Integration von Sequenz- und Genexpressionsdaten als Basis zur prozessorientierten Analyse von Kulturpflanzen

Andreas Stephanik

ITI, Arbeitsgruppe Datenbanken
Otto-von-Guericke Universität Magdeburg
Universitätsplatz 2
39106 Magdeburg
astephanik@googlemail.com

Abstract: Eine Integration von ausgewählten Daten bildet die Grundlage von datenquellen- und datendomänenübergreifenden Analysen. In der vorliegenden Arbeit werden Besonderheiten und Ansätze zur Integration von Life Science Daten insbesondere von Sequenz- und Genexpressionsdaten als Basis von integrierten Analysen im Bereich Kulturpflanzen vorgestellt. Schwerpunkt bildet die materialisierte Integration unter expliziter Berücksichtigung einer Wiederverwendbarkeit in verschiedenen Analyseprozessen. Diese Integration wird auf Schema- und Datenebene durch Ontologien unterstützt.

1 Einleitung

Die Bioinformatik hat sich in den letzten Jahren als essentielle Säule im Forschungsgebiet Biologie etabliert. Grundsätzlich bestehen dabei die Ziele in der Ermittlung von biologischen Strukturen und deren natürlichen Beziehungen sowie der Erforschung von biologischen Prozessen. Nukleotidsequenz- und Genexpressionsdaten stellen dabei zentrale Datendomänen dar, da sie die Basis von biologischen Prozessen widerspiegeln. Die Kulturpflanzen rücken hier immer mehr in den Fokus der Forschung, weil diese eine wesentliche Rolle im Bereich der Ernährung von Mensch und Tier sowie als erneuerbare Energiequelle übernehmen.

Zunächst eine kurze biologische Erläuterung zu diesen Datendomänen: Bestimmte Teilabschnitte von DNA[1]-Sequenzen, die sogenannten Gene, werden unter bestimmten Bedingungen in RNA-Sequenzen überführt. Diese RNA-Sequenzen bilden die Basis für die Biosynthese von Proteinen, welche an verschiedenen Prozessen in Organismen beteiligt sind. Dies repräsentiert die Aktivität der Gene unter den bestimmten Bedingungen und wird als Genexpression bezeichnet. Ziel ist es die Abhängigkeiten der Genexpressionen von verschiedenen Faktoren zu ermitteln. Beispiele dafür sind Vergleiche von Pflanzen mit einer bestimmten Krankheit gegenüber nicht befallenen Pflanzen bzw. zwischen ertragreichen gegenüber weniger ertragreichen Sorten oder die Untersuchung von Genexpressionen bzgl. Trockenheitsresistenz bzw. den Einfluss von anderen äußeren Einflüssen. Weiterhin sind organismenübergreifende Fragestellungen

[1] DNA - **D**eoxyribo**n**ucleic **a**cid, RNA - **R**ibo**n**ucleic **a**cid

von Bedeutung, beispielsweise zwischen verwandten Pflanzenarten aber auch zwischen artfremden Organismen, um spezifische Sequenzen und deren Expression zu erforschen.

Daten zu relevanten Sequenzen und deren Expressionen stammen aus selbst durchgeführten biologischen Experimenten oder können aus anderen frei verfügbaren und kommerziellen Datenquellen extrahiert werden. Die Integration von Genexpressionsdaten aus anderen Datenquellen begründet sich u.a. in der Einsparung von Zeit und Geld gegenüber der Durchführung entsprechender eigener Experimente. Die Ansätze und verwendeten Technologien zur Ermittlung von Expressionswerten sind vielfältig, wobei an erster Stelle die so genannte Array-Technologie zu nennen ist, welche wiederum in verschiedenen Varianten angewandt wird (z.B. Affymetrix).

