Vergleichende Analyse von Transkriptionsfaktor-Genfamilien am Beispiel der R2R3-MYB-Transkriptionsfaktoren aus Ackerschmalwand (*Arabidopsis thaliana*) und Reis (*Oryza sativa*) mit bioinformatischen Methoden

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Martin Werber

aus Hamburg

angefertigt am Max-Planck-Institut für Züchtungsforschung

Köln, 2004

Berichterstatter:	Prof. Dr. Bernd Weißhaar
	Prof. Dr. Diethard Tautz
Prüfungsvorsitzender:	Prof. Dr. Martin Hülskamp

Tag der mündlichen Prüfung: 4.7.2003

Zusammenfassung

Die Proteinfamilie der MYB-Transkriptionsfaktoren ist in eukaryontischen Organismen weit verbreitet und reguliert eine Vielzahl zellulärer Prozesse.
In der vorliegenden Arbeit wurden Mitglieder einer pflanzenspezifischen MYB-Subfamilie, die R2R3-MYBs, mit Hilfe bioinformatischer Analysen identifiziert und strukturell charakterisiert. R2R3-MYBs zeichnen sich durch eine aus zwei Sequenzwiederholungen bestehende stark konservierte DNA-Bindedomäne aus. Mit der Veröffentlichung der pflanzlichen Genomsequenzen von *Arabidopsis thaliana* und *Oryza sativa* bot sich erstmals die Möglichkeit, die Mitglieder dieser Subfamilie für die jeweiligen Spezies zu bestimmen und in Umfang und Struktur zu vergleichen. Um die ständig in Veränderung begriffene und anwachsende Menge von nur teilweise annotierten genomischen Sequenzdaten zu durchsuchen wurden Algorithmen, Verfahren der Bioinformatik und Datenbankkonzepte entwickelt und zur Klassifizierung der R2R3-MYBs eingesetzt.
Als Ergebnis dieser Analysen konnten 24 R2R3-MYB-Gene, drei 3R-MYB-Gene und ein 4R-MYB aus *Arabidopsis thaliana* und 114 R2R3-MYB-Gene, fünf 3R-MYB-Gene und ein 4R Gen aus *Oryza sativa* neu beschrieben und eingeordnet werden. Die Anzahl der R2R3-MYB-Gene und ihre Einordnung in Gruppen unterstützen die These, dass die Amplifikation der R2R3-MYB-Genfamilie vor der Aufspaltung der Pflanzen in Dikotyledonen und Monokotyledonen stattfand. Zudem ergaben vergleichende Analysen der C-terminalen Bereiche der R2R3-MYB-Proteinsequenzen aus beiden Organismen, dass die Motive trotz der Distanz der Organismen zueinander über große Zeiträume der Evolution konserviert worden sind. Programme und Datenbanken wurden so konzipiert, dass sie für weitere Fragestellungen eingesetzt werden können.

Summary

The family of MYB transcription factors is conserved among eukaryotic organisms and regulates a number of cellular processes.

Here, the identification and structural characterization of R2R3-MYBs, a plant-specific subfamily of MYB transcription factors, is described. R2R3-MYBs carry a highly conserved DNA-binding domain of two repetitive sequences. The now available genomic sequences of *Arabidopsis thaliana* and *Oryza sativa* for the first time allowed the complete detection of all R2R3-MYB family members in the corresponding organisms and their comparison in terms of species-dependent structure and abundance. For the handling of increasing and continuously changing information on so far only partially annotated genomic sequence data, algorithms, bioinformatics techniques and database concepts were developed and employed for R2R3-MYB factor classification.

24 R2R3-MYB genes, 3 3R-MYB genes and 1 4R-MYB from *Arabidopsis thaliana* und 114 R2R3-MYB genes, 5 3R-MYB genes and one 4R-MYB gene of *Oryza sativa* were thus newly described and grouped. The number of R2R3-MYB genes and their classification support the hypothesis that amplification of R2R3-MYBs occurred prior to the separation of plants into monocots and dicots. In addition comparative analyses of the C-terminal regions of from both organisms further revealed that motives have remained conserved despite the large evolutionary difference between these species. Software and databases developed for the analyses of R2R3-MYBs are generally applicable and can be used for other analyses.

Vorwort

Die Miniaturisierung, Automatisierung und der Einsatz von parallel ausgelegten Methoden in der Genomforschung hat einen Paradigmenwechsel in den Bio-Wissenschaften bewirkt. Große Datenmengen werden erzeugt, die gespeichert, aufbereitet, analysiert und Wissenschaftlern und Anwendern in geeigneter Form zugänglich gemacht werden müssen. Mit der Veröffentlichung des ersten Pflanzengenoms sind die Voraussetzungen für den Eintritt der Pflanzenbiologie in das Zeitalter der Genomik geschaffen worden. Die Möglichkeit molekulare und biochemische Zusammenhänge in kompletten Genomen zu untersuchen wird zu einer Beschleunigung der Forschung und einer raschen Zunahme von Erkenntnissen führen. Eine Aufgabe der Bioinformatik ist es unter anderem, für die Bearbeitung der Datenmengen geeignete Werkzeuge zu entwickeln und bereitzustellen.

Pflanzliche Genome kennzeichnet das ausgeprägte Vorkommen von großen Genfamilien. Hieran haben besonders die stark amplifizierten Transkriptions-faktorfamilien einen großen Anteil. Bei Genfamilien und deren Untergruppen handelt es sich um sehr ähnliche Gene, deren Proteinsequenz eine oder mehrere konservierte Domänen aufweisen, die wesentlichen Anteil an der biologischen Funktion haben. Eine besondere Aufgabe der Genomforschung im pflanzlichen Bereich ist es, die Funktionen der Mitglieder der großen Genfamilien zu bestimmen. Dabei können die Funktionen auch abhängig vom Zelltyp und Gewebe variieren. Dieser Zustand wird häufig als Redundanz beschrieben, obwohl davon ausgegangen werden kann, dass bis auf Pseudogene jedes Gen seine eigene biologische Relevanz hat. Die systematische Untersuchung von derartigen oft sehr großen Genfamilien setzt ein auf Sequenzmotiven und konservierten Domänen basierendes Klassifizierungsschema für die einzelnen Mitglieder voraus. Anhand sequenzierter Genome von Modellorganismen kann eine entsprechende Klassifizierung entwickelt werden. Dabei ist die Festlegung und gegebenenfalls Verfeinerung der Kriterien zur Definition einer Familie eine eigene biologische und bioinformatische Fragestellung.

Vollständig sequenzierte Genome für Pflanzen werden aber auf absehbare Zeit die Ausnahme bleiben. Es stellt sich daher die Frage, wie sich auf der Basis der vorhandenen Sequenzdaten der Modellorganismen und den mit diesen Daten erstellten Genfamiliendefinitionen Verfahren entwickeln lassen, die das Auffinden und Klassifizieren von Genfamilien in neuen Sequenzsammlungen anderer Pflanzen vereinfachen und beschleunigen.

In der vorliegende Arbeit werden Algorithmen, Verfahren der Bioinformatik und Datenbankkonzepte entwickelt, um die verschiedenen Arbeitsschritte für die Erfassung und Klassifizierung einer großen Genfamilie zu automatisieren und wo dies nicht möglich ist zumindest zu unterstützen.

Auf der Basis dieser erarbeiteten Programme und Datenbankkonzepte erfolgt die vergleichende Charakterisierung der R2R3-MYB-Subfamilie in *Arabidopsis thaliana* und *Oryza sativa.* Programme und Datenbanken wurden so konzipiert, dass sie sich auch für weitere Fragestellungen nutzen lassen. Die in dieser Arbeit entwickelten Programme und Datenbanken wurden in verschiedenen Projekten erfolgreich eingesetzt (Stracke, Werber et al. 2001; Bellin D, Werber M et al. 2002; Heim, Jakoby et al. 2003; Hunger, Di Gaspero et al. 2003).

1 Einleitung

1.1 Genfamilien

Als Genfamilien werden Gruppen von Genen mit sehr ähnlichen Sequenzen und einer oder mehreren konservierten Domänen bezeichnet. Häufig ist die molekulare Wirkungsweise der Mitglieder einer Genfamilie z.B. DNA-Bindung, die durch die charakteristischen Domänen bestimmt wird, sehr ähnlich. Variationen in der Anordnung zueinander und Substitutionen in den Domänen führen zu Veränderungen in der molekularen Funktion. Aufgrund der Abhängigkeit der molekularen Funktion von diesen Merkmalen können die beschriebenen Variationen für die Charakterisierung einer Genfamilie genutzt werden. Zusätzlich lassen sich Genfamilien auch über den Grad der Sequenzähnlichkeit definieren. Stringentere Bedingungen können zu einer Einteilung in Subfamilien, weniger stringente zu Superfamilien führen. Ähnliche Sequenzen können aber auch Ergebnis konvergenter Evolution sein. In diesem Fall werden die Gene nicht in einer Familie zusammengefasst. Homologie ist also zwingende Voraussetzung für die Gruppierung in eine Genfamilie.

1.2 Transkriptionsfaktoren

Alle RNA-Moleküle in der Zelle werden durch einen Prozess gebildet, bei dem die Nukleotidabfolge eines Genabschnittes auf der DNA mit Hilfe von DNA-abhängigen RNA-Polymerasen in RNA transkribiert wird. Für den Prozess der Transkription werden eine Reihe von verschiedenen Proteinen und Proteinkomplexen benötigt (Roeder 1991) (Orphanides, Lagrange et al. 1996). Dabei lassen sich die beteiligten Proteine in verschiedene Klassen einteilen:

Sequenzspezifische DNA-bindende Regulatoren: Diese binden an die Promotoren der Gene und können aktivieren oder reprimieren.

Allgemeine Transkriptionsfaktoren: Diese sind ubiquitär und bilden zusammen mit RNA-Polymerase II den Pre-Initiierungskomplex (PIC).

Co-Faktoren / Regulatoren: Diese binden an die Transkriptionsfaktoren der Klassen I und II und sind so an der Regulation beteiligt.

Eine weitere eigene Klasse sind die Faktoren, die an der Entwindung und Remodellierung der DNA beteiligt sind.

Da die DNA der Transkription räumlich nur zugänglich ist, wenn sie von den Nukleosomen abgelöst und aufgespreizt wird, haben auch diese Faktoren regulative Eigenschaften (Singh 1998).

Die Faktoren der Klasse II bilden den PIC mit einer Masse von 2 MDa, zusammengesetzt aus über 40 verschiedenen Proteinen. Die Faktoren der Klasse III können sowohl an die Proteine des PIC als an Transkriptionsfaktoren der Klasse I binden. Die Faktoren der Klassen II, III und IV sind über die verschiedenen Organismenreiche hochkonserviert (Larkin, Hagen et al. 1999) (Baldwin and Gurley 1996). Im Folgenden bezeichnet der Begriff Transkriptionsfaktoren, soweit nicht explizit darauf hingewiesen wird, Transkriptionfaktoren der Klasse I.

Die Effizienz, Spezifität und Sensitivität der Transkriptionsregulation wird durch die Kombination von Transkriptionsfaktoren aller oben beschriebenen Klassen ermöglicht. Die Transkription von Genen wird abhängig von Gewebe, Zelltyp, Zeitpunkt der Entwicklung, Reaktion auf endogene und exogene Faktoren reguliert. Regulatorische Unterschiede scheinen eine wesentliche Grundlage für die Diversität der Organismen und evolutionäre Entwicklung zu sein. Diese Leistung ist nur möglich, indem die beschriebenen Klassen, unter den verschiedenen Bedingungen kombiniert wirken (Singh 1998). Die Aktivität der Transkriptionsfaktoren kann durch Änderungen an der translatierten Proteinsequenz, so genannte posttranslationale Modifikationen wie z.B. Phosphorylierung beeinflusst werden. Dadurch können die DNA-Bindeeigenschaften und Protein-Interaktionseigenschaften durch einen weiteren Mechanismus reguliert werden (Boyle, Smeal et al. 1991; Hunter and Karin 1992; Hicke, Rempel et al. 1995). Transkriptionsfaktoren weisen bezüglich ihrer Sequenz eine modulare Struktur von höher konservierten Bereichen auf. Neben einer DNA-Bindedomäne sind meist Aktivierungs- und Reprimierungsdomänen vorhanden. Unter anderem auf Basis von strukturellen Ähnlichkeiten der einzelnen Module können Transkriptionsfaktoren in Familien und Subfamilien klassifiziert werden (Singh 1998).

1.3 Spezifische DNA-Erkennungssequenzen

Eine weitere Möglichkeit der Unterteilung in Gruppen bieten die spezifischen DNA-Bindesequenzen (Ghosh 1992; Wingender 1994). Diese sind wesentlich durch die Sequenz der DNA-Bindedomäne bedingt. Bei der genomweiten Analyse von Promotorsequenzen lassen sich jedoch wesentlich mehr potentielle Bindestellen

feststellen, als in vivo tatsächlich genutzt werden. Dieser Überfluss an Bindestellen ist ein weiterer Hinweis auf das kombinierte Wirken einer Vielzahl von Faktoren bei der Erkennung der Bindesequenz, die weit über die bloße Erkennung eines kurzen Nukleotidabschnittes hinausgeht.

1.4 Transkriptionsfaktorfamilien in Pflanzen

Mit der Veröffentlichung der ersten pflanzlichen Genomsequenz von *Arabidopsis thaliana* konnte erstmals die Anzahl der Transkriptionsfaktoren in einer Pflanze abgeschätzt werden.

Nach Analysen der Genomsequenz kodiert die Sequenz von *Arabidopsis thaliana* für 1533 Transkriptionsfaktoren das sind ca. 5.9 % aller Gene (Riechmann, Heard et al. 2000). Diese Zahl ist wahrscheinlich noch unterschätzt, da für viele Gene noch keine Funktion bekannt ist. Im Vergleich zu Tieren, Insekten und Pilzen kodiert das Genom von *Arabidopsis thaliana* eine prozentual höhere Zahl von Transkriptionsfaktoren (Riechmann, Heard et al. 2000). Aufgrund dieser Abschätzung sind die drei größten Transkriptionsfaktorfamilien im Genom von *Arabidopsis thaliana* AP2/EREP, MYB und bHLH. In *Arabidopsis thaliana* ist ein deutlich geringerer Anteil von Zink-Koordinierenden Transkriptionsfaktoren vorhanden (22%) als in Tieren und Pilzen (>55%). Mehrere Transkriptionsfaktorfamilien sind nur in Pflanzen vorhanden. Dazu gehören EREBP, NAC, WRKY, R2R3-MYB, Trihelix-Transkriptionsfaktoren, Aux/IAA-Proteine. Einige Domänen sind einzigartig in ihrem Aufbau und kommen nur in Pflanzen vor. Dazu gehört die AP2 Domäne der AP2/EREP-Familie, bei der eine neue Form der DNA-Erkennung durch ß-Faltblätter besteht. Die im Vergleich zu Tieren und Pilzen am stärksten amplifizierten Genfamilien sind die MYB- und die MADS-Genfamilie. Wenn die Anzahl und Komplexität von Domänenkombinationen mit der Komplexität des Organismus korreliert, sind Pflanzen in ihrer Ausstattung mit Transkriptionsfaktorgenen mindestens so komplex wie Tiere und Pilze. Die prozentual größere Anzahl von Transkriptionsfaktoren und die starke Amplifikation einzelner Familien, kann als pflanzenspezifisch angesehen werden. Sie stellt wahrscheinlich eine besondere Anpassung an den sesshaften Lebenswandel der Pflanzen und das damit verbundene notwendige Reaktionsrepertoire auf exogene Einflüsse dar. Mitglieder der besonders amplifizierten Genfamilien MYB und MADS sind an der Entwicklungssteuerung von pflanzenspezifischen Geweben und Organen beteiligt (Riechmann, Heard et al. 2000).

1.5 MYB-Transkriptionsfaktoren

MYB-Transkriptionsfaktoren repräsentieren eine Familie von Proteinen, die eine spezielle konservierte DNA-Bindedomäne aufweisen, die als MYB-Domäne bezeichnet wird. Als erstes MYB-Gen beschrieben wurde das Oncogen v-MYB. Der Name leitet sich von „avian myoblastosis virus“ (AMV) ab. AMV ist ein oncogener Retrovirus, der in Tieren und Menschen myoblastische Leukämie verursacht und myeloide Zellen verändert. v-MYB ist eine veränderte Variante des zellulären c-MYB-Gens in tierischen Zellen (Klempnauer, Gonda et al. 1982; Klempnauer, Ramsay et al. 1983). Neben dem c-MYB-Gen wurden mit a-MYB und b-MYB zwei weitere MYB-Gene in tierischen Organismen entdeckt (Nomura, Takahashi et al. 1988). c-MYB, a-MYB und b-MYB Gene sind an der Kontrolle der Zellteilung und der Zelldifferenzierung beteiligt (Weston and Bishop 1989; Oh and Reddy 1999). Des Weiteren wurden c-MYB, a-MYB und b-MYB ähnliche Gene in Insekten, Pflanzen, Pilzen und Schleimpilzen beschrieben (Lipsick 1996).

Die DNA-Bindedomäne von c-MYB besteht aus drei nicht perfekten Sequenzwiederholungen von etwa 53 Aminosäuren (R1,R2,R3) (Sakura, Kanei-Ishii et al. 1989). Bei der Entstehung des v-MYB Gens ging der größte Teil der ersten Sequenzwiederholung verloren (R1) (Klempnauer, Gonda et al. 1982). Diese erste Sequenzwiederholung ist nicht direkt in die DNA-Bindung involviert, trägt aber zur Stabilisierungen bei. Der kleinste zur DNA-Bindung fähige Bereich ist für c-MYB auf die zweite und dritte Sequenzwiederholung eingegrenzt worden (Gabrielsen, Sentenac et al. 1991; Dini and Lipsick 1993; Ebneth, Schweers et al. 1994; Ogata, Morikawa et al. 1994). Jede Sequenzwiederholung bildet eine *helix-turn-helix* Struktur aus drei Helices. Drei Tryptophan-Reste in regelmäßigen Abständen von 18-19 Aminosäuren bilden einen hydrophoben Kern und sind charakteristisch für eine Sequenzwiederholung in der MYB-Domäne. Der hydrophobe Kern und die Tryptophan-Reste spielen eine wichtige Rolle bei der sequenzspezifischen DNA-Bindung (Ogata, Morikawa et al. 1994; Sasaki, Ogata et al. 2000).

MYB-Proteine können abhängig von der Anzahl der Sequenzwiederholungen in Subfamilien unterteilt werden. Proteine mit einer Wiederholung werden als 1R, solche mit zwei Sequenzwiederholungen, bei denen die Sequenzwiederholungen homolog zur zweiten und dritten Sequenzwiederholung aus c-MYB sind, werden als R2R3 und solche mit drei Sequenzwiederholungen werde als 3R bezeichnet (Kranz, Denekamp et al. 1998; Kranz, Scholz et al. 2000).

In Tieren und Pilzen wurden nur die oben aufgeführten Proteine mit drei Sequenzwiederholungen beschrieben. Es gibt weiterhin Proteine, die eine MYB-ähnliche DNA-Bindedomäne aufweisen. Diese werden jedoch allgemein nicht als MYB-Proteine, sondern entsprechend ihrer Funktion bezeichnet. MYB-Proteine mit einer Sequenzwiederholung und geringer Ähnlichkeit in der Domäne werden als MYB-ähnlich klassifiziert.

Abbildung 1: Konservierte Sequenzmotive in MYB-**Genen. R1,R2,R3:**MYB-**Sequenzwiederholungen; ACT: Aktivator Element; REPR: Repressor Element; MLR3: MYB-ähnliche Sequenzwiederholung; K: putatives Kinase Motiv; [???]: Subgruppen Motiv**

Die pflanzenspezifischen MYB-Proteine mit zwei Sequenzwiederholungen bilden die größte Subfamilie innerhalb der MYB-Genfamilie mit mehr als 100 Mitgliedern (Martin and Paz-Ares 1997; Kranz, Denekamp et al. 1998; Romero, Fuertes et al. 1998). Die MYB-Domäne der 2R3R MYB-Proteine ist homolog zu den R2 und R3 Sequenzwiederholungen aus c-MYB. Außerhalb der MYB-Domäne im C-terminalen Bereich finden sich weitere kurze konservierte Abschnitte, nach denen sich die Proteine dieser Subfamilie in Untergruppen einteilen lassen (Kranz, Denekamp et al.

1998). Serin und Threonin Reste im C-terminalen Teil der Proteine können Hinweise auf posttranslationale Modifikationen sein (Martin and Paz-Ares 1997).

MYB-Proteine mit drei Sequenzwiederholungen (3R, pc-MYB) wurden für verschiedene Landpflanzen beschrieben (Braun and Grotewold 1999; Kranz, Scholz et al. 2000). NtMYB-Gene aus *Nicotiana tabacum* mit drei Sequenzwiederholungen sind an der Kontrolle der Zellteilung beteiligt (Ito, Araki et al. 2001). MYB-Proteine mit einer Sequenzwiederholung bilden eine funktionell heterogene Gruppe von DNA bindenden Proteinen. Einige Proteine weisen eine verkürzte Bindedomäne auf und können daher besser den MYB-ähnlichen Proteinen zugeordnet werden.

Neben der MYB-Genfamilie gibt es weitere Gene und Genfamilien, die ähnliche DNA-Bindedomänen aufweisen. Hierzu gehören CDC5, MIDA1, GARP und TBP. Diese Familien werden aufgrund der zwar ähnlichen aber divergenten DNA Bindedomäne auch unter der Bezeichnung MYB-ähnliche Proteine zusammengefasst, obwohl sie biologisch in sehr unterschiedlichen Zusammenhängen einzuordnen sind.

Die DNA-Bindedomäne von cdc5+ weist eine signifikante Ähnlichkeit zu MYB-Domänen auf. Unterschiedlich sind jedoch die für MYB beschriebenen hochkonservierte Aminosäuren, die an der DNA-Bindung und an der Interaktion der Helices beteiligt sind (Lys-128 (R2), Lys-182, Asn-183 (R3)). cdc5+ besteht aus drei Sequenzwiederholungen, von denen nur die ersten beiden eine hohe Ähnlichkeit zu c-MYB aufweisen, während die dritte Sequenzwiederholung nur schwach konserviert ist. Das cdc5+ Gen wurde bei einem Screen für Zellteilungsmutanten in Hefe entdeckt. Bei der Zellteilung ist cdc5+ in die G2-Phase involviert. Die Proteinsequenz weist Ähnlichkeiten zu c-MYB auf. Es wurde gezeigt, dass der N-terminale Teil von cdc5+ der die MYB-DNA-Bindedomäne kodiert, ausreichend ist um die Mutante zu komplementieren (Ohi, McCollum et al. 1994). Auch in Pflanzen wurde ein cdc5+-Homolog entdeckt (AtCDC5) (Hirayama and Shinozaki 1996). Aufgrund der hohen Ähnlichkeit und der Fähigkeit Hefemutanten zu komplementieren kann für AtCDC5 eine ähnliche Funktion wie cdc5+ angenommen werden. Während AtCDC5 in der N-terminalen DNA-Bindedomäne eine hohe Ähnlichkeit zu cdc5+ aufweist, sind im C-terminalen Bereich andere Motive vorhanden. Darunter sind mehrere Zielmotive für Serin/Threoin-Kinasen, die auf eine posttranslationale Regulation hinweisen könnten. Durch *In-vitro*-Experimente wurde die spezifische Bindesequenz für AtCDC5 als „CTCAGCG“ bestimmt (Hirayama and Shinozaki 1996).

MIDA1 Proteine sind wahrscheinlich an der Regulation der Zellteilung und des Zellwachstums beteiligt. Sie enthalten N-terminal eine J-Domäne und einen Bereich mit Ähnlichkeit zu Z-DNA Bindeproteinen aus Pilzen (Zuotin). Im C-terminalen Bereich der Proteine befindet sich eine MYB-Domäne mit zwei MYB-Sequenzwiederholungen. Durch *In-vitro*-Experimente konnte die spezifische Bindesequenz als „GTCAAGC“ beschrieben werden. Sie sind in verschiedenen Eukaryonten (Pflanzen, Tieren, Pilzen) nachgewiesen worden (Inoue, Shoji et al. 1999).

Mitglieder der Familie der GARP-Proteine sind an der Regulation der Differenzierung zu photosynthetisch aktiven Zellen (Hall, Rossini et al. 1998) und an der Regulation des Phosphormetabolismus beteiligt (PSR1 C. reinhardtii)(Wykoff, Grossman et al. 1999). Damit sind sie wie die pflanzlichen R2R3-MYB-Proteine Regulatoren von originären Stoffwechseln der Pflanze. Sie enthalten eine C–terminale Sequenzwiederholung mit DNA-Bindeeigenschaften, die Ähnlichkeit zu c-MYB Sequenzwiederholungen aufweist. N-Terminal befindet sich eine „phospho-accepting-receiver“ Domäne. Die MYB-ähnliche Domäne der GARP Familie besteht aus ca. 60 Aminosäuren und unterscheidet sich stark von der Konsensussequenz typischer pflanzlicher MYB-Proteine. Es fehlen die drei Tryptophan-Reste im Abstand von 18-19 AS und auch andere Positionen, die an der DNA-Bindung direkt beteiligt sind, weisen Veränderungen auf. Damit übereinstimmend ist auch die spezifische Bindesequenz gegenüber den klassischen MYB-Proteinen mit „AGAT(TCG/CTT)“ verändert (Riechmann, Heard et al. 2000).

„Telomeric binding proteins“ TBP sind an der Sicherstellung der chromosomalen Integrität beteiligt. Sie besitzen C-terminal eine MYB-ähnliche DNA-Bindedomäne mit einer Sequenzwiederholung. Die DNA-Bindedomäne besitzt drei typische Tryptophanreste. Andere Aminosäuren sind jedoch gegenüber dem MYB-Konsensus verändert (Bilaud, Koering et al. 1996; Yu, Kim et al. 2000). Die spezifische telomerische DNA-Bindesequenz wurde *in-vitro* als „GGTTTAG“ bestimmt.

Die Grundlage für die Analyse der MYB-Genfamilie bilden die oben beschrieben Klassifizierungen nach Anzahl der Sequenzwiederholungen. Dabei muss zunächst zwischen der Situation in Tieren und Pflanzen unterschieden werden. Die Anzahl tierischer MYB-Proteine ist vergleichsweise klein und umfasst ca. zehn Proteine, wenn auch die MYB-ähnlichen Proteine mit einbezogen werden. MYB-Proteine mit nur einer Sequenzwiederholung stellen eine heterogene Gruppe mit geringer

Sequenzähnlichkeit im Vergleich zu allen anderen MYB-Proteinen dar, sie sind für Pilze, Pflanzen und Tiere beschrieben worden. MYB-Proteine vom R2R3-Typ, mit zwei zu c-MYB homologen Sequenzwiederholungen, sind bislang nur in Pflanzen nachgewiesen worden. Hier ist diese Subfamilie mit mehr als 100 Mitgliedern stark amplifiziert. Phylogenetische Studien auf der Basis von c-MYB homologen Proteinen aus verschiedenen Organismen geben Hinweise, dass ein Vorfahre der R2R3-MYB-Proteine drei Sequenzwiederholungen hatte und die erste Sequenzwiederholung verloren ging (Jin and Martin 1999). Die starke Amplifikation der R2R3-Subfamilie konnte sowohl in dikotyledonen als auch in monokotyledonen Pflanzen nachgewiesen werden (Romero, Fuertes et al. 1998; Rabinowicz, Braun et al. 1999). In dem Moos P. patens wurden dagegen bisher nur zwei MYB-Proteine nachgewiesen (Leech, Kammerer et al. 1993). Es kann daher angenommen werden, dass die Amplifikation mit der Entwicklung der Landpflanzen einherging. Ein weiterer Hinweis für diese These sind die bisher nachgewiesenen Funktionen von R2R3-MYB-Proteinen, die im Folgenden beschrieben werden.

1.6 Funktionen pflanzlicher R2R3 MYB-Proteine

Bisher beschriebene R2R3-MYB-Proteine sind an der Regulation von Stoffwechselwegen beteiligt, die für Pflanzen spezifisch sind.

Dazu gehören die Regulation des Sekundärstoffwechsels, Steuerung der Zellform und die Reaktion auf Hormone bzw. auf exogene Faktoren, wie Pathogenbefall oder Trockenstress. ZmMYB1, AmMYB305 und AmMYB340 sind an der Regulation des Antocyaninstoffwechsels beteiligt (Cone, Burr et al. 1986; Paz-Ares, Wienand et al. 1986). PhMYB1 und AmMIXTA regulieren die Bildung der Epidermalzellen von Petalen (Oppenheimer, Herman et al. 1991; Noda, Glover et al. 1994). AtGL1 (AtMYB0) ist an der Bildung von Trichomen und Wurzelhaaren beteiligt (Hülskamp, Miséra et al. 1994). AmPHANTASTICA reguliert das Wachstum und die Dorsoventralität von Blüten (Waites, Selvadurai et al. 1998). Eine Überexpression von AtMYB13 führt zu Veränderungen an der Blüte (Kirik, Kölle et al. 1998). Die Expression von AtMYB77 wird bei der Embryogenese erhöht (Kirik, Kolle et al. 1998). AtMYB2 reguliert den Alkoholdehydrogenase1 Genpromotor und reguliert Stoffwechselreaktionen auf Trockenstress (Hoeren, Dolferus et al. 1998). AtMYB30 ist am sensitiven Zelltod durch Pathogene beteiligt (Daniel, Lacomme et al. 1999).

1.7 Projekte zur Aufklärung der Genomsequenz von Pflanzen

Es gibt öffentliche und kommerzielle, sowie nationale und internationale Bestrebungen die Genomsequenz von geeigneten Pflanzenarten aufzuklären. Bei der Auswahl geeigneter Pflanzen spielen zum einen wissenschaftliche Handhabbarkeit und zum anderen ökonomische Relevanz eine Rolle.

