technische universität
dortmund

fakultät statistik

Integrativer Ansatz zur Identifizierung neuer, prognostisch relevanter Metagene mittels Clusteranalyse

Dissertation von Evgenia Freis

zur Erlangung des Doktorgrades der Naturwissenschaften

vorgelegt der Fakultät Statistik

Technische Universität Dortmund

Dortmund, August 2013

(überarbeitet im Oktober 2014)

Gutachter: Prof. Dr. Katja Ickstadt

Prof. Dr. Jörg Rahnenführer

Tag der mündlichen Prüfung: 10. Dezember 2013

Danksagung

Die vorliegende Dissertation entstand zum größten Teil während meiner Tätigkeit als wissenschaftliche Mitarbeiterin an der Fakultät Statistik der Technischen Universität Dortmund. An dieser Stelle mochte ich mich von Herzen bei all denjenigen bedanken, die mich stets unterstützt und an mich geglaubt haben.

Beginnen möchte ich mit Frau Prof. Dr. Katja Ickstadt, der ich ganz besonders für die Idee zu dieser Arbeit und die umfassende fachliche Betreuung danken möchte. Ich danke ihr recht herzlich für ihre Zeit und Geduld, für die hilfreichen Rückmeldungen und ihre konstruktive Kritik.

Weiterhin bedanke ich mich beim Herrn Prof. Dr. Jörg Rahnenführer für die Betreuung meiner Dissertation als Zweitgutachter, seine ständige Hilfsbereitschaft und fachliche Aufgeschlossenheit.

Mein Dank gilt auch Herrn Prof. Dr. Jan Hengstler, der mir mit seinem medizinischen Fachwissen zur Seite stand. In diesem Zusammenhang bedanke ich mich auch bei den Mitarbeiterinnen des IfADo, Frau Silvia Selinski und Frau Cristina Cadenas für ihre Unterstützung.

Audrey Qiuyan Fu und Bettina Grün sei für die zahlreichen Hilfestellungen gedankt.

Der ganzen Fakultät Statistik der TU Dortmund möchte ich für ein sehr angenehmes Arbeitsklima und die ständige Hilfsbereitschaft danken, insbesondere Arno Fritsch, André König, Leo Geppert, Marco Grzegorczyk, Sabrina Herrmann, Kai Kammers, Claudia Köllmann und Imke Tempelmann für ihre hilfreichen Tipps und Anregungen.

Außerdem bedanke ich mich bei Frau Nicole Gerling und Herrn Alexander Pogoster von Arvato Infoscore für ihre Geduld und Verständnis.

Mein ganz besonderer Dank gilt meiner Familie; vor allem meiner Mutter, die mir stets Mut zugesprochen hat, meinem Mann, der mir immer den Rücken freigehalten hat, und meiner Tochter, die in dieser Zeit öfters auf die Spaziergänge mit ihrer Mutter verzichten musste. Ohne euch wäre meine Arbeit in dieser Form nicht möglich gewesen.

Abbildungsverzeichnis

Tabellenverzeichnis

Inhaltsverzeichnis

1 Einleitung

So eine Arbeit wird eigentlich nie fertig. Man muss sie für fertig erklären,

wenn man nach Zeit und Umständen das Möglichste getan hat.

(J.W. Goethe, Italienische Reise, Caserta, den 16. März 1787)

Brustkrebs ist die zweithäufigste Krebserkrankung bei Frauen mit mehreren Tausend Neuerkrankten jährlich. Das mittlere Alter liegt dabei aktuell bei nur 65 Jahren und die betroffenen Frauen werden leider immer jünger. Obwohl die Medizin kontinuierlich Fortschritte auf diesem Gebiet macht, versterben viele Patientinnen an den Folgen dieser Erkrankung. Eine der Ursachen ist die verspätete Diagnose, denn durch frühzeitiges Erkennen der Karzinome können diese meist geheilt werden. Damit das erhöhte Brustkrebsrisiko frühzeitig erkannt und somit gezieltere Therapien für die betroffenen Patientinnen ermöglicht werden können, wurden in den vergangenen Jahren und werden weiterhin zahlreiche klinische Studien zur Aufdeckung potentieller Risikofaktoren durchgeführt.

Ein Aspekt dabei ist die Untersuchung der Gene. Es wurde erkannt, dass ihre Veränderungen zwar nicht zwangsläufig zum Ausbruch einer Krankheit führen, ihre

Expressionen jedoch näher analysiert werden sollten, um ein Karzinom frühzeitig erkennen zu können. Dabei helfen eine Zuordnung von Genen zu den bekannten Funktionsgruppen und die Simulation von krebsrelevanten Ereignissen bei der Identifizierung von biologischen Prozessen, die klinisch relevant sind.

Mehrere Ansätze wurden schon erfolgreich vorgestellt und umgesetzt. In einigen davon wurden Gene mit ähnlichen Expressionen zu den sogenannten Metagenen zusammengefasst, die sich als prognostische Faktoren für das Überleben der Patienten erwiesen haben. Oft liegen dabei Daten mit kurzen Zeitreihen vor. Das Ziel dieser Arbeit ist die Entwicklung eines neuen integrativen Ansatzes zur Clusterung kurzer Expressionszeitreihen zur Identifizierung neuer prognostisch relevanter Metagene.

Dafür werden im ersten Teil des hier vorgestellten Ansatzes humane Mammakarzinom-Zelllinien MCF7 (Michigan Cancer Foundation 7, vgl. Kapitel 2.2) analysiert. Da die Überexpression der onkogenen Variante der Rezeptortyrosinkinase ErbB2 mit einer schlechteren Prognose assoziiert ist (Ménard et al., 2001) und somit Rückschlüsse auf die biologischen Hintergründe der Entstehung der Brustkrebserkrankung ermöglichen könnte, wurde sie in diesen MCF7-Zelllinien induziert und zu sechs Zeitpunkten nach der Induktion beobachtet. Mit verschiedenen Clustermethoden werden hier Gengruppen mit ähnlichen Expressionsverläufen identifiziert, für die mit Hilfe der Gene Ontology (auch GO genannt) - bzw. Promoteranalyse die biologisch interessantesten ermittelt werden. Zur Verifizierung der hier angewendeten Methoden wird ein weiterer Datensatz mit Expressionswerten kurzer Zeitreihen erfolgreich herangezogen.

Im zweiten Teil des Ansatzes werden für diese Gengruppen Metagene gebildet. Diese werden auf ihre prognostische Relevanz in den Brustkrebsdaten von 766 Patientinnen mittels Überlebenszeitanalyse untersucht, um so neue biologisch und prognostisch relevante Cluster aufzudecken.

Nachdem im Kapitel 2 der Hintergrund und die Datensituation näher erläutert werden, werden im dritten und vierten Kapitel zunächst der neu entwickelte Clusteransatz und die entsprechenden statistischen Methoden vorgestellt, die anschließend in Kapitel fünf zur Identifizierung neuer prognostisch relevanter Metagene angewandt werden. Eine Zusammenfassung der Ergebnisse und ein Ausblick auf die möglichen, an diese Arbeit anknüpfenden Untersuchungen werden im letzten und 6. Kapitel gegeben.

2 Hintergrund und Material

Weltweit erkranken jährlich mehrere Millionen Menschen an einem Karzinom. In Deutschland ist Brustkrebs die häufigste Krebserkrankung bei Frauen mit jährlich mehr als 72 Tausend Neuerkrankten (Robert Koch Institut, 2013). Trotz intensiver medizinischer Behandlungen versterben viele Patientinnen an den Folgen dieser Erkrankung. Eine der Ursachen dafür ist die verspätete Diagnose, denn rechtzeitig erkannte Karzinome sind in den meisten Fällen heilbar. Da von Krebs insbesondere ältere Menschen betroffen sind und unsere Gesellschaft immer älter wird, ist es umso wichtiger, das Krebsrisiko frühzeitig zu erkennen. Dafür werden zahlreiche klinische Studien durchgeführt. Aufgrund dieser wissenschaftlichen Forschungen, der immer besseren Heilungsmethoden und Vorsorgeuntersuchungen liegt die Überlebensrate bei dieser Erkrankungsart heute bei rund 81 Prozent. Eine Steigerung dieser Erfolgsquote ist weiterhin anzustreben.

In den letzten Jahren hat die Untersuchung der Gene immer mehr an Bedeutung gewonnen. Die Wissenschaftler kommen dabei zu der Erkenntnis, dass die veränderten Gene zwar nicht zwangsläufig zum Ausbruch einer Krankheit führen, deren Expressionen jedoch näher analysiert werden sollten, um ein Karzinom frühzeitig zu erkennen und dadurch bessere Therapien zu gewährleisten.

Eine Zuordnung von Genen zu den entsprechenden GO-Gruppen hilft, biologische Prozesse zu identifizieren, die unter bestimmten Bedingungen bzw. krebsrelevanten Ereignissen, wie die Onkogen-Überexpression oder die Entwicklung von Metastasen, hoch- oder runterreguliert werden und somit klinisch relevant sind. Die Transkriptionsfaktoren regulieren die Expression ihrer Zielgene und sind bei der Analyse von Genexpressionsdaten nach Möglichkeit ebenfalls zu berücksichtigen.

In Unterkapitel 2.1 erfolgt eine kurze Einführung in die Gene Ontology Datenbank, die detaillierte Informationen über die Gene und Genprodukte enthält. In den weiteren Unterkapitel 2.2 und 2.3 werden die dieser Arbeit zugrunde liegenden humanen Mammakarzinom-Zelllinien und die Brustkrebsdaten von 766 Patientinnen näher beschrieben. Anschließend wird in Unterkapitel 2.4 ein weiterer Datensatz mit kurzen Zeitreihen vorgestellt, der zum späteren Methodenvergleich herangezogen wird.

2.1 Gene Ontology Datenbank

Das Gene Ontology Konsortium hat 1998 (Ashburner et al., 2000) ein System zur vereinheitlichten Beschreibung von Genprodukten über verschiedene Datenbanken aufgebaut. Die gleichnamige biomedizinische Ontologie-Datenbank ist inzwischen weit verbreitet und wird häufig verwendet und ständig weiterentwickelt. Sie ermöglicht neben einer formalen Beschreibung der Daten und deren Zusammenhänge Rückschlüsse daraus zu ziehen, die helfen komplexe Fragestellungen zu lösen.

Die GO bietet drei Ontologien zur kontrollierten, vereinheitlichten und strukturierten Beschreibung von Genen, Genprodukten und -sequenzen:

- Die **molekulare Funktion** (**MF)** beschreibt mit etwa 7825 Termen auf der molekularen Ebene die von den einzelnen Genprodukten ausgeführten Funktionen (z.B. Bindung).
- Die **biologischen Prozesse (BP)**, an denen die Genprodukte sich beteiligen, sind in der gleichnamigen Ontologie mit ca. 13860 Termen zusammengefasst (z.B. Zellwachstum).

- Die **zelluläre Komponente (CC)** umfasst etwa 1993 Terme und dient der Beschreibung der Substrukturen der Genprodukte (z.B. Chromosomen).

Zu den Anzahlen der Terme liegen unterschiedliche Informationen vor. Die hier vorgestellten wurden der Dissertation von Hummel (2009) entnommen.

Die hierarchische Struktur innerhalb von GO ist als gerichteter azyklischer Graph (Directed Acyclic Graph, DAG) mit einer Wurzel organisiert, deren Ausschnitt in Abbildung 2.1.1 dargestellt ist. Der DAG besteht dabei aus Knoten (GO-Bezeichnungen) und Kanten (Teilmengen-Relationen zwischen denen). Die Knoten sind als „Eltern" - und „Kinder" - Knoten zu betrachten und können die Teilmengen von den jeweils anderen Knoten sein. Anhand dieses Graphs ist ersichtlich, dass so für die verschiedenen GO Terme immer speziellere Terme abgeleitet werden können.

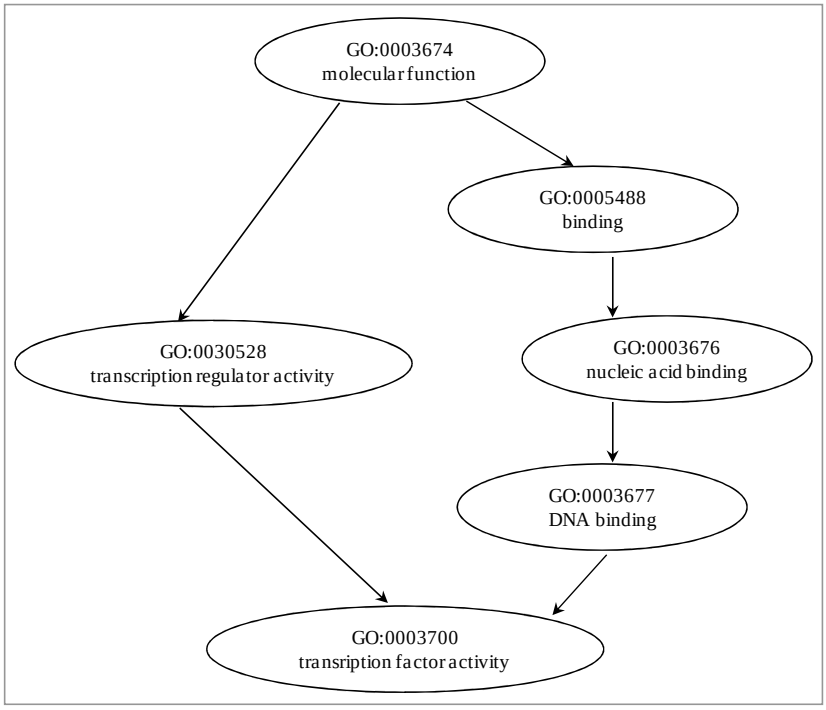

Abb. 2.1.1: Ausschnitt aus dem DAG (aus Hummel, 2009)

Jedes Gen oder Genprodukt kann gleichzeitig zu mehreren Ontologien gehören und wird durch die entsprechende ID (Entrez gene ID) identifiziert. So ist zum Beispiel CDKN1A (cyclin-dependent kinase inhibitor 1A, auch P21 genannt) unter allen drei Genontologien mit der Entrez ID 1026 gelistet, z.B. als „GO:0005737: cytoplasm" in CC, als „GO:0005515: protein binding" in MF oder auch als „GO:0000075: cell cycle

checkpoint" unter den biologischen Prozessen (http://www.genecards.org/). Die entsprechende Zuordnung (Annotation) erfolgt dabei fachmännisch durch erfahrene und qualifizierte Biologen (z.B. Camon et al., 2004) und wird laufend aktualisiert. Zu jeder erfassten ID liegen in der GO-Datenbank weitere Zusatzinformationen wie der Genname, die Probeset-Bezeichnung, mögliche Synonyme oder deren Bedeutung vor, die unter http://www.geneontology.org/ abgefragt werden können.

Da die Zuordnung der Affymetrix Probesets zu den GO Termen möglich ist, bieten sich in R spezielle, vom Bioconductor (Gentleman et al., 2004; www.bioconductor.org) zur Verfügung gestellten Annotationspakete (z.B. *topGO()*, vgl. Alexa und Rahnenführer, 2006), die laufend aktualisiert werden (Hummel, 2009).

Obwohl seit der Gründung der GO-Datenbank im Jahr 1998 mehrere Tausend Gene und Funktionen erfolgreich und vollständig untersucht und erfasst wurden, sind es noch längst nicht alle und das Vokabular wird kontinuierlich mit Informationen zu den weiteren Genprodukten vervollständigt. So kann beispielsweise ein Gen durch seine Entrez ID eindeutig identifiziert werden, durch die Probesets allerdings nicht immer, falls die entsprechende Information fehlt. Somit ist die Vorlage der Entrez ID entscheidend für bestimmte Analysefragestellungen von Microarray Experimenten, um mögliche Rückschlüsse auf biologische Hintergründe ziehen zu können. Bei der Analyse der vorliegenden Daten in Kapitel 5 liegt zu mehreren Probesets keine Entrez ID vor, so dass sie aus dem entsprechenden Analyseteil ausgeschlossen werden (vgl. folgendes Unterkapitel 2.2).

2.2 MCF7/NeuT-Zelllinien

MCF7 steht für Michigan Cancer Foundation 7 und ist eine Zellkultur, die von einer kaukasischen 69-jährigen Frau angelegt wurde. Da sie gut charakterisiert ist und einige Eigenschaften des Brustepithels gut erhalten sind, eignen sie sich sehr gut für Brustkrebsstudien und finden seit 1973 häufig Anwendung in der Krebsforschung (Soule et al., 1973).

ErbB2 (auch HER-2 genannt) ist eine Rezeptortyrosinkinase. Sie wurde in der Krebsforschung als erstes Zielprotein in der Therapie der Krankheit behandelt und hat sich als prognostischer Faktor erwiesen. So ist z.B. ihre Überexpression in den Brustkrebszellen mit einer schlechten Überlebensprognose assoziiert.

Spezielle Expressionssysteme, in denen die ErbB2-Expression gezielt kontrolliert werden kann, werden benutzt, um die Komplexität der damit verbundenen bzw. dadurch ausgelösten Vorgänge genauer untersuchen zu können. So wurde in einer in-vitro-Studie aus den oben erwähnten MCF7-Zelllinien die MCF7/NeuT-Zelllinie erzeugt (vgl. Trost et al., 2005; Hermes, 2007). Um einen Eindruck zu gewinnen, welche Gene zu welchem Zeitpunkt durch die Überexpression von ErbB2 beeinflusst werden, wurde in dieser Zelllinie die Ribonukleinsäure (RNA) zu sechs verschiedenen Zeitpunkten nach Überexpression der onkogenen Variante von ErbB2 (NeuT, welches als Folge einer Punktmutation entstand und eine konstitutiv aktive Variante des ErbB2 Rezeptors darstellt) gewonnen. Als Zeitpunkte wurden 0, 6, 12 und 24 Stunden (h) sowie 3 und 14 Tage (d) gewählt. Bei der anschließenden graphischen Darstellung der Verlaufskurven in Kapitel 5 bzw. im Anhang sollte deswegen beachtet werden, dass die Zeitabstände zwischen den Beobachtungen nicht äquidistant sind. Die Gewinnung der RNA erfolgte dabei in drei aufeinander folgenden Versuchen, so dass für die Genexpressionsanalysen Triplikate zur Verfügung stehen.

Die Analyse der Genexpressionen erfolgte mittels Affymetrix Gene Chip HG-U133-Plus2 und mit Hilfe des Statistik-Softwareprogramms R (Versionen 2.5.0, 2.6.1, 2.11.1 und 2.15.2). Eine detaillierte Beschreibung der Microarray-Versuche ist in der Dissertation von Hermes (2007) zu finden. Die vorliegende MCF7-Datenmatrix enthält für die 54675 Probesets je 18 log_2-transformierte Genexpressionswerte: für jede der drei Triplikate eine Messung zu jeweils 6 Zeitpunkten. Für die in dieser Arbeit vorgestellte Analyse wurde eine Vorauswahl der Probesets getroffen. Mit Hilfe des moderaten t-Tests und unter Kontrolle der False Discovery Rate, adjustiert nach Benjamini/Hochberg (1995), wurden mit Hilfe des R-Pakets *limma* (Version 2.10.5) 54675 Probesets auf 2632 zu mindestens einem der sechs Zeitpunkte differentiell exprimierten Gene beschränkt. Für die Einzelheiten wird an dieser Stelle auf die Diplomarbeit von Krahn (2008) verwiesen.

Die Zuordnung der Probesets zu deren GO-Gruppen erfolgt mit dem R-Paket *topGO* (Version 2.10.0, vgl. Alexa und Rahnenführer, 2006), der zum Zeitpunkt der Auswertung für mehrere Gene keine Entrez ID gefunden hat. Die zu analysierenden Datensätze reduzieren sich um die entsprechende Zeilenanzahl: 1667, 1562 und 1560 für die CC-, BP- bzw. MF-Gruppe.

Wegen der Input-Anforderungen vom R-Paket *GOSim* (Version 1.2.7.7, vgl. Froehlich, 2012), das mit den Gennamen und nicht mit den Probeset-Bezeichnungen arbeitet und für die Berechnung der Ähnlichkeitsmatrix (vgl. Unterkapitel 4.2) eingesetzt wird, werden für die Gene mit den gleichen Entrez IDs die entsprechenden Expressionen zu einem medianen Wert pro Gen zusammengefasst. Die daraus resultierenden Datenmatrizen umfassen dadurch Expressionswerte für 1206 Gene in der CC-Gruppe, 1227 in der BP- und 1157 in der MF-Gruppe.

An dieser Stelle ist ferner zu erwähnen, dass im Folgenden keine Unterscheidung zwischen den Begriffen „Probeset" und „Gen" erfolgt, da dies hier nicht ausschlaggebend ist.

2.3 Brustkrebsdaten

In der vorliegenden Ausarbeitung werden drei Kohorten (einzeln und kombiniert in einer Gesamtkohorte) mit insgesamt 766 lymph-nodal-negativen Brustkrebspatientinnen untersucht und deren metastasenfreie Zeit (MFI) analysiert (Schmidt et al., 2008). Alle Datensätze wurden von der National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) Datenbank (Edgar et al., 2002) mit den zugehörigen GEO Stichprobennummern GSE11121, GSE2034, GSE6532 und GSE7390 heruntergeladen.

Ein Teil der untersuchten Daten sind Ergebnisse einer populationsbasierten Mammakarzinom-Kohortenstudie der Johannes Gutenberg Universität in Mainz, die zwischen 1988 und 1998 durchgeführt wurde. 200 Brustkrebspatientinnen, die noch keine Chemotherapie erhalten hatten und keine befallenen Lymphknoten aufwiesen, wurden in die Studie aufgenommen, operiert und teilweise bestrahlt, erhielten aber keine weiteren Behandlungen wie Immun-, Hormon- oder Chemotherapie.

Das bei den Operationen entnommene Tumorgewebe wurde zur weiteren Analyse zuerst schockgefroren und danach mit Hilfe der Microarray Technologie und dem Affymetrix Gene Chip HG-U133A untersucht. Dabei wurden mehrere klinischen Variablen und die zugehörigen RNA-Expressionen zu 22283 Probesets erhoben (auf die Aufbereitung der Proben wird an dieser Stelle nicht näher eingegangen, da dies nicht der Gegenstand vorliegender Ausarbeitung ist; die Einzelheiten dazu können jedoch Schmidt et al. (2008) entnommen werden). Eine detaillierte Übersicht über die erhobenen klinischen Variablen mit den jeweiligen Definitionen kann in der Arbeit von Köhne (2008) nachgeschlagen werden. An dieser Stelle werden nur die wichtigsten Variablen vorgestellt, die als Risikofaktoren für die Krebsentstehung schon bekannt sind und dementsprechend bei der in Kapitel 3.4 vorgestellten Überlebenszeitanalyse berücksichtigt werden könnten:

- Das Alter der Patientinnen und deren Überlebenszeit.
- Angaben, ob die Patientin gestorben ist oder noch lebt. Im Fall eines Todes wurde zusätzlich der Grund festgehalten. Patientinnen, die nicht aufgrund des Tumors verstarben, wurden zensiert.
- Überlebenszeit, berechnet vom Tag der Diagnose bis zum Tag des Todes.
- Dichotomisierte Angaben zum Auftreten einer Metastase, eines Rezidivs oder eines Zweitkarzinoms.
- Die metastasen-, rezidiv- und zweitkarzinomfreien Zeiten, berechnet vom Tag der Diagnose bis zum Tag der Feststellung der Metastase, eines Rezidivs bzw. dem Auftreten eines Zweitkarzinoms.
- In der so genannten diseasefreien Zeit wurden diese Angaben vereinheitlich, indem hier die Zeit zwischen dem Tag der Diagnose und dem zeitlich zuerst aufgetretenen Ereignis der Variablendefinition zu Grunde liegt.
- Weiterhin wurden sowohl die Tumorgröße als auch der Tumorgrad (Grading) zum Zeitpunkt der Diagnose festgehalten.

Zwei weitere Affymetrix HG-U133A Microarray Brustkrebsdatensätze wurden von der NCBI GEO Datenbank heruntergeladen. Beide enthalten sowohl die gemessenen RNA-Expressionen als auch die klinischen Variablen zu den ebenfalls lymph-nodal-negativen Patientinnen.

Die Rotterdam-Kohorte (vgl. Carroll et al., 2006; Wang et al., 2005) besteht aus den Genexpressionen von 180 rezidivfreien Frauen und 106 Patientinnen mit Fernmetastasen. Allerdings fehlen hier teilweise die kompletten Angaben zu den oben beschriebenen klinischen Variablen, die bei der Überlebenszeitanalyse als relevante Kovariablen in die Analyse mit einbezogen werden könnten. So wurden z.B. die Angaben, ob die festgestellten Todesfälle tumorbedingt aufgetreten sind, komplett vernachlässigt. Auch das Alter der Patientinnen zum Zeitpunkt der Diagnose und die näheren Angaben zum Tumor wie der Tumorgrad oder die Tumorgröße wurden bei der Rotterdam-Kohorte leider nicht erfasst.

Die zweite Kohorte, die Transbig-Kohorte (Desmedt et al., 2007; Loi et al., 2007), enthält Daten zu 280 an Brustkrebs erkrankten Frauen. Wie auch in der Mainz-Kohorte wurden hier das Alter der Patientinnen, die Tumorgröße und der -grad erhoben. Auch das Auftreten einer Metastase oder eines Rezidivs wie die entsprechenden metastasen- bzw. rezidivfreien Zeiten wurden dokumentiert. Obwohl festgehalten wurde, ob die Patientinnen zum Zeitpunkt des Studienendes noch gelebt haben oder gestorben sind, liegen die näheren Angaben zum Todesgrund im zweiten Fall jedoch nicht vor.

Für diese Arbeit ist die metastasenfreie Zeit, berechnet vom Tag der Diagnose bis zum Tag der Feststellung der Metastase, die zentrale und in allen Kohorten erhobene Variable. Die Datenmatrix für die anschließende Überlebenszeitanalyse (beschrieben in Unterkapitel 3.4) enthält für die 766 Patientinnen die Angaben zum Auftreten der Metastase (ja/nein) und der berechneten metastasenfreien Zeit.

Die Mainz-, Transbig- und Rotterdam-Kohorten wurden in zahlreichen früheren Studien zur Erforschung der Zusammenhänge zwischen Brustkrebs und unterschiedlichen Genen untersucht (z.B. Cadenas et al., 2010; Petry et al., 2010; Schmidt et al., 2010) und bilden somit eine fundierte und geeignete Datengrundlage für die vorliegende Arbeit.

Die heruntergeladenen Raw.cel Dateien der GEO Datenbank sind für alle drei Datensätze MAS5.0 (Microarray Analysis Suite) normalisiert. Für die hier folgende Analyse wurden sie mit dem am häufigsten zur Auswertung von Expressionsdaten eingesetzten Verfahren RMA (Robust Multiarray Average, vgl. Irizarry et al., 2003) normalisiert. Die zu analysierende Genexpressionsmatrix umfasst 22283 Zeilen und 766 Spalten mit log_2-transformierten und RMA normalisierten Genexpressionen. Es liegen keine fehlenden Expressionswerte vor.

2.4 AAS-Daten

Zur Verifizierung der in dieser Arbeit angewendeten Methoden wird ein weiterer Datensatz mit kurzen Zeitreihen herangezogen. Dieser wird hier nur kurz vorgestellt, da er nicht der Schwerpunkt der vorliegenden Arbeit ist.

Die Reaktion der Saccharomyces cerevisiae Hefe auf den Aminosäurenentzug (amino acid starvation AAS) wurde zu den Zeitpunkten 0,5, 1, 2, 4 und 6 Stunden nach der Reizung gemessen. Ferner wurden Expressionswerte vor dem Entzug als Kontrollwerte gemessen (entspricht dem Zeitpunkt 0). Eine detaillierte Beschreibung der Microarray-Versuche ist in Gasch et al. (2000) zu finden.

Die Analyse der Genexpressionen erfolgte mittels des Laser-Scanning-Mikroskop GenePix 4000 und ScanAlyze-Programms. Gene, zu denen keine Messung vorlag, die mit einem Fold Change unter 2 oder deren Expressionsniveau sich kaum verändert hat, wurden bei der Zusammenstellung des Datensatzes nicht berücksichtigt. Die unter http://www.benoslab.pitt.edu/astro/Amino_Acid_Starvation.txt zur Verfügung gestellte Datenmatrix enthält dadurch 700 von den ursprünglichen 1652 Zeilen mit den Expressionswerten zu jeweils 5 Zeitpunkten in den Matrixspalten. Zum Zeitpunkt 0 liegen keine Werte vor.

Da bei mehreren Probesets keine Information zu deren Genontologien vorliegt, werden sie in den betroffenen Analyseabschnitten nicht berücksichtigt. So resultiert für die CC-Gruppe eine Genexpressionsmatrix mit 565, für die MF-Gruppe mit 670 und für die BP-Gruppe mit 666 Zeilen (vgl. http://www.yeastgenome.org) und jeweils 5 Spalten.

Zu diesen AAS-Daten liegen aus einsichtigen Gründen keine Überlebenszeitdaten vor. Daher können sie nur zum Vergleich der eingesetzten Clustermethoden im ersten Teil des Clusteransatzes (vgl. Unterkapitel 3.1) herangezogen werden.

3 Statistische Methoden zum Clustern

Die Gene Array Untersuchungen haben große Fortschritte auf dem Gebiet der Tumorfrüherkennung und der daraus resultierenden verbesserten Therapieplanung ermöglicht. Eines der Ziele bei der Analyse von Expressions-Zeitreihen ist die Identifizierung von Genen mit ähnlichen Expressionsprofilen über die Zeit. Dabei gibt es eine große Vielfalt eingesetzter Verfahren.

So wird bei Schmidt et al. (2008) und Eisen et al. (1998) die hierarchische Clusteranalyse zur Analyse der Expressionswerte verwendet. Tavazoie et al. (1999) clustern die Genexpressionen mit der gängigen *k*-means Methode. Kim und Kim (2007), Ernst et al. (2005), Di Camillo et al. (2005), Gerber et al. (2007), Sacchi et al. (2005) und Wu et al. (2007) setzen auf die vereinfachten Strategien, welche auf Trends und/oder die Ausrichtung und Maß der Genexpressionsänderungen beruhen. In einem neuen Ansatz von Springer et al. (2011) wird das beim STEM-Verfahren von Ernst et al. (2005) vorgeschlagene Optimierungskriterium für die Auswahl des Modellprofils durch das Einsetzen von Frame Potential ersetzt.

Zwei weitere Algorithmen werden von Tchagang et al. (2009) vorgestellt. Zum einen ist dies die Methode ASTRO (Analysis of Short-Time-series using Rang Order preservation) unter Ausschluss der Gene mit konstanter Genexpression über alle Zeitpunkte und Aufstellung einer Rang-Matrix. Zum anderen das Verfahren MiMeSR (Minimum Mean

Squares Residue), bei dem im Gegensatz zu ASTRO die Größenordnung der Expressionsänderung und die Informationen über die Transkriptionsfaktoren (TF) berücksichtigt werden. Beide Programme sehen ausschließlich den Zugriff über das Internet vor, so dass sowohl die Daten als auch die zusätzlich erforderlichen Angaben wie die minimale Anzahl der Gene im Cluster, der minimale Fold Change oder die Transkriptionsfaktoren online eingegeben werden müssen.

In seiner Dissertation führt König (2014) eine umfangreiche Simulationsstudie zum Vergleich einiger Methoden zur Zeitreihenanalyse der Genexpressionsdaten durch, wie z.B. maSigFun (Schätzung der linearen Regressionen in zwei Stufen und anschließenden Variablenselektion nach Nueda et al., 2009), 2T-GSA (der 2-Gruppen-Enrichment Test, modifiziert durch den sequentiellen Vergleich mit einer Referenz), 1S-GSA (Abwandlung vom 2T-GSA, indem der Segmentations-Test statt des Enrichment Tests zur Datenanalyse herangezogen wird) und die Kombination von diesen beiden (der Segmentations-Test wird nur im durch einen FDR Threshold festgelegten signifikanten Bereich auf Anreicherung getestet), mit dem Ergebnis, dass für eine explorative Analyse auf Gengruppenebene die sequentiellen Enrichment Tests unter Berücksichtigung der Signifikanz differentieller Expressionen die robustesten und schlüssigsten Ergebnisse erzielten.

Ein weiteres umfangreiches Gebiet mit zahlreichen und effektiven Analysemöglichkeiten bietet die Bayesianische Herangehensweise an die Problematik des Clusterns von Genexpressions-Zeitreihen. In diesem Bereich wurden bereits zahlreiche Publikationen veröffentlicht und es wird auch weiterhin intensiv geforscht (z.B. Medvedovic und Sivaganesan, 2002; Ramoni et al., 2002; Medvedovic et al., 2004; Rasmussen et al., 2009; Wang et al., 2012; Mori et al., 2013). Frühwirth-Schnatter und Kaufmann (2002) beschreiben die ökonometrische Modellierung von Paneldaten unter Einsatz von MCMC-Algorithmen. Sun et al. (2012) stellen eine robuste Bayesianische Clusterung von Genexpressionsdaten mit Wiederholungsmessungen vor, in welcher die Datenstruktur in besonderer Weise einbezogen wird. Wang und Wang (2013) haben einen Gibbs Sampling Algorithmus für hierarchische Dirichlet Prozess Modelle entwickelt. Kormaksson et al. (2012) stellen in ihrer Ausarbeitung einen Clusteransatz vor, der auf zwei aufeinander aufbauenden Algorithmen beruht. Im ersten Schritt wird dabei die hierarchische Clusteranalyse durchgeführt und anschließend werden mit Hilfe des Expectation-Maximization-(EM)-Algorithmus die „wahren" Gruppen initialisiert. Fraley und Raftery

(1999), Fu et al. (2011, 2013) und Grün (2011) haben ihre Methoden in R mit *mclust()*, *DIRECT()* bzw. *BayesMix()* implementiert. Diese sind frei zugänglich. In seiner Dissertation geht Fritsch (2010) auf die bei Medvedovic und Sivaganesan (2002) verwendete Distanzmetrik näher ein und stellt mehrere Methoden zur optimalen Clusterfindung vor, von denen einige auch in dieser Arbeit zum Einsatz kommen.

Eine weitere Möglichkeit zum Clustern von Genexpressions-Zeitreihen unter Einsatz von Bayesianischen Aspekten bietet die so genannte „sphärische" Clusteranalyse. So setzen Dortet-Bernadet und Wicker (2008) endliche Mischungsmodelle mit stereographischen Projektionen von Normalverteilungen auf einer Einheitskugel (unit sphere) ein, um die Expressionsdaten zu clustern und Gengruppen mit bekannten Genen und Funktionen aufzudecken. In einem anderen Artikel stellen Dortet-Bernadet und Fan (2007) endliche Mischungsmodelle mit von-Mises-Fischer Verteilungen (vMF) und gleichzeitiger Variablenselektion vor. Banerjee et al. (2005) benutzen ebenfalls vMF Verteilungen und passen die endlichen Mischungsmodelle mit dem EM-Algorithmus an - hauptsächlich zum Clustern von Textdokumenten.

Bei der Zusammenfassung des Methodenüberblickes ist die Frage des gleichzeitigen Clusterns über Gene und Zeitpunkte aufgekommen. Dazu stellen z.B. Madeira und Oliveira (2004) einige Algorithmen zum Biclustern vor, bei denen gleichzeitig nach Ähnlichkeiten entlang beider Dimensionen gesucht wird. Heard et al. (2005) beschreiben eine Bayesianische modellbasierte agglomerative Methode, die allerdings nicht geeignet ist, falls eine kleine Anzahl an zu findenden Cluster angestrebt wird.

Ferner stellt die Berücksichtigung des biologischen Vorwissens ein weiteres interessantes Vorgehen beim Clustern von Genexpressionsdaten dar. Hierfür werden unterschiedliche Herangehensweisen angeboten. So schlagen Subramanian et al. (2005) den Zweigruppen-Vergleich mit Hilfe einer Gene Set Enrichment Analyse (GSEA) vor, die auch als Programm zur Verfügung gestellt wurde. Chu et al. (1998) machen eine Strukturannahme der Daten, indem 7 wichtige Zeitverlaufsmuster definiert werden, sowohl visuell als auch aus Publikationen, und setzen auf einen Korrelationskoeffizienten für die Clusterzuordnung. Schlicker et al. (2006) stellen ein Maß für die Ähnlichkeit zweier Gene Ontology Terme und aufbauend darauf auch zwischen den Genprodukten vor, welches zur Identifizierung biologisch relevanter Gengruppen herangezogen werden kann. Aber auch die ganz „triviale" Trennung nach GO-Gruppen (http://www.geneontology.org/)

kann zur Berücksichtigung des biologischen Vorwissens in Betracht gezogen werden. Die Analyse der biologischen Hintergrundinformation kann jedoch auch nachgelagert im Anschluss an die Clusterung durchgeführt werden. Shamir et al. (2005) bieten ein für Akademiker frei zugängliches Programm an (EXPANDER, vgl. Unterkapitel 4.1), das die ermittelten Gengruppen im Hinblick auf die Überrepräsentation von Transkriptionsfaktoren oder auch zur Analyse hinsichtlich ihrer GO-Gruppenzugehörigkeit bewertet.

Um die Idee der Clusteranalyse für die Genexpressions-Zeitreihen und somit eine der Ziele dieser Arbeit zu veranschaulichen, kommt an dieser Stelle ein kleiner Motivationseinschub im Hinblick auf die in Unterkapitel 3.2 und 3.3 beschriebenen und hier angewendeten Methoden.

Die Beobachtungen y_{it} sind log_2 - transformierte und RMA normalisierte Genexpressionen mit n Probesets (i = 1, …, n) und T Zeitpunkten (t = 1, …, T). In der Abbildung 3.1 werden zeitliche Expressionsverläufe am Beispiel von drei Genen dargestellt.

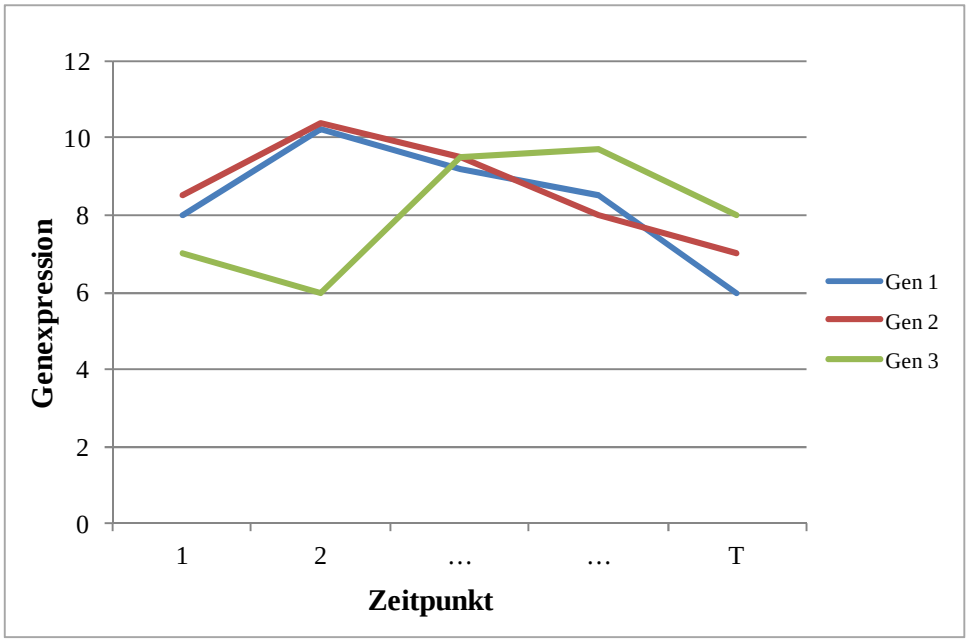

Abb. 3.1: Beispielhafter zeitlicher Verlauf von drei Genen zur Veranschaulichung der Idee beim Clustern von Expressions-Zeitreihen

Das Ziel der in dieser Arbeit angewendeten Verfahren ist die Gene 1 und 2 einem Cluster zuzuordnen, da sie offensichtlich einen ähnlichen Zeitverlauf aufweisen. Im Folgenden werden diese Methoden näher beschrieben.

