of this problem, we visualized the data as a two-dimensional plot (Figure 6.8) for coordinates *mass.1* and *mass.frac.1*. Indeed, even for a ideal situation (infinite resolution mode), the molecule no. 3 occupies a region on the border between lipids and peptides (however, the real RF classifier of course operates on a higher dimensional space, therefore this plot does not necessarily reflects the real causes of the weak classifier performance for molecule no. 3). In addition, we tested within-lipid classifier (trained on reduced feature set of lipids data, where center-masses were rounded to second decimal digits). Both molecules 1 and 2 were correctly classified as glycerophospholipids (GP) with a probability of 98% and 82.7%, respectively.

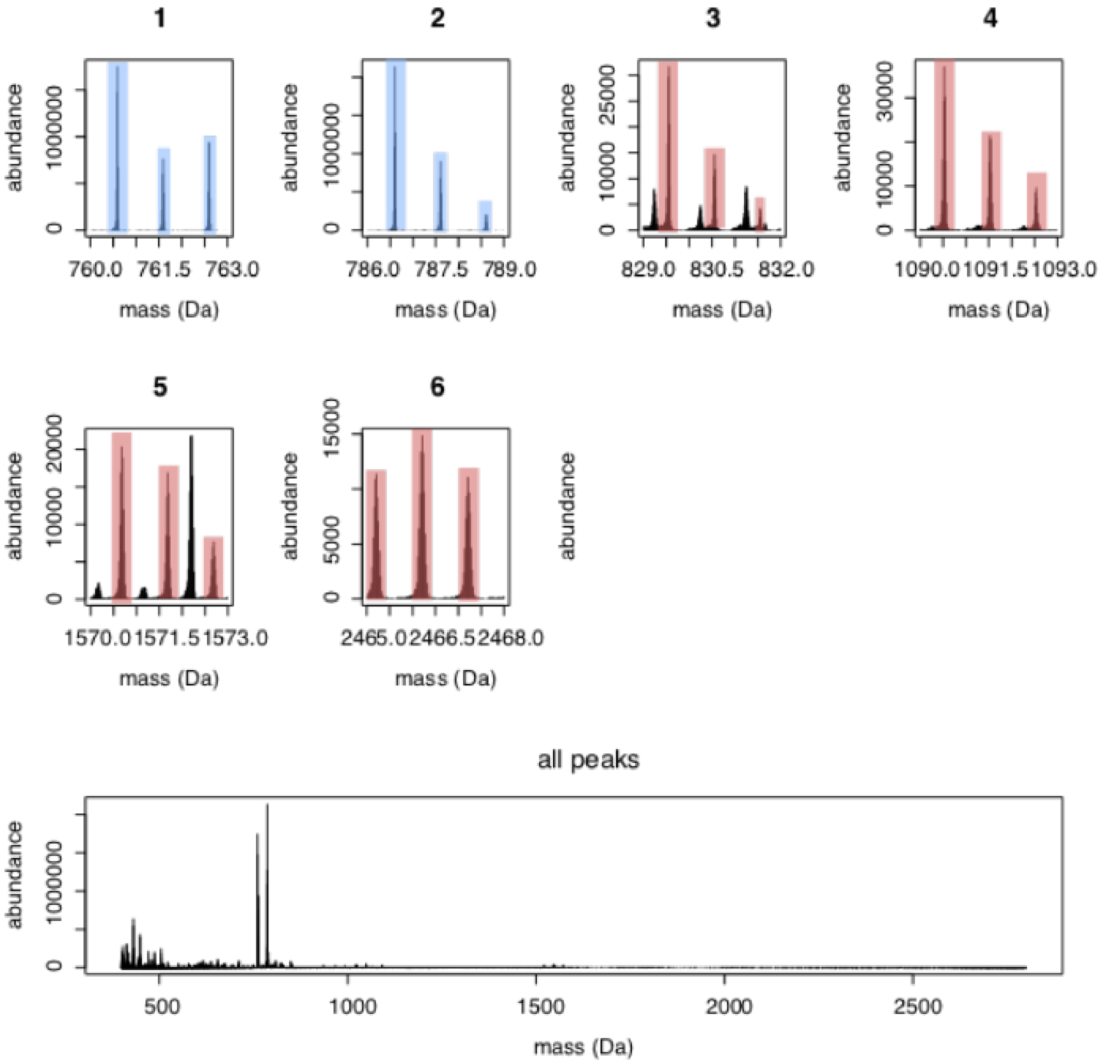

Figure 6.7: The plots with the total raw MS data and zoom-ins to the regions occupied by six molecules found within a mixture (lipids in blue; peptides in red).

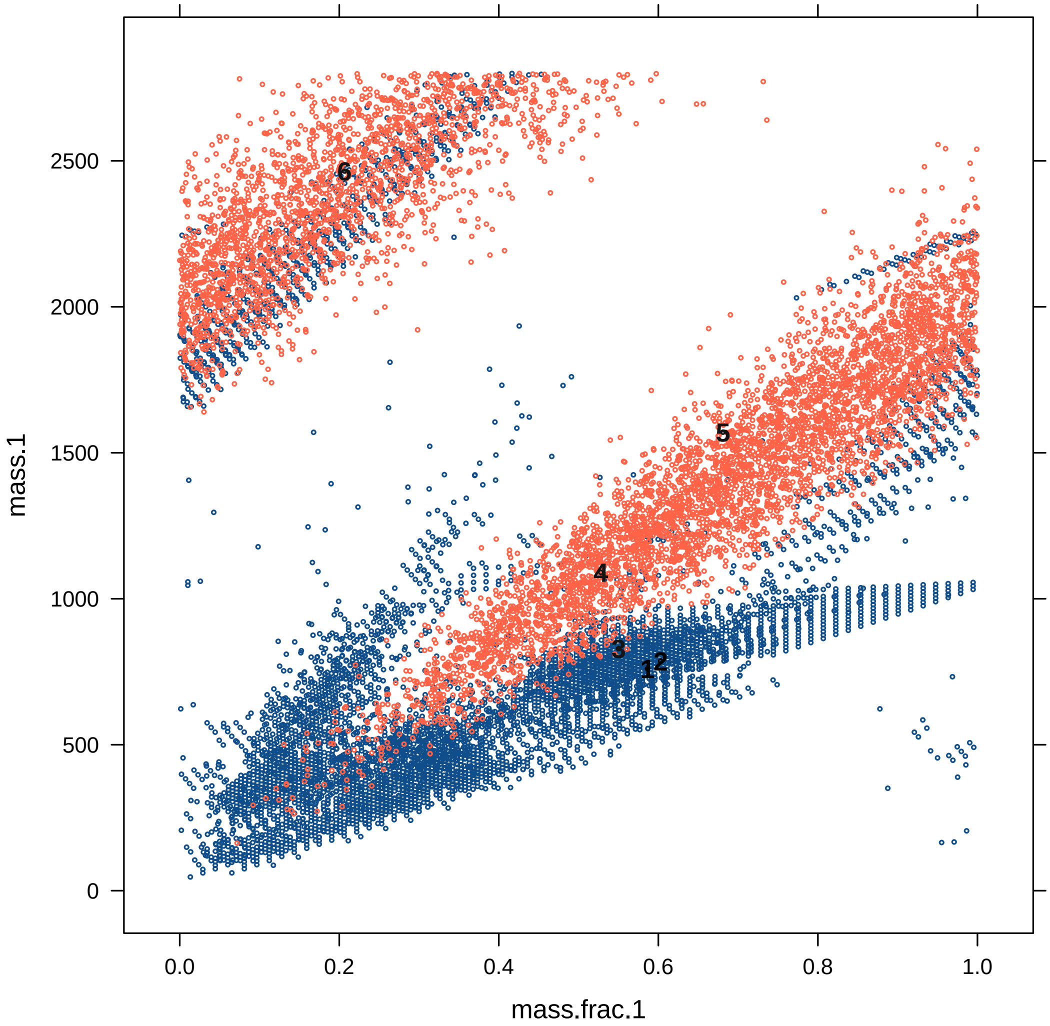

Figure 6.8: Two-dimensional plots with distribution of the analyzed data sets with lipids (blue) and *in silico* digested peptides, where dimensions are *mass.1* and *mass.frac.1*. The six analyzed molecules are denoted in black according to their labels from Table 6.4.