1.7.1 Arabidopsis thaliana

Arabidopsis thaliana ist eine wichtige Modellpflanze für die Identifizierung von Genen und deren Funktion und hat eine Reihe von Eigenschaften, durch die sie für die Pflanzengenomforschung besonders geeignet ist. Dazu gehören eine kurze Generationszeit, eine große Anzahl von Nachkommen, geringe Größe und ein relativ kleines Genom (125 Mb) (The Arabidopsis Genome Initiative 2000). Im Jahr 1996 wurde die Arabidopsis Genome Initiative (AGI) gegründet, die sich die Sequenzierung des Genoms von *Arabidopsis thaliana* Ökotyp Columbia zum Ziel gesetzt hatte. Die Sequenzierstrategie beruhte auf der Verwendung von Klon-Bibliotheken, deren Klone lange Sequenzfragmente enthalten. Die Klone wurden durch „restriction fragment fingerprinting“ Analyse durch PCR und Hybridisierung von „sequence tagged sites“, sowie durch Hybridisierung und „southern blot“ Analysen, physikalisch kartiert. Die Ergebnisse wurden mit den genetischen Karten integriert und lieferten so den sog. „tilling path“ für das Zusammensetzen der Klone zu der kontinuierlichen Gesamtsequenz.

Die Sequenzgenauigkeit wurde durch Vergleiche mit bekannten Sequenzen verifiziert und mit 99,99 % bis zu 99,999 % bestimmt.

Im Dezember 2000 wurde in dem Journal Nature (AGI, 2000) die Fertigstellung der Genomsequenz publiziert (The Arabidopsis Genome Initiative 2000).

In der Veröffentlichung (The Arabidopsis Genome Initiative 2000) wurden auch erste das ganze Genom umfassende Analysen präsentiert. Unter anderem wurden für die Genomsequenz alle kodierenden Bereiche vorhergesagt und mit Hilfe von Sequenzvergleichen annotiert. Dabei wurde die Zahl aller putativ kodierenden Bereiche mit 25500 bestimmt. Davon konnte für 13000 eine Zuordnung zu Interpro-Domänen erfolgen und für etwa 70 % eine ähnliche Sequenz in anderen Organismen gefunden werden. Trotz dieser viel versprechenden Zahlen muss festgehalten werden, dass nur die Sequenzähnlichkeit oder die Existenz von ähnlichen Domänen

keine eindeutige Absicherung der Funktion ist. Sie kann aber als Ausgangspunkt für weitergehende Analysen genutzt werden.

1.7.2 Oryza sativa

Oryza sativa (Reis) ist ein Mitglied der Familie der Gräser, zu der auch Mais, Weizen, Roggen, Hirse und Zuckerrohr gehören. Eine Reihe von Eigenschaften machen *Oryza sativa* zu einem idealen Modellorganismus für die Monokotyledonen und die Familie der Gräser. Die Genomgröße von *Oryza sativa* ist vergleichsweise klein (430-460 MB (Sasaki and Burr 2000; Feng, Zhang et al. 2002; Goff, Ricke et al. 2002)) und es existieren eine Reihe von Werkzeugen für molekularbiologische Arbeiten, wie effiziente Transformationstechniken und genetische Karten. Zudem ist die ökonomische Bedeutung von *Oryza sativa* als Nutzpflanze ein weiterer Grund um diese Pflanze intensiv zu erforschen.

Im Jahr 1998 begann das International Rice Genome Sequencing Project (IRGSP) im Rahmen eines internationalen Projektes mit der Sequenzierung des *Oryza sativa* Genoms. Als Strategie für die Sequenzierung wurde das vergleichsweise langsame schrittweise Sequenzieren von minimal überlappenden Klonen mit großen Insertionen gewählt. Obschon diese Strategie sowohl teuer als auch langsam ist bietet sie die höchste Genauigkeit (99,99 %).

Kurz nach dem Start des Internationalen *Oryza sativa* Genomprojektes wurde an der Universität Washington von Monsanto ein eigenes *Oryza sativa* Genomprojekt initiiert, bei dem mit geringer Abdeckung ein Set von Klonen sequenziert werden sollte, das etwa 260 Mb abdeckt. Obwohl mit dieser Strategie 95 % der Gene in den BACs identifiziert werden konnten, reichte die Abdeckung nicht für das Zusammensetzen der Fragmente. Dennoch wurde das Monsanto Projekt aufgrund der anderen Sequenzierstrategie wesentlich früher abgeschlossen als das des IRGSP. Später hat Monsanto jedoch die Klone und Sequenzdaten dem IRGSP angeboten. Erst später sind zwei weitere Projekte zur Sequenzierung des *Oryza sativa* Genoms an den Instituten Beijing Genomics Institute (BGI) und Torrey Mesa Research Institute (TMRI / Syngenta) gestartet. Beide Projekte haben gemeinsam, dass sie den wesentlich schnelleren und günstigeren Weg der Shotgun-Sequenzierungsmethode von Klonen mit kleinen Insertionen gewählt haben. Syngenta produzierte auf diese Weise in relativ kurzer Zeit einen Entwurf mit sechsfacher Abdeckung für die Subspecies *japonica*.

BGI produzierte Sequenzdaten für zwei verschiedene Varietäten 93-11 und PA64, die die paternale und maternale Quellen für eine Hochleistungs-Hybridsorte bilden. Beide entstammen der *indica* Subspecies. Das BGI Projekt erreichte mit der Varietät 91-121 eine Abdeckung von 4x und mit der Varietät PA64 eine Abdeckung von 1.1x.
Da die Insertlänge bei beiden Projekten kleiner als 7 kb ist, können nur mit diesen Sequenzdaten keine kompletten kontinuierlichen Abschnitte berechnet werden. Deshalb wurde in beiden Projekten damit begonnen, die Klone gegen genetische und physikalische Karten zu kartieren.
Mit der Fertigstellung von zwei Chromosomen im IRGSP lassen sich auch erstmals die qualitativen Unterschiede zu den „draft" Veröffentlichungen von TIMRI und BGI abschätzen. Bei einem Vergleich der 493729 bp langen Sequenz von Chromosom 1 mit 127550 Sequenzen von *indica* aus dem BGI Projekt konnten 78 % der gesamten Region in den *indica* Sequenzen detektiert werden. Es gab aber 65 Lücken in den Contigs und 22 % der Basen aus *japonica* wurden nicht in den *indica* Sequenzen wieder gefunden. Zudem wurde bei einem Vergleich der Genvorhersagen festgestellt, dass nur die Hälfte der Gene mit vollständigen kodierenden Regionen bestimmt worden waren.
Zusammenfassend sind mit den Sequenzdaten aus IRGSP, Monsanto, BGI, und TIMRI die Sequenzdaten für das *Oryza sativa* Genom mehrfach, in unterschiedlicher Qualität und Abdeckung und mit unterschiedlichen Methoden erstellt worden. Zudem wurde mit verschiedenen Subspezies gearbeitet.
Um die vorhandenen Sequenzdaten optimal nutzbar zu machen, wäre eine Integration der verschiedenen Sequenzprojekte in eine Datenbasis wünschenswert, dies ist jedoch in näherer Zeit nicht zu erwarten.
Für die bioinformatische Arbeit mit den Genomsequenzen ist die Verlässlichkeit der Sequenz ein wichtiges Kriterium. Die Abweichung einzelner Basen kann Veränderungen in der Genvorhersage zur Folge haben. Zudem sind weitergehende Analysen aufgrund der abgeleiteten Proteinsequenz sehr zweifelhaft, wenn schon die zu Grunde liegende Nukleotidsequenz eine hohe Fehlerrate aufweist.
Die vorliegende Arbeit verwendet daher die genomischen Sequenzen aus dem IRGSP, die wenn auch noch nicht vollständig fertig gestellt, im Vergleich zu den „shotgun" Sequenzierprojekten eine höhere Verlässlichkeit bieten (Feng, Zhang et al. 2002; Sasaki, Matsumoto et al. 2002).

2 Zielsetzung der Arbeit

Zielsetzung der vorliegenden Arbeit ist die vergleichende Analyse der R2R3-MYB-Subfamilie in *Arabidopsis thaliana* und *Oryza sativa* mit bioinformatischen Methoden auf der Basis der veröffentlichten Genomsequenzen.

Zu diesem Zweck muss eine Bioinformatik-Infrastruktur geschaffen werden. Dazu gehören Datenbanken, in denen Sequenzinformationen und Annotationen erfasst werden und Programme, die das Editieren und Auslesen dieser Daten ermöglichen.

Des Weiteren müssen für die effiziente und umfassende Beschreibung großer Genfamilien die Such- und Annotationsvorgänge automatisiert werden, da die zu durchsuchenden Sequenzdaten kontinuierlicher Veränderung unterliegen und nur durch die Automatisierung eine vollständige Beschreibung auf der Basis aktueller Daten erfolgen kann. Die Vergleichende Analyse einer Subfamilie mit bioinformatischen Methoden umfasst die Analyse von übereinstimmenden und unterschiedlichen Merkmalen auf Sequenzebene, sowie die Bestimmung von Subgruppen.

3 Methoden

Die Charakterisierung von Genfamilien auf der Basis von Sequenzdaten kann mit Hilfe einer Reihe existierender Programme für Sequenzvergleiche, Motivsuche, Genstrukturvorhersage genutzt und kombiniert werden. Für die Suche nach ähnlichen Proteinsequenzen stehen Programme für paarweise Sequenzvergleiche wie BLAST (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997) und FASTA zur Verfügung (Pearson and Lipman 1988). Beide Programme bieten eine schnelle und einfache Möglichkeit um nach ähnlichen Sequenzen in Datenbanken zu suchen. Dabei werden von den Programmen ähnliche Sequenzen in der Datenbank gesucht und dann das optimale Alignment zwischen der Such- und der ähnlichen Sequenz berechnet. Dieses Alignment wird anschließend bewertet („score“). Weiterentwicklungen von Blast nutzen iterative Verfahren, um auch Sequenzen mit geringerer Ähnlichkeit in den Datenbanken zu finden (Altschul, Madden et al. 1997). BLAST und FASTA sind Programme, bei denen die Optimierung der Suchgeschwindigkeit bei weitgehender Erhaltung der Suchgenauigkeit im Vordergrund stand.

Ein sensitiveres Verfahren zur Suche in Sequenzdaten basiert auf dem sog. „hidden markov model“ (HMM) Algorithmus.

HMM sind Wahrscheinlichkeitsmodelle, die allgemein auf lineare Sequenzen von Informationen angewendet werden können, und für eine gleichsam sensitive wie spezifische Analyse von biologischen Sequenzen geeignet sind. Zunächst wurden HMM in der Spracherkennung eingesetzt (Rabiner 1989), einige Jahre später entstanden die ersten Anwendungen für die Analyse von DNA Sequenzen (Churchill 1989; Krogh 1994). Im Gegensatz zu den vorher beschriebenen Sequenzvergleichen basierenden Suchverfahren, besteht der Suchprozess mit HMM aus mehreren Schritten, die vom Benutzer nacheinander durchgeführt werden müssen. Der Vorteil HMM basierter Suchverfahren ist die höhere Sensitivität gegenüber einfachen Vergleichen zwischen zwei Sequenzen.

HMM beschreiben eine Wahrscheinlichkeitsverteilung über einen potentiell beliebig großen Sequenzraum. Die Funktionsweise eines HMM kann am besten als ein Sequenzgenerator verstanden werden. Das HMM gibt dabei die einzelnen Symbole der Sequenz mit jeder „state“ Änderung aufgrund des Wahrscheinlichkeitsmodells ab. Dabei gibt es zwei logische Ebenen, die Symbolebene und die „state“ Ebene. Der

Übergang von einem „state“ zum nächsten entspricht dem Fortschreiten zur nächsten Sequenzposition. Ein „state“ kann aber auch eine Insertion oder Deletion sein. Da die „state“ Ebene verborgen bleibt, wird sie englisch als „hidden“ bezeichnet. Ein HMM kann von einem Sequenzsatz ausgehen, der nicht aligned ist und ein Alignment erzeugen oder auf einem Alignment basieren. Im letzteren Fall ist die Abfolge der „states“ bereits festgelegt und es müssen nur die Übergangswahrscheinlichkeiten berechnet werden („profile HMM“). Die Bewertung und das Alignment der untersuchten Sequenz erfolgt mit standard Alignment-Methoden (Eddy 1998).

In der vorliegenden Arbeit sind, wenn nicht explizit anders beschrieben mit HMM „profile HMM“ gemeint. Es existieren eine Reihe von HMM basierten Programmen für die Analyse von biologischen Sequenzen. Das Programmpaket HMMER (Eddy SR nicht publiziert) basiert auf einem so genannten „Plan 7“ „state“-Modell, dass in der Lage ist alle wichtigen Zustandsübergänge in Alignments biologischer Sequenzen, wie zum Beispiel Insertionen, Deletionen, lokale Alignments zu beschreiben.

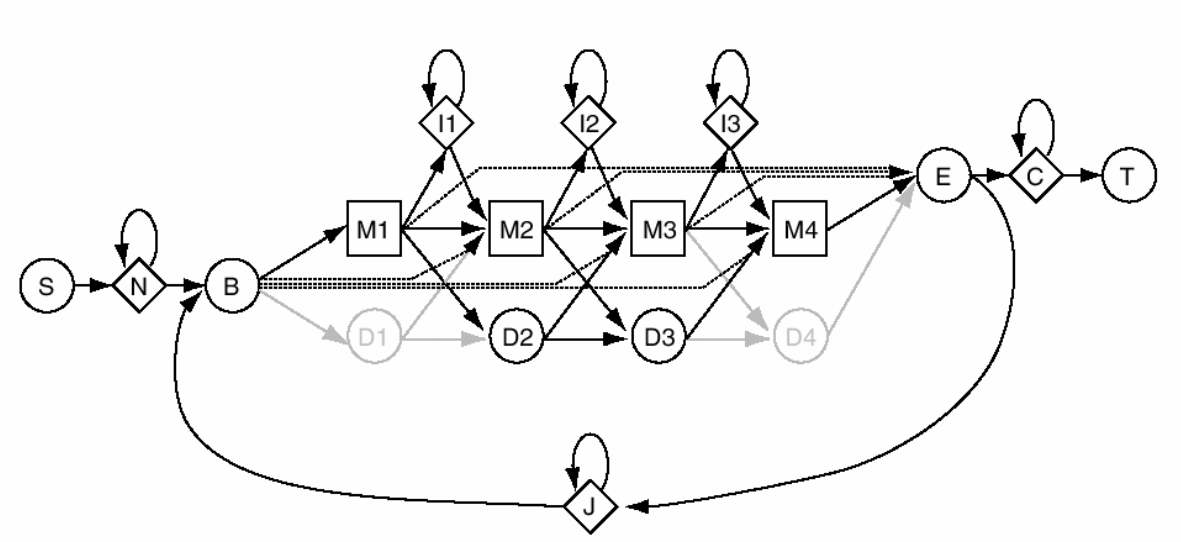


Abbildung 2: Plan 7 HMM. Quadrate entspr. "matches", Rauten entspr. Insertionen, Kreise entspr. Deletionen. S und T , B und E nicht emmitierende Start und Stopp „states“. N und C nicht „aligned“ Anfang und Ende „states“. J Lückenübergang. Aus (Eddy 1998)

Für kleinere Genfamilien werden die oben beschriebenen Programme standardmäßig zur Sequenzsuche eingesetzt. Dabei muss der Wissenschaftler nacheinander die Programme ausführen, die Eingabedaten vorbereiten, die Ergebnisse auswerten und ggf. kombinieren, um so die Eingabe für das nächste Programm zu erstellen. Bei der manuellen Bearbeitung kommt es an verschiedenen Stellen immer wieder zu Entscheidungen, die von dem Wissenschaftler aufgrund

seines Wissens über biologische Zusammenhänge getroffen werden. Für die Charakterisierung von Genfamilien ist die Objektivierung, Erfassung und Transparenz der Entscheidungen, die zu der Aufnahme eines Kandidaten oder seiner Ablehnung führen wichtig. Grundlage für die Systematisierung dieses Analyseprozesses bildet die Automatisierung durch ein zusätzliches Programm, welches die Arbeitsschritte nacheinander durchführt und die Ergebnisse speichert, auswertet und kombiniert.

4 Verwendete Programme:

Meme / Mast	(Bailey and Gribskov 1998)
Blast	(Altschul, Madden et al. 1997)
EMBOSS	(Rice, Longden et al. 2000)
HMMER	HMMER (Eddy SR unpublished)
Phred	(Ewing and Green 1998)
Crossmatch	(Gordon, Abajian et al. 1998)
PERL	http://www.perl.com
PHP	“open source” Skript Sprache http://www.php.net
MySQL	“open source” DBMS. http://www.mysql.com
Interproscan	(Apweiler, Attwood et al. 2001)
Stackpack	(Miller, Christoffels et al. 1999)
MaxdSQL	Relationales Datenbank Schema für Expressionsdaten http://bioinf.man.ac.uk/microarray/maxd/maxdSQL/
Phylip	Programmpaket für phylogenetische Analysen http://evolution.genetics.washington.edu/phylip.html
Clustalw	(Thompson, Higgins et al. 1994)
Jalview	“Multiple Alignment Editor” http://www.ebi.ac.uk/~michele/jalview/download.html
FgeneSH	(Salamov and Solovyev 2000)
GeneMarkR	(Besemer and Borodovsky 1999)

5 Ergebnisse

5.1 Überblick

In der vorliegenden Arbeit wurden Datenbanken und Programme für die vergleichende Charakterisierung von Genfamilien auf der Basis von Genomsequenzen, entwickelt. Im Folgenden wird das Zusammenwirken der Programme beschrieben (siehe Abbildung 1).

GenAgent ist ein Datenbank gestütztes System für die “high throughput” Analyse von DNA Sequenzen. *GenomeDB* ist eine Datenbank für die Integration von genomischen Sequenzen aus externen Quellen. *FamilyBuilder* ist ein Programm für die automatisierte Identifizierung und Klassifizierung von Genfamilienmitgliedern in genomischen Sequenzdaten und bezieht seine Daten aus *GenAgent* und *GenomeDB*. Die durch Einsatz von FamilyBuilder gewonnenen Annotationen werden in der *TF-Workbench* erfaßt und können über eine Oberfläche vom Benutzer eingesehen und editiert werden. Die in der TF-Workbench abgelegten Informationen können über TF-Cards in Form von Web-Seiten externen und internen Benutzern zugänglich gemacht werden.

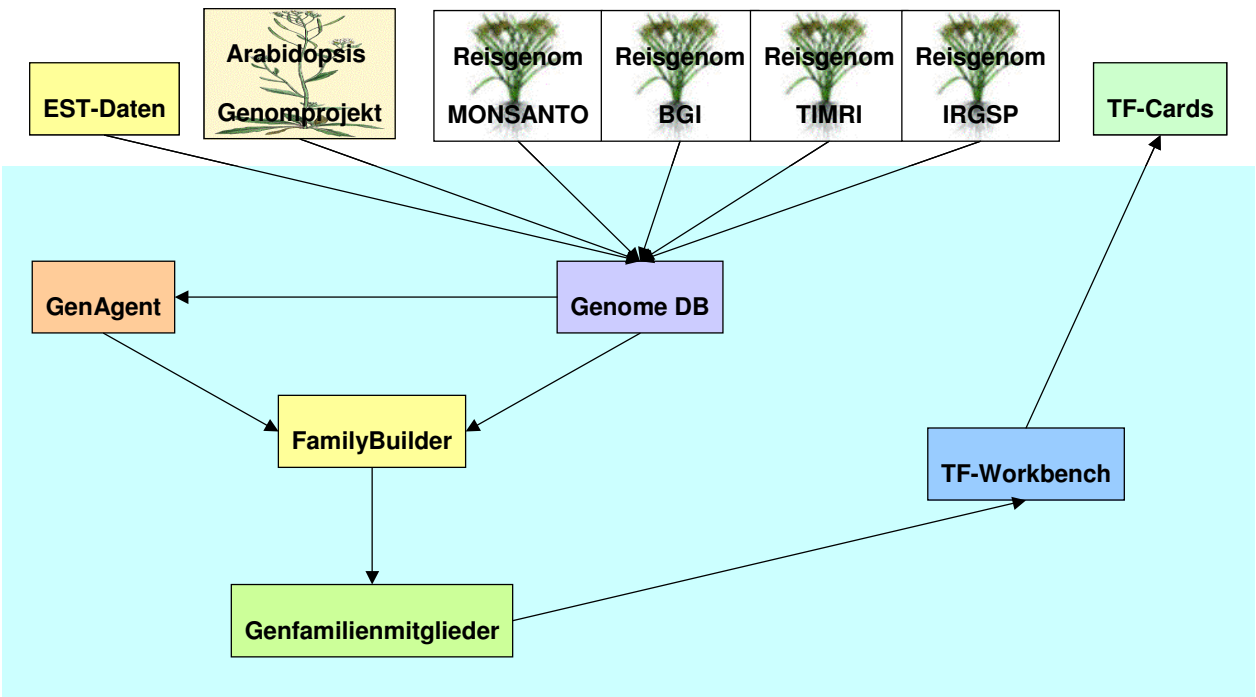


Abbildung 3: Übersicht über Verknüpfungen zwischen den entwickelten Datenbanken und Programmen.

5.2 Datenbanken

Im Rahmen dieser Arbeit wurden mehrere Datenbanken entwickelt, die zum einen biologische Daten organisieren und integrieren, und zum anderen die Datenbasis für Programmentwicklungen, wie Analysewerkzeuge und Benutzeroberflächen für strukturiertes Suchen bereitstellen. Da verschiedene der unten vorgestellten Programme nacheinander aufgerufen werden, um die Daten zu verarbeiten, werden die Datenbanken auch zur Speicherung von Zwischenergebnissen genutzt. Die umfangreiche Erfassung von Zwischenergebnissen ist besonders für die Dokumentation von Auswertungsschritten hilfreich. Diese Art der Nutzung von Datenbanken hat sich bewährt, da auf diese Weise Programm und Datenspeicherung getrennt voneinander entwickelt werden können. Funktionalitäten können, je nach Anforderung, schnell in Form von kleinen Programmen bereitgestellt werden. Zum Beispiel kann ein Programm auf die Sequenzdaten zugreifen, um die Genstruktur vorherzusagen, während ein anderes Programm die erzeugten Daten verwendet, um sie zu visualisieren.

5.2.1 *GenomeDB*: Integration von externen Sequenzdaten in einer relationalen Datenbank

Die Datenbank *GENOMEDB* wurde konzipiert, um für den *GenAgent*, die *TF-Workbench* und den *FamilyBuilder* externe Sequenzdaten zur Vefügung zu stellen. Die Sequenzdaten werden von den Primärdatenbanken, wie NCBI (Wheeler, Church et al. 2003) und von speziellen Datenbanken, wie der MatDB (MIPS München) (Schoof, Zaccaria et al. 2002; Frishman, Mokrejs et al. 2003) für *Arabidopsis thaliana* bzw. IRGSP für *Oryza sativa* (Sakata, Nagamura et al. 2002) in Form von Textdateien angeboten. In diesen Textdateien werden immer wieder andere Formate für die Annotation der Sequenzen verwendet. Zudem enthalten die für *Oryza sativa* angebotenen Sequenzdaten doppelte Einträge. Um die Benennung von Einträgen zu vereinheitlichen und einen nicht-redundanten Sequenzdatensatz bereitzustellen wurden die in dieser Arbeit genutzte Sequenzdaten, wie die Genomsequenzen aus *Arabidopsis thaliana* und *Oryza sativa*, sowie die EST-Daten in einer Datenbank zusammengefasst. In Abbildung 4 ist die vereinfachte Datenbankstruktur grafisch dargestellt. Die *GENOMEDB* wird nicht über eine Benutzeroberfläche angesprochen, sondern ausschließlich von Programmen im Rahmen von Sequenzvergleichen und weiteren Analysen genutzt.

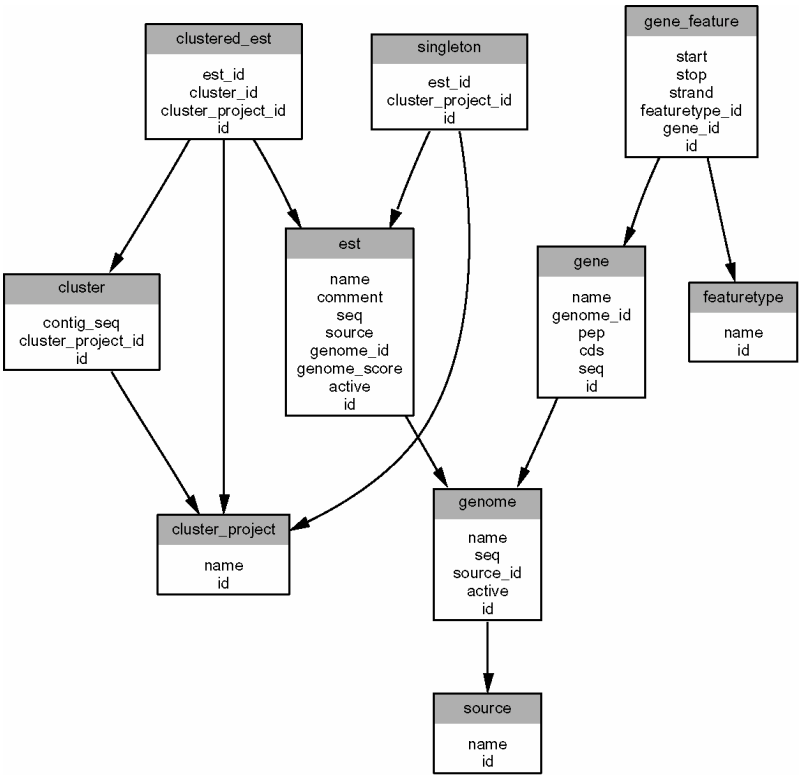


Abbildung 4: Vereinfachte Struktur der GenomeDB Datenbank

5.2.2 GenAgent[1]

Das Programm *GenAgent* ist ein Datenbank gestütztes System für die “high throughput” Analyse von DNA Sequenzen. Der Name *GenAgent* leitet sich aus der Terminologie der Informatik ab, bei der Agenten autonom agierende Programmeinheiten sind, die selbstständig tätig werden. Das Programm besteht aus mehreren `PERL` Skripten zur Verarbeitung der Sequenzen, einer Datenbank als zentralem System zur Speicherung von Sequenzen und der Annotation und `PHP` Skripten, die über dynamisch generierte Webseiten eine Benutzeroberfläche bereitstellen und die Informationen hierfür aus der Datenbank beziehen.

5.2.2.1 Eingangsdatenanalyse

Der erste Schritt im Rahmen des *GENAGENT* Programms ist die Prozessierung von Sequenzdaten. Die Sequenzen werden direkt aus der Sequenzierabteilung ADIS des MPI in Form von Elektroferrogrammen („trace files“) bezogen. Neben der eigentlichen Sequenz werden zusätzliche Informationen zum Sequenziervorgang erfasst. Dazu gehören, Primer, Spur bzw. Kapillarnummer, Zeitpunkt der Erstellung mit Datum und Uhrzeit, die bearbeitende Person, Projektkennziffern, Plattennummern und Plattenkoordinaten. Zunächst wird das ABI-Elektroferrogramm vom Sequenziergerät in einen bestimmten Verzeichnisbereich im NFS hoch geladen. `PERL` Skripte übernehmen dabei die Verteilung anhand von Projektkennziffern und Plattennummern. Anschließend beginnt der Verarbeitungsprozess, der vollautomatisch abläuft.

Anschließend wird in den ABI-Dateien mit dem Programm `PHRED` nochmals die Basenabfolge bestimmt („basecalling“). Zwar sind in der ABI-Datei schon Sequenzauswertungen enthalten, `PHRED` bietet jedoch die Möglichkeit jeder Basenzuordnung außerdem einen Wahrscheinlichkeitswert über die Richtigkeit der Zuordnung zuzuweisen (Ewing and Green 1998). Diese Wahrscheinlichkeitswerte ermöglichen anschließend das Trimmen der Sequenz aufgrund der Qualität. Dabei wird in einem über die Sequenz laufenden Fenster von definierter Größe die

[1] Die für diese Arbeit entwickelten Programme und Datenbanken, wurden von mir mit kurzen Eigenamen benannt. Um für den Leser den Unterschied zwischen den Eigenentwicklungen und den verwendeten externen Programmen darzustellen, sind die Namen der *Eigenentwicklungen* kursiv und die `externen Programme` in Courier gesetzt.

durchschnittliche Qualität bestimmt. Wenn die Qualität über einen definierten Wert steigt, beginnt die qualitative Sequenz, wenn der Wert unter eine bestimmte Schwelle sinkt, wird die Sequenz an dieser Stelle abgeschnitten. Häufig haben Sequenzdaten am Anfang und am Ende Bereiche von schlechter Qualität und diese Bereich können so automatisch detektiert und entfernt werden.
Neben den internen Sequenzdaten können auch Sequenzdaten aus Datenbanken im Internet, wie zum Beispiel NCBI in den GenAgent geladen werden. Hierbei entfallen die vorangehend beschriebenen Schritte.
Im nächsten Schritt der Sequenzverarbeitung werden Vektor- und Linkersequenzen entfernt. Hierzu werden die Sequenzen mit dem Programm `Crossmatch` gegen eine Vektor- und Linker-Sequenzdatenbank verglichen. Die unprozessierten und die prozessierten Sequenzen werden zusammen mit den zusätzlichen Informationen in eine relationale Datenbank geladen, die unter dem DBMS `MySQL` läuft. Hier werden die Sequenzen einerseits als Referenz gespeichert, andererseits dient die Datenbank als Informationsgrundlage für die Benutzeroberfläche, die ein bequemes Abfragen von Informationen ermöglicht.

5.2.2.2 Agenten für Sequenzvergleiche

Die Sequenzen können, nachdem sie in die `MySQL`-Datenbank geladen wurden, automatisch gegen verschiedene Sequenzdatenbanken verglichen werden. Jeder Agent, der sich aus der zu durchsuchenden Datenbank, dem Suchprogramm und den Suchparametern zusammensetzt, bildet eine autonome Einheit, die beliebig viele Sequenzen aus der Datenbank prozessieren kann (s.o. Agent). Es lassen sich beliebig viele solcher Agenten definieren. Die Suchvorgänge können automatisch in regelmäßigen Abständen gestartet werden. Wenn Sequenzen erstmalig in die Datenbank aufgenommen werden, startet automatisch ein bereits eingestellter Agent die Sequenzanalyse.
Die Ergebnisse jeder Suche werden in der Datenbank abgelegt. Dabei wird die komplette Antwort bestehend aus den einzelnen Treffern für die Suchanfrage gespeichert. Nach der Suche wird das Suchergebnis gegen das Vorherige, in der Datenbank abgelegte Ergebnis verglichen. Wenn die Suche ein verändertes Ergebnis aufweist, wird diese Änderung von dem Programm erfasst. Auf diese Weise können Benutzer über aktuelle Veränderungen in der Annotation ihrer Sequenzen benachrichtigt werden.

