LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Bachelorarbeit
in Bioinformatik

Fatty acid pool reconstruction from metabolomics measurements of phosphatidylcholines

Benjamin Drexler

Aufgabensteller:	Prof. Fabian Theis
Betreuer:	Jan Krumsiek
Abgabedatum:	15.09.2010

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15.09.2010 ____________________
Benjamin Drexler

Abstract

Phosphatidylcholines (PCs) consist of a glycerol backbone, a phosphate group with an attached choline and two fatty acid (FA) resdiues. The mass spectrometry technology used for the determination of metabolite concentrations in the KORA and a mouse nutritional challenging study is not able to resolve the FA composition of the PCs. In this thesis, we implemented a method to calculate the FA frequencies which are incorporated in the biosynthesis of phosphatidylcholines. This calculation is based on the measured PC concentrations. The computed FA frequencies would help to understand the FA composition of measured PCs in detail. The computed FA frequencies show significant genetic associations with single nucleotide polymorphisms. It was also possible to detect the effects of a nutritional challenging of mice on the computed FA frequencies. These results support the hypothesis that the computed FA frequencies are meaningfulness with respect to the biological context. An examination of the computed and measured FA frequencies indicate that the free fatty acid pool does not correlate with the fatty acid pool, which is used in the biosynthesis of PCs.

Zusammenfassung

Phosphatidylcholine (PC) bestehen aus Glycerin, einer Phosphatgruppe, Cholin und zwei Fettsäure-Resten. Die Massenspektrometrie-Technik, welche in der KORA Studie und einer Studie mit Mäusen, die unter einer bestimmten Diät gehalten wurden, verwendet wurde, ist nicht in der Lage, die genaue Fettsäure (FS) Zusammensetzung der PCs zu bestimmen. Wir haben in dieser Thesis eine Methode entwickelt, um die einzelnen FS Häufigkeiten, welche bei der Biosynthese von PCs in diese eingebaut werden, zu berechnen. Diese Berechnung beruht auf den gemessenen PC Konzentrationen der jeweiligen Studie. Die berechneten FS Häufigkeiten würden dazu dienen, die FS Zusammensetzung der gemessenen PCs aufzuschlüsseln. Die berechneten FS Häufigkeiten zeigen signifikante genetische Assoziationen mit Einzelnukleotid-Polymorphismen. Außerdem war es möglich, die Auswirkungen der Diät in den berechneten FS Häufigkeiten zu erkennen. Diese Ergebnisse sind ein Indikator dafür, dass die berechneten FS Häufigkeiten als biologisch sinnvoll erachtet werden können. Der Vergleich von den berechneten FS Häufigkeiten und gemessenen Häufigkeiten von freien Fettsäuren lässt darauf schließen, dass der an freien Fettsäuren nicht mit dem an Fettsäuren, welche für die Biosynthese von Phosphatidylcholinen verwendet werden, korreliert.

Acknowledgements

I would like to thank Jan Krumsiek for constant supervision and assistance, and Prof. Fabian Theis for making this thesis possible.

Contents

Chapter 1

Introduction

In the following sections, we explain the fundamental biochemistry and the term *metabolomics* that will be required to understand the biological background of the problem statement.

1.1 Basic Biochemistry

The following two sections cover the basic biochemistry about fatty acids and phospholipids. This includes the structure, denotation, biochemical processes and biological functions.

1.1.1 Fatty Acids

Fatty acids (FA) are hydrocarbon chains with a carboxylic acid group at one end.[1] They vary in chain length and also in the degree of desaturation. These variations influence the fatty acid's properties, e.g. the melting point. A shorter chain or a higher desaturation grade leads to a lower melting point. Fatty acids have various biological functions. They are part of phospholipids and therefore are involved in biological membranes (see section 1.1.2). They are also fuel molecules in an organism's metabolism. Three FAs form ester bonds with the carboxyl groups of a glycerol. The resulting molecule is known as a triglyceride. Derivates of FAs also act as hormones and intracellular messengers.

The FA's carbon atoms are numbered starting at the carboxyl terminus. The second carbon atom is called α and the third is called β. The last carbon atom of the carbon chain is referred to as ω. The presentation for a double bound is *cis*/*trans*-Δ^n with n being the position of the first carbon atom of the both

[1] Fundamental biological backgrounds in this chapter were extracted from one biology textbooks: [1]

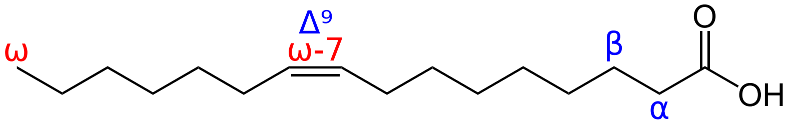

Figure 1.1: Representation of the fatty acid with the common name palmitoleic acid. An systematic denotation for this FA is *cis*-Δ^9-hexadecenoic acid. The second and third carbon atoms are referred to as α and β, that last one as ω. The denotation of the double bound is also shown, e.g. Δ^9 or ω-7.

carbon atoms involved in the double bound. Alternatively, a double bound can be described with the ω-n-denotation. This denotation begins the numeration at the ω carbon atom with 1 and the carbon double bounds is with the position of the first carbon atom, e.g. ω-3-fatty acid. In this work we generally refer a FA to its number of carbon atoms and double bounds. For instance, the FA with the common name palmitoleic acid has 16 carbon atoms and 1 double bound (see figure 1.1). This FA will be denoted with the term C16:1.

The FA synthesis occurs in the cytoplasm and is a cycle, which runs through four main steps. These steps are catalyzed by the enzyme fatty acid synthase (FAD) in human. FAD consists of several functional domains with one them being an acyl carrier protein (ACP) domain. A cycle of the synthesis begins with an aycl group and a malonyl group. In the first cycle, the acyl group is acetyl-CoA. That and malonyl-CoA are being transferred into a thiol linkage with the ACP domain of FAD. This results in the groups acetyl-ACP and malonyl-ACP, which are then condensated to form acetoacetyl-ACP. Step two is a reduction of the double bound of acetoacetyl ACP to a hydroxyl group resulting in β-hydroxylbutyryl-ACP. The third step is a dehydration between the C-2 and C-3 resulting in crotonyl-ACP. In the last step of the first cycle, crotonyl-ACP is reduced to butyryl-ACP. This would be used as the acyl group for the second cycle. Hence an elongation of the acyl group by two carbon atoms occurs in each cycle. The cycle is repeated until the acyl group contains 16 carbon atoms, which is then released as palmitic acid.

Palmitic acid provides the foundation for the elongation to longer fatty acids. The elongation process is catalyzed by elongases. These are proteins, which are associated with the endoplasmic reticulum on the cytosolic side. Malonyl-CoA here also serves as the donor of the two carbon atoms. Hence most natural FAs contain an even number of carbon atoms. The elongation is similar to the FA synthesis, but catalyzed by different protein.

The desaturation of FAs is catalyzed by the following enzymes: NADH-cytochrome b_5 reductase, cytochrome b_5 and desaturase. It is based on several electron transfers between these enzymes and substrates, e.g. NADH, O_2 and a fatty acyl CoA. In the first step, electrons are transferred from NADH to the FAD moiety of

NADH-cytochrome b_5 reductase. This leads to an reduction of the iron atom of cytochrome b_5 to the Fe^{2+} state. Subsequently, the iron atom of the desaturase is converted from the Fe^{3+} into the Fe^{2+} state. This enables the desaturase to interact with O_2 and the fatty acyl CoA substrate, which results in a double bond at the fatty acyl CoA and the formation of two H_2O molecules. The human organism can only perform the desaturation of Δ^5, Δ^6 and Δ^9 and not the desaturation of Δ^{12}. Therefore ω-3- and ω-6-fatty acids can be of particular importance, because some of them are essential FAs of the human. Thus the human has to take in a sufficient amount of these through food.

1.1.2 Phospholipids

Phospholipids form membranes due their amphipathic character. This means, they possess a hydrophilic and hydrophobic part with the head being the hydrophilic and the FA chains being the hydrophobic part. Phospholipids consist of four components:

- a backbone, e.g. glycerol or sphingosine
- one or two fatty acids depending on the backbone
- a phosphate group
- an alcohol group

The backbone can differ between glycerol and sphingosine. This results in the two subclasses, phosphoglycerides and sphingolipids. A phosphoglyceride consists of two fatty acids, which are attached to the hydroxyl groups of the C-1 and C-2 carbon atoms of glycerol through an ester bond. A phosphate group forms another ester bond with the hydroxyl group of the C-3 carbon atom. The resulting molecule is known as diacylglycerol 3-phosphate, which is the simplest phosphoglyceride. Although it is uncommon in membranes it provides a basis for the synthesis of further phosphoglycerides. These phosphoglycerides contain an alcohol group, which form another ester bound with the phosphate group. This part of the phospholipid is also referred to as the head group. Common alcohols incorporated in PC are the amino acids glycerol, ethanolamine, serine and choline (see figure 1.2). The latter, phosphatidylcholines (PC), are the most common phospholipids in membranes and we will deal exclusively with these type of phospholipids in this thesis.

Sphingolipids are the second subclass of phospholipids. Their backbone is sphingosine, which is an amino alcohol with 18 carbon atoms and a desaturated hydrocarbon chain. In contrast to phosphoglycerides sphingolipids contain only one

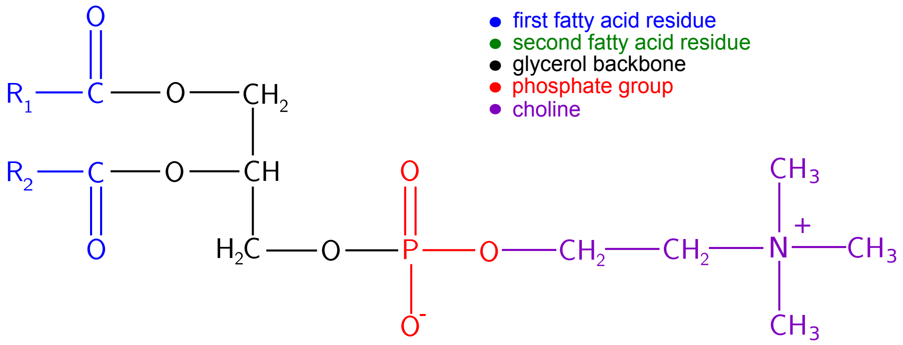

Figure 1.2: Representation of a phosphatidylcholine. R_1 and R_2 indicate the hydrocarbon chain of the first or second fatty acid respectively. *Figure adapted from Wikipedia*

fatty acid, which is amino-linked to the sphingosine. The earlier mentioned alcohols serine, ethanolamine and choline are also utilized as the head group for sphingolipids. This head group is O-linked to the sphingosine.

1.2 Metabolomics

The metabolome describes the abundance of metabolites in a cell, tissue or organism. Metabolites are the products and intermediates of metabolism. The metabolome can change within seconds in contrast to the genome or proteome. [2]. All cellular processes of the the transcriptome and proteome end in metabolites. Therefore the metabolome is considered to be an indicator of an organism's phenotype endpoint. There are around 3,000 known endogenous metabolites of the human organism. Humans take in metabolites through the environment additionally, which increases the number of metabolites up to 100,000 [2].

Metabolomics refers to the comprehensive study of these metabolites and their reactions. There are several methods to determine the metabolites and their according concentrations. The three major determination methods are chromatography, nuclear magnetic resonance (NMR) and mass spectrometry (MS). All of these methods include several submethods with each of them has their assets and drawbacks. Therefore most of them are limited to a specific class of metabolites with specific biochemical properties and a composition of methods is necessary to covering a wide range of metabolites.

Metabolomics is used for various applications in research. For instance, the ratio of phenylalanine and tyrosine can be used as a biomarker of the disease phenylketonuria (PKU) for newborn [4]. Another studied examined the human response to glucose challenging [22]. Illig et al. [12] used metabolomics in a genome

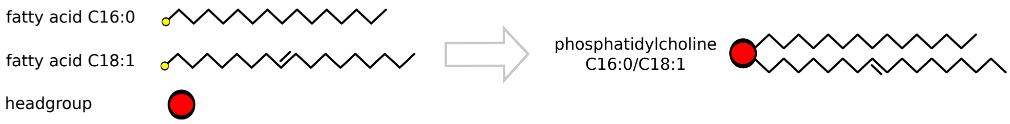

Figure 1.3: Simplified example of the phosphatidylcholine biosynthesis.

wide association study to show that there is association between metabolite frequencies and genetic variations (SNPs).

1.3 Problem Statement

A Phosphatidylcholine (PC) contains two fatty acids (see section 1.1.2 and figure 1.3). Due to the variety of fatty acids (FAs), there are several possibilities for the FA composition of a PC. The mass spectrometry technology which was used to establish the data used in this work can not distinguish between PCs that have the equal sum of carbon atoms and double bounds of the their corresponding FAs. For instance, PC36:4 can resolved to several FA compositions (see figure 1.4). Thus there is an uncertainty of the FA composition of the PCs, whose concentrations were measured with mass spectrometry.

In this thesis, we attempted to develop a method to calculate the FA frequencies which are used for the PC biosynthesis. This calculation is based on the measured PC concentrations. The computed FA frequencies would help to understand the FA composition of measured PCs in detail. The estimation of missing data based on measured data is referred to as imputing [12]. With a more abstract view at this problem, we try to estimate parameters based on experimental data. Such an problem can be referred as an inverse problem [8].

