



universität
wien

DIPLOMARBEIT

Titel der Diplomarbeit

„Bioinformatische Methoden in der ernährungswissenschaftlichen Forschung“

Verfasserin

Bernadette Binder

angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag.rer.nat.)

Wien, 2011

Studienkennzahl lt. Studienblatt:

A 474

Studienrichtung lt. Studienblatt:

Ernährungswissenschaften

Betreuerin / Betreuer:

Univ.-Prof. Dr. Jürgen König

Danksagung

Nachdem die Arbeit nun beendet ist möchte ich mich ganz herzlich bei den Menschen bedanken, die dies erst möglich gemacht haben und mich in unterschiedlicher Weise unterstützt haben.

Mein erster Dank gilt Herrn Univ.-Prof. Dr. Jürgen König für die Betreuung meiner Diplomarbeit. Durch Ihn war es mir möglich dieses faszinierende Thema zu bearbeiten.

Weiters möchte ich dem Institutsvorstand Univ.-Prof. Dr. Ibrahim Elmadfa, sowie allen Professoren des Instituts für Ernährungswissenschaften danken, dass sie mir das Wissen vermittelt haben diese Arbeit zu schreiben.

Ganz besonders möchte ich mich bei meiner Familie bedanken, die mir dieses Studium erst ermöglicht hat. Meine Familie war es auch, die mich immer wieder dazu motiviert hat, diese Arbeit zu Ende zu bringen.

DANKE

Inhaltsverzeichnis

1 Einleitung	1
2 „OMICS“- Technologien	3
2.1 Genomics.....	4
2.2 Transcriptomics.....	5
2.3 Proteomics	5
2.4 Metabolomics.....	6
2.5 Nutrigenomics	7
3 Microarray Analyse	9
3.1 Methode	9
3.2 Anwendungen	9
3.2.1 Lebensmittelsicherheit	10
3.2.2 Adipositasforschung.....	11
3.2.3 Versorgung mit Vitaminen und Spurenelementen.....	11
3.2.4 Proteinmicroarrays	12
3.2.5 Typ II Diabetes.....	13
3.3 Planung von Microarray-Experimenten.....	13
3.3.1 Samplesize oder Fallzahl-Planung.....	15
3.4 Durchführung und Auswertung	17
3.4.1 Vorbereitung der Chips, Hybridisierung und Scannen	17
3.4.2 Normalisierung.....	18
3.4.3 Log Transformation	19
3.4.4 Statistische Auswertung.....	20
3.5 Datenhandling.....	21

3.5.1 MIAME Standard.....	22
3.5.3 MAGE	22
3.5.4 MO	23
3.6 Das Transkriptom mathematisch betrachtet	23
3.7 Vor- und Nachteile von Microarrays.....	24
4 qRT-PCR	24
4.1 Methode.....	24
4.2 Anwendungen.....	25
4.2.1 Lebensmittelanalytik und Lebensmittelsicherheit.....	25
4.2.2 Morbus Crohn	25
4.2.3 Lebensmittelallergien	26
4.2.4 Diätetische Lebensmittel.....	26
4.2.5 Kalorische Restriktion und oxidativer Stress.....	27
4.3 Durchführung und Auswertung	27
4.3.1 Quantifizierung.....	27
4.3.2 Amplifikation und mathematische Modelle.....	28
4.3.3 Statistische Modelle	29
5 Proteom Analysen – Chromatographie, Elektrophorese und Massenspektroskopie	30
5.1 Methoden.....	30
5.2 Anwendung.....	31
5.2.1 Detektion von renalen Schädigungen	31
5.2.2 Peptidbiomarker.....	31
5.2.3 Enzymsuche	32
5.3 Durchführung	32
5.3.1 Tandemmassenspektrometrie	32

5.3.1.1 Identifizierung von Peptiden.....	34
5.3.1.2 Quantifizierung von Peptiden.....	35
5.3.1.3 Validierung der identifizierten Peptide.....	36
5.3.1.4 Validierung der identifizierten Proteine.....	36
5.3.1.5 Verschiedene Untersuchungsansätze	37
5.3.2 Gelelektrophorese und Massenspektroskopie	37
5.4 Schwierigkeiten der Massenspektroskopie	39
5.4.1 Quantifizierung.....	39
5.4.2 Zeitaufwand	39
5.4.3 Isoformen	39
5.4.3. Identifizierung.....	40
5.4.5 „Rauschen“	40
5.5 MIAPE.....	40
6 Metabolom – Analysen	41
6.1 Massenspektroskopie	44
6.1.1 Tandemmassenspektroskopie	45
6.1.2 Gaschromatographie und Massenspektrometrie	45
6.1.4 Flüssigkeitschromatographie und Massenspektrometrie	46
6.2 Kernspinresonanzspektroskopie	47
6.3 Datenhandling.....	47
6.4 Anwendung.....	48
6.4.1 Neugeborenen Screening	48
6.4.2 Metabolisches Syndrom.....	49
6.4.3 Lipide, Lipoproteine und Cholesterolmetabolite in Körperflüssigkeiten ...	49
7 Methoden zur Dimensionsreduktion und Datenreduktion	50
7.1 PCA	51

7.1.1 Hauptkomponentenanalyse mit bekannter Varianz-Kovarianz-Matrix Σ ..	53
7.1.2 Berechnung der Anzahl der Hauptkomponenten:	54
7.1.3 Darstellung im \mathbb{R}^2 :	55
7.1.4 Checkliste für die Durchführung einer Hauptkomponentenanalyse	56
7.2 Multidimensionale Skalierung (MES)	57
7.3 Faktorenanalyse	61
7.4 Clusteranalyse	62
7.4.1 Hierarchische Clusteranalyse:	62
7.4.2 Partitionierte Verfahren:	64
7.4.3 Interpretation von Dendogrammen	65
Die Interpretation eines Dendogramms erfolgt nach folgendem Prinzip:	65
7.4.4 Anwendung von Clusterbildung in der Praxis	66
7.5 Partial Least Squares Analyse	67
7.5.1 Anwendung des PLS in naturwissenschaftlichen Studien.....	67
8 Schlussbetrachtungen	69
9 Quellenverzeichnis:.....	73
10 Zusammenfassung	81
11 Abstract.....	83
12 Curriculum Vitae	85

Abbildungsverzeichnis

Abb. 1	"Omics" in der Ernährungsforschung	3
Abb. 2	„Omics“-Kaskade	4
Abb. 3	Stellung der Proteomics in der Bioinformatik	6
Abb. 4	Einflüsse auf das Metabolom	7
Abb. 5	Zusammenspiel von Ernährung und Omics	8
Abb. 6	Schema des Versuchsablaufs bei Microarray-Studien	14
Abb. 7	Zusammenhang zwischen Sample Size und FDR	16
Abb. 8	Schema einer Microarray Analyse	17
Abb. 9	Quantifizierung in der q-RT-PCR	28
Abb. 10	Schema einer Proteomanalyse	30
Abb. 11	Peptide Mass Fingerprinting und Peptide Fragmentation Fingerprinting	33
Abb. 12	Schema eine Proteom MS/MS	34
Abb. 13	Verknüpfung von Elektrophorese und Massenspektrometrie	37
Abb. 14	Metabolomics Workflow	43
Abb. 15	Weitere Geräte-Kombinationen für Massenspektrometrie	44
Abb. 16	Spannender Baum	56
Abb. 17	Minimal spannender Baum	57
Abb. 18	Verschiedene Dendogramme	64
Abb. 19	Complete Linkage Dendogramm	65
Abb. 20	Dendogramm nach PCA	66

Tabellenverzeichnis

Tab. 1	Falsch positiv und Falsch negativ	21
--------	-----------------------------------	----

1 Einleitung

Die Bioinformatik stellt mit Sicherheit eine der heterogensten naturwissenschaftlichen Disziplinen dar. Sie beinhaltet Aspekte der Informatik, der Molekularbiologie, sowie der Physiologie. Nicht umsonst wird die Bioinformatik als interdisziplinäre Naturwissenschaft bezeichnet. Sie gliedert sich in verschiedene Teilaspekte, wie die Sequenzanalyse, die Strukturbiologie und die integrative Bioinformatik.

Diese wissenschaftliche Disziplin hat in den letzten Jahren auch verstärkt in die ernährungswissenschaftliche Forschung Einzug gehalten, da man erkannt hat, dass es einen komplexen Zusammenhang zwischen der Gesundheit eines Menschen und seinem Genom gibt. Man hat festgestellt, dass jegliche Interaktion zwischen dem menschlichen Körper und seiner Ernährung physiologisch komplex abläuft. Das heißt, dass der Genotyp den Stoffwechselweg der Nährstoffe, die Genexpression, sowie die Bildung von Proteinen eines Individuums bestimmt, und somit letztlich auch für dessen Gesundheitszustand verantwortlich ist. Diese komplexen Wechselbeziehungen sind dafür verantwortlich, dass nicht die Gesamtheit der menschlichen Gene, sondern der Verlauf der Expression und Translation und der damit verbundene Polymorphismus für das unterschiedliche Auftreten von Krankheiten, für die unterschiedliche Verträglichkeit von Medikamenten, für die Verstoffwechslung von Nahrungsbestandteilen und Toxinen ist .

Die Bioinformatik ist wohl die am raschesten wachsende Technologie in sämtlichen Forschungsschwerpunkten der Lebenswissenschaften. Man nimmt an, dass ihr zukünftig eine Vorreiterrolle in der Ernährungs- und Lebensmittelforschung inne sein wird. Durch die Fülle an Daten, die sich durch bioinformatische Untersuchungsmethoden ergeben, vor allem in Folge eines enormen Output trotz zumeist verhältnismäßig kleiner Stichprobe, ist gerade in integrative Bioinformatik einer der am raschesten wachsenden Zweige dieser Wissen-

schaft. Darunter versteht man Speicherung, Organisation und Archivierung der aus verschiedenen Analysen gewonnen Daten in Form von Datenbanken, sowie den Abgleich neu gewonnener Analysedaten mit bestehenden Datenbanken.

Das Ziel dieser Arbeit war es die einzelnen „Omics“-Technologien aus ernährungswissenschaftlicher Sicht zu beleuchten, deren Untersuchungsmethoden zu erfassen und ihren Einsatz in ernährungswissenschaftlichen Studien zu zeigen. Diese Arbeit hat es sich zum Ziel gesetzt, die für die Ernährungswissenschaft relevanten Aspekte der bioinformatischen Forschung zu beleuchten, verschiedene Methoden zu erklären und deren Vor- und Nachteile aufzuzeigen. Es soll gezeigt werden, für welche Bereiche der Ernährungs- und Lebensmittelforschung welche bioinformatischen Untersuchungsmethoden geeignet sind und welche der Analysen welcher Omics-Technologie zuzuordnen ist. Um die Komplexität und Problematik der Datenfülle, welche mit den verschiedenen Analysen einhergeht, zu demonstrieren, werden im letzten Kapitel Maßnahmen zur Datenreduktion demonstriert.

2 „OMICS“- Technologien

Die „Omics“ (Genomics, Transcriptomics, Proteomics und Metabolomics) bieten großes Potential, durch die mit ihnen assoziierten bioinformatischen Methoden den wichtigen wissenschaftlichen Zielsetzungen der Ernährungsforschung, der Rolle der Ernährung bei der Stoffwechselregulation zu eruieren, gerecht zu werden. [ZHANG et al., 2008]

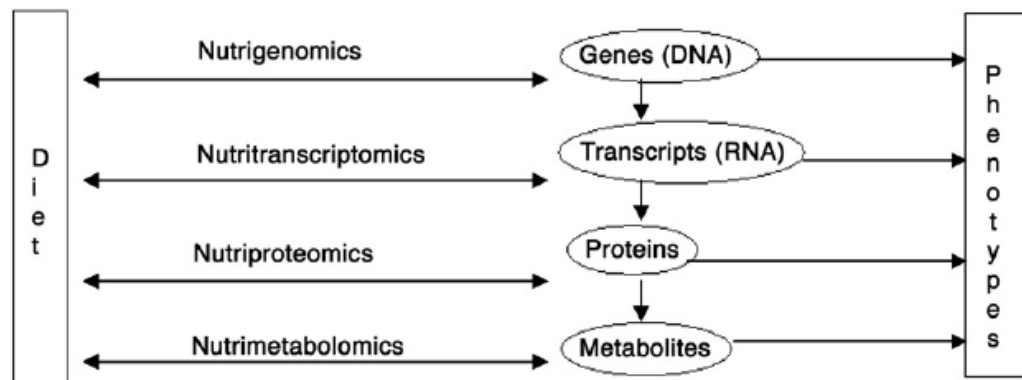


Abbildung 1 "Omics in der Ernährungsforschung [ZHANG et al., 2008]

Die Palette der Omics ist noch weit größer, aber nur wenige sind im Bereich der Ernährungsforschung wirklich gebräuchlich. Die Wortendung „-omics“ oder -omik im deutschen Sprachgebrauch bezeichnen die Untersuchung eines Gebiets, während die Endung –om die Gesamtheit aller Stoffe oder Substanzen dieses Untersuchungsgebiets beschreibt. So wird die Untersuchung des Metaboloms, welches per Definition alle Metabolite umfasst, als Metabolomik bezeichnet. [DETTMER, 2008]

Durch den Einsatz bioinformatischer Methoden in der ernährungswissenschaftlichen Forschung entstanden neue Begriffe und Disziplinen wie die Nutrigenomik, die die Begriffe des Genoms und der Ernährung nicht nur im Wort vereint, sondern deren Wechselwirkung zueinander auch als Erklärung des Begriffes an sich verstanden werden sollte. Auch Transkriptomics, Proteomics, Metabolomics und verwandte Begriffskombinationen sind aus der ernährungswissenschaftlichen Forschung nicht mehr wegzudenken und dienen der Erklärung der neuen Erkenntnisse von komplexen Zusammenhängen. [ZEISEL, 2007]

2.1 Genomics

Die Genomik befasst sich mit der Genomanalyse. Unter Genom ist die Gesamtheit der Erbinformation einer Zelle zu verstehen. Die Genomik stellt den Anfang der omics-Kaskade dar. [BÖCKENHAUER und BONGARTZ, 2003]

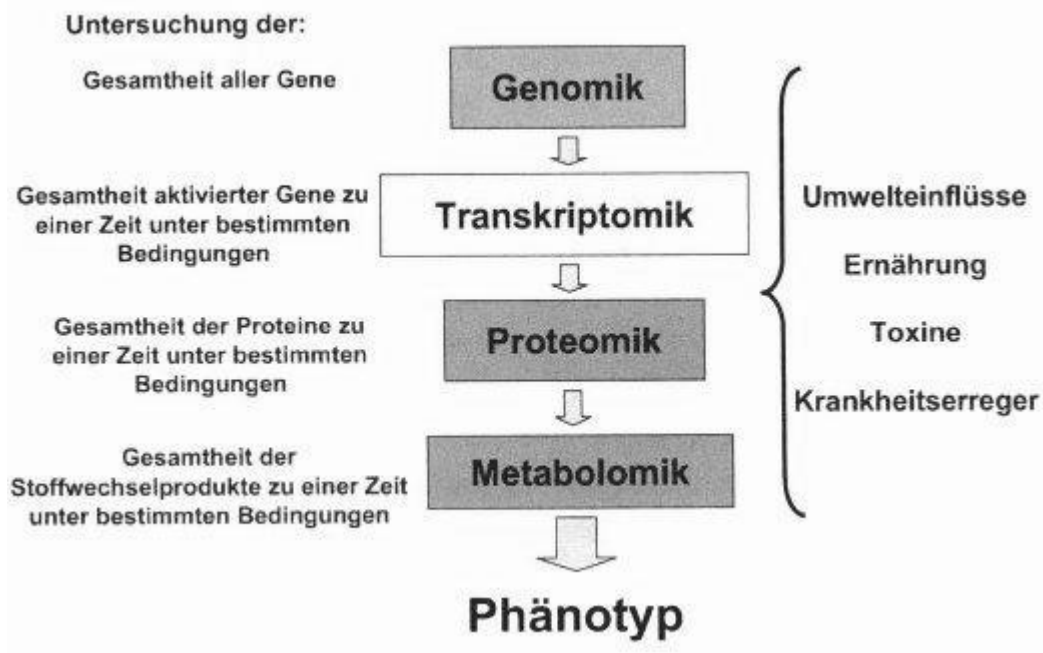


Abbildung 2 "Omics"-Kaskade [DETTMER, 2008]

2.2 Transcriptomics

Abhängig von den Einflüssen aus Lebensbedingungen, Stress, Ernährung und Umwelteinflüssen wird die im Genom gespeicherte Information abgeschrieben. Die Gesamtheit der daraus entstandene mRNA, d.h. alle Transkripte, wird als Transkriptom bezeichnet. [JAHN und SCHOMBURG, 2009]

2.3 Proteomics

Das Proteom wird als das Protein-Pendant zum Genom definiert. Es stellt die Gesamtheit der Proteine zu einem bestimmten Zeitpunkt und zu bestimmten Bedingungen dar. Die Proteomik verfolgt das Ziel Proteine zu untersuchen, um Informationen über ihre Struktur, ihre Funktion und ihre Rolle im Stoffwechsel zu gewinnen. Das Proteom ist dynamisch und daher Veränderungen unterworfen, welche von Ernährung, Stress und Umweltfaktoren beeinflusst werden. [PATTERSON und AEBERSOLD, 2003]

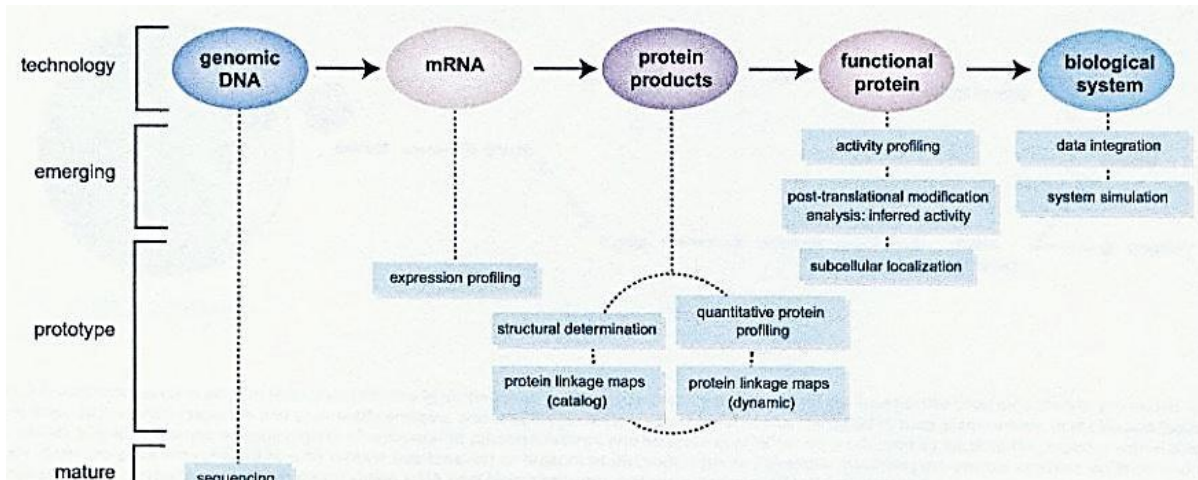


Abbildung 3 Stellung der Proteomics in der Bioinformatik [PATTERSON und AEBERSOLD, 2003]

2.4 Metabolomics

Die Metabolomik versucht die Metabolite, d.h. Substanzen die an Stoffwechselfvorgängen als Produkte, Co-Faktoren, Zwischenprodukte etc. beteiligt sind, zu quantifizieren. Das Metabolom, also die Gesamtheit der Metabolite in einem Organismus, umfasst Substanzen verschiedener Stoffklassen zu Tausenden. Daher sind deren chemische und physikalische Eigenschaften sehr heterogen. Genauso wie ihre Eigenschaften so hat auch der Konzentrationsbereich indem sie vorliegen eine sehr große Spannweite. Das Metabolom wird durch Faktoren wie Ernährung, Stress und Umwelteinflüsse beeinflusst, daher spricht man von einem dynamischen System, welches zeitlichen Schwankungen unterworfen ist. Jede Probennahme und deren Analytik des Metaboloms ist daher als Momentaufnahme zu betrachten. [DETTMER, 2008]

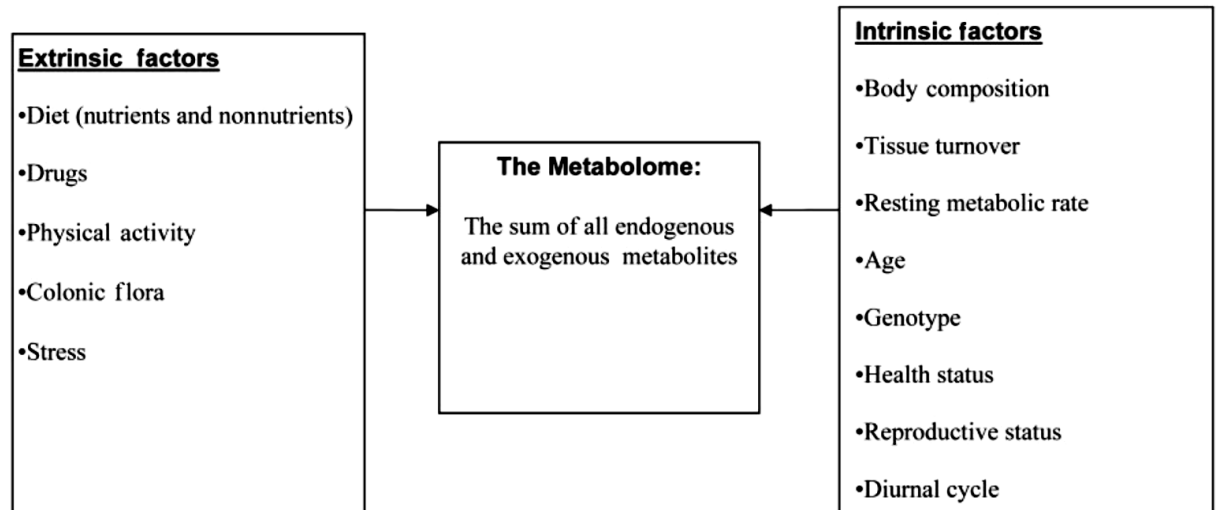


Abbildung 4 Einflüsse auf das Metabolom [GIBNEY et al., 2005]

