LUDWIG-MAXIMILIANS UNIVERSITÄT MÜNCHEN
TECHNISCHE UNIVERSITÄT MÜNCHEN

Institute of Bioinformatics and Systems Biology;

Helmholtz Center Munich

Masters Thesis

in Bioinformatik

A systems biological approach on phenomic and metabolic relationships in the Qatar Metabolomics Study of Diabetes

Ulrich Neumaier

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Dr. Jan Krumsiek, Dr. Gabi Kastenmüller, Prof. Dr. Karsten Suhre
Abgabedatum:	15.10.2013

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

October 11, 2013 _____

Ulrich Neumaier

Acknowledgement

At first I would like to thank Professor Dr. Fabian J. Theis and Professor Dr. Karsten Suhre for the opportunity to do my masters thesis in their lab. I also would like to thank my two supervisors Dr. Jan Krumsiek and Dr. Gabi Kastenmüller for their help and their inspiring impulses during the masters thesis. Furthermore I would like to thank Dr. Noha A.Yousri for the inspiring discussions during my stay in Qatar. Additionally I would like to thank the rest of the group of Professor Theis. Last I would also like to thank the rest of the group of Professor Suhre for their help during my stay in Qatar.

Abstract

With the help of genome-wide association studies, more and more information about the background of various diseases gets discovered, but by far not everything is revealed yet, as the genetic background is not enough to explain diseases. Another part that has to be solved, but is still in the early stages, is the phenomics, as different phenotypes are risk factors for others. In order to explain such a linkage of two or more phenotypes, factors like genes or metabolites, have to be taken into consideration. An example for phenotypes that are risk factors for each other would be obesity, which heightens the risk of getting diabetes which in turn heightens the risk of getting heart diseases.
In order to be able to reveal more of those comorbidities, a systems biological approach like calculating correlation networks or Gaussian Graphical Models is very useful, as the dependencies of the phenotypes can be seen in those networks. This approach is already used at the genomic and metabolic level with great success. However the phenomic datasets bring new challenges with them, as there is a broad range of different measurements for the phenotypes, which means a very heterogeneous dataset. One example, for such a difficulty would be the fact, that phenotype data often contains binary data, which can only be handled by correlation calculations and the more advanced mixed graphical models, but not by Gaussian Graphical models.
In order to examine those problems those two methods were applied to the Qatar Metabolomics Study of Diabetes. The most significant problems which were found during the examination were the fact, that the dataset is very heterogeneous. Therefore three different methods for the calculation of the correlation networks were used in order to differentiate between edges between two binary phenotypes, a binary and a continuous phenotype and two continuous phenotypes. Another problem that came with the binary phenotypes, was the problem of imputing. The problem of finding the best method for imputing is difficult for continuous phenotypes, but for the binary phenotypes, it gets even more difficult. This is due to the fact that it is only possible to impute with two different values, which heavily biases the results, as shown in this thesis. Therefore a complete dataset would be the most important thing, for those researches, but can not be achieved in reality most of the time. Another problem that had to be solved, was whether correcting for co-factors in the case of correlation networks is really an advantage, or whether it is better not to correct for confounders. Based on the results of this work the answer to this question is dependent on the questions that have to be answered with the network. Therefore, sometimes the correction has advantages and sometimes no correction is better.
After solving those problems, the phenotype heart disease was examined. This was done by building a network of all the related phenotypes and metabolites for it and examining the relationships to the other phenotypes, like retinopathy or metabolites, like creatine. Another question was which of the weight related parameters describe the overall health status of a person best. Therefore another network was built, which showed that not one weight parameter is able to describe all the risk factors, but certain risks can be described best by certain anthropometric measures. Therefore it is best to measure more than just one in order to get

the risks for all the weight related diseases.
Ultimately the mixed graphical models would be a very useful tool, in order to discover new and interesting correlations between phenotypes and metabolites, if the dataset allows it. This means, if the dataset is not biased and has not to many missing values, which both lower the specificity of the algorithm drastically. For such a dataset I would recommend a correlation based network, which might not provide the best results, but is more stable towards missing values.

Kurzfassung

Genomweite Assoziationsstudien liefern immer mehr Hintergrundwissen über die verschiedensten Krankheiten, aber dieses Wissen reicht bei weitem noch nicht aus, da Krankheiten nicht nur einen genetischen Grund haben können. Ein weiterer Grund weswegen ein Patient eine Krankheit bekommen kann, ist auf Grund bestimmter Phenotypen. Die Untersuchung solcher Phenotypverbindungen ist jedoch noch am Anfang. Da solche Verbindungen zwischen zwei phenotypen die unterschiedlichsten Gründe haben kann, muss man zusätzliche Faktoren, wie Gene oder Metaboliten mit einbeziehen. Ein Beispiel für eine solche Phenotypenverbindung wäre Fettleibigkeit, welche das Risiko Diabetes zu bekommen deutlich erhöht. Diabetes wiederum erhöht das Risiko auf Herzkrankheiten.
Ansätze der Systembiologie, wie das berechnen von Korrelationsnetzwerken, oder Gaußsche graphische modelle sind sehr hilfreich, um weitere solcher Abhängigkeiten zweier Phenotypen zu finden. Diese Methoden wurden schon mit großem Erfolg auf dem Gebiet der Metaboliten und der Gene verwendet. Jedoch bringt der Phenotypdatensatz neue Herausforderungen mit sich, da dieser mit vielen verschiedenen Methoden gemessen wurde, was einen sehr heterogenen Datensatz zur Folge hat. Eine Herausforderung, welche ein Phenotypendatensatz mit sich bringt, wären zum Beispiel die binären Datenreihen, welche nur durch Korrelationen und durch die gemischten graphischen modelle berechnet werden können, aber nicht mit Gaußschen graphischen Modellen.
Um die Probleme zu untersuchen, wurden diese beiden Ansätze auf der "Qatar Metabolomics Study of Diabetes" angewendet. Das größte Problem, das durch die Untersuchung hervor kam, war die Heterogenität des Datensatzes. Um zwischen den Korrelationen zweier binärer Phenotypen, einem binären zu einem kontinuierlichen Phenotypen und zweier kontinuierlicher Phenotypen zu unterscheiden, wurden drei verschiedene Methoden verwendet. Ein weiteres Problem, das durch die binären Daten verursacht wurde, war ein Problem beim Imputen. Eine richtige imputing Methode für kontinuierliche Daten zu finden ist schwierig, jedoch noch viel schwieriger ist es eine gute Methode für binäre Daten zu finden. Dies hat den Grund, das man bei binären nur mit zwei Variablen imputen kann, was, wie in dieser Thesis gezeigt, die Ergebnisse stark verfälschen kann. Deswegen wäre ein vollständiger Datensatz sehr wichtig für solche Untersuchungen, dies kann in der Realität aber nur selten verwirklicht werden. Ein weiteres Problem war die Korrektur von Co-faktoren und ob es besser ist für solche Co-faktoren zu verbessern, oder nicht. Basierend auf den Resultaten dieser Arbeit, ist dies aber abhängig von der Fragestellung, welche beantwortet werden soll. Deswegen sollteman manchmal eine Korrektur bevorzugen und manchmal nicht.
Nach dem diese Problem behandelt wurden, wurde ein Herzkrankheitsnetzwerk untersucht. Dafür wurde ein Netzwerk mit allen relevanten Phenotypen, wie Retinopathy und Metaboliten, wie Kreatin, für Herzkrankheiten gebaut. Eine weitere Fragestellung, die durch ein weiteres Netzwerk bearbeitet wurde, war welche anthropometrische Messung am besten die Krankheitsrisiken, welche mit dem Gewicht in Verbindung zu bringen sind, zusammenfasst. Aus dem Netzwerk geht hervor, dass man am besten nicht nur eine Messung vornimmt, sondern mehrere antropometrische Messungen.

Letztenendes wären die gemischten graphischen modelle sehr hilfreich um neue und interessante Korrelationen zwischen Phenotypen und Metaboliten zu finden, wenn es der Datensatz zulässt. Für einen Datensatz, der sehr viele fehlenden Einträge hat, würde ich ein Korrelationsnetzwerk empfehlen, da es vielleicht nicht die selben Ergebnisse, wie ein graphisches modell liefert, aber viel robuster gegenüber fehlende Daten ist.

Contents

1 Introduction

1.1 Phenomics and the future challenge for systems biology

In the recent years, genome-wide association studies (GWAS) had a big impact on systems biology, as those studies provided a huge amount of data for genotype to phenotype interactions. These information ended in big databases, which provide some insight in the phenotypes and their genomic background. One major problem that can still not be answered with those networks, is how two people with the same genetic background can have different phenotypes. Therefore not only the genetics have to be examined but also the correlations between phenotypes and the metabolic background. Metabolites and phenotypes can also reflect different influences on a disease, such as environmental factors, which sometimes have a huge impact on diseases. One disease that can be caused through genetic background, as well as metabolomic changes and influences of other phenotypes like weight of the patient would be diabetes. As the GWAS will be able to provide the information for the genomic part in the near future, there are still problems in order to understand diabetes. One would be the metabolomics part and another one would be the phenomics part. The phenomics part is arguably the most challenging part, as there is a broad range of phenotype measurements and therefore the complexity of the phenomic datasets and the processing of the complex phenomic datasets set the major tasks for the systems biology.

1.2 Weight development, obesity and different weight parameters

One of the major health concerns of the 21st century is obesity as it is one of the leading causes of preventable death.[4] This is also getting worse, as obese children become obese adults and face multiple health problems that are associated with obesity at younger ages. This is also shown by the raise in number of obese people in the recent years. Figure 1.1 shows that the mean body mass index (BMI) increased in all the listed countries over the last decade and that the overall highest mean BMI was located up until 2009 in the USA. Thus also the comorbid diseases, that go along with a high BMI are getting more and more of a problem. Those diseases are for example, heart diseases, osteoarthritis, obstructive sleep apnea, certain types of cancer and type 2 diabetes.[25] Thus obesity overall also heightens the risk of death. Figure 1.2 demonstrates that slight overweight is not severe, though very obese people have to face a overall higher mortality rate. Normally the overall mortality rate is also higher in the underweight group, except for one group, the age group 25-59 where a positive effect of having underweight can be seen.

The fact that influences of obesity on human health develops to be a major concern, the

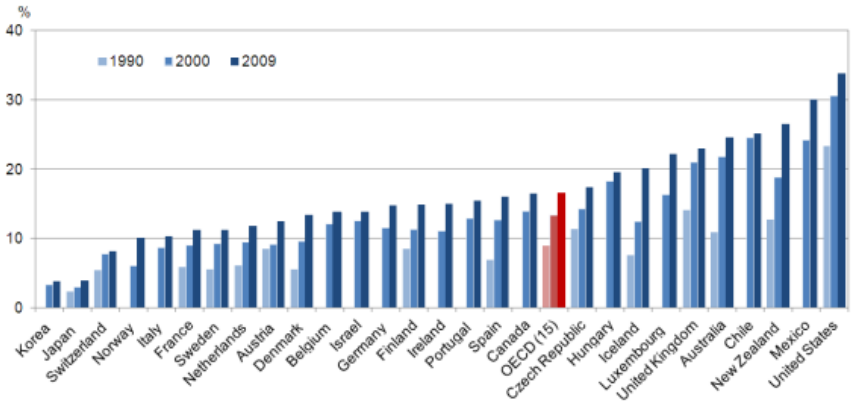

Figure 1.1: The amount of obese people in developed countries. Over the years 1990 to 2009 the percentage in all the shown countries rose drastically. Especially shocking is the fact that some countries have more than 20% of obese people.[60] This figure was taken from "www.downeyobesityreport.com/2012/06/"

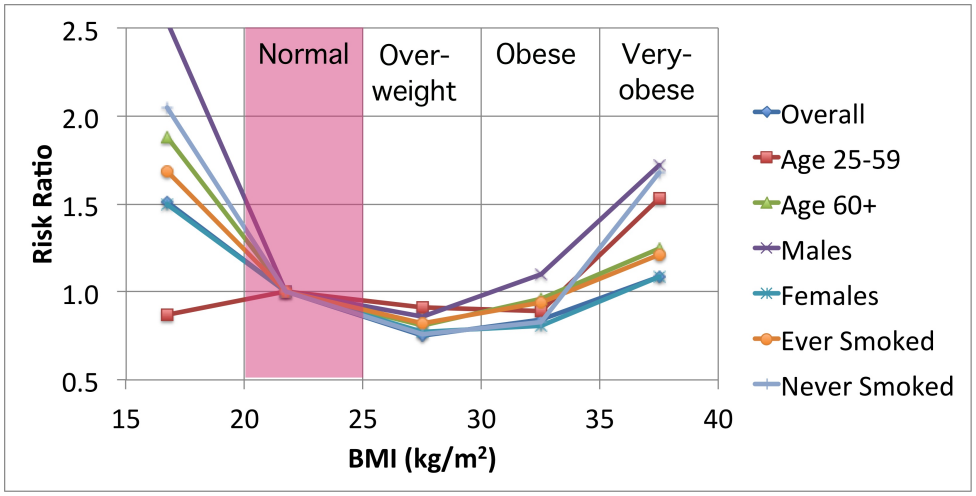

Figure 1.2: The mortality rate for different BMI categories. Additional to the overall population also always two comparable curves are shown. Those would be the two age groups, the two gender groups and the smoking habit. The interesting thing to see here is that for an age between 25 and 59 underweight is not crucial according to the mortality rate, whereas the overall death rate is heightened in all the other groups by underweight. Overall, the graph makes a big "U"-turn, where the two extremes, obesity and underweight, are not healthy for nearly every group.[76] This figure was taken from "http://protonsforbreakfast.wordpress.com/category/obesity/".

health and pharmaceutical industry, developed different health parameters. Those anthropometric measures take different parameters into consideration, are important, in order to be able to categorize the patients. For example, the most used value is the body mass index (BMI), which gives an estimation on how the weight to height relationship is. The BMI value is calculated by the Equation 1.1. Most of the time there are different categories which divide the people into different BMI increments which can be seen in Table 1.1. Those classifications give an overall better understanding, what the BMI values mean.

BMI	Classification
<18.5	underweight
18.5-24.9	normal weight
25.0-29.9	overweight
30.0-34.9	class I obesity
35.0-39.9	class II obesity
>40.0	class III obesity

Table 1.1: Different weight classifications by BMI.[81] Nevertheless there are sometimes little deviations in those categories, as some times other cofactors, like age or sex, are taken into consideration.[29]

$$BMI = \frac{weight}{height^2} \tag{1.1}$$

Another well accepted anthropometric measure is the Waist-to-Hip Ratio (WHR) which gives a good estimate on how the fat is distributed throughout the body. WHR is calculated by dividing a persons waist circumference by the hip circumference (see Equation 1.2). With the help of the WHR one can estimate, whether a person has more of an apple or a pear shaped body (see Figure 1.3).[72] The apple shape is considered for men with a WHR over 1 and for women with an WHR over 0.8 and means, that the fat is mostly located around the stomach area. The pear shaped persons carry most of their fat around the buttocks.[72] The shape is not only a visual difference, but it is also a factor that influences the healthiness of a person, as the pear shaped body form seems to be superior to the apple shaped body form, as in general the pear shaped persons have a lower risk of getting obesity related diseases such as diabetes or heart diseases.

$$WHR = \frac{waist}{hip} \tag{1.2}$$

As the obesity issue got more and more interesting over the last decades, some other anthropometric measures have been investigated, which try to combine previous knowledge about the overall body fat percentage and the allocation of the fat in order to give a good and easy estimate on how high the risks of getting obesity related diseases are for a patient. One of those measures is the *Body Adiposity Index (BAI)* which takes into account the waist circumference and the height, as those two are very correlated to the percentage of body fat.[6] The BAI-value is calculated by the Equation 1.3 and was developed by Bergman et al. Another example is the *"A Body Shape Index" (ABSI)* which not only takes the waist circumference into account, but also the BMI. The ABSI is calculated by the Equation 1.4. The ABSI is a substantial risk factor for premature mortality and tries to unify the most frequently used anthropometric measures, the BMI and the WHR.[36]

$$BAI = \frac{100 \cdot hipCircumference}{height \cdot \sqrt{height}} - 18 \tag{1.3}$$

$$ABSI = \frac{WaistCircumference}{BMI^{\frac{2}{3}} \cdot height^{\frac{1}{2}}} \tag{1.4}$$

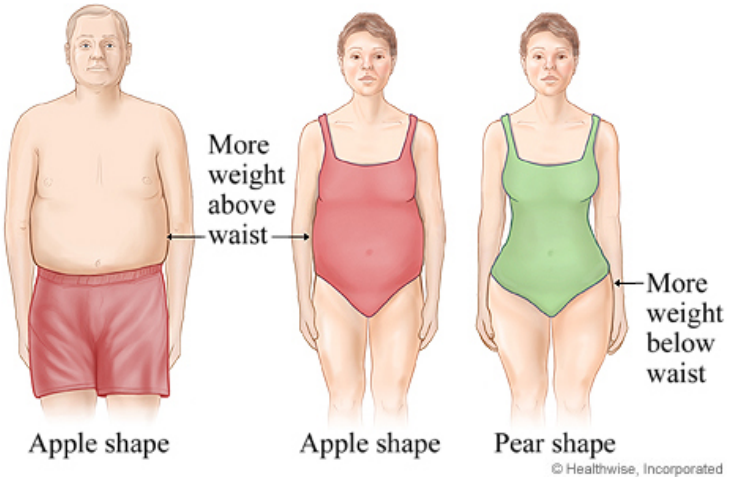

Figure 1.3: Different weight distribution and the apple and pear shape. Whereas the apple shaped body has more fat above the waist and is likely to develop health problems due to it, the pear shape body has more weight below the waist and shows less obesity related problems.[72] This picture was taken from `http://www.uofmhealth.org/health-library/zm6365`.

1.3 Diabetes

Diabetes is a metabolic disease examined by researchers all over the world, but the underlying pathomechanisms are still not fully understood as many different factors can cause someone to develop the disease.[83] One of the symptoms is a high blood sugar concentration, which can be caused either by insufficient insulin production of the pancreas or because the cells do not respond to the produced insulin.[42] The type of diabetes, where the β-cells in the pancreas are lost, which leads to an insulin deficiency, is called type-1-diabetes mellitus (T1D). Apart from that type-2-diabetes mellitus (T2D) is caused by insulin resistance but can also come along, with an overall lower insulin secretion.

1.3.1 Type-2-diabetes mellitus

Diabetes type 2 makes more than 90% of all diabetes ailments worldwide.[43] In the USA, it is the main reason for loss of sight, kidney failure and amputation.[58] Diabetes type 2 is caused by a insulin resistance. Even though the β-cells of the pancreas may have an hyperfunction, there is no response to the hormone in the different tissues although the insulin receptors are viable. One reason for developing diabetes mellitus type 2 would be genetic background, but there are also other disease-promoting factors, like obesity and physical inactivity.[84] In general diabetes can be diagnosed with a blood sugar test.[85] The value which shall not be overstepped is a fasting blood sugar value of 126 $\frac{mg}{dl}$ or higher, measured in the venous plasma. Another test for the diagnosis of diabetes is called oral glucose tolerance test (OGTT). For this test, the patient gets 75 gram of glucose and if the blood sugar value gets over 200 $\frac{mg}{dl}$, after a two hour waiting time, diabetes is confirmed.
In the early phases of diabetes, there is a lack of symptoms, which causes many cases of diabetes to stay undiagnosed in the early phases. Some premonitions that could be seen in the early stages of diabetes, would be frequent urination, feeling of weakness, thirst and dry skin.

However, most of the time diabetes is firstly diagnosed, when the side effects appear. Those comorbid conditions would be for example myocardial infarction, a stroke, renal insufficiency, retina damage, diabetic neuropathy or a diabetic foot.[3] These diseases are most of the time caused by the higher amount of sugar in the blood vessels, which can clump and block the blood flow (see Figure 1.4). If the vessels are attached to nerves this can lead to diabetic neuropathy (see Figure 1.6).[27]

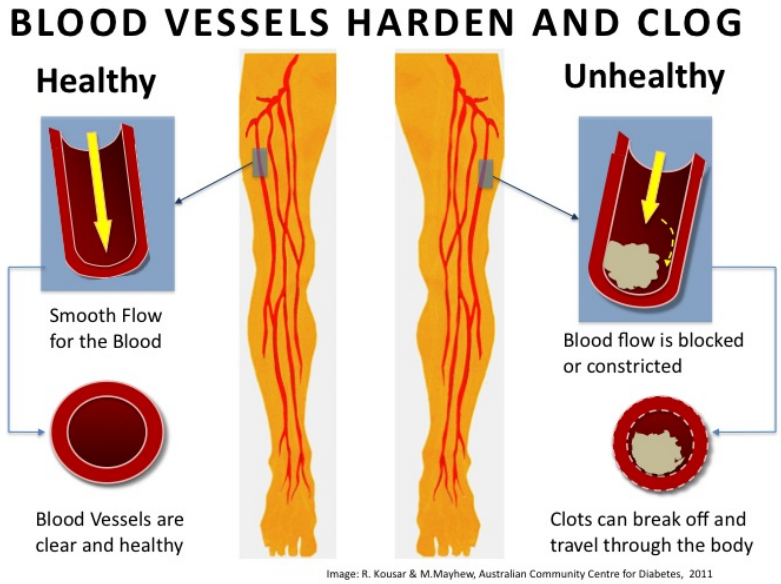

Figure 1.4: The effects of high blood sugar on blood vessels. On the left hand side one can see a healthy blood vessel, where nothing blocks the flow of blood. On the right side however the blood flow is blocked or constricted, which can be caused by the high amount of glucose in the blood of diabetes patients. This leads to the side effects of diabetes like the diabetic foot or diabetic neuropathy, as those blocked vessels are not capable of supporting the attached tissue with sufficient nutrition and oxygen.[63] This figure was taken from `"http://www.diabetesinfo.org.au/webdata/images/Blood20vessels20harden20and20clot.jpg"`

1.4 Heart disease

Heart diseases are the number one reason for deaths in the U.S. and is also cause of disability.[28] The most common reason to develop heart diseases is the blockage or constriction of the coronary arteries, which can be caused for example by a too high amount of glucose in the blood. Therefore heart diseases are a comorbid disease of diabetes.[50] This heart issue is developing slowly, as the clots in the vessels are growing slowly and leads most of the time to heart attacks. Heart diseases can also be an inborn error, like inborn deformities. One example would be a deformity of the cardiac valve, which therefore can not close properly and causes heart issues. This disease is called aortic regurgitation.[52] In order to lower the risks of getting heart diseases, controlling the risk factors is vital. Therefore controlling the blood pressure, lower the cholesterol, work out and reduce smoking is very important to lower the risks of getting heart diseases.[56]

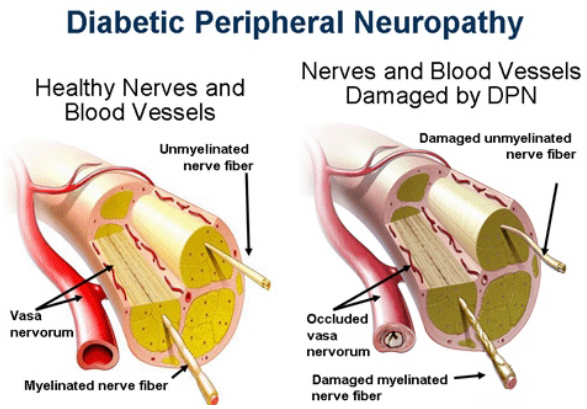

Figure 1.5: Neuropathy caused by damaged blood vessels On the left hand side, there is a healthy blood vessel, which is able to provide enough blood, oxygen and nutrition for the nerve it is attached to. On the right hand side the blood, oxygen and nutrition supply by the vessel is not given any more, as the vessel is blocked by glucose clumps, this leads to severe damage of the nerves, as they lack blood and nutrients.[39] This figure was taken from `"http://www.hyderabadendocrinology.com/content/diabetes-and-neuropathy"`

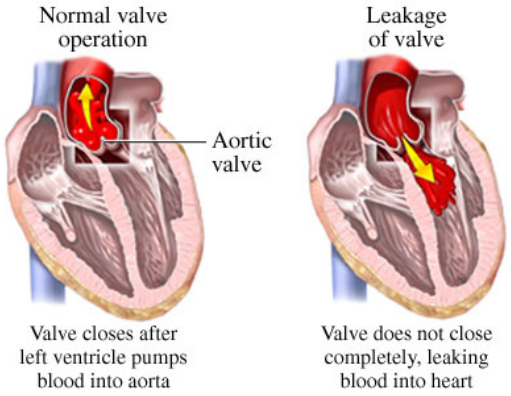

Figure 1.6: Aortic regurgitation On the left hand side a functional aortic valve can be seen, which closes properly after the ventricle pumps blood into the aorta. In contrary on the left hand side, the valve does not close properly and blood leaks back into the heart. This figure was taken from `"http://www.heart-valve-surgery.com/aortic-valve-regurgitation-symptoms.php"`

1.5 Metabolomics

Metabolomics is one of the recently uprising fields that provides new and interesting insights in diseases and their effects on an organism. Metabolomics means measuring concentrations of endogenous and also exogenous metabolites in different tissues or body fluids under certain conditions. The metabolome represents a snapshot of all the metabolites in a biological system and the influences of environmental factors. The metabolites are the smallest subunits, with which the proteins, the RNA and the DNA are put together (see Figure 1.7) by all organisms. However different factors such as nutrition, environmental factors and medical treatment complicate the examination of the metabolome. The nutrition which the patient takes in, can lead to false conclusions, as the metabolites that are contained in the food

can alter the metabolite concentrations drastically. The environmental factors are also able to influence the metabolite concentration, as the body reacts to the environmental changes, which can be seen in the metabolite concentrations. Last but not least, drugs are certainly able to change those concentrations as they target specific pathways in the body and thus can block the production of some metabolites or cause an overproduction of others. As the end product of the genetic setup, it can describe certain phenotypes best, as it provides a functional readout of the physiological state.[9]

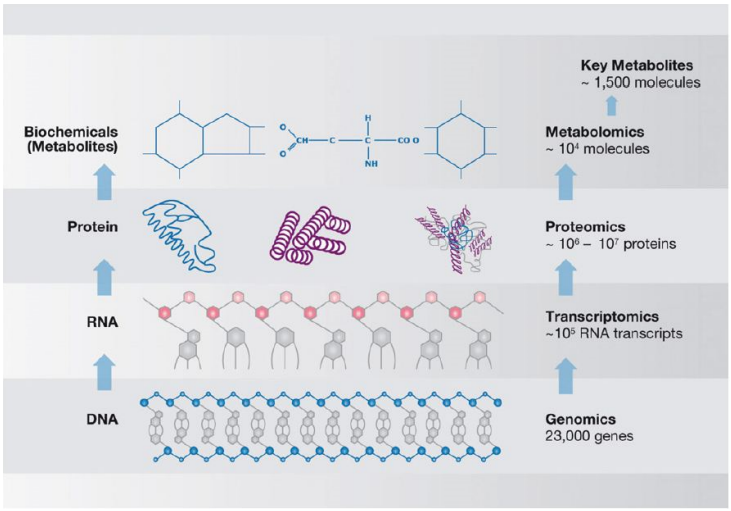

Figure 1.7: Coherences between DNA, RNA, proteins and metabolites: The Metabolites at top are the smallest building blocks of all the compartments in the organism. As the metabolites are the smallest subunit of those four, it offers the best resolution in order to understand the physiological state of any organism. It is also the true functional endpoint of biological events.[55] This figure was taken from Rudnicki et al.