3.1 Ansatz

All diese Verfahren werden in einer Vielfalt von Ansätzen zur Analyse von Expressions-Zeitreihen auf unterschiedliche Weise herangezogen. Bis heute sind allerdings nur solche Vorgehen bekannt, die die folgenden wichtigen Aspekte entweder gar nicht oder nur zum Teil berücksichtigen (z.B. Schmidt et al., 2008; Winter et al., 2007):

- die Zeitreihendynamik mit $T < 10$,
- die biologische Hintergrundinformation und
- die prognostische Relevanz der ermittelten Cluster.

Der hier vorgestellte neue integrative Ansatz zur Clusterung kurzer Expressions-Zeitreihen (vgl. Abbildung 3.1.1) unterscheidet sich von den schon bekannten in der Kombination der Information über die zeitabhängigen Genexpressionen in BRC Zelllinien mit biologischem Wissen. Dies erlaubt es prognostisch relevanten Gengruppen für die anschließende Metagenberechnung und Überlebenszeitanalyse auf eine bisher einzigartige Art und Weise zu erkennen.

Teil I beruht auf der Analyse humaner Mammakarzinom-Zelllinien. Mit den Clusteranalyseansätzen, auf die in den Unterkapiteln 3.2 und 3.3 näher eingegangen wird, werden Gengruppen mit ähnlichen Expressionsverläufen, also co-regulierte Gencluster identifiziert. Anschließend werden hier die biologisch interessanten Cluster ermittelt, wie in Unterkapitel 4 beschrieben.

Für die so identifizierten Gengruppen werden im Teil II des Ansatzes Metagene gebildet (vgl. Unterkapitel 4.3) und mittels Überlebenszeitanalyse, beschrieben in Unterkapitel 3.4, auf ihre prognostische Relevanz untersucht.

In den folgenden Unterkapiteln wird auf die einzelnen Schritte näher eingegangen.

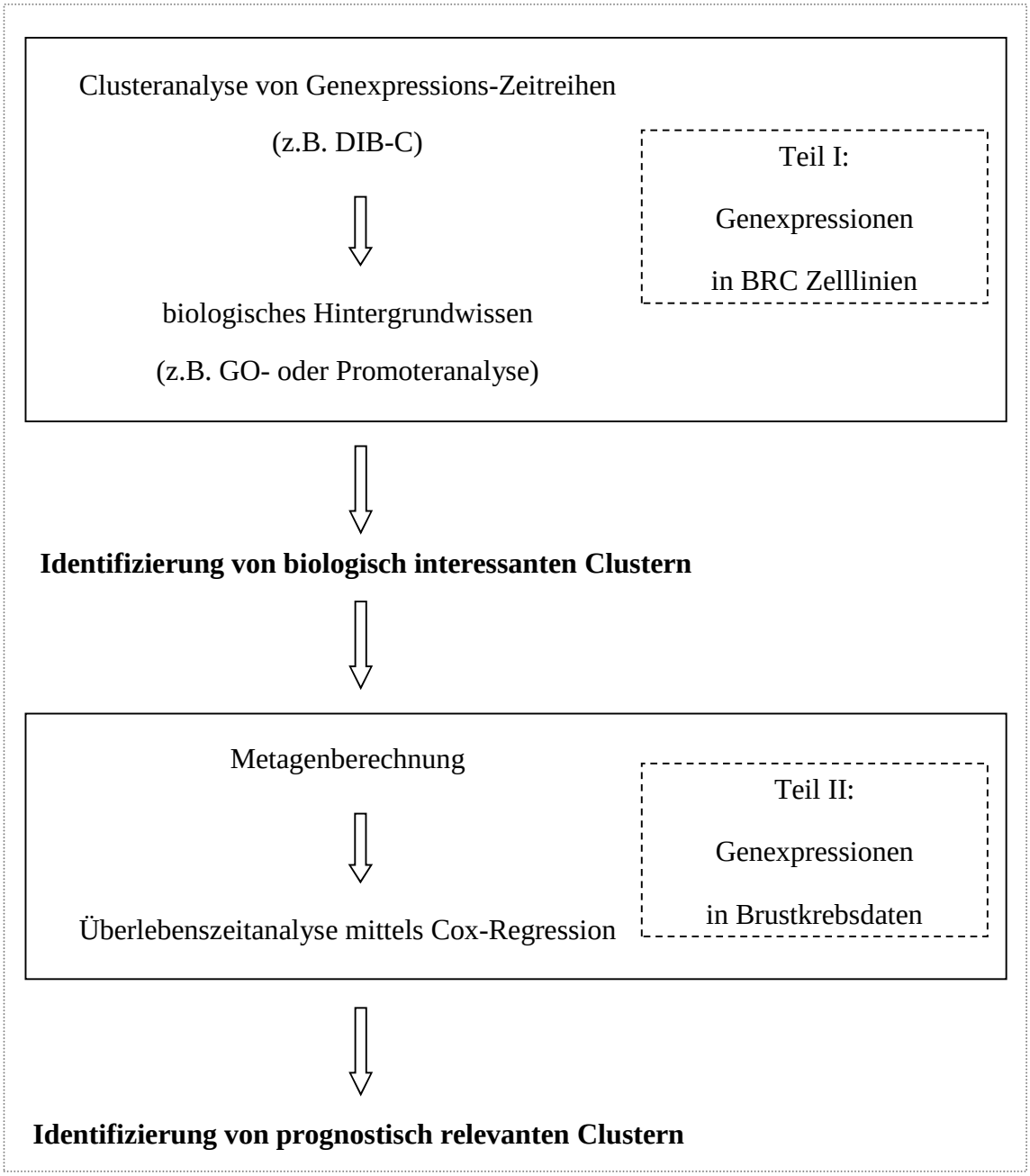

Abb. 3.1.1: Konzept des integrativen Clusteranalyseansatzes zur Identifizierung prognostisch relevanter Metagene. Dabei bilden im Teil I die Genexpressionen in Mammakarzinom-Zelllinien die Datengrundlage; für den zweiten Teil des Ansatzes sind die Brustkrebsdaten die Analysebasis.

3.2 Nicht-modellbasierte Clusteranalyse der Expressionsdaten

Bei dieser Art der Clusteranalyse werden keine Vorkenntnisse über die Daten benötigt. Zur Anwendung modellfreier Methoden müssen außerdem keine Annahmen getroffen werden, so dass diese öfters vorrangig zur Analyse der Daten herangezogen werden, um erste Eindrücke über deren Struktur zu gewinnen.

Im Folgenden werden Methoden beschrieben, die in dieser Arbeit bei der nicht-modellbasierten Clusterung der Daten angewendet werden.

3.2.1 *k*-means

Dieses partitionierende Verfahren, dessen Name auf MacQueen (1967) zurück zu führen ist, basiert auf einfachen Heuristiken zur Gruppierung der Beobachtungen um *k* Mittelwerte und wird daher auch *k*-means genannt. Es benötigt keine Vorkenntnisse über die Daten und ist deswegen die am häufigsten eingesetzte Methode in der Clusteranalyse. *K*-means ist sehr schnell und einfach umsetzbar, seine Ergebnisse hängen jedoch sehr stark von den (u.U. zufälligen) Startpositionen ab.

Das Ziel dieser Methode ist die Minimierung der Varianzsumme innerhalb der Cluster:

$$min \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_{i,j} - c_j \right\|^2.$$

Dabei ist $x_{i,j}$ die Beobachtung *i* im Cluster *j* und c_j das zugehörige Clusterzentrum. Die Iterationsschritte, die der *k*-means Algorithmus durchläuft, sind Anhang D.1 zu entnehmen.

Diese modellfreie Clusterung der Daten kann in R mit Hilfe von der Funktion *kmeans()* und der Angabe der gewünschten Clusteranzahl erfolgen. Die Ausgangsmittelwerte werden hierbei entweder zufällig generiert (Voreinstellung) oder können direkt angegeben werden.

Um die Clusteranzahl *k* optimal zu wählen, gibt es mehrere Möglichkeiten (vgl. Tibshirani et al., 2001; Kaufman und Rousseeuw, 1990; Milligan und Cooper, 1985). In dieser Arbeit werden die folgenden zwei Kriterien zur Bestimmung von *k* herangezogen:

- Der Verlauf der Fehlerquadratsumme (RSS) wird für unterschiedliche ganzzahlige Werte von *k* betrachtet und entsprechend der Stelle in der Kurve, wo diese nicht mehr entscheidend sinkt, wird die optimale Anzahl von Clustern festgelegt.

- Eine weitere Möglichkeit bilden die Silhouetten-Werte. Diese werden für jede Beobachtung i im Cluster durch die Normierung der Differenz des mittleren Abstands zwischen dieser Beobachtung und allen anderen desselben Clusters (mit a_i bezeichnet) und des minimalen mittleren Abstands zu allen Mitgliedern im besten anderen Cluster (vgl. Kaufman und Rousseeuw, 1990), bezeichnet mit b_i, mit deren Maximum berechnet:

$$\frac{b_i - a_i}{max(a_i, b_i)} .$$

  Für jede Clusterung werden anschließend die durchschnittlichen Silhouetten-Werte (average silhouette width) berechnet, deren Maximum die optimale Clusteranzahl angibt.

  Ein Vorteil gegenüber der Bestimmung der optimalen Clusteranzahl mittels RSS ist, dass hier zusätzlich die Beurteilung der geclusterten Daten ermöglicht wird. Dabei gilt, dass bei den Silhouetten-Werten < 0 die Beobachtungen falsch klassifiziert wurden, da b_i dabei kleiner als a_i ist. Je näher die Silhouetten-Werte bei 1 liegen, desto besser ist die Klassifikation der Daten.

Diese beiden Kriterien werden in der vorliegenden Arbeit zur Bestimmung der optimalen Clusteranzahl kombiniert. Da *k*-means ein randomisiertes Verfahren ist und als Folge bei mehreren Durchläufen auch unterschiedliche Ergebnisse liefert, sollte es mehrfach und mit unterschiedlichen Startwerten (z.B. mit der Option *nstart* in *kmeans()*) durchgeführt werden.

3.2.2 Short time-series expression miner (STEM)

Das von Ernst et al. (2005) vorgeschlagene Verfahren STEM wurde zur Identifizierung von Genen mit ähnlichen Expressionsmustern und unter Berücksichtigung der

Zeitreihendynamik der Expressionswerte entwickelt. Es basiert auf vordefinierten Modellprofilen als Muster für mögliche Genexpressionsverläufe, welche unabhängig von den Daten festgelegt werden.

Im ersten Schritt dieser Methode werden Expressionswerte für jedes Gen *i* und Beobachtungswiederholung *j* (Triplikat, vgl. Unterkapitel 2.2) zum Zeitpunkt 0 (y_{ij0}) von den Expressionswerten zu den folgenden *t* Zeitpunkten subtrahiert. Anschließend wird für jedes Gen und Zeitpunkt die mediane Zeitreihe $\tilde{y}_{it}$ berechnet.

Im nächsten Schritt werden Modellprofile festgelegt, welche die zeitlichen Genexpressionsänderungen in den Daten widerspiegeln sollten. Sie werden durch die folgenden zwei Parameter definiert:

- Durch den Parameter c ($c \in \mathbb{Z},\ c \geq 1$) wird die Menge aller möglichen Modellprofile bestimmt, bei denen die Differenz von Expressionswerten zwischen den benachbarten Zeitpunkten *c* nicht überschreitet. So kann z.B. bei c = 3 die Veränderung der Genexpressionen zweier aufeinander folgenden Zeitpunkte in den zugehörigen Modellprofilen höchstens 3 betragen. Dabei kann sie sowohl um 1, 2 oder 3 ansteigen oder abfallen, als auch gleich bleiben. Für *t* Zeitpunkte sind somit $(2c + 1)^{t-1}$ Profile möglich.

- Der zweite Parameter $m < 100$ (vgl. Ernst et al., 2005) definiert die maximale Anzahl möglicher Profile und zwar so, dass der minimale Abstand zwischen den jeweils zwei Profilen in dieser Menge maximiert ist.

Die Gene werden hinsichtlich ihrer Beobachtungen und einer Distanzmetrik *d* den ausgewählten Modellprofilen zugeordnet (vgl. Ernst et al., 2005). Der mediane Genexpressionsprofil $\tilde{y}_{it}$ wird dem Cluster mit der minimalen Distanz $d(\tilde{y}_{it}, m_{kt})$ zugeordnet. m_{kt} bezeichnet dabei das *k*-te zeitabhängige Modellprofil. Bei mehreren Modellprofilen mit dem kleinsten Abstand wird die Anzahl zutreffender Profile (*AP*) ermittelt und das Gen anschließend zu allen diesen Profilen zugeordnet, jedoch mit $1/AP$ gewichtet.

Die Distanzmetrik *d* ist definiert als $d = 1\text{-}\rho$, wobei ρ den Korrelationskoeffizienten nach Bravais-Pearson zwischen den Genexpressions- und Modellprofilen bezeichnet:

$$\rho\left((\tilde{y}_{it})_{t=1,\dots,T}, (m_{kt})_{t=1,\dots,T}\right) = \frac{\sum_{t=1}^{T}(\tilde{y}_{it}-\bar{\tilde{y}}_i)(m_{kt}-\bar{m}_k)}{\sqrt{\sum_{t=1}^{T}(\tilde{y}_{it}-\bar{\tilde{y}}_i)^2 \sum_{t=1}^{T}(m_{kt}-\bar{m}_k)^2}}.$$

Obwohl die Größe 1-ρ die Dreiecksungleichung nicht erfüllt und somit keine Metrik ist, wird deren Einsatz hier durch die Erfüllung der verallgemeinerten Version dieser mit

$$d(x,z) \leq 2\big(d(x,y) + d(y,z)\big) \text{ für alle Modellprofile } x, y, z$$

rechtfertigt (vgl. Ernst et al., 2005).

Signifikante Modellprofile werden durch den Vergleich der beobachteten Anzahl zugewiesener Gene zum Profil und deren erwarteten binomialverteilten Anzahl identifiziert (vgl. Ernst et al., 2005). Die Einhaltung des globalen Signifikanzniveaus α wird durch die Bonferroni-Korrektur gewährleistet.

3.2.3 Difference-based clustering algorithm (DIB-C)

Eine weitere Möglichkeit zur Identifizierung von Genen mit ähnlichen Expressionsverläufen unter Berücksichtigung der Zeitreihendynamik bietet das Verfahren DIB-C, vorgeschlagen von Kim und Kim (2007). Es kann dabei in zwei logische Methodenblöcke aufgeteilt werden: die Charakterisierung der Expressionsänderung für jede Zeitreihe und benachbarte Zeitpunkte sowie die anschließende Clusterzuordnung. Dieses Verfahren wurde speziell für die kurzen Zeitreihen entwickelt und eignet sich dementsprechend sehr gut für die in dieser Arbeit gesetzten Ziele.

Im ersten logischen DIB-C Block wird jedes Gen i (i = 1, …, n) der Genexpressions-Zeitreihe $\tilde{y}_{it}$ mit T Zeitpunkten durch (T - 1) + (T - 2) Symbole charakterisiert (die Einzelheiten werden im Folgenden näher beschrieben). Durch diese Buchstabenreihenfolge werden der erstrangige sowie der zweitrangige Unterschied abgebildet und somit die horizontale bzw. die vertikale Struktur der Daten (vgl. Kim und Kim, 2007). Die Bestimmung der Genexpressionsänderung basiert dabei auf dem moderaten *t*-Test, durchgeführt mit Hilfe des R-Pakets *limma* (Smyth, 2005), indem die

Instabilität der Fehlervarianz einzelner Gene durch ein hierarchisches Bayes-Modell korrigiert wird (vgl. Smyth et al., 2004).

Zuerst wird für den erstrangigen Unterschied der Expressionsänderung die zugehörige Matrix bestimmt, die als Einträge für jedes Gen und für alle benachbarten Zeitpunkte die zugehörigen Werte der moderaten t-Statistik

$$\gamma_{it}^{(1)} = \frac{\beta_{it}}{\tilde{s}_i \sqrt{v_{it}}}$$

zwischen benachbarten Zeitpunkten *t* und (*t* + 1) zum Signifikanzniveau $\alpha^{(1)}$ enthält:

$$Y^{(1)} = \left(\gamma_{it}^{(1)}\right)_{n\times(T-1)}.$$

Dabei bezeichnet β_{it} die zugehörige mittlere Differenz, $\tilde{s}_i$ den a posteriori-Mittelwert des Schätzers der Fehlervarianz und v_{it} die Stichprobenvarianz (vgl. Kim und Kim, 2007).

Nach dem Vergleich dieser Werte mit dem $\left(1 - \alpha^{(1)}/2\right)$ - Quantil der t-Verteilung mit df_{it} Freiheitsgraden, $t = T - 1$ (nach Smith et al., 2004) erfolgt die Einstufung der Expressionssteigung:

- **I** (**I**ncrease ≙ Anstieg), falls $\gamma_{it}^{(1)} > T(1 - \alpha^{(1)}/2\,; df_{it})$,
- **D** (**D**ecrease ≙ Abfall), falls $\gamma_{it}^{(1)} < T(1 - \alpha^{(1)}/2;\ df_{it})$ und
- **N**, falls keine signifikanten Veränderungen festgestellt werden konnten (**N**o change).

Jedem Gen wird somit in diesem Schritt eine (T – 1) - stellige Buchstabenfolge zugewiesen, die die Richtung der Expressionsänderung beschreibt. Dies wird in einer Matrix *E* (**E**rstrangiger Unterschied) festgehalten:

$$E = (f_{it})_{n\times(T-1)}.$$

Analog erfolgt die Bestimmung des zweitrangigen Unterschiedes, der das Krümmungsverhalten der Genexpressions-Zeitreihe angibt. Matrix $Y^{(2)}$ enthält jetzt allerdings die Differenzen zwischen den moderaten t-Statistiken mit $t = T - 2$ zum Signifikanzniveau $\alpha^{(2)}$, die zur Bestimmung des erstrangigen Unterschieds berechnet wurden:

$$Y^{(2)} = \left(\gamma_{it}^{(2)}\right)_{n\times(T-2)}.$$

Der Vergleich dieser Werte mit dem $\left(1-\alpha^{(2)}/2\right)$ - Quantil einer empirischen t'-Verteilung (näheres dazu in Kim und Kim, 2007) beschreibt das Krümmungsverhalten einer Genexpressions-Zeitreihe durch die folgendermaßen definierten $(T-2)$ Buchstaben:

- **V** (con**V**ex ≙ konvex), falls $\gamma_{it}^{(2)} >$ empirisches $(1-\alpha^{(2)}/2)$ - Quantil,
- **A** (conc**A**ve ≙ konkav), falls $\gamma_{it}^{(2)} <$ empirisches $(1-\alpha^{(2)}/2)$ - Quantil und
- **N** für keine signifikanten Veränderungen in der Krümmung (**N**o change),

festgehalten in einer Matrix *Z* (**Z**weitrangiger Unterschied):

$$Z = (f_{it})_{n\times(T-2)}.$$

Die anschließende Kombination der beiden Matrizen *E* und *Z* zu *U* ergibt somit im ersten DIB-C Block die Charakterisierung der Genexpressions-Zeitreihen durch $(T-1) + (T-2) = (2T-3)$ Symbole:

$$U_{n\times(2T-3)} = \left[E_{n\times(T-1)}|Z_{n\times(T-2)}\right].$$

Durch diese $(2T-3)$ Stellen der Buchstabenfolge sowie je drei mögliche Symbole sind 3^{2T-3} verschiedene Cluster möglich. Deswegen werden im zweiten logischen Block des Verfahrens Gengruppen mit einer Genanzahl kleiner als die vom Benutzer vordefinierte Grenze aufgelöst und basierend auf dem Korrelationskoeffizienten nach Bravais-Pearson (s. Distanzmetrik in Unterkapitel 3.2.2) den Cluster mit ähnlichen Probesets zugeordnet (vgl. Kim und Kim, 2007). Die Ähnlichkeit wird dabei über die Korrelation über alle Zelllinien bestimmt. Somit gehören Probesets, die das gleiche Muster an Steigung und Krümmung, also einen ähnlichen Zeitverlauf aufweisen, zum gleichen Cluster.

3.2.4 Penalized frame potential (PFP)

Die Auswahl der Basisprofile bei STEM ist willkürlich und einschränkend, da nur mit ganzzahligen Sprunghöhen in den Zeitpunkten t_i gearbeitet wird. Springer et al. (2011)

haben einen Algorithmus entwickelt, der die Gruppierung der Daten mittels geometrischer Betrachtung verbessert. Die Idee der Entwicklung dieses Verfahrens entstand in der gemeinsamen Diskussion mit den Autoren der bei STEM genannten Nachteile. Es handelt sich hierbei um ein neuartiges Verfahren in einem weitfortgeschrittenen Entwicklungsstadium, zu dem erst eine Veröffentlichung vorliegt. Deswegen erfolgt die folgende Methodenbeschreibung in starker Anlehnung an dieses Paper und wird nur an dieser Stelle zitiert.

Motiviert durch die aktuellen Entwicklungen in der Frame-Theorie (z.B. Kovačevič und Chebira, 2007) wird bei PFP das bei STEM vorgeschlagene Optimierungskriterium für die Auswahl des Modellprofils durch das Einsetzen von Frame Potential ersetzt.

Für ρ in der Distanzmetrik d = 1-ρ (vgl. 3.2.2) gilt jetzt durch die Orthogonal-Projektion π_H von $\tilde{y}_{it}$ auf den $(T-1)$ - dimensionalen linearen Teilraum H:

$$\rho\left((\tilde{y}_{it})_{t=1,\dots,T}, (m_{kt})_{t=1,\dots,T}\right) = \frac{\sum_{t=1}^{T}(\tilde{y}_{it}-\bar{\tilde{y}}_i)(m_{kt}-\bar{m}_k)}{\sqrt{\sum_{t=1}^{T}(\tilde{y}_{it}-\bar{\tilde{y}}_i)^2 \sum_{t=1}^{T}(m_{kt}-\bar{m}_k)^2}} =$$

$$\frac{\langle \pi_H(\tilde{y}_{it}), \pi_H(m_{kt})\rangle}{\|\pi_H(\tilde{y}_{it})\|_2 \, \|\pi_H(m_{kt})\|_2} = \langle \theta_{\tilde{y}_{it}}, \theta_{m_{kt}} \rangle \ ,$$

wobei $\langle \theta_{\tilde{y}_{it}}, \theta_{m_{kt}} \rangle \in S^{T-1} \cap H$ das Skalarprodukt aus den projizierten Daten darstellt. Durch die quadratische Euklidische Distanz zwischen den Projektionen zugehöriger Zeitreihen auf der Einheitskugel S^{T-1} kann die Distanzmetrik aus 3.2.2 somit vereinfacht werden zu

$$d(\tilde{y}_{it}, m_{kt}) = 1 - \langle \theta_{\tilde{y}_{it}}, \theta_{m_{kt}} \rangle.$$

Durch diese Darstellungsart des Ähnlichkeitsmaßes können die Cluster C_l, $1 \leq l \leq m$, als Voronoizellen (vgl. z.B. Saff und Kuijlaars, 1997) auf der Einheitskugel interpretiert werden.

PFP ist in MATLAB (MATrizen LABoratorium, Version 7.14.0) implementiert. Um die optimalen vorzugebenen Startwerte m und α zu berechnen, wird hier ähnlich wie bei STEM ein Greedy-Ansatz (deren klassische Vorgehensweise in Chvátal (1979) beschrieben wurde) angewendet. Hierbei wird in jedem Schritt das Modellprofil mit dem größten Abstand zu den bisher gewählten Profilen bestimmt. Dabei betrachtet der Abstand d die Modellprofile als Punkte auf der Sphäre S^{T-1}.

Seien die auf S^{T-1} normierten Vektoren enthalten in $\Theta_m = \{\theta_t \mid t = 1, \dots, m\} \in S^{T-1}$. Der PFP-Algorithmus berechnet dann für die gegebenen Daten, die in der Matrix *Y* mit den log_2-transformierten und auf der Einheitskugel projizierten Zeitreihen festgehalten werden, $m \in \mathbb{N}$ Modellprofile, indem das Funktional

$$F_\alpha(Y, \Theta_m) = \alpha \frac{T}{m^2} TFP(\Theta_m) + m(1-\alpha) \cdot P(Y, \Theta_m)$$

über dieser Einheitskugel S^{T-1} minimiert wird. Dabei ist das Funktional TFP für $\Theta_m \in S^{T-1}$ definiert durch

$$TFP(\Theta_m) = \sum_{t,l=1}^{m} |\langle \theta_t, \theta_l \rangle|^2.$$

Ist die Anzahl der möglichen Profile $m \leq T$, so gilt:

$$\min_{\Theta_m \in S^{T-1}} TFP(\Theta_m) = m.$$

Ansonsten gilt für $m > T$:

$$\min_{\Theta_m \in S^{T-1}} TFP(\Theta_m) = \frac{m^2}{T}.$$

Das Minimum von Θ_m hat dabei die Eigenschaft, dass die Vektoren einerseits zueinander möglichst weit entfernt und andererseits nahe bei den normierten Daten sein sollen. Das Funktional TFP steuert das gegenseitige „Wegdrücken" der θ_t im Sinne des Coulombschen Gesetzes (vgl. Benedetto und Fickus, 2003). Der Strafterm $P(Y, \Theta_m)$ bestraft große Entfernungen zwischen den Daten und sorgt hier so für die Datennähe. Der Parameter $\alpha \geq 0$ steuert das Verhältnis zwischen dem datenabhängigen und -unabhängigen Teil des Minimierungsproblems, indem ein zu großes α zu einer Art overfitting führen und ein zu kleines α dagegen eventuell die Struktur der Daten zu wenig in Betrachtung ziehen könnte.

Durch diese Art der geometrischen Betrachtung der Clusterung und das einfache Umsetzen ist PFP eine bessere Alternative zu STEM und wird in dieser Arbeit als nicht-modellbasierte Vergleichsmethode eingesetzt. Das zugehörige Programm kann nach Rücksprache mit den Autoren zur Verfügung gestellt werden.

3.3 Modellbasierte Clusteranalyse der Expressionsdaten

Bei den hier vorgestellten modellbasierten Clusteransätzen werden im Gegensatz zu den modellfreien Methoden Wahrscheinlichkeitsmodelle unterstellt und die unbekannten Parameter dieser Modelle inferriert. Dabei wird die Annahme getroffen, dass die vorliegenden Daten den *K* unabhängigen Populationen entstammen, wobei die wahre Anzahl der Populationen unbekannt ist.

In diesem Kapitel werden die Mischungsmodelle vorgestellt, mit deren Hilfe die Clusterstruktur vorliegender Daten untersucht werden kann. Dabei kommen sowohl die endlichen als auch die flexibleren unendlichen Mischungsmodelle zum Einsatz. Hier werden die beiden Typen von Mischungsmodelle Bayesianisch formuliert, so dass in beiden Fällen eine Inferenz mit Hilfe von Markov Chain Monte Carlo (MCMC) Methoden (vgl. Anhang D.2) möglich ist.

Die Grundlage Bayesianischer Modelle ist das Bayes-Theorem (Bayes, 1763; Carlin und Louis, 2000). Durch den Bayes-Ansatz können die unbekannten Parameter der unterstellten Wahrscheinlichkeitsmodelle mit MCMC inferriert werden. Weiterhin können Aussagen über die Hypothesen für die Parameter getroffen werden. Bayesianische Modelle ermöglichen es zudem ggf. vorliegendes Vorwissen (a priori-Wissen) im Modell zu berücksichtigen. Insofern besitzen die hier vorgestellten Verfahren mehrere Vorteile gegenüber den nicht-modellbasierten Methoden, die in Kapitel 3.2 beschrieben wurden.

3.3.1 Finite mixture models

Die Idee von Mischungsmodellen geht auf die Arbeit von Pearson (1894) zurück, der als Erster eine Modellanpassung an die Daten mit zwei Komponenten der Normalverteilung durchgeführt hat. Seitdem werden sie zur Analyse heterogener Daten in verschiedenen Anwendungsgebieten (wie z.B. in der Biologie oder in der Ökonometrie) eingesetzt. Zum Ansatz endlicher Mischungsmodelle (engl.: finite mixture models) wird im Folgenden eine Einführung gegeben; für mehr Details wird an dieser Stelle auf McLachlan und Peel (2000) verwiesen.

Ein endliches Mischungsmodell für eine Zufallsvariable *Y* ist gegeben durch:

$$p(y \mid \theta, \pi) = \sum_{k=1}^{K} \pi_k p(y \mid \theta_k),$$

wobei *K* die Anzahl der Komponenten ist und *k*, *k* = 1, …, *K*, die *k*-te Mischungskomponente darstellt. Ferner ist $\pi = (\pi_1, \dots, \pi_K)$ der Gewichtsvektor und $p(y \mid \theta_k)$ eine parametrische Verteilungsfamilie mit den Parametern θ_k. Die Parameter der *K* Verteilungen können als Vektor zusammengefasst werden, symbolisch: $\theta = (\theta_1, \dots, \theta_K)$. Als Verteilungsfamilie können z.B. die multivariaten Normalverteilungen $p(y \mid \mu_k, \Sigma_k)$ mit dem Erwartungswertvektor μ_k und der Kovarianzmatrix Σ_k verwendet werden.

Für den Gewichtsvektor π gilt dabei:

$$\sum_{k=1}^{K} \pi_k = 1 \text{ und } \pi_k > 0 \quad \forall\, k \in \{1, \dots, K\} \ .$$

Dieser und der Parametervektor θ sind unbekannt und müssen geschätzt werden. Die statistische Schätzung kann z.B. durch die Maximum-Likelihood-Schätzung mit dem Expectation-Maximization-Algorithmus (EM-Methode, vgl. Dempster et al., 1977) oder mit Hilfe der Bayesianischen Schätzung, die im Anhang D.2 beschrieben wird, erfolgen. Wegen der Optimierung der Clusterergebnisse mit der Posterior Similarity Matrix, durch die eine sinnvolle Zusammenfassung von MCMC-Samples von Clusterungen erfolgen kann (vgl. Unterkapitel 3.3.4), wird auf den EM-Algorithmus in dieser Arbeit nicht näher eingegangen.

In R wurden mehrere entsprechende Möglichkeiten zum modellbasierten Clustern implementiert, wie z.B. Pakete *BayesMix* (Grün, 2011) für Mischungen von univariaten und *mclust* (Fraley und Raftery, 1999) für Mischungen von multivariaten Normalverteilungen. In dieser Arbeit kommt das R-Paket *BayesMix* (Version 0.7-2) zum Einsatz, da es die Anpassung der Modelle an die Daten mit Hilfe der MCMC-Methoden ermöglicht und *mclust()* einen EM-Algorithmus zur Parameterschätzung heranzieht.

Hier gilt jetzt für die Parameter θ_k der univariaten Gauss-Verteilung:

$$\theta_k = (\mu_k, \sigma_k^2).$$

Als a priori-Verteilung der Gewichte wird eine Dirichlet-Verteilung unterstellt, die jeder Komponente *k* ein Gewicht π_k zuweist:

$$\pi_1, \dots, \pi_K \sim Dir(e_{0,1}, \dots, e_{0,K}).$$

Dabei können $e_{0,1}, \dots, e_{0,K}$ als die a priori vorhandene Pseudo-Beobachtung interpretiert werden.

Die a priori-Parameter werden bei *BayesMix()* durch eine entsprechende Einstellung generiert und sehen für den Fall konjugierter a priori-Verteilung wie folgt aus:

$$\sigma_k^2 \sim InvGamma\left(\frac{v_{0,k}}{2}, \frac{v_{0,k} S_{0,k}}{2}\right) \text{ und}$$

$$\mu_k \mid \sigma_k^2 \sim N\left(b_{0,k}, B_{0,k}\, \sigma_k^2\right)$$

mit den z.B. nach Raftery (1996) definierten a priori-Parametern $b_{0,k} = \bar{y}_k$, $B_{0,k} = 2{,}6 \cdot R(y_k)^2$, $v_{0,k} = 2{,}56$ und $S_{0,k} = \frac{length(y_k)-1}{length(y_k) var(y_k)}$ mit der robusten Schätzung der Varianz R. Für eine unabhängige, konjugierte a priori-Verteilung ist der Erwartungswert μ_k definiert durch

$$\mu_k \sim N\left(b_{0,k}, B_{0,k}\right).$$

Der Vorteil dabei wäre die Unabhängigkeit dieser a priori-Verteilung von der geschätzten Varianz.

Die Startwerte, die zum Starten von MCMC benötigt werden, können dabei gemäß Frühwirth-Schnatter (2006) initialisiert werden mit den gleichen Werten für π und $\tau = \frac{1}{\sigma^2} = \frac{1}{(R/1{,}34)^2}$ und den entsprechenden $\frac{1}{k+1}$ - Quantilen für μ. Das MCMC-Sampling erfolgt mit einem weiteren R-Paket *rjags* (Version 3.3.0, Plummer, 2009) unter Einsatz von JAGS (Just Another Gibbs Sampler).

Die nächste bei den endlichen Mischungsmodellen aufkommende Frage ist die nach der Anzahl der Komponenten. Falls die Anzahl der Komponenten nicht a priori bekannt ist, können verschiedene Informationskriterien zur Bestimmung von *K* herangezogen werden. Die weitverbreiteten sind dabei das Akaikes Informationskriterium AIC (Akaike, 1973) und das Bayesianische Informationskriterium BIC, auch Schwarzschen Informationskriterium genannt (Schwarz, 1978). Auch andere Kriterien können zur

Bestimmung von *K* herangezogen werden, von denen jedoch keines optimal ist (vgl. Biernacki und Govaert, 1999). Während das Akaikes Informationskriterium die wahre Anzahl der Komponenten zu überschätzen scheint (McLachlan und Peel, 2000), umgeht das BIC diese Verzerrung bei einer korrekten Wahl der Datenverteilung und wird deswegen in mehreren Ausarbeitungen empfohlen (z.B. Fraley und Raftery, 1998) und in dieser Arbeit zur Bestimmung der Komponentenanzahl in den endlichen Mischungsmodellen verwendet. Das Bayesianische Informationskriterium ist definiert durch

$$-2 \cdot log\, p(y \mid \hat{\theta}, \hat{\pi}, M_K) + d_K \cdot log(n)$$

mit den Maximum a posteriori (MAP) - Schätzer des Mischungsmodells M_K , entsprechender Parameteranzahl d_K und n Beobachtungen (vgl. Fritsch, 2010). Beim Vergleich der Modelle mit unterschiedlichen *K*'s anhand des BIC wird das Modell gewählt und somit die Komponentenanzahl *K* bestimmt, wo dieses Kriterium minimal ist. Bei einer anderen Ansicht der Formel, wo alles mit „-1" multipliziert wird, ist als bestes Modell dasjenige zu wählen, wo das Bayesianische Informationskriterium BIC maximal ist (wie z.B. in *mclust()*, Fraley und Raftery, 1999).

3.3.2 Dirichlet Process mixture models

Eine Alternative bieten die unendlichen Mischungsmodelle (engl.: infinite mixture models), bei denen im Vergleich zu den endlichen Mischungsmodellen die Komponentenanzahl nicht vorgegeben werden muss. Im Folgenden werden die wohl bekanntesten Dirichlet Prozess (DP) Modelle vorgestellt, die von Ferguson (1973) eingeführt wurden und auf einem Dirichlet Prozess (Ferguson, 1973; Antoniak, 1974) basieren. Die praktische Umsetzung erfolgt in dieser Arbeit mit Hilfe des R-Pakets *DPpackage* (Version 1.1-6) und der Funktion *DPdensity* (Jara et al., 2012).

Nach Ferguson (1973) ist ein Dirichlet Prozess ein stochastischer Prozess mit Verteilungen über dem Maßraum $(\Omega, \mathcal{A})$, so dass der DP eine Verteilung über Verteilungen ist:

$$G \sim DP(\alpha, G_0)$$

und eine daraus gezogene Zufallsvariable auch eine Verteilung ist. Für jede endliche Menge $(A_1, \dots, A_l)$ auf Ω gilt somit:

$$\big(G(A_1), \dots,\ G(A_l)\big) \sim Dir\big(\alpha G_0(A_1), \dots,\ \alpha G_0(A_l)\big).$$

Dabei ist α $(\alpha > 0)$ der Präzisionsparameter für die Variabilität der Verteilungsfunktion um ihre Basisverteilung G_0 , um die sie zentriert ist. Weiterhin gilt, dass der Erwartungswert und die Varianz durch

$$E\big(G(A)\big) = G_0(A) \text{ bzw.}$$

$$Var\big(G(A)\big) = \frac{G_0(A)\big(1 - G_0(A)\big)}{\alpha + 1}$$

gegeben sind. Die Ähnlichkeit von G zu G_0 ist umso größer, je größer der Präzisionsparameter α ist.

Eine konjugierte a priori-Verteilung für einen Dirichlet Prozess ist wiederum ein Dirichlet Prozess, so dass gegeben n Realisierungen $\theta_1, \dots, \theta_n$ für die zugehörige a posteriori-Verteilung gilt:

$$G \mid \theta \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum\nolimits_{i=1}^{n} \delta_{\theta_i}\right)$$

mit den Punktmassen δ_{θ_i} (vgl. Blackwell und MacQueen, 1973).

Um eine zufällige Verteilungsfunktion G über einen Dirichlet Prozess zu generieren, können über den sogenannten Stick-Breaking-Process die zufälligen Gewichte für Realisierungen aus der Basisverteilung G_0 bestimmt werden. Nach Sethuraman (1994) ist eine formelle Darstellung des Dirichlet Prozesses gegeben durch

$$G = \sum_{k=1}^{\infty} \pi_k\, \delta_{\theta_k} \sim DP(\alpha, G_0)$$

mit $\sum_{k=1}^{\infty} \pi_k = 1$ und $\pi_k \geq 0\ \forall\ k$.

Jara et al. (2012) betrachten mit ihrer Funktion *DPdensity()* eine DP-Mischung von Normalverteilungen für die Dichteschätzung, da sich durch diese Verteilungsart beliebige Verteilungen approximieren lassen (vgl. Escobar und West, 1995):

$$y_i \mid \mu_i, \Sigma_i \sim N(\mu_i, \Sigma_i), \;\; i = 1, \dots, n$$

$$\mu_i, \Sigma_i \mid G \sim G$$

$$G \mid \alpha, G_0 \sim DP(\alpha, G_0)$$

mit der konjugierten Normal-Inverse Wishart-Basisverteilung

$$G_0 = N(\mu \mid m_1, (1/k_0)\Sigma) IW(\Sigma \mid v_1, \psi_1).$$

Die a priori-Parameter werden bei *DPdensity()* durch eine entsprechende Einstellung generiert und sind durch

$$\alpha \mid a_0, b_0 \sim Gamma(a_0, b_0)$$

$$m_1 \mid m_2, s_2 \sim N(m_2, s_2)$$

$$k_0 \mid \tau_1, \tau_2 \sim Gamma(\tau_1/2\,, \tau_2/2)$$

$$\psi_1 \mid v_2, \psi_2 \sim IW(v_2, \psi_2)$$

gegeben (vgl. Jara et al., 2012). Diese können z.B. nach Bensmail et al. (1997) definiert werden mit $k_0 = 1$, $v_1 = 5$ und der Diagonalmatrix mit Stichprobenvarianzen ψ_1. Dieses hat sich in Fritsch (2010) als vorteilhaft erwiesen. Zur Berechnung der a posteriori-Verteilung wird ein Gibbs-Algorithmus herangezogen.

3.3.3 DIRECT

Als modellbasierte Vergleichsmethode wird das Verfahren von Fu et al. (2013) eingesetzt. Es wurde speziell für die Zeitreihen mit wiederholten Messungen entwickelt und basiert auf einem Random-Effects-Mischungsmodell und einem Dirichlet Prozess Prior. Zur besseren Vergleichbarkeit und Nachvollziehbarkeit der Methode werden hier einige Annotationen in Anlehnung an die zugehörige Ausarbeitung (Fu et al., 2013) übernommen.

Es wird angenommen, dass den Genexpressionen die Zufallsvariablen $M_i = (M_{i11}, \dots, M_{iJR})'$ zu Grunde liegen und dass diese gemäß folgender Dichte verteilt sind:

$$f(M_i|\, \mu, \Sigma) = \sum_{k=1}^{K} w_k\, f_k(M_i|\, \mu_k, \Sigma_k).$$

f_k liegt dabei eine multivariate Normalverteilung zu Grunde. Weiterhin wird angenommen, dass $m_i = (m_{i11}, \dots, m_{iJR})'$ die Realisierungen davon sind.