Durch Genexpression und durch Proteomanalysen lässt sich das Potential für physiologische Veränderungen erforschen, aber erst die Messung der Konzentration verschiedener Metabolite in Körperflüssigkeiten, Zellen und Geweben stellen den Endpunkt der physiologischen Regulation dar. [ZHANG et al., 2008]

2.5 Nutrigenomics

Die Ernährungsforschung verwendet die oben genannten Omics-Technologien. Deren Zusammenfassung wird in ernährungswissenschaftlichem Kontext als Nutrigenomik bezeichnet. [DANIEL et al., 2009]

Die Nutrigenomik stützt sich auf Entwicklungen verschiedener naturwissenschaftlicher Disziplinen, die sich über Jahrzehnte stetig weiter entwickelt haben. Sie sieht sich als Schmelztiegel von Molekularbiologie, Genetik, Medizin, Informatik und Chemie. [VERGÈRES und SAGAYA, 2007]

Nahrungsinhaltsstoffe, Mikro- sowie Makronährstoffe, stellen eine Beeinflussung für alle biologischen Prozesse eines Organismus dar. [DANIEL et al., 2009]

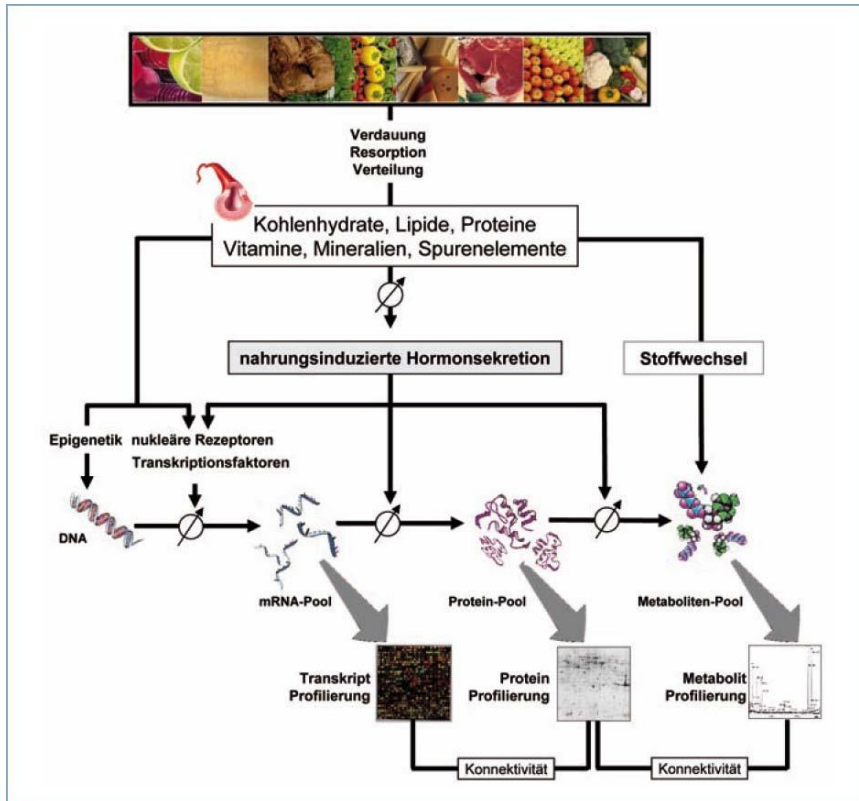


Abbildung 5 Zusammenspiel von Ernährung und Omics [DANIEL et al., 2009]

3 Microarray Analyse

3.1 Methode

Microarray Analysen werden zu den sogenannten *High-Troughput-Screening* oder HTS-Methoden gerechnet. Die daraus gewonnenen Daten sind als Momentaufnahmen der Genexpressionen in einem Organismus zu verstehen, welche gleichzeitig ablaufen. Microarray-Analysen sind daher eng mit der Erforschung von Transkriptomen verknüpft. So kann man durch Arrays untersuchen welche Gene unter welchen Voraussetzungen zeitgleich aktiv sind. Bei diesen Voraussetzungen kann es sich z. B. um ein bestimmtes Krankheitsbild handeln, dem man durch Microarray-Analysen versucht, ein Aktivitätsmuster zuzuordnen. [HÜTT und DEHNERT, 2006]

DNA -Microarrays werden am häufigsten zur Transkriptomanalyse angewandt, da es sich um eine Hochdurchsatzmethode handelt, bei der bis zu 50.000 Transkripte gleichzeitig untersucht werden können. [ZHANG et al., 2008]

Bei Microarrayanalysen ist zu beachten, dass es sich bei dieser Methode um eine Chip-Technologie handelt, genauer gesagt um RNA-Chips. Man unterscheidet zwischen 2 Arten von Chips, abhängig davon wie viele RNA Proben auf den Chip passen, 1 oder 2 Proben. [FU et al., 2010]

3.2 Anwendungen

Die Anwendung von Microarrays ist sehr vielseitig, lässt sich aber wie folgt charakterisieren und einteilen: [VICTOR et al., 2005]

- Analyse von Gen- oder Proteinexpression von Proben unterschiedlicher Populationen oder Versuchsgruppen
- Durchführung einer Clusteranalyse
- Einteilung verschiedener Krankheitsenditäten (Krankheitsbilder)

Der Zweck eines Microarrays stellt also das Auffinden von Unterschieden zwischen verschiedenen Gruppen, das Erkennen von Gruppen (Cluster), sowie der Einsatz zur Diagnose dar. [VICTOR et al., 2005]

3.2.1 Lebensmittelsicherheit

Durch die DNA-Sequenzierung verschiedener pathogener Keime welche für die Lebensmittelsicherheit und Ernährung von Bedeutung sind, wie *Escherichia coli* O157:H7, *Campylobacter jejuni* und *Listeria monocytogenes*, war es möglich ihre Physiologie und ihre pathogene Wirkung besser zu erfassen. Gerade um Unterschiede in der Pathogenität verschiedener Bakterien, ihrer Urformen und Mutationen, zu erfassen, sind Microarrays eine Möglichkeit zur Erforschung der Unterschiede in der bakteriellen Genexpression. [FRATAMICO et al., 2008]

Molekulare Diagnostik gewinnt zunehmend an Bedeutung im Bereich der Lebensmittelsicherheit, besonders im Zusammenhang mit gentechnisch veränderten Lebensmitteln und neuartigen Produkten, sogenannten Nutraceuticals. Diese stellen einen Brückenschlag zwischen Lebensmitteln und pharmazeutischen Produkten dar. [LIU-SRATTON et al., 2004]

Auch in Bezug auf mikrobielle Pathogene wird die cDNA-Microarray- Analyse als zukunftsweisend erachtet, nicht nur bezüglich gentechnisch veränderten Lebensmitteln. Dafür muss aber eine benutzerfreundlichere Anwendung ermöglicht werden, damit Microarrays zur Standardmethode im Bereich Lebensmittelsicherheit werden. [ROY und SEN, 2006]

Es wird vermutet, dass sich die DNA-Microarray-Technologie sich im Bereich der Lebensmittelsicherheit vor allem für die schnelle und genaue Qualitätskontrolle bei häufig mit Salmonellen belasteten Lebensmitteln eignen könnte. [ROY und SEN, 2006]

Als Diagnosetool um potentiell pathogene Keime aufzufinden und zu charakterisieren bieten DANN-Microarrays die Möglichkeit einer relativ schnellen und vor allem simultanen Durchführung, welche es ermöglicht viele Proben gleichzeitig zu analysieren. [FRATAMICO et al., 2008]

3.2.2 Adipositasforschung

Im Rahmen einer kürzlich veröffentlichten Studie wurde insulinresistentes Fett- und Lebergewebe von adipösen Nagern untersucht und so durch die Microarray-Technologie Informationen über die Veränderungen in der Genexpression von Fettleibigen gewonnen, sowie die damit verbundenen metabolischen Erkrankungen. Dabei wurden die Microarray-Daten mit den Ergebnissen von 28 weiteren Studien, welche ebenfalls die Genexpression von schlanken und fettleibigen Tieren mittels dieser Methode untersuchten verglichen. Nur 6 dieser Vergleichsstudien verwendeten menschliches Gewebe. Es konnte gezeigt werden, dass 88 verschiedene Gene bei fettleibigen Organismen anders exprimiert werden als bei schlanken Organismen. [KIM und PARK, 2010]

3.2.3 Versorgung mit Vitaminen und Spurenelementen

In einer anderen Studie wurde an Ratten mittels Microarray ein Kupfer- und Eisenmangel durch Nahrungsrestriktion untersucht. Die Downregulation der Gene, die für die mitochondriale und peroxisomale β -Oxidation von Fettsäuren verantwortlich sind, sowie die Überexpression der Gene, die für den Transport von Cholesterin in Blut verantwortlich sind, konnte festgestellt werden. [TOSCO et al., 2010]

Molekulare Biomarker ermöglichen durch Untersuchungen mittels Microarrays einen Rückschluss auf den Ernährungszustand. Die dabei verwendete mRNA und die Tatsache, dass Erythrozyten eine Lebensdauer von durchschnittlich 120 Tagen bei Menschen haben ermöglichen es so, auf den Ernährungszustand über längere Zeit zu schließen, anstatt nur über die aktuelle Nährstoffaufnahme Auskunft zu geben.

Auf diese Weise wurde bereits der Selen-, sowie der Zinkstatus untersucht. [SUDE, 2010]

3.2.4 Proteinmicroarrays

Obwohl Massenspektroskopie die Hauptuntersuchungsmethode in der Proteomanalytik ist, werden auch Chip-methoden wie Microarray auch zur Erforschung von Protein Expression angewandt. [PATTERSON und AEBERSOLD, 2003]

Microarray-Analysen ermöglichen zwar nur eine Momentaufnahme des Proteoms. Durch ständige Veränderungen der Proteine hinsichtlich Struktur, Funktion, Lokalisation und Turnover-Raten ist diese Untersuchungsmethode mit Schwierigkeiten und Einschränkungen behaftet, denen gegenüber aber die Möglichkeit steht, tausende Proben gleichzeitig zu analysieren. [BERREDE et al., 2010]

Neue Techniken ermöglichen auch eine gleichzeitige Gen- und Proteinexpressionsanalyse derselben Probe. Bei den dafür eingesetzten Arrays handelt es sich um Microbead-Flüssigarrays. [PORSCHESKI und STEINERT, 2006]

3.2.5 Typ II Diabetes

Durch das Zusammenspiel von Daten aus histologischen Befunden, klinischer Chemie und Microarray-Analysen versucht man im Tierexperiment mit Mäusen besser Angriffspunkte für die medikamentöse Behandlung von Diabetikern und neue Diagnosemöglichkeiten zu finden. Durch die Untersuchung der veränderten Genexpression konnten verschiedene Gene bestimmt werden, welche sich als Biomarker für eine frühzeitige Diabetesdiagnose eignen könnten. [BOYCE et al., 2005]

3.3 Planung von Microarray-Experimenten

Die Planung jedes Microarray-Experiments beruht darauf, dass bestimmte aus biometrischer Sicht wichtige Aspekte in den Planungsprozess miteinfließen. So ist es wichtig, sich die Fragestellung vor Augen zu halten um das optimale Studiendesign und die richtige Fallzahl (Samplesize) zu wählen. Es ist in jedem Fall eine Datenreduktion in Betracht zu ziehen um die Auswertung zu erleichtern. Maßnahmen zur Fehlerminimierung sind anzuwenden um eine sinnvolle Auswertung zu ermöglichen, welche für aussagekräftige Ergebnisse notwendig ist. Statistische Verfahren müssen der Fragestellung angepasst sein und gegebenenfalls einen Vergleich zwischen einzelnen Gruppen oder eine Klassifikation zulassen. Die erhaltenen Ergebnisse sind zu interpretieren. Bei der Versuchsplanung ist aber in jedem Fall zu beachten, dass es kein Standardschema gibt, sondern jedes Experiment von der individuellen Fragestellung abhängt. Jedes Versuchsablaufschemata (Abb. 6) zeigt immer nur einen möglichen Versuchsablauf. Einzelne Schritte können je nach Fragestellung auch entfallen, oder ersetzt werden. [VICTOR et al., 2005]

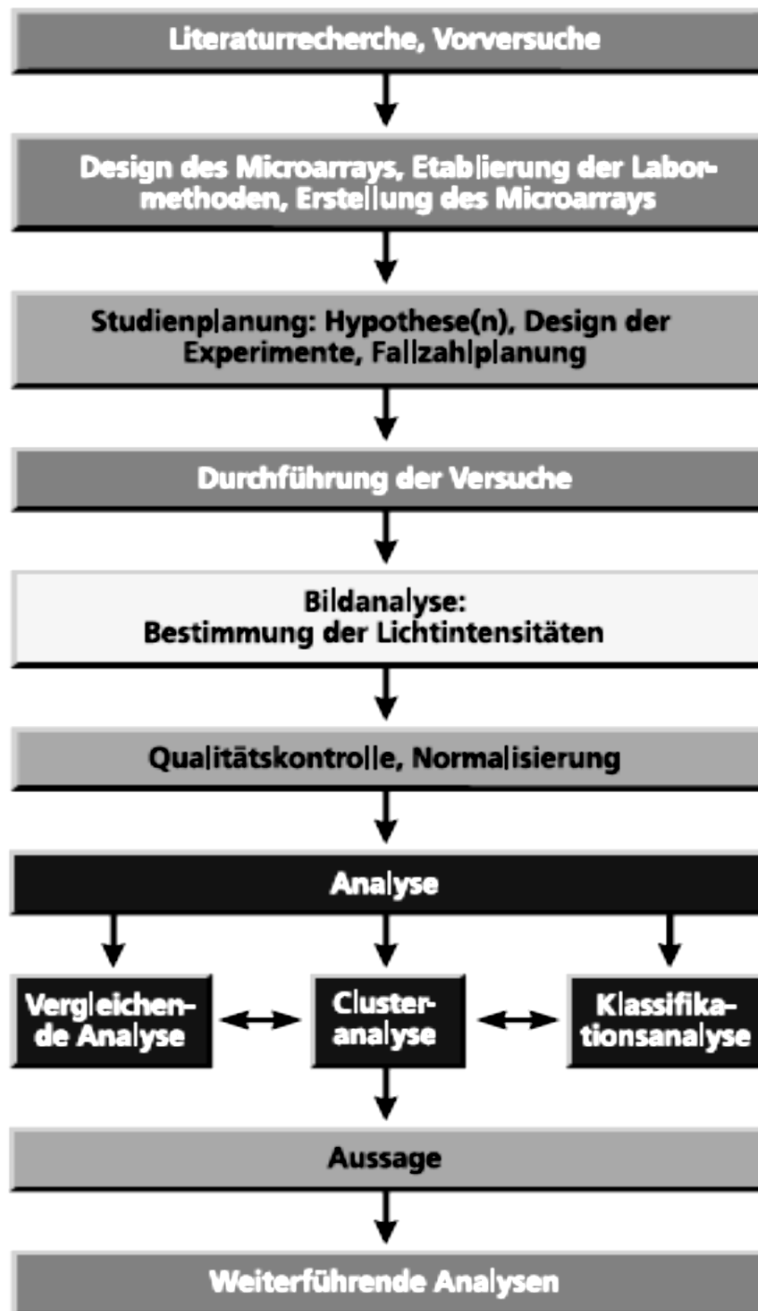


Abbildung 6 Schema des Versuchsablaufs bei Microarray-Studien
[VICTOR et al., 2005]

3.3.1 Samplesize oder Fallzahl-Planung

Generell ist anzumerken, dass eine höhere Fallzahl zu besseren Ergebnissen führt. Da dies aber eine Kostensteigerung und einen nicht unerheblichen Mehraufwand verursacht, so sollte immer eine Fallzahlberechnung, allein schon wegen der wissenschaftlichen Aussagekraft erfolgen. Zwei wichtige Faktoren sind bei der Festlegung der Fallzahl bei Microarray Studien zu beachten, einerseits die Power, die für die Aussagekraft der erhaltenen Ergebnisse entscheidend ist, sowie die FDR (false discovery rate), die Anzahl der falsch positiven und falsch negativen Ergebnisse. Entscheidend ist das Setzen des richtigen Cut-off. Eine Verbesserung der FDR führt immer auch zu einem Verlust der Power. Daher ist ein FDR von 0 nicht möglich. [JORSTAD et al, 2007]

Bei der Untersuchung der Genexpression ist zu beachten, dass geringe Unterschiede in der Expression nur bei großen Fallzahlen gefunden werden können, aber große Unterschiede in der Expression bereits bei geringen Unterschieden. Die Samplesize für Humanstudien liegt im Allgemeinen bei 5 bis 10 Probanden pro Gruppe, während das Minimum für Invitrostudien bei drei liegt. Je höher die Fallzahl desto besser die Ergebnisse. [FU et al., 2010]

Eine fehlende Fallzahlplanung wird wissenschaftlich mit Skepsis betrachtet und spricht im Allgemeinen nicht für eine qualitativ hochwertige Studie, egal ob diese im ernährungswissenschaftlichen, medizinischen oder anderen naturwissenschaftlichen Bereich geschieht. Bei der Planung ist die Fragestellung zu berücksichtigen und die Fallzahl so zu wählen um signifikante Ergebnisse zu ermöglichen. [RÖHRIG et al.,2010]

Wie enorm die Unterschiede im Einfluss der Fallzahlen auf die falsch positiv und falsch negativ Rate wird in Abb. 7 verdeutlicht. Am Beispiel von 5000 unregulierten, 3000 „down“-regulierten und 2000 „up“-regulierten Genen bei unterschiedlicher Fallzahl. [JORSTAD et al, 2007]

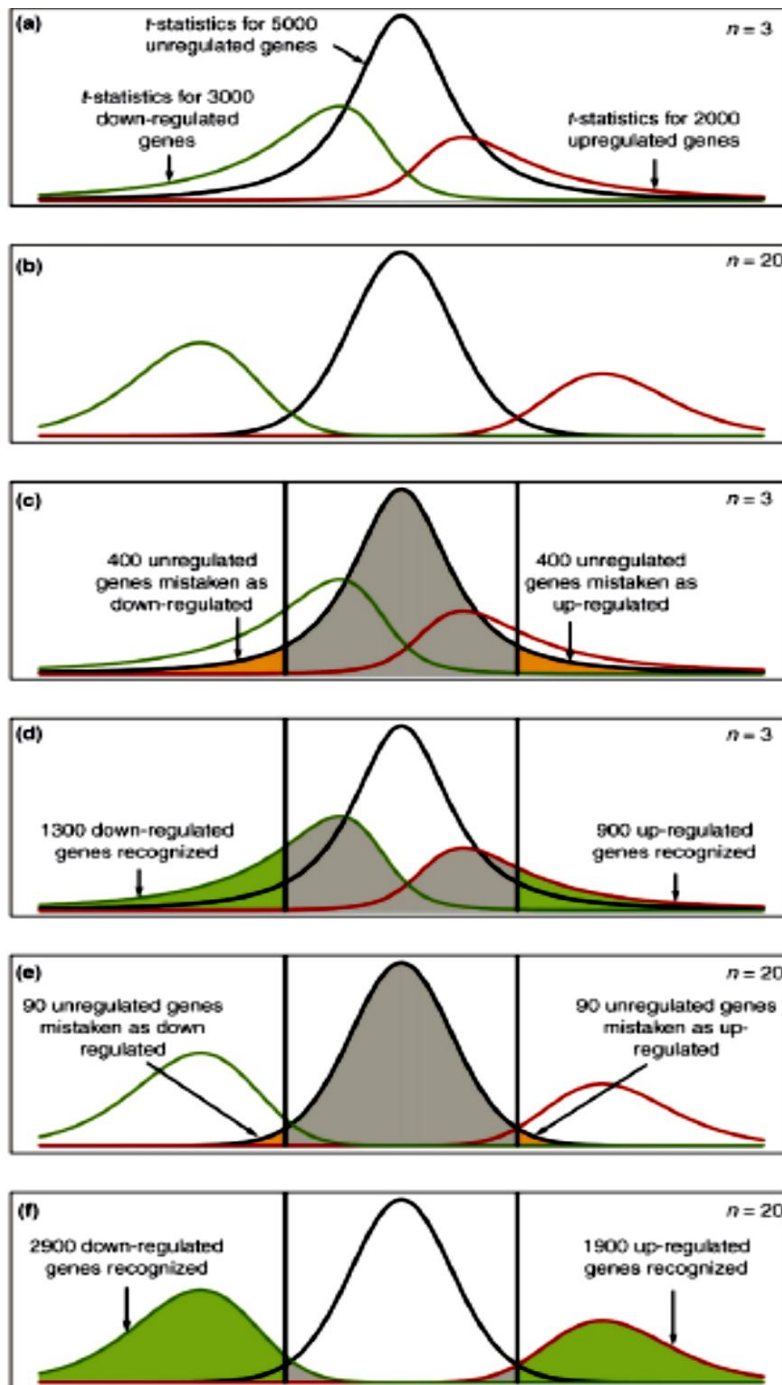


Abbildung 7 Zusammenhang zwischen Sample Size und FDR [JORSTAD et al., 2007]

3.4 Durchführung und Auswertung

Auf das Microarray-„Experiment“ folgen verschiedene Schritte wie die Normalisierung, Datenreduktion und Log-Transformation. Dies ist die Voraussetzung für die eigentliche statistische Auswertung. [FU et al., 2010]

3.4.1 Vorbereitung der Chips, Hybridisierung und Scannen

Auf einer Immobilisationsmatrix werden Nukleinsäuren an genau definierten Positionen zu den zu untersuchenden Genen mit bekannter Sequenz angelagert. Das Hybridisieren erfolgt mit fluoreszenzmarkierten Target-Nukleinsäuren aus zu untersuchendem Gewebe, wodurch bei späterem Scannen durch Fluoreszenz-Scanner der Ort der Hybridisierung auf dem Chip identifiziert wird, es entsteht ein charakteristisches Bild. [SCHOBER, 2002]

