Peter Rieger, Stephan Heymann, Heiko Müller

Datenbankgestützte Wissensakquisition in den Lebenswissenschaften

Die Integration und Analyse molekularbiologischer Daten spielt eine entscheidende Rolle in der lebenswissenschaftlichen Forschung. Zum Zwecke der Wissensakquisition werden dabei Daten aus heterogenen Quellen integriert, analysiert und die Ergebnisse als Ausgangspunkt für einen weiteren Integrations- und Analyseschritt benutzt. Aus den speziellen Eigenschaften der Datenquellen und des oftmals explorativen Vorgehens ergeben sich eine Reihe von Anforderungen an die Datenverarbeitung, die über das klassische Data Warehouse hinaus gehen. Wir haben die Integrations- und Analyseplattform GENE-EYE entwickelt, um diesen spezifischen Anforderungen gerecht zu werden. Sie bietet die Fähigkeit einer einfachen Einbindung heterogener Datenquellen in ein relationales Datenbanksystem. Ein globales Schema exisitert dabei nicht. Stattdessen kann der Anwender eigene Zielschemata definieren und diese mit Daten aus den eingebundenen Quellen instanziieren. Zusätzlich werden eine Reihe von Algorithmen der Bioinformatik bereitgestellt, mit denen auf diesen Daten gearbeitet werden kann. Anhand eines Beispieles wird der praktische Einsatz der vorgestellten Plattform demonstriert.

1 Einleitung – Wissensakquisition in den Lebenswissenschaften

Die Lebenswissenschaften umfassen ein breites Spektrum an Forschungsgebieten, deren Ziel das Verständnis des komplexen Phänomens Leben ist. Die Genomforschung liefert mit ihrer Aufklärung des Aufbaus der Baupläne des Lebens (Genome) die molekularbiologische Grundlage der Lebenswissenschaften. Breitgefächert sind neben den Forschungsgebieten auch die dabei vorkommenden Datentypen. Den größten Anteil nehmen Informationen über die Zusammensetzung, Struktur und Funktion der Biomoleküle ein. Weitere Daten beschreiben die Aktivität einzelner Genomabschnitte (Gene) unter variierenden äußeren Bedingungen (Genexpression), die Ursachen genetisch bedingter Krankheiten sowie das Zusammenwirken der Biomoleküle in biologischen Systemen und Netzwerken. Die Daten werden in einer Vielzahl heterogener Datenquellen bereitgehalten. Mit der Bioinformatik ist eine Disziplin entstanden, deren Ziel es ist, die anfallenden Daten adäquat zu verwalten und Methoden zu deren effizienten Verarbeitung und Analyse bereitzustellen.

Wissensakquisition in den Lebenswissenschaften basiert darauf, die Erkenntnisse der Genomforschung in Zusammenhang zu setzten, daraus Hypothesen z.B. für das Auftreten von Krankheiten und Umweltproblemen abzuleiten, diese zu verifizieren sowie Therapien und Lösungen zu entwickeln. Dies bedingt einen zyklischen Prozess aus Datengenerierung, -integration und -analyse. Die Integration von Daten in den Lebenswissenschaften dient daher wesentlich stärker dem Versuch, Zusammenhänge gestützt auf empirische Beobachtungen zu entdecken, als der Notwendigkeit, die Vielzahl unabhängiger erhobener Fakten in einem einheitlichen Modell zu konsolidieren.

Angesichts der Bedeutung, die Integrationsprojekte für eine erfolgreiche Unterstützung lebenswissenschaftlicher Forschungen haben, sind eine Vielzahl an Systemen und Plattformen entwickelt worden. Im letzten Jahr wurde an dieser Stelle ein ausführlicher Überblick über die Anforderungen und vorgeschlagenen Lösungen zum Thema Integration molekularbiologischer Datenquellen gegeben [LR03]. Der prominenteste Vertreter solcher Systeme ist sicherlich das ‚Sequence Retrieval System' (SRS) [EUA96]. SRS bietet für die Abfrage von mehr als hundert Datenquellen eine einheitliche Abfrageschnittstelle, unterstützt aber leider datenquellenübergreifende Abfragen so gut wie nicht. Dem gegenüber stehen integrierte Lösungen mit globalem Schema wie z.B. Tambis [BBB+98], die es mit Hilfe einer speziellen Ontologie eine feste Anzahl von Datenquellen transparent abzufragen. Diese Lösung hat hauptsächlich im akademischen Bereich Aufmerksamkeit gewonnen, konnte sich aber aufgrund der mangelnden Berücksichtigung von Aspekten wie Systemantwortzeit nicht im praktischen Einsatz durchsetzen.

Ein genereller Nachteil existierender Integrationsansätze ist ihre mangelnde Unterstützung der Analyseprozesse sowie die Verwaltung und Pflege der daraus abgeleiteten Daten. Im Rahmen dieser Arbeit geben wir einen Überblick über die Anforderungen an datenbankgestützte Wissensakquisition in den Lebenswissenschaften und beschreiben eine Architektur zu deren Realisierung. Im folgenden Abschnitt beschreiben wir dazu die Anforderungen an ein System zur datenbankgestützten Wissensakquisition in den Lebenswissenschaften. Abschnitt 3 stellt dann die von uns zu diesem Zwecke entwickelte Plattform vor. In Abschnitt 4 werden anhand eines Beispieles die Abläufe in den einzelnen Schichten der Architektur näher erklärt. Eine Zusammenfassung und ein Ausblick werden in Abschnitt 5 gegeben.

2 Herausforderungen datenbankgestützter Genomdatenanalyse

Wie bereits erwähnt bedarf die Forschung in den Lebenswissenschaften der Integration und Analyse unterschiedlicher molekularbiologischer Datentypen. Über die Jahrzehnte haben sich eine Vielzahl molekularbiologisch relevanter Datensammlungen und -formate herausgebildet. Das Journal *Nucleic Acid Research* gibt in seiner jährlichen Ausgabe „*Database Issue*" [NAR04] hierzu einen guten Überblick. Aus den speziellen Eigenschaften der Datenquellen und -analysen, sowie der oftmals durch unsicheres und unvollständiges Wissen bedingten explorativen Vorgehensweise ergeben sich eine Reihe von Anforderungen an die Verwaltung und Auswertung molekularbiologischer Daten, die teilweise über das klassische Data Warehousing hinaus gehen.