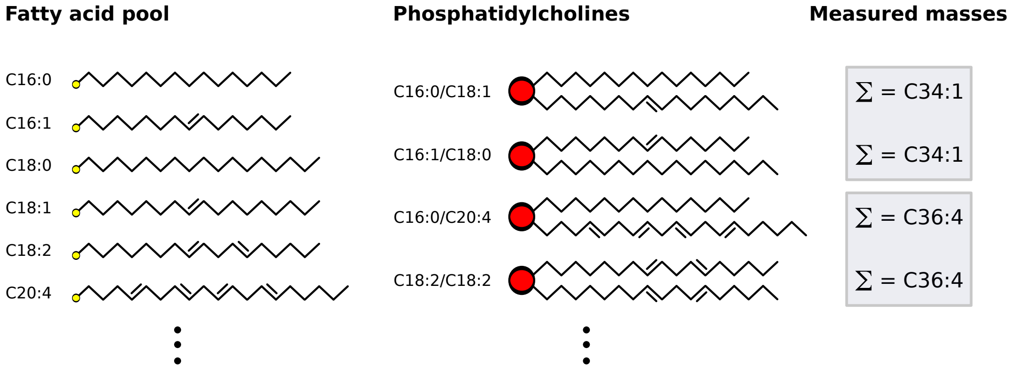

Figure 1.4: Simplified example of the phosphatidylcholine biosynthesis.

Chapter 2

Methods and Materials

This chapter deals with the developed methods and used materials for this thesis.

2.1 Imputing Method

In the following sections, we give an overview about the basic idea, their components and the implementation of the imputing method.

2.1.1 Basic Idea

The imputing method has the goal to find a set of FA frequencies which explains the in mass spectrometry measured PC concentrations best. It is necessary to generate PC frequencies based on the FA frequencies to evaluate a set of FA frequencies. Afterwards the evaluation is performed with a cost function which provides a measurement of distance between the computed PC frequencies and the frequencies which were measured via mass spectrometry. The search space of all FA frequencies is n-th dimensional with n being the number of fatty acids.

2.1.2 Definitions

For the further explanation of the method, we will define some fundamental terms. A relative frequency of a FA is referred to as $x_{c,d}$ with c being the number of carbon atoms and d the number of double bounds. Furthermore, the relative frequency of a PC is denoted as y_{c_1,d_1,c_2,d_2}. c_n and d_n represent the number of carbon atoms and double bounds of the n-th FA in the PC. $z_{c,d}$ denotes the relative frequency of PCs, which their FA composition is not resolved. This is the case for the measured PCs of the used studies. We will refer to relative frequencies of measured PCs as $\overline{z}_{c,d}$.

2.1.3 Basic Assumptions

In this section, we explain some assumptions that are needed as a basis for the imputing method.

Non-negativity

Obviously, the relative frequencies are non-negative:

$$\forall(c,d) : x_{c,d} \geq 0, y_{c_1,d_1,c_2,d_2} \geq 0, z_{c,d} \geq 0$$

Lipid Sum

A lipid sum is the sum of the relative frequencies of the PCs, which have the same number of carbon atoms and double bounds. Lipid sums are the relative frequencies, which are measured by mass spectrometry (see section 1.3), i.e:

$$z_{c,d} := \sum_{\substack{c_1+c_2=c \\ d_1+d_2=d}} y_{c_1,d_1,c_2,d_2} \tag{2.1}$$

Relative Frequencies

We are dealing with relative frequencies in this work and therefore the sum over all relative frequencies equals 1. y_{c_1,d_1,c_2,d_2} is composed of two $x_{c,d}$. With the definition of lipid sums follows:

$$\sum x_{c,d} = 1 \overset{(2.1)}{\Longrightarrow} \sum z_{c,d} = 1$$

2.1.4 Generation Function

The generation function attempts to simulate the incorporation of FAs at the glycerol backbone. It calculates the PC frequencies for given FA frequencies which is accomplished in two steps. First, the generation function computes the frequencies of PCs y_{c_1,d_1,c_2,d_2}. Second, it calculates the sum of PCs, which have the same number of carbon atoms and double bounds. The outcome are the relative frequencies of PCs $z_{c,d}$.

At this point it is not clarified which assumptions or constraints the incorporation of the FAs underlies. For instance, there is a hypothesis that polyunsaturated fatty acids (PUFA) are exclusiveley at the second position in the glycerol backbone [3]. Such assumptions and constraints would have to be included in the generation function. In this we work we assume that the the incorporation of FAs in the glycerol backbone is independent of the respective FA at the other position. The statistical independency leads to the following assumption in the generation process: $y_{c_1,d_1,c_2,d_2} = x_{c_1,d_1} \cdot x_{c_2,d_2}$.

2.1.5 Cost Function

A cost function provides a measurement of distance between computed PC frequencies and the frequencies, which was measured via mass spectrometry, and therefore helps to evaluate a possible solution. There are several possibilities to calculate the distance. We implemented the following variants of a cost function in this work.

Distance-based:

$$C_1(z, \overline{z}) = \sum_{i=1}^{n} (z_i - \overline{z}_i)^2$$

This method is also known as the least square error (LSE) and is one of the most common approaches to calculate a distance. The problem with this method for our purpose is that PCs with a higher relative frequency have a greater influence on the distance than one with a lower frequency. In this work we want to achieve a good estimation for every FA regardless to the scale of the corresponding PCs. Therefore we implemented a variant of this function which normalizes each single component and thus every single result has an equal influence on the overall result. Distance-based normalized:

$$C_2(z, \overline{z}) = \sum_{i=1}^{n} \left(\frac{z_i - \overline{z}_i}{\overline{z}_i} \right)^2$$

We will refer to the result of a cost function as *costs* in this work.

2.1.6 Simulated Annealing

Simulated annealing is a heuristic method to solve an optimization problem [14]. It attempts to find a good approximation of the global minimum in the search space by simulating an annealing process. It repeats the steps of alteration and evaluation of the current putative solution. The evaluation is performed by a given cost function $f(x)$, which needs to be minimized. After each iteration the method picks a new solution which is close to the current one in the search space. If this solution leads to lower cost values, the method will continue with the new solution. Otherwise the method will accept the solution with the following probability:

$$\exp\left(-\frac{f(x_{new}) - f(x_{old})}{T} \right)$$

where $f(x_{new})$ are the cost values of the new and $f(x_{old})$ are the cost values of the old solution. T is the virtual temperature which decreases exponentially. Hence the higher the current temperature, the more likely the method will continue with the worse solution. The exponential decrease of T is shown in the following term:

$$T \leftarrow T \cdot c \tag{2.2}$$

where c is the cooling schedule, which has a value between 0 and 1. Note that a sufficient slow cooling schedule would always result in the global minimum [18]. But the increase of runtime would make such a slow cooling schedule not feasible.

In our case a putative solution is a set of FA frequencies. Therefore the function which will be minimized by simulated annealing consists out of a generation function (see section 2.1.4) and a cost function (see section 2.1.5), because the PC frequencies have to be computed to evaluate the set of FA frequencies.

2.1.7 Construction of the Fatty Acid Pools

The term fatty acid pool (FAP) describes the FAs, which will be used to calculate the PC frequencies and hence whose frequencies will be determined. The FAP great influence on the out coming result, because it determines which PCs can be generated in general and how their FA compositions are in detail. For instance, if we want to generate PC40:5, we will have to put a FA with an uneven number of double bounds, e.g. 1, 3 or 5, into the FAP. Otherwise we would not be able to generate a PC with 5 double bounds and the costs would rise. Therefore we put a lot effort into the examination of the FAP. In this section, we explain how we established the examined FAPs. A general overview of the FAPs is shown in table 2.1.

Original (ORI)

The FAP ORI was provided by an earlier work on this field [15]. This FAP was used as a foundation for the first examinations with the imputing method.

Literature (LIT)

We examined papers which determined the FA composition of PCs [5, 6, 7, 9, 11, 19, 20, 21]. These PCs were mainly taken out of blood samples. This FAP is an union of all FAs that were mentioned in these studies.

Pathway (PATH)

For the FA composition of this FAP, we used the FA biosynthesis pathway of KEGG (pathway id: hsa01040) as a foundation [13]. PATH contains every FA, which is existent in this KEGG pathway. In addition, the FAP was expanded by fatty acids, which were missing so far and have been recognized as important in the development of previous FAPs. For instance, C12:0 and C14:0 were added, because the pathway covers only the biosynthesis of FAs with a number of carbon atoms above 16. Hence, the resulting FAP is not an identical implementation of the KEGG pathway.

Literature + Pathway (LITPATH)

This FAP is an union of LIT and PATH. This FAP was designed with the purpose to evaluate whether these FAPs each for themselves lack in crucial FAs. If this is the case, it will result in an considerable improvement of the performance.

Expert (EXP and EXP260)

The FAP EXP was established in a collaboration with Josef Ecker from the workgroup Schmitz, Universitätsklinikum Regensburg. It was unsure whether C26:0 should be in EXP. Therefore an additional FAP was created to examine both cases. EXP260 contains all FAs of EXP and C26:0 in addition.

Overfitting (OVER)

This FAP contains all possible FAs with 10 - 26 carbon atoms and 0 - 6 double bounds. This includes also FAs with an uneven number of carbon atoms. This FAP was constructed with the purpose to verify the hypothesis that too many FAs can lead to an overfitting by the imputing method.

fatty acid pool	contained fatty acids	additional information
original (ORI)	C12:0, C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, C18:3, C18:4, C20:0, C20:3, C20:4, C20:5, C22:0, C22:4, C22:5, C22:6, C24:0, C24:4, C24:5, C24:6	This FAP was provided by an earlier work [15].
literature (LIT)	C12:0, C14:0, C15:0, C16:0, C16:1, C17:0, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C20:2, C20:3, C20:4, C20:5, C21:0, C22:4, C22:5, C22:6, C23:0	This FAP is based on the examination of literature [5, 6, 7, 9, 11, 19, 20, 21]
pathway (PATH)	C12:0, C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, C18:3, C20:0, C20:3, C20:4, C20:5, C21:0, C22:5, C22:6, C23:0, C24:5, C24:6	This FAP is based on the FA biosynthesis pathway of the KEGG database [13].
literature + pathway (LITPATH)	C12:0, C14:0, C15:0, C16:0, C16:1, C17:0, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C20:2, C20:3, C20:4, C20:5, C21:0, C22:4, C22:5, C22:6, C23:0, C24:4, C24:5, C24:6	This FAP is a union of the FAPs LIT and PATH.
expert (EXP)	C12:0, C14:0, C14:1, C16:0, C16:1, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C20:2, C20:3, C20:4, C20:5, C22:0, C22:4, C22:5, C22:6, C24:0, C24:4, C24:5, C24:6	This FAP was developed in a collaboration with an expert in this field.
expert + C26:0 (EXP260)	C12:0, C14:0, C14:1, C16:0, C16:1, C18:0, C18:1, C18:2, C18:3, C20:0, C20:1, C20:2, C20:3, C20:4, C20:5, C22:0, C22:4, C22:5, C22:6, C24:0, C24:4, C24:5, C24:6, C26:0	It was unsure whether C26:0 should be in EXP. Therefore an additional FAP was created to examine both cases.
overfitting (OVER)	$\begin{matrix} \mathrm{C10{:}0} & \cdots & \mathrm{C10{:}6} \\ \vdots & \ddots & \vdots \\ \mathrm{C26{:}0} & \cdots & \mathrm{C26{:}6} \end{matrix}$	This FAP contains all possible FAs with 10 - 26 carbon atoms and 0 - 6 double bounds.

Table 2.1: Description of the examined fatty acid pools.

2.1.8 Determine the Minimum Number of Runs

As already explained in section 2.1.6, simulated annealing is a heuristic method to solve an optimization problem. The heuristic character induces the problem that not every resulting solution of the annealing process is a viable one. Therefore we need to determine a minimum number of runs which we have to perform to receive at least one viable result with a high probability.

S is a set of costs with $|S| = n$ and s_{opt} being the minimum cost and hence the best solution. S_{viable} is a set with the following elements:

$$S_{viable} := \{s \in S | s \leq s_{opt} \cdot (1 + \epsilon)\}$$

with ϵ being the allowed aberration with respect to the best solution s_{opt}. We will refer ϵ also as *error tolerance*. The set S_{viable} contains all elements which are to be considered as viable.

$$\beta := \frac{|S_{viable}|}{|S|}$$

β is the empirical probability that a solution of the set S is viable one.

$$\delta := (1 - \beta)^n$$

δ is the probability that there is no viable solution in the set S and is referred as the *viability accuracy*. We have to calculate the following term to determine the minimum number of runs n.

$$n = \frac{\ln(\delta)}{\ln(1 - \beta)}$$

So n is dependent on the values of δ and β, whereas β itself is dependent on ϵ and the set of solutions S. To determine the minimum number of runs, we have to assign values for the parameters ϵ and β and examine a sample set of solutions S.

2.1.9 Queue Calculation

All methods were implemented in MATLAB (The Mathworks Inc.). The runtime of the imputing method for one sample is approximately 2.5 seconds. Hence the overall runtime for the KORA dataset (see section 2.3.1) with 931 samples and 100 runs per sample is about 65 hours. Therefore we also implemented the ability to submit the calculation to a GRID engine that can parallelize the calculation on 150 cores. This can accelerate the calculation for the KORA dataset up to 150 times from 65 hours to 44 minutes.

2.2 Toy Data

The usage of toy data is a common approach to evaluate the correctness and robustness of a new methodology. We perform a forward simulation of the toy data to evaluate the imputing method. Hence, the toy data is used as an evaluation dataset. This dataset represents the FA frequencies, which are unknown in the work with real data. The relative PC frequencies are computed with the evaluation data and the according generation function (see section 2.1.4). These PC frequencies are then used to accomplish the process of the imputing. Afterwards an evaluation of the putative imputing solution is possible due to the known correct solution, which is the evaluation data set. In general, the toy data were generated as uniformly distributed random numbers between 0 and 1.

