Veröffentlichung der Ergebnisse von Forschungsvorhaben im BMBF-Programm

Ausbildungs- und Technologieinitiative Bioinformatik

"Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung, und Forschung unter dem Förderkennzeichen 0312 706A gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren."

Forschungsvorhaben: Verbundprojekt: Bioinformatik Centrum Gatersleben-Halle (IPK): Nachwuchsgruppe Plant Data Warehouse.

Förderkennzeichen: 0312 706A

Zuwendungsempfänger: Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Corrensstr. 3, 06466 Gatersleben

Ausführende Stelle: Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Corrensstr. 3, 06466 Gatersleben

Projektleiter: Prof. Dr. Ivo Grosse und Dr. Uwe Scholz

Laufzeit: 1. 5. 2002 – 31. 10. 2007

1 Übersicht

1.1 Aufgabenstellung

Das Ziel der Teilprojektes *Plant Data Warehouse* bestand in der Entwicklung einer flexiblen Softwareplattform zur Analyse von molekularen, phänotypischen und taxonomischen Daten aus den Bereichen der Pflanzen- und Kulturpflanzenforschung sowie von Daten zu pflanzengenetischen Ressourcen mittels *Data Warehouse* Technologie. Dies beinhaltet zum einen die Integration großer Datenmengen verschiedener Domänen aus IPK- und IPB-internen sowie weltweit verteilten Quellen. Zum anderen beinhaltet es die Integration als auch Entwicklung komplexer Anwendungssoftware zur Analyse und Visualisierung der integrierten Daten.

Der Nutzen des *Plant Data Warehouse* besteht im wesentlichen in drei Punkten: Erstens erlaubt es im Sinne des *Data Mining* die Aufdeckung versteckter Korrelationen in vielschichtigen Datenmengen und die Analyse und Visualisierung der integrierten Daten. Zweitens liefert es einen einfachen und direkten Zugriff auf die integrierten Datensätze und ermöglicht damit Entscheidungshilfen bei der Planung und Durchführung neuer Experimente und Forschungsprojekte. Drittens reduziert es den Aufwand zur Installation und Konfiguration verschiedenster Softwarepakete zur Datenanalyse und Visualisierung auf lokalen Rechnern der Anwender.

Eine wichtige Nebenbedingung bei der Entwicklung des *Plant Data Warehouse* bestand darin, die Auswahl der zu integrierenden Daten, die Integration der Daten, die Auswahl der zu integrierenden Software als auch die Entwicklung der Anwendungssoftware zum einen auf die Bedürfnisse der biologisch arbeitenden Gruppen am IPK Gatersleben und IPB Halle und zum anderen auf die der internationalen wissenschaftlichen Gemeinschaft abzustimmen.

1.2 Wissenschaftliche und technische Ausgangspunkte

Das Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) in Gatersleben sowie das Leibniz-Institut für Pflanzenbiochemie (IPB) in Halle gehören zu den großen, international bedeutsamen Zentren der Pflanzenforschung, in denen Probleme der modernen Biologie vorrangig an Kulturpflanzen bearbeitet werden.

Im Zentrum der grundlagen- und anwendungsorientierten, interdisziplinären Forschung des IPK Gatersleben steht die Erarbeitung neuer Erkenntnisse und Technologien mit dem Ziel einer umfassenden Nutzung pflanzengenetischer Ressourcen für eine optimierte Stoffproduktion und für eine umweltverträglichere Landwirtschaft. Mit der bundeszentralen *ex situ* Genbank verfügt das IPK Gatersleben über eine einzigartige Sammlung pflanzengenetischer Ressourcen aus über 3.000 botanischen Arten von ca. 800 verschiedenen Gattungen mit einem Gesamtbestand von etwa 148.000 Kulturpflanzenmustern.

Das Ziel des IPB Halle ist es, die Funktion der großen Vielfalt chemischer Verbindungen, die Pflanzen und höhere Pilze generieren, mit interdisziplinären Forschungsansätzen aufzuklären. Die vier Abteilungen des IPB Halle verbinden auf einzigartige Weise chemische und molekularbiologische Kompetenz zur Analyse dieser komplexen Systeme. Die gewonnenen Erkenntnisse eröffnen neue Wege für eine innovative und nachhaltige Nutzung in Pflanzenproduktion, Pflanzenschutz, Biotechnologie und Wirkstoffentwicklung. Die Speicherung, Auswertung und Verknüpfung der an diesen Instituten generierten Massendaten ist nur mittels Bioinformatik möglich. Insbesondere die Auswertung der Daten zu pflanzengenetischen Ressourcen mit den Genom-, Transkriptom-, Metabolom- und Proteomdaten erfordert die Entwicklung neuer Methoden der Datenauswertung, -verarbeitung und -verknüpfung. Dafür fehlten im Jahr 2002 an beiden Instituten sowohl die Infrastruktur als auch die Kompetenz

auf dem Gebiert der Bioinformatik.

Die wertvollen, im Zuge teurer Experimente generierten Massendaten wurden auf den lokalen Festplatten einiger Servern und verschiedener Arbeitsplatz-PCs verschiedener Arbeitsgruppen und verschiedener Abteilungen gespeichert. Die Dateiformate wurden von den jeweiligen Bearbeiterinnen und Bearbeitern der Daten gewählt mit dem Resultat, dass selbst Daten einer Datendomäne, z. B. Expressionsdaten, in verschiedenen Dateiformaten vorlagen. Auch die Analyseprogramme wurden von den jeweiligen Bearbeiterinnen und Bearbeitern der Daten gewählt mit dem Ergebnis, dass z. B. Expressionsdaten, die in verschiedenen Teilprojekten generiert wurden, auf verschiedene Weise normiert wurden.

Diese Ausgangssituation im Jahr 2002 machte domänenübergreifende Analysen selbst innerhalb der beiden Institute fast unmöglich. Solche domänenübergreifenden Analysen, insbesondere auch über die Institutsgrenzen hinweg, wurden aber im Zuge der Entwicklung der Biotechnologie essentiell wichtig für die moderne Biologie und Züchtungsforschung. Daraus resultierte die Notwendigkeit, die Daten am IPK Gatersleben und am IPB Halle gemeinsam mit weiteren öffentlich verfügbaren Daten zu integrieren und ein Data Warehouse für Kulturpflanzen zur Analyse dieser Daten zu entwickeln.

1.3 Planung und Ablauf des Vorhabens

Die Entwicklung des *Plant Data Warehouse* gliederte sich global in drei überlappende Phasen:

1. die Erstellung von *Operativsystemen* zur Haltung der am IPK Gatersleben und IPB Halle generierten und gesammelten Primärdaten,

2. die Erstellung der *Data Marts* des *Plant Data Warehouse* zur Integration von Daten aus verschiedenen IPK- und IPB-internen sowie weltweit verteilten Quellen,

3. die Integration von Anwendungssoftware sowie die Entwicklung von Anwendungen und Algorithmen zur Analyse der integrierten Daten.

Phase 2 beinhaltet neben der Erstellung der *Data Marts* auch die Integration der Daten in das *Plant Data Warehouse*. Phase 3 beinhaltet neben der Integration und Entwicklung von Anwendungssoftware auch die Durchführung von Analysen der integrierten Daten. Im Projektantrag wurden die folgenden Arbeitspakete (AP) des Teilprojektes *Plant Data Warehouse* spezifiziert, die gemeinsam durch unseren Projektpartner B.I.M.-Consulting mbH in Magdeburg und durch die Arbeitsgruppe *Plant Data Warehouse* am IPK Gatersleben bearbeitet wurden:

1. Evaluierung und Konsolidierung bestehender Datenbestände und Anforderungsanalyse bezüglich der benötigten Daten (AP1)

2. Design des Datenbankschemas des *Plant Data Warehouse* (AP2)

3. Integration von Daten aus den Operativsystemen und externen Datenquellen und Ermöglichung von Interoperabilität (AP3)

4. Konsistenzüberprüfungen und Fehlerkorrektur (AP4)

5. Anforderungsanalyse bezüglich der benötigten Anwendungen und deren Entwicklung und Integration (AP5)

6. Erstellung des *Plant Bioinformatics Portals* als zentrale Präsentationsplattform des BIC-GH und aller in das *Plant Data Warehouse* integrierten Anwendungen (AP6)

Weiterhin wurden nach dem Besuch der BMBF Evaluierungskommission am 25. Mai 2005 die folgenden fünf Arbeitspakete spezifiziert und zusätzlich bearbeitet:

7. Versionskontrolle (AP7)
8. Kurations-System (AP8)
9. Synchronisationsplan und Dokumentation (AP9)
10. Backup-Strategie (AP10)
11. Wartungsplan (AP11)

Die Evaluierung und Konsolidierung bestehender Datenbestände und die Anforderungsanalyse bezüglich der benötigten Daten (AP1) wurde innerhalb der ersten beiden Projektjahre abgeschlossen. Die für das *Plant Data Warehouse* notwendigen Operativsysteme wurden in Abstimmung mit der Arbeitsgruppe Bioinformatik des IPK Gatersleben entwickelt und sind seitdem in Benutzung. Sie bilden die Voraussetzung für alle folgenden Arbeitspakete, insbesondere für die Integration von Daten in das *Plant Data Warehouse* (AP3).

Das Design des Datenbankschemas des *Plant Data Warehouse* (AP2) wurde gemeinsam mit den Kollegen unseres Industriepartners B.I.M.-Consulting mbH entwickelt. Das Grunddesign wurde mit Ende des dritten Projektjahres abgeschlossen, das endgültige Design inklusive aller Feinheiten, wie z. B. dem Kurations-System (AP7), mit Ende des letzten Projektjahres. Das Datenbankschema des *Plant Data Warehouse* wurde aufbauend auf den Ergebnissen der Evaluierung der Operativsysteme und der Anforderungsanalyse (AP1) entworfen und erstellt.

Die Entwicklung von Modulen zur automatischen und semiautomatischen Konsistenzüberprüfung und Fehlerkorrektur (AP4) wurde mit Ende des letzten Projektjahres abgeschlossen. Die Überprüfung von Akzessionsnummern und anderen Schlüsseln sowie einfachen Konsistenzbedingungen wird über Datenbankfunktionalitäten gesichert.

Die Entwicklung und Integration von Anwendungen zur Datenanalyse und Visualisierung (AP5) stand im Zentrum der zweiten Hälfte des *Plant Data Warehouse* Projektes und wurde mit Ende des letzten Projektjahres beendet. Die frühzeitige Fertigstellung von Prototypen ermöglichte den Nutzern des *Plant Data Warehouse* bereits zwei Jahre vor der Beendigung des Projektes erste Analysen. Aus einige dieser Analysen ergaben sich inzwischen neue Erkenntnisse auf den Gebieten der Pflanzengenetik und Züchtungsforschung.

Das *Plant Bioinformatics Portal* (AP6) wurde als zentrale Präsentationsplattform des BIC-GH und aller in das *Plant Data Warehouse* integrierten Anwendungssoftware im vierten Projektjahr erstellt. Es wurde im Frühjahr 2006 auf den neuen *Application Server* portiert und seitdem pro Monat durch mehr als 1000 Anwender aus mehr als 60 Ländern sowie aus gov-, edu-, com- und net-Domänen genutzt. Weitere Verbesserungen, die vor allem aus der Rückkopplung der Nutzer resultierten, wurden gemäß der Planung kontinuierlich bis zum Projektende eingepflegt.

AP7 bis AP11 wurden im vierten und fünften Projektjahr sowie im Rahmen der kostenneutralen Verlängerung des *Plant Data Warehouse* Projektes bearbeitet. Sie garantieren den robusten Betrieb des *Plant Data Warehouse* über die Förderperiode hinaus. Das Feedback- und Kurationssystem erlaubt die effiziente Wartung des *Plant Data Warehouse* und die kontrollierte Bereinigung der integrierten Daten.

Die Ergebnisse aller Arbeitspakete sind ausführlich in Abschnitt 2.1 dargestellt. Das *Plant Data Warehouse* wurde am 31. Oktober 2007 von der Arbeitsgruppe Bioinformatik des IPK Gatersleben übernommen und wird durch sie weiter gewartet, wodurch eine nachhaltige Nutzung des *Plant Data Warehouse* durch Biologen und Bioinformatiker weltweit auch in Zukunft gesichert ist.

1.4 Wissenschaftlicher und technischer Stand

Viele der für das *Plant Data Warehouse* sowie dessen Nutzer wichtigen Daten lagen in IPK- und IPB-internen sowie weltweit verteilten Quellen vor. Ebenfalls existierten viele der *Plant Data Warehouse* benötigten Analyseprogramme, wie z. B. Blast, Blat, GeneSequer, SIM4 oder Spidey, vor. Des weiteren existierten mehrere Softwarepakete zur komplexen Analyse genotypischer und phänotypischer Daten, deren *Workflow* genutzt und für das *Plant Data Warehouse* re-implementiert werden konnte. Erfahrungen auf dem Gebiet der Data Warehouse Technologie bestanden bei unserem Projektpartner B.I.M.-Consulting mbH sowie bei unseren Kooperationspartnern der Universität Leipzig. *SOAP Web Services* für Bioinformatikanwendungen wurden mit Beginn des Projektes an verschiedenen Stellen der Welt entwickelt.

1.5 Zusammenarbeit mit anderen Stellen

Wissenschaftliche Kooperationen sind heutzutage auf den Gebieten der Bioinformatik und der Genom- und Postgenomforschung eine Notwendigkeit. So wurden auch im Rahmen des *Plant Data Warehouse* Projektes Kooperationen auf nationaler und internationaler Ebene geknüpft, die über die initiale Förderperiode hinaus bestehen und sich fruchtbar weiterentwickeln. Enge Kontakte bestanden über die gesamte Projektlaufzeit hinweg mit den verschiedenen Arbeitsgruppen des BIC-GH. Diese Zusammenarbeit führte nicht nur zu gemeinsamen Entwicklungen von Anwendungen für das *Plant Data Warehouse*, sondern auch zu gemeinsamen Publikationen.

Des weiteren entwickelte sich eine sehr intensive Kooperation mit der Arbeitsgruppe Bioinformatik des IPK Gatersleben. Die Grundlage des *Plant Data Warehouse* bilden sauber strukturierte Datenbestände, und diese Vorleistung wurde gemeinsam mit der Arbeitsgruppe Bioinformatik erbracht. Daran anschließend setzte sich die enge Zusammenarbeit beim Aufbau des *Plant Data Warehouse* und abschließend beim Transfer des *Plant Data Warehouse* in die Arbeitsgruppe Bioinformatik fort. Gemeinsame Publikationen resultierten aus dieser fruchtbaren Zusammenarbeit.

Intensive Kooperationen entstanden mit vielen experimentell orientierten Arbeitsgruppen am IPK Gatersleben und IPB Halle. Diese Kooperationen begannen in der Anfangsphase des *Plant Data Warehouse* Projektes bei den Anforderungsanalysen und der Erstellung der *Use Cases* und wurden im Zuge der Entwicklung des *Plant Data Warehouse* und der verschiedenen Anwendungen weiter ausgebaut. In vielen Fällen mündete die Zusammenarbeit in Analysen der von den Experimentatoren generierten Daten mit dem *Plant Data Warehouse* gemeinsam durch die Bioinformatiker des *Plant Data Warehouse* Projektes und unserer Experimentatoren.

Die Arbeitsgruppen am IPK Gatersleben, mit denen wichtige Zusammenarbeiten im Rahmen des *Plant Data Warehouse* Projektes entstanden, sind: Außenstelle Nord, Dr. K. Dehmer; Bioinformatik, Dr. U. Scholz; Dateninspektion, Dr. M. Strickert; Expressionskartierung, Dr. L. Altschmied; Genbankdokumentation, Dr. H. Knüpffer; Genomdiversität, Prof. A. Graner; Genregulation, Dr. H. Bäumlein; Genwirkung, Prof. U. Wobus; Gen- und

Genomkartierung, Dr. M. Röder; Invitro Erhaltung und Cryo-Lagerung, Dr. J. Keller; Molekulare Netzwerke, Dr. F. Börnke; Molekulare Pflanzenphysiologie, Prof. G. Kunze, Prof. U. Sonnewald; Pflanzenstress und Entwicklung, Dr. P. Bauer; Phytoantikörper, Dr. U. Conrad; Quantitative Evolutionäre Genetik, Dr. K. Schmid; Ressourcengenetik und Reproduktion, Dr. A. Börner; Taxonomie Pflanzengenetischer Ressourcen, Dr. R. Fritsch; Transkriptomanalyse, Dr. P. Schweizer.

Auf nationaler Ebene entwickelten sich fruchtbare Kooperationen mit den folgenden Partnern: Bundesanstalt für Züchtungsforschung, Quedlinburg, Dr. L. Freese; Biobase GmbH, Wolfenbüttel, Dr. A. Kel, Dr. O. Kel, Prof. E. Wingender; Freie Universität, Berlin, Prof. K. Reinert; Friedrich Miescher Labor, Tübingen, Dr. G. Rätsch; Humboldt-Universität, Berlin, Prof. H. Herzel, Prof. S. Hougardy, Dr. A. Schmitt; IPB Halle, Dr. S. Rosahl; Martin-Luther-Universität Halle-Wittenberg, Prof. P. Molitor; Max Planck Institut für Molekulare Genetik, Berlin, Dr. A. Schliep; Max Planck Institut für Molekulare Pflanzenphysiologie, Golm, Dr. B. Kersten; Max Planck Institut für Züchtungsforschung, Köln, Dr. H. Schoof; Universität Bielefeld, Bielefeld, J. Baumbach, T. Kohl, Prof. B. Weisshaar; Universität Leipzig, Leipzig, Prof. E. Rahm; Friedrich-Alexander-Universität Erlangen-Nürnberg, Dr. S. Biemelt.

Auf internatonaler Ebene entstanden Zusammenarbeiten vor allem auf dem Gebiet der Entwicklung neuer Analysesoftware. In vielen Fällen erfüllte die weltweit existierende Software nicht die Anforderungen der Nutzer des *Plant Data Warehouse*, und so mussten existierende Anwendungen modifiziert oder neue Anwendungen entwickelt werden. Dies führte zu fruchtbaren Kooperationen vor allem mit den folgenden Gruppen: Berlex Bioscience, San Francisco, USA, Dr. J. Fickett; Ciudad Universitaria, Madrid, Spanien, Prof. J. Vicente Carbajosa; Cold Spring Harbor Laboratory, Cold Spring Harbor, USA, Dr. A. Smith, Dr. D. Ware, Prof. M. Zhang; European Bioinformatics Institute, Hinxton, GB, M. Hoffman; Hebrew University of Jerusalem, Jerusalem, Israel, Prof. H. Margalit; Institut Curie, Paris, Frankreich, P. Neuvial; Massachusetts Institute of Technology, Cambridge, USA, Dr. D. Holste; Max Perutz Labs, Wien, Österreich, Prof. A. v. Haeseler; North Shore LIJ Research Institute, Manhasset, USA, Dr. W. Li; Ohio State University, Columbus, USA, Prof. R. Davuluri; St. Petersburg Polytechnical University, St. Petersburg, Russland, Prof. M. Samsonova; Tel Aviv University, Tel Aviv, Israel, Prof. I. Ben-Gal; University of Rijeka, Rijeka, Kroatien, Prof. B. Podobnik; University of Barcelona, Barcelona, Spanien, Prof. J. Cerquides; University of Evry, Evry, Frankreich, Dr. P.-Y. Bourguignon, Prof. B. Prum University of Massachusetts, Lowell, USA, Prof. K. Marx; University of Pennsylvania, Philadelphia, USA, Prof. A. Hatzigeorgiou; University Pompeu Fabra, Barcelona, Spanien, Dr. R. Castelo; URGV, Evry, Frankreich, Prof. M. Caboche, Dr. A. Lecharny.

Unabhängig zu diesen vielfältigen Kooperationen mit individuellen Gruppen waren für das *Plant Data Warehouse* Projekt Ideenaustausche in größerer Breite wichtig. So wurden verschiedene Konferenzen und Workshops organisiert, wie z. B.:

- das *Minisymposium on Expression Data Analysis* am 2. – 3. September 2004 am IPK Gatersleben,
- das *Minisymposium on Metabolome Analysis in Plants* am 17. – 18. März 2005 am IPK Gatersleben,
- der *Miniworkshop on Optimal Reconstruction of Permuted Markov Models and Bayesian Networks* am 7. Juli 2006 an der Martin-Luther-Universität in Halle,
- die *Konferenz on Data Warehouse Technologies in Bioinformatics (DWTB 2006)* am 14. – 16. Dezember 2006 in Wittenberg,

Für die Doktoranden des *Plant Data Warehouse* Projektes waren ebenfalls die drei in den Jahren 2005, 2006 und 2007 am IPK Gatersleben und IPB Halle stattfindenden *ISC Studentenkonferenzen* ein wichtiger Höhepunkt ihrer wissenschaftlichen Laufbahn und die Möglichkeit, ihre Fortschritte bei der Entwicklung des *Plant Data Warehouse* zu präsentieren als auch mit potentiellen Anwendern des *Plant Data Warehouse* ins Gespräch zu kommen.