2.1 Integrierte Verwendung von Daten und Algorithmen

Das Spektrum der vorhandenen Datenformate reicht von syntaktisch stark überladenen Flat-File Formaten über XML-Dokumente und relationale Datenbanken bis zu speziell für Einzelfragestellungen entwickelte Objekt-Datenbanksysteme. Die Forderung an eine Integrationsplattform besteht nun darin, Methoden und Werkzeuge bereitzustellen, die eine einfache Einbindung verschiedener Quellen und einen einheitlichen Zugriff auf die Daten ermöglichen. Dies reicht von einem Zugriff auf einzelne Datenquellen bis hin zur datenquellenübergreifenden Integration in ein problembezogenes Zielschema.

Neben den Daten und Erkenntnissen sind durch die Bioinformatik eine Vielzahl von Algorithmen und Anwendungen entstanden, die heute unverzichtbar für den Prozess der Wissensakquisition geworden sind. Traditionell sind diese Anwendungen an spezifische Ein- und Ausgabeformate gebunden, was ihre Anwendung auf selbstgenerierte, integrierte Datensätze erschwert. Um die explorative Arbeitsweise geeignet unterstützen zu können, muss neben der flexiblen Einbindung von Datenquellen auch die flexible Anwendung der vorhandenen Analysealgorithmen auf die Daten gewährleistet sein. Es muss deshalb eine einheitliche Bearbeitungsschnittstelle bereitgestellt werden, die es angefangen von der Abfrage und Filterung der Daten bis zur schrittweise Komposition vollständiger Workflows erlaubt, die Arbeitsweise an den Erkenntnisbedürfnissen zu orientieren und nicht an den vorhandenen Skripten zur Transformation der Daten zwischen den einzelnen Ein-/Ausgabeformaten.

2.2 Dokumentation von Datenabhängigkeiten und Analyseprozessen

Genomdaten lassen sich grob in Rohdaten, d.h. Ergebnisse von Laborexperimenten, und abgeleitete Daten (Sekundärdaten) klassifizieren. Letztere sind das Ergebnis von Analyseprozessen oder –workflows, die wiederum auf Rohdaten und Sekundärdaten beruhen. Dadurch entsteht ein Netzwerk aus Datenabhängigkeiten. Diese Abhängigkeiten werden in der Regel ebenso wenig dokumentiert wie die Prozesse der Datengenerierung. Hierdurch wird die Nachvollziehbarkeit und somit die Möglichkeit zur Verifikation der Ergebnisse durch Dritte erschwert. Oftmals werden in den Daten sprachliche Ausschmückungen und Unsicherheiten wie „es könnte sein, dass“ oder „wir vermuten, dass“ unterschlagen. Der Nutzer der Sekundärdaten kann daher in der Regel keine fundierten Rückschlüsse auf die Belastbarkeit einer Aussage treffen. Daher ist die systematische und vollständige Dokumentation der Datenherkunft und der ausgeführten Analyseprozesse eine wesentliche Anforderung an eine Integrations- und Analyseplattform.

2.3 Modellmanagement

Derzeit existieren keine globalen, vereinheitlichten Modelle zur datenquellenübergreifenden Abbildung biologischer Zusammenhänge. In einer Integrationsplattform, die es erlaubt, Datenquellen miteinander zu verknüpfen, ihre Verwendung zu Dokumentieren und Workflows in dieser Infrastruktur zu planen und durchzuführen, sollte auch das Management der Modelle so ausgestaltet werden, dass ein kontinuierliches Anpassen der Modelle an den schrittweise Fortschritt der Erkenntnisse und Einsichten unterstützt wird [BLP00]. Modelle, insbesondere Schemata der zu integrierenden Quellen, sollen dabei nicht nur Dokumentationscharakter haben, sie sollen auch *unmittelbar* die Automatisierung wichtiger Transformations- und Verarbeitungsschritte auf der Datenebene ermöglichen.

2.4 Datenqualität und Datenbereinigung

Ein zusätzliches Problem für die Analyse von Genomdaten ist ihre stark schwankende und teilweise sehr zweifelhafte Qualität. Die Ursache hierfür liegt im Prozess der Datengewinnung und -verwaltung begründet [MNF03]. So kann mangelnde Dokumentation der Datenabhängigkeiten schnell zu veralteten Daten führen, wenn Änderungen in der Datenbasis, die Auswirkungen auf daraus abgeleitete Sekundärdaten haben, unbemerkt bleiben. Da die Daten u.a. als Grundlage millionenschwerer Forschung zur Medikamententwicklung dienen, sollte eine hohe Datenqualität absolute Priorität besitzen. Die Verbesserung der Qualität existierender Daten durch identifizieren und beseitigen von Fehlern und Widersprüchen wird als Datenbereinigung (*data cleansing*) bezeichnet.. Wir unterscheiden grob zwischen syntaktischer und semantischer Datenbereinigung.

Die syntaktische Datenbereinigung erfolgt im Rahmen des ETL-Prozess. Hierbei werden Abweichungen hinsichtlich des Datenformats und der verwendeten Wertebereiche aufgedeckt und beseitigt.

Die weitaus größere Herausforderung stellt die generelle Verifikation der Korrektheit der Daten dar (semantische Datenbereinigung). Nur in wenigen Fällen können gesicherte biologische Erkenntnisse herangezogen werden, um fehlerhafte Informationen aufzudecken. Oftmals kommt es jedoch im Rahmen von Berechnungen vor, dass einzelne Daten als fehlerhaft erkannt werden. Diese in den Analysen gewonnenen Erkenntnisse sollen deshalb direkt zur Korrektur fehlerhafter Daten herangezogen werden. Somit wird der aufwendige Prozess der Datenbereinigung handhabbar. Die bereinigten Daten werden in die Datenbasis eingepflegt und stehen somit für nachfolgende Analysen zur Verfügung. Die Möglichkeit zur Datenbereinigung sollte deshalb integraler Bestandteil einer Integrations- und Analyseplattform sein.