Table 6.4: Six molecules (two lipids and four peptides) found in our real MS dataset.

no.	name	type	formula	m_o			$10^6 \times \frac{m_t - m_o}{m_o}$			iso_o		$iso_t - iso_o$	
				mass.1	mass.2	mass.3				iso.ratio.21	iso.ratio.31		
1	PC 18:1(9z)/16:0	lipid	$C_{42}H_{83}N_1O_8P_1$	760.5864	761.5899	762.6012	-1.0176	-1.2198	-12.0718	0.4383	0.5406	0.0323	-0.41589
2	PC 18:1(9z)/18:1(9z)	lipid	$C_{44}H_{85}N_1O_8P_1$	786.6029	787.6059	788.6091	-2.0544	-1.6214	-1.7385	0.4234	0.0958	0.0690	0.03930
3	PKC substrate	peptide	$C_{34}H_{69}N_{16}O_8$	829.5501	830.5521	831.5536	-2.07462	-1.29673	-0.23811	0.4637	0.1379	-0.0259	-0.02767
4	ACTH 4-11	peptide	$C_{50}H_{72}N_{15}O_{11}S_1$	1090.5279	1091.5308	1092.5339	-2.0880	-2.1639	-4.1225	0.5795	0.2615	0.0372	-0.00658
5	Glu1 Fibrinopeptide B	peptide	$C_{66}H_{96}N_{19}O_{26}$	1570.6810	1571.6838	1572.6852	-2.3047	-2.2854	-1.4752	0.8284	0.3794	-0.0235	-0.00599
6	ACTH 18-39	peptide	$C_{112}H_{166}N_{27}O_{36}$	2465.2065	2466.2102	2467.2114	-3.1081	-3.3878	-2.7322	1.3070	0.9737	0.0369	-0.00342

Table 6.5: The table with molecules expected to be within the analyzed lipid/peptide mixture.

Peptides					
Name	concentration in pmol/μL	formula	Avg.MW	mono mass	mono mass with H^+
Kemptide	2.15	$C_{32}H_{61}N_{13}O_9$	771.918	771.472	772.479
PKC substrate	2.1	$C_{34}H_{68}N_{16}O_8$	829.016	828.541	829.548
ACTH 4-11	1.55	$C_{50}H_{71}N_{15}O_{11}S$	1090.268	1089.518	1090.526
Glu1 Fibrinopeptide B	1.05	$C_{66}H_{95}N_{19}O_{26}$	1570.592	1569.67	1570.677
ACTH 18-39	0.7	$C_{112}H_{165}N_{27}O_{36}$	2465.701	2464.191	2465.199
Lipids					
PC 18:1(9z)/16:0	0.95	$C_{42}H_{82}NO_8P$	760.076	759.578	760.586
PC 18:1(9z)/18:1(9z)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601
PC 16:0/18:0	0.95	$C_{42}H_{84}NO_8P$	762.092	761.593	762.601
PC 18:1 (9trans)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601
PC 18:0/16:0	0.95	$C_{42}H_{84}NO_8P$	762.092	761.593	762.601
PC 16:0/18:1(9z)	0.95	$C_{42}H_{82}NO_8P$	760.076	759.578	760.586
PC 18:1(6z)/18:1(6z)	0.95	$C_{44}H_{84}NO_8P$	786.113	785.593	786.601

7
Isotopic fine structure

In Chapters 4-6, we considered the aggregated isotopic variants. However, when the mass spectrometry resolution increases, we can distinguish several fine peaks. In fact, the experimentalists spent their funds for the instruments with the high-resolution functionality, and they do not want to aggregate them back. However, the actual fine structure of the aggregated peak seems to be very complicated (cf. Figure 7.1), and its huge size prevents from accurate representation. On the other hand, it is useful to analyze not only the center-masses but also the other parameters, such as the spread of the fine distribution for a given aggregated variant. As a consequence, additional questions might arise, e.g. what are the limitations (if they exist) when the consecutive aggregated peaks overlap.

Let us remind that isotopic fine structure distinguishes variants with different molecular mass. In particular, we consider separately the variants composed of different numbers of each of the stable isotopes, while summaric chemical formula ($C_vH_wN_xO_yS_z$) for all these variants remains the same. We would concentrate on the fine distribution for the given aggregated variants. More precisely, we would consider the most abundant aggregated variants, as those are of the most practical significance.

7.1 Variance of the fine distribution

We have already shown in Chapter 4 how to calculate the first moment (expected value) of the fine structure for given aggregated variants, which is called center-mass. Here, using analo-

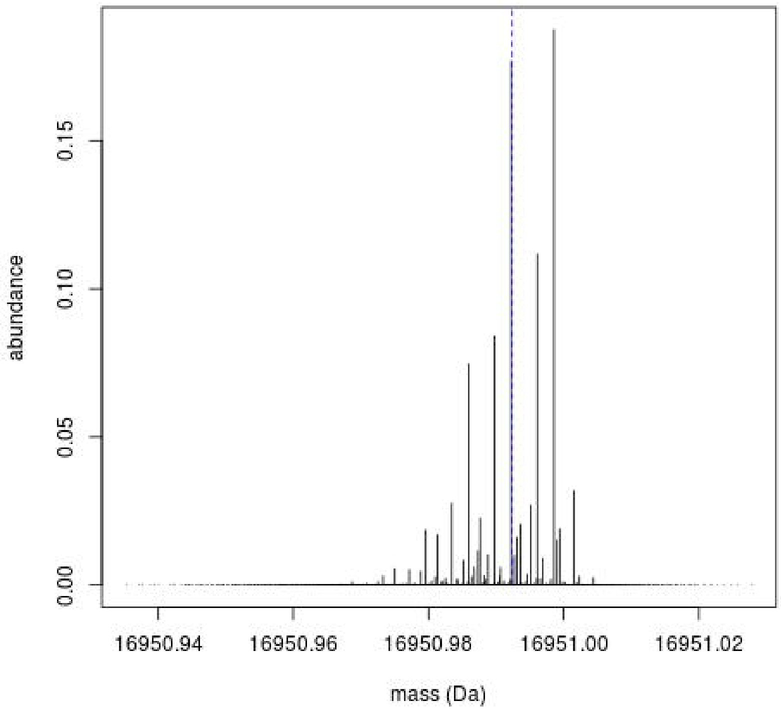

Figure 7.1: The fine structure of the most abundant aggregated peak (its center-mass is depicted with dotted line) for apomyoglobin, for which chemical formula is $C_{769}H_{1212}N_{210}O_{218}S_2$. The fine structure is generated using `isoDalton` software (Snider, 2007).