Des weiteren waren Mitglieder der Arbeitsgruppe *Plant Data Warehouse* an der Organisation der jährlich stattfindenden Klausurtagungen des BIC-GH in Wittenberg, der *European Summer School*[1] *on Plant Genomics and Bioinformatics* im Rahmen der Plant Metanet Initiative am 18. – 29. September 2006 in Potsdam/Golm sowie des *Plant Bioinformatics Symposium* am 24. – 25. September 2007 an der Martin-Luther-Universität in Halle beteiligt. Darüber hinaus wirkten Mitglieder des *Plant Data Warehouse* Teams in den Programmkomitees der folgenden Konferenzen mit: *International Conference on Information Visualisation in Biomedical Informatics (IV07)*, 2. – 6. Juli 2007, Zuerich, *Lernen Wissen Adaption (LWA'07)*, 24. – 26. September 2007, Halle, und *German Conference on Bioinformatics (GCB'07)*, 26. – 28. September 2007, Potsdam.

Die Zusammenarbeit mit der Martin-Luther-Universität Halle-Wittenberg entwickelte sich nicht nur auf dem Gebiet der Forschung, sondern auch auf dem Gebiet der Lehre exzellent. So wurden seit Wintersemester 2003/2004 Vorlesungen im Rahmen des Diplomstudiengangs Bioinformatik und des Master Aufbaustudiengangs Bioinformatik am Institut für Informatik der Mathematisch Naturwissenschaftlich Technischen Fakultät der Martin-Luther-Universität Halle-Wittenberg im Umfang von durchschnittlich 4 bis 6 SWS pro Semester gehalten. Dieses Lehrangebot umfasste die folgenden, teilweise jährlich angebotenen, Vorlesungen:

- Algorithmen der Bioinformatik II
- Angewandte Bioinformatik, gemeinsam mit Dr. Scholz (Arbeitsgruppe Bioinformatik), Dr. Schreiber (Arbeitsgruppe Netzwerkanalyse) und Dr. Seiffert (Arbeitsgruppe Mustererkennung)
- Expressionsdatenanalyse I und II
- Molekulare Phylogenie
- Sequenz- und Expressionsdatenanalyse I und II
- Sequenzanalyse I und II

Diese Lehraktivitäten dienten zum einen dem Ziel, Probleme aus der Praxis in die Lehre einzubringen. Zum anderen resultierte aus ihnen ein verstärktes Interesse der Studentinnen und Studenten an Forschungsprojekten des BIC-GH. Das verstärkte Interesse führte zu verschiedenen exzellenten Diplomarbeiten mit Nutzen für das *Plant Data Warehouse* Projekt. Die enge Kooperation mit der Martin-Luther-Universität Halle-Wittenberg wurde dadurch intensiviert, dass der Leiter des *Plant Data Warehouse* Projektes ab Wintersemester 2006/2007 eine Vertretungsprofessur an der Martin-Luther-Universität Halle-Wittenberg und mit Beginn des Wintersemesters 2007/2008 einen Ruf an die Martin-Luther-Universität Halle-Wittenberg annahm.

[1] ETNA (European Training and Networking Activity)

2 Eingehende Darstellung

2.1 Erzielte Ergebnisse

Alle Arbeitspakete konnten erfolgreich abgeschlossen werden. Das *Plant Data Warehouse* stellt heute eine leistungsfähige Softwareplattform zur Integration und Analyse von molekularen, phänotypischen und taxonomischen Daten sowie von Daten zu pflanzengenetischen Ressourcen aus IPK- und IPB-internen und weltweit verteilten Quellen dar und wird inzwischen monatlich durch mehr als 1.000 Anwender aus dem In- und Ausland genutzt. Die folgenden Absätze beschreiben die wichtigsten Teilergebnisse im einzelnen.

2.1.1 AP1 Evaluierung und Konsolidierung bestehender Datenbestände und Anforderungsanalyse bezüglich der benötigten Daten

2.1.1.1 Crop EST Informationssystem CR-EST Das Crop EST Informationssystem CR-EST wurde in Zusammenarbeit mit den Arbeitsgruppen Bioinformatik und Transkriptomanalyse entwickelt und enthält Informationen über cDNA Bibliotheken, offene Leserahmen, EST-Cluster, Sequenzalignments, funktionelle Annotationen, Abbildungen von ESTs auf metabolische Netzwerke sowie eine Reihe von Anwendungen zu deren integrativer Analyse. CR-EST ist derzeit mit Daten aus Gerste, Erbse, Kartoffel, Weizen und Tabak befüllt und prinzipiell offen für weitere Organismen. CR-EST ist öffentlich und frei verfügbar unter `http://pgrc.ipk-gatersleben.de/cr-est/`, ein Artikel zu CR-EST wurde publiziert [7], und das Informationssystem wird monatlich von ca. 200 – 400 Anwendern genutzt.

2.1.1.2 Molekulare Marker Datenbank MoMa Die Molekulare Marker Datenbank MoMa wurde in Zusammenarbeit mit den Arbeitsgruppen Bioinformatik und Genomdiversität entwickelt. Sie hält Daten über Einzelnukleotidpolymorphismen (SNPs), einfache Sequenzrepeats (SSRs), *Restriction Fragment Length Polymorphisms* (RFLPs) und *Amplified Fragment Length Polymorphisms* (AFLPs) sowie Informationen über Allele, Kartierungspositionen in populationsspezifischen Karten und Konsensuskarten, Positionen von Polymorphismen sowie Primer Paare zur Amplifizierung in 30 Tabellen und 26 Sichten. Die MoMa Datenbank ist unter `http://pgrc.ipk-gatersleben.de/moma/` verfügbar und wurde seit ihrer Fertigstellung intensiv genutzt.

2.1.1.3 Europäischen Gerstendatenbank EBDB Eine der wichtigsten Aufgaben von AP1 war das Re-Engineering der Europäischen Gerstendatenbank EBDB. Die Mitglieder der Gattung *Hordeum* bilden einen Pool mit einem breiten ökologischen Spektrum und sind daher weit verbreitet. So wurde Gerste schon früh als Nahrungsmittel entdeckt und ist eine der wichtigsten und ältesten domestizierten Pflanzen. Ihre Bedeutung für den Menschen liegt heute vor allem in der Nutzung als Futtermittel und als Braugerste. Bereits 1983 wurde von der *Barley Working Group* des *European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR)* die Schaffung einer Datenbank initiiert, die die Passport- und Evaluierungsdaten sämtlicher Gerstenakzessionen in Europäischen Genbanken enthalten und dem Austausch von Informationen zwischen Züchtern, Wissenschaftlern und Genbanken dienen sollte. Die Ziele waren:

- die Erfassung der genetischen Vielfalt von Gersten in Europäischen Genbanken,
- die Bereitstellung von Informationen über verfügbare Akzessionen für Genbanken, Züchter und Wissenschaftler,

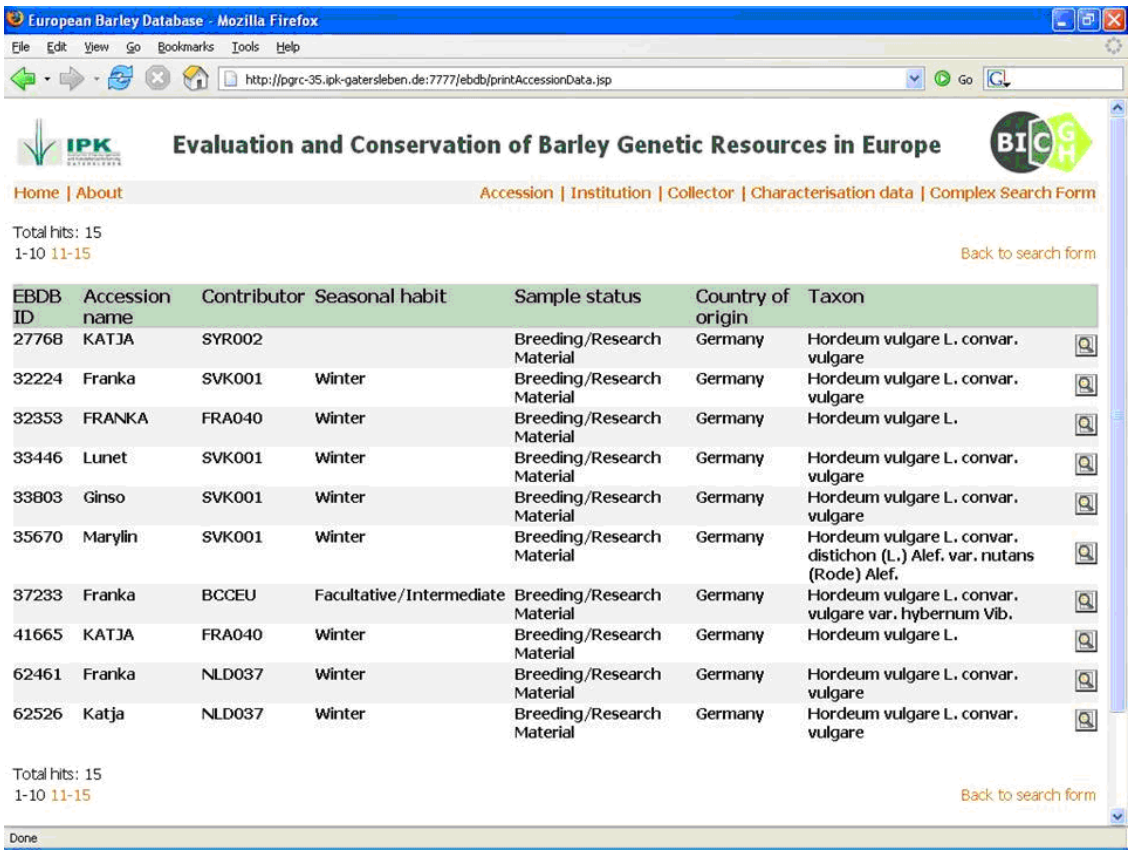

Abbildung 1: Formular für die Suche nach Akzessionen der Europäischen Gerstendatenbank (EBDB).

- die Erkennung und Eliminierung von Duplikaten,
- die Erkennung von "geographischen Lücken" in Europäischen Sammlungen im Hinblick auf künftige Sammelreisen.

Die Europäische Gerstendatenbank(EBDB) enthält derzeit die Daten von mehr als 155.000 Akzessionen aus 35 Europäischen Genbanken sowie einigen außereuropäischen Sammlungen sowie die *Barley Core Collection*. Das Re-Engineering der Europäischen Gerstendatenbank EBDB und der in ihr enthaltenen *Barley Core Collection* BCC wurde in Zusammenarbeit mit der Arbeitsgruppe Genbankdokumentation durchgeführt. Das Re-Engineering umfasste nicht nur die Neumodellierung des ER-Schemas, sondern auch die komplette Neugestaltung der Anwendungsoberfläche, um eine adäquate Außendarstellung zu ermöglichen (Abbildungen 1 und 2). Die Anwendungen wurden als *Java Server Pages* entwickelt, um ihre problemlose Einbettung in den zentralen *Application Server* des IPK zu ermöglichen. Die Datenbank wurde unter dem Datenbankmanagementsystem Oracle auf dem zentralen Datenbankserver des IPK implementiert, was zu einer erheblichen Leistungssteigerung und Reduktion des Wartungsaufwandes führte.

2.1.1.4 Europäische Poadatenbank EPDB Die Europäische Poadatenbank EPDB wurde von der Zentralstelle für Agrardokumentation und -information an das BIC-GH transferiert und in Zusammenarbeit mit der IPK Außenstelle Nord weiterentwickelt. Das Re-Engineering umfasste die Erweiterung des ER-Schemas und die Neuentwicklung der Anwendungsoberfläche mit Hilfe von *Java Server Pages*. Wie im Fall der EBDB und der anderen für

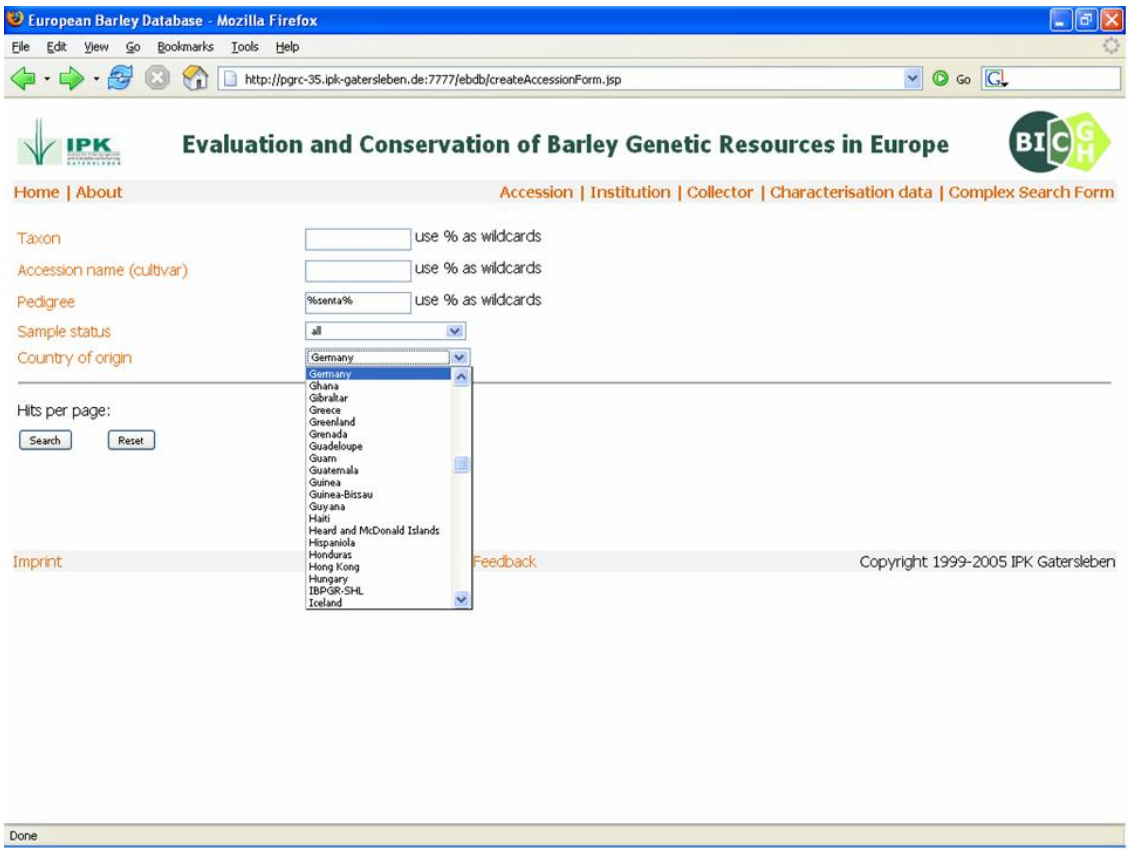

Abbildung 2: Ergebnis der Akzessionssuche in der Europäischen Gerstendatenbank (EBDB).

das *Plant Data Warehouse* entwickelten Operativsysteme läuft die Anwendungskomponente der EPDB auf dem zentralen *Application Server* des BIC-GH und die Datenbankkomponente der EPDB auf dem zentralen Datenbankserver des BIC-GH, was den Wartungsaufwand minimiert.

2.1.1.5 Charakterisierungs- und Evaluierungsdatenbank EVAL Die Charakterisierungs- und Evaluierungsdatenbank EVAL wurde in Zusammenarbeit mit der Arbeitsgruppe Genbankdokumentation entwickelt. Sie wurde mit einem Import-Tool versehen, welches prüft und erzwingt, dass nur solche Daten, die dem IPGRI-Standard entsprechen, in die Datenbank eingespielt werden können. Die Datenbank wurde inzwischen als Modul in das neue Genbankinformationssystem GBIS integriert, und die im *Plant Data Warehouse* benötigten Charakterisierungs- und Evaluierungsdatenbank werden seitdem direkt aus dem Genbankinformationssystem GBIS gelesen.

2.1.1.6 Pyrosequencing Datenbank PSQDB Pyrosequenzierungsdaten sind von fundamentaler Bedeutung für die moderne Züchtungsforschung. Zur strukturierten Haltung von Pyrosequenzierungsdaten und als notwendige Voraussetzung zum Import dieser Daten in das *Plant Data Warehouse* wurde in Zusammenarbeit mit den Arbeitsgruppen Bioinformatik und Genomdiversität die Pyrosequenzierungsdatenbank PSQDB entwickelt. Die PSQDB erlaubt die strukturierte Verwaltung von Sequenzen und Primern, Referenzmarkern der Gerste, SNPs, Häufigkeitsmesswerten der Allele, Definitionen der Akzessionen, nutzerspezifizierbaren Definitionen von Kernkollektionen, Passportdaten, Charakterisierungs- und Evaluierungsdaten sowie dazugehörigen Labor- und Experimentdaten. Die Architektur der

PSQDB wurde so entworfen, dass sie in Zukunft universell zur Speicherung von Pyrosequenzierungsdaten genutzt werden kann. Die PSQDB wurde durch unsere Kooperationspartner, vor allem im Teilprojekt *Einsatz molekularer Marker zur Eliminierung von Duplikaten* innerhalb des Projektes *Aufbau einer bundeszentralen* ex situ *Genbank für landwirtschaftliche und gartenbauliche Kulturpflanzen: Zusammenführung der Genbanken des IPK und der BAZ Braunschweig* intensiv genutzt.

2.1.1.7 Datenbank für Assoziationsstudien PhaSe Assoziationsstudien sind wichtig in vielen Projekten der modernen Pflanzenzüchtung weltweit. Um die in solchen Projekten generierten Daten zentral speichern zu können, wurde in Kooperation mit den Arbeitsgruppen Bioinformatik und Genomdiversität die PhaSe Datenbank entwickelt. Sie ergänzt und verknüpft die MoMa Datenbank, die Sequenzdatenbank SeqDB und Datenbanken für Passportdaten sowie Charakterisierungs- und Evaluierungsdaten. Die PhaSe Datenbank wurde u. a. im Rahmen des Deutsch-Französischen Gabi-Génoplante Projektes *Bridging genomics and genetic diversity: Associations between gene polymorphism and trait variation in cereals* genutzt.

2.1.1.8 Datenbank für Malz- und Brauqualität Gemeinsam mit der Arbeitsgruppe Gen- und Genomkartierung wurde eine Datenbank für Malz- und Brauqualität von Sommergersten aufgebaut. Dabei wurden Daten aus Jahrbüchern der Braugerstengemeinschaft der letzten ca. 20 Jahre erfasst, bereinigt und in Oracle importiert. All diese Daten wurden in das *Plant Data Warehouse* integriert und stehen seitdem für verschiedenste Analysen, z. B. Assoziationsstudien, durch das *Plant Data Warehouse* zur Verfügung.

2.1.2 AP2 Design des Datenbankschemas des *Plant Data Warehouse*

Gemeinsam mit unserem Industriepartner B.I.M.-Consulting mbH wurde zu Beginn des Projektes über Auswahl der Implementierungs- bzw. Lösungsstrategie zur Realisierung des *Plant Data Warehouse* Projekts diskutiert. Das Ergebnis dieser Analysen war die Entscheidung den *Enterprise Data Mart* Ansatz ohne fixes, sondern mit generischen Schemata zu implementieren. Die Gründe für die Auswahl und ein Schema-Beispiel für einen *Data Mart* (Sequenzmart) wurden im Abschlussbericht unseres Industriepartners B.I.M.-Consulting detailliert präsentiert.

Analog zu den generischen Schemata der *Data Marts* wurde für die Verwaltung des Benutzer- und Rechtemanagements ein flexibler Ansatz gewählt, der ebenfalls im Abschlussbericht unseres Industriepartners umfangreich erläutert wurde.

Im *Plant Data Warehouse* werden Sequenz-, Marker- Expressions-, Metabolom-, Passport-, Phänotyp- sowie Wetterdaten gespeichert. In der von uns gewählten Architektur erfolgt die Speicherung dieser Daten in den *Data Marts*. Diese *Data Marts* unterstützen eine anwendungsbezogene Speicherung bzw. Vorbereitung der Daten im *Plant Data Warehouse*. Die *Data Marts* wurden in Zusammenarbeit mit unserem Industriepartner sowie der Arbeitsgruppe Bioinformatik modelliert, implementiert und mit Daten befüllt.

Der Alignmentmart wurde mit Alignments zwischen ESTs verschiedener Gräser und dem Genom von Reis befüllt. Die Berechnungen der Alignments erfolgte mit den Programmen Blast, Blat, GeneSequer, Sim4 und Spidey. Um die Alignments einheitlich in einer Datenstruktur erfassen zu können, war es notwendig, Parser für diese Alignment Programme als auch einen einheitlichen Score zur Bewertung der Alignments zu entwickeln. Eine webbasierte Abfragemaske für die Datenbank wurde mittels *Oracle-APEX* implementiert. Der Transkriptommart wurde in Zusammenarbeit mit der Arbeitsgruppe Bioinformatik entwickelt

und dient der Integration und Analyse von Makroarray-, Mikroarray- und Affymetrixdaten aus IPK- und IPB-internen und weltweit verteilten Quellen.

Die auf statischen HTML Seiten beruhende Präsentation der Wetterdaten des IPK wurde durch eine dynamische Präsentation auf dem zentralen *Application Server* des BIC-GH abgelöst. Der Wettermart des *Plant Data Warehouse* wurde so entwickelt, dass die Wetterdaten des IPK Gatersleben komplett eingespielt werden konnten. Die auf dem *Application Server* implementierten Anwendungen ermöglichen nun eine integrative Analyse von ca. 20 Zeitreihen (Luft- und Bodentemperaturen in verschiedenen Höhen, Luftfeuchtigkeit, Niederschlag, Windstärke, Sonneneinstrahlung, etc.), die teilweise mehr als 50 Jahre zurückreichen und heute im 10-Minuten-Takt erfasst werden. Die auf dem *Application Server* eingesetzte Portal-Technologie ermöglicht es versierten Nutzern, ihre eigenen Analysen zusammenzuklicken und dann auf den aktuellen Daten durchzuführen.