3 Eine Integrations- und Analyseplattform für die Lebenswissenschaften

Im folgenden Abschnitt wird die von uns Entwickelte Plattform GENE-EYE zur Integration und Analyse von Genomdaten näher beschrieben. Es handelt sich dabei um eine mehrschichtige Architektur, welche die wesentlichen Teilschritte Datenzugriff, Integration und Analyse nachbildet und den für die Wissensakquisition notwendigen iterativen Durchlauf dieser Schichten unterstützt [MRTF04]. Die Daten werden dabei physisch in einem Data Warehouse materialisiert. Dies hat

die Vorteile eines performanten Zugriffs, einer besseren Kontrolle bezüglich Änderungen in der Datenbasis sowie einer einfacheren Einbindung von Datenkorrekturen. Durch dieses Vorgehen geht zwar ein Teil der Aktualität verloren, da aber insbesondere bei umfangreichen Projekten, die neben den informationstechnischen Aufgaben auch labortechnische Teilprojekte beinhalten, die Stabilität der Basisdaten von Bedeutung ist, wird dieser Effekt akzeptiert.

3.1 Die GENE-EYE Architektur

GENE-EYE (Abb. 1) ist konzeptionell in drei Ebenen aufgeteilt. Die unterste Ebene (*Genome Data Store*) dient der syntaktischen Homogenisierung und Materialisierung des Datenbestandes in einem relationalen Datenbankmanagementsyste (RDBMS). Sie repräsentiert jede der Datenquellen als Instanz eines relationalen Schemas, welches direkt aus dem jeweiligen Format der Datenquelle abgeleitet wird. Neben den ggf. erforderlichen Datentypwandlungen findet auf dieser Ebene nur eine Bereinigung syntaktischer Fehler in den Daten statt. Es erfolgt noch keine Integration über Datenquellengrenzen hinweg. Die Einbindung von Daten, die bereits im relationalen Format vorliegen oder auf die über eine relationale Schnittstelle (z.B. ODBC) zugegriffen werden kann, ist unproblematisch. Die Herausforderung besteht vielmehr in der Umsetzung der Datenquellen, die lediglich in semistrukturierter Form vorliegen, als Flat-Files oder XML-Dokumente. Für diese Datenquellen muss zuerst ein relationales Schema entwickelt werden und die Abbildung der semistrukturierter Daten auf dieses Schema erfolgen. Dies wird mit Hilfe des in Abschnitt 3.2 beschriebenen Metadatenmodels weitestgehend automatisiert.

Die Komposition von Sichten, die Daten aus beliebigen Quellen enthalten können, erfolgt in der zweiten Ebene (*Genome Database*). Ziel ist es dabei, Schritt für Schritt zu Schemata zu gelangen, die biologische Konzepte und deren Zusammenhänge angemessen abbilden. In dieser Ebene findet somit die semantische Integration der Datenquellen statt. Da alle Daten in relationaler Form vorliegen, kann der volle Sprachumfang von SQL für diese Aufgabe verwendet werden. Zusätzlich erforderliche Operationen, die nicht in der Funktionsbibliothek des RDBMS enthalten sind, als so genannte ‚Benutzerdefinierte Funktionen' in die Verarbeitung einbezogen werden (siehe Abschnitt 3.4).

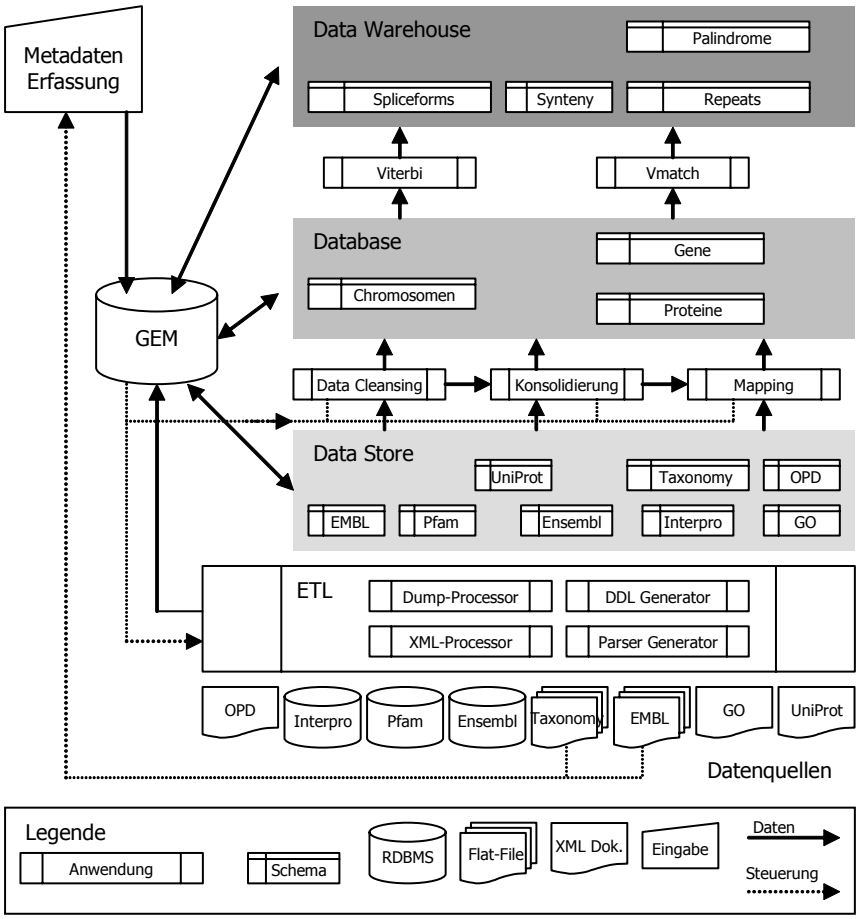

Abb. 1: Architektur der GENE-EYE-Plattform

In der dritten Ebene (*Genome Data Warehouse*) werden schließlich die Ergebnisse aus den Analysen der Artefakte der zweiten Ebene gespeichert und verwaltet. Diese weitere Aufteilung hat im Wesentlichen den Zweck, den Bestand der abgeleiteten Daten vor den möglicherweise nicht synchronisierten Änderungen im Bestand der Daten auf der ersten und zweiten Ebene zu kapseln. Das ermöglicht es, die Abstammungslinien und Herkunft der Ergebnisse über den Zeitverlauf zu verfolgen und zu dokumentieren. Sollen Ergebnisse der dritten Ebene selbst wieder Ausgangspunkt oder Bestandteil weiterer Vorhaben werden, können sie als Basisdaten in die erste Ebene eingestellt werden.

