Bioinformatische, genomweite Analyse genexpressionsregulierender Faktoren in Lebergewebe

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Dr. med. Andreas Teufel, M.Sc.

geb. am 17.10.1971 in Tübingen

Mainz, 2006

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 05. Juli 2006

Inhaltsverzeichnis

1 Einleitung

Die Leber ist ein zentrales Organ des gesamten Stoffwechsels und die größte Drüse des menschlichen Körpers. Die wichtigsten Aufgaben sind die Produktion lebenswichtiger Eiweißstoffe, beispielsweise Gerinnungsfaktoren, die Verwertung von Nahrungsbestandteilen insbesondere bei der Speicherung von Glukose in Form von Glykogen, die Galleproduktion und damit einhergehend der Abbau und die Ausscheidung von Stoffwechselprodukten und Giftstoffen [1]. Um eine regelrechte Leberfunktion zu ermöglichen, ist ein komplexes Zusammenspiel eines ausgedehnten Netzwerks von verschiedensten Proteinen notwendig.

Vor dem Hintergrund der Expressions-Translationsachse ist die Expression eines Gens in einem Gewebe zentrale Voraussetzung für eine konsekutive biologische Wirkungsentfaltung. Da die genomische Information in allen Geweben in gleicher Form vorliegt, verschiedene Gewebe jedoch sehr unterschiedliche Genaktivitäten benötigen, kommt der gewebsspezifischen Regulation der Genexpression offensichtlich eine essentielle Funktion zu. Dabei ist jedoch nach wie vor weitgehend unklar, ob die Gruppe der Gene, die in der Leber stark exprimiert werden, um eine regelrechte Funktion aufrecht zu erhalten, einer zentral gesteuerten „Grundexpression“ in der Leber unterliegen, oder ob diese Gene beeinflußt von individuell die Genexpression regulierenden Faktoren in der Leber exprimiert werden. Für eine Reihe einzelner genetischer Faktoren und Promotorelemente wurde in der Vergangenheit eine individuelle Regulation der Genexpression in der Leber (und anderen Geweben) gezeigt. Aufgrund der Zeit- und Kostenintensität dieser molekularbiologischen Untersuchungen blieben die Analysen jedoch stets auf einzelne Faktoren und Gene beschränkt. Eine Identifikation von zentralen Faktoren wäre allerdings von erheblicher medizinischer und biologischer Relevanz, da diesen Faktoren möglicherweise auch eine zentrale Funktion bei

der Differenzierung von hepatologischen Erkrankungen zukäme und sie somit potentielle neue therapeutische Targets darstellen und neue Therapiemöglichkeiten eröffnen könnten.

1.1 Humanes Genom, Humanes Genom Sequenzierungs-Projekt

Die Entwicklung des Humanen Genomprojekts verlief zunächst schleppend. Nach der Veröffentlichung einer Methode zur DNA Sequenzierung durch Maxam und Gilbert 1977 [2] dauerte es bis 1987, ehe in den USA die Vision eines Humanen Genom Projekts präsentiert und ab 1990 offiziell der Aufbau einer entsprechenden Datenbank begonnen wurde. 1992 kam mit der Einbindung von James Watson und unter seiner Führung dann zunehmend Bewegung in die Bemühungen, das Humane Genom zu entschlüsseln. Nach wie vor erschien die Aufgabe der Sequenzierung des gesamten Genoms jedoch kaum zu bewältigen.

Erst die Etablierung bioinformatischer Methoden des Sequenzvergleichs, insbesondere des BLAST Algorithmus [3, 4], schuf die Voraussetzung für die Entwicklung des sog. Shot-Gun-Sequenzieransatzes durch Craig Venter [5]. Unter der privaten Konkurrenz durch dessen Firma Celera Genomics kam es um die Jahrtausendwende zu einem wahren Wettrennen um die vollständige Sequenzierung des Humanen Genoms zwischen dem öffentlichen Humanen Genom Projekt und Celera Genomics, welches mit der Publikation entsprechender Draft-Sequenzen in den Fachzeitschriften *Nature* und *Science* im Frühjahr 2001 seinen vorläufigen Höhepunkt fand [6, 7].

Seither liegt das Hauptaugenmerk des Humanen Genom Projekts auf der Zusammensetzung der einzelnen, im Rahmen des Shot-Gun-Verfahrens sequenzierten, genomischen Sequenzfragmente, zu einer einzigen zusammenhängenden Sequenz der jeweiligen Chromosomen. Im letzten Update der genomischen Datenbanken (NCBI Build 35) waren bereits die Sequenzen von 14 humanen Chromosomen vollständig verfügbar [8]. Auf den

verbleibenden, noch unvollständig zusammengesetzten Chromosomen, finden sich nur noch wenige lückenhafte Abschnitte, so dass auch hier mit einer baldigen Vervollständigung der Sequenz gerechnet werden kann. Mit der Fertigstellung der Sequenz des Humanen Genoms werden von den großen Sequenzierzentren, aber auch von kleinen Labors, zunehmend auch Sequenzen anderer Organismen, insbesondere die der biologischen Modellorganismen Maus, Ratte, Fisch, Frosch, Fliege und Wurm, in die genomischen Datenbanken eingespeist. Darüber hinaus wurden in den letzten Jahren im Rahmen großer Sequenz-Expressionsstudien Millionen von EST (Expressed Sequence Tag) und SAGE (Serial Analysis of Gene Expression) Tag Sequenzen analysiert und ebenfalls in den genomischen Datenbanken abgelegt [9] [10] [11].

1.2 Genexpression

Die phänotypischen und funktionellen Unterschiede zwischen verschiedenen Zellen und Geweben hängen weitgehend von der unterschiedlichen Expression entsprechender Gene ab. Dabei findet die Regulation dieser Prozesse vorwiegend auf der Ebene der Transkriptionsinitiation statt. Darüber hinaus kann die Genexpression aber auch auf den Ebenen der Chromatinstruktur, der Verarbeitung des RNA-Transkripts, des Transports des Transkripts im Zytoplasma, der Translation der mRNA, der Stabilität der mRNA oder der Proteinstabilität reguliert werden [12].

Eukaryote Genexpression wird durch die RNA Polymerase II initiiert. Im Gegensatz zur prokaryoten RNA Polymerase II benötigt sie allerdings zuvor das Zusammenwirken eines ganzen Komplexes von allgemeinen Transkriptionsfaktoren (Abb. 1). Schon bei der Zusammensetzung dieses initialen Transkriptionskomplexes kann die Transkription eines Gens reguliert werden. Der initiale Transkriptionskomplex bindet an einen kurzen DNA-Abschnitt, die sog. Initiale Promotor Region. Dieser minimale Promotor beinhaltet häufig

eine hochkonservierte kurze Region, die sog. TATA-Box mit der Konsensussequenz TATAAA. Diese TATA-Box findet sich in der Regel 20-30 bp upstream des Transkriptionsstarts [13].

An diese Sequenz bindet das sogenannte TATA-Box Binde-Protein (TBP). Die Bindung des TBP an die TATA-Box ist wichtig für die Ausbildung des Transkriptionsinitiationskomplexes (Transkriptionsfaktoren und RNA-Polymerase II), eines Komplexes von ca. 50 verschiedenen Proteinen. Diese schließen den Transkriptionsfaktor IID (TFIID) ein. Er ist ein Komplex von TATA-Box Binde-Protein (TBP), welches die TATA-Box erkennt, 14 weiteren Faktoren, die TBP binden nicht aber DNA und Transkriptionsfaktor IIB (TFIIB), der sowohl DNA als auch die RNA Polymerase II bindet. Trotz der wichtigen Funktion der TATA-Box wird diese nur bei maximal 70% der eukaryoten Promotoren nachgewiesen [13, 14] [15].

Die generellen Transkriptionsfaktoren sind durch ihre Fähigkeit charakterisert, die Aktivität der RNA Polymerase II zu kontrollieren. Dieser initiale Transkriptionskomplex ist prinzipiell ausreichend für eine Initiierung der Transkription. Die Regulation der Expression ist jedoch im wesentlichen abhängig von weiter upstream liegenden Faktoren, den sog. cis-acting Promotorelementen. Diese cis-acting Promotorelemente finden sich häufig upstream des minimalen Promotors, in größerer Anzahl und orientierungsabhängiger Weise. An diese cis-Elemente können spezifische Transkriptionsfaktoren binden, deren Bindung an den Promotor die Expression des jeweiligen Gens stimuliert oder hemmt (Abb. 1). Die Bedeutung der verschiedenen cis-Elemente kann zwischen verschiedenen Zell- und Gewebetypen sowie unter verschiedenen physiologischen Bedingungen stark variieren. Schließlich kann eine überlappende Lokalisation von Transkriptionsfaktor-Bindungsstellen zu einer kompetitiven Situation zwischen verschiedenen Faktoren führen. Aber auch synergistische Effekte, die von

einer exakten Einhaltung eines gewissen Abstands auf DNA-Ebene abhängig sind, wurden beschrieben [16] [17].

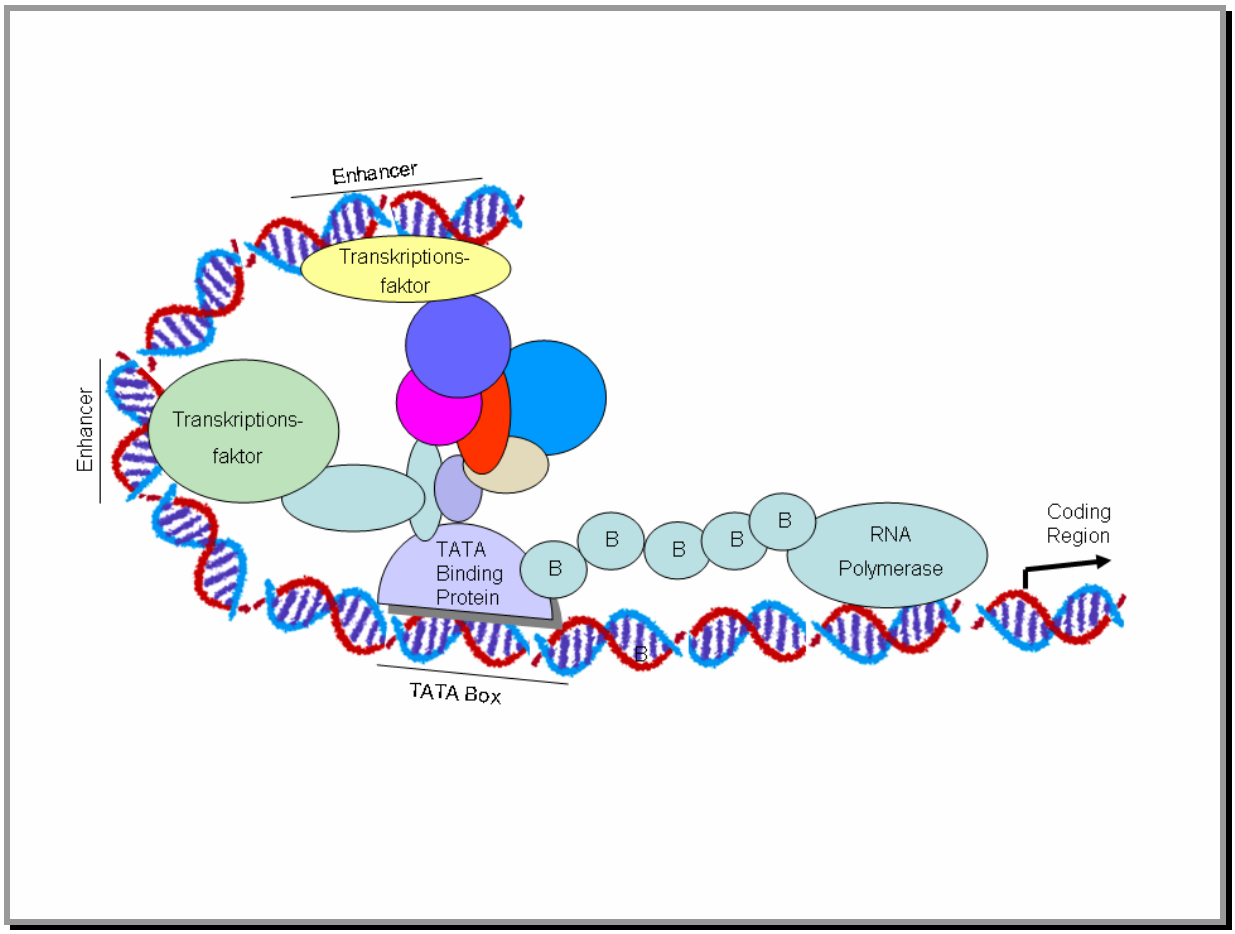

Abbildung 1: Wesentliche Bestandteile der Transkriptionsmaschinerie und deren Beteiligung an der basalen Transkription. TATA binding protein und basale Transkriptionsfaktoren (B) sind wichtig für die positionierung der RNA Polymerase II.

1.3 CpG-Islands

Neben den zuvor beschriebenen cis-Elementen der Promotorregion, die durch eine entsprechende Nukleotidsequenz spezifiziert sind, kommt der Methylierung des DNA-Rückgrats eine potentiell regulatorische Bedeutung zu. Insbesondere spielt dies im Bereich sog. CpG-Islands eine Rolle. CpG-Islands sind genomische Regionen mit statistisch erhöhter CpG-Dinukleotid-Dichte. Dabei sind Cytosine chemisch labil und können einer Desaminierung unterliegen, so dass aus methyliertem Cytosin Uracil wird. Dieser

Mechanismus hat dazu geführt, dass die Frequenz an CpG-Dinukleotiden im Genom nur ca. ein Fünftel so hoch ist, wie man sie sonst erwarten würde. In CpG-Islands ist diese Frequenz dagegen etwa so hoch wie die anderer Dinukleotide. CpG-Islands spielen eine Rolle in der differentiellen Genexpression, indem durch eine Methylierung der CpG-Islands in der Promotorregion die entsprechenden Gene oft vermindert exprimiert werden (Abb. 2). Circa 60% aller menschlichen Gene haben CpG-Islands in ihren Promotorbereichen [18] [19] [20] [21].

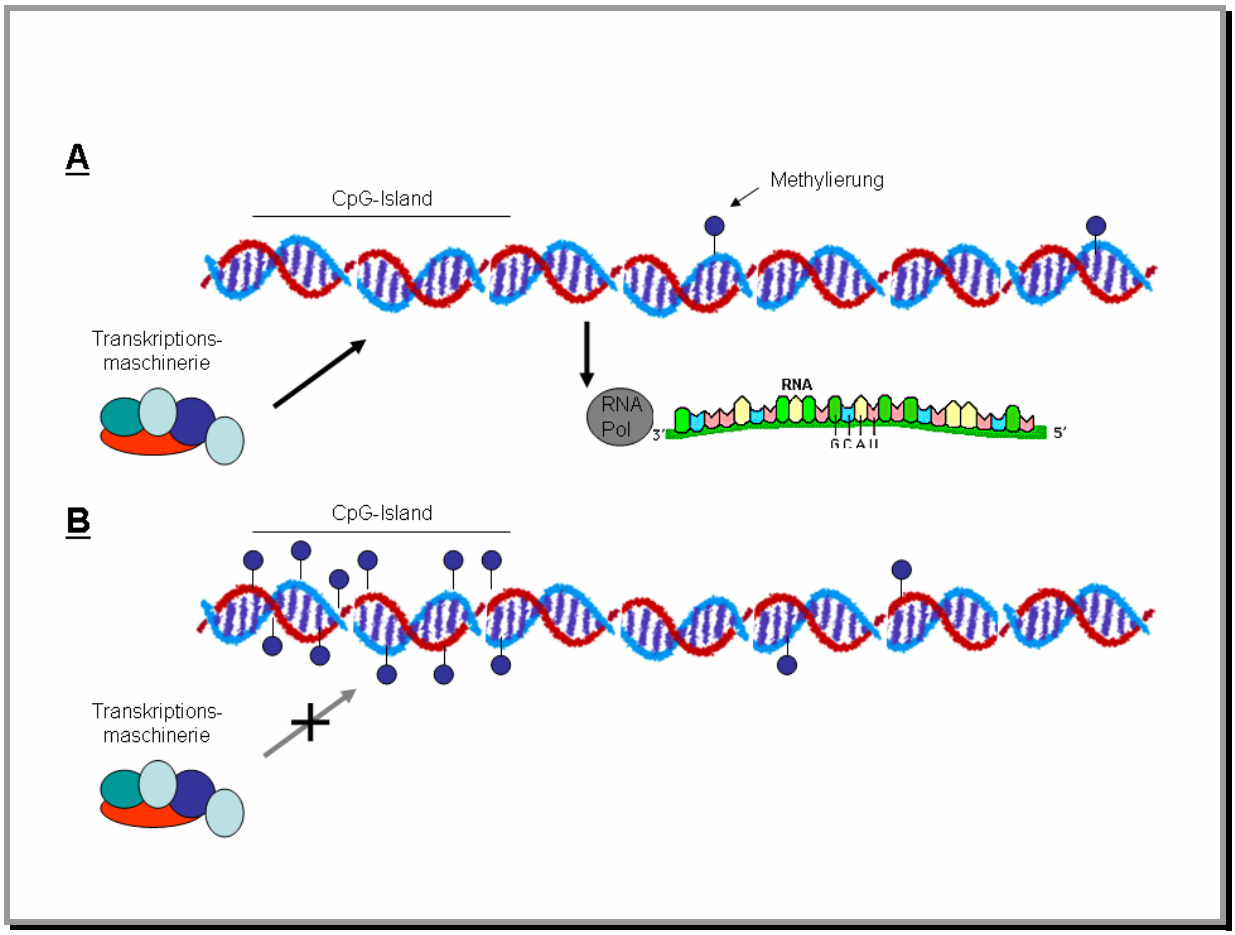

Abbildung 2: CpG-Dinukleotide in CpG-Islands können eine regulatorische Funktion im Bereich von Promotoren besitzen. (A) In unmethyliertem Zustand kann die Transkriptionsmaschinerie an den Promotor binden. (B) Methylierte CpG-Islands bieten andere Bindungsverhältnisse für Transkriptionsfaktoren als unmethylierte Abschnitte, was eine veränderte Expression, häufig eine Minderung, nach sich ziehen kann.