Ein ausführliche Beschreibung des restlichen *Data Marts* ist im Abschlussbericht unseres Industriepartners B.I.M.-Consulting mbH enthalten.

2.1.3 AP3 Integration von Daten aus den Operativsystemen und externen Datenquellen und Ermöglichung von Interoperabilität

Entsprechend der Antragstellung wurden die folgenden Daten in enger Abstimmung mit unserem Industriepartner B.I.M.-Consulting mbH in das *Plant Data Warehouse* integriert:

1. Passportdaten zu Gerste vom IPK Gatersleben,
2. Charakterisierungs- und Evaluierungsdaten zu Gerste und anderen Kulturpflanzen von der IPK Genbank und anderen Europäischen Genbanken,
3. Passport- sowie Charakterisierungs- und Evaluierungsdaten der *Barley Core Collection* der Europäischen Gerstendatenbank EBDB,
4. Molekulare Marker Daten zu Gerste vom IPK Gatersleben,
5. Bilddaten von der BIC-GH Arbeitsgruppe Mustererkennung,
6. Sequenzdaten (ESTs, SNPs, SSRs) zu Gerste vom IPK Gatersleben,
7. Genetische Karten zu Gerste vom IPK Gatersleben und anderen öffentlichen Quellen der Welt,
8. Expressionsdaten zu Gerste und *Arabidopsis thaliana* vom IPK Gatersleben,
9. Proteom- und Metabolomdaten zu *Arabidopsis thaliana* vom IPK Gatersleben und IPB Halle sowie aus anderen öffentlichen Quellen der Welt.

Im *Plant Data Warehouse* sind aktuell Datensätze von 9 Kulturpflanzen aus verschiedenen Bereichen integriert. Die Quellen sind neben internen Operativsystemen (z. B. CR-EST, MoMa, SeqDB, PSQDB, Flarex oder GBIS) externe Systeme wie NCBI Genbank, TIGR, TAIR, PlexDB, AtGenExpress oder SGED. Der Integrationsprozess ist dokumentiert und automatisiert, so dass bei Aktualisierungen in den Quellen ein automatisches Update initiiert werden kann. Somit ist es jetzt und in Zukunft möglich, IPK- und IPB-interne Daten, wie z. B. Gersten-Marker, im Zusammenhang mit öffentlichen Daten, wie z. B. dem Reis-Genom, zu analysieren.

Eine Interoperabilität kann durch den Austausch von Daten (Ergebnisse, Parameter) zwischen den im *Plant Bioinformatics Portal* angebotenen Anwendungen mit verschiedenen Methoden erreicht werden. Mit *SOAP Web Services* können standardisiert Daten ausgetauscht werden. So wurden *SOAP Web Services* für Operativsysteme sowie die im *Plant Data Warehouse* integrierten Daten entwickelt und eine Kooperation mit der Initiative *BioMoby* zum Einsatz von *SOAP Web Services* begonnen. Des weiteren wurden *SOAP Web Services* für verschiedene in das *Plant Data Warehouse* integrierte Anwendungen entwickelt, um diese in übergreifende Netzwerke einzubringen.

Entsprechend dem an MyNCBI (vgl. http://www.ncbi.nlm.nih.gov) angelehnten Konzept myPBP (my *Plant Bioinformatics Portal*) wurde für das *Plant Data Warehouse* ein *DataCart* entwickelt. Er ermöglicht eine flexible Verknüpfung der in das *Plant Data Warehouse* integrierten Anwendungen durch den Austausch von Daten zwischen Anwendungen innerhalb des *Plant Bioinformatics Portal*. Verglichen mit der Nutzung von *SOAP Web Services* kann so ohne hohen Implementierungsaufwand seitens der Anwendungen eine Interoperabilität erreicht werden. Die Daten stellen Ergebnisse von integrierten Anwendungen dar. Die Metadatenverwaltung erfolgt im *Meta Data Repository*.

2.1.4 AP4 Konsistenzüberprüfungen und Fehlerkorrektur

Der Datenimport basiert auf dokumentierten Prozeduren. Dabei wird die Konsistenz von Akzessionsnummern, andere Schlüsseln sowie einfachen Checks, wie korrekte Zahlen, Datentypen, Datumsangaben über die Oracle Datenbankfunktionalität realisiert. Entsprechende Fehler beim Import sind protokolliert und können somit behoben werden. Für nach dem Import festgestellte Fehler ist für Meldung durch die Anwender die Nutzung des Feedbacksystems möglich. Dieses wird im Abschnitt Kurations-System kurz beschrieben.

2.1.5 AP5 Integration von Anwendungen

Die Anforderungsanalysen wurden intensiv in den ersten Projektmonaten und anschließend projektbegleitend bis zum Ende des *Plant Data Warehouse* Projektes durchgeführt. Die aufgenommenen *Use Cases* und Spezifikationen wurden dokumentiert und im nicht-öffentlichen Teil des *Plant Bioinformatics Portal*, der den Entwicklern zugänglich ist, abgelegt. Die entwickelten Anwendungen lassen sich einteilen in Anwendungen zur Analyse und Visualisierung von (a) Sequenzdaten, (b) Sequenz- und Markerdaten, (c) Expressions- und Metabolomdaten, (d) Sequenz- und Expressionsdaten, (e) Sequenz-, Marker-, und Expressionsdaten, (f) Sequenz-, Marker-, Passport-, Phänotyp- und Wetterdaten sowie (g) übergreifende Anwendungen. Viele der Anwendungen sind zwar zur Analyse von Daten konkreter Spezies entwickelt worden, lassen sich aber auf Daten beliebiger Spezies, deren Daten in das *Plant Data Warehouse* integriert sind, anwenden.

2.1.5.1 Analyse von Sequenzdaten

Browser für den Alignmentmart Alignments spielen eine fundamentale Rolle bei der Beantwortung verschiedener biologischer Fragestellungen. Daher wurde in Kooperation mit unseren experimentellen Partnern, vor allem aus der Arbeitsgruppe Genomdiversität, eine Anwendungsoberfläche für den Alignmentmart (Abschnitt 2.1.2) entwickelt, welche komplexe Abfragen an diesen *Data Mart* erlaubt.

Tabelle 1: Prozentsatz der richtig vorhergesagten Transkriptionsfaktorbindungsstellen (Sensitivität) bei einer Spezifität von 99.9%. Die Tabelle zeigt, dass durch die Nutzung der VOM bzw. VOB Modelle die die Sensitivität für jeden der sechs Transkriptionsfaktoren **AP-1**, **CEBP**, **GATA**, **NF-1**, **SP-1** und **Thyroid** (bei gleich bleibender Spezifität) erhöht werden kann.

Modell 1	**Modell 2**	**AP-1**	**CEBP**	**GATA**	**NF-1**	**SP-1**	**Thyroid**
MM	MM	66.7	25.1	77.2	70.4	74.0	50.0
BN	MM	65.0	24.8	70.3	61.8	73.1	45.5
VOMM	VOMM	**67.9**	**25.9**	**79.0**	**71.0**	**75.1**	51.8
VOBN	VOMM	67.5	25.7	78.1	69.2	73.5	**52.0**

Berechnung von Allelhäufigkeiten und des *Polymorphism Information Content (PIC)* Für das *Plant Data Warehouse* wurde ein Programm entwickelt, das in gegebenen Sequenz Alignments nach Einzelnukleotidpolymorphismen (SNPs) und INDELs sucht und anschließend die SNPs hinsichtlich ihrer relativen Allelhäufigkeiten und ihres *Polymorphism Information Contents (PIC)* charakterisiert sowie die Haplotyphäufigkeiten berechnet. Dieses Programm wurde u. a. im Rahmen des Deutsch-Französischen Gabi-Génoplante Projektes *Bridging genomics and genetic diversity: Associations between gene polymorphism and trait variation in cereals* intensiv genutzt.

Erkennung offener Leserahmen in Gersten- und Weizen-ESTs Das beste Programm zur Vorhersage offener Leserahmen (ORFs) in Pflanzen-ESTs ist das Programm Bestorf der Firma Softberry. Da insbesondere bei kurzen ORF-Fragmenten die Vorhersagegenauigkeit von Bestorf die Anforderungen der Nutzer des *Plant Data Warehouse* nicht erfüllt, wurde ein auf Gerste und Weizen spezialisiertes Programms, ESTstriker, entwickelt. ESTstriker unterscheidet sich von konkurrierenden Programmen zur Erkennung von ORFs durch seine Unterteilung der ESTs in verschiedene Längen- und Typenklassen. Ein Vergleich auf annotierten Transkripten von Gerste zeigt nach Kreuzvalidierung die Überlegenheit von ESTstriker im Vergleich zu konkurrierenden Programmen inklusive Bestorf. Insbesondere für kurze ORF-Fragmente erlaubt ESTstriker eine wesentlich genauere Vorhersage von ORFs als Bestorf.

Erkennung cis-regulatorischer Elemente basierend auf annotierten Sequenzen Die Erkennung cis-regulatorischer Elemente ist wichtig für viele Forschungsgruppen des IPK Gatersleben und IPB Halle. Eine neue Methode zur computergestützten Erkennung von Transkriptionsfaktorbindungsstellen wurde in Zusammenarbeit mit der Forschungsgruppe von Prof. Ben-Gal, Tel Aviv University, and der Arbeitsgruppe Bioinformatik der Martin-Luther-Universität Halle-Wittenberg, entwickelt. Der Algorithmus basiert auf einer Kombination von *Variable Order Markov* (VOM) Modellen und *Variable Order Bayes* (VOB) Bäumen und ermöglicht für die Mehrheit der heute bekannten Transkriptionsfaktorbindungsstellen eine genauere Erkennung als existierende Programme (Tabelle 1). Die Software wurde in das Plant Bioinformatics Portal integriert, und verschiedene mit diesen Algorithmen und der Web Anwendung durchgeführte Fallstudien wurden inzwischen veröffentlicht [5, 10, 14].

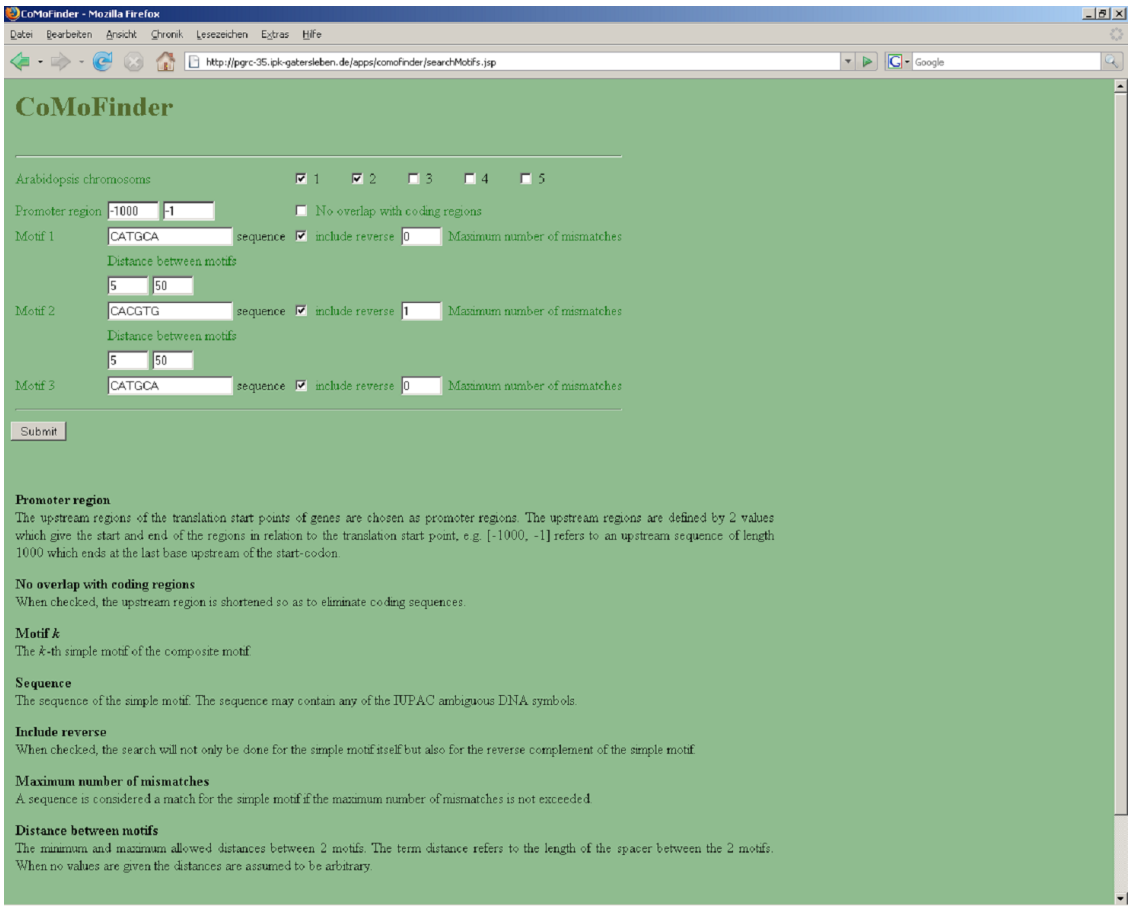

Abbildung 3: Eingabeformular des Programms *CoMoFinder* im *Plant Bioinformatics Portal*.

Vorhersage von Zielgenen bekannter Transkriptionsfaktoren Die Komplexität der Regulation der Genexpression wird in höheren Eukaryonten durch die kombinatorische Interaktion verschiedener Transkriptionsfaktoren und das Vorkommen von *Composite Motifs* erreicht. *Composite Motifs* bestehen aus mehreren verschiedenen Transkriptionsfaktorbindungsstellen, die in unterschiedlichem Abstand und in unterschiedlicher Reihenfolge angeordnet sind. Das im Rahmen des trilateralen Projektes *Arabidoseed* entwickelte Programm *CoMoFinder* wurde im Rahmen des *Plant Data Warehouse* Projektes modifiziert und in das *Plant Data Warehouse* integriert (Abbildung 3).

CoMoFinder wurde dahingehend erweitert, dass die Eingabedaten jetzt direkt aus dem Sequenzmart und dem Annotationsmart des *Plant Data Warehouse* gelesen werden können. Die *Composite Motifs* werden durch die Konsensussequenz der einzelnen Teilmotive und durch Intervalle von erlaubten Abständen zwischen den Teilmotiven spezifiziert. Bei der Spezifikation der Konsensussequenzen wurde die Angabe von nicht-eindeutigen Nukleotiden gemäß dem IUPAC Standard erlaubt. Das Programm wurde in das *Plant Bioinformatics Portal* integriert und seitdem intensiv durch verschiedene biologische Anwender genutzt.

Vorhersage von *de-novo* Transkriptionsfaktorbindungsstellen Die Erkennung neuer cis-regulatorischer Elemente in nicht-alignierten und nicht-annotierten Promotorregionen, die aus mRNA Expressionsexperimenten oder ChIP/chip-Experimenten stammen, ist eine wichtige Aufgabe bei der Analyse von genregulatorischen Netzwerken. Existierende Programme zur Erkennung von *de-novo* Transkriptionsfaktorbindungsstellen sind für viele praktischen Anwendungen im Bereich der modernen Pflanzengenomforschung nur bedingt

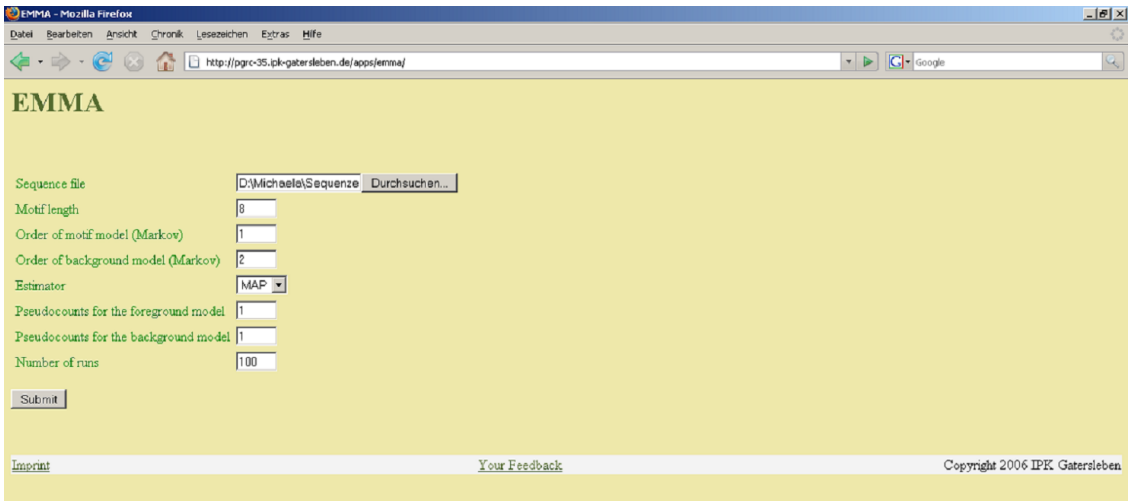

Abbildung 4: Eingabeformular des Programms *EMMA* im *Plant Bioinformatics Portal*.

tauglich, weil sie nicht genügend Anpassungsmöglichkeiten für die konkreten Fragestellungen ermöglichen. Im Rahmen des trilateralen Projektes *Arabidoseed* wurde daher ein neues Programm entwickelt, welches den speziellen Anforderungen der Nutzer auf dem Gebiet der Pflanzengenomforschung Rechnung trägt.

Dieses Programm, *EMMA*, wurde im Rahmen des *Plant Data Warehouse* Projektes weiterentwickelt. Es ermöglicht die Vorhersage von Transkriptionsfaktorbindungsstellen sowie von *cis-regulatorischen Modulen*, die aus mehreren Einzelmotiven bestehen, unter Verwendung von VOM und VOB Modellen (Abschnitt 2.1.5.1). *EMMA* wurde an das *Cluster Execution Framework* des BIC-GH Linuxclusters angeschlossen und in das *Plant Bioinformatics Portal* integriert (Abbildung 4). Seitdem wurde und wird es intensiv von verschiedenen Forschungsgruppen aus verschiedenen Ländern für die Analyse von Sequenzdaten im Kontext von Expressionsdaten und ChIP/chip-Daten genutzt.

Maximum Entropie Modelle *Maximum Entropie* Modelle sind probabilistische Modelle, die von Gene Yeo und Chris Burge, Massachusetts Institute of Technology, für die Vorhersage von kurzen Spleißstellen vorgeschlagen wurden und beliebige Abhängigkeiten modellieren können. Im Gegensatz zu Markov Modellen oder Bayesnetzen lassen sich die Parameter dieser Modelle in den meisten Fällen nicht analytische berechnen. Aus diesem Grund müssen numerische Verfahren zur Ermittlung der optimalen Parameter verwendet werden. Die Laufzeit dieser Verfahren wächst leider exponentiell mit der Sequenzlänge, so dass die Anwendung existierender numerischer Verfahren auf sehr kurze Sequenzen beschränkt ist.

Im Rahmen des *Plant Data Warehouse* Projektes wurde daher ein Verfahren entwickelt, welches Rechengeschwindigkeit der schnellsten bis dato existierenden Verfahren um mehr als das zehnfache übertrifft (Abbildung 5). Dieses Verfahren reduziert den Rechenaufwand von mehreren Wochen auf wenige Tage und ermöglicht damit Datenanalysen, die vorher nicht möglich waren. Darüber hinaus erlaubt der neue *Blockwise Iterative Scaling (BGIS)* Algorithmus die Analyse längerer Sequenzen, was zu einer signifikanten Erhöhung der Sensitivität und Spezifität der Klassifikation führt (Abbildung 6).

