UNIVERSITÄT BIELEFELD

TECHNISCHE FAKULTÄT

ARBEITSGRUPPE BIOINFORMATIK / MEDIZINSCHE INFORMATIK

Untersuchung von Life–Science–Datenbeständen zur Identifikation von Genotyp–Phänotyp–Korrelationen

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

vorgelegt der Technischen Fakultät

der Universität Bielefeld

von Dipl.-Inf. Thoralf Töpel

geb. am 6. Februar 1976 in Magdeburg

Thoralf Töpel:
Untersuchung von Life–Science–Datenbeständen zur Identifikation von Genotyp–Phänotyp–Korrelationen

Der Technischen Fakultät der Universität Bielefeld
am 10. August 2004 vorgelegt,
am 26. November 2004 verteidigt und genehmigt.

Gutachter:
Prof. Dr. Hofestädt, Universität Bielefeld
Prof. Dr. Trefz, Universität Tübingen

Prüfungsausschuß:
Prof. Dr. Ragg, Universität Bielefeld
Prof. Dr. Hofestädt, Universität Bielefeld
Prof. Dr. Trefz, Universität Tübingen
Dr. Büntemeyer, Universität Bielefeld

144 Seiten
36 Abbildungen
12 Tabellen

Gedruckt auf alterungsbeständigem Papier (ISO 9706)

Danksagung

Die vorliegende Arbeit entstand während meiner Arbeit an der Technischen Fakultät der Universität Bielefeld und am Institut für Technische und Betriebliche Informationssysteme der Otto–von-Guericke–Universität Magdeburg. Sie wurde im Rahmen eines vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Forschungsprojektes ermöglicht.

An dieser Stelle möchte ich mich besonders bei Prof. Dr. Ralf Hofestädt für die Betreuung meiner Arbeit bedanken. Durch ihn wurde mein Interesse an der Bioinformatik geweckt und die Möglichkeiten und Freiräume für das erfolgreiche Gelingen der Arbeit geschaffen. Prof. Dr. Friedrich–Karl Trefz von der Klinik für Kinder– und Jugendmedizin des Klinikums Reutlingen danke ich für die beständige Motivation durch die praktischen Bezüge zur klinischen Medizin und die Übernahme des externen Gutachtens. Auch Prof. Dr. Georg Paul von der Universität Magdeburg gilt mein besonderer Dank, da er durch seine Unterstützung entscheidend zum Gelingen dieser Arbeit beigetragen hat.

Allen Kollegen der AG Bioinformatik der Technischen Fakultät in Bielefeld und am Institut für Technische und Betriebliche Informationssysteme in Magdeburg danke ich für das motivierende Arbeitsumfeld, anregende Diskussionen und konstruktive Kritik. Besonderer Dank gilt dabei Nadine Fröhlich, Dr. Uwe Scholz, Matthias Lange, Andreas Freier, Andreas Stephanik und Roland Schnee. Für die ausgezeichnete technische Unterstützung und auch Motivation geht eine großes Dankeschön an Gerd Lange, Fred Kreutzmann und Steffen Thorhauer. Ich danke auch Anke Schneidewind, Daniel Reitz, Daniel Tiedge und Dr. Sören Balko für ihre große Ausdauer bei der täglichen Nordparkrunde.

Für ihre Anregungen und Diskussionen möchte ich den Kollegen des Verbundprojektes „Modellierung von genregulatorischen Netzen“ danken, insbesondere Dr. Ulrike Mischke, Dagmar Scheible und Dr. Stephanie Doehr. Nicht vergessen möchte ich natürlich die Studenten, die durch ihre Arbeit zum Gelingen dieses Vorhabens beigetragen haben. Vielen unerwähnten Freunden und Kollegen danke ich außerdem für die Zeit, die wir miteinander verbringen konnten und die Unterstützung und Anregung, die sie mir auch außerhalb des Campus entgegengebracht haben.

