Hybride Integration von molekularbiologischen Annotationsdaten

Christine Körner†, Toralf Kirsten†, Hong-Hai Do†, Erhard Rahm†‡

†Interdisziplinäres Zentrum für Bioinformatik, Universität Leipzig
http://www.izbi.de

‡Institut für Informatik, Universität Leipzig
http://dbs.uni-leipzig.de

{koerner, kirsten, do}@izbi.uni-leipzig.de
rahm@informatik.uni-leipzig.de

Abstract: Wir präsentieren einen Ansatz, um Annotationsdaten von molekularbiologischen Objekten wie Genen, Proteinen und Pathways aus öffentlichen Datenquellen für datenintensive Expressionsanalysen verwendbar zu machen. Die Expressionsdaten sind mit Experimentbeschreibungen physisch in einem Data Warehouse integriert, um schnelle Auswertungen zu unterstützen. Die öffentlichen Annotationsdaten werden virtuell über einen Mediatoransatz integriert und bedarfsgesteuert für Analysen abgerufen. Für die einheitliche Anbindung der Datenquellen wird das verbreitete Tool SRS (Sequence Retrieval System) der Fa. LION bioscience genutzt. Die Kopplung zwischen dem Warehouse und SRS erfolgt über einen Query-Mediator unter Nutzung explizit gespeicherter Beziehungen (Mappings) zwischen den Instanzen der öffentlichen Datenquellen. Dieser hybride Integrationsansatz wurde als Erweiterung des Leipziger Data Warehouse für Genexpressionsdaten (http://www.izbi.de/GEWARE) implementiert und wird für die Einbindung von GeneOntology, LocusLink und Ensemble in Analysen eingesetzt. Neben der Darstellung des Integrationskonzepts und seiner Realisierung werden auch Ergebnisse erster Performanzmessungen präsentiert.

1 Einleitung

Die rasanten Entwicklungen in der Biotechnologie haben es ermöglicht, dass ganze Genome unterschiedlicher Organismen in kürzester Zeit sequenziert werden konnten. Das Hauptziel der genomischen Forschung besteht nun darin, die Funktionsweise der Gene und deren Produkte auf der genomweiten Ebene zu verstehen. Das Wissen über molekularbiologische Objekte wie Gene, Proteine, Krankheiten etc., wird gesammelt, kontinuierlich aktualisiert und in zahlreichen öffentlich zugängigen Datenquellen verfügbar gemacht. Derzeitig existieren über 500 solcher Quellen [Ga04].

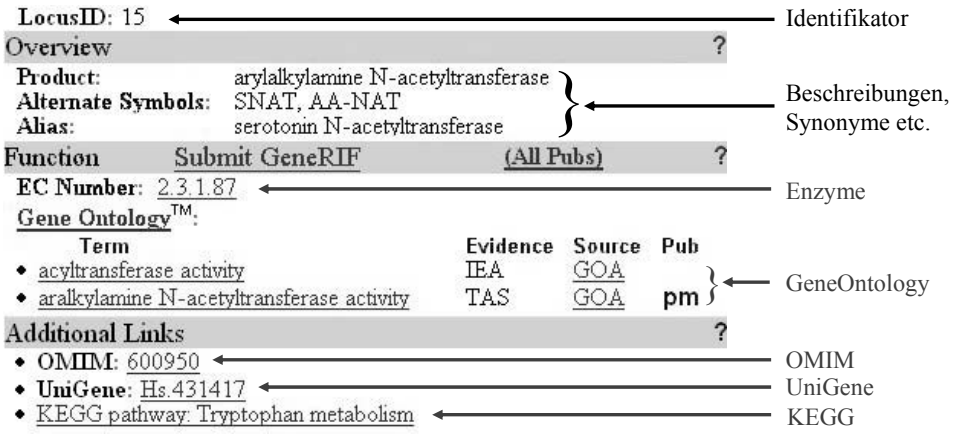


Abbildung 1: Annotations- und Mappingdaten in LocusLink

Zur Verdeutlichung der zu betrachtenden Arten von Daten zeigt Abbildung 1 einen Ausschnitt eines Gen-Eintrages der Referenzquelle LocusLink[1]. Der Eintrag enthält Beschreibungen zu dem Gen mit dem quell-spezifischen Identifikator *15*. Wir unterscheiden dabei zwischen *Annotations*- und *Mappingdaten*. Erstere werden durch quell-spezifische Attribute wie *Product*, *Alternate Symbols* etc. beschrieben und sind oft textueller Natur. Die Mappingdaten umfassen Weblinks zu anderen Datenquellen, welche durch die Identifikatoren (Accessions) aus den entsprechenden Quellen gekennzeichnet sind, z.B. *2.3.1.87* (Enzyme[2]) oder *Hs.431417* (UniGene[3]). Ein *Mapping* umfasst alle Korrespondenzen zwischen den Objekten zweier Quellen. Die Mappings versetzen den Benutzer in die Lage, zwischen den betreffenden Quellen zu navigieren und deren Daten zu verknüpfen, ggf. über mehrere Zwischenstationen hinweg. Im Beispiel können damit bei der Analyse des Gens auch Annotationen der referenzierten Objekte aus GeneOntology und UniGene herangezogen werden.

Viele Applikationen, wie z.B. für Genexpressions- und Proteinanalysen, erfordern eine Integration zahlreicher Annotationsdaten aus unterschiedlichen Datenquellen. Die Nutzung von Weblinks stellt einen einfachen Weg zur Datenintegration dar, welcher bereits weit verbreitet in den einzelnen Datenquellen benutzt wird. Allerdings wird damit nur die interaktive Analyse einzelner Objekte unterstützt, nicht jedoch die gleichzeitige und umfassende Auswertung für zahlreiche Objekte.

In diesem Papier präsentieren wir einen neuen Ansatz zur Integration von Annotationsdaten aus öffentlichen Datenquellen und wenden diesen zur Unterstützung von Expressionsanalysen an. Dabei sind die Expressionsdaten zusammen mit den Experimentbeschreibungen physisch in einem Data Warehouse integriert. Die öffentlichen Annotationsdaten werden virtuell über einen Mediator integriert, wofür das verbreitete Tool

[1] http://www.ncbi.nlm.nih.gov/projects/LocusLink/
[2] http://www.expasy.org/enzyme/
[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

SRS (Sequence Retrieval System) der Fa. LION bioscience eingesetzt wird. Die Kopplung zwischen dem Warehouse und SRS erfolgt über einen eigens implementierten Query-Mediator. Die wesentlichen Beiträge unserer Arbeit sind:

- Wir kombinieren eine materialisierte und virtuelle Integration von Daten in eleganter Weise, um deren Vorteile in einem neuartigen hybriden Ansatz zu vereinen. Einerseits kann für komplexe Analysen experimenteller Daten im Warehouse eine hohe Performanz erreicht werden, andererseits können über den Mediator bedarfsgesteuert aktuelle Annotationsdaten für die Auswertung herangezogen werden.

- Öffentliche Datenquellen werden einheitlich durch das bewährte Mediator-Tool SRS eingebunden, welches Schnittstellen zu zahlreichen öffentlichen (sowohl datei-basierten als auch relationalen) Datenquellen bereitstellt. Dadurch wird eine redundante Implementierung von Importfunktionen vermieden, und es können zahlreiche Datenquellen einfach in unsere Integrationsplattform eingebunden werden.

- Wir extrahieren explizit die Mappings aus den Datenquellen und materialisieren sie in einer eigenen Datenbank, der Mapping-Datenbank. Die explizite Trennung der Mappings von anderen Daten erlaubt, Join-Wege zwischen Datenquellen flexibel zu bestimmen und zur Performanzoptimierung ggf. vorzuberechnen.