Detektion und Korrektur von Annotationsfehlern Die in das *Plant Data Warehouse* importierten Daten stammen aus verschiedenen Referenzdatenbanken. Diese werden häufig durch Eingaben per Hand aus wissenschaftlichen Publikationen gefüllt. Dabei können

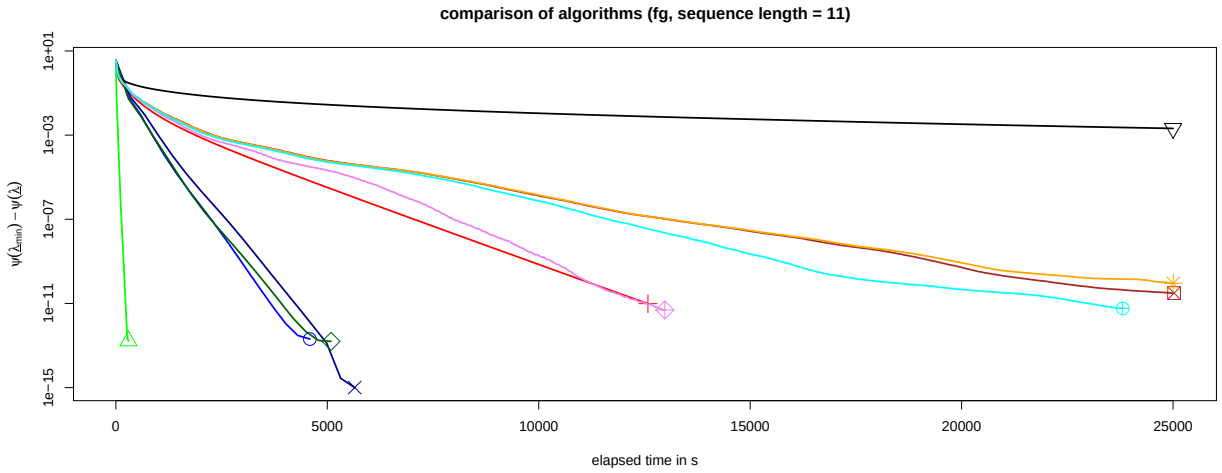

Abbildung 5: Entwicklung der Likelihood (y-Achse) während des Trainings als Funktion der Zeit (x-Achse). Die hellgrüne Kurve zeigt das Konvergenz des *BGIS* Algorithmus, der eine deutlich geringere Laufzeit aufweist als alle anderen Algorithmen (blau = *sequential generalized iterative scaling* Algorithmus im P-Raum, hellgrün = *blockwise generalized iterative scaling* Algorithmus im P-Raum, rot = *generalized iterative scaling* Algorithmus, dunkelblau = *sequential generalized iterative scaling* Algorithmus, dunkelgrün = *blockwise generalized iterative scaling* Algorithmus, schwarz = Gradientenanstieg, braun und orange = *conjugate gradient* Algorithmen, rosa = Quasi Newton Algorithmus nach Broyden Fletcher Goldfarb Shanno, türkis = *limited memory* Quasi Newton Algorithmus nach Broyden Fletcher Goldfarb Shanno mit 10 Vektoren).

verschiedene Fehler auftreten. Im Bereich der Sequenzdaten und Annotationsdaten treten zum einen Tippfehler und zum anderen Verschiebungen der *Features* auf. Im Fall der Annotation von Transkriptionsfaktorbindungsstellen kommt hinzu, dass oftmals der Strang, auf dem das cis-regulatorische Element liegt, auf der Basis der durchgeführten Experimente nicht bestimmt werden kann. Zur Lösung diese Problems und zur Behebung der erstgenannten Fehler nutzt man bisher weitestgehend Expertenwissen gekoppelt mit extensiver Handkuration.

Daher wurde ein Softwarepaket *Motif-Adjuster* entwickelt, welches Experten bei der Detektion und Korrektur dieser Fehler unterstützt und einen Vorschlag für den Strang des untersuchten cis-regulatorischen Elementes unterbreitet. *Motif-Adjuster* basiert auf probabilistischen Modellen für Sequenzmotive, die selbstlernend aus einem Datensatz Sequenzen herausfiltern, die keine Bindungsstelle oder die Bindungsstelle an einer anderen als der annotierten Position enthalten. Des weiteren berechnet *Motif-Adjuster* die Wahrscheinlichkeit für das Motivvorkommen auf beiden Strängen und liefert eine Warnung, falls der Strang mit der höheren Wahrscheinlichkeit nicht mit der Strangannotation übereinstimmt.

Motif-Adjuster ermöglicht in vielen Fällen eine signifikante Erhöhung der Qualität der integrierten Daten. In Zusammenarbeit mit Kollegen von der Universität Bielefeld wurde *Motif-Adjuster* erfolgreich zur Neu-Annotation der Bindungsstellen des Transkriptionsfaktors **NarL** eingesetzt (Abbildung 7). Die neu annotierten Daten bilden die Grundlage für das Training von Parametern verschiedener Modelle zur genomweiten Vorhersage von **NarL** Bindungsstellen. Dadurch ist es uns gelungen, putative Bindungsstellen in Zielgenen zu identifizieren, von denen bekannt ist, dass sie durch **NarL** reguliert werden, für die jedoch bislang keine potentiellen Bindungsstellen identifiziert werden konnten. Erste Experimente zur experimentellen Verifikation der vorhergesagten Bindungsstellen sind an der Universität Bielefeld

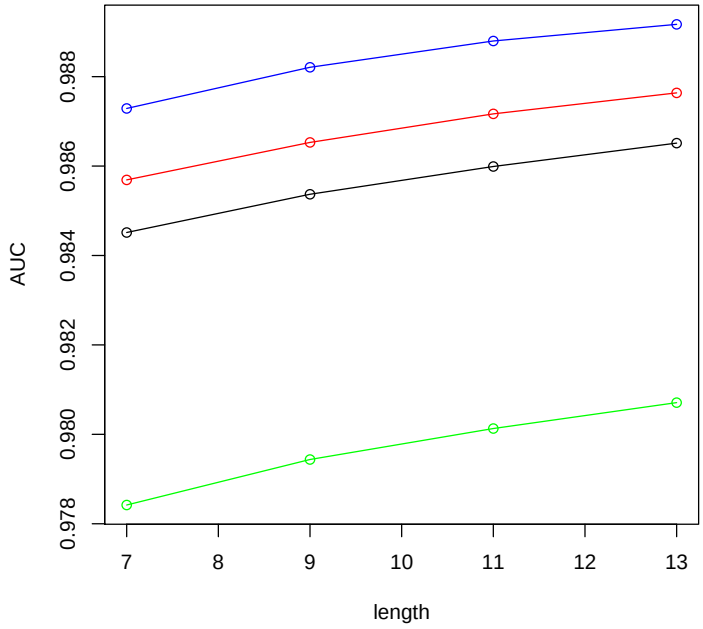

Abbildung 6: Das Gütemaß *Fläche unter der ROC-Kurve* (y-Achse) als Funktion der Sequenzlänge (x-Achse). Die grüne Kurve entspricht dem *PWM* Modell, die schwarze dem Bayes Baum, die rote dem Inhomogenen Markov Modell erster Ordnung und die blaue dem *Maximum Entropie* Modell. Deutlich zu erkennen ist die Verbesserung der Klassifikation durch das *Maximum Entropie* Modell im Vergleich zu den anderen Modellen. Der hier verwendete Datensatz wurde aus der Datenbank *splicedb* extrahiert, die experimentell verifizierte Spleißstellen bereitstellt.

angelaufen, und eine Veröffentlichung von *Motif-Adjuster* unter einer freien akademischen Lizenz ist derzeit in Planung.

2.1.5.2 Analyse von Sequenz- und Markerdaten

SNP2CAPS Einzelnukleotidpolymorphismen (SNPs) sind eine weit verbreitete Art von DNA Variationen und sehr gut zur Entwicklung von Markern geeignet. Ihre Erkennung ist jedoch sehr kostspielig und erfordert moderne Labortechnik, die in vielen, vor allem kleineren, Laboren der Welt nicht verfügbar ist. Im Gegensatz dazu ist die Entwicklung von *Cleaved Amplified Polymorphic Sequence (CAPS)* Markern einfach und preiswert. In Zusammenarbeit mit der Arbeitsgruppe Genomdiversität wurde daher die Analysesoftware *SNP2CAPS* entwickelt, die SNP Marker in CAPS Marker konvertiert. Basierend auf der Analyse von 3.045 Gersten-ESTs ergab sich, dass 90% der SNP Marker in CAPS Marker konvertiert werden konnten. *SNP2CAPS* wurde publiziert [4] und ist in das *Plant Data Warehouse* integriert.

Clusterbasierte Kartierung von Gersten- und Weizen-ESTs Eine häufig auftretende Frage bei der Markerentwicklung ist die, ob zu einem gegeben EST bereits ein Marker existiert, und wenn ja, ob er kartiert wurde. Um bei der Beantwortung dieser Frage

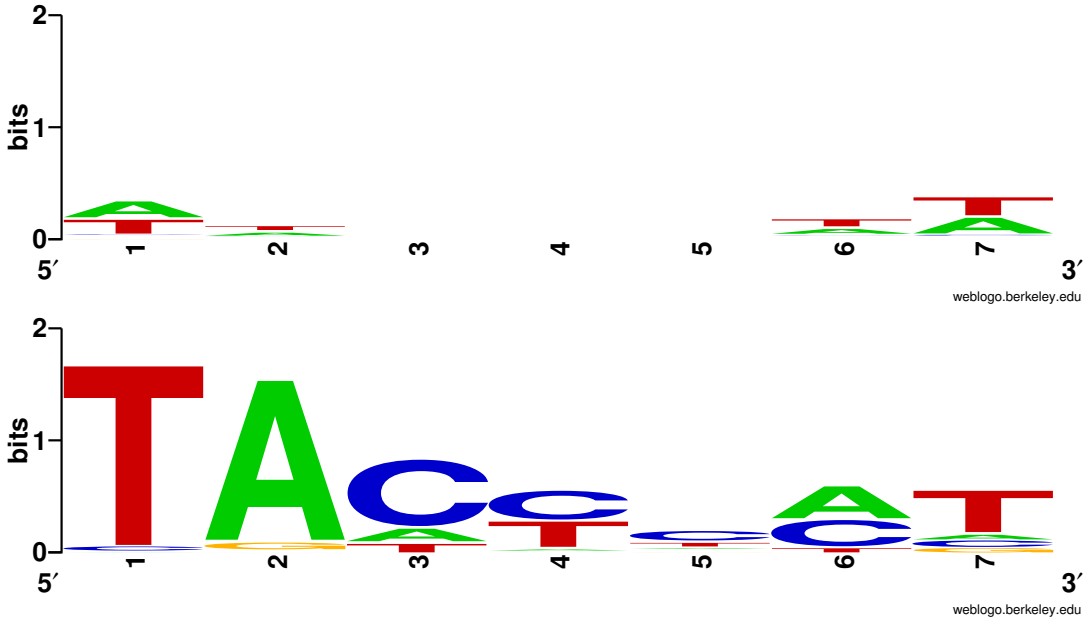

Abbildung 7: Sequenzlogos der **NarL** Bindungsstellen vor (oben) und nach (unten) der automatischen Fehlerkorrektur. Das obere Logo wurde basierend auf den Originalannotationen erstellt. Das untere Logo wurde basierend auf den durch das Programm *Motif-Adjuster* korrigierten Daten erstellt. Man erkennt deutlich, dass das obere Logos wesentlich verrauschter als das untere ist. Daraus ergeben sich stark fehlerbehaftete Vorhersagen von **NarL** Bindungsstellen im Gesamtgenom basierend auf den Originaldaten und wesentlich weniger fehlerbehaftete Vorhersagen von **NarL** Bindungsstellen im Gesamtgenom basierend auf den durch *Motif-Adjuster* korrigierten Daten. Bemerkenswert ist, dass in diese Fallstudie keine Handkuration einfloss.

zu helfen, wurde in Zusammenarbeit mit der Arbeitsgruppe Genomdiversität eine Analysepipeline, der *Sequence Mapping eXplorer (SMeX)*, zur Kartierung von ESTs entwickelt. Die im *Plant Data Warehouse* zur Verfügung stehenden Daten umfassen ESTs von Nutzpflanzen, Clusterdaten, genetische Kartierungspositionen von ca. 1.000 Gerstenmarkern, die überwiegend am IPK entwickelt wurden, sowie ESTs von GenBank, Unigene Cluster vom TIGR und vom NCBI und Markerdaten von GrainGenes. *SMeX* ermöglicht dem Nutzer, erstens Marker und Kartierungspositionen für gegebene ESTs und deren homologen Sequenzen innerhalb einer oder in nahen verwandten Spezies und zweitens ESTs und deren homologen Sequenzen zu einem gegebenen Bereich von Kartierungspositionen zu finden. Zusätzlich sind drittens Blast-Suchen nutzerspezifischer Sequenzen gegen alle in der Datenbank gespeicherten ESTs möglich. *SMeX* ist im *Plant Bioinformatics Portal* verfügbar, und die rechenintensive Blast-Suche ist in das *Cluster Execution Framework* eingebunden und damit auf dem BIC-GH Linuxcluster möglich.

Kartierung von Gersten-ESTs unter Zuhilfenahme der Syntänie zu Reis Eine Erweiterung der Clusterbasierten Kartierung besteht in der Kartierung von ESTs basierend auf Syntänien zu Referenzgenomen. Die genetische Kartierung von ESTs ist für viele molekularbiologische Anwendungen auf dem Gebiet der Kulturpflanzenforschung ein fundamentaler, jedoch auch zeitaufwendiger und teurer Schritt. Daher wurde in Kooperation mit der Arbeitsgruppe Genomdiversität eine Pipeline zur rechnergestützten Kartierung von Gersten-ESTs auf Grundlage der Syntänie zu Reis entwickelt. Diese Pipeline verwendet eine

genetische Karte von Gerste, die derzeit 1001 der rund 130.000 am IPK generierten Gersten-ESTs enthält, sowie die durch TIGR annotierte genomische DNA von Reis. Die Pipeline enthält folgende Schritte:

1. Spliced Alignment der Gersten-ESTs mit der genomischen DNA von Reis mit Spliced Alignment Programmen wie Blat, GeneSeqer, Sim4 und Spidey,
2. Detektion von syntänen Regionen auf Basis der genetisch kartierten ESTs.
3. Vorhersage der Kartierung der restlichen öffentlich verfügbaren Gersten-ESTs basierend auf dem Syntäniemodell.

Die durch die Pipeline benötigten Eingabedaten liegen im Sequenzmart und Markermart vor, die durch die Pipeline berechneten *Spliced Alignments* werden im Alignmentmart abgespeichert, und die Vorhersagen der Kartierungspositionen können zur weiteren Analyse in den DataCart geladen oder im Markermart abgespeichert werden. Die Berechnungen der *Spliced Alignments* sind sehr rechenintensiv und werden daher durch das *Cluster Execution Framework* auf dem BIC-GH Linuxcluster gestartet.

Unter Benutzung der Standardparameter für Spidey, Sim4 und Blat können für 79%, 78% bzw. 66% der Gersten-ESTs Homologe in Reis detektiert werden. Eine *Leave-One-Out* Kreuzvalidierung des Syntäniemodells ergab für den derzeit besten Parametersatz (Spliced Alignment Programm = Spidey, Window Size = 4 Mb $\times$ 20 cM, Local Prediction = Linear Regression), dass die genetische Kartierungsposition von 44% (53%) der öffentlich verfügbaren und mit dem Reisgenom alignierbaren Gersten-ESTs mit einer Genauigkeit von 5 cM (10 cM) rechnergestützt bestimmt werden können.

Identifizierung von Genomduplikationen in Gerste Genomduplikationen sind ein verbreitetes Phänomen in Pflanzengenomen. Sie spielen eine wichtige Rolle bei der Entstehung von Genen mit neuen Funktionen und bei der Artbildung. Darüber hinaus hat die Kenntnis über duplizierte Bereiche im Pflanzengenom Bedeutung bei der Bearbeitung von molekularbiologischen Fragestellungen in der Pflanzenzüchtung. Publizierte Studien basierend auf Kreuz-Hybridisierungs-Experimenten lassen eine ähnliche Genomstruktur innerhalb der Familie der Gräser vermuten, zu denen viele wichtige Kulturpflanzen wie Reis, Mais, Weizen, Gerste oder Roggen gehören.

Die frei verfügbare genomische Sequenz von Reis deutet auf mehrere duplizierte Segmente hin, die ungefähr zwei Drittel des Reisgenoms ausmachen. Bis auf eines dieser Segmente stammen höchstwahrscheinlich alle Segmente von einem *whole-genome duplication* Ereignis ab, das vor der Artbildung innerhalb der Gräser stattgefunden hat. Die gegenwärtig verfügbaren genetischen Karten der Gerste, die nur einige Tausend Marker enthalten, liefern nicht genügend Datenpunkte, um segmentale Duplikationen basierend auf Sequenzhomologien aufzuzeigen.

In Kooperation mit der Arbeitsgruppe Genomdiversität wurden Algorithmen entwickelt, um Duplikationen im Gerstengenom unter Zuhilfenahme des Reisgenoms indirekt aufzeigen zu können. Die Grundidee besteht darin, dass Gersten-EST-Marker einer duplizierten Region in Gerste mit einer Region im Reisgenom alignieren. Wenn also Gersten-EST-Marker zweier unterschiedlicher Regionen in der genetischen Karte von Gerste mit derselben Region auf dem Reisgenom alignieren, ist dies ein Indiz für eine Duplikation dieser Region im Gerstengenom.

Des weiteren sind Alignments von Gersten-EST-Markern zweier unterschiedlicher Regionen in der genetischen Karte von Gerste mit zwei unterschiedlichen Regionen im Reisgenom ein Indiz dafür, dass sowohl im Gerstengenom als auch im Reisgenom eine Duplikation

vorliegt. Basierend auf der genetischen Transkript-Karte von Gerste, die über 1000 Gersten-ESTs enthält, konnten auf diese Weise insgesamt sechs anzestrale Genomduplikationen identifiziert werden, die seit der Artbildung von Gerste und Reis vor ca. 60 Millionen Jahren in beiden Genomen konserviert sind. Weitere Analysen haben gezeigt, dass die Verteilung der putativ orthologen und putative paralogen Gene zwischen Gerste und Reis in Übereinstimmung mit dem erwarteten Muster anzestral duplizierter Segmente ist.

2.1.5.3 Analyse von Expressionsdaten Expressionsdaten spielen eine fundamentale Rolle in vielen Genom- und Postgenomprojekten, und so wurde eine Pipeline zur Normierung, Filterung und Analyse von Makroarray Expressionsdaten in Zusammenarbeit mit den Arbeitsgruppen Bioinformatik, Expressionskartierung, Genomdiversität und Transkriptomanalyse sowie mit der Arbeitsgruppe Bioinformatik der Martin-Luther-Universität Halle-Wittenberg entwickelt. Diese Pipeline, *SMArrT*, liest die Rohdaten aus dem Transkriptommart und ermöglicht dem Nutzer über eine intuitiv zu bedienende graphische Nutzeroberfläche die Normierung und Filterung der eingelesenen Expressionsdaten.

Mikroarrays der Firma Affymetrix werden als Alternative zu Makroarrays häufig zur Untersuchung differentiell exprimierter Gene genutzt. Insbesondere dann, wenn es um die Vergleichbarkeit von Expressionsdaten aus verschiedenen Laboren geht, wird bevorzugt auf Affymetrixdaten zurückgegriffen. Daher wurde in Zusammenarbeit mit dem IPB Halle und dem Max Planck Institut für molekulare Pflanzenphysiologie in Golm eine analoge Pipeline für Affymetrixdaten erstellt. Im Vorfeld der Implementierung dieser Pipeline war es notwendig, die Qualität der verschiedenen existierenden Normierungsverfahren an Hand von pflanzenspezifischen Datensätzen zu überprüfen.

In der Vergangenheit waren bei methodischen Vergleichen von Affymetrixdaten Datensätze aus der Pflanzenforschung stark unterrepräsentiert. Für die Auswertung wurden zahlreiche Methoden entwickelt, die sich jedoch sehr stark unterscheiden. Daher wurden in einer systematischen Studie auf der Basis von Rohdaten des IPB Halle acht etablierte Normierungsmethoden (MAS5.0, RMA, GCRMA, PLIER, dChip, VSN und Varianten) miteinander verglichen. Die Rohdaten stammten aus Experimenten, in denen *Arabidopsis thaliana* mit dem Phytopathogen *Phytophtora infestans* behandelt wurde.

Die Ergebnisse zeigen, dass die Wahl der Normierungsmethode einen gravierenden Einfluss auf die Analyseergebnisse und deren biologische Interpretation hat (Abbildungen 8 und 9). Jede der acht Normierungsverfahren wurde unabhängig von den anderen zur Normierung der Expressionsdaten verwendet, und im Anschluss wurden in jedem der acht Fälle die differentiell exprimierten Gene mit Hilfe von Standardverfahren bestimmt. Abschließend wurde verglichen, wie gut die acht so bestimmten Listen differentiell exprimierter Gene überlappen. Das überraschende Ergebnis war, dass kein einziges Gen (von mehr als 20.000 Genen) in allen acht Listen auftrat. Darüber hinaus wurde ebenfalls kein Gen von sechs oder sieben Verfahren detektiert. Nur fünf Gene wurden von fünf der angewendeten Methoden als differentiell exprimiert gefunden und nur 20 von vier Methoden.

Da im Fall der von uns untersuchten Daten von *Arabidopsis thaliana* verschiedene Normalisierungspipelines zu extrem verschiedenen Analyseergebnissen führten, ergab sich die Konsequenz, im Fall von *Arabidopsis thaliana* in Zukunft keine der Normierungsverfahren einzeln zu verwenden. Stattdessen ergab sich der Vorschlag, alle Normierungspipelines simultan zu nutzen, um am Ende wissenschaftlich verwertbare Ergebnisse zu erhalten. Die resultierende Pipeline basiert auf Funktionen von R bzw. *Bioconductor* und wurde unter Nutzung von *Rsoap* in das *Java Framework* integriert.

2.1.5.4 Analyse von Sequenz- und Expressionsdaten

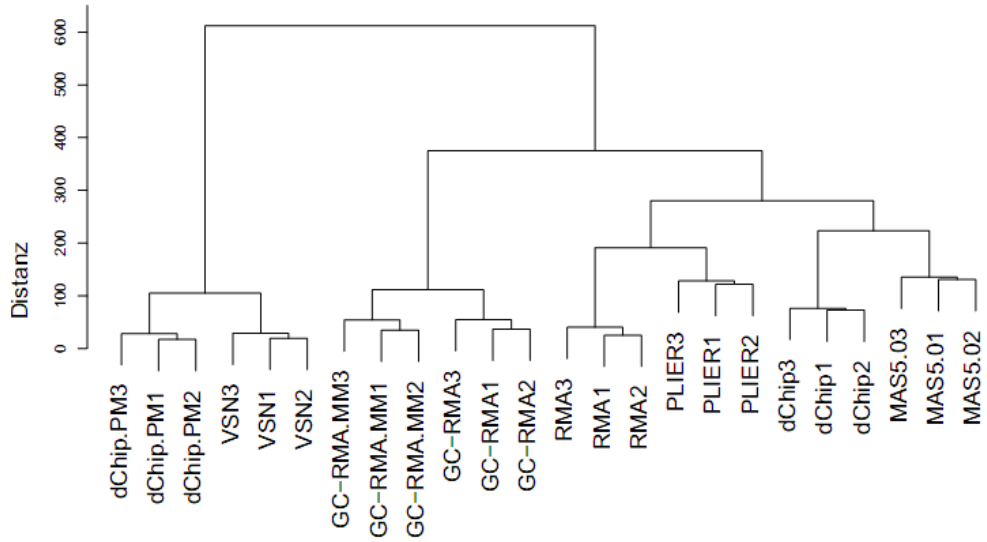

Abbildung 8: Clusterdiagramm für alle Vorverarbeitungsmethoden und jeweils drei Replikate zwölf Stunden nach Behandlung mit *Phytophtora infestans*, Clustermethode: com- plete linkage.