In der Praxis hat sich der Vergleich mit `blastx` gegen die “non-redundant protein database” (nr) bei NCBI und `blastx` gegen Swissprot bewährt. Die nr-Datenbank bietet die Möglichkeit, auf alle Sequenzinformationen zuzugreifen, die derzeit weltweit öffentlich verfügbar sind. Der Nachteil dieser Datenbank ist, dass die nr-Datenbank auch automatisch translatierte Sequenzen enthält, die Annotationen in der Regel automatisch erfolgen und die Datenbank insgesamt redundant ist. Da die Swissprot Datenbank manuell gepflegt wird, sind die Annotationen von besserer Qualität. Die neuesten Ergebnisse werden jedoch erst in den Primärdatenbanken, wie zum Beispiel NCBI oder EBI veröffentlicht.
Neben der Suche mit `BLAST` kann auch das Programm `INTERPROSCAN` eingesetzt werden. `INTERPROSCAN` kombiniert mehrere Mustererkennungs- und HMM- basierte Motivsuchprogramme. Die Ergebnisse werden genauso wie bei den `BLAST` Suchen behandelt und in der Datenbank gespeichert. Die durch `INTERPROSCAN` identifizierten Domänen und Motive werden sowohl durch ihre Verknüpfung zu den Motivdatenbanken beschrieben, als auch durch Verknüpfungen zu Ontologien. Ontologien sind Begriffshierarchien mit kontrolliertem, das heißt eindeutigem Vokabular. Damit können Prozesse oder Objekte in der Biologie in definierter Detailliertheit beschrieben werden. Die Verknüpfung der Motive mit Ontologien ist eine Technologie, die für die Erstellung von “functional categories” bei der Annotation von Arrayexperimenten genutzt werden kann.

5.2.2.3 Datenbankstruktur

Abbildung 5 zeigt die Datenbankstruktur der *GENAGENT* Datenbank. Um die Darstellung übersichtlich zu halten, werden nur die Tabellen gezeigt, die zur Kernfunktionalität beitragen.
Der logische Anknüpfungspunkt für die Speicherung von Sequenzdaten ist der Klon.

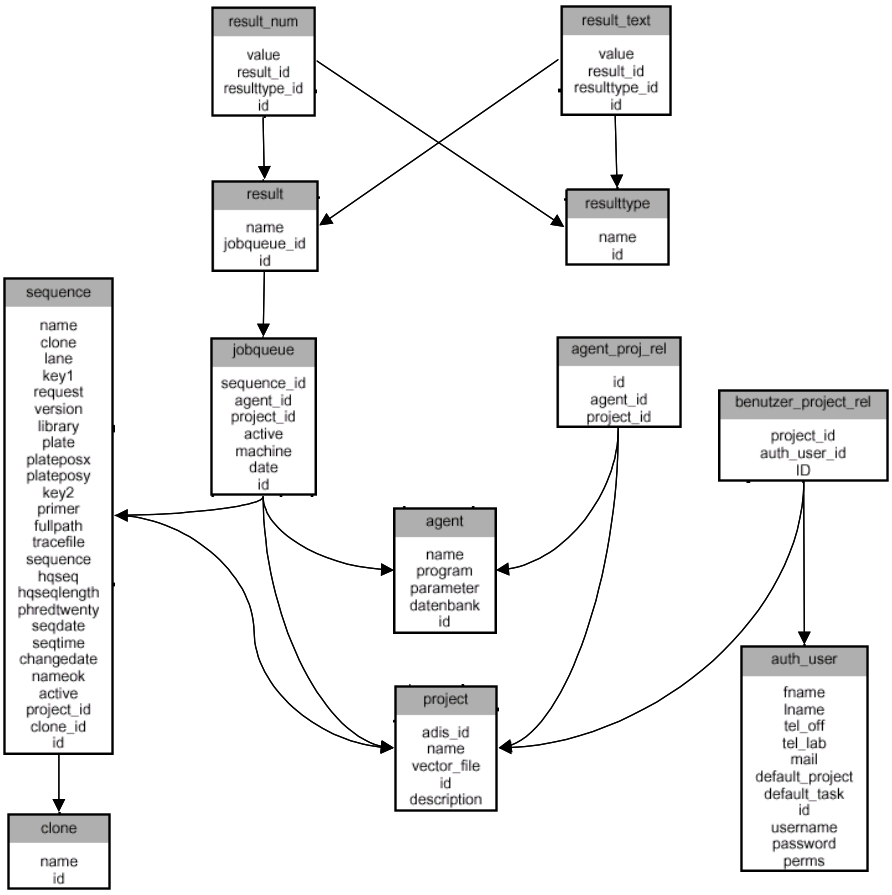


Abbildung 5: Vereinfachte Struktur der *GENAGENT* Datenbank.

Jeder Klon hat einen eindeutigen Namen und innerhalb der DB einen eindeutigen Identifizierer („unique identifier“). Zu einem Klon können mehrere Sequenzen in der Datenbank vorhanden sein. Sequenzen sind Bestandteile von Projekten und Benutzer können an mehreren Projekten beteiligt sein. Beliebig viele Projekte können mit beliebig vielen Agenten verknüpft werden. Jede durch einen Agenten zu bearbeitende Sequenz wird in einer Job-Warteschlange erfasst. Die Job-Warteschlange wird anhand der Agentenspezifikationen abgearbeitet. Die Abarbeitung kann auf mehreren Rechnern parallel erfolgen. Dabei fragt jeder Rechner den jeweils nächsten zu aktivierenden Job aus der Jobqueue ab. Der Job wird auf „in Bearbeitung“ gesetzt. Wenn der Job fertig gerechnet ist, wird das Ergebnis zurück in die DB geschrieben und der Job auf “erledigt“ gesetzt. Wenn die Bearbeitung unvollendet abbricht, bleibt der Eintrag auf „in Bearbeitung“ stehen. Nach einer spezifizierten Latenzzeit wird der Job auf „unbearbeitet“ zurückgesetzt. Die Tabellen „result“, „result_type“, „result_text“ und „result_num“ sind generisch aufgebaut, d.h. sie können die Ergebnisse von beliebigen Suchanfragen aufnehmen. Die Elemente eines Treffers können entweder numerisch oder Textdaten sein.

Zusätzlich wird eine Bezeichner definiert, der die Semantik des Wertes liefert. Bezeichner können z.B. „score“, „e-value“, „accession nr“ oder ähnliches sein. Aufgrund der generischen Umsetzung der Ergebnisspeicherung kann der *GENAGENT* auf beliebige Sequenzanalysen erweitert werden.

5.2.2.4 Sequenzgruppierung (Clustering)

Die Sequenzdaten im GenAgent können mit Hilfe des `STACKPACK`-Programmpakets (Miller, Christoffels et al. 1999) zu größeren kontinuierlichen Sequenzabschnitten zusammengesetzt werden. Dabei werden von dem `STACKPACK` Programm in einem mehrschrittigen Verfahren, aufgrund der Ähnlichkeiten der Sequenzen zueinander, Gruppen von Sequenzen gebildet (Clustering). Da dieses Verfahren sehr rechenintensiv ist - mit der Zahl der Sequenzen steigt die Anzahl der nötigen Vergleiche exponentiell - sind moderne Clustering-Verfahren in Hinblick auf die Geschwindigkeit der Algorithmen optimiert. `STACKPACK` benutzt dazu im ersten Schritt den besonders leistungsfähigen `D2-Algorithmus` (Burke, Davison et al. 1999), um zunächst anhand von kurzen Sequenzübereinstimmungen möglichst schnell zu einer groben Gruppierung zu kommen. In den späteren Schritten werden die Gruppen genauer analysiert und gegebenenfalls verkleinert oder vergrößert. Im idealen Fall erhält man gruppierte Sequenzen, die die kompletten kodierenden Sequenzabschnitte eines Gens enthalten. In den meisten Fällen lassen sich zumindest größere Teile von Genen rekonstruieren. Ein weiterer Vorteil des Clustering ist, dass die Redundanz in den EST-Datensätzen beseitigt wird. Bei der Analyse von nicht normalisierten Banken bietet Clustering die Möglichkeit, die Häufigkeiten von Transkripten in bestimmten Geweben zu ermitteln. Technisch wird die Anbindung von `StackPack` an den *GENAGENT* über das DBMS gewährleistet. Da `Stackpack` auch mit dem `MySQL` DBMS arbeitet, können die in beiden Systemen eingetragenen Sequenznamen als eindeutige Referenzen auf die Sequenzdaten herangezogen werden. Wenn zu einem Klon mehrere Sequenzen vorliegen, kann die Zusammengehörigkeit von Sequenzen und Clustern zu Klonen durch den *GenAgenten* ausgewertet werden.

5.2.2.5 Benutzeroberfläche

Während im Hintergrund verschiedene Perl Skripte an der Verarbeitung der Sequenzen beteiligt sind, arbeitet der Benutzer mit dem *GENAGENT* über eine Web-Oberfläche.

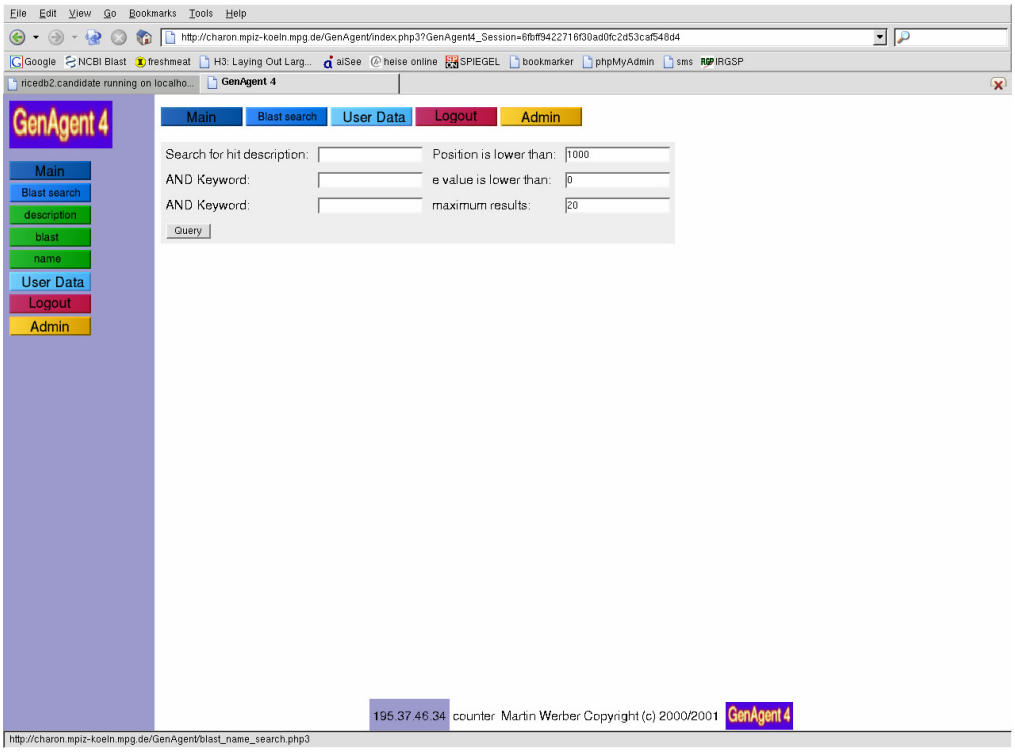


Abbildung 6: Fenster „description“. Suche nach Beschreibungen in den Suchergebnissen

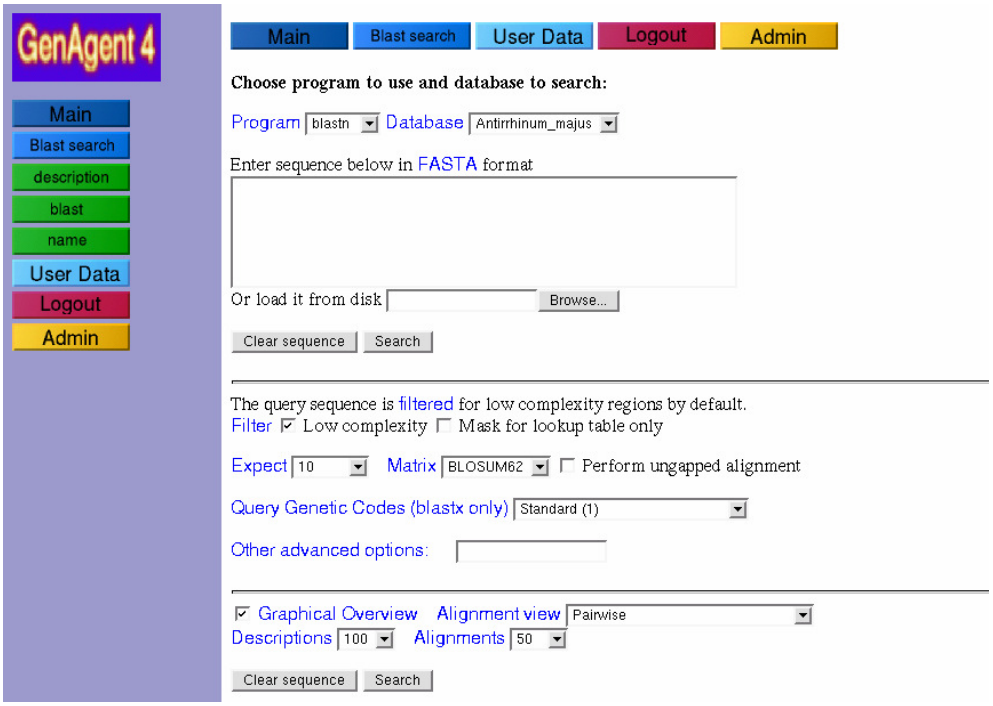


Abbildung 7: Fenster „blast“, Suche mit BLAST in den Sequenzen, die im *GenAgent* gespeichert sind.

Die Benutzeroberfläche bietet verschiedene Möglichkeiten die Analyseergebnisse zu durchsuchen. Bei der Suche nach Stichworten unter dem Menüpunkt “description” (siehe Abbildung 6) können die Annotationen aller Sequenzen durchsucht werden. Dabei kann das Suchergebnis über mehrere, durch logisches „UND“ verknüpfte Schlüsselwörter eingegrenzt werden. Zusätzlich kann die Ergebnismenge durch Beschränkung auf eine bestimmte Ranglistenposition und einen maximalen „*e-value“* definiert werden. Eine weitere Möglichkeit ist die Suche mit BLAST in den Sequenzdaten, die im *GenAgent* gespeichert sind (siehe Abbildung 7).

Die im *GenAgent* gespeicherten Sequenzannotationen und funktionellen Klassifikationen (“functional categories”) können dazu genutzt werden, um Experimente mit Expressionsarrays zu annotieren. Die Verknüpfung zwischen den Expressionsdaten und den Sequenzdaten erfolgt über den Klonnamen als eindeutige Referenz. Um die Expressionsdaten zu erfassen, wurde das Datenbankschema der Arraydatenbank `MaxdSQL` der Bioinformatik Gruppe der Universität Manchester genutzt. Diese Datenbank implementiert das `ArrayExpress` Datenbankschema vom European Bioinformatics Institute (EBI) in einer standardisierten SQL-Form (ANSI SQL 92 Syntax). Aufgrund der standardisierten Implementation konnte es einfach in das bestehende DBMS `MySQL` integriert werden.

Mit nur geringfügigen Anpassungen können die Datenbanken GenAgent und `MaxdSQL` über den Klon als eindeutigen Identifizierer verknüpft werden. Die Kombination von EST-Annotation und Arraydatenauswertung wurde in einem Projekt, bei dem 3000 ESTs aus *Beta vulgaris* untersucht wurden, erfolgreich eingesetzt. Details dieser Anwendung des GenAgent werden in der Diskussion beschrieben.

Insgesamt wird der GenAgent von sechs verschiedenen Projekten, mit derzeit ca. 50.000 EST -Sequenzen intensiv genutzt.

5.2.3 TF-Workbench

Das Programm *TF-WORKBENCH* wurde als eine integrierte Annotationsumgebung für die Analyse und Verwaltung von Daten zu Genfamilienmitgliedern entwickelt. Es besteht aus einer Datenbank, PERL Skripten, die im Hintergrund Analyse- und Verwaltungsaufgaben übernehmen und einer Benutzeroberfläche, die von PHP Skripten dynamisch generiert wird.

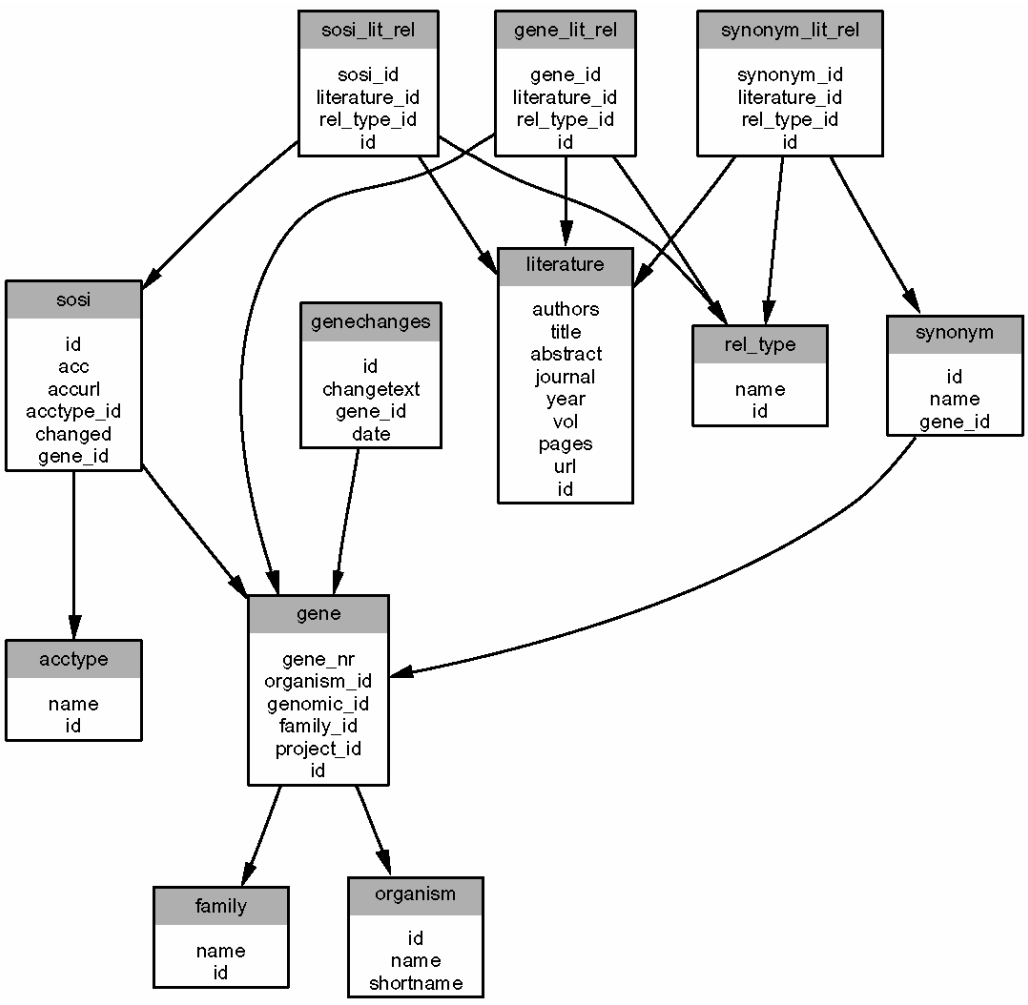


Abbildung 8: Vereinfachte Struktur der TF-Workbench Datenbank. Unterteil Gene und Text-Annotationen

5.2.3.1 Datenbankstruktur

Zentraler Anknüpfungspunkt bei dem Entwurf der TF-Workbench Datenbank ist das Genfamilienmitglied. In den folgenden Abbildungen wurde die Struktur der Datenbank aus Gründen der Übersichtlichkeit in zwei Darstellungen aufgeteilt (siehe Abbildung 8). Die Tabelle “gene” wurde jeweils übernommen, ist aber in der

Datenbank nur einmal vorhanden. Jedes Gen wird durch die Vergabe einer Nummer in Kombination mit einem (Sub-) Familiennamen eindeutig beschrieben. Der angezeigte Namen wird dann aus diesen Angaben automatisch generiert. Wenn zum Beispiel das Feld “gene_nr” die Zahl 1 enthält und das Feld “family_id” auf den Eintrag “MYB” in der Tabelle “family” verweist und “organism_id” auf Arabidopsis, dann generiert das System daraus “AtMYB001”. Die Nummern werden auf drei Stellen mit Nullen aufgefüllt.

Jedes Gen verweist auf den Literatureintrag, der das Gen erstmalig beschreibt. Andererseits können Gene immer wieder in der Literatur behandelt werden. Die Tabelle “gene_lit_rel” erfaßt diese Relationen. In dem Feld “rel_type_id” wird zudem aufgeführt, welche Semantik diese Relation hat, d.h. ob es sich um eine Primärbeschreibung oder um eine zusätzliche Beschreibung handelt. Zu Genen können in der Tabelle “alias” synonyme Bezeichner erfaßt werden. Es ist leider üblich, das selbe Gen in verschiedenen Arbeitsgebieten und Veröffentlichungen unterschiedlich zu benennen.

Die Tabelle “sosi” steht für “sources of sequence information”. Hier können Quellen für Sequenzen unterschiedlicher Arten erfaßt werden.

Dies können “ESTs”, “full length cDNA” Klone oder andere Sequenzen sein. Der Typ der Sequenz wird in “acctype” erfaßt. Sofern vorhanden können auch die “sosi” Einträge durch eine Literaturstelle erfaßt werden.

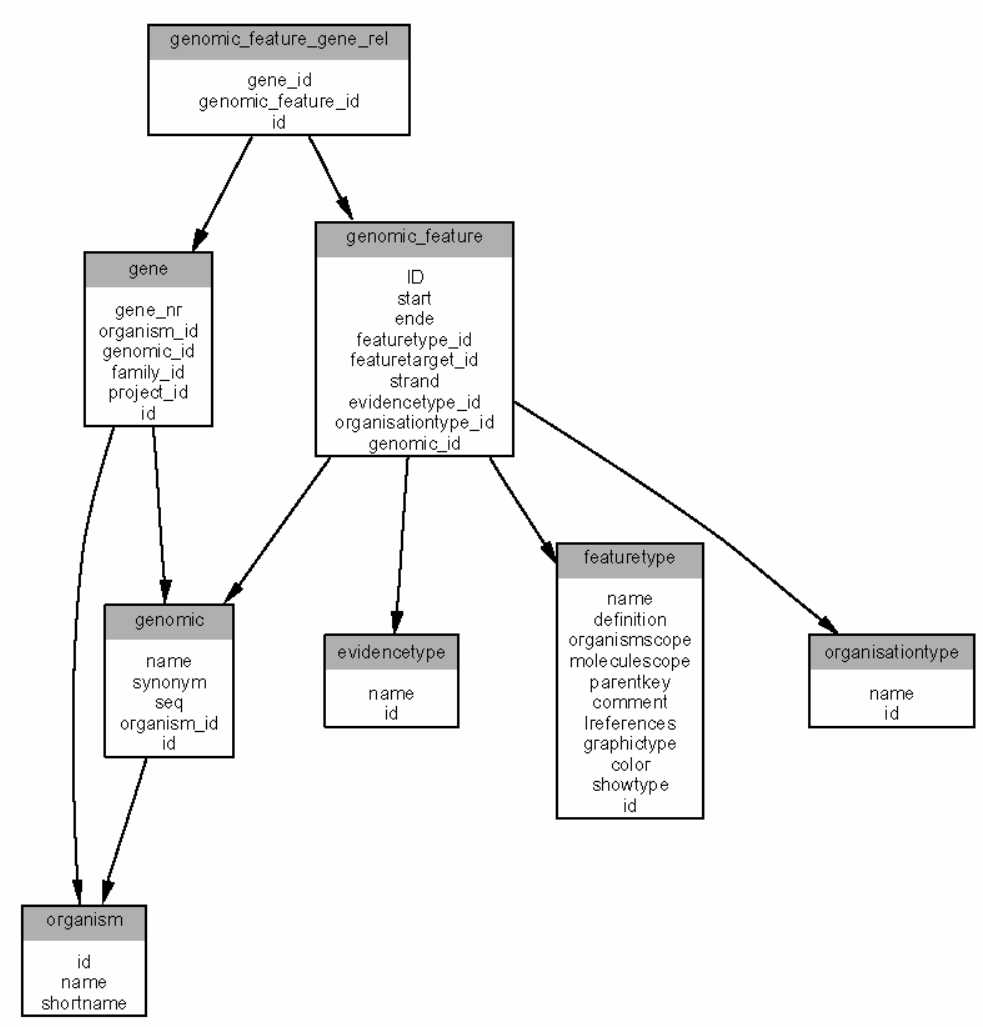


Abbildung 9: Vereinfachte Struktur der TF-Workbench Datenbank. Unterteil Gen-Annotationen

Jedes Gen wird durch eine Referenz auf das Genom eindeutig beschrieben. Zu jedem Gen kann es mehrere Einträge in der Tabelle “genomic feature” geben, die jeweils genetische Elemente auf der Genomsequenz beschreiben. Beispiele sind Exons, Introns und nicht translatierte Bereiche. Die genomische Sequenz, auf die sich die Tabelle “genomic feature” bezieht, bildet die Referenz für alle Sequenzinstanzen, die eine Gen einnehmen kann. Die Sequenzinstanzen umfassen kodierende Sequenz, offenes Leseraster, Promotorbereiche und translatierte Proteinsequenz. Die Sequenzinstanzen werden nicht statisch in den Tabellen abgelegt sondern dynamisch bei der Ausgabe ermittelt. Kodierende Sequenzen werden über die Exonkoordinaten aus der genomischen Sequenz herausgezogen, und Proteinsequenzen werden durch Translation aus der kodierenden Sequenz gewonnen.

5.2.3.2 Benutzeroberfläche

Die Benutzeroberfläche der TF-Workbench besteht aus HTML Seiten, die dynamisch von PHP Skripten generiert werden. Die einzelnen Seiten können von einer Menüzeile ausgewählt werden.

5.2.3.2.1 Genstrukturannotation

Eine zentrale Stellung in der Benutzeroberfläche nimmt die Seite ein, in der die Genstruktur der Genfamilienmitglieder annotiert wird (siehe Abbildung 10).
Um diesen Vorgang zu erleichtern, werden eine Reihe von Hilfestellungen gegeben. Nach dem Namen des Gens werden in einer Tabelle die bereits eingetragenen Sequenzmerkmale “features” mit ihren Koordinaten aufgelistet. „features“ sind Sequenzsegmente, denen eine Eigenschaft zugeordnet wird. Unterhalb der Tabelle findet sich eine grafische Darstellung der Genstruktur. Hier können auch weitere Sequenzelemente, wie zum Beispiel offene Leserahmen dargestellt werden.
Der Ausschnitt der Darstellung und damit die Vergrößerung der Darstellung können verändert werden. Unterhalb der grafischen Darstellung ist ein frei wählbarer Abschnitt der genomischen Sequenz dargestellt. Die Exons sind blau hervorgehoben.
Während der Eingabe von Elementen wird laufend die kodierende Sequenz und die Proteinsequenz berechnet. Auf diese Weise können biologisch unzulässige Intron-Exon Grenzen sofort bemerkt und korrigiert werden.
Auch das Eintragen von Sequenzbereichen mit ihren Koordinaten wurde vereinfacht. Sequenzabschnitte in der genomischen Sequenz können mit der Maus markiert werden. Die Koordinaten des markierten Bereichs werden im Hintergrund durch Javascript berechnet und in den Auswahlfenstern im unteren Bereich der Seite in den Feldern “Start / Stop” angegeben.
Die Verknüpfung “sim4 comparison” öffnet ein neues Fenster, in welches eine kodierende DNA Sequenz eingegeben werden kann. Diese wird dann durch `sim4` (Florea, Hartzell et al. 1998) gegen die genomische Sequenz verglichen. Die Ausgabe liefert die Koordinaten übereinstimmender Bereiche und den Grad der Übereinstimmung in Prozent.
Der Link "genscan prediction" öffnet ein Fenster, in dem, für den im Hauptfenster dargestellten Bereich, eine Genstrukturvorhersage angezeigt wird. Der link “MIPS”

führt zu dem entsprechenden Eintrag in der MatDB des „Munich Information Center for Protein Sequences“ (MIPS).