- Unser Ansatz wurde im Rahmen der *GeWare*-Plattform [KDR04, KDR03] implementiert und integriert mehrere öffentliche Datenquellen zur Unterstützung von Expressionsanalysen. Das web-basierte System ist unter http://www.izbi.de/GEWARE verfügbar. Erste Performanzmessungen für den Zugriff auf die Annotationsdaten wurden durchgeführt und belegen die Praktikabilität des Ansatzes.

Das Papier gliedert sich im Weiteren wie folgt. Kapitel 2 diskutiert verwandte Integrationsansätze und deren Abgrenzung zu unserem Ansatz. Kapitel 3 beschreibt wesentliche Anwendungsszenarien und die daraus resultierenden Integrationsanforderungen. Kapitel 4 stellt unser Integrationskonzept im Überblick vor. Die beiden darauf folgenden Kapitel erläutern zentrale Komponenten im Detail, nämlich Aufbau und Wirkungsweise der Mapping-Datenbank (Kap. 5) und die Anfragebearbeitung im Query-Mediator (Kapitel 6). Kapitel 7 präsentiert erste Performanzmessungen bevor die Zusammenfassung in Kapitel 8 die Arbeit abschließt.

2 Verwandte Arbeiten

Eine Übersicht repräsentativer Ansätze sowie Systeme zur Datenintegration in der Bioinformatik ist in [LC03], [HK04] und [ST03] zu finden. Bisherige Lösungen verfolgen meistens entweder einen physischen (materialisierten) Data Warehouse Ansatz oder einen virtuellen Ansatz mit Mediatoren. Ersterer basiert auf einer Vorabintegration der Daten in einer zentralen für datenintensive Analysen optimierten Datenbank. Im virtuellen Fall erfolgt der Zugriff auf die Datenquellen erst zur Laufzeit, um stets die aktuellsten und nur die benötigten Daten abzurufen. Zu den Vertretern, die mit unserem Ansatz eng verwandt sind, zählen zum einen COLUMBA [Ro04] und GenMapper [DR04], die

einen Ansatz der materialisierten Integration verfolgen, sowie DiscoveryLink [Ha01], Kleisli [CCW03, Wo98] und im Besonderen SRS [EHB03, Zd02]), die einem virtuellen Ansatz zuzurechnen sind. Im Folgenden diskutieren wir die einzelnen Systeme und die Abgrenzung zu unserer Arbeit.

Ähnlich wie in unserem Ansatz streben die Mediatoren DiscoveryLink, Kleisli und SRS keine vollständige semantische Integration der Quellschemata zur Konstruktion eines globalen Schemas an. Sie basieren auf einem einfachen Schema für Quellen und deren Attribute, mit dem neue Datenquellen relativ einfach hinzugefügt werden können. Kleisli bietet z.B. Schnittstellen zu mehr als 60 öffentlichen Datenquellen und SRS stellt Wrapper zur Anbindung von mehr als 700 Quellen bereit. Zur Steigerung der Performanz und Nutzung umfassender Anfragemöglichkeiten können lokale Kopien der Datenquellen angelegt werden, die periodisch zu aktualisieren sind. Gegenüber dieser Flexibilität besteht der Nachteil von DiscoveryLink und Kleisli darin, dass die Datenintegration weitestgehend in der Verantwortung des Nutzers verbleibt. So sind vom Nutzer in Anfragen explizite Join-Bedingungen anzugeben, um Daten aus verschiedenen Quellen zusammenzuführen.

Dieser Nachteil von DiscoveryLink und Kleisli wurde in SRS und unserem Ansatz durch den Einsatz von Mappingdaten adressiert. SRS registriert die verfügbaren Mappings in den Instanzdaten, d.h. zwischen Objekten der verschiedenen Datenquellen, und ermöglicht, die korrespondierenden Objekte in einer anderen Quelle zu identifizieren. Durch eine Indizierung dieser Objekte kann eine höhere Performanz bei der Abfrage von Daten aus unterschiedlichen Datenquellen erzielt werden. SRS wählt dabei immer den kürzesten Pfad zwischen den Datenquellen als Join-Weg aus. Damit sind jedoch Umfang und Qualität der Join-Ergebnisse potentiell beschränkt, da alternative bzw. längere Join-Wege umfassendere oder aktuellere Ergebnisse ermöglichen können. Mit unserem Ansatz erreichen wir eine flexiblere und dennoch effiziente Berechnung der Join-Operationen durch a) Nutzer-wählbare Join-Wege und b) Vorberechnung der Join-Wege der Quellen zu einer zentralen Quelle, wodurch Wege mit einer maximalen Länge von 2 garantiert werden.

COLUMBA [Ro04] verfolgt einen physischen Integrationsansatz, um Proteinannotationen verschiedener Datenquellen in einer lokalen Datenbank zusammenzuführen. Bei der Konstruktion des globalen Schemas werden die Schemata der Datenquellen weitestgehend übernommen, um den Aufwand zur Schemaintegration und des Datenimports gering zu halten. Dabei wird eine Datenquelle, die Protein Data Bank (PDB), als zentrale Quelle identifiziert. Zwischen dieser und allen anderen Quellen wird jeweils ein Mapping generiert, so dass Join-Operationen zwischen zwei beliebigen Datenquellen günstig über die zentrale Quelle berechnet werden. Diese Technik wird auch in unserem Ansatz eingesetzt, um die Mapping-Datenbank zu konstruieren. Anstatt jeweils nur ein Mapping zwischen zwei Quellen vorzuberechnen, unterstützt unser Integrationsansatz alternative Mappings für unterschiedliche Join-Pfade.

GenMapper [DR04] verfolgt ebenfalls eine physische Integration der Annotationsdaten. Es verwendet ein generisches Datenmodell, das GAM (Generic Annotation Model), in dem sowohl die Beziehungen zwischen Objekten innerhalb einer Quelle, wie sie bei-

spielsweise bei Taxonomien und Ontologien auftreten, als auch zwischen Objekten verschiedener Quellen gespeichert werden. Durch mächtige Operatoren können aus der GAM-Repräsentation Annotationssichten für unterschiedliche Analysezwecke generiert werden. Allerdings ist GAM spezifisch auf Mappingdaten beschränkt, während Daten komplexer Strukturen, wie z.B. geometrische Daten der Proteinfaltstrukturen oder die Sequenzen des Genoms, nicht einbezogen werden. Unsere Implementierung benutzt die Mappingdaten des GenMappers, um die Mapping-Datenbank zu konstruieren.

3 Analyseszenarien und Integrationsanforderungen

3.1 Analyseszenarien

Abbildung 2 stellt zwei typische Analyseszenarien aktueller Bioinformatikanwendungen dar, die Expressions- und Annotationsanalyse. Eine Expressionsanalyse ermittelt und vergleicht mit Hilfe unterschiedlicher Technologien, wie z.B. Microarray, den Aktivitätsgrad von Genen oder Proteinen unter unterschiedlichen Bedingungen der Zelle, beispielsweise im normalen bzw. kranken Gewebe. Das Ziel ist i.d.R., Gruppen von Genen / Proteinen mit ähnlichem Expressionsmuster zu identifizieren. Beispielsweise können Gene, die eine hohe Aktivität in den Krebszellen, nicht aber in den gesunden Zellen aufweisen, für die unkontrollierte Teilung der Krebszellen verantwortlich sein. Die Analyse der Annotationen einzelner Gene der Gruppe kann nun Aufschluss darüber geben, ob eventuell auch ein gemeinsames Muster mit ähnlichen molekularen Funktionen vorhanden ist. Umgekehrt können anhand der Suche in Annotationsdaten Gruppen von Genen oder Proteinen mit ähnlichen funktionalen Eigenschaften identifiziert werden. Für diese Gen- / Proteingruppen kann anschließend eine Expressionsanalyse durchgeführt werden, um Einblicke in das Expressionsmuster zu gewinnen.