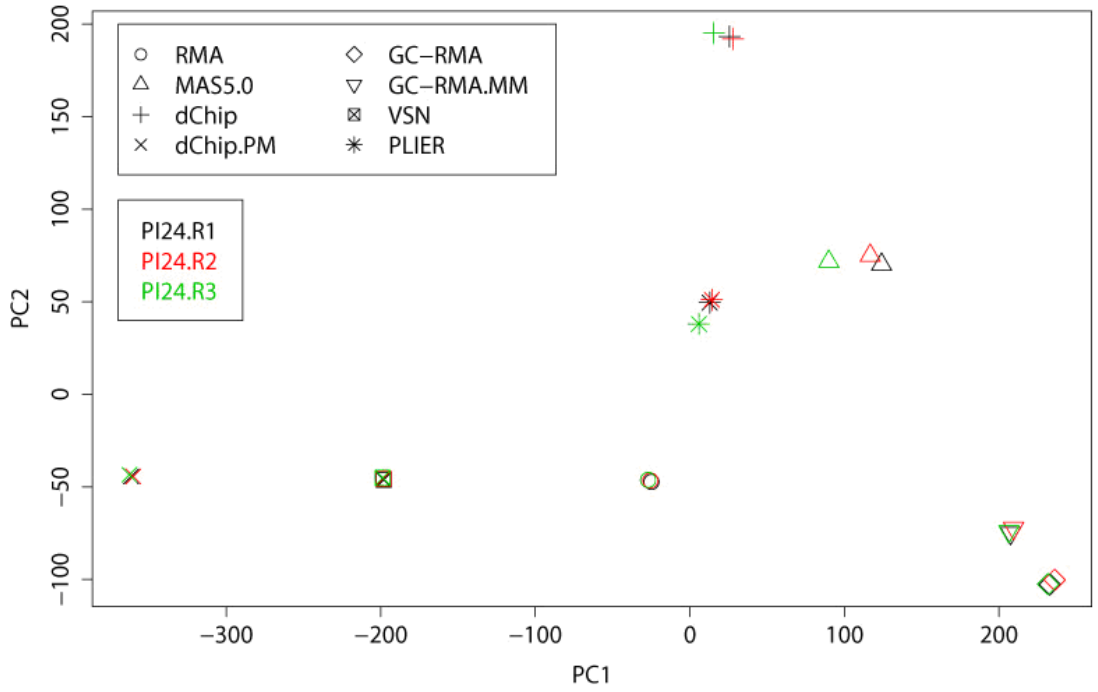

Abbildung 9: Diagramm der ersten beiden Hauptkomponenten für den Zeitpunkt 24 Stunden nach Behandlung von *Arabidopsis thaliana* mit *Phytophtera infestans*. Die Hauptkomponenten erklären 70,2% bzw. 16,7% der Varianz. Es wurden jeweils die drei Replikate aufgetragen (schwarz, rot, grün). Besonders dChip, VSN und RMA liefern reproduzierbare Ergebnisse.

Analyse von Expressionsdaten und cis-regulatorischen Elementen Im Fall der Analyse von Expressionsdaten von vollständig sequenzierten Genomen kann die Pipeline zur Normierung und Filterung von Expressionsdaten (Abschnitt 2.1.5.3) durch eine anschließende Analyse überrepräsentierter cis-regulatorischer Elemente erweitert werden. Die *Transcription Factor Binding Site and Expression Data Analysis* Pipeline wurde in Zusammenarbeit mit den Arbeitsgruppen Expressionskartierung, Genexpression und Molekulare Marker entwickelt, um in Mengen von Promotoren überrepräsentierte Transkriptionsfaktorbindungsstellen (TFBSs) zu finden.

Die Liste von Genen, deren Promotoren analysiert werden, wird Anhand von Expressionsdaten erstellt. Die Transkriptionsfaktoren, deren Bindungsstellen in den Promoterregionen überproportional vorhanden sind, sind putative Kandidaten für Regulatoren der analysierten Gene. Die Promoterregionen werden mit Gewichtsmatrizen der Transfac Datenbank, VOM Bäumen oder VOB Bäumen abgesucht. So kann für jeden Transkriptionsfaktor einfach gezählt werden, welche Gene (putativ) durch ihn beeinflusst werden und welche nicht. Die Transkriptionsfaktoren werden anschließend nach der (Bonferroni-korrigierten) *Mutual Information* oder dem *Mathews Correlation Coefficient* sortiert.

Die Transfac Daten werden aus dem Annotationsmart gelesen, und die Vorhersage der TFBSs wird auf dem BIC-GH Linuxcluster berechnett. Die Ergebnisse werden im Annotationsmart abgelegt und, auf Wunsch, im GFF Format exportiert. Die statistischen Berechnungen geschehen direkt auf dem Datenbankserver, und die Präsentation erfolgt durch eine Web Anwendung auf Basis von *Java Server Pages* im *Plant Bioinformatics Portal*.

Da die Genome von Gerste und Weizen noch nicht vollständig sequenziert sind, sind die Promoterregionen für diese Systeme weitgehend unbekannt. Es ist allerdings möglich, erstens mit dem Alignmentmart homologe ESTs aus Reis für viele der Gersten- und Weizen-ESTs zu identifizieren, zweitens deren Promotoren in Reis vorherzusagen und drittens die *Transcription Factor Binding Site and Expression Data Analysis* Pipeline auf Expressionsdaten von Gerste und Weizen und Genomdaten aus Reis anzuwenden.

Analyse von Expressionsdaten im Kontext genomischer Nachbarschaft Existierende Methoden zur Analyse von Expressionsdaten verarbeiten Expressionsdaten unabhängig von den verfügbaren Sequenzdaten (Abschnitt 2.1.5.3). Gerade in Pflanzen gibt es jedoch eine signifikante Korrelation zwischen den Expressionsdaten von im Genom benachbarten Genen. Um diese Korrelationen bei der Analyse von Expressionsdaten nutzen zu können, wurde eine Pipeline basierend auf *Hidden Markov* Modellen entwickelt und in das *Java Framework* integriert.

Standard *Hidden Markov* Modelle haben den Nachteil, dass keine statistischen Abhängigkeiten höherer Ordnung modelliert werden können, weil die Anzahl der Modellparameter exponentiell mit der modellierten Ordnung wächst, was bei der Analyse der verfügbaren Expressionsdaten zu Übertraining führt. Dieses Problem konnte durch die Entwicklung von *Variable Order Hidden Markov* Modellen gelöst werden. Die Modelle und die für sie benötigten Lernverfahren wurden in das *Java Framework* integriert. Die Pipeline zur Analyse von Expressionsdaten im Kontext genomischer Nachbarschsaft wurde inzwischen vielfach genutzt.

Analyse von ChIP/chip-Daten im Kontext genomischer Nachbarschsaft Derzeit werden verschiedene Transkriptionsfaktoren mit Hilfe der ChIP/chip-Technologie untersucht, um deren Zielgene zu identifizieren. Für die effiziente Auswertung dieser ChIP/chip-Daten wurde in Kooperation mit den Arbeitsgruppen Expressionskartierung, Genregulation und Phytoantikörper sowie mit weiteren externen Kooperationspartnern aus Bielefeld, Paris,

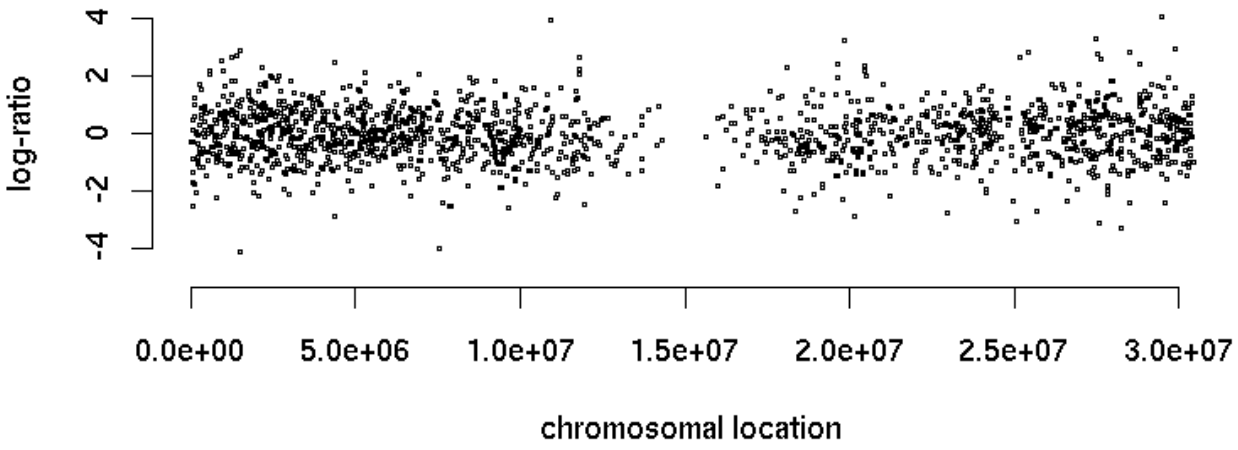

Abbildung 10: *Log-Ratio* Werte im Kontext ihrer genomischen Position auf Chromosom 1 von *Arabidopsis thaliana*. Deutlich zu erkennen ist die Lage des Zentromers.

Madrid und Sevilla eine Analyse-Pipeline entwickelt, die die Normierung und die Selektion von Zielgenen ermöglicht. Diese Pipeline basiert im wesentlichen auf den selben Algorithmen, die für die Analyse von Expressionsdaten im Kontext genomischer Nachbarschsaft entwickelt wurden.

Die biologische Motivation für die Nutzung eines *Hidden Markov* Modells liegt darin, dass Zielgene im allgemeinen nicht zufällig über das gesamte Genom verteilt sind. Insbesondere in Pflanzengenomen findet man eine starke Clusterung der Gene auf Grund gemeinsam genutzter Promotoren zweier benachbarter Gene sowie durch den hohen Anteil tandemduplizierter Gene. Abbildung 10 zeigt das Profil der normierten und logarithmierten Intensitäten in Abhängigkeit von der genomischen Position auf Chromosom 1 von *Arabidopsis thaliana*.

Hidden Markov Modelle mit zwei Zuständen und normalverteilten Ausgaben haben sich für die Modellierung der normierten ChIP/chip-Messwerte als adäquat erwiesen. Abbildung 11 zeigt die Verteilung der normierten und logarithmierten Intensitätsverhältnisse. Man erkennt, dass die ursprünglich nichtnormalverteilten Rohdaten durch die Normierung annähernd normalverteilt werden. Abbildung 12 zeigt die Architektur des *Hidden Markov* Modells. Das *Hidden Markov* Modell wird mit einem modifizierten Baum Welch Algorithmus auf den normierten ChIP/chip-Daten trainiert. Abschließend wird für jedes Gen die *a posteriori* Wahrscheinlichkeit dafür berechnet, dass dieses Gen ein Zielgen des untersuchten Transkriptionsfaktors ist.

Die Rohdaten der ChIP/chip-Experimente werden vorverarbeitet und normiert und kommen als *Log-Ratio* Werte im Kern der Pipeline an. Dort werden sie entweder (i) einer *Log-Fold-Change* Analyse ohne Berücksichtigung von Sequenzinformationen unterzogen oder (ii) durch ein *Hidden Markov* Modell im Kontext von Sequenzinformationen analysiert. Die Standardausgabe beider Verfahren ist eine ranggeordnete Liste der potentiellen Zielgene des untersuchten Transkriptionsfaktors. Diese Liste kann anschließend zur Vorhersage von cis-regulatorischen Elementen mittels *de-novo* Motiv-Suche mittels *EMMA* oder zur Suche nach bereits bekannten cis-regulatorischen Elementen oder cis-regulatorischen Modulen mittels *CoMoFinder* genutzt werden. Des weiteren kann diese Liste genutzt werden, um das Expressionsverhalten der potentiellen Kandidatengene in verschiedenen Geweben und unter verschiedenen Stressbedingungen miteinander zu vergleichen.

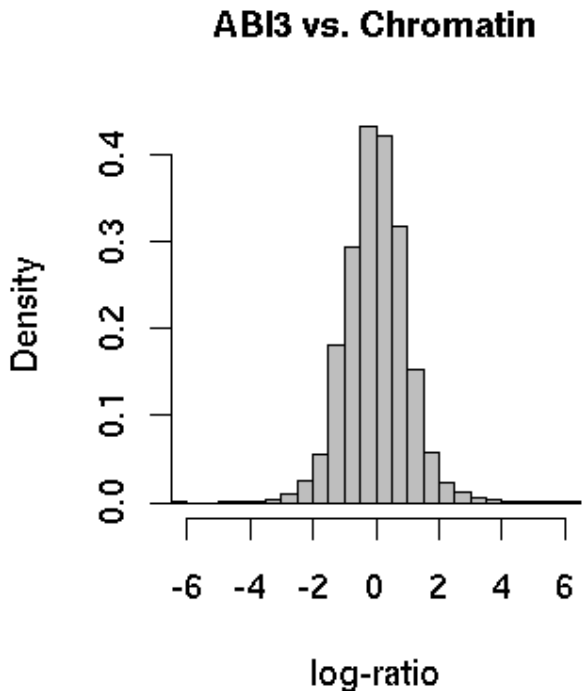

Abbildung 11: Histogramm der normierten und logarithmierten Intensitätsverhältnisse. Man erkennt, dass die *Log-Ratio* Werte annähernd normalverteilt sind.

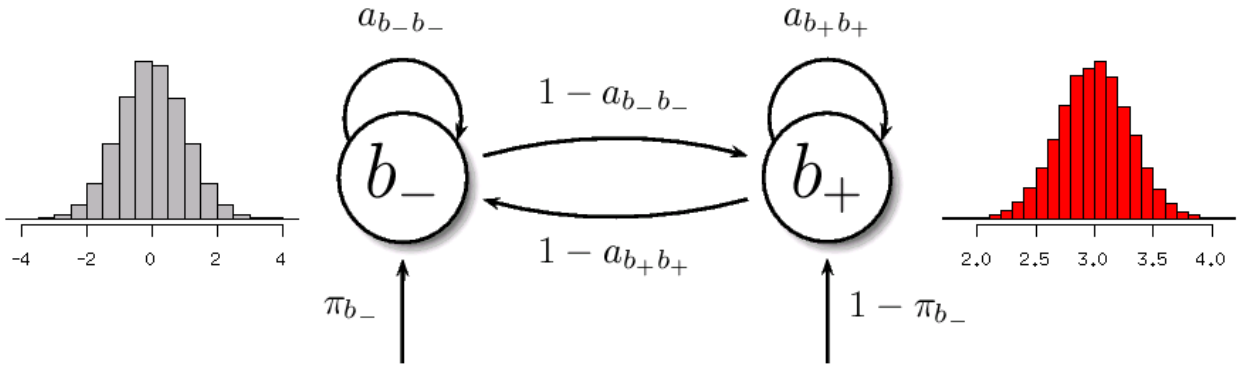

Abbildung 12: Architektur des *Hidden Markov* Modells zur Analyse von ChIP/chip-Daten im Kontext ihrer genomischen Lage. Der Zustand $b-$ repräsentiert potentielle Nicht-Zielgene, und der Zustand $b+$ repräsentiert potentielle Zielgene. Die Emissionen erfolgen durch zustandsspezifische Normalverteilungen.

Integrative Analyse von Sequenz-, Expressions- und ChIP/chip-Daten Für die integrative Auswertung von Sequenz-, Expressions- und ChIP/chip-Daten wurde eine Pipeline entwickelt, die diese Expressions- und ChIP/chip-Daten im Kontext genomischer Nachbarschaft normiert, potentielle Zielgene selektiert und anschließend *de-novo* Motiv-Suchen nach cis-regulatorischen Elementen und cis-regulatorischen Modulen durchführt. Die Pipeline ist in das *Java Framework* integriert und steht damit allen Nutzern frei zur Verfügung. Im Rahmen des trilateralen Projektes *Arabidoseed* wurde diese Pipeline wurde für die Analyse von ChIP/chip-Daten intensiv genutzt. Abbildung 13 zeigt eine schematische Übersicht dieser Pipeline.

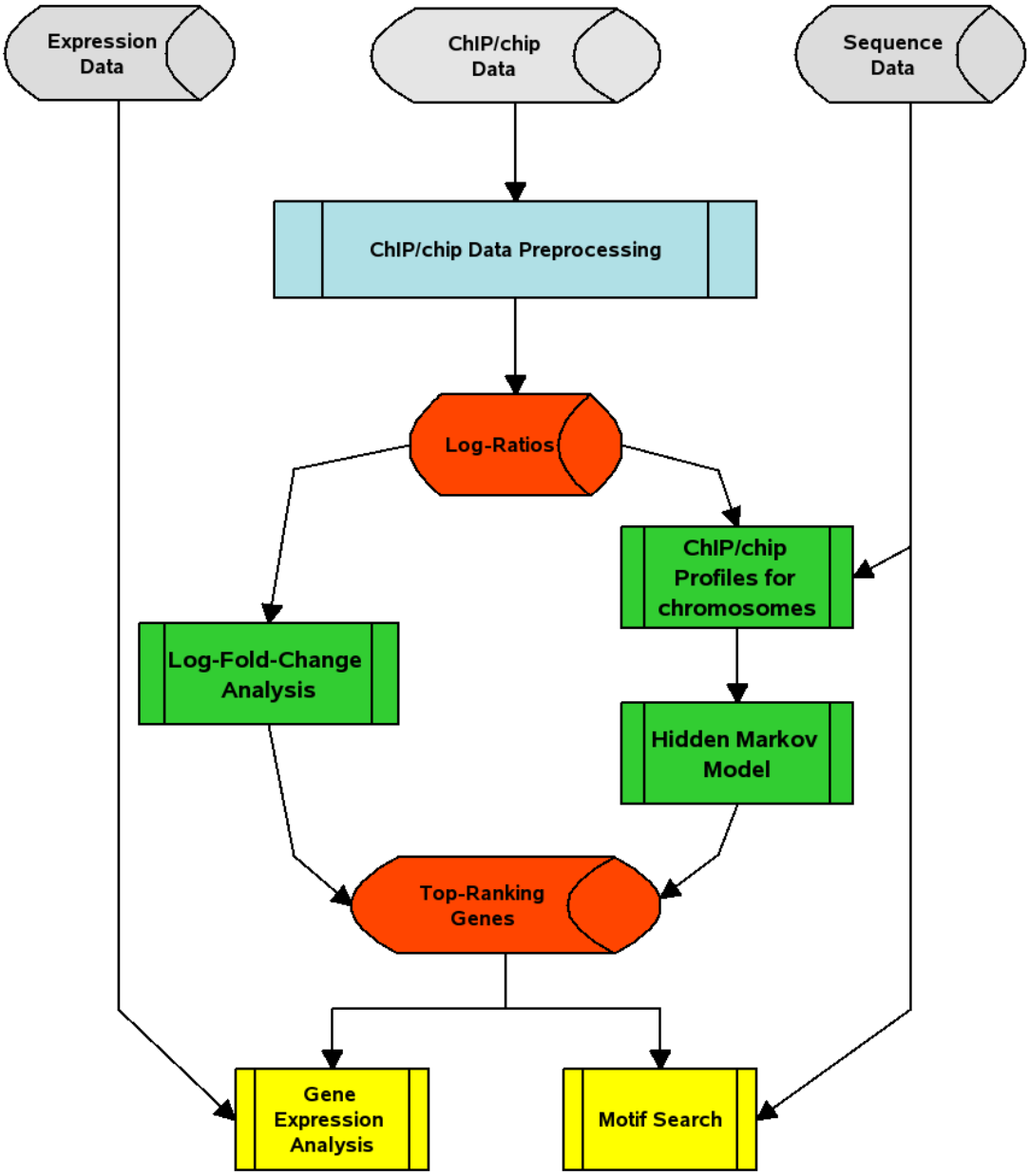

Abbildung 13: Pipeline zur integrativen Analyse von Sequenz-, Expressions- und ChIP/chip-Daten.