Kurzfassung

Die Fortschritte im Bereich der biotechnologischen Forschung der letzten Jahre haben zu einer Vielzahl von unterschiedlichen Datenbanken und Informationssystemen geführt, die ihre Daten für weitergehende Untersuchungen über das World Wide Web bereitstellen. Diese weltweit verteilten Life–Science–Datenquellen beschreiben verschiedene Aspekte biologischer Systeme und verzeichnen ein beständiges Anwachsen des verfügbaren Datenbestandes. Die Zusammenführung der vorhandenen molekularbiologischen und medizinischen Daten und ihre Untersuchung auf Beziehungen und Abhängigkeiten ist für den Nutzer von größtem Interesse. Die dazu im Rahmen dieser Arbeit präsentierten Ergebnisse wurden im Deutschen Humangenomprojekt durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert und in ein Teilprojekt eines Konsortiums aus GBF Braunschweig, GSF München, Universität zu Köln, Universität Bielefeld und Universität Tübingen eingebracht.

Die Vorstellung eines Vorschlages für eine flexible Analyseumgebung, die die Suche nach Korrelationen von Genotyp und Phänotyp bei angeborenen Stoffwechselerkrankungen innerhalb integrierter Datenbestände unterstützt, ist Ziel dieser Arbeit. Dazu werden verschiedene Architekturen zur Datenintegration vorgestellt und bestehende Ansätze anhand bestimmter Merkmale gegenübergestellt. Zur Vorbereitung der Datenintegration werden weiterhin unterschiedliche molekularbiologische und medizinische Datenquellen analysiert und die erforderlichen Datenbestände für den Integrationsschritt ausgewählt. In diesem Rahmen werden auch Anforderungen an eine Datenbank für Mutationen und assoziierte Phänotypen formuliert und umgesetzt, da eine solche Datenquelle derzeit noch nicht in entsprechendem Funktionsumfang verfügbar ist. Auf der Basis dieser klinischen und molekulargenetischen Daten sind fallbasierte Suchanfragen möglich, die bereits Genotyp–Phänotyp–Korrelationen im Kleinen, beispielsweise zur Unterstützung der Differentialdiagnostik, ermöglichen.

Bei der Untersuchung der vielfältigen Zusammenhänge innerhalb und zwischen den einzelnen Komponenten eines biologischen Systemes ist es jedoch auch erforderlich, neben eindeutigen Ergebnissen auch ähnliche Resultate zu ermöglichen. Für die Berechnung dieser Ähnlichkeiten werden sowohl eigene Ansätze als auch bestehende Verfahren vorgestellt und auf ihre Eignung im vorliegenden Szenario untersucht. Der Ausgangspunkt für die Untersuchung von Beziehungen zwischen Genotypen und Phänotypen sind die in einer Integrationsdatenbank zusammengeführten Daten aus verschiedenen Life–Science–Quellen.

Als Ergebnis dieser Arbeit wird neben einem Architekturvorschlag auch ein funktionsfähiger, webbasierter Prototyp des Gesamtsystemes präsentiert. Dabei werden die einzelnen Komponenten des Architekturvorschlages vorgestellt und in ihrer Funktionsweise erläutert. Durch die Integration von medizinischen und molekularbiologischen Daten wird im Rahmen eines Beispielszenarios die Nutzung des Prototypen und das Vorgehen innerhalb der Analyseumgebung verdeutlicht.

Inhaltsverzeichnis

Abbildungsverzeichnis **ix**

Tabellenverzeichnis **xiii**

Abkürzungsverzeichnis **xv**

1 Einleitung **1**

1.1 Motivation und Einordnung 1

1.2 Inhalt und Struktur 5

2 Molekularbiologische und informationstechnische Grundlagen **9**

2.1 Molekularbiologische Grundlagen 9

2.1.1 Die DNS als Träger der genetischen Information 10

2.1.2 Von der DNS zum Protein 12

2.1.3 Metabolismus und Stoffwechselwege 17

2.1.4 Erbkrankheiten und Stoffwechseldefekte 18

2.1.5 Wirkstoffpointing 19

2.2 Informationstechnische Grundlagen 20

2.2.1 Informationssysteme und Relationale Datenbanksysteme 20

2.2.2 Fallbasiertes Schließen 22

2.2.3 Ähnlichkeitsbewertung 25

2.3 Zusammenfassung 26

3 Analyse von Datenquellen und Integrationsansätzen **29**

3.1 Verfügbare Datenquellen 29
3.1.1 Medizinische Datenquellen 31
3.1.2 Genomische Sequenzdatenquellen 33
3.1.3 Proteinsequenz– und Proteinstrukturdatenquellen 34
3.1.4 Metabolische und regulatorische Datenquellen 37
3.1.5 Wirkstoffdatenquellen 38
3.1.6 Zusammenfassende Gegenüberstellung 40
3.2 Integrationsarchitekturen und bestehende Ansätze 40
3.2.1 Architekturen zur Integration von Datenquellen 43
3.2.2 Bewertung von Integrationsansätzen 48
3.2.3 Vorstellung ausgewählter Integrationsansätze 50
3.3 Zusammenfassung 57

