Integration molekularbiologischer Daten

Ulf Leser[1], Peter Rieger[2]

[1]Humboldt-Universität zu Berlin, Wissensmanagement in der Bioinformatik,
leser@informatik.hu-berlin.de
[2]Humboldt-Universität zu Berlin, Datenbanken und Informationssysteme,
rieger@informatik.hu-berlin.de

Abstrakt: Molekularbiologische Forschung ist undenkbar geworden ohne den massiven Einsatz von Computern, sowohl zur Datenanalyse als auch zur Datenverwaltung. Bedingt durch die thematische und räumliche Fragmentierung der weltweiten Forschung in eine Vielzahl von Gruppen, Firmen und Konsortien spielt dabei die Integration von Daten eine herausragende Rolle. Zu diesem Zweck wurden sowohl Lösungen entwickelt, die auf dem integrierten Zugriff auf verteilte Datensammlungen basieren, als auch solche, die das physikalische Kopieren der Ausgangsdaten in ein integriertes System vorsehen. Der folgende Artikel gibt einen Überblick über die spezifischen Probleme der Datenintegration in der Bioinformatik, stellt die wichtigsten Projekte und Produkte in diesem Gebiet vor und weist auf neue Entwicklungen und offene Forschungsthemen hin.

1 Einleitung

Seit dem Beginn der „industriellen" Erforschung molekularbiologischer Fragestellungen durch das Human Genome Projekt gilt die Integration der dabei anfallenden Daten als eine der großen Herausforderungen der Bioinformatik [Doe93; Rob95]. Im Unterschied zu der qualitativen Arbeit vieler kleiner Labore an einzelnen Genen, Sequenzen oder Abschnitten von Chromosomen liefern die heute vorherrschenden Hochdurchsatzverfahren in kurzer Zeit Daten über komplette Genome, wie z.B. Sequenzbruchstücke von allen Genen eines bestimmten Organismus oder Expressionsmuster von Tausenden von Genen in einer Zelle. Eine Analyse dieser Daten ist durch das manuelle Recherchieren von Publikationen und relevanten Datenbanken nicht mehr möglich. Biologen müssen in ihrer Forschungstätigkeit durch Werkzeuge und Verfahren unterstützt werden, welche die Daten der durchgeführten Experimente mit Informationen aus komplementären Datenquellen anreichern und eine Einordnung und Bewertung der experimentellen Daten durch den Vergleich mit Daten anderer Gruppen ermöglicht. Beide Bereiche führen automatisch zu Problemen der Datenbankintegration.

Molekularbiologische Forschung erzeugt eine Vielzahl von Daten, die so unterschiedliche Dinge wie die Sequenz eines Gens, das Aussehen eines Individuums, den Verlauf einer Krankheit oder die räumliche Struktur eines Proteins beschreiben. Die Heterogenität der Originaldaten wird potenziert durch die unterschiedlichen Möglichkeiten, diese in verschiedensten Schemas und Formaten zu modellieren. Diese Heterogenität hat zusammen mit der weltweiten Fragmentierung molekularbiologischer Forschung und der Diversität der untersuchten Fragestellungen zu einer kontinuierlich wachsenden Menge von öffentlich verfügbaren Datenbanken geführt, deren Zahl heute auf ca. 600-1000 geschätzt wird [DBBV00].

Nach einem kurzen Überblick über die wichtigsten biologischen Konzepte werden wir in Kapitel 2 auf einige typische Bioinformatikdatenbanken eingehen und deren Ausrichtung, Modellierung und Zugriffsfunktionalität beschreiben.

Die große Bedeutung der Datenintegration in der Bioinformatik hat bereits Anfang der neunziger Jahre zur Erforschung geeigneter Methoden und der Entwicklung von Forschungsprototypen geführt (siehe z.B. [Karp94; Karp95c]). In Kapitel 3 diskutieren wir die zugrunde liegenden Konzepte. Einige Projekte und Prototypen stellen wir in Kapitel 4 vor. Diese widmen sich unterschiedlichen Aspekten der Datenintegration, wie z.B. die objekt-orientierte Multidatenbank- und Modellierungssprache OPM (siehe Abschnitt 4.2), das Flatfile-Indexierungssystem SRS (siehe Abschnitt 4.1) oder das ontologiebasierte Integrationsprojekt TAMBIS (siehe Abschnitt 4.5).

Eine übergreifende Bewertung der Entwicklung des Gebietes nehmen wir in Kapitel 5 vor. Zum einen leiten wir Rückschlüsse aus den Erfolgen und Misserfolgen der vorgestellten Systeme ab, zum andern weisen wir auf neue Entwicklungen und aktuelle Forschungsfragen im Gebiet der Integration molekularbiologischer Datenbanken hin.

2 Molekularbiologische Daten und Datenbanken

Die Aufgabe der Datenintegration besteht darin, die Vielfalt und Vielzahl der experimentellen und abgeleiteten Daten in einen konsistenten Beschreibungszusammenhang zu bringen. Der zentrale Zusammenhang, der das Rückgrat der meisten Integrationsanstrengungen darstellt, ist das Genom einer Spezies. Eng verknüpft mit Sequenzierungsprojekten, in denen die Abfolge der Erbinformation (DNA) abschnittsweise bestimmt wird, werden sogenannte Karten erstellt, welche die Organisation des Erbgutes auf den Chromosomen widerspiegeln. Auf diesen Karten werden die molekularbiologisch relevanten Merkmale eingetragen. Hauptziel ist die Erkennung aller Gene, also der Sequenzabschnitte, die für den Phänotyp eines Individuums von Bedeutung sind. Entsprechend dem zentralen Dogma der Molekularbiologie werden Gene im Prozess der Expression in RNA-Moleküle übertragen, die schließlich in Proteine übersetzt werden. Proteine sind die Funktionsträger praktisch aller Vorgänge, die in einer Zelle ablaufen.

Ergänzt werden diese Aktivitäten zur Sammlung der „Stammdaten des Lebens“ durch ein breites Spektrum experimenteller Verfahren, die versuchen, dynamische Aspekte wesentlicher Lebensvorgänge zu erfassen. So ermöglichen zum Beispiel sogenannte Microarrayexperimente (siehe Abbildung 1) die relative Aktivierung einzelner Gene in Abhängigkeit von Zelltyp, Gewebe oder Umgebungseinflüssen zu messen. Techniken aus dem Bereich der Proteomforschung bestimmen alle in einer Zelle vorhandenen Proteine oder gestatten Aussagen über die Interaktionsmuster von Proteinen, was zur Untersuchung metabolischer Vorgänge und zellulärer Signalwege führt. Neben den unmittelbar experimentellen Verfahren stellen zunehmend die Ergebnisse bioinformatischer Analysen eine wichtige Datenquelle für die Forschung dar, wie z.B. die Klassifizierung von Proteinen aufgrund von Strukturähnlichkeiten oder das Erkennen von Korrelationen zwischen der Expression von Genen und bestimmten Krankheitsbildern.

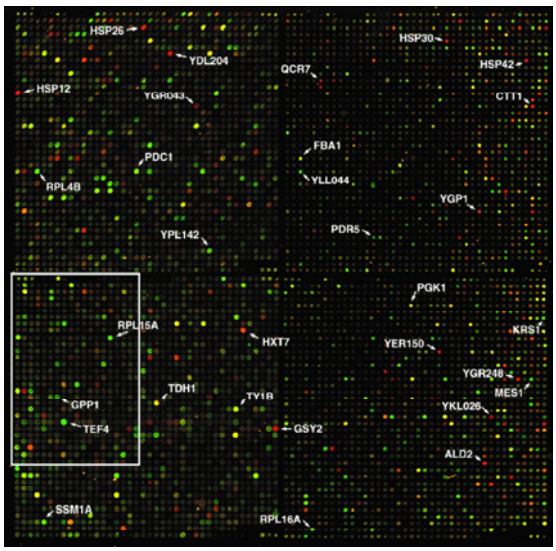

Abbildung 1: Durch Microarrayexperimente ist es möglich, zu einem bestimmten Zeitpunkt die Aktivität von mehreren tausend auf einem Chip befestigten Genen in einer Zelle gleichzeitig zu messen [DIB97]. Jedes Feld repräsentiert ein Gen, die Graustufe (im Original farbig) korreliert mit der Stärke der Expression. Ziel ist beispielsweise eine Diagnostik von Tumoren durch den Nachweis charakteristischer Expressionsmuster. Microarrayexperimente werden von vielen Gruppen durchgeführt; die Ergebnisse lassen sich aber durch Unterschiede in den experimentellen Parametern, in den Skalierungsverfahren, in den Vokabularen zur Beschreibung der untersuchten Zelle und in den benutzten Datenbankschema kaum vergleichen.

