Aus der Abteilung Bioinformatik
(Prof. Dr. rer. nat. E. Wingender)
im Zentrum Informatik, Statistik und Epidemiologie
der Medizinischen Fakultät der Universität Göttingen

**Analyse biologischer
Signaltransduktionsnetzwerke
auf der Grundlage von Genexpressionsdaten**

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
der Medizinischen Fakultät
der Georg-August-Universität zu Göttingen

vorgelegt von

Jost Degenhardt

aus

Bad Harzburg

Göttingen 2009

Dekan: Prof. Dr. med. C. Frömmel

I. Berichterstatter: Prof. Dr. rer. nat. E. Wingender

II. Berichterstatter/-in: Prof. Dr. rer. nat. Beissbarth

III. Berichterstatter/-in: Prof. Dr. Theis

Tag der mündlichen Prüfung: 10.08.2010

Inhaltsverzeichnis

1	Einleitung.....	1
2	Material und Methoden.....	5
2.1	Graphentheorie.....	5
2.2	Biologische Signaltransduktionsnetzwerke.....	6
2.3	Netzwerkcluster.....	8
2.4	Statistische Grundlagen.....	10
2.4.1	Expressionssignifikanz.....	10
2.4.2	Korrelationskoeffizient.....	12
2.5	Relationale Datenbanken.....	13
2.6	Verwendete Softwarepakete.....	15
2.6.1	JAVA.....	15
2.6.2	JUNG.....	15
2.6.3	Graphviz.....	16
2.6.4	GNU R.....	16
2.6.5	PostgreSQL.....	16
2.7	SAGE.....	17
2.7.1	Molekularbiologische Grundlagen.....	17
2.7.2	CGAP.....	20
2.8	TRANSPATH.....	23
3	Versuchsaufbau.....	28
3.1	TRANSPATH als relationale Datenbank.....	28
3.2	SAGE-Daten als relationale Datenbank.....	31
3.3	Netzwerkgenerierung.....	33
3.3.1	Identifikation der Startgene.....	33
3.3.2	Aufteilen der Startgene.....	34
3.3.3	Abbilden der Startgene auf Startmoleküle.....	34
3.3.4	Abbilden der Startmoleküle auf Orthologe.....	35
3.3.5	Erweiterung des Netzwerks.....	36
3.3.6	Färbung des Netzwerks.....	36
3.4	Netzwerkvergleich.....	39
3.5	Signalperkolationscluster.....	41
4	Ergebnisse.....	43
4.1	Korrelation von Netzwerkähnlichkeit und Gewebeeigenschaften.....	43
4.2	Identifikation von Schlüsselmolekülen anhand von Beispielen.....	46
4.2.1	CXCR4.....	46
4.2.2	AKT/MDM2/p53.....	48
5	Zusammenfassung.....	50
6	Anhang.....	52
6.1	Allgemeines Design der Anwendung.....	52
6.2	Handbuch.....	54
6.2.1	Auswahl der Expressionsdaten.....	55
6.2.1.1	SAGE-Daten.....	56

6.2.1.1.1 Auswahl der SAGE-Daten.....	56
6.2.1.1.2 Feinauswahl der SAGE-Daten.....	58
6.2.1.1.3 Auswahl und Konfiguration der Statistischen Methode.....	59
6.2.1.2 Microarray-Daten.....	60
6.2.2 Erstellung des Netzwerks.....	62
6.2.3 Darstellung des Signaltransduktionsnetzwerkes und Verlinkung zu anderen Datenbanken.....	63
6.3 Verwendete TRANSPATH-Attribute.....	66
7 Literaturverzeichnis.....	68

1 Einleitung

Die Sequenzierung des menschlichen Genoms (International Human Genome Sequencing Consortium 2004) sowie die Entwicklung von Highthroughput-Methoden wie SAGE (Velculescu et al. 1995) oder Microarrays (Skena et al. 1995) stellten einen entscheidenden Fortschritt im Verständnis des Genoms und seiner Expression dar. Zum ersten Mal war es möglich, Erkenntnisse über den gesamten Expressionszustand eines Gewebes zu einem definierten Zeitpunkt zu gewinnen. Sowohl zur Sequenzierung ganzer Genome als auch für die Auswertung von Genexpressionsexperimenten wurden computergestützte Rechenmodelle benötigt, was maßgeblich zur Weiterentwicklung der Bioinformatik beitrug. Seitdem konnten diese Methoden vielfach erfolgreich dazu eingesetzt werden, Gene zu identifizieren, deren Expressionsgrad z. B. für einen bestimmten Krankheitszustand spezifisch ist. Neben den zahlreichen Erfolgen werfen jedoch vor allem die enormen Datenmengen, die durch diese Methoden erzeugt werden, noch heute diverse Probleme auf. Demzufolge werden immer ausgefeiltere Verfahren entwickelt, um wichtige von unwichtigen Informationen zu trennen.

Vor diesem Hintergrund erscheint es um so erstaunlicher, daß ein auf den genannten Methoden basierendes Experiment lediglich einen kleinen Teil dessen wiedergibt, was in den untersuchten Geweben tatsächlich geschieht. Schließlich stellt die Expression eines Gens lediglich das Ergebnis bzw. den Ausgangspunkt unzähliger Interaktionen dar, an denen neben Nukleinsäuren auch Proteine sowie niedermolekulare Substanzen wie ATP oder masselose Faktoren wie UV-Licht beteiligt sein können. Bildet man diese Interaktionen auf Signale im Sinne von Aktivierung, Inhibition, etc. ab, bilden sie ein biologisches Signaltransduktionsnetzwerk. Erkenntnisse über diesen Teil der Zellfunktion, auch Interaktom genannt, lassen sich aus Genexpressionsdaten höchstens indirekt gewinnen. Vielmehr werden dazu andere experimentelle Methoden benötigt. Hierzu zählen z.B. Yeast-2-Hybrid-Systeme (Fields und Song 1989) sowie quantitative Verfahren zur Analyse der Proteinexpression wie Proteinarrays (MacBeath und Schreiber 2000, Eckel-Passow et al. 2005). Diese und viele weitere Methoden bzw. die mit ihnen verbundenen Fragestellungen werden oft auch unter der Bezeichnung Proteomik subsumiert (Wasinger et al. 1995, Vemuri und Aristidou 2005). Die Herausforderungen, die dieses Fach an die Forschung in Zukunft stellen wird, läßt sich allein schon dadurch erahnen, daß den ca. 30.000 Genen des menschlichen Genoms ein Proteom gegenübersteht, das sich aus einer weitaus größeren Anzahl von Proteinen zusammensetzt. So belaufen sich die Schätzungen auf 50.000-90.000 einzelne Proteine im menschlichen Proteom (Harrison et al. 2002, Orchard et al. 2005). Auch wenn die Proteomik im Vergleich zur Genetik

sicher noch in den Kinderschuhen steckt, produziert sie ebenfalls bereits, auch in Bezug auf die angesprochenen biologischen Signaltransduktionsnetzwerke, erhebliche Datenmengen, die denen aus Genexpressionsdaten gegenüberstehen.

Um nun ein möglichst umfassendes Bild dessen zu gewinnen, was in einer bestimmten Zelle passiert, wäre demnach die sinnvolle Verknüpfung der beiden oben beschriebenen Datenmengen wünschenswert. Könnte man die in einem SAGE- oder Microarrayexperiment als signifikant über- oder unterexprimiert identifizierten Gene bzw. ihre Genprodukte auf das entsprechende biologische Signaltransduktionsnetzwerk abbilden, ergäben sich neue Möglichkeiten zum Verständnis der Zellfunktion. Wird z. B. ein Gen x als Ergebnis eines Expressionsexperiments als signifikant überexprimiert identifiziert, könnte sein Genprodukt X direkt oder über eine Signalkaskade dazu führen, daß ein Rezeptor Z , dessen zugrundeliegendes Gen z nicht signifikant überexprimiert wurde, in seiner biologischen Funktion verändert wird. Weder das entsprechende Expressionsexperiment noch das Signaltransduktionsnetzwerk würden jeweils für sich genommen diese Information enthalten.

Zur Identifikation von Z , etwa als potenzielles Ziel neuer therapeutischer Interventionsmöglichkeiten, wäre demnach ein Werkzeug in Form einer Software erforderlich, das eine solche Verbindung zwischen ihnen herzustellen vermag.

Eine wichtige Rolle spielt dabei die konkrete Datengrundlage, auf der eine solche Software aufbaut. Auf der Seite der Genexpressionsdaten existieren diverse Datenbanken, die die Ergebnisse von Microarray-Experimenten bereitstellen. Als Beispiele seien die Stanford Microarray Database (Ball et al. 2005) und ArrayExpress (Brazma et al. 2003) genannt. Auf seiten der SAGE-Daten wird vom CGAP-Projekt (Boon et al. 2002) die SAGE-Genie Datenbank (Liang 2002) zur Verfügung gestellt. Die hohe Qualität und Vollständigkeit der SAGE-Genie Datenbank sowie ihr konzeptioneller Aufbau, zu möglichst jedem Gewebe sowohl Daten über den normalen Zustand als auch über möglichst viele neoplastisch veränderte Formen bereitzustellen, lassen sie besonders geeignet erscheinen.

Auf der Seite der Protein-Interaktions-Daten ist die Auswahl an geeigneten Datenbanken, die eine möglichst große Zahl einzelner Interaktionen enthalten, weitaus weniger groß. Hier sei z. B. die DIP-Datenbank (Xenarios et al. 2000) genannt. Sehr viel aussagekräftiger sind jedoch Signaltransduktionsdatenbanken, da sie Reaktionsketten abbilden, die auch Nicht-Protein-Komponenten enthalten. Hier sei, nicht zuletzt aufgrund der hohen Datenqualität TRANSPATH (Krull et al. 2006) genannt.

Die Anforderungen an die zu entwickelnde Software sind somit vielfältig. Zunächst müssen SAGE-Daten mit geeigneten statistischen Verfahren im Hinblick auf die differentielle Expression

einzelner Gene verarbeitet werden. Die auf diese Weise identifizierten Gene müssen daraufhin auf ihre Genprodukte abgebildet werden, die wiederum mit den ihnen entsprechenden TRANSPATH-Einträgen verknüpft werden müssen.

Ausgehend davon muß aus den TRANSPATH-Daten ein entsprechendes Signaltransduktionsnetzwerk aufgebaut werden. Dieses Netzwerk soll wiederum in möglichst intuitiver Weise dargestellt werden, wobei sinnvolle Funktionen zur Verfügung gestellt werden sollen, die die Analyse des jeweiligen Netzwerks im Hinblick auf Schlüsselmoleküle erleichtern. Ebenfalls wichtig ist die Verknüpfung der einzelnen Moleküle und Interaktionen mit entsprechenden Einträgen in externen Datenbanken, die z. B. in einem Webbrowser angezeigt werden können.

Einige dieser Schritte bedingen sehr rechen- und speicherintensive Operationen. Darüberhinaus sollten die verwendeten Datenbanken stets auf dem aktuellsten Stand gehalten werden. Demnach bietet sich beim allgemeinen Design der Software ein Client-Server-Konzept an, bei dem die Datenbanken sowie die ressourcenintensiven Schritte auf einem zentral verwalteten Server ausgeführt werden, auf den mehrere Clients zugreifen können.

Ein Teil der oben beschriebenen Funktionalität ist in bereits vorhandenen Anwendungen wie Cytoscape (Ideker et al. 2002, Shannon et al. 2003) oder TopNet (Hanisch et al. 2004) realisiert worden. Die Unterschiede bestehen zum einen in der Art der verwendeten Daten. Während ähnliche Anwendungen in der Regel auf Genexpressionsdaten aufbauen, die vom jeweiligen Benutzer importiert werden müssen, soll die zu entwickelnde Anwendung mit den SAGE-Genie-Daten bereits einen umfangreichen Ausgangspunkt für entsprechende Experimente mitbringen und laufend auf dem aktuellsten Stand halten können, ohne jedoch auf die Möglichkeit zu verzichten, eigene Genexpressionsdaten mit einzubeziehen. Ein weiterer Unterschied liegt darin begründet, daß die vorhandenen Anwendungen nicht auf SAGE-Daten, sondern auf Microarrays aufbauen und meistens auch keinen Zugriff auf TRANSPATH haben.

Auf der programmtechnischen Ebene stellt die bereits beschriebene Client-Server-Architektur ein weiteres Unterscheidungsmerkmal dar. Außerdem unterscheidet sich die zu erstellende Anwendung in Bezug auf die Analyse der Netzwerke. Die Errechnung der Signifikanzwerte für die einzelnen Moleküle und die Unterteilung des Netzwerks in spezielle *cluster* seien hier als Beispiele genannt, die in den entsprechenden Kapiteln näher erläutert werden.

Ein weiterer möglicher Anwendungsbereich bezieht sich auf die Ähnlichkeit zwischen verschiedenen Netzwerken. Man könnte Netzwerke verschiedener Experimente miteinander vergleichen, um zu untersuchen, ob solche, die aus demselben Gewebetyp oder Krankheitszustand erstellt wurden, untereinander ähnlich sind. Wäre dies der Fall, ließen sich z. B. Algorithmen

entwickeln, die durch den Vergleich eines aus einem Gewebe unklarer Dignität erstellten Netzwerkes mit solchen, die aus neoplastisch veränderten Zellen stammen, eine hypothetische Aussage über dessen Zustand treffen.

Das Ziel dieser Arbeit besteht somit zunächst darin, eine Software zu entwickeln, die die oben geschilderte Funktionalität zur Verarbeitung und Analyse von Genexpressions- und Signaltransduktionsdaten bietet. Als nächstes sollen mit dieser Anwendung die oben aufgestellten Hypothesen überprüft werden:

1. Es soll anhand von Beispielen versucht werden, die Hypothese zu bestätigen, daß Schlüsselmoleküle, deren Gene wie oben beschrieben nicht signifikant differentiell exprimiert wurden, aufgrund ihrer Rolle im Signaltransduktionsnetzwerk als solche identifiziert werden können.
2. Eine große Anzahl von Netzwerken, denen unterschiedliche Gewebe bzw. Krankheitszustände zugrundeliegen, sollen miteinander verglichen werden, um die Hypothese zu überprüfen, daß es eine Korrelation zwischen der Art der zugrundeliegenden Daten und der Netzwerkähnlichkeit gibt.

Die verwendeten Daten sollen dabei auf SAGE-Experimenten beruhen, wobei die speziesspezifische Herkunft auf *Homo sapiens* beschränkt bleiben soll. Die beschriebenen Konzepte lassen sich jedoch genau so auf Daten anwenden, die auf der Grundlage anderer Methoden und für andere Spezies gewonnen wurden.

2 Material und Methoden

2.1 Graphentheorie

Um Netzwerke effektiv modellieren zu können, wurden verschiedene grundlegende Prinzipien der Graphentheorie angewendet. Die verwendeten Begriffe und Definitionen wurden Diestel (2006) entnommen und lauten wie folgt.

Definition 1:

Ein Graph ist ein Paar $G=(V, E)$ disjunkter Mengen mit $E \subseteq [V]^2$. Die Elemente von E sind also 2-Tupel von V . Die Elemente von V werden als Knoten, die Elemente von E als Kanten bezeichnet. Haben die Kanten eine Richtung, können sie also als Pfeil dargestellt werden, ist der Graph gerichtet, ansonsten ist er ungerichtet. Den einzelnen Elementen von V und E können eindeutige Bezeichner (*labels*) zugewiesen werden. Alle hier verwendeten Graphen weisen eindeutige *label* für Elemente von V und E auf, so daß ein Graph keine zwei Knoten mit ein und demselben *label* enthalten kann.

Definition 2:

Ein Graph G' ist ein Teilgraph von $G=(V, E)$, wenn $V' \subseteq V$ und $E' \subseteq E$ mit $E' = V' \times V'$.

Definition 3:

Zwei Graphen $G1=(V1, E1)$ und $G2=(V2, E2)$ heißen isomorph, wenn eine eindeutige Funktion $f: V1 \rightarrow V2$ existiert, so daß

1. für jedes v_i in $V1$ gilt $f(v_i)$ in $V2$
2. für jedes v_i, v_j in $V1: (v_i, v_j) \text{ in } E1 \iff (f(v_i), f(v_j)) \text{ in } E2$

Ein Graph $H=(V, E)$ heißt gemeinsamer, isomorpher Teilgraph zweier Graphen $G1$ und $G2$, wenn ein Teilgraph $H1$ von $G1$ und ein Teilgraph $H2$ von $G2$ existieren, so daß $H1$ und $H2$ isomorph zu H sind. Existiert kein anderer gemeinsamer isomorpher Teilgraph $H'=(V', E')$, für den gilt $|V'| > |V|$ und $|E'| > |E|$, ist H der größte, gemeinsame, isomorphe Teilgraph.

Definition 4:

Einen Sonderfall stellen Graphen dar, deren Knoten sich in zwei disjunkte Teilmengen auftrennen lassen. Verbinden die Kanten eines solchen Graphen stets Knoten aus verschiedenen Teilmengen, spricht man von einem *bipartiten Graphen*.