gous reasoning, we will introduce the generating function for the second moment (variance), referring to the distribution variability from its mean. First, let us remind the basic formula for the variance for j-th aggregated variant:

$$\mathrm{Var}(m_j) = E(m_j^2) - E(m_j)^2. \tag{7.1}$$

Of note, the value of $E(m_j)^2$ can be calculated as the square of the center-masses obtained from the original BRAIN (cf. Equation 4.23). The remaining part can be expanded as:

$$E(m_j^2) = \frac{\sum_k m_{jk}^2 p_{jk}}{\sum_k p_{jk}} \tag{7.2}$$

where the denumerator, analogously as in Equation (4.2), is simply an aggregated isotopic distribution of the j-th aggregated variant, and thus can be also provided by the original BRAIN algorithm. The remainder is the numerator of the Equation (7.2), namely $\sum_k m_{jk}^2 p_{jk}$. We first introduce generating function for this problem:

$$T(I; v, w, x, y, z) = \sum_j \sum_k m_{jk}^2 p_{jk} I^j = \sum_j q_j^{\perp} I^j. \tag{7.3}$$

In addition, we define the polynomials:

$$R_C(I, J, K) = P_{C_{12}} J^{M_{C_{12}}} K^{C_{12}} + P_{C_{13}} J^{M_{C_{13}}} K^{C_{13}} I, \tag{7.4}$$

and

$$W_A^*(I) = \sum_j p_{A,j} m_{A,j}^2 I^j \tag{7.5}$$

for carbon; polynomials for other elements ($R_H(I)$, $\ldots$, $R_S(I)$, $W_H^*(I)$, $\ldots$, $W_S^*(I)$) are defined analogously. Moreover, we would use polynomials $W_C(I), \ldots, W_S(I)$ defined in Equation (4.23).

Let us consider:

$$\begin{aligned} Q^{\perp}(I, J, K; v, w, x, y, z) = \quad & R_C(I, J, K)^v \times R_H(I, J, K)^w \times R_N(I, J, K)^x \times \\ & \times R_O(I, J, K)^y \times R_S(I, J, K)^z, \end{aligned} \tag{7.6}$$

and its standard form:

$$Q^{\perp}(I, J, K; v, w, x, y, z) = \sum_j (\sum_k p_{jk} J^{m_{jk}} K^{m_{jk}}) I^j. \tag{7.7}$$

Polynomial $Q^{\perp}(I, J, K; v, w, x, y, z)$ can be differentiated over J and K, and then we can set $J = K = 1$:

$$\begin{aligned} \frac{\partial^2}{\partial J \partial K} Q^{\perp}(I, J, K; v, w, x, y, z)|_{J=K=1} &= \sum_j (\sum_k m_{jk}^2 p_{jk} J^{m_{jk}-1} K^{m_{jk}-1}) I^j |_{J=K=1} \\ &= \sum_j (\sum_k m_{jk}^2 p_{jk}) I^j = T(I; v, w, x, y, z) \end{aligned} \tag{7.8}$$

where the last equation follows from Equation (7.3). Alternatively, we can differentiate polynomial $Q^{\perp}(I, J, K; v, w, x, y, z)$ over J and K using Equation (7.6), i.e. by applying the formula of differentiation a product, then set $J = K = 1$, and obtaining a final result:

$$\begin{aligned} T(I; v, w, x, y, z) &= \frac{\partial^2}{\partial J \partial K} Q^{\perp}(I, J, K; v, w, x, y, z)|_{J=K=1} \\ &= v \times (v-1) \times Q(I; v-2, w, x, y, z) \times P_C(I)^2 + \\ &+ v \times w \times Q(I; v-1, w-1, x, y, z) \times P_C(I) \times P_H(I) + \\ &+ v \times x \times Q(I; v-1, w, x-1, y, z) \times P_C(I) \times P_N(I) + \\ &+ v \times y \times Q(I; v-1, w, x, y-1, z) \times P_C(I) \times P_O(I) + \\ &+ v \times y \times Q(I; v-1, w, x, y, z-1) \times P_C(I) \times P_S(I) + \\ &+ v \times Q(I; v-1, w, x, y, z) \times P_C^*(I) + \star, \end{aligned} \tag{7.9}$$

where $\star$ replaces the summation involving 24 analogous products of polynomials.

Recall, that the variance can be obtained from Equation (7.1) by using coefficients of polynomial $T(I; v, w, x, y, z)$, center-masses and probabilities of the aggregated isotopic variants. We implemented calculation of $T(I; v, w, x, y, z)$ using two methods to multiply polynomials, both available in R: Fast Fourier Transform (function `fft`) and a standard library `PolynomF` for operations on polynomials. We decided not to use algebraic approach (i.e. BRAIN iterative formulae), as preliminary results showed that the high complexity of Equation (7.9), i.e. a lot of summations and multiplications, does involve numerical errors in practice. The comparison of the two methods is depicted in Figure 7.2. In addition, we observed that the results correspond to the values estimated by `isoDalton` software (Snider, 2007) (cf.Figure7.2(D)).

Furthermore, we processed the Uniprot database and built a linear model to analyze the relationship between the center-mass of the most abundant peak (m_a) and the variance of this center-mass (the visualization suggests a linear trend between the variance and the shift between most abundant and monoisotopic masses, cf. Figure 7.3). As a result, we obtained