Die Ermittlung der Komponentenanzahl *K* erfolgt hier unter Verwendung des Dirichlet Prozess Priors (vgl. Unterkapitel 3.3.2). Seien in γ_i für jede Beobachtung *i* der Mittelwertvektor über die Zeitpunkte und Wiederholungen μ_i und die drei Variabilitätsterme für ϕ_i, τ_i und ϵ_i enthalten:

$$\gamma_i = \{\mu_i, \lambda_{\phi_i}, \lambda_{\tau_i}, \lambda_{\epsilon_i}\}, \qquad i = 1, \dots, n.$$

Dann folgen hier die γ_i's einer Verteilung *G*, die über einen Dirichlet Prozess modelliert wird:

$$\gamma_i \sim G$$

$$G \sim DP(\alpha, G_0).$$

Wird dem Mittelwertvektor eine multivariate Normalverteilung und den drei Variabilitätstermen eine Gleichverteilung zu Grunde gelegt, so ist die Basisverteilung G_0 als Produkt entsprechender Dichtefunktionen über alle *k*'s definiert (Fu et al., 2013).

Sei M_{ijr} der Expressionswert der *i*-ten Beobachtung für den *j*-ten Zeitpunkt in der Wiederholung r mit $i = 1, \dots, n,\ j = 1, \dots, J$ und $r = 1, \dots, R$. Das Random-Effects-Modell ist bei Fu et al. (2013) definiert durch

$$M_{ijr}|\, \{Z_i = k\} = \Theta_j^k + \phi_i^k + \tau_{ij}^k + \epsilon_{ijr}^k$$

mit der Clusterzugehörigkeitsvariablen Z_i.

Dabei stellt Θ_j^k den „wahren" festen Effekt des *j*-ten Zeitpunktes der *k*-ten Mischungskomponente dar. ϕ_i^k und τ_{ij}^k sind die zufälligen Effekte der Beobachtung *i* innerhalb der Cluster bzw. zu einem Zeitpunkt *j*. ϵ_{ijr}^k kann wie eine Wechselwirkung (ein

Wiederholungseffekt) zwischen den Zeitpunkten und Wiederholungen aufgefasst werden. Weiterhin gilt:

$$E\left(M_{ijr} | \{Z_i = k\}\right) = \Theta_j^k,$$

$$\phi_i^k | \{Z_i = k, \lambda_\phi^k\} \sim_{iid} N(0, \lambda_\phi^k),$$

$$\tau_{ij}^k | \{Z_i = k, \lambda_\tau^k\} \sim_{iid} N(0, \lambda_\tau^k),$$

$$\epsilon_{ijr}^k | \{Z_i = k, \lambda_\epsilon^k\} \sim_{iid} N(0, \lambda_\epsilon^k)$$ (vgl. Fu et al., 2013)

mit der Clusterzugehörigkeitsvariablen Z_i und den entsprechenden Variabilitätstermen λ_ϕ^k, λ_τ^k und λ_ϵ^k.

Insgesamt sind die interessierenden Parameter in ξ folgendermaßen zusammengefasst:

$$\xi = \left\{K, \mu_1, \dots, \mu_K, \lambda_\phi^1, \dots, \lambda_\phi^K, \lambda_\tau^1, \dots, \lambda_\tau^K, \lambda_\epsilon^1, \dots, \lambda_\epsilon^K, Z_1, \dots, Z_n, \alpha\right\},$$

zu deren Bestimmung ein MCMC-Algorithmus, bestehend aus zwei Schritten in jeder Iteration, durchgeführt wird:

1. Für jedes *i* wird die Clusterzugehörigkeit Z_i bestimmt/aktualisiert. Dazu schlagen Fu et al. (2013) einen speziellen komponentenweisen Metropolis-Hastings Sampler vor.
2. Anhand erfolgter Clusterung werden anschließend alle weiteren Modellparameter bestimmt.

Diese Aufteilung in zwei Schritte ist von daher sinnvoll, da der zweite Schritt von dem im ersten Schritt bestimmten *K* abhängt. Die Einzelheiten können dem Supplemental Material in Fu et al. (2013) entnommen werden.

Anschließend definieren die Autoren eine *n x K* Posterior Allocation Probability Matrix *P*, die die Wahrscheinlichkeiten p_{ik} (*k* = 1, …, *K*) enthält, dass die Beobachtung *i* zum *k*-ten Cluster gehört:

$$p_{ik} = P(Z_i = k \mid M)$$

mit $M = (M_1, \dots, M_n)$. Die Bestimmung dieser Matrixeinträge erfolgt in zwei Schritten. Im Resampling-Schritt werden Clusterwahrscheinlichkeiten unter Einsatz von MCMC

bestimmt und Schritt für Schritt das endliche Mischungsmodell definiert. Um sicher zu stellen, dass die Bezeichnungen der Cluster über alle MCMC-Stichproben gleich sind, wird der Relabeling-Algorithmus von Stephens (2000) angewendet. Auf die weiteren methodischen Einzelheiten wird an dieser Stelle auf die entsprechende Ausarbeitung von Fu et al. (2013) verwiesen.

Das zugehörige Tool DIRECT (Version 1.0) wurde in R implementiert und in dieser Arbeit zur Analyse der Daten herangezogen.

3.3.4 Wahl einer optimalen Clusterung

Nach der Anpassung der vorgestellten Bayes-Modelle an die Daten mit Hilfe von MCMC-Algorithmen liegen nach *N* Durchläufen anschließend $\boldsymbol{c}^{(1)}, \dots, \boldsymbol{c}^{(N)}$ MCMC-Samples von Clusterungen $\boldsymbol{c} = (c_1, \dots, c_n)$ vor.

Die Matrix der paarweisen a posteriori-Wahrscheinlichkeiten, dass die Gene *i* und *j*, $i, j = 1, \dots, n$, den gleichen Expressionsverlauf aufweisen und somit zum gleichen Cluster gehören, wird mit

$$\pi_{ij} = P(c_i = c_j | y)$$

bezeichnet (vgl. Medvedovic et al., 2002; Dahl, 2006; Fritsch und Ickstadt, 2009). Diese *n* x *n* - Matrix mit den Einträgen π_{ij} ist die Posterior Similarity Matrix (PSM), durch die eine sinnvolle Zusammenfassung von MCMC-Samples von Clusterungen erfolgen kann.

Die in Unterkapitel 5.2 verwendete Distanzmatrix ist gegeben durch 1 - π_{ij} und enthält die Gegenwahrscheinlichkeiten zu π_{ij}, also die a posteriori-Wahrscheinlichkeiten, dass die Gene *i* und *j* nicht dem gleichen Cluster angehören. Diese Wahrscheinlichkeiten definieren ein sinnvolles Distanzmaß. Wie in der Dissertation von Fritsch (2010) gezeigt, erfüllen sie die Eigenschaften einer Pseudo-Metrik.

Die PSM kann sowohl für die finiten als auch für die infiniten Mischungsmodelle für die sinnvolle Zusammenfassung der Clusterergebnisse benutzt werden. In R kann sie durch den Aufruf der *comp.psm()*-Funktion in dem Paket *mcclust* (Version 1.0) berechnet werden.

Es gibt eine Reihe von Ansätzen, um aus der PSM eine optimale Clusterung abzuleiten (vgl. Dissertation Fritsch, 2010). Folgende zwei kommen in dieser Arbeit zum Einsatz:

- MINBINDER basiert auf der Verlustfunktion von Binder (1978)

$$L(\hat{\boldsymbol{c}}, \boldsymbol{c}) = \sum_{i<j} a \cdot I_{\{\hat{c}_i \neq \hat{c}_j\}} I_{\{c_i = c_j\}} + b \cdot I_{\{\hat{c}_i = \hat{c}_j\}} I_{\{c_i \neq c_j\}}$$

  mit positiven Konstanten *a* und *b* und den 0-1 Matrizen $I_{\{\hat{c}_i = \hat{c}_j\}}$ und $I_{\{c_i = c_j\}}$ mit den geschätzten bzw. wahren unbekannten Ähnlichkeitseinträgen. Mit *a* wird die Vollständigkeit und mit *b* die Homogenität der Clusterung gewichtet. Wie bei Hurn et al. (2003) werden auch in dieser Arbeit beide Konstanten auf 1 gesetzt, um beide ähnlich zu gewichten. Wegen $E(I_{\{c_i = c_j\}} \mid y) = \pi_{ij}$ (vgl. Fritsch und Ickstadt, 2009) ist der a posteriori erwartete Verlust dann gegeben durch

$$E(L(\hat{\boldsymbol{c}}, \boldsymbol{c}) \mid y) = \sum_{i<j} I_{\{\hat{c}_i = \hat{c}_j\}} - \pi_{ij} \, .$$

  Für die Wahl der geschätzten Clusterung gilt diese Funktion zu minimieren, was in R mit *minbinder()* erfolgt.

- Wie von Fritsch und Ickstadt (2009) gezeigt, ist die Maximierung des Rand-Index (Rand, 1971) zwischen $\boldsymbol{c}$ und $\hat{\boldsymbol{c}}$ äquivalent mit der Minimierung von Binders Verlustfunktion (mit $a = b = 1$).

  Wegen seiner Nachteile (vgl. Fritsch, 2010) schlagen Fritsch und Ickstadt (2009) jedoch den adjustierten Rand-Index von Hubert und Arabie (1985) vor (MAXPEAR). Von Fritsch (2010) implementiert, sucht *maxpear()* in R nach der Clusterung, die den a posteriori erwarteten adjustierten Rand-Index zwischen der wahren und der geschätzten Clusterung maximiert:

$$\frac{\sum_{i<j} I_{\{\hat{c}_i = \hat{c}_j\}} I_{\{c_i = c_j\}} - \sum_{i<j} I_{\{\hat{c}_i = \hat{c}_j\}} \sum_{i<j} I_{\{c_i = c_j\}} / \binom{n}{2}}{1/2 \left[\sum_{i<j} I_{\{\hat{c}_i = \hat{c}_j\}} + \sum_{i<j} I_{\{c_i = c_j\}} \right] - \sum_{i<j} I_{\{\hat{c}_i = \hat{c}_j\}} \sum_{i<j} I_{\{c_i = c_j\}} / \binom{n}{2}} .$$

  Dabei ist $\binom{n}{2}$ folgendermaßen definiert: Bei der Betrachtung von zwei Clusterungen werden sowohl Übereinstimmungen (zwei Beobachtungen werden dem gleichen bzw. unterschiedlichen Clustern zugeordnet) als auch Unterschiede

zwischen den Clustern ermittelt. Seien Übereinstimmungen in A und Unterschiede in D festgehalten. Dann gilt:

$$A + D = \binom{n}{2}$$ (Rand, 1971).

In dieser Arbeit werden somit optimale Clusterungen basierend auf den an die Daten angepassten Bayes-Modellen sowie der Posterior Similarity Matrix und den darauf aufbauenden Ansätzen MINBINDER und MAXPEAR ermittelt. Allerdings ist bei den endlichen Mischungsmodellen mit einer festen Anzahl von Komponenten hierdurch nicht garantiert, dass die resultierende Clusterung auch ebendiese Clusteranzahl hat.

3.4 Überlebenszeitanalyse mit dem Cox-Modell

Der Begriff Überlebenszeit (oder besser gesagt, ereignisfreie Überlebenszeit) bezeichnet die Zeit zwischen der Aufnahme des Patienten in eine Studie bis zum Eintritt des untersuchten Ereignisses. Bei einer Überlebenszeitanalyse ist die Zeitdauer bis zum Eintreten eines Ereignisses die Zielvariable. Dabei kann diese nicht exakt berechnet werden, falls die Studie vorzeitig abgebrochen wird oder der Patient wegen einer anderen Krankheit als beobachtet verstirbt. In solchen Fällen kann keine exakte Überlebenszeit ermittelt werden und die Daten werden zensiert. Ferner wird zwischen einer Rechts- und einer Linkszensierung unterschieden, je nachdem, welcher der beiden Zeitpunkte bekannt ist. Bei der Untersuchung medizinischer Fragestellungen werden die Daten meist rechtszensiert, z.B. falls das Zielereignis bis zum Studienende nicht beobachtet werden konnte.

Durch das multiplikative Regressionsmodell wird die so genannte Hazard-Rate

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq X \leq t + \Delta t \mid X \geq t)}{\Delta t}, \quad t > 0,$$

modelliert, welche die Wahrscheinlichkeit des Eintretens des untersuchten Ereignisses zum Zeitpunkt *t* angibt, vorausgesetzt, dass das Zielereignis (z.B. Krankheit oder Tod) bis zu diesem Zeitpunkt noch nicht eingetreten ist.

Die bekannteste Methode zur Analyse von Überlebenszeitdaten ist das proportionale Cox-Modell oder auch Proportionales Hazardratenmodell (Cox, 1972), das zur Untersuchung des Einflusses einer oder mehrerer Größen auf die Zielvariable eingesetzt werden kann. Hier wird die Hazard-Rate durch

$$h(t,z) = h_0(t)exp(\beta' z)$$

modelliert mit gegebenem Kovariablenvektor $z = (z_1, \dots, z_j)$ und den zu schätzenden Regressionsparameter dieser Einflussvariablen $\beta = (\beta_1, \dots, \beta_j)$. $h_0(t)$ ist dabei die sogenannte Baseline-Hazard-Rate und gibt das Ausfallrisiko zum Zeitpunkt *t* an, wenn der Kovariablenvektor *z* gleich dem Nullvektor ist.

Sei $L_i(\beta)$ die bedingte Wahrscheinlichkeit des Ereignisauftretens für die Person *i* zum Zeitpunkt t_i, $i = 1, \dots, I$, unter der Voraussetzung, dass keine Bindungen zwischen $t_1, \dots, t_I$ vorliegen. Ferner wird angenommen, dass alle Kovariablenausprägungen der Person *i* im Vektor z_i bekannt sind, so dass $h(t_i, z_i) = exp(\beta' z_i)$. Die Schätzung von β erfolgt dann durch die Maximierung der partiellen Likelihood $L(\beta)$:

$$L(\beta) = \prod_{i=1}^{I} \left(L_i = \frac{exp(\beta' z_i)}{\sum_{k \in N_{t_i}} exp(\beta' z_k)} \right)$$

mit N_{t_i} als der Menge aller Personen unter Risiko zum Zeitpunkt t_i (vgl. Anderson und Keiding, 2006). Treten zwischen den Ereigniszeiten doch Bindungen auf, so wird beispielsweise nach Efron (1977) die partielle Likelihood durch

$$L^*(\beta) = \prod_{i=1}^{I} \frac{exp\left(\beta'\left(\sum_{g \in D_i} z_g\right)\right)}{\prod_{g=1}^{d_i} \left[\sum_{k \in N_{t_i}} exp(\beta' z_k) - \frac{g-1}{d_i} \sum_{k \in D_i} exp(\beta' z_k) \right]}$$

approximiert, mit der Ereignisanzahl d_i zum Zeitpunkt t_i und mit der Anzahl der Personen D_i, bei denen das Zielereignis bis zu diesem Zeitpunkt schon beobachtet werden konnte.

Die Maximierung der partiellen Likelihood erfolgt in zwei Schritten:

- Zuerst wird die zugehörige Log-Likelihoodfunktion $LL(\beta)$ berechnet.

- Anschließend wird mit Hilfe eines Newton-Raphson-Algorithmus folgende Gleichung des zugehörigen Scores $U_i(\beta)$ gelöst:

$$U_i(\beta) = \frac{\partial}{\partial \beta_i} LL(\beta) = 0 \text{ für } 1 \leq i \leq j.$$

  Nach Therneau und Grambsch (2000) ist der hierdurch ermittelte ML-Schätzer $\hat{\beta}$ konsistent und asymptotisch normalverteilt mit einem Erwartungswert β und Varianz $\left(\frac{\partial^2}{\partial \beta^2} LL(\beta)\right)^{-1}$.

Im Cox-Modell wird ein konstanter Variableneffekt auf die Zielgröße über die Zeit angenommen. Die zentrale Annahme ist jedoch die Proportionalität der Hazard-Raten zu einander:

$$\frac{h(t, z_A)}{h(t, z_B)} = \frac{h_0(t) exp(\beta' z_A)}{h_0(t) exp(\beta' z_B)} = exp\big(\beta'(z_A - z_B)\big)$$

für zwei verschiedenen Kovariablenvektorausprägungen z_A und z_B (vgl. Klein und Moeschberger, 2003). Dieser Quotient der Ereignisraten ist somit zeitunabhängig und konstant. Er wird im Ergebniskapitel 5 mit HR bezeichnet und gibt durch die Exponentialfunktion die Richtung der Risikoveränderung pro Variableneinheit.

Um feststellen zu können, welche Metagene einen Einfluss auf die Überlebenszeit ausüben, wird in dieser Arbeit ein nichtparametrischer zweiseitiger Log-Rank (Score) Test herangezogen, der bei dieser Art der Fragestellung häufig eingesetzt wird. Die zu prüfenden Hypothesen hierbei sind

$$H_0\colon\ \beta = \beta_0 \ \text{ vs. } \ H_1\colon\ \beta \neq \beta_0\ ,$$

also ob der Regressionsparameter für das jeweilige Metagen von β_0 (hier gleich Null) von Null verschieden ist (vgl. Klein und Moeschberger, 2003). In dieser Arbeit wird somit der Effekt des einzelnen Metagenes z auf die Prognose $h(t, z)$ in Bezug auf die metastasenfreie Zeit geschätzt.

Unter der Nullhypothese ist die Ableitung der Loglikelihood-Funktion $U(\beta)$ normalverteilt mit dem Erwartungwert 0 und Varianz $I(\beta)$:

$$U(\beta) \sim N\big(0, I(\beta)\big),$$

wobei $I(\beta)$ der beobachteten Fisher-Informationsmatrix bis auf das Vorzeichen der logarithmierten partiellen Likelihood, zweimal abgeleitet nach β, entspricht (Hosmer und Lemeshow, 1999):

$$I(\beta) = -\frac{\partial^2}{\partial \beta^2} LL(\beta) \, .$$

Die zugehörige Teststatistik ist gegeben durch

$$\chi^2 = U(\beta_0)' I^{-1}(\beta_0) U(\beta_0)$$

und ist χ^2- verteilt mit *j* Freiheitsgraden. Die Nullhypothese wird zum Signifikanzniveau α abgelehnt, falls der Wert dieser Teststatistik größer als das $(1-\alpha)$ - Quantil dieser Verteilung ist (Hosmer und Lemeshow, 1999).

4 Biologisch-statistische Verfahren

Nach der erfolgreich durchgeführten Clusteranalyse liegen mehrere Gengruppen mit ähnlichen Expressionsverläufen vor, die im Hinblick auf ihren biologischen Hintergrund vorerst noch keine Information liefern. Da die aufgedeckte Co-Expression dabei das Ergebnis wichtiger Regulationsmechanismen sein könnte, sind diese Gengruppen die interessanten Kandidaten für die anschließenden Metagenberechnung und die Überlebenszeitanalyse.

Das Unterkapitel 4.1 beschreibt zwei Möglichkeiten zur Ermittlung der biologischen Hintergrundinformation aus den schon geclusterten Daten. Im Unterkapitel 4.2 wird ein Ansatz vorgestellt, das biologische Wissen noch vor der eigentlichen Clusteranalyse der Daten zu berücksichtigen. Auf die Metagenberechnung wird in 4.3 eingegangen.

Somit erfolgt in diesem Kapitel eine Vorstellung der biologisch-statistischen Verfahren zur Identifizierung biologisch relevanter Gengruppen.

4.1 EXPANDER

Shamir et al. (2005) haben für alle interessierten Wissenschaftler ein Programm zur Verfügung gestellt, das mehrere Möglichkeiten zur Analyse der Microarray-Daten bietet (http://www.cs.tau.ac.il/~rshamir/expander). Mit EXPANDER (EXPression ANalyzer and DisplayER) können die Expressionsdaten bspw. normalisiert und nach vordefinierten Einstellungen gefiltert werden. Sie können geclustert und/oder in Bezug auf deren biologische Relevanz untersucht werden. Hierbei unterstützt und erkennt EXPANDER die Gen-Annotationen (s. Unterkapitel 2.1) zu mehreren Organismen (inzwischen schon zu 15 mit der Version 6.06, die während der Fertigstellung der vorliegenden Arbeit erschienen ist). Die im Kapitel 5 vorgestellten Ergebnisse werden mit EXPANDER2.0 bzw. EXPANDER5Win erzeugt.

Unter EXPANDER laufen zwei Programme, die in dieser Arbeit zur Promoteranalyse und zur Analyse der potentiell relevanten Cluster hinsichtlich derer GO-Gruppenassoziationen herangezogen werden: PRIMA und TANGO.

PRIMA (PRomoter Integration in Microarray Analysis): Das Ziel der Promoteranalyse ist das Auffinden von Transkriptionsfaktoren (TF), deren Bindungsstellen in einem Set von Genen signifikant überrepräsentiert sind im Vergleich zur vordefinierten Gengruppe. Als Input wird die Gruppeneinteilung der Probesets mit gleichen Expressionsmuster über die Zeit aus der Clusteranalyse, z.B. DIB-C (vgl. Unterkapitel 3.2.3) erwartet. PRIMA sucht nach den Gengruppen, in denen sich gehäuft bestimmte Transkriptions-faktorbindungsstellen finden lassen. Dafür ziehen die Autoren die so genannten „position weight matrices" (PWM) für die bekannten menschlichen TF's (vgl. Wingender et al., 2000) heran und berechnen durch

$$sim(PWM, s_1, s_2, \dots, s_l) = \prod_{j=1}^{l} p(s_j, j)$$

die Ähnlichkeit der zu untersuchten Promoter-Sequenzen $s_1, s_2, \dots, s_l$ zu den bekannten PWMs. Mit $p(i, j)$ bezeichnen die Autoren die Basisfrequenz i an der j-ten Position der entsprechenden PWM der Länge l (hierzu und zu den weiteren Einzelheiten wird auf Elkon et al. (2003) verwiesen).

TANGO (Tool for ANalysis of GO enrichments) bietet eine alternative Möglichkeit zur Identifizierung von biologisch relevanten Clustern unter Berücksichtigung der GO-

Kriterien (z.B. Glahn et al., 2008). Mit Hilfe von hypergeometrischen Anreicherungstests und unter Berücksichtigung multipler Signifikanzniveaus werden hierbei Gene Ontology Gruppen aufgefunden, die in einem Set von Genen signifikant überrepräsentiert sind im Vergleich zur vordefinierten Gengruppe. Dahinter steht das mehrfache Ziehen der Gene zu den Zufallsstichproben der gleichen Größe wie das zu analysierende Cluster und die Berechnung der Verteilung maximaler p-Werte für deren funktionelle Anreicherung, um hierdurch die Schwelle für eine signifikante Anreicherung zu ermitteln (vgl. Shamir et al., 2005; Tanay, 2005). Laut Shamir et al. (2005) sind weitere methodische Einzelheiten zu TANGO in Vorbereitung. Auch hier werden die Gengruppen mit ähnlichen Expressionsverläufen über die Zeit ausgewertet, die nach der vorangegangenen Clusteranalyse ermittelt wurden. In dieser Arbeit sind es dieselben Gruppen wie bei der Promoteranalyse mit PRIMA.

Die nach PRIMA bzw. TANGO ermittelten Gencluster sind somit nicht nur aus statistischer Sicht signifikant, sondern auch aus biologischer Sicht potentiell relevant für die Gewinnung neuer Erkenntnisse bzgl. der Brustkrebsentstehung. Sie werden deshalb zur anschließenden Metagenberechnung und der darauf folgenden Überlebenszeit-analyse herangezogen.

Der Hintergrund, gegenüber dem eine Häufung geprüft wird, die Einschränkung bzgl. der Clustergröße oder auch die relevanten GO-Ontologien können vom Benutzer vorgegeben werden, sowie das Signifikanzniveau und die Entscheidung bzgl. der Bonferroni-Korrektur. Durch die Auswahl des Hintergrundes werden bei PRIMA jedoch einige Einstellungen automatisch generiert (wie z.B. die entsprechende PWM oder die Promoter-Sequenzen).

Auf die Einzelheiten zu PRIMA bzw. TANGO sowie zu den weiteren Programmen, die unter EXPANDER laufen, wie z.B. CLICK (Clusteranalyse) oder SAMBA (Biclustering), wird an dieser Stelle auf die Ausarbeitung von Shamir et al. (2005) verwiesen.

4.2 Ähnlichkeitsmaß der Genprodukte

In diesem Kapitel wird ein von Schlicker et al. (2006) eingeführtes Maß für die Ähnlichkeit zweier Gene Ontology Terme und aufbauend darauf auch zwischen den Genprodukten vorgestellt, das in Kapitel 5 dieser Arbeit zur Identifizierung biologisch relevanter Gengruppen herangezogen wird. Dabei ist dieses Maß nur eines von vielen, auf diesem Gebiet vorgestellten Ansätzen (z.B. Speer et al., 2004; Lee und Lee, 2005; Sevilla et al., 2005; Lord et al., 2003).

In der Arbeit von Schlicker et al. (2006) wird, um die Ähnlichkeit zwischen zwei GO Termen g_1 und g_2 zu ermitteln, zunächst die Häufigkeit ihres Auftretens in den Annotationen der GO-Datenbank (vgl. Unterkapitel 2.1) bestimmt:

$$freq(g_1) = anno(g_1) + \sum_{g \in child(g_1)} freq(g).$$

Dabei bezeichnet $anno(g_1)$ die Anzahl der zu g_1 annotierten Genprodukte und $child(g_1)$ die Menge aller zugehörigen Kinderknoten. Für g_2 gilt analog:

$$freq(g_2) = anno(g_2) + \sum_{g \in child(g_2)} freq(g).$$

Nach Lord et al. (2003) ist ein GO Term für die Wissenschaftler umso interessanter, je weniger Information in der GO-Datenbank darüber enthalten ist, also je unwahrscheinlicher er ist. In Schlicker et al. (2006) ist die Wahrscheinlichkeit seines Auftretens (= Informationsgehalt) durch das Verhältnis seiner Häufigkeit und der Häufigkeit des GO Terms an der Wurzel (vgl. Unterkapitel 2.1) gegeben:

$$p(g_1) = \frac{freq(g_1)}{freq(Wurzel)}.$$

Die Relevanz des GO Terms g_1 kann demzufolge durch $1 - p(g_1)$ geschätzt werden.

Zum Vergleich zweier GO Terme g_1 und g_2 kann jetzt deren Ähnlichkeit nach Resnik et al. (1995) und nach Lin (1998) herangezogen und nach dem Vorschlag von Schlicker et al. (2006) kombiniert werden:

Für den maximalen gemeinsamen Informationsgehalt nach Resnik et al. (1995), wobei der Informationsgehalt hier über - $log_{10}p(g_1)$ bzw. - $log_{10}p(g_2)$, gemessen wird, gilt:

$$sim_{Resnik}(g_1, g_2) = \max_{g \in V(g_1, g_2)} \{-\log p(g)\}.$$

Hierbei ist $V(g_1, g_2)$ die Menge der gemeinsamen Vorfahren beider Genterme. In seinem Paper verwendet Resnik $1 - p(g)$ als Alternative zu $-log\, p(g)$. Der entscheidende Vorteil hierbei ist, dass hierdurch die Ähnlichkeit im Bereich [0, 1] bleibt und somit besser interpretierbar als $-log\, p(g)$ ist.

Lin (1998) benutzt als Ähnlichkeitsmaß den Anteil der in beiden Termen gleichermaßen vorhandenen Information, $2log\, p(g)$ an der Gesamtinformation in g_1 und g_2:

$$sim_{Lin}(g_1, g_2) = \max_{g \in V(g_1, g_2)} \left\{ \frac{2\, log\, p(g)}{log\, p(g_1) + log\, p(g_2)} \right\}.$$

Zur Definition des Ähnlichkeitsmaßes der Relevanz zwischen g_1 und g_2 nach Schlicker et al. (2006) werden diese beiden Ähnlichkeitsmaße nun kombiniert zu

$$sim_{Rel}(g_1, g_2) = \max_{g \in V(g_1, g_2)} \left\{ \frac{2\, log\, p(g)}{log\, p(g_1) + log\, p(g_2)} (1 - p(g)) \right\}.$$

Hieraus ist ersichtlich, dass die Ähnlichkeit eines Terms zu sich selbst durch seine Relevanz gegeben ist. Außerdem ist $sim_{Rel}(g_1, g_2)$ symmetrisch und kann Werte im Bereich von [0, 1] annehmen:

- $sim_{Rel}(g_1, g_2) = sim_{Rel}(g_2, g_1)$,
- $sim_{Rel}(g_1, g_2) \in [0, 1]$ (vgl. Schlicker et al., 2006).

Dieses Ähnlichkeitsmaß ist im R-Paket *GOSim()* implementiert und kann durch die Funktion *getTermSim()* berechnet werden.

Seien $GO^A = \{GO_1^A, \ldots, GO_N^A\}$ und $GO^B = \{GO_1^B, \ldots, GO_M^B\}$ die Mengen aller zu den Genprodukten *A* bzw. *B* annotierten GO Terme. Zum Vergleich dieser Genprodukte wird nun eine *N* x *M* - Matrix *S* berechnet, die als Einträge die paarweisen Relevanzähnlichkeiten (berechnet wie oben beschrieben) zwischen den GO Termen dieser Genprodukte enthält. Diese werden nun mit $s_{ij} = sim_{Rel}(GO_i^A, GO_j^B)$ bezeichnet, $i = 1, \ldots, N$ und $j = 1, \ldots, M$ (vgl. Schlicker et al., 2006).

Die Zeilen bzw. Spalten der Matrix *S* stellen zwei unterschiedliche Vergleiche der Genprodukte dar: *A* im Hinblick auf *B*, definiert durch den Zeilenscore

$$Score_{Zeilen} = \frac{1}{N}\sum_{i=1}^{N} \underset{1 \leq j \leq M}{max}\, s_{ij}$$

und *B* im Hinblick auf *A* analog mit dem Spaltenscore

$$Score_{Spalten} = \frac{1}{M}\sum_{j=1}^{M} \underset{1 \leq i \leq N}{max}\, s_{ij}$$

(vgl. Schlicker et al., 2006).

Auch hier schlagen die Autoren Kombinationen dieser beiden Scores vor, die Werte im Intervall von [0, 1] annehmen. Dazu werden zwei Möglichkeiten vorgestellt: das arithmetische Mittel und das Maximum dieser Scores. Beide Möglichkeiten können in R mit der Funktion *getGeneSim()* und mit der Option *similarity* = „funSimAvg" bzw. *similarity* = „funSimMax" berechnet werden. Dabei soll beachtet werden, dass ein hoher durchschnittlicher Gesamtscorewert auch dann erzielt wird, wenn Genprodukte multifunktionell oder die zugehörigen Annotationen unvollständig sind. In diesem Fall soll nach Empfehlung von Schlicker et al. (2006) die „funSimMax" zum Vergleich von Genprodukten herangezogen werden.

Bei der in diesem Abschnitt vorgestellten Art der Ähnlichkeitsberechnung der Genprodukte wird darauf hingewiesen, dass hierbei nur Gene berücksichtigt werden können, zu denen eine Entrez ID vorliegt und die zur derselben Ontologie gehören.

4.3 Metagenberechnung

Durch die Microarray-Experimente wird ein enormes Datenvolumen generiert. Mittels Metagenbildung kann die Reduktion der Gendimension von mehreren Tausend Genen zu einer deutlich kleineren Anzahl von Metagenen ermöglicht werden. So können mehrere Gene mit ähnlichen Expressionen zueinander in Bezug gebracht werden.

Dadurch wird den Wissenschaftlern die Möglichkeit gegeben, wichtige Zusammenhänge zwischen den einzelnen Genen und deren Einfluss auf Entstehung komplexer Krankheiten, wie bspw. Brustkrebs, besser zu verstehen.

Metagene haben sich schon mehrmals als prognostische Faktoren für das Überleben der Patienten erwiesen (z.B. Rody et al., 2009; Cadenas et al., 2010; Petry et al., 2010). Dabei gibt es eine Vielfalt von Ansätzen für die Metagenberechnung. So wird das Metagen bei Schmidt et al. (2008) als der Median über die Genexpressionen jedes Patienten in der entsprechenden Kohorte, normiert mit dem Median aller beobachteten Expressionswerte eines Probesets, definiert. Wardle et al. (2006) setzen die lineare Interpolation von direkt benachbarten Genen und die Mittelwertbildung zur Metagenkonstruktion ein. Bei Ghazoui et al. (2011) kommen GO-Information und Scoring zum Einsatz. So werden hier die Top 100 Gene, sortiert nach dem berechneten Score, in 2 Metagene aufgesplittert: in den Proliferationsmetagen mit Genen, die nach der GO-Klassifikation mit der Proliferation assoziiert sind, und in das Metagen mit den Medianen übrig gebliebener Gene. Urgard et al. (2011) erklären die Mittelwerte in den Genclustern zu den Metagenen.

Im Folgenden wird das Vorgehen von Schmidt et al. (2008) sowie drei weitere Modifikationen dieser Prozedur vorgestellt.

Jedes Gen wird durch mehrere Probesets *i*, $i = 1, \ldots, n$, repräsentiert und jedes Gencluster besteht aus *k* Genen. Schmidt et al. (2008) haben alle Genexpressionen für den Patienten *j*, $j = 1, \ldots, m$, in der entsprechenden Kohorte mit dem Median aller beobachteten Genexpressionswerte eines Probesets normiert:

$$(a) \qquad x_{ij\_norm\_a} = \frac{x_{ij}}{\tilde{x}_i} \ .$$

Diese Art der Normierung der Expressionswerte kann modifiziert werden und wird in (b) - (d) vorgestellt:

$$(b) \qquad x_{ij\_norm\_b} = \frac{x_{ij}}{\tilde{x}_i} - 1 = \frac{x_{ij} - \tilde{x}_i}{\tilde{x}_i} \ ,$$

$$(c) \qquad x_{ij\_norm\_c} = \frac{x_{ij} - \bar{x}_i}{\sigma_{x_i}} \qquad \text{und}$$

$$(d) \qquad x_{ij\_norm\_d} = \frac{x_{ij} - \tilde{x}_i}{MAD(x_i)} \ .$$

Dabei stellt die Variante (b) die zentrierte Normierung der Werte dar. $x_{ij\_norm\_c}$ ist die in der Statistik am häufigsten verwendete Standardisierung mit dem Mittelwert $\bar{x}_i$ und Standardabweichung σ_{x_i} aller beobachteten Genexpressionswerte eines Probesets *i*. Die Variante (d) basiert auf der standardisierten medianen absoluten Abweichung $MAD(x_i)$

mit $MAD = {\sigma}/{1{,}483}$ (vgl. Dokumentation zu R) und ist für Normalverteilungen die robuste Version von (c).

Der Median über die nach einer dieser Varianten normierten Expressionswerte zusammengehörender Probesets des Patienten *j* wird als Metagen *z* für diese Person und Gencluster *C* bezeichnet:

$$z_{j,C} = \left(\tilde{x}_{ij\_norm}\right)_{i \in C}.$$

In Anlehnung an die Ergebnisse einer früheren Ausarbeitung (vgl. Freis et al., 2009), in der diese vier vorgestellten Metagenberechnungsmöglichkeiten miteinander verglichen wurden, wird in dieser Arbeit die Variante (d) zur Normierung der Genexpressionswerte verwendet.

5 Ergebnisse

In diesem Kapitel erfolgt eine Analyse der Daten mit Hilfe der zuvor beschriebenen Methoden. Die Vorgehensweise dabei erfolgt nach dem in Unterkapitel 3.1 vorgestellten Ansatz, der sich für jedes in dieser Arbeit vorgestellte Clusteranalyseverfahren nur im ersten Schritt unterscheidet und hier im Hinblick auf die folgenden Ergebnistabellen auf eine etwas andere Weise dargestellt wird.

So werden für jede Methode im ersten Schritt der Analyse Gengruppen mit ähnlichen Expressionsverläufen in den humanen Mammakarzinom-Zelllinien identifiziert. Die Anzahl signifikanter, an dieser Stelle noch nicht Bonferroni-korrigierter Cluster wird festgehalten und in der Ergebnistabelle neben dem entsprechenden Verfahren notiert (vgl. z.B. Abb. 5.1.2.1). Dabei werden nur Gengruppen mit mindestens 5 Probesets berücksichtigt.

Im nächsten Schritt werden diese Cluster auf ihre biologische Relevanz mit Hilfe entsprechender Methoden – wie z.B. die GO- bzw. Promoteranalyse – geprüft und somit die biologisch interessanten Cluster ermittelt. Die daraus resultierende, reduzierte Anzahl der Gengruppen wird ebenfalls neben dem entsprechenden Ansatz in der Ergebnistabelle festgehalten. Auch hier sind die Ergebnisse noch nicht korrigiert für multiples Testen.

In der darauf folgenden Analyse dieser MCF7-Cluster bilden die Brustkrebsdaten die Datengrundlage. So werden für diese Gengruppen im nächsten Schritt des Ansatzes die Metagene gebildet, die die mittlere Expression dieser Cluster widerspiegeln.

In der anschließenden Überlebenszeitanalyse wird die prognostische Relevanz dieser Metagene in den Brustkrebsdaten sowohl in der Gesamtkohorte als auch in den einzelnen Kohorten nach den folgenden zwei Szenarien ermittelt:

a) Das Metagen ist in der Gesamtkohorte (GK) mit 766 Patientinnen adjustiert signifikant und in mindestens zwei weiteren Einzelkohorten (wenn auch nicht adjustiert) signifikant.

b) Das Metagen ist in der Gesamtkohorte und in der Rotterdam-Kohorte adjustiert signifikant und in mindestens einer weiteren Kohorte signifikant, adjustiert nach der Anzahl der in der Rotterdam-Kohorte signifikanten Cluster (vgl. Lohr et al., 2012. Hier wird analysiert und diskutiert, in welcher Weise/Reihenfolge für das multiple Testen u.a. auf den dieser Arbeit zugrunde liegenden Daten analysiert werden sollte.).

Die entsprechenden Szenarien werden mit a) und b) in den Ergebnistabellen vermerkt.

Da die zentrale Variable der vorliegenden Arbeit die metastasenfreie Zeit ist, ist das Auftreten einer Metastase je Kohorte von Interesse und kann Abbildung 5.1 entnommen werden:

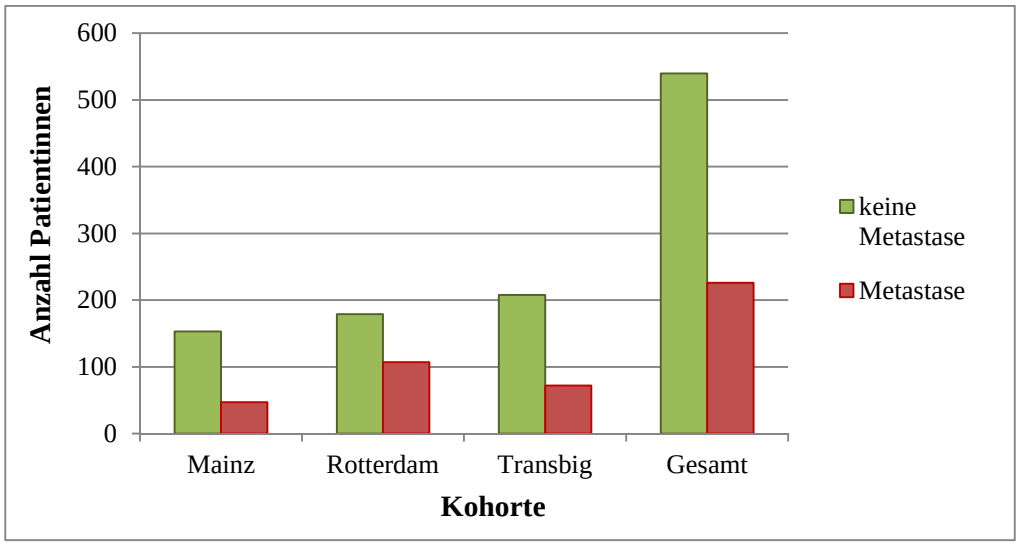

Abb. 5.1: Anzahl Patienten mit festgestellten Metastasen in den vorliegenden Daten je Kohorte

Daraus ist ersichtlich, dass zum Ende der Studie in allen Kohorten außer der Rotterdam-Kohorte bei mehr als der Hälfte der Patientinnen noch keine Metastase beobachtet werden konnte. Der Anteil aufgetretener Todesfälle infolge des Tumors bei den Patientinnen mit einer Metastase ist leider nicht darstellbar, da zu der Rotterdam-Kohorte die entsprechenden Daten komplett fehlen (vgl. Kapitel 2.3).