2.2.1 Noise

Noise appears in all biological measurements, e.g. the determination of metabolite concentrations with mass spectrometry. We implemented the ability to superpose the toy data with noise and thus to simulate the work with real data more realistically. The noise superposition procedure is explained in the following:

$$\tilde{z}_{c,d} = \overline{z}_{c,d} + \overline{z}_{c,d} \cdot r \cdot s \tag{2.3}$$

with x being a relative frequency of a metabolite, s the noise strength and $r \sim \mathcal{N}(0,1)$. Hence the resulting noise superposition is normally distributed and scaled with the metabolite frequency and s. We implemented two versions to specify the noise strength s.

Coefficient of Variation

The coefficient of variation (CV) is defined as follows:

$$CV = \frac{\sigma}{\mu}$$

with σ being the standard deviation and μ being the mean. This version of noise superposition uses the coefficient of variation for each metabolite as r, hence the value differs for each metabolite. The CVs were determined based on the data of the "HuMet" study in collaboration with Gabi Kastenmüller. The determination was enabled by measured concentrations of technical replicates. The use of CV as the value for r allows us to generate a noise superposition that equals the noise superposition of the measured data. This helps to evaluate the robustness of the imputing method specific for the work with the KORA (see section 2.3.1) and mouse (see section 2.3.2) datasets.

Constant Noise Strength

In this version, r has the same value for all metabolites and the value itself is selectable. Additionally, it is possible to set a minimum and a maximum value and also a growth value. r starts with the minimum value and increases with the growth value per each sample till it reaches the maximum value. This gives us the ability to examine the correlation between the increasing noise strength and the resulting increase of the costs and thus evaluate the robustness of the imputing method in general.

2.2.2 Correlated Data

FA frequencies underlie a correlation among each others due to biosynthesis processes, e.g. desaturation and elongation (see section 1.1.1). We have to evaluate, whether this correlation would bias the results of the imputing method. For this purpose, we generated a $m \times n$ matrix M with m being the number of samples and n being the number of FAs. The values of this matrix were uniformly distributed between 0 and 1 and afterwards the values were correlated, which is explained in the following term:

$$M \leftarrow E \cdot D^{\frac{1}{2}} \cdot M^T$$

with E and D being the two first resulting matrices of a singular value decomposition function (MATLAB function *svg*). The correlated data are not distributed uniformly and not between 0 and 1, so we perform a rescaling step.

2.3 Datasets

The following sections provide informations about the studies of the used dataset and what data the dataset contains.

2.3.1 KORA

KORA is a study which examines people from the region of Augsburg, Germany, with respect to the effects of their environment, behavior and genetic variations. This study provides severals thousand independent population-based samples [10]. In this thesis the dataset consists of a subset with 931 samples from participants of the KORA study. This dataset includes the concentration of 151 metabolites such as sugars, amino acids and phospholipids. The latter are the main focus of this work and 36 of these phospholipids are PCs, which are used to calculate the FA frequencies with the imputing method. The PC concentrations were normalized to obtain the relative frequencies $\overline{z}_{c,d}$ before applying the imputing method.

diet	number of week				
	1	2	3	4	5
Control	18	27	27	2	2
Distel	20	25	31	-	-
DistelR	-	-	-	9	6

Table 2.2: Number of mice for every week and diet. 'Control' denotes the standard diet and 'Distel' the safflower oil diet. 'DistelR' describes the group of mice which received a standard diet after three weeks of the safflower oil diet. Thus there are only samples of the fourth and fifth week available.

The metabolite concentrations were determined with the Biocrates AbsoluteIDQ technology via electrospray ionization tandem mass spectrometry (ESI-MS/MS). Another mass spectrometry technology (Metabolon) was used to determine the concentration of 27 free FAs in the samples in addition to the 151 metabolites.

Besides the determination of metabolites, the samples were genotyped using the Affymetrix 6.0 GeneChip. SNPs at 517,480 different loci resulted from this genotyping [12]. A subset of 326 SNPs were used in this work.

2.3.2 Mouse Nutritional Challenging

This study is conducted by the workgroup of Susanne Neschen HMGU and examines the effect of a safflower oil diet on the metabolome of the mouse. There are three mouse strains - C3H, B6J and B6N. The two latter are almost identical except for a mutation in the gene *NNT*. There were two groups of mice for each strain. Group one received a standard diet, whereas group two received a special safflower oil diet. The sample period of this study was three weeks. After every week some samples of each group were chosen to determine the metabolite concentrations. The mice had to be sacrificed for this determination process. The two groups of the C3H strain were examined two additional weeks. The safflower oil group also received the standard diet in this period, giving us the possibility to analyze the changes of the metabolome after dropping the safflower oil diet.

The determination of the metabolite concentrations was performed with the Biocrates AbsoluteIDQ technology via electrospray ionization tandem mass spectrometry (ESI-MS/MS). The dataset contains 167 samples of the 6 groups with 151 metabolite concentrations. Informations about the sample, e.g. the week and the diet, were recorded thus enabling the evaluation of the influence of the diet on the mouse metabolome (see table 2.2).

2.4 Evaluation Methods

2.4.1 Genetic Association

Illig et al. [12] showed in a genome wide association study that there is an association between metabolite frequencies and genetic variations (SNPs). The method of genetic association is based on linear regression and therefore underlies the assumption of an additive genetic model.

In this thesis, we will use their approach of establishing associations to evaluate the results of the different fatty acid pools (see section 2.1.7). For this purpose, we used the computed FA frequencies. As shown by Illig et al., the use of metabolite concentration ratios can lead to stronger associations, e.g. lower p-values. We also apply the ratio calculation when we take a closer look at a single fatty acid pool.

Chapter 3

Results and Discussion

This chapter deals with the results and discussion of the imputing method. At first, it was necessary to establish some general results and to evaluate the correctness and robustness of the imputing method with toy data. Afterwards an analysis of two real datasets was performed.

3.1 General Results of the Imputing Method

In this section, we discuss general results concerning the imputing method. First, we show that there are local minima in the search space and how these local minima affect the imputing method.

3.1.1 Local Minima in the Search Space

The search space is composed of the FA frequencies and therefore it has n dimensions with n being the number of FAs. The imputing method attempts to find the global minimum of this search space with a heuristic approach (simulated annealing). Local minima affect this process, since the imputing method could get stuck in a local minima. Thus, we had to evaluate whether there are local minima in the search space, which then need to be considered in the application of the imputing method.

For this purpose, the imputing method was performed for a few randomly chosen samples of the KORA dataset (see section 2.3.1) with 100 runs of the imputing method per sample. Afterwards we performed a principle component analysis of the computed FA frequencies to visualize the result. Each result was examined individually. An exemplary result of a principle component analysis for one sample is shown in figure 3.1. The enrichment of solutions at several locations

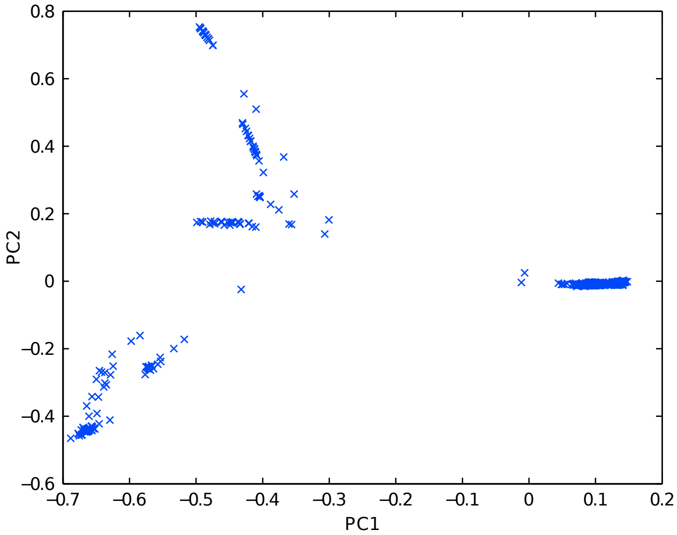

Figure 3.1: First two principle components (PC1 and PC2) for a random sample of the KORA dataset.

indicate the existence of local minima. Thus, we have to consider local minima in the application of the imputing method.

3.1.2 Minimum Number of Runs

With the existence of local minima in the search space (see section 3.1.1) and the heuristic character of the imputing method, we conclude that not every result of the imputing method is a viable one. Therefore we have to determine a minimum number of repetitions of the imputing method for each sample, so that at least one result is viable with a high probability. We will refer to the repetitions of the imputing method for a sample as *runs* in this thesis. In this section, we will discuss how we determined this minimum number of runs.

We chose 50 samples randomly out of the KORA dataset and performed the imputing method 1000 times for each sample. The method described in section 2.1.8 covers the determination for a single sample, but we dealt with 50 samples to get a good random sample of the KORA dataset. The best resulting costs of the samples can vary among themselves up to 120% and thus we can not merge all solutions into one set, because it would bias the calculation. Therefore we determine the minimum number for each sample and the mean of all results will become the overall minimum number, thus treating it like an average case analysis.

The parameters ϵ (error tolerance) and δ (non-viable probability) have a considerable influence on the outcome besides the set of solutions itself (see section 2.1.8). We examined systematically a range of values for both parameters. This gives us the ability to see how changes at these parameters influence the outcome in detail.

The size of the set of costs is another aspect which needs to discussed. The described method requires a set with a sufficient number of costs or it could lead to the following problems. First, it could be the case that s_{opt} is not a viable result. Second, β will be estimated more precisely. Therefore we performed the imputing method 1000 times. Additionally, we will also discuss the result with a number of 100 costs per sample. Thus we can evaluate, how significant this method with a lower number of costs is. This will provide us helpful information for further studies with other datasets.

As expected a lower error tolerance and a lower non-viable probability both lead to an increase of the minimum number (see figure 3.2A). A lowering of the non-viable probability leads to a linear growth of the minimum number, whereas a lowering of the error tolerance results in an exponential increase. The non-viable probability is the logarithm to the base 10, hence the growth is also exponential. There is a broad area with a minimum number of runs between 5 and 150. The comparison with the result which only contains 100 costs per set reveals that broad area is almost the same like in the result with 1000 costs per set (see figure 3.2B). The difference is that the growth along both axis is lower. We conclude that an analysis with a lower number of costs per set will be applicable, if the targeted error tolerance value is > 0.02, hence the exponential growth a long the error tolerance-axis has more influence on the minimum number for low error tolerance values.

Based on this result, we decided that the default number of runs per sample will be 100 for the most of the analyses. This number covers more than 50% of the search space for the error tolerance and non-viable probability (see figure 3.3). For instance, it fulfill the requirements of a non-viable probability of 10^{-5} and a error tolerance of 0.02. This number still provides a good overall runtime for the imputing method (see section 2.1.9).

3.1.3 Cooling Schedule of the Simulated Annealing

The term cooling schedule describes how the virtual temperature of the simulated annealing decreases (see in section 2.1.6 the equation 2.2) and is denoted with a value for c between 0 and 1. The cooling schedule influences the finding process of the global minimum, hence the faster the cooling schedule the faster the virtual temperature decreases. This could lead to the effect that the simulated annealing gets stuck in a local minimum (see section 3.1.1) and thus resulting in higher costs. On the other hand, a slow cooling schedule value could result in an unacceptable runtime of the imputing method. Our target is to find a value, which ensures to get

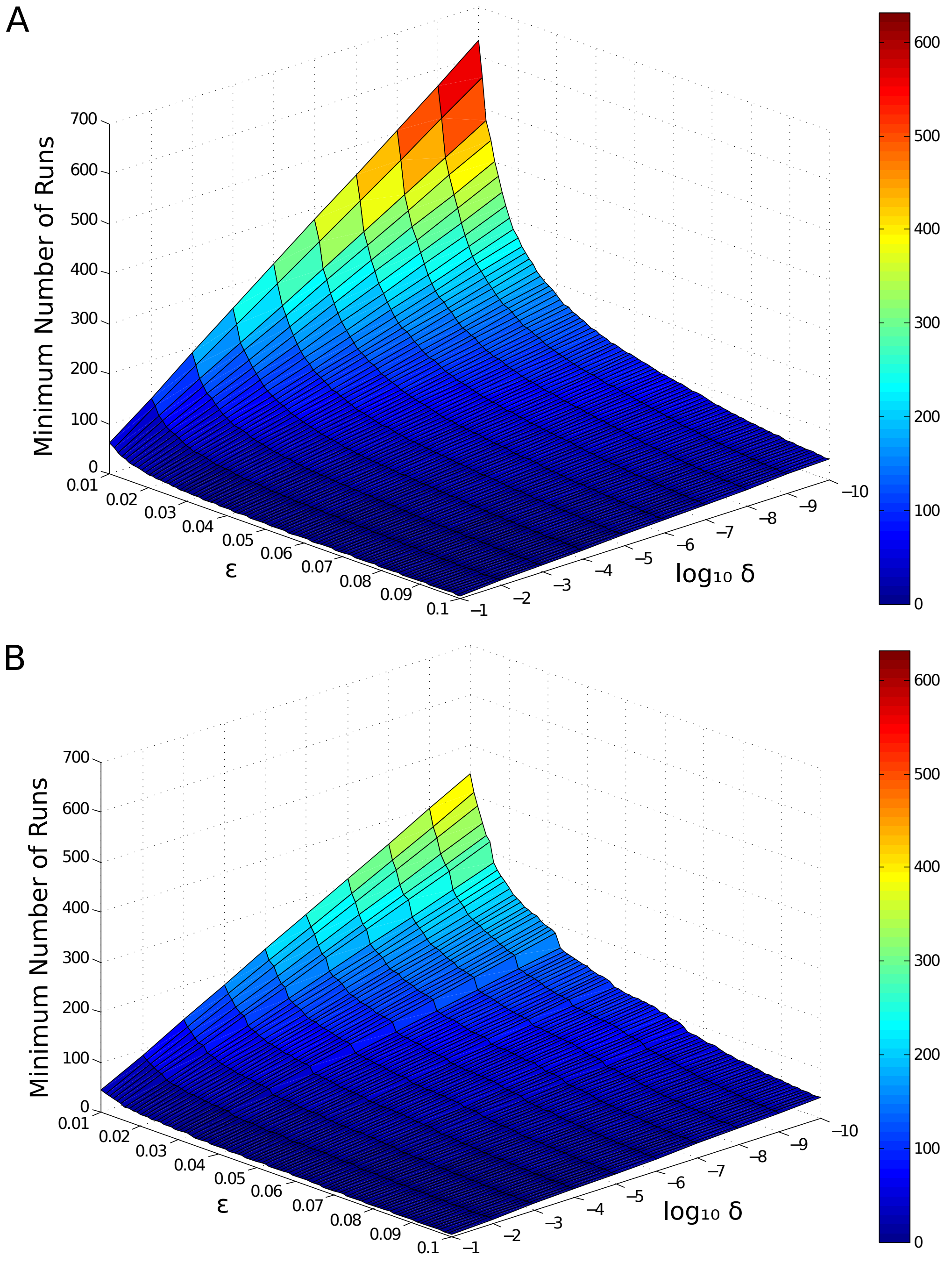

Figure 3.2: Visualization of the minimum number of runs analysis. The error tolerance (ϵ) and non-viable probability (δ) are along the x- and y-axis. The resulting minimum number of runs according to the value of these parameters is on the z-axis. (A) This analysis was performed on 50 samples which were randomly chosen out of the KORA dataset with 1000 runs per sample. A lowering of the non-viable probability leads to a linear growth of the minimum number, whereas a lowering of the error tolerance results in an exponential increase. The non-viable probability is the logarithm to the base 10, hence the growth is also exponential. (B) This analysis was also performed with 50 samples. In contrast to the first analysis, the number of runs per sample is 100. This results in lower exponential growth of the minimum number of runs along the ϵ-axis.