3.2 GENE-EYE Meta-Modell Management

Um die Ableitung eines relationalen Schemas durch das System unterstützen zu können, ist es notwendig, Inhalt, Struktur und Format der semistrukturierten Daten systematisch zu erfassen. Zu diesem Zweck wurde ein Metadatenmodell, GENE-EYE Metadata (GEM), entwickelt. Diese Metadaten bilden dann den Ausgangspunkt für die Automatisierung der Datenintegration.

Die **Charakterisierung der Datenquelle** ist die wichtigste Aufgabe des GEM. Darunter werden die Beschreibung der Struktur, des Inhalts und des Formates sowie die Zuordnung der Strukturelemente zum relationalen Abbild verstanden. Die Erfassung der Charakteristika legt ein einfaches Dokumentenmodell zugrunde: Eine Datenquelle umfasst eine Menge von Einträgen, die jeweils strukturiert bestimmte Eigenschaften (Aspekte) des beschriebenen Objektes oder Sachverhaltes abbilden.

Die Aspekte können untereinander in hierarchisch organisierten Beziehungen stehen und jeweils beliebig häufig auftreten. Ein Aspekt wiederum wird durch eine Menge von Attributen spezifiziert. Da für Flat-File Formate in der Regel keine systematische, maschinenlesbare Dokumentation vorliegt, muss die Erhebung der Daten hier manuell erfolgen. Die Erfassung der notwendigen Metadaten für XML-Dokumente und relationale Datenquellen erfolgt automatisiert.

Weitere Komponenten des Metadatenmodells betreffen die Beschreibung der **Bereitstellung einer Datenquelle** und die Zuweisung der **lokalen Resourcen**. Ersteres umfasst Angaben über den Anbieter der Datenquelle, die Zeitpunkte, zu denen neuen Versionen der Quelle erscheinen, die Zugriffsmethode und die dafür notwendigen Kontoinformationen. Nach dem Bezug einer neuen Version der Datenquelle liegt sie lokal für die physische Integration vor. Durch die Verknüpfung der Bereitstellungsmechanismen mit den eigentlichen Daten kann später die Herkunft und Weiterverarbeitung jedes integrierten Datenwerts nachverfolgt werden. Für die Aufbereitung, Umwandlung und Speicherung der Daten sind lokale Ressourcen erforderlich, deren Zuordnung für die einzelnen Datenquellen ebenfalls über einen Satz an Metadaten geregelt wird. Hierzu zählen u.a. Tabellenbereiche für die Speicherung von Daten, Indexen und großen Objekten.

3.3 Automatisierung der physischen Datenintegration (ETL)

Auf Basis der vollständigen Beschreibung der Datenquelle im Metadatenmodell werden dann folgende Prozesse automatisiert ausgeführt:

- Erzeugung der erforderlichen SQL-DDL Anweisungen, um das Schema für die Integration der Datenquelle anlegen zu können.
- Erzeugung eines Parsers, der die erforderlichen Ladedateien aus den verfügbaren Rohdaten aufbereitet.
- Erzeugung von Verwaltungsskripten, die für die Durchführung der oben benannten Operationen erforderlich sind.

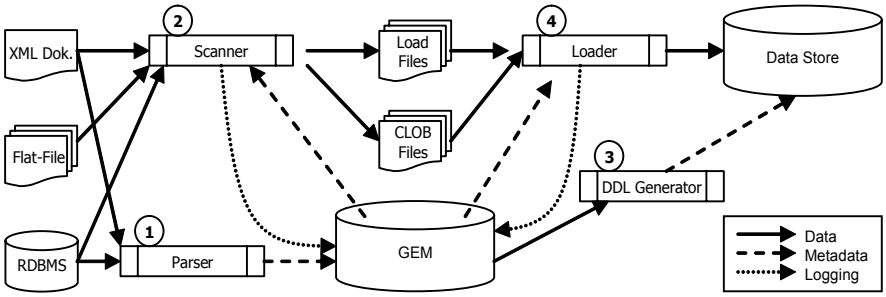

Abb. 2: Ablauf des ETL-Prozesses

- Erkennung und Dokumentation von Fehlern, soweit sie auf Basis der Metadaten bei der Durchführung der oben genannten Operationen festgestellt werden können.

Abb. 2 stellt den Daten- und Metadatenfluss für die Integration von Datenquellen dar. Nachdem die erforderlichen Metadaten erfasst wurden (1), wird der erzeugte Parser verwendet, um die Eingabedaten entsprechend der Struktur des abgeleiteten Schemas zu transformieren (2). Während dieses Prozesses werden zu jedem Attribut statistische Informationen gesammelt, die notwendig sind, um die syntaktische Definition des Zielschemas abschließen zu können. Dabei handelt es sich insbesondere um die Feststellung von Wertebereichen numerischer Attribute sowie die tatsächlich auftretenden Längen von Zeichenketten, da verfügbare Dokumentationen in der Regel nicht von vornherein preisgeben, ob z.B. ein CHAR, VARCHAR oder CLOB als Datentyp eines Attributes zu bevorzugen ist. Mit Hilfe dieser Daten ist es dann erst möglich, die Größe der Speicherseiten für das Schema und die passenden Tabellenbereiche für die Speicherung der Daten festzulegen (3). Da die Struktur des Schemas bereits bekannt ist und die Ladedateien für die Datenbank in einem Format erzeugt werden, das die Feldbreite nicht benötigt, ist es dennoch ausreichend, die Quelldaten ein einziges mal zu durchlaufen. Nach der Erzeugung DDL Anweisungen wird der eigentliche Ladevorgang angestoßen, um die Daten in die *Genome Data-Store* Ebene von GENEEYE zu laden (4).

Im Rahmen der physischen Datenintegration erfolgt auch die Korrektur syntaktischer Probleme in den Datenquellen (syntaktische Datenbereinigung). Die Notwendigen Abfragen und Transformationen können direkt durch die Spezifikation entsprechender Integritätsbedingungen in den quellenspezifischen Parser integriert werden. Sie decken Formatverletzungen auf und standardisieren die Daten. Durch die Verwendung geeigneter Wörterbücher werden Schreibfehler beseitigt, einheitliche Vokabularien erzwungen und Abkürzungen aufgelöst. Im Anschluss werden relationsinterne sowie -übergreifende Integritätsverletzungen aufgedeckt. Hierzu zählen Schlüssel- und Fremdschlüsselbeziehungen bzw. sämtliche Formen von funktionalen Abhängigkeiten. Integritätsverletzungen lassen sich mit Hilfe relationaler Datenbanktechnologie relativ einfach aufdecken, wohingegen die Korrektur oftmals einen manuellen Eingriff erforderlich macht.