Einen weiteren interessanten Aspekt und u.U. eine andere Betrachtungsweise der Analyseergebnisse würden z.B. das Alter der Patientinnen zum Zeitpunkt der Diagnose oder der Tumorgrad bzw. die Tumorgröße liefern. Die entsprechenden Angaben wurden jedoch bei der Rotterdam-Kohorte nicht erfasst, so dass dieser Punkt bei der Analyse der Daten leider nicht betrachtet werden kann.

Alle Tests wurden zum Signifikanzniveau $\alpha = 5\%$ durchgeführt und, wenn nicht explizit anders erwähnt, mit der Anzahl der durchgeführten Tests Bonferroni-korrigiert. Die gesamte Analyse erfolgte mit dem Softwareprogramm R (R Development Core Team, 2010). Die Analyse biologischer Relevanz erfolgte u.a. mit dem Programm EXPANDER (vgl. Shamir, 2005). Als Hintergrund, gegenüber dem eine Häufung geprüft wurde, sind alle menschlichen Gene gewählt worden. Ferner ist hier keine weitere Einschränkung vorgenommen worden. Die Graphiken wurden mit Microsoft Office Excel 2007 erzeugt.

5.1 Nicht-modellbasierte Verfahren

Zur besseren Übersicht über die einzelnen Ergebnisse erfolgt an dieser Stelle deren Aufteilung in die Ergebnisse der nicht-modellbasierten und modellbasierten Verfahren. In diesem Unterkapitel erfolgt eine detaillierte Analyse der Daten mit den nicht-modellbasierten Clustermethoden. Hierbei werden sowohl die Ergebnisse der Clusteranalyse als auch der darauf folgenden Analyse der prognostischer Relevanz der Cluster vorgestellt.

5.1.1 *k*-means

Bei der Auswertung der vorliegenden Zeitreihen mit dem am häufigsten eingesetzten und beliebten Clusterverfahren *k*-means wird im ersten Schritt die Fehlerquadratsumme zur Bestimmung der optimalen Clusteranzahl *k* betrachtet. Deren graphische Darstellung ist Abbildung 5.1.1.1 zu entnehmen:

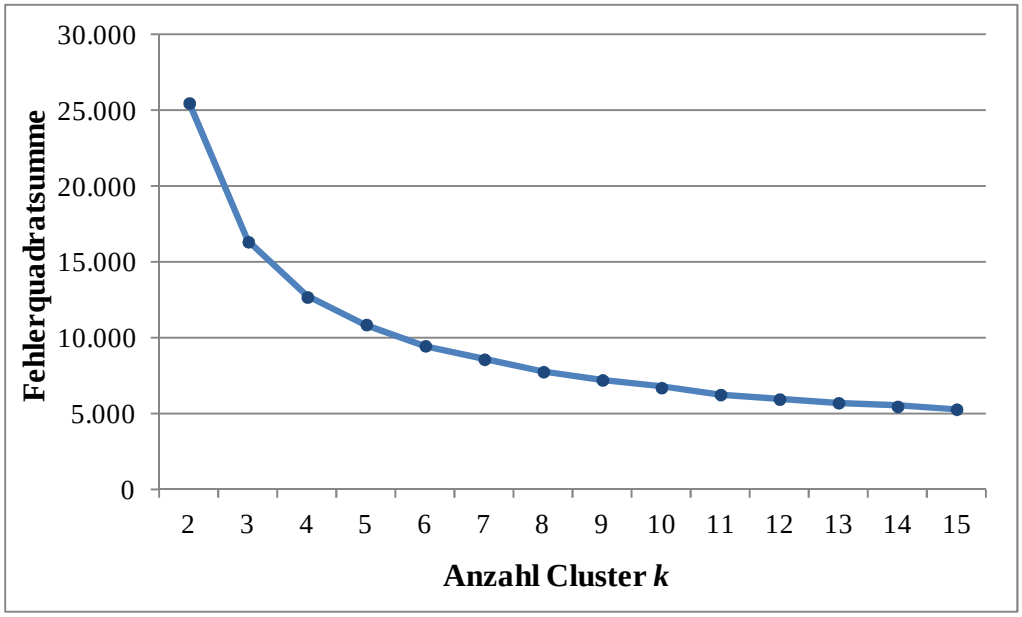

Abb. 5.1.1.1: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (ohne Berücksichtigung der GO-Zuordnung)

Für eine eindeutige Entscheidung sollte in dieser Kurve ein Knick erkennbar sein. Die entsprechende Stelle wäre dann die optimale Anzahl der Cluster. Da dies hier jedoch nicht der Fall ist, würde die Entscheidung rein intuitiv auf *k* zwischen 8 und 10 fallen.

An dieser Stelle werden nur die Ergebnisse für die MCF7-Daten ohne GO-Gruppenzuordnung dargestellt, da die Entscheidungen mit der Berücksichtigung der GO-Information sehr ähnlich ausfallen. Die zugehörigen Abbildungen können dem Anhang B entnommen werden.

Im nächsten Schritt werden nun die Silhouetten-Werte berechnet, um so ggf. doch noch auf die wahre Anzahl der Cluster zu kommen. Diese Werte können Abbildung 5.1.1.2 entnommen werden. Die Entscheidung fällt hier eindeutig auf $k = 2$, da an der entsprechenden Stelle der durchschnittliche Silhouetten-Wert maximal ist. Bei einem Wert von 0,7 kann vorerst von einer guten Datenklassifikation ausgegangen werden.

Diese Entscheidung passt jedoch nicht zu der Schlussfolgerung anhand der Betrachtung der Kurve der Fehlerquadratsumme. Ein möglicher Grund dafür ist die Vernachlässigung der Zeitreihendynamik bei diesem Verfahren.

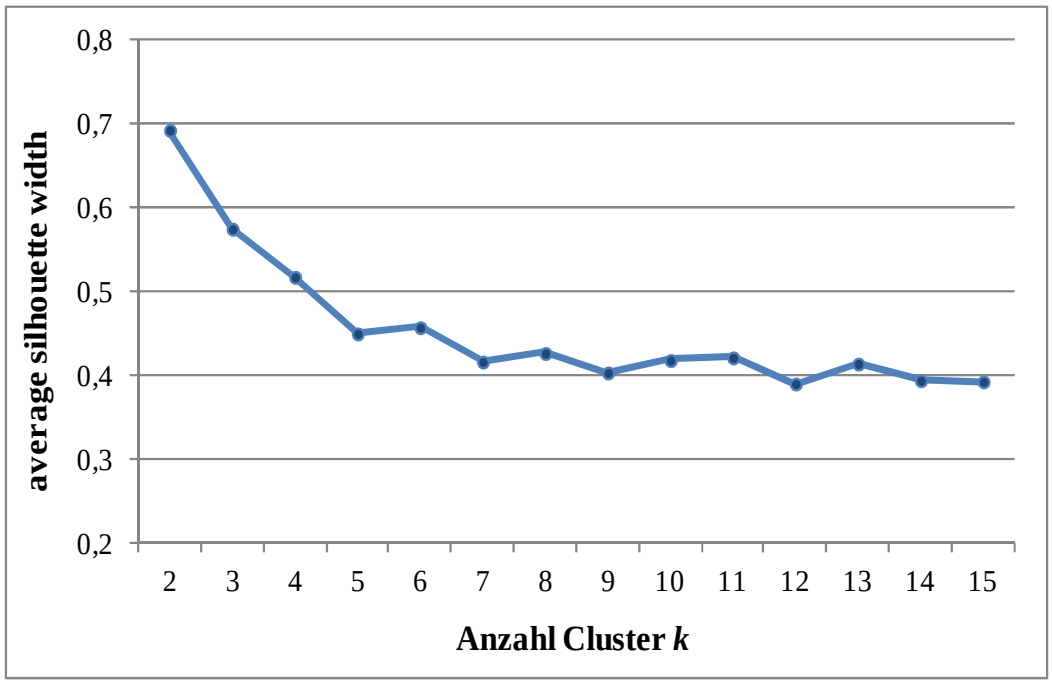

Abb. 5.1.1.2: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (ohne Berücksichtigung der GO-Zuordnung)

Außerdem werden bei 2632 Genen bei $k = 2$ höchstwahrscheinlich keine aussagekräftigen Cluster gebildet. Und da die Konvergenz schon nach 10 Schritten erreicht ist und demzufolge eine Betrachtung von einer größeren Clusteranzahl zwecklos ist, kann an dieser Stelle festgehalten werden, dass eine der gängigsten Clustermethoden – *k*-means – zur Analyse vorliegender komplexer Microarray-Zeitreihendaten wenig geeignet ist. Für die weitere Analyse der Daten wird auf dieses Verfahren deswegen und wegen widersprüchlicher Ergebnisse beider Entscheidungskriterien verzichtet.

5.1.2 STEM

Die vorliegenden Daten wurden ebenfalls mit dem STEM-Verfahren ausgewertet, das in Unterkapitel 3.2.2 vorgestellt wurde. Dabei wurden verschiedene Parameter-einstellungen für die Auswahl der Modellprofile vorgenommen. So wurden für die

Parameter *c* und *m* jeweils die Werte *c* = 1, *c* = 2 und *c* = 3 bzw. *m* = 10, *m* = 50, *m* = 60 sowie *m* = 100 gewählt (vgl. Krahn, 2008). Bei einer Laufzeit von ca. 30 Minuten ergab die anschließende Clusteranalyse 538 signifikante Gengruppen mit mehr als 4 Probesets. Diese sind jedoch noch nicht Bonferroni-korrigiert.

In dem darauf folgenden Schritt wurde der Fokus auf die Cluster gelegt, die nach der Promoteranalyse mit PRIMA signifikant überrepräsentierte Transkriptionsfaktor-bindungsstellen aufwiesen bzw. nach der GO-Analyse mit TANGO mit einer Gene Ontology Gruppe signifikant assoziiert waren. Diese Einschränkung hat die Anzahl der Cluster von 538 auf 58 bzw. 41 reduziert.

Für diese biologisch relevanten Gengruppen mit ähnlichen Expressionsverläufen wurden daraufhin die zugehörigen Metagene in den Brustkrebsdaten von 766 Patientinnen berechnet, die anschließend auf ihre prognostische Relevanz untersucht wurden. Die zusammengefassten Ergebnisse sind Abbildung 5.1.2.1 zu entnehmen.

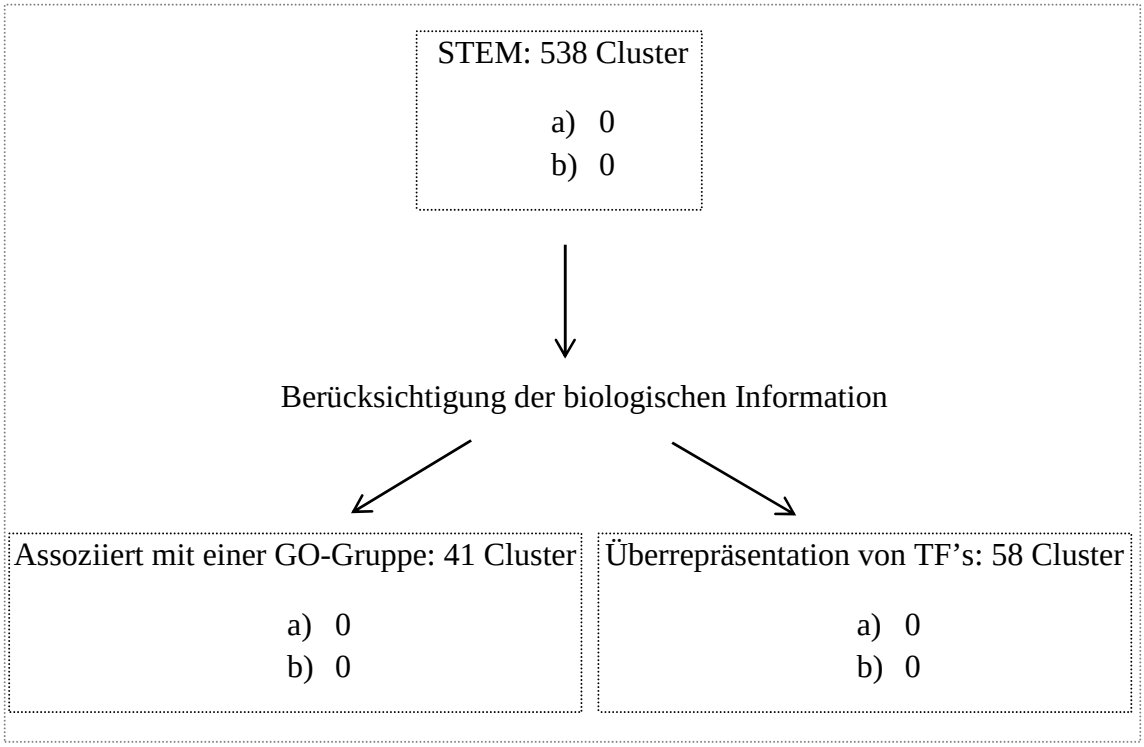

Abb. 5.1.2.1: Ergebnisse der STEM-Analyse: Anzahl signifikanter Cluster je Erfolgsszenario

Daraus ist ersichtlich, dass die Analyse der vorliegenden Daten mit dem STEM-Verfahren keine Cluster liefert, die einem der vorgestellten und erforderlichen Erfolgsszenarien entsprechen.

Zusammenfassend lässt sich sagen, dass mit dieser Methode mehrere Cluster identifiziert wurden, die signifikant mit Prognose assoziiert sind. Einige davon waren nach TANGO bzw. PRIMA zudem noch mit mindestens einer Gene Ontology Gruppe signifikant assoziiert bzw. wiesen signifikant überrepräsentierte Transkriptions-faktorbindungsstellen auf. Die Untersuchung auf deren vordefinierte prognostische Relevanz reduzierte jedoch die Anzahl dieser Cluster auf null. Aus diesem Grund und auch wegen der willkürlichen und einschränkenden Auswahl der Basisprofile bei STEM (s. Diskussion in Unterkapitel 3.2.2) wird die Analyse der Daten mit diesem Verfahren nicht weiter vertieft.

5.1.3 DIB-C

Eine weitere nicht-modellbasierte Methode „difference-based clustering algorithm" wurde zur Klassifikation der Genexpressions-Zeitreihen angewendet. Dazu liegen schon einige eigene Veröffentlichungen vor (Freis et al., 2009; Freis et al., 2012).

Die Implementierung erfolgte in R unter Einsatz von *limma()* mit mehreren Durchläufen und unterschiedlichen Einstellungen für die Signifikanzniveaus des erst- und des zweitrangigen Unterschiedes, so dass für $\alpha^{(1)}$ und $\alpha^{(2)}$ alle möglichen Kombinationen der Werte (0,0001; 0,001; 0,05; 0,01) der Reihe nach verwendet und mit einer zugehörigen Anzahl notiert wurden, z.B. „U1" für $\alpha^{(1)}$ = 0,0001 und $\alpha^{(2)}$ = 0,0001, „U2" für $\alpha^{(1)}$ = 0,0001 und $\alpha^{(2)}$ = 0,001, „U3" für $\alpha^{(1)}$ = 0,0001 und $\alpha^{(2)}$ = 0,05 usw.. Jedem Gen wurde eine 9-stellige Buchstabenfolge zugeordnet. Die ersten 5 Buchstaben beschreiben die Steigung und die letzten 4 die Krümmung der Zeitreihe. Cluster mit weniger als 5 Genen wurden den größeren Klassen zugeordnet, was bis zu 77% aller Cluster betroffen hat (vgl. Krahn, 2008).

Das Verfahren hat nach ca. 30 Minuten Laufzeit im ersten Schritt 692 signifikante Cluster in der Gesamtkohorte ermittelt (vgl. Tab. 5.1.3.1). Nach der Bonferroni-Korrektur hat

sich diese Zahl auf 23 Cluster reduziert, die nach dem Erfolgsszenario a) in der Gesamtkohorte adjustiert und in mindestens zwei einzelnen Kohorten nicht korrigiert signifikant sind. Drei Cluster davon sind auch in der Rotterdam-Kohorte adjustiert signifikant nach dem Erfolgsszenario b).

Die daran anschließende Analyse mit EXPANDER lieferte 21 Gengruppen mit überrepräsentierten Transkriptionsfaktorbindungsstellen und 60 Cluster, die mit einer GO-Gruppe signifikant assoziiert sind. Nach der Korrektur für multiples Testen sind jedoch nur 3 davon nach dem Szenario a) signifikant geblieben. Die zugehörige Tabelle A.1 mit Einzelheiten ist Anhang A zu entnehmen.

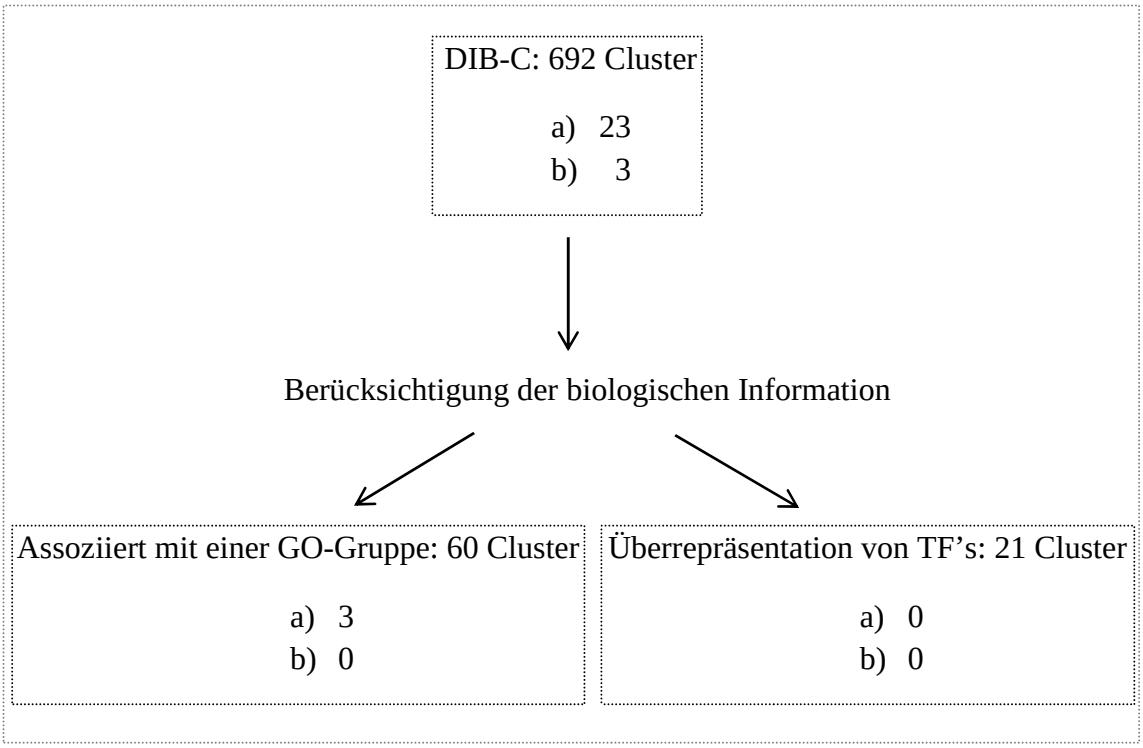

Abb. 5.1.3.1: Ergebnisse der DIB-C-Analyse: Anzahl signifikanter Cluster je Erfolgsszenario

Einige Cluster überschneiden sich sehr stark. So unterscheiden sich Gengruppen „U3.NNNNN,NAVN" und „U7.NNNNN,NAVN" in nur 4 Probesets, die keine Auswirkung auf das zugehörige HR zu haben scheinen (vgl. Tab. A.1). Zwei weitere Cluster „U13.DNNIN,NNNN" und „U14.DNNIN,NNNN" sind bis auf ein Gen gleich mit der gleichen Auswirkung auf die Prognose.

Unter Anwendung dieser Methode wurde ein mutmaßliches Seneszenz-Cluster identifiziert. Die zeitlichen Verläufe zugehöriger Gene sind Abbildung 5.1.3.2 zu

entnehmen. In diesem Cluster ist nach 6 bzw. 12 Stunden nach dem Einschalten der ErbB2-Expression kein signifikanter Unterschied zum Zeitpunkt 0 nachweisbar. Hingegen kommt es nach 24 Stunden zu einer signifikanten Abnahme. Somit ist diese Zeitreihe ähnlich dem Entstehungsverlauf seneszenter Zellen.

Sechs dieser Gene sind mit einer schlechteren Prognose assoziiert. GINS2 und CDCA8 nehmen dabei an der Proliferation teil. RAD51 und ASF1B beteiligen sich an den DNA-Reparaturmechanismen. Zu den Genen LOC146909 und FAM64A liegt z.Zt. noch keine genaue Angabe zu deren Funktionen vor. Das im zweiten Teil des vorgestellten Ansatzes berechnete Metagen dieser Gengruppe ist mit einem HR von 1,71 und dem Bonferroni-korrigierten p-Wert = 0,004 in der Gesamtkohorte signifikant mit einer schlechteren Prognose assoziiert. Aber auch in jeder der drei Einzelkohorten lässt sich dieses Ergebnis verifizieren. Die anschließende Analyse mit PRIMA bzw. TANGO lieferte jedoch keine zusätzliche biologische Relevanz dieses Metagens, da keine GO-Gruppen-Assoziationen bzw. Überrepräsentation von TF's festgestellt werden konnte.

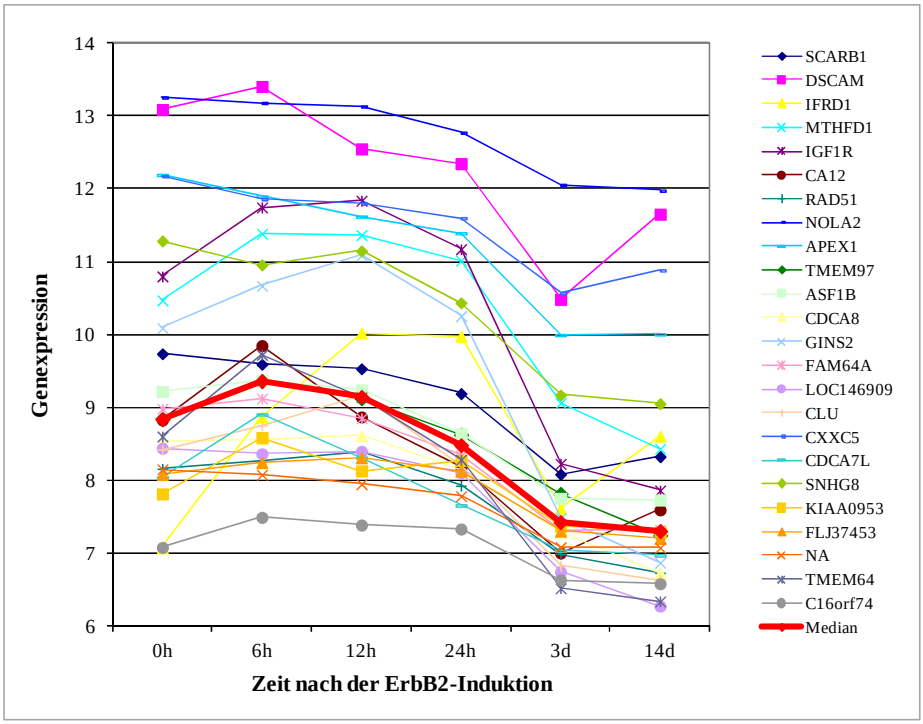

Abb. 5.1.3.2: Zeitlicher Verlauf der Genexpressionen im Seneszenz-Cluster „U12.NNNDN,NNNNV"

Ein weiteres Cluster erschien aus biologischer Sicht sehr interessant, in dem die Bindungsstelle des Transkriptionsfaktors SRF überrepräsentiert ist. Es umfasst sieben Gene, deren Metagen mit einer besseren Prognose bei Brustkrebs assoziiert ist (vgl.

Anhang C.1). Die Genexpressionen steigen dabei gleich nach der ErbB2-Induktion an und fallen 12 Stunden danach signifikant ab. Diese Gengruppe ist signifikant nach einem der vordefinierten Szenarien, jedoch nicht korrigiert für multiples Testen (Einzelheiten dazu s. Anhang C.1).

Abschließend lässt sich sagen, dass mit Hilfe der DIB-C-Methode mehrere Hundert in der Gesamtkohorte signifikanten Cluster identifiziert wurden, von denen viele auch nach den Szenarien a) bzw. b) prognostisch relevant waren. Auffallend viele sind dabei mit einem HR < 0 mit einer besseren Prognose assoziiert. Drei dieser Cluster waren zudem noch nach der EXPANDER-Auswertung mit mindestens einer GO-Gruppe signifikant assoziiert. Dabei überschneiden sie sich sehr stark und könnten in einer weiteren Analyse näher betrachtet werden, wie auch die zwei neuen Metagene, auf die in diesem Kapitel näher eingegangen wurde.

5.1.4 PFP

In diesem Kapitel werden Ergebnisse der Clusteranalyse mit dem PFP-Algorithmus (vgl. Unterkapitel 3.2.4) vorgestellt, um in dieser Arbeit eine weitere nicht-modellbasierte Methode zum Vergleich heranzuziehen.

Die Auswertung der Daten mit diesem Verfahren erfolgte mit dem neu entwickelten PFP-Tool mit unterschiedlichen α - und m - Einstellungen. Dabei wurde α auf 0,2, 2 und 10 und die maximale Anzahl möglicher Profile m auf 35, 50 und 65 gesetzt. Dadurch entstanden 9 mögliche Szenarien, die nach Rücksprache mit den Autoren das gesamte Spektrum gut abdecken. Die zugehörigen Ergebnisse wurden für alle Kombinationsmöglichkeiten zusammenfassend in der Abbildung 5.1.4.1 dargestellt.

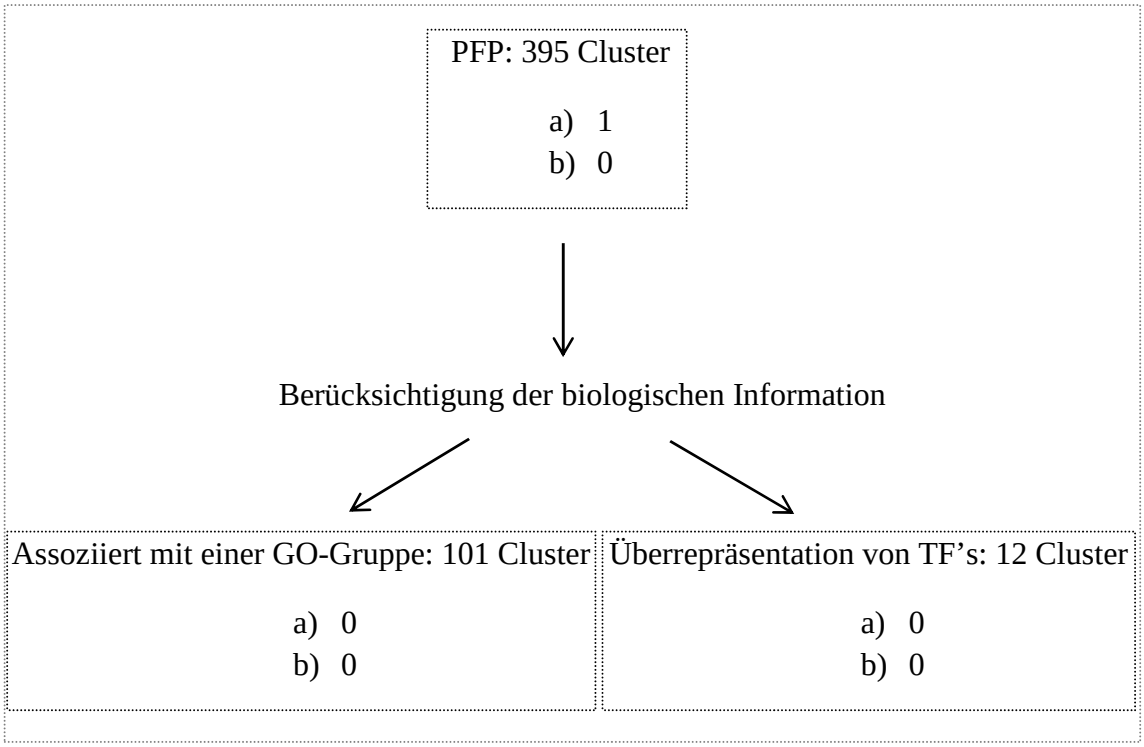

Abb. 5.1.4.1: Ergebnisse der PFP-Analyse: Anzahl signifikanter Cluster je Erfolgsszenario

Die Interpretation dieser Ergebnisse gleicht den vorangegangenen Analysen. So wurden mit PFP 395 (noch nicht adjustierte) signifikante Cluster mit mindestens 5 Genen gefunden. Obwohl 101 davon mit mind. einer GO-Gruppe assoziiert bzw. in 12 Cluster bestimmte Transkriptionsfaktorbindungsstellen überrepräsentiert waren, konnte durch die Überlebenszeitanalyse nur ein Cluster nach dem vordefinierten Erfolgsszenario a) mit $\alpha = 0{,}2$ und $m = 50$ ermittelt werden, der jedoch nach PRIMA bzw. TANGO keine biologische Relevanz aufwies (vgl. Tab. 5.1.4.2). Diese Gengruppe ist in der Gesamtkohorte korrigiert und in den Mainz- und Rotterdam-Kohorten unadjustiert signifikant. Mit einem HR = 0,45 ist die höhere Expression dieses Metagens mit einer längeren metastasenfreien Zeit assoziiert. Diese Ergebnisse sind in der Tabelle 5.1.4.2 zusammengefasst:

Tab. 5.1.4.2: Das einzige nach PFP signifikante Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)		GO	TF
alpha0.2_m50	0,023	0,45	x	-	x	a)	-	-

Dieses Cluster ist mit 78 Probesets sehr unübersichtlich. In den zeitlichen Genexpressionsverläufen (in 5.1.4.3 dargestellt) können zum Teil erhebliche Unterschiede festgestellt werden. So steigt der mediane Verlauf etwa nach 3 Tagen nach der ErbB2-Induktion, bei einigen Genen ist hingegen schon am Anfang der Beobachtungszeit ein Abfall zu erkennen. Dies könnte ein Hinweis auf die prognostische Irrelevanz dieser Gengruppe sein. In dieser Deutung ist jedoch Vorsicht geboten.

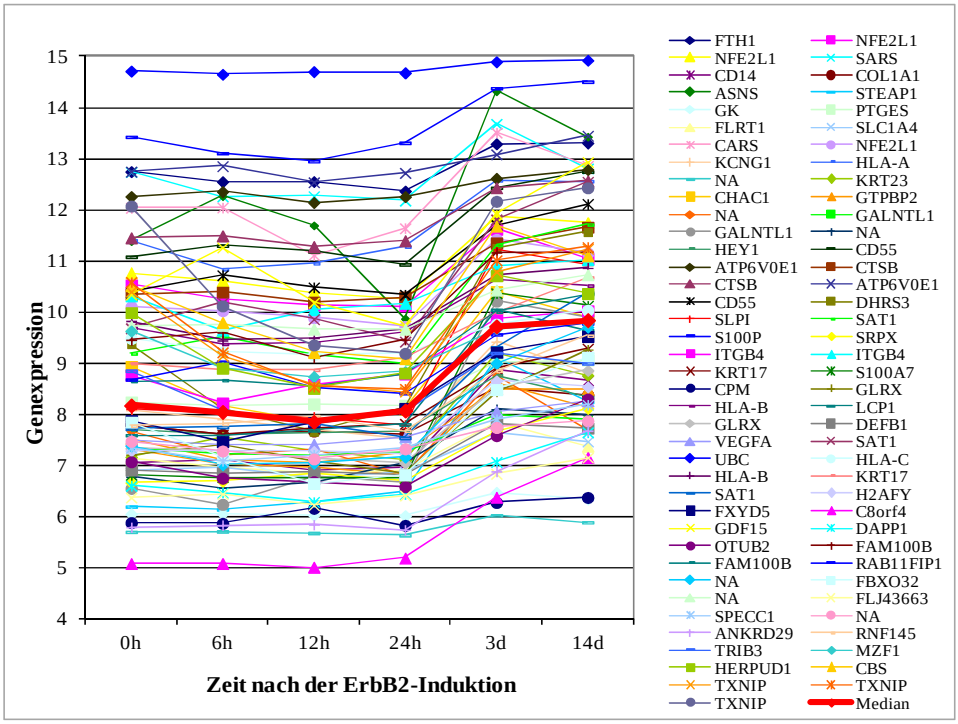

Abb. 5.1.4.3: Zeitlicher Verlauf der Gene im nach PFP signifikanten Cluster mit alpha = 0,2 und m = 50

Im Anhang C.2 sind zusätzliche Ergebnisse der Clusteranalyse mit dem PFP-Algorithmus zu finden. Hierbei werden die Daten noch vor der Auswertung den drei GO-Gruppen (vgl. Unterkapitel 2.1) zugeordnet und liefern dadurch deutlich homogenere Ergebnisse in Bezug auf die Expressionsverläufe.

Mit einer durchschnittlichen Laufzeit von weniger als 1 Minute pro Durchlauf/Einstellung und einer relativ unkomplizierten Umsetzung ist das PFP-Tool sehr einfach in der Anwendung. Obwohl hiermit zunächst sehr viele Cluster mit mindestens 5 Genen identifiziert wurden, ergab die Überlebenszeitanalyse nur eine nach dem Erfolgsszenario a) signifikante Gengruppe, die jedoch sehr unübersichtlich ist, unterschiedliche Zeitverläufe und keine biologische Relevanz aufweist.

5.2 Modellbasierte Verfahren

Dieses Unterkapitel beschreibt Ergebnisse des in dieser Arbeit vorgestellten Ansatzes zur Analyse kurzer Genexpressions-Zeitreihen unter Einsatz modellbasierter Verfahren im ersten Analyseschritt. Es werden sowohl die eigentlichen Ergebnisse der Clusteranalyse als auch die Ergebnisse der Analyse der daraus resultierenden Gengruppen auf deren biologische Relevanz vorgestellt, die mit Hilfe des Programms EXPANDER erfolgte. Ein großer Vorteil dieses Programms ist seine schnelle und unkomplizierte Anwendung. Mit einer Laufzeit von bis zu 40 Minuten pro Durchlauf ist es jedoch sehr zeitaufwändig.

5.2.1 Finite mixture models

Die endlichen Mischungsmodelle wurden an die vorliegenden Daten wie im Unterkapitel 3.3.1 beschrieben angepasst. Dabei wurde wegen der Input-Anforderungen vom R-Paket *BayesMix()* eine Dimensionsreduktion der Daten mit Hilfe der Hauptkomponentenanalyse (PCA) vorgenommen. Nach McLachlan und Peel (2000) sind zwei voneinander gut getrennte Cluster schon durch die Projektion ihrer wenigen Hauptkomponenten ohne Verlust des Informationsgehaltes optimal vertreten. Wie in Abbildung 5.2.1.1 dargestellt, werden die vorliegenden Daten durch ihre erste Komponente schon recht gut beschrieben. Diese wird bei der Clusteranalyse mit *BayesMix()* berücksichtigt, dessen Anwendung eine durchschnittliche Laufzeit von 10 Minuten erfordert. Die zu analysierende Matrix für die Daten ohne Berücksichtigung der biologischen Information enthält somit Expressionswerte von 2632 Genen und der ersten Hauptkomponente.

Werden die Daten noch vor der Clusterung den drei GO-Gruppen (vgl. Unterkapitel 2.1) zugeordnet, so enthalten die entsprechenden Matrizen Expressionswerte in der CC-Gruppe von 1667, in der BP-Gruppe von 1562 und in der MF-Gruppe von 1560 Genen (vgl. Unterkapitel 2.2). Die Anzahl der Spalten ist gemäß der Hauptkomponentenanzahl eins. In Abbildung 5.2.1.3 sind die entsprechenden Ergebnisse im Abzweig „GO-Gruppen” notiert.

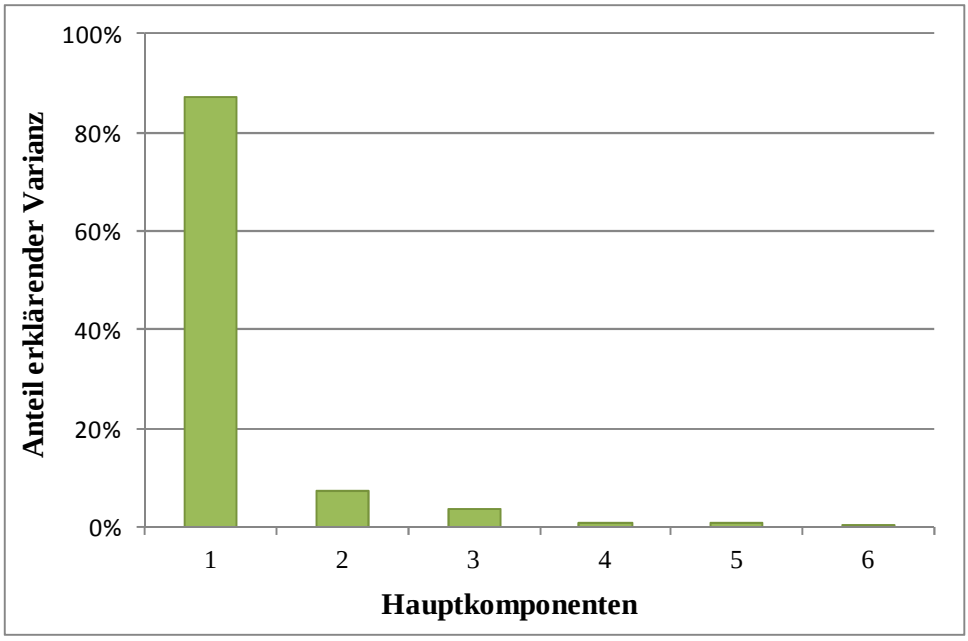

Abb. 5.2.1.1: Darstellung der Anteile erklärender Varianz bei der durchgeführten Hauptkomponentenanalyse zur Dimensionsreduktion

Bei einer anderen Herangehensweise an die Berücksichtigung des biologischen Vorwissens werden die Daten nicht nur den drei GO-Gruppen zugeordnet. Eine Ähnlichkeitsmatrix SimGO gemäß 4.2 geht zusätzlich mit (1-SimGO) in die Distanzmetrik ein. In der Abbildung 5.2.1.3 und in der Tabelle 5.2.1.4 sind die zugehörigen Ergebnisse mit „SimGO" gekennzeichnet. Die Anzahl der Gene, zu denen zum Zeitpunkt der Analyse eine Entrez ID vorlag, ist niedriger und liegt bei 1632, 1528 und 1525 für die CC-, BP- bzw. die MF-Gruppe (vgl. Diskussion in Unterkapitel 2.2). Da bei der Berechnung des GO-Ähnlichkeitsmaßes nach 4.2 Mediane gebildet werden, erfolgt an dieser Stelle ein Verlust an Informationen, die u.U. relevant sein könnten, so dass bei der Ergebnisinterpretation Vorsicht geboten ist.

Für die Wahl der Komponentenanzahl, die bei den endlichen Mischungsmodellen (FMM) vorgegeben werden muss, sind in Tabellen A.6 - A.8 BIC-Werte zu den einzelnen *K*'s aufgeführt. Daraus ist ersichtlich, dass für die Modellanpassung an die vollständigen Daten ohne GO-Betrachtung die Komponentenanzahl $K = 3$ den minimalen BIC-Wert liefert und somit in *BayesMix()* vorgegeben wird. Für die nach der GO-Information eingeschränkten Datensätze ist die Anzahl der Komponenten für die CC- und die MF-Gruppen gleich 2 und für die BP-Gruppe gleich 5.