2 Anforderungen

Bezüglich der Anforderungen an Systeme für prozessorientierte Analysen sind als erstes die potentiellen Akteure zu betrachten. Speziell die Affinität des jeweiligen Akteurs zur IT und insbesondere zu den Komponenten der IT-gestützten Analyseprozesse beeinflusst die Akzeptanz von IT-Systemen zur prozessorientierten Analyse. Dies resultiert aus unterschiedlichen Fähigkeiten zur Nutzung, Erstellung und Erweiterung entsprechender Softwarekomponenten bzw. derer Schnittstellen. Zu den Softwarekomponenten gehören Datenquellen und -senken sowie Funktionen zur Analyse und Visualisierung. In dieser Publikation liegt der Fokus auf den Datenquellen und deren Integration.

2.1 Eigenschaften von Sequenz- und Genexpressionsdaten

Sequenzdaten zur Repräsentation von DNA/RNA zeichnen sich minimal durch einen Identifikator (accession number) und der Kodierung der einzelnen Nukloetide in Form von ACGT/U aus. Zu den Sequenzdaten sind weitere Informationen (Annotationen genannt) verfügbar, welche anhand des Sequenzidentifikators in den Datenquellen adressiert werden können. Zu den Genexpressionsdaten, welche die Aktivität der Sequenzen beschreiben, gehören die numerischen Expressionswerte sowie textuelle Annotationen zu den Experimenten und den untersuchten Materialien.

Die numerischen Expressionswerte werden von den Datenquellen überwiegend als Tab- bzw. kommaseparierte (TSV/CSV) Dateien angeboten und bilden so genannte Expressionsmatrizen. Demgegenüber sind Annotationen von Sequenzen, Experimenten und Materialien in verschiedenen Formaten verfügbar, meist als Freitext (beispielsweise in Text-, PDF- oder Microsoft-Word-Dateien), Schlüssel-Wertepaaren oder als XML (z.B. MAGE-ML [Spe02]).

Neben der Strukturierung der Annotationen ist die Verwendung von Ontologien in den Datenquellen für eine Integration vorteilhaft, da dadurch zum einen auf Schemaebene auf Ontologiekonzepte gemappte Objekttypen und Attribute sowie zum anderen auf

Wertebene Terme von Ontologien eine Homogenisierung unterstützen. Die in diesem Umfeld relevanten Ontologien sind Gene Ontology [Ash00], Plant Ontology [Jai05] und MGED Ontology [Whe06].

Weiterhin wichtig für eine Integration ist die Betrachtung von möglichen Datenänderungen um gegebenenfalls aktualisierte oder neue Daten zu importieren. Bei den Genexpressionsdaten ändern sich einmal erhobene Expressionswerte nicht mehr. Sequenzdaten werden nur selten durch nachfolgende Sequenzierungen angepasst. Sequenzannotationen können erweitert bzw. geändert werden. Die verwendete Strukturierung der Annotationen beeinflusst die automatisierte Aktualisierung. Bei der Verwendung von Freitext sind gegenüber semistrukturierten Formaten wie Schlüssel-Wertepaare oder XML gezielte Aktualisierungen schwieriger, da die jeweiligen Objekttypen und Attribute im Text bestimmt werden müssen. Die Aktualisierung der Annotationen von Samples und der Experimentbeschreibungen unterliegt der gleichen Problematik wie bei den Sequenzannotationen.