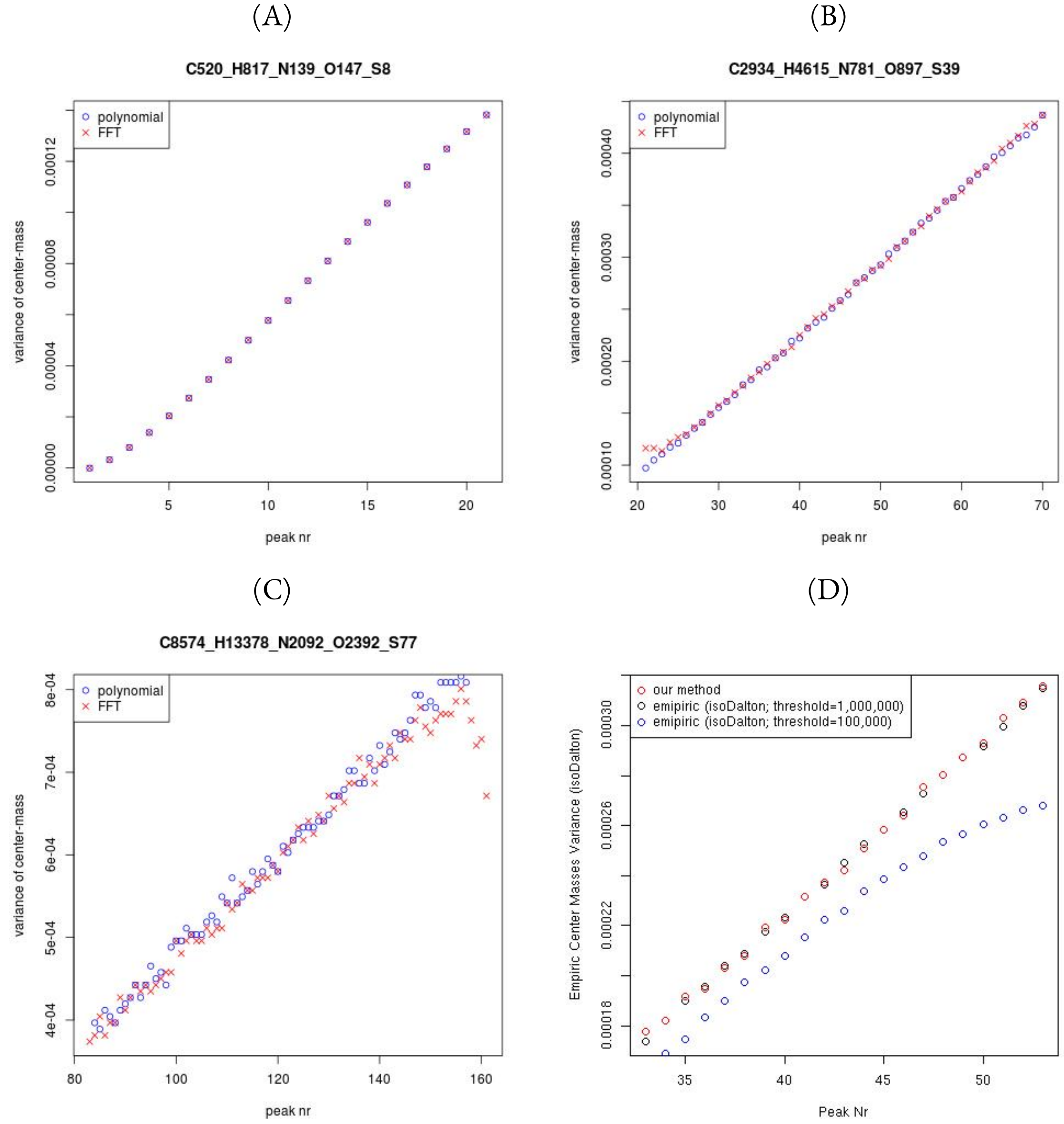

Figure 7.2: The variance of the fine structure of the most abundant aggregated variants for (A) Human insulin, (B) Bovine serum albumin, (C) Renal isoform, subunit Human ATP binding cassette protein (cf. Table 4.2). In addition to the FFT approach, we calculated the Equation 7.9 using the R library *PolynomF* for operations on polynomials. (D) As another method of the validation, we calculated the variance of Bovine serum albumin using `isoDalton` using different parameters for the number of generated peaks (note, this software generates the fine structure of the whole distribution). We observe, that if this parameter is big enough, the results returned by FFT approaches using Equation 7.9 are consistent with those returned by `isoDalton`.

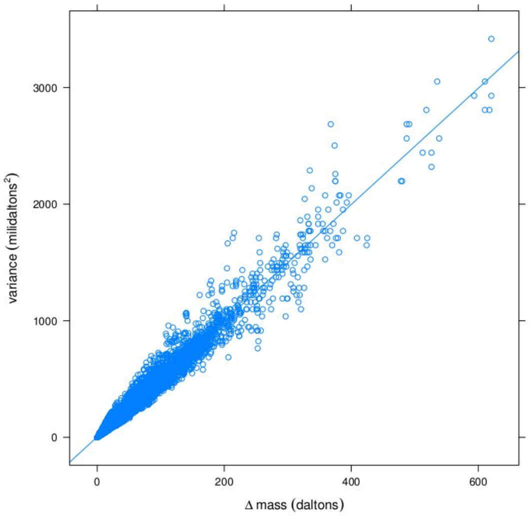

Figure 7.3: The relationship between the variance of the most abundant center mass, and the mass shift between most abundant and monoisotopic peak, calculated for the molecules from Uniprot database. The linear trend can be observed.

the formula

$$\text{variance} = 1.503 \times 10^{-6} + 3.077 \times 10^{-9} \times m_a \tag{7.10}$$

where both coefficients have p-values below 10^{-16}.

7.2 Information theory entropy

The information theory entropy is a measure of the (un)certainty of the random variable of given distribution.

Information theory entropy

Information theory entropy for a discrete random variable X with distribution function $P(X)$ is defined as:

$$H(X) = -E[\log(P(X))]. \tag{7.11}$$

For j-th aggregated variant the information theory entropy, denoted here as $H(j)$, can be calculated as follows (first equation is an application of the Equation (7.11)):

$$\begin{aligned}
H(j) &= -\sum_k \frac{p_{jk}}{\sum_k p_{jk}} \log(\frac{p_{jk}}{\sum_k p_{jk}}) = \frac{-\sum_k p_{jk} \log(\frac{p_{jk}}{\sum_k p_{jk}})}{\sum_k p_{jk}} \\
&= \frac{-\sum_k p_{jk}\{\log(p_{jk}) - \log(\sum_k p_{jk})\}}{\sum_k p_{jk}} \\
&= \frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}} + \frac{\sum_k p_{jk} \log(\sum_k p_{jk})}{\sum_k p_{jk}} \\
&= \frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}} + \frac{(\sum_k p_{jk}) \log(\sum_k p_{jk})}{\sum_k p_{jk}} \\
&= \frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}} + \log(\sum_k p_{jk}) \qquad (7.12)
\end{aligned}$$

Surprisingly, the $\frac{-\sum_k p_{jk} \log(p_{jk})}{\sum_k p_{jk}}$ can be calculated using the Equation (4.2), where m_{jk} is replaced with $-\log p_{jk}$. Moreover, $\log(\sum_k p_{jk}) = \log(q_j)$, where q_j is a probability of j-th aggregated isotopic variant, so the second term in formula for $H(j)$ can be calculated using original BRAIN. As a result, the information theory entropy can be effectively obtained.

7.3 Overlap between the consecutive aggregated variants

The natural question which we can consider when analyzing the aggregated variants is a problem of overlap between the consecutive peaks. As we already noticed in the variance analysis, this value increases for the most abundant peaks with higher molecular sizes. First, we can apply the following theorem:

Chebyshev's inequality

For a random variable X with $E(X) = \mu < \infty$ and $sd(X) = \sigma$, and for any $k \in \mathbb{R}_{>0}$:

$$Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \qquad (7.13)$$

In particular, for $k = 3$ from Chebyshev's inequality, we obtain that approximately 88.9% of a given distribution is within $\pm 3\sigma$ margin from its (finite) mean.

Therefore, using model from Equation (7.10), we can check when $3\sigma > 0.5$:

$$\begin{aligned} 3\sigma > 0.5 \quad &\Leftrightarrow \quad \sigma > \frac{0.5}{3} \\ &\Leftrightarrow \quad variance > (\frac{0.5}{3})^2 \\ &\Leftrightarrow \quad 1.503 \times 10^{-6} + 3.077 \cdot 10^{-9} \times m_a > (\frac{0.5}{3})^2 \\ &\Leftrightarrow \quad 3.077 \times 10^{-9} \times m_a > (\frac{0.5}{3})^2 - 1.503 \times 10^{-6} \qquad (7.14) \end{aligned}$$

As a result, we obtain that $m_a \approx 9$ MDa, and this mass seems to be huge. However, Wang et al. (2012) published the article with a meaningful title (”Increasing the trapping mass range to m/z$= 10^9$ – A major step toward high resolution mass analysis of intact RNA, DNA and viruses”), being a clear signal that MS processing of mega- or even gigadalton particles is not a purely theoretical consideration.

Another approach to the problem of overlapping aggregated variants is to investigate the maximal/minimal mass of the fine peaks within a given variant. Taking into account the average mass per additional neutron (cf. Table 7.1), we can see that the lightest possible aggregated variant should have only ^{15}N heavy isotopes. By analogy, the heaviest possible aggregated variant should be purely composed of ^{2}H heavy isotopes. Then, for variant with j additional neutrons, the mass spread between these extreme masses equals:

$$j \cdot (\mu_{2H} - \mu_{15N}) = j \times 0.0092421 \text{ Da}. \qquad (7.15)$$

Of note, in cases where there are not enough nitrogens or hydrogens within a molecule, Equation (7.15) gives an upper bound for the mass spread. Using this approximation, we estimate that the spread would reach 1 Da for $j \approx 108$.

In summary, two alternative estimations for the overlap between the most abundant isotopic peaks when aggregated variants are considered. First, variance/standard deviation approach, uses the Chebyshev's inequality. However, without adequate approximations of the aggregated peak shape it is difficult to predict the most accurate value of the parameter k used in Equation (7.13). The second approach – based on mass spread approximations – is more conservative, as the extreme fine variants are very tiny and in practice not observable in a spectrum.