Tab. 5.2.1.2: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (ohne Berücksichtigung der GO-Zuordnung)

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	22732,06	7	22964,32	12	22989,08
3	**20658,70**	8	23182,47	13	22770,56
4	23128,66	9	23014,93	14	23126,53
5	22851,66	10	22889,14	15	23186,77
6	22705,64	11	22768,53		

Die Parameter der unabhängigen, konjugierten a priori-Verteilungen wurden in *BayesMix()* durch die Voreinstellung „priorsRaftery" nach Raftery (1996) definiert und der burn-in auf 1000 gesetzt.

Die Anzahl dadurch ermittelter Cluster ist in Abbildung 5.2.1.3 zusammengefasst. So wurden mit den endlichen Mischungsmodellen 6 signifikante Cluster (4 mit der Optimierungsmethode MAXPEAR und 2 mit MINBINDER) identifiziert, die jedoch noch nicht adjustiert nach der Anzahl der Tests sind und keinem der beiden vordefinierten Erfolgsszenarien entsprechen. Diese Zahlen sprechen dafür, dass diese Methode nicht optimal für diese Art der Daten zu sein scheint. Mit 4 bzw. 2 Clustern bei 2632 Genen werden viel zu große Gengruppen gebildet, so dass hier anscheinend nicht sinnvoll getrennt wird.

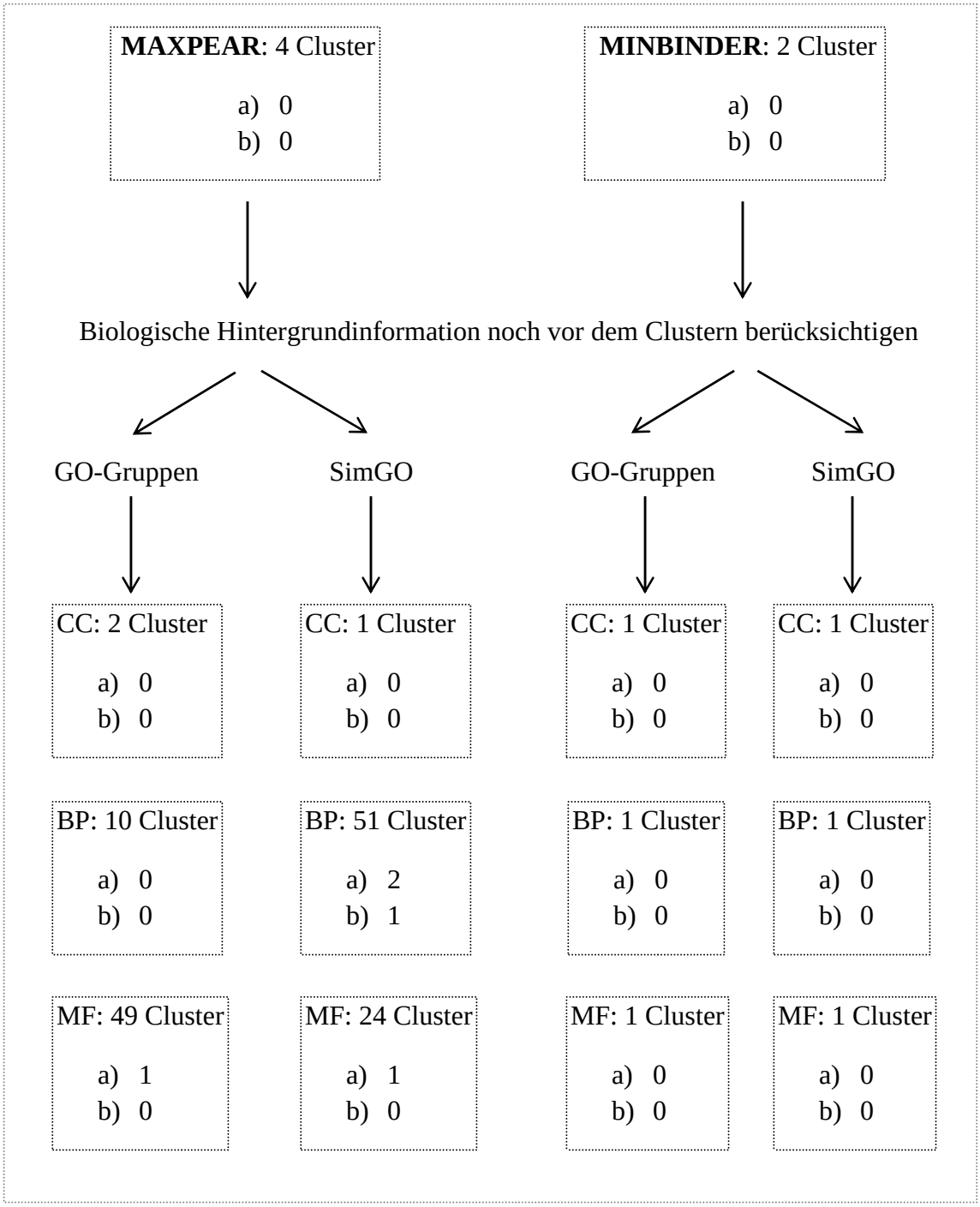

Abb. 5.2.1.3: Ergebnisse der finite mixture models-Analyse: Anzahl signifikanter Cluster je Optimierungsmethode, GO-Gruppe und Erfolgsszenario

Die Berücksichtigung der biologischen Hintergrundinformation ist dagegen mit einem größeren Erfolg verbunden. So werden in allen Gruppen außer in CC deutlich mehr Cluster ermittelt. Die Ergebnisse der Überlebenszeitanalyse sind Tabelle 5.2.1.4 zu entnehmen. Die zusätzliche Analyse mit EXPANDER zeigte keine weitere biologische Relevanz.

Tab. 5.2.1.4: Die nach der finite mixture models-Analyse signifikanten Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)	
MF_Pear_166	< 0,001	1,64	x	x	x	a)
MF_SimGO_Pear_7	0,001	1,64	x	x	x	a)
BP_SimGO_Pear_9	< 0,001	1,96	x (x)	x	x (x)	a) b)
BP_SimGO_Pear_63	< 0,001	1,94	-	x	x	a)

Alle vier Gengruppen sind nach dem Erfolgsszenario a) mit einem adjustierten p-Wert $\leq$ 0,001 und nach der Optimierungsmethode MAXPEAR signifikant. Das Cluster „BP_SimGO_Pear_9" ist zusätzlich noch nach dem zweiten Szenario b) signifikant und wurde (wie der Bezeichnung entnommen werden kann) in der BP-Gruppe unter Berücksichtigung des GO-Ähnlichkeitsmaßes in der Distanzmetrik identifiziert. Das Cluster ist mit 37 Genen jedoch unübersichtlich und vereint einige signifikante Gengruppen aus Auswertung in 5.2.2 (vgl. zugehörige Tabellen im Anhang B).

Auffällig ist hier, dass in allen Clustern das HR deutlich größer 1 ist, so dass die höhere Expression dieser Metagene mit einer kürzeren metastasenfreien Zeit assoziiert ist. Die zugehörigen Graphiken sind im Anhang A zu finden. Dabei ist zu bemerken, dass die zeitlichen Verläufe in diesen Clustern Unterschiede aufweisen, falls das Ähnlichkeitsmaß nach Schlicker et al. (2006) gebildet wird.

5.2.2 Dirichlet Process mixture models

Für die Auswertung in diesem Abschnitt wurde ähnlich 5.2.1 eine Dimensionsreduktion der Daten mit Hilfe der Hauptkomponentenanalyse vorgenommen. Die Datenmatrizen umfassen dementsprechend 2632 bzw. 1667, 1562 oder 1560 Zeilen (vgl. 5.2.1) und zwei Spalten, die für die ersten zwei Hauptkomponenten stehen. Alle Programm- und Modellvoreinstellungen wurden in Anlehnung an Fritsch (2010) vorgenommen, wie z.B.

die a priori-Parameter nach Bensmail et al. (1997) oder auch a_0 = 4 und b_0 = 2 für den Präzisionsparameter α. Mit einer Laufzeit von ca. 1 Stunde pro Durchlauf ist die Datenanalyse mit dieser Methode unter Einsatz von *DPpackage()* sehr zeitaufwändig. Die zugehörigen Ergebnisse sind Abbildung 5.2.2.1 zu entnehmen.

So wurden hier noch ohne Berücksichtigung der GO-Information drei Cluster ermittelt, die nach dem Erfolgsszenario a) signifikant waren – eine MAXPEAR- und zwei MINBINDER-Gengruppen. Die zusätzliche Analyse mit TANGO zeigte eine Assoziation mit vier GO-Gruppen von einem dieser Cluster, dessen Abbildung im Anhang B.4 zu finden ist. Hier ist das Cluster „Bind_10" in der Gesamtkohorte adjustiert und in den Transbig- und Rotterdam-Kohorten unadjustiert signifikant. Mit einem HR von 1,69 ist die höhere Expression des zugehörigen Metagens mit einer kürzeren metastasenfreien Zeit und damit mit einer schlechteren Prognose assoziiert.

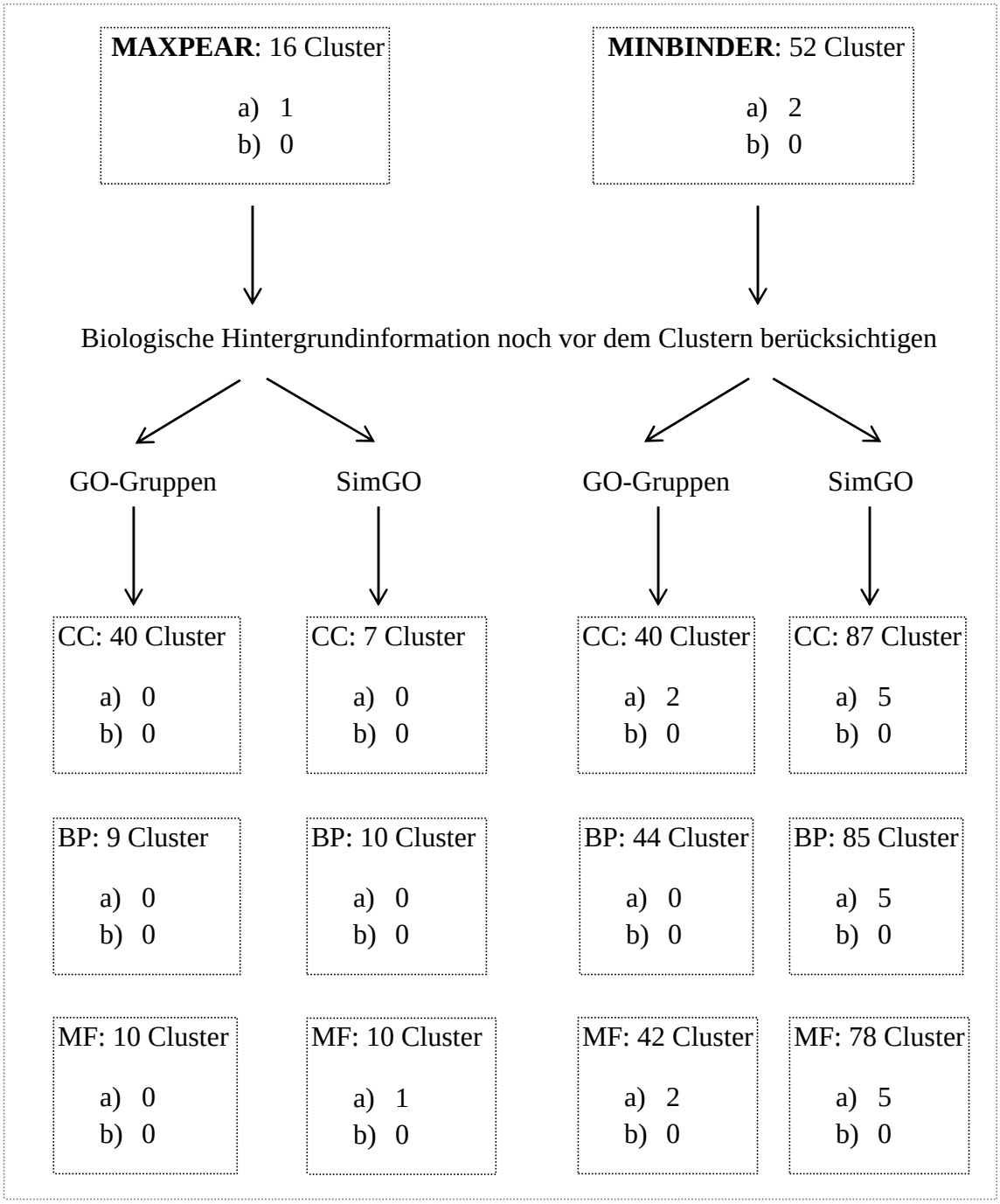

Abb. 5.2.2.1: Ergebnisse der DP mixture models-Analyse: Anzahl signifikanter Cluster je Optimierungsmethode, GO-Gruppe und Erfolgsszenario

Bei der Betrachtung weiterer Ergebnisse unter Berücksichtigung des biologischen Vorwissens, deren Einzelheiten in Tabelle 5.2.2.2 dargestellt sind, fällt auf, dass durch die Minimierung von Binders Verlustfunktion deutlich mehr nach den beiden vordefinierten Erfolgsszenarien signifikante Cluster gefunden werden, als durch die Maximierung des Rand-Index.

Tab. 5.2.2.2: Die nach der DP mixture models-Analyse signifikanten Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)	
Pear_41	0,005	1,34	x	-	x	a)
Bind_10	< 0,001	1,69	-	x	x	a)
Bind_502	< 0,001	1,63	x	x	x	a)
CC_Bind_95	< 0,001	0,28	-	x	x	a)
CC_Bind_317	< 0,001	1,94	x (x)	x	-	a)
MF_Bind_308	< 0,001	2,54	x	x	x (x)	a)
MF_Bind_354	< 0,001	1,64	x	-	x	a)
BP_SimGO_Bind_44	< 0,001	1,65	x (x)	x	x	a)
BP_SimGO_Bind_79	0,001	1,64	x (x)	-	x	a)
BP_SimGO_Bind_111	0,010	1,95	x	x	-	a)
BP_SimGO_Bind_146	< 0,001	1,73	-	x	x	a)
BP_SimGO_Bind_262	< 0,001	1,47	x (x)	x	-	a)
CC_SimGO_Bind_43	< 0,001	1,64	x	x	x	a)
CC_SimGO_Bind_79	0,001	1,65	x (x)	-	x	a)
CC_SimGO_Bind_112	0,011	1,95	x	x	-	a)
CC_SimGO_Bind_144	< 0,001	1,73	-	x	x	a)
CC_SimGO_Bind_259	< 0,001	1,47	x (x)	x	-	a)
MF_SimGO_Bind_37	0,003	1,48	x	x	x	a)

MF_SimGO_Bind_75	0,003	1,42	x	x (x)	-	a)
MF_SimGO_Bind_94	0,006	1,41	x (x)	x	-	a)
MF_SimGO_Bind_106	0,009	1,95	x	x	-	a)
MF_SimGO_Bind_138	< 0,001	1,73	-	x	x	a)
MF_SimGO_Pear_668	0,004	0,22	x	-	x	a)

Einige dieser Cluster überschneiden sich. So ist z.B. das Cluster „MF_SimGO_Bind_37" mit einigen wenigen Ausnahmen im Cluster „BP_SimGO_Bind_44" enthalten, in dem zudem auch das Cluster „CC_SimGO_Bind_43" zu finden ist. „MF_SimGO_Bind_94" ist zum größten Teil im Cluster „BP_SimGO_Bind_262" zu finden.

Mehrere dieser signifikanten Gengruppen unterscheiden sich gar nicht und sind komplett identisch:

- „CC_SimGO_Bind_79" mit „BP_SimGO_Bind_79",
- „CC_SimGO_Bind_112" mit „BP_SimGO_Bind_111" und „MF_SimGO_Bind_106",
- „CC_SimGO_Bind_259" mit „BP_SimGO_Bind_262",
- „CC_SimGO_Bind_144" mit „BP_SimGO_Bind_146" und „MF_SimGO_Bind_138".

Auf die biologisch interpretierbaren Motive, die sie zeigen, wird an dieser Stelle nicht näher eingegangen. Es ist jedoch zu erwähnen, dass zumindest die letzten drei Cluster eine wichtige Bedeutung bei der Erforschung der Brustkrebsentstehung haben. Sie sind auch im Cluster „BP_SimGO_Pear_63" (Unterkapitel 5.2.1) und in „MF_SimGO_Bind_111" mit „BP_SimGO_Bind_119" (Unterkapitel 5.2.3) wieder zu finden. Viele der darin enthaltenen Gene kodieren für die verschiedenen Isoformen des Proteins Tubulin und beteiligen sich an mehreren wichtigen Funktionen wie z.B. an der Zellteilung. Einige davon wurden schon vereinzelt untersucht und anhand der vorliegenden Brustkrebsdaten für prognostisch relevant befunden (vgl. Martin et al., 2012).

In einem dieser Cluster, auf das hier repräsentativ für die anderen näher eingegangen wird, steigt die mittlere Genexpression gleich nach der ErbB2-Induktion und fällt nach etwa 24 Stunden ab, wie Abbildung B.5 im Anhang entnommen werden kann. Da einige wenigen Gene ein anderes Verlaufsmuster als die Mehrheit aufweisen, ist eine Optimierung der Distanzmetrik vorstellbar.

Diese Gengruppen sind in allen Kohorten außer in der Mainzer Kohorte signifikant. Das HR ist dabei deutlich über 1, so dass die höhere Metagenexpression dieser Cluster mit einer kürzeren metastasenfreien Zeit assoziiert ist.

Weiterhin ist zu bemerken, dass nicht nur die Tubulin-Cluster, sondern (mit Ausnahme von zwei Gengruppen) alle Metagene hier mit einer schlechteren Prognose assoziiert sind, da die entsprechenden HR's größer 1 sind.

Zusammenfassend für dieses Unterkapitel lässt sich sagen, dass die Anpassung der Dirichlet Prozess Mischungsmodelle an die vorliegenden Daten der erfolgreichste Ansatz dieser Arbeit ist, der die meisten Ergebnisse in Bezug auf die vordefinierten Szenarien liefert. Auch die daraus resultierenden Gengruppen sind deutlich kleiner und homogener als bei den anderen Verfahren. Durch die MINBINDER-Optimierung der Clusterungen werden im Vergleich zu MAXPEAR deutlich mehr signifikante Ergebnisse geliefert, deren Mehrheit ein HR > 1 aufweist. Wie auch bei den vorangegangenen Analysen, werden bei der Berücksichtigung der biologischen Hintergrundinformation deutlich mehr interessante Cluster identifiziert, vor allem bei der Einbindung dieser in die Distanzmetrik.

Auch aus biologischer Sicht ist dieses Verfahren sehr erfolgreich. Einige Cluster enthalten Gene, die schon als prognostisch relevant in den Brustkrebsdaten bekannt sind, jedoch nicht in der Form eines Metagens. Ferner zeigen auch andere hierbei ermittelten Gengruppen biologisch interpretierbare Motive, die in den weiteren Untersuchungen näher betrachtet werden könnten.

5.2.3 DIRECT

Die vorliegenden Daten wurden ebenfalls mit der Vergleichsmethode DIRECT ausgewertet. Das in R implementierte Tool benötigt keine mediane Zeitreihe, da die

Autoren durch die Voreinstellung „OU" im Programm eine Mittelwertbildung nach Ornstein-Uhlenbeck (vgl. Merton, 1971) implementiert haben, die in Anlehnung an die Ausarbeitungen von Fu et al. (2011, 2013) wie auch alle weiteren Parametereinstellungen übernommen wurde. Da dieses Tool keine Anwendung der PCA bedarf, enthalten die zu analysierenden Datenmatrizen in Abhängigkeit von der berücksichtigten biologischen Information die Zeilen analog 5.2.1 und 18 Spalten (3 Zelllinien und 6 Zeitpunkte).

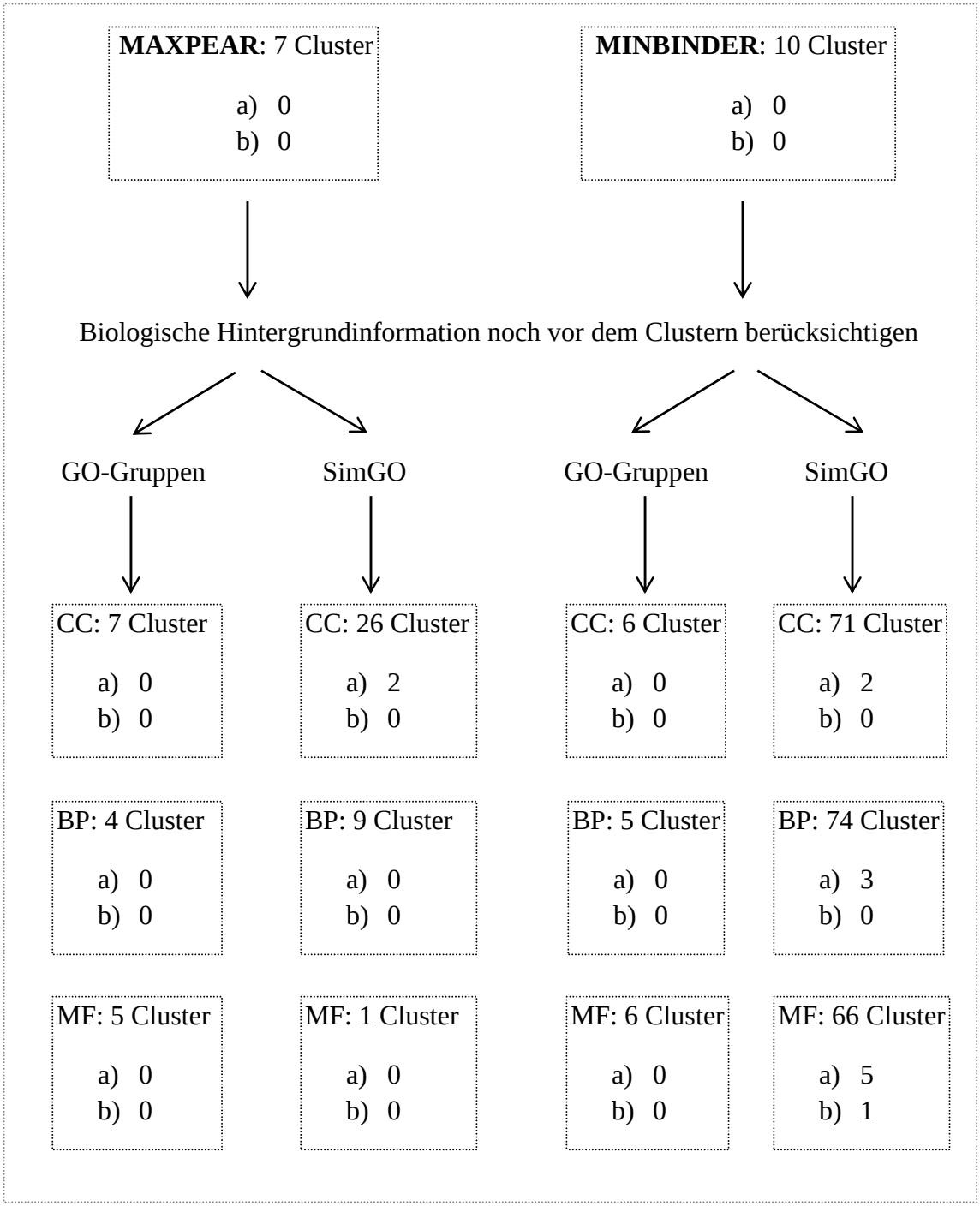

Abb. 5.2.3.1: Ergebnisse der DIRECT-Analyse: Anzahl signifikanter Cluster je Optimierungsmethode, GO-Gruppe und Erfolgsszenario

Bei der Betrachtung der Ergebnisse in Abbildung 5.2.3.1 fällt auf, dass hier im Vergleich zu MAXPEAR mit Hilfe der Optimierungsmethode MINBINDER deutlich mehr signifikanter Cluster gefunden werden, insbesondere bei der Bildung des Distanzmaßes gemäß Unterkapitel 4.2.

Die nach den vordefinierten Erfolgsszenarien a) bzw. b) signifikanten Cluster wurden mit allen zugehörigen Informationen in Tabelle 5.2.3.2 zusammengefasst. Dieser ist zu entnehmen, dass mit Ausnahme von zwei Clustern alle mit einer schlechteren Prognose assoziiert sind, da die zugehörigen HR's über 1 liegen. Ihre Auswertung mit EXPANDER lieferte jedoch keine weitere biologische Relevanz.

Tab. 5.2.3.2: Die nach der DIRECT-Analyse signifikanten Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)	
MF_SimGO_Bind_29	0,001	1,63	x (x)	x	x	a)
MF_SimGO_Bind_36	0,005	1,79	x	x	-	a)
MF_SimGO_Bind_94	0,002	1,73	x	-	x	a)
MF_SimGO_Bind_111	< 0,001	1,73	-	x	x	a)
MF_SimGO_Bind_134	< 0,001	0,57	x (x)	-	x (x)	a) b)
CC_SimGO_Pear_8	0,001	2,15	x (x)	x	x	a)
CC_SimGO_Pear_22	0,005	0,38	x	-	x (x)	a)
CC_SimGO_Bind_34	< 0,001	1,63	x	x	x	a)
CC_SimGO_Bind_118	< 0,001	1,74	-	x (x)	x	a)
BP_SimGO_Bind_9	0,021	1,60	x (x)	-	x	a)
BP_SimGO_Bind_34	< 0,001	1,63	x	x	x	a)
BP_SimGO_Bind_119	< 0,001	1,74	-	x (x)	x	a)

Ein Cluster – „MF_SimGO_Bind_134" – ist sogar nach beiden Szenarien signifikant. Diese Gengruppe wurde im MF-Datensatz mittels Optimierungsmethode MINBINDER und unter Berücksichtigung des GO-Ähnlichkeitsmaßes in der Distanzmetrik identifiziert. Sie ist in allen Kohorten außer der Transbig-Kohorte adjustiert signifikant und mit einem HR = 0,57 mit einer besseren Prognose assoziiert. Die entsprechenden Expressionsverläufe können Abbildung B.3 im Anhang entnommen werden. Daraus ist ersichtlich, dass die zeitlichen Verläufe Unterschiede aufweisen. Das kann auch bei einigen anderen Clustern beobachtet werden, so dass die hier gewählte Art der Distanzmetrikbildung nicht optimal zu sein scheint.

Mehrere Cluster finden sich in der Auswertung 5.2.2 wieder. So überschneiden sich die Cluster „MF_SimGO_Bind_29" und „BP_SimGO_Bind_34" stark mit „MF_SimGO_Bind_37" bzw. „MF_SimGO_Bind_94" bei den DP Mischungsmodellen. Die Gene in „MF_SimGO_Bind_111" und „BP_SimGO_Bind_119" (hier mit Ausnahme von einem Gen) sind sogar dieselben wie im Cluster „MF_SimGO_Bind_138" in Unterkapitel 5.2.1. Diese könnten bei einer weiteren Untersuchung näher betrachtet werden.

Somit führt die Auswertung der Daten mit dem in R implementierten Tool DIRECT zu ähnlichen Ergebnissen wie die DP Mischungsmodelle, was auch plausibel zu sein scheint, da diese Methode ebenfalls auf dem Einsatz von Dirichlet Prozess Prior beruht. Die Anwendung dieser Methode erfordert bei einer durchschnittlichen Laufzeit von ca. 15 Minuten pro Durchlauf zwar deutlich weniger Zeit, liefert aber auch weniger prognostisch relevante Cluster nach den vordefinierten Szenarien. Auch hier ist die Optimierung der Clusterungen mit MINBINDER erfolgreicher im Vergleich zu MAXPEAR, insbesondere bei der Berücksichtigung des biologischen Vorwissens in der Distanzmetrik.

Da sich einige dieser Gengruppen in den Ergebnissen der vorangegangenen Analysen wieder finden (vgl. Unterkapitel 5.2.1 und 5.2.2), liegt es nahe, dass auch hier biologisch interpretierbare Motive vorliegen, die in den weiteren Untersuchungen näher betrachtet werden könnten.

6 Zusammenfassung und Ausblick

Durch die weltweit steigende Anzahl der Krebserkrankungen werden vermehrt Microarray Experimente durchgeführt, um durch die Analyse der veränderten Genexpressionen die möglichen Rückschlüsse auf die biologischen Hintergründe der Entstehung dieser Krankheit ziehen zu können und dadurch bessere und gezieltere Therapien zu ermöglichen.

Ein möglicher Ansatz dazu ist die Untersuchung der einzelnen Gene auf deren prognostische Relevanz in den Krebsdaten. Dieses wird schon erfolgreich umgesetzt und die Ergebnisse werden in zahlreichen wissenschaftlichen Ausarbeitungen veröffentlicht.

In der vorliegenden Arbeit wurde eine neuartige Vorgehensweise an diese Problematik vorgestellt und angewendet. Hierbei wurden mit mehreren Clusteranalyseverfahren nicht die einzelnen Gene, sondern Gengruppen mit ähnlichen Expressionsverläufen identifiziert. Dazu wurde ein bekannter Datensatz mit drei humanen Mammakarzinom-Zelllinien MCF7 herangezogen, in die die onkogene Variante von ErbB2 induziert und zu sechs Zeitpunkten nach der Induktion beobachtet wurde. Der Hintergrund dabei war, dass die Überexpression dieses Onkogens mit einer schlechteren Prognose assoziiert ist.

Die Clustermethoden haben dabei zwei methodische Blöcke gebildet: Zum einen wurden die Daten mit den nicht-modellbasierten Verfahren DIB-C und STEM analysiert. Zum anderen erfolgte eine Analyse mit den modellbasierten finiten und infiniten Mischungsmodellen, die in R mit Hilfe von Paketen *BayesMix()* für Mischungen von univariaten Normalverteilungen bzw. *DPpackage()* für die auf dem Dirichlet Prozess basierten Modellen implementiert wurden. Als Vergleichsmethoden zu den nicht-modellbasierten Algorithmen wurden das bekannte *k*-means und das neu entwickelte PFP-Verfahren herangezogen. Als nicht-modellbasierter Vergleich diente das in R implementierte Tool DIRECT, das auf einem Random-Effects-Mischungsmodell basiert und dem ein Dirichlet Prozess Prior zu Grunde liegt.

Für die ursprüngliche mit diesen Clustermethoden zu analysierende MCF7-Datenmatrix mit 54675 Probesets und mit je 18 log_2-transformierten Genexpressionswerten, die für jeden der drei Triplikate zu jeweils 6 Zeitpunkten stehen, wurde für diese Arbeit eine Vorauswahl der Gene getroffen. Diese erfolgte mit dem R-Paket *limma()* unter Kontrolle der False Discovery Rate und adjustiert nach Benjamini/Hochberg, so dass die zu analysierende Teilmatrix Expressionswerte zu den 2632 zu mindestens einem der sechs Zeitpunkte differentiell exprimierten Gene enthielt. Die Zuordnung der Probesets zu deren GO-Gruppen mit den R-Paketen *topGO()* bzw. *GOSim()* endete in den weiteren Teilmatrizen mit der entsprechenden Zeilenanzahlen für die CC-, BP- bzw. MF-Gruppen: 1667/1206, 1562/1227 und 1560/1157. Je nach Clusteransatz wurden entweder alle 18 Genexpressionswerte berücksichtigt (DIB-C, DIRECT), Mediane gebildet (STEM, *k*-means, PFP) oder es erfolgte eine Dimensionsreduktion mittels Hauptkomponentenanalyse (finite und infinite Mischungsmodelle).

Zusätzlich erfolgte sowohl bei den finiten und infiniten Mischungsmodellen als auch bei der DIRECT-Methode eine sinnvolle Zusammenfassung von MCMC-Samples von Clusterungen durch die Berechnung der Posterior Similarity Matrix. Aus dieser wurden wiederrum optimale Clusterungen durch die Minimierung von Binders Verlustfunktion (MINBINDER) bzw. durch die Maximierung des adjustierten Rand-Index von Hubert und Arabie (MAXPEAR) ermittelt.

Die Berücksichtigung des biologischen Vorwissens erfolgte mittels EXPANDER und den entsprechenden Tools TANGO und PRIMA für die GO- bzw. Promoteranalyse, die nach der erfolgten Clusteranalyse eingesetzt wurden. In einer anderen Kombination der

Verfahren wurden alle Gene noch vor der Clusteranalyse den drei Gene Ontology Gruppen zugeordnet.

Die durch dieses Vorgehen identifizierten Cluster sind die co-regulierten Gengruppen, die an dieser Stelle vorerst „nur" potentiell relevant in der Brustkrebsforschung zu sein scheinen. Für sie wurden Metagene gebildet, die die mittlere Expression dieser Cluster widerspiegeln und anschließend mit Hilfe der Cox-Modelle auf ihre prognostische Relevanz in den Brustkrebsdaten sowohl in der Gesamtkohorte von 766 Patientinnen als auch in den einzelnen Kohorten untersucht wurden. Dabei betrug die Anzahl der an Brustkrebs erkrankten Frauen in der Mainzer Kohorte 200, in der Transbig-Kohorte 280 und in der Rotterdam-Kohorte 286 Patientinnen.

Durch diese neuartige Vorgehensweise wurden neue biologisch relevanten Cluster aufgezeigt. Die prognostische Relevanz der Metagene wurde durch zwei Erfolgsszenarien herausgestellt, die zum einen durch die Diskussion mit Biologen und Medizinern und zum anderen durch die statistische Analyse der Daten folgendermaßen definiert wurden:

Szenario a) Das Metagen ist in der Gesamtkohorte adjustiert signifikant und in mindestens zwei einzelnen Kohorten mindestens nicht adjustiert signifikant.

Szenario b) Das Metagen ist in der Gesamtkohorte und in der Rotterdam-Kohorte adjustiert signifikant und in mindestens einer weiteren Einzelkohorte signifikant, adjustiert nach der Anzahl der in der Rotterdam-Kohorte signifikanten Cluster.

Zusammenfassend ist für die nicht-modellbasierten Methoden in Bezug auf diese Erfolgsszenarien Folgendes festzuhalten:

Bei der Analyse der vorliegenden Daten mit der Methode *k*-means wurden für die Entscheidung bzgl. optimaler Clusteranzahl die Fehlerquadratsumme und die durchschnittlichen Silhouetten-Werte betrachtet. Ergebnisse beider Kriterien konnten gegenseitig nicht verifiziert werden, so dass für eine tiefergehende Analyse der Daten auf dieses Verfahren verzichtet wurde.

Die Analyse mit STEM lieferte mehrere Cluster, die sowohl in den vorliegenden Daten signifikant mit der Prognose assoziiert als auch aus biologischer Sicht interessant waren. Die Untersuchung auf deren prognostische Relevanz nach den vordefinierten Szenarien reduzierte die Anzahl dieser Cluster jedoch auf null. Da auch von der methodischen Seite

aufgrund der willkürlichen Auswahl der Basisprofile ein Optimierungspotential besteht (vgl. Diskussion in 3.2.4), wurde die Analyse der Daten mit diesem Verfahren nicht weiter vertieft.

DIB-C ist die erfolgreichste Methode aus den hier angewandten nicht-modellbasierten Verfahren. Sie lieferte mehrere signifikante und nach den Szenarien a) bzw. b) prognostisch relevante Cluster. Auffallend viele von ihnen waren mit einer besseren Prognose assoziiert. Drei Cluster sind nach der EXPANDER-Auswertung mit mindestens einer GO-Gruppe signifikant assoziiert und somit auch biologisch relevant. Diese Gencluster überschneiden sich sehr stark und könnten in einer weiteren Analyse näher betrachtet werden. Auch die beiden neuen, prognostisch relevanten Cluster, die im Ergebniskapitel vorgestellt wurden, sollten in einer an diese Arbeit anschließenden Analyse näher untersucht werden.

Mit PFP konnten zunächst fast genauso viele nicht adjustierte signifikante Cluster mit mindestens 5 Genen ermittelt werden, wie mit der DIB-C-Methode. Die Überlebenszeitanalyse ergab jedoch nur ein prognostisch relevantes Metagen, deren höhere Expression mit einer längeren metastasenfreien Zeit assoziiert ist. Dieses Cluster ist mit 78 Genen sehr unübersichtlich und weist unterschiedliche Zeitverläufe auf. Da auch nach PRIMA bzw. TANGO keine biologische Relevanz ermittelt werden konnte, ist dieses Metagen nicht weiter interessant. Eine zusätzliche Analyse, bei der die biologische Information noch vor dem tatsächlichen Clustern berücksichtigt wird, lieferte in Bezug auf die vordefinierten Szenarien deutlich kleinere und homogenere Gengruppen, die in einer weiteren Analyse vertieft untersucht werden könnten.

In den Tabellen 6.1 und 6.2 erfolgt eine Übersicht über die nicht-modellbasierten Methoden aus statistischer bzw. biologischer Sicht sowie die Empfehlung des für die Ermittlung prognostisch relevanter Gengruppen nach dem hier vorgestellten Ansatz geeigneten Verfahrens sowohl ohne (Tab. 6.1) als auch unter Berücksichtigung des biologischen Vorwissens (Tab. 6.2):

Tab. 6.1: Nicht-modellbasierte Methoden – Zusammenfassung und Empfehlung in Bezug auf den in dieser Arbeit vorgestellten Ansatz und die vordefinierten Erfolgsszenarien (ohne Berücksichtigung des biologischen Vorwissens)

	***k*-means**	**STEM**	**DIB-C**	**PFP**
Anwendbarkeit	mit *kmeans()* in R schnell umsetzbar	Programmierungs-aufwand	Einsatz vom R-Paket *limma()*, Programmierungsaufwand	mit dem PFP-Tool schnell umsetzbar
Anforderungen zu variablen Komponenten	Anzahl Cluster	Parameter *c* und *m*	Signifikanzniveaus der erst- und des zweitrangigen Unterschiede	Parameter α und *m*
Laufzeit	< 1 Min.	ca. 30 Min.	ca. 30 Min.	< 1 Min.
# signifikante Cluster	0	538	692	395
# prognostisch relevante Cluster	a) 0 b) 0	a) 0 b) 0	a) 23 b) 3	a) 1 b) 0
Größe dieser Cluster	-	-	5 - 101	78
Empfehlung	-	-	+	-

Tab. 6.2: Nicht-modellbasierte Methoden – Zusammenfassung und Empfehlung in Bezug auf den in dieser Arbeit vorgestellten Ansatz und die vordefinierten Erfolgsszenarien (unter Berücksichtigung des biologischen Vorwissens mit EXPANDER)

	***k*-means**	**STEM**	**DIB-C**	**PFP**	**EXPANDER**
Anwendbarkeit	mit *kmeans()* in R schnell umsetzbar	Programmierungs-aufwand	Einsatz vom R-Paket *limma()*, Programmierungsaufwand	mit dem PFP-Tool schnell umsetzbar	schnelle Umsetzung
Anforderungen zu variablen Komponenten	Anzahl Cluster	Parameter *c* und *m*	Signifikanzniveaus der erst- und des zweitrangigen Unterschiede	Parameter α und *m*	Organismus, Liste mit Genen und entspr. Clusterzugehörigkeiten, Hintergrund
Laufzeit	< 1 Min.	ca. 30 Min.	ca. 30 Min.	< 1 Min.	ca. 40 Min.
# signifikante Cluster	0	99	81	113	
# prognostisch relevante Cluster	a) 0 b) 0	a) 0 b) 0	a) 3 b) 0	a) 0 b) 0	
Größe dieser Cluster	-	-	5 - 57	-	
Empfehlung	-	-	+	-	

Die Ergebnisse modellbasierter Methoden lassen sich wie folgt zusammenfassen:

Die Anpassung endlicher Mischungsmodelle an die Daten mit *BayesMix()* führt mit MAXPEAR zur erfolgreicheren Clusterfindung nach den hier vordefinierten Szenarien als mit der Optimierungsmethode MINBINDER, jedoch nur bei der Berücksichtigung des biologischen Vorwissens in der Analyse. Andernfalls wird mit keiner dieser beiden Optimierungsmethoden ein nach den vordefinierten Szenarien interessantes Cluster ermittelt. Auffallend ist, dass hier in allen prognostisch relevanten Gengruppen die höhere Expression der Metagene mit einer kürzeren metastasenfreien Zeit und somit mit einer schlechteren Prognose assoziiert ist. Die biologisch interpretierbaren Motive, die einige dieser Cluster zeigen, sind ein wichtiger Ansatzpunkt für eine weitere an diese Arbeit anschließende Untersuchung der vorliegenden Daten.