1.4 Differentielle Genexpression als Pathomechanismus zur Entstehung von Erkrankungen

Eine entscheidende Rolle der „core Promotorregion“, sowie der regulatorischen Promotorabschnitte und –elemente, wurde in der Vergangenheit mehrfach dokumentiert. Hinsichtlich der Bedeutung der minimalen Promotorregion stellt eines der prominentesten Beispiele die Hämophilie B Leyden dar. Patienten mit dieser Erkrankung weisen einen Mangel des plasmatischen Gerinnungsfaktors IX auf. Als Folge dieser Bildungsstörung des Faktors IX kann es bei den betroffenen Patienten zu schweren Blutungen kommen. Diese Symptomatik bessert sich nach der Pubertät, wobei dann auch die Plasmakonzentration des Faktors IX auf ca. 60% der normalen Konzentration ansteigt.

Verantwortlich für diesen Synthesedefekt dieses Gerinnungsfaktors sind Mutationen innerhalb der unmittelbar upstream des Transkriptionsstarts gelegenen 38 bp. Für mehrere Mutationen im Bereich des Promotors, an den Positionen +13, -5, -6, -20, -26, wurde eine essentielle Bedeutung für die Genexpression des Faktor IX Gens nachgewiesen. *In vitro* Transkriptionsexperimente konnten eine Minderung der Expression durch jede einzelne dieser Mutationen von ca. 40% zeigen. Für den Promotorbereich, in dem diese Mutationen liegen, mit Ausnahme der -26 Mutation, wurde in Footprinting Analysen eine Proteinbindung, als Nachweis für die Bindung entsprechender transkriptionsregulierender Proteine gezeigt. In diesem Abschnitt liegen Bindungsstellen für den Androgenrezeptor sowie den Transkriptionsfaktor HNF4. Nach der Pubertät kann eine Bindung des Androgenrezeptors den Ausfall der HNF4 Bindung kompensieren, weshalb es zu einer Besserung der Symptomatik kommt [22].

1.5 Differentielle Genexpression als therapeutisches Target und pharmakogenetischer Ansatz

Aufgrund der Bedeutung einer differentiellen Genexpression für die Pathogenese von hepatologischen Erkrankungen ist diese zugleich Ansatzpunkt therapeutischer Strategien. Im Rahmen therapeutischer Ansätze ist sowohl die vermehrte Expression von Genen möglich, die eine Krankheitsentstehung hemmen, als auch eine verminderte Expression von Genen, die eine Krankheitsentstehung fördern. Diese Ansätze müssen derzeit noch als experimentell angesehen werden, besitzen aber aufgrund ihres gezielten und neuen Ansatzes ein erhebliches Potential, die bisherigen therapeutischen Strategien weiter zu entwickeln und zu verbessern.

Hauptansatz zur vermehrten Expression von Genen ist derzeit die Expression mittels entsprechend designter viraler Vektoren. Eine Vielzahl von Studien zur gezielten Beeinflussung der differentiellen Genregulation mittels viraler Vektoren wurde hinsichtlich der Behandlung von Lebertumoren initiiert. So kann durch eine virusvermittelte Expression von Endostatin das Wachstum des Hepatozellulären Karzinoms (HCC) gehemmt werden [23]. Ferner kann die adenovirusvermittelte Genexpression von p53 die Sensitivität von HCC Zellen gegen über TRAIL hinsichtlich der Induktion des programmierten Zelltods, der Apoptose, erhöhen [24]. Auch zur Therapie von Hepatitiden bietet dieser Ansatz neue Perspektiven. Eine adenovirusvermittelte vermehrte Expression von Interferon gamma führt zu einer effektiven Kontrolle der HBV Replikation [25]. Das Hauptproblem dieses Ansatzes besteht darin, die Vektoren in jede einzelne Zelle zu schleusen. Dies ist derzeit technisch noch nicht möglich.

Zweiter wesentlicher Ansatz zur Beeinflussung der differentiellen Genregulation bei Lebererkrankungen besteht in der Expression kurzer RNA Abschnitte, sogenannter siRNA (short interfering RNA), die durch gezielte Anlagerung die Expression einzelner Gene

hemmen [26]. Die siRNA Technologie hat in den letzten Jahren einen erheblichen Aufschwung genommen. Zum Beispiel hemmen siRNAs gegen Fibroblast Growth Factor 2 (FGF-2) die Malignität von HCC-Tumorzellen [27]. Im weiteren wird durch eine Hemmung der Expression von IGF1 mittels siRNA die Empfindlichkeit von transformierten Hepatozyten gegenüber Apoptose wieder hergestellt [28]. Auch für die Therapie von Hepatitiden ist dieser Ansatz vielversprechend. So konnte zum Beispiel für siRNA, welche gegen die 5´ untranslatierte Region (5´UTR) des HCV Replicons gerichtet ist, eine deutliche Reduktion des HCV RNA Levels gezeigt werden [29].

1.6 Bioinformatik

Bioinformatische Methoden und Ansätze sind in der modernen Genetik unverzichtbar. Derzeit finden sich in den genomischen Datenbanken ca. 45 Millionen Sequenzeinträge und die Zahl der Einträge wächst derzeit mehr als exponentiell [30] [8]. Diese Flut von neuen Sequenzinformationen ist mit herkömmlichen molekularbiologischen Methoden, insbesondere aufgrund ihres hohen Zeit- und Kostenaufwands, nicht mehr zu bewältigen. Einen Ausweg aus diesem Dilemma bieten bioinformatische Algorithmen und Analysen. Entsprechend gewinnt das Gebiet der Bioinformatik zunehmend an Bedeutung.

Wesentliche Unterstützung bietet die Sparte der Bioinformatik zunächst in der Erfassung und Speicherung der gesammelten Sequenzinformationen verschiedenster Organismen in großen, vielfach über das Internet frei zugänglichen Datenbanken [8] [30]. Neben einzelner Abfrage dieser Datenbanken über eine herkömmliche Internetseite sind diese Datenbanken auch automatisiert über entsprechende, meist individuell in der Programmiersprache Perl [31] geschriebene Programme, abzufragen. Dies stellt hinsichtlich genomweiter Analysen, die mit der Komplettierung der menschlichen DNA Sequenz mehr und mehr von Interesse sind, eine Alternative von zunehmender Bedeutung dar.

Bioinformatische Methoden bieten Unterstüzung im Bereich des Vergleichs einzelner aber auch multipler Sequenzen. Diese Verfahren, z.B. das Basic Local Alignment Search Tool (BLAST, [3, 4]), waren Voraussetzung für die Entwicklung hocheffektiver Sequenzierverfahren, wie besonders des Shot-Gun-Sequenzierens [5]. Beim Shot-Gun-Sequenzieren wird das gesamte Genom zunächst mittles einer PCR-Reaktion vervielfältigt, die kurze Primer verwendet, welche sich zufällig an die DNA anlagern und eine konsekutive Amplifikation des Abschnitts ermöglichen. Auf diese Weise entstehen zusammenhangslos Millionen kurzer Sequenzfragmente, die anschließend, einem Puzzle gleichend, zusammengefügt werden müssen [5]. Diese Aufgabe gelingt nur mit Hilfe bioinformatischer Algorithmen, die gleiche, d.h. sich überlappende Sequenzabschnitte erkennen und konsekutiv zu einer Gesamtsequenz zusammenfügen. Im Rahmen von derzeit aufkommenden, neuen high-throughput Sequenziertechniken [32] [33] mit noch kürzeren Sequenz-Reads wird die Bioinformatik in diesem Bereich in den kommenden Jahren erheblich an Bedeutung zunehmen.

1.7 Genomische Datenbanken

Im Zuge des Humanen Genom Projekts haben sich eine Vielzahl von genomischen Datenbanken etabliert. Als Referenzdatenbank gilt derzeit GenBank [8], die Datenbank der National Institutes of Health (NIH) der USA. Gemeinsam mit den Datenbanken des Europäischen Bioinformatik Institutes (EBI) [30] sowie der Japanischen Datenbank (DDBJ) [34] tauschen diese Datenbanken täglich ihre Sequenzinformationen aus, so dass diese auf verschiedenen Plattformen verfügbar sind. Die entsprechenden Sequenzinformationen sind öffentlich über das Internet abrufbar. Um entsprechende Sequenzinformationen leicht und ohne wesentliche bioinformatische Vorkenntnisse zugänglich zu machen, wurden vermehrt sog. Genom-Browser entwickelt, über die Sequenzinformationen mittels einfacher Text-

Sucheingaben möglich sind (Abb. 3). Ein entsprechender Browser wird zum Beispiel vom National Center of Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov) angeboten [35]. Neben reinen genomischen Sequenzinformationen bieten diese Browser mittlerweile Links zu einer Vielzahl verschiedener Sequenzdatenbanken und bioinformatischer Algorithmen an und ermöglichen so eine effiziente Vernetzung für komplexe biomedizinische Fragestellungen.

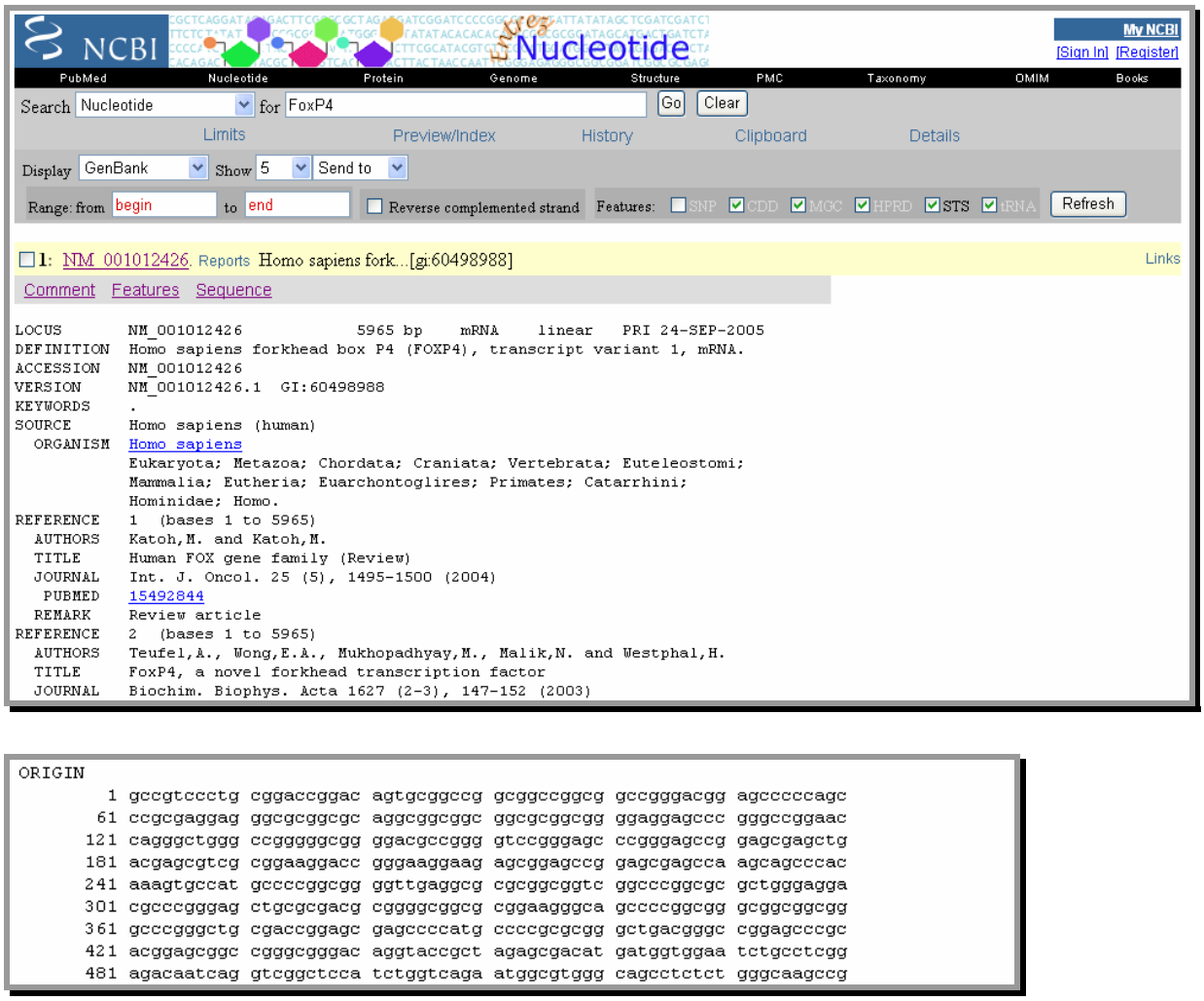

Abbildung 3: NCBI GenBank Browser. Suchergebnis für die Eingabe FoxP4 und Auswahl der entsprechenden Transkriptvariante 1.

1.8 Bioinformatische Analyse von Promtorelementen und TATA-Boxen

Die Aktivität eines Promotors ist im wesentlichen von variablen Promotorelementen abhängig, die Bindungsstellen für Transkriptionsfaktoren darstellen. Diese

Transkriptionsfaktorbindungsstellen sind durch kurze, in den meisten Fällen 5-15 bp lange Nukleotidabschnitte definiert. Die Sequenz dieser Promotorelemente lässt sich aufgrund der ihnen zugrunde liegenden Variabilität nicht allein durch deren individuelle Sequenz beschreiben. Hinsichtlich einer Automatisierung zur bioinformatischen Analyse von Promotorsequenzen bieten sich im wesentlichen zwei Möglichkeiten.

Einerseits kann die Sequenz mittels der IUPAC (International Union of Pure and Applied Chemistry) Consensus-Sequenz beschrieben werden. Die IUPAC Symbole erlauben die Darstellung einer Variablilität mittels zusätzlicher Symbole. So steht zum Beispiel der Buchstabe B für die Nukleotide C, G oder T. Wesentlich variabler ist die Darstellung von Promotorelementen mittels Positional Weight Matrizen. Eine Weight Matrix stellt eine zweidimensionale Matrix von Zahlen dar, die eine Präferenz von Nukleotiden an den jeweiligen Stellen eines Promotorelements widerspiegeln. Diese Weight Matrix wird zuvor durch ein Sequenzalignment von mehreren Bindungsstellen desselben Transkriptionsfaktors erzielt (Abb. 4). Bei der Analyse von TATA-Boxen werden diese wie Promotorelemente behandelt und entsprechend auf dieselbe Weise identifiziert.

Zur Analyse von Promotorelementen und Transkriptionsfaktorbindungstellen stehen mittlerweile verschiedene Algorithmen zur Verfügung. Insgesamt ist die bioinformatische Technologie der Vorhersage und Identifizierung von Promotorelementen noch limitiert, so dass bei der Analyse ein Kompromiss zwischen zu erzielender Sensitivität des Algorithmus und gleichzeitiger Spezifität und Validität der vorhergesagten Elemente eingegangen werden muss. Im Vergleich verschiedener Algorithmen von Promotorelementen an gut definierten Kontroll-Promotorsequenzen wurde PromotorScan als der Algorithmus mit der höchsten Spezifität identifiziert, die jedoch auf Kosten der Sensitivität erzielt wird [36].