Daten aus allen beschriebenen Bereichen werden weltweit in mehreren hundert Datenbanken in einer Vielzahl von Formaten frei verfügbar für die Forschung bereitgestellt. Anstelle einzelner Referenzen sei hier auf die jährliche Januarausgabe der Zeitschrift *Nucleic Acid Research* erwiesen, die Veröffentlichungen zu molekularbiologischen Datenbanken bündelt; eine Übersicht findet man außerdem in

[BK03]. Abbildung 2 zeigt die 129 Datenquellen und deren 278 Datenbankquerbezüge, die zur Zeit über das Integrationssystem SRS (siehe Abschnitt 4.1) am EBI abgefragt werden können.

Von zentraler Bedeutung sind die Sequenzdatenbanken EMBL in Europa, Genbank in den Vereinigten Staaten und DDBJ in Japan, die in einem synchronisierten Verfahren ihre Datenbestände täglich miteinander abgleichen. Alle drei Datenbanken enthalten im Wesentlichen denselben Inhalt, werden aber in unterschiedlichen semistrukturierten Flatfileformaten bereitgestellt. Einträge reichen vom vollständigen Chromosom mit mehreren Millionen Nukleotiden bis zu experimentellen Artefakten, die aus lediglich zwei Basen bestehen. Auch die wichtigsten Proteindatensammlungen (SWISSPROT und TrEMBL) werden primär in einem Flatfileformat verteilt, können aber auch in Form von Exportdateien eines relationalen Datenbankmanagementsystems bezogen werden. Neuere Datenbanken, wie ENSEMBL (komplett annotierte Genome), ArrayExpress (Ergebnisse von Microarrayexperimenten) oder Interpro (Vorhersagen funktioneller Proteinabschnitte), setzen auf (objekt-)relationale Datenbankmanagementsysteme und verwenden häufig XML als Austauschformat.

Neben diesen, auf bestimmte Typen von Daten spezialisierten Datenbanken, gibt es auch speziesspezifische, wie MGD für Mäuse oder SDG für Bäckerhefe, chromsomenspezifische oder krankheitsspezifische Datenbanken. Eine weitere wichtige Informationsquelle sind Sammlungen von Publikationen, wie Medline, oder publikationsähnliche, mit hohem manuellem Aufwand aktuell gehaltene Datensammlungen wie OMIM, die ausführliche Informationen zu jeder bekannten menschlichen Erbkrankheit sammelt. Damit ist OMIM zum einen selber eine integrierte Datenbank, dient zum anderen aber auch als wichtige Quelle für viele weitere Datenintegrationsprojekte. Gerade diese Konstellation ist typisch für molekularbiologische Datenbanken.

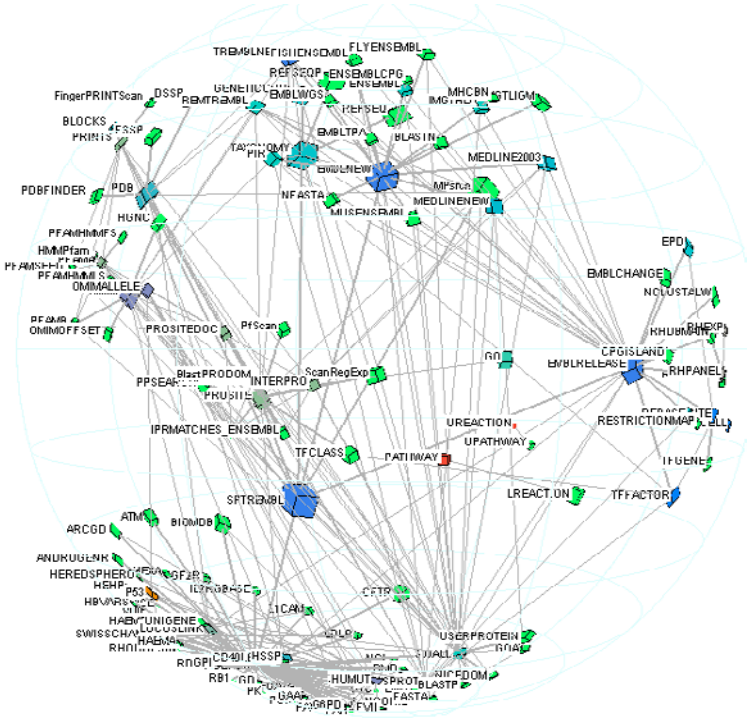

Abbildung 2: Datenquellen und Querbezüge des EBI SRS Servers (Stand 20. Februar 2003). Querbezüge repräsentieren manuell oder automatisch hergestellte, uni- oder bidirektionale Querverweise zwischen Datenbanken.

3 Anforderungen und Probleme der Integration molekularbiologischer Datenquellen

Techniken zur Bereitstellung eines einheitlichen Zugriffs auf eine Menge heterogener, autonomer und verteilter Datenbanken werden seit den 80'er Jahren unter den Begriffen „Föderierte Datenbanken" [SL90], „Multidatenbanken" [KLK91] oder „mediator-basierte Informationssysteme" [Ull97] intensiv untersucht. Daneben hat sich in der Molekularbiologie der Data Warehouse Ansatz zur Informationsintegration etabliert [CPW+01; LLRC98]. Die ersten drei Methoden werden als virtuelle, Data Warehouses dagegen als materialisierte Integrationsansätze bezeichnet.

Grundidee der föderierten Datenbanken ist die Erzeugung der Illusion einer einzigen Datenbank, die zwar nur virtuell existiert, deren Manipulation aber für den Benutzer unbemerkbar in eine Reihe semantisch äquivalenter Operationen auf den zugrunde liegenden Datenquellen übersetzt wird. Dem gegenüber stellen Multidatenbanksprachen eine einheitliche Zugriffssprache für Daten in verteilten Datenbanken bereit, ohne das Problem der semantischen Heterogenität direkt zu adressieren. Mediator-basierte Systeme können als eine Erweiterung von föderierten Systemen angesehen werden, bei der auch semistrukturierte und abfragebeschränkte Quellen, wie beispielsweise Webinterfaces, einbezogen werden. Der Fokus liegt dadurch weniger auf Schemaintegration, sondern auf Anfrageübersetzung. Der Begriff „Data Warehouse“ bezeichnet im Kontext der Bioinformatik Verfahren, die die physikalische Sammlung und Integration aller Daten in eine einheitliche Datenbank zum Ziel haben.

Die verschiedenen Ansätze sind als Reaktion auf eine Reihe von Anforderungen entstanden:

- *Transparenz*. Benutzer des integrierten Systems sollen keine Kenntnisse über Datenorganisation und -abfrage der integrierten Datenquellen benötigen.
- *Vollständigkeit*. Die Daten aller Datenquellen sollen durch das integrierte System uneingeschränkt zugreifbar sein.
- *Semantische Korrektheit und Redundanzfreiheit*. Das Schema des integrierten Systems ist semantisch korrekt und seine Elemente sind eindeutig definiert, d.h., dass Daten aus den Datenquellen korrekt in dieses eingeordnet werden. Für semantisch „gleiche“ Daten aus verschiedenen Quellen existiert ein eindeutiges globales Schemaelement.

Diese Anforderungen sind, obwohl ursprünglich aus einer betriebswirtschaftlichen Perspektive abgeleitet, auch auf Integrationsprojekte in der Bioinformatik übertragen worden. Die Zulässigkeit dieser Übertragung werden wir noch diskutieren (siehe Abschnitt 5.1).