Eine dieser Analysen, basierend auf ChIP/chip-Daten des Transkriptionsfaktors **ABI3**, ergab das folgende interessante Ergebnis: Unter den 70 besten potentiellen **ABI3** Zielgenen wurden einige bekannte Zielgene wie z. B. verschiedene *Napine* identifiziert, was einen ersten Hinweis auf die Funktionalität der entwickelten Pipeline gab. Die Analyse der Expressionsdaten der vorhergesagten Zielgene ergab des weiteren, dass diese überwiegend im Samen exprimiert werden (Abbildung 14). Drittens lieferte die *de-novo* Motiv-Suche mittels *EMMA* das *G-Box* Motiv als das am stärksten konservierte cis-regulatorische Element (Abbildung 15). Viertens ergab die Suche nach *de-novo* cis-regulatorischen Modulen das *RY* Motiv in der Nähe der konservierten *G-Boxen*. Dieses cis-regulatorische Modul kommt statistisch signifikant häufiger in den untersuchten 70 putativen Zielgenen als in Negativkontrollgenen vor. Gegenwärtig werden diese potentiellen Zielgene von unseren Kollegen am IPK Gatersle-

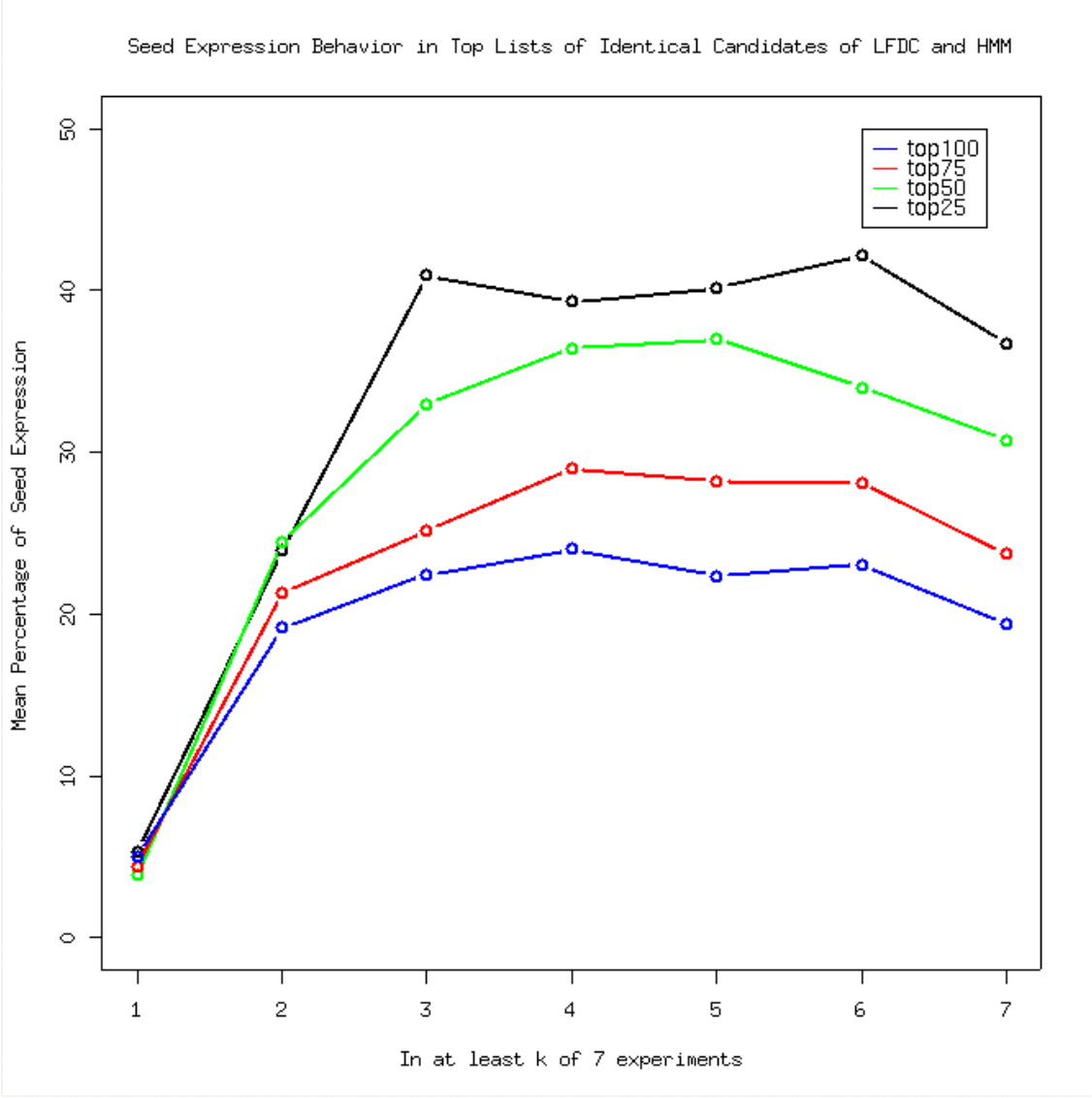

Abbildung 14: Samenexpression unter den besten 25, 50, 75 und 100 **ABI3** Zielgenen. Zufällig ausgewählte Gene zeigen dagegen nur eine durchschnittliche Samenexpression von rund 3%.

Abbildung 15: *G-Box* Motiv in Promotoren von potentiell durch **ABI3** regulierten Genen.

ben in unabhängigen Experimenten überprüft. Vorläufige Ergebnisse für sieben dieser Gene haben bereits jetzt bestätigt, dass **ABI3** an ihrer Regulation beteiligt ist.

2.1.5.5 Analyse von Sequenz-, Marker-, und Expressionsdaten

Konservierung von Genfunktionen in syntänen Regionen In Zusammenarbeit mit der Arbeitsgruppe Pflanzenstress und Entwicklung wurden Sequenzdaten, Markerdaten und Expressionsdaten von *Arabidopsis thaliana* und Tomate analysiert mit dem Ziel, syntäne Regionen von Genen zu identifizieren, die bei der Eisenaufnahme eine Rolle spielen. Die konkrete Frage lautete, ob die Gene, die bei der Eisenaufnahme in Tomate eine Rolle spielen, in Genomregionen liegen, die eine überdurchschnittlich hohe Syntänie mit dem Genom von *Arabidopsis thaliana* aufweisen. Zur Beantwortung dieser Frage wurde eine Pipeline entwickelt, die als Eingabe die folgenden Daten erhält: 1. Tomaten-ESTs, 2. Das Genom von *Arabidopsis thaliana*, 3. Tomaten-Marker inklusiver ihrer Kartierungspositionen und 4. Expressionsdaten von Tomate. Anhand der Expressionsdaten werden die Tomatengene in zwei Teilmengen unterteilt: Gene, die bei der Eisenaufnahme eine Rolle spielen, und solche, die dabei keine Rolle spielen. Anschließend werden die Tomaten-ESTs mit dem Genom von *Arabidopsis thaliana* aligniert. Anhand der Kartierungspositionen der Tomatenmarker und des berechneten Alignments mit dem Genom von *Arabidopsis thaliana* werden die syntänen Regionen zwischen beiden Genomen berechnet. Abschließend wird mit Hilfe von *Fisher's exact test* bestimmt, ob die Gene, die bei der Eisenaufnahme in Tomate eine Rolle spielen, signifikant häufiger in syntänen Region liegen als die Gene der Kontrollgruppe, die bei der Eisenaufnahme in Tomate eine Rolle spielen. Die Pipeline wurde so generisch entwickelt, das sie für analoge Analysen in anderen Organismen genutzt werden kann. Die Analysen der Tomatendaten ergaben, dass hier die Gene, die bei der Eisenaufnahme eine Rolle spielen, verstärkt in syntänen Regionen zwischen Tomate und *Arabidopsis thaliana* liegen [1]

2.1.5.6 Analyse von Sequenz-, Marker-, Passport-, Phänotyp- und Wetterdaten

Allele Mining Es wurde festgestellt, dass Resistenzen gegen das *Barley yellow mosaic virus (BaYMV)* und das *Barley mild mosaic virus (BaMMV)* mit der Sequenzvariabilität des Gersten-Gens Hv-elF-4E zusammenhängen. Bei einem Screening auf Polymorphismen in diesem Gen wurden 672 Akzessionen sequenziert und dabei 30 exonische Polymorphismen gefunden, die zu 44 Haplotypen führen. In Kooperation mit den Arbeitsgruppen Genomdiversität und Quantitative Evolutionäre Biologie wurden Algorithmen entwickelt, um erstens die Anzahl der Haplotypen in der Gesamtkollektion aller Gerstenakzessionen der IPK Genbank abzuschätzen und zweitens besonders aussichtsreiche Kandidaten für die weitere Sequenzierung vorzuschlagen.

Die Gesamtkollektion enthält ca. 21.000 Akzessionen, von denen für 13.799 Passport- und phänotypische Daten zur Verfügung stehen. Analysen unserer Kooperationspartner haben ergeben, dass populationsgenetische Modelle für diese Fragestellung ungeeignet sind. Daher wurden verschiedene alternative Modelle aus der Statistik und Sequenzanalyse implementiert (Abschnitt 2.1.5.7) und auf diese Fragestellung angewendet. Eine Kreuzvalidierung ergab, dass das derzeit erfolgreichste Modell ein Bayes Baum ist (Abbildung 16).

Dieses Modell sagt 270 $\pm$ 20 Haplotypen für eine Population von 13.799 Akzessionen vorher (Abbildung 17). Unter Hinzunahme von phänotypischen und Passportdaten wurde der Datensatz aufgeteilt, um dadurch die Vorhersagegenauigkeit weiter zu erhöhen. Es ergab sich, dass die Teilung der Population nach ihrer Anfälligkeit sowie ihrer Herkunft die

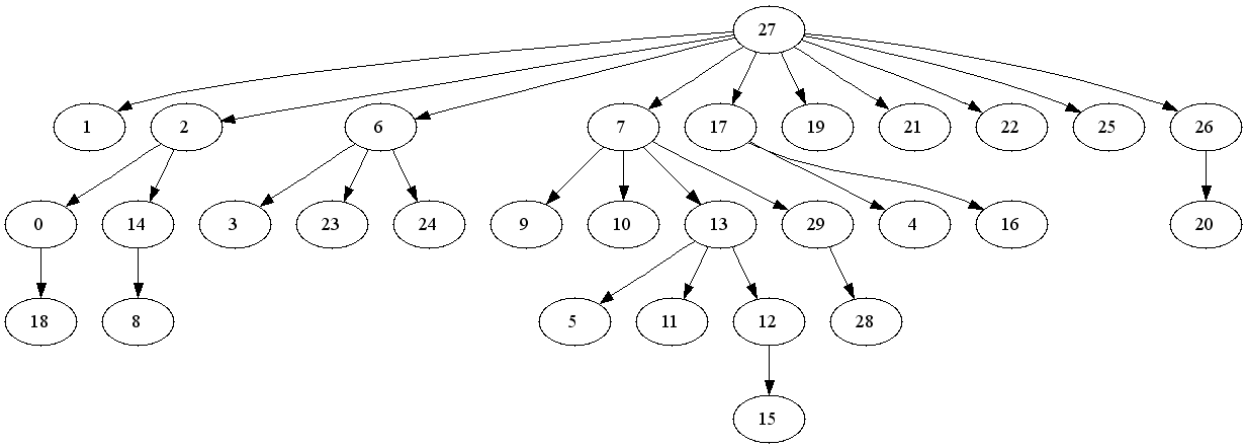

Abbildung 16: Struktur des Bayes Baums, der auf den Sequenzdaten der 672 Akzessionen trainiert wurde. Der Baum zeigt statistische Abhängigkeiten (Kanten) zwischen den verschiedenen polymorphen Positionen. Interessant sind die vielen Kanten zwischen nichtbenachbarten Positionen, die durch ein einfaches Kopplungsungleichgewicht zwischen benachbarten Positionen nur partiell modelliert werden können.

Vorhersagegenauigkeit am stärksten erhöhte. Da die Herkunft der Akzessionen in fast allen Fällen bekannt ist und damit leicht aus dem *Plant Data Warehouse* extrahiert werden kann, wurde diese Teilung genauer untersucht. Dabei stellte sich heraus, dass die Diversität des Hv-elF-4E Gens in den europäischen Akzessionen besonders gering ist, während sie in den ostasiatischen Akzessionen besonders hoch ist.

Passportdatenbrowser Passportdaten bilden einen wichtigen Teil der Daten über pflanzengenetische Ressourcen. Daher wurde eine Anwendung für das *Plant Bioinformatics Portal* entwickelt, um alle in den Passportmart des *Plant Data Warehouse* integrierten Passportdaten durchsuchen zu können. Dieser Passportdatenbrowser stellt eine wichtige Ergänzung zu den Weboberflächen der Europäischen Gerstendatenbank EBDB und der Europäischen Poadatenbank EPDB dar (Abschnitt 2.1.1). Insbesondere ermöglicht er die direkte Verknüpfung der Passportdaten mit den in den Phenomart integrierten Charakterisierungs- und Evaluierungsdaten.

Wetterdaten Die Analyse von phänotypischen Daten im Kontext von molekularen Daten sowie Passport- und Umweltdaten rückt verstärkt in den Fokus der Pflanzenzüchter. Bei der Analyse von Korrelation zwischen molekularen Daten und phänotypischen Daten ist es wichtig, äußere Einflüsse, wie z. B. Wettereinflüsse, zu berücksichtigen. Um diese Wettereinflüsse mathematisch adäquat berücksichtigen zu können, ist es notwendig, die Korrelationen innerhalb der Wetterdaten modellieren zu können. Bei der Untersuchung von Korrelationen innerhalb der Wetterdaten stießen wir auf zwei interessante Beobachtungen, die durch keines der bekannten stochastischen Modelle reproduziert werden konnten. In Zusammenarbeit mit Prof. Podobnik, Universität Zagreb, gelang es uns, für jedes der beiden beobachten Phänomene ein stochastisches Modell zu entwickeln [8, 9].

Diversity Studies Toolkit In Kooperation mit der Arbeitsgruppe Genomdiversität wurde eine Pipeline zur Analyse von Marker-, Phänotyp- und Passportdaten entwickelt. Das *Diversity Studies Toolkit* nutzt für Berechnungen die Statistiksprache R und bietet

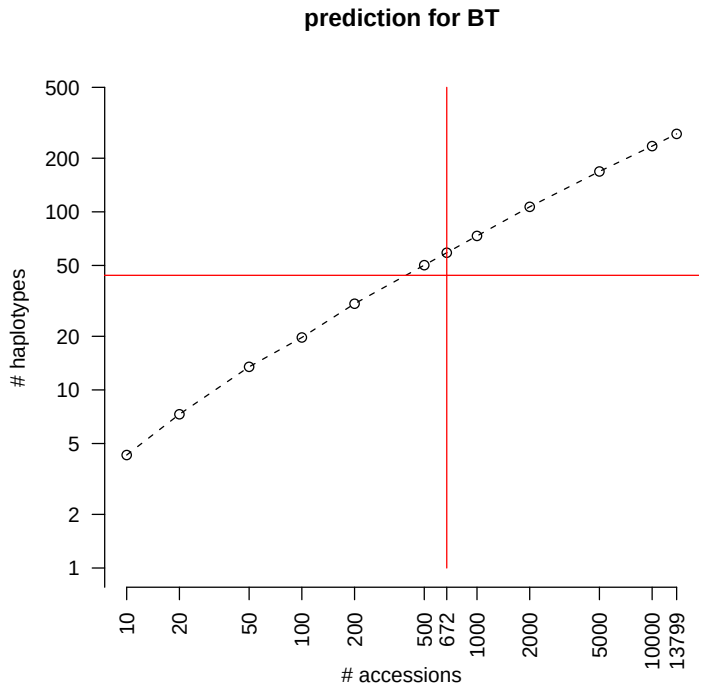

Abbildung 17: Erwartungswert der Anzahl der Haplotypen als Funktion der Anzahl der Akzessionen. Die Kurve wächst monoton und erreicht einen Wert von 270 ± 20 Haplotypen für eine Population von 13.799 Akzessionen. Die Standardabweichung von 20 ergibt sich aus Extrapolationen der Kreuzvalidierungen.

generische Suchformulare zum Zusammenstellen und Klassifizieren von Genbankobjekten, wie Akzessionen, Partien oder Aufwüchse. *DiSTo* bietet derzeit die folgenden Funktionen:

1. Deskriptive Statistiken über eine Auswahl von genetischen, phänotypischen oder Passportattributen
2. Ermittlung populationsgenetischer Parameter durch AMOVA (Analysis of Molecular VAriances)
3. 3. Berechnung von genetischen Abständen, deren Visualisierung sowie Clusterung und Rekonstruktion genetischer Bäume
4. Multivariate Analyse durch Hauptkomponentenanalyse und die 2D- bzw. 3D-Visualisierung der Ergebnisse

Pyrosequenzierungsdaten zu einer 2906 Akzessionen umfassenden Lolium Kollektion wurden aus der *Pyrosequencing Datenbank PSQDB* in den Markermart des *Plant Data Warehouse* integriert. Des weiteren wurden die dazugehörigen Passport- und Charakterisierungs- und Evaluierungsdaten aus dem *Genbank Informationssystem GBIS* in den Passportmart und Phenomart integriert. Der Diversitätsmart wurde erstellt und mit einer strukturierten Teilmenge dieser Daten befüllt, um komplexe und domänenübergreifende Analysen dieser Daten zu ermöglichen.

Im Rahmen des Teilprojektes *Einsatz molekularer Marker zur Eliminierung von Duplikaten* innerhalb des Projektes *Aufbau einer bundeszentralen* ex situ *Genbank für landwirtschaftliche und gartenbauliche Kulturpflanzen: Zusammenführung der Genbanken des IPK*

Gatersleben und der BAZ Braunschweig wurde *DiSTo* intensiv genutzt, um die Pyrosequencing Daten gemeinsam mit den Passport- und Phänotyp-Daten der Lolium Sammlung zu analysieren.

Der Diversitätsmart und *DiSTo* wurden so entworfen, dass sie in Zukunft artenübergreifend zur integrativen Analyse von Pyrosequenzierungs-, Passport- und Evaluierungsdaten genutzt werden können.

Knoblauchkernkollektion Die Gattung Allium umfasst ca. 600 Arten, deren Verbreitungsgebiet sich über die Holarktis, Paleotropis und Neotropis erstreckt. Am IPK Gatersleben wurde während der vergangenen 20 Jahre eine umfassende Lebendsammlung etabliert, die die Grundlage für die morphologische, anatomische, entwicklungsphysiologische, karyologische und molekulare Untersuchung vieler Arten bildet. Innerhalb der Kulturarten-Sammlung nimmt neben der Küchenzwiebel der Knoblauch (*Allum sativum L.*) eine vorrangige Stellung ein. Wegen seiner ausschließlich vegetativen Erhaltungsweise muss diese *Allium*-Art in klonalen Akzessionen gehalten werden, die besonders hohen Betreuungsaufwand erfordern, da sie nicht als Saatgut gelagert werden können. Ihre gute Charakterisierung ist deshalb besonders wichtig. Von den 540 Knoblauch Akzessionen, die derzeit am IPK betreut werden, wurden im Rahmen eines EU-Projektes 124 Akzessionen herausgestellt, die unter dem Gesichtspunkt einer möglichst guten Repräsentanz der Gesamtsammlung nun als Kernkollektion des IPK bezeichnet wird (Abbildung 18).

Um diese am IPK Gatersleben geführte *Garlic Core Collection* im Internet zu präsentieren, wurden Bilddaten der Arbeitsgruppe In vitro Erhaltung und Cryo-Lagerung zu *Allium sativum* mit den dazugehörigen Passport- und Evaluierungsdaten in das *Plant Data Warehouse* integriert. In Zusammenarbeit mit den Arbeitsgruppen Invitro Erhaltung und Cryo-Lagerung, Bioinformatik, Genbankdokumentation und Genomdiversität sowie der Genbankaußenstelle Malchow wurde eine webbasierte Anwendung entwickelt und in das *Plant Bioinformatics Portal* integriert, die dem Nutzer die Eingabe von Akzessionsnummern in ein Textfeld oder die Auswahl von Akzessionen aus einer Auswahlliste ermöglicht (Abbildung 19).

Als Resultat werden Passport- und Evaluierungsdaten der Akzessionen gelistet. Darüber hinaus werden Thumbnail-Bilder verschiedener morphologischer Teile der Knoblauchpflanzen dargestellt (Abbildung 20). Die Bilder sind in die folgenden Kategorien eingeteilt: *outside view of plants*, *bulbs and cloves*, *bulb structure*, *field pictures*, *inflorescence and bulbils* und *other pictures*. Die Details werden per Mausklick in höherer Auflösung zur Verfügung gestellt. Außerdem wird die chronologische Abfolge der Ontogenese von Infloreszenzen und Bulbillen dargestellt.