Definition 5:

Die Anzahl der Kanten, mit denen ein Knoten verbunden ist, wird als *Grad* oder *Inzidenz* eines Knotens bezeichnet. Ein Knoten V und eine Kante E sind zueinander *inzident* wenn sie sich berühren, sprich V ein Teil von E ist. In gerichteten Graphen kann noch zwischen positiver und negativer Inzidenz unterschieden werden, wobei eine Kante $E = V_1, V_2$ positiv inzident zu V_1 und negativ inzident zu V_2 ist.

Definition 6:

Sind zwei Knoten V_1 und V_2 Teil der Kante E , werden sie als adjazent bezeichnet. Zwei Kanten E_1 und E_2 sind zueinander adjazent, wenn sie zu ein und demselben Knoten V inzident sind.

2.2 Biologische Signaltransduktionsnetzwerke

Signaltransduktion beschreibt eine Interaktion, bei der ein bestimmtes Signal, auch Stimulus genannt, in ein anderes Signal umgewandelt wird. Beginnen solche Interaktionen miteinander Ketten und Pfade zu bilden, ergibt sich ein Signaltransduktionsnetzwerk. Die Abläufe des inter- und intrazellulären Interaktions lassen sich als Interaktionen zwischen Proteinen, Second-Messengern, etc. darstellen. Auf diese Weise erhält man ein Signaltransduktionsnetzwerk, das den Informationsfluß widerspiegelt, der sich sowohl in der Zelle selbst als auch zwischen verschiedenen Zellen abspielt.

Die das Signaltransduktionsnetzwerk bildenden Moleküle lassen sich grob in verschiedene Gruppen einteilen. So enthält die Gene Ontology (Harris et al. 2004) 1751 menschliche Gene, die mit dem Term „*signal transducer activity*“ assoziiert sind. Dabei handelt es sich zum Großteil um Kinasen, Phosphatasen, G-Proteine und Rezeptoren. Darüberhinaus spielen Second-Messenger und niedermolekulare Hormone wie z.B. Steroide ebenfalls eine wichtige Rolle.

Ein Netzwerk kann generell als Struktur beschrieben werden, die sich mathematisch als Graph modellieren läßt und sich aus folgenden Beziehungen zwischen den Knoten aufbaut:

Interaktion: Eine binäre Beziehung i zwischen zwei Knoten A und B . Die Graphen, die den hier betrachteten Netzwerken zugrundeliegen, enthalten ausschließlich gerichtete Kanten, sodaß für die von ihnen modellierten Interaktionen $A \rightarrow B$ gilt. A wird hierbei als Vorgänger, B als Nachfolger bezeichnet. In Signaltransduktionsnetzwerken wird A auch Signaldonor bzw. B Signalakzeptor genannt.

Kette (Chain): Eine lineare Folge von Interaktionen $i_1, i_2, \dots, i_n$, so daß der Nachfolger der Interaktion i_{n-1} zugleich der Vorgänger der Interaktion i_n ist. Jede Kette hat einen Startknoten N_0 und einen Endknoten N_n , was auch mit $chain(N_0, N_n)$ beschrieben wird.

Pfad (Pathway): Ein Pfad setzt sich aus mehreren Ketten zusammen, wobei es wieder einen Start- und einen Endpunkt gibt, was mit $pathway(N_0, N_n)$ beschrieben wird. Charakteristisch für Pfade ist, daß im Gegensatz zu Ketten divergente und konvergente Beziehungen zwischen den Knoten erlaubt sind. So können z. B. Feedbackbeziehungen dargestellt werden: Der dargestellte Pfad $P = pathway(A, F)$ enthält z. B. eine Konvergenz von C und D auf E , sowie eine Divergenz von E auf B und F . Die Interaktionen $B \rightarrow C \rightarrow E \rightarrow B$ stellen eine Feedbackbeziehung her.

Eine mögliche Abfolge von Ketten, aus denen sich P zusammensetzen könnte, lautet:
 $chain(A, B) + chain(B, D) + chain(B, E) + chain(E, B) + chain(D, F)$.

Ein Netzwerk enthält also eine Anzahl von Startknoten und eine Anzahl von Endknoten, die über Pfade miteinander verknüpft sind, wobei ein Knoten in mehreren Pfaden enthalten sein kann.

Weitere Begriffe, die häufig zur Beschreibung biologischer Netzwerke gebraucht werden:

Aktivierung/Inhibition: Ein Protein kann im Rahmen einer Interaktion von einem aktiven in einen inaktiven Zustand überführt werden oder umgekehrt. Damit ein solches Phänomen im Experiment beobachtet werden kann, muß die entsprechende Form natürlich eine ausreichend hohe Konzentration erreichen.

upstream/downstream: Ist die Interaktion $A \rightarrow B$ Teil eines der oben beschriebenen Konstrukte, sagt man auch, daß A *upstream* von B , bzw. daß B *downstream* von A liegt. Diese Beziehung kann auch indirekter Natur sein, wenn also die Verbindung zwischen A und B nicht in einer einzelnen Interaktion, sondern einer Kette oder einem Pfad besteht.

Crosstalk: Ist ein Knoten Teil mehrerer Pfade, wird die Verbindung, die er auf diese Weise zwischen ihnen herstellt, auch als „*crosstalk*“ bezeichnet.

2.3 Netzwerkcluster

Zufallsnetzwerke

Eine Möglichkeit, einen Zufallsgraphen bzw. ein Zufallsnetzwerk zu erstellen, besteht in der von Erdos und Renyi (1959) publizierten Methode. Sie basiert auf einer festgelegten Wahrscheinlichkeit p , mit der eine Kante E zwischen zwei Knoten V_1 und V_2 existiert, bzw. der Wahrscheinlichkeit $1-p$, daß sie nicht existiert. Da die Gesamtzahl der möglichen Kanten in einem ungerichteten Graphen G , in dem n Knoten enthalten sind, $n(n-1)/2$ beträgt, ist die durchschnittliche Anzahl der Kanten eines solchen Graphen $n(n-1)p/2$. Um den durchschnittlichen Grad eines Knotens in einem solchen Graphen zu berechnen, muß noch bedacht werden, daß jede Kante zwei Enden hat, so daß sich der durchschnittliche Grad z eines Knotens N nach der Formel

$$z_N = \frac{n(n-1)p}{n} = (n-1)p \approx np$$

berechnet.

Clusteringkoeffizient

Vergleicht man die auf die oben beschriebene Weise generierten Zufallsnetzwerke mit solchen, die auf realen Daten basieren, zeigt sich das Phänomen des *clusterings*. Damit ist gemeint, daß die Wahrscheinlichkeit zweier Knoten N_1 und N_2 , durch eine Kante verbunden zu sein, steigt, wenn ein dritter Knoten N_3 existiert, zu dem beide adjazent sind. Watts und Strogatz (1998) führten das Maß des Clusteringkoeffizienten C ein, der die Wahrscheinlichkeit angibt, mit der zwei zu einem bestimmten Knoten adjazenten Knoten ebenfalls adjazent sind. Da in Zufallsgraphen die

Wahrscheinlichkeit zweier Knoten, über eine Kante miteinander verbunden zu sein, per definitionem unabhängig von weiteren zu ihnen adjazenten Knoten ist, ist in einem solchen Graphen $C = p$ bzw. $C \approx z/n$.

In so gut wie allen realen Netzwerken ist C hingegen erheblich höher, was vermuten läßt, daß in der Verteilung der Knotengrade solcher Netzwerken wichtige Informationen über die ihnen zugrundeliegenden Prinzipien enthalten sind. Ein bedeutendes Merkmal vieler realer Netzwerke sind z. B. sogenannte *hubs*, also Knoten, die eine im Vergleich zur durchschnittlichen Inzidenz erheblich höhere Anzahl von Kanten aufweisen, die zu ihnen inzident sind. *Hubs* spielen wiederum eine große Rolle bei dem sogenannten *small-world*-Phänomen, womit gemeint ist, daß der kürzeste Weg zwischen zwei beliebigen Knoten im Durchschnitt aus sehr viel weniger Kanten besteht, als dies in Zufallsnetzwerken der Fall ist.

Weak Components

Das oben beschriebene Phänomen des *clusterings* führt in vielen realen Netzwerken, so auch bei biologischen Signaltransduktionsnetzwerken, zur Ausbildung sogenannter *weak components*. Hierbei wird der im Falle eines Signaltransduktionsnetzwerkes ja eigentlich gerichtete Graph zunächst als ungerichteter Graph modelliert. Als *weak component* wird nun ein Graph bezeichnet, in dem von jedem Knoten aus jeder andere Knoten des Graphen über eine oder mehrere Kanten erreichbar ist. Besteht ein Graph aus mehreren Teilgraphen, die nicht miteinander verbunden sind, läßt er sich demnach in mehrere *weak components* unterteilen. Ein solches Verfahren, das einen Graphen bzw. das darauf basierende Netzwerk im Hinblick auf die Inzidenz seiner Knoten und deren Erreichbarkeit untereinander in verschiedene Teilgraphen unterteilt, wird auch als Clustern, ein solcher Teilgraph als *cluster* bezeichnet. Wird ein Graph mittels Clustern in *weak components* unterteilt, handelt es sich bei den resultierenden Teilgraphen demnach um *weak component cluster*.

Die Aufteilung eines Signaltransduktionsnetzwerkes in *weak components* kann dazu dienen, das Netzwerk in verschiedene funktionelle Untereinheiten aufzuteilen. Eine offensichtliche Fehlerquelle dieser Vorgehensweise ist jedoch, daß die Modellierung der Protein-Protein-Interaktionen als ungerichtete Kanten den Signalfluß eines solchen Netzwerkes nicht berücksichtigt.

Strong Components

Im Gegensatz zu *weak components* berücksichtigen sogenannte *strong components* die Ausrichtung der Kanten. Als *strong component* wird ein Teilgraph bezeichnet, bei dem von jedem Knoten aus jeder andere Knoten unter Berücksichtigung der Kantenausrichtung erreichbar ist. Im

Falle eines Signaltransduktionsnetzwerkes würde somit die in den Interaktionen enthaltene Information nicht verloren gehen. Bezogen auf das *clustering* großer Netzwerke erscheint diese Methode jedoch zu restriktiv, da zur Ausbildung einer *strong component* die beteiligten Knoten einen geschlossenen Kreis bilden müssen, was in Bezug auf große Netzwerke ein seltenes, auf wenige Knoten beschränktes Phänomen ist. Im Gegensatz zu *weak components* enthält also die Ergebnismenge dieses Verfahrens nur einen kleinen Teil der Kanten des ursprünglichen Graphen.

2.4 Statistische Grundlagen

2.4.1 Expressionssignifikanz

Zur Analyse der differentiellen Expression von SAGE-Tags in zwei verschiedenen Zelltypen wurde die von Lal et al. (1999) beschriebene Methode verwendet. Gegeben seien zwei Zelltypen Y und Z . Eine bestimmte mRNA hat in Y und Z die unbekannten Konzentrationen y und z . Mittels der SAGE-Methode wird aus Y eine Gesamtzahl A Tags sequenziert, aus Z eine Gesamtzahl von B . Darunter befinden sich a bzw. b Tags, die der gesuchten mRNA entsprechen. Das erste Problem besteht nun darin, von A und B auf y und z zu schließen. Daraufhin muß entschieden werden, ob y und z einen signifikanten Konzentrationsunterschied darstellen.

Audic und Claverie (1997) beschreiben einen auf der klassischen Statistik basierenden Ansatz. Sie formulieren als Nullhypothese $H_0: y=z$ und als Alternativhypothese $y \neq z$. Von den beobachteten Daten werden die Formeln abgeleitet, mit denen H_0 mit verschiedenen Konfidenzintervallen abgelehnt wird. Unterscheiden sich a/A und b/B signifikant voneinander, wird H_0 abgelehnt und y und z als verschieden angesehen.

Chen et al. (1998) schlagen hingegen einen alternativen Ansatz vor, der auf dem Bayesschen Wahrscheinlichkeitsbegriff beruht: Betrachtet wird der Quotient

$$x = \frac{y}{y+z}$$

und die dazugehörige „a priori“-Verteilungsfunktion $f(x)$ über das Intervall $[0,1]$. Ist die Gesamtzahl der jeweils sequenzierten Tags gleich, also $A=B$, ist die „a posteriori“-Verteilungsfunktion

$$g(x) = f(x) x^a (1-x)^b.$$

Die Konzentration y übersteigt z um mindestens das F -fache wenn $x \geq L$ ist, wobei $L = F/(F+1)$ gilt. Lal et al. (1999) haben diese Methode auf Fälle ausgeweitet, bei denen $A \neq B$ ist, indem

$$g(x) = f(x) \frac{(x^a(1-x)^b)}{[1 + (A/B - 1)x]^{a+b}}$$

gesetzt wird. Ist $A = B$, ergibt sich hier wieder die bereits weiter oben genannte Formel für $g(x)$.

Es sei darauf hingewiesen, daß für die statistische Auswertung von SAGE-Daten diverse weitere Methoden existieren. Die von Vêncio et al. (2004) ruft beispielsweise den Autoren zufolge weniger falsch positive Ergebnisse hervor. Die Wahl fiel auf die Methode nach Lal et al. (1999), da sie auch von der SAGE-Genie-Webseite, also quasi der Referenzplattform zur Auswertung von SAGE-Daten, verwendet wird.

Zu den Vorteilen des Bayesschen Ansatzes zählt, daß nicht einfach die Nullhypothese angenommen oder abgelehnt wird. Im Gegensatz dazu kann die Wahrscheinlichkeit, daß die Konzentrationen y und z signifikant verschieden sind, für jeden gewünschten Faktor F gewählt werden. Von größter Bedeutung ist hierbei die Wahl der „a priori“-Verteilung $f(x)$. Wie in Chen et al. (1998) beschrieben, eignet sich hierzu $f(x) = x^c(1-x)^c$, so daß $f(x)$ eine symmetrische Funktion mit einem Maximum bei 0,5 ergibt. Je höher c gewählt wird, desto größer muß der Unterschied zwischen a und b sein, um als signifikant zu gelten. Frühere Analysen (Lal et al. 1999, Polyak et al. 1997, Zhang et al. 1997) lassen einen Wert von $c=3$ als angemessen erscheinen. In der vorliegenden Arbeit wurde daher $c=3$ und $F=2,0$ gewählt.

2.4.2 Korrelationskoeffizient

Der Korrelationskoeffizient k nach Pearson ist ein Maß für den Grad des linearen Zusammenhangs zwischen zwei Merkmalen. Er kann Werte zwischen -1 und 1 annehmen. Bei einem Wert von 1 (bzw. -1) besteht ein vollständig positiver (bzw. negativer) linearer Zusammenhang zwischen den betrachteten Merkmalen. Wenn der Korrelationskoeffizient den Wert 0 aufweist, hängen die beiden Merkmale überhaupt nicht linear voneinander ab. Die Formel, nach der sich k berechnen läßt lautet dabei:

$$k = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2)(n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2)}}$$

n : Anzahl der Messwerte
 x, y : Messreihen

Eine Alternative stellt die Rangkorrelation nach Spearman dar, die auch nichtlineare Zusammenhänge erfaßt. Hierzu wird jedem Meßwert mittels einer Funktion $rg(x)$ ein Rang zugeordnet und dieser dann in eine Ordinalskala eingefügt. Weist eine Meßreihe identische Meßwerte auf, teilen diese sich einen Rang. Der resultierende Wert für k nimmt ebenfalls Werte zwischen 1 und -1 an. Die zugehörige Formel lautet:

$$k = 1 - \frac{6 \sum_{i=1}^n (rg(x_i) - rg(y_i))^2}{n(n^2 - 1)}$$

n : Anzahl der Meßwerte
 x, y : Meßreihen

Beide Koeffizienten werden zur Korrelation zwischen Netzwerkähnlichkeit und bestimmten Netzwerkeigenschaften verwendet.

2.5 Relationale Datenbanken

Das Prinzip der relationalen Datenbank basiert grundsätzlich auf der sinnvollen Verknüpfung von Relationen. Eine Relation stellt hierbei eine zweidimensionale Tabelle dar. Die Zeilen einer solchen Tabelle stehen für einzelne Einträge und werden in diesem Zusammenhang auch als Entitäten bezeichnet. Die Spalten stehen für die verschiedenen Attribute. Jede Tabelle sollte dabei mindestens einen Primärschlüssel enthalten. Ein Primärschlüssel ist ein Attribut oder eine Kombination von Attributen, über deren Werte die einzelnen Einträge eindeutig identifizierbar sind. Soll nun eine Verknüpfung zwischen zwei Relationen hergestellt werden, erhält die eine Tabelle eine zusätzliche Spalte, in der das Primärschlüsselattribut für die Entitäten der anderen Tabelle gespeichert wird. Eine solche Spalte bezeichnet man auch als Fremdschlüssel.

Bei der Erstellung einer relationalen Datenbank werden die Daten meistens in mehrere Tabellen aufgeteilt, deren Gesamtheit man auch als Schema bezeichnet. Die Art und Weise, in der die Ausgangsdaten in Tabellen aufgeteilt werden, richtet sich vor allem nach der Kardinalität der Beziehung zwischen den Entitäten. Die Beziehungen werden unter diesem Gesichtspunkt wie folgt eingeteilt:

Kardinalität	Beschreibung	Beispiel
1:1	Die eine Entität steht genau einer anderen Entität gegenüber und umgekehrt.	Jedes Land hat genau eine Hauptstadt und jede Hauptstadt gehört zu genau einem Land.
1:n	Auf der einen Seite steht genau eine Entität einer anderen gegenüber. Umgekehrt steht eine Entität mehreren gegenüber.	in jedem Land gibt es mehrere Städte und jede Stadt gehört zu genau einem Land.
n:m	Eine Entität steht mehreren anderen gegenüber und umgekehrt.	Flüsse können durch mehrere Länder fließen und durch jedes Land können mehrere Flüsse fließen.