Neben diesen „klassischen“ Anforderungen sind für die molekularbiologische Forschung die folgenden Aspekte von besonderer Bedeutung:

- *Aktualität*. Viele Fragestellungen verlangen die Verfügbarkeit möglichst aktueller Daten bzw. können mit aktuelleren Daten effektiver beantwortet werden.
  Während Ansätze zur virtuellen Integration immer höchste Aktualität gewährleisten, hängen materialisierte Verfahren von organisatorischen oder technischen Maßnahme zur Sicherstellung der Aktualität der Daten ab.
- *Performance*. Für viele Anwendungen ist die Performance der Anfragebearbeitung von höchster Bedeutung. Insbesondere die Analyse von Daten aus Hochdurchsatzexperimenten benötigt hohe Performance, um den für Forschungsprojekte typischen explorativen Umgang mit den Daten, der zu sich häufig verändernden und schwer vorhersehbaren Anfragemuster führt, zu unterstützen.
  Virtuelle Integrationsverfahren, insbesondere wenn sie auf verteilte Datenquellen zugreifen, können diese Performance in der Regel nicht gewährleisten. Analyseorientierte Integrationsprojekte verfolgen deshalb in der Regel materialisierte oder hybride Ansätze.
- *Datenintegration*. Viele Integrationsprojekte konzentrieren sich auf die Schemaebene, z.B. auf Algorithmen zur Abfrageübersetzung oder Sprachen zur Formulierung von Schemakorrespondenzen. Für Biologen wichtiger ist aber die Integration der Daten [WB96]: Erkennung und Verschmelzung von Duplikaten in verschiedenen Quellen, Erkennung von Querbezügen zwischen Objekten, Erkennen und Bereinigen von Widersprüchen, etc. Viele Integrationsprojekte verwenden den größten Teil des Aufwandes auf diese Aspekte und beschäftigen dazu häufig eine Vielzahl von Experten, die diese Aufgaben manuell erledigen [ATB+01].
  Datenintegration ist ein hochgradig anwendungsabhängiges und schwierig zu abstrahierendes Problem, das oftmals viele manuelle Schritte benötigt. Für virtuelle Ansätze ist eine hochwertige Datenintegration deshalb kaum durchführbar. Materialisierende Projekte müssen das Problem lösen, sich durch Datenintegrationsprozesse ergebende Veränderungen bei Updates der Datenquellen beizubehalten.

3.1 Technische Integrationsprobleme

Die technische Integration von Datenquellen ist eine Grundvoraussetzung für integrierte Systeme. Unter technischer Integration verstehen wir zum einen den Umgang mit unterschiedlichen Datenformaten und zum anderen die Homogenisierung unterschiedlicher Zugangssprachen. Molekularbiologi-

sche Datenbanken, die von reinen Flatfiles über proprietäre Systeme wie ACeDB [SM99] bis zu relationalen oder objektorientierten DBMS als Basis benutzen, sind in beiden Aspekten äußerst heterogen.

Viele Datenbanken verwenden für den Datenaustausch reine ASCII-Dateien, die einem bestimmten Format gehorchen. Diese Formate sind häufig komplex geschachtelt, machen intensiv von Microsyntax und sprechenden Schlüsseln Gebrauch und verfügen über eine Vielzahl von Textfeldern, deren Inhalt nicht kontrolliert oder standardisiert wird. Trotzdem hat diese Problematik in den letzten Jahren deutlich an Schärfe verloren. Heute werden Datenquellen nahezu ausnahmslos in relationalen Systemen verwaltet, aus denen die nach wie vor als primäres Austauschformat benutzten Flatfiles automatisch erzeugt werden und damit eine zwar komplizierte, aber reguläre Struktur aufweisen. Darüber hinaus werden die proprietären, über Jahre gewachsenen Formate zunehmend durch XML abgelöst – wodurch zwar das Parsing erleichtert, nicht aber das semantische Integrationsproblem gelöst wird[3]. Schließlich existieren durch Open-Source Projekte wie BioJava, BioSQL, BioPerl oder BioPython für viele wichtige Datenbanken fertige und frei verfügbare Parser.

Molekularbiologische Datenbanken sind typischerweise für den Anwender auf zwei Arten zugreifbar: Als Flatfile über FTP oder über eine Webschnittstelle. Flatfiles sind gut geeignet für eine materialisierte Integration und können durch Kopieren auf lokale Systeme und anschließendes Parsen auch in virtuellen Integrationssystemen verwendet werden (siehe z.B. SRS, Abschnitt 4.1). Datenquellen, die nur über Webschnittstellen zugänglich sind, erzwingen dagegen eine virtuelle Integrationsmethode, machen sie aber gleichzeitig sehr schwierig. Webschnittstellen sind in der Regel benutzerorientierte Suchformulare, die nur schwer mit formalen Anfragesprachen beschrieben werden können. Diesem Thema haben sich in der Datenbankforschung der letzten Jahre eine Vielzahl von Arbeiten gewidmet [VP97] und die Ergebnisse werden zunehmend auf Anwendungen in der Bioinformatik übertragen (siehe das Kapseln von Webseiten durch Funktionen in CPL, Abschnitt 4.3, oder das Wrapperkonzept von DiscoveryLink, Abschnitt 4.4).

3.2 Semantische Integrationsprobleme

In der Bioinformatik stellt die semantische Integration von Datenbanken eine besondere Herausforderung dar [FHLM98]. Wir unterscheiden zwischen semantischer Heterogenität auf der Schema- oder Definitionsebene (Was ist ein Gen?) und auf der Datenebene (sind Gen X und Y identisch?); auf letzteres gehen wir in Abschnitt 5.2 noch einmal näher ein.

Bedingt durch die Autonomie der einzelnen Datenprovider und der Tatsache, dass in dem relativ jungen Forschungsgebiet der Molekularbiologie viele Konzepte einer dauernden Revision und Re-Definition unterliegen, ist eine Vielzahl unterschiedlicher und unverträglicher Nomenklaturen, Begriffe und Definitionen entstanden [ABKS98]. Beispielsweise ist ein fundamentales Konzept wie das des Gens in unterschiedlichen Datenbanken nicht einheitlich definiert [Rob94]: so werden Sequenzabschnitte, die zur Regulation eines Genes dienen, von einigen Quellen als Bestandteil eines Gens betrachtet und von anderen nicht. Für viele Bereiche, wie etwa der Beschreibung von mit bestimmten Krankheiten verbundenen Phänotypen, existieren in der Molekularbiologie noch keine Standards (dafür um so mehr in der Medizin), für andere, wie die Beschreibung von Proteinfunktionen auf zellulärer Ebene, sind sie erst im Entstehen [GO01]. Erschwerend kommt hinzu, dass viele Datenquellen ihre Daten bzw. Schemata kaum dokumentieren.

Semantische Integration stellt für materialisierte wie virtuelle Integrationsverfahren ein schwieriges Problem dar. Die eingesetzten Verfahren reichen von virtuellen Tabellen (DiscoveryLink, Abschnitt 4.4) über objektorientierte Views (OPM, Abschnitt 4.2) bis zur Definition von komplexen Ontologien (TAMBIS, Abschnitt 4.5) und bieten dadurch sehr unterschiedliche Niveaus der semantischen Vereinheitlichung. Als Konsequenz aus diesen Schwierigkeiten wird semantische Integration in vielen Projekten entweder nur ad-hoc gelöst, d.h. unter Einsatz einer Vielzahl von PERL-Skripten, oder vom System ignoriert und dem Benutzer überlassen (SRS, Abschnitt 4.1).

3.3 Problem der exponentiell wachsenden Datenmengen

Schließlich wird häufig die schiere Datenmenge und die damit verbundene Größe der Datenbanken als Problem angeführt. Sieht man aber von Datenbanken ab, die experimentelle Rohdaten wie Tracefiles von Sequenziermaschienen oder hochaufgelöste Bilder aus Microarrayexperimenten speichern, so sind

[3] So sind z.B. EMBL, SWISSPROT und InterPro in XML verfügbar. Standardisierungsbemühungen für XML Formate existieren im Bereich der DNA Sequenzen (BSML und Agave).

molekularbiologische Datenbanken nach heutigen Maßstäben nicht besonders groß und können mit den verfügbaren Systemen problemlos verwaltet werden. Das aktuelle Genbank Release (Release 104) ist als unkomprimierte Flatfiles ca. 100 Gigabyte groß; ENSEMBL wird auf dem frei verfügbaren MySQL betrieben. Aufgrund der komplexen Algorithmen bleibt aber die Analyse dieser Datenmengen eine Herausforderung für die Datenbanktechnologie. Nicht ignoriert werden kann auch das exponentielle Wachstum der Datenmengen (siehe Abbildung 3), das darauf hinweist, dass das Datenmengenproblem schon in naher Zukunft neue Relevanz erfahren könnte.