In order to describe those relationships between the metabolites, networks are probably the best approach. As a metabolic network in general is a system of different chemical reactions, that are interconnected. One example for such a chemical reaction system would be the biosynthesis of valine, leucine and isoleucine, which can be seen in Figure 1.8. Many of those metabolic reaction systems are stored and can be viewed via metabolic database systems such as the Kyoto Encyclopedia of Genes and Genomes (KEGG).[34] However one has to always keep in mind that those networks are not absolute and may contain some errors. As well as some are not really errors, but just a generalization problem as there might be certain circumstances, that are able to change those chemical reaction systems. Therefore in a metabolic network in general an edge between two nodes means that there is a chemical reaction or sometimes reaction pathway, where one node, or better one metabolite can react to another metabolite.

Another type of network that is similar to the metabolite network is the signaling network. However in signalling networks, the essential purpose is the regulation of other processes, whereas the energy and mass flow, which is the main purpose of metabolic networks, is just a requirement in order to describe the regulation of the other processes.[37] In order to understand those metabolic networks, one has to combine different information like classical

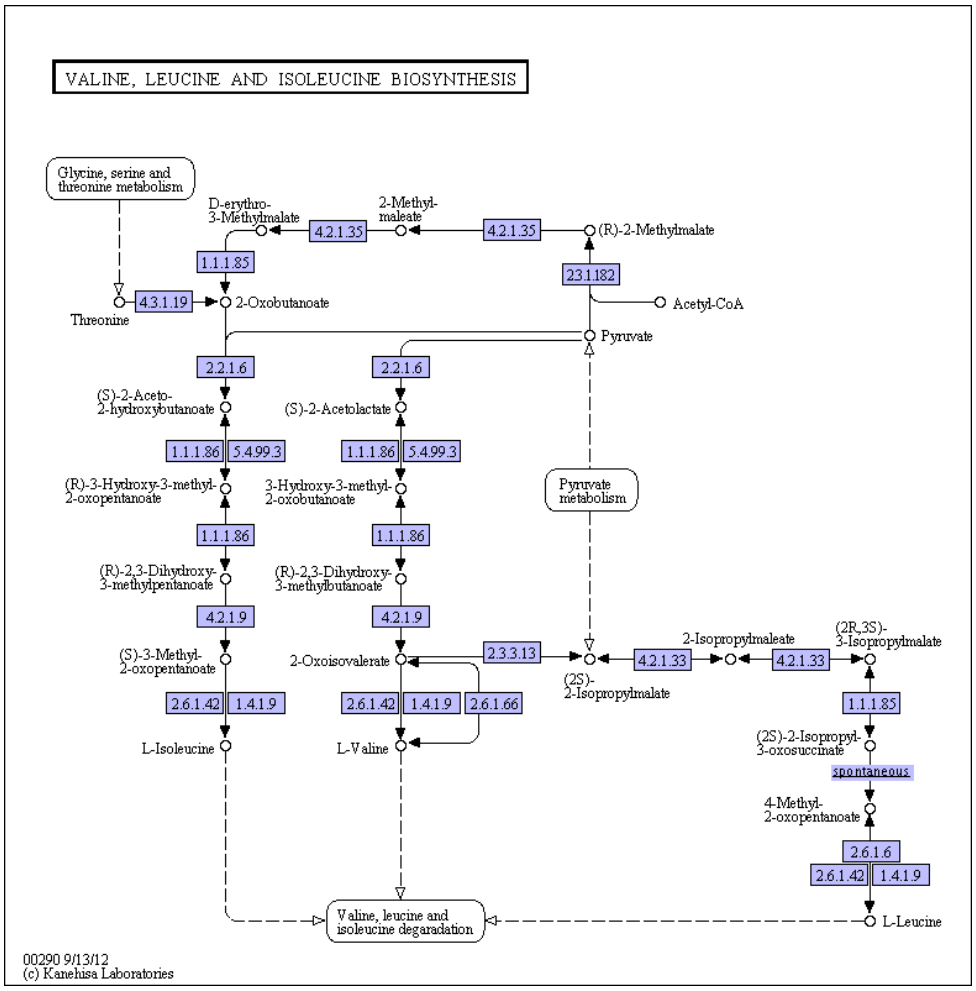

Figure 1.8: Valine, Leucine and Isoleucine biosynthesis. This picture shows the chemical reaction system represented as a network in the KEGG database.[40]

biochemistry, genomics, functional genomics, network analysis and simulations. With the development of measurement kits for metabolites, new databases and datasets that contain the concentration of certain metabolites under certain conditions of an organism, are built. Therefore also the network inferring algorithms get more and more important in the metabolic context.

1.6 From bodyfluid samples to datasets

The first step of the measurement of metabolite concentration with the various methods that can be used, is the collection of samples from the patients and their preprocessing. In order to be able to measure the metabolite concentration later on, the metabolic reactions have to be stopped in the biofluid. This can be done by shock freezing, denaturing of the enzymes and by adding acids or some solvents, like chloroform.[66] This step is very sensible, as a wrong method for stopping the reactions might also influence the metabolite concentrations.
However there are still other, unwanted parts in the solvent, which have to be separated. These unwanted substances are for example proteins, parts of cell walls, DNA, RNA or some salts. Therefore the solvent can be centrifuged in order to separate the metabolites from the

rest of the solvent. Another method for the separation would be a solid phase extraction.[75] The method that is taken is also dependent on the fluid, which is examined, as they have different properties.[19]
The next step after dissolving the metabolites from the rest of the fluids, is to measure their concentrations, which can be done in two general ways. The first one would be the targeted metabolomics and the second one would be the metabolic profiling which is also called non-targeted metabolomics.

1.6.1 Non-targeted metabolomics

Non-targeted metabolomics is the measurement of all metabolites in a solvent, which also includes chemical unknowns. Therefore advanced chemometric techniques, like multivariate analysis, have to be used for non-targeted metabolomics, in order to generate a dataset of a manageable size. Afterwards those signals, that are saved in the database have to be identified. This can be done by the usage of in-silico libraries, which match the signals to a metabolite. Another method for identifying the signals is by experimental investigation. One of the advantages of non-targeted metabolomics is the possibility to find new metabolites which can be later on used for the targeted metabolomics approach. Nevertheless the time which is needed for the detection of those new metabolites and the evaluation of the huge amounts of raw data, which is produced by such an approach, is immense. The two companies, that are using this approach are *metabolon* and *chenomx*.[49][32]

1.6.2 Targeted metabolomics

Targeted metabolomics is defined as the identification of pre defined metabolites, for which the structure and biochemical features are known. In contrast to the non-targeted metabolomics, the targeted metabolomics takes advantage of the fact that the metabolites which are examined are known and therefore also their specific kinetics, end products and pathways to which they contribute. This is why the sample preparation can be optimized for the given set of metabolites, which are examined. Another advantage of targeted metabolomics is that the relative abundancies and concentrations of the predefined metabolites can be determined more precisely than with a nontargeted approach.[74] One company, that uses this approach would be *biocrates*.[9]

1.7 Correlation networks

One way of analysing metabolite and phenotype datasets, is to calculate a correlation network. The calculation of such networks however can be done in many different ways. One method would be to calculate the *Pearson correlation* which can be seen in Equation 1.5. In this equation, $\rho_{X,Y}$ is the correlation coefficient between the two random variables X and Y, μ_X and μ_Y are their expected values, σ_X and σ_Y are their standard deviation and E stands for the expected value. The equation was developed by Karl Pearson.[64]

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y} \tag{1.5}$$

In order to get a network with metabolites or phenotypes as nodes and their correlation as edges, this equation is used for all pairs of metabolites or phenotypes in the dataset, which in

turn results in a correlation network.
Another possibility to calculate the correlation between two metabolites would be the *Spearman* correlation. The *Spearman* correlation is defined as the *Pearson correlation coefficient* between ranked variables.[33] The advantage of the *Spearman* correlation over the *Pearson* correlation, is that the *Spearman* correlation is a nonparametric measure of statistical dependence.
But also more advanced methods that calculate partial correlations, like linear, logistic regression or the Gaussian Graphical Model can be used.

1.8 Gaussian Graphical Models

Gaussian Graphical Models (GGMs) is a recently arising method for studying gene associations and lately also for metabolic interactions. GGMs are also known as "covariance selection" and "concentration graph".[73] GGMs are based on partial correlations and thus are able to determine independence of two genes or metabolites in due consideration of all the other genes or metabolites, that are considered as cofactors by the model. As a consequence, direct and indirect interactions, can be distinguished, whereas a normal correlation network would fail and would possibly predict both, a direct and an indirect interaction.

The differences between those two interaction types are depicted in Figure 1.9, where the red arrow from node A to node C stands for a direct interaction. An indirect interaction is always with any kind of intermediate, which is shown in the figure as blue dashed lines from A to B and then from B to C. An indirect interaction can also have multiple intermediates.[73]

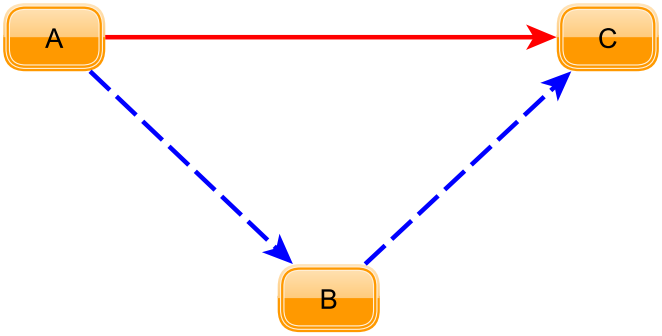

Figure 1.9: Difference between direct and indirect interactions. There are two different pathways that go from A to C. One would be the red edge, which represents a direct interaction. The other one would be the indirect interaction which is from node A to node B and from node B to node C, depicted as blue, dashed lines.

From the biological point of view, a differentiation between a direct and an indirect interaction means, that pathways can be distinguished more precisely, as the GGM filters out many false positive edges. This means that the GGMs can distinguish, whether metabolite A directly interacts with metabolite C, or maybe is degraded into C, or is somehow directly related to metabolite C, or if there is any intermediate. Still there is one disadvantage, that GGMs have compared to correlation networks. This is that correlation networks are able to handle binary data. This is a big disadvantage in the case of phenomic data, as there are many binary variables, like phenotypes, that describe if the patient has a disease or not. Therefore a more

advanced method, the mixed graphical models, which can handle binary data, should be used for phenotype data most of the time.

2 Materials and methods

2.1 Qatar Metabolomics Study of Diabetes

The Qatar Metabolomics Study of Diabetes is a dataset which contains metabolite concentrations and phenotype data. The metabolite concentrations were measured in three different tissues, blood, urine and saliva, with three different methods. Those methods were metabolon, for blood, urine and saliva, biocrates for blood and chenomx for urine. Therefore there were 2473 different measurement series for metabolites for the 375 patients. The 111 different phenotypes which are also gathered in the dataset are wide-ranging, as there are measurements whether patients have diseases or not, measurements of blood parameters, such as the hematocrit level and the percentage of neutrophils, but also different anthropometric measurements, like the body mass index and waist to hip ratio. The metabolic measurements are also wide ranging, as the three kits represent three different ways of measuring metabolites. Therefore, there are many different metabolites measured and sometimes also the same ones from different methods.

2.2 Metabolic measurement kits

The general procedure of creating metabolite datasets would be to differentiate the fluid, that you examine, in order to separate the metabolites from each other and then, to measure the masses of the different metabolites in order to calculate the concentrations of them.[5][67][22][71] Different separation methods would be *Gas chromatography*[67], *High-performance liquid chromatography (HPLC)*[22], or *capillary electrophoreses*[71]. It is also possible to directly examine the fluid, without separating it beforehand, with the *flow injection analysis (FIA)*[77][1]. The mass measurement is mainly done by mass spectrometry, but it can also be done by Nuclear Magnetic Resonance (NMR). Following are three different approaches from three different companies, which were applied to the *Qatar Metabolomics Study of Diabetes*.

2.2.1 Biocrates

Biocrates uses a targeted metabolomics approach, which means, that there is a set of predefined metabolites that are examined with the kit. In the case of the *Qatar Metabolomics Study of Diabetes* the *AbsoluteIDQ p180* kit was used. This kit is able to identify 186 different metabolites from 5 different compound classes. Those are:

1. Acylcarnitines
2. Amino acids
3. Hexoses

4. Phospho- and sphingolipids
5. Biogenic amines

The methods which are used in order to measure the concentration are first, a flow injection analysis with mass spectrometry and secondly a high performance liquid chromatography with mass spectrometry. These methods can also detect metabolites at a very small amount of a sample.[8]

2.2.2 Metabolon

Metabolon uses in contrast to biocrates a non-targeted metabolomic approach, which means that they do not only search for a pre defined set of metabolites, but test the fluid for all its metabolites.[49] Therefore also concentrations of unknown metabolites are measured with this method. In order to examine the samples a multi-step process is used. The first step is a liquid chromatography with mass spectrometry with the use of electrospray ionisation in order to produce ions. The second step is, like the first step, a liquid chromatography with mass spectrometry, but here the metabolites were not ionised by an electrospray ionisation. The third step would be a gas chromatography with a mass spectrometry. With the help of those three methods, the biochemical profiles are created. Then those profiles are compared to a reference database under the usage of retention index and mass spectrum, in order to identify them.

2.2.3 Chenomx

Chenomx uses like metabolon a non-targeted metabolomic approach, but in contrast to metabolon and biocrates, the quantifying step is done by nuclear magnetic resonance (NMR). The advantages of a NMR based detection mechanism is that there is no need to separate the fluid before the andalysis, as NMR is capable of detecting hundreds of metabolites simultaneously. Therefore the identifying step also differs to metabolon and biocrates, as it uses targeted profiling. This means that they use compound signature libraries, which are modeled to behave like the pure spectra of the individual compounds under similar experimental conditions.[32]

2.3 Correlation network for phenotypes

In order to investigate the relationships between phenotypes, correlation networks were built from the phenotypes in the dataset. To this end, the correlation between two phenotypes have to be calculated, which can be done by several different methods. As there were binary and quantitative data parts and due to the fact, that I wanted to correct for confounder as much as possible, I used three different methods for the calculation. All the methods were implemented in Matlab version R2011b and can be found on the additional cd.

2.3.1 Phenotype data preprocessing

In order to apply the correlation methods later on, some preprocessing had to be done, like grabbing out the important phenotypes from the whole dataset as well as handling the missing values.

Phenotype filtering

In general all the information that is provided in a dataset is interesting, but nonetheless some phenotypes can be redundant or can cause issues with some correlation methods and therefore have to be filtered out. First of all redundant phenotypes would be phenotypes, which are derived from another phenotype. One example would be the phenotype *"abnormal cholesterol"*, when we have the *cholesterol* value. The *"abnormal cholesterol"* just takes the cholesterol value and divides it into two groups by a certain cutoff. Thus correlations to the *cholesterol* value is expected to be more precisely and more meaningful than the derived phenotype.
A second matter for neglecting a phenotype, would be if what it describes is not informative. An example for such a phenotype would be *"other complications"* where even if there is a correlation between this phenotype and another one, no true statement can be given, as the phenotype can be anything. The third and last exclusion criterion would be due to lack of data. Examples for extreme cases in Figure 2.1, would be the birthplaces of the grandparents (phenotypes 36 to 39), where there are less than 10 entries which can not give any reasonable results.

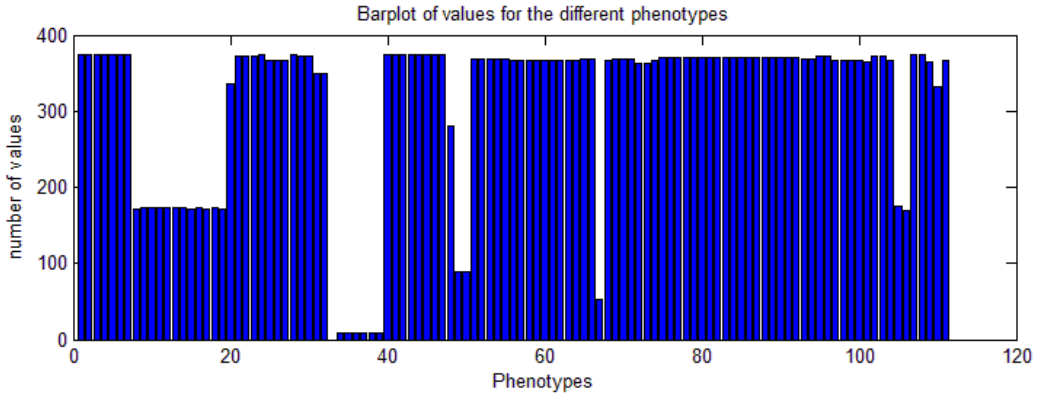

Figure 2.1: True entries in the dataset. There is a large difference in the amount of data available for some phenotypes, which can cause problems for some of the evaluation steps. As some phenotypes had only few true entries, they were excluded from further examination steps.

Handling missing values for the correlation network

As not all the phenotypes, that had missing values could be left out, as it would mean to throw away half of the dataset, I only deleted the phenotypes with many missing values, like the before mentioned birthplace of the grand parents. First of all there were two ways to handle the missing values, on the one hand impute the missing values and on the other hand use only the data which is given. In order to calculate the pairwise correlation between two phenotypes, using only the given data is much better, as there are no errors due to a correlation of the imputed values. A fact, why in some cases the deletion of missing values is not used, is that much power is lost if it is applied to the whole dataset. This is due to the fact that if there is a missing value in any phenotype, the data for either the whole patient with the missing value,or the whole phenotype has to be deleted. This means that there is much information lost. In the case of Table 2.1 a global deletion of the missing values would delete rows one, two and six, as there are missing values. However this would mean, if the correlation between

whole phenotype matrix			
Phenotype1	Phenotype2	Phenotype3	...
1	NaN	NaN	...
2	1	NaN	...
3	3	4	...
4	4	6	...
5	1	2	...
NaN	6	1	...

Table 2.1: Example matrix with missing values. In this made up example matrix, the disadvantages of a globally used deletion of missing values can be seen, as also the second row would be deleted, even though it contains valid data for the correlation calculation between *Phenotype1* and *Phenotype2*.

phenotype and confounder matrix			
Phenotype1	Phenotype2	Confounder1	Confounder2
1	NaN	6	1
2	1	1	3
3	3	3	2
4	4	4	4
5	1	2	5
NaN	6	5	6

Table 2.2: Pairwise correlation matrix This is the matrix, which is assembled from the two phenotypes, for which the correlation shall be calculated and the confounder matrix, which is similar for nearly all the phenotypes. This matrix is assembled and the missing values, here depicted as NaN, are cut out, before the calculation of each pairwise correlation between two phenotypes.

the phenotypes *Phenotype1* and *Phenotype2* is calculated, not only the rows, for which one of the two phenotypes have a missing value, but also the second row, where none of those two phenotypes have a missing value has been deleted. In order to not have this unnecessary data loss, I made a second matrix, which contained the data for the two phenotypes, for which I wanted to calculate the pairwise correlation and the confounders for which I wanted to correct and deleted the missing values there. The resulting matrix for this can be seen in Table 2.2. Through this approach only the data is lost, where one of the phenotypes or confounders have a missing value and not valid data points.

2.3.2 Binary to binary

For the correlation between two binary variables, I used the χ^2-test, which is an approximation for the Fisher's exact test. The Fisher's exact test was not applicable, due to the fact that it can only be used with small numbers.[45] For this part a correction for confounders was not possible, as the χ^2 test does not incorporate confounders. The χ^2-test is calculated by the Formula 2.1 and in general gives a good estimate on how far an observed frequency is apart from the expected frequency.[44] For this calculation I used the Matlab function *crosstab* and built up a matrix with the resulting p-values, that filled the first spots of the whole phenotype

to phenotype matrix.

$$\chi^2 = \sum \frac{(ObservedFrequency - ExpectedFrequency)^2}{ExpectedFrequency} \tag{2.1}$$

2.3.3 Binary to continuous

For this part I used logistic regression, which is able to correct for confounders if needed. This equation is realized in Matlab by the function *glmfit*, which is a method for logistic regression. For this method one has to define the parameters of the Equation 2.2, first. In this case X_1 to X_n are defined as one of the two phenotypes and $n-1$ co-factors. The other phenotype defines the Y. β_1 to β_n is a measure on how strong Y depends on the appropriate X and β_0 is the basis value, without the influences of the variables.

$$Logit(Y_{1/0}|X_i = x_i) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{2.2}$$

2.3.4 Continuous to continuous

The continuous to continuous part could have been calculated by a Gaussian graphical model, as there are only continuous phenotypes. However in order to make it at least some kind of comparable to the rest of the network, I used a linear regression approach. Besides, the GGM can not be used to its full potential, as it only takes into account half of the dataset, the continuous part and therefore the correction which is included in a GGM can only correct for half the dataset, which would still be more than by the linear regression. Therefore calculating this part with a linear regression might not be the best solution, but it makes it more comparable to the rest of the network. In order to include it in the whole phenotype correlation matrix, I calculated this with the Matlab function *corr* and corrected it beforehand with the function *CorrectLM*. Thus the calculation is done by the equation 2.3. Here ϵ is the error of the equation, and β_1 to β_p describe how much the according *X* influences the examined phenotype *Y*.

$$y_i = \beta_1 \cdot x_{i1} + ... + \beta_p \cdot x_{ip} + \epsilon_i \tag{2.3}$$

2.3.5 Assemble the three parts

In order to assemble the results from the different methods, I made one common matrix, where I put all the results in the according spots. Thus the matrix, which was assembled contains three different parts, which can be seen in Figure 2.2. Still the matrix was not complete after the assembling, as the results where biased by multiple testing. This was fixed by using an FDR approach for multiple testing correction, which was applied to the different parts.

2.4 Linear regression for the linkage of phenotypes to metabolites

The linkage between phenotypes and metabolites was calculated by a linear regression. This was done using an R-script which was coded by Professor Karsten Suhre, which calculates a linear correlation and corrects for the inputted co-factors. As the phenotypes are also included

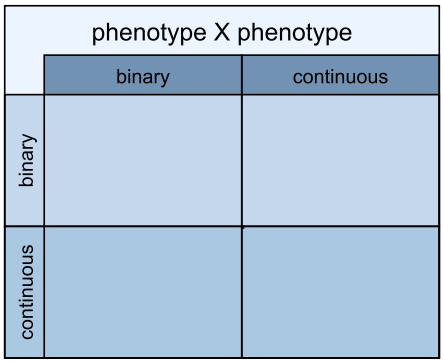

Figure 2.2: Correlation matrix for phenotypes. The matrix was assembled from three different parts, the binary to binary part, the binary to continuous part and the continuous to continuous part.

in this examination, the same assumptions as for the phenotype correlation calculations are valid here as well. This means, only the correlations from the previously selected phenotypes to the metabolites have been calculated, as well as the same co-factor correction was used. At the metabolite level there were only those metabolites excluded which had only few valid values. The calculated values for the correlation between phenotypes and metabolites have been added to the already calculated phenotype to phenotype matrix. Therefore three quarters of the correlation matrix are filled out (see Figure 2.3).

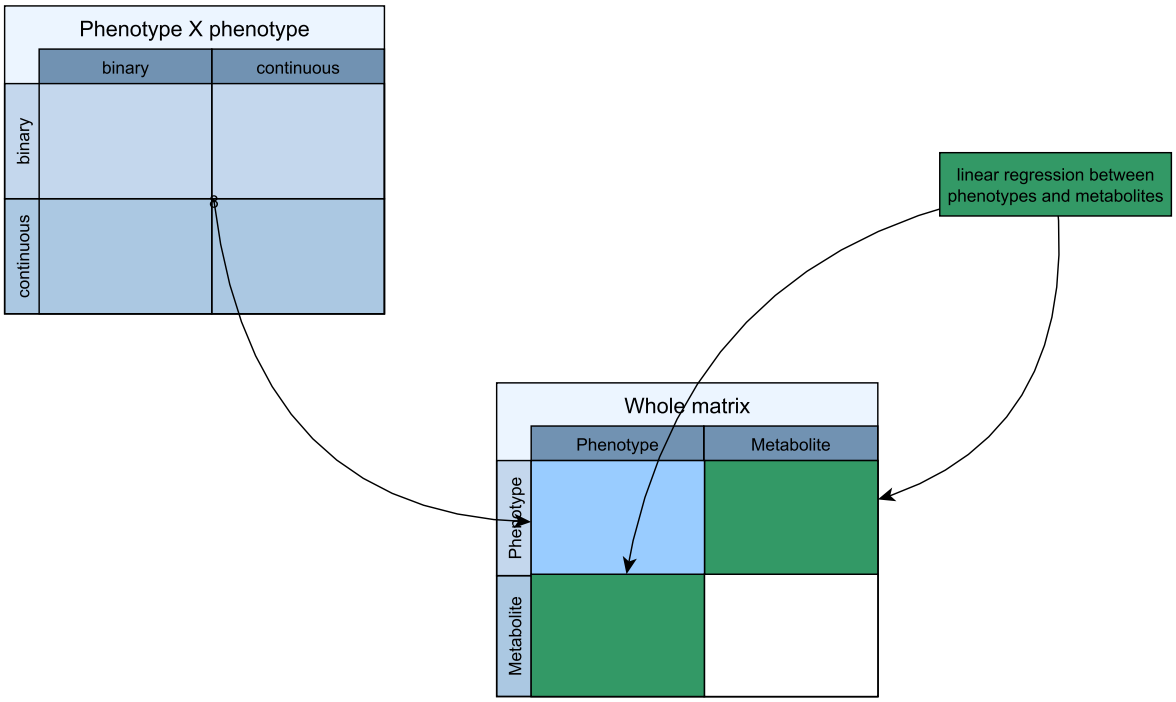

Figure 2.3: Nearly finished correlation matrix for phenotypes and metabolites. With the correlations between the phenotypes and metabolites, which were calculated by linear regression, three quarters of the whole matrix are filled up, only the metabolite to metabolite part is still missing.

2.5 Finish the correlation matrix and build networks

The last quarter of the correlation matrix was filled up by a Gaussian Graphical Model, which was calculated by Kieuh Trinh Do during her masters thesis.[15] Therefore the whole correlation matrix was finished and can be seen in Figure 2.4.