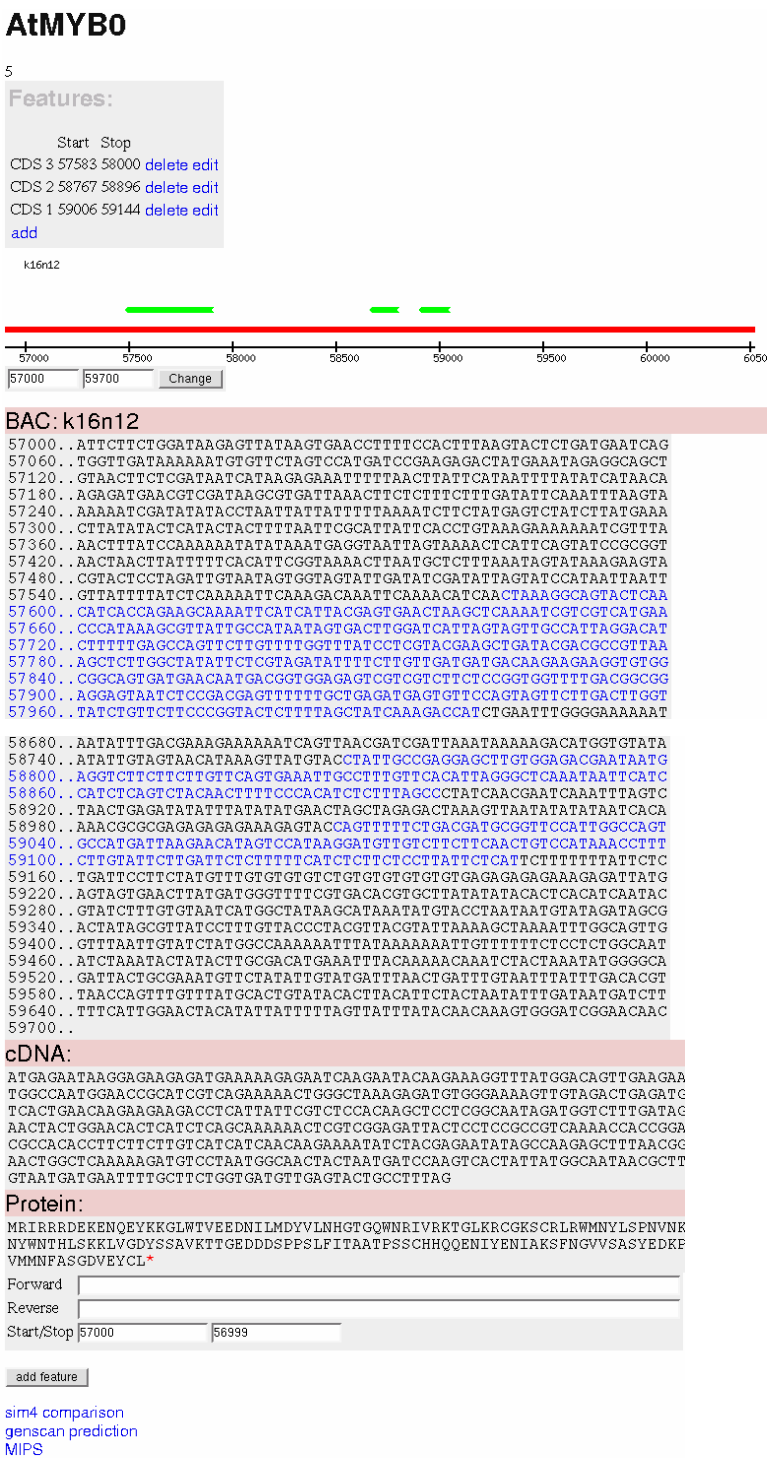


Abbildung 10: Benutzeroberfläche für die Annotation der Genstruktur

5.2.3.2.2 Textannotation

In dem Fenster “Edit” (siehe Abbildung 11) werden zusätzliche Textinformationen zu jedem Gen erfaßt.

Dazu gehören eindeutige laufende Nummer innerhalb der (Sub-) Familie, synonyme Namen, der AGI-Bezeichner (Arabidopsis Genome Initiative identifier), der Name der genomischen Referenz (Klon) und die primäre Literaturreferenz. Zudem können mehrere sog. “sources of sequences information (SOSI)” eingetragen werden. Neben der primären Literaturstelle können auch weitere Literaturstellen erfaßt werden, in denen Informationen zu diesem Gen enthalten sind.

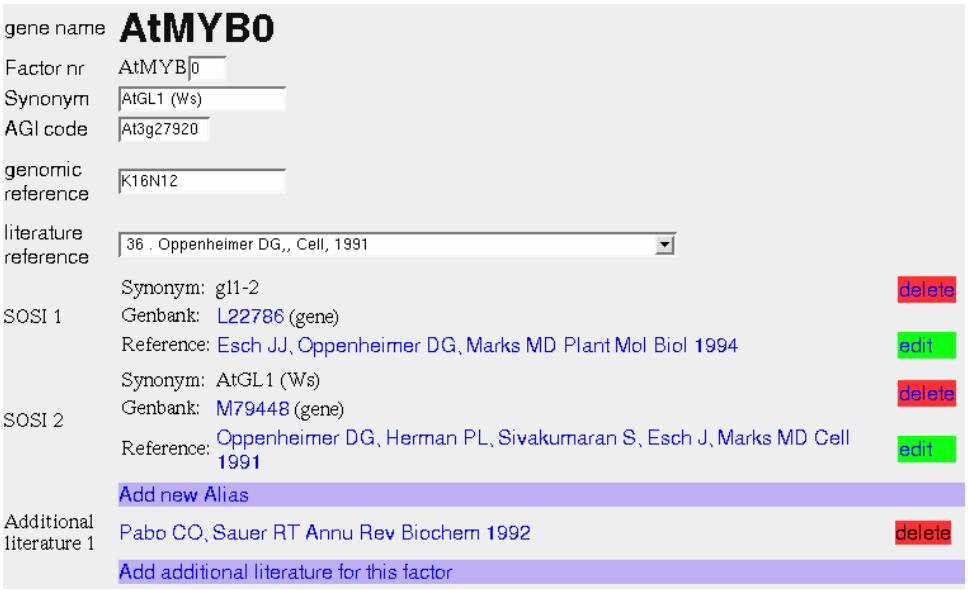


Abbildung 11: Textannotation der Genfamilienmitglieder

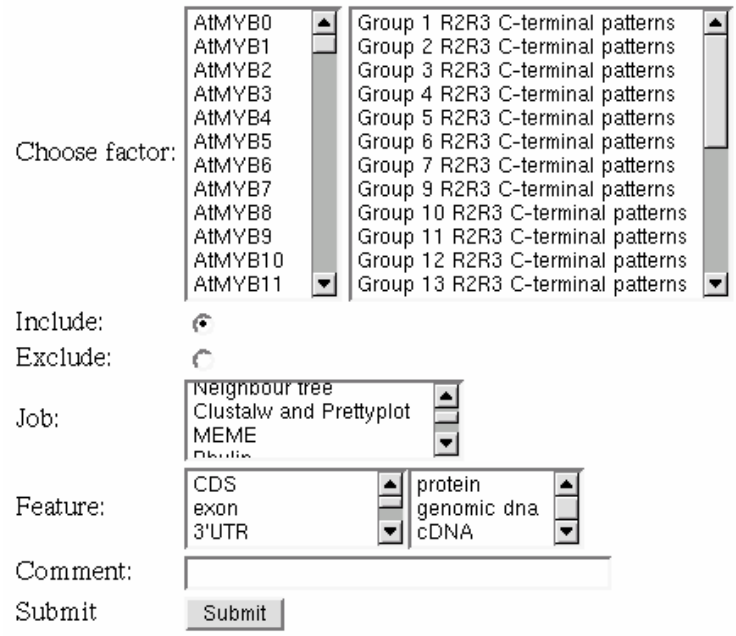


Abbildung 12: „Analysis“: Auswahl von Genen, Gengruppen und Programmen für die Sequenzanalyse.

5.2.3.2.3 Sequenzanalyse

In dem Fenster “Analysis” (siehe Abbildung 12) können mit den Gensequenzen grundlegende Analysen durchgeführt werden. Die Gene werden einzeln oder in Form von Gruppen aus Listen ausgewählt, dabei kann die Auswahl invertiert werden. Zur Analyse stehen Alignment-Programme (`ClustalW`), Motivsuche-Programme (`MEME`) und Programme zur Berechnung von Sequenzdistanzen und phylogenetischen Bäumen aus dem `PHYLIP` Programmpaket zur Vefügung. Die entsprechenden Programme werden im Hintergrund von Perl-Skripten aufgerufen und die Ergebnisse werden - grafisch aufbereitet und in HTML-Seiten eingebunden - dargestellt. Eine besondere Vereinfachung für die Durchführung von Analysen stellt die Möglichkeit dar, die Sequenzinstanz des Gens, kodierende DNA oder Proteinsequenz und einen annotierten Bereich, zum Beispiel die MYB-Domäne, einfach aus der Liste auszuwählen und eine Analyse für diese Auswahl durchzuführen.

5.2.3.2.4 TF-Cards

Die in die *TF-Workbench* eingetragenen Informationen können über die *TF-Cards* externen und internen Benutzern zugänglich gemacht werden (siehe Abbildung 13). Dabei handelt es sich um dynamisch generierte HTML-Seiten, die auf einer Seite alle wichtigen Informationen zu einem Gen zusammenstellen. Dazu gehören Namen, Literaturreferenzen und Sequenzinformationen. Zusätzlich werden in einer Grafik die sechs ähnlichsten Sequenzen aus der Genfamilie dargestellt. Dies kann innerhalb eines Organismus oder über Organismengrenzen hinweg erfolgen.

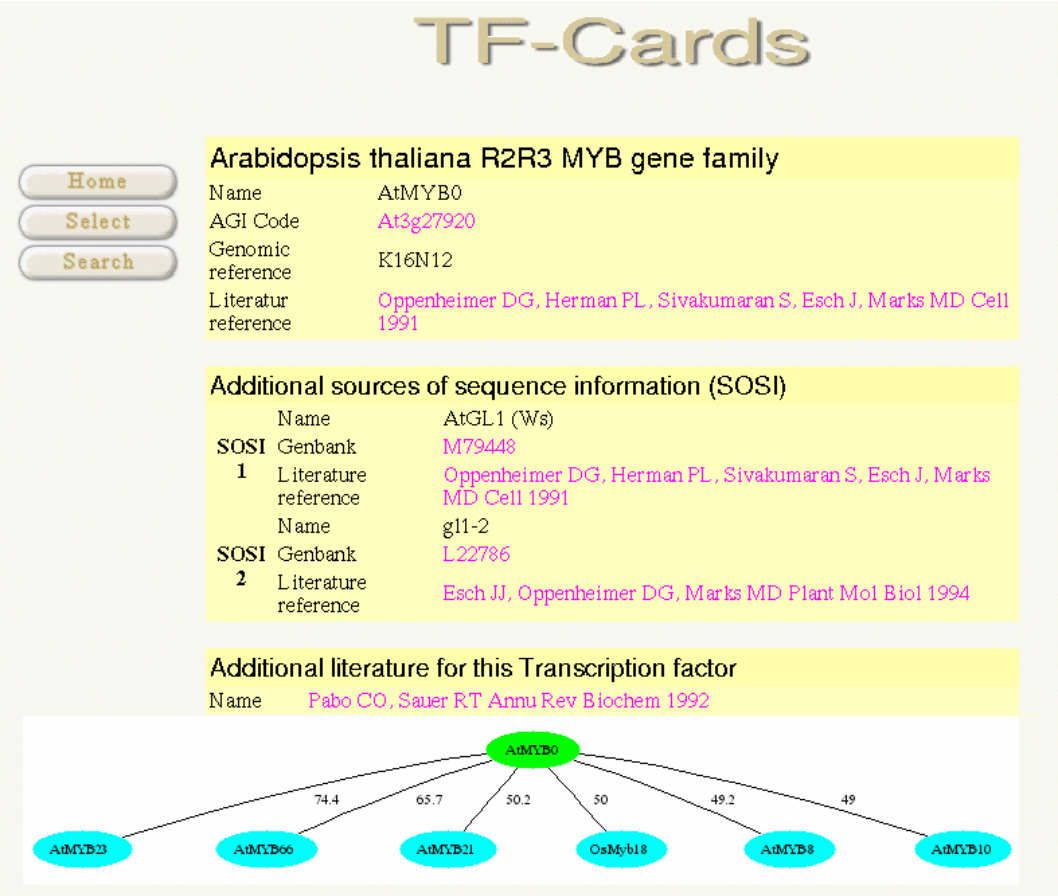


Abbildung 13: *TF-Cards* fassen auf einer dynamisch aus der Datenbank generierten Seite alle wichtigen Informationen zu einem Genfamilienmitglied zusammen. Der Benutzer kann aus verschiedenen Genfamilien und Organismen über ein Menü auswählen. Die Angaben sind weitgehend mit externen Datenbanken verknüpft.

5.3 Programme für die Sequenzsuche und Klassifizierung

Die Suche nach Mitgliedern einer Genfamilie in genomischen Sequenzen ist zeitaufwendig und fehleranfällig. Die durch Fortschritte in den Sequenzierungsprojekten sich ständig verändernde Datenlage wirkt zusätzlich erschwerend. Grundlegende Bestandteile dieser Fragestellung sind die Effizienz der Suchmethode und die algorithmische Klassifizierung von Genfamilienmitgliedern. Im Folgenden werden Programme beschrieben, die den Vorgang der Suche und Klassifizierung weitgehend automatisieren.

5.3.1 Klassifizierung von Genfamilien mit HMM

Für die Klassifizierung von Mitgliedern einer Genfamilie muß zuerst ein Klassifizierungsschema erstellt werden. Für MYB-Gene beginnt dieses hierarchisch gegliederte Schema mit einer Definition der genfamilien-typischen Domäne in der Proteinsequenz und untergliedert sich dann in die verschiedenen Typen von Sequenzwiederholungen der MYB-Domäne bis hin zu Motiven siehe dazu Abbildung 14.

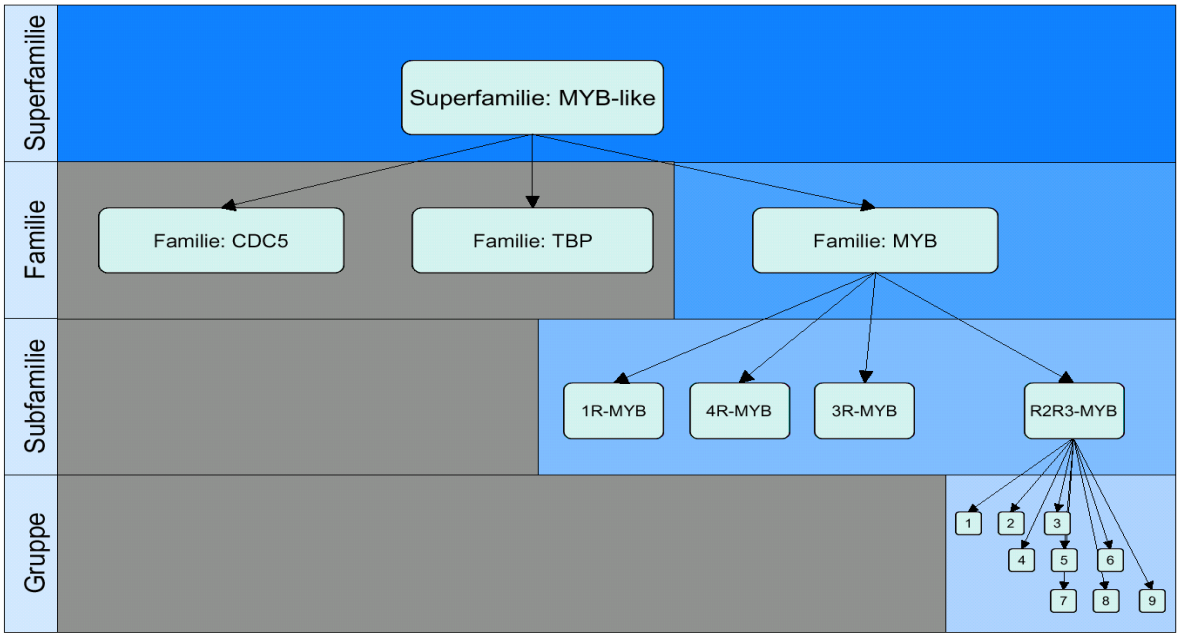


Abbildung 14: Klassifizierung **von Genfamilien**

Für die Erstellung des Klassifizierungsschemas für die MYB-Genfamilie wurden bislang bekannte MYB-Proteinsequenzen mit zwei und drei Sequenzwiederholungen

(Kranz, Denekamp et al. 1998) mit CLUSTALW (Thompson, Higgins et al. 1994) aligned und das Alignment mit dem Programm `Jalview` (Cuff, Clamp et al. 1998) dargestellt. Außerdem wurden die Sequenzen mit dem Programm `MEME` nach konservierten Bereichen durchsucht. Auf Basis der mit `MEME` berechneten Motive und der Untersuchung des Alignment wurden daraus die Bereiche der drei Sequenzwiederholungen ausgeschnitten und wiederum als Alignment gespeichert. Mit dem Programm `hmmbuild` (Parameter –s für „Smith-Waterman local alignment“) aus dem `HMMER` Programmpaket wurde aus dem Alignment ein HMM generiert. Dieser Vorgang wurde für die beiden Sequenzenwiederholungen von 2R3R-MYB-Proteinen und für die drei Sequenzwiederholungen von MYB-Proteinen mit drei Sequenzwiederholungen durchgeführt. Neben den Proteinsequenzen von MYB-Genen wurden nach demselben Verfahren auch Proteinsequenzen von Genen der GARP und MIDA1 Familie analysiert. Die aus diesen „alignments“ berechneten HMM dienen als Negativkontrolle bei der Klassifizierung. Die berechneten HMM können in einer Datei zusammengefasst werden und dienen in den folgenden Schritten als Grundlage für die Klassifizierung.

5.3.2 Motivsuche mit Motif signature cooccurance scan (MSCS)

Die BLAST Suche ist eine einfache und standardmäßig genutzte Methode um nach ähnlichen Sequenzen in einer Sequenzdatenbank zu suchen. Der wesentliche Vorteil von BLAST ist die hohe Geschwindigkeit, mit der auch große Datensätze durchsucht werden können. Bei der Suche von Proteinsequenzen gegen genomische DNA wirkt sich allerdings die Tatsache nachteilig aus, dass Motive durch die Genstruktur unterbrochen sein können.

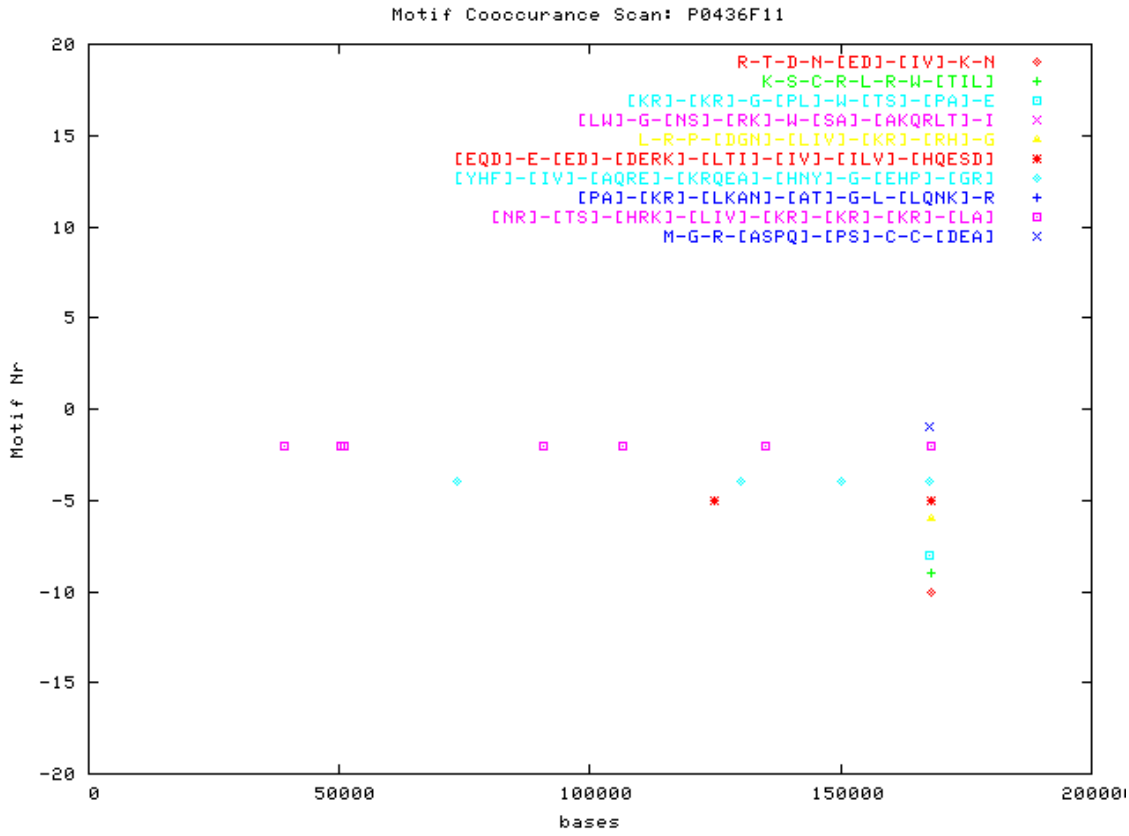


Abbildung 15: Motif Cooccurance Scan: Auf der X-Achse sind Basenpaare aufgetragen. Die Y-Achse hat keine Einheit, die Motive sind einfach in fixen Abständen zueinander aufgetragen. In der Legende oben Rechts sind die Symbole für die Suchmuster im Prosite-Format aufgeführt. Bei ca. 170000 ist eine senkrechte Aufreihung von acht Motivtreffern zu erkennen. An dieser Stelle ist aufgrund der Anhäufung von Treffern ein MYB-Gen zu erwarten.

Das Programm *MSCS* wurde entwickelt, um mit hoher Spezifität und Sensitivität in einem nicht annotierten genomischen Sequenzdatensatz nach Genen einer Familie zu suchen. Durch die Auswahl von Motiven mit hoher Signifikanz für die gesuchte Familie kann die Spezifität der Suche erhöht werden. Durch das Aufteilen in kurze Fragmente, die ähnlich wie kurze Primer an vielen Stellen binden können, wird eine hohe Sensitivität erreicht. Zunächst werden mit MEME (Bailey and Gribskov 1998) aus einem Satz von Proteinsequenzen, die entweder aus dem gleichen Organismus oder

heterolog sein können, kurze konservierte Sequenzmotive gesucht, die in allen Sequenzen vorkommen.

Die von MEME ermittelten Alignments der Motive sind in der Signifikanz abfallend aufgelistet. Für die Suche nach MYB-Genen werden Motive mit einem „e-value“ kleiner als 10^{-20} verwendet. Die von MEME ermittelten Alignments werden mit einem Perl-Skript analysiert und in Prosite-Muster umgewandelt.

Mit den Prosite-Motive werden die genomische Daten anschließend mit dem Programm FUZZTRANS (Rice, Longden et al. 2000) durchsucht.

Insgesamt kommt es so zu einer wenig spezifischen, aber sehr sensitiven Erkennung der kurzen Motivabschnitte in der genomischen Sequenz. Neben korrekt erkannten Abschnitten werden aber, aufgrund der hohen Fehlertoleranz, auch viele Falschpositive erkannt. Daher werden in dem folgenden Schritt die Treffer in Form von erkannten Abschnitten auf der Sequenz gefiltert. Es werden solche Treffer eliminiert, die keine benachbarten Treffer anderer Motive in einem festgelegten Fenster aufweisen. Für das Fenster wurde ein Wert von 3000 bp festgelegt. Zusätzlich wurde eine Mindestanzahl von 3 unterschiedlichen Motiven innerhalb des Prüffensters festgelegt.

Durch diese Filterschritte ist es mit hoher Spezifität möglich, die Abschnitte herauszufiltern, die eine starke Anhäufung von familienspezifischen Motiven in einem kurzen Abschnitt genomischer DNA aufweisen. Die Abbildung 15 zeigt das Suchergebnis für einen Klon vor der Filterung. Nach der Filterung werden die detektierten Sequenzabschnitte dem automatisierten Verfahren des im Folgenden beschriebenen Programms *FamilyBuilder* unterworfen.

5.3.3 AFP

Das Programm „*Aminoacid Frequency Plot*“ (*AFP*) ist ein Visualisierungswerkzeug, das für die Arbeit mit verschiedenen Genfamilien entwickelt wurde. Die Konservierung von Aminosäuren in einem multiplen Alignment wird häufig mit „Sequence Logos“ (Schneider and Stephens 1990) dargestellt. Diese Art der Visualisierung hat den Vorteil, dass stark konservierte Aminosäuren leicht erkennbar sind. *AFP* stellt die Häufigkeit der Aminosäuren an einer Position in einem multiplen Alignment dar. Das Programm kann über eine Weboberfläche genutzt werden.

5.3.4 Kombinierte Suche und Klassifizierung in *FamilyBuilder*

Das Programm *FAMILY BUILDER* automatisiert die Identifizierung und Klassifizierung von Genfamilienmitgliedern in genomischen Daten. Das Programm besteht aus mehreren Perl-Skripten, die automatisch nacheinander aufgerufen werden (siehe Abbildung 16). Die zwei Grundvoraussetzungen für die automatisierte Suche mit *FAMILYBUILDER* sind Beispielsequenzen für Gene der zu charakterisierenden Familie und ein Klassifizierungsschema. Mit den Beispielsequenzen werden HMM für die Klassifizierungseigenschaften generiert. Die HMM werden in einer Datenbank hierarchisch erfaßt. Die Suche nach Genen kann mit verschiedenen Programmen, die eine Suche von Proteinsequenzen gegen DNA zulassen, beginnen. Eine Möglichkeit ist der `tblastn` Algorithmus aus dem `BLAST` Programm. Die Suche kann aber auch mit einfachen Mustersuchen `FUZZTRANS` (Rice, Longden et al. 2000) oder dem im vorangegangenen Abschnitt beschriebenen *MSCS* erfolgen.

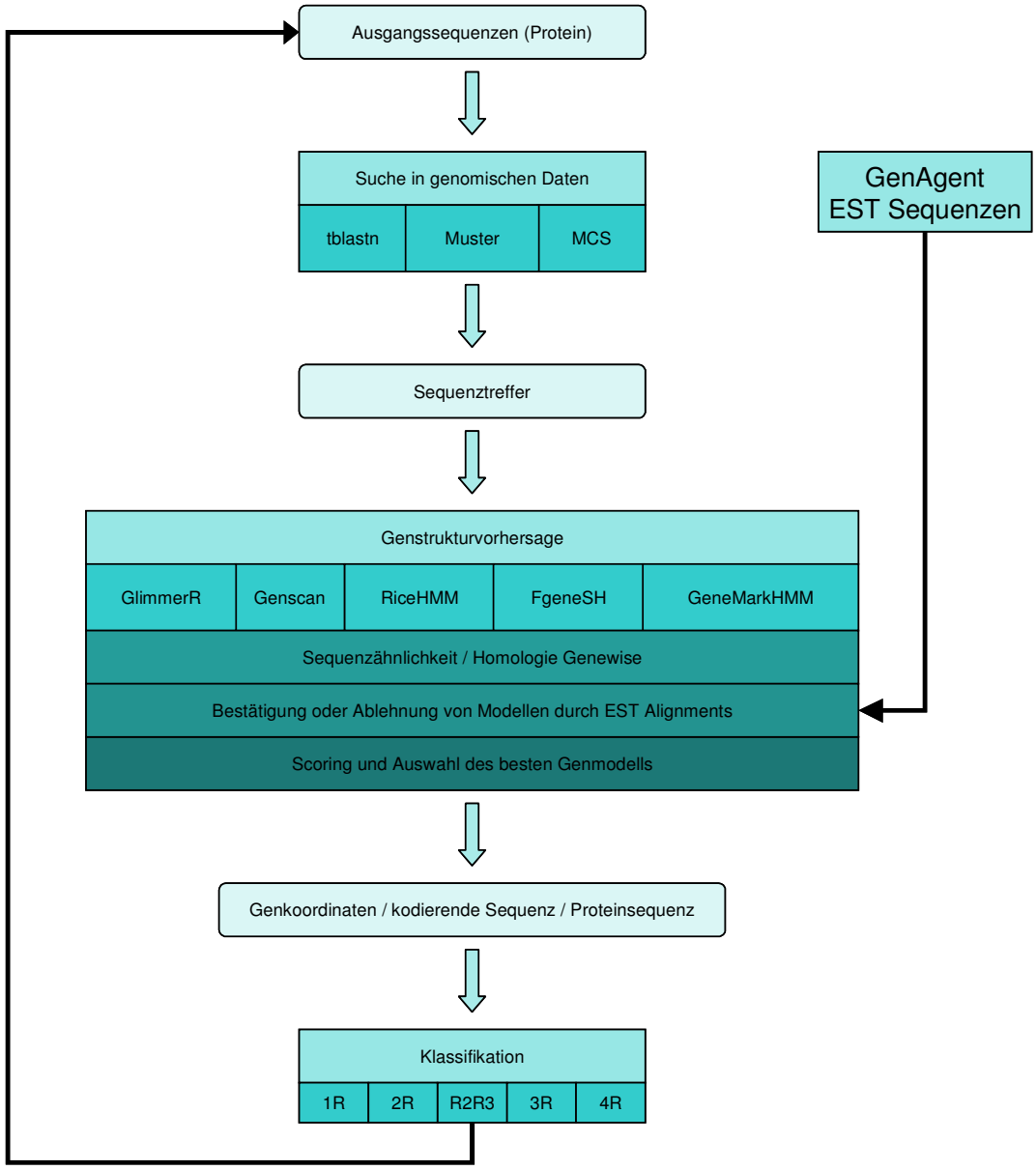


Abbildung 16: Ablaufschema des FamilyBuilder. Programmteile sind als Rechtecke dargestellt, Daten als Rechtecke mit abgerundeten Ecken. Beschreibung des Ablaufs siehe Text.

Wichtig ist bei der Verwendung von `BLAST` als anfängliche Suche, dass die Parameter mit niedriger Stringenz gewählt werden. Die Anfangssuche soll möglichst alle Gene finden, auch wenn dabei (viele) Falschpositive detektiert werden. Die folgenden Schritte, insbesondere die Klassifizierung, erkennen und entfernen die Falschpositiven. Nach der Suche werden die genomischen Treffer weiter analysiert. Der bei der Suche bestimmte ähnliche Sequenzabschnitt wird strangaufwärts und strangabwärts erweitert. Durch diese Maßnahme soll sichergestellt werden, dass der im Folgenden behandelte Sequenzbereich das komplette Gen enthält. Dieser

erweiterte Abschnitt wird dann mit verschiedenen Programmen für die Genstrukturvorhersage bearbeitet. Die Ausgaben der verschiedenen Programme werden in der Datenbank erfaßt. Zusätzlich zu den *ab initio* Vorhersagen wird eine homologie-basierte Genstrukturvorhersage durchgeführt. Dabei wird ein möglichst ähnliches Protein in Sequenzdatenbanken gesucht und mit dem Programm GeneWise gegen die genomische Sequenz aligned. Durch diese Methode ergibt sich auch eine Übersicht über die Genstruktur. Alle Genmodelle werden in der Datenbank abgelegt, und stehen somit auch später noch einer visuellen Inspektion zur Verfügung.

Das *GenAgent* Programm kann genutzt werden, um die EST-Sequenzen zu finden, mit denen die Genmodelle bestätigt werden können. Dazu werden aus externen Datenbanken alle EST-Sequenzen für diesen Organismus geladen und mit `StackPack` geclustert. Die geclusterten Sequenzen werden dann mit `sim4` gegen die genomischen Sequenzabschnitte verglichen.