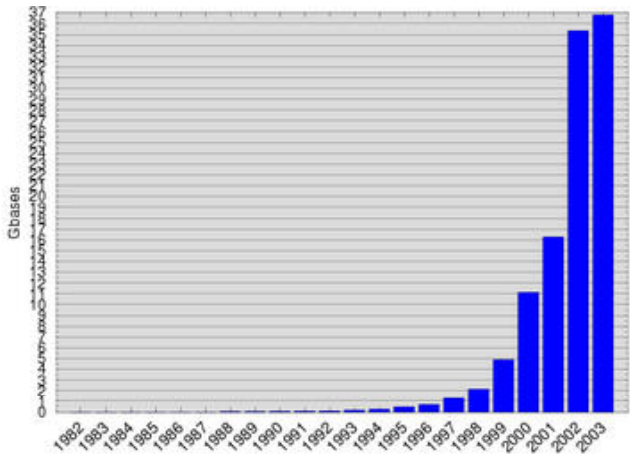

Abbildung 3: Wachstum der Sequenzdatenbank EMBL in Basenpaaren (Stand Feb. 2003).

4 Projekte und Produkte

Integrationsprojekte in der Molekularbiologie haben entweder eine generelle Integrationsmethode zum Ziel, wie z.B. die Entwicklung einer quellübergreifenden Anfragesprache, oder die Lösung eines spezifischen Integrationsproblems, wie das Zusammenführen aller verfügbaren Daten über Stoffwechselwege in einer bestimmten Spezies. Tatsächlich sind die überwiegende Mehrzahl der verfügbaren molekularbiologischen Datenbanken selber „integrierte Datenbanken“ in dem Sinne, dass sie Daten aus oder Referenzen auf andere Datenquellen beinhalten. Die dabei zur Integration verwendeten Methoden sind aber in der Regel hochspeziell, ad-hoc und eher dem Bereich der Datenanalyse als der Datenverwaltung zuzuordnen. Im Folgenden werden wir auf solche Projekte nicht eingehen.

Die erste Generation von Integrationssystemen in der Bioinformatik verfolgte den Anspruch der Integration aller verfügbaren Daten, zielte auf monolithische, eng integrierte Systeme und setzte auf Non-Standard Datenbanksysteme. Prominentes Beispiel dafür ist IGD, die „Integrated Genomics Database“ [Rit94]. Ziel des Projektes war die Integration aller verfügbaren molekularbiologischen Daten in ein zentrales Data Warehouse basierend auf der im Rahmen eines Projekts zum Mappen von *C. Elegans* entwickelten Datenbanksoftware ACEDB [SM99]. Die semantische Integration erfolgte durch eine Menge von manuell erstellten Parsern in PERL. Das Projekt scheiterte an der fehlenden Skalierbarkeit der zugrunde liegenden Technik bei gleichzeitiger enger Verzahnung mit dieser, dem hohen Aufwand zur Pflege der Parser und der fehlenden Zielgerichtetheit der Integration, die für viele Anwender keinen Nutzen erkennbar machte.

Heutige Integrationsprojekte in der Bioinformatik lassen sich grob in vier Klassen einteilen:

- Bei der ersten Klasse handelt es sich um speziell auf die Anforderungen der Bioinformatik zugeschnittene Textindexierungssysteme. Datenquellen, die als Flatfiles vorliegen, werden geparst und die einzelnen Felder indexiert. Über eine webbasierte Benutzerschnittstelle ist die Formulierung von Anfragen möglich, die auch Daten mehrerer Quellen kombinieren können. Vertreter dieser Klasse sind SRS und BioRS.
- Die zweite Klasse basiert auf Multidatenbanksprachen und konzentriert sich auf technische Aspekte der Integration, d.h. die Bereitstellung einer einheitlichen Schnittstelle zum Zugriff auf heterogene Quellsysteme ohne den Anspruch einer semantischen Integration. Beispiele hierfür sind OPM, Kleisli/CPL und DiscoveryLink.

- Die dritte Klasse konzentriert sich auf die semantische Integration von Daten. Ziel ist die Bereitstellung eines integrierten und homogenen, „globalen“ Schemas, das semantische Unterschiede in Konzepten und Klassen der Quellsysteme für den Benutzer auflöst und damit versteckt. Vertreter dieser Klasse sind TAMBIS (siehe unten) und andere ontologiebasierte Ansätze [MEK+00; Sch98].
- Die vierte Klasse zielt auf den Aufbau von integrierten Datenbanken zur Unterstützng komplexer Analyseverfahren. Vertreter dieser Klasse, wie z.B. GIMS [CPW+01], basieren auf materialisierten Ansätzen, um die notwendige hohe Performance erreichen zu können.

4.1 Sequence Retrieval System (SRS)

SRS wurde Anfang der 90ziger Jahre als Zugriffs- und Indexierungssystem für die Sequenzdatenbank EMBL entwickelt [EUA96]. Später wurde SRS zu einem System zum integrierten Zugriff auf beliebige Flatfiles weiterentwickelt, bevor es 1998 von der Firma LION übernommen wurde. Das System ist für akademische Zwecke frei verfügbar (http://www.lionbioscience.com/solutions/products/srs).

SRS verlangt eine lokale Installation von Datenbanken als Flatfile bzw. Sammlung von Flatfiles. In der speziell für SRS entwickelten Skriptsprache Ikarus werden Parser geschrieben, die die hierarchischen Strukturen der Faltfiles in eine Menge von Objekten mit mengenwertigen Attributen und einfacher Schachtelung abbilden. Einen eigentlichen Klassenbegriff kennt SRS aber nicht. Die SRS Anfragesprache ist an typische Information Retrieval Sprachen angelehnt, erweitert diese aber um einige spezifische Operatoren (z.B. das „Absteigen“ in Subobjekte) und erlaubt die Formulierung von Bedingungen wie „größer-gleich“ an numerische Attribute. Daten aus verschiedenen Quellen können unter Ausnutzung von explizit gespeicherten Referenzen miteinander verknüpft werden.

SRS ist ein weit verbreitetes System. Es gibt Wrapper für über 400 Datenquellen und weltweit über 100 Installationen. Alle Datenbanken, die am European Bioinformatics Institute (EBI) verwaltet werden, sind über SRS zugreifbar. Eine SRS Installation beinhaltet die Installation eines leistungsfähigen Webinterfaces (siehe Abbildung 4), das vielen Biologen durch seine Verwendung am EBI vertraut ist. Durch exzessive Indexierung der Daten ist SRS sehr schnell und nach Aussagen seiner Entwickler diesbezüglich relationalen Datenbanken weit überlegen [ZLAE00]. Schnittstellen zum Zugriff auf SRS existieren für gängige Programmiersprachen wie JAVA, Perl und C. Die aktuelle Version von SRS, SRS7, kann neben Flatfiles auch auf relationale Datenbanken zugreifen.

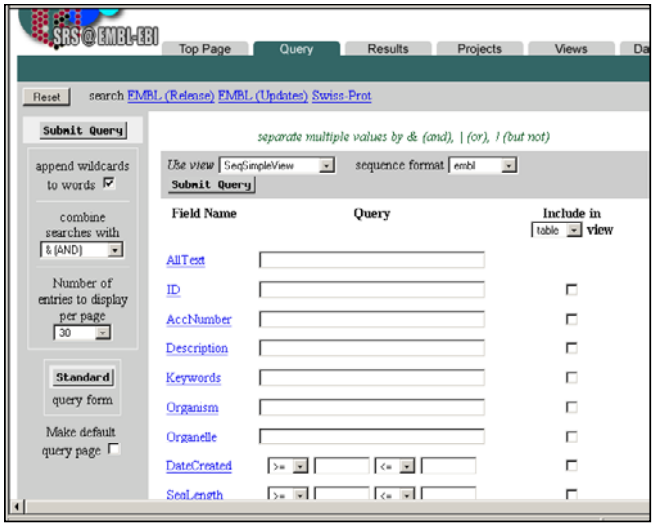

Abbildung 4: Webinterface von SRS7. Das System erlaubt u.a. die gleichzeitige Durchsuchung mehrer Datenbanken, die Formulierung von numerischen Prädikaten, die Definition von individuellen „Views“ und das Speichern von Anfragen für spätere Sessions.

Ein Nachteil von SRS ist der hohe Proprietätsgrad. Große Teile des Systems sind in der außerhalb der SRS-Gemeinde unbekannten Sprache Ikarus geschrieben. Die Methoden zur Übersetzung und Optimierung von Anfragen sind nicht dokumentiert. Das Hinzufügen von neuen Datenbanken in ein laufendes SRS System ist nicht trivial und erfordert das Ändern einer Reihe von nicht oder nur schlecht dokumentierten Konfigurationsdateien. Die SRS Anfragesprache ist bei komplexen Anfragen gewöhnungsbedürftig. Das System erschließt sich damit einem Informatiker nur schwer.