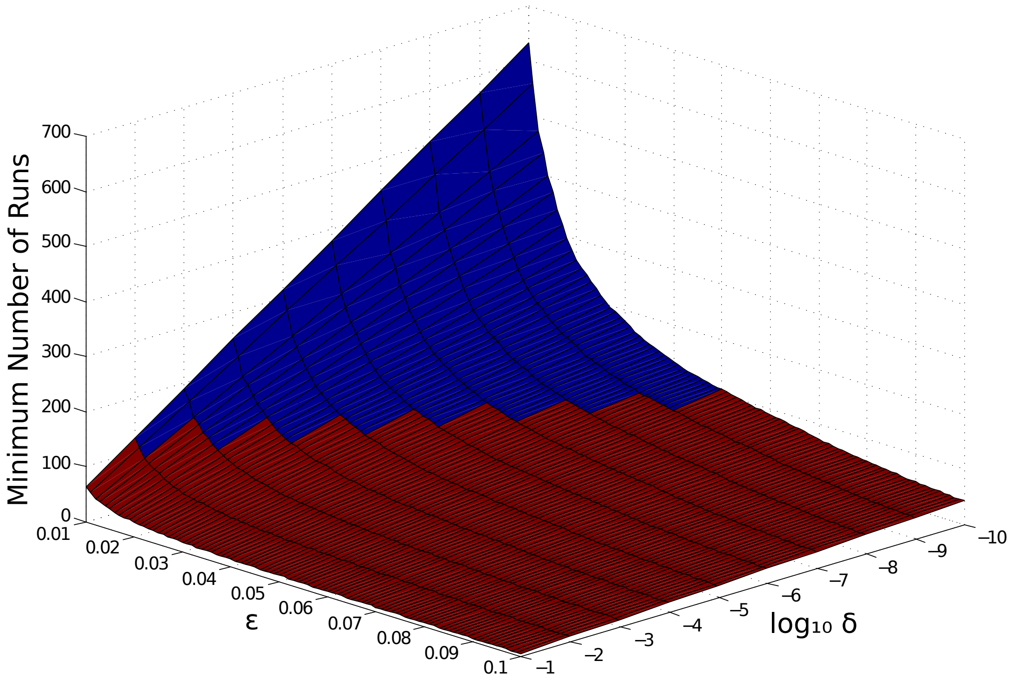

Figure 3.3: Visualization of the minimum number of runs analysis with 50 randomly chosen samples of the KORA dataset and 1000 runs per sample. A red coloring means that these two values of the two parameters ϵ and δ results in a minimum number of runs, which is ≤ 100. Otherwise the surface is colored blue.

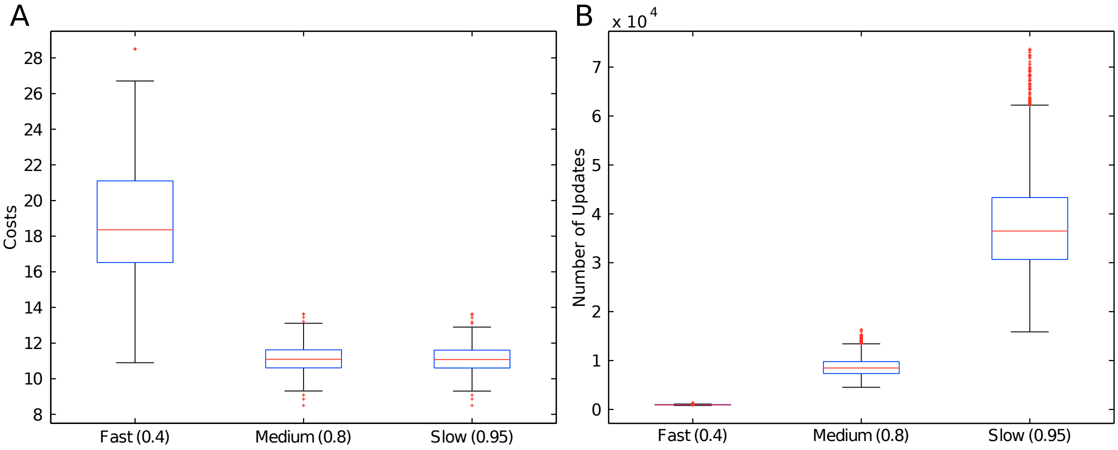

Figure 3.4: Comparison of different cooling schedules. The dataset consists of 100 randomly chosen samples of the KORA study and 100 runs per sample for this analysis. Three different cooling schedules (fast, medium and slow) are along the x-axis. The cooling schedule value c is displayed in the brackets. (A) The costs of the different cooling schedules for the 100 samples are on the y-axis. (B) The number of updates of the virtual temperature is on the y-axis. The total update number for each run of all samples (100 runs $\times$ 100 samples $=$ 10000 update numbers) is displayed.

a viable result with an appropriate runtime of the imputing method. We examined three different cooling schedules (fast, medium and slow) and their effects on the costs. The dataset were 100 randomly chosen samples of the KORA study and the imputing method was performed 100 times for each sample.

The resulting costs of the fast cooling schedule were much higher (approximately 1.75 times) than the costs of the other two (see figure 3.4A). The bad performance is likely to be explained by the fact, that the simulated annealing gets stuck in a local minima too early with a faster decreasing temperature. Therefore we can rule out the fast cooling schedule because of the unacceptable performance. The costs of the medium and the slow cooling schedule were almost identical. The number of updates with the slow cooling schedule is approximately 4 times higher than the number of updates with the medium cooling schedule (see figure 3.4B), which also leads to an increase of the runtime of the imputing method with the same factor. Because there is no increase in the performance with a slow cooling schedule, we use the medium cooling schedule in further analyses.

3.2 Toy Data

We used toy data to evaluate the correctness and robustness of the imputing method. The identical toy data was used for all purposes in the following analyses.

It was generated as a matrix with uniformly distributed numbers between 0 and 1 and consists out of 1000 samples. The number of FAs depends on the used fatty acid pool (see section 2.1.7), which contains 21 FAs in this case. We used the fatty acid pool which was the first established one (ORI), because the analysis of toy data was performed before the analysis of the fatty acid pools. The FAP do not influence the results of toy data like the results of real datasets. Therefore the use of a less sophisticated fatty acid pool would not bias the following analysis.

3.2.1 Effects of Missing Information

The generation function generates the PC frequencies for all possible FA composition with the given fatty acid pool (see section 2.1.4). Real datasets may contain a limited number of PC frequencies, e.g. 36 PCs in KORA and the mouse dataset (see section 2.3.1 and section 2.3.2). Hence the resulting frequencies of the generation function contain PC which were not measured in the real datasets. Therefore, we need to reduce the generated PC frequencies to the frequencies that were measured to apply the imputing method for these datasets. This reduction results in a loss of information for the imputing method and could therefore lead to a worse solution. In this section, we examined how the reduction effects the results on toy data. This helped us to evaluate whether the imputing method is still able to provide viable solutions after the reduction, which is necessary in the work with real datasets.

The toy dataset contained 1000 samples and the imputing method was performed 100 times for each sample. This dataset was examined twice. In the first run, all PC frequencies which were generated by the generation function were used in the imputing process ('complete information'). In the second run, the set of PCs were reduced from 93 to the 36 PCs which are measured in the two studies ('missing information'). Therefore only about one-third of the information was available for the imputing method. We had to calculate the mean costs per PC to compare the results, because the cost function adds the cost of each PC (see section 2.1.5) and the two results had a different number of PCs. Otherwise it would bias the comparison.

First of all, the results for 'missing information' show that the imputing method is applicable, since the mean costs per PC are close to 0 (see figure 3.5). That is, the computed FA frequencies can generate PC frequencies which are almost identical with the measured data. The results of 'missing information' demonstrate that reduction of the PC frequencies lead to an increase of the costs as expected. The difference between the results is significant for the median as shown with the notch representation. Considering the low costs, we conclude that the reduction only has a small influence on the outcome.

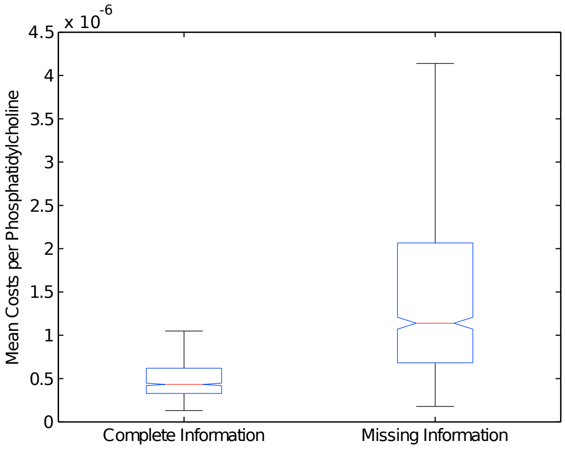

Figure 3.5: Comparison of two results of a toy dataset. The y-axis are the mean costs per phosphatidylcholine. 'complete information' describes that all PCs were used by the imputing method, which were generated by the generation function. Whereas a reduction of this set of PCs was performed before the imputing process for 'missing information'. Outliers were removed for a more accurate resolution of the results.

3.2.2 Noise Analysis

We will analyze the influence of noise on the results of the imputing method in the following sections. We examined two different versions of noise superposition.

Coefficient of Variation

We performed a third analysis on the same toy data set in addition to the two runs in the section 3.2.1. Besides the reduction of the set of PCs, this dataset was applied with a noise superposition. The noise strength for each PC was based on the determined CVs (see section 2.2.1). The imputing method was also performed 100 times per sample. The noise superposition procedure was performed individually before each repetition of the imputing method.

The results of this analysis shows that the noise superposition leads to an considerable increase of the costs (see figure 3.6). In comparison to the difference between 'complete information' and 'missing information', this increase shows that noise has much more influence on the costs than the reduction of the set of PCs. The conclusion of this analysis is that noise is a factor in the work with real datasets that is not negligible.

It was not possible to apply the noise superposition for the dataset with all possible PCs, because the CVs could only be determined for the PCs that were measured. Therefore we did not have the combination of all PCs generated by the generation function and a noise superposition according to the CVs. It would

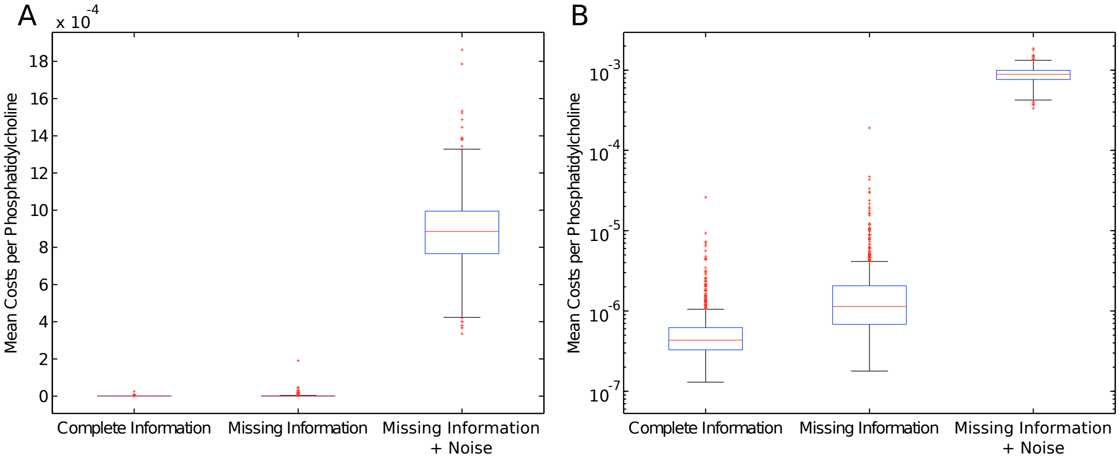

Figure 3.6: Influence of noise on the results of the imputing method. The toy dataset consists of 1000 samples and the imputing method was performed 100 times per sample. The noise strength is given by determined CVs. (A) The mean costs per phosphatidylcholine are along the y-axis. (B) This is the same plot as A, but the axis of the mean costs per phosphatidylcholine is log-scaled

be interesting to see whether the noise superposition has a significantly stronger influence on the results with the reduction of the set of PCs than on the result with all possible PCs.