Wenn zunächst keine EST-Sequenzen ein Genmodell eindeutig verifizieren können wird aus den Genmodellen jenes ausgewählt, das im Vergleich mit den HMM den höchsten “score” liefert. Dieses Verfahren simuliert die Entscheidung, die ein Wissenschaftler trifft, wenn er aus den verschiedenen Genmodellen das Modell auswählt, welches am besten die typische Proteinsequenz dieser Genfamilie beschreibt. Alternativ dazu wurde das Programm `Combiner` (TIGR) getestet, das die Genmodelle verschiedener Vorhersagen zusammenfassen kann. Da aber die Vorhersageprogramme zum Teil mit sehr ähnlichen Verfahren arbeiten, ist eine einfache Mehrheitsentscheidung nicht sinnvoll.

Der nächste Schritt im Programmablauf ist die Klassifizierung. Die Proteinsequenzen werden zunächst als MYB-Superfamilien Proteine identifiziert und bewertet. Dies geschieht mit einem HMM über die gesamte MYB-Domäne. Die Abgrenzung der MYB-Familie von den MYB-ähnlichen Familien GARP und MIDA1 mit zwei Sequenzwiederholungen geschieht durch zwei HMM für diese Familien. Durch den Vergleich mit einem HMM aller Domänen-Bestandteile werden die Bereiche der Sequenzwiederholungen lokalisiert und das am höchsten bewertete Modell ausgewählt. Auf diese Weise werden sowohl die Anzahl der Wiederholungen, als auch der Typ des Motivs ermittelt.

5.4 Biologische Ergebnisse:

5.4.1 Suche nach Genen der R2R3-MYB-Subfamilie in Genomsequenzen von *Arabidopsis thaliana*

Die Suche nach R2R3-MYB-Genen in Genomsequenzen von *Arabidopsis thaliana* mit den vorangehend beschriebenen Programmen ergab 126 R2R3-MYB-Gene, fünf 3R-MYB-Gene und ein 4R-MYB-Gen. Die Grundlage für die Suche bildeten die Sequenzdaten aus der MatDB des „*Munich Information Center for Protein Sequences*“ (MIPS). Die Ergebnisse dieser Analysen wurden in der Teilpublikation (Stracke, Werber et al. 2001) veröffentlicht. In der Veröffentlichung (Kranz, Denekamp et al. 1998) waren 97 Familienmitglieder (102 inklusive der Sequenzfragmente) beschrieben worden. Drei der 3R-MYB-Gene und das 4R-MYB-Gen waren zu Beginn dieser Arbeit noch nicht bekannt und konnten in dieser Arbeit erstmalig beschrieben werden. Somit konnten mit Hilfe der beschriebenen Programme und der Genomsequenz von *Arabidopsis thaliana* als Datengrundlage die Mitglieder der R2R3-MYB-Subfamilie in *Arabidopsis thaliana* vollständig identifiziert werden. Das 4R-MYB-Gen ist bislang das erste beschriebene MYB-Gen, dessen Proteinsequenz vier Sequenzwiederholungen enthält. Neben den vier vollständigen Sequenzwiederholungen sind auch noch Teile einer weiteren Sequenzwiederholung vorhanden. 4R-MYB weist Sequenzähnlichkeiten zu AtCDC5 auf.

Bei allen Genen wurde die Genstruktur annotiert und - soweit vorhanden - die EST-Sequenzen und „full length“ cDNA für die Verifikation der Genstruktur verwendet. Die Referenzinformationen zu den neu beschriebenen MYB-Genen sind in

Tabelle 1 zusammengefasst. Im Anhang sind in Tabelle 2 alle in dieser Arbeit und in vorangegangenen Arbeiten beschriebenen MYB-Gene aus *Arabidopsis thaliana*, die mehrere Sequenzwiederholungen enthalten aufgeführt.

Name	Gen Kode	Synonym	GenBank Acc. Nr.	Klon	Referenz
AtMYB22	At5g40430		AF175986	MPO12	Stracke *et al.*, 2001
AtMYB25	At2g39880		AF175988	T28M21	Stracke *et al.*, 2001
AtMYB42	At4g12350	T4C9.190	AF175999	T4C9	Stracke *et al.*, 2001
AtMYB43	At5g16600	MTG13.4	AF175990	MTG13	Stracke *et al.*, 2001
AtMYB49	At5g54230		AB010695	MDK4.4	Stracke *et al.*, 2001
AtMYB55	At4g01680		AF104919	T15B16.4	Stracke *et al.*, 2001
AtMYB64	At5g11050		AY032854	T5K6	Stracke *et al.*, 2001
AtMYB85	At4g22680		AF175993	T12H17	Stracke *et al.*, 2001
AtMYB88	At2g02820		AF175994	T20F6	Stracke *et al.*, 2001
AtMYB89	At5g39700		AF175995	MIJ24	Stracke *et al.*, 2001
AtMYB97	At4g26930		AF176002	F10M23	Stracke *et al.*, 2001
AtMYB98	At4g18770		AF176003	F28A21	Stracke *et al.*, 2001
AtMYB100	At2g25230		AF176004	T22F11	Stracke *et al.*, 2001
AtMYB103	At1g63910		AF214116	T12P18	Stracke *et al.*, 2001
AtMYB105	At1g69560		AF249308	F24J1	Stracke *et al.*, 2001
AtMYB106	At3g01140		AF249309	T4P13	Stracke *et al.*, 2001
AtMYB107	At3g02940		AF249310	F13E7	Stracke *et al.*, 2001
AtMYB108	At3g06490		AF262733	F5E6	Stracke *et al.*, 2001
AtMYB109	At3g55730		AF262734	F1I16	Stracke *et al.*, 2001
AtMYB110	At3g29020		AF272732	K5K13	Stracke *et al.*, 2001
AtMYB111	At5g49330		AF371977	K21P3	Stracke *et al.*, 2001
AtMYB112	At1g48000	T2J15.9	AY008377	T2J15	Stracke *et al.*, 2001
AtMYB113	At1g66370	T27F4.12	AY008378	T27F4	Stracke *et al.*, 2001
AtMYB114	At1g66380	T24F4.13	AY008379	T24F4	Stracke *et al.*, 2001
AtMYB115	At5g40360		AF334814	MPO12	Stracke *et al.*, 2001
AtMYB116	At1g25340		AF334815	F4F7	Stracke *et al.*, 2001
AtMYB117	At1g26780		AF334816	T24P13	Stracke *et al.*, 2001
AtMYB118	At3g27780		AF334817	MGF10	Stracke *et al.*, 2001
AtMYB119	At5g58850		AF371978	K19M22	Stracke *et al.*, 2001
AtMYB120	At5g55020		AF371979	K13P22	Stracke *et al.*, 2001
AtMYB121	At3g30210		AF371980	MIL15	Stracke *et al.*, 2001
AtMYB122	At1g74080		AF371983	F2P9	Stracke *et al.*, 2001
AtMYB123	At5g35550		AF371981	MOK9	Stracke *et al.*, 2001
AtMYB124	At1g14350		AF371982	F14L17	Stracke *et al.*, 2001
AtMYB4R1	At3g18100		AY033827	MRC8	Stracke *et al.*, 2001
AtMYB3R3	AT3g09370		AF214117	F3L24	Stracke *et al.*, 2001
AtMYB3R4	AT5g11510	F15N18.100	AF371975	dt_e_23	Stracke *et al.*, 2001
AtMYB3R5	At5g02320	T1E22.80	AF371976	T1E22	Stracke *et al.*, 2001

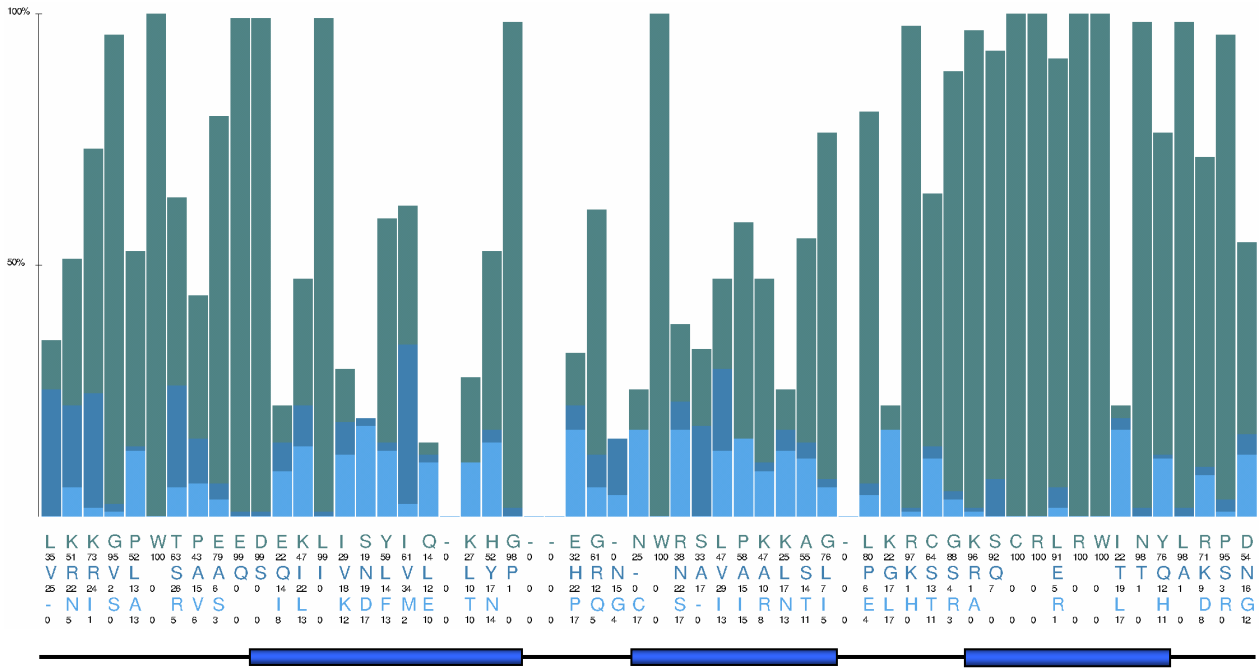
Tabelle 1: Die Tabelle enthält alle in dieser Arbeit identifizierten MYB-Gene aus *Arabidopsis thaliana*, deren Proteinsequenz mehr als eine Sequenzwiederholung aufweist. Die Namensgebung basiert auf der in (Kranz, Denekamp et al. 1998; Romero, Fuertes et al. 1998) eingeführten Nomenklatur. Die eindeutigen Identifizierer („unique identifier") wurden aus der MatDB Datenbank des Munich Information Center for Protein Sequences (MIPS) bezogen.

5.4.2 Suche nach Genen der R2R3-MYB-Subfamilie in Genomsequenzen von *Oryza sativa*

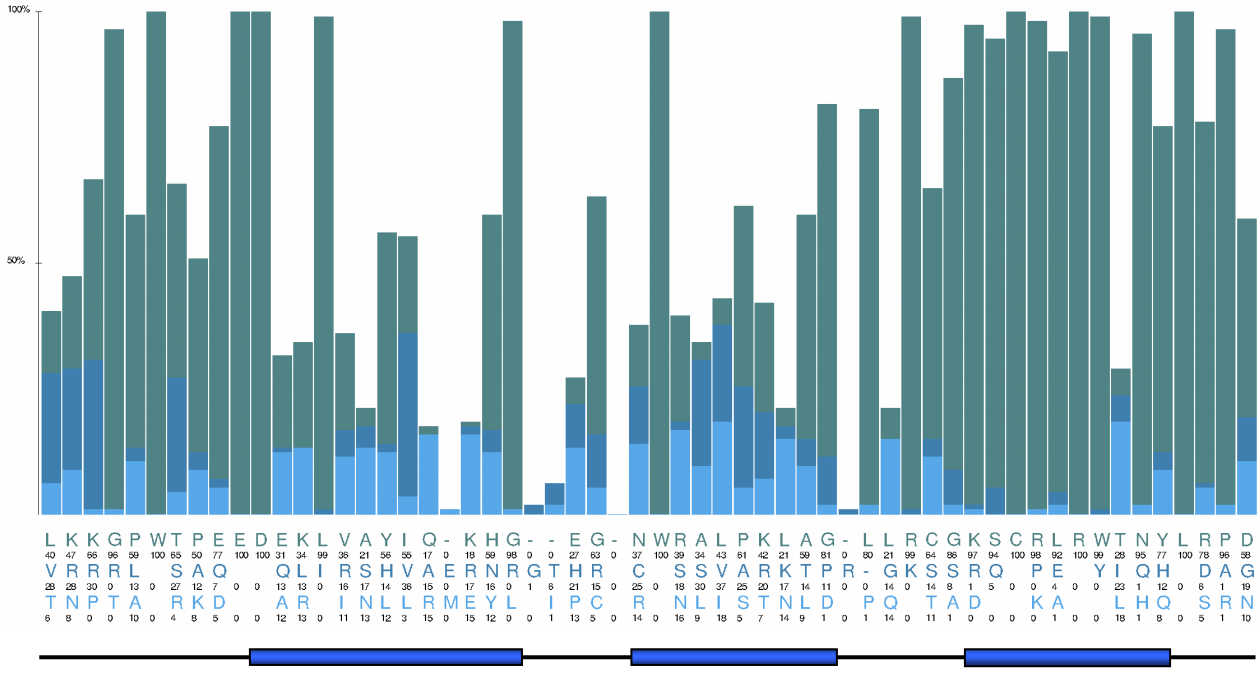
Die Suche nach R2R3-MYB-Genen in den Genomsequenzen wurde entsprechend der Vorgehensweise bei *Arabidopsis thaliana* durchgeführt. Grundlage für die Suche bildete ein Sequenzdatensatz, der aus der NCBI Datenbank am 20.4.03 in die GenomeDB geladen wurde. Es wurden 120 R2R3-MYB-Gene identifiziert. Davon sind 114 bislang nicht beschrieben worden. Des Weiteren wurden fünf 3R- und ein 4R-MYB-Gen identifiziert.
Die Referenzinformationen zu den beschriebenen R2R3-MYB-Genen aus *Oryza sativa* sind in Tabelle 3 im Anhang zusammengefasst.

5.4.3 Vergleichende Untersuchung der R2R3-MYB-Domäne in Proteinsequenzen von R2R3-MYB-Genen aus *Arabidopsis thaliana* und *Oryza sativa*.

Für die Untersuchung der R2R3-MYB-Domäne wurden mit den Proteinsequenzen aller bislang bekannten und der neu identifizierten R2R3-MYB-Gene aus *Arabidopsis thaliana* und *Oryza sativa* ein multiples Alignment errechnet. Aus diesem wurde der Bereich der konservierten Domäne ausgeschnitten und gespeichert. Mit dem Programm *AFP* (siehe 5.3.3) wurden dann die Aminosäurehäufigkeiten berechnet und in Form eines Balkendiagramms dargestellt.
In Abbildung 17 und Abbildung 18 sind die Ergebnisse dargestellt. Die R2R3-MYB-Domäne ist in beiden Organismen stark konserviert. Der Bereich um die dritte Helix ist in R2 und R3 stärker konserviert als der restliche Bereich der Domäne. Zwischen den Helices liegen Aminosäuren mit geringerer Konservierung. Bei den Proteinsequenzen aus *Oryza sativa* sind im Vergleich zu denen aus *Arabidopsis thaliana* häufiger Insertionen von kleinen Aminosäuren in den Bereichen zwischen den Helices festzustellen.



A



B

Abbildung 17: Die drei häufigsten Aminosäuren an jeder Position sind in Form von farbigen Balken aufgetragen. Unterhalb der Balken sind die Aminosäuren und die jeweiligen Häufigkeiten in Prozent angegeben. Die blauen Flächen unterhalb der Diagramme kennzeichnen die drei Helices in der Sequenzwiederholung. (A) R2 Sequenzwiederholung der MYB-**Domäne in Proteinsequenzen aus *Arabidopsis thaliana.* (B) R2 Sequenzwiederholung der** MYB-**Domäne in Proteinsequenzen aus *Oryza sativa.***

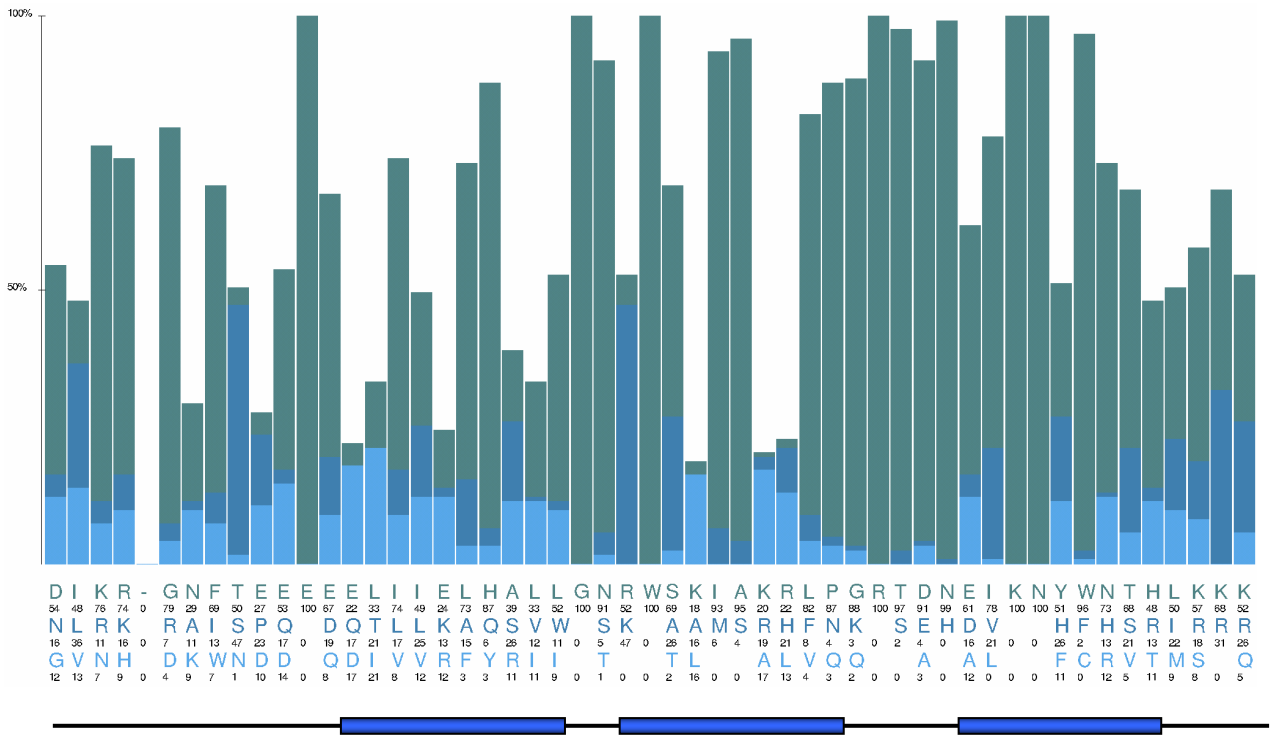


Abbildung 18: Die drei häufigsten Aminosäuren an jeder Position sind in Form von farbigen Balken aufgetragen. Unterhalb der Balken sind die Aminosäuren und die jeweiligen Häufigkeiten in Prozent angegeben. Die blauen Flächen unterhalb der Diagramme kennzeichnen die drei Helices in der Sequenzwiederholung. (A) R3 Sequenzwiederholung der MYB-**Domäne in Proteinsequenzen aus *Arabidopsis thaliana.* (B) R3 Sequenzwiederholung der** MYB-**Domäne in Proteinsequenzen aus *Oryza sativa.***

5.4.4 Analyse von Gruppen in den MYB-Genen aus *Arabidopsis thaliana* mit mehr als einer Sequenzwiederholung in der Proteinsequenz

Die Proteinsequenzen aller R2R3-, 3R und 4R-MYB-Gene wurden analysiert, um Ähnlichkeiten in den Sequenzen festzustellen und die Gene anhand der Ähnlichkeiten in Gruppen zusammenzufassen. (Dieses Verfahren wird im folgenden entsprechend dem englischen Fachterminus als Clustering, bezeichnet)
Für die Clustering-Analyse wurde die Gesamtlänge der Proteinsequenzen verwendet. Mit dem Programm `clustalw` wurde ein multiples Alignment berechnet. Dieses wurde mit `seqboot` einem „bootstrap“ Verfahren mit 500 Wiederholungen unterzogen. Mit dem daraus resultierenden Sequenzdatensatz wurden mit `protdist` die Sequenzdistanzen berechnet. Mit `neighbor` wurde anschließed nach der „neighbor joining“ Methode ein Clustering berechnet, aus dem mit `consense` ein Konsensusbaum nach der Mehrheitsmethode erstellt wurde. Die „boostrap“ Werte der zur Einteilung in Gruppen verwendeten Teilbäume wurden überprüft und waren größer als 250. Der Konsensusbaum wurde mit `drawgram` visualisiert und anschließend wurden die Gruppen mit farbigen Flächen und Symbolen annotiert.

5.4.5 Identifizierung von Motiven im C-Terminus der Proteinsequenzen von R2R3-MYB-Genen aus *Arabidopsis thaliana*

Die Proteinsequenzen der MYB-Gene wurden mit `meme` analysiert, und die in `meme` identifizierten Motive in multiplen Alignments der betreffenden Sequenzen manuell überprüft. Die Motive wurden mit den in (Kranz, Denekamp et al. 1998) beschriebenen Motiven verglichen und wenn möglich angepasst. Aufgrund der in dieser Arbeit identifizierten MYB-Gene konnten zwei Subgruppen neu hinzugefügt werden. Mehrere Motive konnten in diesem Ansatz nicht mehr berücksichtigt werden, da die Analyse im Gegensatz zu (Kranz, Denekamp et al. 1998) ausschließlich auf MYB-Genen aus *Arabidopsis thaliana* beruhte. Die Ergebnisse der Analysen aus 5.4.4 und 5.4.5 sind in der folgenden Grafik zusammengefasst (Abbildung 19)

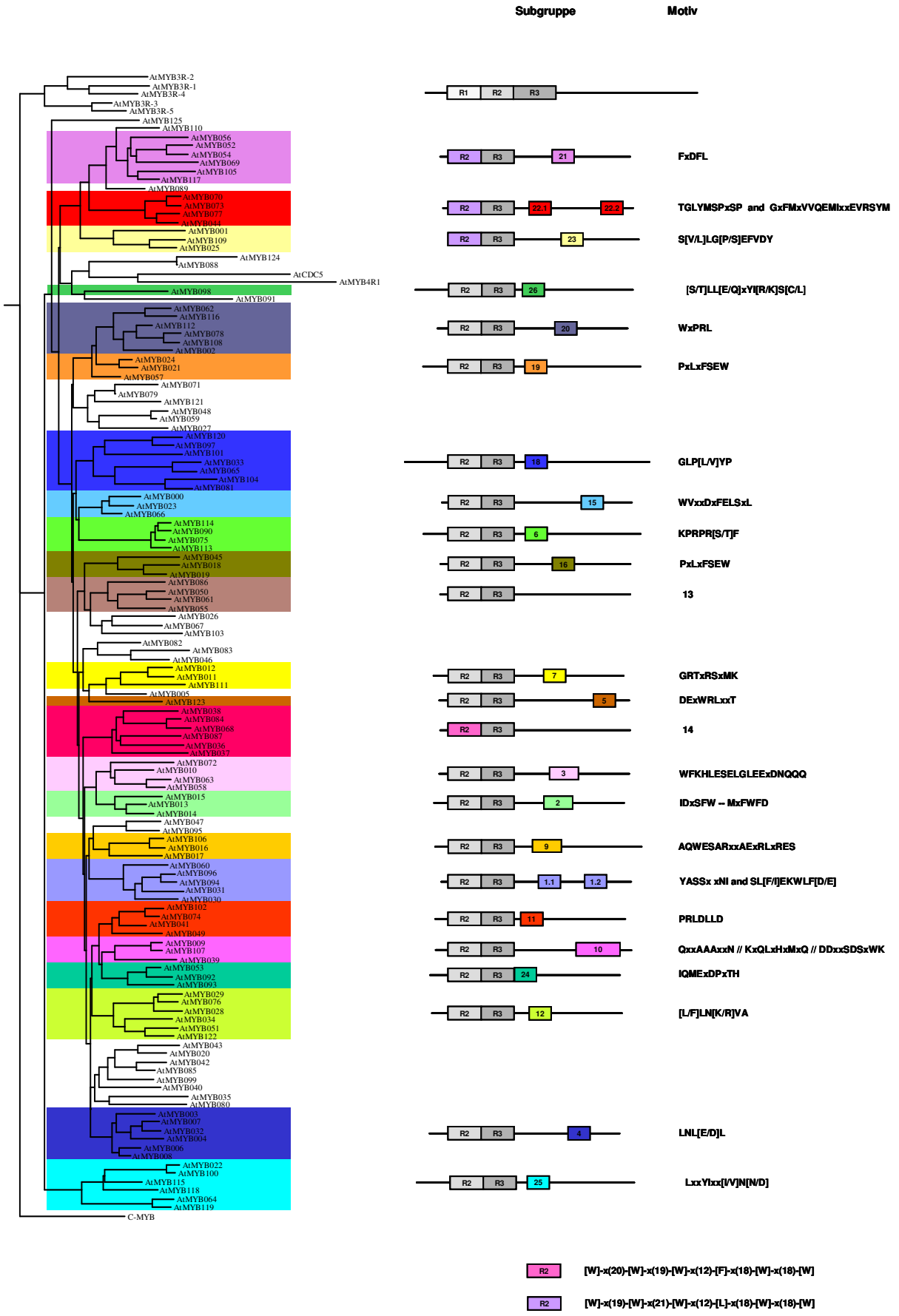


Abbildung 19: Clustering-Analyse der R2R3, 3R und 4R Proteinsequenzen von MYB-Genen aus *Arabidopsis thaliana*. Die farbigen Flächen kennzeichnen die Zugehörigkeit zu einer Motivgruppe. Rechts neben den Gruppen sind schematisch die Domänen und die Position der Motive dargestellt. Daneben sind bei den C-terminalen Motiven die Konsensussequenzen der Motive angegeben.

5.4.6 Clustering-Analyse und Identifizierung von Motiven im C-Terminus der Proteinsequenzen von R2R3-MYB-Genen aus *Oryza sativa*

Die Proteinsequenzen aller R2R3-, 3R- und 4R-MYB-Gene, wurden wie unter 5.4.4 und 5.4.5 beschrieben analysiert um Ähnlichkeiten in den Sequenzen festzustellen und die Gene anhand der Ähnlichkeiten in Gruppen zusammenzufassen.
Die Ergebnisse der Analyse sind in der folgenden Grafik zusammengefasst (Abbildung 20).