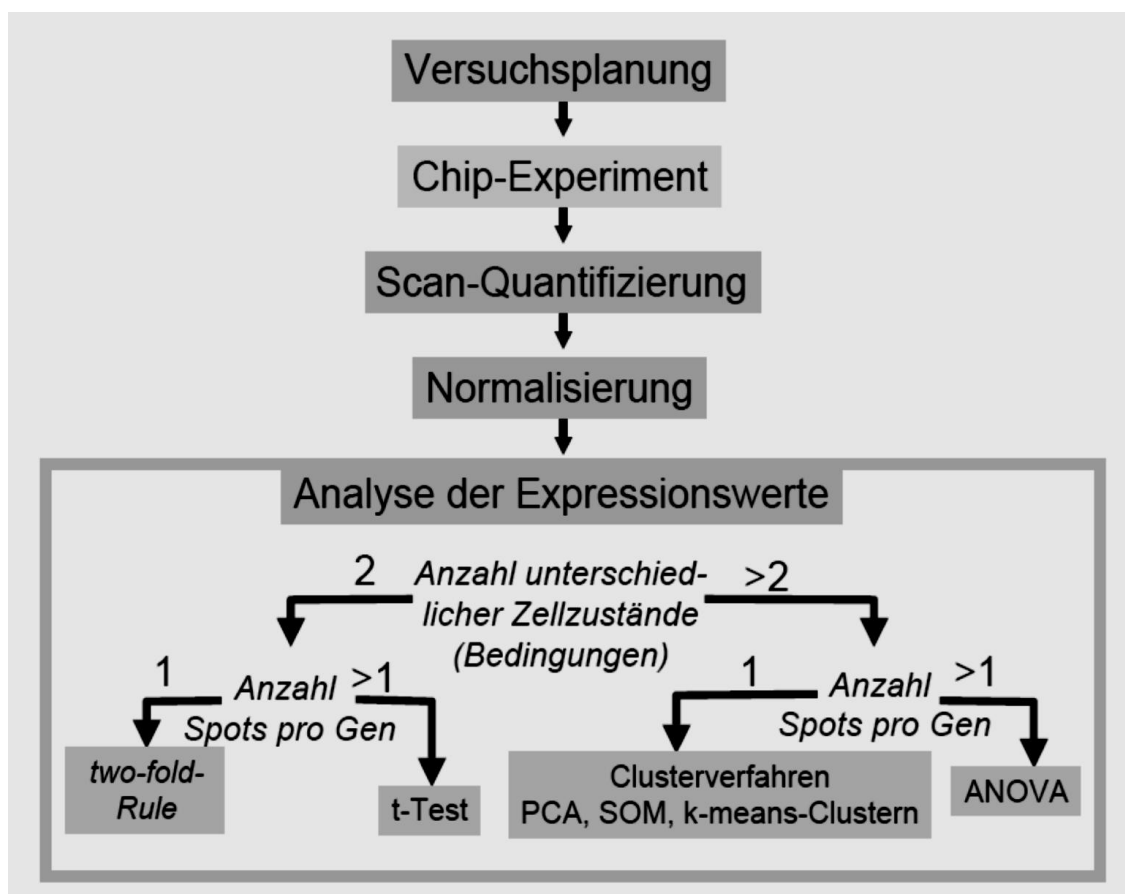


Abbildung 8 Schema einer Microarray Analyse [REPENNING et al., 2009]

3.4.2 Normalisierung

Die Normalisierung ist eine computergestützte Methode zur Korrektur von systematischen Fehlern. Diese Datentransformation ermöglicht einen Vergleich von Chips unterschiedlicher Einrichtungen, sowie unterschiedlicher Scanning-Parameter. Sie bildet die Voraussetzung für einen anschließenden Datenvergleich und erfolgt nach vorangegangener Hybridisation und Scannen der Arrays. [FU et al., 2010]

Von allen Teilen der Auswertung ist die Normalisierung der komplexeste Teil. Mit ihr werden Unterschiede in der Technik sowie Unterschiede zwischen verschiedenen Chips ausgeglichen. Der Sinn dahinter ist es einen tatsächlichen biologischen Unterschied sichtbar zu machen in dem Intensitätsunterschiede anderer Herkunft eliminiert. [REPENNIG et al., 2009]

Die Notwendigkeit einer Normalisierung ergibt sich aus der Tatsache heraus, dass gemessene Intensitäten keine absolute Größe sind, sondern versuchsbedingten Schwankungen unterliegt. Neben oben erwähntem Programm gibt es natürlich noch andere Möglichkeiten der Normalisierung. So besteht die Möglichkeit einer Nullkontrolle über die Hybridisierung der Housekeeping Gene, deren Intensität von denen zu untersuchenden Gene abgezogen wird, ähnlich dem Leerwert bei spektroskopischen Untersuchungen. Andererseits besteht auch die Möglichkeit alle gemessenen Intensitäten zur Normalisierung heranzuziehen. Dies geschieht durch die einen Abzug der mittleren Intensität. Im Gegensatz zur Verwendung von Housekeeping Genen, die gegenüber den zu untersuchenden deutlich in der Minderheit sind fließen hier sehr viel mehr Intensitätswerte ein. [VICTOR et al., 2005]

Bei der Normalisierung werden im Wesentlichen zwei Strategien unterschieden, die within-Normalisierung und die between-Normalisierung. Die within-Normalisierung dient dem Ausgleich von Microarray spezifischen Unterschieden, die beispielsweise durch unterschiedliche Chargen entstehen oder Farb-

stoffbedingt sind. Die within-Normalisierung kann durch verschiedene Normalisierungsarten durchgeführt werden: globale Normalisierung, intensitätsabhängige lineare oder nicht-lineare Normalisierung. Alle Methoden der within-Normalisierung stützen sich auf die Annahme, dass die Expression aller Gene im Mittel gleich ist. Wenn hingegen angenommen wird, dass die Expression sehr verschieden ist, um gezielt Gene selektieren zu können kann diese Art der Normalisierung nicht angewandt werden. Die between-Normalisierung setzt an den schon bereinigten Daten an. Diese wird häufig durch eine Datenzentrierung durchgeführt. Die between-Normalisierung dient dem Ausgleich von Abweichungen zwischen den Microarrays. [REPENNIG et al., 2009]

Es gibt verschiedene Computerprogramme um eine Normalisierung vollautomatisch durchzuführen. Häufig eingesetzte Methoden sind MAS 5 (Microarray Suite 5.0) von Affymetrix und gcRMA von Bioconductor, dabei wird aus den gescannten Bildern mittels der Computersoftware die Expressionssignale und der detection call errechnet. Wird ein Gen exprimiert so ist der detection call „P“ (present), wird es nicht exprimiert lautet er „A“ (absent). Im Falle dass die Genexpression nur schwach ist, ist der detection call „M“ (marginal). [FU et al., 2010]

3.4.3 Log Transformation

Es handelt sich dabei, sofern durchgeführt, um den letzten Schritt vor der statistischen Auswertung. Eine Transformation ist nicht zwingend erforderlich, kann vorteilhaft sein. Bei 1-Proben-Chips bringt eine Log Transformation üblicherweise keinen nennenswerten Unterschied. [FU et al., 2010]

Bei der Log Transformation bedeutet, dass exponentielle Beziehungen zu linearen umgewandelt werden. In weiterer Folge führt dies zu einer Begradigung einer schiefen Verteilung und zu kleineren Varianzen. Es gibt auch noch andere Transformationsarten wie Standardisierung (Z-Transformation). Dabei handelt es zu einer Transformation zur Standardnormalverteilung. [LEE, 2010]

3.4.4 Statistische Auswertung

Bei der statistischen Auswertung kommen verschiedene Modelle und Testverfahren zum Einsatz. Am häufigsten kommen der Zweistichproben-t-Test, die einfaktorielle und die zweifaktorielle Anova zum Einsatz, wobei es sich um die Varianzanalyse handelt. Dabei geht wird eine Normalverteilung der statistischen Fehler angenommen, die nicht immer korrekt ist, aber sich bei kleinen Stichprobenumfängen als ideal erwiesen hat. Zumeist wird die Abhängigkeit von Stichproben übersehen, da hierfür anderen Testverfahren angewendet werden müssen. [FU et al., 2010]

Die eigentliche Herausforderung jeder Statistischen Auswertung besteht darin, dass die Vielzahl an gemessenen Expressionen aus einer geringen Anzahl an Proben stammt. Oftmals wird beim Gruppenvergleich der Expressionsstärke der „fold change“ betrachtet. Problematisch dabei ist, dass häufig höhere Intensitätswerte und Varianzen beobachtet werden. Die Verwendung statistischer Tests auf unterschiede wie. Z.B. der t-Test, ist daher sinnvoll und notwendig. [VICTOR et al., 2005]

Das „fold change“ Verfahren nutzt die Tatsache, dass der Unterschied des Logarithmus eines Wertes zu seinem Kehrwert ausschließlich im Vorzeichen zu finden ist. Bei diesem Verfahren werden zwei Zustände untersucht. Ein einzelner Spot steht für ein Gen. Diese Art der Betrachtung ist wie oben erwähnt recht fehleranfällig und daher nur als Notlösung zu betrachten. Für zuverlässige Aussagen sollten mehrere Spots pro Gen auf einem Chip einkalkuliert werden, was natürlich sowohl zeit- und kostenintensiver ist. [REPENNING et al., 2009]

3.4.5 Auftreten von Fehlern

Wie eingangs erwähnt gibt es die FDR (False discovery rate), also falsch positiv oder falsch negative Fehler. Hierbei handelt es sich um Fehler der 1. und der 2. Art. [FAHMEIER et al., 1997]

Falsch positiv bedeutet ein Fehler 1. Art, d. h. ein Gen wurde als unterschiedlich exprimiert erkannt, obwohl keine Unterschied vorlag. Falsch negativ ist ein Fehler 2. Art, was so viel bedeutet wie, dass es nicht gelingt einen Unterschied aufzudecken, obwohl er vorhanden ist. [LIN et al., 2010]

	Entscheidung für	
	H0	H1
H0 wahr	Richtig	Falsch Fehler 1. Art
H1 wahr	Falsch Fehler 2. Art	Richtig

Tabelle 1 Falsch positiv und Falsch negativ (mod. nach [FAHRMEIR et al., 1997])

3.5 Datenhandling

Durch die Vielzahl verschiedener Technologien und den daraus resultierenden Parametern, wird der quantitative Vergleich von Ergebnissen aus verschiedenen Experimenten erschwert. Sogar die Reproduzierbarkeit innerhalb eines Experimentes kann mitunter schwierig sein. Dem kann nur über eine Standardisierung der Ergebnisse aus Microarray-Studien und die Datenzugänglichkeit über online verfügbare Datenbanken entgegengewirkt werden. [SCHOBBER, 2002]

Die MGED (Microarray Gene Expression Data) Gesellschaft hat es sich seit den späten 90er Jahren zum Ziel gesetzt den Datenaustausch bei Arraystudien zu

erleichtern. Die MGED hat daher Standards geschaffen, welche das Datenhandling erleichtern sollen, einerseits durch den MIAME Standard, andererseits durch MAGE und MO. [BALL und BRAZMA, 2006]

3.5.1 MIAME Standard

Durch MIAME (Minimum Information About Microarray Experiments), einer Richtlinie, die vom MGED-Konsortium (Microarray Gene Expression Database) 1999 geschaffen wurde [SCHOBBER, 2002] um Ergebnisse von Studien leichter zu interpretieren und vergleichen zu können. MIAME gibt das Informationsminimum an, welches zu jeder Microarrayanalyse gespeichert werden muss um die Reproduzierbarkeit und Vergleichbarkeit der Analysedaten zu gewährleisten. [LIU-STRATTON et al.,2004]

Als Folge des MIAME Standards gibt es mittlerweile an die 30 Reporting Standards, nicht nur für Genexpressionsdaten, sondern auch für Daten aus Proteom und Metabolomstudien, sowie für andere Hochdurchsatzmethoden. [QUACKENBUSH, 2009]

3.5.3 MAGE

MAGE (Microarray Gene Expression) besteht aus drei Teilen:

- MAGE-OM (Microarray and Gene Expression Object Model), Standardobjektmodell für Microarrayexperimente
- MAGE-ML (Microarray Gene Expression Markup Language), Standardaustauschformat für Microarrayexperimente
- MAGE-Stk (Microarray and Gene Expression Software Toolkit), ein Tool für die Erstellung von MAGE-ML [BALL und BRAZMA, 2006]

3.5.4 MO

MO steht für MGED Ontology und definiert allgemeine Begriffe und Regeln für Annotation, Datenanalyse und Datenaustausch. MO wird in MAGE-ML Dateien dazu verwendet Microarrayexperimente und die daraus resultierenden Daten zu beschreiben. [BALL und BRAZMA, 2006]

3.6 Das Transkriptom mathematisch betrachtet

Im Rahmen einer deutschen Studie wurde vor einigen Jahren ein mathematisches Modell des Transkriptoms erstellt. Festgestellt wurde hierbei, dass die analysierten Datensätze eine auffällig schiefe Verteilung zeigten. Bei den verwendeten Daten handelte es sich um Signale von mehr als 1100 Genen von 5 Probanden ohne bekannte Vorerkrankungen in fünffacher Ausfertigung. Ermittelt wurden die Signale mittels Oligonukleotid Microarrays aus den weißen Blutkörperchen im venösen Blut der Probanden. Gesamt wurden mehr als 28.000 Signale analysiert. [HOFFMANN, 2006]

Anhand der Daten ist es gelungen das menschliche Transkriptom mathematisch wie folgt zu beschreiben: $C_g = \alpha * P / \lambda$

Dies Gleichung versteht sich so: C_g (Gleichgewichtskonstante), α (Transkriptionskoeffizient, P (Polymerase-II-Aktivität), λ (Zerfallskonstante). Dieses mathematische Modell nimmt eine Gleichverteilung der Variablen an.

Dieses Modell zeigt, dass viele ähnliche Expressionswerte auf einer Aufhebung der Schwankungen der Einzelvariablen durch gegenseitige Beeinflussung beruhen. Auch der Phänomen von extrem hohen Genexpressionen wurde untersucht. Dabei spielt der mRNA Abbau eine wesentliche Rolle, da vor allem niedrige Zerfallskonstanten für hohe Genexpression verantwortlich sind. Überraschen ist dabei, dass deren Einfluss höher ist, als die von hohen Syntheseraten. [HOFFMANN, 2006]

Transkriptomdaten werden oft mit der Genaktivität verknüpft, dabei spielt die Synthese eigentlich eine untergeordnete Rolle. Vielmehr ist die Hauptrolle beim Abbau zu sehen, welcher in dieser Art der Interpretation vollkommen vernachlässigt wird. Geringe Genexpression ist also nicht unbedingt mit einer geringen Transkriptionsaktivität gleichzusetzen, sondern kann auch bei hoher Aktivität durch hohe Abbauraten entstehen. [HOFFMANN, 2006]

3.7 Vor- und Nachteile von Microarrays

Durch die Anwendung der Microarray-Technologie ist es möglich durch parallel ablaufende Experimente die Expression verschiedener Genabschnitte zeitgleich zu untersuchen. Durch diese Methode ist es nicht nur möglich eine Vielzahl von Analysen relativ schnell durchzuführen, sondern durch die vielen gleichzeitigen Analysen wird auch das Verständnis für das Zusammenspiel auf molekularer Ebene verbessert. Dadurch, dass es auch bei sehr kleiner Probandenzahl zur Generierung vieler Ergebnisse kommt, ist aber mit einem gleichzeitigen Anstieg der Fehlerquote zu rechnen. So steigt z. B. das Risiko für falsch positive Resultate. Eine Absicherung der Ergebnisse mittels qRT-PCR (real time PCR) sollte daher immer in Betracht gezogen werden. [VICTOR et al., 2005]

4 qRT-PCR

4.1 Methode

Bei der Analyse von Microarray-Daten kommt es oftmals zu unerwünschten Nebeneffekten. Dieses Rauschen kann bei großen Datenmengen aus einer Vielzahl an Genen und Proben zu Falschidentifizierungen von Genen führen. Daher werden die Ergebnisse von Microarray-Studien oftmals mittels qRT-PCR überprüft. Die Probandenanzahl bei PCR-Analysen ist aber kleiner als bei der dazugehörigen Microarray-Studie. [FU et al.,2009]

Die qRT-PCR oder quantitative Polymerase-Kettenreaktion funktioniert wie die klassische PCR [Böckenhauer und Bongratz, 2003], indem sie der Vervielfältigung von DNA dient, dabei wird aber auch eine Quantifizierung der so gewonnenen DNA durch Fluoreszenzmessungen durchgeführt. [FU et al.,2009]

Man unterscheidet zwischen absoluter und relativer Quantifizierung unterscheiden, welche entweder über Kalibrationskurven oder mit Hilfe von Referenzgenen durchgeführt werden. [PFAFFL et al., 2004]

4.2 Anwendungen

4.2.1 Lebensmittelanalytik und Lebensmittelsicherheit

Auf Molekularebene lassen sich gentechnische Veränderungen und Kontaminationen durch Mikroorganismen nicht nur durch Microarrays, sondern auch durch PCR nachweisen. Mit PCR kann daher im Rahmen der Lebensmittelsicherheit die Qualität, sowie das Einhalten von Spezifikationen und gesetzlichen Vorgaben geprüft werden. [FISCHER und HAASE, 2006]

4.2.2 Morbus Crohn

Zwischen 2004 und 2007 wurde in einer portugiesischen Studie [Ferreira et al., 2010] die Wechselwirkung der Fettaufnahme mit dem Krankheitsverlauf von Morbus Crohn mittels PCR untersucht. Aus Blutproben wurde DANN extrahiert und der Polymorphismus (SNPs) der apoptotischen Gene (CASP9, FASLG, PPARy) von 99 ambulanten Patienten mit Morbus Crohn, sowie 116 Gesunden untersucht. Die Fettaufnahme wurde mittels Fragebögen über Verzehrshäufigkeiten (das letzte Jahr vor der Befragung) erhoben. Es konnte gezeigt werden,

dass ein hoher Konsum an ungesättigten Fettsäuren mit aktiver Krankheit assoziiert ist. [FERREIRA et al., 2010]

4.2.3 Lebensmittelallergien

In einer italienischen Studie wurden kürzlich Samen und Lebensmittel, welche häufig Haselnuss Spuren enthalten auf ihr allergisches Potential, durch PCR-Analyse ihrer wichtigsten Allergen codierenden Gene (Cor a 1, Cor a 8, Cor a 14) untersucht. Es wurde festgestellt, dass vor allem bei Proben von komplexen Lebensmittel, die PCR Amplifikation des Cor a 8 Gens besser ist als die der beiden anderen Genen. Daher ist es möglich über dieses Allergen codierende Gen, bereits geringe Spuren von Haselnuss in Lebensmittel nachzuweisen. [D'ANDREA et al., 2011]

Eine österreichische Studie untersuchte die DNA von Allergenen mittels RT-PCR in weißem Senf. Getestet wurden sowohl eigens hergestellte Proben, sowie handelsübliche Fertigprodukte. Die PCR-Analyse wurde für vier Allergen codierende Gene durchgeführt. Es konnte DNA Spuren von weißem Senf ab einer Menge von 5 pg nachgewiesen werden. [FUCHS et al., 2010]

4.2.4 Diätetische Lebensmittel

Shaw et al. untersuchten in einer englischen Studie die Auswirkungen von verschiedenen Fettsäuren (Arachidonsäure, Eicosapentaensäure, Decosahexaensäure, Linolsäure, Ölsäure und Palmitinsäure) auf die Expression von Genen, welche entzündliche Proteine codieren in menschlichen Nabelvenenendothelzellen mittels qRT-PCR. Es zeigte sich, dass die einzelnen Fettsäuren einen unterschiedlichen Einfluss auf die Genexpression und somit auch auf die Entzündungsprozesse in Epithelzellen haben. Arachidonsäure und Decosahexaensäure zeigten abhängig von der Konzentration und Inkubationszeit eine leichte Down-Regulation der Genexpression bzw. hatten keinen Einfluss auf die Gen-

expression. Die anderen Fettsäuren führten zur vermehrten Expression einiger Gene. Eine gesteigerte Genexpression inflammatorischer Proteine durch Arachidonsäure konnte in dieser Studie nicht festgestellt werden. [SHAW et al., 2007]

4.2.5 Kalorische Restriktion und oxidativer Stress

Eine spanische Studie untersuchte die Auswirkungen von Energie reduzierter Kost auf die mitochondrielle Genexpression bei Übergewichtigen (durchschnittlicher BMI 32,5) mittels qRT-PCR zu untersuchen. Untersucht wurde die Expression der Gene COX15 (Cytochrom-c-Oxidase), NDUFS2 (NADH-Coenzym Q Reductase), MGST2 (Microsomale Gluthadion S-Transferase), TNF α (Tumornekrosefaktor α assoziiertes Gen) in PBMC (Peripheral Blood Mononuclear Cell oder dt. mononukleäre Zellen des peripheren Blutes). Diese Studie zeigte, dass sich COX15 und MGST2 als Biomarker für oxidativen Stress bei Energie-restriktion besonders gut eignen und daher auch für nutrigenomische Adipositasstudien. [CRUJEIRAS et al., 2009]

4.3 Durchführung und Auswertung

4.3.1 Quantifizierung

Während der Reaktion wird die entstandene Menge an PCR-Produkt gemessen und daraus die ursprünglich in die Reaktion eingegangene Menge an DNA berechnet. So wird die Anzahl der durch die Polymerase Kettenreaktion gebildete Kopien berechnet. Die Messung erfolgt fluorimetrisch über unterschiedlich markierte Oligonukleotide. Nach der Anregung mit kurzwelligem Licht kommt es zu einer messbaren Emission. [FISCHER und HAASE, 2006]