Beispiel im Anwendungsumfeld: Die Genexpressionen bzgl. der Keimfähigkeit in der Modellpflanze Arabidopsis thaliana sollen untersucht werden. Neben Expressionswerten aus eigenen Experimenten müssen weitere Expressionswerte aus externen Datenquellen heran gezogen werden. Beispiele für relevante Datenquellen sind die frei verfügbaren AtGenExpress [Sch05] (Affymetrix), AFGC [Fin02] (Zweifarben-Microarray) und NASCArrays [Car04] (Affymetrix) sowie das kommerzielle Genevestigator [Zim08]. Die relevanten Daten können in den Datenquellen anhand der textuellen Experimentbeschreibung gesucht werden. Zur Integration ist zunächst ein Mapping der Sequenzen durchzuführen, damit diese Sequenzen für die Analysen vergleichbar sind. Im Idealfall werden gleiche Sequenzidentifikatoren verwendet. Oft sind jedoch Algorithmen anzuwenden, die ein Mapping auf Sequenzebene ermöglichen. Schließlich müssen die Expressionswerte vergleichbar sein, wofür eine Normalisierung durchgeführt werden muss. In den Datenquellen sind zumeist bereits technologiespezifisch (Affymetrix, Zweifarben-Microarray) normalisierte Expressionswerte enthalten. Für eine experiment- bzw. datenquellenübergreifende Analyse sind weitere Transformationen erforderlich.

2.2 Anforderungen zur Integration

Mit den in den vorherigen Abschnitten geschilderten Eigenschaften werden folgend Anforderungen für die Integration von Sequenz- und Genexpressionsdaten aufgeführt. Aufgabe der Integration von Sequenz- und Expressionsdaten ist die Schaffung einer Vergleichbarkeit der Daten, insbesondere der Expressionswerte, um übergreifende Analysen zu ermöglichen.

I. Flexible Datenmodellierung: Die Erstellung eines umfassenden Schemas zur Integration aller relevanten Attribute ist aufgrund der Erweiterungen von Annotationen nicht vollständig möglich. Dagegen können Sequenzdaten mit Identifikator und Sequenzkodierung sowie Expressionswerte bei der Modellierung als spezifische Attribute angenommen werden.

II. Effizientes Datenmanagement: Effizient bezieht sich hier auf die Speicherung von Daten, um bereits im Integrationssystem (d.h. datenbankgestützt) analyserelevante Datentransformationen vornehmen zu können. Vor allem Expressionswerte sollen als numerische Einzelwerte abgespeichert werden, um bereits in der Datenbank Normalisierungen und weitere Berechnungen durchführen zu können.

III. Schnittstellen zum Datenzugriff: Neben der Anforderung bereits im Integrationssystem Berechnungen vornehmen zu können, müssen weitere Anwendungen für Analysen oder zur Visualisierung auf die integrierten Daten zugreifen.

IV. Möglichkeiten der Datenkuration: Integrierte Daten sollen manuell durch Experten oder automatisiert mittels entsprechender Komponenten korrigiert werden können.

3 Ansatz zur Datenintegration von Sequenz- und Genexpressionsdaten

Grundsätzlich sind zwei Ansätze zur Datenintegration möglich: Die virtuelle Integration, bei der die Daten zur Anfragezeit aus den Datenquellen importiert und ggf. transformiert werden. Vorteil dieses Ansatzes ist die Aktualität der Daten. Die in diesem Anwendungsumfeld relevanten Daten obliegen jedoch, wie bereits erwähnt, keiner bzw. geringer Änderungen. Kritisch sind auch die möglichen Antwortzeiten zu betrachten. Aufgrund fehlender geeigneter Schnittstellen der Datenquellen sind die Daten überwiegend aus HTML-Seiten oder Dateien zu extrahieren. Eine Möglichkeit der Datenkuration ist ohne eine Zwischenspeicherung nicht gegeben.

Im Folgenden wird der zweite Ansatz - die materialisierte Integration - unter Berücksichtigung der aufgeführten Anforderungen für das Anwendungsumfeld näher betrachtet. Die Anforderung I - Flexible Datenmodellierung - erfordert eine Eigenschaft des Integrationsansatzes, um beim Hinzufügen von noch nicht modellierten Objekttypen oder Attributen Änderungen zu vermeiden. Dies betrifft in erste Linie Schemaänderungen in der Datenbank sowie von aufsetzenden Schnittstellen. Ein Entity-Attribute-Value-Ansatzes (EAV [NB98]) ermöglicht die Vermeidung von Schemaänderungen, indem Objekttypen bzw. Entities und Attribute auf Datenebene hinzugefügt werden. In Abbildung 1 ist ein an EAV angelehntes flexibles Schema dargestellt.