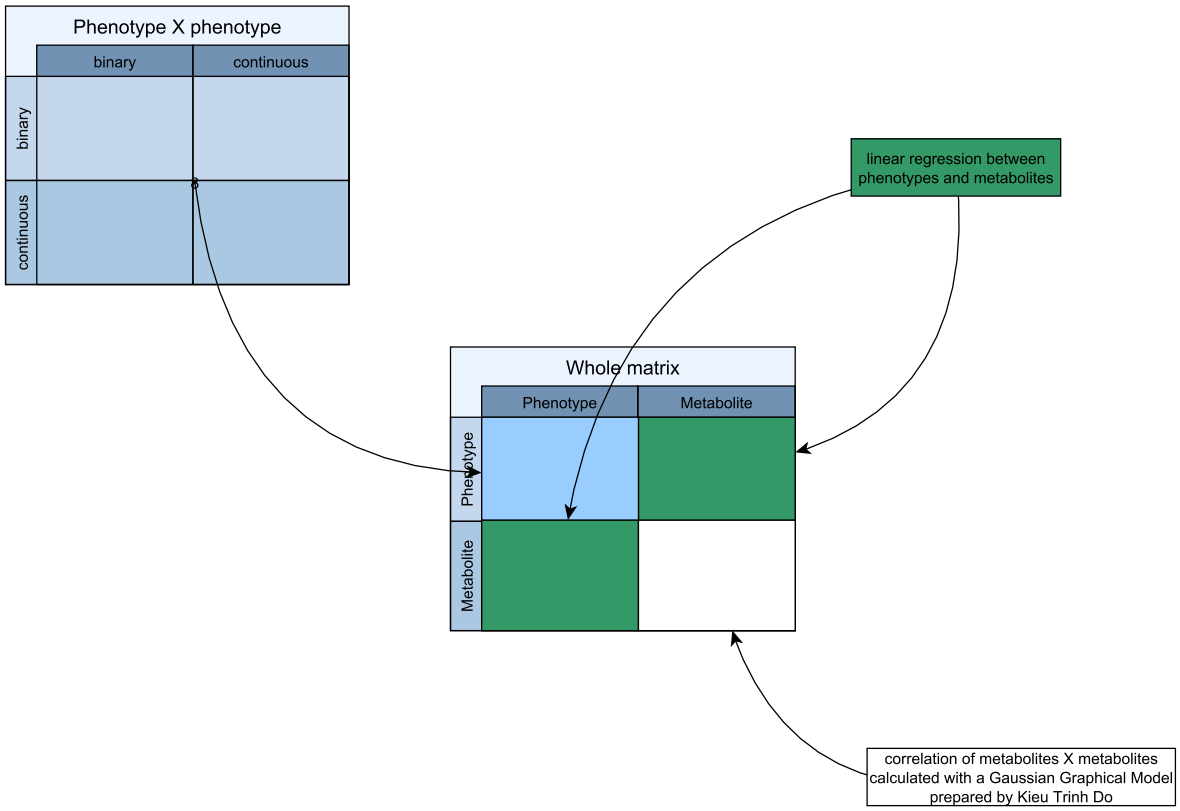

Figure 2.4: Finished correlation matrix for phenotypes and metabolites. With the results of Gaussian Graphical Model, which Kieu Trinh Do calculated for the different metabolites, the whole phenotype to metabolite matrix was filled up.[15]

In order to make networks from the whole calculated matrix, I compiled a binary matrix from the correlation matrix. The cells of the binary matrix were filled with a 1 if there was a significant correlation in the according cell of the correlation matrix and a 0 when there was no significance. A significance cut-off decided whether a value was significant or not. As the matrix is assembled by three parts, it is also possible to take three different cut-offs, one for the phenotype to phenotype part, one for the phenotype to metabolite part and another one for the metabolite to metabolite part. Those cut-offs were adjusted to the question that had to be answered and whether a very sparse or a very densely connected network is better for answering the question.
The network then afterwards was drawn according to the binary network, where a 1 meant that an edge was drawn and a 0 means no edge had to be drawn. This step was done by a script called writeYED, which is a tool, that was coded by the group of *Fabian Theis* at the *Helmholtz institute munich* and writes an *.graphml*-file which can be visualized by the graph editor yEd.

2.6 Size and Node reduction in the networks

In order to examine different parts of the network, as well as for the examination of specific phenotypes, I shrinked the network according to the phenotypes of interest. This was done by simply taking the whole correlation matrix, with all the relationships between phenotypes and metabolites and a 1 for a relationship between two nodes and a 0 for no relationship and look for directly related phenotypes and metabolites. A dummy network is shown in Figure 2.5.

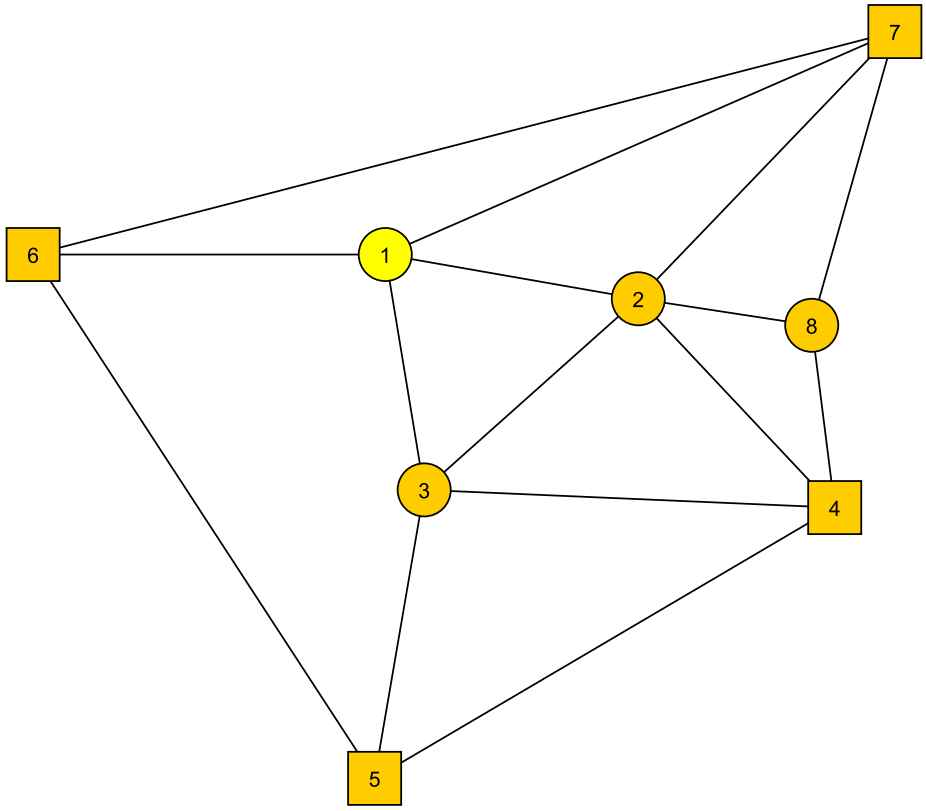

Figure 2.5: Dummy network Here the phenotypes are shown as circles and metabolites as squares. The light yellow phenotype is the phenotype of interest.

The algorithm can be divided into three different steps. The first step would be to find all the phenotypes that are directly related to the phenotype of choice. Figure 2.6 shows the results of the first step, here node 1 is the phenotype of interest.

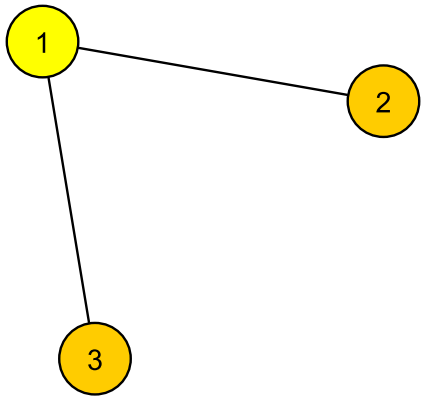

Figure 2.6: The first step of the network extraction algorithm. In this step, the phenotypes (2 and 3) that are directly related to the phenotype of interest (1) are added to the graph.

In the second step all the metabolites, which were related to the phenotype, or phenotypes of interest where added to the network as well as the connections between the metabolites. The resulting network is shown in Figure 2.7.

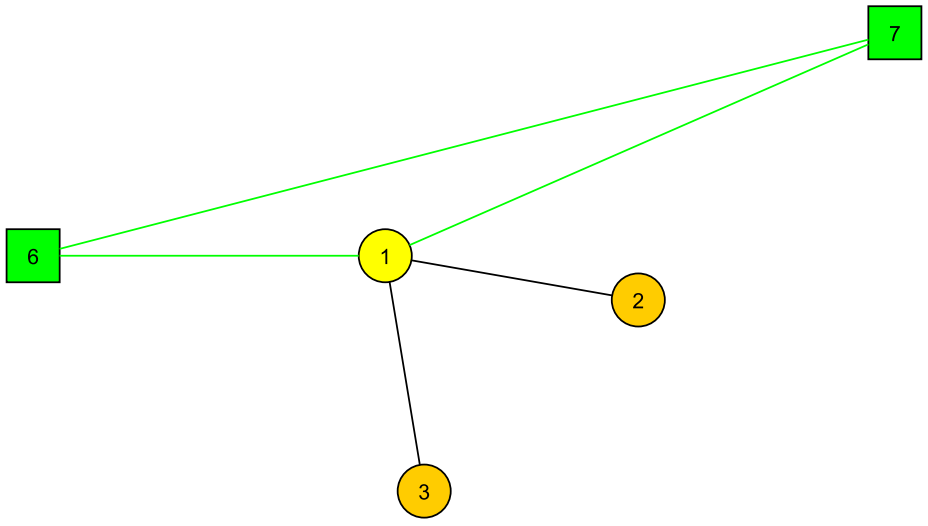

Figure 2.7: The second step of the network extraction algorithm. In this step of the algorithm, the directly linked metabolites are added to the graph as well as the connections between the metabolites. The added parts are shown in green.

In the third and last step, I looked for the connections between the phenotypes which were added in the first step and also between the phenotypes and the metabolites. This means, there are no new nodes added in this step, but the existing nodes are just connected to each other. The resulting network can be seen in Figure 2.8.

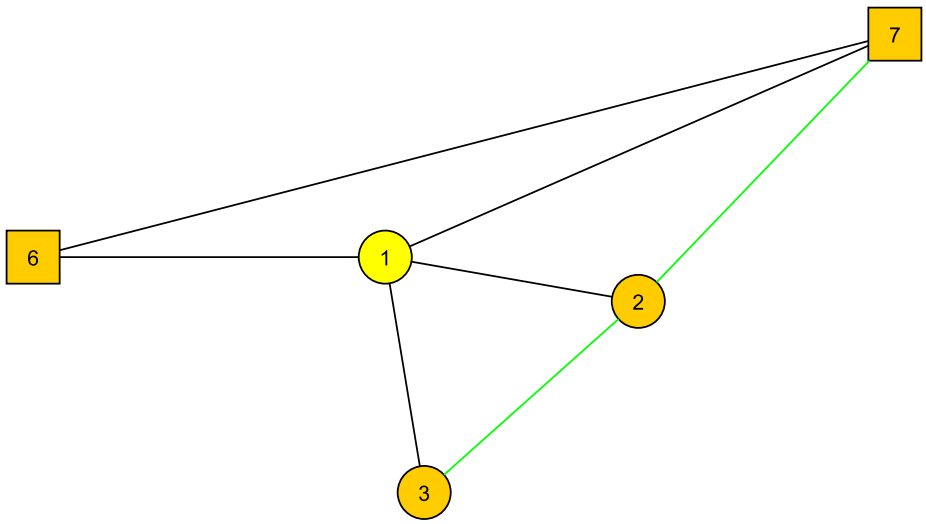

Figure 2.8: The third step of the network extraction algorithm. Here, the first order phenotypes and metabolites are connected to each other. The added edges are depicted in green.

2.7 yED

In order to construct, visualize, and explore networks an Editor is needed, as looking onto the correlation matrix alone is very incomprehensible.[86] For this purpose there are some different editors, the one I chose is yEd. The yEd editor is a freely downloadable program, which can be found at the yWorks homepage and its custom layout is shown in figure 2.9. It is a Java based graph editor, developed and constantly updated by yWorks. The output file is a so called *.graphml*-file but the graphs themself can also be exported as a graphic file, like *.jpg* or as *.pdf*. This makes it a useful tool for systemsbiology, as the networks, which are inferred through the different algorithms, can be written as *.graphml*-file and thus visualized and post processed in the editor and then exported afterwards for example as pdf.

2.8 Graphmodel and graph explanation

In order to make the networks easier understandable and also to point out some features of the network some different visual properties between nodes and edges are necessary. In order to make something visible in a network there are different possibilities:

1. colour of the node
2. shape of the node
3. size of the node

In my opinion the colour code is the best for the viewer as too many shapes are hard to see for the eye, if they are small enough, which is the case if one generally looks at least at several hundred nodes. Thus the only thing that can be still seen at large scale is the colour. Therefore I made two colour codes, one for the edges and one for the nodes.

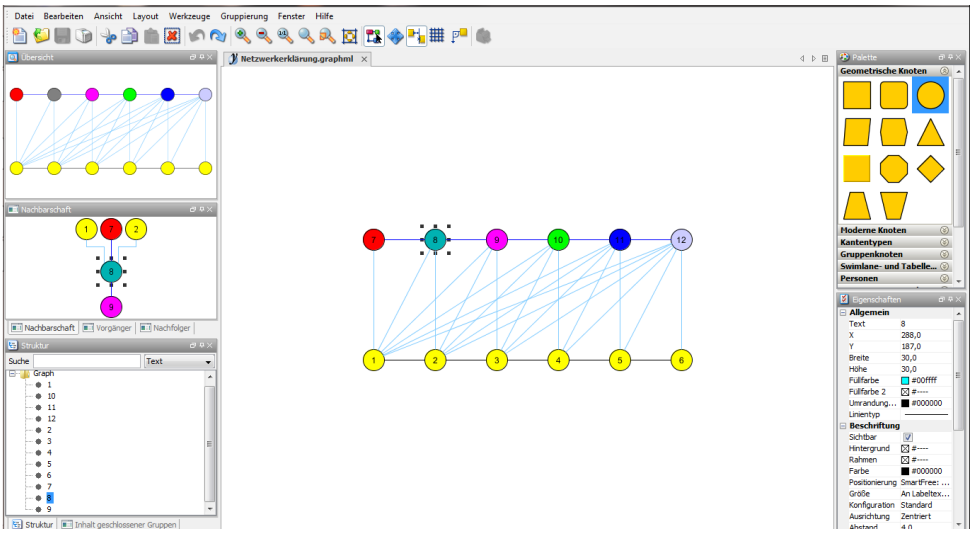

Figure 2.9: Layout of the yEd graph editor. In the center, there is the editing panel where your network is located. In the upper left, there is an overview panel, which is useful if you zoom in on larger networks, as you can see, which part of the network you currently look at. At the left side in the middle, the direct neighbourhood of your currently selected node can be seen. Whereas on the bottom left side, there is a list of all the nodes sorted by the name. Here you can also search for certain nodes by name. At the top right there are the different nodes which can be drawn. On the bottom right you can see the properties, like colour, name or location, of your currently selected node or edge, or whatever you selected in the graph panel.

2.8.1 Node colour code

First of all I emphasized the phenotypes with a yellow colour in order to be able to distinguish easily between metabolites and phenotypes. As phenotypes can be divided so differently in various clusters, due to many overlaps like some similar characteristics, same risk factors or similar genetic reasons, I decided not to cluster the phenotypes any further. Most of the time the questions that have to be answered with this networks, are in the following fashion:
"The following two or more phenotypes are quite related to each other, which metabolites are different and which are related to both?"
Therefore I made the color code of the metabolites according to the amount of phenotypes they are linked to. Thus I can easily distinguish between metabolites, that are only related to one phenotype and which one not. This results in the following colour code for metabolites:

- Metabolite related to one single phenotype: red. (see Figure 2.10)
- Metabolite related to two phenotypes: turquoise. (see Figure 2.12)
- Metabolite related to three phenotypes: purple (see Figure 2.12)
- Metabolite related to four phenotypes: light green(see Figure 2.12)
- Metabolite related to five phenotypes: dark blue (see Figure 2.12)
- Metabolite related to six or more phenotypes: grayish (see Figure 2.8.1)

Figure 2.10: Metabolite connected to one single phenotype. Metabolites with only one connection to a phenotype(metabolite to metabolite connections are not taken into account) are coloured red.

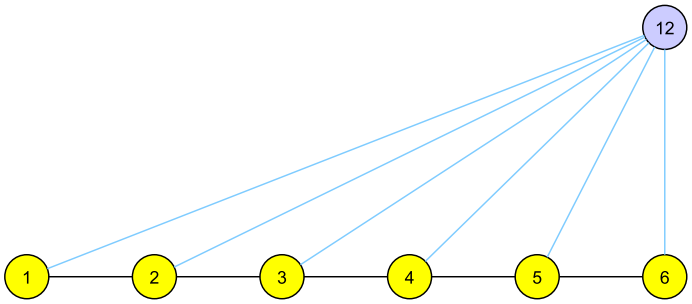

Figure 2.11: Metabolite connected to six different phenotypes. Metabolites with six or more connections to phenotypes (metabolite to metabolite connections are not taken into account) are coloured grayish.

2.8.2 Edge colour

But not only the nodes can be differentiated from each other, also the edges can have a different colour and a different breadth. In general the breadth of an edge is according to the correlation strength, but as in my case the very heterogeneous methods for calculating makes it difficult to compare the values, I decided not to use the correlation values as scale for the line breadth. However the colour of the edges was adjusted to which kind of relationship it is. There are three different ones:

1. Phenotype to phenotype relationship, which is coloured in black (see Figure 2.12)
2. Phenotype to metabolite relationship, which is coloured in turquoise (see Figure 2.12)
3. Metabolite to metabolite relationship, which is coloured in dark blue. (see Figure 2.12)

All the different characteristics of the edges and nodes can be seen in the summary dummy network in Figure 2.13

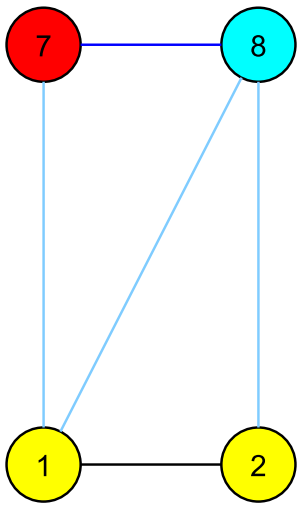

Figure 2.12: Colour code of the edges in the networks. There are three different ones, which are, an edge between two phenotypes, which is coloured in black, an edge between a phenotype and a metabolite, which is coloured in turquoise and an edge between two metabolites, which is coloured in dark blue.

2.9 Mixed graphical models with random forests

Mixed graphical models can, in contrary to regular Gaussian Graphical Models, not only handle continuous data, but also mixed data with continuous, categorical and binary data. Thus, this type of model should be very useful for phenomics data, as they contain multiple different types of variables. There are some different approaches to the mixed graphical models which calculate the partial correlation in many different ways. I used the approach from Fellinghauer et al, which used a random forest approach and for the cutoff stability selection.[7]

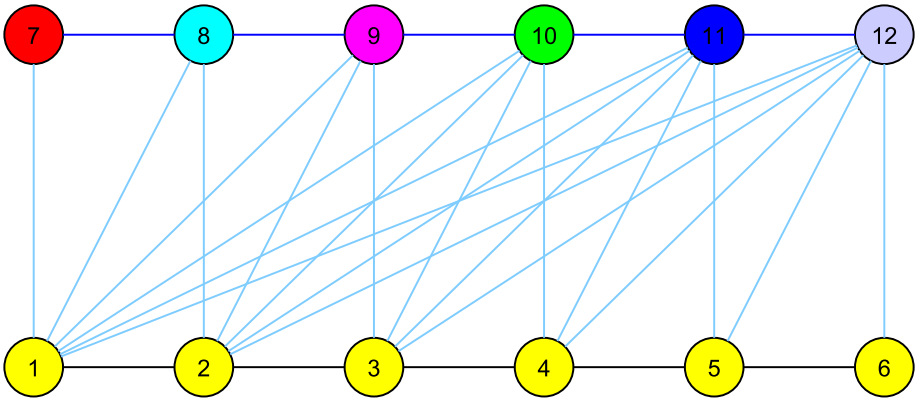

Figure 2.13: Overview network which shows the different colour codes for edges and nodes. This network sums up all the characteristics of the network, which means the two different color codes, of the edges and the nodes.

2.9.1 Stability selection

Stability selection is a variable selection algorithm which uses a group lasso in an subsampling approach.[47] [2] The first step is the subsampling step, where the algorithm performs B lasso regressions on subsamples of size $\lfloor \frac{n}{2} \rfloor$. Based on this estimation it calculates the probability of including a variable $k \in \{1, \cdots, p\}$ with Equation 2.4, where B is how often the lasso step was repeated for the $\lfloor \frac{n}{2} \rfloor$ samples and $\hat{\pi}_k^\eta$ is the regularization parameter.

$$\hat{\pi}_k^\eta = \frac{1}{B} \sum_{b=1}^{B} 1[\hat{\beta}_{b,k}^\eta \neq 0] \tag{2.4}$$

In this equation η is a tuning parameter that controls the amount of regularization and $\mathbf{1}$ denotes the indicator function. Because it is a lasso regression normally one chooses $\eta = \lambda$. Anyway, $\hat{\pi}_k^\eta$ can be alternatively calculated in due consideration of the complexity of the model. If you then select the regularization parameters $\eta \in E$, the variables selected like this, are shown in equation 2.5

$$\{k : \underset{\eta \in E}{max} \hat{\pi}_k^\eta \geq \pi_{thr}\} \tag{2.5}$$

In the algorithm one can also get an upper bound for the expected number of false discoveries E(V). This is calculated by equation 2.6. Here q_E is the average number of selected variables over the subsamples when η is varied in E. Thus one can also control the upper bound of the family-wise-error rate (FWER).

$$E(V) \leq \frac{1}{2 \cdot \pi_{thr} - 1} \cdot \frac{q_E^2}{p} \tag{2.6}$$

3 Results and Discussion

In order to examine the Qatar Metabolomics Study of Diabetes different methods have been tested. Especially the phenotype aspect of the cohort has priority here.

3.1 Phenomics in a systems biological approach

As the field of phenomics in the systems biological approach is growing fast lately, some questions arose.

1. What is the best way of calculating a correlation between two phenotypes?
2. Should you correct for a confounder and which one would be interesting?
3. What to do with missing values?
4. How to decide, which relationship is significant and which one is insignificant?

3.1.1 Phenotypes in the Qatar Metabolomics Study of Diabetes

In the dataset, the Qatar Metabolomics Study of Diabetes, there are 111 different phenotypes. 28 of those phenotypes were neglected due to lack of data, because they were derived from other phenotypes or other reasons. These reason for each phenotype can be seen in Table 3.1. Still 83 of the 111 initial phenotypes have been used for the calculation of the networks.

3.1.2 Correction or not?

Comorbidity between two phenotypes can be a big problem for a systems biological approach on such data, as the comorbidity can invoke wrong correlations and can suppress right ones. Thus correction for such confounders can make a huge difference in inferring phenotype networks. Therefore some deliberations have to be made, like when does correction make no sense and when would it be necessary to correct.

When to correct for co-factors

The correction does make sense, if it is interesting to know the correlation between two phenotypes without any comorbid conditions. One case, where this could be interesting is when three phenotypes, such as Parkinson's disease, Gaucher's disease and depression, are examined. Both, depression and Gaucher's disease are highly correlated to Parkinson's disease and are therefore also somehow correlated to each other.[48][10] The question now would be if Gaucher's disease and depression are still correlated when the effects of Parkinson's disease are corrected out. This would mean, are those two correlated independently from Parkinson's disease? In order to answer such questions correcting for the co-factors, in this case Parkinson's disease, is absolutely necessary.

neglected phenotype	reason
1 Valid metabolomics and phenotype data	no real phenotype
2 WCMC-Q subject ID	no real phenotype
5 Diabetes state according to patient	more precise value: 6 Diabetes state adjusted for HbA1c
7 Diabetes catagory	more precise value: 6 Diabetes state adjusted for HbA1c
18 Other dermatological complication	no real statement can be given with this phenotype, as it is too abstract
19 Other complications	no real statement can be given with this phenotype, as it is too abstract
33 Birth place patient	ethnicities are the more reasonable phenotypes
34 Birth place mother	not sufficient data
35 Birth place father	not sufficient data
36 Birth place maternal grandmother	not sufficient data
37 Birth place maternal grandfather	not sufficient data
38 Birth place paternal grandmother	not sufficient data
39 Birth place paternal grandfather	not sufficient data
40 Ethnic group	different binary ethnicities are easier to handle, as an ethnic group of 2 does not mean you are twice as "ethnic" as group 1.
45 Indian or Philippines ethnicity	Indian and Philippine ethnicity already in the dataset
48 cigarettes per day	not sufficient data
53 Abnormal cholesterol	derived from: 52 Cholesterol (mmol/L)
55 Abnormal TG	derived from: 54 Triglycerides (mmol/L)
57 Abnormal HDL	derived from: 56 HDL-C (mmol/L)
59 Abnormal LDL	derived from: 58 LDL-C (mmol/L)
61 Abnormal chol/hdl ratio	derived from: 60 Chol / HDL ratio
64 Calcium (mmol/L)	corrected value taken: 63 Corrected calcium (mmol/L)
67 eGFR	not sufficient data
85 Neutrophils (10*3 /uL)	90 Neutrophils (%) was taken
86 Lymphocytes (10*3 /uL)	91 Lymphocytes (%) was taken
87 Monocytes (10*3 /uL)	92 Monocytes (%) was taken
88 Eosinophils (10*3 /uL)	93 Eosinophils (%) was taken
89 Basophils (10*3 /uL)	94 Basophils (%) was taken

Table 3.1: Neglected phenotypes and the reason for it. Those phenotypes were neglected, as they would not add any information to the network and would only be time consuming, when the network is calculated.

When can correction cause problems?

However in the case of phenotype data, correction is not always right, as a correlation between two phenotypes can be valid, even though it is just caused by comorbid conditions. If the same correlation is examined, it is still true, that one might have a higher risk of getting Gaucher's disease and depressions at the same time, even though it is caused by the common comorbid condition, the Parkinson's disease. Another point that can be critical with correction is that there is always a chance of overcorrection and some true edges are deleted therefore.

When do I correct

Therefore one has to always remind, which question has to be answered with the network later on. As sometimes it is better to correct for certain co-factors and sometimes it is better not to correct for them. In my work I most of the time corrected for the known confounders age, gender, BMI and ethnicity. Although I did not correct for all of those variables all the time, as it makes no sense to correct for BMI if the weight related parameters are examined, as those variables are too closely related. A summary for the correction I used can be seen in Table 3.2. The correction for diabetes has been adjusted to the question that I wanted to answer, so whether I wanted to see correlations to diabetes or not.

Phenotype	corrected for
Weight related (like WHR and BMI)	Age, Gender, Ethnicity (and Diabetes)
Age	Gender, BMI, Ethnicity (and Diabetes)
Gender	Age, BMI, Ethnicity (and Diabetes)
Ethnical groups	Age, Gender, BMI (and Diabetes)
Diabetes	Age, Gender, BMI, Ethnicity
All other Phenotypes	Age, Gender, BMI, Ethnicity (and Diabetes)

Table 3.2: Certain phenotypes and their correction The diabetes phenotype was not always used for correction as it was also sometimes interesting how some phenotypes are related to diabetes. So this correction was added or left out, depending on the question that had to be answered.

3.1.3 Missing values

The next question which has to be answered, before you can calculate a proper phenotype network is, how to handle missing values. There are plenty of different methods, that can be used. These approximation methods would be taking the mean, the minima or maxima or a certain value dependent on another phenotype. The easiest thing to do would be to pick one of those methods and apply it to the whole dataset. Therefore all phenotypes would be treated in the same way, but with the heterogeneity of the phenomic data, every method had big disadvantages and thus could not be applied to the whole dataset.

The mean for example is not applicable to binary variables as this would lead to a third value in nearly all cases and the variable would no longer be binary and thus only the minima or maxima would be the two globally usable methods left over. The minima and maxima are also not right in certain cases, like the minima would be wrong, if you just have no measurement for certain groups of patients, which then would cause a correlation due to the missing value

correction. This is the same for the maxima, which is also wrong for another reason, because if one assumes that the missing values are just not measurable because their value is below the detection value of a test, taking the maxima would be a false assumption.
Still another globally applicable method would be to only use the patients which have data for all phenotypes, the drawback here would be that this lowers your power very drastically. Thus the most time consuming, but best method would be to differentiate between the phenotypes and to impute using foreknowledge. For example in the dataset I had a binary phenotype called smoker, which was described whether a patient smokes (1) or not (0) and another phenotype, whether cotinine was detected (1) or not (0). There were missing values in the smoker phenotype which I imputed according to the cotinine detected phenotype, as cotinine is a strong indicator for smoking.[16] This was not as easy for some other cases, but at least some missing values were imputed this way.