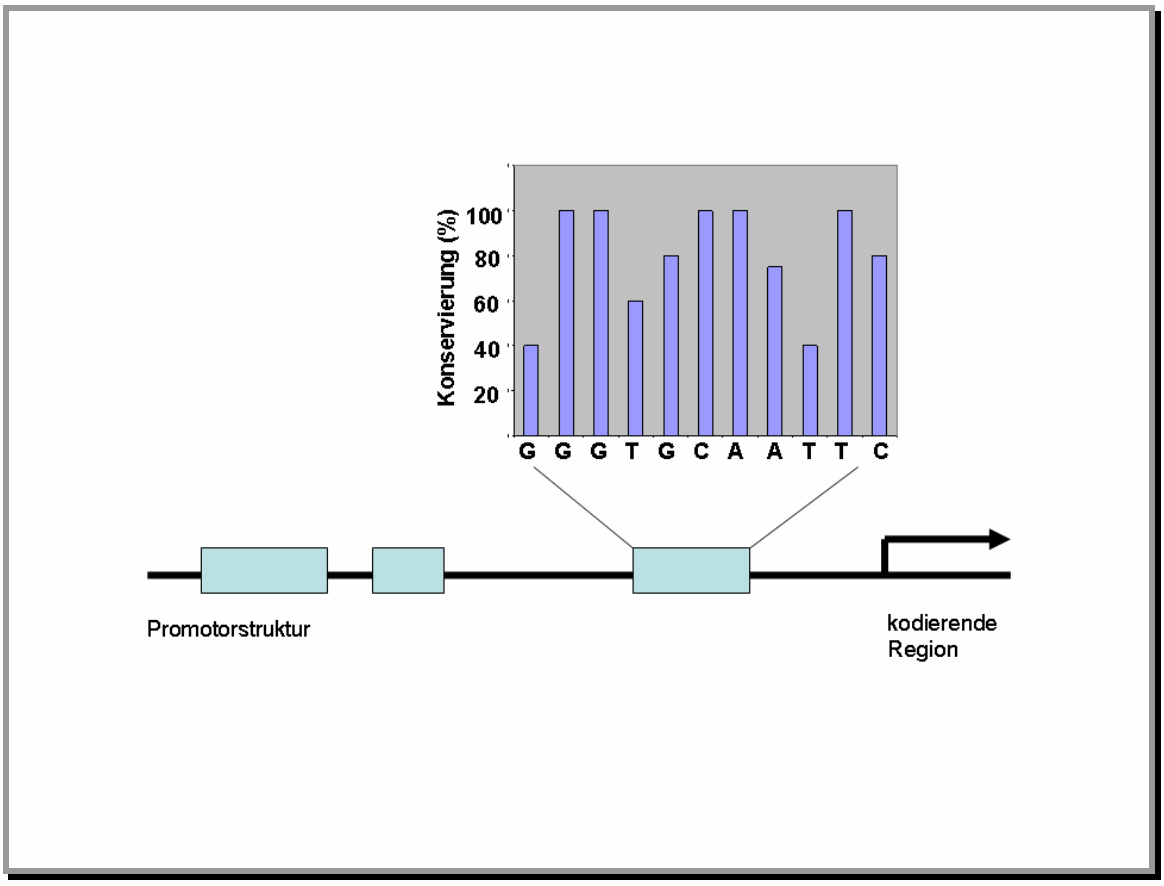

Abbildung 4: Schematische Darstellung der Promotorstruktur. Beispielhaft ist eine Transkriptionsfaktorbindungsstelle skizziert. Die Konservierung der einzelnen Nukleotide im Vergleich verschiedener Bindungsstellen desselben Transkriptionsfaktors ist als Säulendiagramm dargestellt.

1.9 Bioinformatische Analyse von CpG-Islands

Die Cytosine der meisten CpG-Dinukleotide des humanen Genoms sind methyliert. Methylierte Cytosine sind chemisch labil und können sich in der Zelle zu Uracil umwandeln. Ein so entstandenes Uracil paart sich während der DNA-Replikation mit Adenin und bei der nächsten DNA-Replikation mit Thymin. Dieser Mechanismus führt dazu, dass die Frequenz an CpG-Dinukleotiden im Genom ungefähr nur ein Fünftel der zu erwartenden Frequenz beträgt. CpG-Islands sind Regionen, die eine erhöhte Frequenz von CpG-Dinukleotiden aufweisen. Sie liegen häufig in Promotorbereichen, können aber auch auf das erste Exon übergreifen oder in gänzlich anderen Abschnitten des Humanen Genoms lokalisiert sein. Die CpG-Dinukleotide dieser Regionen sind meist nicht methyliert und entgehen daher dem

Mutationsdruck. Liegt ein CpG-Island im Promotorbereich eines Gens, so wird dieses Gen nur exprimiert, wenn keine Methylierung vorliegt. Silencing der Genexpression durch Methylierungen wird bei der Zellteilung vererbt und ist im wesentlichen irreversibel. Insgesamt besitzen ca. 60-80% aller menschlichen Gene CpG-Islands in ihren Promotorbereichen.

Das Auffinden solcher CpG-Islands mit potentiell regulatorischer Funktion geschieht auf bioinformatischem Weg typischerweise mittels eines statistischen Modells, dem Hidden Markov Modell. Ein Markov Modell bezieht sich auf eine Abfolge von Beobachtungen, bei der die Wahrscheinlichkeit einer Beobachtung von den vorausgegangenen Beobachtungen abhängig ist. Eine DNA Sequenz lässt sich als statistisches Problem einer Markov Kette darstellen, da die Wahrscheinlichkeit, eine bestimmte Base an einer Position vorzufinden, beim Durchgang der DNA Sequenz von 5´ nach 3´ von den vorausgegangenen Basen abhängig sein kann. Insbesondere in kodierenden Regionen ist es gut belegt, dass die Wahrscheinlichkeit einer Nukleotidbase von den fünf vorhergehenden Basen abhängt, was die Abhängigkeiten zusammenhängender Aminosäure-Codons widerspiegelt. In nichtkodierenden Regionen bestehen diese Abhängigkeiten nicht. Analysiert man eine unbekannte Sequenz kontinuierlich von 5´nach 3´, so lässt sich aufgrund der Nukleotidabfolge und im Vergleich mit der zu erwartenden Wahrscheinlichkeit bestimmen, ob diese Sequenz eher einer kodierenden (Exon) oder einer nicht-kodierenden (Intron) Region entspricht. In simultaner Weise kann auch das Vorkommen von CpG-Islands vorhergesagt werden [21].

1.10 Zielsetzung der vorliegenden Arbeit

Die Expression eines Gens in einem Gewebe ist zentrale Voraussetzung für eine konsekutive biologische Wirkungsentfaltung. Für eine Reihe einzelner genetischer Faktoren und Promotorelemente wurde in der Vergangenheit gezeigt, dass die Genexpression in der Leber

und auch in anderen Geweben einer individuell spezifischen Regulation unterliegt. Aufgrund der Zeit- und Kostenintensität dieser molekularbiologischen Untersuchungen blieben die Analysen jedoch stets auf einzelne Faktoren und Gene beschränkt.

Mit der Verfügbarkeit des gesamten humanen Genoms sowie dessen Expressionsdaten in großen Microarray- und SAGE-Datenbanken bietet sich die Möglichkeit, solche Regulationsmechanismen in großem, genomweitem Maßstab zu untersuchen. Dabei stellt sich insbesondere die Frage, ob es übergeordnete Faktoren gibt, die eine Expression speziell in der Leber fördern oder hemmen oder ob jedes Gen von einer unabhängigen Kombination von Faktoren reguliert wird, in dessen Summe die Expression des individuellen Gens in der Leber am stärksten erfolgt (Abb. 5). Sollten sich übergeordnete, eine Expression in der Leber stimulierende Faktoren finden, wären diese potentiell interessant für die Entwicklung neuer Therapien bei einer Reihe von Lebererkrankungen.

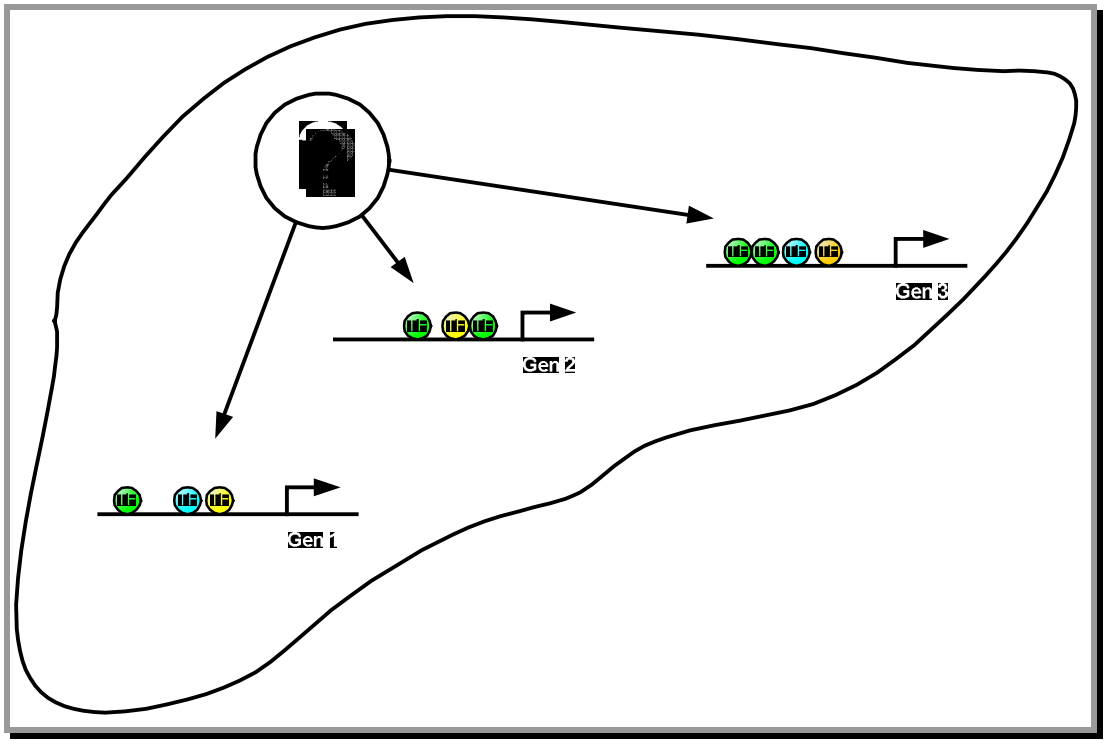

Abbildung 5: Fragestellung der Arbeit: Gibt es übergeordnete, eine Expression speziell in der Leber fördernde oder hemmende Faktoren?

2 Material und Methoden

2.1 Computer Hardware

Sämtliche bioinformatischen Programmierarbeiten und Daten-Analysen wurden auf einem handelsüblichen Personal Computer der Firma Dell mit einem Intel P4 Prozessor und 1GB RAM ausgeführt.

2.2 Datenbank

Expressionsdaten, Promotorsequenzen, Promotorelemente und TATA-Boxen wurden in einer postgreSQL Datenbank gespeichert. Auf diese wurde von einem postgreSQL Server zugegriffen (www.postgresql.org).

2.3 Microarray Daten

Microarray Expressionsdaten von 12 verschiedenen Geweben (Knochenmark, Gehirn, Herz, Niere, Leber, Lunge, Bauchspeicheldrüse, Prostata, Skelettmuskel, Rückenmark, Milz und Thymus) waren von der GeneNote Website des Weizman Institute of Science [37] verfügbar. Die Daten wurden mittels eines speziell zu diesem Zweck geschriebenen Programms heruntergeladen und in der Programmiersprache Perl [31] implementiert (“genegrabber.pl“). Kernstück dieses Programms ist das WWW::Mechanize Modul, das auf dem Webserver des Comprehensive Perl Archive Network (http://www.cpan.org) frei verfügbar ist. Unter Zuhilfenahme dieses Moduls können sämtliche Expressionsprofile der obigen zwölf Gewebe, d.h. insgesamt 62.839 individuelle Spots pro Gewebe, eines Affymetrix Array GeneChips HG-U95A-E heruntergeladen werden.

2.4 Zuordnung der Microarray-Probesets zu individuellen Genen

Die Zuordnung der Microarray-Probesets zu den individuellen Genen erfolgte über entsprechende Tabellen, welche auf der Affymetrix Website (http://www.affymetrix.com/support/technical/byproduct.affx?product=hgu95) frei zugänglich waren. In diesen Tabellen war jedes Probeset einem individuellen Gen zugeordnet. Pro Gen fanden sich minimal ein und maximal 22 Probesets auf dem Microarray.

2.5 Genomische Datenbanken, Promotorsequenzen

Genomische Sequenzen sowie RNA- und Proteinsequenzen standen über die Datenbanken des European Bioinformatics Institute (EBI)/Ensembl [30] und des National Center of Biotechnology Information (NCBI)/GenBank [8] zur Verfügung. Transkriptionstart und 5´ Promotorsequenzen wurden von der EBI/Ensembl Database geladen.

2.6 PromotorScan, Analyse von Promotorsequenzen

Promotorsequenzen wurden mittels des PromoterScan Algorithmus analysiert, welcher über einen Internetbrowser verfügbar war (http://0-thr.cit.nih.gov.portia.nesl.edu/molbio/proscan). Die Analyse der Promotorsequenzen geschah automatisiert mittels eines eigens in der Programmiersprache Perl geschrieben Programms. Kernstück dieses Programms war ebenfalls das WWW::Mechanize Modul, mit dem Informationen aus Internetseiten herausgelesen werden können.

PromoterScan [38] wurde auf der Grundlage von sog. Positional Weight Matrizen entwickelt, wobei graduelle Abstufungen hinsichtlich einer möglichen Bindungsstelle und deren optimaler Basenfolge angegeben werden. Aus diesem Grund werden gelegentlich dieselben Transkriptionsfaktorbindungsstellen erneut angegeben, allerdings um eine oder wenige Base(n) verschoben. Um eine auf dieser doppelten Vorhersage beruhende Falschzählung der

Promotorelemente zu vermeiden, wurden alle Vorhersagen desselben Elements innerhalb eines Abschnitts von 8 bp up- oder downstream ignoriert.

2.7 Selektion von Genen, deren Expression in der Leber am höchsten ist

Aus allen Probesets und aus beiden Arrays pro Gewebe wurden zunächst pro Gen die Mittelwerte gebildet. Aus diesen wurde dann der Mittelwert aller Gewebe und die diesbezügliche Standardabweichung berechnet. Danach wurden der Mittelwert der Expression des jeweiligen Gens in der Leber (Leber-Mittelwert) und der Mittelwert aller Gewebe addiert und durch die Standardabweichung geteilt [(Leber-Mittelwert + Mittelwert aller Gewebe)/Standardabweichung]. Fand sich das Ergebnis >= +1,96 so galt das entsprechende Gen als in der Leber am höchsten exprimiert, bei =< -1,96 als in der Leber am niedrigsten exprimiert.

2.8 Statistik

Aufgrund der großen Menge einzelner Promotorelementbindungsstellen und einer Vielzahl von Nukleotidkontingenten wurde eine statistische Analyse zur Erarbeitung eines statistischen Signifikanzniveaus angestrebt. Für den Fall der Zählung und Berücksichtigung aller Elemente, also auch einer Mehrfachzählung von Elementen, die wiederholt auf demselben Promotor vorkommen, läßt sich allerdings kein statistisches Verfahren zur Beurteilung einer Signifikanz anwenden.

Wird lediglich das Auffinden eines Elements auf einem Promotor ausgewertet, also das Mehrfachvorkommen von Elementen auf einem Promotor ignoriert, so läßt sich ein exakter Fischer Test zur Beurteilung der Signifikanz anwenden. Als Signifikanzniveau zur

Beurteilung der statistischen Signifikanz wurde 0,05 gewählt. Im Fisher Test errechnete P-Values < 0,05 gelten somit als signifikant.

Bei der Vielzahl der durchgeführten einzelnen Tests war eine False Discovery Rate zu berücksichtigen. Diese wurde berechnet, indem zunächst die Elemente anhand des im Fisher Test erzielten P-Values absteigend sortiert wurden. Für diese Faktoren wurde dann das Signifikanzniveau ermittelt, indem in derselben Reihenfolge für jeden Faktor einzeln ein um eine False Discovery Rate bereinigtes Signifikanzniveau berechnet wurde, indem das initiale Niveau von 0,05 absteigen durch 182 bis 1 dividiert wurde. Auf diese Weise wurde jedem Faktor ein neues Signifikanzniveau zugeschrieben. Lag der P-Value unter dem Signifikanzniveau, welches um die False Discovery Rate bereinigt war, so galt der Faktor als signifikant unterschiedlich verteilt.

3 Ergebnisse

In der vorliegenden Arbeit wurde untersucht, ob die Genexpression speziell in der Leber übergeordneten Faktoren unterliegt, oder ob jedes Gen von einer unabhängigen Kombination von Faktoren reguliert wird, als dessen Summe die Expression des individuellen Gens in der Leber am stärksten ist. Sollten sich übergeordnete, die Expression in der Leber stimulierende Faktoren finden, wären diese potentiell interessante, neue therapeutische Targets für multiple Lebererkrankungen unterschiedlicher Genese.

3.1 Download von Expressionsdaten

Zur Untersuchung dieser Fragestellung wurden zunächst aus einem Affymetrix Microarray Datenset die Expressiondaten für 12 individuelle Gewebe (Knochenmark, Gehirn, Herz, Niere, Leber, Lunge, Bauchspeicheldrüse, Prostata, Skelettmuskel, Rückenmark, Milz und Thymus) von jeweils insgesamt 62.839 Arrayspots extrahiert (Abb. 6). Dabei waren die Ergebnisse für jedes Gewebe doppelt verfügbar, was eine höhere Zuverlässigkeit der Daten für jedes Gewebe bedeutet. Die zugehörigen Bezeichnungen für die einzelnen Spots des Affymetrix Array GeneChips HG-U95A-E waren auf der Affymetrix Website verfügbar. Sie wurden initial komplett von dieser gewonnen und in eine postgreSQL Datenbank geschrieben. Für jedes einzelne dieser individiellen Datensets und Spots wurde die GeneNote Website [37] mittels eines eigens geschriebenen Programms abgefragt, implementiert in der Programmiersprache Perl [31]. Nach erfolgreicher Abfrage der Website wurde aus dieser das Ergebnis extrahiert und in die postgreSQL Datenbank geschrieben. Anschließend wurden die einzelnen Spots, für die mittels des Arrays Expressionsdaten verfügbar waren, individuellen Genen zugeordnet.