Die Grundlage bildet der Bereich *controlled vocabulary* in dem aus Ontologien Termkategorien und Terme für Objekttypen bzw. Attribute (Bereich *objects*) und Attributwerte importiert werden, um bereits definierte Begriffe und Beziehungen zu verwenden. Basis bildet die MGED Ontology, in der beispielsweise der Ontologiebegriff „Biomaterials“ und der Unterbegriff „Organism“ in *term category* zur Beschreibung des äquivalenten Objekttypen bzw. Attributes heran gezogen werden. Eine Instanz (*object*) des Objekttypen „Biomaterials“ hat als Beispiel das Attribut (*object property*) „Organism“ mit dem Wert (*term*) „Arabidopsis thaliana“. Die Herkunft der Elemente wird durch *external databases* gespeichert. Idealerweise nutzen Datenquellen bereits eine oder mehrere der Ontologien, wie z.B. ArrayExpress [Par05], so dass Abbildungen

auf Schema- und Instanzebene bei der Verwendung derselben Ontologie direkt erfolgen können.

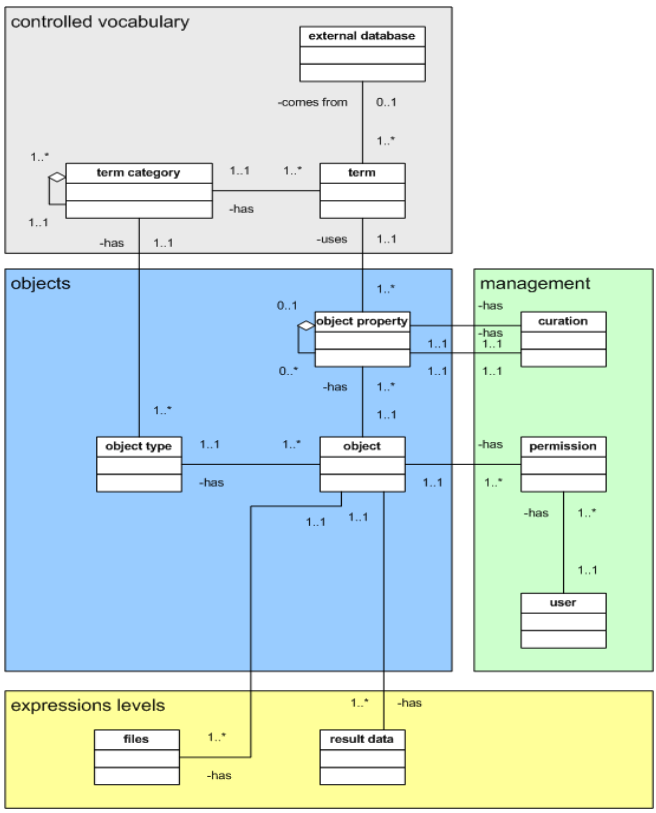

Abbildung 1: Schema zur flexiblen Speicherung von Sequenz- und Genexpressionsdaten

Aufgrund einer materialisierten Integration stehen die Daten direkt in einer oder mehreren Datenbanken zur Verfügung und können während sowie nach der Integration analysespezifisch transformiert werden, wie es durch die Anforderung II beschrieben wird. Dazu kann die Funktionalität des entsprechenden Datenbank Managementsystems (DBMS) genutzt werden. Voraussetzung ist die geeignete Speicherung als Einzelwerte. Weiterhin können bei einer materialisierten Integration die Optimierungsmöglichkeiten der Datenbanksysteme zur Steigerung der Effizienz bei Anfragen genutzt werden, was bei der virtuellen Integration praktisch schwer möglich ist.