3.4 Genomic Toolkit - ein molekularbiologischer Werkzeugkasten

Neben der physischen Integration heterogener Datenquellen in ein Data Warehouse muss auch die Verarbeitung und Analyse der Genomdaten durch die Infrastruktur unterstützt werden. In einem RDBMS fehlen hierfür insbesondere Funktionen für die Verarbeitung des zentralen Genomdatentyps, der Sequenz. Wir haben daher eine Reihe der wichtigen Operationen als so genannte Benutzer definerte Funktionen (*User Defined Function*, UDF) implementiert. Hierzu zählen die Erzeugung von Statistiken über die Zusammensetzung einer Sequenz und Transformationen wie *reverse_complement* zur Bestimmung komplementärer DNA-Stränge und *translate* zum umschreiben einer DNA/RNA-Sequenz in eine Aminosäuresequenz (Protein). Des Weiteren wurden die gängigen Algorithmen zum paarwei-

sen approximativen Vergleich von Sequenzen (*Needleman-Wunsch*) [NW70] bzw. von Teilsequenzen (*Smith-Waterman*) [SW81] sowie ihre heuristischen Varianten *FASTA* [PL88] und *BLAST* [AGM+90] implementiert. Letztere erlauben als Eingabe auf der einen Seite eine einzelne Anfragesequenz und auf der anderen Seite eine ganze Datenbank von Sequenzen.

Die Implementierung der „einfachen“ Sequenzmanipulationen bereiten dem gegenüber keine Laufzeitprobleme. Beim Laufzeitvergleich der wichtigen Algorithmen ‚*Needleman-Wunsch*’ und ‚*Smith-Waterman*’ zeigte sich, dass unsere UDF-Implementierung sogar um einen konstanten Faktor schneller ist, als eine der ‚klassischen’ Implementierungen (Abb. 3) in der Programmbibliothek EMBOSS [RLB00]. Die Umsetzung der komplexen Programme *BLAST* und *FASTA* bedarf momentan noch einiger Optimierungen, um die Effizienz der UDF-Varianten in den Bereich der seit Jahren optimierten Flat-File Originale zu bringen. Der Vorteil, der bereits jetzt für diese Implementierung spricht, ist die Möglichkeit die Vergleichsoperationen gegen genau die Menge von Sequenzen durchführen zu können, die für eine wissenschaftliche Fragestellung von Bedeutung ist. Die frei verfügbaren Installationen (Web-Interfaces) bieten dem Anwender nur die Auswahl zwischen wenigen, meist sehr großen Zusammenstellungen von Sequenzdaten. Dies führt neben der Berechnung eigentlich unnötiger Ergebnisse noch dazu, dass der Anwender mit hohem Aufwand aus einer langen Liste die Ergebnisse identifizieren muss, die für ihn interessant sind.

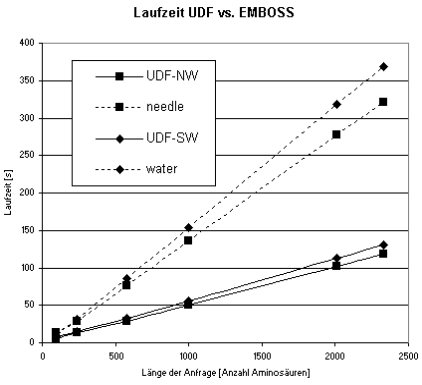

Abb. 3: UDF vs. Flat-File

4 *In silico* Experimente – Mehr Wissen über die Vielfalt der Proteine

Der praktische Einsatz der vorgestellten Plattform wird im folgenden Abschnitt anhand eines in Bearbeitung befindlichen biologischen Forschungsprojektes beschrieben. Das Projekt stellt die Bildung von Proteinvarianten anhand gegebener Geninformation nach und charakterisiert diese Varianten hinsichtlich Norm und Pathologie.

Die Darstellung ist als Arbeitsbericht über die technisch-inhaltlichen Aspekte eines realen bioinformatischen Projektablaufs angelegt. Sie verfolgt nicht das Ziel, die Fülle der biologischen Faktenlage zu kommunizieren, deren Auswertung im übrigen auch noch nicht abgeschlossen sein kann, weil sich aus der Projektbearbeitung heraus immer wieder neue Frage-Antwort-Zyklen entwickeln.

4.1 Biologischer Hintergrund und Ziele

Drei wesentliche Substanzklassen, mit denen sich die Molekularbiologie beschäftigt; sind die polymeren Kettenmolekülen DNA, RNA und Proteine. Der Träger der Erbinformation ist die DNA. Sie speichert die notwendige Information zur Synthese der Proteine, die ihrerseits die funktionell wirksamen Bestandteile lebender Organismen darstellen. Im Prozess des Informationstransfers von DNA zu Protein werden zunächst feste Abschnitte auf der DNA (Gene) in das als Mittler fungierende Kettenmolekül RNA abgeschrieben (Transkription). In vielzelligen Organismen wird die RNA anschließend prozessiert (Spleißen), wobei Zwischenabschnitte (Introns) entfernt und die verbleibenden Abschnitte (Exons) linear verknüpft werden. Im Anschluss erfolgt die Umschrift gespleißter RNA in ein Protein (Translation). Proteinvarianten entstehen in der Zelle hauptsächlich durch *Alternatives Spleißen*, weil nicht immer dieselben Abschnitte als Exons bzw. Introns behandelt werden. Daraus resultieren Spleißformen (= voneinander abweichende Exon-Konkatenierungsmuster), und in der Folge entstehen Proteinvarianten mit unterschiedlicher Bioaktivität.