3.1.4 Significance cutoff and multiple testing correction

The last thing which is important, to think about, before you can calculate a network is how to determine, whether a correlation between two nodes is significant or not. There are many different possibilities, that are generally approved, especially for the multiple testing correction. Multiple testing correction is necessary in this case, because I performed multiple statistical tests which may cause one attribute to be significant just by chance.[45] This means for *n* repeats, the significance level $\overline{\alpha}$ also called family wise error rate (FWER) is given by equation 3.1.[54][57]

$$\overline{\alpha} = 1 - (1 - \alpha_{\{perComparison\}})^n \tag{3.1}$$

Thus $\overline{\alpha}$ increases as the number of comparisons increases. Therefore I used an FDR correction after Benjamini Hochberg which is based on a ranking according to the p-value (ranking from smallest to largest) and the False Discovery Rate, which gives an estimate of the predicted proportion of false positives among all taken results. In this procedure one searches for the highest rank, where the inequation 3.2 is still true. Here the PValue is the p-value of the examined rank, i is the current rank, m is the number of tests and QValue is the significance cutoff level.[45]

$$PValue <= \frac{i}{m} QValue \tag{3.2}$$

3.2 Heart disease related networks

The following examinations are based on the heart disease phenotype. This phenotype was taken from a questionnaire, which was handed to the diabetes patients in the Qatar Metabolomics Study of Diabetes. Thus the correlation to heart disease are for diabetes patients, which also have problems with their heart. As these results are taken from a questionnaire, there is no specification, which heart problems the patients have, thus those heart problems might be anything from strong heart attacks and strokes, over coronary heart disease and diabetic cardiomyopathy to cardialgia.

3.2.1 Phenotype network

The phenotype network (Figure 3.2) shows all the phenotypes, which significantly correlate to heart disease. On the one hand a group of weight related phenotypes like BMI, waist-to-hip-ratio and weight and on the other hand, diabetic neuropathy. The data, on which the network is based, is the whole dataset. Although all the correlations between heart disease and the other phenotypes and metabolites were only calculated for diabetes patients, as heart disease was only measured in diabetes patients.

Diabetic neuropathy and heart disease

Diabetic neuropathy is a disease which is caused by high levels of glucose in the blood. These high levels of glucose damage the blood vessels, which are attached to the nervous system, which in turn slows down the rate of nutrition and oxygen supply (see Figure 1.4).[27] Due to this, the nerves won't be able to work properly and thus take severe damage or even mortify. The symptoms are pain, tingling or numbness.[53] As diabetic neuropathy is a nerve problem, it can occur in every organ system, like sex organs, the digestive tract and also in the heart. This means, that those heart issues could sometimes be not really caused by the heart itself, but the patients just could not tell, whether they have heart problems, or if it was just a side effect of their diabetic neuropathy. This would of course result in the seen correlation between heart disease and diabetic neuropathy.
There might also be other explanations for this correlation, as the blood vessels, which are related to both diseases can cause many issues throughout the body. Thus the same clots which cause the diabetic neuropathy can also cause heart issues, when those clots are in the vessels around the heart.

Weight related phenotypes and heart disease

A well known risk factor for heart diseases is the weight. High BMI and even more, high waist to hip ratio are associated with an increased risk of heart disease which can be also seen in the Figure 3.1.[41][17] As there are many different weight related phenotypes linked to heart disease in the correlation network, this study also supports the overall accepted fact, that a high weight is a risk factor for heart diseases.

3.2.2 Phenotypes and metabolites related to heart disease

In order to be able to explain the reasons behind an edge between two phenotypes, which is important for further medical and pharmaceutical examinations, a metabolite component was added to the phenotype network. A full network, which contains all this information is shown in figure 3.3. First of all, the phenotypes which are directly linked to the heart disease phenotype are less interconnected with the metabolites, which are related to heart disease. This might indicate, that the majority of metabolic pathways which are correlated to the heart disease phenotype, are not related to the comorbid phenotypes. But there are still three metabolites, which might be important for the relationship between the phenotypes and heart disease.

Metabolites connected to multiple phenotypes

There are three metabolites that are connected to heart disease and also to another phenotype, which is correlated to heart disease. This would be two sphingomylins (SM.C16.1 and

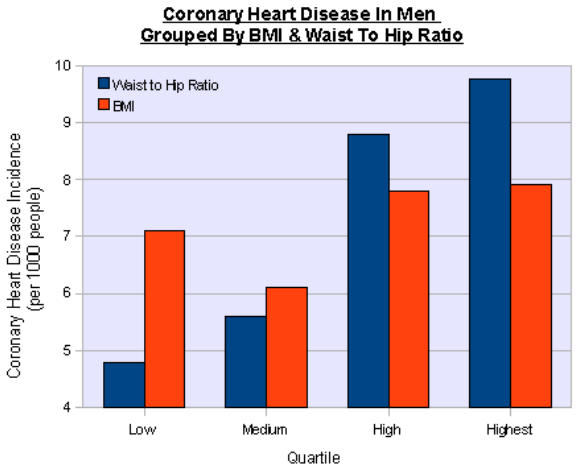

Figure 3.1: Relationship between BMI, WHR and heart disease. This figure clearly shows that WHR is the much better indicator for heart diseases compared to BMI. This can be seen as the heart disease risk is going up steadily, the higher the WHR and the BMI goes up and down.[17] This figure was taken from "`http://healthhubs.net/images/waisthipratio.gif`"

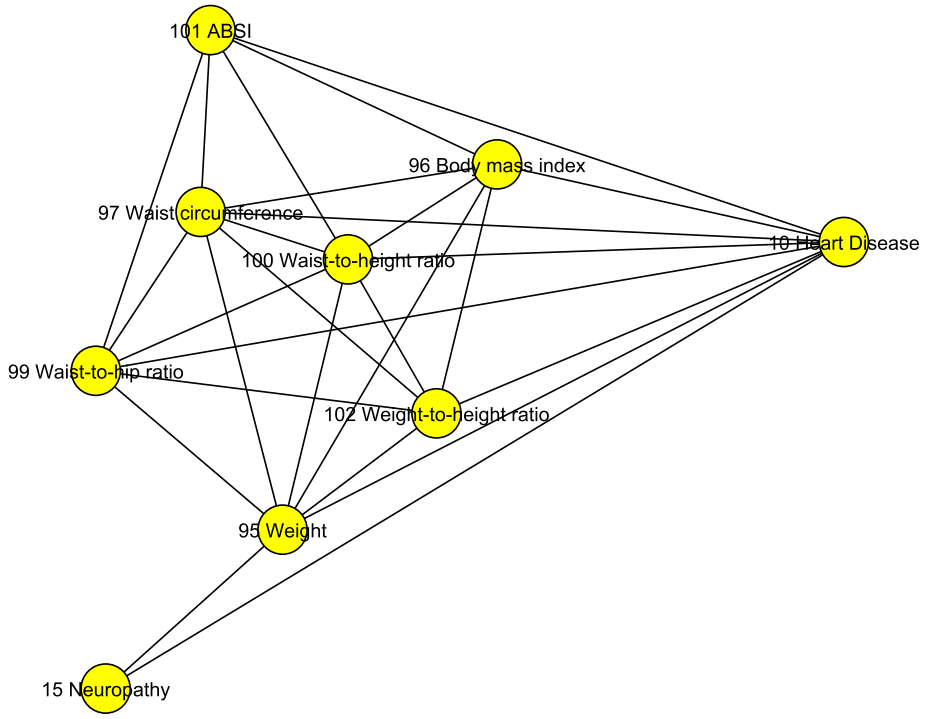

Figure 3.2: Phenotypes related to heart disease. The related phenotypes are interconnected among each other. The significance level in this graph is 0.05 after a multiple testing correction.

SM.C18.0) and an unknown (X-12740).

Spingomyelins: The two sphingomyelins are connected to the heart disease and the weight

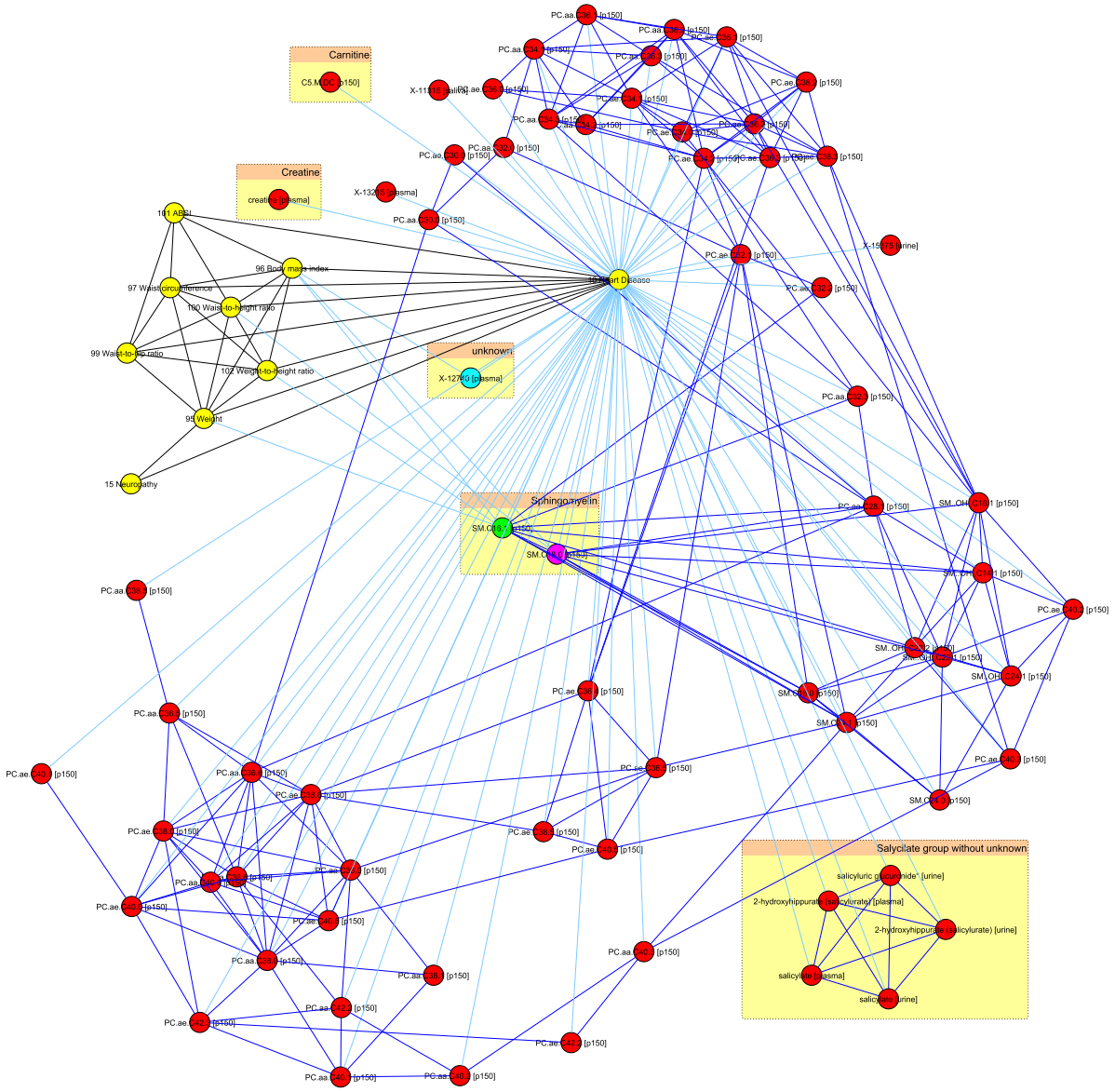

Figure 3.3: Heart disease network with all the connections to metabolites and phenotypes. There are 2 different significance cutoffs, which is 0.05 for phenotype to phenotype relationships as well as phenotype to metabolite relationships. For the metabolite to metabolite cutoff I chose 0.0001 as the network would have been to interconnected and confusing with a less stringent cutoff, as the phosphatidylcholins are very dense connected.

related phenotypes, like BMI, weight to height, waist to height and weight itself. Sphingomyelins in general are sphingolipids, that can be found in animal cell membranes. They are especially important in the myelin sheath, which coats axons and neurons. The myelin sheath is vital for the right function of the nervous system.[61] Sphingomyelin can be structurally divided in three different parts which are shown in 3.4, in the different colours:

1. Sphingosine group (black)

2. Phosphocholin group (red)

3. Fatty acid group (blue)

The correlation between heart diseases and sphingomyelins was already found by Chen et al who examined the correlation between Sphingomyelin levels in plasma and Coronary Heart Disease (CHD).[11] Interestingly in our data only 2 (SM.C16.1 and SM.C18.0) of the 15 sphingomyelins, which were examined by biocrates, were correlated to the heart disease and are also connected to BMI. The correlation between the sphingomyelins and BMI was not observed in the study of Chen et al., as they had a balanced BMI for the sphingomyelins. As they only looked on Sphingomyelin in general, this correlation might been overseen and might be interesting for further examinations, as those two could link the phenotype heart disease to the anthropometric measures.

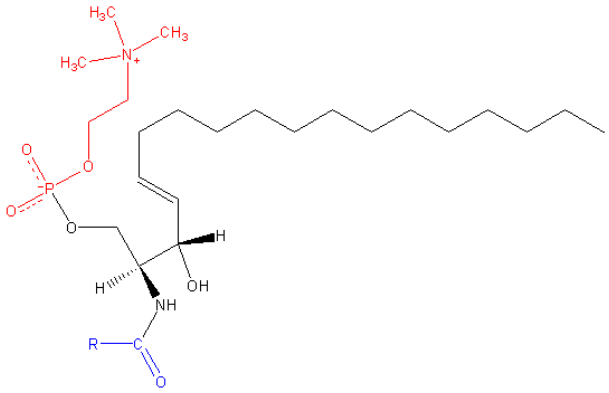

Figure 3.4: Chemical structure of Sphingomyelin This chemical structure shows the different parts of the Sphingomyelin, which can be divided in a phosphocholine group (red), a sphingosine group (black) and a fatty acid group (blue).[82] This picture was taken from `http://en.wikipedia.org/wiki/Sphingomyelincite_note-Voet-1`

Unknown X-12740 This unknown is correlated to *heart disease* and *BMI*. As it is an unknown one can only give guesses what it might be and what it might have to do with the two diseases. The first thing to mention is, that it is also correlated to *salicylate* in urine, which would be a sign for being related to Aspirin. This correlation can not be seen in Figure 3.3, but in Figure 3.6, as this correlation is no more significant at a significance cutoff of 0.0001 but at a significance cutoff of 0.05. As the salicylate cycle is not connected to BMI, but the unknown X-12740 is highly correlated to BMI, it is more probable, that it is some artifact from co-medication.

Salicylate part

Salicylate and its derivatives salicylurate and salicyluric glucuronide, were all found to correlate with heart disease. Another derivative of Salicylate is Acetylsalicylate (shown in Figure 3.5) which is better known as Aspirin. Salicylate in general is known as anti-inflammatory drug and is used in many different derivatives in different drugs.[42] Therefore in general also Aspirin is given as an anti-inflammatory and pain-relieving drug, in lower concentrations however, Aspirin is given as a blood thinner to CHD-patients.[59] This would explain the seen correlation in figure 3.6. As the patients with heart issues are prescribed Aspirin, which is able to dilute the blood, which in turn reduces the formation of blood clots and therefore counteracts the problems of many heart diseases. Another explanation would be the correlation between heart

diseases and headache or migraine, but this is not the case, as shown by Cook et al. [12] Cook et al. examined whether there is a correlation between migraine and heart diseases. In his study however he came to the conclusion that migraine is an independent risk factor for CHD.

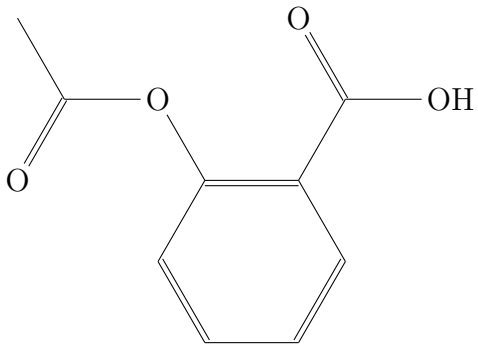

Figure 3.5: Chemical structure of Aspirin.

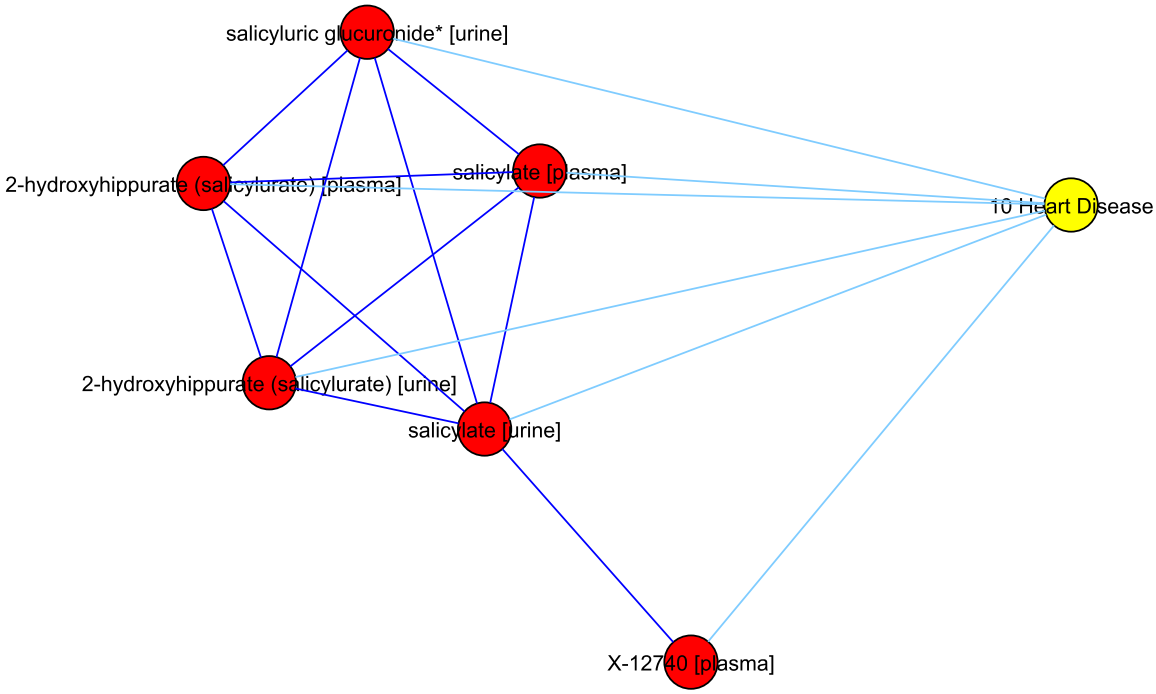

Figure 3.6: Heart disease, salicylate and its derivatives. Heart disease is correlated to salicylate and its derivatives salicylurate, salicyluric glucuronide and an unknown, X-12740. This network was cut out from the heart disease related network with an overall cutoff of 0.005.

Creatine and heart disease

Creatine is one of the most important metabolites for the contraction of muscles, as it provides the needed energy. The process of energy extraction in the muscles is a ATP to ADP reaction, where the needed ATP is produced by the reaction of creatinephosphate to creatine. This

reaction is shown in Equation 3.3.[21] Thus it is also quite important for the heart, as it is also a muscle which needs much energy. Therefore it is first of all interesting whether the high correlation between creatine and heart disease is positive or negative, which means if the heart disease patients have overall higher creatine values or lower creatine values. Figure 3.7 shows that the overall creatine concentrations are much smaller in heart disease patients. This means that the correlation between heart disease and creatine is not due to creatine intake as pharmaceutical medicament.

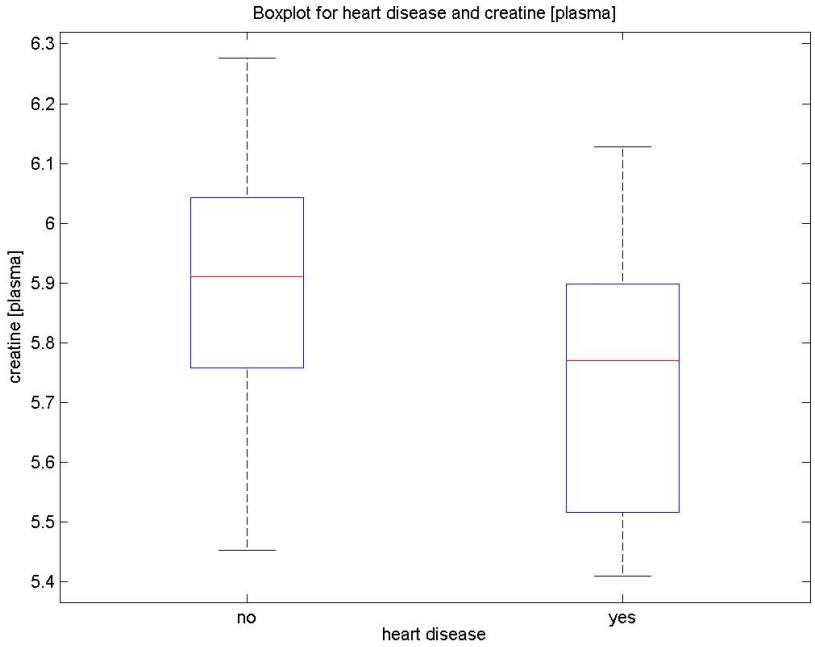

Figure 3.7: Boxplot for heart disease and the creatine levels in the plasma. The negative correlation between creatine and heart disease means, that the overall creatine concentration is lower in heart disease patients.

$$\text{Creatinephosphate} + \text{ADP} \rightleftharpoons \text{Creatine} + \text{ATP} \tag{3.3}$$

As result, it might be possible, that the production of creatine is disrupted in heart disease patients. The production of creatine however takes not place in the muscles, because there one needed enzyme, the transmethylase is not present. This means that the concentration in the blood shows how much creatine can be used in the muscles, as they are not able to produce creatine themself, but are produced in other tissues like the liver.[20] The production is a two step biosynthesis. In the first step a guanidino group is transferred from Arginin to Glycin. The resulting Guanidinoacetat is then methylated with the help of the transmethylase. The result is creatine, for which the structure is shown in Figure 3.8. Therefore one possible explanation would be, that this pathway is broken somewhere, or not efficient enough to supply enough creatine for the muscles and therefore also for the heart which in turn causes problems with the heart.

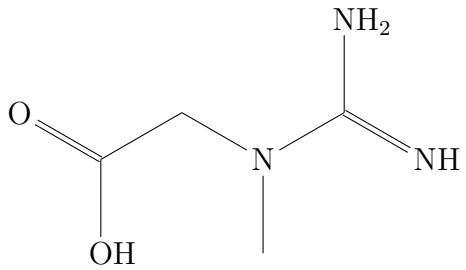

Figure 3.8: Chemical structure of creatine. Creatine is a very important detergent for the muscle contraction.

A second possibility for this correlation would be that the transportation of creatine to the blood and therefore to the muscles fails. In order to test this, it would be interesting to know, if the overall concentration of creatine in the tissues where the creatine is produced, like liver, are higher in patients with heart disease than in other patients.
Important to keep in mind is, that the correlation seen in those networks are only for diabetes patients, that have a heart disease. This means that also the blood vessels, which are attached to the tissues, which produce the creatine, might be damaged due to the diabetes. However this should result in more than just one deficiency symptom. Therefore this assumption is not as likely.
More likely however is, that the people with heart diseases are not as sportive, as those without a heart disease. This is caused by the fact that most of the time effort, like sport, worsens the heart issues, which causes people with heart diseases to do less sport. This would lead to an overall lower creatine concentration in the blood of heart disease patients, as the body will not produce as much creatine, if the muscles do not need it. Because less sport, leads to less muscle mass which in turn leads to less energy needed for those muscles. Thus the lower levels of creatine would not be the cause of those heart diseases, which the diabetes patients have, but the creatine levels would be a sideeffect of the heart disease.

Carnitines and their relation to heart disease

One carnitine (C5.M.DC [p150]) can also be seen in the Network 3.3. Carnitines in general are transporter of fatty acids and therefore the carnitines bind those fatty acids.[42] From the chemical point of view carnitines are zwitterionic alcohols and their structure can be seen in Figure 3.9.
However not all the carnitines that can be measured in the biocrates dataset are represented in the network and thus correlated to the heart disease, but only one specific, the Methylglutaryl-L-carnitine. The structure of Methylglutaryl-L-carnitine is shown in figure 3.10. This special form of carnitine is mainly known as an indicator for the Reye like syndrome.[65] The main effects of the Reye like syndrome, low levels of blood sugar in brain and liver and vomiting, are not directly connected to heart diseases. Interesting by the Reye like disease is its very high correlation to Aspirin intake, which means, that this might also be the connection between the heart disease, as the patients that take Aspirin due to the heart issues develop a form of the Reye like syndrome and therefore the Methylglutaryl-L-carnitine household is impaired, which leads to the given correlation.[23]

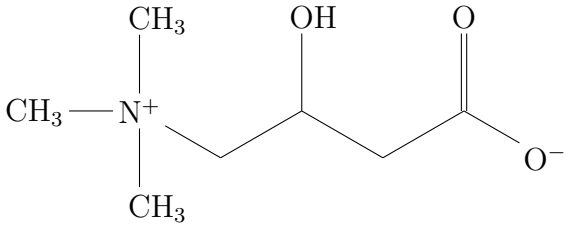

Figure 3.9: This is the chemical structure of a Carnitine in its zwitterionic form.

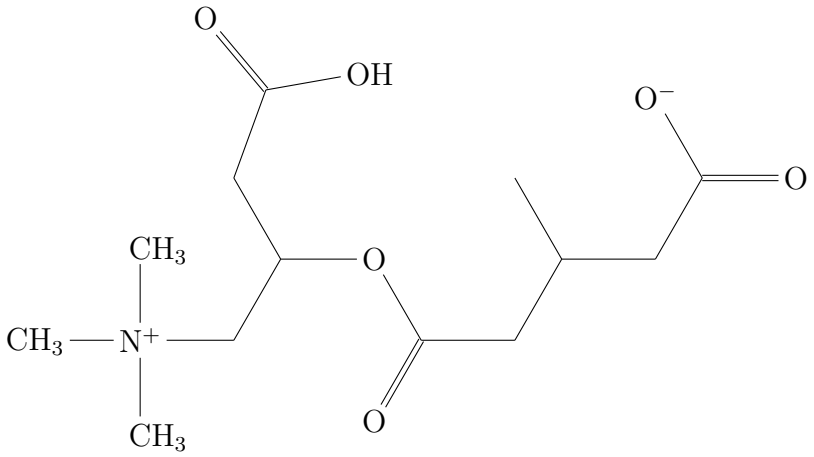

Figure 3.10: Methylglutaryl-L-carnitine. The carnitine derivative, which is related to the Reyes like syndrom and heart disease

Phosphatidylcholines and heart disease

One well represented group of metabolites in the network, are the phosphatidylcholines. There are 47 of the 77 phosphatidylcholines, which are measured with the biocrates kit, contained in this network. Phosphatidylcholines are phosphoglycerides which are located in the membrane. The hydrophilic headgroup of phosphatidylcholines contains a glycerin and a choline which is attached to it. The structure of the hydrophilic headgroup is shown in Figure 3.12. The residues R_1 and R_2 are two fatty acid and at the same time the lipophilic part which is needed as part of a membrane. As the composition of the membrane is very important in order to be able to retain its functions, the phosphoglycerides can not only be synthesized de novo, but they can be converted into each other through a reaction mechanism.[20] Therefore a certain equilibrium between the phosphoglycerides can be obtained. This also makes the concentrations of all those phosphoglycerides very dependent on each other, but very stable. This is also what one can see in the Network 3.11 where all the phosphatidylcholines that are connected to the heart disease are shown. This shows that even with a high cutoff still many phosphatidylcholines are densely connected.