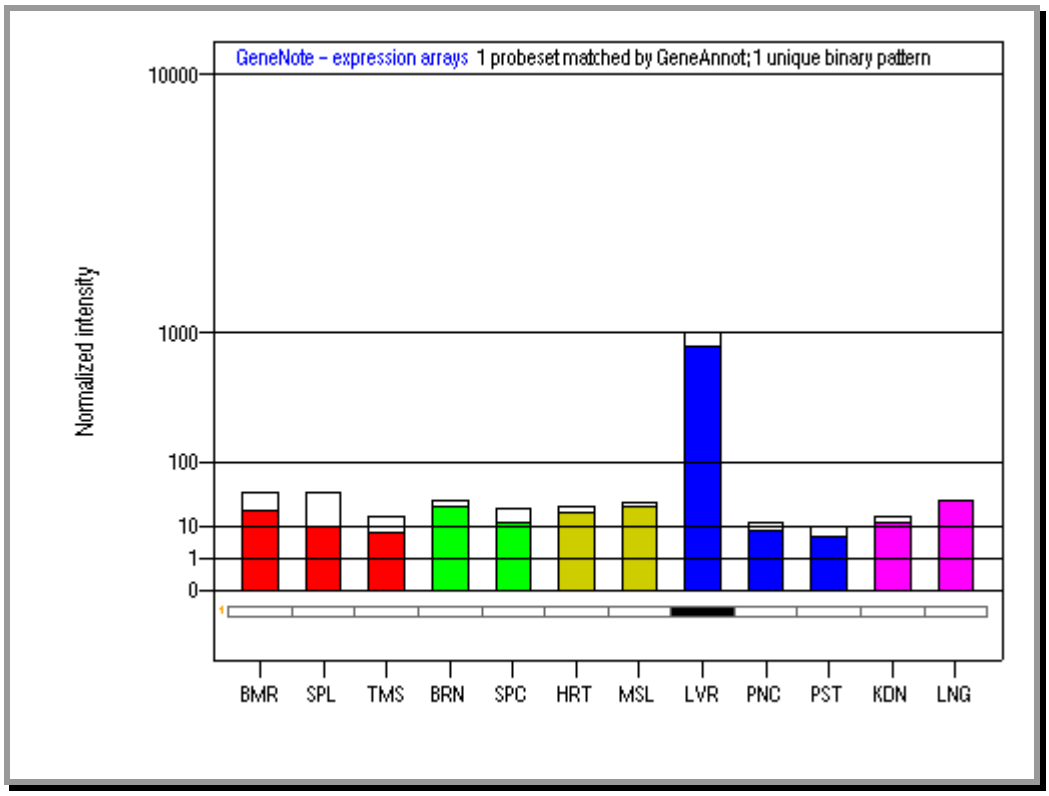

Abbildung 6: GeneNote Expressionsprofil für ASGR1, welches eine mit Abstand höchste Expression des Gens in der Leber (LVR) zeigt.

Eine entsprechende Zuordnung ist ebenfalls auf der Affymetrix Website verfügbar. Sie wurde von der Website extrahiert und den bestehenden Zuordnungen entsprechend in die vorhandene postgreSQL Datenbank geschrieben. Anhand der von Affymetrix bereitgestellten Zuodnungstabelle wurden die 62.839 Microarray-Probesets des Arrays insgesamt 15.472 individuellen Genen zugeordnet. Somit waren die meisten Gene auf dem Array in mehreren Microarray-Probesets repräsentiert, wobei die Anzahl der pro Gen vorhandenen Spots unterschiedlich war und aus mindestens einem, maximal 22 Microarray-Probesets bestand.

3.2 Gene mit höchster Expression in Lebergewebe

Um die individuellen Gene den Geweben mit ihrer höchsten Expression zuordnen zu können, wurde zunächst aus allen für das jeweilige Gen zur Verfügung stehenden Spots eine mittlere Expression errechnet. Dieser Mittelwert wurde zusätzlich in die postgreSQL Datenbank

aufgenommen. Anschließend wurden alle Gene der Datenbank danach beurteilt, ob deren höchste Expression in der Leber oder in einem anderen Gewebe zu finden war. Um einen deutlichen Unterschied zwischen den beiden zu untersuchenden Gruppen von Genen, also der Gruppe mit höchster Expression in der Leber und der mit höchster Expression in anderen Geweben, zu erhalten, wurde eine klare Abgrenzung der Gene, welche am höchsten in der Leber exprimiert sind wünschenswert. Deshalb wurde ein restriktives Kriterium für die Zuordnung zu einer höchsten Expression in der Leber gewählt: Als Gene mit höchster Expression in der Leber wurden nur solche gewertet, deren Expressionswert in der Leber nach Abzug des durchschnittlichen Expressionswerts dieses gleichen Gens in allen anderen Geweben und nach Teilung durch die Standardabweichung aller Proben >= +1,96 war. War der entsprechende Wert <= -1,96, so wurde er als niedrigste Expression in Lebergewebe berücksichtigt. Anhand dieser Kriterien wurden insgesamt 778 Gene als am höchsten in Lebergewebe exprimiert und nur 14 als am niedrigsten in Lebergewebe exprimiert eingestuft.

3.3 Algorithmus der Datenanalyse

Zum Vergleich des Expressionsniveaus mit potentiell die Genexpression regulierenden Faktoren, wurden in einem zweiten Schritt die Promotorsequenzen der einzelnen zugehörigen Gene in dieselbe Datenbank geschrieben. Da Promotoren bzw. das Vorkommen regulatorischer Elemente in einer upstream des Transkriptionsstarts liegenden Promotorregion nicht definiert sind, wurde im folgenden für die Analyse von Promotorelementen eine Region 1000 bp upstream des Transkriptionsstarts von der genomischen Datenbank des Europäischen Bioinformatik Instituts (EBI, [30]) in der postgreSQL Datenbank gespeichert. Lediglich für die Analyse der CpG-Islands wurden Regionen von 5000 bp upstream des Transkriptionsstarts zusätzlich in der Datenbank abgelegt und ausgewertet.

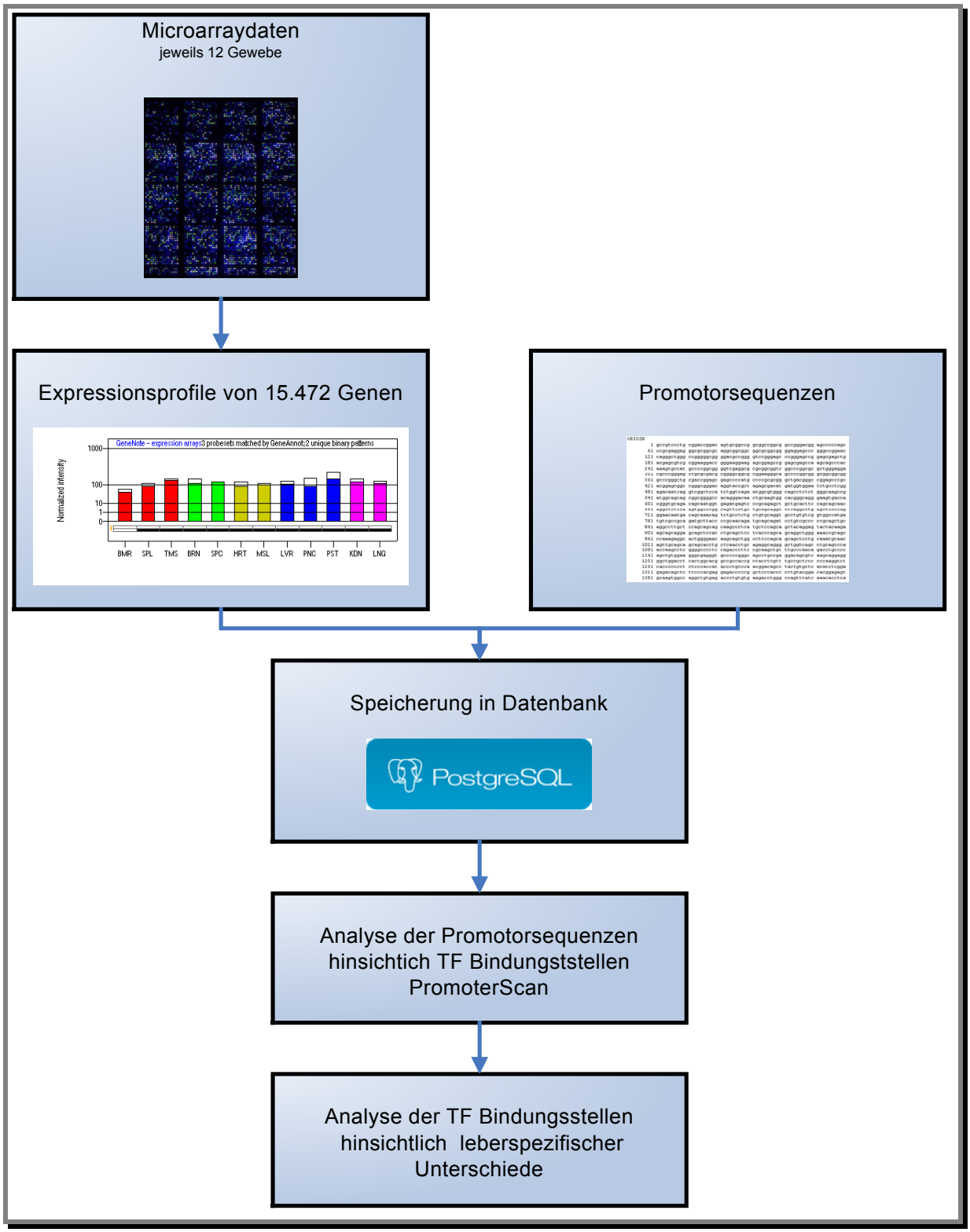

Abbildung 7: Speisung der postgreSQL Datenbank mit Expressions- und Sequenzdaten zur Korrelation von Promotorelementen und Genexpression.

Nicht für alle Gene waren in den genomischen Datenbanken Promotorsequenzen verfügbar. Insgesamt konnten Promotorregionen für 14.644 Gene aus der Datenbank des EBI geladen werden. Erneut wurden Sequenzen automatisiert heruntergeladen mittels eines eigens in Perl implementierten Programms, dessen Kernstück erneut das WWW::Mechanize Perl-Modul der CPAN Bibliothek war.

Die heruntergeladenen Promotorsequenzen wurden auf potentielle Transkriptionsfaktorbindungsstellen und TATA-Boxen untersucht. Zur Untersuchung wurde der PromotorScan Algorithmus [38] gewählt, welcher Promotorsequenzen unter Verwendung von Positional Weight Matrizen hinsichtlich des Vorkommens potentieller Promotorelemente analysiert. Der entsprechende Algorithmus war über einen Web Browser verfügbar, so dass die einzelnen Promotorsequenzen automatisiert durch diese Website geschleust und die Ergebnisse aus der entsprechenden Website extrahiert wurden. Die Speicherung dieser Ergebnisse erfolgte ebenfalls in der postgreSQL Datenbank. Dabei wurden sämtliche vorhergesagten Transkriptionsfaktorbindungsstellen in die Analyse einbezogen, also sowohl Promotorelemente auf dem forward als auch auf dem reverse Strang (Abb. 7).

3.4 Analyse von Promotorsequenzen auf das Vorkommen potentieller Transkriptionsfaktorbindungsstellen

Auf insgesamt 7042 Promotorsequenzen wurden Transkriptionsfaktorbindungsstellen identifiziert. Da PromotorScan [38] auf der Grundlage von Positional Weight Matrizen entwickelt wurde, kommt es bei der Vorhersage von möglichen Bindungsstellen zu einer graduellen Abstufung der Bindungswahrscheinlichkeit eines an die Sequenz bindenden Faktors. Dabei kann es zu einer mehrfachen Vorhersage derselben Bindungsstelle kommen, verschoben um lediglich einige wenige Basen. Um eine auf dieser mehrfachen Vorhersage beruhende Falschzählung der Promotorelemente zu vermeiden, wurden alle Vorhersagen

desselben Elements innerhalb eines Abschnitts von 8 bp up- oder downstream ignoriert. Somit fand sich eine Gesamtzahl von 241.984 Transkriptionsfaktorbindungsstellen.

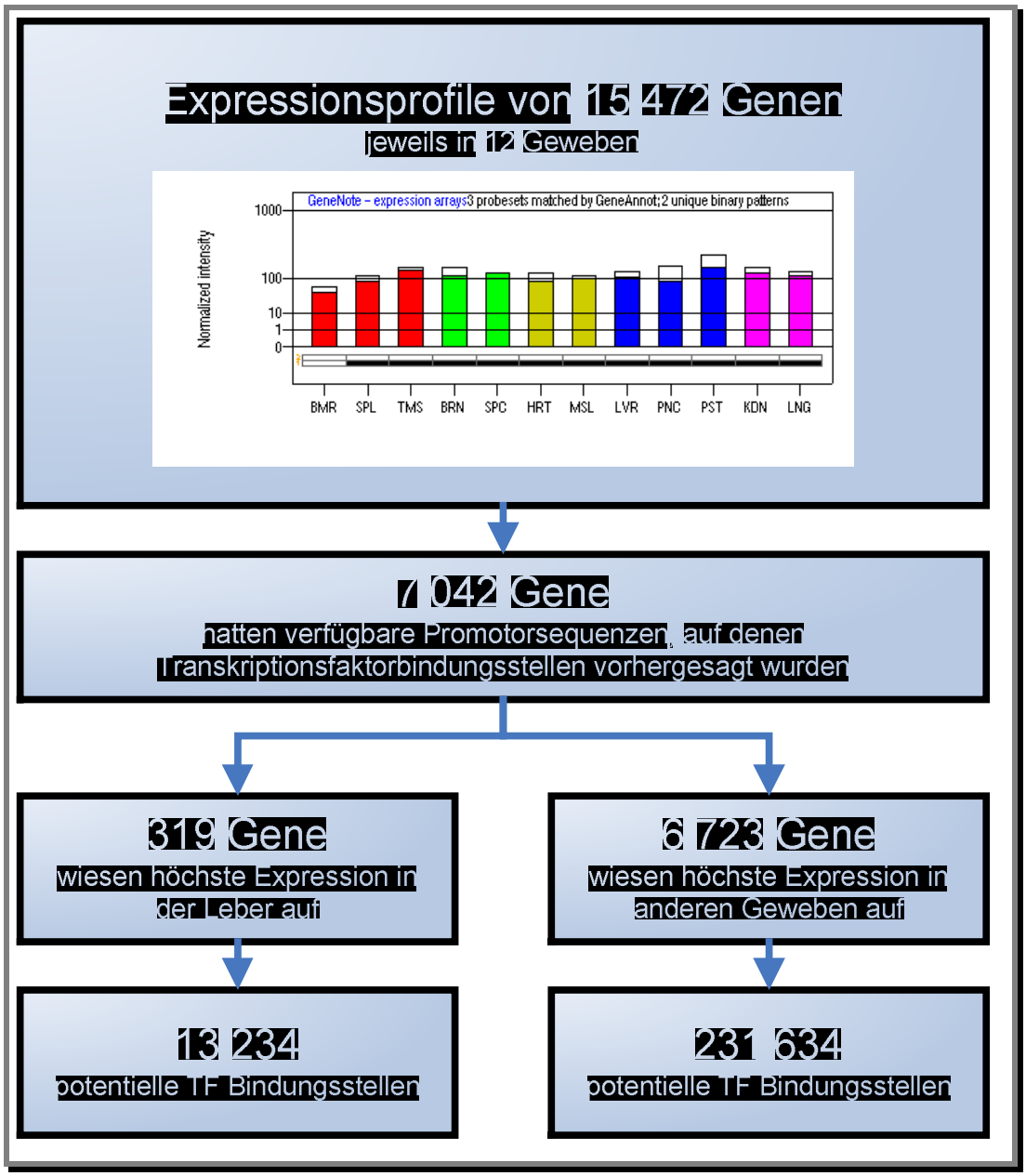

Abbildung 8: Selektionierung von Genen mit höchster Expression in der Leber und Auswertung von Transkriptionsfaktorbindungsstellen.

Anhand der Microarray-Expressionsdaten wurde die Gesamtgruppe der verfügbaren Gene und Promotoren in zwei Gruppen unterteilt. Die erste Gruppe beinhaltet Gene, deren höchste

Expression in der Leber zu finden war. Die zweite Gruppe beinhaltet die Gene, die in anderen Geweben am höchsten exprimiert waren. Danach wurde jeder potentiell bindende Transkriptionsfaktor auf unterschiedliches Vorkommen in diesen beiden Gruppen untersucht (Abb. 8). Dies geschah unter der Vorstellung, dass übergeordnete Faktoren, die eine Expression in der Leber stimulieren, in der Gruppe der Gene, die in der Leber am höchsten exprimiert sind, verhältnismäßig wesentlich häufiger zu finden sein sollten. Eine solche unterschiedliche Verteilung zwischen diesen beiden Gruppen war jedoch für keinen einzigen Faktor nachzuweisen. Insgesamt wurden auf diese Weise 182 individuelle, potentiell bindende Transkriptionsfaktoren, die in PromotorScan [38] verfügbar waren, auf Unterschiede in diesen beiden Gruppen untersucht (Abb. 9) .