Die Betrachtung der DP Mischungsmodelle mit *DPpackage()* lieferte die meisten Erfolge nach den vordefinierten Erfolgsszenarien. Einige hiermit ermittelten Cluster lassen sich mittels anderer Methoden verifizieren und mindestens ein Metagen ist auch aus medizinischer Sicht mit dem Brustkrebs relevant assoziiert. Dieses (hier so genannte) Tubulin-Metagen enthält Gene, die sich an mehreren wichtigen biologischen Funktionen beteiligen und schon als prognostisch relevant bekannt sind, jedoch nicht in der Form eines Metagens. Dies ist ein Zeichen dafür, dass diese Methode für die in dieser Arbeit gesetzten Ziele gut geeignet ist. Durch die MINBINDER-Optimierung der Clusterungen werden im Vergleich zu MAXPEAR deutlich mehr signifikante Ergebnisse geliefert. Bei der Berücksichtigung der biologischen Hintergrundinformation werden jedoch deutlich mehr interessante Cluster identifiziert werden. Dies zeigt sich insbesondere, wenn sie in die Distanzmetrik eingebunden werden.

Die Auswertung mit dem in R implementierten Tool DIRECT führt zu ähnlichen Ergebnissen wie die DP Mischungsmodelle. Auch hier führt die Optimierung mit MINBINDER zur erfolgreicheren Clusterfindung nach den beiden vordefinierten Erfolgsszenarien bei der Berücksichtigung des biologischen Vorwissens in der Distanzmetrik, deren deutliche Mehrheit mit einer schlechteren Prognose assoziiert ist.

Eine zusammengefasste Übersicht über die modellbasierten Methoden sowie die Empfehlung eines geeigneten Verfahrens zur Identifizierung prognostisch relevanter Gengruppen mit Hilfe des hier vorgestellten Ansatzes ist den Tabellen 6.3 und 6.4 zu entnehmen:

Tab. 6.3: Modellbasierte Methoden – Zusammenfassung und Empfehlung in Bezug auf den in dieser Arbeit vorgestellten Ansatz und die vordefinierten Erfolgsszenarien (ohne Berücksichtigung des biologischen Vorwissens)

	Finite mixture models		**DP mixture models**		**DIRECT**	
	MINBINDER	**MAXPEAR**	**MINBINDER**	**MAXPEAR**	**MINBINDER**	**MAXPEAR**
Anwendbarkeit	Programmierungsaufwand für die Optimierungsmethoden					
	mit *BayesMix()* und *rjags()* in R einfach umzusetzen, PCA erforderlich		Einsatz von *DPpackage()* in R, PCA erforderlich		mit *DIRECT()* in R einfach umzusetzen	
Anforderungen zu variablen Komponenten	Anzahl Cluster, MCMC-Einstellungen, Definition der a priori-Parameter		MCMC-Einstellungen, Definition der a priori-Parameter		MCMC-Einstellungen, Modelleinstellungen	
Laufzeit	ca. 10 Min.	ca. 10 Min.	ca. 1 Std.	ca. 1 Std.	ca. 15 Min.	ca. 15 Min.
# signifikante Cluster	2	4	52	16	10	7
# prognostisch relevante Cluster	a) 0 b) 0	a) 0 b) 0	a) 2 b) 0	a) 1 b) 0	a) 0 b) 0	a) 0 b) 0
Größe dieser Cluster	-	-	6 - 14	9	-	-
Empfehlung	-	-	+	+	-	-

Tab. 6.4: Modellbasierte Methoden – Zusammenfassung und Empfehlung in Bezug auf den in dieser Arbeit vorgestellten Ansatz und die vordefinierten Erfolgsszenarien (unter Berücksichtigung des biologischen Vorwissens mit EXPANDER, Zuordnung der Gene zu den GO-Gruppen und unter Berücksichtigung des GO-Ähnlichkeitsmaßes in der Distanzmetrik (Ergebnisse in Klammern))

	Finite mixture models		**DP mixture models**		**DIRECT**		**EXPANDER**
	MINBINDER	**MAXPEAR**	**MINBINDER**	**MAXPEAR**	**MINBINDER**	**MAXPEAR**	
Anwendbarkeit	Einsatz von R-Paketen *topGO()* und *GOSim()* für die GO-Gruppenzugehörigkeit bzw. Berechnung des Ähnlichkeitsmaßes der Genprodukte, Programmierungsaufwand für die Optimierungsmethoden						Schnelle Umsetzung
	mit *BayesMix()* und *rjags()* in R einfach umzusetzen, PCA erforderlich		Einsatz von DPpackage() in R, PCA erforderlich		mit *DIRECT()* in R einfach umzusetzen		
Anforderungen zu variablen Komponenten	Anzahl Cluster, MCMC-Einstellungen, Definition der a priori-Parameter		MCMC-Einstellungen, Definition der a priori-Parameter		MCMC-Einstellungen, Modelleinstellungen		Organismus, Liste mit Genen und entspr. Clusterzugehörigkeiten, Hintergrund
Laufzeit	ca. 10 Min.	ca. 10 Min.	ca. 1 Std.	ca. 1 Std.	ca. 15 Min.	ca. 15 Min.	ca. 40 Min.
# signifikante Cluster	3 (3)	61 (76)	126 (250)	59 (27)	17 (211)	16 (36)	
# prognostisch relevante Cluster	a) 1 b) 0	a) 1 (3) b) 0 (1)	a) 4 (15) b) 0	a) 0 (1) b) 0	a) 0 (10) b) 0 (1)	a) 0 (2) b) 0	
Größe dieser Cluster	5	14 - 82	5 - 33	218	7 - 23	48 - 94	
Empfehlung	-	(+)*	+	-	(+)*	-	* nach der Berücksichtigung des GO-Ähnlichkeitsmaßes in der Distanzmetrik

Die Einbindung des verwendeten GO-Ähnlichkeitsmaßes in die Distanzmetrik ist bei allen Methoden nur eingeschränkt zu empfehlen, da durch die Medianbildung der Gene Information verloren geht und hierdurch u.U. auch verzerrte mediane Zeitverläufe betrachtet werden.

Zum Vergleich dieser Clustermethoden wurde ein weiterer Genexpressionsdatensatz mit kurzen Zeitreihen herangezogen. In den AAS-Daten wurde zu 5 Zeitpunkten die Reaktion der Saccharomyces cerevisiae Hefe auf den Aminosäurenentzug gemessen. Die Zuordnung dieser Gene zu den GO-Gruppen erfolgte durch http://www.yeastgenome.org und resultierte für die CC-Gruppe in einer Genexpressionsmatrix mit 565, für die MF-Gruppe mit 670 und für die BP-Gruppe mit 666 Zeilen. Da für diese Daten keine Überlebenszeitdaten vorliegen, konnte an dieser Stelle nur ein Vergleich der nach der Clusteranalyse resultierenden Clusterstrukturen erfolgen. Dabei war der Einsatz von DIB-C und DIRECT nicht möglich, da für die AAS-Daten keine Wiederholungen vorliegen, die für die erfolgreiche Methodenumsetzung erforderlich wären.

Bei der Fragestellung, ob die Clusterstrukturen für die MCF7- und AAS-Daten ähnlich sind oder ob bei einem Datensatz vermehrt kleinere und bei dem anderen vermehrt größere Gengruppen gefunden werden, konnten keine großen Unterschiede festgestellt werden. Während der Einsatz der endlichen Mischungsmodelle keine brauchbaren Ergebnisse für beide Datensätze liefert, scheinen STEM, DP Mischungsmodelle und das PFP-Verfahren in gleicher Weise besser zu trennen. Dabei findet PFP deutlich mehr kleinere Cluster und Gengruppen mit signifikant überrepräsentierten TF-Bindungsstellen.

Durch die hier vorgestellte Analyse bildet diese Arbeit eine fundierte Grundlage für weitere Untersuchungen sowohl der vorliegenden als auch ähnlicher Gen-expressionsdaten. So haben mehrere hier identifizierte Cluster einige Gene gemeinsam und es ist denkbar, diese einzeln auf ihre prognostische Relevanz zu untersuchen (wie es bspw. in der Medizin üblich ist) und zu einem Cluster zusammenzufassen. Denkbar ist auch die Überprüfung, wie sich die zugehörigen Genexpressionen je nach Mammakarzinom-Gruppe – wie z.B. die positiven und die negativen ErbB2-Tumore – oder nach anderen biologisch wichtigen Aspekten getrennt verhalten. So wie für das SRF-Metagen, bei dessen Untersuchung im nächsten Schritt von Interesse ist, ob die zugehörigen Gene in ErbB2 positiven Mammakarzinomen stärker exprimiert sind als in den ErbB2 negativen Tumoren. Falls dies der Fall ist, würde sich die Frage anschließen,

ob dieses Cluster eine Subgruppe ErbB2 positiver Karzinome definiert, welche eine bessere Prognose haben. Dies ist ein interessanter Punkt für eine an diese Arbeit anknüpfende Publikation. Außerdem wurden nicht alle identifizierten Gengruppen im Einzelnen näher betrachtet, so dass hier u.U. noch weitere neue Metagene gefunden, aber noch nicht als solche identifiziert wurden.

Auch das Überdenken der vordefinierten Erfolgsszenarien ist denkbar, so dass der Erfolg nicht nur medizinisch (Szenario (a)), sondern auch wissenschaftlich (ähnlich Szenario (b)) begründet werden könnte. Unter Umständen würden sich dadurch noch weitere interessanten Metagene identifizieren lassen.

Aktuell läuft an der TU Dortmund (Fakultät Statistik) eine Untersuchung der Sensitivität der Einstellungen in PRIMA. Je nach Ergebnis wäre eine weitere Untersuchung der ermittelten Cluster auf die Überrepräsentation von Transkriptionsfaktorbindungsstellen von Interesse.

Da die Berücksichtigung des biologischen Wissens eine entscheidende Rolle bei der Identifizierung prognostisch relevanter Metagene zu spielen scheint, ist die Einbindung weiterer entsprechender Datenbanken (ähnlich MIPS, vgl. Mewes et al., 2006) möglich. Wegen unterschiedlicher zeitlicher Verlaufsmuster in einigen Gengruppen ist eine Optimierung der hier verwendeten Distanzmetrik bzw. Einsatz einer ganz anderen denkbar.

Ein weiterer interessanter Aspekt ist die Adjustierung nach relevanten klinischen Variablen. Bei den Brustkrebsdaten kämen z.B. das Alter der Patientinnen oder die Tumorgröße in Frage. Da bei den vorliegenden Daten zu der Rotterdam-Kohorte diese wie auch viele weiteren Angaben jedoch komplett fehlen, konnte dieser Punkt hier leider nicht berücksichtigt werden. Dies könnte aber ein zusätzlicher Anhaltspunkt bei der Analyse ähnlicher Krebsdaten mit dem hier vorgestellten Ansatz werden.

Ferner ist die im Methodenüberblick schon erwähnte so genannte Bayesianische „sphärische” Clusteranalyse eine vielversprechende Möglichkeit zum Clustern von Genexpressions-Zeitreihen, auf die in den anschließenden Untersuchungen näher eingegangen werden kann. Dabei sind die von-Mises-Fisher (vMF) Verteilungen eine weit verbreitete Verteilungsart beim Clustern auf der Sphäre. Ob durch Mischung von diesen ähnlich der Normalverteilung auch alle anderen Verteilungen sich approximieren

lassen, sollte dann jedoch geprüft werden. Zumindest lehnen Dortet-Bernadet und Wicker (2008) diese in ihrer Arbeit ab, da die Iso-density Linien von den vMF Verteilungen zirkulär auf der Sphäre sind und von den darauf projizierten Normalverteilungen unterschiedliche Formen annehmen können, was eine flexiblere Clusterbildung ermöglicht.

Weiterhin könnte eine Gegenüberstellung von der Posterior Allocation Probability Matrix von Fu et al. (2013), die die Wahrscheinlichkeiten enthält, dass die Beobachtung *i* zum *k*-ten Cluster gehört, und der Posterior Similarity Matrix von Frisch (2010) mit den a posteriori-Wahrscheinlichkeiten, dass die Gene *i* und *j* dem gleichen Cluster angehören, angegangen werden.

Abschließend lässt sich sagen, dass diese Arbeit eine fundierte Grundlage für die Erforschung der Brustkrebsentstehung bildet und somit fortgesetzt werden sollte, indem die angesprochenen Ansatzpunkte in einer weiteren, an diese Arbeit anschließenden Untersuchung näher betrachtet werden.

Literatur

Akaike, H. (1974): „A new look at statistical model identification.” *IEEE Transactions on Automatic Control*, 19:716-723.

Alexa, A., Rahnenführer, J. (2006): „Gene set enrichment analysis with topGO.” R-Paket topGO.

Alexa, A., Rahnenführer, J., Lengauer, T. (2006): „Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.” *Bioinformatics*, 22:1600-1607.

Anderson, P. K., Keiding, N. (2006): *Survival and event history analysis*. John Wiley & Sons Ltd, London.

Antoniak, C. E. (1974): „Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.” *Annals of Statistics*, 2:1152-1174.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000): „Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.” *Nat Genet*, 25(1):25-9.

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S. (2005): „Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research*, 6:1-39.

Bayes, T. (1763): „An essay towards solving a problem in the doctrine of chances." *Philosophical Transactions of the Royal Society*, 370-418.

Benedetto, J. J., Fickus, M. (2003): „Finite normalized tight frames." *Adv Comput Math*, 18:357-385.

Bensmail, H., Celeux, G., Raftery, A. E., Robert, C. P. (1997): „Inference in model-based cluster analysis." *Statistics and Computing*, 7:1-10.

Benjamini, Y., Hochberg, Y. (1995): „Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B*, 57:289-300.

Biernacki, C., Govaert, G. (1999): „Choosing models in model-based clustering and discriminant analysis." *J Stat Comput Simulation*, 64:49-71.

Binder, D. A. (1978): „Bayesian cluster analysis." *Biometrika*, 65:31-38.

Blackwell, D., MacQueen, J. B. (1973): „Ferguson distributions via Pólya urn schemes." *Annals of Statistics*, 1:353-355.

Cadenas, C., Franckenstein, D., Schmidt, M., Gehrmann, M., Hermes, M., Geppert, B., Schormann, W., Maccoux, L. J., Schug, M., Schumann, A., Wilhelm, C., Freis, E., Ickstadt, K., Rahnenführer, J., Baumbach, J. I., Sickmann, A., Hengstler, J. G. (2010): „Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer." *Breast Cancer Research*, 12:R44.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R. (2004): „The Gene Ontology annotation (GOA) database: sharing knowledge in uniprot with Gene Ontology." *Nucleic Acids Res*, 32:D262-D266.

Carlin, B. P., Louis, T. A. (2000): *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Carroll, J. S., Meyer, C. A., Song, J., Li, W., Geistlinger, T. R., Eeckhoute, J., Brodsky, A. S., Keeton, E. K., Fertuck, K. C., Hall, G. F., Wang, Q., Bekiranov, S., Sementchenko, V., Fox, E. A., Silver, P. A., Gingeras, T. R., Liu, X. S., Brown, M. (2006): „Genome-wide analysis of estrogen receptor binding sites." *Nat Genet,* 38:1289–97.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., Herskowitz, I. (1998): „The transcriptional program of sporulation in budding yeast." *Science,* 282:699-705.

Chvátal, V. (1979): „A greedy heuristic for the set cover problem." *Mathematics of Operations Research,* 4:233-235.

Cox, D. R. (1972): „Regression models and life tables." *Journal of the Royal Statistical Society, Series B,* 34(2):187-220.

Dahl, D. B. (2006): „Model-based clustering for expression data via a Dirichlet process mixture model." In Do, K. A., Müller, P., Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics,* 201-218, Cambridge University Press.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977): „Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, Series B,* 39:1-38.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., D'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., Sotiriou, C., TRANSBIG Consortium (2007): „Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series." *Clin Cancer Res,* 13:3207-14.

Di Camillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S. K., Trajanoski, Z., Cobelli, C. (2005): „A quantization method based on threshold optimization for microarray short time series." *BMC Bioinformatics,* 6(4):S11.

Dortet-Bernadet, J. L., Fan, Y. (2007): „Simultaneous clustering and variable selection with respect to correlation." School of Mathematics and Statistics, University of New South Wales, Sidney.

Dortet-Bernadet, J. L., Wicker, N. (2008): „Model-based clustering on the unit sphere with an illustration using gene expression profiles.” *Biostatistics*, 9(1):66-80.

Edgar, R., Domrachev, M., and Lash, A. E. (2002): „Gene expression Omnibus: NCBI gene expression and hybridization array data repository.” *Nucleic Acids Res*, 30:207-10.

Efron, B. (1977): „The efficiency of Cox’s likelihood function for censored data.” *Journal of the American Statistical Association*, 72(359):557-565.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998): „Cluster analysis and display of genome-wide espression patterns.” *Proc Natl Acad Sci USA*, 95:14863-14868.

Elkon, R, Linhart, C., Sharan, R., Shamir, R., Shiloh, Y. (2003): „Genome-wide in silencio identification of transcriptional regulators controlling the cell cycle in human cells.” *Genome Research*, 13(5):773-780.

Ernst, J., Nau, G. J., Bar-Joseph, Z. (2005): „Clustering short time-series gene expression data.” *Bioinformatics*, 21(1):159-168.

Escobar, M. D., West, M. (1995): „Bayesian density estimation and inference using mixtures.” *J Amer Statist Assoc*, 90:577-588.

Fahrmeir, L., Kaufmann, H. L., Ost, F. (1981): *Stochastische Prozesse*. Carl Hanser Verlag, München.

Ferguson, T. S. (1973): „A Bayesian analysis of some nonparametric problems.” *Annals of Statistics*, 1:209-230.

Fraley, C., Raftery, A. E. (1998): „How many clusters? Which clustering method? Answers via model-based cluster analysis.” *Computer J*, 41:578-588.

Fraley, C., Raftery, A. E. (1999): „MCLUST: Software for model-based cluster analysis.” *J Classification*, 16:297-306.

Freis, E., Selinski, S., Hengstler, J. G., Ickstadt, K. (2012): „Cluster analytic strategy for identification of metagenes relevant for prognosis of node negative breast cancer.” In L. Schmidt-Thieme, A. Geyer-Schulz, W. Gaul (Eds.), *Challenges concerning the*

DATA ANALYSIS – COMPUTER SCIENCE – OPTIMIZATION interface. Studies in Classification, Data Analysis, and Knowledge Organization, 10:475-483, Springer.

Freis, E., Selinski, S., Weibert, B., Krahn, U., Schmidt, M., Gehrmann, M., Hermes, M., Maccoux, L., West, J., Schwender, H., Rahnenführer, J., Hengstler, J. G., Ickstadt, K. (2009): „Effects of metagene calculation on survival: An integrative approach using cluster and promoter analysis." In *Sixth International Workshop on Computational Systems Biology* (pp. 47–50), TICSP series, 48. Tampere, Finland.

Fritsch, A. (2010): „Bayesian mixtures for cluster analysis and flexible modeling of distributions." Dissertation, TU Dortmund.

Fritsch, A., Ickstadt, K. (2009): „Improved criteria for clustering based on the Posterior Similarity Matrix." *Bayesian Analysis*, 4:367-392.

Froehlich, H. (2012): „Computation of functional similarities between GO terms and gene products; GO enrichment analysis." R-Paket GOSim.

Frühwirth-Schnatter, S. (2006): *Finite mixture and Markov switching models*. Springer, Berlin.

Frühwirth-Schnatter, S., Kaufmann, S. (2002): „Bayesian clustering of many short time series." *Working Paper*, Vienna University of Economics and Business Administration.

Fu, A. Q., Russell, S., Bray, S. J., Tavare, S. (2011): „Bayesian Clustering of Multivariate Data Under the Dirichlet-Process Prior." R-Paket DIRECT.

Fu, A. Q., Russell, S., Bray, S. J., Tavare, S. (2013): „Bayesian Clustering of Replicated Time-Course Gene Expression Data with Weak Signals." *Annals of Applied Statistics*, 7(3):1334-1361.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O. (2000): „Genomic expression programs in the response of yeast cells to environmental changes." *Molecular Biology of the Cell*, 11:4241-4257.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2004): *Bayesian data analysis*. Chapman & Hall, London.

Geman, S., Geman, D. (1984): „Stochastic relaxation, Gibbs distributions and the Bayesian restauration of images." *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 6:721-741.

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

Gerber, G. K., Dowell, R. D, Jaakkola, T. S., Gifford, D. K. (2007): „Automated discovery of functional generality of human gene expression programs." *PLoS Comput Biol*, 3(8):e148.

Ghazoui, Z., Buffa, F. M., Dunbier, A. K., Anderson, H., Dexter, T., Detre, S., Salter, J., Smith, I. E., Harris, A. L., Dowsett, M. (2011): „Close and stable relationship between proliferation and a hypoxia metagene in aromatase inhibitor-treated ER-positive breast cancer." *Clin Cancer Res*, 17(9):3005-3012.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996): *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Glahn, F., Schmidt-Heck, W., Zellmer, S., Guthke, R., Wiese, J., Golka, K., Hergenroder, R., Degen, G. H., Lehmann, T., Hermes, M., Schormann, W., Brulport, M., Bauer, A., Bedawy, E., Gebhardt, R., Hengstler, J. G., Foth, H. (2008): „Cadmium, cobalt and lead cause stress response, cell cycle deregulation and increased steroid as well as xenobiotic metabolism in primary normal human bronchial epithelial cells which is coordinated by at least nine transcription factors." *Arch Toxicol*, 82:513-524.

Grün, B. (2011): „bayesmix: bayesian mixture models with JAGS." R-Paket BayesMix.

Hastings, W. K. (1970): „Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57:97-109.

Heard, N. A., Holmes, C. C., Stephens, D. A., Jand, D. J., Dimopoulos, G. (2005): „Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges." *PNAS*, 102(47):16939-16944.

Hermes, M. (2007): „Konditionale Expression von Her2/NeuT: Einfluss auf die Zell- und Tumorenentwicklung." Dissertation, Universität Leipzig.

Hosmer, D. W. Jr., Lemeshow, S. (1999): *Applied survival analysis regression modelling of time to event data*, John Wiley & Sons, New York.

Hubert, L., Arabie, P. (1985): „Comparing partitions." *Journal of Classification*, 2:193-218.

Hummel, M. B. (2009): „Strategien für die Expressionsanalyse in funktionellen Gengruppen." Dissertation, Universität München.

Hurn, M., Justel, A., Robert, C. P. (2003): „Estimating mixtures of regressions." *Journal of Computational and Graphical Statistics*, 12:55-79.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., Speed, T. P. (2003): „Summaries of Affymetrix GeneChip probe level data." *Nucleic Acids Res*, 31(4):e15.

Jara, A., Hanson, T., Quintana, F. A., Müller, P., Rosner, G. L. (2012): „DPpackage: Bayesian nonparametric modeling in R." R-Paket DPpackage.

Kaufman, L., Rousseeuw, P. J. (1990): „Finding groups in data – An introduction to cluster analysis." John Wiley & Sons, New York.

Kim, J., Kim, J. H. (2007): „Difference-based clustering of short time-course microarray data with replicates." *Bioinformatics*, 8:253.

Klein, J. P., Moeschberger, M. L. (2003): *Survival analysis: Techniques for censored and truncated data*. Springer, New York.

Köhne, A.-K. (2008): „Integrative Datenanalyse von Phosphorproteinen und RNA–Expressionsdaten beim Mammakarzinom." Diplomarbeit, TU Dortmund.

König, A. (2014): „Temporal Activation Profiles of Gene Sets for the Analysis of Gene Expression Time Series." Dissertation, TU Dortmund.

Kormaksson, M., Booth, J. G., Figueroa, M. E., Melnick, A. (2012): „Integrative model-based clustering of microarray methylation and expression data." *Annals of Applied Statistics*, 6(3):1327-1347.

Kovačevič, J., Chebira, A. (2007a): „Life beyond bases: the advent of frames: part 1." *IEEE SP Mag*, 24(4):86-104.

Kovačevič, J., Chebira, A. (2007b): „Life beyond bases: the advent of frames: part 2." *IEEE SP Mag*, 24(5):115-125.

Krahn, U. (2008): „Identifikation von Clustern in Genexpressions-Zeitreihen zur Analyse der Zellentwicklung." Diplomarbeit, TU Dortmund.

Lee, P. H., Lee, D. (2005): „Modularized learning of genetic interavction networks from biological annotations and mRNA expression data." *Bioinformatics*, 21(11):2739-2747.

Lin, D. (1998): „An informativ-theoretic definition of similarity." *Proceedings of the 15th International Conference on machine Learning* (pp.296-304).

Lohr, M., Köllmann, C., Freis, E., Hellwig, B., Hengstler, J.G., Ickstadt, K., Rahnenführer, J. (2012): „Optimal strategies for sequential validation of significant features from high-dimensional genomic data." *J Toxicol Environ Health A, 75*(8-10):447-460.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P. Haris, A., Bergh, J., Foekens, J. A., Klijn, J.G., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M. J., Sotiriou, C. (2007): „Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade." *J Clin Oncol*, 25:1239-46.

Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A. (2003): „Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." *Bioinformatics*, 19(10):1275-83.

MacQueen, J. B. (1967): „Some methods for classification and analysis of multivariate obseravations." In *Proceedings of 5th Berkeley Symposium on Mathematical statistics and probability*, 281-297, University of California Pres.

Madeira, S. C., Oliveira, A. L. (2004): „Biclustering algorithms for biological data analysis: A survey." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24-45.

Martin, M., Müller, K., Cadenas, C., Hermes, M., Zink, M., Hengstler, J. G., Käs, J. A. (2012): „ErbB2 overexpression triggers transient high mechanoactivity of breast tumor cells." *Cytoskeleton*, 69:267-277.

MATLAB (2012): The MathWorks Inc, Natick, Massachusetts.

McLachlan, G., Peel, D. (2000): *Finite mixture models*. Wiley, New York.

Medvedovic, M., Sivaganesan, S. (2002): „Bayesian infinite mixture model based clustering of gene expression profiles." *Bioinformatics*, 18:1194-206.

Medvedovic, M., Yeung, K. Y., Burngarner, R. E. (2004): „Bayesian infinite mixture model based clustering of replicated microarray data." *Bioinformatics*, 20:1222-32.

Ménard, S., Fortis, S., Castiglioni, F., Agresti, R., Balsari, A. (2001): „HER2 as a prognostic factor in breast cancer." *Oncology*, 61(2):67-72.

Merton, R. C. (1971): „Optimum consumption and portfolio rules in a continuous-time model." *J. Econom. Theory*, 3:373-413.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953): „Equations of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21:1087-1092.

Mewes, H. W., Frishman, D., Noubibou, O., Pagel, P., Rattei, T., Oesterheld, M., Ruepp, A., Stümpflen, V. (2006): „MIPS: analysis and annotation of proteins from whole genomes in 2005." *Nucleic Acids Res*, 34: D169-D172.

Milligan, G. W., Cooper, M. C. (1985): „An examination of procedures for determining the number of clusters in a data set." *Psychometrika*, 50:159-179.

Mori, K., Oura, T., Noma, H., Matsui, S. (2013): „Cancer outlier analysis based on mixture modeling of gene expression data." *Computational and Mathematical Methods in Medicine*, 2013:693901.

Nueda, M., Sebastián, P., Tarazona, S., Garcia-Garcia, F., Dopazo, J., Ferrer, A., Conesa, A. (2009): „Functional assessment of time course microarray data." *BMC Bioinformatics*, 10(6):S9.

Pearson, K. (1894): „Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London A*, 185:71-110.

Petry, I. B., Fieber, E., Schmidt, M., Gehrmann, M., Gebhard, S., Hermes, M., Schormann, W., Selinski, S., Freis, E., Schwender, H., Brulport, M., Ickstadt, K., Rahnenfuhrer, J., Maccoux, L., West, J., Kolbl, H., Schuler, M., Hengstler, J. G. (2010): „ErbB2 induces an antiapoptotic expression pattern of Bcl-2 family members in node-negative breast cancer." *Clin Cancer Res*, 16(2):451-460.

Plummer, M. (2009): „rjags: Interface to the JAGS MCMC library." R-Paket.

R Development Core Team (2010): „*R: A language and environment for statistical computing*." R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

Raftery, A. (1996): „Hypothesis testing and model selection." In „Markov chain Monte Carlo in practice." (pp. 163-187), Gilks, W.R., Richardson, S., Spiegelhalter, D. J. (Eds.), Chapman & Hall, London.

Ramoni, M. F., Sebastiani, P., Kohane, I. S. (2002): „Cluster analysis of gene expression dynamics." *Proc Natl Acad Sci USA*, 99(14):9121-26.

Rand, W. M. (1971): „Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66:846-850.

Rasmussen, C. E., de la Cruz, B. J., Ghahramani, Z., Wild, D. L. (2009): „Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures." *IEEE/ACM T Comput Bi*, 6:615-628.

Resnik, P. (1995): „Using information content to evaluate semantic similarity in a taxonomy." *Proc 14th Int'l Joint Conf Artifical Intelligence*, 48-453.

Resnik, P. (1999): „Semantic similarity in a taxonomy: An information-based measure and its applicationto problems of ambiguity in natural language." *J Artif Intell Res*, 11:95-130.

Robert, C. P., Castella, G. (1999): *Monte Carlo statistical methods*. Springer, London.

Robert-Koch-Institut (2013): Zentrum für Krebsregisterdaten, http://www.rki.de/Krebs/DE/Content/Krebsarten/Brustkrebs/brustkrebs_node.html.

Rody, A., Holtrich, U., Pusztai, L., Liedtke, C., Gaetje, R., Ruckhaeberle, E., Solbach, C., Hanker, L., Ahr, A., Metzler, D., Engels, K., Karn, T., Kaufmann, M. (2009): „T-cell metagene predicts a favourable prognosis in estrogen receptor negative and HER2 positive breast cancers." *Breast Cancer Res*, 11:R15.

Sacchi, L., Bellazzi, R., Larizza, C., Magni, P., Curk, T., Petrovic, U., Zupan, B. (2005): „TA-clustering: Cluster analysis of gene expression profiles through temporal abstractions." *Int J Med Inform*, 74(7-8):505-517.

Saff, E., Kuijlaars, A. B. J. (1997): „Distributing many points on a sphere." *Math Intell*, 19:5-11.

Schlicker, A., Domingues, F. S., Rahnenführer, J., Lengauer, T. (2006): „A new measure for functional similarity of gene products based on Gene Ontology." *BMC Bioinformatics*, 7:302.

Schmidt, M., Bohm, D., von Torne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H. A., Hengstler, J. G., Kolbl, H., Gehrmann, M. (2008): „The humoral immune system has a key prognostic impact in node-negative breast cancer." *Cancer Res*, 68:5405-5413.

Schmidt, M., Petry, I. B., Böhm, D., Lebrecht, A., von Törne, C., Gebhard, S., Gerhold-Ay, A., Cotarelo, C., Battista, M., Schormann, W., Freis, E., Selinski, S., Ickstadt, K., Rahnenführer, J., Sebastian, M., Schuler, M., Koelbl, H., Gehrmann, M., Hengstler, J. G. (2010): „Ep-CAM RNA expression predicts metastasis-free survival in three cohorts of untreated node-negative breast cancer." *Clin Cancer Res*, *16*(2):451-460.

Schwarz, G. (1978): „Estimating the dimension of a model." *Annals of Statistics*, 6:461-464.

Sethuraman, J. (1994): „A constructive definition of Dirichlet priors." *Statistica Sinica*, 4:639-650.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., Corrales, F. J., Rubio, A. (2005): „Correlation between gene expression and GO semantic similarity." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4):330-338.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y., Elkon, R. (2005): „EXPANDER - an integrative program suite for microarray data analysis." *BMC Bioinformatics*, 6:232.

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., McGuire, W. L. (1987): „Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene." *Science*, 235:177-182.

Smyth, G. K. (2004): „Linear models and empirical Bayes methods for assessing differential expresssion in microarray experiments." *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article3.

Smyth, G. K. (2005): „Limma: Linear models for microarray data." *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Gentleman, R.C., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (eds), Springer, New York, 397-420.

Soule, H., Vazquez, J., Long, A., Albert, S., Brennan, M. (1973): „A human cell line from a pleural effusion derived from a breast carcinoma." *Journal of the National Cancer Institute*, 51(5):1409-1416.

Speer, N., Spieth, C., Zell, A. (2004): „A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology." In *Proceedings of the CICB 2004*.

Springer, T., Ickstadt, K., Stöckler, J. (2011): „Frame potential minimization for clustering short time series." *Adv Data Anal Classif*, 5:341-355.

Stephens, M. (2000): „Bayesian analysis of mixture models with unknown number of components – an alternative to reversible jump method." *Ann Stat*, 24:40-74.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S, Mesirov, J. P. (2005): „Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profile." *PNAS*, 102(43):15545-15550.

Sun, J., Garibaldi, J. M., Kenobi, K. (2012): „Robust Bayesian clustering for replicated gene expression data.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(5):1504-1514.

Tanay, A. (2005): „Computational analysis of transcriptional programs: Function and evolution.” Dissertation, Universität Tel Aviv.

Tavazoie, S., Hughes, J. D, Campbell, M. J., Cho, R. J., Church, G. M. (1999): „Systematic determination of genetic network architecture.” *Nature Genetics Letters*, 22.

Tchagang, A. B., Bui, K. V., McGinnis, T., Benos, P. V. (2009): „Extracting biologically significant patterns from short time series gene expression data.” *BMC Bioinformatics*, 10:255.

Therneau, T., Grambsch, P. (2000): *Modelling Survival Data, Extending the Cox Model*. Springer, New York.

Thomas, A., O'Hara, B., Ligges, U., Sturtz S. (2006): „Making BUGS Open.” *R News*, 6:12-17.

Tibshirani, R., Walther, G., Hastie, T. (2001): „Estimating the number of clusters in a data set via the gap statistic.” *J R Statistic Soc B*, 63(2):411-423.

Trost, T. M., Lausch, E. U., Fees, S. A., Schmitt, S., Enklaar, T., Reutzel, D., Brixel, L. R., Schmidtke, P., Maringer, M., Schiffer, I. B., Heimerdinger, C. K., Hengstler, J. G., Fritz, G., Bockamp, E. O., Prawitt, D., Zabel, B. U., Spangenberg, C. (2005): „Premature senescence is a primary fail-safe mechanism of ErbB2-driven tumorigenesis in breast carcinoma cells.” *Cancer Res*, 65:840-849.

Wang, K., Ng, S. K., McLachlan, G. J. (2012): „Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects.” *BMC Bioinformatics*, 13:300.

Wang, L., Wang, X. (2013): „Hierarchical Dirichlet process model for gene expression clustering.” *EURASIP Journal on Bioinformatics and System Biology*, 2013:5.

Wang, X., Wu, M., Li, Z., Chan, C. (2008): „Short time-series microarray analysis: Methods and challenges.” *BMC Syst Biol*, 2:58.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D. Tommermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D., Foekens, J. A. (2005): „Gene-expression profiles to predict distant metastasis of lymphnodenegative primary breast cancer." *Lancet*, 365:671-79.

Wardle, F. C., Odom, D. T., Bel, G. W., Yuan, B., Danford, T. W., Wiellette, E. L., Herbolsheimer, E., Sive, H. L., Young, R. A., Smith, J. C. (2006): „Zebrafish promoter microarrays identify actively transcribed embryonic genes." *Genome Biology*, 7:R71.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., Schacherer, F. (2000): „TRANSFAC: An intergrated system for gene expression regulation." *Nucleic Acids Res*, 28:316-319.

Winter, S. C., Buffa, F. M., Silva, P., Miller, C., Valentine, H. R., Turley, H., Shah, K. A., Cox, G. J., Corbridge, R. J., Homer, J. J., Musgrove, B., Slevin, N., Sloan, P., Price, P., West, C. M., Harris, A. L. (2007): „Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers." *Cancer Res*, 67(7):3441-3449.

Wu, H., Yuan, M., Kaech, S., Halloran, M. (2007): „A statistical analysis of memory CD8 T cell differentiation: An application of a hierarchical state space model to a short time course microarray experiment." *Annals of Applied Statistics*, 1(2):442-458.

Urgard, E., Vooder, T., Võsa, U., Välk, K., Liu, M, Luo, C., Hoti, F., Roosipuu, R., Annilo, T., Laine, J., Frenz, C. M., Zhang, L., Metspalu, A. (2011): „Metagenes Associated with Survival in Non-Small Cell Lung Cancer." *Cancer Informatics*, 10:175-183.