4.2 Object Protocol Model (OPM)

OPM als Datenintegrationswerkzeug ist die Weiterentwicklung eines Werkzeugkastens zur objektorientierten Modellierung von Datenbanken [CM95]. Das „Object Protocol Model“ ist ein objektorientiertes Datenmodell mit Vererbung und Assoziationen, berechneten und mengenwertigen Attributen und speziellen Beziehungstypen zur Modellierung von Protokollen (siehe Abbildung 5). OPM Modelle sowie Anfragen in der Sprache OPM-QL werden in relationale Schemata bzw. Anfragen übersetzt. Ausgehend von einem OPM Schema können automatisch Web-Sites mit Formularen für die Abfrage und Ausgabe von Objekten erzeugt werden. Zusätzlich zu einem graphisches Modellierungswerkzeug ist ein „Retrofitting Tool“ zum nachträglichen Erstellen von OPM Modellen für bestehende relationale Schemata verfügbar.

OPM wurde zur Modellierung verschiedener Datenbanken im Umfeld des Human Genome Project eingesetzt, wie z.B. der Genome Database (GDB) oder der Genome Sequence Database (GSDB), und erlangte dadurch hohe Popularität. Die Software wurde 1997 von der Firma GeneLogic Inc. übernommen und kommerzialisiert (http://www.genelogic.com/opm.htm).

Die Vorteile von OPM liegen in dem semantisch reichhaltigen Datenmodell. Dies ermöglicht eine intuitive Modellierung von Daten zur Chromosomenkartierung, eine Aufgabe, die in der frühen Phase des Human Genome Project eine dominierende Rolle spielte. Eine Erweiterung von OPM-QL zur Multidatenbanksprache OPM*QL ermöglicht außerdem die Formulierung von Anfragen an mehrere OPM Schemata. Voraussetzung dafür ist aber die Definition eines OPM Schemas für jede beteiligte Quelle in Form einer Reihe relationaler Views.

```
OBJECT CLASS     Chromosome isa* Genomic Segment
                 ATTRIBUTE cellularCompartement: [1,1] CompartmentDict
OBJECT CLASS     Amplimer isa* Genomic Segment
                 ATTRIBUTE isExpressed: [1,1] YesNoUnknown_UnkDict
                 ATTRIBUTE sequence: list-of [0,] VARCHAR(255)

SELECT Name = GSDB:Gene.name, Reason = HGD:Gene.reason,
       Annotation = HGD:Gene.annotation
FROM   GSDB:Gene, HGD:Gene
WHERE  HGD:Gene.accessionID = GSDB:Gene.gdb_xref
```

Abbildung 5: Definition eines OPM Schemas. Klassen können von anderen Klassen (isa*) erben und mengenwertige und komplexe Attribute besitzen. Darunter ist eine OPM*QL Anfrage zur Verknüpfung von GDB und GSDB abgebildet. Voraussetzung zur Benutzung von Datenbanken in OPM*QL Anfragen ist die Definition eines OPM Views auf die Datenbank.

OPM ist nicht als Integrationssystem konzipiert worden, auch wenn es in vielen Publikationen als solches bezeichnet wird. Werkzeuge z.B. zur Integration von OPM Schemata oder zum Einbinden von Datenintegrationsfunktionen fehlen, und die Optimierung verteilter Anfragen ist nur prototypisch erfolgt. Daraus ergibt sich unter anderem der größte Nachteil von OPM, die mangelnde Performance. Neben dem ungenügenden Optimierer ist dies insbesondere durch den starren Schemaübersetzungsalgorithmus verursacht, der die Ausrichtung von relationalen Schemata an spezielle Anforderungen von Projekten nicht ermöglicht. Mit der wachsenden Bedeutung der Analyse großer Datenmengen erweist sich dies zunehmend als Hindernis.

4.3 Kleisli/K2 und Collection Programming Language (CPL)

Kleisli wurde als Datentransformationssystem an der University of Pennsylvania entwickelt [BDHO95]. Es basiert auf dem Nested Relational Calculus und der Anfragesprache CPL (Collection Programming Language). Das zugrunde liegende Datenmodell erlaubt geschachtelte Strukturen und mengenwertige Attribute und kann dadurch viele der typischen Flatfileformate in der Bioinformatik auf natürliche Weise repräsentieren.

Kleisli erlaubte als erstes System mit einer Anfrage und ohne lokale Installation der Originalquellen die Beantwortung eines Teils der „12 unanswerable queries“, die in einem Report des Department of Energy 1993 als Herausforderung an die Bioinformatik formuliert wurden [Doe93]. Es wurden Wrapper für die wichtigsten molekularbiologischen Datenbanken entwickelt und eine Reihe von Pharmakonzernen lizenzierten das System [LNW03]. Das sich auf semantische Aspekte konzentrierende Projekt TAMBIS benutzte Kleisli als technische Infrastruktur zum Zugriff auf Datenquellen

(siehe unten). Auch Kleisli wurde kommerzialisiert und wird heute von der Firma GeneticXChange Inc. vertrieben (http://www.geneticxchange.com/).

Kleisli ist durch eine lange Entwicklung zu einem technisch ausgereiften System geworden. Semantische Integration wird nicht originär adressiert; CPL unterstützt aber die Definition von Funktionen, die relationalen Views entsprechen und bis zu einem gewissen Grad Integrationsaufgaben vor einem Benutzer verdecken können.

Häufig kritisiert wurde Kleisli für die gewöhnungsbedürftige Syntax und Non-standard Semantik der Sprache CPL, die eine Benutzung durch Endbenutzer praktisch unmöglich macht und auch für Informatiker einen Einlernprozeß erfordert. Im Nachfolgesystem K2 wurde CPL deshalb durch die an SQL angelehnte Sprache sSQL ersetzt.

4.4 DiscoveryLink

DiscoveryLink entstand aus dem IBM Forschungsprojekt Garlic [HKWY97], das ursprünglich zum Zugriff auf verteilte und heterogene Multimediadaten entwickelt wurde. Die Technologie von Garlic besteht im Kern aus einer Integration externen Quellen in die Datenbank DB2. Die Datenquellen werden in DB2 als virtuelle Tabelle registriert und können danach wie „normale“ Tabellen benutzt und in beliebigen SQL-Anfragen verwendet werden (siehe Abbildung 6). Der tatsächliche Zugriff erfolgt über speziell zu entwickelnde Wrapper. Verwendet eine Anfrage gegen eine solche virtuelle Tabelle Operatoren, die ein Wrapper bzw. die zugrunde liegende Datenquelle nicht bereitstellt, so wird dies von DB2 automatisch ausgeglichen.

```
CREATE WRAPPER ChemWrapper LIBRARY 'libchem.a'
CREATE SERVER ChemHits
  WRAPPER    ChemWrapper
             OPTIONS (NODE 'X.Y.com', PORT '2003', VERSION 'Z.ZZ')
CREATE NICKNAME PROTEINS {
             ID           VARCHAR(30) NOT NULL,
             NAME         VARCHAR(60),
             PROT_FAMILY  VARCHAR(256),
             REL_DISEASE  VARCHAR(256)
  } SERVER ChemHits
```

Abbildung 6: Definition von Datenquellen in DiscoveryLink. Datenquellen werden durch Server adressiert, die technologisch auf einem vorab definierten Wrapper basieren. Das Mapping einer Tabelle eines Servers auf eine virtuelle DB2 Tabelle erfolgt über Nicknames. Diese können dann in Anfragen wie originäre DB2 Tabellen verwendet werden.

DiscoveryLink wird von IBM als Integrationslösung für die Lebenswissenschaften vermarktet (http://www-3.ibm.com/solutions/lifesciences). Semantische Datenintegration wird nicht adressiert, ist aber, wie bei vielen anderen Systemen, durch die Definition von Views in begrenztem Umfang möglich. Der Schwerpunkt des Projektes liegt auf Performancemaximierung, die durch ein spezielles Anfrageplanungsverfahren zusammen mit einem detaillierten Kostenmodell erreicht wird. In [HSK+01] wird gezeigt, dass die DiscoveryLink Middleware keinen nennenswerten Overhead für rein relationale Integrationssysteme mit sich bringt; Messungen für Systeme mit integrierten Flatfiles werden, vermutlich mangels sinnvoller Vergleichsdaten, nicht angegeben.