Tabelle 1: Beziehungen zwischen Entitäten

Kennt man die Kardinalität der einzelnen Beziehungen, geht man wie folgt vor:

1:1-Beziehung: Der Tabelle für die eine Entität wird eine Spalte hinzugefügt, die für die zweite steht. Die zweite Entität kann dabei entweder in einer eigenen Tabelle gespeichert werden, so daß die neue Spalte den Fremdschlüssel enthält, der auf den Primärschlüssel dieser Tabelle verweist. Als Alternative können auch die Attribute der zweiten Entität in solche der ersten Entität umgewandelt und als weitere Spalten in die Tabelle eingefügt werden.

1:n-Beziehung: Der Tabelle für die Entität auf der n-Seite wird eine weitere Spalte hinzugefügt, die wiederum einen Fremdschlüssel enthält. Alternativ können die Attribute der zweiten Entität als Spalten hinzugefügt werden, was jedoch zu erheblichen Redundanzen führen kann.

n:m-Beziehung: Es wird eine zusätzliche Tabelle erstellt, die zwei Spalten enthält. Die eine enthält dabei den Fremdschlüssel zur Entität auf der n-Seite, die andere entsprechend den zur Entität auf der m-Seite. Eine solche Tabelle wird auch als Linktabelle bezeichnet.

Eine Erweiterung des relationalen Datenbankkonzepts ist das der Objektrelationalität. Hier werden die Tabellen in eine objektorientierte Vererbungshierarchie eingeordnet. Einzelne Tabellen können also Entitäten und deren Attribute von anderen Tabellen erben. So könnte man z. B. zwei Tabellen erstellen, die für Groß- bzw. Kleinstädte stehen. Würde man nun diese beiden Tabellen von einer Tabelle für Städte ableiten, enthielte sie beide Arten von Städten.

2.6 Verwendete Softwarepakete

2.6.1 Java

Java ist eine objektorientierte Programmiersprache. Java-Programme werden in einer speziellen Umgebung, der Java-Laufzeitumgebung oder Java-Plattform ausgeführt, deren wichtigster Bestandteil die *Java Virtual Machine* (JVM) ist. Dazu werden Java-Programme in *bytecode* übersetzt, der von der JVM ausgeführt wird. Ein entscheidender Vorteil gegenüber vielen anderen Programmiersprachen wie z. B. C++ ist es daher, daß Java-Programme auf verschiedenen Plattformen, für die eine JVM existiert (Linux, Windows, etc.) ausgeführt werden können, ohne für jede dieser Plattformen neu kompiliert werden zu müssen („*write once, run anywhere*“).

Darüberhinaus weist Java eine Vielzahl von Eigenschaften auf, die es als besonders geeignet für die hier entwickelte Anwendung erscheinen lassen. Das Serialization-API zum Speichern und übermitteln von JAVA-Objekten, sowie das Reflection-API zur einfachen Implementierung einer Plugin-Funktionalität seien als Beispiele genannt.

Ein weiterer Vorzug von Java ist es, daß neben der mitgelieferten Klassenbibliothek zahlreiche weitere Open-Source-Bibliotheken existieren, die die Erstellung wissenschaftlicher Anwendungen erleichtern (z. B. JUNG).

2.6.2 JUNG

JUNG steht für *Java Universal Network/Graph Framework* und ist komplett in Java geschrieben. Es bietet einen Rahmen für die Modellierung, Analyse und die Visualisierung von Graphen bzw. Netzwerken aller Art. Die von JUNG bereitgestellten Klassen, allen voran diejenigen, die zur Modellierung von Graphen und deren Knoten und Kanten bestimmt sind, konnten problemlos auf die Bedürfnisse der entwickelten Anwendung angepaßt bzw. erweitert werden. Es existieren außerdem Schnittstellen zur Implementierung von Algorithmen oder Layouts, etc., die z. B. bei der Aufteilung von Netzwerken in *cluster* und bei der Einbindung von Graphviz genutzt wurden.

2.6.3 Graphviz

Graphviz (Ganser und North 2000) ist ein von AT&T und den Bell-Labs entwickeltes Open-Source-Programmpaket zur Visualisierung von gerichteten und ungerichteten Graphen. Es ist in C implementiert und somit streng genommen nicht plattformunabhängig. Es liegen jedoch ausführbare Dateien für alle gängigen Betriebssysteme vor.

Zu Graphviz existiert ein Java-Frontend namens Grappa, das Java-Klassen zur Modellierung eines Graphviz-Graphen bereitstellt.

Die entwickelte Anwendung nutzt Graphviz (via Grappa) zur Erstellung des Layouts der Netzwerke, wofür die von JUNG definierte Layout-Schnittstelle genutzt wird. Der Nutzen von Graphviz beschränkt sich somit auf die höhere ästhetische Qualität der Layouts im Vergleich zu denen, die JUNG selbst erstellen kann.

2.6.4 GNU R

GNU R (R Development Core Team 2006) ist eine auf vielen Plattformen verfügbare, freie Statistik-Software. Bei der Konzeption orientierten sich die Entwickler an der in den Bell Laboratories entwickelten Sprache zur Verarbeitung statistischer Daten S. GNU R ist in C geschrieben und somit nicht plattformunabhängig, es stehen jedoch ausführbare Dateien für alle gängigen Plattformen bereit. Da R mittlerweile bei der Implementierung wissenschaftlicher Algorithmen so etwas wie ein Standard geworden ist, liefern viele Autoren von Publikationen eine Implementierung des von ihnen vorgestellten Algorithmus in Form eines R-Skriptes mit. R-Skripte, wie die auf Vêncio et al. (2004) beruhenden, können von der entwickelten Anwendung für den Vergleich von Metabibliotheken als Alternative zum Standardverfahren genutzt werden.

2.6.5 PostgreSQL

PostgreSQL (Stonebraker et al. 1990) ist ein freies objektrelationales Datenbankverwaltungssystem (ORDBMS). PostgreSQL ist konform mit dem SQL92/SQL99 Standard und stellt einen JDBC-Treiber bereit, über den man von Java-Anwendungen aus auf PostgreSQL-Datenbanken zugreifen kann. Die entwickelte Anwendung nutzt PostgreSQL, um TRANSPATH und die SAGE-Daten als objektrelationale Datenbank zu speichern.

2.7 SAGE

Am Anfang jedes Expressionsexperiments steht die Wahl der Methode, mit der man Erkenntnisse über den Expressionszustand einzelner Gene gewinnen möchte. Heutzutage stehen sich mehrere solcher Methoden gegenüber. Dabei handelt es sich zum einen um solche, die auf der Hybridisierung von speziell markierter DNA oder RNA mit den entsprechenden Nukleotidketten der untersuchten Zellen basieren und als Microarrays bezeichnet werden (Schena et al. 1995). Darüberhinaus existiert die SAGE-Methode, die bereits 1995 von Velculescu et al. beschrieben wurde (Velculescu et al. 1995).

SAGE steht hierbei für *Serial Analysis of Gene Expression*. Da die hier verwendeten Daten ausschließlich aus SAGE-Experimenten stammen, wird die Beschreibung der biologischen Grundlagen und der Weiterverarbeitung der erhobenen Daten auf diese Methode beschränkt. Es sei jedoch darauf hingewiesen, daß die in Kapitel 3 und 4 beschriebenen Prinzipien ebenso Expressionsdaten als Ausgangspunkt haben können, die mit anderen Methoden erzeugt wurden.

Ein Grund für die Entscheidung, SAGE-Daten zu verwenden, war die gute Verfügbarkeit qualitativ hochwertiger Daten, vor allem durch CGAP. CGAP ist die Abkürzung für *Cancer Genome Anatomy Project* und wird vom NCI, also dem *National Cancer Institute* unterstützt. Das Hauptziel der CGAP-Initiative ist es, eine möglichst vollständige Sammlung von Expressionsdaten aus sowohl gesundem als auch neoplastisch verändertem Gewebe zur Verfügung zu stellen, aus deren Vergleich heraus sich Erkenntnisse über Gene gewinnen lassen, die eine Schlüsselrolle in der Krebsentstehung spielen könnten. Hierzu stellt CGAP einen öffentlich zugänglichen Server zur Verfügung, der als zentraler Speicherort für die SAGE-Daten diverser Laboratorien dient. Die Daten werden zu SAGE-Bibliotheken zusammengefasst, die sich üblicherweise auf ein bestimmtes SAGE-Experiment beziehen. Die darin enthaltenen Daten werden wiederum mit anderen Datenbanken verknüpft und stehen als Grundlage für die Weiterverarbeitung, also z. B. den Vergleich verschiedener SAGE-Bibliotheken, zur Verfügung.

2.7.1 Molekularbiologische Grundlagen

Die biologischen Prinzipien, die der SAGE-Methode zugrunde liegen, unterscheiden sich deutlich von denen der Microarray-Methode. Wie im Folgenden beschrieben wird, werden z. B. keine speziell markierten Nukleotidketten zur Hybridisierung benötigt. Eine bemerkenswerte Folge dieses Umstandes ist es, daß auch die Expression völlig unbekannter Gene oder solcher, deren Expression nicht erwartet wurde, gemessen werden kann.

Ein SAGE-Experiment läßt sich in folgende Schritte unterteilen:

1. Die mRNA wird über ihren Poly(A)-Schwanz an magnetische Beads gebunden, aus denen Oligo(T)-Sequenzen herausragen.
2. Da mRNA für die folgenden Schritte zu instabil ist, muß sie erst zu komplementärer cDNA übersetzt werden, die daraufhin in doppelsträngige DNA umgewandelt wird, die ihrerseits an die Beads gebunden ist.
3. Im nächsten Schritt wird die DNA durch ein Restriktionsenzym gespalten, das als „Ankerenzym“ bezeichnet wird. Ein Beispiel für ein solches Ankerenzym ist NlaIII mit der Restriktionssequenz „GTAC“. Es liegen somit unterschiedlich lange DNA-Stränge vor, die an ihrem 5'-Ende die gleiche Sequenz aufweisen und mit ihrem 3'-Ende an die Beads gebunden sind.
4. Die Probe mit der an die Beads gebundenen DNA wird in zwei Hälften aufgeteilt. Daraufhin werden die 5'-Enden der DNA, die in der ersten Probe enthalten sind, mit Linker A ligiert. Die DNA der zweiten Probe entsprechend mit Linker B.
5. Beide Linker enthalten eine Restriktionssequenz für ein „Tagging-Enzym“. Ein Tagging-Enzym spaltet DNA in einer bestimmten Entfernung von seiner Restriktionssequenz. Übliche Tagging-Enzyme spalten z. B. nach 10 oder 17 Nukleotiden, wobei sie glatte Enden erzeugen. Nach diesem Schritt liegen somit DNA-Stränge einer definierten Länge vor, die keine klebrigen Enden enthalten und nicht mehr an die Beads gebunden sind.
6. Daraufhin wird jeweils das dem Linker abgewandte Ende eines DNA-Stranges, der mit Linker A ligiert wurde, mit dem entsprechenden Ende eines an Linker B gebundenen DNA-Stranges ligiert. Somit liegt ein DNA-Strang vor, an dessen Enden sich jeweils die Linker befinden, wobei Sequenzen der ursprünglichen cDNAs dazwischen liegen. Ein solches Konstrukt bezeichnet man auch als „Ditag“.
7. Die Ditags werden nun per PCR amplifiziert, wobei Primer eingesetzt werden, die komplementär zu den beiden Linker-Sequenzen sind.
8. Nach der Amplifizierung werden die Ditags wieder mit dem Ankerenzym gespalten, so daß die Linker wieder von der cDNA entfernt werden.
9. Da die Ditags nun an beiden Seiten klebrige Enden aufweisen, die zu denen anderer Ditags komplementär sind, können sie zu einem sogenannten „Concatemer“ aneinander gehängt werden.
10. Die Concatemere werden daraufhin in Vektoren eingebracht, um sie klonieren und sequenzieren zu können.

Von den im Concatemer enthaltenen Sequenzen der einzelnen Tags läßt sich somit auf die Gene schließen, die im jeweils untersuchten Gewebe exprimiert wurden. Darüber hinaus gibt die Anzahl der jeweils zu einem Gen gehörenden Tags Informationen über das quantitative Ausmaß der Expression. Abbildung 1 gibt das allgemeine Prinzip noch einmal verkürzt wieder.

Abbildung 1: Schematische Darstellung der SAGE-Methode

1. *Bindung der mRNA an magnetische Beads*
2. *Umwandlung der mRNA in cDNA*
3. *Restriktionsverdau der cDNA durch das Ankerenzym und Aufteilung in Probe A und Probe B*
4. *Ligation mit Linker A bzw. Linker B*
5. *Restriktionsverdau durch das Tagging-Enzym*
6. *Erzeugung der Ditags*
7. *Amplifikation der Ditag mittels PCR*
8. *Erneuter Restriktionsverdau durch das Ankerenzym*
9. *Erzeugung des Concatemers*

2.7.2 CGAP

Die Zuordnung eines 10 oder 17 Basenpaare langen Tags zu einem bestimmten Gen stellt ein gravierendes Problem dar, das die sorgfältige Analyse der aus den SAGE-Experimenten stammenden Rohdaten erforderlich macht. Das CGAP versucht durch die vor allem von Boon et al. (2002) entwickelten Verfahren eine möglichst verlässliche Abbildung von Tags auf Gene zu erzeugen und stellt die Ergebnisse auf seiner SAGE-Genie-Webseite zur Verfügung (<http://cgap.nci.nih.gov/SAGE>).

Zunächst werden aus der Gesamtzahl der bekannten Tags solche aussortiert, die gewissen Kriterien nicht genügen. Beispielsweise werden alle Tags entfernt, die nur ein einziges Mal gefunden wurden, was einen Fehler in der Detektion, z. B. einen Sequenzierungsfehler, nahelegt. Auf diese Weise wird eine Liste verlässlicher Tags erzeugt.

Als nächstes werden aus den in Datenbanken wie z. B. UniGene oder Refseq enthaltenen Sequenzen „virtuelle Tagdatenbanken“ erstellt, indem sie nach potentiellen Tags durchsucht werden. War das Ankerenzym z. B. *NlaIII*, werden die Sequenzen nach dessen Restriktionssequenz „CAGT“ durchsucht, wobei nur die letzten vier in 3'-Richtung gelegenen Fundorte berücksichtigt werden. Wird eine solche Sequenz gefunden, definieren die darauffolgenden Basen ein neues „Virtuelles Tag“ (VT). Je nach verwendeter Taglänge handelt es sich typischerweise um 10 oder 17 Basen. Hat man nun beispielsweise auf einer Sequenz, die aus einem Unigene-Cluster stammt, ein solches VT gefunden, läßt sich so eine VT-zu-UniGene Beziehung herstellen.

Bei dieser Zuordnung werden die Sequenzen zusätzlich in verschiedene Gruppen eingeteilt, je nachdem, ob sie einen Poly(A)-Schwanz, ein Poly(A)-Signal, beide oder keines dieser Merkmale aufweisen. Anhand dieser Einteilung läßt sich also abschätzen, wie verlässlich das 3'-Ende der Sequenz definiert ist bzw. wie sicher die Restriktionssequenz und somit das eigentliche Tag tatsächlich enthalten sind. Die virtuellen Tagdatenbanken werden also in jeweils vier Untergruppen aufgeteilt. Die Verlässlichkeit dieser Untergruppen wird daraufhin zusätzlich bestimmt, in dem der Prozentsatz der in ihnen enthaltenen VTs ermittelt wird, die sich in der zuvor definierten Liste der

verlässlichen Tags wiederfinden. Tabelle 2 gibt einen beispielhaften Auszug aus den Daten wieder, die Boon et al. (2002) anhand des Vergleichs mehrerer virtueller Tagdatenbanken mit einer Liste von 194126 verlässlichen Tags erhielten.

Datenbank	Virtuelle Tagdatenbank (Einträge)	Poly(A)-Signal	Poly(A)- Schwanz	% der VTs in Liste verlässlicher Tags
Refseq	RefSeqP5S (3135)	+	+	90,9
	RefSeqP5R (1760)	-	+	88,4
	RefSeqSR (4480)	+	-	87,0
	RefSeqR (4663)	-	-	68,3
Unigene (die jeweils längste Sequenz eines Clusters)	20KP5S (5040)	+	+	90,0
	20KSR (5433)	+	-	86,0
	20KP5R (3629)	-	+	85,9
	20KR (5430)	-	-	69,7

Tabelle 2: Vergleich virtueller Tagdatenbanken mit 194126 verlässlichen Tags (Boon et al. 2002)

Auf diese Weise ist es möglich, den ermittelten VT-zu-Sequenz-Verbindungen und somit den sich daraus ergebenden VT-zu-Gen-Beziehungen relative Verlässlichkeitswerte zuzuordnen und somit für jedes Gen ein „bestes Tag“ und für jedes Tag ein „bestes Gen“ zu bestimmen. Es müssen jedoch einige Besonderheiten und Fehlerquellen beachtet werden.