4 Datenbank für Mutationen und assoziierte Phänotypen 59

4.1 Motivation 59
4.2 Anforderungen 60
4.3 Diskussion vorhandener Mutationsdatenbanken 63
4.3.1 PAH–Mutationsdatenbank 63
4.3.2 Mutationsdatenbank für Tetrahydrobiopterin–Mangel (BIODEF und BIOMDB) 64
4.3.3 ARPKD–Mutationsdatenbank 64
4.3.4 Zusammenfassende Gegenüberstellung 64
4.4 Architekturvorschlag 66
4.5 Realisierung 68
4.5.1 Dateneingabekomponente 68
4.5.2 Datenauswertungskomponente 71
4.5.3 Fallbasierte Anfrageschnittstelle 71
4.5.4 Datenbanksystem 78
4.6 Zusammenfassung 78

5 Ähnlichkeiten und Beziehungen in Life–Science–Datenbeständen 81

5.1 Ähnlichkeit auf Domänenebene 82

5.1.1 Domäne der klinischen Phänotypen 83

5.1.2 Domäne der biochemischen Reaktionen und Reaktionsketten . . . 85

5.1.3 Domäne der genomischen Sequenzen 91

5.1.4 Zusammenfassende Gegenüberstellung 94

5.2 Genotyp–Phänotyp–Korrelation auf Szenarioebene 95

5.3 Zusammenfassung 101

6 Vorstellung des Prototypen des Gesamtsystemes 103

6.1 Architektur und Komponenten 104

6.1.1 Architektur im Überblick 104

6.1.2 Replikationssteuerung 105

6.1.3 Domänendatenverwaltung 107

6.1.4 Genotyp–Phänotyp–Analyse 109

6.2 Vorgehen und Anwendung am Beispiel 109

6.3 Zusammenfassung 116

7 Zusammenfassung und Ausblick 117

Anhang 123

A WWW-Adressen ausgewählter molekularbiologischer Datenquellen 123

B Ausgewählte Datenquellen im Detail 125

B.1 EMBL 125

B.1.1 Originaldatensatz als Flatfile 125

B.1.2 Originaldatensatz im XML–Format 126

B.1.3 Adapterschema 128

C Beispiel für einen Fallbericht 129

D Glossar **133**

Literaturverzeichnis **135**

Abbildungsverzeichnis

1.1 Wachstum des Datenvolumens in der Nukleotidsequenzdatenbank EMBL 2

1.2 Übersicht zum Aufbau des DHGP–Projektes „Modellierung genregulatorischer Netzwerke“ 4