Die Technologie von DiscoveryLink baut auf den etablierten Methoden relationaler Datenbanken auf und ist damit für Informatiker leicht zugreifbar – nicht aber für Biologen. Ein Nachteil von DiscoveryLink ist das Fehlen von fertigen Wrappern für auch nur die wichtigsten molekularbiologischen Datenquellen. Diese sollen mit der nächsten Version verfügbar sein, zusammen mit einem „Wrapper Development Kit“ für die Entwicklung eigener Wrapper.

4.5 Transparent Access to Molecular Biology Databases (TAMBIS)

Im Unterschied zu den bisher vorgestellten Systemen, die sich ganz auf technische Aspekte konzentrieren, widmet sich das an der University of Manchester entwickelte TAMBIS [BBB+98] ausschließlich den semantischen Problemen der Integration heterogener Datenquellen in der Molekularbiologie. Fragen des Datenzugriffs, der Anfragemöglichkeiten oder des Datenformats werden nicht adressiert, sondern durch die Benutzung von Kleisli (siehe oben) als technische Infrastruktur umgangen.

Kern von TAMBIS ist eine 1800 Begriffe umfassende Ontologie molekularbiologischer Begriffe. Diese wurde über einen Zeitraum von zwei Jahren in der Beschreibungslogik GRAIL entwickelt und später in das weiter verbreitete und semantisch ausdrucksstärkere OIL portiert [SGHB02]. Das Onto-

logiedesign verfolgte eine gemischt Top-Down / Bottum-Up Strategie: Zunächst wurde eine Kernontologie Top-Down modelliert, in der die wichtigsten Begriffe der Anwendungsdomäne in einem multihierarchischen Konzeptbaum angeordnet sind. Darauf aufbauend wurden konkrete Anfragen an Quellsysteme in den Begriffen der Kernontologie beschrieben und durch den Subsumptionsmechanismus der Beschreibungslogik automatisch in die Kernontologie eingeordnet. Die daraus resultierenden Konzepte der Ontologie entsprechen CPL-Funktionen, die die Auswertung von Anfragen und die Transformation der Ergebnisse kapseln.

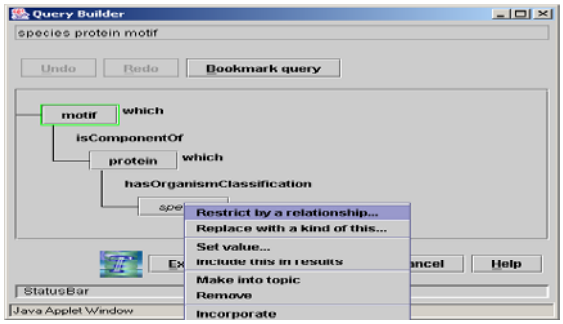

Abbildung 7: Webbasierte Benutzerschnittstelle von TAMBIS. Eine Anfrage wird interaktiv „zusammengeklickt“, wobei TAMBIS die in der Ontologie hinterlegten Definitionen der Konzepte zur Vorschlaggenerierung ausnutzt. Die fertige Anfrage wird durch Subsumption in die Ontologie eingeordnet und relevante Datenquellen abgeleitet. Diese werden über CPL-Funktionen angesprochen.

Anfragen in TAMBIS werden ebenfalls als Konzepte beschrieben, die in einem ersten Schritt durch Subsumption in die Ontologie eingeordnet werden. In einem zweiten Schritt werden dann alle semantisch subsumierten Konzepte, die an konkrete Quellanfragen gebunden sind, ausgewertet. Zur Optimierung wird ein einfacher greedy Algorithmus verwendet. Um Benutzern den Zugang zum System zu erleichtern, wurde eine graphische, interaktive Anfrageschnittstelle entwickelt (siehe Abbildung 7).

TAMBIS ist zweifellos das umfangreichste Projekt zur Adressierung semantischer Heterogenität in der Bioinformatik. Die entstandene Ontologie vereinigt eine Vielzahl von Konzepten in einer homogenen Begriffswelt mit klarer und computerlesbarer Abgrenzung der einzelnen Klassen untereinander.

Das Projekt TAMBIS ist ausgelaufen und wird nicht fortgesetzt. Erfahrungen aus der Perspektive tatsächlicher Benutzer sind bisher nicht veröffentlicht (wohl aber aus verwandten Projekt; siehe [RZS+99]). Aufgrund der quälenden Langsamkeit und der notwendigen Einarbeitung in eine komplexe Begriffswelt, die den Zugriff auf den Biologen bestens bekannte Datenbanken eher verschleiert als erleichtert, muss die Benutzbarkeit des Prototypen bezweifelt werden.

4.6 Weitere Projekte

Neben den genannten Projekten bzw. Produkten gibt es eine Reihe weiterer Systeme zur Integration biomedizinischer Datenbanken. Das BioRS-System wurde ursprünglich am Max-Planck-Institut für Proteinstrukturen entwickelt und wird heute von der Firma BioMax vertrieben (http://www.biomax.de). BioRS ähnelt SRS im Aufbau; ursprünglich wurde es als Flatfile-Indexierungs und –retrievalsystem entwickelt, kann heute aber ebenso auf relationale Datenbanken zugreifen. Die Software besteht aus mehreren Modulen, die über CORBA miteinander kommunizieren. Die Datenbanken des US-amerikanischen Zentrums zur Verwaltung molekularbiologischer Datenbanken, dem NCBI (siehe http://www.ncbi.nlm.nih.gov/Entrez/), sind durch das System Entrez [SEOK96] zugreifbar. Entrez ist das primäre Zugriffssystem für GenBank und PubMed, einer öffentlich zugänglichen Variante von Medline. Ein Integrationssystem, dass nur auf der Speicherung und Berechnung von Crossreferenzen zwischen Objekten in unterschiedlichen Datenbanken basiert, ist DBGET [FGM+98]. Das „Genome Information Management System“ (GIMS) ist ein Data Warehouse zur Unterstützung der Analyse von Microarrayexperimenten und integriert dazu eine Reihe von externen Datenbanken [CPW+01]. Diese werden in quellspezifische relationale Schema geparst. Das System nimmt keine semantische Integration vor, erlaubt aber natürlich Anfragen, die Daten aus den verschiedenen Schema adressieren.

Eng verwandt mit der Datenintegration sind Projekte zur Applikationsintegration. Die Life Science Research Initiative (http://www.omg.org/lsr) der OMG standardisiert in über 15 Workgroups Schnittstellen für den Zugriff aus Daten aus unterschiedlichen Bereichen der Lebenswissenschaften, wie Se-

quenz oder Kartierungsdaten, bibliographische Informationen oder chemische Verbindungen. Diese von der OMG koordinierten Entwicklungen haben aber in den letzten Jahren deutlich an Bedeutung verloren gegenüber einer Reihe von OpenSource-Projekten wie BioSQL, BioPerl oder EMBOSS, die ebenfalls Tools und Methoden zur Daten- und Applikationsintegration implementieren.

5 Entwicklungen und Forschungsthemen

Eine Betrachtung der Datenintegrationsprojekte in der Bioinformatik über die letzte Dekade erlaubt eine Reihe von Rückschlüssen, auf die wir im Folgenden eingehen. Nach dieser retrospektiven Betrachtung geben wir einen Ausblick auf aktuelle Themen und Forschungsgebiete.

5.1 Retrospektive Betrachtung der Entwicklung

Es lassen sich zwei große Entwicklungslinien erkennen.

Zum einen erfolgte eine Kommerzialisierung nahezu aller entwickelter Systeme, die ja ursprünglich als Forschungsprojekte begannen. Mit Ausnahme von TAMBIS wurden alle vorgestellten Systeme durch Biotechnologie- oder Bioinformatikfirmen akquiriert und werden heute als kostenpflichtige Produkte inklusive Installation, Beratung und Schulung vermarktet. Andererseits ist es interessant, dass – mit Ausnahme von IBM und DiscoveryLink – kein etablierter Softwarehersteller versuchte, ein Produkt zu diesem Thema selbst zu entwickeln. Bioinformatikfirmen, deren einziges Geschäftsfeld die Datenintegration ist, konnten sich nicht am Mark behaupten, auch wenn deren Entwicklungen kurzzeitig weltweit als wegweisend eingeschätzt wurden (Double Twist). Ungeachtet dessen versuchen Anbieter wie Oracle oder Compaq/Hewlett Packard durch die Gründung eigener „Life Science“ Abteilungen in den zukunftsträchtigen Markt des Managements biomedizinischer Daten einzutreten.