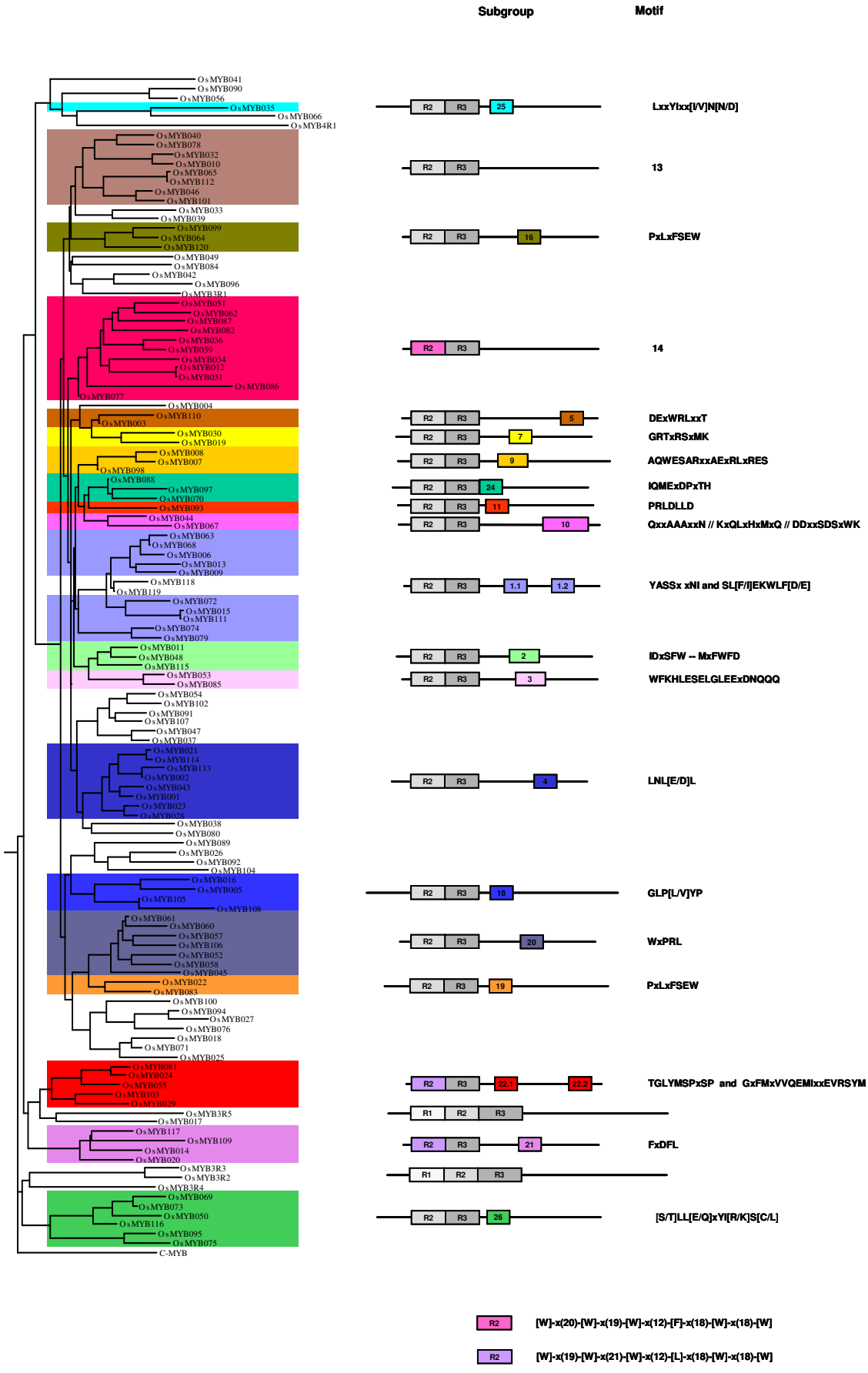
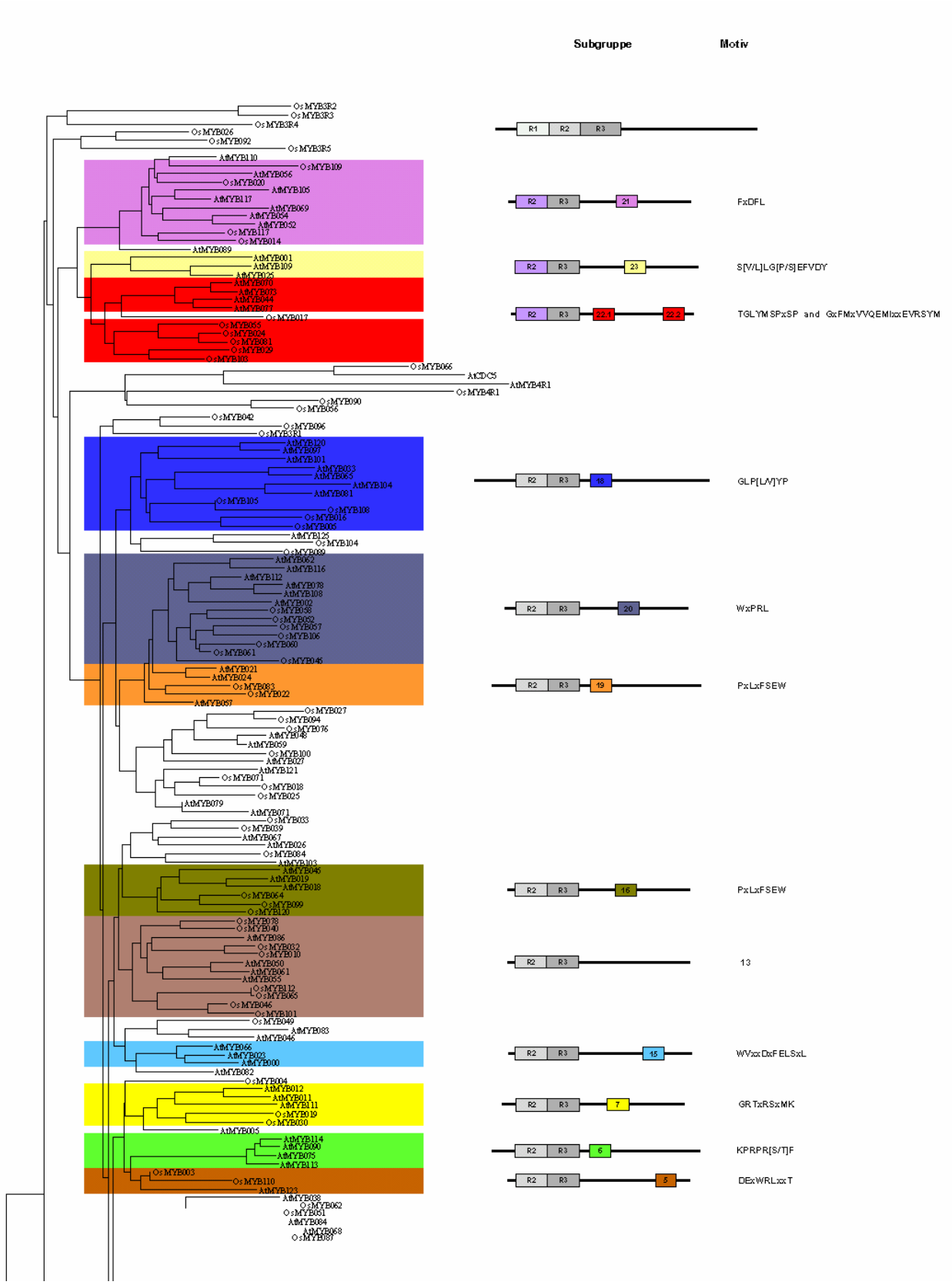


Abbildung 20: Clustering-Analyse der R2R3, 3R und 4R Proteinsequenzen von MYB-Genen aus *Oryza sativa*. Die farbigen Flächen kennzeichnen die Zugehörigkeit zu einer Motivgruppe. Rechts neben den Gruppen sind schematisch die Domänen und die Position der Motive dargestellt. Daneben sind bei den C-terminalen Motiven die Konsensussequenzen der Motive angegeben.

5.4.7 Vergleichende Analyse der Proteinsequenzen aus *Arabidopsis thaliana* und *Oryza sativa*.

Nachdem sowohl in den Sequenzen aus *Arabidopsis thaliana*, als auch in den Sequenzen aus *Oryza sativa* übereinstimmende Motive identifiziert und beschrieben werden konnten, wurde untersucht, ob die Gruppen bei einer Analyse über alle Sequenzen aus beiden Organismen erhalten bleiben. Die Vorgehensweise entspricht der aus 5.4.4 und 5.4.5. Die Ergebnisse der Analyse sind in der Grafik auf den folgenden zwei Seiten zusammengefasst.


Subgruppe
Motiv
R1
R2
R3
FxDFL
S[V/L]LG[P/S]EFVDY
TGLYMSPxSP and GxFMxVVQEMIxxEVRSYM
GLP[L/V]YP
WxPRL
PxLxFSEW
PxLxFSEW
13
WVxxDxFELSxL
GRTxRSxMK
KPRPR[S/T]F
DExWRLxxT

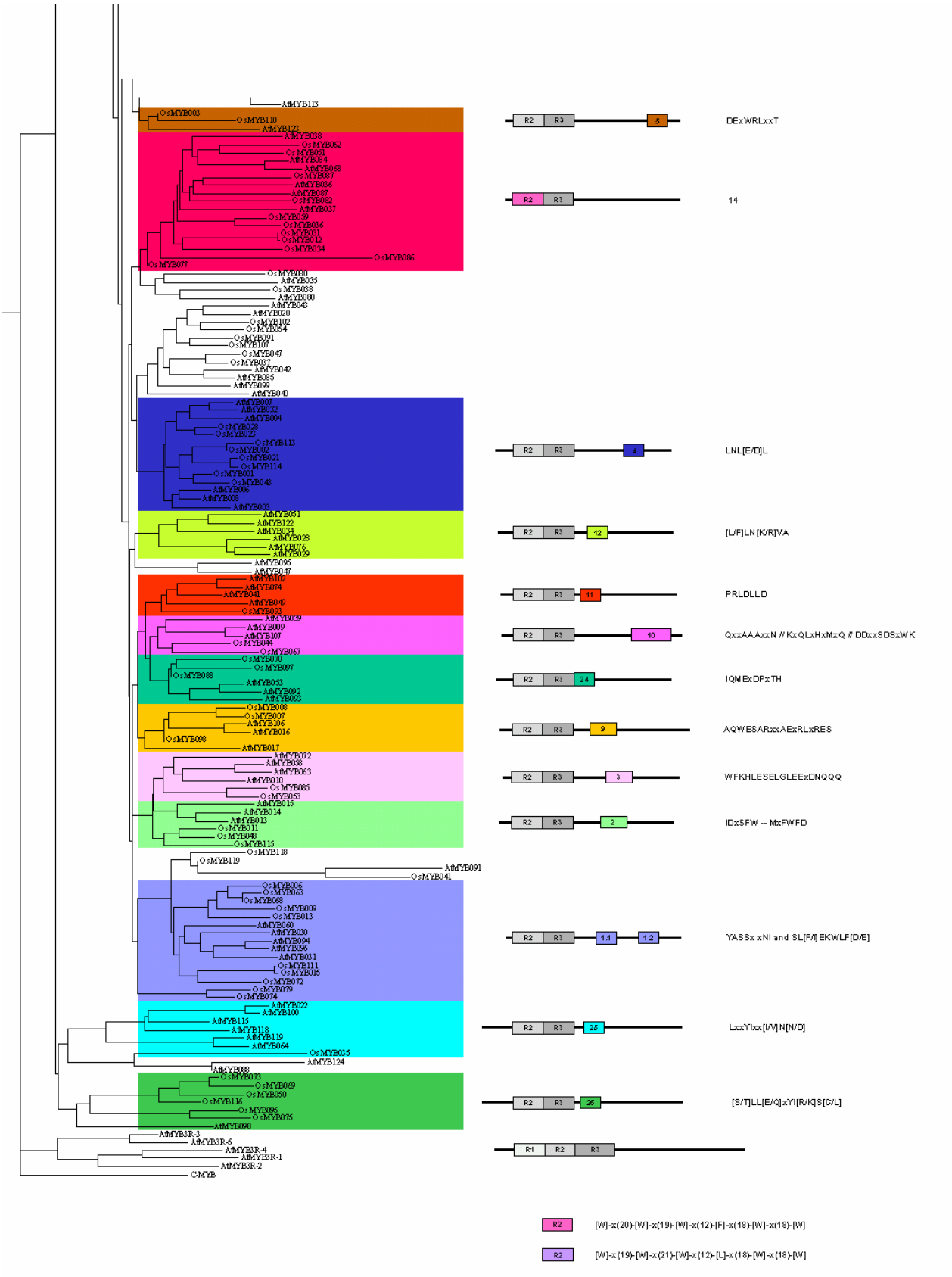


Abbildung 21: Clustering-Analyse der R2R3, 3R und 4R Proteinsequenzen von MYB-Genen aus *Arabidopsis thaliana* und *Oryza sativa*. Die farbigen Flächen kennzeichnen die Zugehörigkeit zu einer Motivgruppe. Rechts neben den Gruppen sind schematisch die Domänen und die Position der Motive dargestellt. Daneben sind bei den C-terminalen Motiven die Konsensussequenzen der Motive angegeben.

6 Diskussion

In der vorliegenden Arbeit wurden Programme und Datenbanken für die automatische Identifizierung und Klassifizierung von Mitgliedern der R2R3-MYB-Subfamilien aus *Arabidopsis thaliana* und *Oryza sativa* entwickelt und zu deren weiterer Untersuchung eingesetzt.

Aus den bisher vorliegenden *Arabidopsis thaliana* und *Oryza sativa* Sequenzdaten von AGI und IRGSP (Sasaki and Burr 2000; The Arabidopsis Genome Initiative 2000) wurden alle Mitglieder der oben genannten Subfamilie identifiziert. Die abgeleiteten Proteinsequenzen der gefundenen Gene wurden anschließend verglichen, und aufgrund ihrer spezifischen Motive in der MYB-Domäne und am C-Terminus in Gruppen eingeteilt.

Eine besondere Herausforderung für die Bewältigung dieser Aufgabe waren die sich ständig ändernden und anwachsenden Sequenzinformationen, die im Zuge der laufenden Genom-Sequenzierungsprogramme zur Verfügung gestellt wurden. Im Folgenden möchte ich daher Möglichkeiten zur Integration dieser Datenmengen mit Hilfe der bioinformatischen Analyse diskutieren.

6.1 Potential der Pflanzengenomprogramme

Mit der Veröffentlichung der Genomsequenz von *Arabidopsis thaliana* ergab sich erstmals die Möglichkeit, Genfamilien eines pflanzlichen Organismus vollständig zu beschreiben und zu charakterisieren. In allen bisherigen Untersuchungen wurden die Gene einer Familie mit experimentellen Methoden, wie PCR, oder Hybridisierungs-Reaktionen identifiziert. Keine dieser Vorgehensweisen schließt jedoch aus, dass manche Gene - etwa aufgrund nicht geeigneter Versuchsbedingungen - nicht nachgewiesen, und so auch nicht als Mitglieder einer Familie erkannt werden. Mit Hilfe der Bioinformatik und der vorliegenden Genomsequenz ist es nun möglich Sequenzabschnitte, die MYB-Gene kodieren, zu identifizieren und zu analysieren. Aber auch bei dem Einsatz von Methoden der Bioinformatik entscheidet die Festlegung von Stringenz und definierenden Eigenschaften über die Annotation oder Abweisung eines Gens als Familienmitglied. Dies ist insbesondere von Bedeutung, wenn derartige Analysen in großem Umfang automatisch vorgenommen werden. Daher muß der Analyse mit Methoden der Bioinformatik die Definition und Klassifizierung einer Genfamilie und ihrer Mitglieder vorangehen.

Die Veröffentlichung der „draft“ Genomsequenzen zweier Subspezies aus *Oryza sativa* und die Fortschritte des IRGSP Projektes verändert die Lage erneut grundlegend. Hiermit ist neben der genomweiten Charakterisierung von Genfamilien einer monokotyledonen Pflanze nun auch die vergleichende Untersuchung zweier Pflanzengenome möglich.

6.2 Pflanzengenomprojekte zu Beginn der Arbeit und heute

In den Jahren 2000 bis 2003 hat sich die Informationslage mit hoher Geschwindigkeit verändert. Zum Ende des Jahres 2000 wurde mit dem Arabidopsis Genom die erste pflanzliche Genomsequenz veröffentlicht (The Arabidopsis Genome Initiative 2000). Seither wurde die öffentlich zugängliche Annotation der Gene und ihrer Funktionen laufend verbessert. Schon 1996 begannen die Arbeiten innerhalb des „rice genome research program“ (RGP) (Sasaki, Yano et al. 1996) mit dem Ziel einer öffentlich verfügbaren Genomsequenz aus *Oryza sativa* von hoher Qualität. Das später gegründete IRGSP (Sasaki and Burr 2000) integrierte die vorhandenen Ergebnisse in ein international koordiniertes Projekt, an dem zehn Länder beteiligt sind. Obwohl später gestartet, veröffentlichten Monsanto und Syngenta im April 2000 und im Februar 2001 unabhängig voneinander zwei „draft“ Versionen der Genomsequenz von *Oryza sativa sp. japonica* (Dickson 2000; Davenport 2001), die mit der „shotgun“ Methode in relativ kurzer Zeit fertig gestellt worden waren. Die Daten von Monsanto und Syngenta sind jedoch nur unter Einschränkungen und einer speziellen Lizenz zugänglich. Mit Hilfe von `BLAST` konnten die Sequenzdaten über eine Webseite durchsucht werden. Dabei war die Suche aber auf eine bestimmte Anzahl von Sequenzen pro Woche beschränkt. Beide Genomprojekte stellten keine kontinuierlichen Sequenzen, sondern nur kurze Sequenzabschnitte von wenigen kb zur Verfügung. Weder die Genstrukturen noch die Genfunktionen waren mit den beiden „draft“ Publikationen öffentlich verfügbar. Das vierte *Oryza sativa* Genomprojekt wurde am Bejing Genomics Institute (BGI) gestartet und hatte die „shotgun“ Genomsequenzierung von zwei Ökotypen der Subspezies *Oryza sativa indica* zum Ziel. Die Ergebnisse wurden zeitgleich mit der Fertigstellung des Syngenta-Projektes veröffentlicht (Yu, Hu et al. 2002). Während dieser Arbeit ist somit nicht nur die erste pflanzliche Genomsequenz veröffentlicht worden, es liegen vielmehr sechs unabhängig erstellte Genomsequenzen zweier verschiedener Pflanzenarten vor, die sich in Sequenzqualität und Annotation unterscheiden.

6.3 Informationslage in den Pflanzengenomprojekten

Für die Qualität der erstellten Genomsequenz ist die Sequenzierstrategie entscheidend. Mit dem „shotgun“ Ansatz wird eine „draft“ Sequenz in kurzer Zeit kostengünstig produziert. Eine wichtige Fehlerquelle hierbei ist jedoch der Zusammenbau der Genomfragmente zu einer größeren Sequenz. Aufgrund von repetetiven Sequenzen können dabei falsche Teile zusammengesetzt werden. Im Gegensatz dazu erhält die Strategie entlang eines „tiling path“ Klon für Klon zu Sequenzieren die Anordnung längerer Genomabschnitte und bietet daher eine größere Verlässlichkeit.

Die Sequenzierstrategie und der Zusammenbau größerer Abschnitte sind Bedingungen, die im Rahmen dieser Arbeit natürlich nicht beeinflusst werden konnten. Die mit der „shotgun“ Methode erstellten Genomsequenzen, sind im Sinne der mathematischen Abdeckung und des „draft“ Status vollständig, aber auf Ebene des einzelnen Gens wesentlich ungenauer. Daher wurden im Hinblick auf die folgenden Analysen die Genomsequenzen des IRGSP für *Oryza sativa* als Grundlage ausgewählt.

Neben Erstellung und Assemblierung der Sequenz gehört zu der Veröffentlichung eines Genoms für die Wissenschaftsgemeinschaft auch eine erste grundlegende Analyse.

6.3.1 Problematik der Genstrukturvorhersage

Der erste Schritt nach der Sequenzierung und Assemblierung eines Genoms ist die Vorhersage aller Gene und deren Strukturen. Im Fall des Arabidopsis-Genomprojektes wurden anfänglich 25500 Gene identifiziert, die für Proteine kodieren (The Arabidopsis Genome Initiative 2000). Zum heutigen Zeitpunkt sind es ca. 26600 (MIPS). Vergleichbar ist die Situation bei den 2002 veröffentlichten Genomsequenzen aus *Oryza sativa*. Die Schätzungen über die Anzahl der Gene in *Oryza sativa* reichen von 32000 bis zu 65000 (Goff, Ricke et al. 2002; Yu, Hu et al. 2002). Diese Annotationen stellen aber nur erste Analysen dar, die weitgehend vollautomatisch durchgeführt wurden.

Die Genstrukturvorhersage in Eukaryonten-Sequenzen stellt generell eine besondere Problematik dar. (Pavy, Rombauts et al. 1999) ermittelten in einer vergleichenden Analyse verschiedener Genstrukturvorhersageprogramme, dass nur 67 von 168 getesteten Genen korrekt vorhergesagt worden waren. (Pertea and Salzberg 2002)

testeten mit einem Datensatz von 1131 Gensequenzen aus *Arabidopsis thaliana* die Sensitivität und Spezifität von drei verschiedenen Genstruktur Vorhersage-programmen. Dabei wurden von 337 bis zu maximal 514 Gene exakt erkannt. Diese Zahlen belegten außerdem, dass besonders die automatisch mit *ab initio* Ansätzen annotierten Gene in sequenzierten Genomen in großer Zahl nicht korrekt annotiert sind.

Der Erfolg von Genstrukturvohersagen ist auch abhängig vom Organismus aus dem die Sequenzen stammen. So liegen für Sequenzen aus Monokotyledonen bisher wenige experimentell bestätigte Genmodelle vor. Dies bedeutet für die Programme, dass weniger Daten für das Training der Algorithmen vorhanden sind. Ein höherer GC-Gehalt und ein Gradient des GC-Gehaltes in monokotyledonen Gensequenzen erschweren die Analyse zusätzlich.

Ein Vorschlag zur Verbesserung der Genstrukturvorhersage ist der Einsatz von mehreren Vorhersageprogrammen und eine Kombination dieser Ergebnisse. (Pertea and Salzberg 2002) weisen darauf hin, dass die Erkennungsleistung durch den Einsatz von mehreren Vorhersageprogrammen verbessert werden kann. Schwierigkeiten bereiten bei dieser Art der Vorgehensweise die Ähnlichkeit der Algorithmen in den Programmen und die nicht dokumentierten Trainingsdatensätze. Eine reine Mehrheitsentscheidung aufgrund der Ergebnisse verschiedener Programme ist also irreführend. Viel versprechender sind der Einsatz von homologie-basierten Vorhersageprogrammen und die Nutzung von EST-Sequenzen zur Verifikation der Ergebnisse. (Pavy, Rombauts et al. 1999; Pertea and Salzberg 2002) diskutieren in ihren Publikationen Strategien um die Situation zu verbessern.

6.3.2 Problematik der Funktionsannotation

Die mit der Veröffentlichung der Genomsequenz von *Arabidopsis thaliana* vorliegende Annotation, bei der mit Hilfe von Sequenzvergleichen das putative Proteom berechnet wurde, wurde weitgehend automatisch durchgeführt. Bis heute finden sich in den von MIPS (Schoof, Zaccaria et al. 2002) veröffentlichten Sequenzdatensätzen von ca. 26600 identifizierten Genen über 15000 Gene mit dem Kommentar „unknown / putative / hypothetical protein“. Neben diesen Funktionsannotierungen liegen zunehmend auch Annotierungen vor, die auf der Erkennung konservierter Motive beruhen (Apweiler, Attwood et al. 2001). Diese sind aber nicht gleichbedeutend mit dem Wissen über eine Funktion des Proteins. Viele

Proteinsequenzen weisen mehrere konservierte Domänen auf, deren Gesamtheit die molekulare Funktion des Proteins bestimmt. Besonders deutlich wird dies am Beispiel der MYB-Superfamilie. Bei der automatisierten Annotation der Genomprojekte werden diese aufgrund ihrer MYB-Domäne als „MYB-protein“ oder „MYB-like“ annotiert. Auf diese Weise werden auch Gene anderer Genfamilien, wie zum Beispiel GARP-Gene oder Gene, die in einem anderen biologischen Zusammenhang stehen, miteinander gruppiert. Eine biologisch sinnvolle Annotation sollte aber die Anzahl der Sequenzwiederholungen und weitere Sequenzmotive bei der Klassifizierung berücksichtigen. Eine Annotation nur aufgrund einer konservierten Domäne kann also als Hinweis dienen, lässt aber nur eine grobe Einordnung zu.

6.4 Bioinformatik für die vergleichende Analyse von Genfamilien

Die vorangegangene Diskussion der Informationslage in den Genomprojekten macht deutlich, dass ein erfolgreiches Arbeiten mit mehreren Genomen und die Charakterisierung einer großen Genfamilie in diesen Sequenzen nur mit dem intensiven Einsatz von Methoden der Bioinformatik zu leisten ist. Die Bioinformatik umfasst dabei zum einen Lösungen für die Datenintegration, zum anderen automatisierte Verfahren für die Analysen.

Bei der Charakterisierung von Genfamilien können die Sequenzdaten aus den Genomprojekten nicht isoliert betrachtet werden, sondern müssen mit biologischen Daten in Zusammenhang gebracht werden. Zu diesen nicht genomischen Daten gehören EST-Sequenzen, Expressionsdaten (Transkriptom) und experimentelle Daten zur funktionellen Charakterisierung von Mitgliedern einer Genfamilie. Dazu gehören beispielsweise Protein-Protein-Interaktionsstudien, Analyse von partiellen oder vollständigen Verlustmutanten bzw. transgenen Pflanzen. Deshalb wurden in dieser Arbeit mehrere Datenbanken entwickelt, die diese verschiedenen Bereiche abdecken.

Die erstellten Datenbanken *GenAgent*, *GenomeDB* und *TF-Workbench* haben gemeinsam das Ziel externe und interne biologische Daten nicht redundant zu erfassen und die einzelnen Datenobjekte logisch zu verknüpfen. Zudem wurde die Erfassung externer Daten weitgehend automatisiert, um der ständig in Veränderung befindlichen Informationslage gerecht zu werden.

Das Programm *GenAgent* wurde für die Verarbeitung und Annotation von intern erstellten EST-Sequenzen und externen Sequenzdaten entwickelt. Die Sequenzannotationen können mit den Ergebnissen aus Expressionsexperimenten verknüpft werden. Die *GenomeDB* wird nicht über eine Benutzeroberfläche angesprochen, sondern dient den Analyseprogrammen als Datenbasis. Erfaßt werden Sequenzen aus den verschiedenen Genomprojekten und deren Annotation. Die *TF-Workbench* ist nicht nur eine Datenbank, sondern ein System aus Datenbank, Benutzeroberfläche und Analysewerkzeugen für die Bearbeitung von Genfamilien.

6.4.1.1 GenAgent

Das Programm *GenAgent* ist ein Datenbank gestütztes System für die "high throughput" Analyse von DNA Sequenzen. Es besteht aus einer relationalen Datenbank auf der Basis des DBMS `MySQL` und mehreren PERL Skripten, die Sequenzvergleiche und Annotation automatisch durchführen, sowie einer HTML-basierten Benutzeroberfläche. Mit dem Entwurf der Datenbankstruktur wurde eine generische Erfassung von Sequenzen und Sequenzvergleichsergebnissen implementiert. Durch diesen Ansatz ist eine Erweiterung auf neue Programme für die Sequenzanalyse vorgesehen.

Für die vergleichende Charakterisierung der R2R3-MYB-Subfamilie wurden *Arabidopsis thaliana* und *Oryza sativa* EST-Sequenzen aus der *GenomeDB* in den GenAgent geladen, einem "clustering" mit `STACKPACK` unterzogen und mit Hilfe des *GenAgent* annotiert. Die EST-Sequenzen wurden bei der Verifikation von Genmodellen benutzt.

Durch die Nutzung eines relationalen DBMS sind Verknüpfungen mit weiteren Datenbanken problemlos möglich. Tatsächlich wurde das *GenAgent* Programm erfolgreich in mehreren Projekten eingesetzt, bei denen die Funktionalität für die spezifischen Anforderungen des Projektes erweitert wurde.

Für ein Projekt zur Charakterisierung von vorwiegend in der Wurzel exprimierten Genen aus *Beta vulgaris* (Bellin D, Werber M et al. 2002) 3000 EST-Sequenzen mit dem GenAgent annotiert und durch "clustering" mit `StackPack` (Miller, Christoffels et al. 1999) ein nicht redundanter Satz von 2048 Sequenzen bestimmt. Macroarrays wurden mit Proben der amplifizierten c-DNA bedruckt und mit drei verschiedenen ^{33}P markierten Proben aus Blättern, Blüten und der Rübe hybridisert. Die Hybridisationssignale wurden mit einem Phosphorimager (STORM, Amersham

Biosciences, Heidelberg, Deutschland) gemessen und mit dem Programm Arrayvision (Imaging Research Inc., Haverhill, UK) ausgewertet. Diese Auswertung wurden mit Hilfe von PERL Skripten mit den Annotationen im *GenAgent* verknüpft und so einer integrierten Auswertung zugänglich gemacht. Um die EST-Sequenzen funktionell zu klassifizieren wurden die Annotationen um eine INTERPROSCAN Analyse (Apweiler, Attwood et al. 2001) erweitert. Die Verknüpfung der Ergebnisse dieser Analyse mit Gen-Ontologien (Ashburner, Ball et al. 2000) ermöglichte eine Einteilung in funktionale Klassen. Auf diese Weise konnte ein Gesamtbild über die Funktionen des untersuchten EST-Sequenzdatensatzes erstellt werden und mehrere Klassen von überwiegend in den Wurzeln exprimierten Genen identifiziert werden. Für ein weiteres Projekt, das auf denselben EST-Sequenzen basiert, wurden die Suchfunktionen des GenAgent genutzt, um die in der relationalen Datenbank abgelegten Annotationen zu durchsuchen. Auf diese Weise konnten 29 „disease-resistance genes“ (R genes) identifiziert und durch „mapping“ und Bestimmung der Transkriptionsstärke weitere charakterisiert werden (Hunger, Di Gaspero et al. 2003).

6.4.1.2 GenomeDB

Die Datenbank *GENOMEDB* wurde entwickelt um Sequenzdaten aus den verschiedenen Genomprojekten nicht-redundant in einer Resource zu integrieren.
Hierdurch wird gewährleistet, dass die verschiedenen Programme (*FamilyBuilder*, MCS, BLAST), die für die Charakterisierung der R2R3-MYB-Subfamilie verwendet wurden, über eine einheitliche Schnittstelle auf die gleichen Daten zugreifen. Mit Hilfe von Perl Skripten werden die Daten fortwährend aktualisiert. Durch einfaches Starten der Analyseprogramme können so neue Informationen in die Berechnungen einbezogen werden. Diese Funktionalität hat sich insbesondere bei der Arbeit mit den Genomsequenzen aus *Oryza sativa* bewährt, da die Ausgangsdaten dieses Genomprojektes bis heute ständigen Änderungen unterliegen.

6.4.1.3 TF-Workbench und TF-Cards

Das Programm *TF-WORKBENCH* wurde im Rahmen dieser Arbeit als eine integrierte Annotationsumgebung für die Analyse und Verwaltung von Daten zu Genfamilienmitgliedern entwickelt. Es besteht aus einer Datenbank und PERL Skripten, die im Hintergrund Analyse- und Verwaltungsaufgaben übernehmen, sowie einer HTML Benutzeroberfläche, die von PHP Skripten dynamisch generiert wird.

Während der Bearbeitung der R2R3-MYB-Subfamilie wurden mit Hilfe der *TF-Workbench* Genstrukturen annotiert, in der Literatur beschriebene Funktionen erfaßt und die Zusammenarbeit innerhalb der Arbeitsgruppe zwischen der Bioinformatik einerseits und der Biologie andererseits organisiert. Alle Genfamilienmitglieder wurden nach aktuellem Stand über die Bentzeoberfläche in die Datenbank eingetragen und stehen so anderen Benutzern zur Verfügung. Insbesondere für externe Nutzer ist die Funktion der *TF-Cards* von Interesse, da hier auf einer HTML-Seite alle wichtigen Informationen zu einem Gen zusammengefasst werden. Die Zusammenarbeit mit externen Kooperationspartnern wird so erleichtert.
Ähnlich wie bei der *GenomeDB* steht auch bei der *TF-Workbench* der Grundgedanke im Vordergrund, aktuelle Informationen als eine Referenz für verschiedene Nutzungen zur Verfügung zu stellen. Bei der *TF-Workbench* kann die Nutzung dabei einerseits durch Programme erfolgen, die auf den bereitgestellten Daten Berechnungen durchführen, andererseits können Wissenschaftler die *TF-Workbench* nutzen, um ihre Daten zu organisieren und sich über einzelne Gene detailliert zu informieren.
Neben der in dieser Arbeit behandelten R2R3-MYB-Faktor Genfamilie sind in der *TF-Workbench* wurden in der *TF-Workbench* weitere Genfamilien erfasst. Für ein Projekt zur strukturellen Charakterisierung der bHLH-Transkriptionsfaktorfamilie in *Arabidopsis thaliana* wurde die *TF-Workbench* bei der Annotation der Genstrukturen und bei der Erfassung von Textannotationen eingesetzt. Die Ergebnisse dieser Arbeiten sind zusammen mit Expressionsstudien zu den Genfamilienmitgliedern veröffentlicht worden (Heim, Jakoby et al. 2003).