1.3 Untersuchung von Genotyp–Phänotyp–Korrelationen durch zwei unterschiedliche Ansätze 6

2.1 Vom Genotyp zum Phänotyp 9

2.2 Basenpaarung 10

2.3 Grafische Darstellung von Chromosomenmutationen (nach [Bro99]) . . . 12

2.4 Steuerung einer Reaktionskette im Stoffwechsel (nach [Ste87]) 17

2.5 Pharmakologisches Prinzip des Wirkstoffpointings 20

2.6 Prozeßmodell des CBR–Zyklus (nach [AP94]) 24

3.1 Wissenswerte Informationen zu einem Gen (nach [GJ02]) 30

3.2 Klassifikation für Multidatenbanksysteme (nach [SL90]) 44

3.3 Referenzarchitektur für Multidatenbanken (nach [LMR90]) 46

3.4 Allgemeine Architektur föderierter Datenbanksysteme (nach [Con97]) . . 47

3.5 DiscoveryLink–Architektur (nach [HSK+01]) 51

3.6 Architektur des FRIDAQ–Frameworks als UML–Klassendiagramm (nach [Sch02]) 52

3.7 Vorgehensmodell zur Anwendung des FRIDAQ–Frameworks als UML–Aktivitätsdiagramm (nach [Sch02]) 53

3.8 Architektur der Mediator–basierten Prototypen des BioDataServers (nach [FHL+02]) 54

3.9 TAMBIS–Modell (nach [GSN+01]) 56

4.1 Architektur des Ramedis–Systemes mit unterschiedlichen Nutzern, Analyse– und Eingabekomponente, Datenbanksystem 66

4.2 Vereinfachtes Relationenschema der Ramedis–Datenbank 69

4.3 Allgemeine Daten zu einem Fallbericht in der Dateneingabekomponente von Ramedis 70

4.4 Darstellung des Ausschnittes eines Fallberichtes und der Visualisierung des Wachstumsparameters Länge in perzentiler Darstellung in der Auswertungskomponente von Ramedis 72

4.5 Darstellung des MAC/FAC–Modell (nach [FGL95]) 75

4.6 Algorithmus für die Suche von ähnlichen Fällen in Ramedis anhand der Parameter Symptom, Laboruntersuchung und ethnische Herkunft 78

5.1 Darstellung der Beispielanfrage nach Nutzung des PathAligner im WWW 91

5.2 Darstellung verwandtschaftlicher Beziehungen zwischen Organismen (nach [Kem69]) 94

5.3 Darstellung der Graphenstruktur für einen Beispieldatenbestand 98

5.4 Algorithmus für die Breitensuche im Graphen X bei gegebenem Subgraphen J und dem Startknoten u 99

5.5 Mögliche Genotyp–Phänotyp–Korrelationen als Graph mit der Kennzeichnung der untersuchten Domänen mit ihren englischen Bezeichnungen und den Identifikatoren der beteiligten Datensätze 100

6.1 Architektur des Prototypen mit externen Analysewerkzeugen und den integrierten, externen Datenquellen 105

6.2 Darstellung der Zielkonflikte der Replikationskontrolle (nach [BD96]) . . 107

6.3 Vorgehen zur Erzeugung der Informationen über nutzerspezifische Domänen innerhalb der Domänendatenverwaltung als UML–Aktivitätsdiagramm 108

6.4 Verschiedene molekularbiologische Datenquellen und ihre zu integrierenden Datenbestände 111

6.5 Nutzerspezifische Integrationsdatenbank mit den ausgewählten Inhalten und den definierten Domänen über den integrierten Daten 112

6.6 Screenshots der webbasierten, grafischen Nutzerschnittstelle mit (a) Anfragemaske mit vordefinierten Informationsdomänen, (b) Beispiel eines vom Nutzer ausgewählten Pfades vom Genotyp (Sequence) zum Phänotyp (Patient), (c) Darstellung von Datensätzen der Originalrelation in Tabellenform, (d) Auswahlmöglichkeit zwischen verschiedenen Fremdschlüsselbeziehungen innerhalb des integrierten Datenbestandes ausgehend von einem bestimmten Attribut 113

6.7 Darstellung des Datenumfanges der Integrationsdatenbank am Beispiel Diabetes mellitus MODY 1 mit (a) einem Überblick der über dem Datenbestand angelegten Informationsdomänen und den enthaltenen Daten sowie (b) der Darstellung einer Auswahl von einzelnen Datensätzen zur Beispielerkrankung, die aus verschiedenen Quellen integriert wurden . . . 115

Tabellenverzeichnis

2.1 Nukleotidbasen und ihre Ein–Buchstaben-Abkürzungen 11

2.2 Aminosäuren und ihre Drei– und Ein–Buchstaben–Abkürzungen 14

2.3 Die Codons aus Nukleotidbasen und ihre zugehörigen Aminosäuren . . . 15

3.1 Gegenüberstellung verschiedener medizinischer und molekularbiologischer Datenquellen anhand von drei Merkmalen 41

3.2 Gegenüberstellung verschiedener molekularbiologischer Integrationsansätze anhand von zehn Merkmalen (nach [Sch02]) 49