The correlation between heart disease and phosphatidylcholines in general was already found by Wang et. al who stated, that cardiovascular heart diseases are not only determined by a genetic factor but also by how the micro-organisms that live in the digestive tract metabolize phosphatidylcholine.[79] This means that a difference between the metabolism of phosphatidylcholines might result cardiovascular disease.

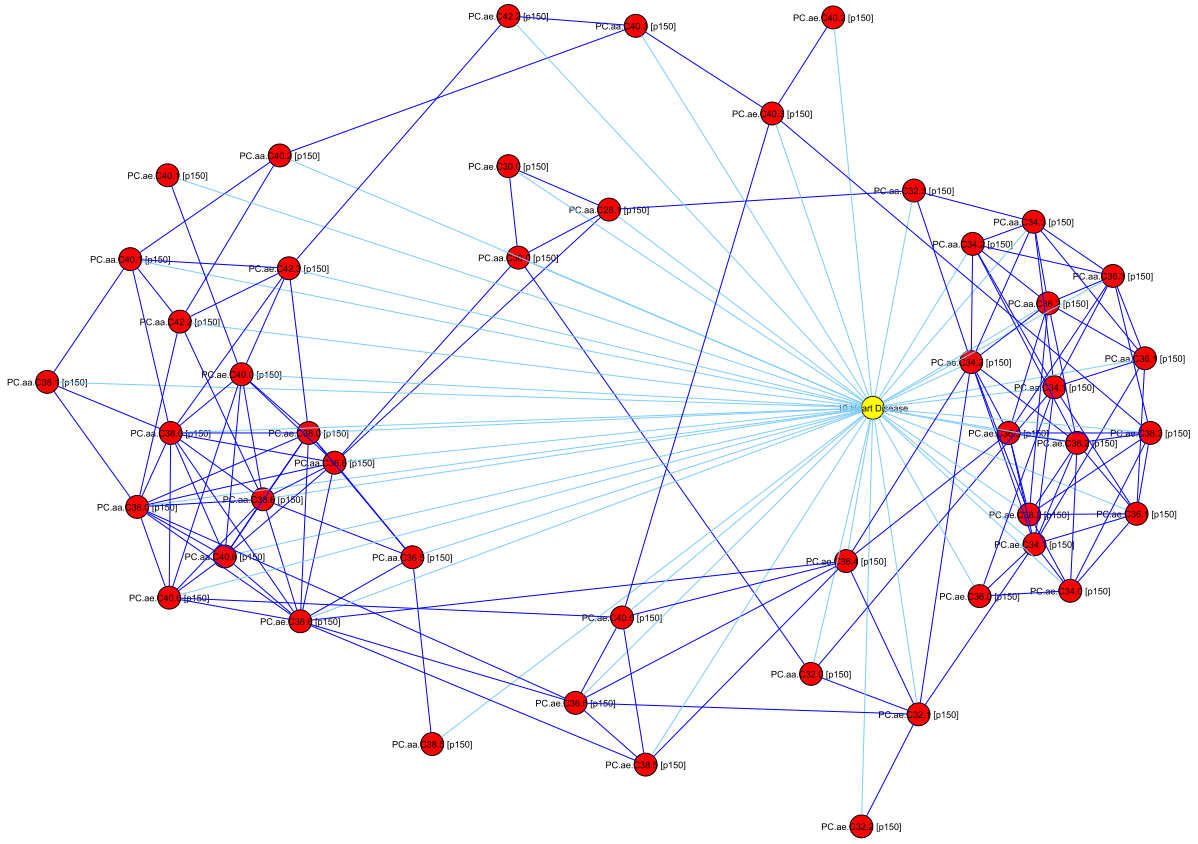

Figure 3.11: Phosphatidylcholines correlated to heart disease. If one compares it to the network 3.3 one can clearly see that the main part of the metabolites that are correlated are phosphatidylcholines. This also shows the importance of those phosphatidylcholines for the development of heart diseases.

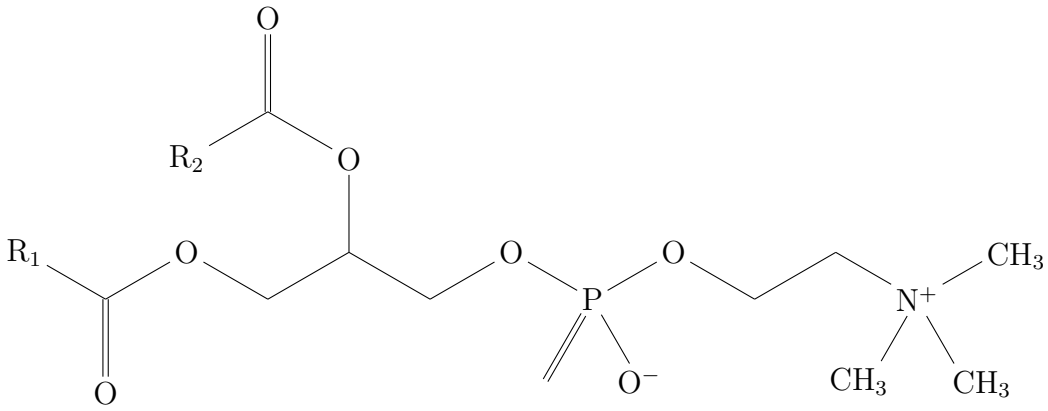

Figure 3.12: Chemical structure of phosphatidylcholins Here, R_1 and R_2 are different fatty acids which are specific for certain phosphatidylcholines.[42]

3.3 BMI-related networks

A well known risk factor for diabetes is obesity, which can be measured through different parameters that are also contained in the dataset.

3.3.1 Phenotype and metabolite network

Since the beginning of weight examination, there was always the question, which parameter best describes the relative weight, in due consideration to the stature of a patient. Thus I also looked at the different weight related phenotypes which were included in the dataset. The parameters I took are:

1. A height parameter as negative control
2. The Body Adiposity Index (BAI) which is calculated by the Equation 1.3
3. The hip circumference
4. The waist circumference
5. The waist to height ratio (WHR)
6. The weight
7. The weight to height ratio
8. The body mass index (BMI), which is calculated by the equation 1.1.
9. The waist to hip ratio
10. The ABSI (A Body Shape Index)[36] which is calculated by the Equation 1.4.

These were the ten phenotypes with which I started my examination and built up a network around those nodes (see Figure 3.13) which contains all the related phenotypes and all the related metabolites. All together the network contains 185 different metabolites and phenotypes with 1763 edges. Even though this network is highly connected with approximately 10 edges per node, there are still some metabolites, that are only related to specific phenotypes. Still only one node is related to only one single phenotype. This would be putrescine (coloured in red) which is only correlated to weight to height ratio. In addition to this, three nodes are connected to only two phenotypes (coloured in turquoise) and nine metabolites, that are correlated to 3 different phenotypes. This shows, that there are not many nodes only connected to few nodes and few nodes connected to many other nodes, which would be a so called power-law distribution in the network, but a relatively balanced amount of edges from each node.

3.3.2 Men vs women

Typically if one looks at the body mass index, there are some arguments, which suggest to divide the dataset in a men and a women dataset. One would be, that normally the percentage of fat is higher in women, that have the same BMI as men, which is caused by an overall higher muscle portion in men.[24] Another justification for the separation of men and women, would be the mean BMI, which is overall higher in men, than in women.[51] Thus I also divided the dataset into a men and a women dataset, to see, if there is some kind of evidence for the differences in the BMI values on the phenotype and metabolite level.

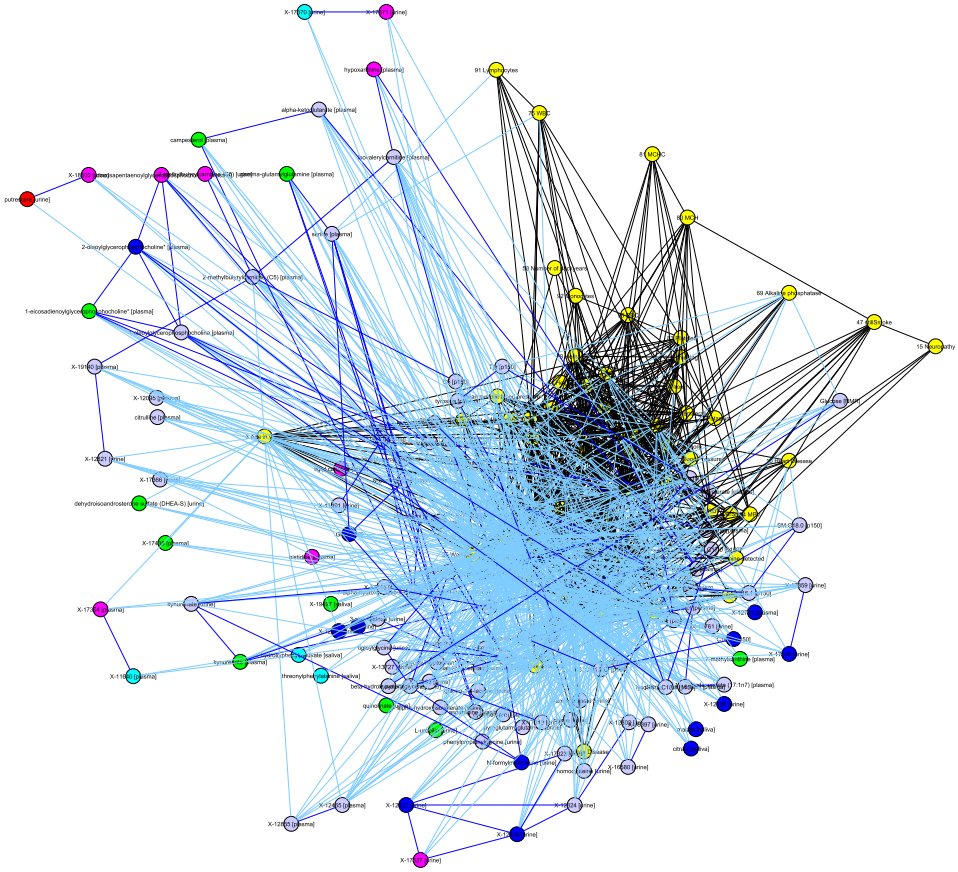

Figure 3.13: Full graph for the 10 different weight related phenotypes. Here I took a 0.05 cutoff at the phenotype to phenotype and phenotype to metabolite level and a 0.0001 cutoff at the phenotype to phenotype level. Still one problem here is, that the first order related phenotypes and metabolites are very highly interconnected and that it is hard to visualize it in a way, one can easily understand the graph.

Preinvestigation

In this dataset however, the BMI values are not distributed as expected, because the men do not have an overall higher BMI, but the BMI of the women is more spread as the mens BMI. This can be seen in the beeswarm plot (Figure 3.15). This phenomenon might be caused by either a overall different BMI distribution or some bias in the dataset.

Nonetheless looking at the different phenotype/metabolite networks can still be interesting, as the two BMI distributions are significantly different and the waist to hip ratio, which is another measurement for the healthiness of a person, is as expected overall higher in men, than in women (Figure 3.16).

Overview of the network

As it is still not very well investigated, which of the anthropometric measurements is the best in order to reflect the healthiness of a population, not only BMI and WHR were included in this examination, but also some other factors. The following list are the abbreviations of the BMI-related phenotypes, which were included in the examination.

- **zHEI:** Z-normalized height

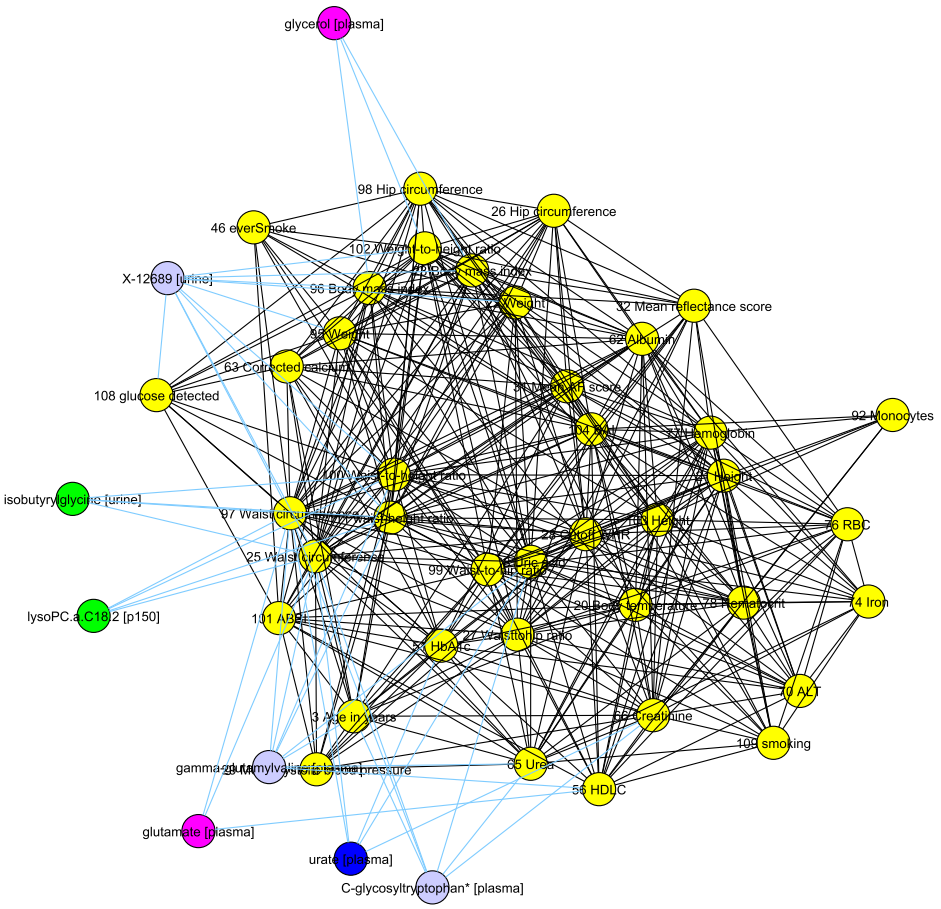

Figure 3.14: Weight related network with low cutoff. The cutoff here is significantly lower than in Figure 3.13 which means a 0.0001 cutoff at all Levels. Here the phenotypes are still too closely related and the network is still too dense to easily understand it.

- **zBAI:** Z-normalized Body adiposity index
- **zHIP:** Z-normalized hip circumference
- **zWAIST:** Z-normalized waist circumference
- **zWTHT:** Z-normalized waist to height ratio
- **zWEI:** Z-normalized weight
- **zWHtR:** Z-normalized weight to height ratio
- **zBMI:** Z-normalized body mass index
- **zWHR:** Z-normalized waist to hip ratio
- **zABSI:** Z-normalized ABSI-value (A Body Shape Index).[36]

The first difference between the men-BMI-network 3.17 and the women-BMI-network 3.18, is the different amount of metabolite nodes, as there are 3 metabolite nodes in the men-BMI-network, but 35 metabolites in the women-BMI-network. One explanation for such a

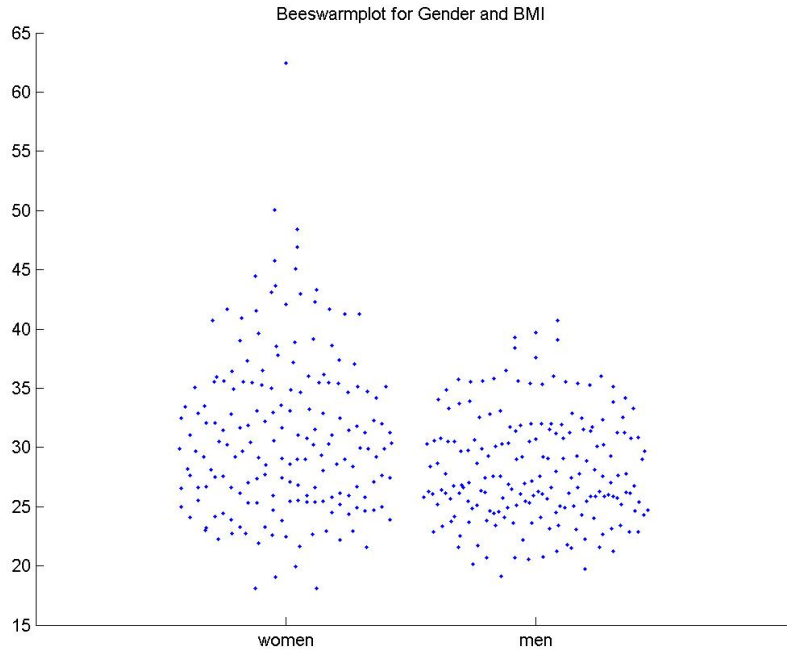

Figure 3.15: Beeswarm plot for the BMI distribution of men and women. The dots represent the different patients, which were in the dataset

Figure 3.16: Beeswarm plot of the waist to hip ratio of men and women. In this plot, the point cloud is as expected higher in men than in women, even though some outliers can be seen in the womens beeswarm.

result would be an imbalanced dataset, but this is not the case, as there are 191 males and

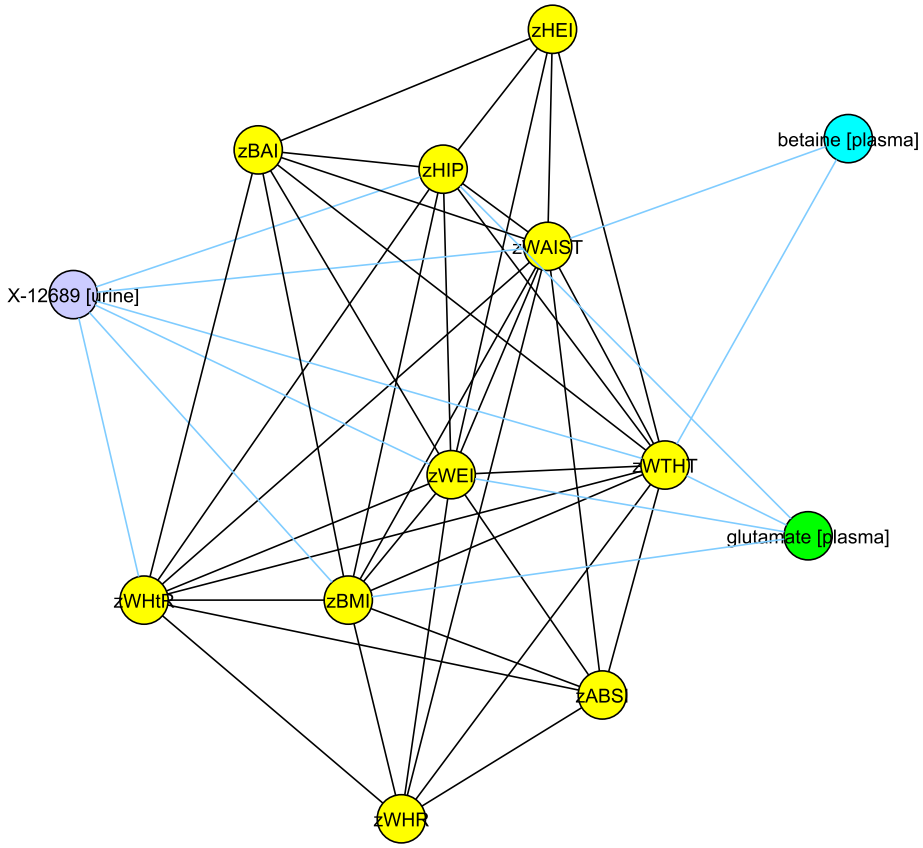

Figure 3.17: Network of weight related phenotypes for men only. The significance cutoff was set to 0.05 here.

183 females. Thus an imbalance in the power of the two datasets is not the reason for the differences in the networks. Another reasonable explanation would be caused by the rigid cutoff, with which one can not determine how significant the edge in one network truly is. This means, if the P-value of an edge in one network is slightly below the cutoff, but in the other network it is slightly above the cutoff, the difference in the P-value is not that big, but nonetheless the edge will be drawn in one network, but won't be in the other network. Figure 3.19 is an extract from the list, where I compared the Q-Values (P-Values after multiple testing correction), for the edges from the men and the women network. It shows, that the differences between the men (Figure 3.17) and women (Figure 3.18) networks are not due to a cutoff problem, as there are big differences in the Q-values for men and women.

Glutamate in the men-BMI-network

Glutamate is one of the most used flavor enhancer. Most of the time glutamate is used in the form of monosodium glutamate, as it then has preferable physical properties like soluble in water and easy storable. These properties are especially important for the food industry. In recent years, it became more and more popular, due to the fact, that it is on the one hand evaluated as safe food ingredient for the general population by the U.S. Food and Drug Administration (FDA) [18], but on the other hand its correlation to BMI was shown by multiple research teams.[26] This correlation can also be seen in the men-BMI-network, as it is linked to not only zBMI but also to zHIP, zWTHT and zWEI. Interestingly these relationships can

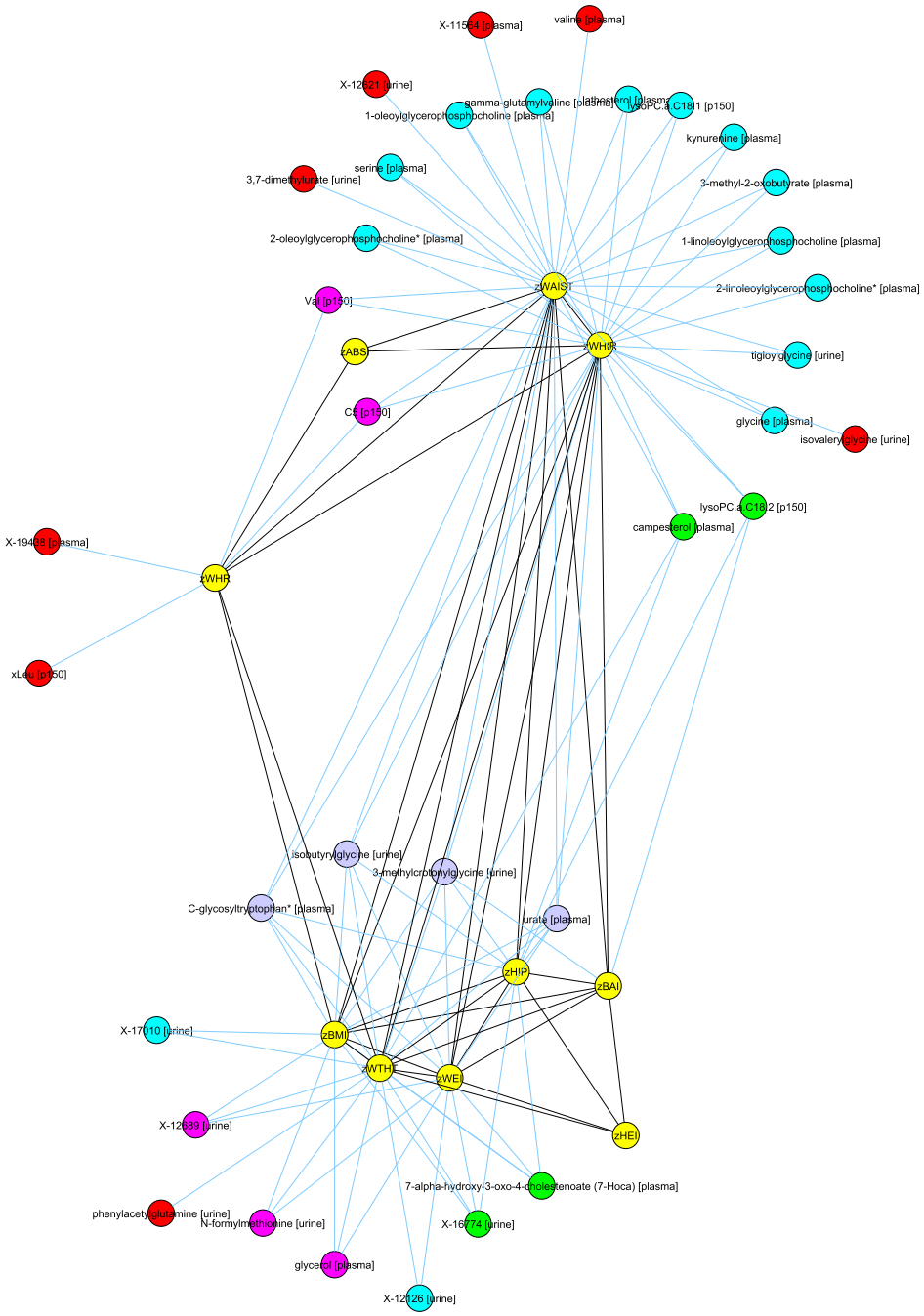

Figure 3.18: Network of weight related phenotypes for women only. The significance cutoff was set to 0.05 here.

not be seen in the women-BMI-network. This fact leads to the question, whether there was a difference in the nutrition or if there might be differences in glutamate tolerance or rather in the degradation pathway. In order to determine the true reason for this difference further data and further examinations have to be done, as there is not sufficient data in order to test this.

MenQValue	WomenQValue	Metabolite	Phenotype
0.0356757993487959	1	betaine [plasma]	ZWAIST
0.039997825165817	0.95083688873137	betaine [plasma]	ZWTHT
0.00945222306056998	0.488761453345308	glutamate [plasma]	ZWEI
0.00323841440791806	0.419024299515113	glutamate [plasma]	ZBMI
0.0468204508461569	0.348457454110644	glutamate [plasma]	ZHIP
0.00267876032242873	0.442631293464675	glutamate [plasma]	ZWTHT
0.438207734145222	0.00507157326564679	glycerol [plasma]	ZWEI
0.489990576127211	0.00444429255208364	glycerol [plasma]	ZBMI
0.441216148271033	0.00424834931906201	glycerol [plasma]	ZWTHT
0.243027268195796	0.0248398542876371	glycine [plasma]	ZWAIST
0.490628216749173	0.0389179520193849	glycine [plasma]	ZWHtR
0.258117933002314	0.0107358397390783	kynurenine [plasma]	ZWAIST
0.490628216749173	0.0301107157617205	kynurenine [plasma]	ZWHtR
0.934261611035843	0.0280237940256269	lathosterol [plasma]	ZWAIST
0.77728371573724	0.0163166259383663	lathosterol [plasma]	ZWHtR
0.468645895071387	0.0107358397390783	serine [plasma]	ZWAIST
0.593723500047184	0.0159370444660522	serine [plasma]	ZWHtR
0.48460817399461	0.0115023082337791	urate [plasma]	ZWEI
0.600462944447922	0.0301376236600071	urate [plasma]	ZBMI
0.455039304344014	0.00179220555968701	urate [plasma]	ZWAIST
0.657077602085315	0.00582779589905268	urate [plasma]	ZHIP
0.586372417509166	0.00822411416893469	urate [plasma]	ZWHtR
0.495446465304767	0.0175951009409408	urate [plasma]	ZWTHT
0.491369158115878	0.0348395812961519	valine [plasma]	ZWAIST
0.549153922500515	0.0255320841432817	C5 [p150]	ZWAIST
1	0.00160582138218288	C5 [p150]	ZWHR
0.490628216749173	0.0389179520193849	C5 [p150]	ZWHtR
0.789653124887203	0.0227479429567032	Val [p150]	ZWAIST
1	0.0249215996354335	Val [p150]	ZWHR
0.614943059257142	0.0389179520193849	Val [p150]	ZWHtR
1	0.0483892567697981	xLeu [p150]	ZWHR
0.438004682639854	0.00592380928887321	lysoPC.a.C18.1 [p150]	ZWAIST
0.766497365991909	0.00367057825390756	lysoPC.a.C18.1 [p150]	ZWHtR
0.350412411540306	0.00179220555968701	lysoPC.a.C18.2 [p150]	ZWAIST
0.657077602085315	0.0201461510520057	lysoPC.a.C18.2 [p150]	ZHIP
0.644917736983569	0.0016369393399565	lysoPC.a.C18.2 [p150]	ZWHtR
1	0.0482357297675375	lysoPC.a.C18.2 [p150]	ZBAI

Figure 3.19: Extract from the table which compares QValues for men and women. This is an extract from the whole table for the comparison on whether the differences between the two networks, Figure 3.17 and 3.18 is just a cutoff problem.