Bezüglich der Anforderung III - Schnittstellen zum Datenzugriff - wird die Anfrageschnittestelle des Datenbanksystems genutzt. Darauf basierend sind Application Programming Interfaces (APIs) für bestimmte Programmiersprachen sowie interoperable Schnittstellen wie SOAP- oder REST-konforme Web-Services zu implementieren. Für Analysen in der Bioinformatik sind zahlreiche Anwendungen vor allem in den Programmiersprachen PERL, Java und R verfügbar. Entsprechend der Anforderung IV

dient eine der Anwendungen der Datenkuration, wobei für eine manuelle durch Experten eine grafische Schnittstelle auf den angebotenen APIs angeboten werden sollte.

3.1 Architektur und Implementierung

Im letzten Abschnitt wird kurz die Architektur des Data-Warehouse-Systems BATEx[3] vorgestellt (siehe Abbildung 2), welches am IPK Gatersleben entwickelt wird.

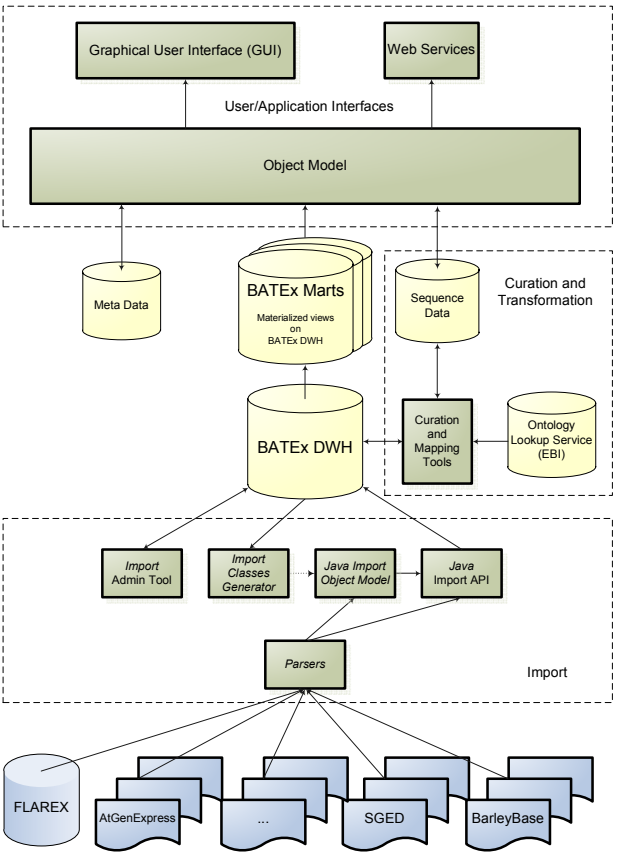

Abbildung 2: Architektur des Data-Warehouse-Systems BATEx

Der Kern des Systems ist das BATEx DWH (Data Warehouse), das den gesamten integrierten Datenbestand beinhaltet. Das generische Datenbankschema (vgl. Abbildung 1 und die Erläuterungen im vorherigen Abschnitt) ermöglicht es, zunächst analyse-unabhängig Daten zu homogenisieren. Darauf aufsetzend werden Data Marts erstellt, die ein analysespezifisches Datenschema besitzen und dafür effizientere Anfragen unterstützen. Der Import in das BATEx DWH erfolgt über ein Java-Datenmodell, welches anhand von initial importierten Ontologien generiert werden kann. Theoretisch

[3] http://pgrc.ipk-gatersleben.de/batex/

können auch gängige ETL-Tools zum Importieren von Daten in das BATEx DWH genutzt werden. Anwendungen und Schnittstellen für Analysen und zur Datenkuration setzen direkt auf das BATEx DWH oder auf die Data Marts auf.

Derzeit sind in BATEx 137 Experimente beispielsweise aus den bereits genannten Systemen AFGC und AtGenExpress importiert.