Bei der absoluten Quantifizierung wird eine Kalibrationskurve herangezogen. Die relative Quantifizierung vergleicht die Genexpression des zu untersuchenden Gens mit der Genexpression von einem so genannten Housekeeping Gen (HKG Gen), welches nicht regulierend ist. Auch die Verwendung eines Referenz-

renzindex mit mehreren HKG Genen ist möglich. Dieser Schritt wird als Normalisierung bezeichnet. [PFAFFL et al., 2004]

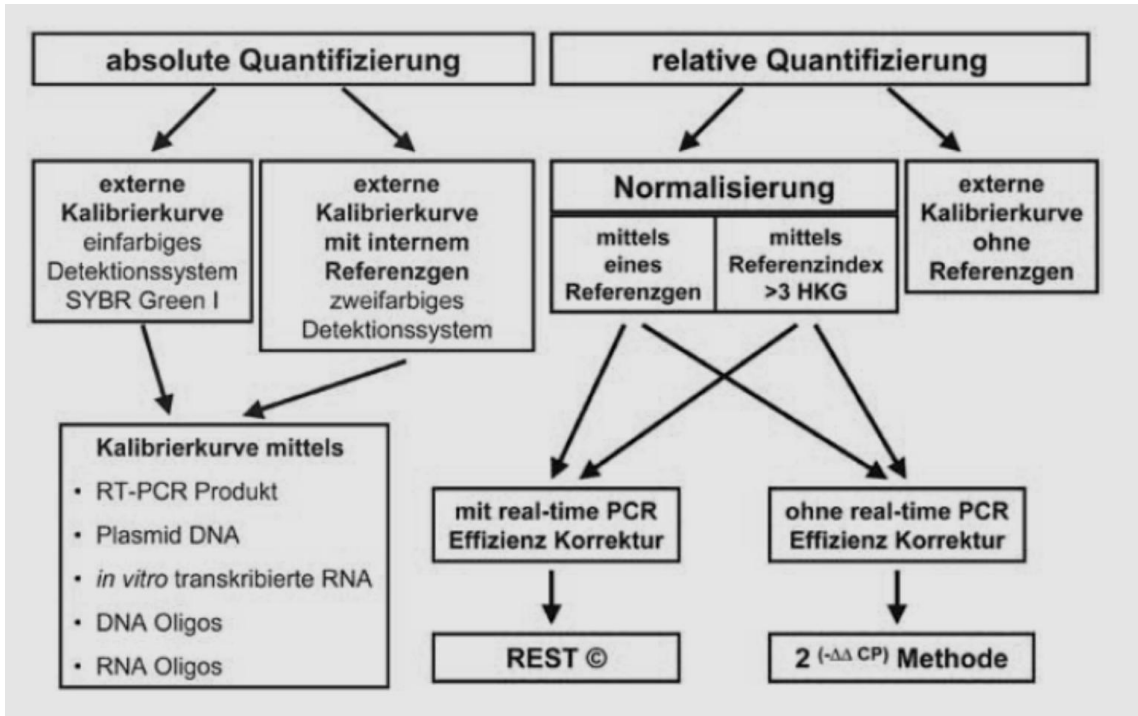


Abbildung 9 Quantifizierung in der q-RT-PCR [PFAFFL et al., 2004]

4.3.2 Amplifikation und mathematische Modelle

Durch die RT-PCR werden DNA-Sequenzen vermehrt. Bei diesem Vorgang spricht man von Amplifikation. Da die Vermehrung exponentiell verläuft, lässt sich die Expression in Abhängigkeit von der Vermehrung durch folgende Gleichung beschreiben: $Y_n = Y_0 \cdot 2^n$, wobei Y_0 die Genexpression zu Beginn und Y_n die Genexpression nach n Vermehrungszyklen darstellt. Da diese Gleichung nur theoretisch zutrifft und unter Versuchsbedingungen die Amplifikation im Experiment einer gewissen Streuung unterliegt, findet eine weitere Gleichung Verwendung: $Y_t = Y_0 \cdot (1+e)^t$. Die Dauer der kontinuierlichen Verstärkung wird durch die Variable t beschrieben und e die Effizienz der Verstärkung mit einem

Wert zwischen 0 und 1, wobei 1 die höchste Effizienz darstellt. Diese Gleichung findet sowohl für die Zielgene, die untersucht werden, sowie für Referenzgene Anwendung. [FU et al., 2010]

Der Wirkungsgrad, sprich die Effizienz, muss hoch und vergleichbar sein, damit eine zuverlässige Quantifizierung der zu untersuchenden Gene mittels q RT-PCR möglich ist. [TÖWE et al., 2010]

Es zeigt sich, dass ein Aktivitätsverlust der Polymerase zu einer Abnahme der Effizienz, durch die Verringerung der Wahrscheinlichkeit ein Molekül zu duplizieren innerhalb eines Zyklus, führt und es zu einem Sättigungseffekt kommt. [LALAM et al., 2005]

4.3.3 Statistische Modelle

Es haben sich nur zwei statistische Methoden für die Auswertung von PCR Studien durchgesetzt, einerseits das ANOVA (Varianzanalyse) basierende GEE Modell und die Kovarianz Analyse (ANCOVA). Das GEE Modell eignet für Verlaufsanalysen mit mehreren Probennahmen zu unterschiedlichen Zeitpunkten, z. B. um die Genexpression während einer Intervention zu messen. [FU et al., 2010]

Unter GEE (generalized estimating equations) werden verallgemeinerte Schätzgleichungen verstanden, wodurch es auch ermöglicht wird Ergebnisse von Probanden mit einer unterschiedlichen Zahl an Messungen miteinander zu vergleichen. [BENDER et al., 2007]

Üblicherweise wird für die Auswertung eine Statistiksoftware wie SAS oder ähnliches verwendet. [FU et al., 2010]

5 Proteom Analysen – Chromatographie, Elektrophorese und Massenspektroskopie

5.1 Methoden

Es zeigt sich, dass die Proteomanalysen und ihr zentrales Ziel, die Identifizierung von Proteinen, schwieriger und komplexer ist als vergleichsweise Transkriptom Analysen wie Microarrays. Im Gegensatz zu Microarrays, wo sich jedes Gen an einem Platz des Arrays befindet, gibt es diese genaue Lokalisation bei Proteomuntersuchungen nicht. Hier findet die Peptididentifizierung mittels Massenspektroskopie statt, wobei diese durch überlappende Peaks verkompliziert wird. [FU et al.,2009]

Die Aufgabe der MS in der Proteom Analyse liegt im Wesentlichen in der Identifizierung und Quantifizierung, während die vorangehende Separation durch chromatografische Methoden, Elektrophorese oder Gelfiltration erfolgt. [PATTERSON und AEBERSOLD, 2003; SPRENGER und JENSEN, 2010]

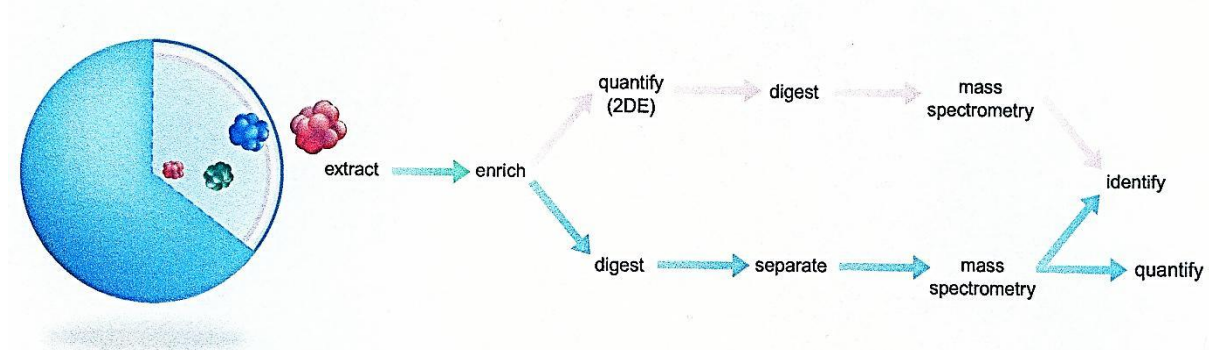


Abbildung 10 Schema einer Proteomanalyse [PATTERSON und AEBERSOLD, 2003]

In der Proteinanalytik findet vor allem die MS/MS (Tandem-Massenspektroskopie) Anwendung. Der erste Scan dient der Quantifizierung, während der zweite der Identifizierung dient. [FU et al.,2009]

Vorangehende Separationsmethoden können bei der MS/MS (Tandemmassenspektroskopie) entfallen. [BÖCKER, 1997]

5.2 Anwendung

5.2.1 Detektion von renalen Schädigungen

Mittels ESI-TOF Massenspektrometrie wurde in einer deutschen Studie die Proteinmuster im Harn von gesunden Probanden, Patienten mit nephrotischem Syndrom und Diabetikern mit/ohne Mikroalbuminurie bzw. Proteinurie untersucht. Es konnte gezeigt werden, dass durch fehlende Proteine bei Patienten mit nephrotischem Syndrom eine sichere Diagnose auch ohne Nierenbiopsie möglich ist, da bestimmte Proteine bei dieser Erkrankung nicht im Harn vorliegen. Es konnten auch Veränderungen im Proteinmuster von Diabetikern festgestellt werden, wodurch man davon ausgehen kann, dass die Diagnose von renalen Schädigungen bereits im symptomlosen Frühstadium möglich ist, in welchem eine diätetische Beeinflussung noch möglich ist. [HAUBITZ et al., 2004]

5.2.2 Peptidbiomarker

Ein nicht unbedeutender Focus der Proteomics richtet sich auf die Identifizierung von Biomarkern aus minimal invasiv gewinnbaren Proben. Dabei handelt es sich üblicherweise um Speichel-, Urin-, Plasma- oder Serumproben. Beim Einsatz von Peptidbiomarkern ist zu beachten, dass es durch proteolytische Umsetzung, zur Modifizierung des Markers kommen kann. Andererseits kann es zur Proteinstabilisierung und in weiterer Folge auch zur Stabilisierung des Biomarkers kommen. Für die Ernährungswissenschaften von Bedeutung können die Biomarker verschiedener Klassen zur Beurteilung von diätetischen

Maßnahmen bei ernährungsassoziierten Erkrankungen oder im Rahmen von Interventionsstudien z. B. bei Diabetikern oder Patienten mit metabolischem Syndrom sein. [JÜRGENS et al., 2005]

5.2.3 Enzymsuche

Die Massenspektrometrie bietet auch die Möglichkeit noch unbekannte Enzyme mit definierter katalytischer Funktion zu finden. Diese Methode wird MES (Massenspektrometrie unterstütztes Enzymscreeningsystem) genannt. Über die Reaktionsprodukte wird hierbei das Enzym mittels MALDI-MS analysiert. Hierbei werden Proteine immobilisiert und mit Enzymsubstraten inkubiert. Der Nachweis auf das Vorhandensein eines Enzyms geschieht über Veränderungen des Substrates, welche durch MS nachgewiesen werden. Auch die Wirkung von Inhibitoren können getestet werden. Gegenüber Enzymassays konventioneller Art besteht der Vorteil, dass eine Kontrolle der Substratumsetzung vorliegt, wodurch die Anzahl der falsch positiven Ergebnisse verringert wird. [SCHLÜTER, 2006]

5.3 Durchführung

5.3.1 Tandemmassenspektrometrie

Die MS/MS ergibt sich durch die Kombination zweier Massenspektrometer und einer zwischengeschalteten Stoßaktivierung, welche zum weiteren Zerfall führt. Diese Technik führt zum Anstieg der Selektivität, welche es ermöglicht auch Proben mit ausgeprägter Matrixbelastung zu identifizieren und zu quantifizieren. [BÖCKER, 1997]

Kommt die Tandemmassenspektroskopie bei der Proteinanalyse zur Anwendung so wird diese Methode als PFF (Peptide Fragmentation Fingerprinting)

bezeichnet. Auf diese Weise können einzelne Peptide im MS gesondert betrachtet werden. Aus den durch die MS bestimmten Molekulargewichte, kann ein Fingerprint generiert werden. [BLÜGGEL, 2002]

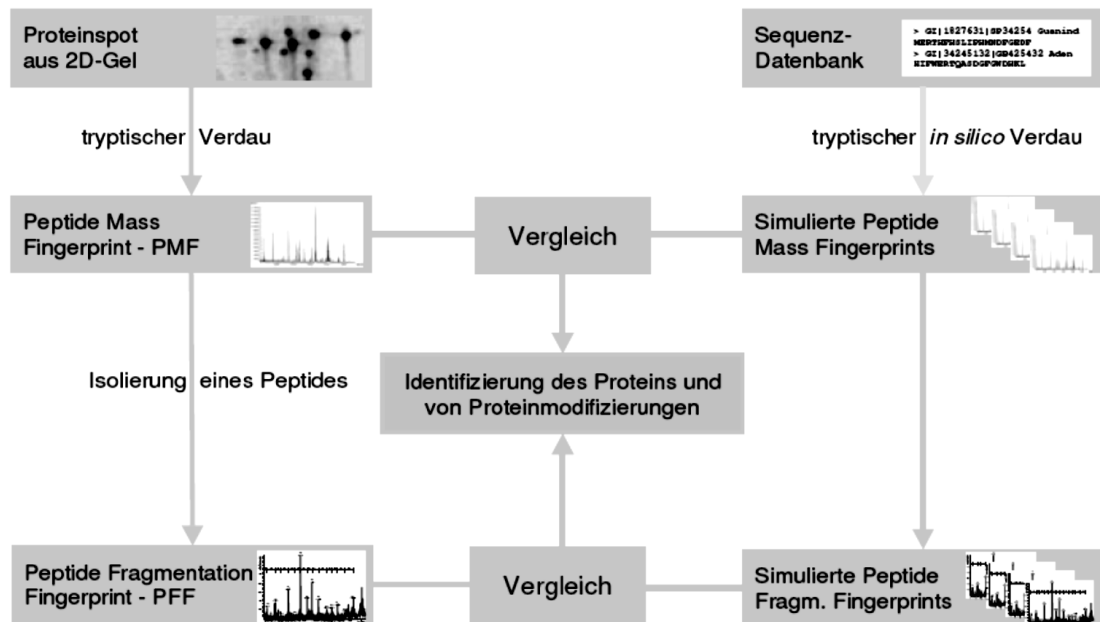


Abbildung 11 Peptide Mass Fingerprinting und Peptide Fragmentation Fingerprinting [BLÜGGEL, 2002]

Der Einsatz von MS/MS ermöglicht es auch eine vorangehende Auftrennung entfallen zu lassen, das bedeutet, dass auf chromatographische Trennmethode verzichtet werden kann. [Böcker, 1997] Eine LC ESI MS/MS kann aber für besonders intensive Analysen von Vorteil sein, da sie besonders umfangreiche Informationen liefert. Hierbei werden die mittels HPLC aufgetrennten Peptidgemische direkt ins Massenspektrometer gesprüht (ESI) und fragmentiert. [WATTENBERG und BLÜGGEL, 2002]

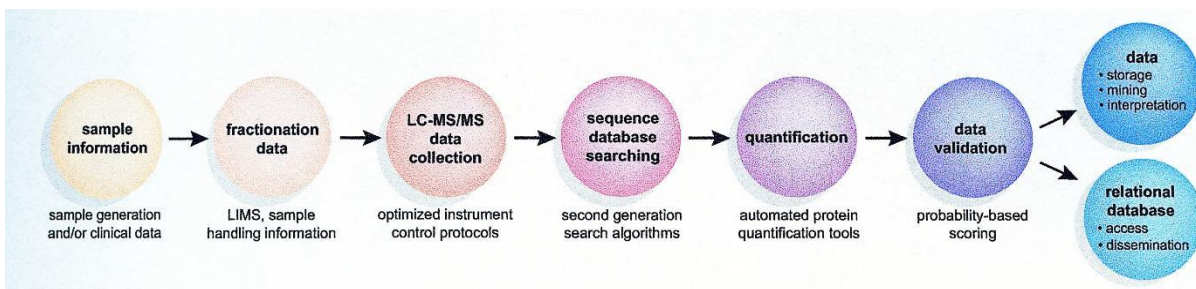


Abbildung 12 Schema eine Proteom MS/MS [PATTERSAON und AEBERSOLD, 2003]

5.3.1.1 Identifizierung von Peptiden

Die Basis der Identifizierung beruht auf die Annahme, dass die Aufspaltung jedes Peptides bekannt ist. Das impliziert, dass die Masse nach der Aufspaltung (im zweiten Scan) und die ursprüngliche Masse, eine Rekonstruktion zulässt, wodurch man auf die Struktur im ersten Scan schließen kann. [FU et al.,2009]

Die häufigste Art der Identifizierung von Peptiden bei der Verwendung von Hochdurchsatzmethoden wie MS/MS ist ein Abgleich mit verschiedenen Datenbanken. Dabei wird das experimentelle Spektrum mit theoretischen Spektren aus diesen Datenbanken verglichen. Dabei erfolgt der Abgleich nur mit Massenspektren von Peptiden, welche die selbe Masse wie die gemessene Masse +/- eines Toleranzbereiches haben. Die theoretischen Spektren werden durch gängige Regeln der Fragmentation ermittelt. Verschiedene Tools zur Datenbanksuche wie SEQUEST, MS-Tag, MASCOT, Sonar, X-Proteo, X! Tandem oder Probid können zur Peptididentifikation genutzt werden. [NESVIZHSKII und AEBERSOLD, 2004]

Die möglichen Peptide werden durch die Suchmaschinen gemäß ihren Score gelistet, welcher den Grad der Übereinstimmung angibt. Unabhängig von der gewählten Suchmaschine ist die Anzahl der identifizierten Spektren zumeist recht niedrig. Bei vielen Experimenten liegt die Quote unter 20%. Die Identifizie-

rung hängt wesentlich von der Qualität der Spektren und vom Gerätetyp ab, aber eine Identifizierungsquote von 50% wird auch bei qualitativ hochwertigen Untersuchungen kaum überschritten. [DEUTSCH et al., 2008]

Die oben beschriebenen Sequenzsuchmaschinen sind nicht die einzige Identifizierungsmöglichkeit. Es gibt auch die Möglichkeit mit Konsensspektren aus früheren Experimenten zu arbeiten. Es werden echte beobachtete Spektren verwendet als vereinfachte theoretische. Die dritte Möglichkeit ist die De-Novo-Sequenzierung. Auf diese Weise wird softwaregestützt die Identifizierung eines MS/MS Spektrums ohne Vorkenntnisse versucht. [DEUTSCH et al., 2008]

5.3.1.2 Quantifizierung von Peptiden

Man kann zwei verschiedene Arten der Identifizierung von Peptiden unterscheiden, wobei eine mit Hilfe von Isotopen durchgeführt wird. [FU et al.,2009]

Eine mögliche Methode die die ICPL (Isotope Coded Protein Label). Bei dieser werden alle freien Aminogruppen eines Proteingemisches markiert. Hierbei macht man sich den Umstand zu Nutze, dass Isotope, obwohl chemisch ident, ein unterschiedliches Molekulargewicht aufweisen. Bei Proteomanalysen werden mindestens zwei Zustände einer Zelle untersucht werden, da das Proteom im Gegensatz zum Genom dynamisch ist. Daher werden für jeden zu analysierenden Zustand andere Isotopenmarkierungen verwendet. Die Quantifizierung ist also auch nach Wiedervereinigung der Proben möglich. Diese Wiedervereinigung ist statistisch wichtig, weil sich jede Veränderung, die nach Durchmischung auftritt, beide Gruppen gleichermaßen betrifft. [LOTTSPREICH, 2006; FU et al.,2009]

5.3.1.3 Validierung der identifizierten Peptide

Die größte Schwierigkeit ist nicht eine passendes Peptid mit diversen Tools in einer Datenbank zu finden, sondern unter den vielen scheinbar möglichen, das richtige oder das passendste auszuwählen. Eines der Probleme ist, dass viele Spektren in Datenbanken von schlechter Qualität sind, es könnte daher auch das richtige Peptid mangels Daten als falsch erkannt werden. Es kann natürlich auch sein, dass es viele scheinbar passende Treffer gibt, ohne dass sich das tatsächlich richtige Peptid in der Datenbank befindet. Es ist aber möglich die Sensitivität der Identifizierung zu erhöhen. Dies kann durch Clustering redundanter Spektren erfolgen. Gänzlich ist eine automatische Validierung aber nie fehlerlos, da sich die Probleme nicht vollständig eliminieren lassen. Prinzipiell wäre auch eine manuelle Validierung denkbar, aber für Hochdurchsatzmethoden die eine Vielzahl von Spektren liefern vom Zeitaufwand her nicht machbar. [NESVIZHSKII und AEBERSOLD, 2004]

5.3.1.4 Validierung der identifizierten Proteine

Das Ziel der Anwendung von Hochdurchsatzmethoden in der Proteomik ist die Proteinidentifizierung in Proben. MS/MS Spektren werden aber von Peptiden und nicht von Proteinen erstellt. Daher basieren alle Aussagen über dieses Protein auf der korrekten Identifizierung von Peptiden. Auch geringe falsch-positive Raten bei der Peptididentifikation können sich negativ auf die Proteinidentifikation auswirken, da die Fehler aufsummiert werden. Sehr viele Peptididentifikationen können daher auch viele Fehler bergen, genauso schwierig ist es aber ein Protein mit nur einer Peptididentifikation oder nur sehr wenigen zu bestimmen. [NESVIZHSKII und AEBERSOLD, 2004]

5.3.1.5 Verschiedene Untersuchungsansätze

Massenspektrometrische Analysen in Kombination mit chromatografischen Trennmethode können nach zwei Gesichtspunkten hin durchgeführt werden. Man unterscheidet zwischen unfokussierter MS (Shotgun Methode) und fokussierter MS. Bei ersterer wird das intensivste MS-Spektrum zur Identifizierung herangezogen. Bei letzterer Methode werden zuerst die informationsreichsten Peptide untersucht, die in einem zweiten Durchgang gezielt untersucht werden, die Signalstärke bleibt dabei unbeachtet. [SCHMIDT et al., 2008]