2.1.5.7 Übergreifende Anwendungen

Cluster Execution Framework Viele Bioinformatikanwendungen sind sehr rechenintensiv. Ihre Laufzeit würde auf einem einzigen Prozessor oft Monate übersteigen. Zur Berechnung solch rechenintensiver Jobs steht am BIC-GH der im Rahmen des *Plant Data Warehouse* Projektes beschaffte Linuxcluster zur Verfügung. In Kooperation mit der Arbeitsgruppe Bioinformatik wurde für das *Plant Bioinformatics Portal* ein Softwaresystem, das *Cluster Execution Framework*, entwickelt, welches es ermöglicht, die über das *Plant Bioinformatics Portal* gestarteten Anwendungen auf dem Linuxcluster laufen zu lassen. Dieses Softwaresystem ermöglicht außerdem die datenbankgestützte Verwaltung aller Jobs der verschiedenen Bioinformatikanwendungen über eine webbasierte Nutzeroberfläche. Weiterhin

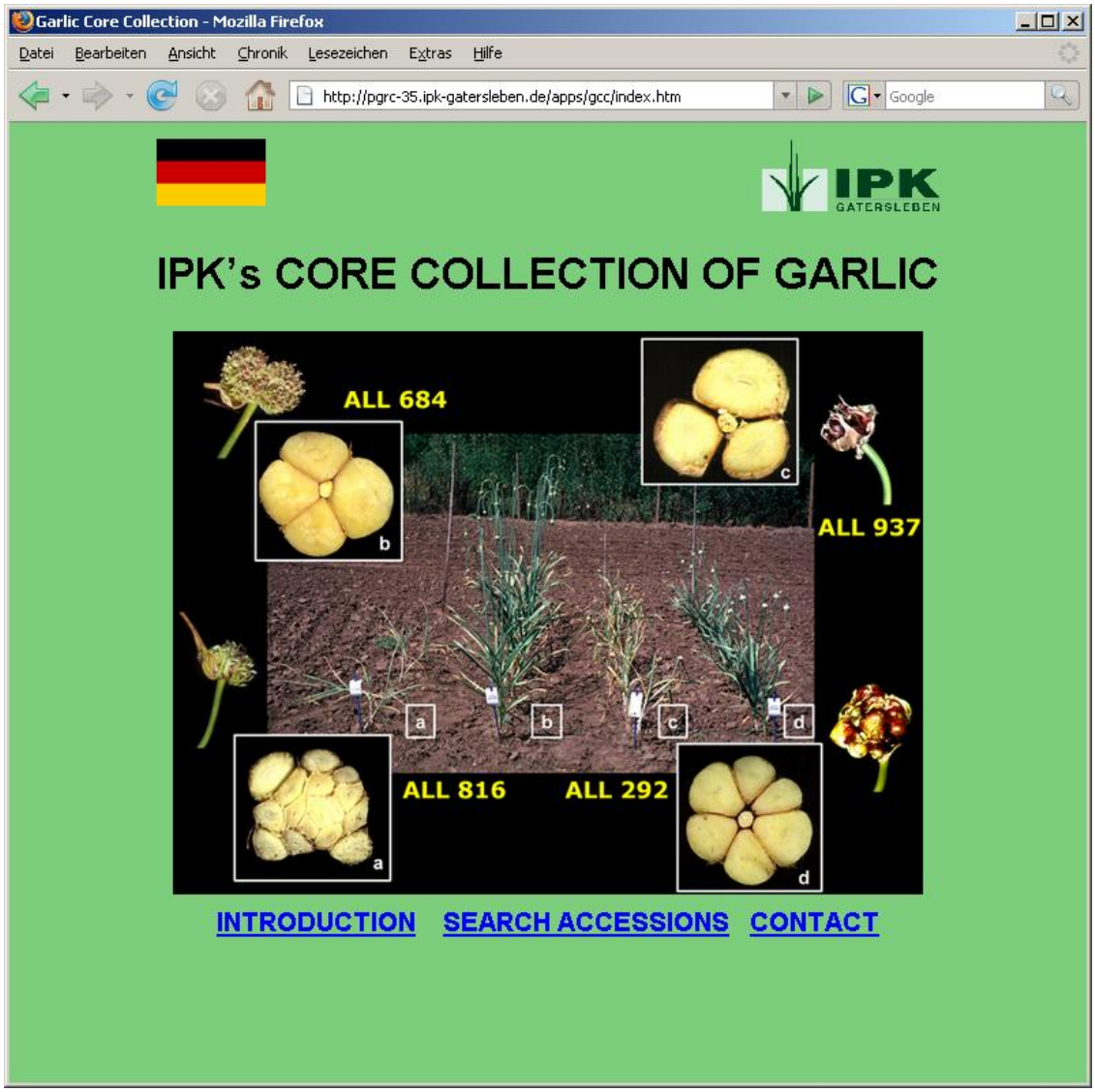

Abbildung 18: Startseite der *Garlic Core Collection* im *Plant Bioinformatics Portal*.

wurden *SOAP Web Services* entwickelt, die den Zugriff auf viele Funktionen des Systems ermöglichen. Die generische Nutzeroberfläche bietet ohne weiteren Mehraufwand Eingabe- und Steuermasken für neue Anwendungen, die über eine Verwaltungskomponente eingebunden werden können. Das *Cluster Execution Framework* ermöglicht z. B. die Berechnung von Blast und MEME sowie Blat, GeneSeqer, Primer3, Sim4 und Spidey.

Java Framework Viele der erhobenen *Use Cases* erfordern Datenanalysen mit Hilfe von graphischen Modellen. Dafür wurde in Zusammenarbeit mit der Arbeitsgruppe Bioinformatik der Martin-Luther-Universität Halle-Wittenberg ein *Java Framework* erstellt, das die Implementierung, Testung und Nutzung dieser Modelle ermöglicht. Durch die Nutzung der Programmiersprache *Java* ist es möglich, das *Framework* plattformunabhängig zu nutzen.

Neben den häufig benötigten Modellen der Sequenzanalyse, wie z. B. homogenen und inhomogenen Markov Modellen, *Permuted Markov* Modellen, Bayesnetzen oder *Maximum Entropie* Modellen, enthält das *Java Framework* auch allgemeine Implementierungen für deren Mischmodelle und Klassifikatoren. Des weiteren wurden verschiedene Lernalgorithmen für all diese Modelle und Modellkombinationen implementiert sowie verschiedene Möglich-

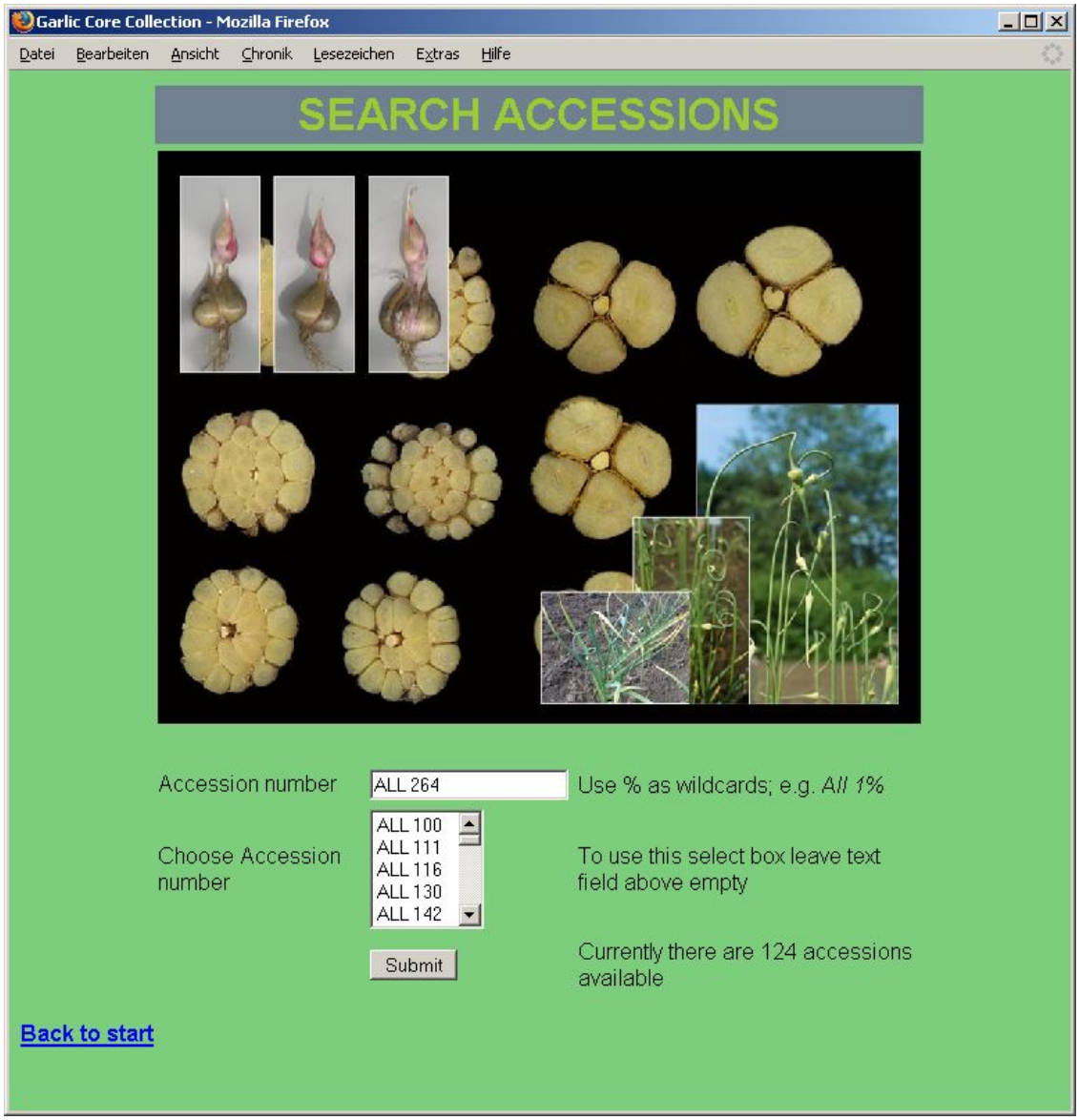

Abbildung 19: Formular für die Suche nach Akzessionsnummern von Knoblauch-Akzessionen der *Garlic Core Collection*.

keiten zur Auswertung von Klassifikatoren (Kreuzvalidierung, *stratified holdout sampling*, verschiedene Gütemaße, etc.). Die allgemeine Implementierung erlaubte und erlaubt es, mit sehr geringem Aufwand neue Problemstellungen zu bearbeiten, indem die vorhandenen Klassen miteinander kombiniert werden.

Das *Java Framework* konnte bereits innerhalb der Projektlaufzeit erfolgreich für verschiedene Anwendungsfälle eingesetzt werden. Viele der im Folgenden genannten Analyseprogramme konnten nur mit Hilfe dieses *Java Framework* effizient implementiert werden. Darüber hinaus ermöglichte das *Java Framework* auf dem Gebiet der Diversitätsforschung die für die IPK Genbank sehr wichtige Berechnung der erwarteten Anzahl von Haplotypen des eIF4E-Gens in der mehr als 20.000 Akzessionen umfassenden Gerstenkollektion des IPK Gatersleben.

Diskriminative Lernverfahren Initiiert durch den Gastaufenthalt von Jesús Cerquides, Universität Barcelona, am BIC-GH im Sommer 2006 und durch moderne Entwicklungen

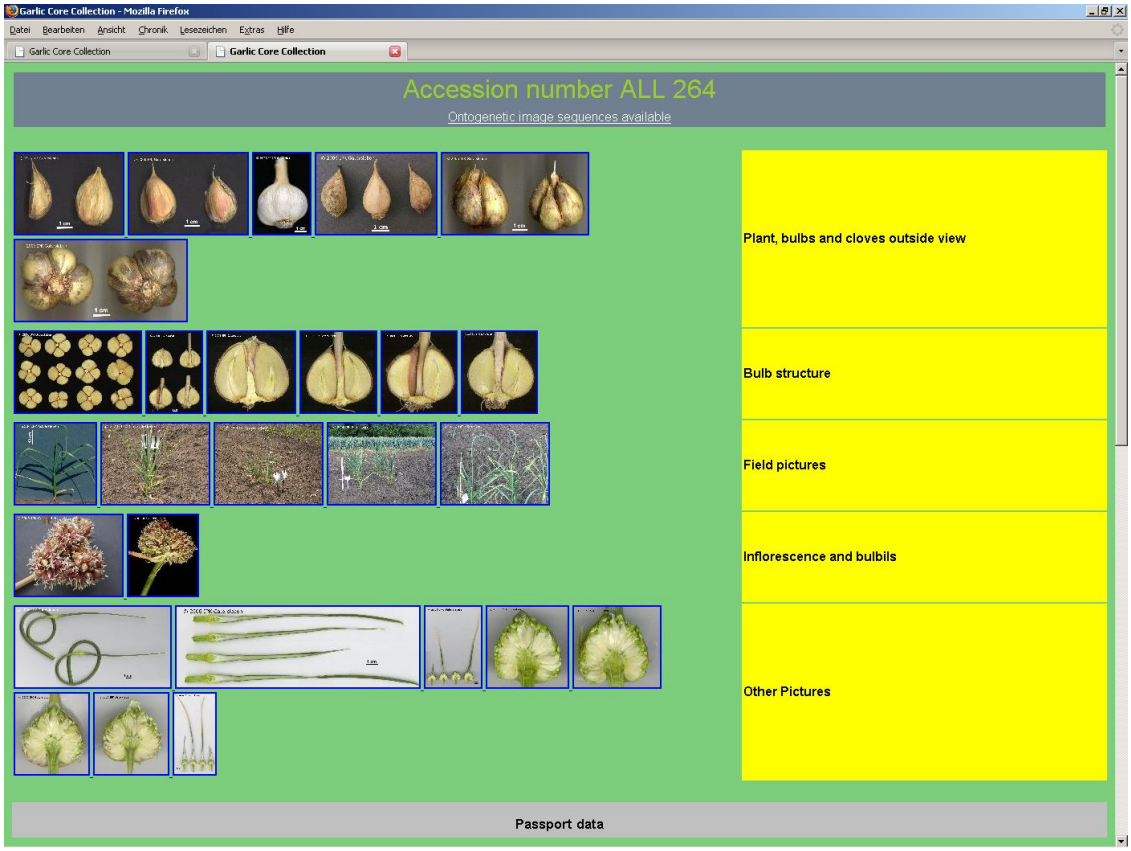

Abbildung 20: Ergebnis der Suche mit Darstellung der Thumbnail-Bilder und Angabe von Passport- und Evaluierungsdaten der Knoblauch-Akzessionen.

auf dem Gebiet des maschinellen Lernens wurden Methoden des diskriminativen Lernens für die im *Java Framework* implementierten Modelle entwickelt. Diese Verfahren, wie z. B. das *Maximum Conditional Likelihood* Verfahren oder das *Supervised Posterior* Verfahren, erfordern einen weit höheren Rechenaufwand als z. B. das *Maximum Likelihood* Verfahren oder das *Maximum A Posteriori* Verfahren, liefern aber dafür signifikant genauere Vorhersagen [21, 22, 23]. Abbildung 21 zeigt die höhere Genauigkeit des *Maximum Conditional Likelihood* Verfahrens bei der Erkennung von Spleißstellen im Vergleich zum *Maximum Likelihood* Verfahren und zum derzeit weltbesten Verfahren basierend auf *Maximum Entropie* Modellen. Abbildungen 22 und 23 zeigen das mit Hilfe des *Supervised Posterior* Verfahrens gefundene Sequenzlogo sowie die Unterschiede zum Sequenzlogo, das mit dem herkömmlichen *Maximum A Posteriori* Verfahren gefunden wurde. Durch die Integration dieser Lernverfahren in das *Java Framework* stehen sie nun allen Nutzern frei zur Verfügung.

2.1.6 AP6 Erstellung des *Plant Bioinformatics Portals* als zentrale Präsentationsplattform des BIC-GH und aller in das *Plant Data Warehouse* integrierten Anwendungen

Das *Plant Bioinformatics Portal* (http://www.bic-gh.de) bildet den zentralen Einstiegspunkt für Nutzer des *Plant Data Warehouse*. Es wurde in Kooperation mit unserem Industriepartner B.I.M.-Consulting entwickelt. Zum Einsatz kam dabei die Oracle Portal Technologie. Die detaillierte Vorstellung des *Plant Bioinformatics Portal* erfolgt im Abschlussbericht unseres Industriepartners.

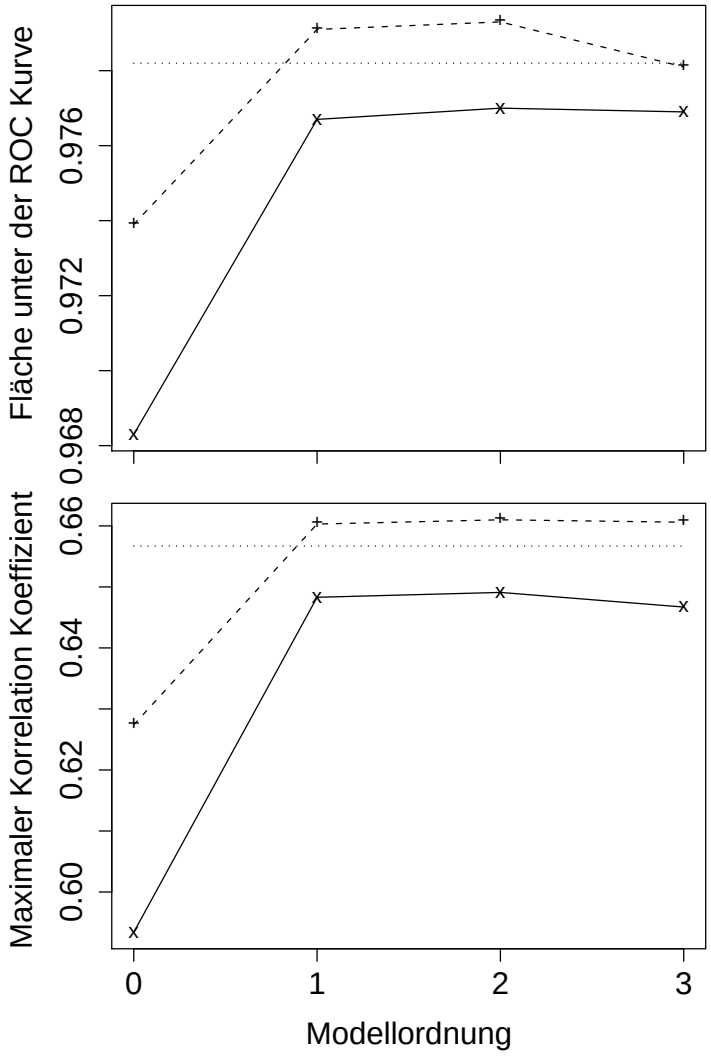

Abbildung 21: Genauigkeit der Klassifikation von Spleißenstellen, quantifiziert durch die Fläche unter der ROC Kurve (oben) und den maximalen Korrelationskoeffizienten (unten), als Funktion der Modellordnung. Die durchgezogenen Kurven zeigen die Klassifikationsgenauigkeiten für Modellkombinationen, deren Parameter mit dem klassischen *Maximum Likelihood* Lernverfahren bestimmt wurden. Die gestrichelten Kurven zeigen die Klassifikationsgenauigkeiten für Modellkombinationen, deren Parameter mit dem diskriminativen *Maximum Conditional Likelihood* Lernverfahren bestimmt wurden. Die horizontalen Linien zeigen die Klassifikationsgenauigkeiten des derzeit weltbesten Modells zur Erkennung von Spleißstellen. Man erkennt, dass unabhängig vom gewählten Maß der Genauigkeit der Klassifikation das diskriminative *Maximum Conditional Likelihood* Lernverfahren in allen Fällen eine signifikante Verbesserung der Klassifikation von Spleißstellen gegenüber dem klassischen *Maximum Likelihood* Lernverfahren liefert . Man erkennt außerdem, dass das diskriminative *Maximum Conditional Likelihood* Lernverfahren eine genauere Klassifikation als das derzeit weltbeste Modell zur Erkennung von Spleißenstellen ermöglicht.

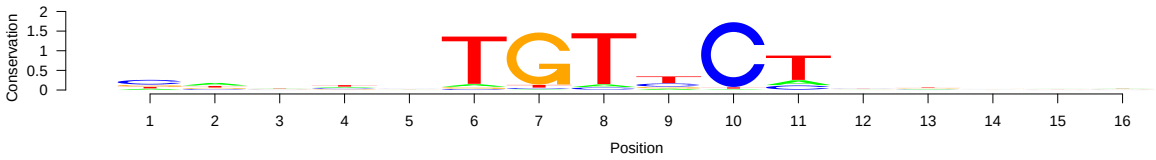

Abbildung 22: Sequenzlogo des diskriminativ gelernten *PWM* Modells des Transkriptionsfaktors **AR/GR/PR**. Die konservierten Positionen treten deutlicher hervor als im Sequenzlogo des generativ gelernten *PWM* Modells, was zu einer genaueren Erkennung von **AR/GR/PR** Bindungsstellen in genomweiten Motiv-Suchen führt.

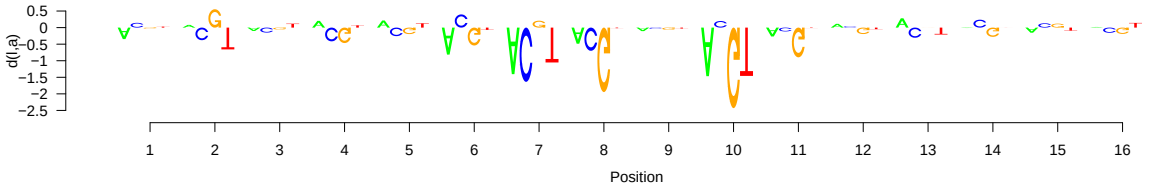

Abbildung 23: Differenz des neuen (diskriminativ gelernten) und alten (generativ gelernten) Sequenzlogos. Man erkennt deutlich die Positionen, an denen sich das neue von dem alten Sequenzlogo unterscheidet. Interessanterweise sind die Unterschiede im Kern der Bindungsstelle, d. h. im Bereich der größten Konserviertheit, am stärksten.