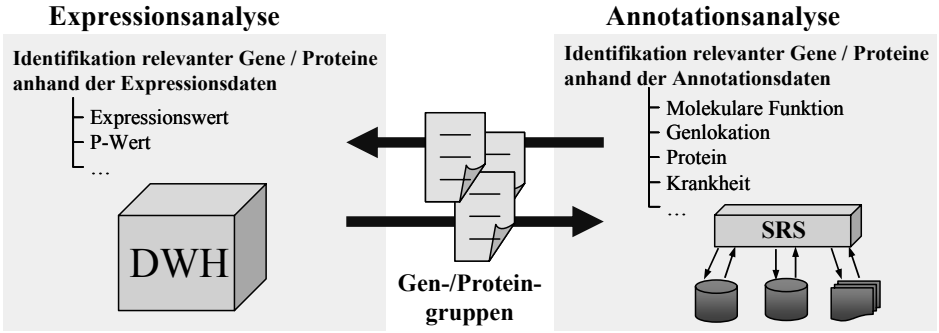


Abbildung 2: Analyseszenarien

3.2 Integrationsanforderungen

In Hinsicht auf die Integration der Annotationsdaten und deren Einbeziehung in Analysen sind folgende Anforderungen relevant:

- *Flexibilität und Anpassbarkeit*: Die öffentlichen Datenquellen unterliegen i.d.R. einem ständigen Entwicklungsprozess, so dass Änderungen sowohl auf Schema- als auch auf Instanzebene stattfinden. Deshalb ist es erforderlich, die integrierten Daten in einfacher Art und Weise zu aktualisieren, wenn sich die ursprünglichen Quellen geändert haben. Ferner ist es aufgrund der hohen Anzahl verfügbarer Datenquellen notwendig, mit möglichst wenig Aufwand Daten aus neuen Quellen zu integrieren. Diese Aspekte sprechen für eine virtuelle Integration von Annotationsdaten, wie sie beispielsweise von SRS unterstützt wird.

- *Datenqualität:* Typischerweise beinhalten die Datenquellen nicht nur Korrespondenzen zu anderen Quellen, sondern auch Daten, die durch die Ableitung von Daten anderer Quellen entstehen. Beispielsweise benutzen Gendatenbanken Daten aus Sequenzdatenbanken für die Voraussage von Kandidatengenen oder die Erkennung von ähnlichen Genen (homologe bzw. orthologe Gene). Durch die Verwendung unterschiedlicher Algorithmen und zeitlichen Aktualisierungsabständen weisen die Datenquellen eine unterschiedliche Datenqualität auf, welche wiederum maßgeblichen Einfluss auf die Vertrauenswürdigkeit und damit auf die Akzeptanz der Datenquelle hat. Deshalb ist es erforderlich, die Datenquelle für den Benutzer ersichtlich nachzuweisen.

- *Analyseunterstützung:* Eine Expressionsanalyse benutzt sehr große Datenmengen, die z.B. durch Microarray-Experimente erzeugt werden und mit leistungsfähigen Analyse- und Mining-Methoden auszuwerten sind. Die zur Interpretation erforderlichen Annotationsdaten variieren oft stark, da sie vom Fokus einzelner Untersuchungen und Nutzer abhängen. Annotationen sollten möglichst transparent und unabhängig von ihren Ursprungsquellen mit beliebigen Objekten assoziiert werden können. Beispielsweise soll es möglich sein, die Genfunktionen, die als Begriffe ähnlich einem Vokabular in GeneOntology definiert sind, mit Genen aus beliebigen Gendatenbanken, wie z.B. LocusLink, UniGene und NetAffx zu assoziieren. Ferner sind Filterbedingungen auf Annotationen, die eine einfache Zeichenkettensuche unterstützen, bis hin zur Kombination mehrerer solcher Bedingungen, notwendig, um Objekte mit interessanten Merkmalen zu identifizieren. Letztlich ist es sinnvoll, aufgrund des hohen Grades der Vernetztheit der Datenquellen, alternative Join-Pfade zu unterstützen.

- *Performanz*: Im Analyseprozess, insbesondere der interaktiven Analyse, kommt der Performanz eine tragende Rolle in Hinsicht auf die Benutzerakzeptanz zu. Für Expressionsanalysen ist daher ein physischer Integrationsansatz über ein Data Warehouse vorzuziehen. Bei der virtuellen Anbindung unterschiedlicher Annotationsdaten ist die Ausführung der Join-Operationen zwischen mehreren Quellen besonders leistungskritisch. Daher sollten entsprechende Mechanismen, wie z.B. Indexierung oder die Vorberechnung von Join-Pfaden, zur Performanzoptimierung eingesetzt werden.

4 Architektur im Überblick

4.1 Komponenten

Aufgrund der im letzten Abschnitt beschriebenen Integrationsanforderungen haben wir eine hybride Integrationslösung entwickelt. Abbildung 3a zeigt die Architektur unseres Integrationsansatzes. Die beteiligten Komponenten sind:

- *GeWare*, ein Data Warehouse zur Expressionsanalyse, dient als Integrations- und Testplattform für unseren Ansatz. Die Annotationsdaten aus öffentlichen Datenquellen werden bei der Interpretation der Analyseresultate eingesetzt, wie z.B. bei der Suche nach Gemeinsamkeiten von als aktiviert (exprimiert) identifizierten Genen.

- *SRS* dient als Mediator zu verschiedenen öffentlichen Datenquellen. Zurzeit werden die Annotationsdaten aus LocusLink [PM02], GeneOntology [As00] und Ensembl [Bi04, Po04] integriert. Darüber hinaus sind die Identifikatoren (Accessions) der Datenquellen UniGene [PWS03, Wh03] und NetAffx [Ch04] verfügbar. NetAffx ist eine öffentliche Annotationsquelle für Gene, mit denen Microarrays der Fa. Affymetrix bestückt sind und zum Nachweis der Genexpression verwendet werden.

- Der *Query-Mediator* dient zur Kopplung von *GeWare* und SRS. Seine Aufgabe besteht darin, die Nutzeranfragen in eine oder mehrere SRS-Abfragen zu übersetzen, sie durch SRS ausführen zu lassen und die Ergebnisse im Anschluss zu kombinieren.

- Die *Mapping-Datenbank* speichert die Mappings zwischen den Objekten der integrierten Quellen zur Join-Berechnung. Um den Bestand der Mappings klein zu halten, wird eine sternförmige Anordnung der Quellen um eine zentrale Quelle vorgenommen.

- Die *ADM-Datenbank* (ADM=Administration) speichert die Metadaten über die integrierten Quellen, z.B. die Namen der Quellen, deren Attribute und die verfügbaren Mappings. Anhand dieser Metadaten wird die Web-Oberfläche zur Anfrageformulierung automatisch generiert.

Der Aufbau der einzelnen Komponenten wird im Kapitel 5 (Metadatenverwaltung in der Mapping- und ADM-Datenbank) sowie im Kapitel 6 (Anfragebearbeitung im Query-Mediator) diskutiert. In den beiden folgenden Abschnitten beschreiben wir das Zusammenspiel der Komponenten in zwei Nutzungsprozessen: die Anbindung der Datenquellen und die Anfrageverarbeitung.