Betaine in the men-BMI-network

Betaine, also called Trimethylglycine (TMG) is an N-trimethylated amino acid which normally occurs in plants. As it was first discovered in sugar beets, it was simply called betaine, but now many other betaines have been discovered and one has to specify more exactly, which one is meant.[68] In this dataset however TMG was meant, as the listed betaine had the same molar weight as TMG.
TMG normally looks like shown in Figure 3.20. This structure shows TMG at the isoelectric point, where it is in its bipolar form.[42] In the body however most of the metabolites are not present in their zwitterion form. TMG is a known medication for homocystinuria,[30] which is a recesssive metabolic disorder of the sulfur metabolism.[46] Up until now, I found no direct relationship between the bmi-related parameters and TMG.

Interestingly the correlation between TMG and BMI can only be seen in the men-BMI-

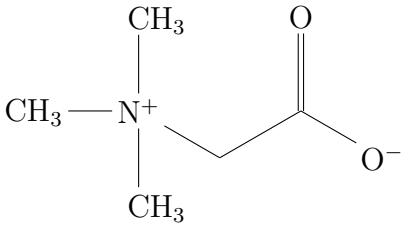

Figure 3.20: Chemical structure of trimethylglcine The structure is depicted in its bipolar form which is obtain in the range of the isoelectric point.

network and not in the women-BMI-network. Therefore the distribution has to be different in men and women. In men there is an overall higher TMG concentration which can be seen in Figure 3.21. Furthermore the correlation which can be seen in the men-BMI-network is a negative one, which means that the higher the TMG concentration is in a patient, the lower his BMI should be. Of course there are also outliers, which can be seen in Figure 3.22.

Figure 3.21: Beeswarmplot for betaine level in plasma between men (on the right hand side) and women (on the left hand side). This plot shows that the betaine levels are significantly higher in men, than in women.

The negative correlation which was shown before, was already seen in pigs, where a higher amount of TMG reduced the amount of adipose tissues.[78] The follow up study in humans however did not find any correlation between TMG and BMI.[69] Though the examined group of Schwab et al. was biased towards women, as there were only 14 men, but 28 women in the test group. Thus it might be possible, that the reduction in adipose fat tissues through TMG medication might still work in male patients but does not work in women. Another possibility would be different nutrition, as TMG, as already mentioned, can be found in sugar beets and different other plants, like Spinach [80] and can be incorporated through the nutrition.

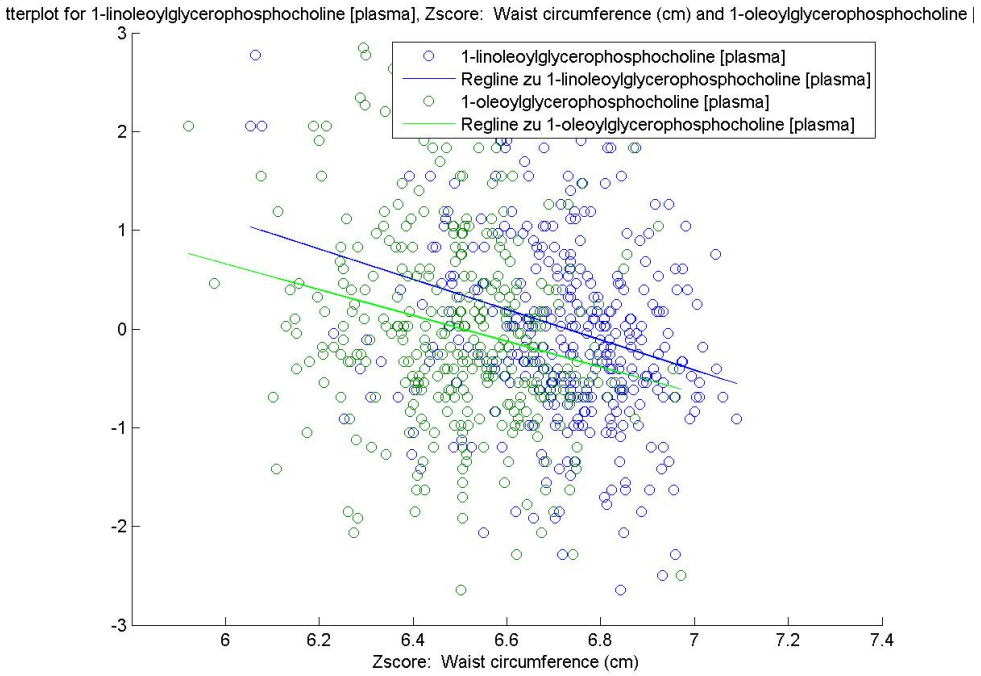

Figure 3.22: Scatter plot for BMI and betaine in men and women. The scatter plot for bmi and betaine in men and its corresponding regression line in blue and the scatter plot for bmi and betaine in women and its corresponding regression line in green. Here one can clearly see, that the dots for men are more correlated than the ones for women which are more scatter over the plot.

What also might add up to the seen relationship between TMG and the weight related phenotypes, is that choline (structure in figure 3.23) which can be degraded through a simple oxidation to betaine, is a medication against fat liver.[13] However one can argue, that a high choline concentration induces weight gain, which should lead to a positive correlation between not only choline and the weight related parameters, but also between the weight related parameters and betaine, as more choline should mean more betaine.[62] Though if the choline is oxidised to betain and therefore the negative effect of weight gain of choline is abolished, the graphs make sense. Thus choline intake might add up to the negative correlation between betaine and the weight related parameters.

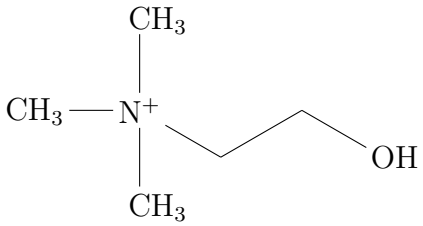

Figure 3.23: Chemical structure of choline. Choline can be oxidised in one step to trimethylglcine, which is shown in figure 3.20

Separate phenotypes in the women-BMI-network

Despite the interconnection between the phenotypes is highly similar between men-BMI-network and the women-BMI-network the metabolites, which are connected to the phenotypes are different. This is also what separates the different weight parameters from each other, as the metabolites are building groups around the phenotypes. This can be seen, if one compares the women-BMI-network (Figure 3.18) and the men-BMI-network (figure 3.17). The two main groups, that can be seen, are the waist-related group on the one hand and on the other hand the phenotypes, that are related to the overall weight, like BMI and weight. The waist-related group, but especially the two phenotypes *zWAIST* and *zWHtR* are really delimited, as there are many different metabolites, that are only related to those two phenotypes. This can be seen in Figure 3.24. The two other waist related phenotypes, which would be *zABSI* and *zWHR* are also very close to those two phenotypes and have some metabolites in common. Even though *zABSI* is not related to any metabolites, it can be seen in this group, as it is only connected to those other three phenotypes. This can be seen in Figure 3.25.

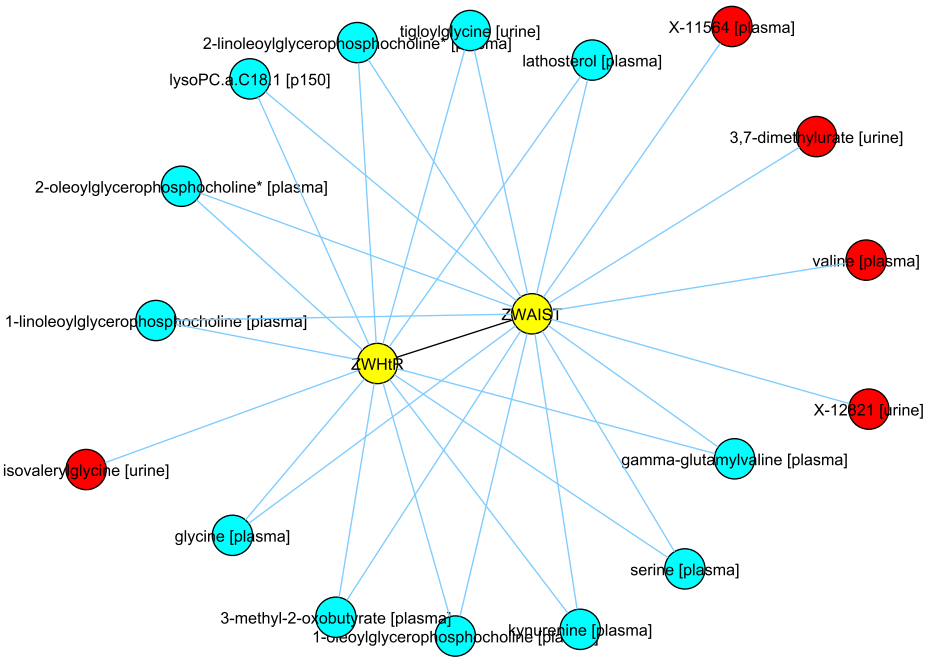

Figure 3.24: Correlated metabolites from the women-BMI-network to the two phenotypes *ZWAIST* and *ZWHtR*. Here one can see, that there are many Metabolites, that are only related to those two phenotypes, which clearly zones those two metabolites away from the rest of the network.

The first thing that is remarkable about the metabolites, that are correlated only to the two phenotypes *zWAIST* and *zWHtR* is, that they are not really structurally related. This means, that there is not only one single degradation or synthesis pathway, that is strongly correlated to the waist parameter, but there have to be multiple ones. The range of structures is from very small metabolites, like glycine (Figure 3.26) and serine (Figure 3.27), to metabolites like gamma-glutamylvaline (Figure 3.28), to relatively big metabolites like 1-oleoyl lysophosphatidylcholine (Figure 3.29) and lathosterol (Figure 3.30).

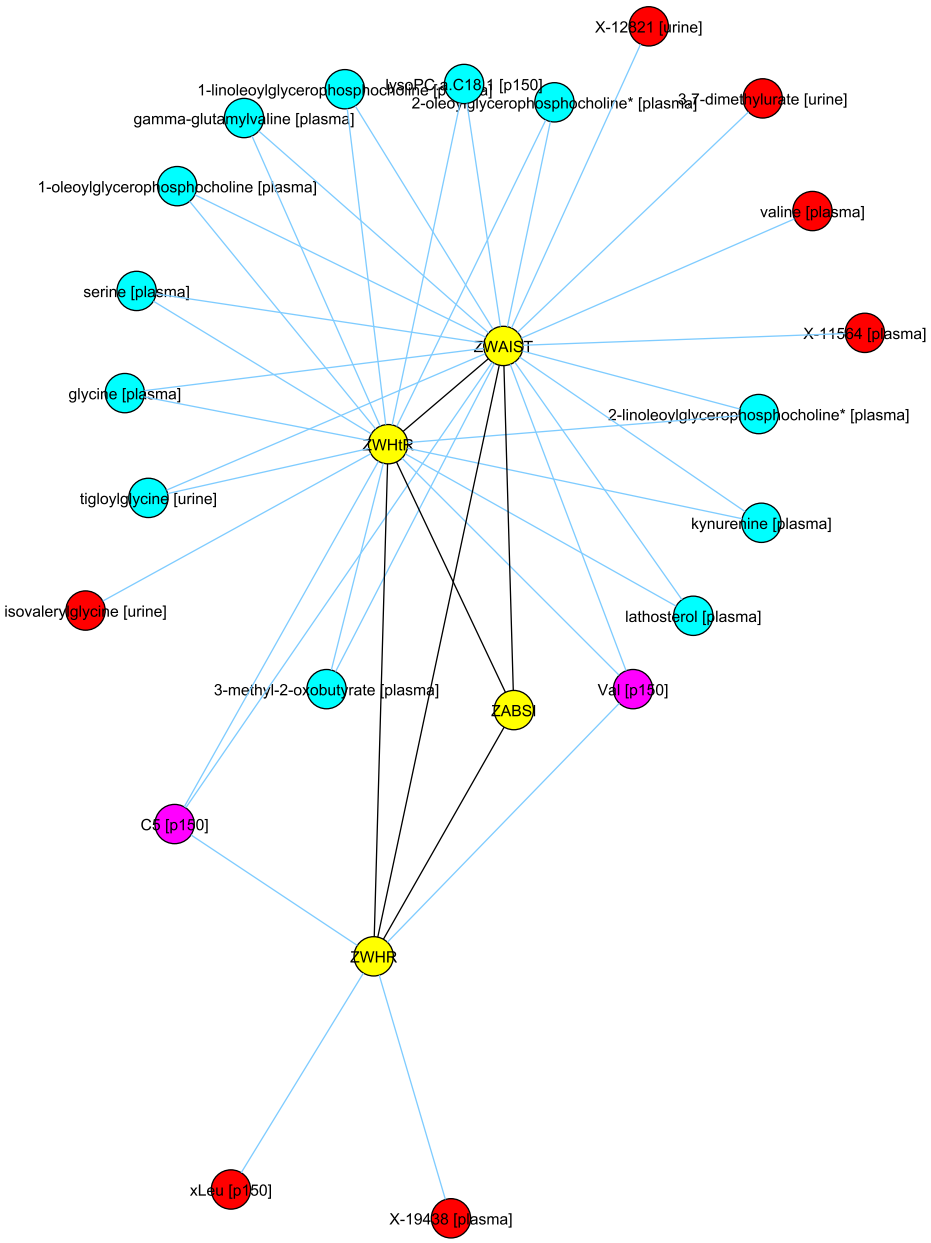

Figure 3.25: Expansion to the network 3.24. The network was expanded by the two phenotypes *ZABSI* and *ZWHR* and the metabolites, that are related to those two phenotypes and *ZWAIST* and *ZWHtR*.

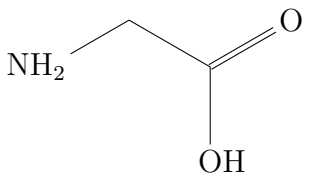

Figure 3.26: Chemical structure of glycine.

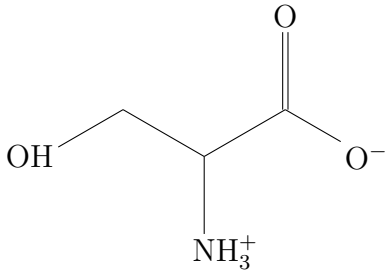

Figure 3.27: Chemical structure of serine

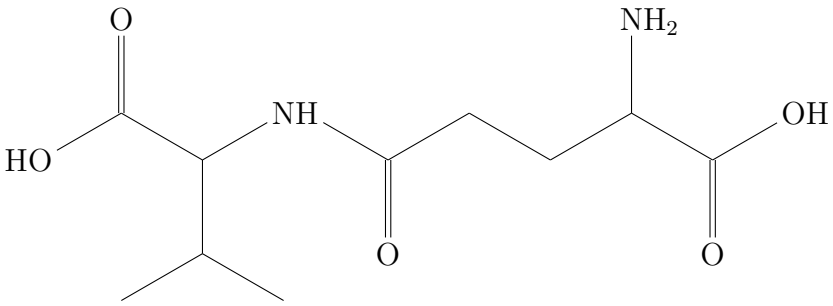

Figure 3.28: Chemical structure of gamma-glutamylvaline

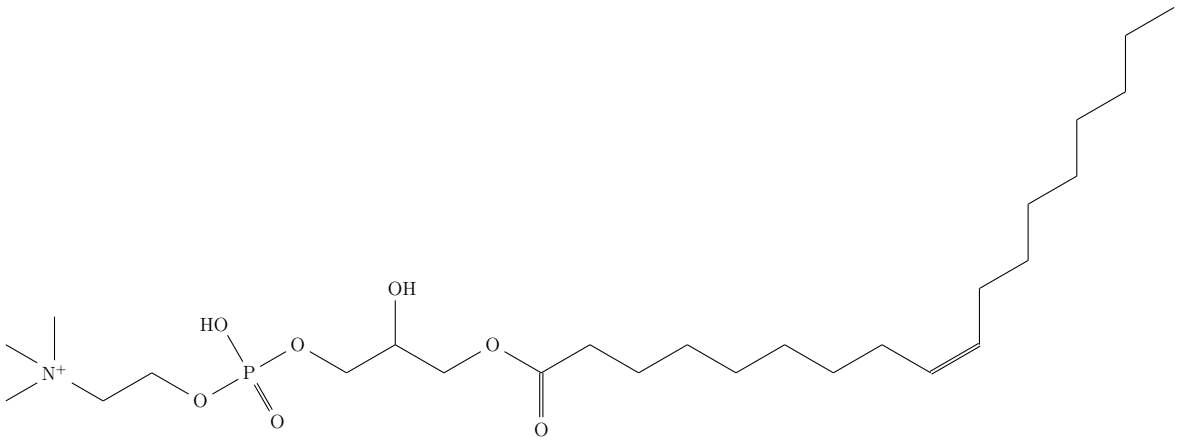

Figure 3.29: Chemical structure of 1-oleoyl lysophosphatidylcholine

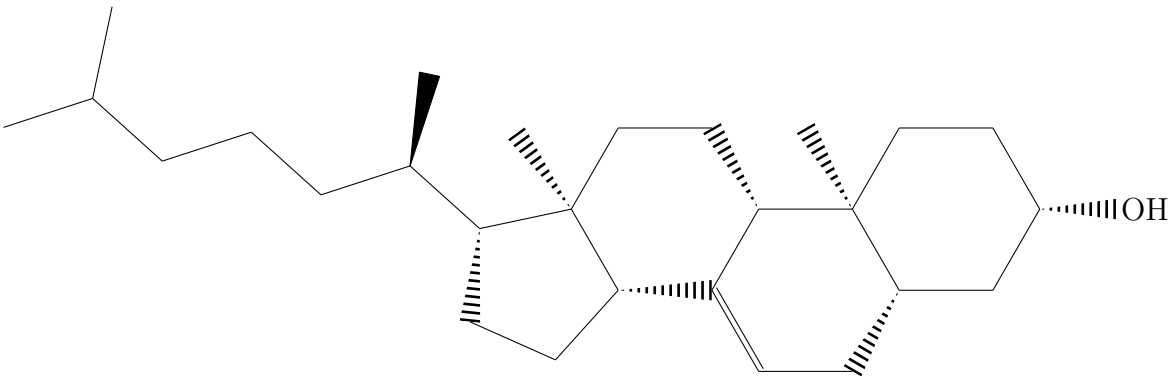

Figure 3.30: Chemical structure of lathosterol

Even though there are such different structures, one can still group many of those metabolites together. One first group would be a glycerophospholcholine group which contains the following metabolites:

- 1-linoleoylglycerophosphocholine [plasma]
- 1-oleoylglycerophosphocholine [plasma]
- 2-linoleoylglycerophosphocholine* [plasma]
- 2-oleoylglycerophosphocholine* [plasma]

The members of this group are very similar to each other, as the difference between the linoleoylglycerophosphocholine and the oleoylglycerophosphocholine is just one double bound which comes with the linoleic acid and the position, where the fatty acid binds. It is still unclear, why there is such a relationship to the waist related parameters, but not to the weight related phenotypes.
This could be explained through a 2-step pathway. The first step is, that glycerophosphocholines enhances the secretion of a growth hormone in the plasma.[35] The second step would be, as shown by Hong et al. that human growth hormones are able to reduce the waist circumference significantly, but not the overall weight.[31] This theory would first of all explain, why those glycerophosphocholines are only correlated to the waist related parameters and secondly it is also supported by the fact, that there is a negative correlation between the waist circumference and the glycerophosphocholines, which can be seen in Figure 3.31.

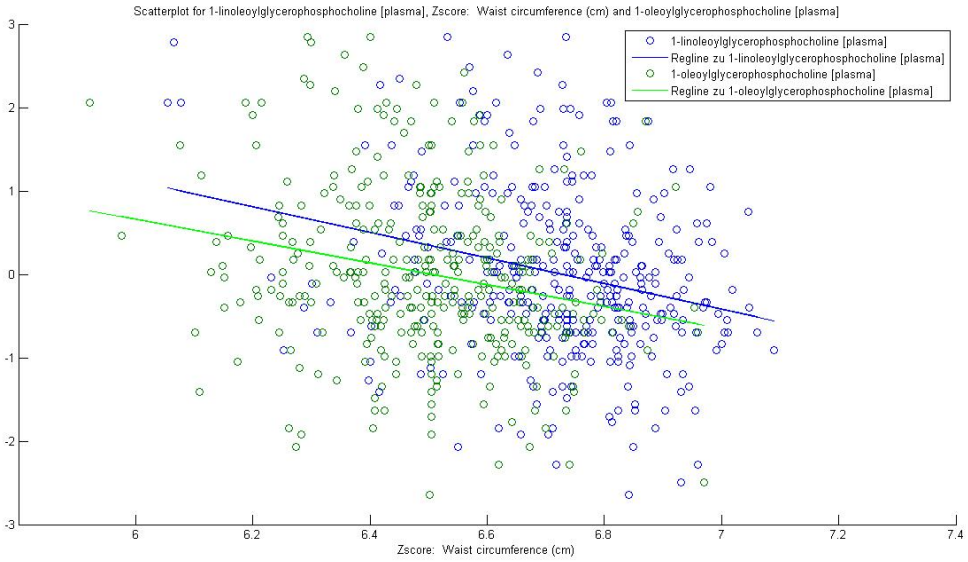

Figure 3.31: Double scatter plot for 1-linoleoylglycerophosphocholine and waist circumference as well as 1-oleoylglycerophosphocholine [plasma] and waist circumference. Both glycerophosphocholines have a negtive correlation to waist circumference, which supports the theory, that glycerophosphocholines are able to lower the waist circumference.

A second group would be a group of small amino acids. These amino acids, that are present in the Network 3.24 are:

- Glycine
- Serine
- Valine

But there are not only those single small amino acids, one can also add to this group some small molecules, which contain those amino acids, like:

- Gamma-glutamylvaline
- Isovalerylglycine
- Tigloylglycine

This group would then be certainly in the same, or a similar pathway as amino acids are needed in order to synthesize those molecules and also the three amino acids can be converted into each other. For example the serine can be converted to glycine through a pyridoxalphosphate (PALP) dependent dissociation of the hydroxymethyl group.[20] One reason for the correlation might also be that the biosynthesis is directly linked to the glycolysis, as the molecule, from which the biosynthesis of those amino acids is a intermediate of the glycolysis.

Metabolites which link all the weight related phenotypes together in the women-BMI-network

In the women-BMI-network however, there are not only metabolites, that are relevant for either only the waist related phenotypes or the weight related measurements, like BMI, but also some that are relevant for all, or at least for many of the weight measurements. Those metabolites would be:

- C-glycosyltryptophan* [plasma]
- Isobutyrylglycine [urine]
- 3-methylcrotonylglycine [urine]
- Urate [plasma]

The interesting part of this, is that there are also two glycines in this part. This shows, that glycine like metabolites do not only determine the shape of the body and are therefore related to the waist related phenotypes only, but are also important in terms of the overall weight.
High rates of urate in the plasma can be caused by lower excretion of urate from the kidney,[70], a higher synthesis or both.[14] This might be caused by the overall high amount of fat and therefore urate is related to nearly all anthropometric measures.

3.4 Mixed graphical models

In this section I used a mixed graphical model approach from Bühlmann et. al in order to see whether it really improves the performance in terms of the phenotype network.[7]

3.4.1 Getting started with mixed graphical models

In order to get started with the mixed graphical models that are based on Stability Selection and Random Forests, some dummy networks were created. In order to create the data for the given weight matrix, the provided R-methods were used. As input I gave the program a 8×8 weight matrix, which provides information for the following Directed Acyclic Graph (DAG) 3.32.

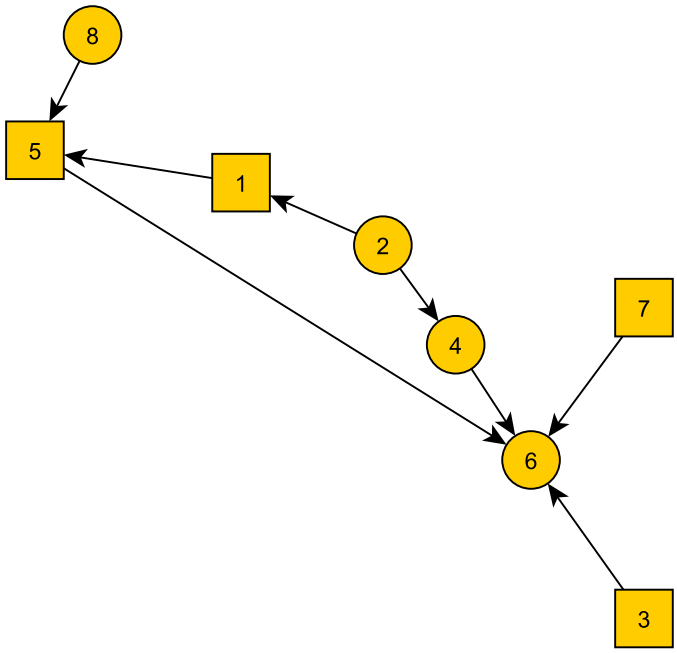

Figure 3.32: Directed acyclic graph, which was given to the mixed graphical model in order to recalculate it. The circles are phenotypes, that have categorical data and the rectangular ones have continuous data. The data was made by the provided method, which made every second parameter a categorical one.

The program then compiles mixed data according to the given weight matrix, takes out the directions from the edges, because the program is not capable of recalculating directions and makes every second parameter categorical thus a mixed dataset can be provided. The program also adds a noise level to the data, thus one can not expect to get a perfect recalculated network. There were different parameter to adjust, for example the maximal sample size, which can have a major impact on the result, if a too low sample size is taken. For the resulting network I used a sample size of 100 which occurred to be enough to get proper results, but not to huge, to get a long runtime. Other parameters would be variation which was added to the data, or the maximum amount of different values in a categorical phenotype. One of the resulting networks can be seen in Figure 3.33.

In the Network 3.34 one can see, that the program was not able to recalculate the network perfectly but nonetheless, it almost got the network right. The mistakes the program made, was that it did not predict the edge between node 5 and 6 but therefore saw a correlation between 1 and 8 and 4 and 7, where there should be none. These mistakes however can be explained through the random noise, that was automatically added to the data.

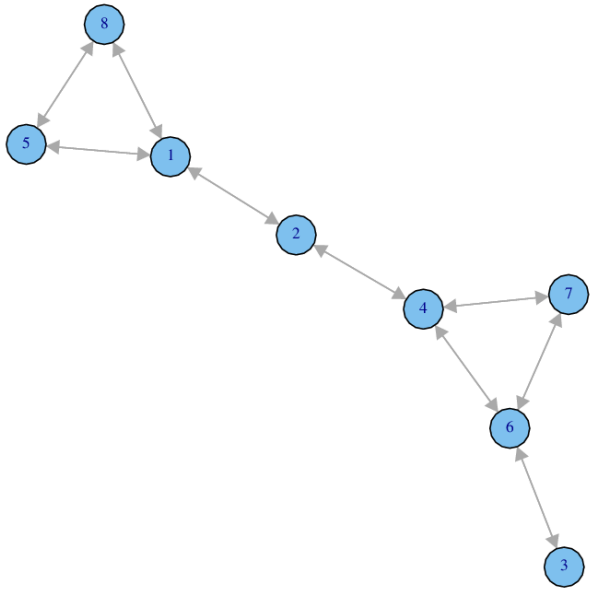

Figure 3.33: Recalculated network The mixed graphical model recalculated this network from the weight matrix, which is according to the Network 3.32.