Promotorelement	Leber	andere Gewebe	Quotient
MT/G-box	2	1	42,1504702
tf-LF1	1	1	21,0752351
HNF-3	2	2	21,0752351
PRDI-BF1	2	4	10,5376176
HNF-1	4	8	10,5376176
PEA2	3	8	7,90321317
OBF	3	13	4,86351579
Y_box_(1)	1	5	4,21504702
MRE_CS6	2	12	3,51253918
ICSBP	5	31	3,39923147
Albumin_US2	3	20	3,16128527
NFMHCIIB	2	14	3,01074787
HNF-4	1	8	2,63440439
LF-A2	1	10	2,10752351
TRF	2	21	2,00716525
GATA-1	10	109	1,93350781
MRE_CS4	5	56	1,88171742
ApoE_A	3	34	1,85957957
GH0	2	23	1,83262914
Ga	6	69	1,83262914
MT-I.4	1	12	1,75626959
gERE	1	12	1,75626959
MT-I.3	3	37	1,70880285

OTF-2A	5	62	1,69961573
retroviral_TATA	10	126	1,67263771
Ig_cd	6	81	1,56112853
TFII-I	4	55	1,53274437
AABS_CS2	12	173	1,46186602
CTF/NF-1	52	776	1,41225802
MEF-I	1	15	1,40501567
Y	31	483	1,35265484
junB-US2	46	724	1,33903427
CP1	12	189	1,33811017
APRT-CHO_US	59	932	1,33416188
ZDNA_CS	1	16	1,31720219
dc_box	4	64	1,31720219
Oct	6	97	1,30362279
HNF1	16	259	1,30194503
UCE.1	20	330	1,27728698
Ig_dc	2	33	1,27728698
NFI	56	926	1,27452826
CTF	123	2039	1,27133591
NRE_Box2_CS	5	83	1,26959248
HSV-tk-2nd_distal_si	9	150	1,26451411
H4TF-2	2	34	1,23971971
PEBP2	49	901	1,14615596
IFN-inducible_CS	3	56	1,12903045
glucagon-G3A	12	227	1,11410935
HLA_DQ-beta_US	1	19	1,1092229
PEA1	9	172	1,10277393
NF-I	2	39	1,08078129
element_II_rs-3	10	196	1,0752671
TFIID	55	1079	1,07427056
EGR-2	48	956	1,0581708
SP-C	1	20	1,05376176
2-Oct	1	20	1,05376176
1-Oct	2	40	1,05376176
ISGF1	1	20	1,05376176
DBP	1	20	1,05376176
NF-E2	2	41	1,02806025
HSV_IE_repeat	49	1009	1,02347524
HC1	4	83	1,01567398
(Sp1)	650	13733	0,99751714
EARLY-SEQ1	512	10923	0,9878715
T-Ag	553	11932	0,97675201
AP-2	1518	32798	0,97543164
C/EBP	3	65	0,97270316
APRT-mouse_US	262	5684	0,97144821
Sp1	2198	47714	0,97085482
myosin-specific	50	1086	0,97031469

PuF	142	3141	0,95278045
NF-kB	83	1841	0,95015998
GCF	910	20398	0,94021296
JCV_repeated_sequenc	714	16011	0,93983623
CAC-BP	12	273	0,92638396
ENKTF1	3	69	0,91631457
NF-GMa	4	93	0,90646173
(SRF)	19	444	0,90186817
H4TF1	40	944	0,89301844
IgHC.2	8	189	0,89207344
USF	3	71	0,89050289
UCE.2	702	16840	0,87855196
SDR_RS	39	944	0,87069298
KROX24	144	3489	0,86982914
MLTF	30	728	0,86848496
AP-1	39	950	0,86519386
SIF	50	1220	0,86373914
MBF-I	52	1271	0,86224408
SRF	6	147	0,86021368
ETF	51	1250	0,85986959
INF.1	25	619	0,85118074
NF-S	16	403	0,8367339
EGR-1	82	2075	0,83285266
MT-I.6	7	178	0,82880138
Oct-factors	10	255	0,82647981
EF-1A	2	52	0,81058597
alpha-1-AT-X-RBP	1	26	0,81058597
GH-TRE	5	131	0,80439829
USF-IIEII	9	236	0,80371659
NGFI-C	13	357	0,76744554
TTR_inverted_repeat	122	3372	0,76250851
MyoD	1	28	0,75268697
EIIF	4	114	0,73948193
NF-D	4	116	0,72673225
ATF	71	2126	0,70382958
Lymphokine_CS	3	90	0,70250784
CREB	101	3079	0,69132795
OBP	4	124	0,67984629
ICSbf	2	62	0,67984629
ApoE_B2	2	63	0,66905508
Lam_B2_US	9	286	0,6632067
element_II_rs-4	6	192	0,6586011
c-fos.5	11	363	0,63864349
c-Myc	3	105	0,60214957
E4TF1	20	709	0,59450593
E4F1	41	1471	0,58741308
EIIaE-A	2	74	0,56960095

ATF/CREB	45	1690	0,5611749
EivF	27	1023	0,55623788
c-fos_US5	25	948	0,55578152
EivF/CREB	30	1167	0,54177982
oct-B1A	1	39	0,54039064
beta-pol_CS	26	1078	0,50830808
TCR-beta_decamer	2	84	0,50179131
AP-4	1	46	0,45815728
HC5	2	95	0,44368916
E2F	9	433	0,43805339
beta-glob_US	1	49	0,43010684
beta-pol_RS	7	344	0,42885653
MRE-BP	1	102	0,20661995
XY_CS.2	0	11	0
TEF1	0	22	0
T3R	0	1	0
SP-E	0	5	0
SOD-1-GC-hexamer	0	8	0
RFX	0	1	0
RF-X	0	1	0
R2	0	1	0
PRDI-BF	0	1	0
PR	0	21	0
Pit-1	0	1	0
PGK-1-US	0	1	0
PGK-1-NF1-like_US	0	1	0
NGFI-B	0	10	0
NF-Y	0	8	0
NF-I/Sp1	0	1	0
NF-InsE2	0	4	0
NF-IL2-D	0	3	0
NF-IL2-A	0	3	0
NFe	0	8	0
NF1	0	22	0
NF-1	0	20	0
Myc-CF1	0	3	0
MGF	0	10	0
LIT-1	1	0	0
Ker1	0	33	0
ITF	0	1	0
ISRE_CS	0	4	0
IgNF-A	0	53	0
Ig_core	0	21	0
IBP-1	0	13	0
hsp70_US	0	22	0
HLAII.Xbox	0	1	0
GR	0	1	0

GHF-1-GH-site_2	0	1	0
GHF-1	0	1	0
GArC3	0	30	0
GArC1	0	29	0
EIIaE-B'	0	28	0
CYP21B_US	0	1	0
CD28RC	0	1	0
CBF	0	14	0
CATT-BP	0	17	0
b-globin	0	26	0
B2	0	12	0
B1	0	19	0
ApoCIIIp1	0	19	0
ApoCIII_NE	1	0	0
APF	1	0	0
AP-3	0	3	0
(AP-1)	0	1	0
alpha-1-AT-X-Hapt	0	9	0

Abbildung 9: Häufigkeiten der einzelnen Promotorelemente in Genen, die am höchsten in der Leber exprimiert sind (Leber) und Genen die am höchsten in anderen Geweben exprimiert sind (andere Gewebe). In der Spalte Quotient sind ist das relative Vorkommen unter Berücksichtigung der unterschiedlichen Größe der Gengruppen berechnet.

Eine statistische Auswertung der so generierten Daten war nicht möglich. Es gibt keinen statistischen Test, der das gegebene Problem adäquat abbildet. Eine statistische Aufarbeitung ist nur für den Fall möglich, dass das Vorkommen eines Elements auf dem Promotor lediglich mit „ja“ oder „nein“ bewertet wird, jedoch die absolute Häufigkeit des Vorkommens einer Transkriptionsfaktorbindungsstelle auf einem Promotor ausser Acht bleibt.

Deshalb wurden im nächsten Schritt die Transkriptionsfaktorbindungsstellen lediglich im Hinblick auf ein Vorkommen, also nicht auf die absolute Zahl der Elemente auf den Promotoren der entsprechenden Gengruppen untersucht. Zur Untersuchung der Signifikanz des Unterschieds für eine jeweilige Transkriptionsfaktorbindungsstelle wurde anschließend ein exakter Fischer Test angewandt. In einer initialen Analyse fanden sich mehrere Faktoren,

die unter diesem Aspekt signifikant unterschiedlich verteilt waren (Abb. 10, mit Stern markiert). Bei 182 einzelnen Testungen muß jedoch eine False Discovery Rate berücksichtigt werden, die berechnet wurde, indem das im Fisher Test zu erzielende Signifikanzniveau in absteigender Reihenfolge durch Faktoren zwischen der Gesamtzahl der analysierten, individuellen Faktoren und 1 dividiert wurde. Lag der p-Value anschließend unter dem berechneten Signifikanzniveau, war von einer Signifikanz auszugehen.

Zieht man die False Discovery Rate in Betracht, kann keiner der zuvor noch signifikant unterschiedlich verteilten Faktoren als signifikant angesehen werden (Abb. 10). Insgesamt konnte somit in der Analyse bekannter Promotorelemente durch den PromotorScan Algorithmus [38] kein signifikanter Unterschied in der Verteilung von Transkriptionsfaktorbindungsstellen nachgewiesen werden.

Promoterelement	P-Value	Signifikanzniveau mit False Discovery Rate
HNF-1	0,001532*	0,000282486
MT/G-box	0,005953*	0,000284091
PEA2	0,01158*	0,000285714
ICSBP	0,0119*	0,000287356
TTR_inverted_repeat	0,01686*	0,000289017
EivF/CREB	0,02065*	0,000274725
ATF/CREB	0,02413*	0,000290698
PRDI-BF1	0,02719*	0,000276243
E4TF1	0,02792*	0,000292398
Y	0,03273*	0,000277778
OBF	0,03326*	0,000294118
E4F1	0,03348*	0,00027933
APF	0,0453*	0,000295858
ApoCIII_NE	0,0453*	0,000280899
LIT-1	0,0453*	0,000297619
E2F	0,04954*	0,000299401
EivF	0,05642	0,000301205
c-fos_US5	0,05945	0,00030303
CTF	0,07023	0,000304878

beta-pol_CS	0,08266	0,000306748
Albumin_US2	0,08352	0,000308642
CREB	0,08722	0,000310559
tf-LF1	0,08855	0,0003125
MRE-BP	0,09175	0,000314465
GATA-1	0,1133	0,000316456
HNF-3	0,1299	0,000318471
MRE_CS6	0,1302	0,000320513
EGR-1	0,1347	0,000322581
retroviral_TATA	0,1368	0,000324675
KROX24	0,1369	0,000326797
Sp1	0,1444	0,000328947
UCE.2	0,1477	0,000331126
Ga	0,1536	0,000333333
NFMHCIIB	0,162	0,00033557
ATF	0,1637	0,000337838
IgNF-A	0,1754	0,000340136
CTF/NF-1	0,1898	0,000342466
AABS_CS2	0,1988	0,000344828
MRE_CS4	0,1997	0,000347222
PuF	0,2188	0,00034965
OTF-2A	0,219	0,000352113
Y_box_(1)	0,2429	0,00035461
HNF1	0,2591	0,000357143
TRF	0,2629	0,000359712
USF-IIEII	0,2859	0,000362319
Ig_cd	0,2867	0,000364964
CP1	0,2892	0,000367647
junB-US2	0,2959	0,00037037
NGFI-C	0,3105	0,000373134
MBF-I	0,3131	0,00037594
GH0	0,314	0,000378788
HC5	0,3269	0,000381679
Lam_B2_US	0,3341	0,000384615
HNF-4	0,3413	0,000387597
TFII-I	0,3419	0,000390625
SDR_RS	0,3465	0,000393701
NFI	0,3473	0,000396825
SIF	0,3541	4,00E-04
beta-pol_RS	0,3793	0,000403226
element_II_rs-4	0,386	0,000406504
c-fos.5	0,3941	0,000409836
ApoE_A	0,3971	0,000413223
ETF	0,3995	0,000416667
LF-A2	0,3997	0,000420168

INF.1	0,4215	0,000423729
UCE.1	0,4219	0,00042735
AP-1	0,4259	0,000431034
MT-I.3	0,4269	0,000434783
HSV-tk-2nd_distal_si	0,4309	0,000438596
PEBP2	0,4375	0,000442478
gERE	0,4529	0,000446429
MT-I.4	0,4529	0,00045045
APRT-CHO_US	0,4594	0,000454545
Oct	0,4646	0,000458716
SRF	0,4824	0,000462963
MEF-I	0,5241	0,00046729
NF-S	0,5359	0,000471698
ZDNA_CS	0,5457	0,00047619
dc_box	0,5505	0,000480769
T-Ag	0,5516	0,000485437
MLTF	0,5533	0,000490196
DBP	0,5663	0,00049505
NRE_Box2_CS	0,5682	5,00E-04
H4TF1	0,5703	0,000505051
HLA_DQ-beta_US	0,586	0,000510204
TCR-beta_decamer	0,59	0,000515464
ISGF1	0,6228	0,000520833
02. Okt	0,6228	0,000526316
SP-C	0,6228	0,000531915
hsp70_US	0,623	0,000537634
TEF1	0,623	0,000543478
CAC-BP	0,6252	0,000549451
b-globin	0,6298	0,000555556
EIIaE-B	0,633	0,000561798
c-Myc	0,6337	0,000568182
EGR-2	0,6383	0,000574713
GArC1	0,6408	0,000581395
GArC3	0,6452	0,000588235
APRT-mouse_US	0,6467	0,000595238
H4TF-2	0,6718	0,00060241
Ig_dc	0,6718	0,000609756
NF-E2	0,6839	0,000617284
NF-I	0,7091	0,000625
PEA1	0,7154	0,000632911
01. Okt	0,7155	0,000641026
AP-4	0,7229	0,000649351
beta-glob_US	0,726	0,000657895
element_II_rs-3	0,7325	0,000666667
IFN-inducible_CS	0,7334	0,000675676

HSV_IE_repeat	0,7397	0,000684932
ENKTF1	0,7464	0,000694444
EIIaE-A	0,776	0,000704225
GCF	0,7847	0,000714286
EARLY-SEQ1	0,7921	0,000724638
HC1	0,7941	0,000735294
Lymphokine_CS	0,8002	0,000746269
OBP	0,819	0,000757576
EIIF	0,8215	0,000769231
NF-D	0,8215	0,00078125
GH-TRE	0,8342	0,000793651
AP-2	0,8352	0,000806452
Oct-factors	0,8536	0,000819672
JCV_repeated_sequenc	0,87	0,000833333
glucagon-G3A	0,8729	0,000847458
NF-kB	0,8831	0,000862069
TFIID	0,9322	0,000877193
alpha-1-AT-X-Hapt	1	0,000892857
alpha-1-AT-X-RBP	1	0,000909091
(AP-1)	1	0,000925926
AP-3	1	0,000943396
ApoCIIIp1	1	0,000961538
ApoE_B2	1	0,000980392
B1	1	0,001
B2	1	0,001020408
CATT-BP	1	0,001041667
CBF	1	0,00106383
CD28RC	1	0,001086957
C/EBP	1	0,001111111
CYP21B_US	1	0,001136364
EF-1A	1	0,001162791
GHF-1	1	0,001190476
GHF-1-GH-site_2	1	0,001219512
GR	1	0,00125
HLAII.Xbox	1	0,001282051
IBP-1	1	0,001315789
ICSbf	1	0,001351351
Ig_core	1	0,001388889
IgHC.2	1	0,001428571
ISRE_CS	1	0,001470588
ITF	1	0,001515152
Ker1	1	0,0015625
MGF	1	0,001612903
MT-I.6	1	0,001666667
Myc-CF1	1	0,001724138

MyoD	1	0,001785714
myosin-specific	1	0,001851852
NF-1	1	0,001923077
NF1	1	0,002
NFe	1	0,002083333
NF-GMa	1	0,002173913
NF-IL2-A	1	0,002272727
NF-IL2-D	1	0,002380952
NF-InsE2	1	0,0025
NF-I/Sp1	1	0,002631579
NF-Y	1	0,002777778
NGFI-B	1	0,002941176
oct-B1A	1	0,003125
PGK-1-NF1-like_US	1	0,003333333
PGK-1-US	1	0,003571429
Pit-1	1	0,003846154
PR	1	0,004166667
PRDI-BF	1	0,004545455
R2	1	0,005
RF-X	1	0,005555556
RFX	1	0,00625
SOD-1-GC-hexamer	1	0,007142857
(Sp1)	1	0,008333333
SP-E	1	0,01
(SRF)	1	0,0125
T3R	1	0,016666667
USF	1	0,025
XY_CS.2	1	0,05

Abbildung 10: Transkriptionsfaktoren, vorhergesagt durch PromotorScan. Errechnung eines P-Values mittels des Exact Fisher Tests, unter Vernachlässigung der Mehrfachvorhersage auf demselben Promotor, zeigt einige Faktoren, die ein positives Signifikanzniveau erreichen (*). Eine solche Signifikanz wäre für einen P-Value < 0,05 erreicht (Signifikanznivaeu 0,05). Nach Berücksichtigung einer False Discovery Rate sind diese Faktoren jedoch nicht mehr signifikant unterschiedlich verteilt.