Eine zweite Entwicklungslinie ist die relative Erfolglosigkeit der Ansätze zur semantischen Integration auf Schemaebene. Obwohl semantische Heterogenität der Schema und Modelle kontinuierlich als dringendes Problem postuliert und die Forderung nach transparentem und homogenen Zugriff aufgestellt wird [Hen03], muss bezweifelt werden, ob dies für die Molekularbiologie tatsächlich zutrifft:

- *Transparenz* verdeckt beispielsweise die Tatsache, dass molekularbiologische Datenbanken aus der Sicht eines Biologen nicht aus einer rein technischen Perspektive betrachtet werden dürfen. Statt dessen ist die Frage nach der Eignung einer Quelle für eine konkrete Fragestellung in erster Linie abhängig von nicht-technischen Faktoren, wir die Reputation der Organisation, die sie aufgebaut hat und betreut, oder die Methoden und experimentellen Protokolle, mit denen die Daten erzeugt wurden – und damit vor allem von der erwarteten Qualität der Daten bzgl. Aspekten wie Fehlerfreiheit, Vollständigkeit und Aktualität [NLF99]. Transparenz in Bezug auf Datenquellen ist deshalb oft ein Hindernis für die Benutzung eines Integrationssystems.
- Die Forderung nach *semantischer Korrektheit und Redundanzfreiheit* ist nicht realistisch, da die Semantik vieler Begriffe in der Molekularbiologie nicht eindeutig definiert ist und Datenquellen die Semantik ihrer Schemaelemente oft bewusst unscharf halten, um einer zu starken Einengung der Forschung zu entgehen. Diese Unschärfe sollte nicht als Unterlassung der Datenbankentwickler betrachtet werden, sondern als Folge eines sich in ständiger Bewegung befindlichen Anwendungsgebietes. Die semantische Integration bzgl. eines Konzeptes wie „Gen“ ist für Biologen deshalb oftmals suspekt, da für die Beurteilung von Objekten einer solchen Klasse viele externe Faktoren beachtet werden müssen – ein Gen, dessen Existenz durch Populationsstudien mit anschließendem direkten Nachweis durch Clonierung nachgewiesen wurde hat einen anderen Stellenwert als ein Gen, das durch Clustering von fehleranfälligen EST-Sequenzen vorhergesagt wurde.

Das Problem der semantischen Heterogenität ist selbstverständlich nicht verschwunden. Die Auflösung der Konflikte erfolgt beim derzeitigen Wissensstand aber meist projektspezifisch. Herausforderung an die Datenbankforschung ist die Bereitstellung flexibler Mechanismen zur „Integration on Demand“. Globale Schemata werden von den Benutzern dagegen wenig akzeptiert.

5.2 Aktuelle Forschungsgebiete

Neben den aufgezeigten Entwicklungen hat eine ganze Reihe neuer Themen an Bedeutung gewonnen.

Data Warehousing und Propagierung von Updates
Mit der Verfügbarkeit immer größerer und umfassenderer Datensammlungen ist eine Interessenverschiebung vom reinen Datenmanagement zur Datenanalyse zu beobachten. Diese Datenanalyse erfolgt oftmals auf einer integrierten Datenbasis, da die Kombination von Daten verschiedener experimenteller Techniken neue Aufschlüsse verspricht. Beispielsweise werden Expressionsdaten zunehmend in Kombination mit textuellen Annotationen und Veröffentlichungen betrachtet. Dies, zusammen mit der drastischen Kostensenkung der Sekundärspeicher, bringt eine Hinwendung zu materialisierten Integrationsstrategien bzw. Data Warehousing, da nur so ausreichende Performance für komplexe Analysen erreicht werden kann. Erkauft wird die erhöhte Performance mit Problemen der Aktualisierung von Datenquellen; das Problem wird verschärft, wenn Datenquellen vor der Verwendung zur Datenanalyse nicht nur 1:1 gespiegelt, sondern auch einem aufwändigen Integrations- und Data Cleansing-Prozess unterzogen werden. Das Propagieren von Änderungen in Quellen durch komplexe Aufbereitungsprozesse ist eine offene und wichtige Forschungsfrage.

Datenqualität
Die Qualität der Daten ist für eine experimentelle Wissenschaft wie die Biologie ein intrinsisch wichtiges Thema. Viele Projekte haben die Verbesserung der Qualität der Ergebnisse zum Ziel, da nahezu alle experimentellen Techniken in der molekularbiologischen Forschung Daten mit einer gewissen Unschärfe erzeugen: Clonierung beinhaltet den Umgang mit mehr oder weniger stabilen Wirtszellen, Hybridisierung hängt von den zu verbindenden, aber a-priori unbekannten, Sequenzen ab, Sequenzierungen ergeben bei weitem nicht immer ein klares Bild von den jeweiligen Basen, etc. Darüber hinaus ist das Matching von biologischen Objekten wie Clone – zur Crossvalidierung von Ergebnissen – ein schwieriges Problem [LLRC98]; selbst für essentielle Objekte wie Gene gibt es bis heute keinen allgemein akzeptierten Benennungsstandard. Die systematische und computerlesbare Erfassung der Datenqualität in Sekundärdatenbanken wie SWISSPROT und EMBL ist lange Zeit unterblieben, was dazu geführt hat, dass man von einem hohen Anteil falscher Sequenzannotationen in den öffentlichen Datenbanken ausgeht [Bre99]. Dieses Problem ist bekannt, und Standardisierungsverfahren speziell für die Annotation von Genen und Proteinen finden zunehmend Bedeutung (siehe Abschnitt über Ontologien weiter unten). Eine systematische Qualitätssicherung molekularbiologischer Datenbanken, insbesondere auch der manuell annotierten und korrigierten, ist notwendig, aber noch nicht etabliert. Ein weiteres wichtiges Thema ist die transparente Berücksichtigung unterschiedlicher Datenqualität in Anfrage- und Analyseprozessen.

Ontologien und Datenintegration
Die Verwendung von Ontologien zur Datenbankintegration in der Bioinformatik wird seit mehreren Jahren diskutiert. Während Ansätze wie [Sch98] oder [BBB+98] Ontologien zum Mapping von Schemaelementen und Anfragen benutzen und sich bisher nicht durchsetzen konnten, haben Ontologien große Popularität als Standards für Annotation gewonnen. Insbesondere die GeneOntology, die von einem internationalen Konsortium gepflegt und weiterentwickelt wird [GO01], hat sich innerhalb weniger Jahre als Standard zur funktionalen Beschreibung von Genen und Genprodukten etabliert. Viele Datenbanken ersetzen ihre bisherigen Freitextannotationen oder Controlled Vocabularies durch Annotationen basierend auf der GeneOntology (siehe z.B. http://www.ebi.ac.uk/GOA), da durch die Verwendung eines gemeinsamen Vokabulars eine Analyse unterschiedlicher Annotationen in integrierten Datenbanken wesentlich erleichtert wird. Interessante Forschungsfragen in diesem Umfeld sind automatische Verfahren zur Annotation von (eventuell bereits annotierten) Daten mit Hilfe von Ontologien, Fragestellungen zum verteilten Management und der Evolution von Ontologien sowie Verfahren zur integrierten Analyse von experimentellen und textuellen Daten.

Benutzerschnittstellen: Integrierte Sichten versus Integrationsprotokolle
Biologen denken oftmals in Protokollen. Protokolle definieren den Ablauf von Handlungsschritten in Experimenten und werden publiziert, ausgetauscht und ständig verbessert. In ähnlicher Weise gehen viele Biologen Fragen der Datenanalyse und der Datenintegration an, nämlich aus der Sichtweise einer Reihe aufeinander abfolgender und voneinander abhängiger Verarbeitungsschritte. Die Analyse einer neu gefundenen Sequenz wird in sogenannten Annotationsworkflows durch den sequentiellen Aufruf einer Reihe von Programmen (z.B. zur Suche nach ähnlichen Sequenzen) oder das Absetzen von Anfragen an Datenbanken (z.B. zur Literaturrecherche oder zur Suche nach funktionalen Bestandteilen) beschrieben [BFS+98]. Im Unterschied dazu verlangt ein datenbankbasiertes, integriertes System wie

DiscoveryLink die Formulierung von Anfragen, um Daten miteinander zu verknüpfen bzw. zu integrieren. Hier öffnet sich ein interessantes Forschungsgebiet im Bereich der Benutzerschnittstellen für integrierte Systeme, die eine eher prozessorientierte Sichtweise ermöglichen und damit auf höhere Akzeptanz bei Biologen hoffen können.