Untersucht man in lebenswissenschaftliche Datenbanken die einander zugeordneten Gen- und Proteindaten, so fällt die beträchtliche Anzahl der Gene auf, denen mehr als nur ein Protein zugeordnet ist. Einige dieser Proteinvarianten besitzen eine starke pathologische Wirkung. So mehren sich Experimentalbefunde, denen zufolge in Krebszellen Proteinvarianten auftreten, die ansonsten nirgendwo nachzuweisen sind. Da diese Varianten oft nur in kleinen Kopienzahlen vorkommen, werden sie von den gängigen experimentellen Erhebungsverfahren leicht übersehen. Die eigentliche Menge an existierenden Proteinvarianten ist also als weit größer als bisher bekannt einzuschätzen. Die Kenntnis des Raums der Proteinvarianten ist jedoch essentiell für theoretische, molekularmedizinische und praktische Fragestellungen. Ziel des Projektes ist es, ein Verfahren zu etablieren, das die anzunehmende Spannbreite funktioneller Proteine unbeschadet ihrer Kopienzahl in verschiedenen Geweben ableitet. Die Darstellung einer solchen Obermenge erleichtert es, im Experiment gezielt nach Minorkomponenten und pathologischen Varianten zu fahnden. Minorkomponenten lassen sich ungleich treffsicherer detektieren, wenn sich der Experimentator beim Sondendesign nach syntaktischen Vorgaben richten kann. Für Medikamentenentwicklungen und bei der Suche nach anderen geeigneten Therapieansätzen können die Zielmoleküle präziser ermittelt werden.

4.2 Realisierung

Die Spleißformengenerierung und –bewertung werden in einem mehrstufigen Ablauf vereinigt. Mit Hilfe einer Modifikation des Viterbi-Algorithmus’ [Vit67] wird sichergestellt, dass nur Spleißformen mit der höchsten Funktionsbewertung zur Aufzeichnung kommen. Die näheren Einzelheiten hierzu sind in der Originalarbeit [HBH+04] ausgeführt. Abb. 4 stellt die Bearbeitungsschritte des Projekts in ihrer logischen Verkettung dar und ordnet sie den beschriebenen Ebenen von GENE-EYE zu. Der Projektablauf wird dabei von unten nach oben nachgezeichnet.

Er beginnt mit der lokalen Spiege-

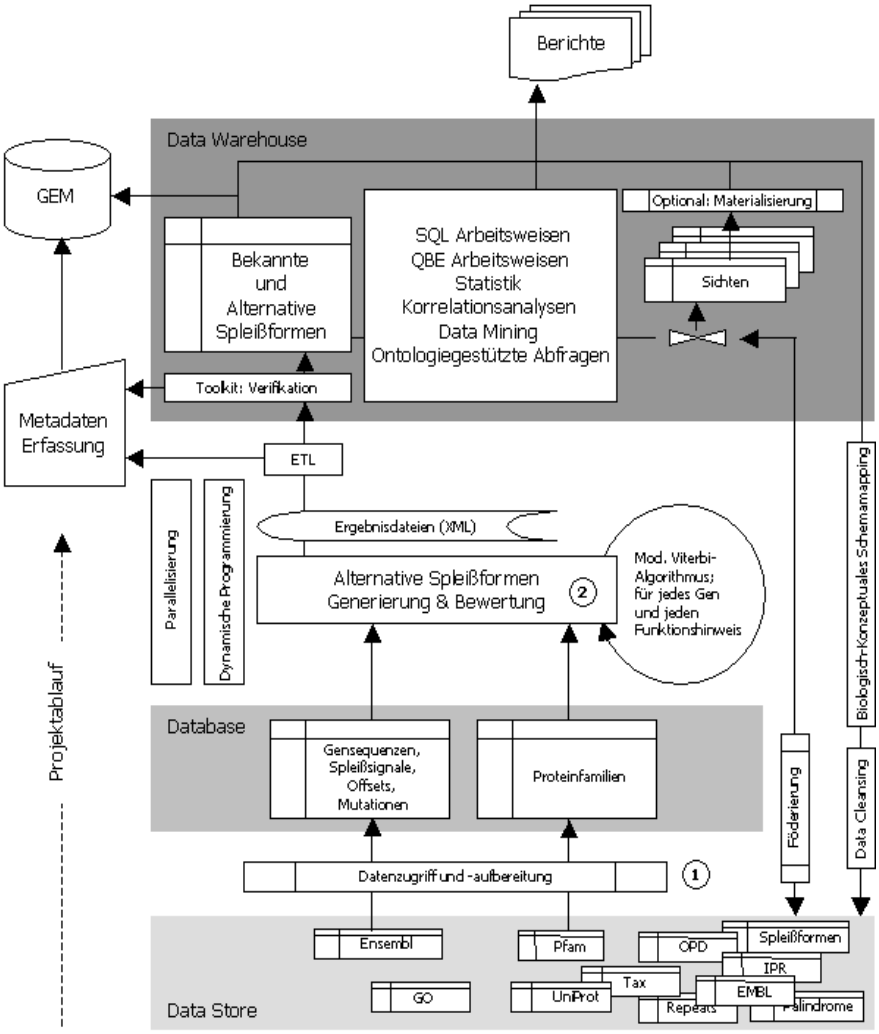

Abb. 4: Projektablauf Alternatives Spleißen

lung der projektrelevanten Datenquellen, die über den ETL-Prozess homogenisiert als relationale Daten plus Metadaten in *Data Store Layer* von GENE-EYE eingepflegt werden. Es sind dies: ENSEMBL [Bir+04], PFAM [Edd00], EMBL [Sto+01], UniProt [Apw+04], OPD [Pri+04], INTERPRO [Mul+04], TAXONOMY [Whe+00] und GO [GO04].

Es folgt Gen für Gen der Zugriff (1) auf die zu analysierenden Humansequenzen. In die Datenbank-Schicht werden Gene übernommen, deren Spleißformenbildung nicht wegen Überlappung oder zu enger Nachbarschaft durch etwaige dritte Effekte überlagert werden kann. Für jede abgegriffene Sequenz werden die Spleißsignale bestimmt.

Die Berechnung der Proteinvarianten, die den Alternativen Spleißformen entsprechen (2), erfolgt auf verteilten Ressourcen (8 Knoten RS600, Linux-Cluster) [HBH+04]. Die funktionellen Bewertungskriterien werden dem lokalen Datenbestand entnommen.