Table 7.1: The table with average mass per additional neutron calculated for all heavy (i.e. not the lightest) isotopic variants for carbon, hydrogen, oxygen and sulphur. We observe the highest value for ^2H, and the smallest value for ^{15}N.

isotope	average mass per additional neutron (Da)
^{17}C	1.003355
^{2}H	1.006277
^{15}N	0.9970349
^{17}O	1.004217
^{18}O	1.002123
^{33}S	0.9993877
^{34}S	0.9978980
^{36}S	0.9987525

7.4 Distribution shape

The final step of our the analysis would be to asses the fine distribution deviance from normality. Of note, the distribution is multinomial, which however, has a bell-shape for *averagine* molecules. Therefore, a gaussian curve might be its good approximation. To compare these two distributions we use the relative entropy and the cross-entropy concepts.

Relative entropy

The relative entropy (also known as Kullback–Leibler divergence or Kullback–Leibler distance) between two distributions, P and Q, is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i) \tag{7.16}$$

Cross-entropy

For two distributions, P and Q the cross-entropy between them is defined as:

$$\mathrm{H}(P,Q) = \mathrm{H}(P) + D_{\mathrm{KL}}(P\|Q) \tag{7.17}$$

Of note, for $P = Q$ we obtain $D_{\mathrm{KL}}(P\|Q) = 0$ and $\mathrm{H}(P,Q) = \mathrm{H}(P)$. In our case, the first considered distribution is a fine distribution of the most abundant aggregated variants of the nine *averagines*, generated according to Equation (6.2). Namely, the proportions from Equation (6.2) were multiplied by $n = 50, 250, 500, 750, 1000, 2000, 3000, 4000, 5000,$

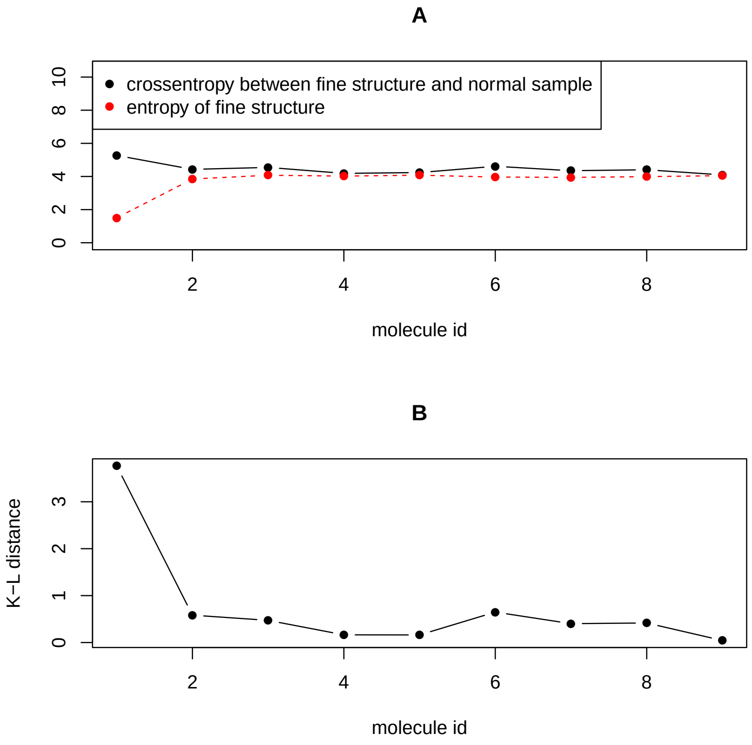

Figure 7.4: (A) Information theory entropy and the cross-entropy between the fine structure of the most abundant peaks of the analyzed averagines, and normal densities of mean and standard deviation as for fine structure. The bigger the molecule, the closer two values are.(B) The corresponding relative entropies which has (not strictly) descending trend.

and the obtained number of atoms were rounded to the integer values. This distribution P is simulated via simple Monte-Carlo (MC) approach – we are able to estimate its mean and variance from the simulated sample or from theoretical formulas presented already in this Chapter. Then, we sampled from the corresponding normal distribution Q of the same mean and variance to obtain the discretized normal distribution (we cannot compare explicitly discrete and continuous distributions with each other, and MC simulation obviously returns discretized result). We observe empirically that as the *averagine* size increases, the cross-entropy $\mathrm{H}(P, Q)$ resembles the $\mathrm{H}(P)$ (Figure 7.4(A)), while the relative entropy $D_{\mathrm{KL}}(P\|Q)$ tends to zero (Figure 7.4(B)). This suggests that, at least for large proteins, the normal distribution is a quite good approximation of the isotopic fine structure of the most abundant aggregated variant.

8
Further works and concluding remarks

In this dissertation, we presented a wide range of methods that can be applied in both mass spectrometry and genetics research involving large-scale data analyses. Here, we provide some perspectives in this field, including our ongoing projects.

Genome (in)stability caused by NAHR

Our results on NAHR prevalence (Figure 2.7) can be compared to the previous study made by Cooper et al. (2011), which involved over $15,000$ children with developmental delay tested by CMA in Signature Genomics Laboratories (SGL). Of note, the sets of six most common recurrent deletions in the two research (our and Cooper et al. (2011)) are consistent. Moreover, the investigation of *de novo* CNVs in $2,312$ patients with intellectual disabilities (ID) was performed by Girirajan et al. (2012), and revealed the high frequencies of deletions in 22q11.21 and 16p11.2 autism loci, which is similar to our observations (Figure 2.8). Additionally, to the results presented in Chapter 2, in our database we have found three somatic mosaicisms that are potentially mediated by NAHR (as franked by direct paralogous LCRs), and were confirmed by FISH analysis (Dittwald et al., 2013c). It should be noted, that also mitotic NAHR events have been suggested as a potential cancer cause-causing mechanism (Gu et al., 2008), and therefore are an interesting topic for future research. Also, next generation sequencing data can be used to systematically identify and analyze recurrent rearrangements (both meiotic and mitotic) previously missed by aCGH assays.

Furthermore, it has been already shown that NAHR can be also caused by other homologous elements, e.g. in Shuvarikov et al. (2013) we identified 3q13.2-q13.31 deletions mediated by Human Endogenous Retrovirus (HERV) elements. Moreover, we have already communicated a genome-wide map of potential genomic instability via HERVs as a poster during the ASHG conference - Piotr Dittwald, Ian M. Campbell *et al.*; Human Endogenous Retroviral Elements (HERVs) Mediate Multiple Genomic Rearrangements Suggestive of Nonallelic Homologous Recombination (NAHR), 63^{rd} American Society of Human Genetics Annual Meeting, Boston, October 2013. Moreover, my colleague from the University of Warsaw, Michał Startek, is working with BCM on similar studies involving long interspersed elements (LINEs). Of note, the UCSC Browser recently published a new version of genome build (hg38; December 2013), that can be considered (after some time needed for recalculating necessary data) as a new reference point for genome-wide maps for NAHR-prone regions.

MS workflow - its complexity and limitations

First, we should underline that their utility for the practical applications should be constantly considered. As mentioned in the introduction, the very important step in MS data analyses is the preprocessing step. First, it is done internally be the instruments, and unfortunately, the regular user has only a little influence (and in fact also a very poor knowledge) on the detailed procedures. Therefore, our models cannot assume that they operate on completely raw data. Secondly, we can have an influence onto steps such as baseline correction, smoothing, and peak picking, therefore a good understanding of the available algorithms might be useful in the further data processing (at the level of the aggregated distribution). Also, an awareness of limits, such as those investigated in Chapter 7, is helpful for accurate data modeling.