Fehlende Gene

Natürlich ist es möglich, wenn auch unwahrscheinlich, daß ein Gen keine Restriktionssequenz für das benutzte Ankerenzym enthält. Der Anteil solcher Gene am menschlichen Genom wird in

Bezug auf *NlaIII* auf unter 1% geschätzt (Boon2002). Tabelle 3 gibt entsprechende Werte für die in Tabelle 2 enthaltenen virtuellen Tagdatenbanken an.

Virtuelle Tagdatenbank	% Sequenzen ohne <i>NlaIII</i>-Restriktionssequenz
RefSeqP5S	0,8
RefSeqP5R	0,3
RefSeqSR	0,5
RefSeqR	2,0
20KP5S	0,7
20KSR	0,6
20KP5R	0,5
20KR	2,5

Tabelle 3: Anteil von Sequenzen ohne Restriktionssequenz für NlaIII (Boon et al. 2002)

Redundante Gene

Enthalten zwei Sequenzen anschließend an eine Restriktionssequenz die gleiche Nukleotidabfolge, würde das entsprechende Tag auch auf beide Gene verweisen. Der Anteil solcher Gene scheint jedoch sehr gering zu sein. So reichen 10 Basenpaare lange Tags aus, um anhand von UniGene-Clustern 98% der Sequenzen eindeutig zu bestimmen, wenn die jeweils längste Sequenz des Clusters zugrunde gelegt wurde.

Interne Tags

Wie bereits weiter oben erwähnt, werden neben dem dem 3'-Ende am nächsten gelegenen Tag die drei weiter in 5'-Richtung gelegenen, sofern sie vorhanden sind, ebenfalls berücksichtigt. Es gibt mehrere Gründe, warum diese „internen Tags“ in einem SAGE-Experiment beobachtet werden könnten. Alternatives Splicing im Bereich des 3'-Endes und alternative Poly(A)-Signale gehören zu den biologischen Ursachen dieses Phänomens. Auf der anderen Seite könnte z. B. ein unvollständiger Restriktionsverdau durch das Ankerenzym oder eine nicht am Poly(A)-Schwanz ansetzende cDNA-Synthese ebenfalls eine Ursache dafür sein, daß weiter *upstream* gelegene Tags detektiert werden.

Vergleicht man das relative Vorhandensein der drei internen Tags in der Liste der verlässlichen Tags mit dem des Tags am 3'-Ende, zeigt sich, daß letzteres mit Abstand am stärksten repräsentiert ist, egal welche virtuelle Tagdatenbank man zugrunde legt (Boon et al. 2002).

2.8 TRANSPATH

Die immer größer werdende Zahl an Erkenntnissen über die Interaktionen in der Zelle, also das Interaktom, macht es erforderlich, die gesammelten Daten in einer dafür geeigneten Form zu speichern, um effizient darauf zugreifen zu können. Hierzu eignet sich das Konzept der relationalen Datenbank, das sich bereits in vielen anderen Bereichen, sowohl innerhalb als auch außerhalb der Bioinformatik, bewährt hat. Ein weiteres Argument für diese Form der Datenrepräsentation ist das Vorhandensein ausgereifter und hoch effizienter Verwaltungssysteme auf diesem Gebiet, wie z.B. PostgreSQL.

Neben dem Problem, die Daten in geeigneter Form zu speichern, stellt sich zunächst jedoch die Frage, wie aus der enormen Anzahl der wissenschaftlichen Publikationen die relevanten Informationen extrahiert werden können. Trotz immer ausgefeilterer Methoden des Textminings, also der automatisierten Analyse von Textdokumenten im Hinblick auf bestimmte Informationen, kann auch heutzutage ein größtmöglicher Qualitätsstandard nur durch manuelle Annotation der Daten erreicht werden, die durch qualifizierte Fachkräfte durchgeführt wird.

Mit TRANSPATH (Krull et al. 2006) liegt eine Datenbank für biologische Signaltransduktion vor, mit der die geschilderten Konzepte realisierbar sind. Die einzelnen Einträge werden erstellt, indem erfahrene Biologen Publikationen nach Erkenntnissen über Interaktionen zwischen Proteinen oder anderen biologischen Funktionsträgern durchsuchen. Es handelt sich also um eine von Hand annotierte Datenbank. TRANSPATH wird von der BIOBASE GmbH gepflegt und weiterentwickelt.

Auch wenn TRANSPATH im Rahmen des internen Gebrauchs bei BIOBASE als relationale Datenbank existiert, liegen die Daten, die Lizenznehmern zum Download zur Verfügung gestellt werden, zunächst als sogenannte Flatfiles vor. Es handelt sich also um Textdateien, die die einzelnen Einträge in einem festgelegten Format enthalten. Bereits auf dieser Ebene wird die Trennung zwischen „Molecules“, „Reactions“ und „Genes“ vorgenommen. Die von TRANSPATH verwendeten Bezeichnungen decken sich nicht immer mit den naheliegenden Bedeutungen. So entspricht, wie weiter unten noch näher erklärt werden wird, ein „Molecule“-Eintrag in TRANSPATH nicht immer einem Molekül. Aus diesem Grund wird im Folgenden den entsprechenden Ausdrücken das Präfix „TP“ vorangestellt, um eine klare Unterscheidung zu ermöglichen.

TPMoleküle: Als TPMolekül wird alles bezeichnet, was als Signaldonor bzw. -akzeptor fungieren kann. Daher muß der Eintrag für ein TPMolekül nicht unbedingt ein tatsächliches Molekül, also z. B. ein Protein widerspiegeln. Es kann sich vielmehr auch um einen Komplex aus mehreren Proteinen, eine Proteinfamilie, den bestimmten Zustand eines Proteins, ein chemisches Element oder einen masselosen Signalgeber wie z. B. UV-Licht handeln.

Im Kontext einer TPReaktion (s. u.) kann ein TPMolekül als Edukt, Enzym, Modulator oder Produkt fungieren. In den ersten drei Formen agiert es dabei als Signaldonor, während Produkte Signalakzeptoren sind.

TPMoleküle werden ferner in verschiedene Typen aufgeteilt, die sich wiederum in eine festgelegte Hierarchie einfügen (siehe Abbildung 2). So wird auf jeder Ebene dieser Hierarchie zwischen speziesspezifischen Einträgen und Orthologen, also solchen, die sich auf keine bestimmte Spezies beziehen, unterschieden. Letztere erhalten das Präfix „Ortho“.

Die höchste Hierarchieebene stellt die der TPMolekülfamilien dar, unter der ganze Klassen von TPMoleküle subsumiert werden. Ein Beispiel ist die Familie der Transkriptionsfaktoren.

Die nächst tiefer gelegene Ebene faßt TPMoleküle in Gruppen zusammen. So gehören beispielsweise alle Splicevarianten eines Genes zur gleichen TPMolekülgruppe. Mit diesen Einträgen können somit Informationen verknüpft werden, für die nicht bekannt ist, welche der Isoformen einen beobachteten Effekt hervorruft.

Die grundlegendste Ebene wird durch die Einträge vom Typ „basic“ gebildet. Hierbei handelt es sich um tatsächlich existierende Genprodukte mit einer definierten Aminosäuresequenz.

Außerhalb dieser Hierarchie werden Komplexe geführt, die durch mehrere TPMoleküle gebildet werden und somit eine eigenständige Singalentität darstellen. Die Unterscheidung zwischen Orthologen und speziesspezifischen Einträgen bleibt jedoch auch hier erhalten.

Abbildung 2: Hierarchische Einteilung der Molecule-Einträge in TRANSPATH. Es wird zum einen zwischen Orthologen und speziesspezifischen Einträgen unterschieden. Zudem wird innerhalb dieser Gruppen zwischen einzelnen Abstraktionsschichten unterschieden. Komplexe aus mehreren TPMolekülen werden ebenfalls gesondert aufgeführt.

Neben den oben genannten Attributen ist jeder TPMolekül-Eintrag mit weiteren Attributen versehen, die zu seiner näheren Beschreibung dienen (siehe Anhang).

TPGene: TPGene waren ursprünglich als TPMoleküle annotiert und spiegeln Gene wider. Um eine klare Trennung zwischen Genen und ihren Genprodukten zu erreichen, werden sie mittlerweile jedoch gesondert aufgeführt. Die Verbindung zu TPMolekülen wird durch TPReaktionen hergestellt (s. u.), die für die Beeinflussung der Gene durch Transkriptionsfaktoren oder die Expression der entsprechenden Genprodukte stehen. Auf diese Weise wird es ermöglicht, komplette Signalwege aus TRANSPATH zu extrahieren, die z. B. bei der Aktivierung eines Rezeptors beginnen und bei der Regulation von Genen durch aktivierte Transkriptionsfaktoren enden. Ferner ist jedes TPGen mit weiteren Datenbanken, wie z.B. TRANSFAC oder UniGene verknüpft.

TPMoleküle und TPGene spiegeln somit die in TRANSPATH enthaltenen Signalentitäten wider.

TPReaktionen: Eine TPReaktion ist die Verbindung zweier Signalentitäten, also TPGenen bzw. TPMolekülen, zu einer Interaktion. Eine TPReaktion wird in dem Graphen, der dem Netzwerk zugrunde liegt, als eigenständiger Knoten modelliert, um weitere Attribute definieren zu können. So erhält jede TPReaktion ein Attribut „effect“, das die Art der Interaktion näher beschreibt. Zusätzlich wird mit „type“ zwischen „semantischen“ und „mechanistischen“ TPReaktionen unterschieden. Wird eine TPReaktion als semantisch annotiert, spiegelt sie die allgemeine Interaktion der beteiligten Signalentitäten im Sinne von Aktivierung und Inhibition im Rahmen des Signaltransduktionsnetzwerkes wider. Ein Pfad, der sich aus solchen TPReaktionen zusammensetzt, gibt also einen Überblick über den allgemeinen Signalfluß.

Mechanistische TPReaktionen hingegen beschreiben die tatsächlichen biochemischen Abläufe. Eine beteiligte Signalentität kann also nicht nur als Aktivator oder Inhibitor, sondern z. B. auch als Katalysator agieren.

Aus der Verknüpfung der oben genannten Elemente ergibt sich nun das Signaltransduktionsnetzwerk. Hierbei wird immer ein Knoten, der für ein TPMolekül oder ein TPGen steht, mit einem TPReaktions-Knoten verknüpft. Dieser enthält dann wiederum die

Verbindung zur nächsten Signalentität. Es stehen sich also zwei Klassen von Knoten gegenüber, die jeweils ausschließlich mit Knoten der anderen Klasse verbunden sind. Es handelt sich demnach um einen bipartiten Graphen. Abbildung 3 zeigt ein Beispiel.

Abbildung 3: TRANSPATH als bipartiter Graph. In dem abgebildeten Schema ist deutlich zu erkennen, daß TPGene bzw. TPMoleküle ausschließlich mit TPReaktionen verknüpft sind und umgekehrt.

3 Versuchsaufbau

3.1 TRANSPATH als relationale Datenbank

Wie bereits unter 2.8 erwähnt lässt sich TRANSPATH als lizenziertes Produkt aus dem Internet herunterladen. Die Daten liegen dabei zunächst in Form von Flatfiles vor. Diese Textdateien enthalten die Daten in einem festgelegten Format, sodaß sie automatisiert ausgelesen und weiter verarbeitet werden können. Listing 1 zeigt einen Ausschnitt einer solchen Datei. Dieses wie auch die folgenden Beispiele beziehen sich auf die Einträge für TPMoleküle. Die beschriebenen Dateiformate und Vorgehensweisen lassen sich jedoch analog auf TPGene und TPReaktionen anwenden.

```
AC  MO000019548
[...]  
NA  p53 (h)  
[...]  
TY  basic.  
HP  <MO000000277>; p53.  
  
DR  {GENOMIC}<UNIGENE:Hs.103997>.  
DR  {GENOMIC}<UNIGENE:Hs.141979>.  
DR  {GENOMIC}<UNIGENE:Hs.148111>.  
DR  {GENOMIC}<UNIGENE:Hs.152790>.
```

Listing 1: Exemplarischer Auszug aus einem Flatfile. Die Daten liegen in Form von Tupeln vor, die aus einem Bezeichner und einem Wert bestehen.

Die Daten liegen also in Form von Tupeln vor, die aus einem Bezeichner (z.B. „AC“) und einem zugeordneten Wert bestehen. Der Bezeichner steht dabei für ein bestimmtes Attribut. So steht „AC“ für das Accession-Attribut eines Eintrags. In diesem Beispiel lautet also die Accession des gezeigten Eintrags „MO000019548“. Tabelle 4 zeigt einige Beispiele für solche Accessions.

Accession	Art des Eintrags	Beschreibung
G001075	TPGen	Humanes p53-Gen
XN000006576	TPReaktion	Expression von humanem p53
MO000019548	TPMolekül	Humanes p53
MO000000277	TPMolekül	Ortholog zu p53
MO000045303	TPMolekül	Ortho-Komplex aus CtBP2, Mdm2 und p53
XN000027217	TPReaktion	Hemmung von Auror-A durch p53

Tabelle 4: Beispiele für TRANSPATH-Accessions und den ihnen zugeordneten Einträgen

Analysiert man nun die Datei Zeile für Zeile, beziehen sich alle folgenden Tupel per definitionem auf den Eintrag mit der zuletzt eingelesenen Accession, bis das nächste Tupel mit „AC“ als Bezeichner erreicht wird. Anderen Bezeichnern sind entsprechend andere Attribute zugeordnet, wobei die im Anhang enthaltenen Tabellen 11 bis 13 nur diejenigen auflisten, die hier tatsächlich verwendet wurden. Diese Art, Informationen aus Textdateien auszulesen, bezeichnet man auch als parsen, den Teil einer Software, der dieser Aufgabe übernimmt, als Parser.

Theoretisch wäre es demnach möglich, eine Anwendung mittels eines Parsers direkt auf diese Dateien zugreifen zu lassen, um TRANSPATH abzufragen. Unter praktischen Gesichtspunkten, vor allem denen der Performanz einer solchen Anwendung, erscheint diese Vorgehensweise jedoch nicht praktikabel. Vielmehr bietet sich zu diesem Zweck eine Repräsentation der Daten in Form einer relationalen Datenbank an. Auf diese Weise können auch die enormen Datenmengen, die sowohl TRANSPATH als auch Genexpressionsdaten darstellen, in akzeptabler Geschwindigkeit abgefragt werden.

Hierbei wurde auch Gebrauch von den objektrelationalen Fähigkeiten von PostgreSQL gemacht. Es werden z. B. alle TPMoleküle, die für ein Ortholog stehen, in „moleculeortho“ gespeichert. In „moleculebasic“ werden hingegen TPMoleküle abgelegt, die nicht für Orthologe stehen. Da beide Tabellen von der Tabelle „molecule“ abgeleitet werden, enthält diese sowohl Orthologe als auch Nicht-Orthologe. Will man nun im Rahmen einer Datenbankabfrage z. B. lediglich die TPMoleküle erhalten, die Orthologe darstellen, kann man sich einfach den Inhalt der „moleculeortho“-Tabelle zurückgeben lassen. So läßt sich die Art der zurückgegebenen TPMoleküle auf sehr einfache und performante Weise einschränken. Abbildung 4 stellt den entsprechenden Teil des Datenbankschemas dar.

Abbildung 4: Darstellung des Datenbankschemas, in dem die TRANSPATH-Daten gespeichert wurden. Es existieren folgende Tabellen:

1. „moleculeortho“: Enthält Einträge, die Orthologe repräsentieren.
2. „moleculebasic“: Enthält Einträge, die weder Orthologe noch TPMoleküle vom Typ „other“ darstellen (s.u.).
3. „moleculemol“: Enthält alle Einträge aus „moleculeortho“ und „moleculebasic“
4. „moleculeother“: Enthält alle Einträge, die in TRANSPATH vom Typ „other“ sind. Hierbei handelt es sich hauptsächlich um niedermolekulare Moleküle wie z. B. ATP oder um keine tatsächlichen Moleküle wie z. B. UV-Licht.
5. „molecule“: Enthält alle in TRANSPATH enthaltenen TPMoleküle.

3.2 SAGE-Daten als Relationale Datenbank

CGAP bietet die Möglichkeit, die Daten, die auf die unter 2.7.2 beschriebene Art aufbereitet wurden, über die SAGE-Genie Webseite zu nutzen. Es können z. B. einzelne Gen-zu-Tag Beziehungen abgefragt werden, oder auch ganze SAGE-Bibliotheken in Hinblick auf die differentielle Expression von Genen miteinander verglichen werden. Um die Daten aus der entwickelten Anwendung heraus in möglichst optimal angepasster Form abfragen zu können, erweist sich dieses Webinterface jedoch als ungeeignet. Für solche Zwecke steht die gesamte Datenbank auch als Download auf der FTP-Seite des CGAP zur Verfügung (<ftp://ftp1.nci.nih.gov/pub/SAGE/>). Es handelt sich um mehrere Flatfiles. Dabei entspricht eine Zeile dem Eintrag in einer Tabelle der Datenbank, wobei die Spaltenattribute des Eintrags durch Tabulator-Zeichen von einander getrennt sind. Tabelle 5 beschreibt die verwendeten Dateien.