6 Zusammenfassung

Die Integration von molekularbiologischen Daten aus heterogenen, autonomen und verteilten Datenbanken ist seit Jahren ein wichtiges Thema und wird es auch auf Jahre hinaus bleiben. Die Integration von Daten in diesem Bereich zielt auf verbesserte Datenanalysen (durch mehr und besser validierte Daten), kürzere Entwicklungszeiten (durch Vermeidung von überflüssiger Arbeit) und neue Erkenntnisse (durch das Entdecken von Korrelationen zwischen vorher unverbundenen Datensätzen). Erschwert wird die Integration durch die große Heterogenität zwischen den verschiedenen Datenquellen, sowohl auf semantischer als auch auf technischer Ebene.

Eine Reihe von Systemen wurde zur Lösung dieser Probleme entwickelt und hier kurz vorgestellt. Auffallend daran ist, dass sich die am weitesten verbreiteten Systeme auf technische Aspekte konzentrieren. Nach unserer Einschätzung liegt dies nicht am Fehlen ausreichend detaillierter Begriffssysteme, sondern an der relativen Unreife des Forschungsgebietes sowie seiner Komplexität, die ausreichend umfassende „globale“ Definitionsgerüste unbenutzbar machen.

Literaturangaben

[ABKS98] Altman, R. B., Benton, D., Karp, P. D. and Schulze-Kremer, S., Eds. (1998). "Workshop on Semantic Foundations for Molecular Biology". Montreal, Canada, ISMB'98 Workshop.

[ATB+01] Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., et al. (2001). "The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites." Nucleic Acids Research 29(1): 37-40.

[BFS+98] Bailey, C., Fischer, S., Schug, J., Crabtree, J., Gibson, M. and Overton, G. C. (1998). "GAIA: Framework Annotation of Genomic Sequences." Genome Research 8: 234-250.

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

[Bre99] Brenner, S. E. (1999). "Errors in Genome Annotation." Trends in Genetics 15(4): 132-133.

[BK03] Bry, F. and Kröger, P. (2003). "A Molecular Biology Database Digest." International Journal On Distributed and Parallel Databases 13(1): 7-42.

[BDHO95] Buneman, P., Davidson, S., Hart, K. and Overton, G. C. (1995). "A Data Transformation System for Biological Data Sources". 21st Conference on Very Large Data Bases, Zuerich, Switzerland.

[CM95] Chen, I. A. and Markowitz, V. M. (1995). "An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools." Information Systems 20(5): 393-418.

[CPW+01] Cornell, M., Paton, N. W., Shengli, W., Goble, C. A., Miller, C. J., Kirby, P., Eilbeck, K., Brass, A., Hayes, A. and Oliver, S. G. (2001). "GIMS - A Data Warehouse for Storage and Analysis of Genome Sequence and Function Data". 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda.

[Doe93] Department of Energy (1993). "Report on the Invitational DOE Workshop on Genome Informatics". Baltimore, Maryland, USA Department of Energy.

[DIB97] DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science 278(5338): 680-6.

[DBBV00] Discala, C., Benigni, X., Barillot, E. and Vaysseix, G. (2000). "DBcat: a catalog of 500 biological databases." Nucleic Acids Research 28(1): 8-9.

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

[FHLM98] Frishman, D., Heumann, K., Lesk, A. and Mewes, H.-W. (1998). "Comprehensive, comprehensible, distributed and intelligent databases: current status." Bioinformatics 14(7): 551-561.

[FGM+98] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Y., A. and Kanehisa, M. (1998). "DBGET/LinkDB: an Integrated Database Retrieval System". 3rd Pacific Symposium on Biocomputing.

[GO01] GeneOntology Consortium, T. (2001). "Creating the gene ontology resource: design and implementation." Genome Research 11(8): 1425-33.

[HKWY97] Haas, L. M., Kossmann, D., Wimmers, E. L. and Yang, J. (1997). "Optimizing Queries across Diverse Data Sources". 23rd Conference on Very Large Database Systems, Athens, Greece.

[HSK+01] Haas, L. M., Schwarz, P. M., Kodali, P., Kotlar, E., Rice, J. and Swope, W. C. (2001). "DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources." IBM Systems Journal 40(2): 489-511.

[Hen03] Hendler, J. (2003). "Science and the Semantic Web." Science 299: 520-521.

[Karp94] Karp, P. D. (1994). "Report of the Workshop on Interconnection of Molecular Biology Databases", SRI International Artificial Intelligence Center, Stanford, California.

[Karp95c] Karp, P. D., Ed. (1995). "2nd Meeting on Interconnection of Molecular Biology Databases". Cambridge, UK, Electronic Proceedings, available at http://www.ai.sri.com/people/pkarp/mimdb.html.

[KLK91] Krishnamurthy, R., Litwin, W. and Kent, W. (1991). "Language Features for Interoperability of Databases with Schematic Discrepancies". ACM SIGMOD 1991, Denver, Colorado.

[LLRC98] Leser, U., Lehrach, H. and Roest Crollius, H. (1998). "Issues in Developing Integrated Genomic Databases and Application to the Human X Chromosome." Bioinformatics 14(7): 583-590.

[LNW03] Li, J., Ng, S.-K. and Wong, L. (2003). "Bioinformatics Adventures in Database Research". 9th International Conference on Database Theory, Siena, Italy, Springer Verlag, LNCS 2672.

[MEK+00] McEntire, R., Karp, P. D., Abernethy, N. F., Benton, D., Helt, G., DeJongh, M., Kent, R., Kosky, A. S., Lewis, S., Hodnett, D., et al. (2000). "An Evaluation of Ontology Exchange Languages for Bioinformatics". 8th Int. Conf. on Intelligent Systems for Molecular Biology, La Jolla / San Diego, CA, USA, AAAI.

[NLF99] Naumann, F., Leser, U. and Freytag, J. C. (1999). "Quality-driven Integration of Heterogeneous Information Systems". 25th Conference on Very Large Database Systems, Edinburgh, UK.

[RZS+99] Rector, A. L., Zanstra, P. E., Solomon, W. D., Rogers, J. E., Baud, R., Ceusters, W., Claassen, W., Kirby, J., Rodrigues, J. M., et al. (1999). "Reconciling users' needs and formal requirements: issues in developing a reusable ontology for medicine." IEEE Transactions on Information Technology in Biomedicine 2(4): 229-42.

[Rit94] Ritter, O. (1994). The Integrated Genomic Database (IGD). Book "The Integrated Genomic Database (IGD)". Suhai, S. New York, Plenum Press, pp.: 57-74.

[Rob94] Robbins, R. J. (1994). Representing Genomic Maps in a Relational Database. Book "Representing Genomic Maps in a Relational Database". Suhai, S. New York, Plenum Press, pp.: 85-96.

[Rob95] Robbins, R. J. (1995). "Information Infrastructure for the Human Genome Project." IEEE Engineering in Medicine and Biology 14(6): 746-759.

[SEOK96] Schuler, G. D., Epstein, J., Ohkawa, H. and Kans, J. A. (1996). "Entrez: Molecular Biology Database and Retrieval System." Methods in Enzymology 266: 141-161.

[Sch98] Schulze-Kremer, S. (1998). "Ontologies for Molecular Biology". 3rd Pacific Symposium on Biocomputing.

[SL90] Sheth, A. and Larson, J. A. (1990). "Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases." ACM Computing Survey 22(3): 183-236.

[SM99] Stein, L. and Mieg, J. T. (1999). " AceDB: A Genome Database Management System." IEEE Computing in Science and Engineering 1(3): 44-52.

[SGHB02] Stevens, R., Goble, C., Horrocks, I. and Bechhofer, S. (2002). "OILing the way to machine understandable bioinformatics resources." IEEE Trans Inf Technol Biomed 6(2): 129-34.

[Ull97] Ullman, J. D. (1997). "Information Integration using Logical Views". 6th Int. Conference on Database Theory; LNCS 1186, Delphi, Greece, LNCS 1186, Springer.

[VP97] Vassalos, V. and Papakonstantinou, Y. (1997). "Describing and Using Query Capabilities of Heterogeneous Sources". 23rd Conference on Very Large Database Systems, Athens, Greece.

[WB96] Weissenbach, J. and Bentolila, S. (1996). "Integrated maps require integrated data." Nature Biotechnology 14: 678.

[ZLAE00] Zdobnov, E. M., Lopez, R., Apweiler, R. and Etzold, T. (2000). "The EBI SRS Server - Recent Developments." Bioinformatics 18(2): 368-373.