Constant Noise Strength

In this section, we want to examine how the noise strength correlates with an increase of the costs in detail. The noise strength is the same for all PCs in contrast to the noise strength which was based on determined CVs. The examined toy dataset consisted of 70 samples, which were identical before the noise was applied. The noise strength underlie a linear growth for each sample, starting with 0 and increasing 0.005 for each sample. Hence the applied noise superposition value was 0.0345 for the last sample. In contrast to other analyses, we did not record the best score for a sample. Instead, we took the average score over all runs of the imputing method for each sample. This was necessary, because the noise underlies a normal distribution and this could lead to samples that were affected by significant less noise than the others. Therefore the consideration of the best result would have biased the overall result of this analysis. Hence, we increased the number of runs from 100 to 1000 to get a better sample size for this average case analysis.

It was expected that the increase of the costs would be continuous. Thus the result is not quite as expected, because there are fluctuations in the increase of the costs (see figure 3.7). Overall the growth of higher costs can be observed,

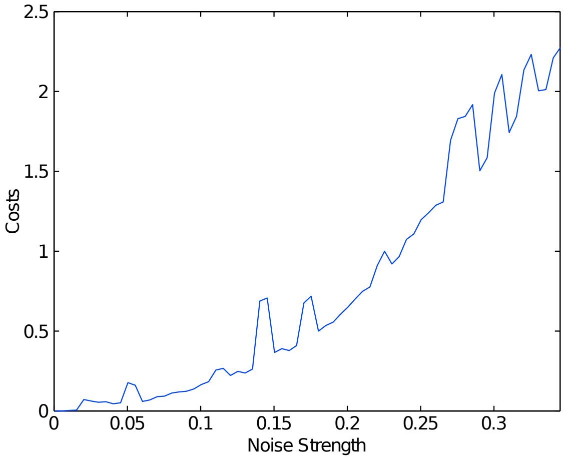

Figure 3.7: Representation of the costs of the imputing method with toy data and noise superposition. The samples were identical before the noise superposition and 1000 runs per sample were performed. The noise strength is equal for each phosphatidylcholine, but was increased for each sample.

but it is unsure how the fluctuations can be explained. It is noticeable that the frequency and also the amplitude of the fluctuations increases with a higher noise strength. Hence we conclude that a higher noise strength increases the variance of the imputing method. It would be interesting to see if a higher number of runs of the imputing method for each sample would lead to a continuous linear growth.

3.2.3 Effects of Correlated Data

The correlated data was generated as explained in section 2.2.2 with $m = 100$ and $n = 3$. Hence the toy data was limited to 100 samples and 3 FAs. Two of the three fatty acids have a correlation coefficient (CC) of 0.5. The remaining fatty acid is uncorrelated (CC = 0). The imputing method was performed with 100 runs per sample. The resulting costs of the correlated data were almost identical (see figure 3.8) to the costs of the uncorrelated data. Overall the costs were slightly lower for the correlated data, but the median is not significant different as shown with the notch representation. Therefore we can conclude that correlated data alone do not bias the results of the imputing method. It would be interesting to see if a combination of correlated data and noise superposition would lead to a different conclusion, because this case would be the most similar to the examination of biological datasets.

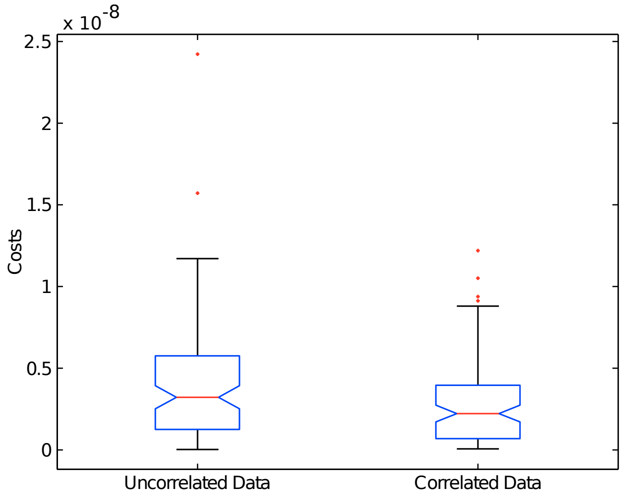

Figure 3.8: Representation of the effects of correlated data by a comparison of the costs. The imputing method was performed 100 times for each of the 100 samples.

3.3 KORA Dataset

This section deals with the examination of the KORA dataset. After the evaluation of the fatty acid pools, we analyzed the result of a single FAP in more detail.

3.3.1 Evaluation of the Fatty Acid Pools

We established fatty acid pools (FAP) based on different resources (see section 2.1.7). In this section, we will evaluate the resulting FAPs. The evaluation of a FAP can not only be based on the resulting costs of the imputing method. For instance, a FAP which contains almost every fatty acid, despite whether the FAs are meaningful with respect to the biological context, will get a good result, because it has too much degrees of freedom in the fitting process. This effect is known as overfitting and will be shown with a concrete FAP. Therefore we have to perform a external validation, which is independent of the fitting process.

This is external validation is provided by the available genotyping data of the KORA dataset (see section 2.3.1). We establish a genetic association with the genotype of the 326 SNPs and the corresponding FA frequencies of the imputing method to evaluate the FAP. A genetic association is assessed with a p-value. We count the number of genetic associations and their p-values which are significant after Bonferroni correction. A good FAP will result in low costs of the imputing method and the corresponding FA frequencies will also show a high number of genetic associations. This can also be considered as a trade off between these two evaluation methods.

Costs of the Different Fatty Acid Pools

The imputing method was performed for each FAP with the whole KORA dataset (931 samples, see section 2.3.1). The number of runs per sample was 100 as specified in 3.1.2.

We exclude the result of the FAP OVER in this discussion of the results for now. Its results will be discussed in detail later. Among the remaining results, the costs differ in the median from 8 to 16 (see figure 3.9). The distribution of the costs around the median are almost identical for all FAPs. This could be due to the fact that a single sample is stable with respect to its cost. This is, a sample with a low score in relation to the other samples of the dataset with the FAP A will be also assigned a low score with the FAP B. This hypothesis would need further analysis to be confirmed. Only EXP and EXP260 have an enrichment of outliers above the median. Five out of these six medians of FAPs share an even closer range of costs (from 8 to 11), but still every median is significant different to the others, besides LIT and LITPATH. The one FAP is PATH, which falls out of this group. Its median is about 45% worse than the worst of the other five, which is a noticeable difference. The lowest costs are achieved by EXP and EXP260.

Genetic Associations of the Different Fatty Acid Pools

The resulting FA frequencies of the imputing method for each FAP were used to establish genetic associations with the available genotyping data. We used linear regression (see section 2.4.1) to calculate the association and the corresponding P-value. For the evaluation of the different fatty acid pools, we only used the FA frequencies itself and not additionally the ratios of the FA frequencies. Three different significance levels were used to evaluate, whether a genetic association is significant after Bonferroni correction. The number of tests were *number of FAs* × *number of SNPs* for each FAP.

The FAP ORI performs quantitative and qualitative the best in this evaluation. It has the highest number of significant genetic associations and the most significant genetic associations were also achieved with this FAP (see figure 3.10 and table 3.2). LIT, LITPATH, EXP and EXP260 perform almost equal considering the number of significant genetic associations. A closer look at their lowest p-value shows that the two latter have more significant associations. PATH stands between ORI and the mentioned group of 4 FAPs with its performance. OVER does not establish a significant association at all.

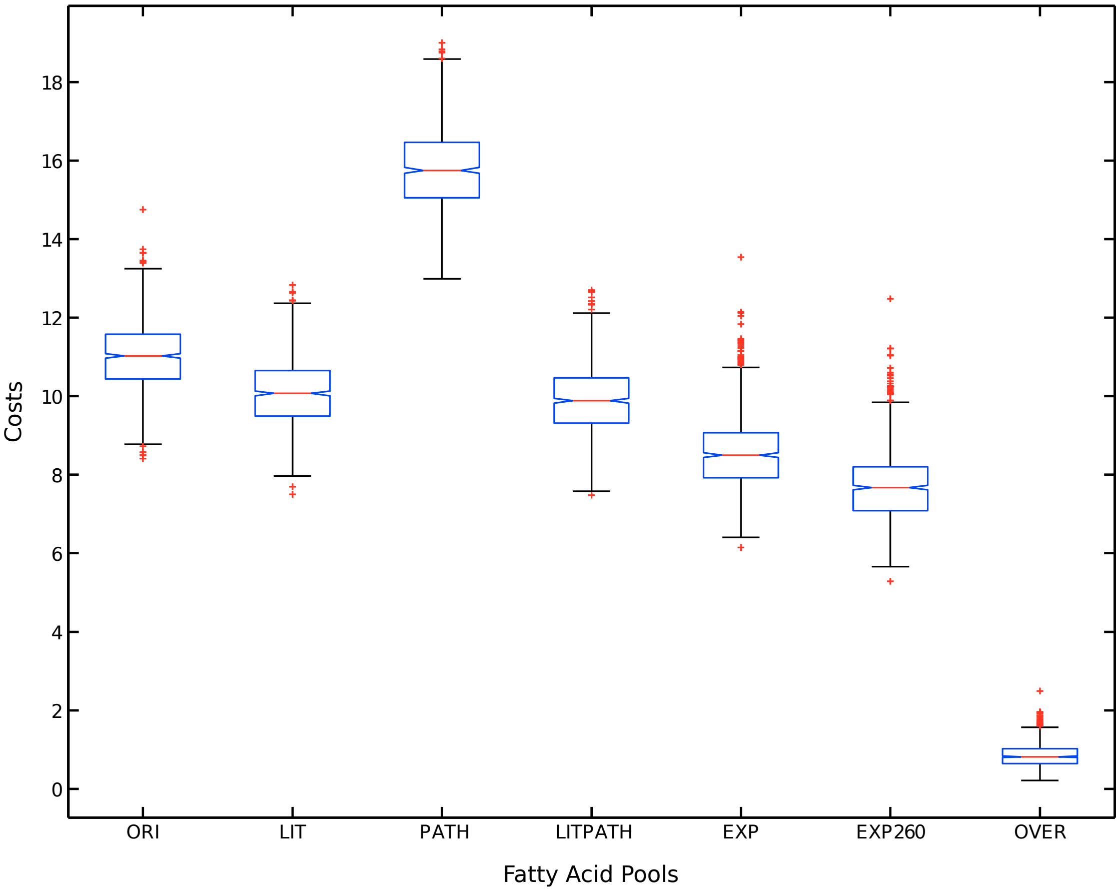

Figure 3.9: Comparison of the costs of the different fatty acid pools (FAP). The imputing method was performed with each FAP for the whole KORA dataset with 100 runs per sample.

fatty acid pool	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	mean of costs ± std
ORI	54	46	41	11.03 ± 0.89
LIT	24	19	16	10.08 ± 0.85
PATH	29	28	22	15.77 ± 1.03
LITPATH	20	16	11	9.91 ± 0.85
EXP	19	16	13	8.57 ± 0.95
EXP260	18	16	13	7.72 ± 0.93
OVER	0	0	0	0.86 ± 0.31

Table 3.1: Numbers of significant genetic associations for the different fatty acid pools (FAP) after Bonferroni correction and the mean costs for all samples ± the standard deviation (std). α denotes the applied significance level.

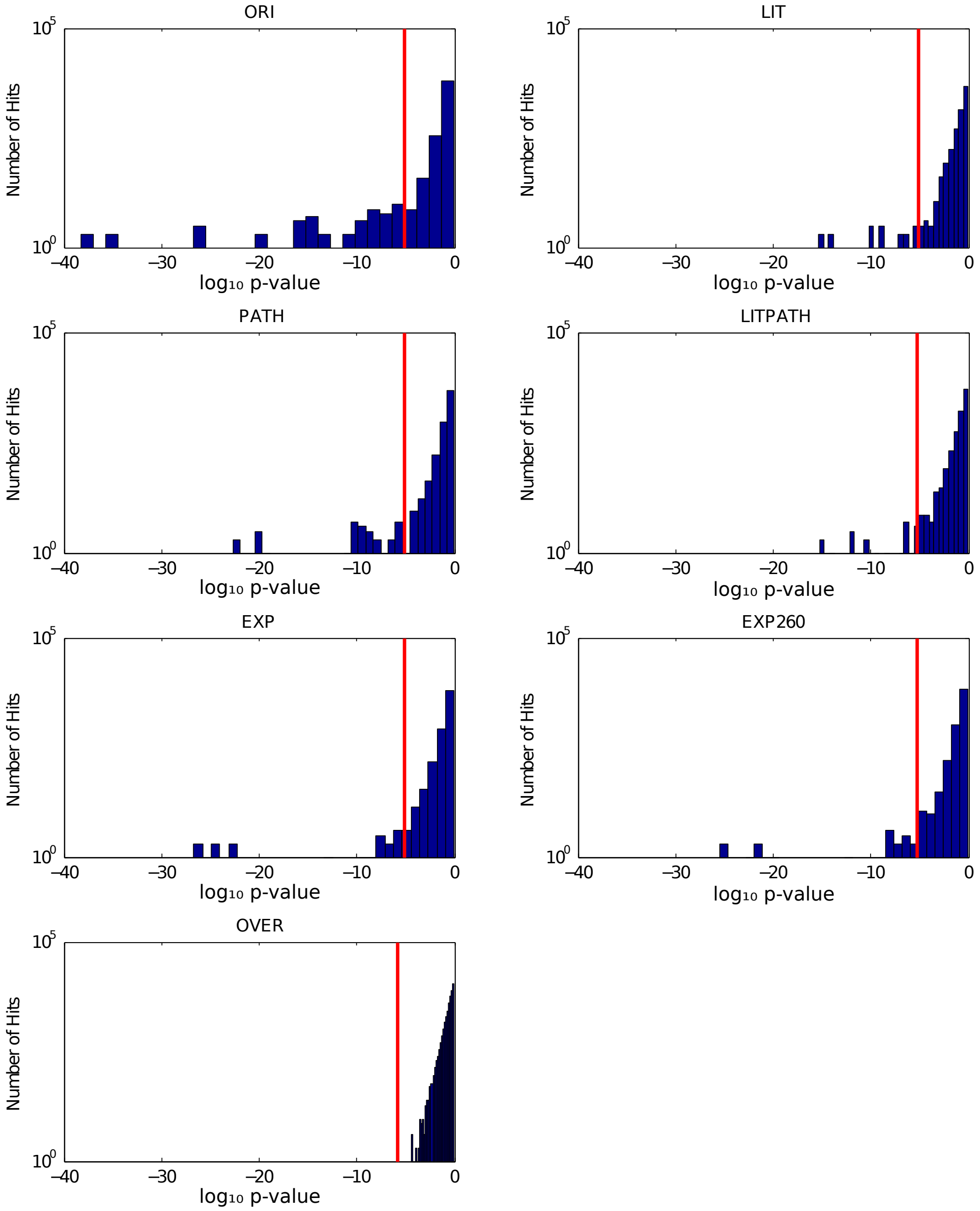

Figure 3.10: Distribution of the genetic associations and their corresponding p-values for the different fatty acid pools. The x-axis denotes the logarithm of the p-values to the base 10. The number of hits are along the y-axis, which is log-scaled. The red line marks a significance level of $\alpha = 0.05$ after Bonferroni correction. All tests below this threshold are considered to be significant.