3.2 Diskussion

Mit dem Schwerpunkt der Analyse von Expressionswerten ist vor allem die atomare Speicherung von Expressionswerten eine wichtige Anforderung. Ein Beispiel für ein Informationssystem mit einem umfangreichen Bestand an Expressionsdaten aber diesbezüglich ungeeigneter Expressionswertespeicherung ist ArrayExpress [Par05], in dem die Expressionswerte Tab-separiert als CLOB verwaltet werden. Neben der Analyse der Expressionswerte sind für datendomäneübergreifende Analysen die Sequenzannotationen von Bedeutung. In den Systemen für Expressionsdaten sind die Sequenzen oft zu entsprechenden Datenquellen von Sequenzannotationen direkt verlinkt, beispielsweise zu den Systemen EMBL [Coc08] oder Gramene [Lia08]. Alternativ können anhand des Sequenzidentifikators oder durch Sequenzvergleiche andere Datenquellen abgefragt werden. Das vorgestellte System bietet durch das generische Schema des zentralen DWH die Möglichkeit weitere Objekttypen bzw. Attribute zu berücksichtigen, ohne Schemaänderungen an dieser Komponente vornehmen zu müssen. Für Sequenzannotationen ist auch die Verknüpfung mit anderen Ansätzen möglich (z.B. GenMapper [DR04]); Sequenzannotationen bilden allerdings nicht den Schwerpunkt dieser Arbeit.

Danksagung

Die Dissertation zum Thema „Integration von Daten und Methoden zur Unterstützung einer prozessorientierten Analyse von Pflanzen“ erfolgt als externer Doktorand an der Universität Magdeburg in der Arbeitsgruppe Datenbanken des Instituts für Technische und Betriebliche Informationssysteme in der Zusammenarbeit mit dem Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) Gatersleben. Mein Dank gilt Prof. Dr. G. Saake und Dr. E. Schallehn für die Betreuung seitens der Uni Magdeburg sowie Dr. U. Scholz, Dr. M. Lange, Dr. S. Weise, Herrn C. Künne, Herrn T. Rutkowski und Herrn B. Steuernagel für die Unterstützung seitens des IPK Gatersleben.

Literaturverzeichnis

[Ash00] Ashburner, M. et al.: Gene Ontology: tool for the unification of biology. Nature Genetics, 25, 2000, S. 25-29.

[Coc08] Cochrane,G. et al.: Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. Nucleic Acids Research, 36 (Database issue), 2008, D5-D12.

[Cra04] Craigon, D.J. et al.: NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucl. Acids Res., 32, 2004, D575-577.

[DR04] Do, H.H.; Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach Proc. EDBT 2004, Heraklion, Greece, Springer LNCS, March 2004, 2004-03.

[Fin02] Finkelstein, D. et al.: Microarray data quality analysis: lessons from the AFGC project. Plant Molecular Biology, Springer, 48, 2002, 119-132.

[Jai05] Jaiswal, P. et al.: Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comparative and Functional Genomics, 6, 2005, S. 388-397.

[Lia08] Liang, C., et al.: Gramene: a growing plant comparative genomics resource. Nucleic Acids Research 36 (Database issue), 2008, D947-D953.

[NB98] Nadkarni, P.; Brandt, C.:. Data Extraction and Ad Hoc Query of an Entity-Attribute-Value Database. Journal of the American Medical Informatics Association, 5, 1998, S. 511-527.

[Par05] Parkinson, H. et al.: ArrayExpress-a public repository for microarray gene expression data at the EBI. Nucl. Acids Res., 33, 2005, D553-555.

[Sch05] Schmid, M. et al.: A gene expression map of Arabidopsis thaliana development. Nature Genetics, 37, 2005, S. 501-506.

[Spe02] Spellman, P. et al.: Design and implementation of microarray gene expression markup language (MAGE-ML) Genome Biol, 3(9), 2002

[Whe06] Whetzel, P. et al.: The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics, 22, 2006, S. 866.

[Zim08] Zimmermann, P. et al.: Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Advances in Bioinformatics, 2008