Anhang A

Zusätzliche Tabellen

Tab. A.1: Die nach der DIB-C-Analyse signifikanten Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)			GO	TF
U12.NNNDN,NNNV	0,004	1,71	x (x)	x	x	a)		-	-
U5.NNNDN,NNNN	0,005	2,47	x	x	x	a)		x	-
U11.NNDNN,NAVN	< 0,001	0,62	x	x	x (x)	a)		-	-
U12.NNDDN,NANV	0,015	1,63	x	x	x	a)		-	-
U14.DNNIN,NNNN	< 0,001	0,46	x	x (x)	-	a)	b)	-	-
U13.DNNIN,NNNN	< 0,001	0,46	x	x (x)	-	a)	b)	-	-

U16.DNNDN,VNNN	0,009	1,66	x	x (x)	-	a)		-	-
U16.DNNIN,NNVA	0,002	0,46	x	x (x)	-	a)	b)	-	-
U15.DNNIN,NNNN	0,001	0,52	x	x (x)	-	a)		-	-
U7.NNNNN,NAVN	0,045	0,57	x	x	x	a)		-	-
U3.NNNNN,NAVN	0,045	0,57	x	x	x	a)		-	-
U2.NNNNN,NNAV	0,005	1,91	x	x	x	a)		-	-
U8.NNNDN,NNAV	0,002	2,39	x	x	x	a)		x	-
U9.NNDDN,NNNN	0,048	1,74	x	-	x	a)		-	-
U15.ININN,NNNN	0,026	1,61	x	-	x	a)		-	-
U10.NNDDN,NNNN	0,033	1,73	x	-	x	a)		-	-
U12.NNNDN,NNNN	< 0,001	1,69	x	x	-	a)		x	-
U16.INDDN,ANNN	< 0,001	1,47	x	x	-	a)		-	-
U12.NNNNN,NNNA	0,001	0,24	x	x	-	a)		-	-
U4.NNNNN,NNNA	0,012	0,31	x	x	-	a)		-	-
U16.NINNN,VNNN	0,034	0,61	x	x	-	a)		-	-
U12.DNNNN,VNNN	0,036	2,09	x	x	-	a)		-	-
U8.NNNNN,NNNA	0,037	0,33	x	x	-	a)		-	-

Tab. A.2: Die nach der DIB-C-Analyse signifikanten Cluster mit zugehörigen Genen

Cluster	Gene
U12.NNNDN,NNNV	APEX1, ASF1B, C16orf74, CA12, CDCA7L, CDCA8, CLU, CXXC5, DSCAM, FAM64A, FLJ37453, GINS2, IFRD1, IGF1R, KIAA0953, LOC146909, MTHFD1, NA, NOLA2, RAD51, SCARB1, SNHG8, TMEM64, TMEM97
U5.NNNDN,NNNN	BIRC5, C16orf33, CCND1, CDC20, CDCA5, CLIC3, CYB5A, DTNA, DYNLL1, EEF2K, FAM38D, FBXO9, FLJ45983, FOXM1, GATA3, GEMIN5, GINS2, GPR87, IGF1R, IHPK2, KIAA0953, KRT8, LOC120364, LOC645431, LRP8, MSX2, MYBL2, NA, NME1, NOL3, OSR2, PFAS, PKIA, POLR3K, RRm2, SCFD2, SH3BGRL, SIDT1, SLC5A6, SPC24, SPOCK1, TIMP3, TMEM46, TMEM48, TMEM64, TYMS, WDR77, ZWINT
U11.NNDNN,NAVN	CREM, EEF1A1, LANCL2, NLK, PURB, TMPRSS3, TRAF3
U12.NNDDN, NANV	EMP2, F2RL1, GEMIN5, INPP4B, MCM5, MCM7, TMEM46
U14.DNNIN,NNNN	CBS, HERPUD1, ISL2, ITGB4, MAP2K5, MTX1, MZF1, SEMA4D, SGCG, TBC1D2, TSC22D3, TXNIP
U13.DNNIN,NNNN	CBS, HERPUD1, ITGB4, MAP2K5, MTX1, MZF1, SEMA4D, SGCG, TBC1D2, TSC22D3, TXNIP
U16.DNNIN,NNVA	AARS, CBS, DDIT4, HERPUD1, ISL2, MTX1, MZF1, SEMA4D, SDCD, TBC1D2, TRIB3, TSC22D3, TXNIP
U16.DNNDN,VNNN	ABCC5, ESPL1, KCNJ8, NR2F2, PCDH19, PHF15
U15.DNNIN, NNNN	AARS, ISL2, MZF1, SEMA4D, SGCG, TBC1D2, TSC22D3, TXNIP
U7.NNNNN,NAVN	ANKH, CD59, CPM, EEF1A1, GOLM1, LOC285812, OSMR, PURB, RFTN1, SMURF1, SPRR1A, TMPRSS3, TRAF3, UGCGL2, ZNF275
U3.NNNNN,NAVN	ANKH, CD59, CPM, CREM, EEF1A1, GOLM1, LANCL2, LOC285812, NA, NKL, OSMR, PURB, RFTN1, SMURF1, SPRR1A, TMPRSS3, TRAF3, UGCGL2, ZNF275
U8.NNNDN,NNAV	BIRC5, BMP7, CCND1, CDC20, CDCA5, CLIC3, CLU, EEF2K, FAM38D, FLJ13236, FLJ45983, FOXM1, GATA3, GPR87, IHPK2, KRT8, LOC120364, MYBL2, NA, NOL3, OSR2, PFAS, PKIA, RRM2, SCFD2, SH3BGRL, SIDT1, SPC24, TGM2, TMEM48, TYMS, ZWINT

U2.NNNNN,NNAV	CLIC3, EEF2K, FAM38A, FLJ13236, IHPK2, KRT8, LRP8, NOL3, OSR2, SNRPA, SPOCK1, ST3GAL1
U10.NNDDN,NNNN	ARHGAP36, F2RL1, GEMIN5, GFRA1, H2AFX, INSIG1, MCM5, MCM7, PEBP1, PRSS23
U9.NNDDN,NNNN	EMP2, F2RL1, GEMIN5, GFRA1, GINS2, H2AFX, HSPC111, INPP4B, INSIG1, MCM5, MCM7, PEBP1, PRSS23, RP13-102H20.1, TMEM46
U15.ININN,NNNN	EXT1, GBR2, GRB10, IRAK2, IRGQ, JMJD6, KIAA1609, MGLL, RASD1, TMEM45B, TTC9, TUBA4A, ULBP2
U12.NNNDN,NNNN	CCNB1, CDT1, CISD1, COX4NB, EPR1, KIF2C, MGP, POP1, PTTG1, RNASEH2A, SORD
U12.NNNNN,NNNA	ACTN1, APLP2, B3GNT5, BRAF, CALB2, CARS, CCNA1, CEBPB, CFLAR, CHST11, CLIC4, DUSP6, EDF1, EPS8L1, ERN1, ETFDH, FAAH2, FURIN, FYN, GART, GLIPR1, GLTSCR2, GPR37, HOXC9, HPCAL1, IK, KIAA1217, LAT2, LIPH, LRP10, MAP2K5, MORN4, MZF1, NA, NFYC, NMNAT2, NUP62CL, OSBPL6, PDLIM5, S10A9, SERPINB1, TIMP4, TPM4, TRIM15, TSC22D3, TUFT1
U16.INDDN,ANNN	CDC25A, CES2, DKC1, NAT10, SNORA9
U4.NNNNN,NNNA	ACTN1, APLP2, B3GNT5, BRAF, CALB2, CARS, CCNA1, CEBPB, CFLAR, CHST11, CLIC4, DUSP6, EDF1, EPS8L1, ERN1, ETFDH, FAAH2, FURIN, FYN, GALNTL1, GART, GLTSCR2, GPR37, HEY1, HLA-A, HLA-B, HOXC9, HPCAL1, IK, KIAA1217, LAMC2, LAT2, LIPH, LRP10, MAP2K5, MARCKS, MORN4, MZF1, NA, NFYC, NMNAT2, NUP62CL, OSBPL6, PAQR5, PDLIM5,PRPF6, PRSS22, PRGES, S100A9, SDC4, SERPINB1, SLC31A1, SPECC1, TBC1D2, TIMP4, TRIM15, TSC22D3
U16.NINNN,VNNN	CDH2, DSG2, EPAS1, FXYD5, IFNGR1, ITGB4, MARCKS, OSMR, SP100, STC1, TMEM165
U12.DNNNN,VNNN	ABCC5, ADCY1, DBP, ESPL1, FRAT2, GPER, HEY2, HOXC6, KCNJ8, LOC401052, MAFB, MYH11, NA, NR2F2, NUMA1, PCDH10, PRPF6, PSRC1, SALL2, SMAD6, SSTR5-AS1, TNRC18, ZMYM3
U8.NNNNN,NNNA	ACTN1, APLP2, B3GNT5, BRAF, CALB2, CFLAR, CHST11, CLIC4, CYSTM1, DUSP6, EDF1, EPS8L1, ETFDH, FURIN, FYN, GALNTL1, GART, GLTSCR2, GPR37, HEY1, HFAAH2, HLA-A, HLA-B, HOXC9, HPCAL1, IK, KIAA1217, LAMC2, LAT2, LIPH, LRP10, MAP2K5, MARCKS, MORN4, MZF1, NA, NFYC, NMNAT2, NUP62CL, OSBPL6, PAQR5, PDLIM5, PRPF6, PRSS22, PTGES, S100A9, SDC, SERPINB1, SLC31A1, SPECC1, TBC1D2, TIMP4, TMP4, TRIM15, TUFT1

Tab. A.3: Die nach der finite mixture models-Analyse signifikanten Cluster mit zugehörigen Genen

Cluster	Gene
MF_Pear_166	BIRC5, CUEDC2, SULT1A3, TK1, TMEM109
MF_SimGO_Pear_7	APEX1, ASF1B, AURKB, BIRC5, BUB3, CCNA1, CCNB1, CCND1, CDC25A, CDCA8, CDK4, CDKN2C, CDT1, CHAF1A, CHMP1B, CIT, CIZ1, CKS1B, CKS2, DCLRE1B, DDX11, EHMT2, FANCG, GADD45A, GAS2L1, GINS2, GTSE1, H1FX, H2AFV, H2AFX, H2AFY, H3F3B, HAUS4, HAUS7, HIRIP3, HIST1H1C, HIST1H2BD, HIST1H2BH, HIST3H2A, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, MCM2, MCM3, MCM5, MCM7, MCM9, MGMT, MUTYH, NCAPD2, NUMA1, OBFC1, PLK1, PMS2P1, POLD2, POLM, PPP2R2D, PSRC1, PZZG3P, RAB11FIP3, RAD51, RBMS1, RECQL4, REPIN1, RFC2, RNASEH2A, RRM2, RUVBL1, SAC3D1, SEPT10, SERTAD2, SPAG5, TIMELESS, TIPIN, TK1, TP53TG1, TREX1, TYMS, XRCC1, ZWINT
BP_SimGO_Pear_9	AURKB, BIRC5, BUB3, CCNA1, CCNB1, CCND1, CDC25A, CDCA8, CDKN2C, CIT, CKS1B, CKS2, GAS2L1, GTSE1, HAUS4, HAUS7, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, MLF1, NCAPD2, NUMA1, PLK1, PPP2R2D, PRNP, PSRC1, RBM38, RGCC, RPRM, SAC3D1, SEPT10, SERTAD2, SPAG5, TIMELESS, ZWINT
BP_SimGO_Pear_63	DNAL4, KIF21B, KTN1, POMP, TBCD, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5

Tab. A.4: Die nach der DP mixture models-Analyse signifikanten Cluster mit zugehörigen Genen

Cluster	Gene
Pear_41	CTSL2, EXOSC3, FABP5, HETR3, LGALSL, PDSS1, PRKAG2, QRSL1, TMEM185B
Bind_10	ACTB, ACTG1, ATP6V0E1, CSTB, EIF1, H3F3B, IER2, PSMC5, RAN, S100A10, TMSB10, TUBA1B, TUBA1C, TUBB4B
CC_Bind_95	ALDH3A2, ARHGDIB, ASPH, ATP2B1, BRAF, CFLAR, CLDN1, CORO2B, DEFB1, DLAT, DUSP6, EHBP1L1, ETFDH, FGB, FGG, FLRT1, GPR37, GULP1, ITGA6, LGALS8, MARK14, NA, NAB1, NFYC, PLD1, RIPK1, SAMD4A, SLC1A4, SYNPO, TIMP4, TRIB2, TRIO, VEGFA

Bind_502	ESPL1, HIST3H2A, HOXB7, NA, RPS15A, TIMELESS
CC_Bind_317	ACSF2, CCNF, CES2, FAM64A, HADH, INSIG1, PSRC1, RARA, SOX12
MF_Bind_308	BMP7, CDCA8, CDT1, CES2, COL18A1, HADH, HOXC6, INSIG1, MSX2, RAD51, RARA, RET, SLC29A3, SMAD6, ZNF32
MF_Bind_354	CRTAP, FAM64A, HPSE, IGFBP5, PSRC1
BP_SimGO_Bind_44	AURKB, BIRC5, BUB3, CCNB1, CDC20, CDCA8, CIT, CKS2, ESPL1, HAUS4, HAUS7, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC3D1, SEPT10, SPAG5, ZWINT
BP_SimGO_Bind_79	CDKN2C, E2F5, MLF1, MYBL2, RPRM
BP_SimGO_Bind_111	B3GNT3, B4GALT5, GALNT2, GALNT6, MUC5AC, POMT1, ST3GAL1
BP_SimGO_Bind_146	KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5
BP_SimGO_Bind_262	DNAL4, KIF18B, KIF22, KIF2C, KIFC1
CC_SimGO_Bind_43	AURKB, BIRC5, BUB3, CCNB1, CDC20, CDCA8, CIT, ESPL1, HAUS4, HAUS7, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC3D1, SEPT10, SPAG5, ZWINT
CC_SimGO_Bind_144	KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5
CC_SimGO_Bind_79	CDKN2C, E2F5, MLF1, MYBL2, RPRM
CC_SimGO_Bind_112	B3GNT3, B4GALT5, GALNT2, GALNT6, MUC5AC, POMT1, ST3GAL1
CC_SimGO_Bind_259	DNAL4, KIF18B, KIF22, KIF2C, KIFC1
MF_SimGO_Bind_37	AURKB, BIRC5, CDC20, CDCA8, CHMP1B, CIT, CKS2, HAUS4, HAUS7, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC3D1, SEPT10, SPAG5, ZWINT
MF_SimGO_Bind_75	SPRR1B, SFN, PPL, KRT16, SPRR1A

MF_SimGO_Bind_94	RRS1, MPHOSPH6, KIF22, KIF2C, KIFC1, KIF18B
MF_SimGO_Bind_106	B3GNT3, B4GALT5, GALNT2, GALNT6, MUC5AC, POMT1, ST3GAL1
MF_SimGO_Bind_138	KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5
MF_SimGO_Pear_668	ABLIM1, ACOT9, ACSL3, AGA, AHR, AKAP12, AKAP13, APLP2, AQP3, ARHGDIB, ASCL1, ASPH, ATP2B1, BDKRB2, CALU, CAMSAP1, CAP1, CASP4, CD14, CEACAM1, CFLAR, CHAC1, CHIC2, CHMP1B, CHST12, CLIC4, COL1A1, COMT, CORO2B, CPE, CSNK2A1, CST6, CTNNAL1, CYP1B1, DENND1A, DHRS3, DLAT, DUSP4, DUSP5, DUSP6, DUSP13, EDEM1, EGFR, EGR4, ENTPD4, EPHA2, EPS8L1, EXT1, FAM129A, FGB, FGD6, FLRT1, FSTL3, FXYD5, GADD45A, GBF1, GCH1, GCLM, GEM, GFPT1, GFPT2, GLRX, GPR37, GPR87, GPX3, GRB10, GULP1, H2AFY, HBEGF, HERPUD1, HEY1, HLA-C, HOMER2, HPGD, HSPB8, IDH3A, IFNGR1, IFNGR2, IGHG1, IL11, IL15RA, IMPAD1, INPP1, ITGA2, ITGA5, JAG1, JUN, KCNG1, KIF21B, KLF2, KLF6, KLK6, KPNA1, KRT17, KTN1, LAMA3, LAT2, LCP1, LIMA1, LIPG, LMO4, LONRF3, LOXL2, MAK16, MAP1LC3B, MAP2K5, MAP4K4, MAP7, MAPRE1, MARCKS, MCL1, MET, MMD, MMP9, MPZL2, NFKB2, NPC1, NPTN, NRP1, NT5E, OCRL, OGT, OPN3, OSMR, P2RX5, P4HA2, PAPOLA, PCSK6, PDCD11, PDIA6, PDLIM5, PEX13, PI3, PIGH, PLAUR, PLEKHG6, PLSCR1, PLXNA2, PPARG, PPL, PPP2CB, PPP4R1, PRKCSH, PRBP, PRPF3, PRPS1, PTGES, QPCT, RAB27A, RBFOX2, RBM28, RBMS1, RIPK1, RPS6KA3, S100A2, S100A7, SAT1, SC5DL, SEMA3B, SERPINB1, SERPINB8, SERPINE2, SERTAD2, SGPL1, SH2D3A, SHQ1, SLC1A4, SLC20A1, SLC22A4, SLC31A1, SLPI, SMURF1, SNRPA1, SP100, SPINK1, SPOCK1, SPRED2, SPRR1B, SPTAN1, SPTLC2, ST3GAL1, STAT3, STC1, STEAP1, STK39, STX18, SYNPO, TARS, TBC1D2, TBC1D8, TES, TGFBR2, TGM2, TMPRSS4, TNFAIP8, TNFRSF21, TPM4, TRIB1, TRIB2, TRIP10, TSC22D3, TUBA4A, TUSC3, TXNRD1, UBAP2L, UBE2K, ULBP2, UMPS, UPP1, VDR, VEGFA, VRK3, WWTR1, YKT6, ZFAND5, ZNF202, ZNF862

Tab. A.5: Die nach der DIRECT-Analyse signifikanten Cluster mit zugehörigen Genen

Cluster	Gene
MF_SimGO_Bind_29	AURKB, BIRC5, BUB3, CCNB1, CDCA8, CHMP1B, CIT, CKS2, HAUS4, HAUS7, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC2D1, SEPT10, SPAG5, TIMELESS, ZWINT

MF_SimGO_Bind_36	ASF1B, CHAF1A, DDX11, H1FX, H2AFV, H2AFX, H2AFy, H3F3B, HIRIP3, HIST1H1C, HIST1H2BD, HIST1H2BH, HIST3H2A, MCM2, PTTG3P, RAD51, RUVBL1
MF_SimGO_Bind_94	MCM9, RBMS1, RECQL4, REPIN1, RNASEH2A, RRM2, TK1, TREX1, TYMS
MF_SimGO_Bind_111	KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5
MF_SimGO_Bind_134	ANXA9, CD44, FLRT1, ICAM3, MPZL2, NPTN
CC_SimGO_Bind_34	AURKB, BIRC5, BUB3, CCNB1, CDCA8, CIT, HAUS4, HAUS7, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC3D1, SEPT10, SPAG5, TIMELESS, ZWINT
CC_SimGO_Bind_118	DNAL4, KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5
CC_SimGO_Pear_8	APEX1, AURKB, BIRC5, BUB3, CCNA1, CCNB1, CCND1, CDC25A, CDCA8, CDKN2C, CDT1, CENPM, CHAF1A, CIT, CIZ1, CKS1B, DCLRE1B, EHMT2, FANCG, GAS2L1, GINS2, GINS3, GTF3C4, GTSE1, H1FX, H2AXV, H2AFx, H2AFY, H3F3B, HAU4, HAUS7, HIRIP3, HIST1H1C, HIST1H2BD, HIST1H2BH, HIST3H2A, HMG20B, HMGN2, HMGN3, INTS3, KDM6B, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, MCM2, MCM3, MCM5, MCM7, MGMT, MLF1, MUTYH, NCAPD2, NUMA1, NUP160, NUP85, NYNRIN, OBFC1, OGT, PHF15, PLK1, POLD2, PPP2R2D, PRNP, PSRC1, PTTG3P, RAD51, RBM38, RBMS1, RECQL4, REPIN1, RFC2, RGCC, RPRM, RRM2, RUVBL1, SAC3D1, SEH1L, SEPT10, SERTAD2, SETMAR, SLC2A8, SLC37A4, SPAG5, TIMELESS, TIPIN, TK1, TMEM48, TP53TG1, TREX1, TYMS, XRCC1, ZWINT
CC_SimGO_Pear_22	ABHD2, ACTN1, BLNK, C1QBP, CALU, CAP1, CD55, CD59, CLCF1, CLU, FGB, FGG, FLNA, GATA3, HERC5, HLA-A, HLA-B, HLA-C, HLA-DMA, IFITM1, IFITM2, IFNGR1, IFNGR2, IGHG1, KLF6, KLK6, MAPK14, NFKB2, PFN1, PLLP, PMS2P1, POLM, POTEKP, PRCP, S100A14, S100A7, S100A8, S100A9, SDC2, TFPI, TFPI2, THBD, TIMP1, TNFSF13, TRAF3, ULBP2, VCL, WDR1
BP_SimGO_Bind_9	CCNA1, CCND1, CDC25A, CDKN2C, CKS1B, CKS2, RGCC
BP_SimGO_Bind_34	AURKB, BIRC5, BUB3, CCNB1, CDCA8, CIT, HAUS4, HAUS7, KIF18B, KIF22, KIF2C, KIFC1, MAPRE1, NCAPD2, PLK1, PPP2R2D, SAC3D1, SEPT10, SPAG5, TIMELESS, ZWINT
BP_SimGO_Bind_119	DNAL4, KIF21B, KTN1, TUBA1A, TUBA1B, TUBA1C, TUBA4A, TUBB2A, TUBB3, TUBB4B, TUBB6, TUBBP5

Tab. A.6: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – MF)

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**12310,58**	7	12519,47	12	13240,06
3	12418,38	8	12924,75	13	13552,15
4	13467,91	9	13008,92	14	13106,66
5	12946,26	10	13176,21	15	13211,57
6	13122,55	11	13316,79		

Tab. A.7: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – BP)

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	12469,39	7	13324,27	12	12607,58
3	12490,99	8	13105,22	13	13141,42
4	12512,73	9	13084,04	14	12783,57
5	**12442,96**	10	12581,97	15	13291,64
6	12891,16	11	13418,26		

Tab. A.8: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – CC)

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**13218,29**	7	14197,53	12	14401,89
3	13952,30	8	14313,44	13	14321,26
4	13280,00	9	14285,75	14	14255,53
5	14378,07	10	13738,9	15	14314,69
6	14315,51	11	14156,15		

Tab. A.9: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (ohne Berücksichtigung der GO-Zuordnung) für die AAS-Daten

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**3932,74**	7	4031,75	12	4108,08
3	3948,62	8	4053,92	13	4065,28
4	3975,09	9	4069,79	14	4088,34
5	3994,98	10	4097,05	15	4097,39
6	4015,14	11	4117,96		

Tab. A.10: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – MF) für die AAS-Daten

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**3771,97**	7	3828,65	12	3898,34
3	3798,94	8	3847,54	13	3911,98
4	3811,59	9	3856,48	14	3960,27
5	3833,06	10	3873,16	15	3935,07
6	3880,41	11	3937,40		

Tab. A.11: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – BP) für die AAS-Daten

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**3750,78**	7	3854,18	12	3900,19
3	3769,42	8	3823,79	13	3888,69
4	3791,96	9	3857,68	14	3900,29
5	3809,32	10	3849,69	15	3916,48
6	3826,18	11	3857,09		

Tab. A.12: Übersicht über die BIC's zur Festlegung der optimalen Komponentenanzahl *K* (mit Berücksichtigung der GO-Zuordnung – CC) für die AAS-Daten

K	**BIC**	***K***	**BIC**	***K***	**BIC**
2	**3078,21**	7	3236,33	12	3201,09
3	3098,85	8	3286,40	13	3204,73
4	3120,13	9	3154,68	14	3289,01
5	3196,14	10	3171,71	15	3255,40
6	3162,11	11	3183,71		

Anhang B

Zusätzliche Abbildungen

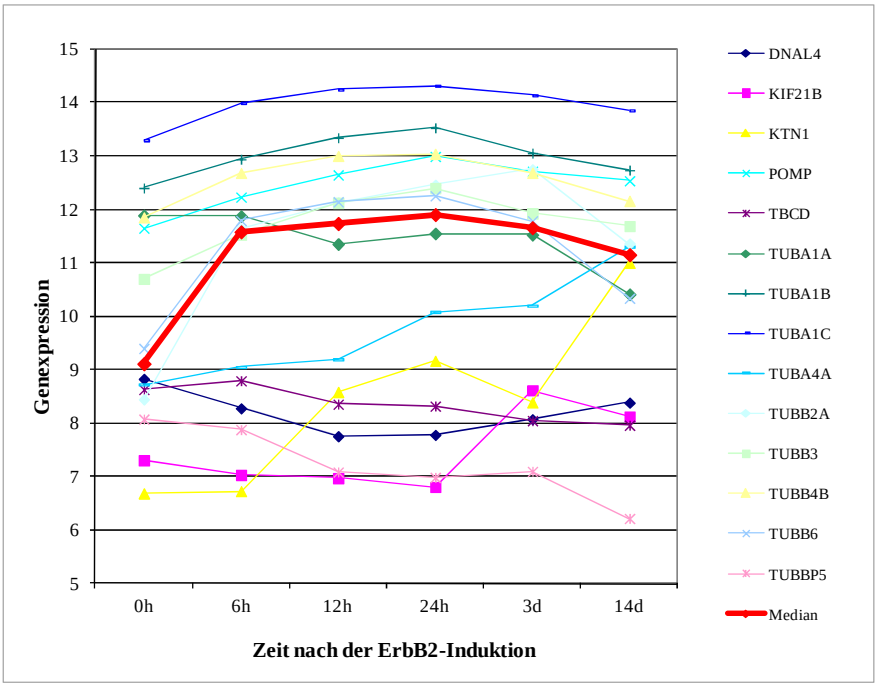

Abb. B.1: Zeitliche Genexpressionsverläufe im Cluster „BP_SimGO_Pear_63” (finite mixture models)

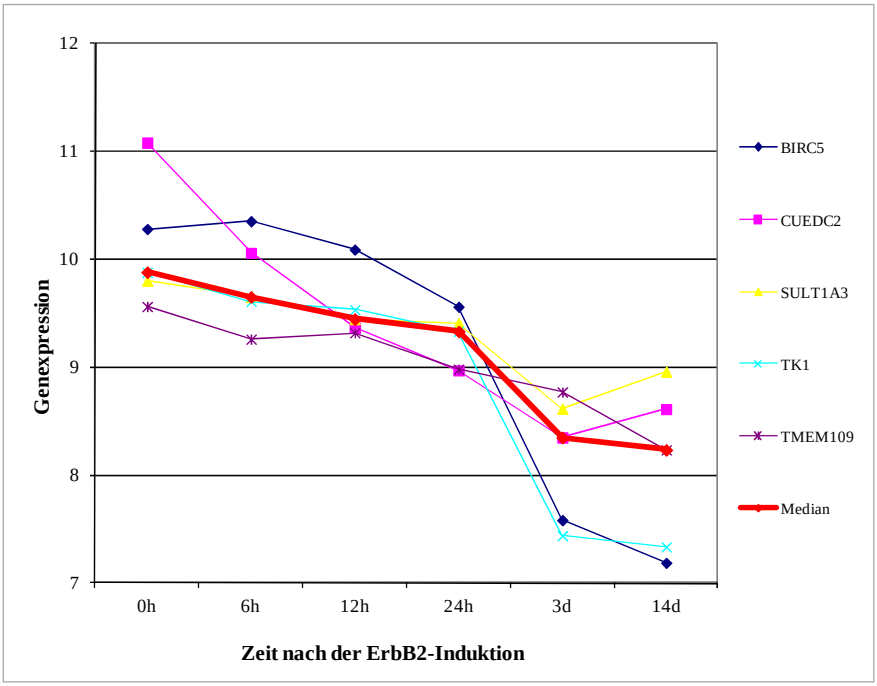

Abb. B.2: Zeitliche Genexpressionsverläufe im Cluster „MF_Pear_166” (finite mixture models)

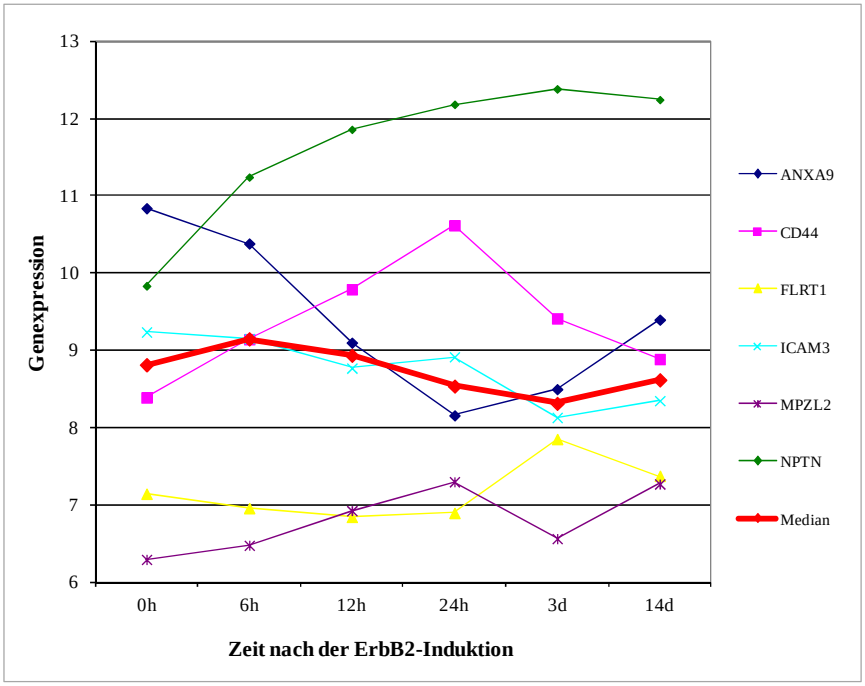

Abb. B.3: Zeitliche Genexpressionsverläufe im Cluster „MF_SimGO_Bind_134” (DIRECT)

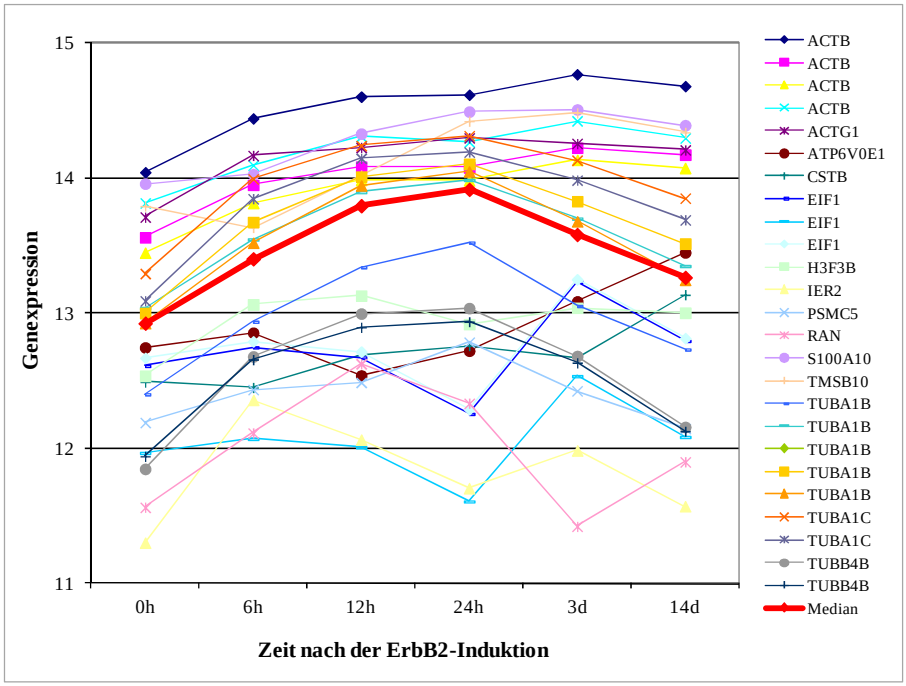

Abb. B.4: Zeitliche Genexpressionsverläufe im Cluster „Bind_10" (DP mixture models)

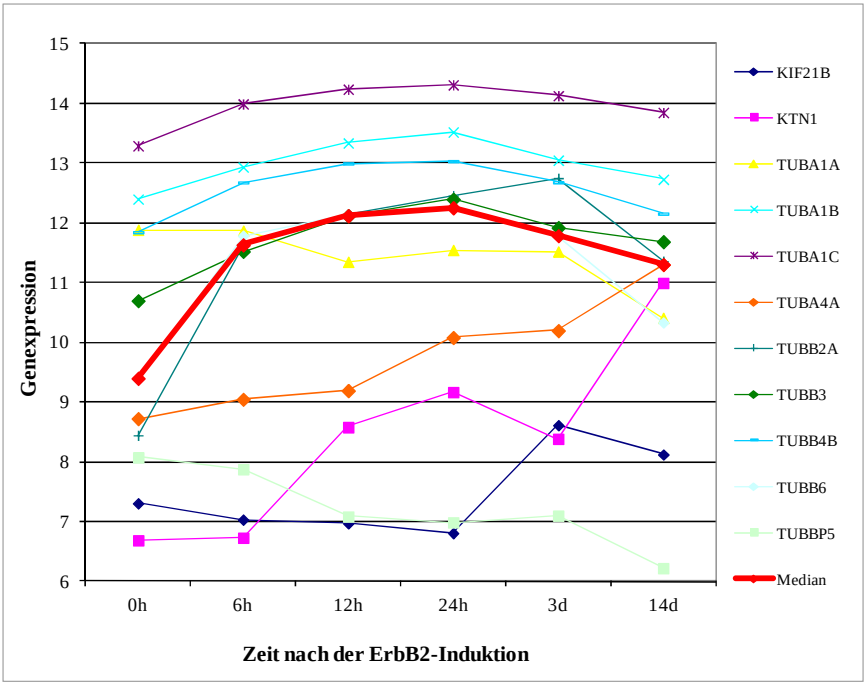

Abb. B.5: Zeitliche Genexpressionsverläufe im Cluster „BP_SimGO_Bind_146" (DP mixture models)

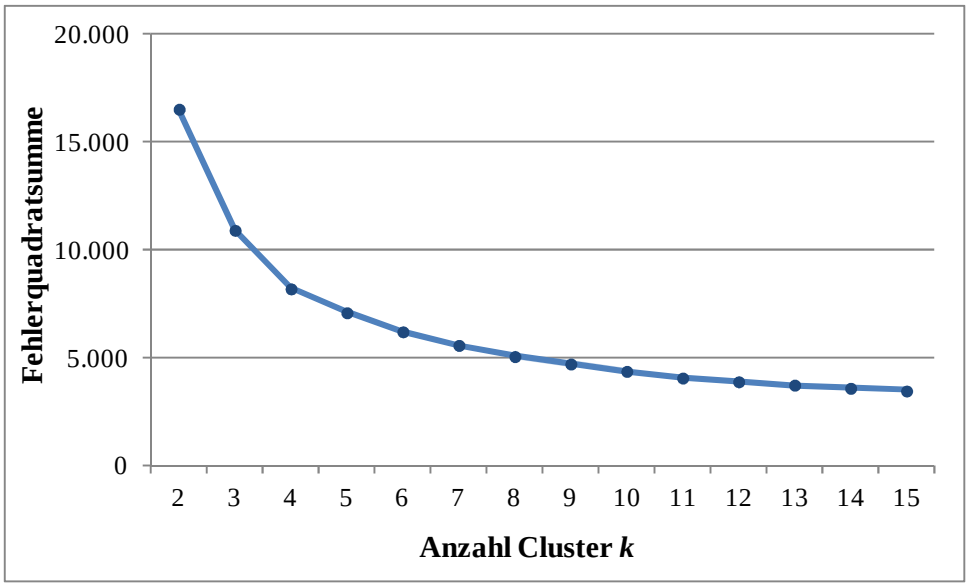

Abb. B.6: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – CC)

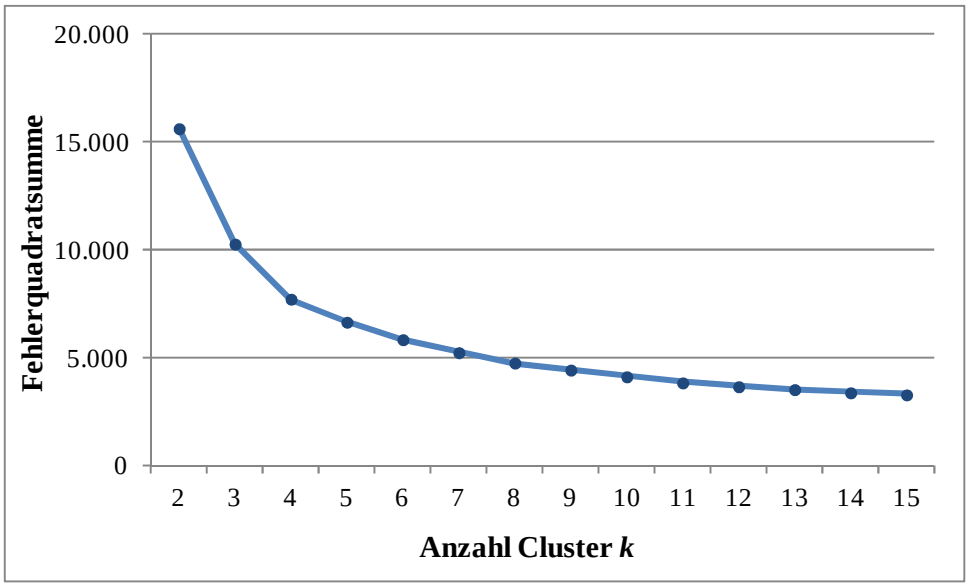

Abb. B.7: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – MF)

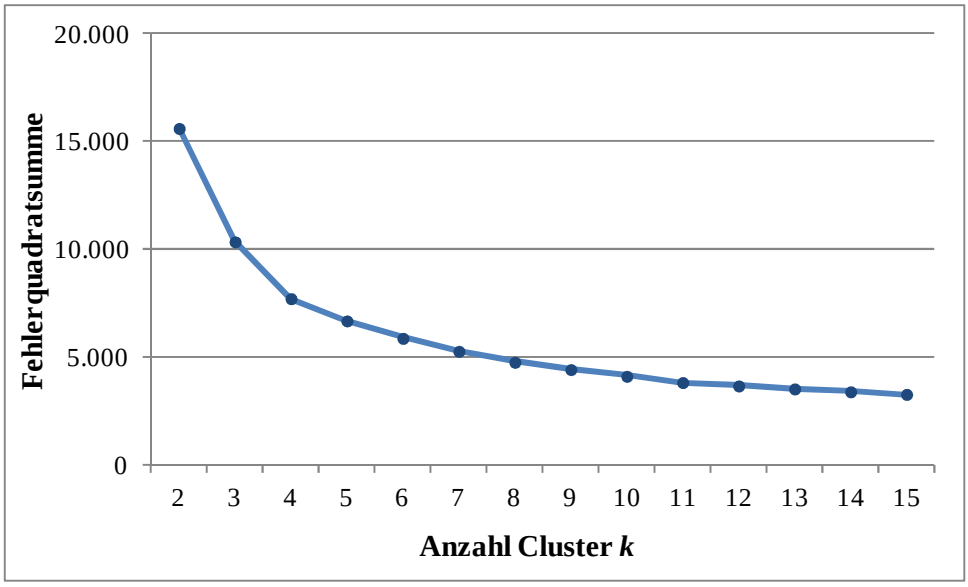

Abb. B.8: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – BP)

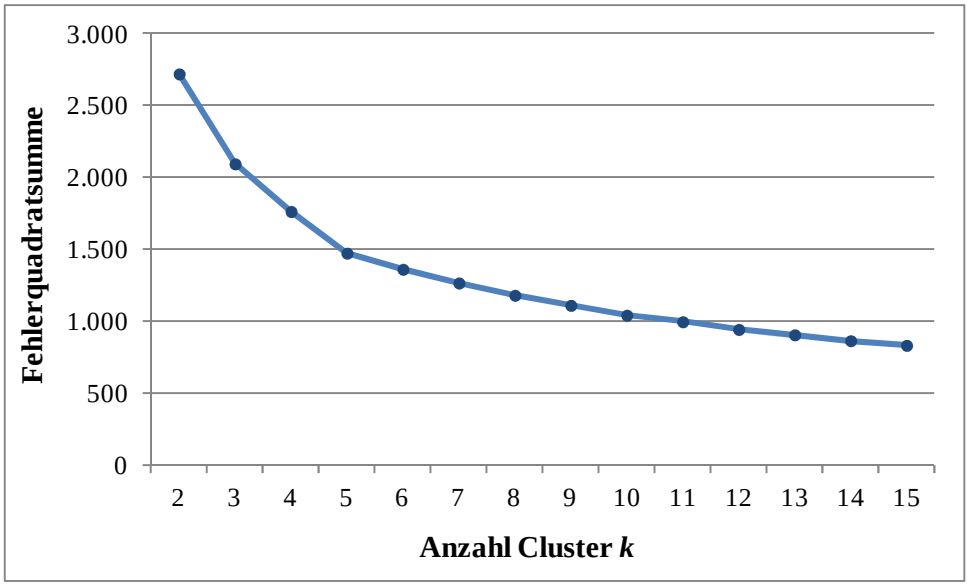

Abb. B.9: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (ohne Berücksichtigung der GO-Zuordnung)

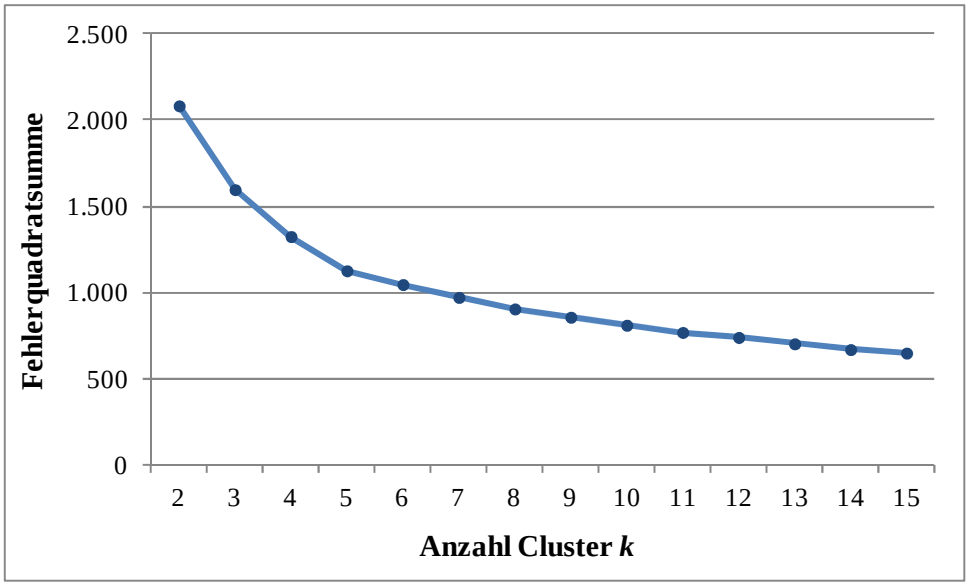

Abb. B.10: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – CC)