4.1 Gegenüberstellung existierender Mutationsdatenbanken anhand ausgewählter Anforderungen 65

4.2 Übersicht zum aktuellen Datenvolumen in Ramedis (Stand August 2004) 79

5.1 Übersicht der Merkmale für den klinischen Phänotyp innerhalb der Ramedis–Datenbank mit Antworttypen nach [Goo96] 84

5.2 Beispiel für den Vergleich von klinischen Phänotypen 84

5.3 Berechnung des Elternschaftskoeffizienten für ein Beispiel (nach [Kem69]) 94

5.4 Gegenüberstellung von Verfahren zur Ähnlichkeitsbewertung anhand verschiedener Merkmale 95

5.5 Gegenüberstellung von Objekten innerhalb der Problemstellung und den äquivalenten Elementen der Graphentheorie 97

Abkürzungsverzeichnis

API	-	Application Programming Interface
BLOB	-	Binary Large Object
BMBF	-	Bundesministerium für Bildung und Forschung
CAS	-	Chemical Abstract Service
CBR	-	Case-Based Reasoning
CDS	-	Coding Sequence
CORBA	-	Common Object Request Broker Architecture
DB	-	Datenbank
DBS	-	Datenbanksystem
DBMS	-	Datenbankmanagementsystem
DDBJ	-	DNA Database of Japan
DDL	-	Data Definition Language
DHGP	-	Deutsches Humangenomprojekt
DNA	-	Desoxyribonucleinacid
DNS	-	Desoxyribonukleinsäure
DTD	-	Document Type Definition
EBI	-	European Bioinformatics Institute
ER	-	Entity–Relationship
EMBL	-	European Molecular Biology Laboratory
FTP	-	File Transfer Protocol
GBF	-	Gesellschaft für Biotechnologische Forschung
GSF	-	Gesellschaft für Strahlenforschung
GUI	-	Graphical User Interface
HGMD	-	Human Gene Mutation Database

Fortsetzung auf der nächsten Seite

Fortsetzung von der vorherigen Seite

HGP	-	Human Genom Project
HTML	-	Hypertext Markup Language
HTTP	-	Hypertext Transfer Protocol
IS	-	Informationssystem
JDBC	-	für Java Database Connectivity
JIPID	-	Japan International Protein Information Database
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
MIPS	-	Munich Information Center for Protein Sequences
MODY	-	Maturity-onset Diabetes of the Young
MMDB	-	Molecular Modeling Database
NCBI	-	National Center for Biotechnology Information
NCGR	-	National Center for Genome Resources
NIH	-	National Institutes of Health
ODBC	-	Open Database Connectivity
OMG	-	Object Management Group
OMIM	-	Online Mendelian Inheritance in Man
OQL	-	Object Query Language
OTC	-	Ornithin–Transcarbamylase
PAH	-	Phenylalaninhydroxylase
PDB	-	Protein Data Bank
PIR	-	Protein Information Resource
PKU	-	Phenylketonurie
PSD	-	Protein Sequence Database
RCSB	-	Research Collaboraty for Structural Bioinformatics
RNA	-	Ribonucleinacid
RNS	-	Ribonukleinsäure
SCOP	-	Structural Classification of Proteins
SI	-	Systéme International d’ Unités
SIB	-	Swiss Institute of Bioinformatics
SOAP	-	Simple Object Access Protocol
SQL	-	Structured Query Language
SRS	-	Sequence Retrieval System

Fortsetzung auf der nächsten Seite

Fortsetzung von der vorherigen Seite

TCP/IP	-	Transmission Control Protocol / Internet Protocol
TrEMBL	-	Translation of EMBL nucleotide sequence database
UML	-	Unified Modeling Language
URL	-	Uniform Resource Locator
WWW	-	World Wide Web
XML	-	Extensible Markup Language

1 Einleitung

In diesem Kapitel werden neben der Motivation für die vorliegende Arbeit auch die Zielstellung und die Gliederung der Arbeit vorgestellt. Dazu werden die Anforderungen und Architekturen zur Integration von Life–Science-Datenquellen betrachtet und ein Vorschlag skizziert, wie die integrierten medizinischen und molekularbiologischen Daten zur Identifikation von Genotyp–Phänotyp–Korrelationen genutzt werden können. Ein Teil der prototypischen Realisierung der Ergebnisse dieser Arbeit wurde im Rahmen einer Förderung innerhalb des Deutschen Humangenomprojektes ermöglicht.