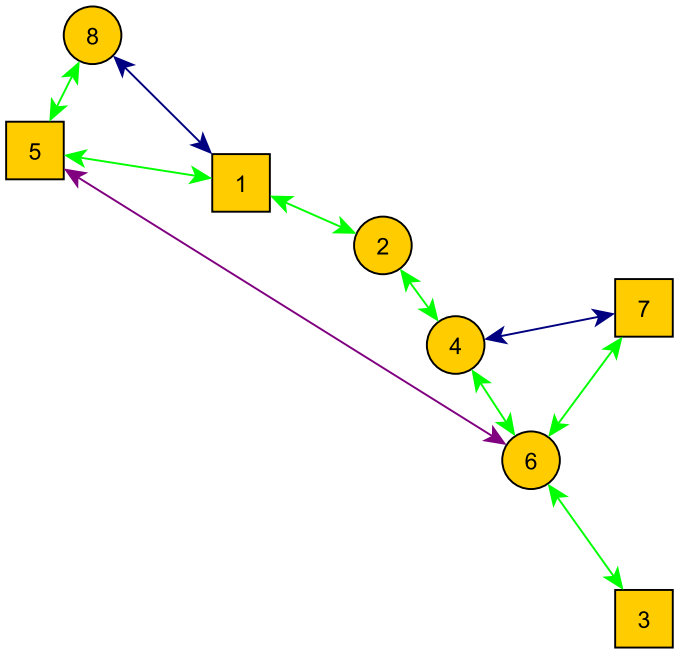

Figure 3.34: Differences between the calculated graph and the input graph. The green edges are correctly predicted ones, the violet ones, are not predicted by the program and the dark blue ones are miss calculated ones.

3.4.2 Imputing at the phenotype level

The first thing, that had to be done, before committing real data to the program was to decide how to impute at the phenotype level, as there were many missing values. In order to impute, I had to differentiate between binary and continuous variables, due to the fact that

for example taking the mean value as imputed value is wrong in the binary cases. In order to do so a simple Matlab query, which looked for all the variables with two or less variables (NaN values excluded) was sufficient to separate the binary from the continuous variables.

Imputing the binary variables

For the binary variables it was hard to decide for a global strategy, which would be one of the four:

1. Setting all missing values to 0, or false
2. Setting all missing values to 1, or true
3. Setting the missing values to 0 and 1 according to the distribution of the measured variables
4. Deleting those with missing values

There were several binary variables, which were only measured in diabetes patients. Which excluded the "deleting" strategy, as I would delete half of my dataset that way. Also setting the values to 0 or 1 would not be good, as I would first of all cause an correlation between diabetes and those phenotypes, which would not be correct in some cases. But as most of those phenotypes are said to be only present in diabetes patients, I decided to impute with 0 for the non-diabetes patients. However another method was taken in the case of for example "smoker" or "not smoker", as there was also a measurement for all patients, of the cotinine level, which is a strong indicator for smoking, thus all that have a cotinine value higher than 0 were imputed as "smoker" (1) and all that had a cotinine value of 0 were rationed as "nonsmoker". In the binary case however not only the NaN-values had to be imputed, but also the "-1". Where "-1" stands for a binary value, that could not be calculated, due to missing values in the parameters, they were calculated from. Nevertheless, the problem resolved itself, as the few "-1"-values got deleted through the deletion of the NaNs in the other parameters.

Imputing continuous variables

Recently the imputation of continuous variables got more and more in the spotlight, as there are different opinions on which way is the best. However, those discussions are most of the time on gene or metabolite datasets, which have to be imputed. Nonetheless, most of the times, the arguments are also viable in the case of phenomics. One fraction reasons, that most of the time the missing values are caused by concentrations, below the limit of detection and thus one should impute the NaNs as the minimum value. The other fraction states that the best best way to impute is the mean value, as one could invoke a correlation, were there should be no correlation. A made up example for this, is shown in Table 3.3. Where on the one hand, if one imputes with the minimum, *Phenotype1* and *Phenotype3* have a high correlation, but not *Phenotype2* and *Phenotype3*. On the other hand this correlation might just be invoked by the fact, that *Phenotype1* was not measured in positive *Phenotype3* patients, which in turn would cause a correlation, where there should not be one.

However I decided to cut most of the missing values, as they did not have an huge impact on the power of the dataset as still the data of 334 of the 375 patients was left in the dataset. This means around 10% loss of the dataset which is acceptable. Nevertheless for some phenotypes,

	Phenotype 1	Phenotype 2	Phenotype 3
Patient 1	NaN	0.2	0
Patient 2	0.8	1.6	1
Patient 3	NaN	NaN	0
Patient 4	0.7	NaN	1
Patient 5	NaN	NaN	0
Patient 6	0.2	1.6	1

Table 3.3: This table shows a very small imputing example.

imputing would have been nonsense, as they only had true values for less than half of the data. Thus imputing them with any value could cause a correlation easily, as well as deleting the missing values would cause to loose more than half of the dataset and its power. Thus I neglected those phenotypes, as the results which I would get from them were to uncertain.

3.4.3 Imputing or not

As already discussed earlier also in the case of the mixed graphical models one has to be careful, whether imputing is the right choice. In this case one of the bigger problems is, that the imputing might turn out to invoke false positive edges, by filling the same values which can not be avoided with binary variables, as there are only two possible values, which can be used for imputing. This is especially crucial for the biary phenotypes, which are only measured in diabetics patients, as imputing there with the same value causes a correlation to each other and also to diabetes. This can be seen in Figure 3.35 which is a combined network. It contains the information of two networks each with the same phenotypes, but different methods of imputation. In the one network I just deleted all the data rows, where there was a missing value in any of the phenotypes. In the other one I imputed according to previous knowledge. In the network, the green lines mark the edges, which are only in the network with the imputation, the blue ones are only in the network without imputation and the black ones are in both.
What one can see clearly is that on the one hand with the deleting of rows with missing values, also all the relationships to diabetes get lost. On the other hand by imputing one invokes some relationships between phenotypes, for example the correlation between Retinopathy and High Blood Pressure, which only are due to the imputing. But as the deletion of the missing values causes the loss of so many true datapoints and also loss of so much power, I decided to use imputation but therefore I had to be careful if I look at certain phenotypes, namely those with many missing values.

3.4.4 Compare the resulting networks

The comparison of the results of the mixed graphical models and the method that I used beforehand, the multiple correlation network with mixed graphical models for the metabolites, was harder than it should be. This was due to a different cutoff model. The cutoff model, that I used in my program was based on a P-value cutoff with a multiple testing correction after Benjamini Hochberg. However the mixed graphical model from Bühlmann et al. used another approach. They used stability selection which is an approach based on FWER. Thus it can not be easily compared to the usually used P-value cutoff approach, as the comparison

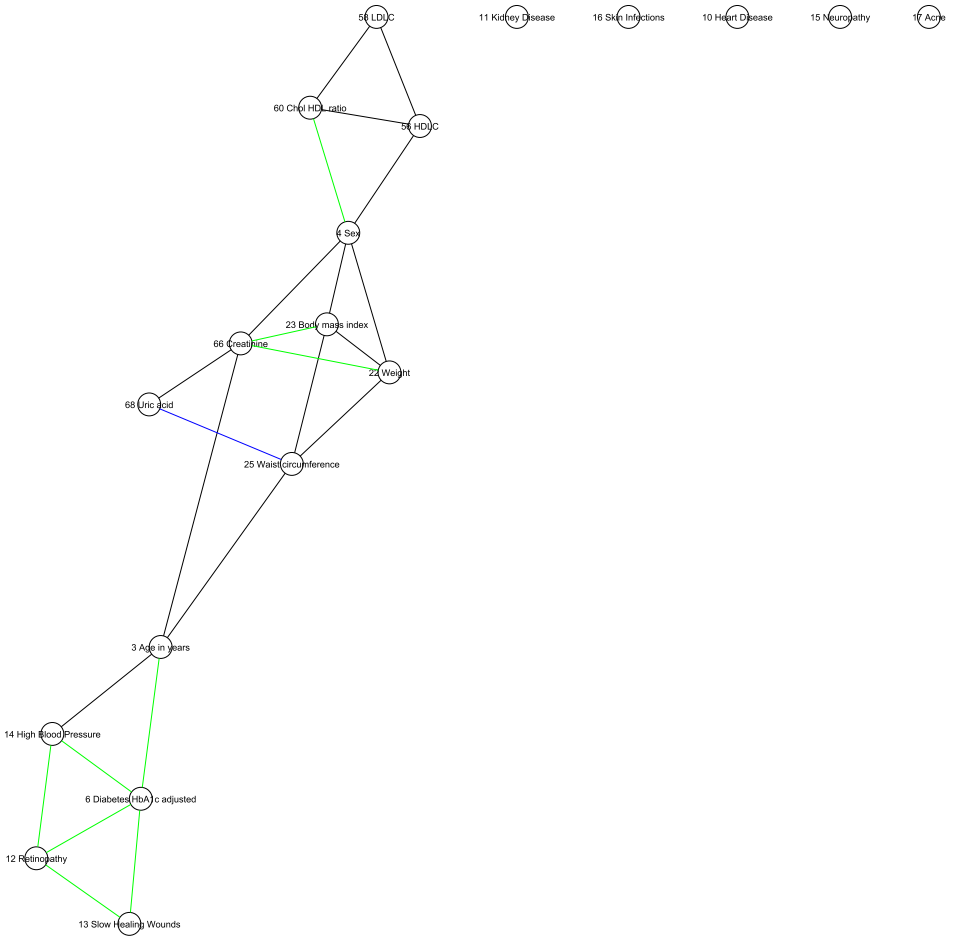

Figure 3.35: Excerpt of the phenotypes and the effects of imputing on the phenotype network This network combines two network, where one is with imputed values for the different phenotypes and the other one is without imputing, but deleting the rows with missing values, which means a severe loss of datapoints. The green edges here are the edges, which are only in the network which was imputed, the blue lines show the ones, which are only in the not imputed network and the black ones are those lines, which are in both networks.

for very small networks already takes a very long time.[2]
Therefore one can not say, whether the differences in the networks are due to a different cutoff or due to differences in the algorithms. Hence I made a network with the mixed graphical model for some phenotypes and took the same amount of edges in my approach to look at which edges are different. Thus i got the Network 3.36 with the mixed graphical model approach. This network has 23 edges and thus I also made two phenotype networks with the same amount of edges, one with the correction for co-factors (see Figure 3.37) and another one without correction (see Figure 3.38).

Comparison of the mixed graphical network with the correlation network without correction for co-factors

In order to be able to compare the inferred networks I made a differential network for the mixed graphical model and the correlation network without a correction for co-factors. The

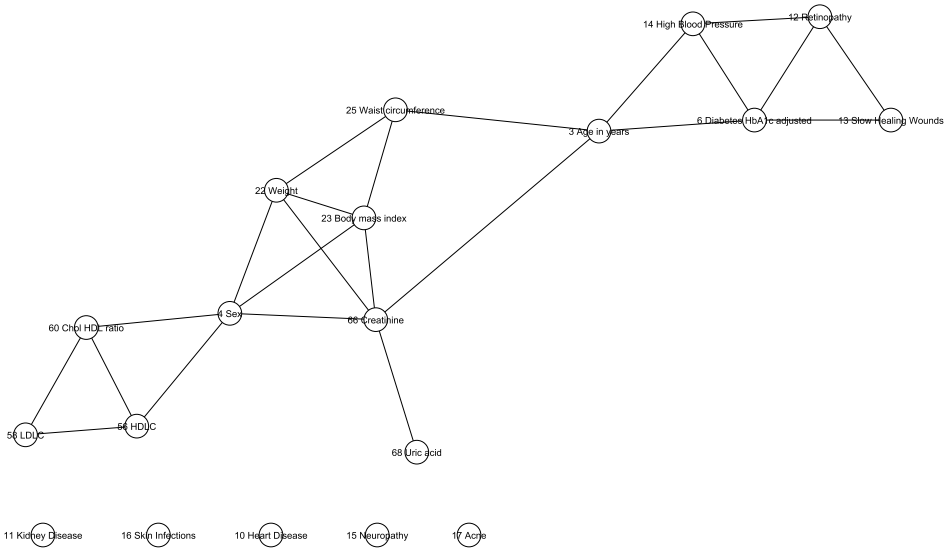

Figure 3.36: Results of the mixed graphical model for a small amount of phenotypes. The phenotypes were predetermined in order to have a good variety of interesting phenotypes.

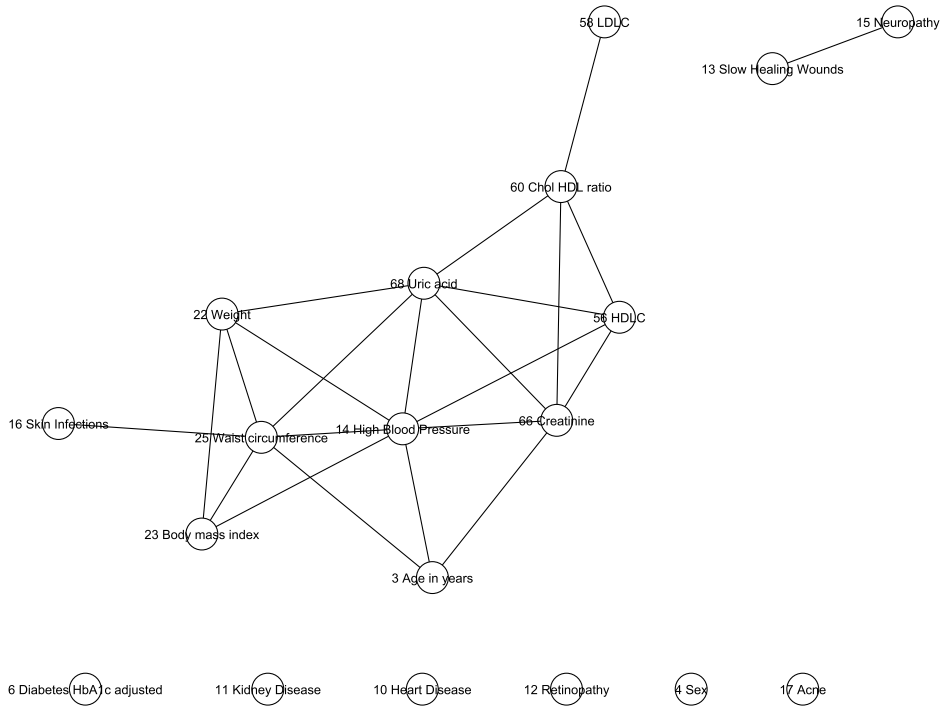

Figure 3.37: Resulting network, inferred with the correlation methods. The cutoff in this network was determined by the number of edges, which are 23 like in the network which was inferred with the mixed graphical model approach (see Figure 3.36). Therefore there was no true P-value cutoff, but a cutoff by the number of edges, in order to be able to compare it to the mixed graphical model approach.

resulting network can be seen in Figure 3.39. For the network the best 23 edges were taken from the correlation algorithm as well as from the mixed graphical model. The edges, which

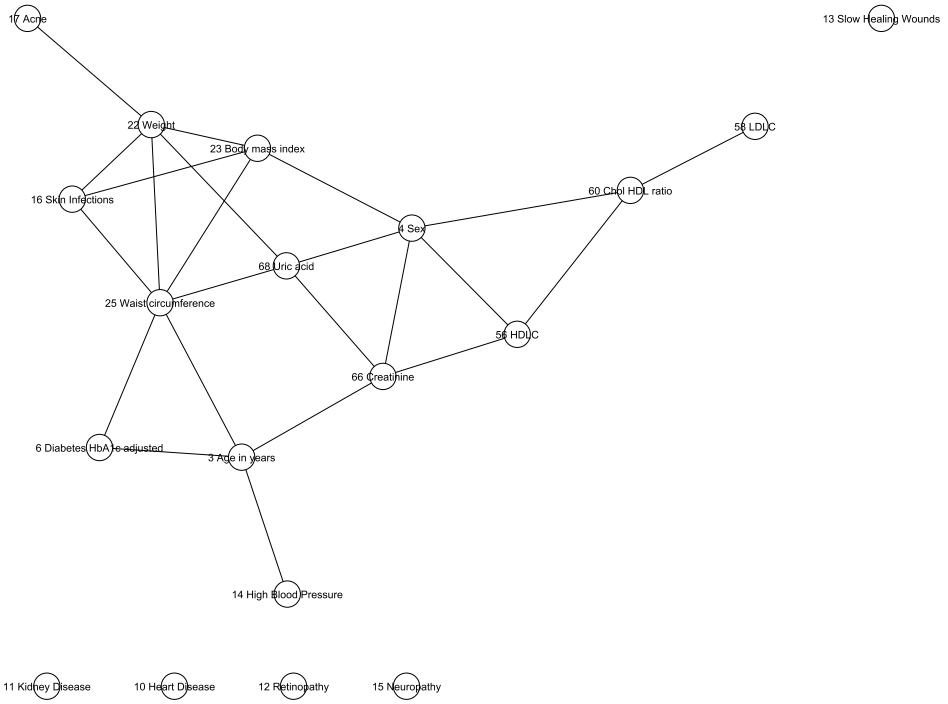

Figure 3.38: Correlation network without correction This network is similar to the network in Figure 3.37, as it also contains 23 edges and is inferred by the correlation methods, but this time it is without any correction for co-factors. Like in Figure 3.37, also in this network, the cutoff was determined by the number of edges, in order to make it comparable to the mixed graphical model approach.

were in both networks are coloured in black, the edges from the mixed graphical model are coloured in green and the edges, which are only in the correlation network are coloured blue.

One thing, that is noticeable in the network are the edges from the diabetes phenotype. As there the effects of the imputing can be seen, which would be a higher correlation of the binary variables that are only measured in diabetes patients to the diabetes phenotype. These correlations are invoked through the imputation of the missing values, as there, all the non diabetics get the same value. As those correlations can not be found in the correlation network, there have to be some other edges, that are added instead in the correlation network. Thus the networks overlay is not that big, as for every edge, that can only be found with the one approach, another edge is taken by the other approach.
Therefore from the 23 best hits of both algorithms, only 14 were similar. Still there are the 5 edges between diabetes and the imputed phenotypes, that were measured in diabetes patients and whose correlation is only due to the imputing and can only be seen in the mixed graphical model. Another influence of imputing can be seen at the phenotype "16 skin infection" which is another phenotype that was measured in diabetes patients and only has correlations to three different anthropometric measurements, weight, BMI and waist circumference in the correlation network. These edges can most likely not be seen in the Gaussian Graphical model, as with the imputed values, the correlation to the anthropometric measures gets less significant.

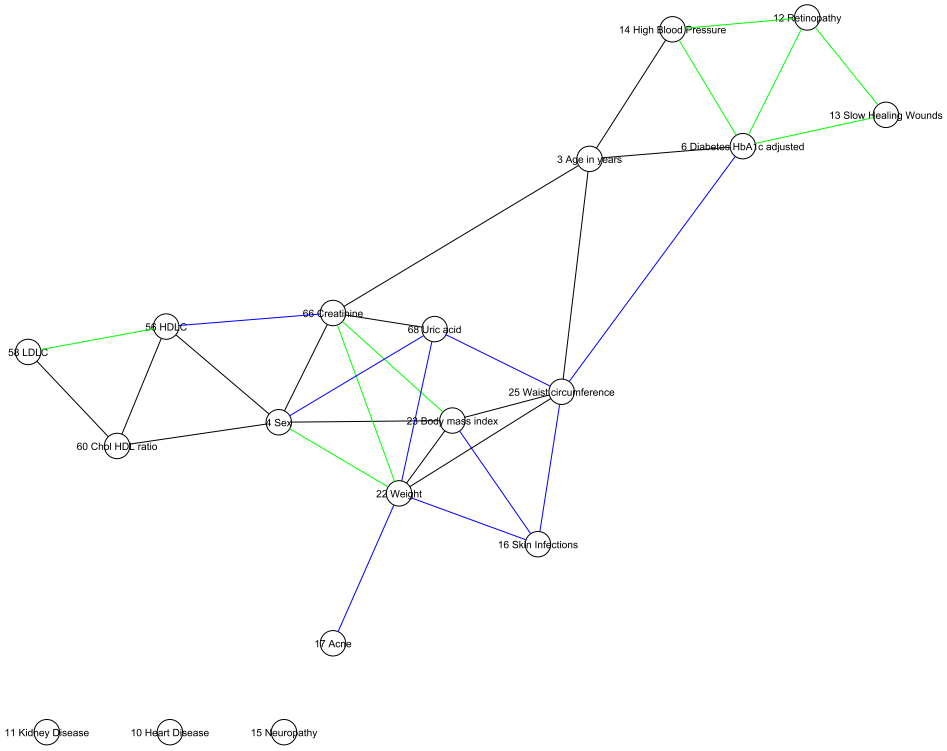

Figure 3.39: Differential network for two networks, each with 19 phenotypes and the 23 best rated edges. The edges where on the one hand rated by the correlation methods and on the other hand by the mixed graphical model. Here a green edge means, that the edge is only found in the network which was inferred by the mixed graphical model, the blue edges are only in the correlation network and the black ones, are in both networks.

Comparison of the mixed graphical network with correlation network with correction for co-factors

In order to compare the mixed graphical model, with the correlation network which was corrected for co-factors, another differential network has been built and can be seen in Figure 3.40. Here the difference between the two networks is even higher, as only 9 of the top hits are in both networks and therefore coloured black. Thus the correction for the co-factors Age, Gender and Diabetes, enlarges the gap between the correlation network and the mixed graphical model even further. The impact of the correction can be seen best at the phenotype "4 Sex", which has 5 edges only in the mixed graphical model, but 0 in the correlation network, as they were corrected out.

3.4.5 Adding metabolites to the mixed graphical models

In the metabolite area also many phenotypes were missing and due to the previous results, which showed, that imputing is not always the best way, of dealing with them, I decided not to use all the metabolite data for the mixed graphical model networks. Therefore I used all the phenotypes, which were previously selected for the correlation networks and the biocrates data for the metabolites, as those had only few missing values. The resulting network can be

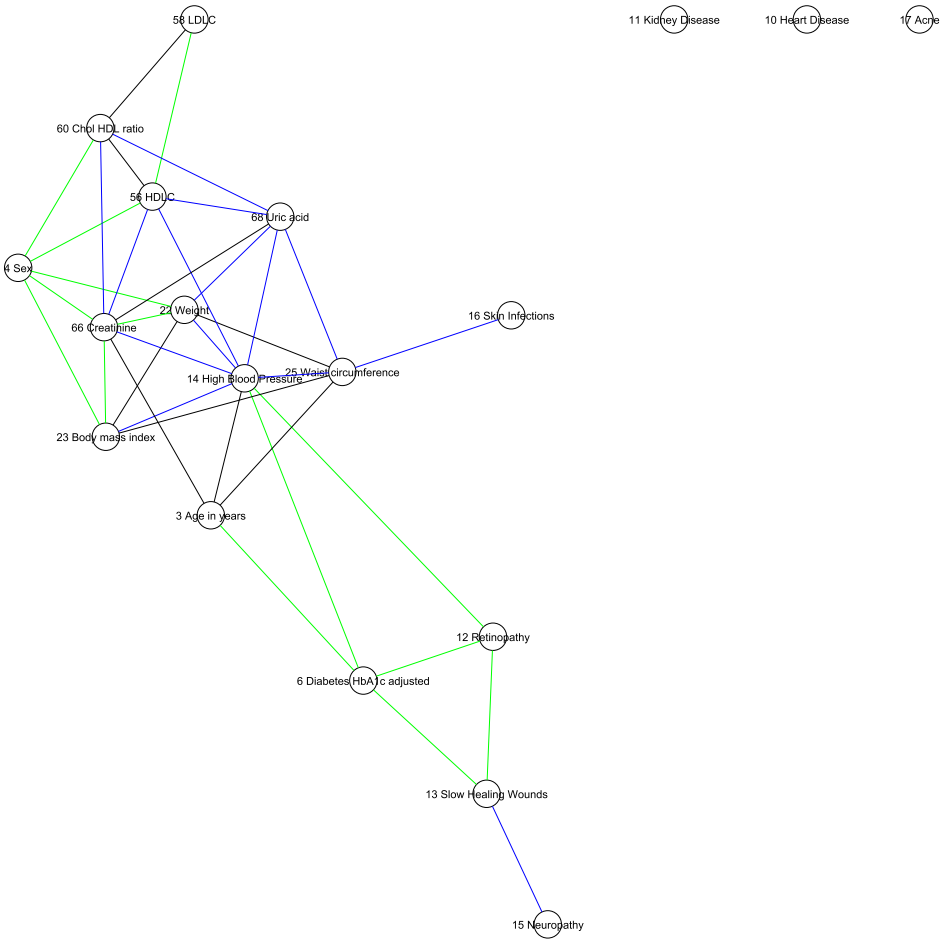

Figure 3.40: Differential network for the mixed graphical model and my correlation model with correction for co-factors.

seen in Figure 3.41.
Interestingly there are not so many different phenotypes to metabolite edges only some can be found. The most intersting part of this network is the phenotype "heart disease", which was very densely connected to many metabolites from the biocrates dataset in the correlation network, in this network however it is completely unconnected. Therefore it is also not in the network, as i deleted all the nodes with no edge. This might again be due to the imputing of the phenotype, as the imputed values have lowered the correlation to those nodes.
The network around the weight related phenotypes however stays solid, at least for the phenotypes (see Figure 3.42). As for the metabolites, most of the metabolites that were related to these phenotypes in the correlation network were in the metabolon dataset and thus can not be found in this network.