Better models for monoisotopic mass prediction

In Chapter 6, we showed the model for predicting the monoisotopic mass from the observed mass peaks. Although the model does not reveal a good accuracy, we already suggested it has a potential for a more adequate performance. Indeed, the better model, called MIND, i.e. MonoIsotopic mass liNear preDictor, was presented as a proof-of-concept in a poster at the ASMS conference in 2013 (Piotr Dittwald, Frederik Lermyte, Frank Sobott, Anna Gambin, Dirk Valkenborg; MIND: a soft-sensor to improve mass accuracy in high-resolution top-down proteomics. 61^{st} ASMS Conference on Mass Spectrometry and Allied Topics, Minneapolis, June 2013; DV was a poster presenter). It should be noted that this story is not yet published as a research study and needs some further verification.

Lipid centrifuge

The proposed approach for the lipid-vs.-peptide classification has to be further validated for more data samples. Moreover, developing better algorithms for retrieving the aggregated structure from raw data files is also a challenging task. Nevertheless, the Lipid Centrifuge workflow can potentially be an interesting alternative for the physicochemical fractionation techniques (e.g. liquid chromatography based methods) that are commonly used in mixture analysis, however, they introduce additional noise and variance into the measurements. Additional application of the method is in mass spectrometry imaging, where MS experiments are used to visualize the spatial distribution of the sample components (Stoeckli et al., 2001; Van de Plas et al., 2007; Van de Plas, 2010).

Effective modeling isotopic fine structure

As already mentioned in Chapter 7, the isotopic fine structure can be highly complex and therefore practically impossible for exact modeling. Recently, we have been working on developing an effective algorithm for simulating isotopic fine structure for given aggregated variant. The method called McFine is based on Monte-Carlo approach, and will be communicated as a poster during ASMS conference in 2014 (Piotr Dittwald, Dirk Valkenborg, Alan L. Rockwood, Anna Gambin; McFine - an algorithm to approximate the isotope fine structure of peptides and proteins, accepted as poster for 62^{nd} ASMS Conference on Mass Spectrometry and Allied Topics, Baltimore, June 2014).

Concluding words

As shown in this dissertation, the interdisciplinary approach is often inevitable in the biomedical studies. However, a need for deeper understanding of the analyzed problems has been experienced as a fascinating challenge by the author. Finally, the ethical issues that arise in the context of the research should be wisely considered, especially when dealing with a mystery of life.

References

Aebersold, R., Mann, M., 2003, Mass spectrometry-based proteomics, Nature, 422, 198–207

Albers, C. A., Paul, D. S., Schulze, H., Freson, K., Stephens, J. C., Smethurst, P. A., Jolley, J. D., Cvejic, A., Kostadima, M., Bertone, P., Breuning, M. H., Debili, N., Deloukas, P., Favier, R., Fiedler, J., Hobbs, C. M., Huang, N., Hurles, M. E., Kiddle, G., Krapels, I., Nurden, P., Ruivenkamp, C. A., Sambrook, J. G., Smith, K., Stemple, D. L., Strauss, G., Thys, C., van Geet, C., Newbury-Ecob, R., Ouwehand, W. H., Ghevaert, C., 2012, Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome, Nature Genetics, 44, 435–439

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., Eichler, E. E., 2002, Recent segmental duplications in the human genome, Science, 297, 1003–1007

Böcker, S., Letzel, M., Lipták, Z., Pervukhin, A., 2009, Sirius: decomposing isotope patterns for metabolite identification, Bioinformatics, 25, 218–224, http://bioinformatics.oxfordjournals.org/content/25/2/218.full.pdf+html

Bondeson, M. L., Dahl, N., Malmgren, H., Kleijer, W. J., Tonnesen, T., Carlberg, B. M., Pettersson, U., 1995, Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome, Human Molecular Genetics, 4, 615–621

Brancia, F. L., 2006, Recent developments in ion-trap mass spectrometry and related technologies, Expert Review of Proteomics, 3, 143–151

Breiman, L., 2001, Random forests, Machine Learning, 45, 5–32

Brownawell, M., Fillippo, J., 1982, A program for the synthesis of mass spectral isotopic abundances, Journal of Chemical Education, 59, 663–665

Bruce, C., Shifman, M. A., Miller, P., Gulcicek, E. E., 2006, Probabilistic enrichment of phosphopeptides by their mass defect, Analytical Chemistry, 78, 4374–4382

Brunetti-Pierri, N., Berg, J. S., Scaglia, F., Belmont, J., Bacino, C. A., Sahoo, T., Lalani, S. R., Graham, B., Lee, B., Shinawi, M., Shen, J., Kang, S. H., Pursley, A., Lotze, T., Kennedy, G., Lansky-Shafer, S., Weaver, C., Roeder, E. R., Grebe, T. A., Arnold, G. L., Hutchison, T., Reimschisel, T., Amato, S., Geragthy, M. T., Innis, J. W., Obersztyn, E., Nowakowska, B., Rosengren, S. S., Bader, P. I., Grange, D. K., Naqvi, S., Garnica, A. D., Bernes, S. M., Fong, C. T., Summers, A., Walters, W. D., Lupski, J. R., Stankiewicz, P., Cheung, S. W., Patel, A., 2008, Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities, Nature Genetics, 40, 1466–1471

Chandramouli, K., Qian, P. Y., 2009, Proteomics: challenges, techniques and possibilities to overcome biological sample complexity, Human Genomics and Proteomics, 2009

Chial, H., 2008, Cytogenetic methods and disease: Flow cytometry, CGH, and FISH, Nature Education, 1

Claesen, J., Dittwald, P., Burzykowski, T., Valkenborg, D., 2012, An efficient method to calculate the aggregated isotopic distribution and exact center-masses, Journal of the American Society for Mass Spectrometry, 23, 753–763

Cole, R., 1997, Electrospray ionization mass spectrometry: fundamentals, instrumentation, and applications., Wiley, New York

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B. C., Shaffer, L. G., Eichler, E. E., 2011, A copy number variation morbidity map of developmental delay, Nature Genetics, 43, 838–846

Cotter, R. J., 1994, Time-of-flight mass spectrometry., American Chemical Society, Columbus, OH

Cravatt, B. F., Simon, G. M., Yates, J. R., 2007, The biological impact of mass-spectrometry-based proteomics, Nature, 450, 991–1000

Crick, F., 1970, Central dogma of molecular biology, Nature, 227, 561–563

Dittwald, P., Valkenborg, D., 2014, BRAIN 2.0: Time and Memory Complexity Improvements in the Algorithm for Calculating the Isotope Distribution, Journal of the American Society for Mass Spectrometry, 25, 588–594

Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D., Gambin, A., 2013a, BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry, Analytical Chemistry, 85, 1991–1994

Dittwald, P., Gambin, T., Gonzaga-Jauregui, C., Carvalho, C. M., Lupski, J. R., Stankiewicz, P., Gambin, A., 2013b, Inverted low-copy repeats and genome instability–a genome-wide analysis, Human Mutation, 34, 210–220

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M. Y., Rodriguez Rojas, L. X., Elton, L. E., Scott, D. A., Schaaf, C. P., Torres-Martinez, W., Stevens, A. K., Rosenfeld, J. A., Agadi, S., Francis, D., Kang, S. H., Breman, A., Lalani, S. R., Bacino, C. A., Bi, W., Milosavljevic, A., Beaudet, A. L., Patel, A., Shaw, C. A., Lupski, J. R., Gambin, A., Cheung, S. W., Stankiewicz, P., 2013c, NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits, Genome Research, 23, 1395–1409

Eidhammer, I., Flikka, K., Martens, L., Mikalsen, S.-O., 2007, Computational Methods for Mass Spectrometry Proteomics, Wiley-Interscience