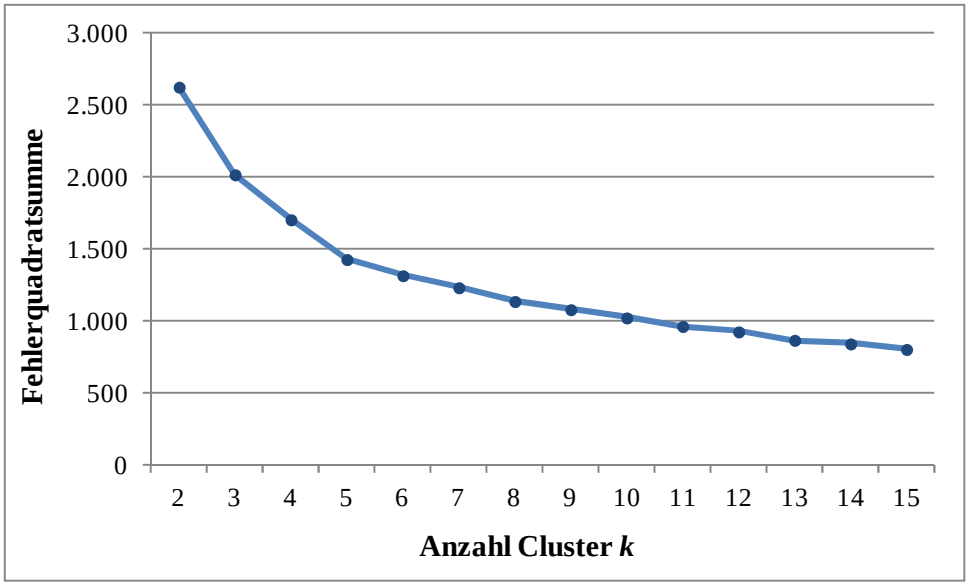

Abb. B.11: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – MF)

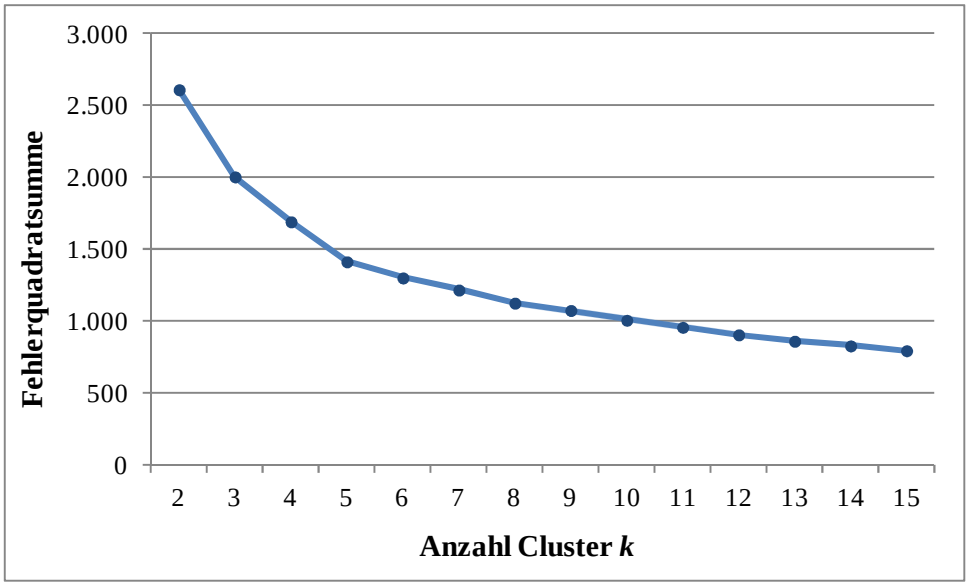

Abb. B.12: Fehlerquadratsumme in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – BP)

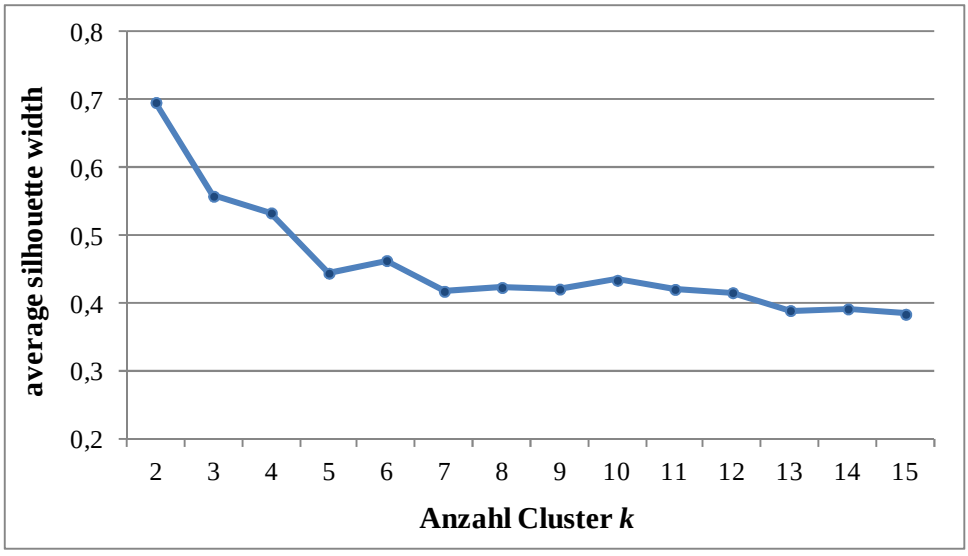

Abb. B.13: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – CC)

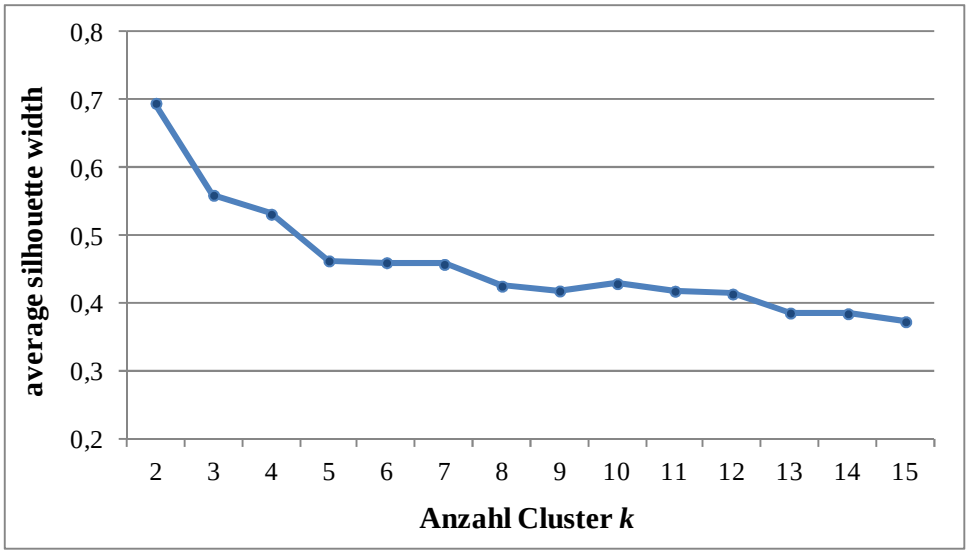

Abb. B.14: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – MF)

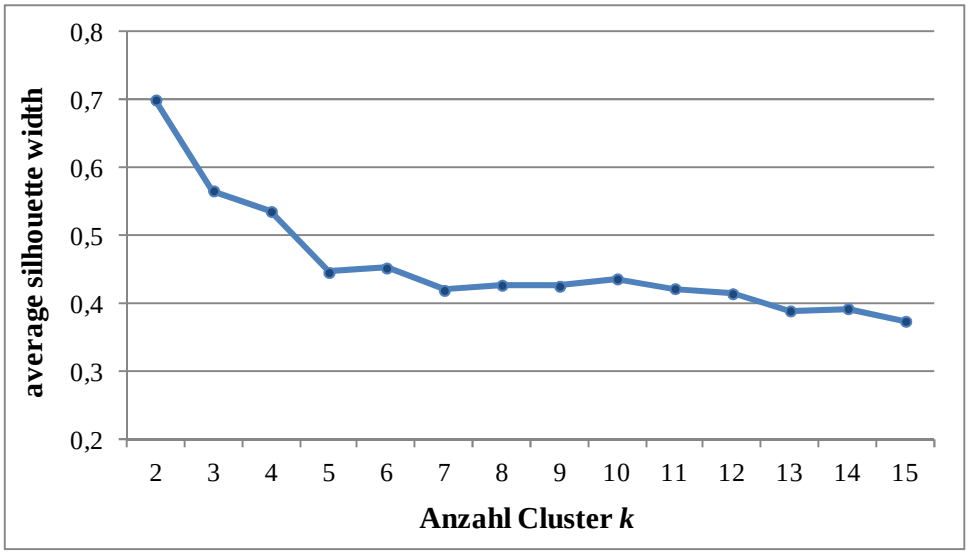

Abb. B.15: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die MCF7-Daten (mit Berücksichtigung der GO-Zuordnung – BP)

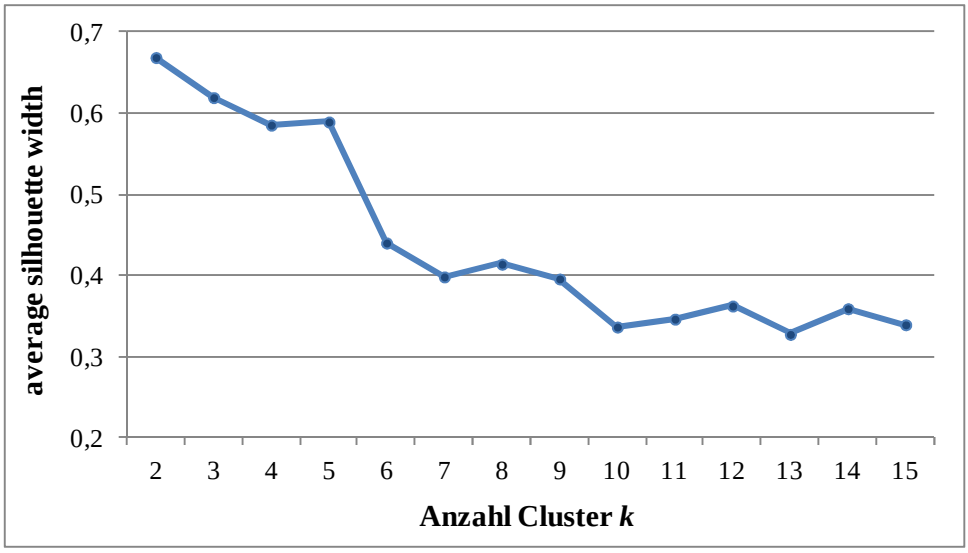

Abb. B.16: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (ohne Berücksichtigung der GO-Zuordnung)

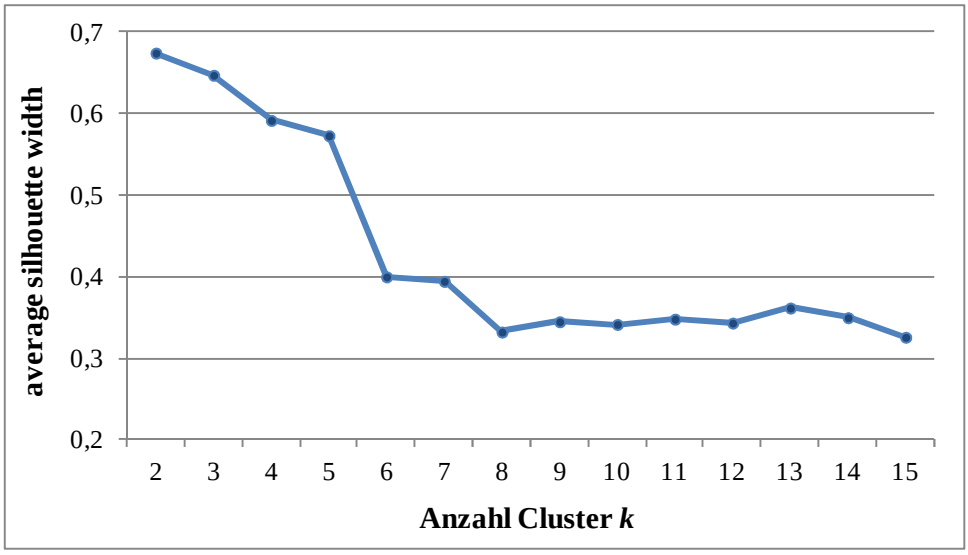

Abb. B.17: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – CC)

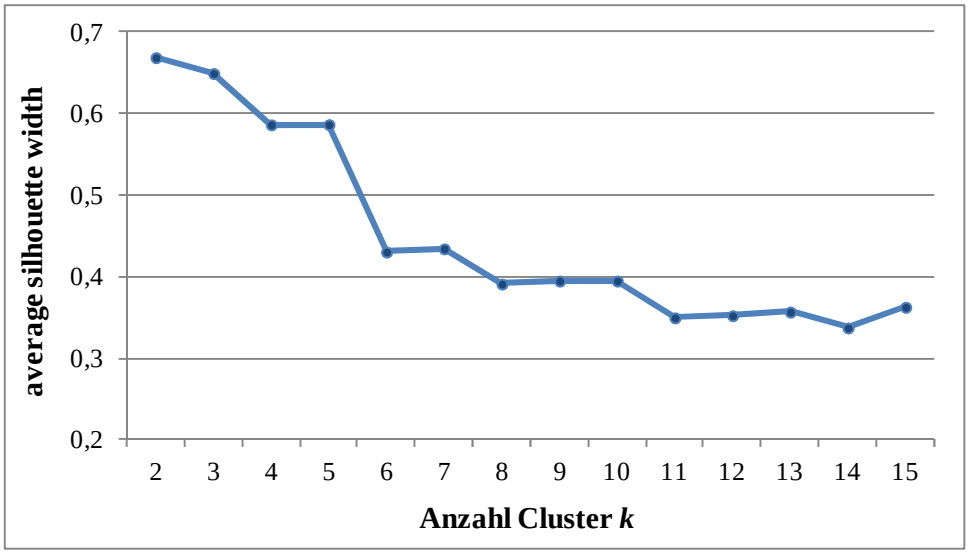

Abb. B.18: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – MF)

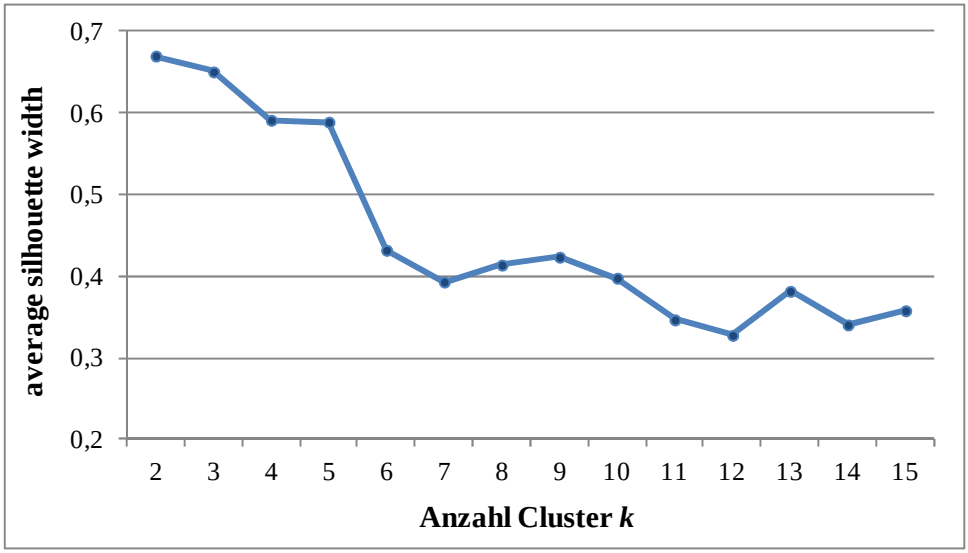

Abb. B.19: Die durchschnittlichen Silhouetten-Werte in Abhängigkeit von der Clusteranzahl *k* für die AAS-Daten (mit Berücksichtigung der GO-Zuordnung – BP)

Anhang C

Zusätzliche Ergebnisse

C.1 SRF-Cluster

Das SRF-Cluster weist eine überrepräsentierte TF-Bindungsstelle für SRF auf. Es ist signifikant sowohl in der Gesamtkohorte als auch in den einzelnen Kohorten Mainz und Rotterdam (vgl. Tab. C.1.1), jedoch nicht adjustiert für multiples Testen.

Tab. C.1 1: Signifikanz des SRF-Clusters in den einzelnen Kohorten

	p	HR_{GK}	Mainz	Transbig	Rotterdam		GO	TF
U10.INDNN,ANNN	0,001	0,70	x	-	x	a)	-	x

In diesem Cluster kommt es nach 6 Stunden nach dem Einschalten der ErbB2-Expression zum signifikanten Anstieg und nach 24 Stunden zur Abnahme der Genexpressionen. Zu weiteren Zeitpunkten ist kein signifikanter Unterschied nachweisbar.

Das Cluster umfasst sieben Gene – EGR1, EGR2, EZR, KAZN, TMPRSS3 und zwei weitere, die zum jetzigen Zeitpunkt noch nicht eindeutig zugeordnet werden können – und ist mit einem HR von 0,70 mit einer besseren Prognose bei den Brustkrebspatientinnen assoziiert. Der zugehörige zeitliche Verlauf ist in der Abbildung C.1.2 dargestellt:

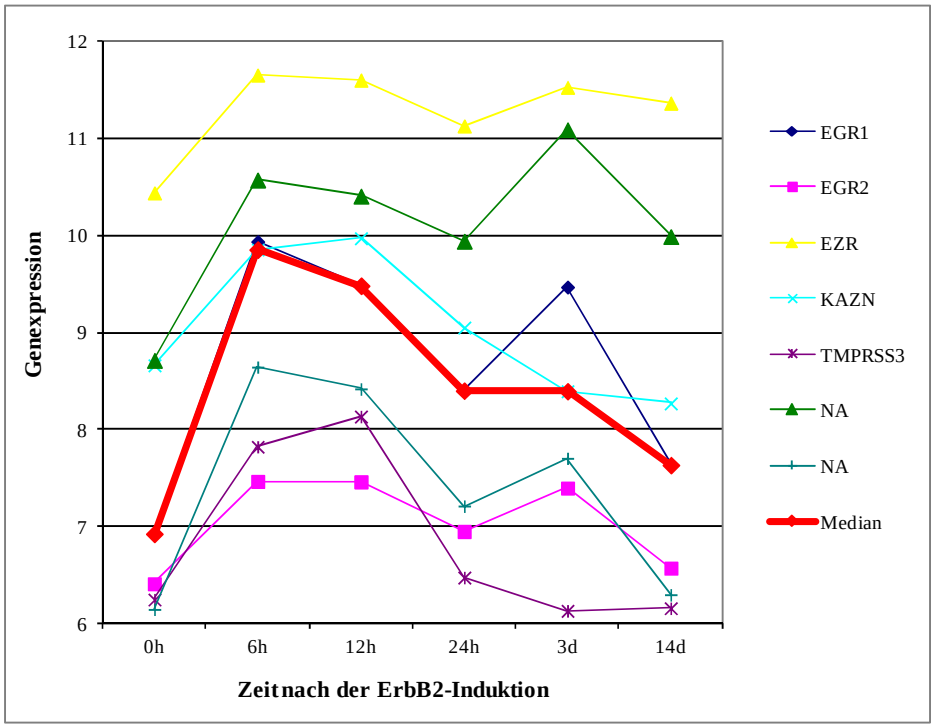

Abb. C.1.2: Zeitliche Genexpressionsverläufe im SRF-Cluster „U10.INDNN,ANNN"

SRF wird durch viele Signaltransduktionswege aktiviert, auch durch ErbB2 (oder seine onkogene Variante NeuT). Und da in den in dieser Arbeit vorgestellten MCF7-Zellen ErbB2 aktiviert wurde, scheint es plausibel zu sein, dass SRF als überrepräsentierter TF nachgewiesen wird.

Auch die Richtung der Prognose ist plausibel, soweit es die bekannten Funktionen der Gene zulassen. EGR1 ist ein Transkriptionsfaktor, der auch als Tumorsuppressor fungiert. EGR2 vermittelt die Apoptose (den programmierten Zelltod), und eine hohe Apoptose-Aktivität vermittelt bessere Prognose. Die Gene EZR und KAZN tragen zur

Zelladhäsion bei und Zellen mit ausgeprägter Adhäsion führen weniger wahrscheinlich zu Metastasen. Die Rolle der Serinprotease TMPRSS3 bei der Tumorentwicklung ist noch nicht geklärt.

Zusammenfassend kann festgestellt werden, dass dieses Cluster mit EGR1 und EGR2 zwei bekannte, durch SRF regulierte Gene enthält. Die Genfunktionen der im SRF-Cluster vertretenen Faktoren lassen eine Assoziation mit besserer Prognose plausibel erscheinen. Zur Überprüfung der in vivo Relevanz wäre es von Interesse, ob diese Gene in den ErbB2 positiven Mammakarzinomen stärker exprimiert sind als in den ErbB2 negativen Tumoren.

C.2 PFP-Auswertung unter Berücksichtigung der Gene Ontology-Zuordnung

Eine zusätzliche Auswertung der Daten mit dem PFP-Verfahren unter Berücksichtigung des biologischen Hintergrundwissens ergab 4 weitere Cluster, die in der Gesamtkohorte adjustiert und in mindestens zwei Einzelkohorten unadjustiert signifikant sind. Dies ist folgender Ergebnisaufstellung zu entnehmen:

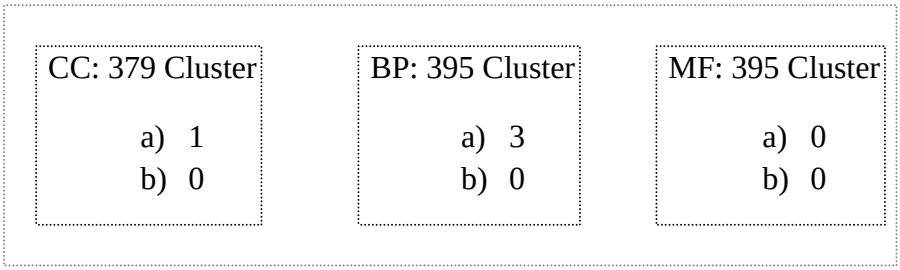

Abb. C.2.1: Ergebnisse einer zusätzlichen PFP-Analyse der nach GO-Gruppen getrennten Daten: Anzahl signifikanter Cluster je Erfolgsszenario

Tabellen C.2.2 bzw. C.2.3 enthalten die zugehörigen Einzelheiten bzw. Genauflistungen:

Tab. C.2.2: Die nach der zusätzlichen PFP-Analyse (unter Berücksichtigung der GO-Zuordnung) signifikanten Cluster mit Signifikanzangaben je Kohorte

	p_{adj}	HR_{GK}	Mainz (kor)	Transbig (kor)	Rotterdam (kor)	
alpha2_m65_BP_36	0,013	2,01	x	-	x	a)
alpha0.2_m50_CC_3	0,001	0,35	x (x)	-	x (x)	a)
alpha0.2_m50_BP_10	0,002	2,13	x	-	x	a)
alpha0.2_m50_BP_21	0,013	0,47	x	-	x	a)

Tab. C.2.3: Die nach der zusätzlichen PFP-Analyse (unter Berücksichtigung der GO-Zuordnung) signifikanten Cluster mit zugehörigen Genen

Cluster	Gene
alpha2_m65_BP_36	ACTR3B, ADCY3, CLCF1, CTGF, EIF3B, ELK4, EMP2, EN2, GADD45A, GCFC2, HNRNPD, MCAM, NAT10, NOP56, NUP160, PDCD11, RFC2, SETMAR, SCL25A15, TEX2, URB1, WASF3
alpha0.2_m50_CC_3	AARS, CARS, CHAC1, COL1A1, DEFB1, DHRS3, FKBP8, FLRT1, FTH1, GLRX, HERPUD1, HEY1, HLA-B, HLA-C, IGBP1, KCNG1, KIF21B, MZF1, NFE2L1, PFXIP1, PTGES, SGCG, TBC1D2, TXNIP, UBC, ZNF862
alpha0.2_m50_BP_10	ADSL, APOBEC3B, ARL4C, ASF1B, ATIC, BIRC5, BMP7, C1QBP, CA2, CCNB1, CCNF, CDC25A, CDCA8, CES2, CHPT1, CISD1, CKS1B, CKS2, CUTC, CYB5A, DDX39A, DKC1, DPH5, DTYMK, EIF2D, EIF3B, ESR1, EXOSC2, FBL, FGFR1, FGGY, FKBP4, GFRA1, GINS2, H2AFX, HADH, HAUS7, HMGN3, HNRNPA1, IGF1R, IGSF1, INSIG1, INTS3, KIF18B, LAS1L, LOXL1, MARC1, MARC2, MCM3, MGP, MRP63, MRPL24, MRTO4, MSX2, MTHFD1, NA, NAT10, NDUFA10, NOL6, NOP2, NOP56, NPM3, PASK, PCCB, PDCD11, PEBP1, PFAS, POLD2, POLR3K, PPIH, PRPF19, PTDSS1, PTTG1, RAD51, RFC2, RNASEH2A, RPL24, RPS21, RRS21, RRS1, RUVBL1, SCARA3, SDC2, SF3B3, SH3BP5, SORD, STMN1, TATDN2, TCTN3, TH1L, TIMM8B, TMEM97, TRMT2B, TXNL4A, TYMS, UBE2C, WDR77, ZNF32, ZNF629, ZWINT
alpha0.2_m50_BP_21	C8orf4, CEACAM1, CEACAM6, CFLAR, CHMP1B, CPE, CST6, DDIT4, DUSP13, EXT1, FGD6, FXYD5, GPX3, GRB10, HLA-A, HLA-B, IK, ITGA5, ITGB4, KRT17, MARCKS, OGT, OSMR, PLD1, PLIN2, PNPLA6, PPL, PRSS22, PTGES, S100P, SP100, TFPI, TRIP10, TUBA4A, UBC, ULBP2, VEGFA, VRK3

Es fällt auf, dass in der Transbig-Kohorte keines dieser Cluster signifikant ist. Weiterhin ist zu bemerken, dass drei von vier Gengruppen mit einem sehr kleinen α = 0,2 und m = 50 möglichen Modellprofilen identifiziert wurden. Diese Ergebnisse sind jedoch mit Vorsicht zu betrachten, da durch das alpha das Verhältnis zwischen dem datenabhängigem und -unabhängigem Teil des Minimierungsproblems gesteuert wird, so dass ein zu kleines alpha eventuell die Datenstruktur zu wenig in Betrachtung zieht.

In Abbildungen C.2.4 bzw. C.2.5 sind beispielhaft Genexpressionsverläufe über die Zeit von zwei dieser Cluster dargestellt.

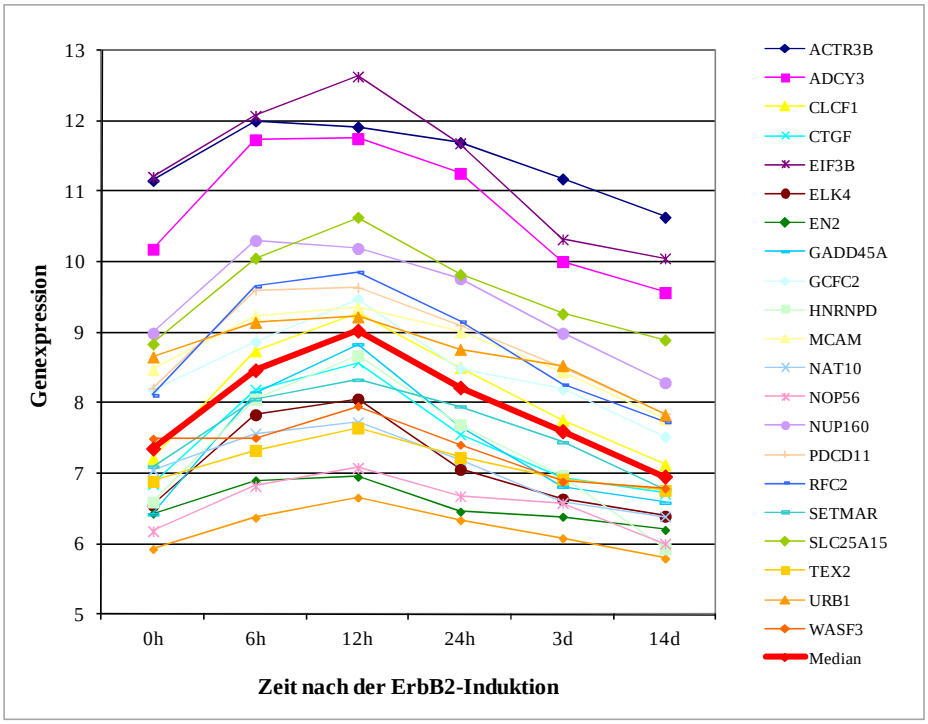

Abb. C.2.4: Zeitliche Genexpressionsverläufe im PFP-Cluster „alpha2_m65_BP_36” nach der zusätzlichen Analyse

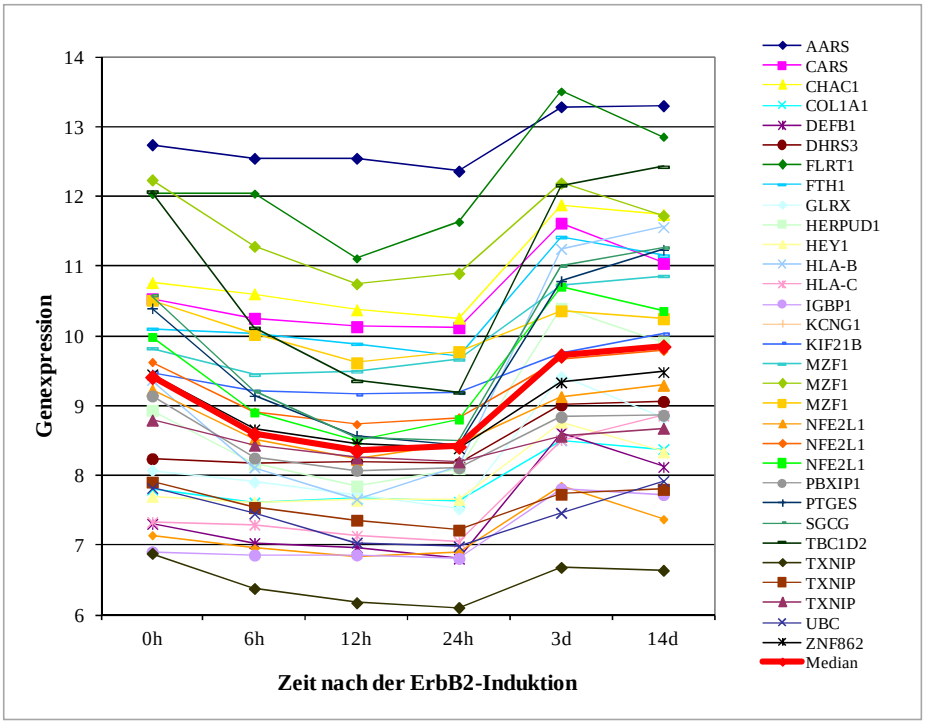

Abb. C.2.5: Zeitliche Genexpressionsverläufe im PFP-Cluster „alpha0.2_m50_CC_3" nach der zusätzlichen Analyse

Daraus ist ersichtlich, dass im Vergleich zur Auswertung ohne GO-Zuordnung (vgl. Unterkapitel 5.1.4) hier deutlich homogenere Verlaufsmuster in den Genexpressionen beobachtet werden können.

Anhang D

Zusätzliche methodische Einzelheiten

D.1 Iterationsschritte in *k*-means

Der *k*-means Algorithmus durchläuft folgende Iterationsschritte:

1. Im ersten Schritt des Verfahrens wird die Clusteranzahl *k* bestimmt bzw. vorgegeben.
2. Die *k* Clusterzentren werden entweder zufällig gesetzt oder anderweitig vorgegeben.
3. In diesem Schritt wird jede Beobachtung demjenigen Cluster zugeordnet, zu dessen Zentrum sie den kleinsten euklidischen Abstand hat.
4. Danach werden alle Clusterzentren neu berechnet.
5. Anschließend werden Schritte 3 und 4 bis zur Konvergenz wiederholt.

D.2 MCMC Methoden

Im Gegensatz zu der klassischen frequentistischen Statistik, stellt der unbekannte Parameter θ bei der Bayesianischen Vorgehensweise formal eine Zufallsvariable dar, da die verfügbare Information darüber durch eine a priori-Verteilung $p(\theta)$ ausgedrückt wird. Nach dem Satz von Bayes gilt dann für die a posteriori-Verteilung

$$p(\theta \mid y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y \mid \theta)\, p(\theta)}{p(y)},$$

wobei $p(y \mid \theta)$ die Likelihood-Funktion ist und $p(y)$ eine Normalisierungskonstante darstellt. Für $p(y)$ gilt dabei:

$$p(y) = \int p(y, \theta) d\theta$$

mit $p(y, \theta)$ als der gemeinsamen Verteilung von *y* und θ.

Die a priori-Verteilung $p(\theta)$ wird vor der Anwendung fest vorgegeben und kann sowohl informativ als auch nicht-informativ (also wenn keine Vorinformation über θ vorliegt) gewählt werden. Falls die a posteriori-Verteilung $p(\theta \mid y)$ derselben Familie von Verteilungen angehört wie die a priori-Verteilung $p(\theta)$, wird von einer konjugierten a priori-Verteilung gesprochen. In diesem Fall stellt die Bestimmung der a posteriori-Verteilung kein Problem dar, da die a priori-Verteilung und die Likelihood $p(y \mid \theta)$ von derselben Form sind (vgl. z.B. Gelman et al., 2004).

Wenn die Lösung des Integrals in $p(y)$ analytisch nicht möglich ist, muss die a posteriori-Verteilung bei der Anpassung der Bayes-Modelle an die Daten entweder approximiert werden, z.B. durch eine Normalverteilung oder durch die so genannten Markov Chain Monte Carlo Methoden inferriert werden. Mit Hilfe von MCMC Methoden werden Stichproben gemäß der a posteriori-Verteilung erzeugt und diese MCMC Stichproben werden benutzt, um die a posteriori-Verteilung zu approximieren.

MCMC Methoden umfassen mehrere Verfahren zur Erzeugung von Stichproben gemäß der gewünschten a posteriori-Verteilung (z.B. Gilks et al., 1996; Robert und Castella, 1999) und sind oft relativ einfach umzusetzen (z.B. mit OpenBUGS, Thomas et al., 2006).

Bei MCMC wird eine Markovkette generiert, deren stationäre Verteilung die gewünschte a posteriori-Verteilung ist. Die Simulation dieser Markovkette liefert dann eine abhängige Stichprobe der a posteriori-Verteilung aus einer so genannten Markovkette.

Eine Markovkette (erster Ordnung) ist ein stochastischer Prozess, also eine Folge von Zufallsvariablen $(X_t)_{t=0}^{\infty}$, mit einem beliebigen Startwert X_0. Jedes Element X_t wird dabei aus der bedingten Verteilung $p(X_t \mid X_{t-1})$ gezogen. Die Verteilung jedes Elementes X_t ist somit nur von dem vorherigen Element X_{t-1} abhängig. Hierbei sollte bei der Konstruktion der Markovkette geachtet werden, dass sie die a posteriori-Verteilung π als ihre stationäre Verteilung besitzt (vgl. Fahrmeir et al., 1981). Dazu gibt es verschiedene Möglichkeiten und eine davon ist der Metropolis-Hastings-(MH)-Algorithmus (Hastings, 1970; Metropolis et al., 1953).

Sei X_0 (X_t mit $t = 0$) der Startwert einer Markovkette der Länge N mit einer Vorschlagsdichte q (z.B. multivariate Normalverteilung) mit $\int q(Y \mid X_t) dX_t = 1$. Der MH-Algorithmus zieht in jedem Iterationsschritt einen potentiellen Nachfolgepunkt gemäß dieser Vorschlagsdichte (engl.: proposal distribution) und verwirft bzw. akzeptiert diesen als den Nachfolgezustand mit einer berechneten Akzeptanzwahrscheinlichkeit. Dabei wird für $t = 1, \ldots, N$ iterativ vorgegangen:

1. Für X_{t+1} wird aus der Vorschlagsdichte $q(Y \mid X_t)$ ein Zustandskandidat Y generiert, der auf den letzten Wert der Kette X_t bedingt ist.

2. Im nächsten Schritt wird die Akzeptanzwahrscheinlichkeit berechnet, dass dieser Y als neuer Zustand akzeptiert wird:

$$\alpha(X_t, Y) = min\left\{\frac{\pi(Y)q(X_t \mid Y)}{\pi(X_t)q(Y \mid X_t)}, 1\right\}.$$

3. Für den Wert X_{t+1} gilt nun:

$$X_{t+1} = \begin{cases} Y, & falls\ R \leq \alpha(X_t, Y), \quad R \sim U(0,1) \\ X_t, & sonst \end{cases}.$$

4. Der MH-Algorithmus wird an dieser Stelle beendet, falls $t = N$ ist, ansonsten läuft er alle Schritte wieder durch mit $t = t + 1$.

Falls durch die Vorschlagsdichte symmetrische Kandidaten generiert werden mit $q(Y \mid X) = q(X \mid Y)$, vereinfacht sich die Akzeptanzwahrscheinlichkeit nach Metropolis et al. (1953) zu

$$\alpha(X, Y) = \ min\left\{\frac{\pi(Y)}{\pi(X)}, 1\right\}.$$

Eine weitere Möglichkeit zur Konstruktion einer Markovkette bietet der so genannte komponentenweise Metropolis-Hastings-Algorithmus (vgl. Gelman et al., 2004). Die Idee dabei ist, den Zufallsvektor in *n* Komponenten $X_1, \dots, X_n$ zu zerlegen und jede Zustandskomponente einzeln neu zu ziehen. Die restlichen Komponenten bleiben dabei unverändert.

Sei eine weitere Komponente $X_{.-i}$ mit $(X_{.1}, \dots, X_{.i-1}, X_{.i+1}, \dots, X_n)$ definiert und sei $X_{t.i}$ der Zustand der *i*-ten Komponente $X_{.i}$ am Ende vom Schritt *t*. Der komponentenweise MH-Algorithmus läuft dann für *t* = 1, …, *N* die folgenden Schritte durch:

1. Für alle Komponenten $X_1, \dots, X_n$ wird aus der Vorschlagsdichte $q_i(Y_i \mid X_{.i}, X_{.-i})$ ein Zustandskandidat $Y_{.i}$ gezogen.

2. Im nächsten Schritt wird die zugehörige Akzeptanzwahrscheinlichkeit berechnet:

$$\alpha(X_{.-i}, X_{.i}, Y_{.i}) = \ min\left\{\frac{\pi(Y_{.i} \mid X_{.-i}) q_i(X_{.i} \mid Y_{.i}, X_{.-i})}{\pi(X_{.i} \mid X_{.-i}) q_i(Y_{.i} \mid X_{.i}, X_{.-i})}, 1\right\}$$

   mit $\pi(X_{.i} \mid X_{.-i})$ als Randverteilung der *i*-ten Komponente gegeben die restlichen Komponenten. Ein Spezialfall des komponentenweisen Metropolis-Hastings-Algorithmus ist der Gibbs-Sampler (Geman und Geman, 1984; Gelman et al., 2004). Bei Gibbs Samplern ist die Akzeptanzwahrscheinlichkeit durch die Vorschlagsdichte $q_i(Y_{.i} \mid X_{.i}, X_{.-i}) = \pi(Y_{.i} \mid X_{.-i})$ gleich 1, so dass die vorgeschlagenen Kandidaten immer akzeptiert werden.

3. Die folgenden Schritte sind gleich denen im normalen Metropolis-Hastings-Algorithmus. Für die neue Komponente $X_{t+1,i}$ wird der Zustand $Y_{.i}$ zu den gleichen Bedingungen wie bei dem normalen angenommen bzw. abgelehnt. Dabei ist es hier im Gegensatz dazu möglich, dass für einige Komponenten der neue Zustand akzeptiert und für die anderen dagegen abgelehnt wird.

4. Der Index t wird um 1 erhöht und der Algorithmus läuft alle Schritte erneut durch, falls $t < N$. Ansonsten wird er an dieser Stelle beendet.

Da der Startwert X_0 dabei beliebig gesetzt werden kann, konvergiert die konstruierte Markovkette erst nach einiger Zeit gegen die stationäre Verteilung, so dass in der Praxis der Bereich am Anfang der Kette – auch burn-in genannt – verworfen werden muss (vgl. Gelman et al., 2004).

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass diese Arbeit von mir selbstständig verfasst und angefertigt wurde und ich keine anderen als die im Literaturverzeichnis angegebenen Quellen verwendet habe. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Dortmund, 25. Oktober 2014 _____

Evgenia Freis