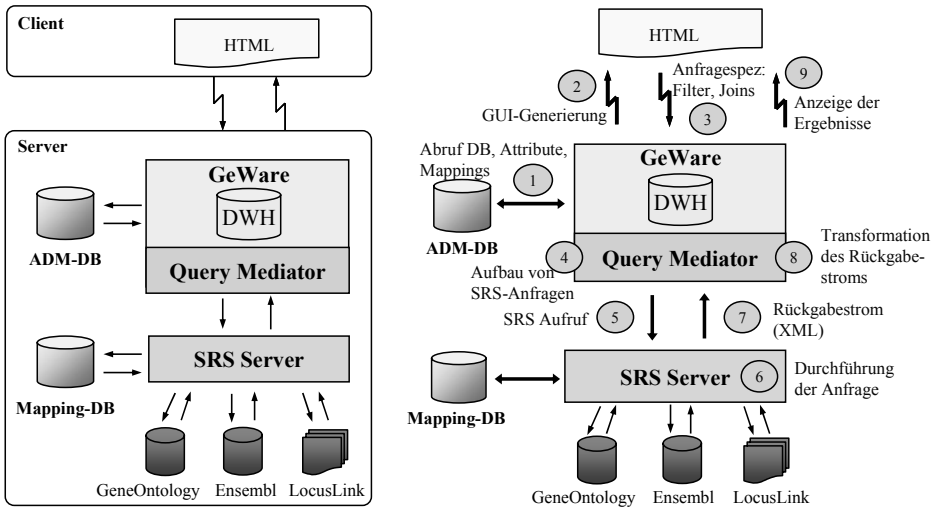


Abbildung 3: Integrationsansatz und Komponenten im Überblick

4.2 Anbindung der Datenquellen

Die umfangreiche Bibliothek verfügbarer Wrapper von SRS gestattet es, fast jede Datenquelle einzubinden. Aus Performanzgründen werden die Datenquellen i.d.R. als lokale Kopien angelegt, auf die über die Wrapper zugegriffen wird. In unserer Testinstallation sind die datei-basierte Quelle LocusLink sowie zwei relationale (MySQL) Datenbanken, Enseml und GeneOntology, derartig integriert. Ferner werden Metadaten über die Quellen, insbesondere die Namen der Datenquellen und deren Attribute erfasst und in der ADM-Datenbank von *GeWare* gespeichert.

Um den Aufwand für die Join-Anfragen gering zu halten, werden die Datenquellen in einem sternförmigen Graphen organisiert. Eine Datenquelle wird dabei als zentrale Datenquelle identifiziert, von der zu jeder anderen Quelle ein oder mehrere Mappings zur Verfügung stehen bzw. berechnet werden können. In der Genexpressionsanalyse ist LocusLink eine etablierte Referenzquelle für Genannotationen und wird deshalb als zentrale Quelle in unserer Integrationslösung verwendet. Für den Aufbau der Mapping-Datenbank werden die Mappings zwischen LocusLink und den anderen Quellen, wie Ensembl und GeneOntology benötigt. Diese werden im GenMapper-Tool abgefragt und in die Mapping-Datenbank importiert. Für jede Verbindung zwischen der zentralen Quelle und einer anderen können mehrere Mappings importiert werden. Sie werden in der ADM-Datenbank erfasst und stehen in *GeWare* als alternative Join-Pfade zur Verfügung. Die Mapping-Datenbank wird durch die relationale Schnittstelle in SRS registriert und integriert.

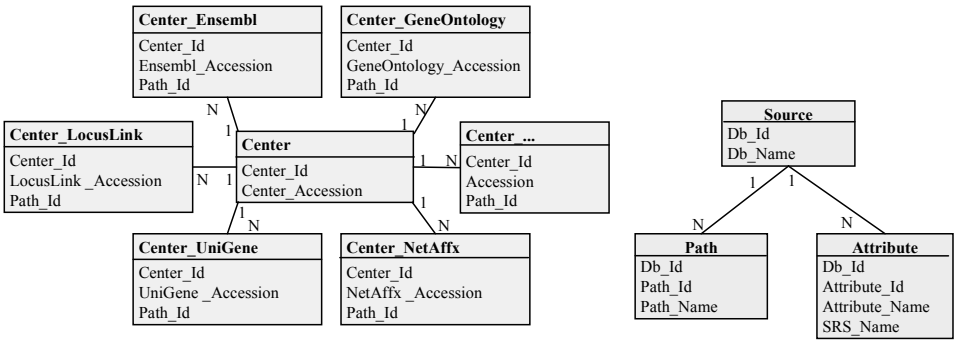
4.3 Anfrageverarbeitung

Abbildung 3b zeigt den allgemeinen Ablauf der Anfrageverarbeitung in unserem System und wird im Kapitel 6 detaillierter beschrieben. Im ersten Schritt (1) werden Metadaten zu den verfügbaren Quellen, deren Attribute und Mappings von der ADM-Datenbank abgerufen. Diese werden dazu genutzt, um eine Web-Oberfläche automatisch zu generieren (2). Auf der Web-Oberfläche kann der Nutzer Anfragen formulieren, indem er die relevanten Attribute und Datenquellen auswählt sowie Filter- und Join-Bedingungen spezifiziert (3). Die Anfragen werden an den Query-Mediator weitergegeben. Dieser interpretiert die Anfragen und generiert daraus einen Anfrageplan (4), welcher in eine oder mehrere SRS-spezifischen Abfragen umgesetzt wird. Der SRS-Server wird aufgerufen, um die generierten Abfragen zu bearbeiten (5). Die jeweiligen Selektionen und Projektionen für die ausgewählten Attribute werden in den Datenquellen ausgeführt, während die Join-Operationen an die Mapping-Datenbank zur Bearbeitung geschickt werden (6). Die Ergebnisse werden von SRS in einem XML-Datenstrom zurückgeliefert (7). Dieser Datenstrom wird vom Query-Mediator entgegengenommen, dessen Daten extrahiert (8) und in ein Ausgabeformat, z.B. HTML für die Anzeige im Webbrowser oder CSV für den Download, konvertiert (9).

5 Metadatenverwaltung

5.1 Die Mapping-Datenbank

Integrationssysteme wie SRS und GenMapper ermitteln korrespondierende Objekte zweier Datenquellen durch eine Mehrwege-Join-Operation entlang des kürzesten Pfades zwischen den entsprechenden Quellen. Dabei sind verschiedene Probleme zu beobachten. Erstens ist der kürzeste Weg nicht immer der sinnvollste für bestimmte Nutzer oder Anwendungen. Zweitens können die Pfade immer noch sehr lang sein, was zu Performanzproblemen bei einer Komposition zur Laufzeit führen kann. Mappings zwischen beliebigen Quellen können zwar zur Performanzoptimierung vorberechnet und materiali-



a) Schema der Mapping-Datenbank **b) Schema der ADM-Datenbank**

Abbildung 4: Metadatenverwaltung in ADM- und Mapping-Datenbank

siert werden. Jedoch führt diese Vorgehensweise zu einem nicht mehr handhabbaren Datenbestand (Komplexität von n^2 bei n Quellen). Wir adressieren diese Probleme einerseits durch die Unterstützung alternativer Pfade, die in Zusammenarbeit mit den Nutzern festgelegt werden können, und andererseits durch die Vorberechnung der Mappings zu einer zuvor ausgewählten zentralen Quelle.

Die Datenquellen werden in unserem Ansatz ähnlich wie in COLUMBA sternförmig (multidimensional) miteinander verbunden. Die Mapping-Datenbank verwendet das in Abbildung 4a gezeigte Schema. Es benutzt den Identifikator einer zentralen Quelle (*Center_Accession* in der Tabelle *Center*), der mit den einzelnen quell-spezifischen Mappingtabellen über den Fremdschlüssel *Center_Id* verbunden ist. Jede Mappingtabelle beinhaltet die Mappings zwischen den Identifikatoren der zentralen Quelle und einer bestimmten Annotationsquelle. Verschiedene Beziehungen zwischen denselben Instanzen der beiden Quellen werden anhand des Mappingpfades (*Path_Id*) unterschieden. Diese Pfade, deren Bezeichnungen sich in der Tabelle *Path* der ADM-Datenbank (siehe Abschnitt 5.2) wieder finden, ergeben sich einerseits direkt aus den Mappingdaten einer Datenquelle oder andererseits aus der Komposition von Mappings über verschiedene Datenquellen hinweg. Eine solche Mappingkomposition wird beim Import der Mappingdaten ausgeführt und damit vorberechnet in der jeweiligen Mappingtabelle gespeichert.