El-Hattab, A. W., Fang, P., Jin, W., Hughes, J. R., Gibson, J. B., Patel, G. S., Grange, D. K., Manwaring, L. P., Patel, A., Stankiewicz, P., Cheung, S. W., 2011, Int22h-1/int22h-2-mediated Xq28 rearrangements: intellectual disability associated with duplications and in utero male lethality with deletions, Journal of Medical Genetics, 48, 840–850

Elinati, E., Kuentz, P., Redin, C., Jaber, S., Vanden Meerschaut, F., Makarian, J., Koscinski, I., Nasr-Esfahani, M. H., Demirol, A., Gurgan, T., Louanjli, N., Iqbal, N., Bisharah, M., Pigeon, F. C., Gourabi, H., De Briel, D., Brugnon, F., Gitlin, S. A., Grillo, J. M., Ghaedi, K., Deemeh, M. R., Tanhaei, S., Modarres, P., Heindryckx, B., Benkhalifa, M., Nikiforaki, D., Oehninger, S. C., De Sutter, P., Muller, J., Viville, S., 2012, Globozoospermia is mainly due to DPY19L2 deletion via non-allelic homologous recombination involving two recombination hotspots, Human Molecular Genetics, 21, 3695–3702

Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R. H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M. J. O., Dennis, E. A., 2009, Update of the lipid maps comprehensive classification system for lipids., Journal of Lipid Research, 50 Suppl, S9–14

Fenn, J., 2002, Electrospray Wings for Molecular Elephants (Nobel Lecture), www.nobelprize.org, Nobel Foundation

Fernandez-de Cossio Diaz, J., Fernandez-de Cossio, J., 2012, Computation of Isotopic Peak Center-Mass Distribution by Fourier Transform, Analytical Chemistry, 84, 7052–7056

Gentleman, R. C., Carey, J. V., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., Zhang, J., 2004, Bioconductor: open software development for computational biology and bioinformatics, Genome Biology, 5(10), R80

Girirajan, S., Campbell, C. D., Eichler, E. E., 2011, Human copy number variation and complex genetic disease, Annual Review of Genetics, 45, 203–226

Girirajan, S., Rosenfeld, J. A., Coe, B. P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R. A., McConnell, J. S., Angle, B., Meschino, W. S., Nezarati, M. M., Asamoah, A., Jackson, K. E., Gowans, G. C., Martin, J. A., Carmany, E. P., Stockton, D. W., Schnur, R. E., Penney, L. S., Martin, D. M., Raskin, S., Leppig, K., Thiese, H., Smith, R., Aberg, E., Niyazov, D. M., Escobar, L. F., El-Khechen, D., Johnson, K. D., Lebel, R. R., Siefkas, K., Ball, S., Shur, N., McGuire, M., Brasington, C. K., Spence, J. E., Martin, L. S., Clericuzio, C., Ballif, B. C., Shaffer, L. G., Eichler, E. E., 2012, Phenotypic heterogeneity of genomic disorders and rare copy-number variants, The New England Journal of Medicine, 367, 1321–1331

Gross, M., Pramanik, B. N., Ganguly, A. K., 2002, Applied electrospray mass spectrometry., Marcel Dekker, New York

Gu, W., Zhang, F., Lupski, J. R., 2008, Mechanisms for human genomic rearrangements, Pathogenetics, 1, 4

Hastings, P. J., Ira, G., Lupski, J. R., 2009, A microhomology-mediated break-induced replication model for the origin of human copy number variation, PLOS Genetics, 5, e1000 327

Hernandez-Martin, A., Gonzalez-Sarmiento, R., De Unamuno, P., 1999, X-linked ichthyosis: an update, British Journal of Dermatology, 141, 617–627

Hu, H., Dittwald, P., Zaia, J., Valkenborg, D., 2013, Comment on the computation of isotopic peak center-mass distribution by fourier transform, Analytical Chemistry, 85, 12 189–12 192

Huang, N., Lee, I., Marcotte, E. M., Hurles, M. E., 2010, Characterising and predicting haploinsufficiency in the human genome, PLOS Genetics, 6, e1001 154

IHGSC, 2004, Finishing the euchromatic sequence of the human genome, Nature, 431, 931–945

Jacobson, N., 2007, Basic algebra 1, Dover

Kirchner, M., Timm, W., Fong, P., Wangemann, P., Steen, H., 2010, Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments, Bioinformatics, 26, 791–797

Klopocki, E., Schulze, H., Strauss, G., Ott, C. E., Hall, J., Trotier, F., Fleischhauer, S., Greenhalgh, L., Newbury-Ecob, R. A., Neumann, L. M., Habenicht, R., Konig, R., Seemanova, E., Megarbane, A., Ropers, H. H., Ullmann, R., Horn, D., Mundlos, S., 2007, Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome, The American Journal of Human Genetics, 80, 232–240

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., Marra, M. A., 2009, Circos: an information aesthetic for comparative genomics, Genome Research, 19, 1639–1645

Lakich, D., Kazazian, H. H., Antonarakis, S. E., Gitschier, J., 1993, Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A, Nature Genetics, 5, 236–241

Lee, J. A., Carvalho, C. M., Lupski, J. R., 2007, A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders, Cell, 131, 1235–1247

Liu, P., Carvalho, C. M., Hastings, P., Lupski, J. R., 2012, Mechanisms for recurrent and complex human genomic rearrangements, Current Opinion in Genetics & Development, 22, 211–220

Lupski, J. R., 1998, Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits, Trends in Genetics, 14, 417–422

Lupski, J. R., 2009, Genomic disorders ten years on, Genome Medicine, 1, 42

Makarov, A., 2000, Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis, Analytical Chemistry, 72, 1156–1162

Marshall, A. G., Hendrickson, C. L., Jackson, G. S., 1998, Fourier transform ion cyclotron resonance mass spectrometry: a primer, Mass Spectrometry Reviews, 17, 1–35

McElduff, F., Cortina-Borja, M., Chan, S. K., Wade, A., 2010, When t-tests or Wilcoxon-Mann-Whitney tests won't do, Advances in Physiology Education, 34, 128–133

Mefford, H. C., Sharp, A. J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V. K., Crolla, J. A., Baralle, D., Collins, A., Mercer, C., Norga, K., de Ravel, T., Devriendt, K., Bongers, E. M., de Leeuw, N., Reardon, W., Gimelli, S., Bena, F., Hennekam, R. C., Male, A., Gaunt, L., Clayton-Smith, J., Simonic, I., Park, S. M., Mehta, S. G., Nik-Zainal, S., Woods, C. G., Firth, H. V., Parkin, G., Fichera, M., Reitano, S., Lo Giudice, M., Li, K. E., Casuga, I., Broomer, A., Conrad, B., Schwerzmann, M., Raber, L., Gallati, S., Striano, P., Coppola, A., Tolmie, J. L., Tobias, E. S., Lilley, C., Armengol, L., Spysschaert, Y., Verloo, P., De Coene, A., Goossens, L., Mortier, G., Speleman, F., van Binsbergen, E., Nelen, M. R., Hochstenbach, R., Poot, M., Gallagher, L., Gill, M., McClellan, J., King, M. C., Regan, R., Skinner, C., Stevenson, R. E., Antonarakis, S. E., Chen, C., Estivill, X., Menten, B., Gimelli, G., Gribble, S., Schwartz, S., Sutcliffe, J. S., Walsh, T., Knight, S. J., Sebat, J., Romano, C., Schwartz, C. E., Veltman, J. A., de Vries, B. B., Vermeesch, J. R., Barber, J. C., Willatt, L., Tassabehji, M., Eichler, E. E., 2008, Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes, The New England Journal of Medicine, 359, 1685–1699

