Data-Warehouse- und Mapping-basierte Datenintegrationsplattformen in der Bioinformatik

Diese Dissertation beschäftigt sich mit der Konzeption und dem Aufbau von Plattformen zur Integration von Daten im Bereich der Bioinformatik, mit denen eine effiziente und zielgerichtete Datenanalyse unterstützt wird. Die Daten sind dabei einerseits das Resultat verschiedener molekularbiologischer Experimente und andererseits Inhalt verschiedenster Datenquellen.

Im *Genetic Data Warehouse* (*GeWare*) werden experimentelle Daten zentral zusammengefasst, die mit Hochdurchsatz-Technologien zur Untersuchung der Genexpression erzeugt wurden. Assoziiert zu diesen experimentellen Daten, kann *GeWare* Metadaten speichern, die das Experiment aus aufbau- und ablauforganisatorischer Sicht beschreiben und damit nachvollziehbar und reproduzierbar machen. Dazu bietet die Plattform so genannte *Annotation Templates*, die eine Menge von Kategorien strukturieren, für die die atomaren Annotationswerte in Bezug auf die experimentelle Beschreibung erfasst werden. Das Konzept der *Templates* kombiniert eine für den Benutzer größtmögliche Flexibilität bei der Definition, Modifikation und Nutzung der Templates mit einem für die Plattformadministration vernachlässigbaren Aufwand. Mit ihnen können (selbst nachträgliche) Anpassungen an experimentspezifische Erfordernisse vorgenommen werden, ohne dass Änderungen am zugrunde liegenden Datenmodell notwendig werden. Darüber hinaus eignet sich das Konzept der *Annotation Templates* auch zur Aufnahme klinischer Parameter, z.B. aus einem Studienverwaltungssystem. Zusätzlich ist die *GeWare*-Plattform mit einem Mediator gekoppelt, der Daten aus ausgewählten Quellen virtuell integriert und damit eine kombinierte und iterative Analyse der experimentellen Daten mit denen der öffentlich verfügbaren Quellen ermöglicht. In diese Integrationslösung ist die Software SRS (Sequence Retrieval System) eingebunden, die auf Basis einer umfangreichen Wrapper-Bibliothek vor allem den Zugriff auf die angebundenen Datenquellen sicherstellt. Kern der Integrationslösung ist eine zentrale Mapping-Datenbank, die Mengen von Korrespondenzen (Mappings) zwischen den Objekten/Instanzen der integrierten Quellen aufnimmt und der effizienten Anfrageverarbeitung dient.

Die *BioFuice*-Plattform nutzt Mappings, um Daten aus privaten und frei verfügbaren Datenquellen sowie Ontologien im Bereich der Bioinformatik zu integrieren. Die Mappings repräsentieren hierbei Korrespondenzen zwischen Objekten spezifischer Typen (z.B. Gen, Protein), die sowohl innerhalb einer Datenquelle als auch zwischen unterschiedlichen Quellen bestehen. Mengenbasierte Operatoren übernehmen die Ausführung der Mappings und können in Skripten zur Abbildung von ad-hoc Workflows zusammengefasst werden. *BioFuice* bietet ein mächtiges GUI, in der Anfragen verschiedenartig formuliert und ausgeführt werden. Dazu zählen neben der freien Skriptprogrammierung und der Abarbeitung von parametrisierten Skripten insbesondere die Formulierung und Ausführung einer Stichwortsuche sowie modellbasierter Anfragen. Letztere erfordern eine automatische Transformation in ausführbare Skripte. Darüber hinaus bietet *BioFuice* einen Datenexport in für die Bioinformatik spezifische Datenformate und eine Schnittstelle zur statistischen Software R, mit der die integrierten Daten einer statistischen Analyse zugeführt werden können.

Mit *GeWare* und *BioFuice* wurden auf Basis unterschiedlicher Anforderun-

gen zwei Plattformen konzipiert und aufgebaut, die im Bereich der Bioinformatik Daten aus unterschiedlichen Quellen integrieren und für umfangreiche Analysen nutzbar machen. Die Plattformen wurden in verschiedenen Projekten verwendet und konnten die dort gestellten Anforderungen abdecken.

Leipzig, Dezember 2007 Toralf Kirsten