1.1 Motivation und Einordnung

Mit der Publikation der Rohsequenz und einer ersten Analyse des menschlichen Genoms in der Fachzeitschrift NATURE [Con01] durch das öffentlich geförderte, internationale Humangenomprojekt im Frühjahr des Jahres 2001 wurde das Interesse der Öffentlichkeit wieder verstärkt auf diese Bemühungen zum Verstehen der komplexen molekularbiologischen und biochemischen Vorgänge im Organismus gelenkt. Zur gleichen Zeit veröffentlichten auch der Amerikaner VENTER und die Firma Celera Genomics ihre Sequenz des humanen Genoms in der wissenschaftlichen Zeitschrift SCIENCE [Ven01]. Dabei stellt sich heraus, daß das menschliche Genom 3,2 Milliarden Bausteine umfaßt und 30000 bis 40000 Gene enthält. Dies war jedoch erst ein vorläufiges Ergebnis. Die verschiedenen Gruppen von Wissenschaftlern füllen weiterhin kontinuierlich die Datenbanken mit sequenzierter Desoxyribonukleinsäure (DNS).

Das Verständnis dieses Bauplanes des Menschen, des menschlichen Genoms, bietet der Medizin und Biotechnologie bisher nicht gekannte Möglichkeiten. Es könnte somit geklärt werden, warum manche Menschen seltene Krankheiten bekommen oder für bestimmte Erkrankungen besonders anfällig sind. Dies kann bereits heute auf bestimmte Fehlfunktionen in den menschlichen Genen zurückgeführt werden. Für Patienten mit genetisch bedingten Erkrankungen bedeutet das eine Chance auf präzise und spezifische Früherkennung, Diagnose und Therapie. Doch nicht nur Menschen mit Erbkrankheiten sind von der Weiterentwicklung betroffen. Durch das Verständnis der Erbinformation könnte bald zielgerichtet in die Proteinsynthese eingegriffen werden und so die pharmakologische Forschung weitergebracht werden.

So wurden im Dezember 2002 in der Fachzeitschrift SCIENCE die vielversprechendsten Forschungsergebnisse des Jahres 2002 vorgestellt. Auf den ersten Platz wurden Veröffentlichungen über die Bedeutung der Ribonukleinsäure (RNS) bei der Genregulation gewählt [Cou02]. Entgegen der vorherrschenden Meinung stellte sich heraus, daß eine bestimmte Klasse von RNS–Molekülen nicht nur genetische Informationen und andere Moleküle transportiert, sondern eine wichtige Rolle bei der Steuerung von Zellprozessen spielt. Die Bedeutung der Forschung auf dem Gebiet der Molekularbiologie wurde außerdem durch den Drittplazierten unterstrichen [The02b]. Dabei handelte es sich um die Entschlüsselungen des Erbgutes der Reispflanze, des Moskito und des Malariaerregers, die nunmehr die Hoffnungen wecken, zielgerichtetere Züchtungsversuche mit Reis und neue Therapien gegen Malaria durchführen zu können.

Der Träger des Erbgutes, die DNS, ist in Form einer Reihe einzelner Chromosomen im Zellkern gelagert. Im menschlichen Körper findet ein ständiger Prozeß des Abschreibens, Kopierens und Übersetzens dieser Erbinformationen statt. Anhand dieser Daten werden so neben anderen Proteinen bestimmte Enzyme synthetisiert. Diese werden für den Ablauf der Stoffwechselprozesse in jeder Körperzelle benötigt.

Die Informationen über die verschiedenen Gene, Enzyme und Stoffwechselvorgänge sind bereits in unterschiedlichen Datenquellen im Internet verfügbar. So bieten EMBL [KAA+04] und GenBank [BKL+04] die DNA–Sequenzen des menschlichen Genoms sowie SwissProt [BBA+03] und PIR [BGH+01] Informationen über Proteine. Daten über Enzyme und metabolische Informationen sind beispielsweise in BRENDA [SCE+04] verfügbar. Außerdem wurde die bekannte Boehringer Wandtafel der Stoffwechselwege im KEGG–System [KGK+04] für die Nutzung über das Internet realisiert. Die meisten der gespeicherten Daten in den verschiedenen Systemen sind jedoch intern in unterschiedlichen Präsentationen vorhanden und werden dem Nutzer auch oft auf recht heterogene Weise zugänglich gemacht. Außerdem steigt der Datenbestand dieser Datenquellen ständig an. Die Abbildung 1.1 illustriert dieses Wachstum am Beispiel von EMBL.