5.3.2 Gelelektrophorese und Massenspektroskopie

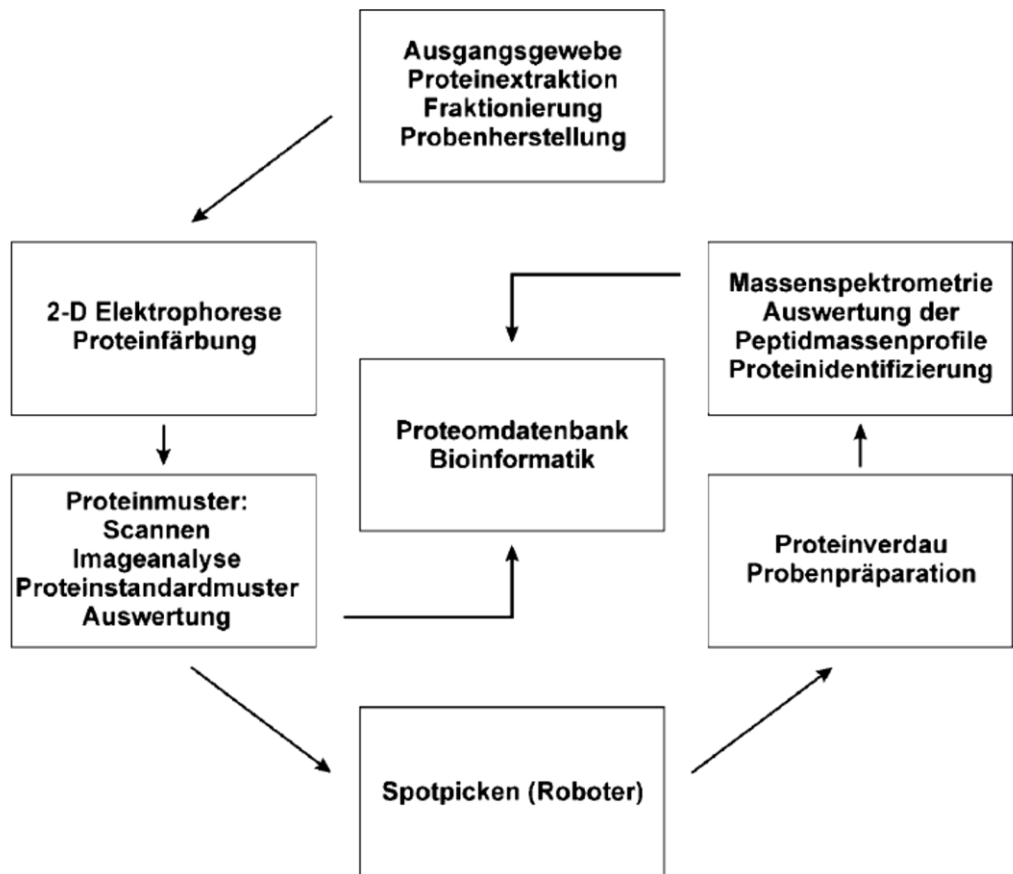


Abbildung 13 Verknüpfung von Elektrophorese und Massenspektrometrie [KLOSE, 2002]

Wird nur eine MS durchgeführt wird muss die Auftrennung von Proteinextrakten aus Zellen und Geweben in jedem Fall zuvor durchgeführt werden. Eine hierfür bewährte Methode ist die 2 D-Gelelektrophorese. Durch diesen vorgeschalteten Schritt wird gleichzeitig eine Auftrennung sowie die Quantifizierung durchgeführt, wodurch der nachfolgenden MS nur noch die Identifizierung obliegt. Um eine MS durchführen zu können müssen noch 2 Zwischenschritte (Picken und Verdau) durchgeführt werden. [WATTENBERG und BLÜGGEL, 2002]

Wird kein 2-D Gel sondern ein 1-D Gel verwendet enthalten die Banden häufig mehr als ein Protein, d.h. die Auftrennung ist nicht vollständig. Eine Identifikation kann daher oft nur mittels Tandemmassenspektroskopie durchgeführt werden. [ASHMAN und GAUSEPOHL, 2002]

Im Gegensatz zur klassischen Gelelektrophorese bietet die FFE (Free Flow Elektrophorese) die Möglichkeit einer Proteomanalyse bei komplexen Membranproteinen. Dies ist deshalb möglich weil es die FFE erlaubt, auch hydrophobe oder stark basische Proteine zu trennen. Es ist möglich Multiproteinkomplexe aus Zell- oder Organellenmembranen zu analysieren, da durch diese Trennmethode sowohl Moleküle wie auch Organellen oder ganze Zellen getrennt werden können. [UEFFING, 2006]

Sind posttranslationale Veränderungen bei den zu untersuchenden Proteinen zu erwarten, so werden diese ebenfalls mittels MS/MS untersucht. [ASHMAN und GAUSEPOHL, 2002]

Im klinischen Bereich hat sich ESI-TOF Massenspektrometrie mit vorangehender Kapillarelektrophorese zur Untersuchung von Plasma- und Proben etabliert, dass hierbei eine Vielzahl an Polypeptiden erfasst werden muss. [HAUBITZ et al., 2004]

5.4 Schwierigkeiten der Massenspektroskopie

5.4.1 Quantifizierung

Im Gegensatz zur Chipmethoden wird die Quantifizierung bei MS/MS durch überlappende Peaks erschwert. [FU et. al, 2009] Eine absolute Quantifizierung durch die 2-D Gelelektrophorese ist auch nicht möglich, wodurch die Reproduzierbarkeit dieser Methode limitiert ist. [BLÜGGEL, 2002] Darüber hinaus hat kaum ein in der Gelelektrophorese eingesetzter Farbstoff alle erforderlichen Eigenschaften (gute Empfindlichkeit, homogene Färbung der Proteinbanden, langer linearer Bereich etc.) um eine optimale Quantifizierung zu gewährleisten. [KLOSE, 2002]

5.4.2 Zeitaufwand

Methoden die sehr umfangreiche Informationen liefern (LC ESI MS/MS) eignen sich zwar am besten für sehr komplexe Peptidgemische, haben aber einen relativ großen Zeitaufwand und sind daher keine Hochdurchsatzmethoden. Für großen Probenumfang eignen sich Proteinarrays besser. [WATTENBERG und BÜGGEL, 2002]

5.4.3 Isoformen

Verschiedene Isoformen von Proteinen können vorliegen und die zu Grunde liegenden posttranslationalen Veränderungen sind variabel und dynamisch. [PATTERSON, 2004] Die 2-D Gelelektrophorese ermöglicht die Erfassung von Proteinisoformen und Polymorphismen und mit ausreichender Genauigkeit auf die Proteinkonzentration im Gewebe zu schließen. [KLOSE, 2002]

5.4.3. Identifizierung

Die Tandemmassenspektrometrie identifiziert nur Peptide, nicht aber Proteine. Diese Peptidspektren können mit bekannten Spektren aus Datenbanken verglichen werden. Zumeist werden aber bei einem Suchlauf aber verschiedene mögliche Peptide für ein einzelnes MS/MS Spektrum gefunden. Dieses Problem lässt sich nur durch unterstützende Software lösen, welche nach bestimmten Suchalgorithmen in verschiedenen Datenbanken gleichzeitig suchen. Auf diese Weise gelangt man durch Treffer für verschiedene Peptide in unterschiedlichen Datenbanken zu einem Protein. Durch spezielle Suchalgorithmen wird dabei die Falsch-positive Rate minimiert. [THIELE et al., 2010]

5.4.5 „Rauschen“

Wie bei den meisten Spektren, zeigen auch die Massenspektren von Proteinen ein Rauschen. Daher sind statistische Maßnahmen zur Trennung der erwünschten Signale vom Rauschen vor der eigentlichen Auswertung. Mittlerweile gibt es verschiedene Software um Massenspektrometriedaten aufzubereiten, wodurch das Verhältnis von Signal zu Rauschen verbessert werden kann. Häufig werden dafür Softwarepackages von Biocanductor verwendet. [LEE, 2010]

5.5 MIAPE

Ähnlich der Standards für Microarraydaten (MIAME) gibt es auch einen Standard für Proteomdaten, MIAPE (Minimum Information about a Proteomics Experiment). Da Proteomanalysen mittels unterschiedlicher Analysemethoden gewonnen werden, gibt der MIAPE-Standard Richtlinien für die verschiedenen Trennmethode (Elektrophorese, Gaschromatographie, etc.), für Massenspektroskopie, aber auch für die statistische Auswertung der Analyseergebnisse vor. Dabei gibt es verschiedene MIAPE-Module, angepasst an die verschiedenen Untersuchungsmethoden in der Proteomik. Eines der Module ist MIAPE-MS, welches die Dokumentationsstandards für Massenspektrometrie beschreibt. Da

bei ist aber zu beachten, dass es sich bei allen MIAPE Modulen um Richtlinien handelt, welche die Datenspeicherung, den Datenaustausch und die Reproduzierbarkeit der Analysen unterstützen sollen, handelt, nicht aber um eine Durchführungsanweisung. [TAYLOR et al., 2007]

6 Metabolom – Analysen

Bisher sah man das größte Potential von Metabolomuntersuchungen in der Pharmakologie und Toxikologie. Verglichen damit ist der Einsatz von Metabolomuntersuchungen in ernährungswissenschaftlichen Studien noch eher rückständig, obwohl ihre Anwendung auf die menschliche Ernährung enormes Potential besitzt. [GIBNEY et al., 2005]

Durch eine Fehlende Anbindung an das Genom muss sich die Analyse des Metaboloms, der Gesamtheit der Metabolite, auf die Messung der Konzentration eindeutig identifizierbarer Metabolite stützen. [NIELSEN und OLIVER, 2006]

Die enorme stoffliche Vielfalt der Metabolite lässt es derzeit nicht zu sie mit nur einer analytischen Methode zu erfassen, weswegen sich unterschiedliche Untersuchungsmethoden bei Metabolic Profiling etabliert haben. Metabolic Profiling oder Metabolomic Fingerprinting ist nichts anderes als eine quantitative Bestimmung verschiedener Metabolite eines Stoffwechselforgangs. [DETTMER, 2008]

Analysiert werden endogene Metabolite mit niedrigem Molekulargewicht aus verschiedenen Geweben und Körperflüssigkeiten. [HUANG und REGNIER, 2008]

Üblicherweise werden zur Untersuchung des Metaboloms verschiedene Analysemethoden wie Massenspektroskopie (MS), entweder als alleinige Methode oder gekoppelt an Gaschromatographie (GC) bzw. Elektrophorese, verwendet.

Die MS kann wie bei der Proteomuntersuchung wieder gekoppelt als Tandem-massenspektroskopie (MS/MS) eingesetzt werden. Als weitere Analysemethoden kommen die NMR (Kernspinanalyse) oder die PCA (Principal Component Analysis) zur Anwendung. [NIELSEN und OLIVER, 2006]

Es ist nicht das Ziel des Metabolomic Fingerprinting alle Metaboliten zu identifizieren, sondern Metabolitenmuster in Form von Spektren oder Chromatogrammen verschiedener Proben zu vergleichen. [DETTMER, 2008]

Metabolic Profiling eignet sich nicht zu diagnostischen Zwecken, sondern dient alleine der Auffindung neuer Biomarker. Für eine biomedizinische Forschung oder Diagnose sind Kenntnisse der biochemischen Stoffwechselwege erforderlich. Targeted Metabolomics, eine klassenweise Quantifizierung der Stoffwechselprodukte, beruht darauf. Ziel ist es hierbei anhand der Metabolit Konzentrationen den Stoffwechselweg zu bestimmen. Daher ist es auf diese Weise möglich auf alternative Stoffwechselwege und die damit einhergehenden Erkrankungen zu schließen, weswegen sich diese Methode zur klinischen Diagnose eignet. [WEINBERGER, 2008]

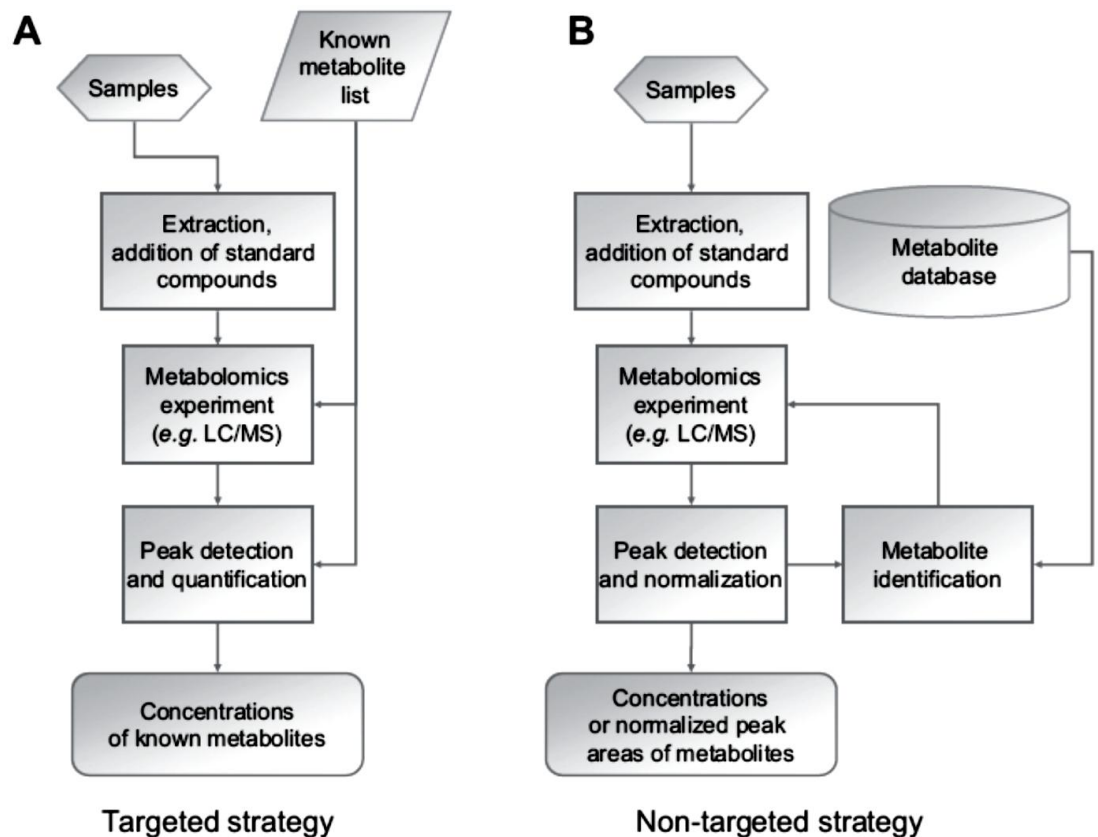


Abbildung 14 Metabolomics Workflow [ORESIC, 2009]

Im Wesentlichen gibt es drei Kriterien für die Auswahl der Trenn- und Analyse-methode:

- Das Ziel der Studie samt dazugehöriger Hypothese: wenn ein spezieller Aspekt des Metaboloms untersucht werden, so ist sicherzustellen, dass gewählten Methoden dafür geeignet sind
- Für den Fall, dass es nicht möglich ist, eine Hypothese aufzustellen, ist es nicht möglich Targeted Metabolomics zu wählen. Da man nicht genau weiß, was zu erwarten ist, ist das Metabolic profiling die richtige Strategie.
- Entscheidend ist auch die zu Verfügung stehende Probenmenge, welche die Auswahlmöglichkeit minimiert. [ORESIC, 2009]

6.1 Massenspektroskopie

Zumeist wird die MS als MS/MS oder in Kombination mit verschiedenen vorangehenden Trennmethode angewandt, obgleich sie auch ohne Trennschritt für die Metabolomanalyse verwendet werden kann. Dabei ist es aber nicht möglich isobare Metabolite zu trennen und auch Begleitphänomene wie Ionenunterdrückung durch Matrixbestandteile treten vermehrt auf. [DETTMER, 2008]

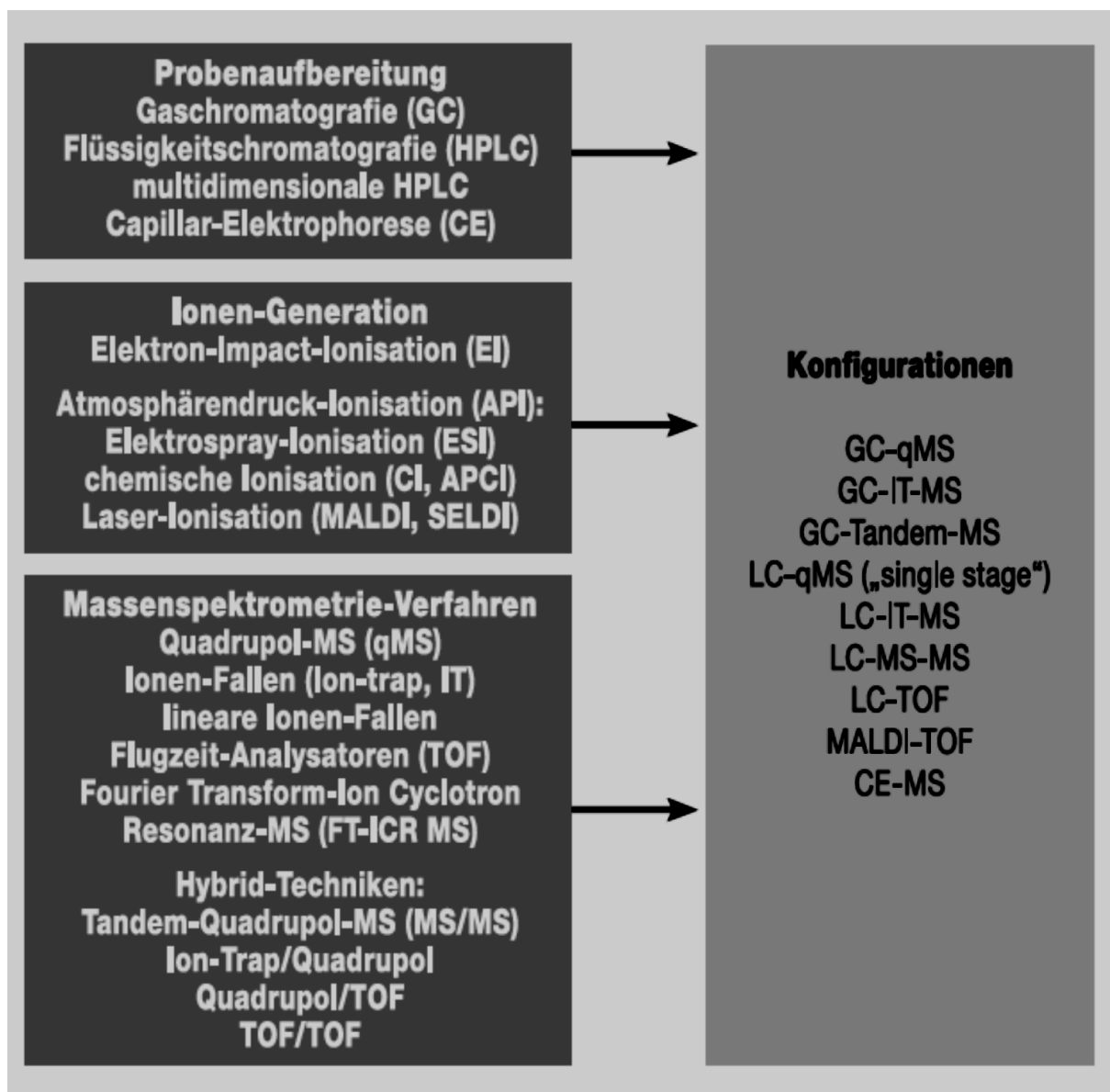


Abbildung 15 Weitere Geräte-Kombinationen für Massenspektrometrie [VOGESSER et al., 2007]

6.1.1 Tandemmassenspektroskopie

Für megaparametrische Analysen, bei denen sehr viele einzelne Metabolite mit einmal erfasst werden sollen, ist die LC-MS/MS, welche schon lange im Neugeborenen-Screening auf angeborene Stoffwechselerkrankungen eingesetzt wird, die Methode der Wahl. Sie ermöglicht es, eine simultane Quantifizierung mehrerer hundert Einzelstoffen in relativ kurzen Analyseläufen durchzuführen. Gerade zur Erstellung eines Profils von niedermolekularen Metaboliten ist diese Methode gut geeignet.

[VOGESSER et al., 2007]

6.1.2 Gaschromatographie und Massenspektrometrie

Bei dieser Technik dient im Wesentlichen die Gaschromatographie zur Auftrennung und die Massenspektroskopie der Quantifizierung. Häufig kommt die Kombination GC-TOF-MS zur Anwendung. Prinzipiell wäre auch eine Trennung mittels Flüssigchromatographie (LC) möglich, aber bei sehr ähnlichen Analysaten hat sich die GC als besser geeignet erwiesen. [JIANG et al., 2010]

Flugzeit-Massenspektrometer (TOF-MS oder time-of-flight-specrometer) haben einen unbegrenzten Massenbereich, wodurch sie gerade im biologischen Bereich zur Analyse von großen Molekülen gut geeignet sind. [BÖCKER, 1997]

Speziell für kleine, eher unpolare Metabolien ist die GC-MS als Standard zu betrachten. Da aber die meisten Metabolite eher polar sind muss vor der Trennung mittels GC eine Derivatisierung vorgenommen werden. Häufig wird hierbei eine Zwei-schrittderivatisierung angewandt, bei welcher zuerst eine Methoximierung stattfindet, welcher eine Silylierung folgt. Eine vollautomatische Derivatisierung durch Autosampler ist möglich. [DETTMER, 2008]

Häufig ist die Trenneffizienz durch die GC nicht ausreichend, wodurch sich eine zweidimensionale GC (GCxGC) vor allem bei sehr komplexen Proben als sinnvoll erweisen kann. Kombiniert wird diese Methode wieder mit anschließender TOF-MS.

