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