Choice of a Fatty Acid Pool

We will discuss in this section, how we evaluated the results of the costs and genetic associations in detail to specify a FAP, which fulfills our requirements. As already mentioned the weighting of these two methods is kind of a trade off.

First, we will discuss an extreme example to demonstrate this trade off. The FAP OVER contains 119 FAs, the most of them are not meaningful with respect to the biological context. The resulting costs of OVER are close to 0. Hence this could lead us to the conclusion that we can explain the measured PC frequencies almost perfect with this FAP and the corresponding FA frequencies. The results of the genetic associations show that there was no significant association. In contrast, the results of the other FAPs demonstrate that significant associations can be established. Thus we can conclude that the resulting set of FA frequencies of OVER are not meaningful with respect to the biological context and the application of this FAP would be useless for further analyses. This results signal that OVER lead to an overfitting in the imputing method. This FAP was designed with the purpose to demonstrate this effect.

The number and the significance of the associations established by PATH were sufficient. On the other hand its costs of the imputing method were extraordinary high in comparison to the other FAPs. A part of these FAPs achieved similar results considering the associations. Therefore we also can rule out PATH for further analyses. This FAP was based on the pathway database of KEGG. The results suggests that this FAP is missing some meaningful FAs, which can be conducted to two reasons. One, the database is missing some reactions. Two, the human takes in FAs through the food. Both of these aspects are not covered in this FAP.

LITPATH, EXP and EXP260 achieved quite similar results in both evaluation methods. EXP and EXP260 (expert FAPs) performed better in the costs aspect than LITPATH and their p-values also show a higher significance. Therefore we would choose one of the expert FAPs over LITPATH. A further comparison between these two FAPs is not necessary, because ORI achieved an outstanding result considering the genetic associations. The number of significant associations is up to 2 - 3 times higher and also the significance of these associations is better. The expert FAPs performed approximately 20-35% better in the costs aspect. At this point the weighting of the evaluation methods came into play. We decided that the better result of the genetic associations overweigh the costs and therefore we used ORI for all following analyses with the KORA and mouse dataset. It cannot be ruled out that on of the expert FAPs would have been a better choice. We assume that either one out of these three FAPs would have been fulfill our requirements.

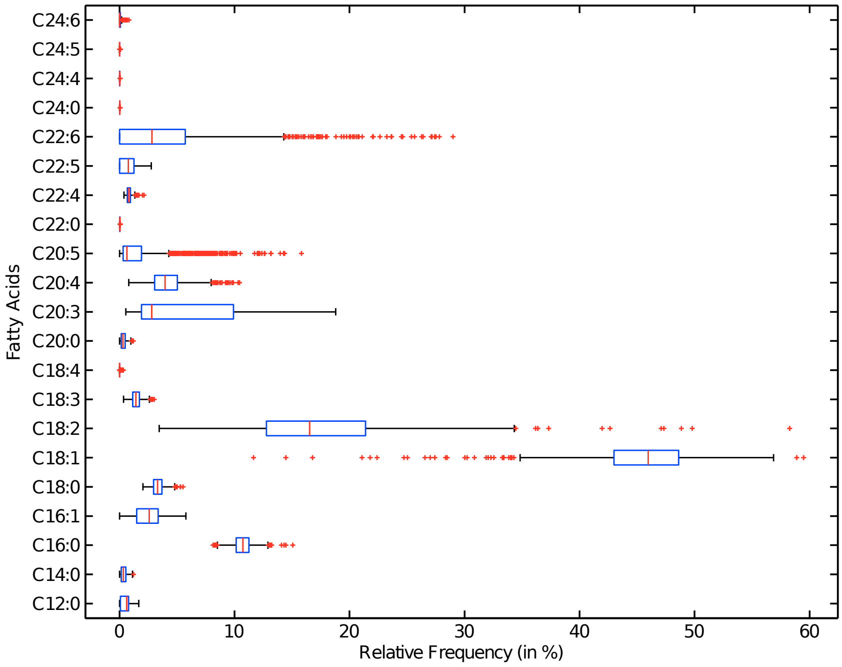

Figure 3.11: Overview of the computed fatty acid frequencies of the imputing method for the KORA dataset. The ORI fatty acid pool and 100 runs per sample were used for this computation.

3.3.2 Analysis of the Computed Fatty Acid Frequencies

The imputing method was performed for the KORA dataset with 100 runs per sample (see section 3.1.2) and ORI as the FAP (see section 2.1.7). First of all, we discuss the resulting FA frequencies. The most abundant FAs are C18:1, C18:2 and C16:0 (see figure 3.11). These three FAs cover approximately 75% of all FAs frequencies. This observation corresponds with the coverage of these FA frequencies (approximately 65%) that were determined experimentally in a study [9]. But the descending order according to the fatty acids abundance of this study differentiate from our result. C16:0 is the most abundant one, whereas it is the 3rd most common among the computed FA frequencies. We have to consider that Hodson et al. used phospholipids and not exclusively PCs for the determination, hence a difference had to be expected.

Another way to look at the result of the imputing method is the relative error for the PC frequencies (see figure 3.12). These errors display how well the PC frequencies can be explained by the FA frequencies. To this, the PC frequencies need to be calculated by the generation function and the computed FA frequencies. The relative error is the difference between the measured and the computed PC normalized with the measured frequency. A relative error above 0 denotes an underestimation and vice versa a relative error below an overestimation of the PC frequency.

Some PCs can be explained very well (C40:6, C42:5, C42:5). Others have a tendency to over- or underestimation. There also PCs which can not be explained at all, thus resulting in an relative error of close to 100%, e.g. PC42:0 or PC26:0. It is interesting to examine, why the FA composition of these PCs could not be generated properly. It is remarkable that most of these PCs do not have a double bound. Hence they need two FAs, which both have also zero double bounds. This results in a small subset of the originally FAP. With the basic understanding, how the evaluation process of the imputing method is proceeded, we can assume that an alteration of the needed FA frequencies lead to an increase of the costs of the remaining PCs. This increase had to be greater than the cost, which contains relative PC errors close to 100%. Otherwise the imputing method had continue with the altered FA frequencies. In order to come that true, the small subset of FAs which have zero double bounds has to be incorporated in other FAs in an overwhelming proportion. Based on this conclusion, we also assume that the PCs with low relative errors have a FA composition, which at least one FA being rarer in other PCs.

Investigation of Genetic Associations

We already used genetic associations to perform the evaluation of the different FAPs (see section 3.3.1). For this purpose, we exclusively used the computed FA frequencies. Illig et al. showed in their work that the use of ratios of concentrations can lead to an even stronger signal of the associations, e.g. lower p-values. We also applied the approach of the ratio calculation, whereas we had to work with frequencies instead of concentrations. First, we take a look at the associations that were established solely by the computed FA frequencies. We refer to this as the non-ratio associations. The used FA frequencies were the result of the imputing method with the FAP ORI.

The strongest non-ratio association is established between the SNP rs174547 and the FA C20:3 (see table 3.2). This SNPs occurs in the intronic region of the gene *FADS1*. The enzyme FADS1 catalyzes the desaturation of C20:3 to C20:4 (see KEGG pathway id 'hsa01040'). Hence this strong association support the meaningfulness of the results of the imputing method. There are also two other FAs (C18:3 and C12:0) that have a stronger association with this SNP than any of the 151 metabolites of the KORA study. C18:3 is the substrate of the enzyme FADS2. Hence, the association between C18:3 and FADS1 could be due to the homology of FADS2 to FADS1. An association between a SNP of FADS2 (rs174550) and C18:3 is not quite that strong than the mentioned associations above, but it is still stronger than any association with one of the 151 metabolites

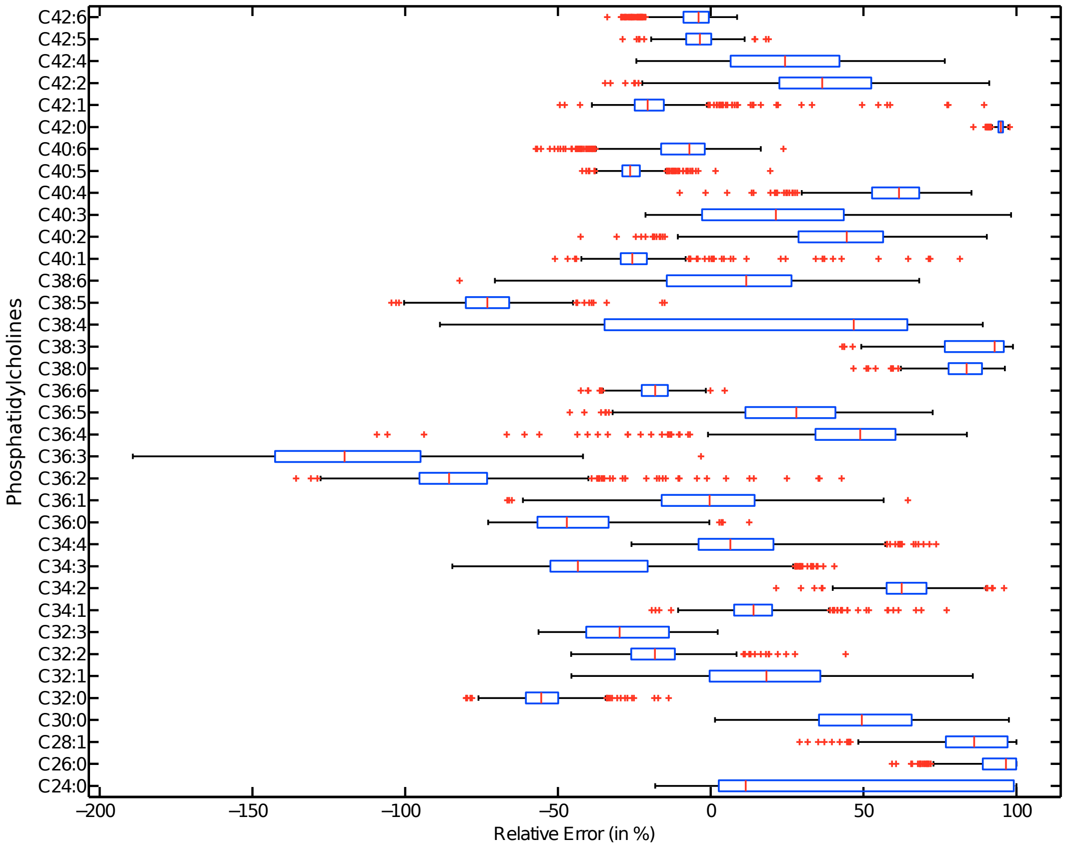

Figure 3.12: Overview of the relative error for the PCs. The PC frequencies were calculated by the generation function and the computed FA frequencies of the KORA dataset. The relative error is the difference between the measured and the computed PC normalized with the measured frequency. A relative error above 0 denotes an underestimation and vice versa a relative error below an overestimation of the PC frequency.

SNP information		computed FA			KORA metabolite		
Identifier	Locus	Association with	p-value	R^2	Association with	p-value	R^2
rs174547	FADS1	C20:3 C18:3 C12:0	6.1×10^{-39} 1.9×10^{-35} 4.9×10^{-27}	16.8% 15.3% 11.8%	lysoPC a C20:4	4.2×10^{-27}	11.8%
rs174548	FADS1	C20:3 C18:3 C12:0	7.4×10^{-37} 7.4×10^{-35} 2.3×10^{-26}	15.9% 15.1% 11.4%	PC aa C38:4	1.1×10^{-25}	11.1%
rs174570	FADS2	C18:3 C20:3	2.1×10^{-10} 4.2×10^{-9}	4.3% 3.7%	PC aa C38:4	1×10^{-9}	4%
rs11849760	-	C24:0	1.1×10^{-9}	3.9%	PC aa C42:1	9×10^{-6}	2.1%
rs3100919	-	C24:0	1.9×10^{-8}	3.4%	PC aa C42:1	5.9×10^{-6}	2.2%

Table 3.2: Top associations between a SNP (SNP information) and FA frequency (computed FA), which have a lower p-value than an association between the SNP and one of the 151 measured metabolites in the KORA study (KORA metabolite). R^2 is the proportion of the variance that can be explained by the genotype in the linear model. 'PC aa' denotes a diacyl bonding between the FAs and the glycerol backbone of the phosphatidylcholine. The term 'lysoPC' describes PCs with only one FA.

of the KORA study. The association of C12:0 could be explained with a homology of *stearoyl-CoA desaturase (SCD)* to FADS1. SCD catalyzes the desaturation of C12:0 to C12:1. But it is more likely that this association is wrong.