Bei der zweidimensionalen Gaschromatographie sind zwei Säulen über einen Modulator gekoppelt. Dieser Modulator führt zu einer Zerlegung des Eluats aus der ersten Trennsäule, also zu einer Fraktionierung. Die eigentliche Trennung erfolgt in der kürzeren zweiten Säule, welche eine Reihe von Chromatogrammen liefert. Aus diesen Einzelchromatogrammen kann ein dreidimensionales Chromatogramm erstellt werden. [DETTMER, 2008]

6.1.3 Elektrophorese und Massenspektrometrie

Im Gegensatz zur Gaschromatographie können durch die Kapillarelektrophorese (CE) auch thermolabile oder polare Metabolite getrennt werden. Bei Anwendung der CE ist keine Derivatisierung notwendig, wodurch sie besonders gut für hydrophile Verbindungen geeignet ist. Auch hierbei handelt es sich um einen Trennschritt für eine nachfolgende TOF-MS. Als Ionisationstechnik findet die ESI (Elektrosprayionisierung) häufig Anwendung. Nachteilig sei aber das geringe Injektionsvolumen erwähnt. [DETTMER, 2008]

6.1.4 Flüssigkeitschromatographie und Massenspektrometrie

Genau wie die CE-MS eignet sich die LC-MS vor allem für thermolabilen oder polare Metabolite, daher ist diese Untersuchungstechnik in der Metabolomforschung ähnlich weit verbreitet wie die CE-MS. Im Gegensatz zur GC ist es zwar möglich polare Substanzgemische zu trennen, aber bei sehr komplexen Gemischen kann es vorkommen, dass darin enthaltene sehr polare Verbindungen mit der Durchflusszeit eluieren. Selektive Derivatisierung ist nur eine der möglichen Methoden um dieses Problem zu eliminieren. Ein weiteres unter Umständen beobachtbares Phänomen ist die Ionenunterdrückung durch Matrixbestandteile die ebenfalls eluiert werden. Eine Behebung ist durch isotope-

markierte interne Standards für viele Metabolite möglich und für eine Quantifizierung erforderlich. [DETTMER, 2008]

6.2 Kernspinresonanzspektroskopie

Die NMR war lange Zeit vorrangig in der Metabolomuntersuchung. Die NMR benötigt nur minimale Probenvorbereitung und ermöglicht eine simultane Erfassung aller protonenhaltigen Metabolite. [DETTMER, 2008]

Für diese Methode spricht auch, dass sie im Gegensatz zur Massenspektroskopie die Probe nicht zerstört. Mittels NMR ist es auch möglich Aussagen über die Lokalisation eines Metaboliten zu treffen und seine Wechselwirkungen mit anderen Zellbestandteilen zu bestimmen. [NOBELI und THORNTON, 2006]

Aufgrund ihrer vielen Nachteile wurde die NMR hinsichtlich ihrer Leistungsfähigkeit längst von anderen Methoden überholt und kann daher nicht als Routinemethode zur Metabolomanalyse verwendet werden. Eines der Hauptprobleme, die ungenügende Auflösung, wäre durch technische Weiterentwicklungen und modernerer Geräte, die durch höhere Feldstärken gekennzeichnet sind, in den Griff zu bekommen. Was letztlich aber für den limitierten Einsatz ausschlaggebend ist, ist die Tatsache, dass das Problem der eingeschränkten Empfindlichkeit. [Weinberger, 2008]

Durch die eingeschränkte Nachweisempfindlichkeit scheinen Spurenkomponenten im NMR-Spektrum nicht auf. [DETTMER, 2008]

6.3 Datenhandling

Metabolomanalysen liefern im Allgemeinen Massenspektren aus LC-MS, GC-MS etc. welche durch geeignete Software aufbereitet werden muss. Dabei steht nicht nur das Finden von Peaks und deren Quantifizierung im Vordergrund,

sondern auch die Umwandlung der gewonnenen Informationen in geeignete Datenformate, welche sich auch zum Abgleich mit diversen Datenbanken eignen. [LEE, 2010]

Obwohl die Untersuchungsmethoden immer besser werden, fehlt es oft noch am nötigen Vergleichsmaterial, da viele NMR und MS-Spektren von kleinen Molekülen oft nicht frei verfügbar sind. Dabei wird die [GIBNEY et al., 2005]

6.4 Anwendung

6.4.1 Neugeborenen Screening

Es ist schon länger bekannt, dass Metabolomuntersuchungen zur Diagnose von Stoffwechselerkrankungen geeignet in der Pädiatrie geeignet sind. Ursprünglich wurden zu Diagnose von angeborenen Stoffwechselerkrankungen mehrfache monoparametrische Test durchgeführt. Diese Tests wurden inzwischen durch multiparametrische Analysen mittels Massenspektrometer abgelöst, welche es ermöglichen 40 bis 80 Parameter simultan zu bestimmen. [WEINBERGER et al., 2006]

Die Verwendung multiparametrischer Marker, hat sich hinsichtlich der Spezifität als positiv erwiesen. Die Sensitivität monoparametrischer Marker wäre zwar hinreichend gegeben, allerdings ist hierbei auch eine Störanfälligkeit in Bezug auf Einflüsse aus der Ernährung gegeben, welche bei multiparametrischen Markern nicht vorliegt. [WEINBERGER, 2008]

Bei der Diagnose von PKU (Phenylketonurie) bei Säuglingen führt der monoparametrische Marker (Phenylalaninspiegel im Serum) häufig zu falsch positiven Diagnosen, da z. B. nach postprandialer Blutabnahme alle Aminosäuren im Blut erhöht sind. Bei multiparametrischen Markern werden hingegen verschiedene Aminosäuren, deren Spiegel in unterschiedlicher Weise erhöht oder erniedrigt sein kann, als Diagnosekriterium herangezogen. [WEINBERGER, 2008]

Seit der Nutzung der Tandemmassenspektroskopie im Neugeborenen Screening ist nicht nur möglich PKU früher zu diagnostizieren, sondern auch viele andere angeborene Stoffwechselerkrankung. Eine dieser Störungen ist der MCAD-Defekt (Medium-Chain Acyl Dehydrogenase). Diese Erkrankung kann nahrungsbedingt zu einer Stoffwechsellentgleisung führen. Durch frühzeitige Diagnosestellung sind Folgen wie ein hypoketonisches und hypoglykämisches Koma, welches letal enden kann, vermeidbar. [SANDER et al., 2000]

6.4.2 Metabolisches Syndrom

Metabolomics eignet sich auch zur Diagnose des metabolischen Syndroms, da diese multiparametrischen Biomarker auch eine Diagnose bei einer leichten Ausprägung dieser Erkrankung ermöglichen. [WEINBERGER et al., 2006]

6.4.3 Lipide, Lipoproteine und Cholesterometabolite in Körperflüssigkeiten

Ein wichtiger Teil der Metabolomics sind Lipide, da ihnen eine zentrale Rolle in diversen Stoffwechselerkrankungen zukommt. Lipiduntersuchungen (Lipidomics) erfolgen zumeist mittels MS/MS nach vorangehender HPLC. Viele Informationen über Interaktionen und Schlüsselfunktionen von Lipiden in pathobiochemischen Prozessen, wie Entzündungen, Krebs, Diabetes und anderen Stoffwechselerkrankungen konnten bereits gewonnen werden, wodurch ein starkes Indiz für die Eignung von Lipiden als Biomarker gegeben ist. [GRIF-FITH et al. 2010]

Im Gegensatz zu Lipiden hat sich bei Cholesterometaboliten, die GC als Trennmethode der Wahl erwiesen. Cholesterometaboliten sind als Metabolomics Forschungsgebiet deshalb so interessant, weil Störungen der Steroidbiosynthese sowie der Gallensäuresynthese sich in mannigfaltigen Krankheitsbildern äußern können. Gerade im Bereich der Arteriosklerose Forschung wur-

den diverse Studien mit Cholesterolmetaboliten durchgeführt. Auch der Zusammenhang der Konzentration dieser Intermediate und neurodegenerativen Erkrankungen wurde bereits an einigen Studien untersucht. [GRIFFITH et a., 2010]

7 Methoden zur Dimensionsreduktion und Datenreduktion

Die Reduktion der Dimensionalität ist im Umgang mit multidimensionalen Daten, welche häufig das Ergebnis von Untersuchungen mit Hochdurchsatzmethoden wie sie in der Metabolomik angewandt werden, eine wichtige Technik. [YAMAMOTO et al., 2009]

Die Dimensionsreduktion ist über verschiedene Methoden erreichbar:

- Hauptkomponentenanalyse (PCA)
- Multidimensionale Skalierung (MDS)
- Faktoren Analyse [LEE, 2010]

Als weitere Methode der Datenreduktion sind Clusteranalysen zu erwähnen, deren haupteinsatzgebiet aus bioinformatischer-ernährungswissenschaftlicher Sicht im Bereich der Microarrayanalysen zu finden ist.

Zum Datamining gehört auch das Auffinden bestimmter Strukturen, welche eine Reduktion der Datenflut auf wirklich aussagekräftige Daten ermöglichen. Hierbei kommt die Clusterbildung als Verfahren zur Gruppierung zum Einsatz. Zusammengefasst wird beispielsweise nach der Ähnlichkeit der Expression bei Microarraydaten. Da das Expressionsprofil zu Clusterbildung herangezogen wird, werden vor allem co-exprimierte Gene geclustert, deren Regulation und Funktion ähnlich ist. Dabei werden aber auch Gene, deren Funktion nicht bekannt ist, auf Grund ihrer Expression in denselben Cluster mit Genen, deren Funktion bekannt ist, zusammengefasst. Dies wiederum ermöglicht Rückschlüsse auf die Funktion der Gene. Die Bildung von Clustern ist nicht einheit-

lich sondern es kommen verschiedene Verfahren zum Einsatz. [SCHOBER, 2002]

7.1 PCA

Die PCA (Principal Component Analysis oder Hauptkomponentenanalyse) hat die Datenreduktion zum Ziel. Sie ist eine Methode der multivariaten Statistik. Dabei werden aus den ursprünglichen Daten Faktoren berechnet, die sogenannten Hauptkomponenten. Die Hauptkomponenten sind durch die linearen Projektionen der Eigenvektoren definiert. Sinn und Zweck dieser Datenreduktionsmaßnahme ist es, mit möglichst wenigen Faktoren alle Ausgangsdaten zu beschreiben. [LU und PANDOLFO, 2011]

Die PCA ist an sich keine Methode zur Visualisierung, aber die Projektion der Hauptkomponenten, sprich Eigenvektoren, schafft eine kompakte Visualisierung der ursprünglich multidimensionalen Daten in Form einer niedrig-dimensionalen Projektion. Die PCA ist also eine Methode zur Dimensionsreduktion. Es erfolgt eine Reduktion auf relevante Dimensionen. Dabei wird die Eigenwertzerlegung genutzt, bei welcher Daten aus einem p -dimensionalen Raum in einem Raum mit kleinerer Dimension projiziert werden. Dabei werden zwei oder drei Hauptkomponenten oder Eigenvektoren zur kompakteren Visualisierung von eigentlich multidimensionalen Daten herangezogen. [LEE, 2010]

Eine Datenmatrix mit mehr als zwei quantitativen Merkmalen bildet den Ausgangspunkt der Hauptkomponentenanalyse.

Der erste Schritt ist die Bestimmung des Mittelwertes Mittelwert:

$$\bar{x}_i = \frac{1}{p} \sum_{j=1}^p x_{ij}$$

Dabei ist stellt x_{ij} den Wert des Merkmals j beim Objekt i dar. Es wird also der Mittelwert aller zu einem Objekt gehörigen Merkmale bestimmt. [HANDL, 2002]

Die meisten multivarianten Verfahren, so auch die PCA setzen zentrierte Daten voraus. Dabei ist der Mittelwert gleich 0.

Entsprechende dieser Datenzentrierung ergibt sich folgende Matrix: [HANDL, 2002]

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

Dieser Schritt wird auch als Normalisierung der Daten bezeichnet, da die Ausgangsmatrix $\mathbf{X} \in \mathbf{R}^{N \times m}$ auf einen Mittelwert von 0 gesetzt wird. Dies kann auch in folgender Form dargestellt werden:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} \in \mathbf{R}^{N \times m}$$

Dabei gilt, $\mathbf{x}_i \in \mathbf{R}^m$, $i=1, \dots, N$. [DING et al., 2010]

Eine PCA kann sowohl bei bekannter als auch bei unbekannter Varianz-Kovarianz-Matrix durchgeführt werden. Zumeist ist die Varianz-Kovarianz-

Matrix Σ nicht bekannt und wird daher geschätzt und als empirische Varianz-Kovarianz-Matrix \mathbf{S} bezeichnet. [HANDL, 2002]

$$\mathbf{S} = \frac{1}{n-1} \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}}) (\mathbf{x}_k - \bar{\mathbf{x}})'$$

7.1.1 Hauptkomponentenanalyse mit bekannter Varianz-Kovarianz-Matrix

Σ

Die 1. Hauptkomponente (1. Eigenvektor) ist die Linearkombination $\mathbf{a}'_1 \mathbf{X}$ (alternativ: X_{a_1}). Dabei müssen folgende Bedingungen erfüllt sein:

$$Var(\mathbf{X}) = \Sigma$$

Es wird die Linearkombination mit der größten Varianz gesucht unter welcher die Nebenbedingung

$$\mathbf{a}'_1 \mathbf{a}_1 = 1$$

Gilt.

Aus der Gleichung

$$Var(\mathbf{a}'_1 \mathbf{X}) = \mathbf{a}'_1 \Sigma \mathbf{a}_1 = \mathbf{a}'_1 \lambda \mathbf{a}_1 = \lambda \mathbf{a}'_1 \mathbf{a}_1 = \lambda$$

Ergibt sich der Eigenwert λ zum Eigenvektor \mathbf{a}_1 . Da als Bedingung die Linearkombination mit der größten Varianz festgelegt wurde ist der Eigenvektor mit dem größten Eigenwert als 1. Hauptkomponente zu wählen.

Die 2. Hauptkomponente (2. Eigenvektor) ist die Linearkombination $\mathbf{a}'_2\mathbf{X}$ (**alternativ:** X_{a_2}). Dabei müssen die Nebenbedingungen wie folgt erfüllt sein:

$$\mathbf{a}'_2\mathbf{a}_2 = 1 \quad \mathbf{a}'_2\mathbf{a}_1 = 0$$

und

Es ist zu beachten, dass \mathbf{a}_1 und \mathbf{a}_2 orthogonal sind. [HANDL, 2002]

Die 1. Hauptkomponente hat den Eigenvektor mit dem größten Eigenwert, die zweite den zweitgrößten. Die Varianzen entsprechen den Eigenwerten und werden als Prozent erklärter Varianz bezeichnet, welche einen Anteil der Gesamtvarianz darstellt. Dies bedeutet, dass sich die Varianz auf alle Achsen verteilt, und jeweils als Prozentsatz der Gesamtvarianz angegeben werden. Die Einzelvarianzen hängen eng mit der Struktur der Ursprungsdaten zusammen. Die kumulierten Varianzen der einzelnen Achsen sollten mindestens 50% der Gesamtvarianz betragen. Je höher dieser Wert liegt, desto besser. Sehr kleine Werte weisen darauf hin, dass das Ziel der Dimensionsreduktion durch die PCA verfehlt wurde. [LEYER und WESCHE, 2007]

7.1.2 Berechnung der Anzahl der Hauptkomponenten:

Ausschlaggebend für die minimale Zahl an Hauptkomponenten ist die Struktur der Ausgangsdaten und damit verbunden auch die empirische Varianz-Kovarianz-Matrix \mathbf{S} , sowie deren Eigenwerte λ_1 bis λ_p , mit λ_1 als größtem Eigenwert.

Wichtiger Punkt dabei ist die Summe der Stichprobenvarianzen der Hauptkomponenten, wodurch festgestellt werden kann wie stark eine Hauptkomponente Anteil an der Gesamtstreuung hat:

$$tr(\mathbf{S}) = \sum_{i=1}^p s_i^2$$

Für die Berechnung der idealen Zahl an Hauptkomponenten ist der Wert α (0,75 -0,85) mit entscheidend. Dieser Wert wird vorgegeben und bewegt sich üblicherweise in zuvor genanntem Rahmen. Die Anzahl errechnet sich aus folgender Formel bei gegebenem Wert α für den kleinstmöglichen Wert r : [HANDL, 2002]

$$\frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p \lambda_i} \geq \alpha$$

Alternativ besteht die Möglichkeit nur diejenigen Hauptkomponenten zu verwenden, deren Eigenwert λ größer als der Mittelwert aller Eigenwerte ist. Diese Möglichkeit wird als Kaiser-Kriterium bezeichnet [HANDL, 2002]

$$\bar{\lambda} = \frac{1}{p} \sum_{i=1}^p \lambda_i$$

7.1.3 Darstellung im \mathbb{R}^2 :

Die Darstellung erfolgt über einen zusammenhängenden Graphen. Dabei ist zu Achten, dass kein spannender Baum entsteht. Grundlage eines Graphen ist die Verbindung zweier Punkte, welche auch als Kante bezeichnet wird. Sind mehrere Punkte miteinander durch Kanten verbunden so wird diese Darstellung als

Graph bezeichnet. Sind alle Punkte durch Kanten verbunden spricht man von einem zusammenhängenden Graph. Ist ein Punkt Ursprung mehrerer Kanten so wird dies als Kreis bezeichnet. Eine Darstellung ohne Kreis wird als spannender Baum bezeichnet.

[HANDL, 2002]

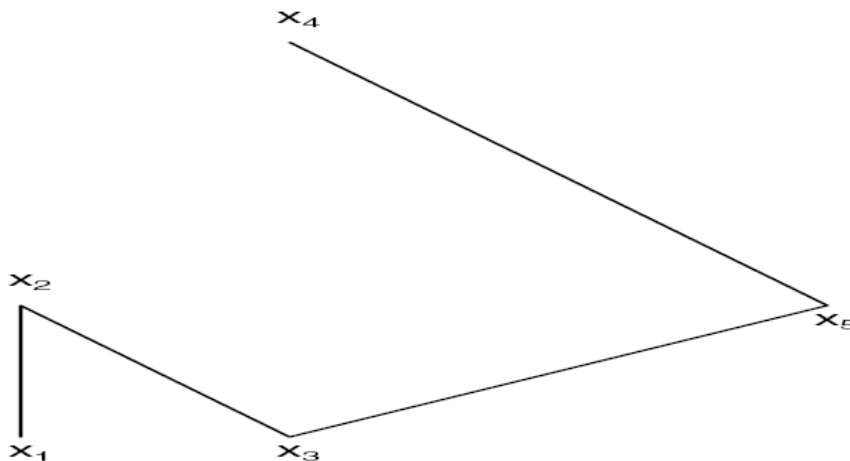


Abbildung 16 Spannender Baum [HANDL, 2002]

7.1.4 Checkliste für die Durchführung einer Hauptkomponentenanalyse

- Vorliegen von hochdimensionalen Daten
- Ausschließlich quantitative Merkmale
- Vorliegen der Daten als Varianz-Kovarianz-Matrix oder Korrelationsmatrix
- Entscheidung welche Matrix für die PCA verwendet wird, an Hand der Varianzen der Merkmale
- Bestimmung der Eigenwerte und Eigenvektoren der verwendeten Matrix
- Berechnung der Anzahl der Hauptkomponenten
- Darstellung in \mathbb{R}^2 , mit minimal spannendem Baum
- Interpretation [HANDL, 2002]

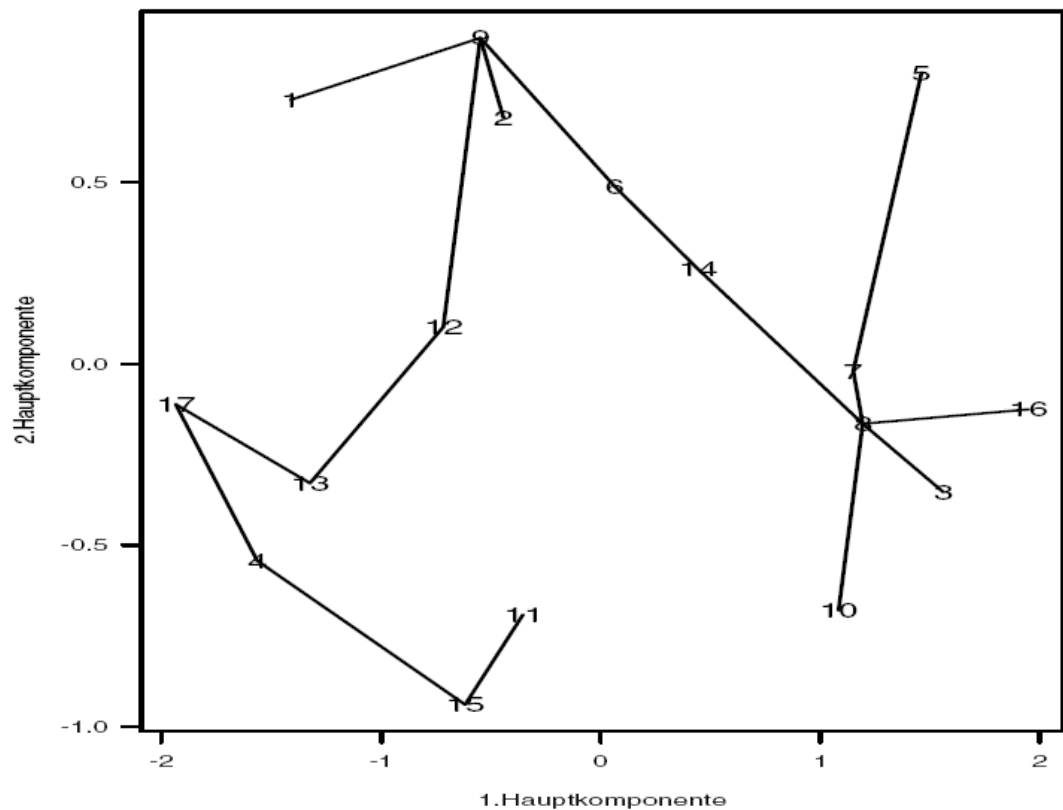


Abbildung 17 Minimal spannender Baum [HANDL, 2002]