Damit beispielsweise Annotationsdaten aus UniGene und Ensembl miteinander verknüpft werden können, ist ein Mapping zwischen beiden Datenquellen notwendig. Jedoch verfügt weder UniGene noch Ensembl über Korrespondenzen zur jeweils anderen Datenquelle. Damit ist kein direktes Mapping ableitbar und es muss ein Join-Weg, der weitere Quellen einschließt, gewählt werden. Ein möglicher Join-Weg besteht in UniGene – LocusLink – NetAffx – Ensembl. Da in unserem Ansatz LocusLink als zentrale Quelle gewählt wurde, enthält die Mappingtabelle *Center_UniGene* das direkte Mapping LocusLink – UniGene. Analog dazu existiert die Mappingtabelle *Center_Ensembl*. Da zwischen LocusLink und Ensembl kein direktes Mapping besteht, beinhaltet diese Mappingtabelle vorberechnete Mappings zwischen beiden Quellen (u.a. auch LocusLink – NetAffx – Ensembl), die anhand des Pfades (Join-Weg) unterschieden werden. Durch die sternförmige Anordnung der Mappingtabellen innerhalb der Mapping-Datenbank ist die Anzahl der zu materialisierenden Mappings linear beschränkt, auch wenn zusätzliche alternative Mappings unterstützt werden (k*n Mappings bei n Quellen und durchschnittlich k alternativen Mappings pro Quelle). Die Verknüpfung von Objekten zweier beliebiger Datenquellen kann günstig durch die Verknüpfung mit dem zentralen Identifikator durchgeführt werden. Die Auswahl der für die Verknüpfung zu verwendenden Mappings wird durch den Nutzer mit der Angabe des Mappingpfades auf der Web-Oberfläche getroffen.

Neue Annotationsquellen können flexibel hinzugefügt werden, indem die entsprechenden Mappings in eine neue Mappingtabelle eingefügt werden. Dadurch werden Annotationsquellen integriert, ohne dass dies zu einer Erhöhung der Laufzeitkomplexität führt, da der Join-Weg in der Mapping-Datenbank zwischen zwei Quellen nie mehr als 2 beträgt (Source – Center – Source). Umgekehrt können für nicht mehr benötigte Annotationsquellen die jeweiligen Mappingtabellen leicht entfernt werden. Die Speicherung der Mappingdaten je Datenquelle in eigenen Mappingtabellen erleichtert zudem die Admi-

nistration; insbesondere können die Mappingdaten je nach Aktualisierungszyklen der Quellen rasch erneuert werden. Der Aktualisierungsprozess verläuft entsprechend dem GenMapper Ansatz durch Löschen der alten und Import der neuen Mappingdaten [DR04]. Die lokalen Kopien der Annotationsquellen können unabhängig davon erneut repliziert werden.

Die Integration einer neuen Datenquelle setzt zudem voraus, dass mindestens ein Mapping zwischen dem Identifikator der zentralen Quelle und der neuen Quelle existiert, unabhängig davon, ob es ein direktes Mapping ist oder aus einer Komposition abgeleitet und vorberechnet wurde. Deshalb hat die Auswahl der zentralen Quelle und somit die Bildung des zentralen Identifikators maßgeblichen Einfluss auf diese Art der Datenintegration. Neben Aktualität, Redundanzgrad, und Akzeptanz der Objekte ist die Anzahl der in der Quelle des zentralen Identifikators enthaltenen Mappings zu berücksichtigen. Solche können in die Mapping-Datenbank importiert und zur Verknüpfung mit anderen Quellen benutzt werden. Beispielsweise ist LocusLink eine Referenz-Datenquelle für Gendaten, während sich SwissProt für Protein-bezogene Daten eignet.

5.2 Die ADM-Datenbank

Abbildung 4b zeigt einen Ausschnitt aus dem Schema der ADM-Datenbank, der zur Verwaltung von Metadaten der integrierten Datenquellen verwendet wird. Zurzeit werden diese Metadaten teilweise manuell, teilweise automatisch durch Datenbankskripte aus den entsprechenden Datenquellen extrahiert und importiert. Die Quellen werden in der *Source*-Tabelle mit einer eindeutigen Nummer (*Db_Id*) und einem Namen verwaltet. Zusätzlich werden die Attribute der einzelnen Quellen für deren Auswahl auf der Web-Oberfläche gespeichert. Dazu dient die *Attribute*-Tabelle, in der die jeweiligen Attribute der Datenquellen mit einer eindeutigen Nummer (*Attribute_Id*), einem dem Nutzer verständlichen Namen, dem SRS-internen Bezeichner sowie der zugehörigen Quelle enthalten sind. Die Pfade, die zur Berechnung der in die Mapping-Datenbank importierten Mappings benutzt wurden, beinhaltet die *Path*-Tabelle. Der Pfadname setzt sich aus den Namen der beteiligten Quellen zusammen, so dass der Nutzer später zwischen alternativen Mappings unterscheiden sowie sich über die Semantik einzelner Mappings informieren kann. Im Abschnitt 6.2 diskutieren wir, wie Web-Oberflächen aus diesen Metadaten für die Anfragespezifikation automatisch generiert werden.

6 Anfragebearbeitung im Query-Mediator

6.1 Anfragetypen

Ausgehend von den spezifischen Anforderungen der Expressions- und Annotationsanalysen unterscheidet der Query-Mediator derzeit zwei Typen von Anfragen, *Projektions-* und *Selektionsanfragen:*

- Projektionsanfragen unterstützen die Expressionsanalyse und erlauben die gemeinsame Darstellung der durch den Benutzer spezifizierten Attribute verschiedener Annotationsquellen für eine Gruppe von Genen.

- Selektionsanfragen unterstützen die Annotationsanalyse und suchen Gene mit bestimmten funktionalen Eigenschaften anhand von spezifizierten Filterbedingungen. Das Ergebnis mündet in einer Gengruppe, die dann in *GeWare* zur Analyse des Expressionsverhaltens weiter verwendet werden kann.

Die zwei Anfragetypen unterscheiden sich nur in den Ein- und Ausgaben. Projektionsanfragen benötigen eine Gengruppe als Eingabe, während Selektionsanfragen eine Gengruppe als Ausgabe erzeugen. Die Verarbeitung ist in beiden Fällen ähnlich und konzentriert sich auf die Assoziation der Gene mit den Attributen aus entsprechenden Datenquellen. Im folgenden Abschnitt wird dieser Prozess ausführlicher diskutiert.

6.2 Anfrageformulierung

Die Anfragen werden im Webbrowser durch die Auswahl der relevanten Attribute (Projektion) und die Spezifikation der Filterbedingungen (Selektion) formuliert. Aus den Metadaten der ADM-Datenbank (siehe Abschnitt 5.2) wird die Web-Oberfläche, wie sie die Abbildung 5 für eine einfache Selektionsanfrage zeigt, zur Formulierung der Anfrage automatisch generiert. Gesucht sind z.B. alle Gene der Datenquelle *NetAffx (Set U95)*, d.h. Gene des Microarray-Sets *U95*, die bestimmte Eigenschaften besitzen: sie befinden sich auf dem Chromosom *vier* und sind mit dem biologischen Prozess *„cell migration“* assoziiert.

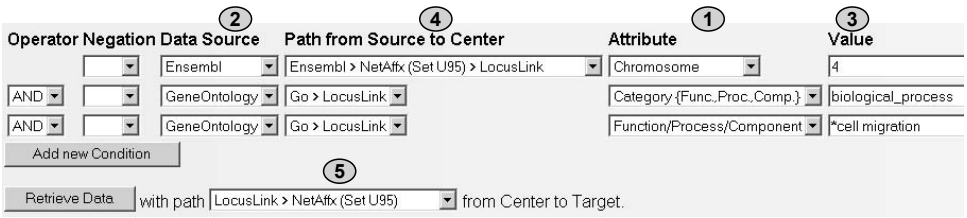


Abbildung 5: Anfrageformulierung auf der automatisch generierten Web-Oberfläche

Auf der Web-Oberfläche können beliebig viele Filterbedingungen zur Selektion bzw. Attribute zur Anzeige von Annotationen zu den Genen spezifiziert werden. Die Bedingungen bestehen aus einem Attribut (1) einer Datenquelle (2), für das ein Filterwert (3) für eine exakte oder Ähnlichkeitssuche angegeben werden kann. Ferner ist ein Mapping zwischen der gewählten Datenquelle und der zentralen Quelle, hier LocusLink, durch einen entsprechenden Pfad (4) auszuwählen. Die einzelnen Bedingungen können beliebig mit den logischen Operatoren OR, AND, und NOT kombiniert werden, wobei OR die niedrigste und NOT die höchste Priorität bei der Evaluation der Anfrage hat. Letztlich ist ein Mapping von der zentralen Quelle zur Zielquelle, im Beispiel NetAffx (Set U95), durch einen entsprechenden Pfad festzulegen (5).