2.1.7 AP7 Versionskontrolle

Es ist voranzustellen, dass Daten aus Operativsystemen des IPK Gatersleben und IPB Halle sowie aus öffentlichen Datenbanken keiner Versionskontrolle des *Plant Data Warehouse* unterliegen. Im *Plant Data Warehouse* wird beim Import-Prozess die jeweilige Version dokumentiert. Es wird davon ausgegangen, dass bei öffentlichen Quellen ein Zugriff auf die entsprechende Version jederzeit beim Anbieter des Systems möglich ist. Für die Verwaltung dieser Versionsbeschreibungen kommt das *Meta Data Repository* zum Einsatz. Eine Beschreibung des *Meta Data Repository* erfolgt im Abschlussbericht unseres Industriepartners B.I.M.-Consulting mbH. Für die Analyseprogramme wird das Standardversionierungssystem CVS verwendet.

2.1.8 AP8 Kurations-System

Für die Datenkuration wurde *Feedbacksystem* mit unserem Industriepartner entwickelt. Dieses ist direkt im *Plant Bioinformatics Portal* integriert und ermöglicht die Kommunikation der Anwender mit den Anwendungsentwicklern bzw. Portalbetreibern. Innerhalb des *Feedbacksystem* können je nach Anfrage- bzw. Kurationswunsch (z. B. Fehlermeldung für ein Analysewerkzeug bzw. Probleme mit den im *Plant Data Warehouse* integrierten Daten) die eingehenden Anfragen verschiedenen Bearbeitern zugewiesen werden. Der gesamte Prozess wird im Kurations-System dokumentiert und ist so jederzeit nachvollziehbar. Als Reaktion erfolgt eine Bereinigung der Daten innerhalb des *Plant Data Warehouse*. Eine detaillierte Beschreibung des Kurations-Systems liegt im Abschlussbericht unseres Industriepartners vor.

2.1.9 AP9 Synchronisation und Dokumentation

Die Datenimports in das *Plant Data Warehouse* sind im Dokumentationsportal detailliert beschrieben. Parallel dazu sind die Prozeduren im *Oracle Warehouse Builder* gespeichert. Die Nutzung dieser Quellen ermöglicht eine Synchronisation beim Datenimport.

Alle entwickelte Anwendungen wurden dokumentiert. Konkret wurden alle *Java* Klassen inklusive aller Methoden, Konstruktoren und Variablen im Quelltext dokumentiert. Alle Klassen wurden in *Java Packages* zusammengefasst, und jedes *Package* wurde in englischer Sprache dokumentiert. Durch *javadoc* können alle Dokumentationen in ein HTML-lesbares Format umgewandelt werden. Die dadurch generierten HTML-Seiten enthalten alle Informationen über die einzelnen Klassen, die *Packages*, die Ableitungshierarchie sowie einen Gesamtüberblick und einen Index. Um eine nachhaltige Nutzung der Dokumentation zu gewährleisten, wurden alle Dokumentationen in englischer Sprache verfasst.

Alle Quelltexte und Dokumentationen liegen im zentralen *CVS Repository* des BIC-GH. Dies ermöglicht eine parallele Weiterentwicklung aller Anwendungen durch mehrere Programmierer bzw. Programmierteams. Außerdem erlaubt es eine Versionskontrolle und das Verfolgen der *History*. Zudem wurde ein *Wiki* aufgesetzt, welches den Programmieren und Anwendern eine Diskussionsplattform bietet, um ihre Fragen und Probleme effektiv zu kommunizieren sowie auf Fragen und Probleme anderer Programmierer und Anwender zu antworten.

Weiterhin wurden alle Datenintegrationsprozesse dokumentiert und alle Abhängigkeiten erfasst. Für die hauptsächlich biologischen Nutzer wurden detaillierte Beschreibungen der Funktionalitäten des *Plant Data Warehouse*s sowie der Prozessflüsse in Form eines *User Guide* erstellt Für die Sicherstellung des Weiterbetriebs sowie mögliche Erweiterungen wurde ein *Developer Guide* erstellt. Des weiteren wurden Tutorien entwickelt, die den Nutzern eine intuitive Dokumentation anbieten. Zur interaktiven Benutzung der Anwendungen wurden ebenfalls *Wizards* entwickelt, die die Führung durch alle Analyseprozesse ermöglichen und dem Anwender Vorschläge über mögliche weitere Schritte innerhalb eines Analyseprozesses unterbreiten.

2.1.10 AP10 Backup-Strategie

Alle Daten, die im *Plant Data Warehouse* integriert sind, werden auf dem zentralen Datenbankserver des BIC-GH gespeichert. Dieser ist in das zentrale Backup-System des IPK Gatersleben integriert. Neben der Sicherung der Datenbank sind alle weiteren Server, die für den Betrieb des *Plant Bioinformatics Portal* verantwortlich sind, in die Datensicherung des IPK Gatersleben eingebunden. Die Administrations-Verantwortlichkeiten liegen dabei bei hausinternem Personal, was eine langfristige Datensicherung auch in Zukunft ermöglicht.

Die entwickelte Datensicherungsstrategie hat sich in der Praxis bereits bewährt. Sowohl für die Datenbank als auch für den *Application Server* wurde aufgrund von Hardwaredefekten eine der Daten aus dem Backup notwendig. Bei allen in der Vergangenheit aufgetretenen Hardwaredefekten wurde die Datenrücksicherung erfolgreich von den Systemadministratoren realisiert.

2.1.11 AP11 Wartungsplan

Um die Wartung des *Plant Data Warehouse* nach Auslaufen der Förderphase abzusichern, wurden alle Datenbestände in die zentrale Oracle-Datenbank des IPK Gatersleben migriert (Abschnitt 2.1.10), deren Administration durch Personal des IPK Gatersleben abgesichert

Abbildung 24: Teilansicht des in englischer Sprache dokumentierten *Java Framework*.

ist. Durch die vorhandene Dokumentation ist es dem vorhandenem Personal möglich, die Wartungsmaßnahmen sowie neue Datenimporte durchzuführen.

Die Pflege der Serversysteme auf Betriebssystemebene wird ebenfalls durch Personal des IPK Gatersleben abgedeckt. Für die Wartung der Oracle Portal Software ist aktuell ein externer Serviceprovider verantwortlich.

2.1.12 Abschlussbemerkungen

Die Arbeiten der Arbeitsgruppe *Plant Data Warehouse* wurden bislang in mehreren Artikeln in wissenschaftlichen Zeitschriften und referierten Konferenzbänden publiziert. Besonders hervorzuheben ist, dass die Arbeiten nicht nur mit anderen Bioinformatik-Gruppen, sondern

in vielen Fällen mit biologischen Anwendern entstanden sind. Diese Publikationen zeigen, dass sich das *Plant Data Warehouse* als eine Integrationsplattform für pflanzliche Daten und das *Plant Bioinformatics Portal* mit den integrierten Anwendungen als ideale Plattform für die tägliche Arbeit der biologischen Anwender entwickelt hat. Das *Plant Bioinformatics Portal* wurde seit der Freischaltung am 23. März 2006 20.575 mal durch Anwender aus 64 Ländern genutzt.

2.2 Voraussichtlicher Nutzen, Verwertbarkeit

Das wissenschaftlich-technische Hauptergebnis des Vorhabens ist die Entwicklung des *Plant Data Warehouse*, welches domänenübergreifende Analysen von Sequenz-, Expressions-, und Metabolomdaten, taxonomischen, genotypischen und phänotypischen Daten sowie Daten zu pflanzengenetischen Ressourcen ermöglicht. Diese Analysen sind von großem Nutzen für die Biologen und Züchter auf dem Gebiet der Kulturpflanzenforschung weltweit als auch für die Wissenschaftler am IPK Gatersleben und IPB Halle.

Das *Plant Data Warehouse* erlaubt beispielsweise (i) Diversitätsstudien von Lolium, Poa oder anderen Nutzpflanzen durch Integration von Pyrosequenzierungs-, Passport- sowie Charakterisierungs- und Evaluierungsdaten, (ii) die integrative Analyse von Sequenz- und Markerdaten, die gegenwärtig in verschiedenen Großprojekten auf dem Gebiet der Pflanzengenomforschung produziert werden, oder (iii) die Analyse von Sequenz-, Expressions- und ChIP/chip-Daten von *Arabidopsis thaliana* und anderen vollständig sequenzierten Pflanzengenomen.

Die öffentliche Freigabe des *Plant Data Warehouse* und der darin integrierten Analysesoftware zur kostenlosen Nutzung durch alle Wissenschaftler der Welt hat dazu geführt, dass das *Plant Data Warehouse* inzwischen von mehr als 1000 Nutzern pro Monat verwendet wird. Des weiteren werden viele der für das *Plant Data Warehouse* entwickelten Anwendungen, z. B. auf dem Gebiet der Erkennung cis-regulatorischer Elemente, inzwischen weltweit genutzt. Eine Kommerzialisierung des *Plant Data Warehouse* und seiner Komponenten ist derzeit nicht geplant.

2.3 Fortschritte bei anderen Stellen

Auf dem Gebiet der Entwicklung einzelner Anwendungen zur Analyse und Visualisierung von Hochdurchsatzdaten der modernen Genom- und Postgenomforschung hat es in den letzten fünf Jahren Fortschritte durch viele verschiedene Gruppen gegeben. Diese können hier im einzelnen nicht aufgezählt, sondern nur an Hand der folgenden vier Beispiele illustriert werden.

Die automatische Annotation von Spleißstellen ist von großem Interesse nicht nur auf dem Gebiet der Pflanzengenomforschung. Durch Gene Yeo und Chris Burge vom Massachusetts Institute of Technology, Cambridge, USA, wurde ein neues Modell zur Erkennung von Spleißstellen entwickelt und in der Fachzeitschrift *Journal of Computational Biology* im Jahr 2004 publiziert. Das *Maximum Entropie* Modell erlaubt statistische Abhängigkeiten auch zwischen nicht-benachbarten Positionen einer Spleißstelle, was zu einer Verbesserung der computergestützten Vorhersage von Spleißstellen führt. Dieses Modell konnte jedoch zum einen durch die Verwendung diskriminativer Lernverfahren (Abschnitt 2.1.5.7) und zum anderen durch die Entwicklung von *Maximum Entropie* Mischmodellen und des *BGIS* Algorithmus (Abschnitt 2.1.5.1) verbessert werden [23].

Zur Erkennung kurzer DNA Motive, wie z. B. Transkriptionsfaktorbindungsstellen, haben Xiaoyue Zhao und Kollegen von der University of California, Berkeley, USA, ein neues stocha-

stische Modell entwickelt und im Jahr 2005 in der Fachzeitschrift *Journal of Computational Biology* publiziert. Das *Variable Order Permuted Markov* Modell stellt eine Verallgemeinerung der *Variable Order Markov* Modelle und *Permuted Markov* Modelle dar und ermöglicht eine genauere Erkennung von kurzen DNA Motiven als andere bis dahin entwickelte Modelle. Die im Rahmen des *Plant Data Warehouse* Projektes entwickelten VOB Modelle (Abschnitt 2.1.5.1) gehen jedoch über dieses Modell hinaus und ermöglichen eine weitere Steigerung der Genauigkeit der Erkennung kurzer DNA Motive [5, 10, 14].

Auf dem Gebiet der Analyse von *arrayCGH* Daten im Kontext genomischer Nachbarschaft wurde durch Oscar Rueda und Ramon Diaz-Uriarte vom Spanish National Cancer Centre, Madrid, Spanien, der *RJaCGH* Algorithmus entwickelt und im Jahr 2007 in der Fachzeitschrift *PLoS Computational Biology* publiziert. Dieser Algorithmus stellt eine Weiterentwicklung der *Hidden Markov* Modelle dar und wurde spezifisch für die Erkennung von Segmentduplikationen optimiert, reicht jedoch nicht an die im Rahmen des *Plant Data Warehouse* Projektes entwickelten Methoden zur Analyse von Expressionsdaten und ChIP/chip-Daten im Kontext genomischer Nachbarschaft heran (Abschnitt 2.1.5.4).

Ein neuer Algorithmus zur diskriminativen Suche von Sequenzmotiven in DNA- und Proteinsequenzen wurde durch Emma Redhead und Timothy Bailey von der University of Queensland, Brisbane, Australien im Oktober 2007 in der Fachzeitschrift *BMC Bioinformatics* publiziert. Dieser Algorithmus wendet das *Maximum Conditional Likelihood* Verfahren (Abschnitt 2.1.5.7) zur Vorhersage von *de-novo* Transkriptionsfaktorbindungsstellen (Abschnitt 2.1.5.1) an. Der *EMMA* Algorithmus ist jedoch in verschiedener Hinsicht flexibler als der *DEME* Algorithmus. So erlaubt der *EMMA* Algorithmus unter anderem die Suche nach cis-regulatorischen Modulen sowie die Verwendung von komplexeren Modellen für die DNA Motive und den Sequenzhintergrund, von sequenz-spezifischen Wahrscheinlichkeiten für das Vorhandensein von Transkriptionsfaktorbindungsstellen und von Strang- und Positionspräferenzen (Abschnitte 2.1.5.1 und 2.1.5.7).

Im Gegensatz zu diesen und vielen weiteren Fortschritten bei der Entwicklung einzelner Anwendungen ist auf Gebiet der integrativen Analyse von molekularen Daten und Daten zu pflanzengenetisch Ressourcen kein dem *Plant Data Warehouse* vergleichbares System auf dem Gebiet der Kulturpflanzenforschung entstanden.

2.4 Veröffentlichungen

Referierte Artikel in Zeitschriften

[1] Petra Bauer, Thomas Thiel, Marco Klatte, Zsolt Bereczky, Tzvetina Brumbarova, Rudiger Hell, and Ivo Grosse. “Analysis of Sequence, Map Position, and Gene Expression Reveals Conserved Essential Genes for Iron Uptake in Arabidopsis and Tomato”. In: *Plant Physiology* 136.4 (2004). Pp. 4169–4183. DOI: 10.1104/pp.104.047233.

[2] Wentian Li, Fengzhu Sun, and Ivo Grosse. “Extreme Value Distribution Based Gene Selection Criteria for Discriminant Microarray Data Analysis Using Logistic Regression”. In: *Journal of Computational Biology* 11.2–3 (2004). Pp. 215–226. DOI: 10.1089/1066527041410445.

[3] Dang D. Long, Ivo Grosse, and Kenneth A. Marx. “Coding and non-coding DNA thermal stability differences in eukaryotes studied by melting simulation, base shuffling and DNA nearest neighbor frequency analysis”. In: *Biophysical Chemistry* 110.1–2 (2004). Pp. 25–38. DOI: 10.1016/j.bpc.2004.01.001.

[4] Thomas Thiel, Raja Kota, Ivo Grosse, Nils Stein, and Andreas Graner. "SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development". In: *Nucleic Acids Research* 32.1 (2004). e5. DOI: 10.1093/nar/gnh006.

[5] Irad Ben-Gal, Ayala Shani, Andre Gohr, Jan Grau, Sigal Arviv, Armin Shmilovici, Stefan Posch, and Ivo Grosse. "Identification of transcription factor binding sites with variable-order Bayesian networks". In: *Bioinformatics* 21.11 (2005). Pp. 2657–2666. DOI: 10.1093/bioinformatics/bti410.

[6] Andreas Graner, Thomas Thiel, Hangning Zhang, Elena Potokina, Manoj Prasad, Dragan Perovic, Raja Kota, Rajeev K. Varshney, Uwe Scholz, Ivo Grosse, and Nils Stein. "Molecular mapping in barley: shifting from the structural to the functional level". In: *Czech Journal of Genetics and Plant Breeding* 41.3 (2005). Pp. 81–88. URL: http://www.cazv.cz/default.asp?ch=54&typ=1&val=40112&ids=2747.

[7] Christian Künne, Matthias Lange, Thomas Funke, Heiko Miehe, Thomas Thiel, Ivo Grosse, and Uwe Scholz. "CR-EST: a resource for crop ESTs". In: *Nucleic Acids Research* 33.suppl_1 (2005). Pp. D619–621. DOI: 10.1093/nar/gki119.

[8] Boris Podobnik, Plamen C. Ivanov, Katica Biljakovic, Davor Horvatic, H. Eugene Stanley, and Ivo Grosse. "Fractionally integrated process with power-law correlations in variables and magnitudes". In: *Physical Review E* 72 (2005). P. 026121. DOI: 10.1103/PhysRevE.72.026121.

[9] Boris Podobnik, Plamen C. Ivanov, Vojko Jazbinsek, Zvonko Trontelj, H. Eugene Stanley, and Ivo Grosse. "Power-law correlated processes with asymmetric distributions". In: *Physical Review E* 71 (2005). 025104(R). DOI: 10.1103/PhysRevE.71.025104.

[10] Jan Grau, Irad Ben-Gal, Stefan Posch, and Ivo Grosse. "VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees". In: *Nucleic Acids Research* 34.suppl_2 (2006). W529–533. DOI: 10.1093/nar/gkl212.

[11] Rajeev K. Varshney, Ivo Grosse, Urs Hähnel, Ralf Siefken, Manoj Prasad, Nils Stein, Peter Langridge, Lothar Altschmied, and Andreas Graner. "Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome". In: *TAG Theoretical and Applied Genetics* 113.2 (2006). Pp. 239–250. DOI: 10.1007/s00122-006-0289-z.

[12] Stephan Weise, Ivo Grosse, Christian Klukas, Dirk Koschützki, Uwe Scholz, Falk Schreiber, and Björn H. Junker. "Meta-All: a system for managing metabolic pathway information". In: *BMC Bioinformatics* 7 (2006). P. 465. DOI: 10.1186/1471-2105-7-465.

[13] Boris Podobnik, Jia Shao, Nikolay V. Dokholyan, Vinko Zlatic, H. Eugene Stanley, and Ivo Grosse. "Similarity and dissimilarity in correlations of genomic DNA". In: *Physica A: Statistical and Theoretical Physics* 373 (2007). Pp. 497–502. DOI: 10.1016/j.physa.2006.05.041.

[14] Stefan Posch, Jan Grau, Andre Gohr, Irad Ben-Gal, Alexander E. Kel, and Ivo Grosse. "Recognition Of Cis-Regulatory Elements With VOMBAT". In: *Journal of Bioinformatics and Computational Biology (JBCB)* 5.2b (2007). Pp. 561–577. DOI: 10.1142/S0219720007002886.

[15] Anna Schallau, Irina Kakhovskaya, Anne Tewes, Andreas Czihal, Jens Tiedemann, Michaela Mohr, Ivo Grosse, Renate Manteuffel, and Helmut Bäumlein. "Phylogenetic footprints in fern spore- and seed-specific gene promoters". In: *The Plant Journal* OnlineEarly Articles (2007). DOI: 10.1111/j.1365-313X.2007.03354.x.

[16] Nils Stein, Manoj Prasad, Uwe Scholz, Thomas Thiel, Hangning Zhang, Markus Wolf, Raja Kota, Rajeev K. Varshney, Dragan Perovic, Ivo Grosse, and Andreas Graner. "A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics". In: *TAG Theoretical and Applied Genetics* 114.5 (2007). Pp. 823–839. DOI: 10.1007/s00122-006-0480-2.

[17] Stephan Weise, Siegfried Harrer, Ivo Grosse, Helmut Knüpffer, and Evelin Willner. "The European Poa Database (EPDB)". In: *FAO-Bioversity Plant Genetic Resources Newsletter* 150 (2007). Pp. 64–70.

Referierte Artikel in Konferenzbänden

[18] Wentian Li and Ivo Grosse. "Gene selection criterion for discriminant microarray data analysis based on extreme value distributions". In: *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*. 1-58113-635-8Berlin, Germany. New York, NY, USA: Association for Computing Machinery, 2003. Pp. 217–223. DOI: http://doi.acm.org/10.1145/640075.640103.

[19] Stephan Weise, Norbert Biermann, Steffen Flemming, Jörn Vorwald, Theo J. L. van Hintum, Helmut Knüpffer, and Ivo Grosse. "Proposal for improvement of PGR data exchange through XML". In: *Proceedings of the EPGRIS final conference*. Prague, Czech Republic 2003.

[20] Sven Mielordt, Ivo Grosse, and Jürgen Kleffe. "Data Integration in the Life Sciences". In: Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006. Chap. Data Structures for Genome Annotation, Alternative Splicing, and Validation, pp. 114–123. ISBN: 978-3-540-36593-8. DOI: 10.1007/1179951110.1007/11799511_11.

[21] Jan Grau, Jens Keilwagen, Ivo Grosse, and Stefan Posch. "On the relevance of model orders to discriminative learning of Markov models". In: *LWA '07: Lernen - Wissen - Adaption, Workshop Proceedings*. Ed. by Alexander Hinneburg. Halle, Germany: Martin Luther University Halle-Wittenberg, 2007. Pp. 61–66. ISBN: 9783860109076. URL: http://lwa07.informatik.uni-halle.de/.

[22] Jan Grau, Jens Keilwagen, Alexander Kel, Ivo Grosse, and Stefan Posch. "Supervised posteriors for DNA-motif classification". In: *GCB '07: German Conference on Bioinformatics*. Ed. by Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron, and Dirk Walther. Potsdam, Germany 2007. Pp. 123–134. URL: http://www.gcb2007.de/.

[23] Jens Keilwagen, Jan Grau, Stefan Posch, and Ivo Grosse. "Recognition of splice sites using maximum conditional likelihood". In: *LWA '07: Lernen - Wissen - Adaption, Workshop Proceedings*. Ed. by Alexander Hinneburg. Halle, Germany: Martin Luther University Halle-Wittenberg, 2007. Pp. 67–72. ISBN: 9783860109076. URL: http://lwa07.informatik.uni-halle.de/.

Weiterhin wurden mehr als 100 Vorträge und Poster auf nationalen und internationalen Fachtagungen, Workshops und Sommerschulen präsentiert.