Mefford, H. C., Cooper, G. M., Zerr, T., Smith, J. D., Baker, C., Shafer, N., Thorland, E. C., Skinner, C., Schwartz, C. E., Nickerson, D. A., Eichler, E. E., 2009, A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease, Genome Research, 19, 1579–1585

Myers, S., Freeman, C., Auton, A., Donnelly, P., McVean, G., 2008, A common sequence motif associated with recombination hot spots and genome instability in humans, Nature Genetics, 40, 1124–1129

Narsky, I., Porter, F. C., 2013, Statistical Analysis Techniques in Particle Physics. Fits, Density Estimation and Supervised Learning, Wiley-VCH

Naylor, J., Brinke, A., Hassock, S., Green, P. M., Giannelli, F., 1993, Characteristic mRNA abnormality found in half the patients with severe haemophilia A is due to large DNA inversions, Human Molecular Genetics, 2, 1773–1778

Naylor, J. A., Green, P. M., Rizza, C. R., Giannelli, F., 1992, Factor VIII gene explains all cases of haemophilia A, Lancet, 340, 1066–1067

Naylor, J. A., Buck, D., Green, P., Williamson, H., Bentley, D., Giannelli, F., 1995, Investigation of the factor VIII intron 22 repeated region (int22h) and the associated inversion junctions, Human Molecular Genetics, 4, 1217–1224

O'Connor, C., 2008, Fluorescence in situ hybridization (FISH), Nature Education, 1

Olson, M., Yergey, A., 2009, Calculation of the isotope cluster for polypeptides by probability grouping, Journal of the American Society for Mass Spectrometry, 20, 295–302

Parsons, J. D., 1995, Miropeats: graphical DNA sequence comparisons, Computer Applications in the Biosciences, 11, 615–619

Peter-Katalinic, J.; Hillenkamp, F., 2007, MALDI MS: A Practical Guide to Instrumentation, Methods and Applications., Wiley-VCH

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 2007, Numerical Recipes: The Art of Scientific Computing, New York: Cambridge University Press

Rockwood, A., 1995, Relationship of fourier transforms to isotope distribution calculations, Rapid Communications in Mass Spectrometry, 9, 103–105

Rockwood, A., Haimi, P., 2006, Efficient calculation of accurate masses of isotopic peaks, Journal of the American Society for Mass Spectrometry, 17, 415–419

Rockwood, A., Van Orden, S., 1996, Ultrahigh-speed calculation of isotope distributions, Analytical Chemistry, 68, 2027–2030

Rockwood, A., Van Orden, S., Smith, R., 1995, Rapid calculation of isotope distributions, Analytical Chemistry, 67, 2699–2704

Rockwood, A., Van Orden, S., Smith, R., 1996, Ultrahigh resolution isotope distribution calculations, Rapid Communications in Mass Spectrometry, 10, 54–59, ISSN 1097-0231

Rosman, K., Taylor, P., 1997, Isotopic compositions of the elements 1997, Pure and Applied Chemistry, 70, 217–235

Senko, M., Beu, S., McLafferty, F., 1995a, Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions, Journal of the American Society for Mass Spectrometry, 6, 229 – 233

Senko, M. W., Beu, S. C., McLafferty, F. W., 1995b, Automated assignment of charge states from resolved isotopic peaks for multiply charged ions, Journal of the American Society for Mass Spectrometry, 6, 52–56

Séroul, R., 2000, Programming for Mathematicians, Berlin: Springer-Verlag

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., Eichler, E. E., 2005, Segmental duplications and copy-number variation in the human genome, The American Journal of Human Genetics, 77, 78–88

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., Fitzpatrick, C. A., Segraves, R., Richmond, T. A., Guiver, C., Albertson, D. G., Pinkel, D., Eis, P. S., Schwartz, S., Knight, S. J., Eichler, E. E., 2006, Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome, Nature Genetics, 38, 1038–1042

Shmakov, S., 2011, A universal method of solving quartic equations, International Journal of Pure and Applied Mathematics, pp. 251–259

Shuvarikov, A., Campbell, I. M., Dittwald, P., Neill, N. J., Bialer, M. G., Moore, C., Wheeler, P. G., Wallace, S. E., Hannibal, M. C., Murray, M. F., Giovanni, M. A., Terespolsky, D., Sodhi, S., Cassina, M., Viskochil, D., Moghaddam, B., Herman, K., Brown, C. W., Beck, C. R., Gambin, A., Cheung, S. W., Patel, A., Lamb, A. N., Shaffer, L. G., Ellison, J. W., Ravnan, J. B., Stankiewicz, P., Rosenfeld, J. A., 2013, Recurrent HERV-H-mediated 3q13.2-q13.31 deletions cause a syndrome of hypotonia and motor, language, and cognitive delays, Human Mutation, 34, 1415–1423

Snider, R., 2007, Efficient calculation of exact mass isotopic distributions, Journal of the American Society for Mass Spectrometry, 18, 1511–1515

Stankiewicz, P., Lupski, J. R., 2002, Genome architecture, rearrangements and genomic disorders, Trends in Genetics, 18, 74–82

Stankiewicz, P., Lupski, J. R., 2010, Structural variation in the human genome and its role in disease, Annual Review of Medicine, 61, 437–455

Stoeckli, M., Chaurand, P., Hallahan, D. E., Caprioli, R. M., 2001, Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues, Nature Medicine, 7, 493–496

Valkenborg, D., Mertens, I., Lemière, F., Witters, E., Burzykowski, T., 2012, The isotopic distribution conundrum, Mass Spectrometry Reviews, 31, 96–106

Van de Plas, R., 2010, Tissue Based Proteomics and Biomarker Discovery - Multivariate Data Mining Strategies for Mass Spectral Imaging, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium)

Van de Plas, R., Ojeda, F., Dewil, M., Van Den Bosch, L., De Moor, B., Waelkens, E., 2007, Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis, Pacific Symposium on Biocomputing, pp. 458–469

Vinberg, E. B., 2003, A course in algebra, American Mathematical Society, Providence

Vissers, L. E., Stankiewicz, P., 2012, Microdeletion and microduplication syndromes, Methods in Molecular Biology, 838, 29–75

Wang, X., Chen, H., Lee, J., Reilly, Peter, T., 2012, Increasing the trapping mass range to m/z $= 10^9$ – A major step toward high resolution mass analysis of intact RNA, DNA and viruses, International Journal of Mass Spectrometry, pp. 28–35

Watson, J. D., Crick, F. H., 1953, Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid, Nature, 171, 737–738

Yamamoto, H., McCloskey, J., 2012, The UniProt Consortium. reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Research, 40, D71–D75

Yamamoto, H., McCloskey, J. A., 1977, Calculations of isotopic distribution in molecules extensively labeled with heavy isotopes, Analytical Chemistry, 49, 281

Yang, C., He, Z., Yu, W., 2009, Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis, BMC Bioinformatics, 10, 4

Yates, J. R., Kelleher, N. L., 2013, Top down proteomics, Analytical Chemistry, 85, 6151

Yergey, J., 1983, A general approach to calculating isotopic distributions for mass spectrometry, International Journal of Mass Spectrometry and Ion Physics, 52, 337–349

Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., Scherer, S. W., 2006, Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome, Cytogenetic and Genome Research, 115, 205–214

Zhang, Y., De, S., Garner, J. R., Smith, K., Wang, S. A., Becker, K. G., 2010, Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information, BMC Medical Genomics, 3, 1

Zubarev, R. A., Makarov, A., 2013, Orbitrap mass spectrometry, Analytical Chemistry, 85, 5288–5296

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.