Another SNP of FADS1 (rs174548) establish also strong associations with these three FAs (C20:3, C18:3 and C12:0). This supports the meaningfulness of our results further. In addition, there are two SNPs (rs11849760, rs3100919), which are not annotated with a gene locus, that show considerable associations with the FA C24:0. It is unsure how we have to evaluate these associations. These associations might be used as an indicator for annotation in the future.

Only five ratio based associations could be established, which were stronger than the strongest non-ratio association (see table 3.3). Additionally, the results are quite controversial with respect to the FAs that were used to calculate the ratios. It is rarely the case that the two FAs are nearby in the biosynthesis pathway of unsaturated FAs as it would had been expected. One exception is the association between a SNP of FADS1 (rs174547) with C18:1 / C18:3. Overall the ratio associations are dominated by SNPs of FADS1. In fact, the 50 strongest associations were established with either one of these SNPs. Therefore, we have to conclude that the ratio approach was not quite that useful as it has been for

SNP identifier	Locus	Association with	p-value	p-gain	R^2
rs174547	FADS1	C12:0 / C20:4	9.7×10^{-51}	5.1×10^{23}	21.5%
		C20:3 / C18:1	8.2×10^{-41}	7.5×10^{1}	17.6%
		C18:1 / C18:3	3.9×10^{-40}	4.9×10^{4}	17.3%
rs174548	FADS1	C12:0 / C20:4	1.4×10^{-49}	1.6×10^{23}	21.0%
		C18:2 / C22:4	3.7×10^{-40}	9.7×10^{21}	17.3%

Table 3.3: Listing of the associations between a SNP and the ratio of two FA frequencies, which show a stronger association than non-ratio associations. The p-gain is calculated by the lower p-value of the non-ratio association of the both FAs with the SNP dividedÄ by the p-value of the ratio association. R^2 is the proportion of the variance that can be explained by the genotype in the linear model.

Illig et al. We have two possible explanations for this fact. One, the computed FA frequencies lack in correctness which is revealed by the ratio calculation. Two, it might be the case that the ratio approach is only applicable for concentrations and not for frequencies.

Correlations between Fatty Acids

Fatty acids and their frequencies underlie a correlation due to their biosynthesis, e.g. desaturation and elongation of FAs (see section 1.1.1). Therefore we evaluated whether the computed frequencies of the imputing method show such a correlation. This would support the meaningfulness of the computed FA frequencies even further.

The Pearson product-moment correlation coefficient (CC) was calculated for the computed FA frequencies (see figure 3.13A). The two FAs C22:0 and C24:0 show the highest CC. This correlation could be explained by an elongation process. The second best correlation is between C20:5 and C24:6. There is a huge distance between these two FAs in the biosynthesis pathway. Hence this correlation is less likely to be justified by elongation and desaturation processes.

Overall 11% of the combinations show a CC above 0.6 and a significance level greater than 0.05 after Bonferroni correction (see figure 3.13B). For a large proportion, these combinations can not be directly referred to the biosynthesis processes as it would had been expected. Hence we conclude based on this result that there is a need for improvement of the computed FA frequencies. Another factor is the usage of relative frequencies for the calculation. I could be the case, that relative frequencies bias this calculation, because there are already dependencies in between these frequencies. For instance, an increase of the FA A automatically leads to a decrease of FA B.

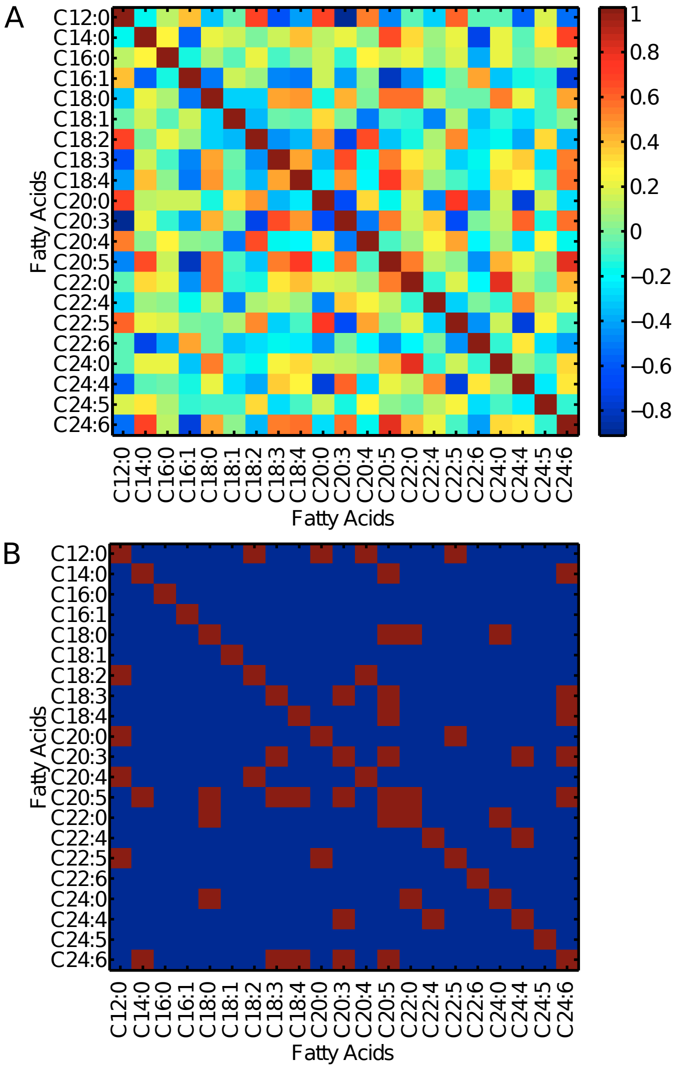

Figure 3.13: Correlation between the computed FA frequencies. (A) The cells are colored by the Pearson product-moment correlation coefficient (CC) of the two corresponding FAs. (B) Boolean representation of the correlation. A cell is colored red, when it fulfill the following conditions: (1) CC > 0.6 (2) $\alpha < 0.05$ after Bonferroni correction. Otherwise it is colored blue.

fatty acid	correlation coefficient	95% CI
C12:0	0.1404	0.0769 – 0.2028
C14:0	0.1802	0.1174 – 0.2417
C16:0	0.0463	-0.0180 – 0.1102
C16:1	0.1866	0.1238 – 0.2479
C18:0	0.0421	-0.0222 – 0.1061
C18:1	-0.0083	-0.0725 – 0.0560
C18:2	0.0108	-0.0535 – 0.0750
C18:3	0.0816	0.0175 – 0.1451
C18:4	0.2758	0.2154 – 0.3342
C20:4	0.0364	-0.0279 – 0.1004
C20:5	0.4645	0.4126 – 0.5134
C22:4	0.1667	0.1036 – 0.2285

Table 3.4: Listing of the Pearson product-moment correlation coefficient between the measured and the computed fatty acid frequencies. CI denotes the confidence interval.

Correlation with the Measured Free Fatty Acid Concentrations

The measured frequencies of FAs (see section 2.3.1) were used to examine the correlation to the FA frequencies which were computed by the imputing method. The set of measured frequencies contains 27 FAs. Only 12 of these FAs were also existent in the used FAP ORI and therefore only 12 correlations could be examined. In general, these two sets of frequencies show almost no correlation (see table 3.4). The average Pearson product-moment correlation coefficient (CC) was 0.1353.

The frequencies of the FA C20:5 obtained the highest CC of 0.4645 (see figure 3.13A). This could be due to the enrichment of frequencies close to 0. The lowest CC of -0.083 is achieved by the FA C18:1 (see figure 3.14B). This is very close to 0 and thus there is almost no linear relationship between these sets of frequencies.

Our results of the genetic associations show that the computed frequencies are meaningful with respect to the biological context. In combination with the results of this examination, we conclude that the pool of free FA (measured FA frequencies) does not correlate with the pool of FAs which are used for the biosynthesis of PCs (computed FA frequencies).

3.4 Mouse Nutritional Challenging Dataset

Besides the KORA dataset, we examined a second dataset which consists out of real data. This dataset was based on the study of a nutritional challenging with mice (see section 2.3.2). The imputing method was performed for 167 samples

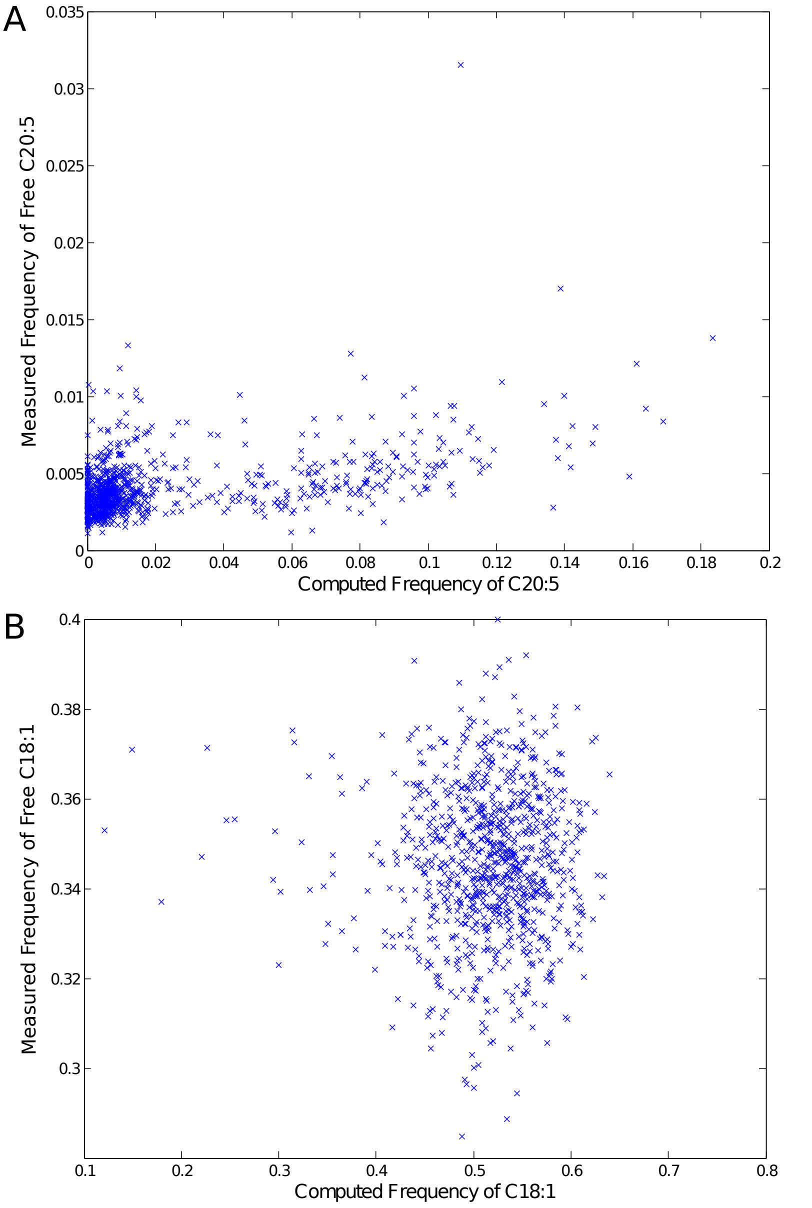

Figure 3.14: Exemplary correlations between computed and measured fatty acids. (A) The correlation of FA frequencies of C20:5 has the CC of 0.4645. (B) The two sets of C18:1 frequencies obtain a CC of -0.083.

diet	number of week				
	1	2	3	4	5
Control	18	27	27	2	2
Distel	20	25	31	-	-
DistelR	-	-	-	9	6

Table 3.5: Number of mice for every week and diet. 'Control' denotes the standard diet and 'Distel' the safflower oil diet. 'DistelR' describes the group of mice which received a standard diet after three weeks of the safflower oil diet. Thus there are only samples of the fourth and fifth week available.

with 100 runs per sample. The used FAP was ORI (see section 2.1.7 which was also used in the analysis of the KORA dataset. It might have been necessary to evaluate the FAPs with the mouse data a second time or to establish a new FAP. This was not possible due to the absence of genotyping data of the mice. An evaluation of the FAPs only based on the resulting costs can be quite difficult and leads to false conclusions.

The focus of this section is to evaluate whether we can observe differences in the resulting FA frequencies between the mice which received different diets. This would support the fact that the results of the imputing method are meaningful in a biological context. For this purpose, we divided the mice into three groups ('Control', 'Distel' and 'DistelR') according to their received diet. The additional information about the mouse strains is of no relevance of our examination and therefore not considered in the grouping. The number of mice is shown in table 3.5. It is important to emphasize the small number of samples in the fourth and fifth week that received a standard diet. We have to consider this small sample size in the evaluation of the results.

The computed frequencies of the FA C18:4 show the most noticeable difference between the mice which received the standard and safflower oil diet (see figure 3.15). The safflower oil diet apparently leads to an increase of the frequency of C18:4. After dropping the safflower oil diet (week four and five) the frequencies went down to the level of the mice which received a standard diet. Before we will discuss an explanation of this effect, we point out the result of the FA C18:2. The frequencies of C18:4 of 'Control' cover a broad range from 0.1 to 0.5 besides week four and five, but this could be due to the limited number of samples. 'Distel' shows almost exclusively values above 0.5 (see figure 3.16) with week 1 being the exception. This could be due to the fact that the diet only lasted for one week at this point. Nevertheless an increase of the median is still noticeable.