6.4.2 Programme für die automatisierte Identifizierung und Klassifizierung von Genfamilienmitgliedern

In dieser Arbeit wurden mehrere Programme entwickelt, die alle dazu dienen, die Annotation und Analyse der R2R3-MYB-Subfamilie zu unterstützen. Um die Aufgabe der vergleichenden Charakterisierung der R2R3-MYB-Genfamilie zu bewältigen, wurde mit dem *FamilyBuilder* ein Programm erstellt, das dem Zustand der Sequenzdaten und ihrer Annotation aus den Genomprojekten Rechnung trägt, und die einzelnen Schritte der Identifizierung und Klassifizierung in einem automatisierten Verfahren zusammenfasst.

Mit den Programmen “motif cooccurrence scan” (MSCS) und “aminoacid frequency plot” (AFP) wurden Werkzeuge entwickelt, mit denen neue Wege für die Suche nach Genen und die vergleichende Charakterisierung von Sequenzen beschritten werden.

6.4.2.1 FamilyBuilder

FamilyBuilder ist ein Programm für die automatisierte Identifizierung und Klassifizierung von Genfamilienmitgliedern in genomischen Sequenzdaten. Der Unterschied zu anderen interaktiven Ansätzen, wie PSI-Blast oder ähnlichen Suchwerkzeugen ist hierbei, dass der Benutzer nicht auf den Annotationsfortschritt innerhalb des Genomprojektes angewiesen ist, sondern direkt mit den genomischen Sequenzen arbeiten kann. Dieser Punkt war für die vorliegende Arbeit wichtig, da die genomweiten Annotationen für die Sequenzierprojekte von *Oryza sativa* nicht vollständig sind und auch nicht in absehbarer Zeit abgeschlossen sein werden. Um daher die genomischen Sequenzdaten direkt nutzen zu können, musste der Schritt der Genstrukturvorhersage in den automatischen Programmablauf integriert werden. Die wesentlichen Abläufe des Programms FamilyBuilder sind daher die Suche, die Genvorhersage und die abschließende Klassifizierung der Proteinsequenz.
Bei der Suche nach genomischen Sequenzabschnitten, die für R2R3-MYB-Gene kodieren, wurde mit dem Suchprogramm MSCS (siehe 6.4.2.2) eine neue Methode entwickelt, um Gene spezifisch und sensitiv in genomischen Sequenzen zu identifizieren.
Für die Genstrukturvorhersage wurde mit dem Konzept der Kombination von Ergebnissen mehrerer Programme eine Strategie gewählt, die derzeit als leistungsfähigste Lösung beschrieben wird. Die Entscheidungen für das richtige Genmodell wurden homologie-basiert und gestützt durch Vergleiche mit EST-Sequenzen getroffen und haben sich in der konkreten Anwendung bei der R2R3-MYB-Subfamilie als erfolgreich erwiesen. Ein möglicher Nachteil dieser Vorgehensweise ist jedoch, dass Gene ohne typische Motivstruktur aufgrund des Vergleichs mit homologen Proteinsequenzen falsch annotiert werden könnten. Für die Annotation der MYB-Gene, die sich durch hoch konservierte Domänen auszeichnen, war dies aber nicht der Fall. Für die Funktionalität des *FamilyBuilder* war wichtig, dass die Genauigkeit der Genvorhersage ausreicht, um aus einem Abschnitt genomischer DNA einen Bereich bestimmen zu können, dessen gespleißte und translatierte Sequenz dem Klassifizierungsschritt zugeführt werden kann. Die so

gewonnenen Informationen über die Genstruktur wurden für die weiteren Analysen in dieser Arbeit manuell überprüft. Die Überprüfung wurde durch die Visualisierungswerkzeuge in der *TF-Workbench* unterstützt. Zehn Kandidaten aus dem *Oryza sativa* Sequenzdatensatz mussten manuell korrigiert werden.

6.4.2.2 MSCS

Das Programm *MSCS* wurde entwickelt, um mit hoher Spezifität und Sensitivität nach Genen einer Familie in einem nicht annotierten genomischen Sequenzdatensatz zu suchen. Die Auswahl von Motiven mit hoher Signifikanz für die gesuchte Genfamilie gewährleistet dabei die Spezifität. Das Aufteilen der Motivsequenzen in kurze Fragmente, die ähnlich wie Primer an viele Stellen binden können, erhöht die Sensitivität. Des Weiteren hat sich dieses Suchwerkzeug in der Praxis vor allem durch die Möglichkeit der Visualisierung der Suchergebnisse bewährt. So können Mitglieder einer Genfamilie selbst in 100 kb langen Sequenzen leicht identifiziert werden. Gleichzeitig sieht der Benutzer, welche konservierten Motive in diesem Gen vorhanden sind, und kann so eine erste Einschätzung geben, ob eine nähere Untersuchung sinnvoll ist. Für die Charakterisierung der R2R3-MYB-Genfamilie war hierbei von Vorteil, dass Sequenzen, die zwar eine MYB-ähnliche Domäne aufweisen, nicht aber der R2R3-MYB-Subfamilie zuzurechnen sind, einfach identifiziert werden konnten. Das Programm eignet sich somit sowohl für die automatisierte Suche von Genen in genomischen Sequenzen, als auch als Kontrollwerkzeug für die schnelle Validierung von Ergebnissen.

6.4.2.3 AFP

Das Programm „Aminoacid Frequency Plot“ (*AFP*) ist ein kleines Visualisierungswerkzeug, das für die Arbeit mit verschiedenen Genfamilien entwickelt wurde. Häufig wird die Konservierung von Aminosäuren in einem multiplen Alignment mit „Sequence Logos“ (Schneider and Stephens 1990) dargestellt. Diese Art der Visualisierung hat den Vorteil, dass stark konservierte Aminosäuren leicht erkennbar sind. Visualisierung ist jedoch auch immer eine subjektive Entscheidung und hängt von den Präferenzen des Betrachters ab. So ist ein Nachteil der „Sequence Logos“ in diesem Zusammenhang das unübersichtliche Gesamtbild, das sich dem Betrachter bietet. Mit Hilfe von AFP sollten deshalb sowohl die Stärke der Konservierung

einzelner Aminosäuren, als auch das Gesamtbild der Konservierung über die Länge der Sequenz übersichtlich darstellt werden.

6.5 Vergleichende Analyse der R2R3-MYB-Subfamilie in *Arabidopsis thaliana* und *Oryza sativa*

Zu Begin der Arbeit waren in *Arabidopsis thaliana* 102 R2R3-MYB-Gene beschrieben worden (Kranz, Denekamp et al. 1998; Romero, Fuertes et al. 1998). Mit Hilfe der in dieser Arbeit entwickelten Programme konnten zusätzlich weitere 24 R2R3-MYB-Gene, drei 3R-MYB-Gene und ein 4R-MYB-Gen beschrieben werden. Detaillierte Informationen sind in
Tabelle 1 aufgeführt.
In (Riechmann, Heard et al. 2000) wird eine Angabe von 131 R2R3-MYB-Genen gemacht. Obwohl die genaue Vorgehensweise bei der Suche, die zu diesen Angaben führte, einer Überprüfung nicht zur Verfügung stand, kann nach nunmehr drei Jahren intensiver Bearbeitung der Genfamilie auf der Basis der Genomsequenz von *Arabidopsis thaliana* davon ausgegangen werden, dass derartige Unterschiede in der Beschreibung der Anzahl der Gene nicht auf Fehler bei der Suche zurückzuführen sind. Vielmehr hängt die Anzahl der identifizierten Gene unmittelbar mit der Stringenz der Suchbedingungen zusammen. In der vorliegenden Arbeit wurde die R2R3-MYB-Subfamilie sehr stringent nach dem Vorhandensein von zu der zweiten und dritten Sequenzwiederholung von c-MYB homologen Sequenzwiederholungen definiert. Hierfür sind mit *MEME* die Proteinsequenzen verschiedener Arten von MYB-Genen analysiert worden. Neben R2R3-Subfamilie wurden andere Subfamilien beschrieben, bei denen zwei Sequenzwiederholungen vorliegen, die jedoch entweder beide zu der zweiten Sequenzwiederholung von c-MYB oder zur ersten und zweiten Sequenzwiederholung homolog sind. Die in der Analyse von (Riechmann, Heard et al. 2000) verwendeten Interpro MYB Domänen Signaturen identifizieren, wie auch in ergänzenden Materialien der Veröffentlichung angegeben ist, Falschpositive. Inzwischen ist die Sequenzierung des Genoms von *Arabidopsis thaliana* bis auf wenige Korrekturen abgeschlossen, so dass davon ausgegangen werden kann, dass mit dieser Analyse alle MYB-Gene mit mehr als zwei Sequenzwiederholungen in diesem Organismus beschrieben werden konnten.

6.5.1 Analyse der R2R3-MYB-Domäne in *Arabidopsis thaliana* und *Oryza sativa*

Die Domäne der R2R3-MYB-Proteine besteht aus zwei Sequenzwiederholungen mit Tryptophanresten im Abstand von 18 bis 20 Aminosäuren. Um die Domäne zu analysieren, wurde aus einem Alignment über alle 126 R2R3-MYB-Proteinsequenzen ein Balkendiagramm mit dem in dieser Arbeit entwickelten Programm *AFP* berechnet. Dabei sind sowohl in der zweiten als auch in der dritten Sequenzwiederholung nicht nur die Tryptophanreste hoch konserviert, sondern auch weitere Positionen. Unterhalb der Aminosäurehäufigkeiten ist die Anzahl der verschiedenen Aminosäuren an der jeweiligen Position innerhalb der Sequenz in Prozent aufgetragen. Augenfällig ist die geringe Variabilität in der dritten Helix von R3. An der achten Position in R3 ist an Stelle des Tryptophans ein Phenylalanin hochkonserviert. Anstelle des Tryptophans kann aber auch Isoleucin (13%) oder Tryptophan (7%) auftreten. Insgesamt ist die dritte Sequenzwiederholung etwas stärker konserviert als die zweite Sequenzwiederholung. Die MYB-Domäne der untersuchten 120 Sequenzen aus *Oryza sativa* weist über die gesamte Sequenz eine hohe Ähnlichkeit zu der Domäne der untersuchten *Arabidopsis thaliana* Proteinsequenzen auf. Unterschiede bestehen in den weniger stark konservierten Bereichen zwischen den Helices. An schwächer konservierten Bereichen sind zum Teil die häufigsten mit den zweithäufigsten Aminosäuren vertauscht.

6.5.2 C-terminale Motive in *Arabidopsis thaliana* und *Oryza sativa*

In (Kranz, Denekamp et al. 1998) wurden erstmals die C-Terminalen Bereiche der R2R3-MYB-Proteinsequenzen umfassend analysiert. Obwohl diese Sequenzabschnitte insgesamt kaum Ähnlichkeiten zueinander aufweisen, konnten kurze Motive identifiziert werden, die innerhalb von Subgruppen, die vorher mit Clustering nach der „neighbor joining“ Methode bestimmt worden waren, auftreten. Da in der Analyse von (Kranz, Denekamp et al. 1998) noch nicht alle R2R3-MYB-Gene berücksichtigt werden konnten, wurden in dieser Arbeit die Analysen wiederholt. Dabei konnten zwei neue Subgruppen (23,25) beschrieben werden. Für sechs Gruppen konnten keine Motive identifiziert werden.

Für *Oryza sativa* waren als Ausgangssituation sechs R2R3-MYB-Gene beschrieben worden (Suzuki, Suzuki et al. 1997) (Lee, Qi et al. 2001). Mit Hilfe der in dieser Arbeit

entwickelten Programme konnten zusätzlich weitere 114 R2R3-MYB-Gene, 5 3R-MYB-Gene und 1 4R-MYB-Gen beschrieben werden. Detaillierte Informationen sind in Tabelle 3 aufgeführt. In (Dias, Braun et al. 2003) wird eine Abschätzung von 200 R2R3-MYB-Genen für Mais und verwandte Monokotyledonen angegeben. Damit wäre die in dieser Arbeit festgestellte Anzahl deutlich niedriger. Obwohl das IRGSP noch nicht vollständig abgeschlossen ist, kann nicht davon ausgegangen werden, dass in der verbleibenden Sequenz noch 100 R2R3-MYB-Gene existieren. In der Veröffentlichung des *Oryza sativa* Genomprojektes von Syngenta (Goff, Ricke et al. 2002) werden erste genomweite Genfamilien-Analysen durch Sequenzvergleiche vorgenommen. Bei diesen Vergleichen ergibt sich eine Zahl von 156 Genen für die MYB-Superfamilie. Wenn von diesen 156 Genen ein Teil noch Proteine mit nur einer Sequenzwiederholung kodiert, dann ist die in dieser Arbeit ermittelte Anzahl von R2R3-MYB-Genen vergleichbar hoch.

6.5.3 Ergebnisse der vergleichenden Analyse

In der Clustering-Analyse der Proteinsequenzen aus beiden Pflanzen sind die meisten Gruppen (20) mit Genen aus *Arabidopsis thaliana* und *Oryza sativa belegt*. Lediglich vier Gruppen enthalten nur Gene aus *Arabidopsis thaliana.*

Es konnte gezeigt werden, dass Subgruppen von in der Proteinsequenz ähnlichen MYB-Genen, zwischen *Arabidopsis thaliana* und *Oryza sativa* konserviert sind. Die Analyse der C-terminalen Bereiche von Proteinsequenzen aus beiden Arten hat gezeigt, dass sowohl in *Arabidopsis thaliana* als auch in *Oryza sativa* kurze konservierte Sequenzabschnitte vorhanden sind. Die beschriebenen Motive korrelieren dabei mit der Einteilung in Subgruppen, die mit der Cluster-Analyse berechnet wurden. In einer kürzlich erschienenen Publikation wird unter Verwendung der Substitutionsanalyse für die Domäne und die C-terminale Bereiche von R2R3-MYB Proteinsequenzen aus verschiedenen Pflanzen ein leichter Selektionsdruck auf die Bereiche der C-terminalen Motive festgestellt (Dias, Braun et al. 2003). Die Konservierung der C-terminalen Motive in den Subgruppen - in einem ansonsten kaum konservierten Sequenzabschnitt - kann als Hinweis für die Existenz dieser Motive vor der Entstehung der monokotyledonen und der dikotyledonen Pflanzen gesehen werden.

.

6.6 Ausblick

In der vorliegenden Arbeit wurden Programme erstellt, mit denen Genfamilien in kurzer Zeit in bisher nicht- oder nur teilweise annotierten genomischen Sequenzen identifiziert und beschrieben werden können. Damit steht nun auch ein Instrumentarium für die Analyse weiterer Genfamilien bereit. Die Identifizierung der R2R3-MYB-Gene aus *Arabidopsis thaliana* und der R2R3-MYB-Gene aus *Oryza sativa* bildet eine Grundlage für weitere bioinformatische Analysen auf Sequenzebene. Zusätzlich kann nun mit der funktionellen Charakterisierung der R2R3-MYB-Gene aus *Oryza sativa* begonnen werden. Aufgrund der in dieser Arbeit vorgelegten Beschreibungen der MYB-Gene können beispielsweise Mutanten-Populationen durchsucht werden.

Ein weiterer Ansatz beruht auf den ähnlichen Funktion von Genen innerhalb der Subgruppen der R2R3-MYB-Genfamilie in *Arabidopsis thaliana*. (Oppenheimer, Herman et al. 1991; Lee and Schiefelbein 1999). Seit Begin dieser Arbeit sind weitere Funktionen von R2R3-MYB-Genen in *Arabidopsis thaliana* publiziert worden. AtMYB30 ist an der Abwehr von Pathogenen durch die positive Regulation des Zelltods beteiligt (Vailleau, Daniel et al. 2002). AtMYB2 an der Reaktion auf Trockenstress (Abe, Urao et al. 2003). In verschieden Publikationen konnten ähnliche Funktionen von Genen, die einer Subgruppe angehören nachgewiesen werden (Aharoni, De Vos et al. 2001; Gocal, Sheldon et al. 2001). Die für *Arabidopsis thaliana* beschriebenen Funktionen können wertvolle Ansatzpunkte für experimentelle Ansätze in *Oryza sativa* bilden. Aufgrund der Konservierungen kann angenommen werden, dass ähnliche Funktionen auch innerhalb der Subgruppen von *Oryza sativa* vorliegen.

7 Zusammenfassung

Durch die Regulation der Gen-Expression auf der Ebene der Transkription werden in der Zelle viele wichtige Prozesse kontrolliert. Zu der Gruppe von Faktoren, die an der Transkription beteiligt sind, gehören verstärkende oder abschwächende Transkriptionsfaktoren, die an die DNA in sequenzspezifischer Weise binden und so die Frequenz der Transkriptionsinitiierung beeinflussen. MYB-Transkriptionsfaktoren bilden eine Proteinfamilie mit einer aus ein bis vier Sequenzwiederholungen bestehenden konservierten MYB-DNA-Bindedomäne. Im Unterschied zu Tieren und Pilzen besitzen Pflanzen eine stark amplifizierte Subfamilie mit zwei hochkonservierten Sequenzwiederholungen (R2R3). Gene dieser Subfamilie sind an der Regulation einer Vielzahl zellulärer Prozesse beteiligt. Dazu gehören u.a. die Regulation des Phenylpropanoid-Stoffwechsels (Borevitz, Xia et al. 2000), die Kontrolle der Zelldifferenzierung (Oppenheimer, Herman et al. 1991) und die Reaktion auf exogene Reize (Jin and Martin 1999).

Die Veröffentlichung der Genomsequenz von *Arabidopsis thaliana* bietet erstmals die Möglichkeit, die Mitglieder dieser pflanzenspezifischen Subfamilie vollständig zu bestimmen und zu charakterisieren. Durch die Veröffentlichung mehrerer Entwürfe aus verschiedenen Projekten für die Genomsequenz aus *Oryza sativa* und den fast vollständigen Sequenzdaten des „International Rice Genome Sequencing Project“ kann erstmals der Umfang und die Struktur der R2R3-MYB-Genfamilie zwischen zwei Pflanzen analysiert und verglichen werden.

Um die ständig in Veränderung begriffene und anwachsende Menge von nur teilweise annotierten genomischen Sequenzdaten zu durchsuchen, wurden in der vorliegenden Arbeit Algorithmen, Verfahren der Bioinformatik und Datenbankkonzepte entwickelt, um die verschiedenen Arbeitsschritte für die Erfassung und Klassifizierung einer großen Genfamilie zu automatisieren und - wo dies nicht möglich ist - zumindest zu unterstützen.

Auf Basis dieser erarbeiteten Programme und Datenbankkonzepte erfolgte die vergleichende Charakterisierung der R2R3-MYB-Subfamilie in *Arabidopsis thaliana* und *Oryza sativa.*

Als Ergebnis dieser Analysen konnten in *Arabidopsis thaliana* 24 R2R3-MYB-Gene, drei 3R-MYB-Gene und ein 4R-MYB-Gen neu beschrieben und strukturell eingeordnet werden. In *Oryza sativa* konnten 114 R2R3-MYB-Gene, fünf 3R-MYB-

Gene und ein 4R-Gen neu beschrieben und strukturell eingeordnet werden. Die Anzahl der R2R3-MYB-Gene in beiden Organismen und die Einordnung von Genen in Gruppen anhand konservierter Motive im C-terminalen Bereich der Proteinsequenz unterstützen die These, dass die Amplifikation der R2R3-MYB-Genfamilie vor der Aufspaltung der Pflanzen in Dikotyledonen und Monokotyledonen stattfand.

Vergleichende Analysen der C-terminale Bereiche der R2R3-MYB-Proteinsequenzen aus beiden Organismen ergaben, dass diese Motive trotz der Distanz der Organismen zueinander über große Zeiträume der Evolution konserviert worden sind. Dies unterstreicht die These, dass die C-terminalen Motive die Funktion der Proteine beeinflusst. Die Ergebnisse dieser Arbeit konnten vorangegangene Studien ergänzen bzw. korrigieren. Auf Basis einer vollständig charakterisierten Genfamilie in *Arabidopsis thaliana* und einer nahezu vollständig charakterisierten Genfamilie in *Oryza sativa* können Experimente zur funktionellen Charakterisierung geplant werden.

Die Programme und Datenbanken wurden so konzipiert, dass sie sich auch für weitere Fragestellungen nutzen lassen. Die in dieser Arbeit entwickelten Programme und Datenbanken wurden in verschiedenen Projekten erfolgreich eingesetzt.

Ich danke der Norddeutschen Pflanzenzucht Hans-Georg Lembke KG für die finanzielle Förderung meiner Arbeiten und die interessanten Einblicke in die Projekte eines Pflanzenzuchtunternehmens.

Herrn Prof. Dr. Heinz Saedler und Herrn Prof. Dr. Francesco Salamini danke ich für die Möglichkeit, die Arbeiten zu meiner Dissertation am Max-Planck-Institut für Züchtungsforschung durchführen zu können.

Herrn Prof. Dr. Diethard Tautz danke ich für die Übernahme des Koreferates.
Herrn Prof. Dr. Martin Hülskamp danke ich für die Übernahme des Prüfungsvorsitzes.

Mein besonderer Dank gilt Prof. Dr. Bernd Weißhaar für die engagierte Betreuung und intensive Unterstützung dieser Arbeit sowie stets hilfreichen Diskussionen.

Allen ehemaligen und jetzigen Mitarbeitern der Arbeitsgruppe Weißhaar, sowie der ADIS danke ich für die angenehme Arbeitsatmosphäre, die fachliche Unterstützung meiner Arbeit, sowie für schöne Stunden bei gemeinsamen Aktivitäten.

Meinen Eltern, meiner Schwester Ulrike, Maria, Bettina und Hartmut danke ich für Unterstützung, Rat und Spaß.

Besonders Maria danke ich für Geduld, Unterstützung und viel mehr was mir in dieser Zeit sehr geholfen hat.

Elisa ☺ danke ich für Ameisenfütterungen im Sommer 2002 im Sonnenschein.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.
Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Bernd Weißhaar betreut worden.

Teilpublikationen:

The R2R3-MYB-gene family in *Arabidopsis thaliana*
Stracke R, Werber M, Weisshaar B.
Curr Opin Plant Biol. 2001 Oct;4(5):447-56. Review

The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity.
Heim MA, Jakoby M, Werber M, Martin C, Weisshaar B, Bailey PC
Mol Biol Evol 2003 May;20(5):735-47

Isolation and linkage analysis of expressed disease-resistance gene analogues of sugar beet (Beta vulgaris L.).
Hunger S, Di Gaspero G, Mohring S, Bellin D, Schafer-Pregl R, Borchardt DC, Durel CE, Werber M, Weisshaar B, Salamini F, Schneider K.
Genome 2003 Feb;46(1):70-82

EST Sequencing, Annotation and Macroarray Transcriptome Analysis Identify Preferentially Root-Expressed Genes in Sugar Beet
Bellin, D.; Werber, M.; Theis, T.; Schulz, B.; Weisshaar, B.; Schneider, K.
Plant Biology; 06, 2002

Anhang

Tabelle 2: R2R3-, 3R- und 4R-MYB-**Gene aus *Arabidopsis thaliana*. Beschreibung siehe unten.**

Name	Gen Kode	Synonym	GenBank Acc. Nr.	Klon	Referenz
AtMYB0	At3g27920	AtGL1 (Ws)	M79448	K16N12	Oppenheimer *et al.*, 1991
AtMYB1	At3g09230		D10936	F3L24	Shinozaki *et al.*, 1992
AtMYB2	At2g47190		D14712	T8I3	Urao *et al.*, 1993
AtMYB3	At1g22640		AF062859	T22J18	Kranz *et al.*, 1998
AtMYB4	At4g38620		AF062860	F20M13	Kranz *et al.*, 1998
AtMYB5	At3g13540	Atmyb5	U26935 (Ler)	MRP15	Li *et al.*, 1996
AtMYB6	At4g09460		U26936	T15G18	Li *et al.*, 1995
AtMYB7	At2g16720	myb7	U26937	T24I21	Li *et al.*, 1995
AtMYB8	At1g35515		Z95803 (Ler)	F15O4	Romero *et al.*, 1998
AtMYB9	At5g16770		AF062861	F5E19	Kranz *et al.*, 1998
AtMYB10	At3g12820		AF062862	MKB21	Kranz *et al.*, 1998
AtMYB11	At3g62610		AF062863	F26K9	Kranz *et al.*, 1998
AtMYB12	At2g47460		AF062864	T30B22	Kranz *et al.*, 1998
AtMYB13	At1g06180	AtMYBIfgn (L1)	Z50869 (Ler)	F9P14	Kirik *et al.*, 1998
AtMYB14	At2g31180		AF062865	F16D14	Kranz *et al.*, 1998
AtMYB15	At3g23250	AtY19	X90384	K14B15	Quaedvlieg *et al.*, 1996
AtMYB16	At5g15310	AtMIXTA	X99809	F8M21	Rabinowicz *et al.*, 1996
AtMYB17	At3g61250		AF062866	T20K12	Kranz *et al.*, 1998
AtMYB18	At4g25560		AF062867	M7J2	Kranz *et al.*, 1998
AtMYB19	At5g52260		AF062868	F17P19	Kranz *et al.*, 1998
AtMYB20	At1g66230		AF062869	T6J19	Kranz *et al.*, 1998
AtMYB21	At3g27810		AF062870	K16N12	Kranz *et al.*, 1998
AtMYB22	At5g40430		AF175986	MPO12	Stracke *et al.*, 2001
AtMYB23	At5g40330	AtMYBrtf	Z68158 (Ler)	MPO12	Kirik *et al.*, 2001
AtMYB24	At5g40350		AF175987	MPO12	Stracke *et al.*, 2001
AtMYB25	At2g39880		AF175988	T28M21	Stracke *et al.*, 2001
AtMYB26	At3g13890		Z95749 (Ler)	MCD16	Romero *et al.*, 1998
AtMYB27	At3g53200		AF062871	T4D2	Kranz *et al.*, 1998
AtMYB28	At5g61420		Z95751 (Ler)	MFB13	Romero *et al.*, 1998
AtMYB29	At5g07690		AF062872	MBK20	Kranz *et al.*, 1998
AtMYB30	At3g28910		AF062873	MDL15	Kranz *et al.*, 1998
AtMYB31	At1g74650	AtY13	X90383	F1M20	Quaedvlieg *et al.*, 1996
AtMYB32	At4g34990		AF062874	M4E13	Kranz *et al.*, 1998
AtMYB33	At5g06100		AF062875	K16F4	Kranz *et al.*, 1998
AtMYB34	At5g60890	ATR1	U66462		Bender *et al.*, 1998
AtMYB35	At3g28470		AF062877	MFJ20	Kranz *et al.*, 1998
AtMYB36	At5g57620		AF062878	MUA2	Kranz *et al.*, 1998
AtMYB37	At5g23000		AF062879	T20O7	Kranz *et al.*, 1998
AtMYB38	At2g36890		AF062880	T1J8	Kranz *et al.*, 1998
AtMYB39	At4g17780	dl4925c	Z97344 (Ler)	AtFCAcontig	Bevan *et al.*, 1998
AtMYB40	At5g14340		AF062881	F18O12	Kranz *et al.*, 1998
AtMYB41	At4g28110	T13J8.220	AF062882	T13J8	Kranz *et al.*, 1998
AtMYB42	At4g12350	T4C9.190	AF175999	T4C9	Stracke *et al.*, 2001
AtMYB43	At5g16600	MTG13.4	AF175990	MTG13	Stracke *et al.*, 2001
AtMYB44	At5g67300	K8K14.2		K8K14	AGI *et al.*, 2000
AtMYB45	At3g48920		AF062883	T2J13	Kranz *et al.*, 1998
AtMYB46	At5g12870		AF062884	T24H18	Kranz *et al.*, 1998
AtMYB47	At1g18710		AF062885	F6A14	Kranz *et al.*, 1998
AtMYB48	At3g46130		Z95772 (Ler)	F12M12	Romero *et al.*, 1998
AtMYB49	At5g54230		AB010695	MDK4.4	Stracke *et al.*, 2001
AtMYB50	At1g57560		AF062886	F25P12	Kranz *et al.*, 1998
AtMYB51	At1g18570		AF062887	F25I16	Kranz *et al.*, 1998
AtMYB52	At1g17950		AF062888	F2H15	Kranz *et al.*, 1998
AtMYB53	At5g65230		AF062889	MQN23	Kranz *et al.*, 1998