Der verwendete PromotorScan-Algorithmus erkennt nur bisher molekularbiologisch bereits identifizierte und bekannte Transkriptionsfaktorbindungsstellen. In einer ersten Analyse der bekannten Transkriptionsfaktorbindungsstellen waren keine signifikanten Unterschiede in der Verteilung zwischen der Gruppe der Gene, welche am höchsten in der Leber exprimert sind

und der Gruppe der Gene, die in anderen Geweben am höchsten exprimiert sind, nachweisbar. Im folgenden wurde untersucht, ob möglicherweise bisher nicht identifizierte Bindungsstellen für eine differentielle Genexpression in der Leber verantwortlich sein könnten. Diese Transkriptionsfaktorbindungsstellen sind typischerweise zwischen 5 und 15 bp lang. Um auszuschließen, dass mit dem zuvor verwandten PromotorScan-Algorithmus Transkriptionsfaktorbindungsstellen nicht erfaßt wurden, die bisher nicht bekannt sind, wurden die Häufigkeiten sämtlicher vorkommender 8 bp (maximal 4^8) und 10 bp (maximal 4^{10}) Nukleotidkombinationen in diesen Promotorsequenzen untersucht.

Hierbei ergaben sich in der Auswertung der Verteilung aller Nukleotidkombinationen, also aller potentiell neuen Transkriptionsfaktorbindungsstellen, keine wesentlichen Unterschiede zwischen Promotoren der Gene, die in der Leber am höchsten exprimiert sind und Promotoren der Gene, die in anderen Geweben am höchsten exprimiert sind. Unterschiede wurden lediglich für Faktoren beobachtet, die in sehr geringer Anzahl auf den Promotoren zu finden waren, so dass kleine Unterschiede der absoluten Zahlen des Vorkommens in einem vergleichsweise großen relativen Unterschied resultierten. Aufgrund der geringen absoluten Zahlen dieser Transkriptionsfaktorbindungsstellen kann das Vorkommen dieser Nukleotidkombinationen jedoch als biologisch irrelevant vernachlässigt werden. Für Nukleotidkombinationen, die in wesentlicher Anzahl zu finden waren, fanden sich keine wesentlichen Unterschiede. Für den Fall der Zählung aller indiviuduellen Promotorelemente auf jedem Promotor findet sich kein statistischer Test, der zur Berechnung eines Signifikanzniveaus herangezogen werden könnte. Es fanden sich zwischen den beiden Gruppen keine biologisch relevanten Unterschiede.

Deshalb wurden die Nukleotidkombinationen auch hier im weiteren lediglich im Hinblick auf das Vorkommen, also nicht auf die absolute Zahl der Elemente auf den Promotoren der

entsprechenden Gengruppen untersucht. Bei diesem Vorgehen konnte zur Untersuchung der Signifikanz einer unterschiedlichen Verteilung einer jeweiligen Nukleotidkombination ein exakter Fischer Test angewandt werden. In einer initialen Analyse fanden sich, den Ergebnissen der Promotorelementanalyse entsprechend, mehrere signifikant unterschiedlich verteilte Faktoren. Bei potentiell 4^8 bzw 4^{10} unterschiedlichen Nukleotidkontingenten und entsprechenden einzelnen Testungen muss mit 5%iger Wahrscheinlichkeit davon ausgegangen werden, dass die Vorhersage falsch positiv ist. Daher muss eine False Discovery Rate berücksichtigt werden (Abb. 11).

A				B			
contig	**liver**	**non liver**	**ratio**	**contig**	**liver**	**non liver**	**ratio**
ggtttcatac	4	1	74.8371467025572	aaaaaaaaaa	655	13482	0.908958817129784
ttatcggttt	3	1	56.1278600269179	tttttttttt	502	11676	0.804390365807719
tgctcctata	3	1	56.1278600269179	tgtgtgtgtg	162	2881	1.05203208658576
tgcggtgtag	3	1	56.1278600269179	gtgtgtgtgt	160	2920	1.02516639318572
tgattatagc	3	1	56.1278600269179	tgggattaca	111	2192	0.947413695709837
tcgtcatgga	3	1	56.1278600269179	gctgggatta	109	2056	0.991883388932239
tccttactac	3	1	56.1278600269179	ggattacagg	105	2200	0.892943227700967
gtatcagtac	3	1	56.1278600269179	ctgggattac	103	2009	0.95921181064751
ggtcaattga	3	1	56.1278600269179	gggattacag	102	2146	0.889257800985652
ggcatatccc	3	1	56.1278600269179	gattacaggc	98	1874	0.978393860305577
gctatatcat	3	1	56.1278600269179	ctcagcctcc	94	1570	1.12017385191726
gcgaatagct	3	1	56.1278600269179	cctcagcctc	90	1556	1.08215668432361
gccgcaaact	3	1	56.1278600269179	gccccgcccc	83	1961	0.791876998509976
gcaaacgatt	3	1	56.1278600269179	ggggcggggc	81	1760	0.861052398140218
cggagaattg	3	1	56.1278600269179	ccccgccccc	80	1398	1.07063156942142
ccttccgaac	3	1	56.1278600269179	tgctgggatt	77	1390	1.03641372231959
cccctacatg	3	1	56.1278600269179	gcctcccaaa	77	1532	0.940349265028868
ccaatcgtat	3	1	56.1278600269179	taatcccagc	74	1456	0.950884075547602
cacggcatcg	3	1	56.1278600269179	atatatatat	73	1104	1.23711768779137
atagcccacg	3	1	56.1278600269179	tcccaaagtg	72	1488	0.905288064950289
atacctctcg	3	1	56.1278600269179	ccagcctggg	72	1570	0.858005503596197
agtggtatgc	3	1	56.1278600269179	ccaaagtgct	72	1365	0.986863473000754
acttcggtaa	3	1	56.1278600269179	agtgctggga	72	1317	1.02283116222174
acgcaaacag	3	1	56.1278600269179	tatatatata	71	1070	1.24145734015924
aagctatgac	3	1	56.1278600269179	aaaaaaaaaa	655	13482	0.908958817129784
aacaatcgtt	3	1	56.1278600269179	tttttttttt	502	11676	0.804390365807719
ttttggtgcg	2	1	37.4185733512786	tgtgtgtgtg	162	2881	1.05203208658576

Abbildung 11: Häufigkeiten individueller Nukleotidkontigente. (A) Kontingente mit deutlichem relativen Unterschied weisen eine geringe absolute Anzahl auf. (B) Kontingente mit hoher, biologisch relevanter Anzahl kommen in beiden Gruppen mit gleicher relativer Häufigkeit vor.

Bei Berücksichtigung der False Discovery Rate, kann keine der zuvor als vermeintlich unterschiedlich verteilten Nukleotidkombinationen noch als signifikant angesehen werden. Insgesamt konnte somit auch bei der erweiterten Suche nach potentiell neuen, bisher nicht identifizierten Transkriptionsfaktorbindungsstellen kein signifikanter Unterschied in der Verteilung solcher potentieller Transkriptionsfaktorbindungsstellen nachgewiesen werden.

3.5 Analyse von Promotorsequenzen hinsichtlich des Vorkommens von TATA-Boxen

Nachdem sowohl für bekannte als auch für potentiell neue Transkriptionsfaktorbindungstellen kein wesentlicher Unterschied in der Verteilung zwischen Genen, die in der Leber am höchsten exprimiert sind und solchen, die in anderen Geweben am höchsten exprimiert sind, gefunden wurde, galt das Augenmerk der Analyse möglicherweise die Transkription beeinflussender Faktoren im weiteren den TATA-Boxen. Der TATA-Box kommt bei der Transkriptionsinitiierung eine wichtige Rolle zu, indem über sie die Bindung des initialen Transkriptionskomplexes vermittelt wird. TATA-Boxen wurden ebenfalls mittels PromotorScan [38] vorausgesagt und jede einzelne Promotorsequenz automatisiert durch die PromotorScan Website geschleust. Im Falle der Identifizierung einer TATA-Box wurde diese aus der entsprechenden Website extrahiert und in der postgreSQL Datenbank gespeichert.

	TATA-Box vorhanden	**TATA-Box nicht vorhanden**
Leber	57	721
nicht Leber	976	13.718

Abbildung 12: Verteilung von TATA-Boxen auf Promotoren von Genen, die in der Leber oder in anderen Geweben am höchsten exprimiert sind.

Durch den stringenten Algorithmus gelang die Identifizierung einer Gesamtzahl von 1033 TATA-Boxen; 57 dieser TATA-Boxen fanden sich auf Promotoren von Genen, die in der Leber am höchsten exprimiert waren, 976 auf Promotoren von Genen, die in anderen Geweben am höchsten exprimiert waren (Abb. 12). Unter Berücksichtigung der unterschiedlichen Größe dieser beiden Gruppen fanden sich im Vergleich keine signifikanten Unterschiede in der Verteilung von TATA-Boxen zwischen diesen beiden Gruppen. Für die Berechnung einer möglichen statistischen Signifikanz wurde erneut ein exakter Fisher Test verwandt, der negativ ausfiel.

3.6 Analyse von Promotorsequenzen hinsichtlich des Vorkommens von CpG-Islands

Schließlich wurde die Bedeutung von CpG-Islands für eine potentiell differentielle Regulation untersucht. CpG-Islands sind genomische Regionen mit statistisch erhöhter CpG-Dinukleotid-Dichte. CpG-Islands spielen eine Rolle in der differentiellen Genexpression, indem durch eine Methylierung der CpG-Islands in der Promotorregion die entsprechenden Gene oft vermindert exprimiert werden [39] [19].

CpG-Islands lassen sich mittels eines statistischen Modells zur Analyse der Nukleotidverteilung, einem Hidden Markov Modell, bioinformatisch gut analysieren. Nach lokaler Implementierung eines entsprechend trainierten Algorithmus [21] wurden insgesamt 8742 CpG-Islands in einem Bereich von bis zu 5 kb upstream des Transkriptionsstarts der jeweiligen Gene identifiziert. Davon fanden sich 364 CpG-Islands auf Promotoren von Genen, die am höchsten in der Leber exprimiert und 8378 auf Promotoren von Genen, die in anderen Geweben am höchsten exprimiert waren (Abb. 13). Die Analyse der CpG-Islands bezüglich potentieller Unterschiede zwischen den Genen, die am höchsten in der Leber bzw. in anderen Geweben exprimiert werden, wurde analog zur Analyse der Transkriptionsfaktorbindungsstellen durchgeführt. Zum einen wurde die absolute Zahl der

vorhergesagten CpG-Islands berücksichtigt, zum anderen im Gegensatz dazu lediglich das Vorhandensein von CpG-Islands auf einem Promotor. Es ergaben sich keine signifikanten Unterschiede in der Verteilung von CpG-Islands auf Promotoren dieser beiden Gengruppen.

	CpG Island-vorhanden	CpG-Island nicht vorhanden
Leber	364	414
icht Leber	8378	6316

Abbildung 13: Verteilung von CpG-Islands auf Promotoren von Genen, die in der Leber oder in anderen Geweben am höchsten exprimiert sind.

3.7 Analyse der Nukleotidzusammensetzung aller Transkripte

Die Nukleotidverteilung bzw. die Struktur der Promotoren weisen in keiner Hinsicht Unterschiede auf, die eine differentielle Expression von Genen in der Leber begünstigen könnten. Deshalb wurden im weiteren die Sequenzen und die Struktur der Transkriptions- und Translationsprodukte der jeweiligen Gene untersucht.

Um die Nukleotidverteilung in den RNA Sequenzen untersuchen zu können, wurden zunächst sämtliche RNA Sequenzen der individuellen Gene vom EBI/Ensembl Genom Server [30] heruntergeladen und in der postgreSQL Datenbank gespeichert. Da sich für viele Gene in der EBI Datenbank multiple Transkripte finden ließen, wurde die Länge der RNA Sequenzen verglichen und jeweils nur die längste Sequenz übernommen. Dies geschah unter der Vorstellung, dass diese Sequenz am vollständigsten in der Datenbank verfügbar war. Anschließend wurde das Vorkommen der einzelnen Nukleotide, Adenin (A), Cytosin (C), Guanin (G) und Thymin (T), ausgezählt. Hierbei fand sich eine hohe Übereinstimmung der Verteilung aller vier Nukleotide zwischen der Gruppe von Genen, die in der Leber am höchsten exprimiert werden, und den Genen, die in anderen Geweben am höchsten exprimiert

werden. Alle vier Nukleotide fanden sich jeweils zu ca. 25% anteilig an der Gesamtmenge der Nukleotide (Abb.14).

	A	C	G	T
Leber	25,27%	25,01%	25,28%	24,44%
andere Gewebe	25,84%	24,88%	25,28%	24,00%

Abbildung 14: Relative Häufigkeit der einzelnen Nukleotide in sämtlichen RNAs der analysierten Gene aufgeteilt nach Genen, die in der Leber (Leber) oder in anderen Geweben (nicht Leber) am höchsten exprimiert sind.

3.8 Analyse der Aminosäurezusammensetzung aller Translationsprodukte

Schließlich wurde die Aminosäurenverteilung in den Proteinsequenzen untersucht. Sämtliche translatierte Proteinsequenzen der individuellen Gene wurde vom EBI/Ensembl Genom Server heruntergeladen und in die postgreSQL Datenbank geschrieben. Da sich für viele Proteine multiple Translationsprodukte in der Datenbank fanden, wurde die Länge der Proteinsequenzen verglichen und jeweils nur die längste Sequenz übernommen. Dies geschah unter der Vorstellung, dass diese Sequenz am vollständigsten in der Datenbank verfügbar war. Anschließend wurde das Vorkommen der einzelnen Aminosäuren im Rahmen ihrer Zughörigkeit zu biochemisch unterschiedlichen Aminosäuregruppen ausgezählt.

Aminosäuregruppe	Höchste Expression in Leber	Höchste Expression in anderen Geweben
Aliphatische Aminosäuren	41%	41%
Aromatische Aminosäuren	9%	7%
Saure Aminosäuren	11%	12%
Basische Aminosäuren	13%	14%
Hydroxylische Aminosäuren	13%	14%
Schwefelhaltige Aminosäuren	5%	4%
Amidhaltige Aminosäuren	8%	8%

Abbildung 15: Relative Häufigkeit der einzelnen Aminosäuregruppen in sämtlichen translatierten Proteinsequenzen der analysierten Gene, aufgeteilt nach der höchsten Expression der korrespondierenden Gene in der Leber (Ja) oder in anderen Geweben (Nein).

Es fand sich eine hohe Übereinstimmung der Verteilung aller Aminosäuren und Aminosäuregruppen zwischen der Gruppe von Genen, die in der Leber am höchsten exprimiert, und den Genen, die in anderen Geweben am höchsten exprimiert werden (Abb.15).

4 Diskussion

Eine regelrechte physiologische Leberfunktion ist lebensnotwendig. Um diese zu ermöglichen, ist ein komplexes Zusammenspiel eines regelrechten Netzwerks von Genen notwendig [1]. Zentrale Voraussetzung für eine biologische Wirkungsentfaltung eines Gens ist seine Expression. Da die genomische Information in allen Geweben und Zellen in gleicher Form vorliegt, kommt der gewebsspezifischen Regulation der Genexpression eine essentielle Bedeutung zu.

In diesem Zusammenhang wurde in den zurückliegenden Jahren wiederholt dokumentiert, dass einzelnen Genen eine differentielle Regulation zukommt bei der Ausbildung von Erkrankungen oder dem Ansprechen auf Medikamente [40] [41] [42] [43] [44] [45] [46]. Diese Untersuchungen belegen die Bedeutung einer differentiellen Genregulation in der Leber. Aufgrund der Zeit- und Kostenintensität dieser molekularbiologischen Untersuchungen, blieben die Analysen jedoch stets auf einzelne Faktoren und Gene beschränkt. Hinsichtlich der Gesamtheit des Transkriptoms ist nach wie vor weitgehend unklar, ob die Gruppe von Genen, die in der Leber exprimiert wird, um dort eine regelrechte Funktion aufrecht zu erhalten, einer zentral gesteuerten Grundexpression in der Leber unterliegt, oder ob diese Gruppe von Genen zufällig, beeinflußt von individuellen, die Genexpression in der Leber regulierenden Faktoren, exprimiert wird. Ließen sich übergeordnete, zentrale Faktoren nachweisen, die eine Genexpression in der Leber übergeordnet regulieren, so wäre dies von erheblicher medizinischer und biologischer Relevanz. Diesen Faktoren könnte möglicherweise eine zentrale Funktion bei der Differenzierung von hepatologischen Erkrankungen zukommen. Sie könnten potentielle, neue therapeutische Targets darstellen.

4.1 Genexpressionsdaten

Zur Untersuchung einer möglichen Beteiligung von Transkriptionsfaktorbindungsstellen, TATA-Boxen, RNA- und Proteinsequenzen an einer übergeordneten Regulation der Genexpression im Lebergewebe wurden aus einem Affymetrix Microarray Datenset die Expressiondaten für 12 Gewebe (Knochenmark, Gehirn, Herz, Niere, Leber, Lunge, Bauchspeicheldrüse, Prostata, Skelettmuskel, Rückenmark, Milz und Thymus) von jeweils insgesamt 62.839 Datensets extrahiert. Mittels Microarray generierte Expressionsdaten weisen generell eine geringe Fehlerrate auf, so dass im allgemeinen eine Validierung relevanter differentiell exprimierter Gene mittels RT-PCR oder Northern-Blot anzustreben ist. Bei 62.839 individuellen Spots ist eine solche Validierung mit molekularbiologischen Methoden nicht möglich. Es gibt jedoch zum gegenwärtigen Zeitpunkt keine andere Methode zur Generierung genomweiter Expressionsdaten (abgesehen von SAGE, was aber erheblich teurer und mit noch größerer Variabilität verbunden ist). Methodisch muss somit also eine geringe Variabilität in Kauf genommen werden. Hinsichtlich einer solchen möglichen Variabilität ist jedoch von einer unselektionierten, gleichmäßigen Verteilung auf die beiden zu vergleichenden Gruppen von Genen auszugehen. Somit kann angenommen werden, dass diese Unschärfe die endgültigen Ergebnisse nicht wesentlich beeinflußt. Zur Minimierung der Ungenauigkeit der Expressionsdaten waren die Ergebnisse für jedes Gewebe zumindest doppelt verfügbar. Eine Vielzahl von Genen war mehrfach auf den betreffenden Arrays gespottet, was wiederum insgesamt eine höhere Validität der Expressionsdaten für jedes Gewebe bedeutet. Die 62.839 Spots des Arrays wurden insgesamt 15.472 individuellen Genen zugeordnet.