Die *in silico* Verifikation erfolgt durch Screening der Basisdaten aus EMBL, UniProt und OPD mit Routinen aus dem Genomic Toolkit unter GENE-EYE. Dabei werden Belege für die Fusionsstellen entfernter Gensequenzabschnitte gesammelt.

Die positiv evaluierten Spleißformen werden in ein Zielschema eingebunden. Der ETL-Prozess hinterlegt gleichzeitig die Metadaten im GENE-EYE Metamodell. Die Auswertung und Interpretation erfolgt in der Data Warehouse Schicht. Dabei werden Sichten auf alle relationalen Daten unter GENE-EYE erzeugt, einschließlich der in diesem und weiteren, parallel bearbeiteten Projekten neu erzeugten Daten. Die Data Warehouse Schicht unterstützt explorative Arbeitsweisen, Hypothesenbildung und – validierung, Datenbereinigungsmaßnahmen und Modellmanagement u. a. dadurch, dass Sichten nicht nur generiert, sondern auch auf alle bzw. die jeweils relevanten Komponenten der Plattform redisloziert bzw. dort dokumentiert werden.

4.3 Projektergebnisse

Die Berechnung der Spleißformen eines Gens mittlerer Länge benötigte durchschnittlich 4 Stunden auf einem Prozessor der von uns eingesetzten IBM eServer pSeries 660 Model 6M1. In parallelisierter Bearbeitung unter Ausnutzung aller acht zur Verfügung stehenden Prozessoren beansprucht das Projekt rund 17 Monate. Darin sind die Kosten für die Verifikation und Reintegration noch nicht enthalten. Diese Prozessschritte sind erst teilweise in eine automatisierte Pipeline eingestellt. Zur Verkürzung der gesamten Projektlaufzeit wird ein erheblicher Anteil der Gene bei unseren Kooperationspartnern an der Universität Jena auf einem Linux-Cluster berechnet.

Das Projekt startete bei einer Anzahl von 26.100 bekannten Spleißformen (das sind etwa 1,3 pro Gen) mit insgesamt 250.000 Struktur- und Funktionshinweisen. Jetzt, etwa zur Halbzeit des Projekts, umfasst die errechnete Obermenge der Spleißformen zwischen 20 und 800 (!) weiterer Spleißform-Kandidaten pro Gen und stützt sich auf etwa ebenso viele zusätzliche Funktionshinweise. Diese Zahl mag unerwartet hoch erscheinen, ist jedoch bei detaillierter Kontextanalyse gar nicht unrealistisch. Unabhängige Schätzungen der Anzahl unterschiedlicher Proteine anhand hochsensibler Trennverfahren liegen im Millionenbereich.

Jedes bearbeitete Gen offenbart das Potenzial, über eine Schar von Proteinen wesentlich differenziertere Funktionsfacetten auszuprägen, als seine bisherige Nominalfunktion abdeckte. Diese vorläufige Hauptaussage wird projektbegleitend mit Fachwissenschaftlern an vielen, vor allem medizinrelevanten Fallstudien evaluiert. Dabei erweist sich als Trend, dass eine Reihe an sich bekannter Beobachtungen, sei es in der Symptomatik von Erkrankungen oder in einem vermuteten

Wirkzusammenhang, in Gestalt spezieller Spleißformen molekulare Plausibilität annehmen. In diesem Sinne ergänzen sich erklärende und arbeitshypothetische Schlüsse.

Mit wachsender Anzahl bearbeiteter Gene treten mengenwertige Betrachtungen an die Seite von Einzelgenstudien. Dafür werden in semantisch begründeten Sichten relationale Daten aus dem GENE-EYE Gesamtbestand und für Explorationszwecke *ad hoc* auch neu verfügbare öffentliche oder proprietäre Daten herangezogen. In der Spleißformenvielfalt spiegeln sich auf Funktionsebene genübergreifende Umbau-, Einschub- und Verlustereignisse wider, die durch chromosomenweite syntaktische DNA-Analysen im Rahmen von Parallelprojekten (es handelt sich dabei um Duplikatanalysen und sog. Syntäniestudien, auf die hier nicht eingegangen werden kann) an den gleichen Basisdaten und mit Hilfe der Funktionen aus dem Genomic Toolkit untersucht werden. Dadurch erfahren die Einsichten in das Sozialverhalten der Moleküle auf Wirkebene ihre genetische Untermauerung. Mit Hilfe von GENE-EYE lassen sich in *projektübergreifender* Arbeitsweise viele sub- und supragenische Sequenzveränderungen der menschlichen Stammesgeschichte mit dem Spleißgeschehen korrelieren. Das hat neben dem domänenspezifischen Wissenszuwachs auch Konsequenzen hinsichtlich der Datenqualität: Das Projekt deckte eine Reihe wiederkehrender Konflikte (*contradiction pattern)* auf, die zur Bereinigung fehlannotierter Genabschnitte im Basisdatenbestand herangezogen werden können und somit Gegenstand der semantischen Datenbereinigung sind.

Während des Projektverlaufes traten häufig Situationen auf, in denen die Qualität der Daten und die Vollständigkeit der Annotationen nicht die Erwartungen erfüllten und darüber hinaus auch nicht den Anforderungen der Fragestellung genügten. Mit Hilfe der Plattform war es möglich, diese Lücken durch das Hinzuziehen weiterer Datenquellen zu schließen, ohne dass Eingriffe in die bestehende Analyse-Pipeline erforderlich wurden. Insbesondere konnten alternative Vorgehensweise einfach aufgrund ihrer Ergebnisse beurteilt werden, ohne dass dadurch ein wesentlicher Mehraufwand entstand. Darüber hinaus entfiel der sonst übliche Aufwand für die Pflege und Weiterentwicklung der Parser für die Aufbereitung der sich häufig ändernden Datenquellen. Alterntive Sichten auf Bestandteile aus verschiedenen Datenquellen konnten über die Abänderung von View-Defintionen realisiert werden. Diese konnten nach kurzer Einarbeitungszeit in ein QBE-Interface durch die Molekularbiologen selbst durchgeführt werden.