Unabhängig von der Zielquelle, von der die relevanten Objekte in der Ergebnismenge enthalten sein sollen, können beliebige Attribute verschiedener Datenquellen abgefragt werden. Dies ist eine wichtige Verbesserung gegenüber SRS, das nur Filterbedingungen für Attribute einer einzelnen Quelle unterstützt. Ferner wird auch die Projektion verschiedener Attribute aus unterschiedlichen Quellen in einer gemeinsamen Anfrage unterstützt, was mit SRS in dieser flexiblen Art noch nicht möglich ist.

6.3 Generierung des Anfrageplans

Anhand der Nutzerspezifikation auf der Web-Oberfläche (siehe Abbildung 5) generiert der Query-Mediator eine SRS-spezifische Abfrage, welche in Abbildung 6 (Schritt 3) gezeigt wird. Diese Abfrage wird anschließend durch SRS abgearbeitet, wobei deren Performanz vom Anfrageplan abhängt. Der Query-Mediator optimiert bei der Erstellung der SRS-spezifischen Abfrage den Anfrageplan mit den folgenden Schritten:

1. *Blockbildung:* Die Filterbedingungen zur Selektion werden in Hinsicht auf den logischen Operator OR in einzelne Blöcke unterteilt. Innerhalb dieser Blöcke können anschließend Optimierungen der Abfrage durchgeführt werden. In unserem Beispiel sind die einzelnen Filterbedingungen einheitlich durch den logischen Operator AND verknüpft. Deshalb wird, wie in Abbildung 6 (Schritt 1) zu sehen, nur ein einziger Block gebildet, der aus den drei Attributen *Chromosome*, *Category* und *Process* besteht. Im Gegensatz dazu bilden die ausgewählten Attribute zur Projektion immer einen Block.

2. *Zusammenfassung quellspezifischer Attribute:* Für jeden resultierenden Block werden die Attribute und Filterbedingungen nach Datenquellen sowie nach den zu verwendenden Mappings sortiert. Damit ist es möglich, Attribute bzw. Filter, die dieselbe Datenquelle und dasselbe Mapping benutzen, so zusammenzufassen, dass sie in einer gemeinsamen Anfrage an die Datenquelle verwendet werden können. Im Beispiel stammen die beiden Attribute *Category* und *Process* aus der Quelle GeneOntology und benutzen dasselbe Mapping zum zentralen Identifikator. Abbildung 6 (Schritt 2) zeigt die Zusammenfassung dieser beiden Attribute, wohingegen sich das Attribut *Chromosome* der Quelle Ensembl abgrenzt.

3. *Zusammensetzung der SRS-Abfrage(n):* Die Namen der Quellen und Attribute werden durch die internen SRS-Bezeichner ersetzt. Die ausgewählten Mappings werden mit den Identifikatoren aus der Mapping-Datenbank versehen. Im Beispiel ist die zweite und dritte auf der Web-Oberfläche spezifizierte Filterbedingung (siehe Abbildung 5) in der Zeile 3 der in Abbildung 6 (Schritt 3) gezeigten SRS-Abfrage zu sehen. Die Quelle *GeneOntology* sowie die Attribute *Category* und *Process* wurden in die SRS-internen Namen *GoTerm* und *typ* bzw. *tna* übersetzt. Ebenfalls wurde die Nummer des Mappings (Nummer 1 im Beispiel) zwischen GeneOntology und LocusLink identifiziert. Die Anfrage kann nun zur Abarbeitung an SRS durch den Aufruf des Interpreters "*getz*" geschickt werden.

1. Schritt: Blockbildung

Block	*Pfad*	*Quelle*	*Attribut*	*Filterwert*
1	Ensembl>NetAffx(Set U95)>LocusLink	Ensembl	Chromosome	4
1	GeneOntology>LocusLink	GeneOntology	Category	biological_process
1	GeneOntology>LocusLink	GeneOntology	Process	*cell migration

2. Schritt: Zusammenfassung quellspezifischer Attribute

Block	*Zsfg.*	*Pfad*	*Quelle*	*Attribut*	*Filterwert*
1	a	Ensembl>NetAffx(Set U95)>LocusLink	Ensembl	Chromosome	4
1	b	GeneOntology>LocusLink	GeneOntology	Category	biological_process
1	b	GeneOntology>LocusLink	GeneOntology	Process	*cell migration

3. Schritt: Zusammensetzung der SRS-Abfrage

```
1 getz -vf "accession" "([Mapping-pid:5]
2   < (Center < ([Mapping-pid:2]<([EnsemblGene-cnm:4]))
3   < ([Mapping-pid:1]<([GoTerm-typ: biological_process] & [GoTerm-tna:*cell migration])))))
```

Abbildung 6: Schritte zur Erstellung des Anfrageplans

Aus der SRS-Abfrage in Abbildung 6 (Schritt 3) geht hervor, dass die Objekte von *EnsemblGene* bzw. *GoTerm* zuerst mit den jeweiligen Filtern identifiziert (Zeilen 2 und 3) und anschließend auf die *Center_Id* (zentraler Identifikator) abgebildet werden (Zeile 2). Die resultierenden zentralen Identifikatoren werden anschließend mit dem Mapping zwischen der zentralen Quelle und der Zielquelle (Mapping Nr. 5, Zeile 1) auf NetAffx abgebildet. Im Ergebnis wird eine Liste von „Accessions“ (Identifikatoren) der entsprechenden NetAffx-Objekte zurückgeliefert.

6.4 Extraktion und Transformation der Ergebnisse

SRS liefert als Antwort für jede Abfrage (Aufruf des "*getz*" Interpreters) einen XML-Datenstrom zurück, aus dem die notwendigen Daten durch den Query-Mediator extrahiert werden. Komplexe Anfragen, wie z.B. Projektionen mit vielen Attributen aus unterschiedlichen Datenquellen, können in mehrere SRS-Abfragen unterteilt werden, woraus ebenso viele XML-Datenströme als Antwort resultieren. Die extrahierten Daten werden anschließend im Query-Mediator zusammengesetzt. Ferner werden fehlende Anfragefunktionen einiger Quellen, die in SRS noch nicht berücksichtigt sind, wie z.B. der Mengendurchschnitt in MySQL, durch zusätzliche Transformationen im Query-Mediator übernommen. Das Ergebnis wird anschließend in das HTML-Format konvertiert und ausgegeben. Eine Gengruppe als Ergebnis einer Anfrage kann als Eingabe neu-

er Anfragen verwendet werden. Ebenso kann das Ergebnis einer Projektionsanfrage um weitere Attribute von verschiedenen Datenquellen erweitert werden. Dies ermöglicht eine iterative Analyse.