Die Summe der Kantengewichte wird für die Beurteilung des spannenden Baums herangezogen. Daher wird unter einem minimal spannenden Baum ein spannender Baum mit kleinster Summe an Kantengewichten verstanden. Dabei wird ausgehend von der kürzesten Kante der Graph erstellt, bis ein spannender Baum entstanden ist. [HANDL, 2002]

7.2 Multidimensionale Skalierung (MES)

Ein möglicher Nachteil der PCA ist die Tatsache, dass sich mit ihr nur quantitative Merkmale darstellen lassen. Sind aber nicht alle Merkmale in einer Daten-

matrix quantitativ, so ist durch die MDS eine zweidimensionale Darstellung möglich. [HANDL, 2002]

Diese Art der Darstellung geht genau wie die PCA von einer Distanzmatrix aus und ermöglicht eine Übertragung der Objekte in einen neuen Raum durch Projektion, wobei die ursprünglichen Distanzen möglichst wenig verzerrt werden sollen. Grundsätzlich ist es möglich die MDS in zwei wesentliche Verfahren zu unterteilen, in die metrische multidimensionale Skalierung und in die nicht metrische multidimensionale Skalierung. [LEYER und WESCHE, 2007]

7.2.1 Metrische Mehrdimensionale (Multidimensionale) Skalierung:

Diese Methode wird auch als Hauptkoordinatenanalyse bezeichnet (principal coordinate analysis, PCO oder PCoA). Sie wird auf Grund ihrer Ähnlichkeit zur PCA auch als verallgemeinerte Form dieser verstanden. Bei dieser Methode werden Projektionen gesucht die eine möglichst unverzerrte Abbildung im neuen Raum ermöglichen. Daher spricht man von einer metrischen Skalierung. Verglichen mit der PCA ist diese Methode in den einzelnen Schritten, vor allem in der Berechnung nicht unähnlich, allerdings liegt dieser Methode ein anderes Modell zu Grunde. [LEYER und WESCHE, 2007]

Das Ziel der MDS ist ähnlich der PCA eine Darstellung von Objekten in einem zweidimensionalen Raum. Daher zählt auch diese Methoden zu den dimensionsreduzierenden Maßnahmen. Im Gegensatz zur Hauptkomponentenanalyse, bei der eine Datenmatrix aus quantitativen Merkmalen der Ausgangspunkt ist, basiert die mehrdimensionale oder multidimensionale Skalierung auf einer Distanzmatrix. Bei Bedarf ist es mathematisch möglich aus einer Datenmatrix eine Distanzmatrix zu generieren. [HANDL, 2002]

Ausgehend von einer Datenmatrix [HANDL, 2002]

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

werden die quadratischen euklidischen Distanzen d_{rs}^2 zwischen den Zeilenvektoren nach folgender Formel:

$$d_{rs}^2 = \sum_{j=1}^p (x_{rj} - x_{sj})^2 = \sum_{j=1}^p x_{rj}^2 + \sum_{j=1}^p x_{sj}^2 - 2 \sum_{j=1}^p x_{rj}x_{sj}$$

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs}$$

Mathematisch lässt sich die multidimensionale Skalierung in 4 Schritte unterteilen: [HANDL, 2002]

- Bestimmung der Matrix $A = (a_{rs})$, für die gilt:

$$a_{rs} = -0.5 d_{rs}^2$$

- Bestimmen der Matrix $B = (b_{rs})$, für die gilt:

$$b_{rs} = a_{rs} - \frac{1}{n} \sum_{r=1}^n a_{rs} - \frac{1}{n} \sum_{s=1}^n a_{rs} + \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n a_{rs}$$

Oder vereinfacht:

$$b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}$$

- Bestimmung der Eigenwerte λ der Matrix B, sowie die entsprechenden Eigenvektoren.
- Durch eine Spektralzerlegung von B und Bildung der Diagonalmatrix Λ kann anhand der dazugehörigen Eigenwerte festgestellt werden, ob die Darstellung in \mathbb{R}^2 korrekt ist. [HANDL, 2002]

7.1.2 Nichtmetrische Mehrdimensionale Skalierung:

Hierbei sind nicht die Distanzen an sich, sondern die Ordnung der Distanzen entscheidend. [HANDL, 2002]

Ausgehend von einer Unähnlichkeitsmatrix lässt sich die Rangfolge von paarweisen Abständen errechnen. [LEYER und WESCHE, 2007]

Diese Unähnlichkeits- oder Distanzmatrix wird mit Δ bezeichnet.

$$\Delta = \begin{pmatrix} \delta_{11} & \dots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \dots & \delta_{nn} \end{pmatrix}$$

Daraus ergeben sich die so genannten Monotoniebedingungen. Diese stellen die Ordnung der Elemente unterhalb der Hauptachse in der oben gezeigten Matrix. [HANDL, 2002]

Bei der nichtmetrischen multidimensionalen Skalierung (NMDS) wird nicht die absolute Distanz als Maßzahl herangezogen. Entscheidend ist nur die Einhaltung der Monotoniebedingungen in den Beziehungen. Daher spricht man bei der MDS von einem rangbasiertem Verfahren. Das Prinzip der Monotonie ist so zu verstehen, dass der niedrigste Rang mit dem kleinsten Abstand in der ursprünglichen Matrix einhergeht. Dieser paarweise Abstand mit dem niedrigsten Rang soll auch im neuen Raum den kleinsten Wert besitzen. Der nächst höhere Rang, besitzt im neuen Raum daher idealerweise den nächst höheren Wert. [LEYER und WESCHE, 2007]

7.3 Faktorenanalyse

Auch diese Methode ist der PCA nicht unähnlich, daher wird der Begriff der Faktorenanalyse oftmals als Synonym für die PCA verwendet. Dieser Umstand ist allerdings nicht korrekt auch wenn beider Methoden in mathematisch in vielen Teilen ident sind. Im Gegensatz zur Hauptkomponentenanalyse, bei welcher nur die gemeinsame Varianz von Bedeutung ist und die eigenständige Varianz nicht berücksichtigt wird, ist die eigenständige Varianz bei der Faktorenanalyse sehr wohl ein wichtiger Punkt. So gesehen kann man nicht von einem Synonym sprechen, aber die Faktorenanalyse als erweiterte Form der PCA oder umgekehrt die PCA als vereinfachte Faktorenanalyse sehen. Die Entscheidung für oder gegen eine Faktorenanalyse oder PCA ist primär von der Fragestellung abhängig, auch wenn die Faktorenanalyse praktisch eine eher untergeordnete Rolle einnimmt. [LEYER und WESCHE, 2007]

7.4 Clusteranalyse

Ziel einer Clusteranalyse ist das Zusammenfassen von Objekten zu homogenen Gruppen bzw. die Durchführung einer Gruppeneinteilung. Diese Objekte können Individuen, Aggregate oder Merkmale darstellen. Die Art des Algorithmus der zur Clusteranalyse herangezogen wird hängt von der Art des Objekts ab. [Bacher et al., 2010]

Zwei ganz wesentliche Ansätze im Clustering sind zu erwähnen, die „unsupervised cluster analysis“ und die „supervised cluster analysis“. Bei der ersten passiert die Clusterbildung ohne weitere Informationen, nur auf Grund der Ähnlichkeit der Einzelobjekte. Bei der zweiten Methode werden zusätzliche Gruppierungsparameter herangezogen. Im Gegensatz zur PCA werden Clusterverfahren auch bei Genexpressionsdaten verwendet. [REPENNING et al., 2006]

Im Wesentlichen kann das Clustering in hierarchische Verfahren und partitionierende Verfahren unterschieden werden:

Alle Arten der Clusterbildung basieren auf einem der folgenden Verfahren:

- Nächste-Nachbarn-Verfahren: Single-Linkage, Complete-Linkage
- Mittelwertmodelle: Average-Linkage
- Clusterzentren
- Klassifikationsobjekte : K-Means [BACHER et al., 2010]

7.4.1 Hierarchische Clusteranalyse:

Hierarchische Clusteranalysen sind gruppenbildende Verfahren, bei denen die errechneten Gemeinsamkeiten als Dendogramm visualisiert werden. Da dies eine sehr flexible Darstellungsmöglichkeit ist, ist bei einem Datensatz immer

mehr als ein Dendrogramm möglich, ohne dass von vorne herein bekannt ist welches das richtige ist. [LEYER und WESCHE, 2007]

Als hierarchische Verfahren werden verschiedene Algorithmen verstanden, die zu unterschiedlichen Dendogrammen führen:

- Single-Linkage-Verfahren: Dieses Verfahren wählt die Distanz zwischen zwei Klassen die verschmolzen werden sollen so, dass die kleinste Distanz der Elemente der beiden Klasse gewählt wird.
- Complete-Linkage-Verfahren: Hierbei wird die größte Distanz zwischen den Elementen der beiden Klassen verwendet.
- Average-Linkage-Verfahren: Bei diesem Verfahren werden alle Distanzen zwischen den Elementen beider Klassen berücksichtigt und der Mittelwert zur Klassenverschmelzung herangezogen. [HANDL, 2002]

Sowohl das Single-Linkage-Verfahren (Methode des nächsten Nachbarn) als auch das Complete-Linkage-Verfahren (Methode des weitest entfernten Nachbarn) gehören der Klasse der Nächste-Nachbarn-Verfahren an. Hierbei erfolgt die Clusterbildung folgendermaßen:

- Jedes zu klassifizierende Objekt hat eine bestimmte Zahl nächster Nachbarn innerhalb des Clusters in dem es selbst ist.
- Die Bezeichnung Nachbar trifft dann zu, wenn sich zwei Klassifikationsobjekte ähnlich sind.

- Jedes zu klassifizierende Objekt hat mindestens einen weiter entfernten Nachbarn, also einen der weniger ähnlich ist. [BACHER et al., 2010]

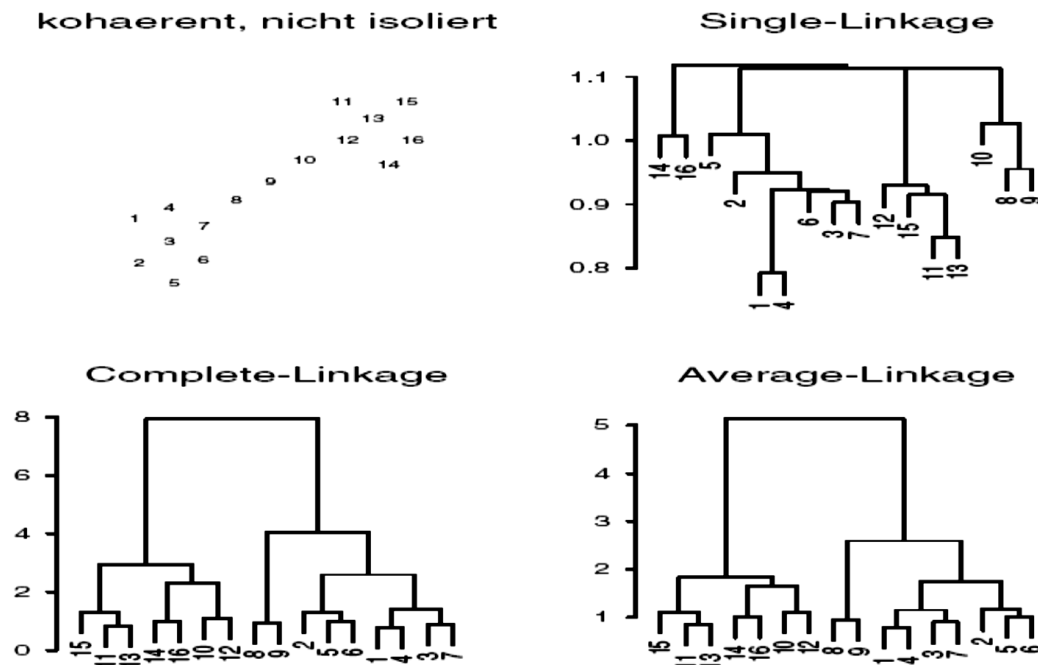


Abbildung 18 Verschiedene Dendogramme [HANDL, 2002]

7.4.2 Partitionierte Verfahren:

Objekte verbleiben sobald sie verschmolzen oder geflüstert wurden innerhalb einer Klasse, zumindest bei hierarchischen Verfahren. Dies ist bei partitionierten Verfahren nicht der Fall. Da wichtigste partitionierte Verfahren ist der k-Merans Algorithmus. [HANDL, 2002]

Bei diesem Verfahren wird die Clusterzahl k bereits vorher definiert. Dafür sind zusätzliche Informationen notwendig (Supervisor Cluster Analysis). Hierbei werden willkürlich Gruppen gebildet und durch Verschiebeprozesse die optimalen Cluster bestimmt. Dabei sollen Unterschiede zwischen den Objekten in einer Gruppe möglichst gering und die Unterschieden zwischen den einzelnen

Gruppen möglichst hoch sein. Das Hauptkriterium ist das Verhältnis der Varianz innerhalb der einzelnen Gruppen zur Varianz zwischen den Gruppen. Gegenüber anderen Clusteranalysen hat dies den Vorteil, dass eine nachträgliche Fehlerkorrektur noch möglich ist und die Cluster hier kein starres System darstellen. [LEYER und WESCHE, 2007]

7.4.3 Interpretation von Dendogrammen

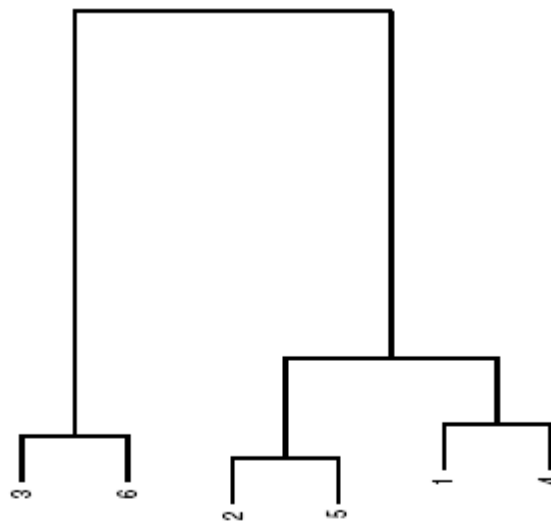


Abbildung 19 Complete Linkage Dendrogramm
[HANDL, 2002]

Die Interpretation eines Dendogramms erfolgt nach folgendem Prinzip:

Die Leserichtung ist immer von unten nach oben. Die Objekte, die sich am ähnlichsten sind werden zu einem Cluster verschmolzen (z. B. 3+6, 2+5 oder 1+4 im oben gezeigten Beispiel). Theoretisch könnte in diesen Cluster noch ein Objekt eingefügt werden. Die nächste Ebene zeigt die Zusammenfassung zweier Cluster, die zuvor gebildet wurden, wieder wegen der größten Ähnlichkeit. [Bacher et al., 2010]

7.4.4 Anwendung von Clusterbildung in der Praxis

In der Praxis sind Dendogramme wesentlich komplexer, als bisher beschrieben. Des Weiteren ist zu beachten, dass die Clusteranalyse häufig nicht isoliert angewandt wird. Ein gutes Beispiel ist die Studie von Bernini et al., bei welcher zwischen 2005 und 2008 Urinproben zur Erstellung von metabolischen Fingerprints verwendet wurde. Diese Studie verwendete unterschiedliche Methoden der multivariaten Statistik zur statistischen Aufbereitung der NMR Daten. Dabei kamen sowohl die PCA zur Dimensionsreduktion, die Kanonische Analyse, sowie eine hierarchische Clusteranalyse zur Anwendung. Das Endergebnis konnte in bekannter Weise als Dendogramm dargestellt werden. [BERNINI et. al, 2009]

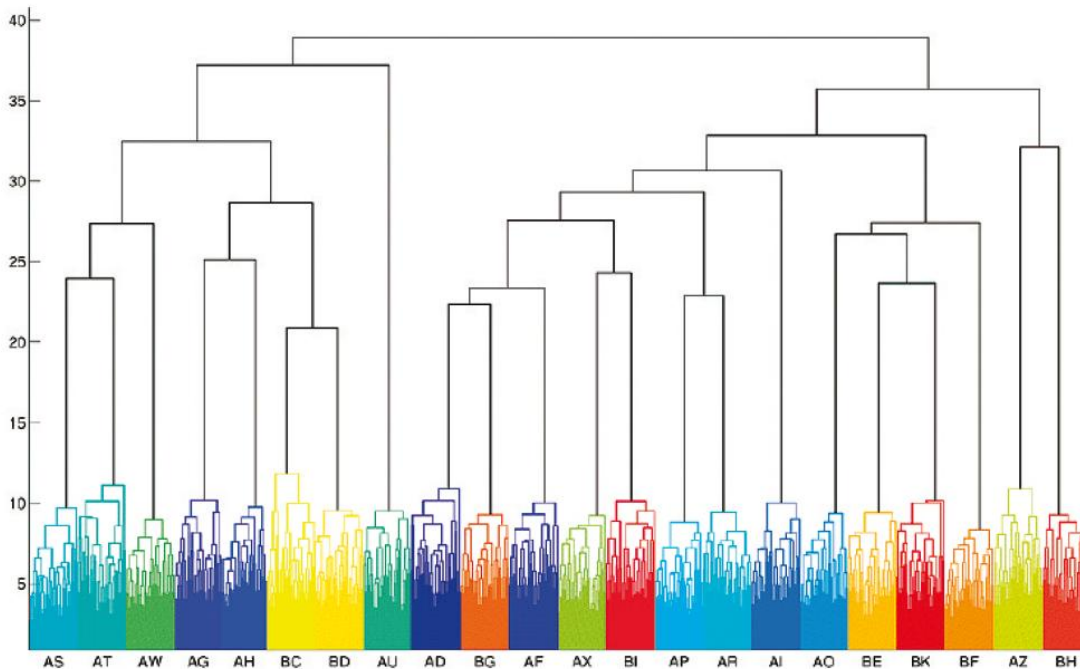


Abbildung 20 Dendogramm nach PCA [BERNINI et al., 2009]

7.5 Partial Least Squares Analyse

Die Partial Least Squares Analyse, kurz PLS genannt, gehört genauso wie die PCA zu Methoden der Datenreduktion. Des Weiteren werden beide den chemometrischen Methoden zugeordnet. Die Chemometrie hat sich aus der Notwendigkeit entwickelt, dass das Datenvolumen bei chemischen Analysen stark angestiegen ist, ein Trend der seit rund 30 Jahren immer stärker zunimmt. Daher kann die Chemometrie als mathematisches Teilgebiet der Chemie angesehen werden. Mit der Anzahl der Variablen, oft mehr als 100 oder 200, steigt auch die Schwierigkeit der Interpretation. [KETTANEH et al., 2005]

Im Gegensatz zur verwandten PCA ist die PLS eine „supervised“ Methode.

7.5.1 Anwendung des PLS in naturwissenschaftlichen Studien

Auch wenn es wesentlich mehr Anwendungen von PLS im wirtschaftlichen Bereich gibt, so gibt es immer wieder auch naturwissenschaftliche Studien, die PLS anwenden, wie die Studie von SONG und OTTO beweist. Im Rahmen von Lebensmitteluntersuchungen wurden wertbestimmende Komponenten in Wurstwaren untersucht. Für die Kalibration und Überprüfung der Schätzgenauigkeit kamen sowohl PCA wie auch PLS zum Einsatz. [SONG und OTT, 1995]

Eine Studie aus 2006 von XIE et al., zeigt, dass die PLS für NIR (nahes Infrarot) als auch für HPLC Studien gleichermaßen geeignet ist. In dieser Studie wurde der Gehalt verschiedener Zucker in Früchten des Lorbeerbaums bestimmt. Für die Entwicklung von geeigneten Modellen zur Quantifizierung wurde die PLS Regression herangezogen. [XIE et al., 2009]

Auch die PLC findet Anwendung im Bereich des Metabolic Profilings auf Basis der NMR. Hierbei findet eine Abwandlung der klassischen PLS die PLS-DA (partial least square discriminant analysis) Anwendung. Diese Methode kann nach vorangehender PCA als Methode zum Supervised Clustering verwendet werden. Die Studie von Ramadan et. al. Zeigt, dass PLS-DA eine Clusterbildung erleichtert auch wenn eine vorangehende Datenreduktion mittels PCA ihr Ziel verfehlt hat. [RAMADAN et al, 2006]

8 Schlussbetrachtungen

Bioinformatische Methoden sind auf Grund ihrer vielseitigen Anwendungsmöglichkeiten aus der ernährungswissenschaftlichen Forschung nicht mehr wegzudenken. Alle Teilbereiche der Omics-Wissenschaften, wie Transcriptomics, Proteomics und Metabolomics, sind in der Ernährungswissenschaft von derartiger Wichtigkeit, dass der Begriff der Nutrigenomics eigens dafür geschaffen wurde.

Es gibt bereits eine Vielzahl an Studien aus verschiedenen für Ernährungswissenschaft er interessanten Forschungsgebieten, aber Omics-Technologien und die gesamte bioinformatische Forschung steckt im Bereich der Ernährungswissenschaft noch in den Kinderschuhen.

Einzelne Teilbereiche, wie das Neugeborenen Screening, die Mutter der Metabolomics sind schon weit gediehen. In den meisten anderen bioinformatischen Disziplinen ist das enorme Potential für die Ernährungswissenschaften bekannt. Die Omics-Technologien sind zukünftig vor allen im Bereich Ernährungssicherheit und bei Stoffwechselerkrankungen, wie Diabetes und metabolisches Syndrom von Bedeutung, was durch eine Vielzahl von Studien bestätigt wird. Nachdem das Screening von Neugeborenen mittels multiparametrischen Hochdurchsatzmethoden wie MS/MS etabliert wurde, ist eine Ausdehnung dieser Verfahren zur Diagnose und zur Kontrolle des Therapieverlaufs anderer Stoffwechselerkrankungen, des Aminosäure-, Lipid- oder Glukosestoffwechsel, das Ziel.