5 Zusammenfassung und Ausblick

Ausgehend von den Spezifika lebenswissenschaftlicher Forschungsprozesse haben wir wichtige Kernanforderungen an den datenbankzentrierten Zweig der Bioinformatik aufgezeigt. Die vorgestellte Plattform GENE-EYE hat das Ziel, diesen Anforderungen infrastrukturell gerecht zu werden. Hatte die Bioinformatik, historisch gesehen, in Perl ihre erste „Integrationsplattform“ [Ste96], so meistern in der jetzigen Entwicklungsphase datenbanktechnologische Kompetenzen eine weitaus höhere Komplexitätshürde. Das technische, konzeptuale und prozedurale Instrumentarium der Datenbanktechnologien stellt sich auf die Verwaltung und kognitive Durchdringung des sich ständig diversifizierenden Datenaufkommens mit ausbaufähigen Lösungen ein. Das GENE-EYE Genome Data Warehouse ist auf die schrittweise Integration von externen und proprietären Informationen auf Daten-, Konzept- und Wissensebene und auf performante Ausführung ständig wiederkehrender Operationen in Projekt-Workflows ausgelegt. Maßgeblich war für uns die Leitlinie, eine homogeniserende, offene und explorative Vorgehensweise zu unterstützen. GENE-EYE ist auf neue umfängliche und semantisch andersgeartete Datenbestände eingerichtet. So kommt z.B. durch die experimentellen Fortschritte der Massenspektrometrie eine nächste Welle von Moleküldaten in einer Art ’Gegensequenzierung der Proteinwelt’ auf die Bioinformatik zu. Gen- und Proteindaten treffen dann von zwei orthogonalen Seiten kommend aufeinander. Was im Rahmen des dargelegten Projekts noch gewebs- und zustandsunabhängigen Obermengencharakter hat, wird dann systematisch auf Zelltypen, Gewebe und Zustände aufteilbar, in denen jeweils ein bestimmter Satz von Genen aktiviert ist und ein spezifisches Spleißformenmuster liefert. Nach diesen Angaben wird schon jetzt von Seiten der Arzneimittelentwickler verlangt.

Mit jeder hinzukommenden Untersuchungsdimension müssen bestehende Integrationslösungen in der nächst höheren Anspruchs- und Integrationsebene aufgehen – oder sie erübrigen sich. Zur Reduzierung des Integrationsaufwandes sollten isoliert entstandene Data Warehouses deshalb in Zukunft ’*lernen*’, nach dem Vorbild des Grid Computings miteinander zu kommunizieren, sprich ihre Informationsbestände und Prozeduren gegenseitig dienstbar zu machen. Hierdurch kann die Datenbanktechnologie ihre große Stärke im Umgang mit statischen Stammdaten voll ausspielen, während die im Bereich alles Lebendigen so wichtigen Bewegungsdaten, Prozesse und Flüsse in anderen Informatiklösungen angesiedelt werden. Mit Blick auf ambitiöse Projekte wie Systembiologie oder virtuelle Zelle deuten sich solche Anforderungshorizonte an. Es wird auch in Zukunft spannend und für Überraschungen gesorgt bleiben.

Danksagung

Diese Arbeit wurde durch die Deutsche Forschungsgemeinschaft (DFG) unterstützt (FR1142/1-3).

Literaturverweise

[AGM+90]Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol., Vol. 215, 403-410, 1990.

[Apw+04]Apweiler, R. et al.: UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32: D115-D119, 2004.

[BBB+98]Baker, P. G., Brass, A., Bechhofer, S., Goble, C., Paton, N., Stevens, R.: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. 6th Int. Conf. on Intelligent Systems for Molecular Biology, Montreal, Canada, AAAI Press, Menlo Park, 1998.

[Bir+04]Birney, E. et al.: Ensembl 2004, Nucleic Acids Research 32, Database issue D468-D470, 2004.

[BLP00]Bernstein, P., Levy, A.Y., Pottinger; R.A.: A Vision for Management of Complex

Models. Technical Report 2000-53, Microsoft Corporation, 2000.

[Edd00]Eddy, S.R.: Profile hidden Markov models. Bioinformatics 14, 755-763, 1998.

[EUA96]Etzold, T., Ulyanov, A., Argos, P.: SRS: Information Retrieval System for Molecular Biology Data Banks. Methods in Enzymology 266: 114-128, 1996.

[GO04]Gene Ontology Consortium, The Gene Ontology (GO) database and informatics resource, Nucleic Acids Research 32, Database issue D258-D261, 2004.

[HBH+04]Hiller, M., Backofen, R., Heymann, S., Busch, A., Gläßer, T.M., Freytag, J.-C.: Efficient prediction of alternative splice forms using protein domain homology. In Silico Biology 4, 0017, 2004.

[LR03]Leser, U., Rieger, P.: Integration molekularbiologischer Daten. Datenbankspektrum Ausgabe 6, 56-66, 2003.

[MNF03]Müller, H., Naumann, F., Freytag, J.-C.: Data Quality in Genome Databases. Proceedings of the Conference on Information Quality (IQ 03), Boston, 2003.

[MRTF04]Müller, H., Rieger, P., Tham, K., Freytag, J.-C.: Dynamic information fusion for genome annotation. Informatik 2004 Workshop über Dynamische Informationsfusion, Ulm, Germany, 2004

[Mul+04]Mulder, N.J. et al.: The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Research 31, 315-318, 2003.

[NAR04]Nucleic Acid Research - Database Issue 2004. http://nar.oupjournals.org/content/vol32/suppl_1/

[NW70]Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., Vol. 48, 433-453, 1970.

[PL88]Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, Vol. 85, 2444-2448, 1988.

[Pri+04] Prince, J.T. et al.: The need for a public proteomics repository. Nature Biotechnology 22, 471 – 472, 2004.

[RLB00]Rice, P., Longden, I., Bleasby, A.: EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, Vol 16, No 6., 276-277, 2000.

[Ste96] Stein, L.: How Perl Saved the Human Genome Project. http://www.stanford.edu/class/gene211/handouts/How_Perl_HGP.html, 1996.

[Sto+01]Stoesser, G. et al.: The EMBL Nucleotide Sequence Database. Nucleic Acids Research 29, 17-21, 2001.

[SW81]Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol., Vol. 147, 195-197, 1981.

[Vit67] Viterbi, A.J: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13, 260-269, 1967.

[Whe+00]Wheeler, D.L. et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 28, No. 1, 10-14, 2000.

Erratum

Auf Seite 4 ist in Zeile 10 das Wort *Parser* durch *Scanner* zu ersetzen