The increase of C18:2 and C18:4 for the 'Distel' group can be conducted to the composition of safflower oil. It consists of 81.4% C18:2, 11.1% C18:1, 5.5% C16:0,

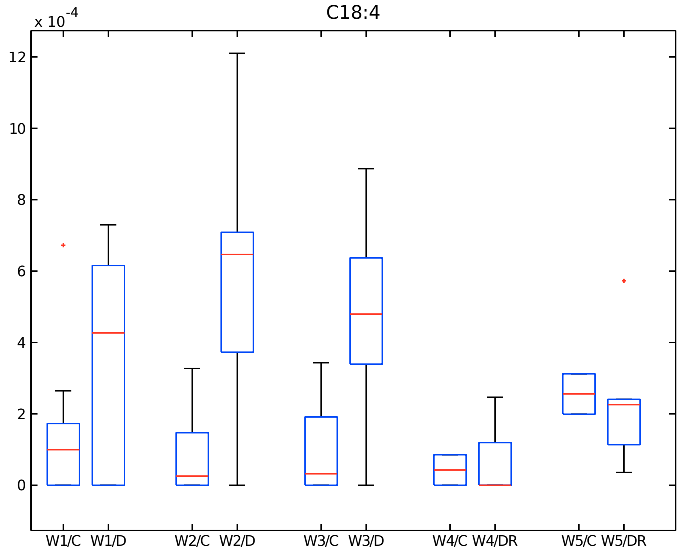

Figure 3.15: Comparison of the C18:4 FA frequencies of the groups of mice which received different diets. WX denotes the number of week, C represents the diet 'Control', D describes 'Distel' and DR stands for 'DistelR'.

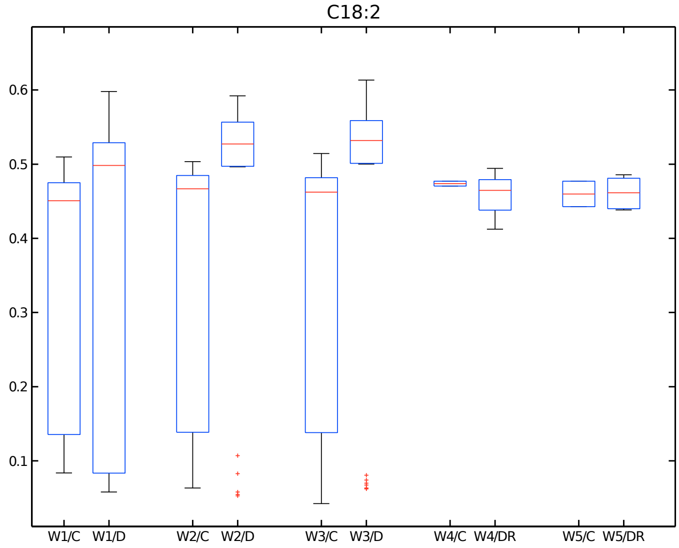

Figure 3.16: Comparison of the C18:2 FA frequencies of the groups of mice which received different diets. WX denotes the number of week, C represents the diet 'Control', D describes 'Distel' and DR stands for 'DistelR'.

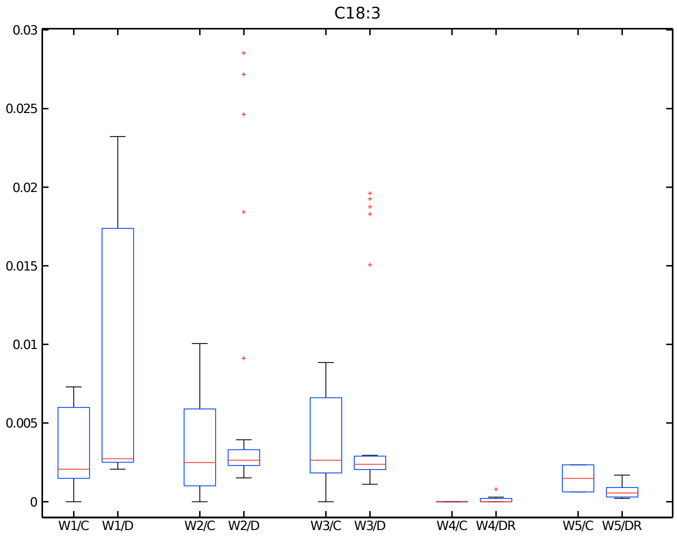

Figure 3.17: Comparison of the C18:3 FA frequencies of the groups of mice which received different diets. WX denotes the number of week, C represents the diet 'Control', D describes 'Distel' and DR stands for 'DistelR'.

1.6% C18:0 and 0.4% C18:3 [17]. Therefore, the results of C18:2 are directly associated with the high share in safflower oil. C18:4 do not occur in safflower oil, but still our results suggest an increase of it. We assume that the higher abundance of C18:2 lead to an increase of the desaturation process to C18:3 and afterwards to C18:4. This increase is not quite that noticeable for median frequencies of C18:3, but the outliers support our thesis (see figure 3.17).

Overall we have to conclude that it is tough to evaluate the significance of this results due to the limited number of samples. It could be the case that more samples would lead to an even stronger signal of the safflower oil diet.

Chapter 4

Summary and Outlook

Phosphatidylcholines (PCs) consist of a glycerol backbone, a phosphate group with an attached choline and two fatty acids (FAs). The mass spectrometry technology used for the determination of metabolite concentrations in the KORA and mouse nutritional challenging studies is not able to resolve the FA composition of the PCs. In this thesis, we attempted to develop a method to calculate the FA frequencies which are used for the biosynthesis of phosphatidylcholines. This calculation is based on the measured PC concentrations and the computed FA frequencies would help to understand the FA composition of measured PCs in detail. For this purpose, we implemented an imputing method which relies on the heuristic approach of simulated annealing.

After establishing some fundamental results, e.g. the existence of local minima in the search space and how these local minima affect the imputing method in the finding process of the global minima, we showed the correctness of our method with the usage of toy data. Toy data was used as an evaluation dataset and the influence of noise, correlated data and missing PCs frequencies was demonstrated, because these three factors occur in the work with real datasets.

Before we were able to analyze the available datasets with real data (KORA and mouse nutritional challenging), we had to do an evaluation of different fatty acid pools (FAP). The FAP is the set of FA whose frequencies are computed by the imputing method. The evaluation of the FAPs was not only based on the resulting costs of the imputing method, because there is the possibility of overfitting. Therefore we considered an external validation which was provided by genotyping data. The genotyping data was used to establish genetic associations. The quality and quantity of these associations were used as an indicator of the meaningfulness of the resulting computed FA frequencies in respect to the biological context.

The results of the FAP ORI for the KORA and mouse nutritional challenging datasets were analyzed in more detail. The meaningfulness of the computed FA frequencies was supported by the results of the genetic associations, e.g. the association of C20:3 with a SNP of FADS1. It was shown by the correlation of the computed FA frequencies among each others and the genetic associations with ratios of FA frequencies, that there is still a need for improvement in the computing process of the FA frequencies. The comparison of the computed and measured FA frequencies indicate that the free fatty acid pool does not correlate with the fatty acid pool, which is used in the biosynthesis of PCs. The analysis of the mouse nutritional challenging dataset revealed that the effects of the diet could be observed in the computed FA frequencies. This is another indicator for the meaningfulness of the computed FA frequencies.

The analysis of the results for this thesis revealed several aspects which could be examined in the future. Some of these were already mentioned in the discussion of the results. For instance, we only examined the influence of noise with a reduced set of generated PCs. It would be interesting to see whether the combination of noise superposition and the complete set of generated PCs results in significant lower costs. We also separated the examination of noise and correlated data. Real datasets contain both factors. Therefore it would simulate the work with a real dataset the most, if we would have combined these factors and applied on the toy data to evaluate the effects. A new consideration would be the extension of the existing framework with a gradient descent function based on NMF [16], which is currently under development in our workgroup. The gradient descent function would be applied on the result after imputing and could lead to an improvement of the resulting costs.

Besides these small aspects of existing results, there are further possibilities to accumulate new results which could provide helpful insights. One, we performed the whole analysis for the data of dicayl PCs in this thesis. The KORA dataset also contains concentrations of acyl-alkyl PCs. The application of the imputing method on this data would generate more results, which could be used to improve the performance of the imputing method. Two, there is another dataset in prospect that contains measured FA frequencies of PCs. This dataset could be used as an evaluation of the computed FA frequencies. This could optimize the generation function of the imputing method. Finally, further improvements, investigations and expansions of the imputing method and their results could help to understand intracellular processes of PC biosynthesis.

Bibliography

[1] Berg, J.M., Tymoczko, J.L., and Stryer, L. *Biochemistry.* W. H. Freeman, sixth edition edition, 2006. ISBN 0716787245.

[2] Blow, N. Metabolomics: Biochemistry's new look. *Nature*, 455(7213):697–700, 2008. doi:10.1038/455697a.

[3] Brockerhoff, H., Hoyle, R.J., and Wolmark, N. Positional distribution of fatty acids in triglycerides of animal depot fats. *Biochim Biophys Acta*, 116(1):67–72, 1966.

[4] Chace, D.H., Sherwin, J.E., Hillman, S.L., Lorey, F., and Cunningham, G.C. Use of phenylalanine-to-tyrosine ratio determined by tandem mass spectrometry to improve newborn screening for phenylketonuria of early discharge specimens collected in the first 24 hours. *Clin Chem*, 44(12):2405–2409, 1998.

[5] Conquer, J.A., Tierney, M.C., Zecevic, J., Bettger, W.J., and Fisher, R.H. Fatty acid analysis of blood plasma of patients with alzheimer's disease, other types of dementia, and cognitive impairment. *Lipids*, 35(12):1305–1312, 2000.

[6] Crowe, F.L., Allen, N.E., Appleby, P.N., Overvad, K., Aardestrup, I.V., Johnsen, N.F., Tjonneland, A., Linseisen, J., Kaaks, R., Boeing, H., Kröger, J., Trichopoulou, A., Zavitsanou, A., Trichopoulos, D., Sacerdote, C., Palli, D., Tumino, R., Agnoli, C., Kiemeney, L.A., de Mesquita, H.B.B., Chirlaque, M.D., Ardanaz, E., Larranaga, N., Quiros, J.R., Sanchez, M.J., Gonzalez, C.A., Stattin, P., Hallmans, G., Bingham, S., Khaw, K.T., Rinaldi, S., Slimani, N., Jenab, M., Riboli, E., and Key, T.J. Fatty acid composition of plasma phospholipids and risk of prostate cancer in a case-control analysis nested within the european prospective investigation into cancer and nutrition. *Am J Clin Nutr*, 88(5):1353–1363, 2008.

[7] Dougherty, R.M., Galli, C., Ferro-Luzzi, A., and Iacono, J.M. Lipid and phospholipid fatty acid composition of plasma, red blood cells, and platelets and how they are affected by dietary lipids: a study of normal subjects from italy, finland, and the usa. *Am J Clin Nutr*, 45(2):443–455, 1987.

[8] Engl, H.W., Flamm, C., Kügler, P., Lu, J., Müller, S., and Schuster, P. Inverse problems in systems biology. *Inverse Problems*, 25, 2009.

[9] Hodson, L., Skeaff, C.M., and Fielding, B.A. Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake. *Prog Lipid Res*, 47(5):348–380, 2008. doi:10.1016/j.plipres.2008.03.003.

[10] Holle, R., Happich, M., Löwel, H., Wichmann, H.E., and Group, M.O.N.I.C.A.O.R.A.S. Kora–a research platform for population based health research. *Gesundheitswesen*, 67 Suppl 1:S19–S25, 2005.

[11] Holub, B.J., Wlodek, M., Rowe, W., and Piekarski, J. Correlation of omega-3 levels in serum phospholipid from 2053 human blood samples with key fatty acid ratios. *Nutr J*, 8:58, 2009. doi:10.1186/1475-2891-8-58.

[12] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B.S., Mewes, H.W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D., Adamski, J., and Suhre, K. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2):137–141, 2010. doi: 10.1038/ng.507.

[13] Kanehisa, M. and Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[14] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. doi:10.1126/science.220.4598.671.

[15] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F.J. High-throughput metabolite profiling reveals metabolic pathway signatures in blood serum samples. *submitted*, 2010.

[16] Lee, D.D. and Seung, H.S. Algorithms for non-negative matrix factorization. *Citeseer*, 2001.

[17] Lee, Y.C., Oha, S.W., Changa, J., and Kim, I.H. Chemical composition and oxidative stability of safflower oil prepared from safflower seed roasted with different temperatures. *Food Chemistry*, 84:1–6, 2004.

[18] Lundy, M. and Mees, A. Convergence of an annealing algorithm. *Math. Program.*, 34(1):111–124, 1986. ISSN 0025-5610. doi: http://dx.doi.org/10.1007/BF01582166.

[19] McKeone, B.J., Osmundsen, K., Brauchi, D., Pao, Q., Payton-Ross, C., Kilinc, C., Kummerow, F.A., and Pownall, H.J. Alterations in serum phosphatidylcholine fatty acyl species by eicosapentaenoic and docosahexaenoic ethyl esters in patients with severe hypertriglyceridemia. *J Lipid Res*, 38(3):429–436, 1997.

[20] Phillips, G.B. and Dodge, J.T. Composition of phospholipids and of phospholipid fatty acids of human plasma. *J Lipid Res*, 8(6):676–681, 1967.

[21] Raatz, S.K., Bibus, D., Thomas, W., and Kris-Etherton, P. Total fat intake modifies plasma fatty acid composition in humans. *J Nutr*, 131(2):231–234, 2001.

[22] Shaham, O., Wei, R., Wang, T.J., Ricciardi, C., Lewis, G.D., Vasan, R.S., Carr, S.A., Thadhani, R., Gerszten, R.E., and Mootha, V.K. Metabolic profiling of the human response to a glucose challenge reveals distinct axes of insulin sensitivity. *Mol Syst Biol*, 4:214, 2008. doi:10.1038/msb.2008.50.