4.2 Verwendete Software

Die im Rahmen dieses Projekts notwendigen Programme und Software-Module wurden in der Programmiersprache Perl implementiert. Perl ist in idealer Weise auf die Notwendigkeiten

bioinformatischer Fragestellungen ausgelegt. Nicht zuletzt kommt dies in einer mittlerweile eigens verfügbaren Suite „BioPerl“ zum Ausdruck. Die frei verfügbare CPAN Bibliothek bietet eine Vielzahl zur unmittelbaren Verwendung vorgefertigter Programmbausteine an, die die Ansteuerung bioinformatischer Websites und die Extraktion von Ergebnissen in automatisiert verwertbarer Form erleichtern. Grundsätzlich wäre eine Programmierung der notwendigen Software-Module auch in anderen Programmiersprachen denkbar. Die Wahl der Programmiersprache hat aber insgesamt keinerlei Einfluß auf die erzielten Ergebnisse, da sie in keiner Hinsicht limitierend bezüglich der zu beantwortenden Fragestellungen ist.

Zum Aufbau der Datenbank wurde postgreSQL, eine kostenlose, open source Software verwendet. Die entsprechende Datenbank hätte in verschiedenen Programmiersprachen implementiert werden können, ohne dass hierbei ein Einfluß auf die erzielten Ergebnisse entstanden wäre. PostgreSQL wurde aufgrund seiner Benutzerfreundlichkeit gewählt, die insbesondere bei den in dieser Untersuchung durchzuführenden umfassenden Abfragen zum Tragen kommt.

4.3 Selektion von Genen mit höchster Expression in der Leber

Die Einteilung der Gene in zu vergleichende Gruppen in Abhängigkeit von ihrer Expression in Lebergewebe erfolgte unter vergleichsweise stringenten Kriterien. Als Gene mit höchster Expression in der Leber wurden in der vorliegenden Arbeit solche Gene gewertet, deren mittlere Expression in der Leber nach Addition des durchschnittlichen Expressionswerts in allen anderen Geweben und dividiert durch die Standardabweichung aller Proben größer oder gleich 1,96 war. Fand sich die Expression des Gens gleich -1.96 oder kleiner, so galten diese Gene als im Lebergewebe am niedrigsten exprimiert. Eine Vielzahl von Selektionskriterien wäre denkbar gewesen. Man hätte etwa einfach die absoluten Expressionwerte heranziehen und den höchsten Expressionswert in einem Gewebe zur Zuordnung verwenden können.

$$\frac{\text{Leber-Mittelwert} + \text{Mittelwert aller Gewebe}}{\text{Standardabweichung}}$$

Abbildung 16: Berechnung des Werts zur Einschätzung einer höchsten Expression in der Leber.

Die Aufteilung der im Array gespotteten Gene ist hinsichtlich einer Identifizierung einer umschriebenen Gruppe von genetischen Faktoren, die die Genexpression in der Leber regulieren jedoch essentiell. Vor dem Hintergrund der etwas geringer einzustufenden Validität mittels Microarray generierter Expressionsdaten, erscheint es sinnvoll, eine kleinere, dafür aber validere und besser definierte Gruppe auf gemeinsame, gehäuft auftretende Merkmale zu untersuchen. Daher wurde für die Auswahl der in der Leber am höchsten exprimierten Gene ein stringentes Kriterium angelegt. Dadurch wurden insbesondere die Gene nicht für die „Leber"-Gruppe berücksichtigt, die ein sehr gleichmäßiges Expressionsniveau in allen Geweben haben und nur eine unbedeutend höhere Expression in der Leber. Formal weisen diese Gene zwar eine höchste Expression in der Leber auf, im Vergleich mit anderen Geweben ist diese nur gering erhöhte Expression in der Leber jedoch biologisch meist nicht relevant im Sinne einer Charakterisierung des Lebergewebes oder ist für dessen Funktion nicht spezifisch.

Anhand der in dieser Arbeit gewählten Vorgehensweise wurden insgesamt 778 Gene als am höchsten in Lebergewebe exprimiert eingestuft. Das entspricht 5% der gesamten, auf dem Array verfügbaren Gene entspricht und stellt damit eine ausreichende Anzahl für eine vergleichende Auswertung dar. Hingegen fanden sich nur 14 Gene als am niedrigsten in der Leber exprimiert. Angesichts einer Gesamtzahl von 15.472 Genen auf dem Array ist die Gruppe der am niedrigsten in der Leber exprimierten Gene nicht den anderen Genen in vergleichenden Untersuchungen sinnvollerweise gegenüberzustellen. Eine negative

Regulation von Genen in der Leber durch übergeordnete Faktoren wurde deshalb in der vorliegenden Arbeit nicht untersucht.

4.4 Extraktion von Promotorsequenzen

Eine Regulation der Transkription kann an verschiedensten Stellen stattfinden. Wiederholt wurde gezeigt, dass die Mehrzahl der regulatorischen Eingriffe im Rahmen der Transkriptionsinitiierung geschehen, also auf der Ebene der Promotorregulation. In der vorliegenden Arbeit wurden deshalb Promotorstruktur und mögliche Bindungsstellen regulatorischer Elemente untersucht. Promotoren oder das Vorkommen regulatorischer Elemente in einer upstream des Transkriptionsstarts liegenden Promotorregion sind in ihrer Ausdehnung nicht definiert. Promotorelemente können sowohl mehrere 10 kb upstream des Transkriptionsstarts ihre Wirkung ausüben, als auch downstream des Transkriptionsstarts lokalisiert sein, z.B. im ersten Intron [47]. Zur automatisierten Auswertung, welche aufgrund des genomweiten Ansatzes unverzichtbar war, ist eine Limitierung notwendig. Solche Limitierungen gibt es auch bei molekularbiologischen Versuchen, da auch hier eine vollständige Klonierung eines Promotors aus den gleichen Gründen nicht möglich ist. Im Rahmen molekularbiologischer Promotorstudien war vielfach eine Region von 1000 bp als wesentlich für eine suffiziente Promotorfunktion analysiert worden. In der vorliegenden Arbeit wurde die Analyse auf eine „core promotor“ Region von 1000 bp upstream des Transkriptionsstarts limitiert. Ausnahme davon erfuhr lediglich die Analyse von CpG-Islands, da für diese häufig eine Lokalisation von mehreren Kilobasen upstream des Transkriptionsstarts gezeigt worden war.

Im Rahmen eines automatisierten Downloads der Promotorsequenzen vom EBI/Ensembl Server konnten in dieser Untersuchung für 14.644 Gene Sequenzen upstream des Transkriptionsstarts identifiziert werden. Bei den verbleibenden 828 Genen waren bisher

unvollständige Charakterisierungen der jeweiligen Gene und somit fehlende oder unvollständige Verfügbarkeit des Transkriptionsstarts und des ersten Exons als Ausgangspunkt zur Bestimmung der Lokalisation der zugehörigen Promotorregion für die fehlende Verfügbarkeit der Promotorsequenz verantwortlich.

4.5 Analyse der Promotorsequenzen

Die Promotorsequenzen wurden auf das Vorkommen potentieller Trankriptionsfaktorbindungsstellen analysiert. Zur Analyse von Promotorelementen und Transkriptionsfaktorbindungstellen stehen verschiedene Algorithmen zur Verfügung. Insgesamt ist die bioinformatische Technologiebezüglich der Vorhersage und Identifizierung von Promotorelementen noch als limitiert anzusehen, so dass hier ein Kompromiss zwischen zu erzielender Sensitivität des Algorithmus und gleichzeitiger Spezifität und Validität der vorhergesagten Elemente eingegangen werden muss.

In der vorliegenden Arbeit wurde der stringente Algorithmus (PromoterScan) mit hoher Spezifität gewählt, um im weiteren eine gut definierte, wenn auch dadurch etwas kleinere Gruppe von Genen zu erhalten und diese im Vergleich mit der restlichen Gruppe der Gene auf ein unterschiedliches Vorkommen von regulatorischen Elementen untersuchen zu können.

Der PromotorScan Algorithmus analysiert Promotorsequenzen unter Verwendung von Positional Weight Matrizen. Auf diese Weise zeigten sich auf 7.042 Promotoren Transkriptionsfaktorbindungsstellen. Dies entspricht einer Quote von 48%.

Der PromotorScan Algorithmus wurde im Vergleich verschiedener Vorhersage-Algorithmen an gut definierten Kontroll-Promotorsequenzen als der Algorithmus mit der höchsten Spezifität identifiziert. Dies geschah jedoch auf Kosten der Sensitivität, weshalb insgesamt

der Verbleib einer zu analysierenden Gruppe von 7.042 Promotoren hingenommen wurde. Die Auswahl dieser Promtoren erfolgte allein durch die Vorhersage von Elementen des Algorithmus. Deshalb kann davon ausgegangen werden, dass die ausgewerteten 48% der Promotoren repräsentativ für die Gesamtgruppe der initial identifizierten Promotoren ist.

PromotorScan wurde auf der Grundlage von Positional Weight Matrizen entwickelt. Bei der Vorhersage von möglichen Transkriptionsfaktorbindungsstellen erfolgt daher eine graduelle Abstufung, wobei es zu einer mehrfachen Vorhersage derselben Bindungsstelle kommen kann, verschoben um lediglich einige wenige Basen. Um eine auf dieser mehrfachen Vorhersage beruhende Falschzählung der Promotorelemente auszuschließen, wurden alle Vorhersagen desselben Elements innerhalb eines Abschnitts von 8 bp up- oder downstream ignoriert. Mit diesem Ansatz fand sich eine Gesamtzahl von 241.984 Transkriptionsfaktorbindungsstellen. Unter der Vorstellung, dass übergeordnete Faktoren, die eine Expression in der Leber stimulieren, in der Gruppe der Gene, die in der Leber am höchsten exprimiert sind, verhältnismäßig wesentlich häufiger zu finden seien, wurde das Vorkommen der vorhergesagten Transkriptionsfaktorbindungsstellen zwischen der Gruppe der Gene, die in der Leber am höchsten exprimiert waren und den restlichen Genen untersucht. Dabei zeigten sich jedoch keine wesentlichen Unterschiede.

Der PromotorScan-Algorithmus erkennt nur bisher molekularbiologisch bereits identifizierte und bekannte Transkriptionsfaktorbindungsstellen. Deshalb wurde im folgenden untersucht, ob möglicherweise bisher nicht identifizierte Bindungsstellen für eine differentielle Genexpression in der Leber verantwortlich sein könnten. Die Transkriptionsfaktorbindungsstellen sind typischerweise zwischen 5 und 15 bp lang. Um auszuschließen, dass mit dem zuvor verwandten PromotorScan-Algorithmus Transkriptionsfaktorbindungsstellen nicht erfaßt wurden, die bisher nicht bekannt sind,

wurden die Häufigkeiten sämtlicher vorkommender 8 bp (maximal 4^8) und 10 bp (maximal 4^{10}) Nukleotidkombinationen in diesen Promotorsequenzen untersucht.

Auch bei Berücksichtigung aller potentiellen Transkriptionsfaktorbindungsstellen ergaben sich keine wesentlichen Unterschiede zwischen den Genen, welche am höchsten in der Leber exprimert sind und den Genen, die in anderen Geweben am höchsten exprimiert sind. Unterschiede wurden lediglich für Nukleotidkombinationen bzw. potentielle Transkriptionsfaktorbindungsstellen beobachtet, die in sehr geringer Anzahl auf den Promotoren zu finden waren, so dass zahlenmäßig kleine Unterschiede, z.B. 3 zu 1, in einem vergleichsweise großen relativen Unterschied resultierten. Aufgrund der geringen absoluten Menge dieser Transkriptionsfaktorbindungsstellen können diese als biologisch irrelevant vernachlässigt werden. Für Transkriptionsfaktorbindungstellen, die in größerer Anzahl zu finden waren, fanden sich keine wesentlichen Unterschiede. Die praktisch gänzliche Übereinstimmung des relativen Vorkommens der jeweiligen Nukleotidkontingente spricht darüber hinaus für die sehr gute Datenqualität. Für die Zählung aller individuellen Promotorelemente auf jedem Promotor findet sich kein statistischer Test, der zur Berechnung eines Signifikanzniveaus herangezogen werden könnte.

4.6 Analyse von TATA-Boxen

Cis-regulatorische Elemente der Promotoren zeigten keine relevanten Unterschiede hinsichtlich des Vorkommens bzw. der Verteilung zwischen Promotoren von Genen, die in der Leber am höchsten exprimiert werden und Genen, die in anderen Geweben exprimiert werden. Als weiteres wesentliches Element der Analyse eukaryoter Promotoren wurde die Verteilung von TATA-Boxen zwischen diesen beiden Gruppen von Genen untersucht.

Die TATA-Box ist ein Sequenzabschnitt, der in vielen Promotoren eukaryotischer Gene zu finden ist, mit der Konsensus Sequenz TATAAA. Diese Sequenz befindet sich meist 25 bis 30 bp upstream der Transkriptionsstartstelle. An diese Sequenz bindet das sogenannte TATA-Box Binde-Protein (TBP) [14]. Die Bindung des TBP an die TATA-Box spielt eine Rolle bei der Ausbildung des Transkriptionsinitiationskomplexes. Es handelt sich um einen Komplex von Transkriptionsfaktoren und RNA-Polymerase II, der ca. 50 verschiedene Proteine beinhaltet. Diese schließen den Transkriptionsfaktor IID (TFIID) ein. Er ist ein Komplex von TATA-Box Binde-Protein (TBP), welches die TATA-Box erkennt, 14 weiteren Faktoren, die TBP binden nicht aber DNA und Transkriptionsfaktor IIB (TFIIB), der sowohl DNA als auch die RNA Polymerase II bindet [15]. TATA-Boxen werden trotz ihrer wichtigen Funktion nur in maximal 70% der eukaryoten Promotoren nachgewiesen. Die Vorhersage und Identifikation von TATA-Boxen mit bioinformatischen Methoden gestaltet sich nach wie vor schwierig [38]. Dies erklärt die vergleichsweise geringe Zahl der in den beiden verglichenen Gruppen identifizierten TATA-Boxen. Unter Berücksichtigung der unterschiedlichen Größen der zu vergleichenden Gengruppen fanden sich keine Unterschiede hinsichtlich der Verteilung bzw. Existenz von TATA-Boxen.

4.7 Analyse der CpG-Islands

Eine weitere Möglichkeit, mit der Gene reguliert werden, ist die DNA Methylierung. Nachdem bei Promotorelementen, TATA-Boxen und cis-acting Elementen in der vorliegenden Arbeit keine wesentlichen Unterschiede in der Verteilung zwischen „Leber-“ und „Nicht-Leber-Genen“ nachweisbar waren, ist somit eine Bedeutung als übergeordnete Faktoren der Regulation nicht anzunehmen. Deshalb wurde eine unterschiedliche Methylierung der DNA in Promotoren untersucht. Dabei hängt das Enzym DNA-Methyltransferase eine Methylgruppe an das Cytosin. Die beiden Cytosine in einem CpG-Dinukleotid auf dem sense und antisense Strang sind im menschlichen Genom meist

methyliert. Methylierte Cytosine sind chemisch labil und können sich in der Zelle zu Uracil umwandeln. Ein so entstandenes Uracil paart sich während der DNA-Replikation mit Adenin, und bei der nächsten DNA-Replikation mit Thymin. Nach der bekannten Basenpaarungsregel erhält man daher Punktmutationen. Dieser Mechanismus führt dazu, dass die Frequenz an CpG-Dinukleotiden im Genom nur ca. ein Fünftel der zu erwartenden Frequenz ist. CpG-Islands sind Regionen, die eine erhöhte Frequenz von CpG-Dinukleotiden aufweisen. Sie liegen oft in Promotorbereichen, können aber auch auf das erste Exon übergreifen oder in gänzlich anderen Abschnitten des humanen Genoms lokalisiert sein. Die CpG dieser Regionen sind meist nicht methyliert und entgehen daher dem Mutationsdruck. Liegt ein CpG-Island im Promotorbereich eines Gens, wird dieses Gen nur exprimiert, wenn keine Methylierung vorliegt [20]. Silencing durch CpG-Islands wird bei der Zellteilung vererbt und ist im wesentlichen irreversibel. Insgesamt besitzen ca. 60-80% aller menschlichen Gene CpG-Inseln in ihren Promotorbereichen [18].