Datei	Beschreibung
Hs.libraries	Informationen über die einzelnen SAGE-Bibliotheken, wie z. B. das verwendete Ankerenzym oder den Verweis auf Publikationen, die mit ihr in Verbindung stehen.
Hs.short.frequencies/ Hs.long.frequencies	Die Tagcounts, aufgeteilt in Longtags, also Tags der Länge 17 und Shorttags mit 10 Basenpaaren.
Hs.short.best_gene/ Hs.long.best_gene	Die jeweils verlässlichste Abbildung eines Tags auf ein bestimmtes Gen für Longtags und Shorttags.
Hs.short.best_tag/ Hs.long.best_tag	Die jeweils verlässlichste Abbildung eines Gens auf ein bestimmtes Tag für Longtags und Shorttags.

Tabelle 5: Die verwendeten CGAP-Flatfiles. Mit Ausnahme der Informationen über die verwendeten SAGE-Bibliotheken werden auf dieser Ebene die Daten aufgrund der Länge der Tags unterteilt, die den jeweiligen Experimenten zugrundeliegen.

Um die Textdateien in Tabellen einer relationalen Datenbanken umzuwandeln, könnte man nun einfach die einzelnen Zeilen der Dateien in Zeilen der entsprechenden Tabellen umwandeln. Die Dateien Hs.long.frequencies und Hs.short.frequencies enthalten z. B. pro Zeile drei durch Tabulator-Zeichen getrennte Werte:

- tag: Das Tag, dessen Anzahl angegeben ist.
- libid: Die Bibliothek, in der das Tag in der angegebenen Anzahl gefunden wurde.
- freq: Die Anzahl des Tags in der angegebenen Bibliothek.

Würde man diese Dateien aneinanderhängen und zu einer einzigen großen Tabelle, in der alle Tagcounts enthalten sind, umwandeln, erhielte man eine Tabelle mit beinahe 9 Millionen Zeilen (Stand 09/2006). Da die Tags jedoch hauptsächlich in Bezug auf ihre Häufigkeit in einer bestimmten Bibliothek abgefragt werden sollen, bietet sich eine sinnvollere Strukturierung der Daten an.

Es wurde zunächst für jede Bibliothek eine einzelne Datei erstellt, in die nur diejenigen Einträge geschrieben wurden, die sich auf die in ihr beobachteten Tagcounts beziehen. Es ergaben sich somit bei 333 Bibliotheken die entsprechende Anzahl Dateien im oben genannten Format, die dann als einzelne Tabellen eingelesen wurden. Hierbei wurde auch von den objektrelationalen Fähigkeiten des verwendeten Datenbankmanagementsystems PostgreSQL Gebrauch gemacht, indem sich alle diese Tabellen von ein und der selben Tabelle ableiten, so daß aus dieser Elterntabelle die Tagcounts auch über die Grenzen einer bestimmten Bibliothek hinaus abgefragt werden können.

Ferner wurden weitere Optimierungen vorgenommen, um die Abfrage der enormen Datenmengen in einer akzeptablen Geschwindigkeit zu ermöglichen. Beispielsweise erhält jedes Tag eine numerische ID, über die es datenbankintern effizienter identifiziert werden kann als über die Zeichenkette, die für seine Sequenz steht. Abbildung 5 gibt einen Teil des Datenbankschemas wieder.

Abbildung 5: SAGE-Daten als relationale Datenbank (Auszug). Die Tabellen mit dem Präfix „freqlib“ stehen für die jeweilige SAGE-Bibliothek mit der entsprechenden Nummer. Die Tabelle „freqs“ dient als Elterntabelle der auf die einzelnen Bibliotheken bezogenen Tabellen.

3.3 Netzwerkgenerierung

3.3.1 Identifikation der Startgene

Am Beginn der Erstellung eines Netzwerkes steht die Selektion der Expressionsdaten, die zugrunde gelegt werden sollen.

Beabsichtigt man z. B. gesundes Brustgewebe mit neoplastisch verändertem zu vergleichen, wählt man die Daten der entsprechenden Experimente aus, auf die diese Kriterien passen und erhält eine Anzahl von SAGE-Datensätzen, die auf gesundem Brustgewebe basieren und eine Anzahl, die aus neoplastisch veränderten Zellen gewonnen wurden.

Der nächste Schritt besteht darin, die erhaltenen Datensätze miteinander zu vergleichen, um Tags zu identifizieren, die für jeweils eines der beiden zugrundeliegenden Gewebe signifikant über- bzw. unterexprimiert wurden. Dabei stehen zwei Vorgehensweisen zur Verfügung:

1. Man vergleicht die Datensätze jeweils einzeln miteinander, sodaß man bei N Datensätzen insgesamt $N(N-1)/2$ Vergleiche durchführt.
2. Man gruppiert die Datensätze in zwei Untergruppen, also z. B. „gesund“ und „neoplastisch“. Diese Untergruppen lassen sich dann zu zwei „Metabibliotheken“ verschmelzen, die man daraufhin miteinander vergleicht, sodaß lediglich ein einziger Vergleich nötig ist.

Desweiteren läßt sich noch ein Schwellenwert zwischen 0 und 1 angeben, der sich auf die Signifikanz des Expressionsunterschieds bezieht. Je höher der Wert ist, desto restriktiver ist die Auswahl der Tags.

Werden zwei Datensätze *Dat1* und *Dat2* miteinander verglichen, enthält das Resultat eines solchen Vergleichs eine Liste von Tags, denen jeweils ein Signifikanzwert bzw. p-Wert zugeordnet ist.

Je kleiner der jeweilige Expressionsunterschied ist, desto kleiner ist zunächst auch entsprechend den in 2.4.1 beschriebenen Verfahren der p-Wert p_T für ein Tag T . Der Anschaulichkeit halber wird p_T daraufhin nach der Formel $p'_T = (1,0 - p_T)$ umgewandelt und mit einem negativen Vorzeichen versehen, falls T in *Dat2* überexprimiert ist. Der Betrag von p'_T ist demnach um so höher, desto signifikanter das jeweilige Tag exprimiert wurde, wobei das Vorzeichen anzeigt, ob es in *Dat1* oder in *Dat2* überexprimiert wurde. Der anfangs angegebene Schwellenwert bezieht sich auf den Betrag von p'_T . Das Resultat ist als Liste darstellbar, wie sie

in Tabelle 6 skizziert ist.

Es sei hier ausdrücklich darauf hingewiesen, daß die Signifikanzwerte, wie sie die in 2.4.1 beschriebenen statistischen Verfahren erzeugen, nicht mit den vergleichbar sind, die z. B. ein konventioneller t-Test produzieren würde, sondern auf dem Baysschen Wahrscheinlichkeitsbegriff basieren.

Tag	p-Wert
ggaccttgga	0,93
taaccaagag	0,25
..	...
cacacagttt	-0,62

Tabelle 6: Beispiele für Tags, denen Signifikanzwerte zugeordnet wurden

3.3.2 Aufteilen der Startgene

Bei der gegebenen Aufgabenstellung ist es sinnvoll, die identifizierten Startgene in zwei Gruppen zu unterteilen. Die Gruppe Nr. 1 enthält dabei diejenigen Startgene, die in Datensatz Nr. 1 überexprimiert wurden, also einen positiven p-Wert aufweisen. Entsprechend erhält Gruppe Nr. 2 die Startgene mit negativem p-Wert, so daß die beiden Gruppen zwei disjunkte Mengen darstellen. Ausgehend davon werden nach der im weiteren beschriebenen Methode zwei Netzwerke erstellt und miteinander verglichen.

3.3.3 Abbilden der Startgene auf Startmoleküle

Nach dem in 3.3.1 beschriebenen Schritt liegt somit eine Menge von Paaren aus jeweils einem Tag (z.B. taaccaagag) und einem p-Wert (z. B. 0,98) vor. Das Problem, das sich nun ergibt besteht darin, die Tags auf Genprodukte abzubilden, die in TRANSPATH enthalten sind. Um die Verbindung zwischen Tags und anderen Datenbanken herzustellen, stellt das SAGE-Projekt eine Abbildung von Tags auf UniGene-Sequenzen zur Verfügung (Lash et al. 2000). Da TRANSPATH ebenfalls eine Annotation von Molekülen zu UniGene bietet, läßt sich über UniGene die benötigte Verbindung zwischen Tags und TRANSPATH-Molekülen herstellen.

Es ergeben sich hierbei jedoch zwei grundsätzliche Probleme:

1. Es besteht die Möglichkeit, daß mehrere Tags auf ein und das selbe TPMolekül abgebildet werden, wobei sich die Frage ergibt, wie mit den einzelnen p-Werten verfahren werden soll. Hier wurde der Mittelwert aus den p-Werten mit dem entsprechenden TPMolekül assoziiert.

Die Tabelle 7 veranschaulicht die Vorgehensweise.

<i>Tag</i>	<i>p-Wert</i>	<i>TPMolekül [p-Wert]</i>
ggaccttgga	0,93	P50(h) [0,93]
taaccaagag	0,25	RHOB(h)
cacacagttt	-0,62	$[(0,25-0,62)/2 = -0,185]$
taaccaagag	0,25	TCTP(h) [0,25]

Tabelle 7: Abbilden der Startgene auf Startmoleküle. Konvergieren mehrere Tags auf ein TPMolekül, wird das arithmetische Mittel ihrer p-Werte mit ihm assoziiert.

2. Aufgrund der im Vergleich zu den bekannten Genen relativ kleinen Menge an Proteinen, die in TRANSPATH annotiert sind, kann ein gewisser Anteil der gefundenen Gene nicht auf ihre Produkte abgebildet werden. Die Bedeutung dieser Einschränkung wird sich demnach um so mehr verringern, als die Anzahl der in TRANSPATH enthaltenen Genprodukte wachsen wird. So lassen sich mit der hier verwendeten Abbildungsmethode und der TRANSPATH-Version 7.1 ca. 12% aller in den SAGE-Daten enthaltenen Tags auf ein oder mehrere TPMoleküle abbilden. Dieser Wert kann je nach Auswahl der Datensätze schwanken.

3.3.4 Abbilden der Startmoleküle auf Orthologe

TRANSPATH unterscheidet, wie bereits beschrieben, verschiedene Typen von TPMolekülen. In Bezug auf das Abbilden von Genexpressionsdaten auf TPMoleküle muß demnach bedacht werden, daß eine Abbildung, wie sie unter 3.3.3 beschrieben wurde, TPMoleküle vom Typ „basic“ ergibt. Dabei handelt es sich um Einträge, die einer bestimmten Spezies zugeordnet sind. Um einen biologisch möglichst umfassenden Graphen zu erhalten, ist es jedoch wünschenswert, die erhaltenen TPMoleküle auf Orthologebene abzubilden. Da die „basic“-TPMoleküle in TRANSPATH mit ihren jeweiligen Orthologen verknüpft sind, ist eine solche Abbildung trivial.

3.3.5 Erweiterung des Netzwerks

Nach dem in 3.3.4 beschriebenen Schritt liegt also eine Liste vor, wie sie in Tabelle 8 angedeutet ist.

<i>Molekül</i>	<i>p-Wert</i>
p50	0,93
RhoB	-0,185
...	...
TCTP	0,25

Tabelle 8: Beispiele für TPMoleküle und die ihnen zugeordneten Signifikanzwerte

Jedes dieser TPMoleküle ist in TRANSPATH mit TPReaktionen verknüpft, die wiederum zu anderen TPMolekülen führen. In diesem bipartiten Graphen, wie er in Abbildung 3 (Kapitel 2.8) dargestellt ist, lassen sich die TPReaktionen zur Veranschaulichung auch durch Kanten zwischen den TPMolekülen ersetzen. In den im Folgenden behandelten Netzwerken kann man sich demnach die TPMoleküle als Knoten und die TPReaktionen als gerichtete Kanten vorstellen. Die in Tabelle 8 dargestellte Liste läßt sich somit als kantenloses Netzwerk modellieren, in dem die gefundenen TPMoleküle die Knoten darstellen, welche jeweils einen p-Wert enthalten. Von jedem dieser Anfangsknoten aus werden nun sowohl *upstream* als auch *downstream* die mit ihm verknüpften Knoten bzw. TPMoleküle gesucht, bis die zuvor festgelegte Suchtiefe erreicht ist.

3.3.6 Färbung des Netzwerks

Die p-Werte können auch als Farbe dargestellt werden, indem der p-Wert eines Knotens einer Funktion $Col(x)$ übergeben wird, die die Farbe aufgrund des p-Werts berechnet. In diesem Fall färbt $Col(x)$ einen Knoten umso intensiver grün, desto stärker positiv sein p-Wert ist. Eine rote Färbung steht entsprechend für negative p-Werte. Knoten mit einem p-Wert nahe 0 werden demnach gelblich eingefärbt.

Das aus dem in 3.3.5 geschilderten Schritt hervorgegangene Netzwerk beinhaltet neben den Anfangsknoten diverse weitere Knoten, die TPMoleküle darstellen. Die Knoten sind über TPReaktionen darstellende gerichtete Kanten verbunden. Zu diesem Zeitpunkt sind nur die Anfangsknoten mit p-Werten versehen bzw. weisen eine Färbung auf.

Um den Rest des Netzwerks einzufärben, wird nun, ausgehend von jedem Anfangsknoten, das

Netzwerk *downstream* durchlaufen. Bei jedem Knoten N , der dabei erreicht wird, geschieht folgendes:

1. Ein Koeffizient $k = 1/d$ wird errechnet, wobei d die Anzahl der Schritte ist, die benötigt wurden, um N zu erreichen.
2. N enthält ein Feld P , dem ein p-Wert p hinzugefügt wird, der sich aus $p = p_a \cdot k$ errechnet, wobei p_a der p-Wert des Anfangsknotens ist, von dem aus N erreicht wurde.
3. N enthält außerdem ein Feld K , dem k selbst angehängt wird.

Dabei ist es natürlich möglich, daß ein Knoten von mehreren Anfangsknoten aus erreicht wird, weswegen nach Abschluß aller Durchläufe jeder Knoten ein Feld P_0^c und ein Feld K_0^c der Größe c enthält. Der Wert c ist dabei die Anzahl der Anfangsknoten, von denen aus der jeweilige Knoten erreicht wurde. Ist $c = 0$, wurde der Knoten also von keinem einzigen Anfangsknoten aus erreicht, wird dieser Knoten aus dem Graphen entfernt. Dieser Fall ist möglich, weil die unter 3.3.5 beschriebene Netzwerkerweiterung ja sowohl *upstream* als auch *downstream* durchgeführt wird. Im Gegensatz dazu spiegelt die Färbung des Graphen den kausalen Zusammenhang zwischen den einzelnen Knoten in Bezug auf Aktivierung und Inhibition wider, sodaß der Graph hierbei lediglich *downstream* durchlaufen wird.

Der p-Wert eines Knotens N errechnet sich dann aus der Formel:

$$p_n = \frac{\sum_{i=0}^{i=c} P_i}{\sum_{i=0}^{i=c} K_i}$$

Der errechnete p-Wert p_n wird dann wieder der Funktion $Col(x)$ übergeben, um die Färbung von N zu berechnen.

Die Abbildungen 6 bis 8 verdeutlichen die einzelnen Schritte.

Abbildung 6: Insulin und WOX1 wurden aufgrund des Vergleiches zwischen zwei Genexpressionsdatensätzen und anschließender Abbildung auf TPMoleküle als signifikant unterschiedlich exprimiert identifiziert. Die grüne Färbung von Insulin bedeutet dabei, daß es in Datensatz Nr. 1 überexprimiert wurde. Die rote Färbung von WOX1 entspricht hingegen einer Überexpression in Datensatz Nr. 2.

Abbildung 7: Das Netzwerk wurde um die Suchtiefe 1 erweitert. Die Anfangsknoten sind daran zu erkennen, daß sie im Durchmesser etwas größer sind. Eine blaue Färbung bedeutet, daß dem entsprechenden Knoten noch kein p-Wert zugeordnet wurde.

Abbildung 8: Die Knoten wurden auf die in 3.3.6 beschriebene Art eingefärbt. Durch die Farbgebung wird deutlich, daß z. B. p73beta sowohl von Insulin als auch von WOX1 beeinflusst wurde. IGF-1R hingegen wurde lediglich von Insulin aus erreicht, während Src von keinem der beiden Anfangsknoten erreicht wurde. Src würde also im nächsten Schritt entfernt.

3.4 Netzwerkvergleich

Das Ziel eines Vergleiches zwischen zwei verschiedenen SAGE-Bibliotheken kann es sein, die signifikant unterschiedlich exprimierten Gene zu identifizieren, die für eines der jeweiligen Gewebe besonders spezifisch sind.

Bezogen auf die Signaltransduktionsnetzwerke, die auf Grundlage eines solchen Vergleichs erstellt werden können, stellt sich darüber hinaus die Frage, welchen Einfluß die Herkunft der SAGE-Bibliotheken auf die Ähnlichkeit zwischen den aus ihnen erstellten Netzwerken hat. Man könnte z. B. annehmen, daß zwei Netzwerke, die aus zwei SAGE-Bibliotheken entstanden sind, die ihrerseits aus demselben Gewebetyp stammen, eine höhere Ähnlichkeit aufweisen als zwei Netzwerke, bei denen die zugrundeliegenden SAGE-Bibliotheken aus verschiedenartigen Geweben erstellt wurden.