Die Abbildung 7a zeigt das Ergebnis der in Abschnitt 6.2 eingeführten Selektionsanfrage, die eine Menge von Genen (Affymetrix Probesets) identifiziert, die auf dem Chromosom *vier* lokalisiert und mit dem biologischen Prozess *cell migration* entsprechend der GeneOntology Klassifikation assoziiert sind. *GeWare* bietet die Möglichkeit, alle oder eine benutzerspezifische Auswahl dieser Gene in einer Gengruppe unter Angabe eines Gruppennamens abzuspeichern. Mit dieser Gruppe kann anschließend z.B. eine Projektionsanfrage durchgeführt werden, deren Ergebnis in Abbildung 7b gezeigt ist. Insbesondere wurden die korrespondierenden Identifikatoren von UniGene, die Gennamen von LocusLink sowie die assoziierten Funktionsnamen von GeneOntology abgefragt. Interessante Gene können hier ebenfalls ausgewählt und in einer neuen Gengruppe für weiterführende Analysen abgelegt werden.

a) Ergebnis einer Selektionsanfrage

Please specify a gene group name and select the probe sets you wish to save.

Gene Group Name: [ ] Save as gene group Unselect all

To download the results please use this link

In addition, annotation can be viewed for the selected probe sets. View annotation

Select?	Probe Sets
☑	39634_at
☑	56938_at
☑	159_at
☑	56940_g_at
☑	59308_at
☑	1934_s_at

Select?	Probe Sets	UniGene: UniGene Accession	Locuslink: Gene Name	GeneOntology: Function/Process/Component
☑	56938_at	Hs.29802	slit homolog 2 (Drosophila)	neurogenesis protein binding glia cell migration extracellular space motor axon guidance olfaction biological_process unknown neuronal cell recognition mesoderm migration chemorepellant activity calcium ion binding
☑	159_at	Hs.79141	vascular endothelial growth factor C	substrate-bound cell migration regulation of cell cycle membrane cell proliferation growth factor activity lymph gland development angiogenesis

b) Ergebnis einer Projektionsanfrage

Abbildung 7: Ergebnisse für Projektions- und Selektionsanfragen

7 Performanz

7.1 Testumgebung

Den Performanzmessungen lag eine Intel-Server-Plattform mit der folgenden Konfiguration als Testumgebung zugrunde.

Hardware	
CPU:	4 x Intel Xeon 2,5 GHz
Hauptspeicher:	8 GB DDR-RAM
Software	
Betriebssystem:	Linux, Fedora 2.4.22
Datenbanken:	IBM DB2 8.1.0
	MySQL, Version 4.0.17-max
SRS-Server:	SRS Relational 7.3.1 für Linux
Java:	Java 2 SUN-Plattform, Standard Edition Version 1.4.2

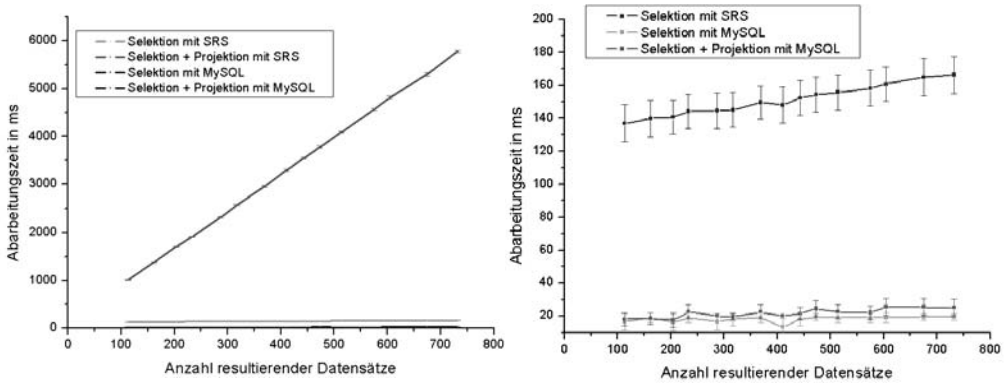
Das Data Warehouse wie auch die ADM- und die Mapping-Datenbank benutzen das relationale System DB2 von IBM in der angegebenen Version. Die Programmlogik des *GeWare*-Systems sowie des Query-Mediators wurden auf Basis der Java 2 SUN-Plattform implementiert. In SRS wurden die Mapping-Datenbank (IBM DB2, RDBMS) sowie die Annotationsquellen LocusLink (Datei), Ensembl (MySQL, RDBMS) und GeneOntology (MySQL, RDBMS) integriert.

Die Performanzmessungen wurden während des regulären Betriebes von *GeWare* durchgeführt. Da das Data Warehouses von den Benutzern in unregelmäßigen Abständen genutzt wird, verteilt sich die Belastung sehr unterschiedlich zwischen Spitzenzeiten und Ruhephasen. Alle Tests wurden unter geringer oder keiner Belastung ausgeführt und unter der Gewährleistung von mindestens 90% freier Prozessorkapazität.

7.2 Ausgewählte Messergebnisse

Aus einer Reihe von Performanzmessungen fokussieren wir hier auf zwei Auswertungen, die die Abarbeitungszeiten von Anfragen in Abhängigkeit zur Anzahl resultierender Datensätze untersuchen. Abbildung 8 zeigt die Ergebnisse der ersten Messreihe. Diese Messreihe soll klären, inwieweit Performanzeinbußen von SRS gegenüber RDBMS, wie bspw. MySQL, bei der Selektion und Projektion auftreten. Dazu wurden jeweils 15 Anfragen an die lokale Kopie der Annotationsquelle Ensembl, deren Daten in der relationalen Datenbank MySQL vorliegen, untersucht. Alle Anfragen benutzen einheitlich das Attribut *des* (Genbeschreibung), unterscheiden sich jedoch im angegebenen Filterwert und dadurch in der Anzahl der resultierenden Datensätze. Darüber hinaus wird zwischen Selektionsanfragen und einer Kombination von Selektion und Projektion unterschieden. Während erstere lediglich den Identifikator von Ensembl zurückgeben, liefern letztere zusätzlich zum Identifikator das gefilterte Attribut zurück. Jeder Messpunkt repräsentiert den Mittelwert aus 20 Wiederholungen einer Anfrage. Die Standardabweichung ist zu jedem Messpunkt als Fehlerbalken aufgetragen. Jede Abarbeitungszeit wurde unabhän-

gig von der Web-Oberfläche evaluiert, um störende Faktoren weitestgehend auszuschließen und um eine Vergleichbarkeit herzustellen.



a) Abarbeitungszeiten von SRS und MySQL **b) Ausschnitt der Abarbeitungszeiten aus a)**

Abbildung 8: Performanz von Projektions- und Selektionsanfragen in SRS und MySQL

Abbildung 8a stellt die Abarbeitungszeiten für die untersuchten Anfragearten, die Selektion sowie die Kombination von Selektion und Projektion, jeweils für SRS und MySQL in Abhängigkeit zur Größe der Ergebnismenge dar. Dabei zeigt sich, dass sowohl SRS als auch MySQL sehr geringe Abarbeitungszeiten (< 200 ms) zur Selektion benötigen, die mit zunehmender Anzahl an resultierenden Datensätzen nur schwach linear ansteigen (siehe Abbildung 8b). Auf die Abarbeitungszeiten für die Kombination von Selektion und Projektion in MySQL trifft dies ebenso zu. Dagegen führt die Kombination von Selektion und Projektion in SRS zu einem starken Anstieg der Abarbeitungszeit, die zudem mit dem Umfang der Ergebnismenge linear zunimmt.

Abbildung 9 stellt die Ergebnisse der zweiten Messreihe zur Performanzbewertung unserer Integrationslösung dar. Dazu wurden Selektionsanfragen an die Annotationsquelle Ensembl verwendet, für die die Abarbeitungszeiten nach jedem einzelnen Schritt des Anfrageplans (Selektion Ensembl, Mapping Ensembl – LocusLink, Mapping LocusLink – NetAffx, Projektion des NetAffx Identifikators) gemessen wurden. Dazu wurden 11 Anfragen generiert, die einheitlich das Attribut *des* (Genbeschreibung) der Datenquelle Ensembl mit verschiedenen Filterwerten benutzen. Wie bei der ersten Messreihe repräsentiert jeder Messpunkt den Mittelwert aus 20 Wiederholungen, wobei die Abarbeitungszeiten unabhängig von der Web-Oberfläche evaluiert wurden. Die Fehlerbalken charakterisieren die Standardabweichung.