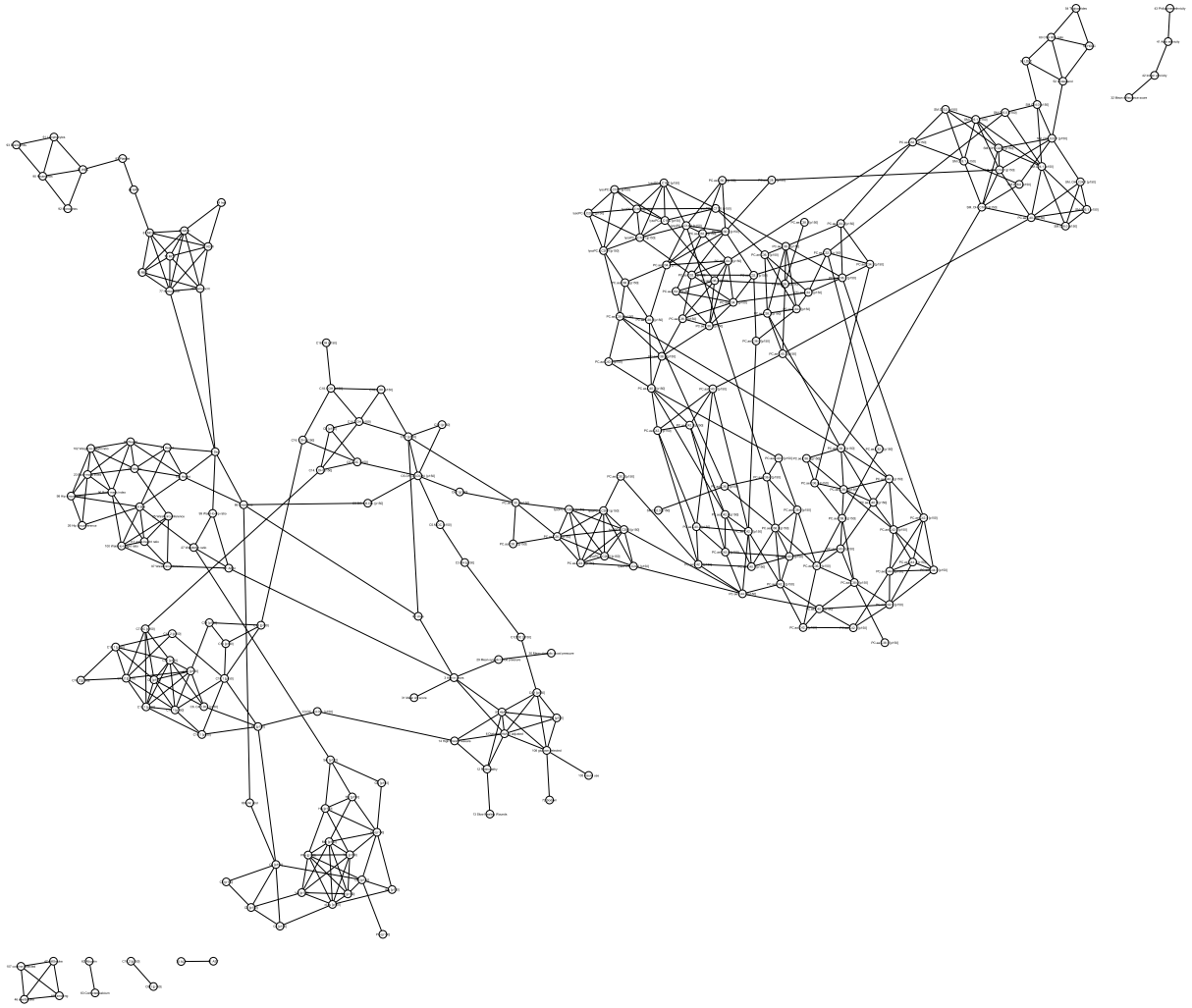

Figure 3.41: Mixed graphical model for phenotypes and the biocrates metabolite data. This network shows the network which was inferred from the phenotypes and biocrates metabolites with the mixed graphical model approach.

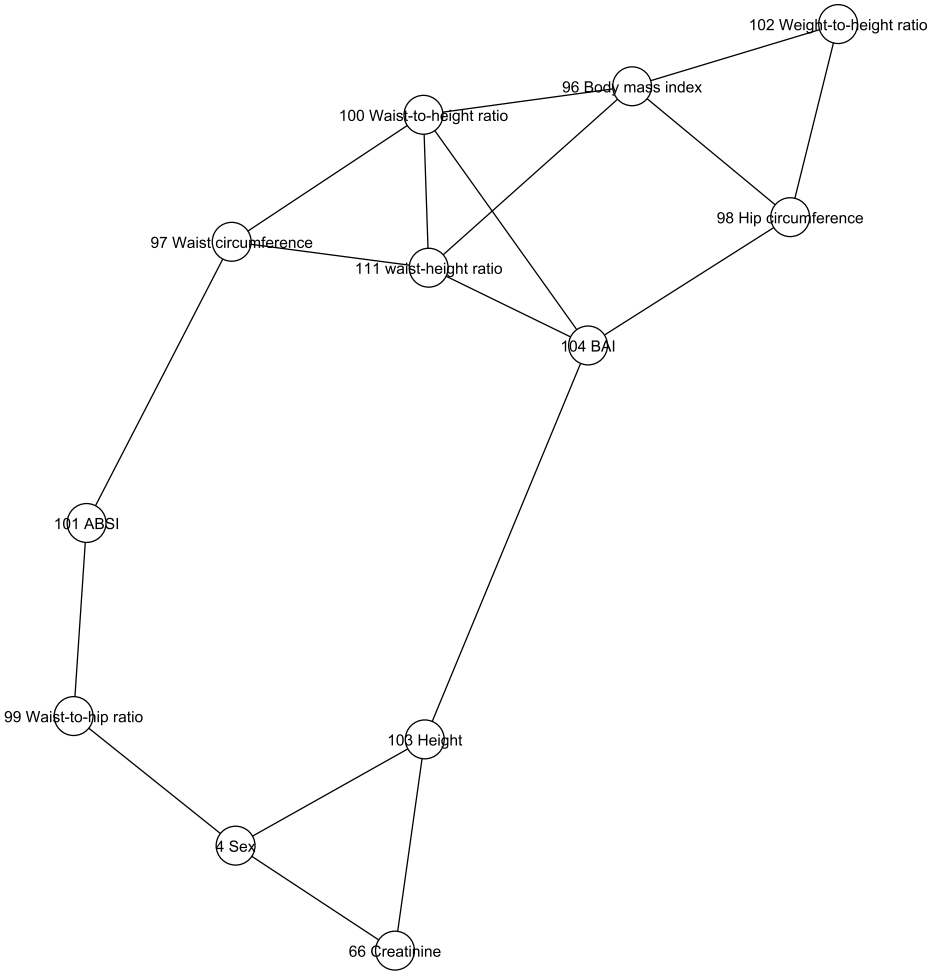

Figure 3.42: The result of the mixed graphical model for the anthropometric measurements. This network is an excerpt from Network 3.41 and shows the correlations from the anthropometric measurements.

4 Conclusion and Outlook

4.1 BMI or WHR

Which one is the best anthropometric measure, still there is no answer to this. Maybe also the body fat percentage can add another kind of information, or maybe is a better measure for overweight than the currently used ones. In my opinion, none of the weight related measures alone can completely describe which risk a patient has to develop obesity related diseases. This is why I would suggest to use multiple measurements, like BMI and WHR and to cluster different phenotypes, which are related to obesity, to the different measures. This way, if for example BMI is measured, the risk for one obesity related disease is measured and by measuring the WHR, another one. Still in order to make such clear statements by just looking at an anthropometric measure, more examinations have to be done.

4.2 Imputing in phenotype data

In the case of phenotype data, I would certainly say, imputing is not good, as this can cause so many issues. The main issue, why imputing values is not good for phenotype data, are in my opinion the binary phenotypes, because the correlations of those can be severely influenced by only few imputed values. Then again, if an interesting phenotype was only measured in half of the dataset, like the *heart disease* phenotype in the *Qatar Metabolomics Study of Diabetes*, throwing away half of the dataset, or the phenotype is even worse than imputing this phenotype. Therefore sometimes satisfying results can not be achieved due to the missing values.

4.3 Correct for co-factors or not

The correction for the different co-factors depends on the question which has to be answered with the inferred networks. In my cases I sometimes corrected for diabetes, as I wanted to know, whether two other phenotypes are related to each other even without the diabetes. Therefore the correction question can not be answered generally and it has to be solved every time a new network is generated.

4.4 Correlation networks for phenotype datasets

In general correlation networks are a solid way of inferring networks, even though Gaussian Graphical models provide a better performance for metabolic data, as shown by Krumsiek et al in their paper, where they compared GGMs with correlation networks.[38] The performance of the GGMs however is heavily dependent on the quality of the dataset, as the GGMs do not allow missing values. As the phenotype data from the Qatar Metabolomics Study of Diabetes

has many missing values in some of the very interesting phenotypes, I think using a correlation approach is the better way of inferring phenotype networks. Another advantage of correlations over the GGMs Krumsiek et al used, is that those GGMs can not handle binary data and therefore the binary phenotype data would have to be left out.

4.5 Mixed Gaussian Graphical models or correlation networks

In order to use the graphical models, which would be able to calculate partial correlations also for the phenotype data, mixed graphical models have to be used, as they can handle binary, categorical and continuous data at the same time. This solves the problem of not being able to calculate GGMs for binary data, but still there is the problem with the full data matrix which is needed for the calculation of a GGM and for a mixed graphical model. This again means that there are three possibilities. The first possibility is to leave out the data for the patient who has a missing value, which lowers the power of the study. The second possibility is to leave out the phenotype with the missing value. The third one would be to impute the missing values. As none of the three possibilities is satisfying, I think that for a dataset like the Qatar Metabolomics Study of Diabetes the correlation network is still the best option.

List of Tables

List of Figures

Bibliography

[1] Biocrates Life Sciences AG. Biocrates platform, 2012-2013. URL http://www.biocrates.com/technology/our-targeted-metabolomics-platform.

[2] Ismaïl Ahmed, Anna-Liisa Hartikainen, Marjo-Riitta Järvelin, and Sylvia Richardson. False discovery rate estimation for stability selection: application to genome-wide association studies. *Stat Appl Genet Mol Biol*, 10(1), 2011. doi: 10.2202/1544-6115.1663. URL http://dx.doi.org/10.2202/1544-6115.1663.

[3] American Diabetes Association. Living with diabetes: Complications, 1995-2013. URL http://www.diabetes.org/living-with-diabetes/complications/.

[4] Lewis A. Barness, John M. Opitz, and Enid Gilbert-Barness. Obesity: genetic, molecular, and environmental aspects. *Am J Med Genet A*, 143A(24):3016–3034, Dec 2007. doi: 10.1002/ajmg.a.32035. URL http://dx.doi.org/10.1002/ajmg.a.32035.

[5] Olaf Beckonert, Hector C. Keun, Timothy M D. Ebbels, Jacob Bundy, Elaine Holmes, John C. Lindon, and Jeremy K. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for nmr spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc*, 2(11):2692–2703, 2007. doi: 10.1038/nprot.2007.376. URL http://dx.doi.org/10.1038/nprot.2007.376.

[6] Richard N. Bergman, Darko Stefanovski, Thomas A. Buchanan, Anne E. Sumner, James C. Reynolds, Nancy G. Sebring, Anny H. Xiang, and Richard M. Watanabe. A better index of body adiposity. *Obesity (Silver Spring)*, 19(5):1083–1089, May 2011. doi: 10.1038/oby.2011.38. URL http://dx.doi.org/10.1038/oby.2011.38.

[7] Martin Ryffel Michael von Rhein Jan D. Reinhardt Bernd Fellinghauer, Peter Bühlmann. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. 2013.

[8] 6020 Innsbruck Austria Biocrates Life Sciences AG, Eduard-Bodem-Gasse 8. Absoluteidq p180 kit, 2012-2013. URL http://www.biocrates.com/products/research-products/absoluteidq-p180-kit.

[9] 6020 Innsbruck Austria Biocrates Life Sciences AG, Eduard-Bodem-Gasse 8. About metabonomics, 2012-2013. URL http://www.biocrates.com/technology/about-metabolomics.

[10] Tessa N. Campbell and Francis Y M. Choy. Gaucher disease and the synucleinopathies: refining the relationship. *Orphanet J Rare Dis*, 7:12, 2012. doi: 10.1186/1750-1172-7-12. URL http://dx.doi.org/10.1186/1750-1172-7-12.

[11] Xueying Chen, Aijun Sun, Yunzeng Zou, Junbo Ge, Jason M. Lazar, and Xian-Cheng Jiang. Impact of sphingomyelin levels on coronary heart disease and left ventricular systolic function in humans. *Nutr Metab (Lond)*, 8(1):25, 2011. doi: 10.1186/1743-7075-8-25. URL http://dx.doi.org/10.1186/1743-7075-8-25.

[12] Nancy R. Cook, Isabela M. Bensenor, Paulo A. Lotufo, I-Min Lee, Patrick J. Skerrett, Marilyn J. Chown, Umed A. Ajani, JoAnn E. Manson, and Julie E. Buring. Migraine and coronary heart disease in women and men. *Headache*, 42(8):715–727, Sep 2002.

[13] Karen D. Corbin and Steven H. Zeisel. Choline metabolism provides novel insights into nonalcoholic fatty liver disease and its progression. *Curr Opin Gastroenterol*, 28(2):159–165, Mar 2012. doi: 10.1097/MOG.0b013e32834e7b4b. URL http://dx.doi.org/10.1097/MOG.0b013e32834e7b4b.

[14] Erick Prado de Oliveira and Roberto Carlos Burini. High plasma uric acid concentration: causes and consequences. *Diabetol Metab Syndr*, 4:12, 2012. doi: 10.1186/1758-5996-4-12. URL http://dx.doi.org/10.1186/1758-5996-4-12.

[15] Kieu Trinh Do. Metabolomics analysis of multiple bodyfluids in the qatar metabolomics study of diabetes., 2013.

[16] Ana Florescu, Roberta Ferrence, Tom Einarson, Peter Selby, Offie Soldin, and Gideon Koren. Methods for quantification of exposure to cigarette smoking and environmental tobacco smoke: focus on developmental toxicology. *Ther Drug Monit*, 31(1):14–30, Feb 2009. doi: 10.1097/FTD.0b013e3181957a3b. URL http://dx.doi.org/10.1097/FTD.0b013e3181957a3b.

[17] A. R. Folsom, J. Stevens, P. J. Schreiner, and P. G. McGovern. Body mass index, waist/hip ratio, and coronary heart disease incidence in african americans and whites. atherosclerosis risk in communities study investigators. *Am J Epidemiol*, 148(12):1187–1194, Dec 1998.

[18] U.S. Food and Drug Administration. Questions and answers on monosodium glutamate (msg), 2012. URL http://www.fda.gov/Food/IngredientsPackagingLabeling/FoodAdditivesIngredients/ucm328728.htm.

[19] Scripps Center for Metabolomics. Services: Targeted metabolomics, 2013. URL http://masspec.scripps.edu/services/metabolomics/smplprep.php.

[20] Petro E.Petrides Georg Löffler. *Biochemie und Pathobiochemie*, volume 6.Auflage. Springer Verlag Heidelberg, 1998.

[21] Peter Vaupel Gerhard Thews, Ernst Mutschler. *Anatomie Physiologie Pathophysiologie des Menschen*, volume 5th Edition. Wissenschaftliche Verlagsgesellschaft mbH Stuttgart, 1999.

[22] Helen G. Gika, Georgios A. Theodoridis, Julia E. Wingate, and Ian D. Wilson. Within-day reproducibility of an hplc-ms-based method for metabonomic analysis: application to human urine. *J Proteome Res*, 6(8):3291–3303, Aug 2007. doi: 10.1021/pr070183p. URL http://dx.doi.org/10.1021/pr070183p.

[23] Jayaprakash A. Gosalakkal and Vishwanath Kamoji. Reye syndrome and reye-like syndrome. *Pediatr Neurol*, 39(3):198–200, Sep 2008. doi: 10.1016/j.pediatrneurol.2008.06.003. URL http://dx.doi.org/10.1016/j.pediatrneurol.2008.06.003.

[24] Steven B. Halls. The bmi gap between men and women, 2008. URL http://www.halls.md/bmi/gap.htm.

[25] David W. Haslam and W Philip T. James. Obesity. *Lancet*, 366(9492):1197–1209, Oct 2005. doi: 10.1016/S0140-6736(05)67483-1. URL http://dx.doi.org/10.1016/S0140-6736(05)67483-1.

[26] Ka He, Liancheng Zhao, Martha L. Daviglus, Alan R. Dyer, Linda Van Horn, Daniel Garside, Liguang Zhu, Dongshuang Guo, Yangfeng Wu, Beifan Zhou, Jeremiah Stamler, and I. N. T. E. R. M. A. P Cooperative Research Group . Association of monosodium glutamate intake with overweight in chinese adults: the intermap study. *Obesity (Silver Spring)*, 16(8):1875–1880, Aug 2008.

[27] Clearly Health. Nerve damage (neuropathy), 2012. URL http://www.clearlyhealth.com/videos/diabetes/neuropathy.

[28] Melonie Heron, Donna L. Hoyert, Sherry L. Murphy, Jiaquan Xu, Kenneth D. Kochanek, and Betzaida Tejada-Vera. Deaths: final data for 2006. *Natl Vital Stat Rep*, 57(14):1–134, Apr 2009.

[29] Universität Hohenheim. Interaktives bmi (body mass index). URL https://www.uni-hohenheim.de/wwwin140/info/interaktives/bmi.htm.

[30] Pål I. Holm, Per M. Ueland, Stein Emil Vollset, Øivind Midttun, Henk J. Blom, Miranda B A J. Keijzer, and Martin den Heijer. Betaine and folate status as cooperative determinants of plasma homocysteine in humans. *Arterioscler Thromb Vasc Biol*, 25(2):379–385, Feb 2005. doi: 10.1161/01.ATV.0000151283.33976.e6. URL http://dx.doi.org/10.1161/01.ATV.0000151283.33976.e6.

[31] J. W. Hong, J. K. Park, C-Y. Lim, S. W. Kim, Y-S. Chung, S. W. Kim, and E. J. Lee. A weekly administered sustained-release growth hormone reduces visceral fat and waist circumference in abdominal obesity. *Horm Metab Res*, 43(13):956–961, Dec 2011. doi: 10.1055/s-0031-1291246. URL http://dx.doi.org/10.1055/s-0031-1291246.

[32] Chenomx Inc. Chenomx, metabolite discovery and measurement, 2009-2013. URL http://www.chenomx.com/.

[33] Arnold D. Well Jerome L. Myers. *Research Design and Statistical Analysis*. 2003.

[34] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, Jan 2004. doi: 10.1093/nar/gkh063. URL http://dx.doi.org/10.1093/nar/gkh063.

[35] Takashi Kawamura, Takeshi Okubo, Koji Sato, Satoshi Fujita, Kazushige Goto, Takafumi Hamaoka, and Motoyuki Iemitsu. Glycerophosphocholine enhances growth hormone secretion and fat oxidation in young adults. *Nutrition*, 28(11-12):1122–1126, 2012. doi: 10.1016/j.nut.2012.02.011. URL http://dx.doi.org/10.1016/j.nut.2012.02.011.

[36] Nir Y. Krakauer and Jesse C. Krakauer. A new body shape index predicts mortality hazard independently of body mass index. *PLoS One*, 7(7):e39504, 2012. doi: 10.1371/journal.pone.0039504. URL http://dx.doi.org/10.1371/journal.pone.0039504.

[37] Per Kraulis. Lecture notes: Structural biochemistry and bioinformatics 2001. metabolic networks, 2001. URL http://www.avatar.se/strbio2001/metabolic/what.html.

[38] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5:21, 2011. doi: 10.1186/1752-0509-5-21. URL http://dx.doi.org/10.1186/1752-0509-5-21.

[39] Dr. Ravi Kumar. Diabetes and neuropathy, 2013. URL http://www.hyderabadendocrinology.com/content/diabetes-and-neuropathy.

[40] Kanehisa Laboratories. Valine, leucine and isoleucine biosynthesis, 2012. URL http://www.genome.jp/kegg-bin/show_pathway?org_name=rn&mapno=00290&mapscale=&show_description=hide.

[41] S. Lamon-Fava, P. W. Wilson, and E. J. Schaefer. Impact of body mass index on coronary heart disease risk factors in men and women. the framingham offspring study. *Arterioscler Thromb Vasc Biol*, 16(12):1509–1515, Dec 1996.

[42] J.L.Tymoczko L.Stryer, J.M.Berg. *Stryer Biochemie*. 2007.

[43] MediLexicon International Ltd. Medical news today: All about diabetes, 2013. URL http://www.medicalnewstoday.com/info/diabetes/.

[44] Phillip McClean. The chi-squared test, 2000. URL http://www.ndsu.edu/pubweb/~mcclean/plsc431/mendel/mendel4.htm.

[45] John H. McDonald. *Handbook of biological statistics*. 2009.

[46] Victor A. McKusick. Homocystinuria due to cystathionine beta-synthase deficiency, 2012. URL http://omim.org/entry/236200.

[47] N. Meinshausen and P Bühlmann. Stability selection. 2010.

[48] M. A. Menza, D. E. Robertson-Hoffman, and A. S. Bonapace. Parkinson's disease and anxiety: comorbidity with depression. *Biol Psychiatry*, 34(7):465–470, Oct 1993.

[49] Inc. Metabolon. Metabolon, solutions beyond the data..., 2013. URL http://www.metabolon.com/.

[50] Frank Mo, Lisa M. Pogany, Felix C K. Li, and Howard Morrison. Prevalence of diabetes and cardiovascular comorbidity in the canadian community health survey 2002-2003. *ScientificWorldJournal*, 6:96–105, 2006. doi: 10.1100/tsw.2006.13. URL http://dx.doi.org/10.1100/tsw.2006.13.

[51] A. Must, G. E. Dallal, and W. H. Dietz. Reference data for obesity: 85th and 95th percentiles of body mass index (wt/ht2) and triceps skinfold thickness. *Am J Clin Nutr*, 53(4):839–846, Apr 1991.

[52] PDR Network. Heart valve disease, 2013. URL http://www.pdrhealth.com/diseases/heart-valve-disease.

[53] NIDDK. Diabetic neuropathies: The nerve damage of diabetes, 2012. URL http://diabetes.niddk.nih.gov/dm/pubs/neuropathies/.

[54] William S Noble. How does multiple testing correction work? *Nat Biotech*, 27(12): 1135–1137, December 2009. ISSN 1087-0156. URL http://dx.doi.org/10.1038/nbt1209-1135.

[55] Dr. Guido Dallmann OA Dr. FASN Michael Rudnicki. Omics-technologien zur analyse von chronischen nierenerkrankungen.

[56] U.S. National Library of Medicine. Heart diseases, 2013. URL http://www.nlm.nih.gov/medlineplus/heartdiseases.html.

[57] Genome Sciences University of Washington. Lecture 10: Multiple testing. URL www.gs.washington.edu/academics/courses/akey/56008/lecture/.

[58] Nicole Ostrow. Vision loss increasing the u.s. as diabetes rates rise. 2012.

[59] Medline Plus. Blood thinners, 2013. URL http://www.nlm.nih.gov/medlineplus/bloodthinners.html.

[60] Andrew Pollack. Prescription drug to aid weight loss wins f.d.a. backing, 2012. URL http://www.downeyobesityreport.com/2012/06/.

[61] Donald J. Voet; Judith G. Voet; Charlotte W. Pratt. *"Lipids, Bilayers and Membranes". Principles of Biochemistry, Third edition*. 2008.

[62] Peter J. Raubenheimer, Moffat J. Nyirenda, and Brian R. Walker. A choline-deficient diet exacerbates fatty liver but attenuates insulin resistance and glucose intolerance in mice fed a high-fat diet. *Diabetes*, 55(7):2015–2020, Jul 2006. doi: $10.2337/\mathrm{db06\text{-}0097}$. URL http://dx.doi.org/10.2337/db06-0097.

[63] M.Mayhew R.Kousar. Blood vessels, 2011. URL http://www.diabetesinfo.org.au/webdata/images/Blood%20vessels%20harden%20and%20clot.jpg.

[64] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. pages 59–66, 1988.

[65] C. R. Roe, D. S. Millington, and D. A. Maltby. Identification of 3-methylglutarylcarnitine. a new diagnostic metabolite of 3-hydroxy-3-methylglutaryl-coenzyme a lyase deficiency. *J Clin Invest*, 77(4):1391–1394, Apr 1986. doi: $10.1172/\mathrm{JCI112446}$. URL http://dx.doi.org/10.1172/JCI112446.

[66] Ute Roessner. Sample prep for metabolomics experiments -plant tissues-, 2009.

[67] Nicolas Schauer, Dirk Steinhauser, Sergej Strelkov, Dietmar Schomburg, Gordon Allison, Thomas Moritz, Krister Lundgren, Ute Roessner-Tunali, Megan G. Forbes, Lothar Willmitzer, Alisdair R. Fernie, and Joachim Kopka. Gc-ms libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett*, 579(6):1332–1337,

Feb 2005. doi: 10.1016/j.febslet.2005.01.029. URL http://dx.doi.org/10.1016/j.febslet.2005.01.029.

[68] Clarke M. Schiweck, H. and G. Pollach. *Sugar. Ullmann's Encyclopedia of Industrial Chemistry*. 2007.

[69] Ursula Schwab, Anneli Törrönen, Leena Toppinen, Georg Alfthan, Markku Saarinen, Antti Aro, and Matti Uusitupa. Betaine supplementation decreases plasma homocysteine concentrations but does not affect body weight, body composition, or resting energy expenditure in human subjects. *Am J Clin Nutr*, 76(5):961–967, Nov 2002.

[70] M.D. Scott C.Litin. *Mayo Clinic: Family Health Book*, volume Fourth Edition. 2010.

[71] Tomoyoshi Soga, Yoshiaki Ohashi, Yuki Ueno, Hisako Naraoka, Masaru Tomita, and Takaaki Nishioka. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J Proteome Res*, 2(5):488–494, 2003.

[72] Healthwise Staff. Body fat distribution, 2010. URL http://www.uofmhealth.org/health-library/zm6365.

[73] Korbinian Strimmer. Notes: Graphical gaussian models for genome data, 2008. URL http://strimmerlab.org/notes/ggm.html.

[74] Karsten Suhre, Christa Meisinger, Angela Döring, Elisabeth Altmaier, Petra Belcredi, Christian Gieger, David Chang, Michael V. Milburn, Walter E. Gall, Klaus M. Weinberger, Hans-Werner Mewes, Martin Hrabé de Angelis, H-Erich Wichmann, Florian Kronenberg, Jerzy Adamski, and Thomas Illig. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One*, 5(11):e13953, 2010. doi: 10.1371/journal.pone.0013953. URL http://dx.doi.org/10.1371/journal.pone.0013953.

[75] SUPELCO. *Guide to Solid Phase Extraction*, 1998.

[76] The Kubrick Theme. Archive for the 'obesity' category, 2013. URL http://protonsforbreakfast.wordpress.com/category/obesity/.

[77] Julian F. Tyson. Flow injection analysis techniques for atomic-absorption spectrometry. a review. *Analyst*, 110(5):419–429, 1985. ISSN 0003-2654. URL http://dx.doi.org/10.1039/AN9851000419.

[78] Cambell R. Virtanen E. *Handbuch der tierischen Veredlung*. 1994.

[79] Zeneng Wang, Elizabeth Klipfell, Brian J. Bennett, Robert Koeth, Bruce S. Levison, Brandon Dugar, Ariel E. Feldstein, Earl B. Britt, Xiaoming Fu, Yoon-Mi Chung, Yuping Wu, Phil Schauer, Jonathan D. Smith, Hooman Allayee, W H Wilson Tang, Joseph A. DiDonato, Aldons J. Lusis, and Stanley L. Hazen. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, 472(7341):57–63, Apr 2011. doi: 10.1038/nature09922. URL http://dx.doi.org/10.1038/nature09922.

[80] P. Weigel, E. A. Weretilnyk, and A. D. Hanson. Betaine aldehyde oxidation by spinach chloroplasts. *Plant Physiol*, 82(3):753–759, Nov 1986.

[81] wikipedia. Obesity, 2013. URL http://en.wikipedia.org/wiki/Obesity#cite_note-HaslamJames-2.

[82] Wikipedia. Sphingomyelin, 2013. URL http://en.wikipedia.org/wiki/Sphingomyelin#cite_note-Voet-1.

[83] www.diabetes ratgeber.net. Diabetes-forschung, 2010. URL http://www.diabetes-ratgeber.net/forschung.

[84] www.diabetes ratgeber.net. Diabetes mellitus typ 2, 2013. URL http://www.diabetes-ratgeber.net/Diabetes-Typ-2.

[85] www.diabetes ratgeber.net. Die diagnose, 2013. URL http://www.diabetes-ratgeber.net/Diabetes-Typ-2/Die-Diagnose-11704_4.html.

[86] yWorks. yworks home, 2013. URL http://www.yworks.com/de/index.html.