Als Kriterium für die Gewebeähnlichkeit wird zum einen die Organzugehörigkeit (Pankreas, Leber, etc.) und zum anderen die Dignität (gesund, neoplastisch) gewählt.

Dann werden n SAGE-Bibliotheken ausgewählt, die daraufhin miteinander verglichen werden. Wie bereits in 3.3.2 beschrieben, werden nach jedem dieser Vergleiche die gefundenen Startgene in zwei Gruppen eingeteilt. Hierbei enthält die eine Gruppe diejenigen Startgene, die in der ersten SAGE-Bibliothek überexprimiert wurden, die zweite entsprechend diejenigen, die in der zweiten verstärkt exprimiert wurden. Daraufhin werden, ausgehend von jeder der beiden Gruppen, nach der in 3.3.5 beschriebenen Methode die jeweiligen Netzwerke erstellt. Die dabei entstandenen Netzwerke bilden nun die Grundlage für den Netzwerkvergleich.

Da die Netzwerkvergleiche kommutativ sind, also Netzwerk Nr.1 verglichen mit Netzwerk Nr. 2 dasselbe Ergebnis erzielt wie der umgekehrte Vergleich und der Vergleich eines Netzwerks mit sich selbst ebenfalls keinen Sinn ergeben würde, ergibt sich eine Gesamtzahl von $n(n-1)/2$ Netzwerkvergleichen. Jeder dieser Vergleiche wird mit drei Ähnlichkeitswerten verknüpft.

1. S_o : Ähnlichkeit in bezug auf die Organzugehörigkeit: Liegt beiden Netzwerken Gewebe aus dem gleichen Organ zugrunde ist $S_o=1$, ansonsten $S_o=-1$.
2. S_d : Ähnlichkeit in bezug auf die Dignität: Liegt beiden Netzwerken Gewebe der gleichen Dignität zugrunde ist $S_d=1$, ansonsten $S_d=-1$.

3. S_g : Ähnlichkeit in Bezug auf den dem Netzwerk zugrunde liegenden Graphen.

Das Maß, in dem sich die beiden verglichenen Netzwerke in Bezug auf die Graphen ähneln, durch die ihre jeweiligen Knoten und Kanten definiert sind, berechnet nach der Formel:

$$S_g = \frac{2V_{GC}}{(V_{G1} + V_{G2})}$$

$G1$: Netzwerkgraph Nr. 1

$G2$: Netzwerkgraph Nr. 2

GC : größter, gemeinsamer, isomorpher Teilgraph von $G1$ und $G2$

V_{G1} : Anzahl der Knoten von $G1$

V_{G2} : Anzahl der Knoten von $G2$

V_{GC} : Anzahl der Knoten von GC

Es ergibt sich somit für zwei identische Netzwerke $S_g = 1$ bzw. für zwei komplett verschiedene $S_g = 0$.

3.5 Signalperkolationscluster

Aus den unter 2.3 beschriebenen Gründen sind sowohl *weak components* als auch *strong components* zum *clustern* eines Signaltransduktionsnetzwerkes nur bedingt geeignet.

Wünschenswert wären *cluster*, die den Signalfluß der Interaktionen bzw. die Ausrichtung der Kanten des zugrundeliegenden Graphen berücksichtigen, ohne daß dabei der Großteil der Kanten verloren geht. Im konkreten Fall soll der Signalfluß von den Startknoten aus betrachtet werden.

Zu diesem Zweck wurde ein Verfahren angewandt, dessen einzelne Schritte wie folgt beschrieben werden können:

1. Es wird eine Liste L_c erstellt, die zu Beginn leer ist.
2. Von einem Startknoten N_x aus wird der Graph unter Berücksichtigung der Kantenausrichtungen durchlaufen. Jede Kante wird dabei dem durch N_x definierten *cluster* C_x hinzugefügt. Sind keine weiteren Kanten mehr erreichbar, wird C_x in der Liste L_c abgelegt. Wird dieser Schritt nun für einen weiteren Startknoten N_y zur Erstellung des *clusters* C_y wiederholt und dabei N_x erreicht, werden die Kanten von C_x zu C_y hinzugefügt und C_x aus L_c entfernt.
3. Nachdem Schritt 2 für alle Startknoten durchlaufen wurde, sind in L_c somit nur noch die

cluster enthalten, deren Startknoten nicht von anderen Knoten aus erreicht wurden. Eine Konsequenz dieser Vorgehensweise ist, daß ein Knoten in mehreren *clustern* enthalten sein kann.

Der Name Signalperkolationscluster wurde gewählt, da sich die oben geschilderte Vorgehensweise anhand eines Netzwerkes veranschaulichen läßt, bei dem die Kanten durch Kanäle und die Knoten durch Gefäße gebildet werden, die von den Kanälen durchsetzt sind, wobei die Ausrichtung der Kanten die Flußrichtung im entsprechenden Kanal wiedergibt. Die Bestimmung der *cluster* geschieht innerhalb dieses Beispiels, indem das Netzwerk nacheinander von jedem Startknoten bzw. dem ihm entsprechenden Gefäß aus mit Wasser aufgefüllt wird. Am Ende des Vorgangs werden die *cluster* durch diejenigen Startknoten bzw. Gefäße definiert, die nicht von anderen Gefäßen aus durchflossen wurden, wie es in der nebenstehenden Abbildung bei den Knoten A und B der Fall ist. Der

Knoten C hingegen gehört sowohl zu dem von A definierten *cluster* als auch zu dem, der von B aus definiert wurde.

Weitere Variationen dieser Methode könnten natürlich darin bestehen, die Anzahl der Knoten, von denen aus dem Signalfluß gefolgt werden soll, aufgrund eines anderen Kriteriums einzuschränken, etwa dem der Funktion des jeweils dargestellten TPMoleküls.

Es sei hier ausdrücklich daraufhin gewiesen, daß die Aufteilung eines Signaltransduktionsnetzwerkes in Signalperkolutionscluster weniger mathematischer Natur ist. Sie orientiert sich vielmehr an Gesichtspunkten der kausalen Verknüpfung einzelner Entitäten eines Signaltransduktionsnetzwerkes.

4 Ergebnisse

4.1 Korrelation von Netzwerkähnlichkeit und Gewebeeigenschaften

Aufgrund der bereits in der Einleitung formulierten Überlegungen, wurde die Hypothese aufgestellt, daß ein Zusammenhang zwischen der Ähnlichkeit zweier Netzwerke und der Art des Gewebes existiert, aus denen die jeweils zugrundeliegende Expressionsdaten stammen.

Zur Überprüfung dieser Hypothese wurden zunächst 110 zufällig ausgewählte SAGE-Bibliotheken bestimmt, die aus diversen Gewebearten bzw. Gewebezuständen gewonnen wurden. Daraufhin wurden $110 \cdot 109 / 2$ Bibliothekpaare gebildet und jeweils auf der Grundlage der unter 3.4 beschriebenen Prinzipien miteinander verglichen, um die differentiell exprimierten Gene zu bestimmen. Als Schwellenwert für die Signifikanz wurde dabei $p=0,95$ gewählt.

Daraufhin wurden die für die jeweilige Bibliothek signifikant überexprimierten Gene auf TPMoleküle abgebildet, sodaß zwei Graphen vorlagen, deren Knoten theoretisch zwei disjunkte Mengen von TPMolekülen repräsentierten. De facto konnte es hierbei jedoch aus den unter 3.3.3 diskutierten Gründen zu Überschneidungen kommen, wenn verschiedene Tags auf ein und das selbe TPMolekül abgebildet wurden.

Als nächstes wurden die beiden Netzwerke bis zu der Suchtiefe 6 erweitert. Hierbei muß beachtet werden, daß im Gegensatz zu der in 3.3 beschriebenen Methode, die TPReaktionen nicht durch Kanten ersetzt wurden. Nach jeweils zwei Erweiterungsschritten wurde die Netzwerkähnlichkeit ermittelt.

Es ergaben sich also $110 \cdot 109 / 2 \cdot 4 = 23980$ Netzwerkvergleiche. Daraufhin wurde zum einen die Korrelation k_O zwischen Netzwerkähnlichkeit und der Organzugehörigkeit und zum anderen die Korrelation k_D zwischen Netzwerkähnlichkeit und Dignität der betreffenden Gewebe errechnet. Beides wurde sowohl für die Gesamtheit der Netzwerkvergleiche, als auch getrennt nach Suchtiefen durchgeführt. Die Berechnung des Korrelationskoeffizienten nach Pearson und die Rangkorrelation nach Spearman ergaben dabei nahezu identische Werte, weswegen hier nur die Ergebnisse nach der Pearson-Methode wiedergegeben werden. Tabelle 9 gibt einen Überblick über die Ergebnisse.

Suchtiefe	k_O	k_D
0	0,08	0,07
2	0,07	0,04
4	0,03	0,05
6	0,03	0,04
alle	0,03	0,03

Tabelle 9: Korrelation zwischen Netzwerkähnlichkeit und Organzugehörigkeit bzw. Dignität des zugrundeliegenden Gewebes, abhängig von der Suchtiefe. Für keins der beiden Kriterien ergab sich eine nennenswerte Korrelation.

Wie sich zeigte, gab es weder in Bezug auf eine bestimmte Suchtiefe noch über alle Netzwerkvergleiche gesehen eine nennenswerte Korrelation, so daß die oben formulierte Hypothese verworfen werden mußte.

Auffällig war bei beiden Korrelationswerten, daß sie mit steigender Suchtiefe abnahmen. Tabelle 10 zeigt die Korrelation von Suchtiefe und den entsprechenden Korrelationswerten.

Korrelationswert	Korrelation zur Suchtiefe
k_O	-0,93
k_D	-0,73

Tabelle 10: Korrelation von Suchtiefe und den zuvor errechneten Korrelationen. Es zeigte sich eine deutlich negative Korrelation zwischen Suchtiefe und den auf den Netzwerkvergleichen beruhenden Korrelationswerten.

Es zeigte sich also eine deutlich negative Korrelation zwischen Suchtiefe und jeweiliger Korrelation. Ein möglicher Grund für diese Abschwächung der ohnehin nicht signifikanten Korrelationen wird klar, wenn für jede Suchtiefe der Mittelwert der Netzwerkähnlichkeit ermittelt wird, was in Tabelle 11 wiedergegeben ist.

Suchtiefe	Mittelwert von
0	0,06
2	0,07
4	0,18
6	0,41
alle	0,18

Tabelle 11: Die durchschnittliche Netzwerkähnlichkeit , abhängig von der Suchtiefe

Hier zeigte sich deutlich, daß mit steigender Suchtiefe auch die durchschnittliche Netzwerkähnlichkeit anstieg. Eine naheliegende Erklärung für dieses Phänomen dürfte in der bereits unter 2.3 diskutierten Eigenschaft realer Netzwerke liegen, sogenannte *hubs*, also Knoten einer extrem großen Inzidenz, zu bilden. Wird im Laufe der Netzwerkerweiterung von zwei, bis zu dieser Suchtiefe relativ unähnlichen Netzwerken, ein solcher *hub* erreicht, werden beiden Netzwerken im nächsten Schritt eine große Anzahl der selben Knoten hinzugefügt, sodaß die Ähnlichkeit beider Netzwerke sprunghaft ansteigen wird. Da von diesem Punkt an die folgenden Netzwerkerweiterungen zu einem großen Anteil auf diesen identischen Knoten aufbauen werden, pflanzt sich der Effekt über die Grenzen einer Suchtiefe hinaus fort. Da mit steigender Suchtiefe natürlich auch die Wahrscheinlichkeit steigt, daß es zu einem solchen Ereignis kommt, kann das oben geschilderte Phänomen damit erklärt werden. Die geschilderten Ergebnisse liefern somit auf indirekte Weise eine Bestätigung dafür, daß es sich bei den aus TRANSPATH extrahierten Netzwerken um solche handelt, die ein typisches Charakteristikum realer Netze aufweisen.

4.2 Identifikation von Schlüsselmolekülen anhand von Beispielen

Das Abbilden von Expressionsdaten auf biologische Netzwerke eröffnet zusätzliche Möglichkeiten, die Daten eines Experiments zu interpretieren. Wie bereits in der Einleitung beschrieben, können z. B. Moleküle existieren, deren zugrundeliegende Gene zwar nicht signifikant differentiell exprimiert wurden, denen jedoch aufgrund ihrer Rolle im Signaltransduktionsnetzwerk eine besondere Bedeutung zukommt. Im folgenden werden Beispiele dafür gegeben, daß die Identifikation solcher Schlüsselmoleküle mit der vorliegenden Anwendung möglich ist. Dazu wurden Moleküle gewählt, deren Schlüsselstellung bereits experimentell bestätigt wurde.

4.2.1 CXCR4

CXCR4 gehört zur Familie der Chemokin-Rezeptoren, die sich vor allem auf der Oberfläche von Leukozyten findet. *CXCR4* wurde unter anderem als Co-Rezeptor identifiziert, den *HIV-1* in Verbindung mit CD4 benötigt, um T-Zellen zu infizieren. Studien haben darüber hinaus ergeben, daß CXCR4 eine wichtige Rolle bei der Metastasierung von Brustkrebs zu spielen scheint (Müller et al. 2001). So zeigte sich, daß sein Ligand CXCL12/SDF-1alpha in genau den Organen stark exprimiert wird, in die Brustkrebszellen besonders oft metastasieren. Desweiteren konnte im Tierversuch gezeigt werden, daß die Behandlung mit CXCR4-Inhibitoren sowohl das Tumorstadium als auch die Ausbildung von Metastasen hemmen kann.

Als Ausgangsdaten wurden die SAGE-Bibliotheken gewählt, die auf der Publikation von Porter et al. (2001) basieren und verschiedene Zustände von Brustgewebe widerspiegeln. Die Bibliotheken wurden in die Gruppen „gesund“ und „krank“ unterteilt, wobei die letzteren auf verschiedenen Brustkrebsstadien basierten. Daraufhin wurden zwei Metabibliotheken gebildet und miteinander verglichen, um die signifikant unterschiedlich exprimierten Gene bzw. deren Produkte zu identifizieren. Als Schwellenwert wurde hierbei 0,95 gewählt.

Es zeigte sich, daß CXCR4 nicht zu den Molekülen zählte, deren Gene als signifikant differentiell exprimiert identifiziert wurden. Daraufhin wurde ausgehend von den Molekülen, deren Gene signifikant über- oder unterexprimiert wurden, das entsprechende Signaltransduktionsnetzwerk erstellt. Als Suchtiefe wurde hierbei 3 gewählt.

Abbildung 9 zeigt einen Ausschnitt des entsprechenden Netzwerks. Es ist deutlich zu erkennen, daß CXCR4 als stark signifikant identifiziert wurde. Die rote Färbung bedeutet in diesem Fall, daß es für krankes Gewebe spezifisch ist. Das Ergebnis bestätigt also die oben geschilderte Erkenntnis, daß CXCR4 eine wichtige Rolle bei der Behandlung von Brustkrebs spielen könnte.

Abbildung 9: Identifikation von CXCR4 als Schlüsselmolekül. Ausgehend von dem in Krebsgewebe signifikant überexprimierten MMP2 (rote Färbung), kommt es über den Liganden SDF-1 zu einer Interaktion mit CXCR4. Die in Kapitel 3 näher beschriebene Netzwerkeinfärbung ergab einen deutlich negativen p-Wert für CXCR4 (-1,0). Im Kontext des Signaltransduktionsnetzwerkes kann somit für CXCR4 eine spezifische Funktion in Brustkrebszellen postuliert werden.

4.2.2 AKT/MDM2/p53

Die Funktion von p53 als Tumorsuppressorprotein in der Zellantwort auf Streß oder DNA-Schäden ist seit langem bekannt. Hierbei reguliert p53 über die Aktivierung von Transkriptionsfaktoren zahlreiche zelluläre Prozesse wie beispielsweise die Kontrolle des Zellzyklus, DNA-Reparaturmechanismen, Apoptose, etc (Agarwal et al. 1997, Bates und Vousden 1999, Levine 1997). So konnte gezeigt werden, daß das *knock out* des *p53*-Gens eine deutlich erhöhte Häufigkeit diverser Tumoren bedingt (Donehower et al. 1992, Harvey et al. 1993). Darüberhinaus wurden in vielen Tumoren Mutationen des *p53*-Gens beobachtet (Hollstein et al. 1991, Greenblatt et al. 1994). Die Regulation von p53, als zentralen Bestandteil des Zellstoffwechsels, durch andere Proteine ist somit von besonderem Interesse.