AtMYB54	At1g73410		AF062890	T9L24	Kranz *et al.*, 1998
AtMYB55	At4g01680		AF104919	T15B16.4	Stracke *et al.*, 2001
AtMYB56	At5g17800		AF062891	MVA3.16	Kranz *et al.*, 1998
AtMYB57	At3g01530		AF062892	F4P13	Kranz *et al.*, 1998
AtMYB58	At1g16490		AF062893	F3O9	Kranz *et al.*, 1998
AtMYB59	At5g59780		AF062894	MTH12	Kranz *et al.*, 1998
AtMYB60	At1g08810		AF062895	F22O13	Kranz *et al.*, 1998
AtMYB61	At1g09540		AF062896	F14J9	Kranz *et al.*, 1998
AtMYB62	At1g68320		AF062897	T22E19	Kranz *et al.*, 1998
AtMYB63	At1g79180		AF062898	YUP8H12R	Kranz *et al.*, 1998
AtMYB64	At5g11050		AY032854	T5K6	Stracke *et al.*, 2001
AtMYB65	At3g11440		AF062899	F24K9	Kranz *et al.*, 1998
AtMYB66	At5g14750		AF062900	T9L3	Kranz *et al.*, 1998
AtMYB67	At3g12720	AtY53	X90386	MBK21	Quaedvlieg *et al.*, 1996
AtMYB68	At5g65790		AF062901	MPA24	Kranz *et al.*, 1998
AtMYB69	At4g33450		AF062902	F17M5	Kranz *et al.*, 1998
AtMYB70	At2g23290		AF062903	T20D16	Kranz *et al.*, 1998
AtMYB71	At3g24310		U62743	K7M2	Xia *et al.*, 1996
AtMYB72	At1g56160		AF062905	F14G9	Kranz *et al.*, 1998
AtMYB73	At4g37260		AF062906	C7A10	Kranz *et al.*, 1998
AtMYB74	At4g05100		AF062907	C17L7	Kranz *et al.*, 1998
AtMYB75	At1g56650		AF062908	F25P12	Kranz *et al.*, 1998
AtMYB76	At5g07700		Z95799 (Ler)	MBK20	Romero *et al.*, 1998
AtMYB77	At3g50060	AtMYBr2	Z54137 (Ler)	F3A4	Kirik *et al.*, 1998
AtMYB78	At5g49620		AF062909	K6M13	Kranz *et al.*, 1998
AtMYB79	At4g13480		AF062910	T6G15	Kranz *et al.*, 1998
AtMYB80	At5g56110		Z95804 (Ler)	MDA7	Romero *et al.*, 1998
AtMYB81	At2g26960		AF062911	T20P8	Kranz *et al.*, 1998
AtMYB82	At5g52600		AF062912	F6N7	Kranz *et al.*, 1998
AtMYB83	At3g08500		Z95806 (Ler)	T8G24	Romero *et al.*, 1998
AtMYB84	At3g49690		Y14209 (Ler)	T16K5	Romero *et al.*, 1998
AtMYB85	At4g22680		AF175993	T12H17	Stracke *et al.*, 2001
AtMYB86	At5g26660	AtMYB4	AB005889	F21E10	Noji *et al.*, 1998
AtMYB87	At4g37780		AF062914	T28I19	Kranz *et al.*, 1998
AtMYB88	At2g02820		AF175994	T20F6	Stracke *et al.*, 2001
AtMYB89	At5g39700		AF175995	MIJ24	Stracke *et al.*, 2001
AtMYB90	At1g66390		AF062915	T27F4	Kranz *et al.*, 1998
AtMYB91	At2g37630	AtPHAN	AC004684	F13M22	Timmermans *et al.*, 1999
AtMYB92	At5g10280		AF062916	F18D22	Kranz *et al.*, 1998
AtMYB93	At1g34670		AF062917	F21H2	Kranz *et al.*, 1998
AtMYB94	At3g47600		AF062918	F1P2	Kranz *et al.*, 1998
AtMYB95	At1g74430	AtMYBCP66	AF101048 (Ler)	F1M20	siehe GenBank (direct submission)
AtMYB96	At5g62470	mybcov1	AJ011669	K19B1	Geri *et al.*, 1999
AtMYB97	At4g26930		AF176002	F10M23	Stracke *et al.*, 2001
AtMYB98	At4g18770		AF176003	F28A21	Stracke *et al.*, 2001
AtMYB99	At5g62320	AtMYBCU15	AF101050 (Ler)	MMI9	siehe GenBank (direct submission)
AtMYB100	At2g25230		AF176004	T22F11	Stracke *et al.*, 2001
AtMYB101	At2g32460	AtM1	X90379	T26B15	Quaedvlieg *et al.*, 1996
AtMYB102	At4g21440	AtM4	X90382	F18E5	Quaedvlieg *et al.*, 1996
AtMYB103	At1g63910		AF214116	T12P18	Stracke *et al.*, 2001
AtMYB104	At2g26950		U26934 (Ler)	T20P8	Li *et al.*, 1999
AtMYB105	At1g69560		AF249308	F24J1	Stracke *et al.*, 2001
AtMYB106	At3g01140		AF249309	T4P13	Stracke *et al.*, 2001
AtMYB107	At3g02940		AF249310	F13E7	Stracke *et al.*, 2001
AtMYB108	At3g06490		AF262733	F5E6	Stracke *et al.*, 2001
AtMYB109	At3g55730		AF262734	F1I16	Stracke *et al.*, 2001
AtMYB110	At3g29020		AF272732	K5K13	Stracke *et al.*, 2001
AtMYB111	At5g49330		AF371977	K21P3	Stracke *et al.*, 2001
AtMYB112	At1g48000	T2J15.9	AY008377	T2J15	Stracke *et al.*, 2001
AtMYB113	At1g66370	T27F4.12	AY008378	T27F4	Stracke *et al.*, 2001
AtMYB114	At1g66380	T24F4.13	AY008379	T24F4	Stracke *et al.*, 2001
AtMYB115	At5g40360		AF334814	MPO12	Stracke *et al.*, 2001
AtMYB116	At1g25340		AF334815	F4F7	Stracke *et al.*, 2001

AtMYB117	At1g26780		AF334816	T24P13	Stracke *et al.*, 2001
AtMYB118	At3g27780		AF334817	MGF10	Stracke *et al.*, 2001
AtMYB119	At5g58850		AF371978	K19M22	Stracke *et al.*, 2001
AtMYB120	At5g55020		AF371979	K13P22	Stracke *et al.*, 2001
AtMYB121	At3g30210		AF371980	MIL15	Stracke *et al.*, 2001
AtMYB122	At1g74080		AF371983	F2P9	Stracke *et al.*, 2001
AtMYB123	At5g35550		AF371981	MOK9	Stracke *et al.*, 2001
AtMYB124	At1g14350		AF371982	F14L17	Stracke *et al.*, 2001
AtMYB125	At3g60460		AF469468	rr_c_30	siehe GenBank (direct submission)
AtCDC5	At1g09770		D58424	F21M12	Hirayama *et al.*, 1996
AtMYB4R1	At3g18100		AY033827	MRC8	Stracke *et al.*, 2001
AtMYB3R1	AT4g32730	PC-MYB1	AF151646	F4D11	Braun *et al*, 1999
AtMYB3R2	At4g00540	PC-MYB2	AF151647	F6N23	Braun *et al*, 1999
AtMYB3R3	AT3g09370		AF214117	F3L24	Stracke *et al.*, 2001
AtMYB3R4	AT5g11510	F15N18.100	AF371975	dt_e_23	Stracke *et al.*, 2001
AtMYB3R5	At5g02320	T1E22.80	AF371976	T1E22	Stracke *et al.*, 2001

Die Tabelle enthält alle bislang veröffentlichten und die in dieser Arbeit identifizierten MYB-Gene, deren Proteinsequenz mehr als eine Sequenzwiederholung aufweist. Die Namensgebung basiert auf der in (Kranz, Denekamp et al. 1998; Romero, Fuertes et al. 1998) eingeführten Nomenklatur. Die eindeutigen Identifizierer („unique identifier") wurden aus der MatDB Datenbank des Munich Information Center for Protein Sequences (MIPS) bezogen. Alle Einträge in der Tabelle beziehen sich auf den Ökotyp Columbia, die Ausnahmen sind mit Ler, Landsberg errecta; WS, Wassilewskija) gekennzeichnet.

Tabelle 3: R2R3-, 3R- und 4R-MYB-**Gene in *Oryza sativa.* Beschreibung siehe unten.**

Name	Klon/Bac	Start	Stop
OsMYB001	**AC130600.1**	**81124**	**82006**
OsMYB002	**AL732638.4**	**17441**	**18563**
OsMYB003	**AC113433.7**	**80234**	**83148**
OsMYB004	**NT_036332.1**	**1310527**	**1311447**
OsMYB005	**AC129718.1**	**33568**	**35761**
OsMYB006	**AP005152.1**	**91055**	**92194**
OsMYB007	**AP004766.1**	**8867**	**10921**
OsMYB008	**AL606460.2**	**93406**	**95496**
OsMYB009	**AP004260.2**	**55895**	**57013**
OsMYB010	**NT_036361.1**	**163464**	**164965**
OsMYB011	**AL731616.2**	**34214**	**38136**
OsMYB012	**AC090973.1**	**35782**	**37136**
OsMYB013	**NT_036365.1**	**151343**	**155511**
OsMYB014	**AP000836.1**	**5026**	**6911**
OsMYB015	**BX000509.1**	**91925**	**92993**
OsMYB016	**AP003570.3**	**13**	**2950**
OsMYB017	**AP003242.3**	**113298**	**114412**
OsMYB018	**AP004083.1**	**78200**	**79758**
OsMYB019	**NT_036354.1**	**136091**	**142104**
OsMYB020	**AF377946.3**	**59921**	**61349**
OsMYB021	**AC135794.2**	**125602**	**127824**
OsMYB022	**AC120990.2**	**9024**	**10677**
OsMYB023	**AP003912.2**	**10994**	**11801**
OsMYB024	**AP003990.1**	**109897**	**110802**
OsMYB025	**AC137697.1**	**1191**	**2551**
OsMYB026	**AP003798.1**	**36402**	**37004**
OsMYB027	**AL731784.2**	**138131**	**142857**
OsMYB028	**AP006174.1**	**33987**	**34899**
OsMYB029	**AC136221.1**	**101184**	**102472**
OsMYB030	**AC137267.1**	**94696**	**100511**
OsMYB031	**AP002746.2**	**17192**	**18546**
OsMYB032	**AC137621.1**	**88005**	**89714**
OsMYB033	**AP004006.3**	**30230**	**31792**
OsMYB034	**AP003229.3**	**12645**	**16625**
OsMYB035	**AP004306.3**	**22456**	**26380**
OsMYB036	**NT_036332.1**	**764111**	**765253**
OsMYB037	**AP006265.1**	**49934**	**52358**
OsMYB038	**AL606441.2**	**41886**	**43365**
OsMYB039	**NT_036333.1**	**157982**	**159520**
OsMYB040	**NT_036373.1**	**336005**	**337448**
OsMYB041	**AL772425.4**	**110597**	**111974**
OsMYB042	**AP002912.2**	**118619**	**119509**
OsMYB043	**AP004669.3**	**146965**	**147910**
OsMYB044	**AL662952.2**	**101608**	**107117**
OsMYB045	**NT_036323.1**	**825752**	**827090**
OsMYB046	**AP003245.4**	**86124**	**87253**
OsMYB047	**AP005092.1**	**121037**	**124843**
OsMYB048	**AP005070.1**	**34921**	**35815**
OsMYB049	**AL713900.3**	**12368**	**15343**
OsMYB050	**AP003442.1**	**146640**	**150153**
OsMYB051	**AP005495.1**	**84244**	**85493**

OsMYB052	**AP002883.2**	**35906**	**40503**
OsMYB053	**AP004786.1**	**11180**	**12297**
OsMYB054	**AP005321.1**	**77747**	**78755**
OsMYB055	**AP005821.1**	**88266**	**89222**
OsMYB056	**AP005200.1**	**56526**	**57104**
OsMYB057	**AC134230.2**	**7960**	**9429**
OsMYB058	**AC118288.1**	**26782**	**28149**
OsMYB059	**AC129717.1**	**155199**	**156263**
OsMYB060	**AC135190.1**	**94881**	**97275**
OsMYB061	**AL928753.3**	**7550**	**12224**
OsMYB062	**AP004057.1**	**47976**	**51962**
OsMYB063	**AP005159.1**	**111941**	**113129**
OsMYB064	**AL663006.2**	**51301**	**53356**
OsMYB065	**AP004231.3**	**3034**	**8160**
OsMYB066	**AL662963.2**	**46176**	**49493**
OsMYB067	**AP001552.1**	**122248**	**123443**
OsMYB068	**AP004567.1**	**31742**	**32744**
OsMYB069	**AP005197.2**	**54155**	**56300**
OsMYB070	**AP004376.1**	**66375**	**68239**
OsMYB071	**AL662969.2**	**103719**	**105631**
OsMYB072	**AC136956.3**	**82724**	**84815**
OsMYB073	**AP003754.3**	**124106**	**125885**
OsMYB074	**AP004788.1**	**38477**	**40661**
OsMYB075	**AB026295.2**	**113277**	**115577**
OsMYB076	**AC108504.1**	**42448**	**44028**
OsMYB077	**AC092075.6**	**149132**	**149902**
OsMYB078	**AC087220.9**	**103625**	**105029**
OsMYB079	**AL663006.2**	**72996**	**74394**
OsMYB080	**AC118670.2**	**46944**	**50715**
OsMYB081	**AP003573.1**	**114453**	**115388**
OsMYB082	**AP005686.1**	**108608**	**109949**
OsMYB083	**AP003607.3**	**33239**	**34037**
OsMYB084	**AP004461.3**	**115702**	**117873**
OsMYB085	**AL731579.2**	**10536**	**12674**
OsMYB086	**AC079888.9**	**54034**	**60186**
OsMYB087	**AC133450.4**	**30763**	**32021**
OsMYB088	**AP003488.1**	**167648**	**168422**
OsMYB089	**AP003723.1**	**41731**	**45038**
OsMYB090	**AP005113.1**	**48109**	**48955**
OsMYB091	**AP005319.1**	**116404**	**117401**
OsMYB092	**AC137696.2**	**39620**	**40348**
OsMYB093	**AP005195.1**	**134521**	**135870**
OsMYB094	**AC134045.1**	**147859**	**148779**
OsMYB095	**AL732638.4**	**57188**	**59271**
OsMYB096	**AL607104.1**	**60648**	**63341**
OsMYB097	**AP005012.1**	**38768**	**40613**
OsMYB098	**AP004666.2**	**78286**	**80193**
OsMYB099	**AP004788.1**	**56240**	**58925**
OsMYB100	**NT_036352.1**	**962033**	**963074**
OsMYB101	**AC098833.1**	**60694**	**63941**
OsMYB102	**AP003926.1**	**63556**	**64721**
OsMYB103	**AP003228.3**	**73922**	**74674**
OsMYB104	**AL663000.2**	**125246**	**126989**
OsMYB105	**AP003629.1**	**38637**	**41894**
OsMYB106	**AP003816.2**	**28642**	**29615**
OsMYB107	**AP005527.1**	**132129**	**133122**
OsMYB108	**P0481H08**	**35190**	**42970**

OsMYB109	**P0413H11**	**25912**	**28154**
OsMYB110	**OSJNBb0015B15**	**54284**	**55574**
OsMYB111	**OSJNBa0056E15**	**76162**	**77191**
OsMYB112	**OSJNBa0026J14**	**3034**	**4853**
OsMYB113	**OSJNBa0010K22**	**14271**	**15709**
OsMYB114	**OSJNBa0019D06**	**159407**	**160578**
OsMYB115	**AC037425.7**	**37868**	**40063**
OsMYB116	**AP005197**	**51034**	**51472**
OsMYB117	**AC123523**	**66358**	**69180**
OsMYB118	**AP003737**	**65819**	**67293**
OsMYB119	**AP003849**	**28856**	**29440**
OsMYB120	**AP004867**	**69064**	**70548**
OsMYB3R1	**AP002912.2**	**106538**	**107736**
OsMYB3R2	**NT_036358.1**	**103701**	**108638**
OsMYB3R3	**AL731872.2**	**98490**	**107045**
OsMYB3R4	**NT_036342.1**	**1160739**	**1165282**
OsMYB3R5	**AC105770.1**	**14695**	**22917**
OsMYB4R1	**AL607095.1**	**143082**	**149495**

Die Tabelle enthält alle in dieser Arbeit identifizierten MYB-Gene aus *Oryza sativa*, deren Proteinsequenz mehr als eine Sequenzwiederholung aufweist. Die Namensgebung basiert auf der in (Kranz, Denekamp et al. 1998; Romero, Fuertes et al. 1998) eingeführten Nomenklatur wobei für den Organismus der eingeführte Bezeichner „Os“ für *Oryza sativa* verwendet wurde. Ein- und zweistelligen Nummern wurden mit Nullen auf drei Stellen aufgefüllt. OsMYB1 bis OsMYB5 (Suzuki, Suzuki et al. 1997). OsMYB7 (Locatelli, Bracale et al. 2000).

Literaturverzeichnis

Abe, H., T. Urao, et al. (2003). "Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) Function as Transcriptional Activators in Abscisic Acid Signaling." PLANT CELL **15**(1): 63-4629.

Aharoni, A., C. H. R. De Vos, et al. (2001). "The strawberry FaMYB1 transcription factor suppresses anthocyanin and flavonol accumulation in transgenic tobacco." The Plant Journal **28**(3): 319-332.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**: 403-410.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-402.

Apweiler, R., T. K. Attwood, et al. (2001). "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." Nucleic Acids Research **29**(1): 37-40.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Bailey, T. L. and M. Gribskov (1998). "Methods and statistics for combining motif match scores." J Comput Biol **5**(2): 211-21.

Baldwin, D. A. and W. B. Gurley (1996). "Isolation and characterization of cDNAs encoding transcription factor IIB from Arabidopsis and soybean." Plant J **10**(3): 561-8.

Bellin D, Werber M, et al. (2002). "EST Sequencing, Annotation and Macroarray Transcriptome Analysis Identify Preferentially Root-Expressed Genes in Sugar Beet." Plant Biology **4**(6): 700-710.

Besemer, J. and M. Borodovsky (1999). "Heuristic approach to deriving models for gene finding." Nucl. Acids. Res. **27**(19): 3911-3920.

Bilaud, T., C. E. Koering, et al. (1996). "The telobox, a Myb-related telomeric DNA binding motiv found in proteins from yeast, plants and human." Nucleic Acids Research **24**(7): 1294-1303.

Borevitz, J. O., Y. Xia, et al. (2000). "Activation Tagging Identifies a Conserved MYB Regulator of Phenylpropanoid Biosynthesis." PLANT CELL **12**(12): 2383-4629.

Boyle, W. J., T. Smeal, et al. (1991). "Activation of protein kinase C decreases phosphorylation of c-Jun at sites that negatively regulate its DNA-binding activity." Cell **64**: 573-584.

Braun, E. L. and E. Grotewold (1999). "Newly discovered plant *c-myb*-like genes rewrite the evolution of the plant *myb* gene family." Plant Physiology **121**: 21-24.

Burke, J., D. Davison, et al. (1999). "d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences." Genome Res. **9**(11): 1135-1142.

Churchill, G. (1989). "Stochastic models for heterogeneous DNA sequences." Bull of Math Biol **51**: 79.

Cone, K. C., F. A. Burr, et al. (1986). "Molecular analysis of the maize anthocyanin regulatory locus *C1*." Proceedings of the National Academy of Sciences of the United States of America **83**: 9631-9635.

Cuff, J., M. Clamp, et al. (1998). "JPred: a consensus secondary structure prediction server." Bioinformatics **14**(10): 892-893.

Daniel, X., C. Lacomme, et al. (1999). "A novel myb oncogene homologue in *Arabidopsis thaliana* related to hypersensitive cell death." The Plant Journal **20**(1): 57-66.
Davenport, R. J. (2001). "RICE GENOME: Syngenta Finishes, Consortium Goes On." Science **291**(5505): 807a-.
Dias, A. P., E. L. Braun, et al. (2003). "Recently Duplicated Maize R2R3 Myb Genes Provide Evidence for Distinct Mechanisms of Evolutionary Divergence after Duplication." Plant Physiology **131**(2): 610-4629.
Dickson, D. (2000). "Royalty-free rice arrives on the web." Nature **406**(6796): 549.
Dini, P. and J. Lipsick (1993). "Oncogenic truncation of the first repeat of c-Myb decreases DNA binding in vitro and in vivo." Mol. Cell. Biol. **13**(12): 7334-507.
Ebneth, A., O. Schweers, et al. (1994). "Biophysical characterization of the c-Myb DNA-binding domain." Biochemistry **33**: 14586-14593.
Eddy, S. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.
Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using *Phred*. II. Error probabilities." Genome Research **8**(3): 186-194.
Ewing, B. and P. Green (1998). "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." Genome Res. **8**(3): 186-194.
Feng, Q., Y. Zhang, et al. (2002). "Sequence and analysis of rice chromosome 4." Nature **420**(6913): 316-20.
Florea, L., G. Hartzell, et al. (1998). "A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence." Genome Res. **8**(9): 967-974.
Frishman, D., M. Mokrejs, et al. (2003). "The PEDANT genome database." Nucl. Acids. Res. **31**(1): 207-211.
Gabrielsen, O. S., A. Sentenac, et al. (1991). "Specific DNA binding by c-Myb: evidence for a double helix-turn-helix-related motif." Science **253**: 1140-1143.
Ghosh, D. (1992). "TFD: the transcription factors database." Nucleic Acids Research **20**(Supplement): 2091-2093.
Gocal, G. F. W., C. C. Sheldon, et al. (2001). "GAMYB-like Genes, Flowering, and Gibberellin Signaling in Arabidopsis." Plant Physiology **127**: 1682-1693.
Goff, S. A., D. Ricke, et al. (2002). "A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica)." Science **296**(5565): 92-100.
Gordon, D., C. Abajian, et al. (1998). "Consed: a graphical tool for sequence finishing." Genome Research **8**(3): 195-202.
Hall, L. N., L. Rossini, et al. (1998). "GOLDEN 2: A novel transcriptional regulator of cellular differentiation in the maize leaf." The Plant Cell **10**: 925-936.
Heim, M. A., M. Jakoby, et al. (2003). "The Basic Helix-Loop-Helix Transcription Factor Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity." Mol Biol Evol **20**(5): 735-747.
Hicke, B. J., R. Rempel, et al. (1995). "Phosphorylation of the *Oxytricha* telomere protein: Possible cell cycle regulation." Nucleic Acids Research **23**(11): 1887-1893.
Hirayama, T. and K. Shinozaki (1996). "A *cdc5+* homolog of a higher plant, *Arabidopsis thaliana*." Proceedings of the National Academy of Sciences of the United States of America **93**(23): 13371-13376.
Hoeren, F. U., R. Dolferus, et al. (1998). "Evidence for a role for AtMYB2 in the induction of the *Arabidopsis thaliana* alcohol dehydrogenase gene (*ADH1*) by low oxygen." Genetics **149**: 479-490.
Hülskamp, M., S. Miséra, et al. (1994). "Genetic dissection of trichome cell development in Arabidopsis." Cell **76**: 555-566.

Hunger, S., G. Di Gaspero, et al. (2003). "Isolation and linkage analysis of expressed disease-resistance gene analogues of sugar beet (Beta vulgaris L.)." Genome **46**(1): 70-82.
Hunter, T. and M. Karin (1992). "The regulation of transcription by phosphorylation." Cell **70**: 375-387.
Inoue, T., W. Shoji, et al. (1999). "MIDA1, an Id-associating protein, has two distinct DNA binding activities that are converted by the association with Id1: a novel function of Id protein." Biochem Biophys Res Commun **266**(1): 147-51.
Ito, M., S. Araki, et al. (2001). "G2/M-Phase-Specific Transcription during the Plant Cell Cycle Is Mediated by c-Myb-Like Transcription Factors." The Plant Cell **13**(8): 1891-1905.
Jin, H. and C. Martin (1999). "Multifunctionality and diversity within the plant *MYB*-gene family." Plant Molecular Biology **41**(5): 577-585.
Kirik, V., K. Kolle, et al. (1998). "Two novel *MYB* homologues with changed expression in late embryogenesis-defective *Arabidopsis* mutants." Plant Molecular Biology **37**(5): 819-827.
Kirik, V., K. Kölle, et al. (1998). "Ectopic expression of a novel *MYB* gene modifies the architecture of the *Arabidopsis* inflorescence." The Plant Journal **13**: 729-742.
Klempnauer, K. H., T. J. Gonda, et al. (1982). "Nucleotide sequence of the retroviral leukemia gene v-*myb* and its cellular progenitor c-*myb*: the architecture of a transduced oncogene." Cell **31**: 453-463.
Klempnauer, K.-H., G. Ramsay, et al. (1983). "The product of the retroviral transforming gene *v-myb* is a truncated version of the protein encoded by the cellular oncogene *c-myb*." Cell **33**: 345-355.
Kranz, H., K. Scholz, et al. (2000). "*c-MYB* oncogene-like genes encoding three MYB repeats occur in all major plant lineages." The Plant Journal **21**(2): 231-235.
Kranz, H. D., M. Denekamp, et al. (1998). "Towards functional characterisation of the members of the *R2R3-MYB* gene family from *Arabidopsis thaliana*." The Plant Journal **16**: 263-276.
Krogh, A. (1994). "Hidden Markov models in computational biology: Application to protein modeling." J. Mol. Biol. **235**: 1501-1531.
Larkin, R. M., G. Hagen, et al. (1999). "Arabidopsis thaliana RNA polymerase II subunits related to yeast and human RPB5." Gene **231**(1-2): 41-7.
Lee, M. M. and J. Schiefelbein (1999). "WEREWOLF, a MYB-related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning." Cell **24**(5): 473-483.
Lee, M. W., M. Qi, et al. (2001). "A novel jasmonic acid-inducible rice myb gene associates with fungal infection and host cell death." Molecular Plant Microbe Interactions **14**(4): 527-535.
Leech, M. J., W. Kammerer, et al. (1993). "Expression of *myb*-related genes in the moss *Physcomitrella patens*." The Plant Journal **3**: 51-61.
Lipsick, J. S. (1996). "One billion years of Myb." Oncogene **13**: 223-235.
Locatelli, F., M. Bracale, et al. (2000). "The Product of the Rice myb7 Unspliced mRNA Dimerizes with the Maize Leucine Zipper Opaque2 and Stimulates Its Activity in a Transient Expression Assay." J. Biol. Chem. **275**(23): 17619-17625.
Martin, C. and J. Paz-Ares (1997). "MYB transcription factors in plants." Trends in Genetics **13**(2): 67-73.

Miller, R. T., A. G. Christoffels, et al. (1999). "A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base." Genome Res **9**(11): 1143-55.
Noda, K.-I., B. J. Glover, et al. (1994). "Flower colour intensity depends on specialized cell shape controlled by a Myb-related transcription factor." Nature **369**: 661-664.
Nomura, N., M. Takahashi, et al. (1988). "Isolation of human cDNA clones of *myb*-related genes, A-*myb* and B-*myb*." Nucleic Acids Research **16**: 11075-11089.
Ogata, K., S. Morikawa, et al. (1994). "Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices." Cell **79**: 639-648.
Oh, I. H. and E. P. Reddy (1999). "The myb gene family in cell growth, differentiation and apoptosis." Oncogene **18**(19): 3017-3333.
Ohi, R., D. McCollum, et al. (1994). "The *Schizosaccharomyces pombe cdc5+* gene encodes an essential protein with homology to c-Myb." The EMBO Journal **13**: 471-483.
Oppenheimer, D. G., P. L. Herman, et al. (1991). "A *myb* gene required for leaf trichome differentiation in Arabidopsis is expressed in stipules." Cell **67**: 483-493.
Orphanides, G., T. Lagrange, et al. (1996). "The general transcription factors of RNA polymerase II." Genes & Development **10**(21): 2657-2683.
Pavy, N., S. Rombauts, et al. (1999). "Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences." Bioinformatics **15**(11): 887-899.
Paz-Ares, J., U. Wienand, et al. (1986). "Molecular cloning of the *c1* locus of *Zea mays*: a locus regulating the anthocyanin pathway." The EMBO Journal **5**: 829-834.
Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proceedings of the National Academy of Sciences of the United States of America **85**(8): 2444-2448.
Pertea, M. and S. L. Salzberg (2002). "Computational gene finding in plants." Plant Mol Biol **48**(1-2): 39-48.
Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proc. IEEE **77**(2): 257-286.
Rabinowicz, P., E. Braun, et al. (1999). "Maize R2R3 *Myb* genes: Sequence analysis reveals amplification in the higher plants." Genetics **153**(1): 427-444.
Rice, P., I. Longden, et al. (2000). "EMBOSS: The European Molecular Biology Open Software Suite." Trends in Genetics **16**(6): 276-277.
Riechmann, J. L., J. Heard, et al. (2000). "Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes." Science **290**(5499): 2105-2110.
Roeder, R. G. (1991). "The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly." Trends in Biochemical Sciences **16**: 402-408.
Romero, I., A. Fuertes, et al. (1998). "More than 80 *R2R3-MYB* regulatory genes in the genome of *Arabidopsis thaliana*." The Plant Journal **14**(3): 273-284.
Sakata, K., Y. Nagamura, et al. (2002). "RiceGAAS: an automated annotation system and database for rice genome sequence." Nucl. Acids. Res. **30**(1): 98-102.
Sakura, H., C. Kanei-Ishii, et al. (1989). "Delineation of three functional domains of the transcriptional activator encoded by the c-myb protooncogene."

Proceedings of the National Academy of Sciences of the United States of America **86**(15): 5758-5762.
Salamov, A. A. and V. V. Solovyev (2000). "Ab initio gene finding in Drosophila genomic DNA." Genome Res **10**(4): 516-22.
Sasaki, M., K. Ogata, et al. (2000). "Backbone dynamics of the c-Myb DNA-binding domain complexed with a specific DNA." J. Biochem. (Tokyo) **127**(6): 945-4629.
Sasaki, T. and B. Burr (2000). "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome." Current Opinion in Plant Biology **3**(2): 138-142.
Sasaki, T. and B. Burr (2000). "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome." Curr Opin Plant Biol **3**(2): 138-41.
Sasaki, T., T. Matsumoto, et al. (2002). "The genome sequence and structure of rice chromosome 1." Nature **420**(6913): 312-6.
Sasaki, T., M. Yano, et al. (1996). "The Japanese Rice Genome Research Program." Genome Res **6**(8): 661-6.
Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-100.
Schoof, H., P. Zaccaria, et al. (2002). "MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome." Nucl. Acids. Res. **30**(1): 91-93.
Singh, K. B. (1998). "Transcriptional regulation in plants: the importance of combinatorial control." Plant Physiol **118**(4): 1111-20.
Stracke, R., M. Werber, et al. (2001). "The R2R3-MYB gene family in Arabidopsis thaliana." Current Opinion in Plant Biology **4**: 447-456.
Suzuki, A., T. Suzuki, et al. (1997). "Cloning and expression of five myb-related genes from rice seed." Gene **198**(1-2): 393-398.
The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*." Nature **408**(6814): 796-815.
Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.
Vailleau, F., X. Daniel, et al. (2002). "A R2R3-MYB gene, AtMYB30, acts as a positive regulator of the hypersensitive cell death program in plants in response to pathogen attack." PNAS **99**(15): 10179-4629.
Waites, R., H. R. N. Selvadurai, et al. (1998). "The *PHANTASTICA* gene encodes a MYB transcription factor involved in growth and dorsoventrality of lateral organs in Antirrhinum." Cell **93**: 779-789.
Weston, K. and J. M. Bishop (1989). "Transcriptional activation by the v-myb oncogene and its cellular progenitor, c-myb." Trends in Genetics **5**: 323.
Wheeler, D. L., D. M. Church, et al. (2003). "Database resources of the National Center for Biotechnology." Nucl. Acids. Res. **31**(1): 28-33.
Wingender, E. (1994). "Recognition of regulatory regions in genomic sequences." Journal of Biotechnology **35**(2-3): 273-280.
Wykoff, D. D., A. R. Grossman, et al. (1999). "Psr1, a nuclear localized protein that regulates phosphorus metabolism in Chlamydomonas." Proc Natl Acad Sci U S A **96**(26): 15336-41.

Yu, E. Y., S. E. Kim, et al. (2000). "Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1." Journal of Biological Chemistry **275**(31): 24208-14.
Yu, J., S. Hu, et al. (2002). "A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica)." Science **296**(5565): 79-92.