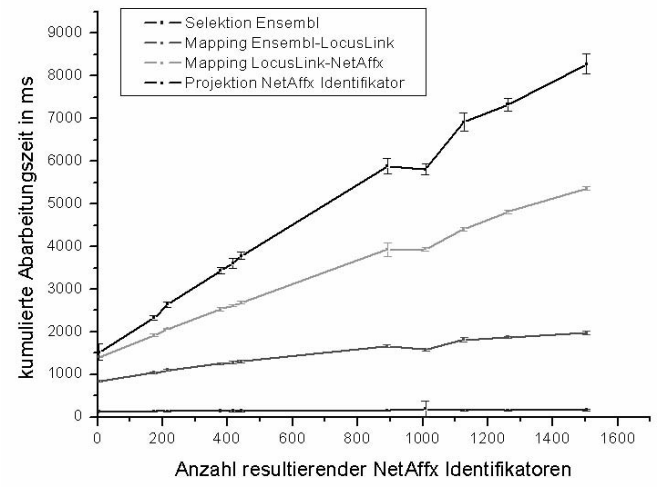


Abbildung 9: Performanz von Selektionsanfragen an die Annotationsquelle Ensembl

Abbildung 9 stellt die Abarbeitungszeiten für einzelne Schritte des Anfrageplans kumulativ dar. Das bedeutet, die Abarbeitungszeit eines betrachteten Schrittes beinhaltet bereits die Zeiten aller vorherigen Schritte und die Gesamtzeit ergibt sich mit der Projektion der NetAffx Identifikatoren. Insgesamt steigt die Bearbeitungszeit wiederum linear mit der Anzahl der Ergebnissätze und liegt für typische Größenordnungen in einem akzeptablen Bereich. Die Selektion der Ensembl-Identifikatoren benötigt die kürzeste Zeit. Dagegen ist die Abbildung dieser Ergebnismenge auf die Mapping-Datenbank noch sehr zeitaufwendig und deutet auf ein Potential zur Performanzverbesserung. Gegenwärtig wird dieser Schritt bzw. alle mit der Mapping-Datenbank korrespondierenden Schritte durch SRS übernommen. Einerseits soll in der neuen SRS Version 8 die Anbindung von relationalen Datenquellen bedeutend verbessert worden sein (allerdings nicht für DB2). Andererseits könnten sich weitere Verbesserungen erzielen lassen, wenn der Query-Mediator direkt auf die Mapping-Datenbank über ein definiertes API zugreift, um die Mappingkomposition *Datenquelle* → *LocusLink* → *Datenquelle* zur Laufzeit in einem Schritt auszuführen.

8 Zusammenfassung

Es wurde ein hybrider Integrationsansatz vorgestellt, um Annotationsdaten von molekularbiologischen Objekten wie Genen, Proteinen und Pathways aus öffentlichen Datenquellen für datenintensive Expressionsanalysen verwendbar zu machen. Die Expressionsdaten sind mit Experimentbeschreibungen physisch in einem Data Warehouse integriert, um schnelle Auswertungen zu unterstützen. Die öffentlichen Annotationsdaten werden virtuell über einen Mediatoransatz integriert und bedarfsgesteuert für Analysen abgerufen. Für die einheitliche Anbindung der Datenquellen wird das verbreitete Tool SRS (Sequence Retrieval System) der Fa. LION bioscience genutzt. Die Kopplung zwischen dem Warehouse und SRS erfolgt über den Query-Mediator. Wir extrahieren die

vorhandenen Mappings, die zwischen den einzelnen Datenquellen bestehen, und speichern sie in einer eigenen Datenbank zur semantischen Integration der Annotationsdaten. Dieser hybride Integrationsansatz wurde als Erweiterung des Leipziger Data Warehouse für Genexpressionsdaten (http://www.izbi.de/GEWARE) implementiert und wird für die Einbindung verschiedener öffentlicher Datenquellen in Annotations- und Expressionsanalysen eingesetzt. Die Integrationslösung ist jedoch nicht auf das Anwendungsfeld der Genexpression beschränkt, sondern kann auch in anderen Bereichen, wie z.B. zur strukturellen Klassifikation ganzer Genome und der Analyse von Proteinstrukturen, Verwendung finden. Die ersten Performanzanalysen zeigen die Praktikabilität des Ansatzes, jedoch auch die Abhängigkeit von der Bearbeitungsgeschwindigkeit von SRS.

Danksagung

Wir danken der Fa. LION bioscience, insbesondere Thure Etzold und Ceara Rea, für die Bereitstellung der Software SRS und für die fachliche Unterstützung bei der Installation und Konfiguration sowie der Einarbeitung in die Software.

Literaturverzeichnis

[As00] Ashburner, M. et al.: Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29, 2000.

[Bi04] Birney, E. et al.: An Overview of Ensembl. Genome Research 14: 925-928, 2004.

[CCW03] Chen, J.; Chung, S.Y.; Wong, L.: The Kleisli Query System as a Backbone for Bioinformatics Data Integration and Analysis. In [LC03]: 147-187.

[Ch04] Cheng, J. et al.*:* NetAffx gene ontology mining tool: a visual approach for microarray data analysis. Bioinformatics 20(9), 1462-3, 2004

[DR04] Do, H.-H.; Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. the 9th Int. Conf. on Extending Database Technology, Heraklion (Greece) 2004. Springer LNCS, 2004.

[EHB03] Etzold, T.; Harris, H.; Beaulah, S.: SRS: An Integration Platform for Databanks and Analysis Tools in Bioinformatics. In [LC03]: 109-145.

[Ga04] Galperin, M.Y.: The Molecular Biology Database Collection - 2004 update. Nucleic Acids Research 32, Database issue, 2004.

[Ha01] Haas, L. et al.: DiscoveryLink – A System for Integrated Access to Life Sciences Data Sources. IBM System Journal 40 (2), 2001.

[HK04] Hernandez, T.; Kambhampati, S.: Integration of Biological Sources: Current Systems and Challenges Ahead. SIGMOD Record 33(3), 2004.

[KDR03] Kirsten, T.; Do, H.-H.; Rahm, E.: A Multidimensional Data Warehouse for Gene Expression Analysis. In: Proc. German Conference on Bioinformatics, Munich 2003.

[KDR04] Kirsten, T.; Do, H.-H.; Rahm, E.: A Data Warehouse for Multidimensional Gene Expression Analysis. Technischer Report, IZBI Universität Leipzig, 2004.

[LC03] Lacroix, Z.; Critchlow T. (Hrsg.): Bioinformatics: Managing Scientific Data. Morgan Kaufmann, 2003.

[PM02] Pruitt, K.D.; Maglott, D.R.: RefSeq and LocusLink: NCBI Gene-centered Resources. Nucleic Acids Research 29 (1), 2001.

[Po04] Potter, S.C. et al.: The Ensembl Analysis Pipeline. Genome Research 14: 934-941, 2004.

[PWS03] Pontius, J.U.; Wagner, L.; Schuler, G.D.: UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information, 2003.

[Ra04] Rahm, E. (Hrsg): Proceedings 1st Intl. Workshop Data Integration in the Life Sciences (DILS) 2004. LNBI 2994, Springer-Verlag, 2004.

[Ro04] Rother, K. et al.: COLUMBA: Multidimensional Data Integration of Protein Annotations. In [Ra04]: 156-171.

[St03] Stein, L.: Integrating Biological Databases. Nature Review Genetics 4(5): 337-345, 2003.

[Wh03] Wheeler D.L. et al.: Database Resources of the National Center for Biotechnology. Nucleic Acids Research 31: 28-33, 2003.

[Wo98] Wong, L.: Kleisli, a Functional Query System. Journal of Functional Programming, 1 (1): 1-000, 1998.

[Zd02] Zdobnov, E.M. et al.: The EBI SRS server – recent developments. Bioinformatics 18: 368-373, 2002.