Das MDM2-Protein wurde bereits von Momand et al. (1992) als Onkoprotein identifiziert, dessen Funktion eng mit der von p53 verknüpft zu sein scheint. Die Expression von MDM2 wird durch p53 aktiviert. Daraufhin inhibiert es das *p53*-Gen, indem es daran bindet und auf diese Weise die Rekrutierung von Transkriptionsfaktoren verhindert, die zur weiteren Expression von p53 benötigt werden. Die Interaktion zwischen MDM2 und p53 stellt somit einen negativen Feedbackmechanismus dar (Freedman et al. 1999). Da p53 seine Wirkung im Nucleus entfaltet, muß MDM2, das zunächst im Cytosol vorliegt, in den Zellkern transloziert werden. Mayo und Donner (2001) konnten zeigen, daß hierzu die Phosphorylierung von MDM2 durch die AKT-Kinase erforderlich ist, deren Rolle als *survival protein* schon seit längerem bekannt war (Ahmed et al. 1997, Kennedy et al. 1997). Es ist also zu vermuten, daß eine erhöhte Aktivität des Akt/MDM2-*pathways* zu einer gestörten Funktion von p53 und somit einem erhöhten Risiko maligner Entartung führt.

Es wurden drei Bibliotheken aus gesundem Lungengewebe und fünf Bibliotheken, die aus neoplastisch verändertem Gewebe der Lunge stammten, zu zwei Metabibliotheken vereint. Daraufhin wurden die Gene identifiziert, die mit einem p-Wert von 1,0 differentiell exprimiert wurden. Weder das *AKT*-Gen noch jenes für MDM2 befanden sich darunter. Daraufhin wurde das entsprechende Netzwerk mit der Suchtiefe 4 erstellt. Hierbei zeigte sich, daß beide Proteine, sowohl AKT als auch MDM2 als deutlich signifikant für das neoplastische Gewebe erkannt wurden ($p = -1,0$). Dieses Ergebnis könnte somit ein Hinweis darauf sein, daß der AKT/MDM2-*pathway* ein potentiell Ziel bei der Therapie des Lungenkarzinoms darstellen könnte.

Abbildung 10: Identifikation von AKT und MDM2 als Schlüsselmoleküle. Die rote Färbung von DNA-PKcs sowie der etwas größere Durchmesser zeigen an, dass sein Gen als signifikant überexprimiert in neoplastischem Gewebe identifiziert wurde. Nachdem das Netzwerk nach der in Kapitel 3 beschriebenen Methode erweitert und eingefärbt wurde, zeigte sich, daß sich sowohl für AKT als auch MDM2 ein p-Wert von -1,0 errechnete. Beide scheinen somit eine spezifische Aktivität in den untersuchten Lungenkrebszellen aufzuweisen.

5 Zusammenfassung

Die stets anwachsende Flut von Daten unterschiedlicher Art und Herkunft, die von Forschungszweigen wie Genetik oder Proteomik produziert werden, stellen immer neue Anforderungen an die bioinformatischen Verfahren, die zu ihrer Verarbeitung und Analyse benötigt werden. Eine besondere Herausforderung besteht darin, die Daten sinnvoll miteinander zu verknüpfen. Diese Form der Datenintegration ist heute eine der Hauptaufgaben der Bioinformatik.

Als Konsequenz dieser Entwicklung besteht in immer stärkerem Maße die Möglichkeit aber auch die Notwendigkeit, die enormen Datenmengen in einem eher holistischen Zusammenhang zu erfassen, der sich über die reduktionistische Unterteilung der Wissenschaft in einzelne Kategorien hinwegsetzt. Auf diese Weise wird es zunehmend möglich sein, eine Zelle nicht nur unter dem jeweiligen Gesichtspunkt dieser einzelnen Teilbereiche, sondern auf der Ebene des Systems zu interpretieren, das ihre Funktionen und ihre Eigenschaften ausmacht. Die Erforschung und Modellierung solcher Systeme hat den relativ jungen Forschungszweig der Systembiologie hervorgebracht, der auf die Entwicklung geeigneter Anwendungen zur Verknüpfung biologischer Daten angewiesen ist.

Eine Fragestellung, die in diesem Zusammenhang natürlicherweise aufgeworfen wird, besteht in der Integration von Erkenntnissen über den Expressionszustand des Genoms mit solchen über das Interaktom einer Zelle. Das Ziel dieser Arbeit besteht demnach darin, eine Anwendung zu entwickeln, die Ergebnisse aus Genexpressionsexperimenten mit dem Inhalt von Signaltransduktionsdatenbanken in sinnvoller Weise zu integrieren vermag.

Anhand der dargestellten Ergebnisse konnte gezeigt werden, daß mit der entwickelten Anwendung die Schlüsselrolle, die Molekülen wie *MDM2* oder *CXCR4* unter bestimmten Bedingungen im jeweiligen Signaltransduktionsnetzwerk zukommt, nachgewiesen werden kann. Die Anwendung kann somit z. B. als Werkzeug zur Generierung neuer Hypothesen verwendet werden, die daraufhin *in vivo* oder *in vitro* überprüft werden können.

Ein weiterer möglicher Anwendungsbereich besteht darin, einen bereits beobachteten Effekt besser nachvollziehen zu können, indem die Interaktionen, die mit ihm verknüpft sind, besser verstanden werden.

Diese Fähigkeiten wurden nur dadurch erreicht, daß der quantitative Charakter der SAGE-Daten auf das zunächst nur deskriptive Interaktom aus TRANSPATH abgebildet wurde, wodurch sich beide Datenarten synergistisch ergänzen konnten. Die Integration bereits bestehender Erkenntnisse über unterschiedliche Teilbereiche der Zellfunktion hat demnach zu einem

zusätzlichen Erkenntnisgewinn geführt.

Sicher steht man mit Anwendungen wie der hier vorgestellten noch ganz am Anfang der Entwicklung. Die Visionen, die sich mit Systembiologie verbinden, gehen jedoch bis hin zu *in silico* simulierten Zellen oder auch ganzen Organismen. Auf diese Weise könnten die komplexen Vorgänge, wie sie bei Krankheiten wie Krebs oder Diabetes angenommen werden müssen, nicht bloß anhand von Einzelbefunden wie Laborparametern oder der Regulation einzelner Gene, sondern als ganzes erfaßt werden. Der damit verbundene Erkenntnisgewinn über die Ursachen dieser Krankheiten, sowie die Möglichkeiten in Bezug auf ihre Prävention und Behandlung, wären wohl nahezu grenzenlos.

6 Anhang

6.1 Allgemeines Design der Anwendung

Die extrem großen Datenmengen und die teilweise sehr rechenintensiven Operationen, die auf sie angewendet werden müssen, um die hier beschriebenen Prinzipien anzuwenden, stellen nicht unbeträchtliche Anforderungen an die Ressourcen, auf die eine entsprechende Anwendung Zugriff haben muß. Will man eine solche Anwendung einer größeren Anzahl von Benutzern zur Verfügung stellen, wäre es unrealistisch, von jedem zu erwarten, sich beispielsweise die benötigten Datenbanken herunterzuladen und ständig auf dem neuesten Stand zu halten. Es bietet sich vielmehr an, die einzelnen Komponenten der Anwendung in einen Client und einen Server aufzuteilen. Der Vorteil dieses Konzepts liegt darin begründet, daß sich die ressourcenintensiven Elemente, wie beispielsweise die benötigte Datenbank, auf einem zentral verwalteten Server befinden, während sich der Benutzer lediglich einen Client installieren muß, über den er die Parameter für das jeweilige Experiment konfigurieren und an den Server schicken kann. Auf dem Server werden daraufhin die eigentlichen Berechnungen durchgeführt, woraufhin er das Ergebnis wiederum an den Client zurückschickt, der es für den Benutzer darstellt.

Als Übertragungsweg, auf dem der Client und der Server dabei kommunizieren, bietet sich natürlicherweise das Internet an. Hierbei wurde darauf geachtet, daß nur Technologien verwendet werden, die sowohl ausgereift als auch weit verbreitet sind. So werden die Daten in Form von serialisierten, also zur Übertragung und Speicherung geeigneten, JAVA-Objekten übermittelt. Die Serverkomponente wurde mit dem JAVA-Servlet-API als J2EE-kompatible Webapplikation implementiert, so daß die Anwendung mit jedem der J2EE-Spezifikation entsprechenden Servletcontainer (Tomcat, Jetty, etc.) verwendbar ist.

Ein weiteres Problem vieler Client-Server-Anwendungen besteht darin, daß sich gerade in großen Firmen- und Universitätsnetzwerken die Clients hinter einer sogenannten Firewall befinden, die ihnen lediglich einen eingeschränkten Zugang zum Internet gewähren. Um die Funktion der Anwendung auch in diesen Fällen zu gewährleisten, wird die Verbindung zwischen Client und Server auf dieselbe Weise realisiert, wie sie zwischen einem Webbrowser und einem Webserver geschieht, daher durch HTTP und über Port 80. Die Konsequenz daraus ist, daß jeder Benutzer, der über seinen Browser auf das Internet zugreifen kann, auch Zugriff auf den Anwendungs-Server hat. Abbildung 11 schematisiert das angewandte Konzept.

Abbildung 11: Client-Server-Design der Anwendung

Weiterhin wurde darauf geachtet, daß die Anwendung leicht um zusätzliche Funktionen erweitert werden können, bzw. ihre einzelnen Komponenten als Framework verwendbar sind, in dem sich ähnliche Anwendungen realisieren lassen. Hierbei wurde auch von vielen Merkmalen der JAVA-Plattform, wie beispielsweise dem Reflection-API, Gebrauch gemacht. Um z. B. den TPMolekülen oder TPReaktionen eine zusätzliche Annotation zu einer weiteren Datenbank hinzuzufügen, müssen lediglich zwei Klassen, die die notwendigen Schnittstellen implementieren, in die entsprechenden Verzeichnisse kopiert werden, ohne daß die gesamte Anwendung neu kompiliert werden müßte.

6.2 Handbuch

Im Folgenden soll die Funktionsweise der Anwendung aus der Sicht des Benutzers dargestellt werden. Hierzu wird der Ablauf und die Auswertung eines Experiments Schritt für Schritt beschrieben. Als Name der Anwendung wurde „DEEP – Differential Expression Effector Prediction“ gewählt.

Die entsprechende Homepage befindet sich unter:

<http://www.bioinf.med.uni-goettingen.de/services/deep>

Der JAVA-Client, auf den sich diese Anleitung bezieht, kann auch direkt unter folgender URL bezogen werden:

<http://www.bioinf.med.uni-goettingen.de/fileadmin/upload/services/deep/deep.jnlp>

6.2.1 Auswahl der Expressionsdaten

Im ersten Schritt wird entschieden, welcher Art die Daten sind, die als Grundlage des Experiments dienen sollen. Neben SAGE-Daten können auch solche aus Microarrays verwendet werden.

Abbildung 12: Auswahl der Datenart

1. Art der Daten auswählen. Es können SAGE-Daten oder Microarray-Daten verwendet werden.

6.2.1.1 SAGE-Daten

Die Verwendung von SAGE-Daten unterteilt sich in mehrere Schritte. Zunächst werden zwei Ausgangsdatensätze erstellt. Danach werden Art und Parameter der statistischen Methode eingestellt, auf deren Grundlage beide Sätze miteinander verglichen werden sollen.

6.2.1.1.1 Auswahl der SAGE-Daten

Um zwei SAGE-Datensätze zu vergleichen, müssen zunächst die SAGE-Bibliotheken ausgewählt werden, aus denen sich der jeweilige Datensatz bzw. die Metabibliothek aufbaut. Die Daten können anhand verschiedener Kriterien selektiert werden. Zum einen können SAGE-Bibliotheken ausgewählt werden, die aus dem Gewebe eines bestimmten Organs stammen. Darüberhinaus kann die Dignität der entsprechenden Gewebe eingeschränkt werden, wobei analog zur SAGE-Genie-Webseite zwischen den Kriterien „gesund“, „neoplastisch“ und „tumorassoziiert“ unterschieden wird. Ein weiteres Kriterium stellt die Länge der Tags dar, die in der jeweiligen Bibliothek zusammengefaßt wurden.

Neben den auf CGAP-Daten basierenden Bibliotheken, können für jeden Datensatz auch eigene Daten verwendet werden. Hierzu müssen eine oder mehrere Textdateien angegeben werden, in denen sich zwei durch Tabulator-Zeichen getrennte Spalten befinden. Die erste Spalte muß hierbei die UniGene-IDs, die zweite die Anzahl der Tagcounts enthalten, die für das entsprechende Gen gefunden wurden.

Compare SAGE Libraries

Select the SAGE Libraries to compare

DEEP
Differential Expression Effector Prediction

Pool A

Data

Libs from CGAP (1) shorttag (2)

☒ include ☐ exclude (3)

Tissue

kidney
liver
lung (4)
lymph node
mammary gland

Histology

cancer (5)
tumor associated
normal

Name

Pool A (6)

Pool B

Data

Lib From File (7)

add remove (8)

I:\data\lib1.dat
I:\data\lib2.dat (9)

Name

Pool B (10)

< Back Next > Finish Cancel

Abbildung 13: Auswahl der Ausgangsdaten

1. Auswahl der Herkunft der Daten für Datensatz A (in diesem Fall wurden Server-Daten ausgewählt).
2. Länge der in den auszuwählenden Bibliotheken enthaltenen SAGE-Tags
3. Legt fest, ob die festgelegten Werte ein Auschlusskriterium sind oder nicht.
4. Auswahl der Gewebeart
5. Auswahl der Dignität
6. Bezeichnung für Datensatz A
7. Auswahl der Herkunft der Daten für Datensatz B (In diesem Fall eigene Daten).
8. Hinzufügen oder Entfernen eines eigenen Datensatzes
9. Liste der ausgewählten eigenen Datensätze
10. Bezeichnung für Datensatz B

6.2.1.1.2 Feinauswahl der SAGE-Daten

Die SAGE-Bibliotheken, auf die die zuvor angegebenen Kriterien zutreffen, werden ermittelt und dem Benutzer angezeigt. Hier können einzelne Bibliotheken aus dem jeweiligen Datensatz entfernt werden, falls dies wünschenswert erscheint.

Abbildung 14: Feinauswahl der Ausgangsdaten

1. Feinauswahl der Bibliotheken für Datensatz A. Die einzelnen Spalten der Tabelle zeigen grundlegende Angaben über die jeweilige Bibliothek. Durch Klicken von „Show“ lässt sich ein Browser-Fenster zur SAGE-Genie Webseite öffnen, das detaillierte Informationen anzeigt. Durch deaktivieren des entsprechenden Häkchens lässt sich die entsprechende Bibliothek aus dem Datensatz entfernen.
2. Die unter 1. beschriebenen Funktionen für Datensatz B

6.2.1.1.3 Auswahl und Konfiguration der Statistischen Methode

Es existieren zahlreiche Methoden, SAGE-Bibliotheken im Hinblick auf die differentielle Expression einzelner Gene zu untersuchen, wie bereits in Kapitel 2.4.1 erläutert wurde. Aus diesem Grund wurde die Anwendung so konzipiert, daß die Auswahl zwischen verschiedenen Methoden möglich ist. Die entsprechenden Parameter für die ausgewählte Methode können ebenfalls hier eingestellt werden.

The screenshot shows a software window titled "Select Statistics" with a subtitle "Select the Statistical Method and Parameters". In the top right corner, there is a logo for "DEEP" (Differential Expression Effector Prediction). The main content area includes a dropdown menu labeled "SAGE-Genie Method." (marked with a circled 1), a text input field containing "Method based on Lal et al. 1999" (marked with a circled 2), and three numerical input controls: "C Value" (marked with a circled 3) set to 3, "F Value" (marked with a circled 4) set to 2, and "S Value" (marked with a circled 5) set to 1. At the bottom of the window, there are four buttons: "< Back", "Next >", "Finish", and "Cancel".

Abbildung 15: Auswahl der statistischen Methode

1. Name der Methode
2. Beschreibung der Methode
3. C-Wert (siehe 2.4.1)
4. F-Wert (siehe 2.4.1)
5. P-Wert (siehe 2.4.1)

6.2.1.2 Microarray-Daten

Obwohl der Fokus der vorliegenden Arbeit auf der Auswertung von SAGE-Daten liegt, lassen sich natürlich auch Signaltransduktionsnetzwerke aus Microarray-Daten erstellen. Die entsprechende Funktion wurde in die Anwendung integriert. Da Microarray-Daten naturgemäß bereits die Information über die Signifikanz der differentiellen Expression der einzelnen Gene enthalten, kann man sich hierbei auf die Importierung der entsprechenden Daten beschränken. Hierzu müssen die Daten in Form einer Textdatei vorliegen, in denen die einzelnen Spalten durch Tabulator-Zeichen getrennt sind. Die erste Zeile sollte eine Beschreibung der jeweiligen Spalte enthalten. Es läßt sich dann die Spalte angeben, in der sich die Werte mit den Bezeichnern für die einzelnen Gene befinden, sowie die Spalte in der die Signifikanzwerte stehen.