CpG-Islands lassen sich mittels eines statistischen Modells zur Analyse der Nukleotidverteilung, einem sog. Hidden Markov Modell, bioinformatisch sehr gut analysieren [21]. Auf diese Weise wurden in der vorliegenden Untersuchung insgesamt 8742 CpG-Islands in einem Bereich von bis zu 5 kb upstream des Transkriptionsstarts identifiziert. Davon fanden sich 364 CpG-Islands auf Promotoren von Genen, die am höchsten in der Leber exprimiert waren, 8378 auf Promotoren von Genen, die in anderen Geweben am höchsten exprimiert waren. Bei der Analyse der CpG-Islands hinsichtlich potentieller Unterschiede zwischen diesen Gruppen wurde zum einen die absolute Zahl der vorhergesagten CpG-Islands berücksichtigt, zum anderen lediglich das Vorkommen von CpG-Islands auf einem Promotor. Dabei ergaben sich keine signifikanten Unterschiede in der Verteilung von CpG-Islands auf Promotoren der beiden Gengruppen.

4.8 Analyse von RNA- und Proteinsequenzen

Hinsichtlich der herkömmlichen Promotorelemente wurden keine wesentliche Unterschiede in Bezug auf die Verteilung zwischen Genen, die in der Leber am höchsten exprimiert sind, und solchen, die in anderen Geweben am höchsten exprimiert sind, gefunden. Ein Hinweis auf die Existenz übergeordneter regulatorischer Promotorelemente ließ sich somit nicht nachweisen. Im nachfolgenden wurden deshalb die Sequenzen des Transkriptoms und Proteoms auf Hinweise zur Vorhersage einer hohen Expression durch Bestimmung ihrer Nukleotid- bzw. Aminosäurenzusammensetzung untersucht.

Angestoßen wurden diese Untersuchungen durch die Möglichkeit, die subzelluläre Lokalisation eines Proteins allein aufgrund seiner Aminosäurenzusammensetzung vorherzusagen [48] [49]. Die dazu erforderlichen bioinformatischen Algorithmen sind mittlerweile sehr gut etabliert und ermöglichen eine Vorhersagegenauigkeit von bis zu 80% [48]. Sie können sich daher sehr gut mit strukturellen Verfahren messen. Bei diesen wird die Lokalisation anhand von strukturellen Merkmalen, wie zum Beispiel eines Nuklear Localization Signals (NLS), vorhergesagt. Somit stellt sich die Frage, ob auch andere biologisch relevante Merkmale, wie zum Beispiel das Expressionsverhalten, allein an der Sequenz abzulesen sind.

Die Analyse von RNA- und Proteinsequenzen im Hinblick auf Unterschiede in der Verteilung von Nukleotiden oder Aminosäuren erbrachte in der vorliegenden Arbeit keine wesentlichen Unterschiede. Alle vier Nukleotide waren zwischen der Gruppe von Genen, die in der Leber am höchsten exprimiert werden, und den Genen, die in anderen Geweben am höchsten exprimiert werden, gleich verteilt und machten jeweils ca. 25% an der Gesamtmenge der Nukleotide aus. Ebenso fand sich eine hohe Übereinstimmung bei der Verteilung der Aminosäuren und Aminosäurengruppen zwischen der Gruppe von Genen, die in der Leber am

höchsten exprimiert werden, und den Genen, die in anderen Geweben am höchsten exprimiert werden. Somit lässt auch die Analyse der RNA- und Proteinsequenzen zur Vorhersage des Expressionsverhalten von Genen in der Leber keine Rückschlüsse auf dieses Verhalten zu.

4.9 Schlussfolgerung

Die Zusammenschau der Ergebnisse zeigt, dass die Regulation der einzelnen Gene im Lebergewebe im wesentlichen individuell erfolgt. Im Rahmen der bioinformatischen Analysen fanden sich keine übergeordneten genetischen „Leberfaktoren“, die vor allem eine Expression von Genen in der Leber stimulieren. Eine durchaus vorstellbare, mögliche Basisstimulation zur Expression in der Leber scheint somit nicht vorhanden zu sein. Potentielle neue, auf eine Regulation der Genexpression in der Leber zielende therapeutische Ansätze müssen sich somit auch weiterhin auf die Beeinflussung individueller Gene konzentrieren.

5 Zusammenfassung

Die Expression eines Gens in einem Gewebe ist zentrale Voraussetzung für eine konsekutive biologische Wirkungsentfaltung. Für eine Reihe einzelner genetischer Faktoren und Promotorelemente wurde in der Vergangenheit eine Regulation der Genexpression in der Leber (und auch in anderen Geweben) gezeigt. Aufgrund der Zeit- und Kostenintensität dieser molekularbiologischen Untersuchungen blieben die Analysen jedoch stets auf einzelne Faktoren und Gene beschränkt.

Mit der Verfügbarkeit des gesamten humanen Genoms sowie dessen Expressionsdaten in großen Microarray- und SAGE-Datenbanken bietet sich die Möglichkeit, solche Regulationsmechanismen in großem, genomweitem Maßstab zu untersuchen. Dabei geht diese Arbeit der Frage nach, ob es übergeordnete, eine Expression speziell in der Leber fördernde oder hemmende Faktoren gibt (Abb. 1) oder ob jedes Gen von einer unabhängigen Kombination von Faktoren reguliert wird, in dessen Summe die Expression des individuellen Gens in der Leber am stärksten ist. Sollten sich übergeordnete, eine Expression in der Leber stimulierende Faktoren finden, wären diese interessant für die Entwicklung neuer Behandlungskonzepte bei Lebererkrankungen.

Zur Untersuchung dieser Fragestellung wurden zunächst aus einem Affymetrix Microarray Datenset für 12 Gewebe die Expressiondaten von insgesamt jeweils 15.472 Genen extrahiert. Dies geschah mittels eines eigens geschriebenen Programms, implementiert in der Programmiersprache Perl. Die Daten wurden in einer Datenbank gespeichert. In einem zweiten Schritt wurden zusätzlich die Promotorsequenzen der einzelnen zugehörigen Gene, definiert als eine 1000 bp Region upstream des Transkriptionsstarts, in dieselbe Datenbank abgelegt. Die Promotorsequenzen wurden mittels eines eigens geschriebenen Programms über den PromotorScan-Algorithmus analysiert. Es handelt sich um einen Algorithmus, der

eukaryote Promotoren voraussagt und auf potentielle Trankriptionsfaktorbindungsstellen und TATA-Boxen untersucht. Auf diese Weise wurden Transkriptionsfaktorbindungsstellen auf 7042 der Promotoren identifiziert. Es fand sich eine Gesamtzahl von 241.984 Transkriptionsfaktorbindungsstellen. Anhand der Microarray-Expressionsdaten wurde die Gesamtgruppe der verfügbaren Gene und Promotoren in zwei Gruppen unterteilt, nämlich in die Gruppe der Gene, deren Expression in der Leber deutlich am höchsten gefunden wurde und in die Gruppe der Gene, die in anderen Geweben am höchsten exprimiert waren. Jeder potentiell bindende Transkriptionsfaktor wurde auf unterschiedliches Vorkommen in diesen beiden Gruppen hin untersucht. Dies geschah unter der Vorstellung, dass übergeordnete Faktoren, die eine Expression in der Leber stimulieren in der Gruppe der Gene, die in der Leber am höchsten exprimiert sind, verhältnismäßig wesentlich häufiger zu finden sein könnten. Eine solches häufigeres Vorkommen ließ sich jedoch für keinen einzigen Faktor nachweisen.

Der verwendete PromotorScan-Algorithmus erkennt nur bisher molekularbiologisch bereits identifizierte und bekannte Transkriptionsfaktorbindungsstellen. Solche Transkriptionsfaktorbindungsstellen sind typischerweise zwischen 5 und 15 bp lang. Um auszuschließen, dass mit dem verwendeten PromotorScan-Algorithmus Transkriptionsfaktorbindungsstellen, die bisher nicht bekannt sind, nicht übersehen wurden, wurden die Häufigkeit sämtlicher möglicher 8 bp (4^8) und 10 bp (4^{10}) Nukleotid-Kombinationen in diesen Promotoren untersucht. Biologisch relevante Unterschiede fanden sich zwischen den beiden Gruppen nicht.

In gleicher Weise wurde auch die Bedeutung von TATA-Boxen untersucht. TATA-Boxen kommt bei der Transkriptionsinitiierung eine wichtige Rolle zu, indem über sie die Bindung des initialen Transkriptionskomplexes vermittelt wird. Insgesamt 1033 TATA-Boxen wurden

ebenfalls mittels PromotorScan vorausgesagt. Dabei waren 57 auf Promotoren von Genen, die in der Leber überexprimiert waren und 976 auf Promotoren von Genen, die in anderen Geweben überexprimiert waren. Der Vergleich dieser beiden Gruppen ließ keine signifikant unterschiedliche Häufigkeit an TATA-Boxen erkennen.

Im weiteren wurde die Bedeutung von CpG-Islands für eine potentiell differentielle Regulation untersucht. Insgesamt wurden 8742 CpG-Islands in einem Bereich von bis zu 5 kb upstream des Transkriptionsstarts identifiziert, 364 davon auf Promotoren von Genen, die am höchsten in der Leber exprimiert waren, 8378 auf Promotoren von Genen, die in anderen Geweben am höchsten exprimiert waren. Signifikante Unterschiede in der Verteilung von CpG-Islands auf Promotoren dieser beiden Gengruppen ließen sich nicht nachweisen.

Schließlich wurden die RNA- und Proteinsequenzen des Transkriptoms und Proteoms hinsichtlich ihrer Zusammensetzung aus einzelnen Nukleotiden bzw. Aminosäuren analysiert. Auch hierbei fanden sich keine signifikanten Unterschiede in der Verteilung zwischen beiden Gengruppen.

Die Zusammenschau der Ergebnisse zeigt, dass die Regulation der einzelnen Gene im Lebergewebe im wesentlichen individuell erfolgt. Im Rahmen der vorgelegten bioinformatischen Analysen fanden sich keine übergeordneten genetischen „Leberfaktoren“, die speziell eine Expression von Genen in der Leber stimulieren. Neue therapeutische Ansätze, die auf eine Regulation der Genexpression in der Leber zielen, werden somit auch weiterhin auf die Beeinflussung individueller Gene fokussiert bleiben.

6 Literatur

1. Corless JK, Middleton HM, 3rd: Normal liver function. A basis for understanding hepatic disease. *Arch Intern Med* 1983, 143(12):2291-2294.
2. Maxam AM, Gilbert W: A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 1977, 74(2):560-564.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.
5. Venter JC, Smith HO, Hood L: A new strategy for genome sequencing. *Nature* 1996, 381(6581):364-366.
6. Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. *Nature* 1999, 401(6756):921-923.
7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: The sequence of the human genome. *Science* 2001, 291(5507):1304-1351.
8. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank: update. *Nucleic Acids Res* 2004, 32(Database issue):D23-26.
9. Sikela JM, Auffray C: Finding new genes faster than ever. *Nat Genet* 1993, 3(3):189-191.
10. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr., Hieter P, Vogelstein B, Kinzler KW: Characterization of the yeast transcriptome. *Cell* 1997, 88(2):243-251.

11. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. *Science* 1995, 270(5235):484-487.

12. Villard J: Transcription regulation and human diseases. *Swiss Med Wkly* 2004, 134(39-40):571-579.

13. Carey M: Mechanistic advances in eukaryotic gene activation. *Curr Opin Cell Biol* 1991, 3(3):452-460.

14. Johnson SA, Dubeau L, White RJ, Johnson DL: The TATA-binding protein as a regulator of cellular transformation. *Cell Cycle* 2003, 2(5):442-444.

15. Kim J, Iyer VR: Global role of TATA box-binding protein recruitment to promoters in mediating gene expression profiles. *Mol Cell Biol* 2004, 24(18):8104-8112.

16. Deutschman CS: Transcription. *Crit Care Med* 2005, 33(12 Suppl):S400-403.

17. Maniatis T, Goodbourn S, Fischer JA: Regulation of inducible and tissue-specific gene expression. *Science* 1987, 236(4806):1237-1245.

18. Antequera F, Bird A: Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* 1993, 90(24):11995-11999.

19. Jones PA, Rideout WM, 3rd, Shen JC, Spruck CH, Tsai YC: Methylation, mutation and cancer. *Bioessays* 1992, 14(1):33-36.

20. Larsen F, Gundersen G, Lopez R, Prydz H: CpG islands as gene markers in the human genome. *Genomics* 1992, 13(4):1095-1107.

21. Takai D, Jones PA: The CpG island searcher: a new WWW resource. *In Silico Biol* 2003, 3(3):235-240.

22. Picketts DJ, Mueller CR, Lillicrap D: Transcriptional control of the factor IX gene: analysis of five cis-acting elements and the deleterious effects of naturally occurring hemophilia B Leyden mutations. *Blood* 1994, 84(9):2992-3000.

23. Liu H, Peng CH, Liu YB, Wu YL, Zhao ZM, Wang Y, Han BS: Inhibitory effect of adeno-associated virus-mediated gene transfer of human endostatin on hepatocellular carcinoma. *World J Gastroenterol* 2005, 11(22):3331-3334.

24. Inoue H, Shiraki K, Murata K, Sugimoto K, Kawakita T, Yamaguchi Y, Saitou Y, Enokimura N, Yamamoto N, Yamanaka Y *et al*: Adenoviral-mediated transfer of p53 gene enhances TRAIL-induced apoptosis in human hepatocellular carcinoma cells. *Int J Mol Med* 2004, 14(2):271-275.

25. Dumortier J, Schonig K, Oberwinkler H, Low R, Giese T, Bujard H, Schirmacher P, Protzer U: Liver-specific expression of interferon gamma following adenoviral gene transfer controls hepatitis B virus replication in mice. *Gene Ther* 2005, 12(8):668-677.

26. Vidal L, Blagden S, Attard G, de Bono J: Making sense of antisense. *Eur J Cancer* 2005, 41(18):2812-2818.

27. Maret A, Galy B, Arnaud E, Bayard F, Prats H: Inhibition of fibroblast growth factor 2 expression by antisense RNA induced a loss of the transformed phenotype in a human hepatoma cell line. *Cancer Res* 1995, 55(21):5075-5079.

28. Ellouk-Achard S, Djenabi S, De Oliveira GA, Desauty G, Duc HT, Zohair M, Trojan J, Claude JR, Sarasin A, Lafarge-Frayssinet C: Induction of apoptosis in rat hepatocarcinoma cells by expression of IGF-I antisense c-DNA. *J Hepatol* 1998, 29(5):807-818.

29. Seo MY, Abrignani S, Houghton M, Han JH: Small interfering RNA-mediated inhibition of hepatitis C virus replication in the human hepatoma cell line Huh-7. *J Virol* 2003, 77(1):810-812.

30. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F *et al*: Ensembl 2005. *Nucleic Acids Res* 2005, 33(Database issue):D447-453.

31. www.perl.org.

32. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437(7057):376-380.

33. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005, 309(5741):1728-1732.

34. http://www.ddbj.nig.ac.jp/intro-e.html.

35. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W *et al*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005, 33(Database issue):D39-45.

36. Stormo GD: DNA binding sites: representation and discovery. *Bioinformatics* 2000, 16(1):16-23.

37. Shmueli O, Horn-Saban S, Chalifa-Caspi V, Shmoish M, Ophir R, Benjamin-Rodrig H, Safran M, Domany E, Lancet D: GeneNote: whole genome expression profiles in normal human tissues. *C R Biol* 2003, 326(10-11):1067-1072.

38. Prestridge DS: Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* 1995, 249(5):923-932.

39. Baylin SB: DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol* 2005, 2 Suppl 1:S4-11.

40. Alaoui-Jamali MA, Dupre I, Qiang H: Prediction of drug sensitivity and drug resistance in cancer by transcriptional and proteomic profiling. *Drug Resist Updat* 2004, 7(4-5):245-255.

41. Burczynski ME, Oestreicher JL, Cahilly MJ, Mounts DP, Whitley MZ, Speicher LA, Trepicchio WL: Clinical pharmacogenomics and transcriptional profiling in early phase oncology clinical trials. *Curr Mol Med* 2005, 5(1):83-102.

42. Piquemal D, Commes T, Manchon L, Lejeune M, Ferraz C, Pugnere D, Demaille J, Elalouf JM, Marti J: Transcriptome analysis of monocytic leukemia cell differentiation. *Genomics* 2002, 80(3):361-371.

43. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP *et al*: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002, 1(2):203-209.

44. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Bernards R, Friend SH: Expression profiling predicts outcome in breast cancer. *Breast Cancer Res* 2003, 5(1):57-58.

45. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002, 415(6871):530-536.

46. Zou TT, Selaru FM, Xu Y, Shustova V, Yin J, Mori Y, Shibata D, Sato F, Wang S, Olaru A *et al*: Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene* 2002, 21(31):4855-4862.

47. Muller M, Wilder S, Bannasch D, Israeli D, Lehlbach K, Li-Weber M, Friedman SL, Galle PR, Stremmel W, Oren M *et al*: p53 activates the CD95 (APO-1/Fas) gene in response to DNA damage by anticancer drugs. *J Exp Med* 1998, 188(11):2033-2045.

48. Huang Y, Li Y: Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* 2004, 20(1):21-28.

49. Fujiwara Y, Asogawa M: Prediction of subcellular localizations using amino acid composition and order. *Genome Inform Ser Workshop Genome Inform* 2001, 12:103-112.

7 Lebenslauf

8 Danksagung