Auch andere ernährungswissenschaftliche Forschungsschwerpunkte können von diesen Techniken profitieren. So gibt es bereits Studien zur Versorgung mit Vitaminen und Spurenelementen und ihre Auswirkung auf die Genexpression.

Da bioinformatische Untersuchungsmethoden in der Ernährungswissenschaft ein sehr rasch wachsendes Forschungsgebiet sind, ist man bestrebt immer

neue Einsatzmöglichkeiten zu finden. Dieser Umstand zeigt sich durch die große Anzahl an wissenschaftlichen Publikationen im Bereich ernährungswissenschaftlicher Forschung, welche bioinformatische Methoden wie Microarrays, PCR, MS/MS, NMR und PCA zum Einsatz bringt.

Vorteile bioinformatischer Methoden zeigen sich in der Schnelligkeit der Analysen und der Tatsache, dass eine Vielzahl an Parallelanalysen möglich ist. Dieser Umstand führt auch zur Bezeichnung Hochdurchsatzmethoden. Dadurch, dass viele Parallelanalysen möglich sind, führt aber zu einer enormen Datenflut. Diese Datenflut kann in der Ernährungswissenschaft nur durch Software gestützte Auswertung mittels geeigneten Statistikprogrammen ausgewertet werden.

Die enorme Datenflut als Folge von Hochdurchsatzmethoden hat dazu geführt dass Methoden zur Datenreduktion oftmals unerlässlich sind. Es gibt eine Vielzahl Methoden, die diesen Zweck erfüllen. Besonders die PCA ist im Bereich der ernährungswissenschaftlichen Omics mittlerweile als etabliert zu betrachten.

Der Datentransfer und die Bewertung der Analyseergebnisse sind ohne Datenbanken für die einzelnen Omics-Disziplinen nicht möglich. Daher wurden und werden Standards für den Datenaustausch geschaffen. Diese Datenstandards wie MIAME für Microarray Daten und MIAPE für Proteomics Forschung, sollen ein einheitliches Datenhandling gewährleisten und einen erleichterten Datenaustausch ermöglichen. Erst diese Datenstandards und die Datenbanken ermöglichen den Vergleich von Daten aus verschiedene ernährungswissenschaftlichen Studien.

Schlussendlich kann man nur sagen, dass die Omics-Technologien mit ihrer Vielzahl an Einsatzmöglichkeiten und ihrer raschen Entwicklung die Zukunft der ernährungswissenschaftlichen Forschung darstellt. Das diese Technologie in den Ernährungswissenschaften allerdings noch in den Kinderschuhen steckt, ist

noch viel Forschungsbedarf vorhanden um alle möglichen Einsatzgebiete zu bestimmen. Trotz des rasanten Wachstums dieser Technologien und den enormen Fortschritten in der Anwendung wird erst mit der Zeit klar werden, in welchen Gebieten der Ernährungsforschung sich diese Technologien als Standardmethode etablieren werden. Erst durch die Verbesserung der Datenauswertung und des Datenhandlings, durch die Verbesserung der Datenstandards und die damit verbundene Erleichterung im Datenaustausch ermöglicht eine optimale Etablierung von bioinformatischen Methoden in der Ernährungsforschung und somit den Routineeinsatz.

9 Quellenverzeichnis:

ASHMAN K, GAUSEPOHL H. Automatische Proteinidentifizierung. Biospektrum – Spezialheft: Proteomics & Drug Development, 2002, 8, 525-526

BACHER J, PÖGE A, WENZIG K. Clusteranalyse – Anwendungsorientierte Einführung in Klassifikationsverfahren, Oldenbourg Verlag, München, 2010

BALL C, BRAZMA A. MGED Standarts: Work in Progress. OMICS A Journal of Integrative Biology, 2006, 10 (2), 138-144

BENDER R, GROUVEN U, ZIEGLER A. Varianzanalyse für Messwertwiederholungen. Dtsch Med Wochenschr, 2007, 132, 61-64

BERNINI P, BERTINI I, LUCHINAT C, NEPI S, SACCENTI E, SCHÄFER H, SCHÜTZ B, SPRAUL M, TENORI L. Individual Human Phenotypes in Metabolic Space and Time. Journal of Proteome Research, 2009

BERREDE L, GARCIA A, CAMARERO J. Protein Microarray – Novel Developments and Applications. Pharm Res, 2010

BLÜGGEL M. Bioinformatik für die Proteomanalyse. Biospektrum – Spezialheft: Proteomics & Drug Development, 2002, 8, 468-488

BÖCKENHAUER H-J, BONGARTZ D. Algorithmische Grundlagen der Bioinformatik – Modelle, Methoden und Komplexität, Teubner, Wiesbaden, 2003

BÖCKER J. Spektroskopie, Vogel, Würzburg, 1997

BOYCE K, KRIETE A, NAGATOMI S, KELDER B, COSCHIGANO K, KOPCHICK J. Phenotypical Enrichment Strategies for Microarray Data Analysis Applied in a Type II Diabetes Study. Omics A Journal of Integrative Biology. 2005, 9 (3), 251-265

CRUJEIRAS a, PARRA D, GOYENECHEA E, ABETE I, MARTINEZ J. Tachyphylaxis effects on postprandial oxidative stress and mitochondrial-related gene expression in overweight subjects after a period of energy restriction. *Eur J Nutri*, 2009, 48, 341-347

D'ANDREA M, COISSON J, LOCATELLI M, GARINO C, CERETI E, ARLORIO M. Validating allergen coding genes (Cor a 1, Cor a 8, Cor a 14) as target sequences for hazelnut detection via Real-Time PCR. *Food Chemistry*, 2010, 124, 1164-1171

DANIEL H, RIST M, RUBIO-ALIAGA I. Nutrition goes omics – Profilierungstechniken in der Ernährungs- und Lebensmittelforschung. *Biospektrum*, 2009, 15, 266-270

DETTMER K. Methoden für die Metabolomik. *Nachrichten aus der Chemie*, 2008, 56, 1043-1047

DEUTSCH E, LAM H, AEBERSOLD R. Data analysis and bioinformatics tools for tandem mass spectrometry. *Physiol Genomics*, 2008, 33, 18-25

DING S, ZHANG P, DING E, YIN S, DENG P, GUI W. On the Application of PCA Technique to Fault Diagnosis. *Tsinghua science and Technology*, 2010, 15 (2), 138-144

FAHRMEIR L, KÜNSTLER R, PIGEOT I, TUTZ G. *Statistik*, Springer Heidelberg, 1997

FERREIRA P, CRAVO M, GUERREIRO C, TAVARES L, SANTOS P, BRITO M. Fat intake interacts with polymorphisms of Caspase9, FasLigand and PPARgamma apoptotic genes in modulating Crohn's disease activity. *Clinical nutrition*, 2010, 29, 819-823

FISCHER M, HAASE I. PCR in der Lebensmittelanalytik. *GIT Laborfachzeitschrift*, 2006, 3, 206-209

FRATAMICO P. The Application of "Omics" Technologies for Food Safety Research. *Foodborne Pathogenes and disease*, 2008, 4 (4), 369-370

FU W, STROMBERG A, VIELE K, CAROLL R, WU G. Statistics and bioinformatics in nutritional sciences: analysis of complex data in the era of systems biology. *Journal of Nutritional Biochemistry*, 2010, 21, 561-572

FUCHS M, CICHNA-MARKL M, HOCHEGGER R. Development and Validation of a Real-Time PCR Method for the Detection of White Mustard (*Sinapis alba*) in Foods. *Journal of Agricultural and Food Chemistry*, 2010, 58, 11193-11200

GIBNEY M, WALSH M, BRENNAN L, ROCHE H, GERMAN B, VAN OMMEN B. Metabolomics in human nutrition: opportunities and challenges. *Am J Clin Nutr*, 2005, 82, 497-503

GRIFFITH W, KOAL T, WANG Y, KOHL M, ENOL D, DEIGNER H. Targeted Metabolomics in der Biomarkerforschung. *Angew. Chem.*, 2010, 122, 5554-5575

HANDL A. *Multivariate Analysenmethoden*. Springer, Heidelberg, 2002

HAUBITZ M, FLISER D, HALLER H. Proteomanalyse – eine neue Perspektive für die klinische Diagnostik. *Deutsches Ärzteblatt*, 2004, 101 (21), 1514-1517

HOFFMANN G. Ein mathematisches Modell des Transkriptoms. *Klinische Chemie Mitteilungen*, 2006, 37(3), 64-72

HUANG X, REGNIER F. Differential Metabolomics Using Stable Isotope Labeling and Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry. *Analytical Chemistry*, 2008, 80 (1), 107114

HÜTT M-T, DEHNERT M. *Methoden der Bioinformatik: Eine Einführung*, Springer, Berlin, 214-220

JAHN D, SCHOMBURG D. Quo vadis, Systembiologie, *Biospektrum*, 2009, 15, 810-811

JIANG W, QIU Y, NI Y, SU M, JIA W, DU X. An Automated Data Analysis Pipeline for GC-TOF-MS Metabolomics Studies. *Journal of Proteome Research*, 2010, 9, 5974-5981

JORSTAD T, LANGAAS M, BONES A. Understanding sample size: what determines the required number of microarrays für an experiment?. *Trends in Plant Science*, 2007, 12 (2) 46-50

JÜRGENS M, MENZEL C, SCHULZ-KNAPPE P, ZUCHT H-D. Peptid-Biomarker in Körperflüssigkeiten. *Biospektrum*, 2005, 11, 778-779

KETTANEH N, BERGLUND A, WOLD A. PCA and PLS with very large data sets. *Computational Statistics & Data Analysis*, 2005, 48, 69-85

KIM Y, PARK T. DNA microarrays to define and search for genes associated with obesity. *Biotechnol. J.*, 2010, 5, 99-112

KLOSE, J. Proteomanalyse. *Biospektrum*. 2002, 8, 91-94

LALAM N, JACOB C, JAGERS P. Estimation of the PCR efficiency based on a size-dependent modelling of the amplification process, *C.R. Acad. Sci*, 2005, 341, 631-634

LEE J. *Statistical bioinformatics – For Biomedical and life Science Researchers*. Wiley-Blackwell, Hoboken, 2010, 180-181

LEYER I, WESCHE K. *Multivariate Statistik in der Ökologie*, Springer, Berlin, Heidelberg, 2007

LIN W-J, HSUEH H-M, CHEN J. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 2010, 11 (48), 1-9

LIU-STRATTON Y, ROY S, SEN C. DNA microarray technology in nutraceutical and food safety. *Toxicology Letters*, 2004, 150, 29-42

LOTTSPREICH F. Eine neue Methode zur quantitativen Proteomanalyse: Isotope Coded Protein Label – ICPL. In: Highlights der Proteomforschung, BMBF und Forschungszentrum Jülich, Jülich – Berlin, 2006, 6-7

LU B, PANDOLFO L. Quasi-objective nonlinear principal component analysis. Neural Networks, 2011, 24, 159-170

NESVIZHISKII A, AEBERSOLD R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. DDT, 2004, 9 (4), 173-181

NIELSEN J, OLIVER S. The next wave in metabolome analysis. Trends in Biotechnology, 2006, 23 (22), 544-546

NOBELI I, THORNTON J. A bioinformatician's view of the metabolome. BioEssays, 2006, 28, 534-545

ORESIC M. Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. Nutrition, Metabolism & Cardiovascular Diseases, 2009, 19, 816-824

PATTERSON S. Data Analysis – the Achilles heel of Proteomics. Natur Biotechnology, 2003, 21, 221-222

PATTERSON S, AEBERSOLD R. Proteomics: the first decade and beyond. nature genetics supplement, 2003, 33, 311- 323

PFAFFL M. Real-time RT-PCR: Neue Ansätze zur exakten mRNA Quantifizierung, Biosepektrum, 2004, 10 (1), 92-95

PORSCHESKI P, STEINERT K. Eine Array-Technologie-Plattform zur parallelen Gen- und Proteinexpressionsanalyse. In: Highlights der Proteomforschung, BMBF und Forschungszentrum Jülich, Jülich – Berlin, 2006, 40-41

QUACKENBUSH J. Data reporting standarts: making the things we use better. Genome Medicine. 2009, 1 (111)

RAMADAN Z, JACOBS D, GRIGOROV M, KOCHHAR S. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta*, 2006, 68, 1683-1691

REPENNIG C, RECK M, STAHL F, SCHEPER T, HITZMANN B. Aufgaben und Verfahren zur Microarray-Analyse, *Biospektrum*, 2009, 15, 646-648

RÖHRIG B, DU PREL J-P, WACHTLIN D, KWIECIEN R, BLETTNER M. Fallzahlplanung in klinischen Studien. *Deutsches Ärzteblatt*, 2010, 107 (31-32), 552-556

ROY S, SEN C. cDNA-Microarray screening in food safety. *Toxicology*, 2006, 221, 128-133

SANDER J, JANZEN N, SANDER S, MELCHIORS U, STEUERWALD U. Tandemmassenspektrometrie – Beitrag zum Neugeborenenenscreening auf angeborene Störungen des Stoffwechsels. *Monatsschr Kinderheilkd*, 2000, 148, 771-777

SCHLÜTER H. MES – Ein Massenspektrometrie-unterstütztes Screeningsystem zur Enzymsuche. In: *Highlights der Proteomforschung, BMBF und Forschungszentrum Jülich*, Jülich – Berlin, 2006, 44-45

SCHMIDT A, PICOTTI P, AEBERSOLD R. Proteomanalyse und Systembiologie. *Biospektrum*, 2008, 14, 44-46

SCHÖBER D. Microarrays, Genexpressionsanalyse und Bioinformatik. *Biospektrum*, 2002, 8,

SHAW D, HALL W, JEFFS N, WILLIAMS C. Comparative effects of fatty acids on endothelial inflammatory gene expression. *Eur J Nutri*, 2007, 46, 321-328

SPRENGER R, JENSEN O. Proteomics and the dynamic plasma membrane: Quo vadis. *Proteomics*, 2010, 10, 3997 – 4010

SONG C, OTTO R. Schnellbestimmung von wertbestimmenden Komponenten in Wurst mittels Nahe-Infrarot-Transmissionsspektroskopie. *Z Lebensm Unters Forsch*, 1995, 201, 226-229

SUDE R. mRNA transcripts as molecular biomarkers in medicine and nutrition. *Journal of Nutrition Biochemistry*, 2010, 21, 665-670

TAYLOR C, PATON N, LILLEY K, BINZ P, JULIAN R, JONES A, ZHU W, APWEILER R, AEBERSOLD R, DEUTSCH E, DUNN M, HECK A, LEITNER A, MACHT M, MANN M, MARTENS L, NEUBERT T, PATTENSON S, PING P, SEYMOUR S, SOUDA P, TSUGIA A, VANDEKERCKHOVE J, VONDRISKA T, WHITELEGGE J, WILKENS M, XENARIOS I, YATES J, HERMJAKOB H. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 2007, 25 (8), 887-893

THIELE H, GLANDORF J, HUFNAGEL P. Bioinformatics Strategies in Life Sciences: From Data Processing and Data Warehousing to Biological Knowledge Extraction. *Journal of Integrative Bioinformatics*, 2010, 7 (1), 141-149

TÖWE S, KLEINEIDAM K, SCHLOTTER M. Differences in amplification efficiency of standart curves in quantitative real-time PCR assays and consequences for gene quantification in environmental samples. *Journal of Microbiological Methods*, 2010, 82 (3), 338-341

TOSCO A, FONTANELLA B, DANISE R, CIATIELLO L, GROBER O, RAVO M, WEISZ A, MARUZULLO L. Molecular bases of copper and iron deficiency-associated dyslipidemia: a microarray analysis of the rat intestinal transcriptome. *Genes Nutr*. 2010, 5, 1-8

UEFFING M. Funktionelle Proteomanalyse von Membranproteinen mit Free Flow Elektrophorese. In: *Highlights der Proteomforschung, BMBF und Forschungszentrum Jülich*, Jülich – Berlin, 2006, 16-17

VERGÈRES G, SAGAYA F. Nutrigenomik: Sience oder Fiction. *ALP Science*, 2007, 503, 2-11

VICTOR A, KLUG S, BLETTNER M. cDNA-Microarrays – Strategien zur Bewältigung der Datenflut. Deutsches Ärzteblatt, 2005, 102, 355-360

VOGESSER M, KOBOLD U, SEIDEL D. Massenspektrometrie in der Medizin - Stellenwert der molekularen Analytik. Deutsches Ärzteblatt, 2007, 104 (31-32) 2194-220

WATTENBERG A, BLÜGGEL M. Automation von Proteomanalysen. Biospektrum – Spezialheft: Proteomics & Drug Development, 2002 ,8, 521-522

WEINBERGER K. Einsatz von Metabolomics zur Diagnose von Stoffwechselkrankheiten. Therapeutische Umschau, 2008, 65, 487-491

WEINBERGER K, GRABER A, KATZENBERGER J. Targeted Metabolomics. Biospektrum, 2006, 12 (2), 231-232

YAMAMOTO H, YAMAJI H, ABE Y, HARADA K, WALUYO D, FUKUSAKI E, KONSO A, OHNO H, FUKUDA H. Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. Chemometrics and Intelligent Laboratory Systems, 2009, 98, 136-142

ZEISEL S H. Nutrigenomics and metabolomics will change clinical nutrition and public health practice: insights from studies on dietary requirements for choline. Am J Clin Nutr , 2007, 86, 542-548

ZHANG X, YAP Y, WEI D, CHEN G, CHEN F. Novel omics technologies in nutrition research. Biotechnology Advances 2008, 26, 169-176

10 Zusammenfassung

Die Bioinformatik ist eine der wichtigsten naturwissenschaftlichen Disziplinen und einer raschen Entwicklung unterworfen, der sich auch die ernährungswissenschaftliche Forschung nicht entziehen kann. In Zeiten wo die ernährungswissenschaftliche Forschung sich nicht mehr mit der Analyse von einzelnen Nährstoffen zufriedengibt und die Omics-Technologien das zentrale Thema sind, ist der Einsatz von bioinformatischen Methoden unerlässlich geworden. Durch die Weiterentwicklung von Analysemethoden und das Zusammenwachsen der unterschiedlichen naturwissenschaftlichen Disziplinen und die damit verbundenen neuen Einsatzmöglichkeiten der verschiedenen Analysemethoden auch in anderen Bereichen werden viele Untersuchungen schneller und effizienter. Eine Folge der Entwicklung der Analysen in der ernährungswissenschaftlichen Forschung und der Etablierung von Hochdurchsatzmethoden wie die Anwendung von Microarrays führen zu einem Zuwachs an Analysedaten, welche die Verwendung von Methoden der multivariaten Statistik notwendig macht. Ziel hierbei ist eine Dimensionsreduktion und eine Datenreduktion ohne die Aussagekraft, der in einer Studie gewonnenen Daten zu verringern. Dabei kommen Methoden wie die PCA, PLS, MDS, Faktorenanalyse oder Clusteranalysen zum Einsatz. Eine Vielzahl an unterschiedlichen Studien aus verschiedenen ernährungswissenschaftlich interessanten Bereichen (Stoffwechselerkrankungen, Biomarkerforschung, Lebensmittelsicherheit, Enzymforschung, Vitaminstatus, Lebensmittelallergien etc.) wurden bereits erfolgreich unter Verwendung von bioinformatischen Methoden und unter Verwendung von Datenreduktionsmaßnahmen durchgeführt.

11 Abstract

Bioinformatics is one of the major scientific disciplines and rapidly changing subject and the nutritional research can't ignore this development. The Aim of nutritional research is no longer the analysis of single nutrients. The so called "omics"-Technologies are the central theme now. Therefore the use of computational methods has become essential. The development of analytical methods and the convergence of different scientific disciplines led to an expanded field of use of individual methods. New applications of analytical methods made nutritional studies faster. On the other hand the establishment of high throughput methods such as the use of microarrays in nutritional research caused more data that must be post-processed by multivariate statistics. The aim in this case is a dimension reduction and also a data reduction, without the loss of information. Methods such as PCA, PLS, MDS, factor analysis and cluster analysis are used for this post-processing. A variety of different studies from different nutritional interesting areas (metabolic diseases, biomarker research, food safety, enzyme research, vitamin status, food allergies, etc.) have already been successfully carried out using computational methods and using data reduction measures.

12 Curriculum Vitae

Persönliche Daten

Name	Bernadette Binder
Geburtsdatum	17.12.1980
Geburtsort	8700 Leoben
Wohnort	2560 Berndorf, Kruppstraße 22
Familienstand	ledig
Nationalität	Österreich
Religionsbekenntnis	röm.-kath.

Schulbildung

1987-1990	Volksschule Leoben Seegraben
1990-1991	Volksschule Berndorf
1991-1999	Gymnasium Berndorf

Hochschulausbildung

1999-2001	Studium der techn. Chemie TU Wien
2001-2004	Humanmedizin Universität Wien bzw. Medizinische Universität Wien
2004-dato	Ernährungswissenschaften Universität Wien

Praktika

11/2009-12/2009
Institut für Ernährungswissenschaften
Projektmitarbeit Lebensmittelkennzeichnung

11/2010-12/2010
Institut für Ernährungswissenschaften
Projektmitarbeit österr. Ernährungsbericht

Berufserfahrung

07/1997
Berndorf Band
Praktikantin

07/1999
Billa
Praktikantin

07/2000-08/2000
Billa
Praktikantin

08/2001
Billa
Praktikantin

07/2002
Österr. Post AG
Urlaubsvertretung

07/2003
Österr. Post AG
Urlaubsvertretung

07/2007-08/2004
Österr. Post AG
Urlaubsvertretung

08/2005	Berndorf Metall- und Bäderbau Praktikantin im Vertrieb
08/2006	Berndorf Metall- und Bäderbau Praktikantin im Vertrieb
08/2007-09/2007	Berndorf Metall- und Bäderbau Praktikantin im Vertrieb
08/2008-09/2008	Berndorf Metall- und Bäderbau Praktikantin im Vertrieb
02/2010-dato	Berndorf Metall- und Bäderbau Assistentin im Marketing