Import Array Data

Import Microarray data

DEEP
Differential Expression Effector Prediction

Library File (1): ... I:\data\array.dat

S Threshold (2): 0

UniGene ID	example
RESULT.SPOT as A	13
Clone ID	IMAGE:753234
Gene Symbol	ZFX
Gene Name	zinc finger protein, X-linked
Cluster ID	Hs.2074
Accession	AA406372
Preferred name	
Locuslink ID	7543
Name	13762
Sequence Type	CDNA
X Grid Coordinate (withi...	13
Y Grid Coordinate (withi...	1
Sector	1
Failed	0
Plate Number	19372

S Value	example
Channel 1 Mean Intensit...	Jul 98
Channel 2 Background (...)	216
Channel 2 Normalized (...)	14.752
Diameter of the spot	
G/R (Mean)	1.202
G/R Normalized (Mean)	.948
Log(base2) of R/G Nor...	.076
Log(base2) of R/G Nor...	
Normalized Ch2 Intensit...	0
Normalized Ch2 Net (Me...	-266
Number of Background ...	859
Number of Spot Pixels	137
Overall Intensity A (sqrt(...	3.562.247
Overall Intensity A (sqrt(...	
R/G (Mean)	.832

Lib A (5): Lib A

Lib B (6): Lib B

< Back Next > Finish Cancel

Abbildung 16: Importierung von Microarray-Daten

1. Pfad zu der Datei, die die Microarray-Daten enthält
2. Schwellenwert für den Signifikanzwert.
3. Die Spalte, die die Bezeichner (also UniGene-Cluster Ids) enthält
4. Die Spalte, die die Signifikanzwerte enthält
5. Bezeichnung für Probe A
6. Bezeichnung für Probe B

6.2.2 Erstellung des Netzwerks

Wurden die differentiell exprimierten Gene identifiziert und auf TPMoleküle abgebildet, kann abermals ein Schwellenwert für den Signifikanzwert derjenigen TPMoleküle festgelegt werden, die als Startknoten zur Netzwerkerstellung dienen sollen. Daraufhin kann das Netzwerk mit der gewünschten Suchtiefe generiert werden.

Abbildung 17: Erstellung des Netzwerks

1. Liste der TPMoleküle und ihrer Signifikanzwerte, die bei aktuell ausgewähltem Schwellenwert als Startknoten dienen würden
2. Einstellen des Signifikanzschwellenwerts
3. TPMoleküle aus der Liste der Startknoten manuell entfernen oder wieder hinzufügen
4. Aus der Liste der Startknoten manuell entfernte TPMoleküle
5. Die Suchtiefe des zu erstellenden Netzwerkes

6.2.3 Darstellung des Signaltransduktionsnetzwerkes und Verlinkung zu anderen Datenbanken

Wurde das Netzwerk entsprechend der gewählten Einstellungen berechnet, wird es gemeinsam mit diversen Zusatzinformation graphisch dargestellt. Anhand der Größe der einzelnen Knoten läßt sich erkennen, bei welchen es sich um Startknoten handelt, da diese einen größeren Durchmesser aufweisen. Die Farbe der einzelnen Knoten spiegelt ihren Signifikanzwert wider. Jeder Knoten und jede Kante kann angeklickt werden, woraufhin zusätzliche Informationen angezeigt werden. Darüberhinaus werden verschiedene Funktionen angeboten, um die Darstellung des Netzwerkes anzupassen, bzw. das Netzwerk selbst zu erweitern oder anderweitig zu modifizieren.

Abbildung 18: Darstellung des Netzwerkes

1. Wird ein TPMolekül oder eine TPReaktion ausgewählt, werden hier verschiedene Informationen angezeigt. Einige dieser Informationen sind anklickbar und stellen somit Links zu entsprechenden Internet-Ressourcen her, die in einem Webbrowser angezeigt werden (siehe 2. und 3.). Darüberhinaus werden der Name, der Typ des ausgewählten Elements, eventuell vorhandene Synonyme sowie Informationen zur Signifikanzberechnung angezeigt.
2. Der Link zur TRANSPATH-ID des ausgewählten TPMoleküls bzw. der ausgewählten TPReaktion, der zur entsprechenden Internetseite führt.
3. Verschiedene Links zu Einträgen in anderen Datenbanken, die mit dem ausgewählten Element verbunden sind.
4. Hier können die TPMoleküle anhand der Annotationen gefiltert werden, die mit ihnen verknüpft sind. Links kann festgelegt werden, ob aufgrund aller Annotationen gefiltert werden soll oder ob sich die Filterkriterien auf eine bestimmte Annotation (OMIM, Pubmed, etc.) beziehen sollen. Rechts kann der Wert eingegeben werden, nach dem gefiltert wird. Sollen z. B. nur die TPMoleküle angezeigt werden, die in einer der mit ihnen verknüpften Annotationen die Zeichenkette „cancer“ enthalten, würde man links „all“ auswählen und rechts „cancer“ eingeben.
5. Die einzelnen TPMoleküle des aktuell angezeigten Netzwerks. Neben dem jeweiligen Namen wird der Signifikanzwert und die Suchtiefe, bei der das entsprechende TPMolekül eingefügt wurde, angezeigt.
6. Die graphische Darstellung des Netzwerks. Sowohl die Knoten bzw. TPMoleküle als auch die Kanten bzw. TPReaktionen sind anklickbar, woraufhin zusätzliche Informationen zum entsprechenden Eintrag angezeigt werden (s. o.). Außerdem führt ein Rechtsklick zu einem zusätzlichen Menü, das verschiedene Funktionen in Bezug auf die Darstellung und Modifikation des Netzwerks bietet. Die Farbgebung der Knoten gibt den entsprechenden Signifikanzwert wieder.
7. Eine kurze Beschreibung der Bibliotheken, die miteinander verglichen wurden. Handelt es sich um SAGE-Bibliotheken, führt ein Klick zu der entsprechenden Internetseite bei SAGE-Genie.
8. Die Auswahl der Methode, nach der die einzelnen Cluster des Netzwerks berechnet werden. In der darunter stehenden Tabelle werden die entsprechenden Cluster angezeigt. Klickt man auf einen Cluster, wechselt die unter 6. beschriebene graphische Darstellung, so daß nur noch dieser Cluster angezeigt wird. Deselektiert man ihn, wechselt die Anzeige wieder zum gesamten Netzwerk.

ID	MO000000046
NAME	PAK1
TYPE	orthogroup
SYNONYMS	alphaPAK; p21-activated kinase 1; p65Pak; PAK; Rac effector kinase
ANNOTATIONS	pubmed omim
A VALUE	0.5
S SOURCES	VEGFR-3(h) (1.0) DNA-PKcs-xbb2(h) (1.0) TSP-1(h) (-1.0)

Abbildung 19: Beeinflussung der Netzwerkfärbung

Da die einzelnen Kanten mit zusätzlichen Informationen über die Art der TPReaktion verknüpft sind, kann der Einfluß, den die jeweilige Reaktionsart auf die Übertragung der Signifikanzwerte hat, festgelegt werden. Die Standardeinstellung besteht darin, die Werte nach dem bereits geschilderten Prinzip von einem Knoten zum anderen zu übertragen. Durch Rechtsklick auf den Graphen kann dieses Verhalten über den oben gezeigten Dialog geändert werden. Im dargestellten Beispiel wird festgelegt, daß alle Signifikanzwerte, die über eine Kante übertragen werden, die mit einer der beiden Reaktionsarten „destabilization“ oder „inhibition“ versehen sind, mit -1.0 multipliziert werden.

6.3 Verwendete TRANSPATH-Attribute

<i>Bezeichner</i>	<i>Attribut</i>	<i>Beschreibung</i>	<i>Beispiel</i>	<i>Kardinalität</i>
AC	ID	Eindeutige ID	G001075	1:1
NA	name	Klartextname	p53(h)	1:n
TY	type	Typenbezeichnung	basic	1:n
SY	synonyms	Synonyme für den jeweiligen Eintrag	tumor protein p53 (Li-Fraumeni syndrome); p53; tp53; TRP53	1:n
XB	Upstream-Reactions	Reactions, die upstream des Molecules liegen	XN000009533	n:m
XA	Downstream-Reactions	Reactions, die downstream des Molecules liegen	XN000006576	n:m
HP	Superfamily	Nächst höhere Ebene der Hierarchie	G010432	1:n
DR	Externer Datenbanklink	Die entsprechende ID in TRANSFAC	G001075	1:1
		Zugangsnummern zu anderen Datenbanken (z.B. UNIGENE), die mit dem Molecule verknüpft sind.	Hs.408312	n:m

Tabelle 12: Verwendete TPGenattribute

<i>Bezeichner</i>	<i>Attribut</i>	<i>Beschreibung</i>	<i>Beispiel</i>	<i>Kardinalität</i>
AC	ID	Eindeutige ID	MO000019548	1:1
NA	name	Klartextname	p53(h)	1:n
TY	type	Typenbezeichnung	basic	1:n
SY	synonyms	Synonyme für den jeweiligen Eintrag	ASp53; NSp53; p53; p53as; RSp53	1:n
XB	Upstream-Reactions	Reactions, die upstream des Molecules liegen	XN000006576	n:m
XA	Downstream-Reactions	Reactions, die downstream des Molecules liegen	XN000025855	n:m
HP	Superfamily	Nächst höhere Ebene der Hierarchie	MO000000277	1:n
RX	Pubmed	PMID-Nummern die zu Publikationen gehören, die mit dem Molecule in Verbindung stehen	9529249	n:m
DR	Externer Datenbanklink	Zugangsnummern zu anderen Datenbanken (z.B. UniGene), die mit dem Molecule verknüpft sind.	Hs.408312	n:m

Tabelle 13: Verwendete TPMolekülattribute

<i>Bezeichner</i>	<i>Attribut</i>	<i>Beschreibung</i>	<i>Beispiel</i>	<i>Kardinalität</i>
AC	ID	Eindeutige ID	XN000004932	1:1
NA	name	Klartextname	p53(h) -> Siah-1(h)	1:n
TY	type	Typenbezeichnung	molecular evidence; indirect	1:n
EF	effect	Effekt bzw. Art der Reaktion	activation	1:n

Tabelle 14: Verwendete TPReaktionsattribute

7 Literatur

(Agarwal et al. 1997) = Agarwal ML, Agarwal A, Taylor WR, Wang ZQ, Wagner EF, Stark GR (1997): Defective induction but normal activation and function of p53 in mouse cells lacking poly-ADP-ribose polymerase. *Oncogene* 15, 1035-41

(Ahmed et al. 1997) = Ahmed NN, Grimes HL, Bellacosa A, Chan TO, Tsichlis PN (1997): Transduction of interleukin-2 antiapoptotic and proliferative signals via Akt protein kinase. *Proc. Natl. Acad. Sci. U.S.A.* 94, 3627-32

(Audic und Claverie 1997) = Audic S, Claverie JM (1997): The significance of digital gene expression profiles. *Genome Res.* 7, 986-95

(Ball et al. 2005) = Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernandez-Boussard T, Jin H, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G (2005): The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* 33, D580-2

(Bates und Vousden 1999) = Bates S, Vousden KH (1999): Mechanisms of p53-mediated apoptosis. *Cell. Mol. Life Sci.* 55, 28-37

(Boon et al. 2002) = Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, Riggins GJ (2002): An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11287-92

(Brazma et al. 2003) = Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA (2003): ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68-71

(Chen et al. 1998) = Chen H, Centola M, Altschul SF, Metzger H (1998): Characterization of gene expression in resting and activated mast cells. *J. Exp. Med.* 188, 1657-68

(Diestel 2006) = Diestel R: Graphentheorie; Springer-Verlag, Heidelberg 2006

(Donehower et al. 1992) = Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA, Butel JS, Bradley A (1992): Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* 356, 215-21

(Eckel-Passow et al. 2005) = Eckel-Passow JE, Hoering A, Therneau TM, Ghobrial I (2005): Experimental design and analysis of antibody microarrays: applying methods from cDNA arrays. *Cancer Res.* 65, 2985-9

(Erdos und Renyi 1959) = Erdos P, Renyi A (1959): On random graphs. *Publicationes Mathematicae* 6, 290-7

(Fields und Song 1989) = Fields S, Song O (1989): A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-6

(Freedman et al. 1999) = Freedman DA, Wu L, Levine AJ (1999): Functions of the MDM2 oncoprotein. *Cell. Mol. Life Sci.* 55, 96-107

(Ganser und North 2000) = Ganser E, North S (2000): An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* 30, 1203-33

(Greenblatt et al. 1994) = Greenblatt MS, Bennett WP, Hollstein M, Harris CC (1994): Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* 54, 4855-78

(Hanisch et al. 2004) = Hanisch D, Sohler F, Zimmer R (2004): ToPNet--an application for interactive analysis of expression data and biological networks. *Bioinformatics* 20, 1470-1

(Harris et al. 2004) = Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium (2004): The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-61

(Harrison et al. 2002) = Harrison PM, Kumar A, Lang N, Snyder M, Gerstein M (2002): A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* 30, 1083-90

(Harvey et al. 1993) = Harvey M, McArthur MJ, Montgomery CA, Butel JS, Bradley A, Donehower LA (1993): Spontaneous and carcinogen-induced tumorigenesis in p53-deficient mice. *Nat. Genet.* 5, 225-9

(Hollstein et al. 1991) = Hollstein M, Sidransky D, Vogelstein B, Harris CC (1991): p53 mutations in human cancers. *Science* 253, 49-53

(Ideker et al. 2002) = Ideker T, Ozier O, Schwikowski B, Siegel AF (2002): Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1, S233-40

(International Human Genome Sequencing Consortium 2004) = International Human Genome Sequencing Consortium (2004): Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-45

(Kennedy et al. 1997) = Kennedy SG, Wagner AJ, Conzen SD, Jordán J, Bellacosa A, Tsichlis PN, Hay N (1997): The PI 3-kinase/Akt signaling pathway delivers an anti-apoptotic signal. *Genes Dev.* 11, 701-13

(Krull et al. 2006) = Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E (2006): TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* 34, D546-51

(Lal et al. 1999) = Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, Papadopoulos N, Vogelstein B, Kinzler KW, Strausberg RL, Riggins GJ (1999): A public database for gene expression in human cancers. *Cancer Res.* 59, 5403-7

(Lash et al. 2000) = Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF (2000): SAGEmap: a public gene expression resource. *Genome Res.* 10, 1051-60

(Levine 1997) = Levine AJ (1997): p53, the cellular gatekeeper for growth and division. *Cell* 88, 323-31

(Liang 2002) = Liang P (2002): SAGE Genie: a suite with panoramic view of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11547-8

(Lueking et al. 1999) = Lueking A, Horn M, Eickhoff H, Büssow K, Lehrach H, Walter G (1999): Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* 270, 103-11

(MacBeath und Schreiber 2000) = MacBeath G, Schreiber SL (2000): Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760-3

(Mayo und Donner 2001) = Mayo LD, Donner DB (2001): A phosphatidylinositol 3-kinase/Akt pathway promotes translocation of Mdm2 from the cytoplasm to the nucleus. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11598-603

(Momand et al. 1992) = Momand J, Zambetti GP, Olson DC, George D, Levine AJ (1992): The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 69, 1237-45

(Müller et al. 2001) = Müller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, McClanahan T, Murphy E, Yuan W, Wagner SN, Barrera JL, Mohar A, Verástegui E, Zlotnik A (2001): Involvement of chemokine receptors in breast cancer metastasis. *Nature* 410, 50-6

(Orchard et al. 2005) = Orchard S, Hermjakob H, Apweiler R (2005): Annotating the human proteome. *Molecular & Cellular Proteomics* 4, 435-440

(Polyak et al. 1997) = Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B (1997): A model for p53-induced apoptosis. *Nature* 389, 300-5

(Porter et al. 2001) = Porter DA, Krop IE, Nasser S, Sgroi D, Kaelin CM, Marks JR, Riggins G, Polyak K (2001): A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res.* 61, 5697-702

(R Development Core Team 2006) = R Development Core Team: R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing, Wien 2006

(Schena et al. 1995) = Schena M, Shalon D, Davis RW, Brown PO (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467-70

(Shannon et al. 2003) = Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003): Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498-504

(Stonebraker et al. 1990) = Stonebraker M, Rowe LA, Hirohama M (1990): The Implementation of POSTGRES. *IEEE Trans Knowl and Data Eng* 2, 125-42

(Velculescu et al. 1995) = Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995): Serial analysis of gene expression. *Science* 270, 484-7

(Vemuri und Aristidou 2005) = Vemuri GN, Aristidou AA (2005): Metabolic engineering in the -omics era: elucidating and modulating regulatory networks. *Microbiol. Mol. Biol. Rev.* 69, 197-216

(Vêncio et al. 2004) = Vêncio RZ, Brentani H, Patrão DF, Pereira CA (2004): Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics* 5, 119

(Wasinger et al. 1995) = Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I (1995): Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* 16, 1090-4

(Watts und Strogatz 1998) = Watts DJ, Strogatz SH (1998): Collective dynamics of 'small-world' networks. *Nature* 393, 440-2

(Xenarios et al. 2000) = Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000): DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289-91

(Zhang et al. 1997) = Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW (1997): Gene expression profiles in normal and cancer cells. *Science* 276, 1268-72

Veröffentlichung aus dieser Arbeit:

Degenhardt J, Haubrock M, Dönitz J, Wingender E, Crass T (2007): DEEP--A tool for differential expression effector prediction. Nucleic Acids Res. 35(Web Server issue):W619-24.

Danksagung

Mein Dank gilt Herrn Prof. Dr. Edgar Wingender und den Mitgliedern seiner Arbeitsgruppe, insbesondere Dr. Torsten Crass und Martin Haubrock, für die großartige Betreuung und Unterstützung.

Außerdem danke ich meiner Familie, auf deren moralischen und materiellen Rückhalt ich mich immer verlassen konnte.