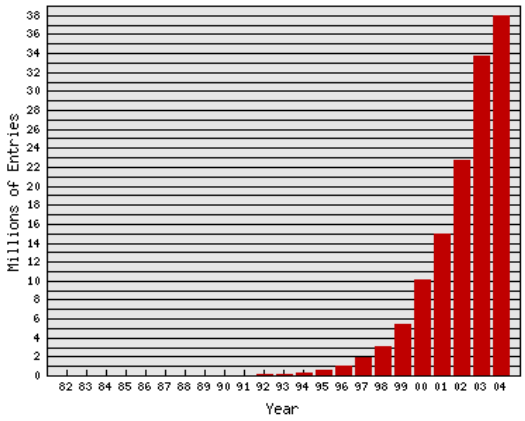

Abbildung 1.1: Wachstum des Datenvolumens in der Nukleotidsequenzdatenbank EMBL

Es liegt somit eine große Anzahl von wertvollen Datenquellen vor, die spezielle Einblicke in spezifische Aspekte von biologischen Systemen geben. Diese reichen von den bereits beschriebenen Nukleotidsequenzen des menschlichen Genoms bis zu den klinischen Daten von einzelnen Patienten. Durch die Entstehung dieser Datenquellen im Rahmen von untereinander abgegrenzten Forschungsprojekten sind die meisten der verschiedenen Life–Science–Datenbestände wenig miteinander verbunden oder aufeinander abgestimmt. Dennoch müssen diese verteilten, heterogenen Datenquellen gemeinsam genutzt werden, um Daten für die verschiedensten Anwendungen der Bioinformatik zu liefern. Eine manuelle Recherche in relevanten Datenbanken und Informationssystemen ist jedoch bei der großen Menge an verfügbaren Daten nicht mehr möglich [LR03]. Diese Originaldaten sind durch eine hohe Heterogenität in Bezug auf die beschriebenen biologischen Aspekte, verwendete Schemata und Formate gekennzeichnet.

Zur einheitlichen Verwaltung aller von einer Anwendung benötigten Daten wird die Datenintegration verwendet. Sie ermöglicht nach [HS97] eine kontrollierte nicht–redundante Datenhaltung des gesamten relevanten Datenbestandes. Abhängig von der realisierten Architektur werden im Idealfall beispielsweise benutzergerechte Anfragesprachen angeboten, die Anfragen ohne Rücksicht auf die interne Realisierung der Datenspeicherung zulassen. Ein effizienter Zugriff auf die Datenbestände wird dabei durch eine interne Optimierung ermöglicht. Ziele und Anforderungen an eine Integration biologischer Daten wurden in mehreren Aufsätzen [Kar95, DOB95, MR95] beschrieben. Die Integration und Analyse der verfügbaren molekularbiologischen und medizinischen Daten ist Konsequenz dieser Entwicklung.

Im Rahmen des Deutschen Humangenomprojektes wurde ein Konsortium von fünf Partnern gebildet, das durch die Nutzung von medizinischen und molekularbiologischen Datenquellen und speziellen Analysemethoden die Modellierung genregulatorischer Netzwerke untersucht. Als Partner waren an diesem Vorhaben die folgenden Forschungseinrichtungen und Firmen beteiligt: Universität Göttingen (Prof. Wingender), Biobase Biological Databases GmbH, GSF München (Dr. Werner), Genomatix Software GmbH, Universität zu Köln (Prof. Schomburg), Universität Bielefeld (Prof. Hofestädt) und die Universität Tübingen (Prof. Trefz). Zur Bearbeitung des Projektes ist die Nutzung und Analyse vorhandener Life–Science–Datenquellen, deren Inhalt in den letzten Jahren erheblich an Qualität und Quantität gewonnen hat, unbedingt erforderlich. Dieses Vorhaben wurde nach [DEF+02] in drei Teilbereiche gegliedert, die nachfolgend erläutert werden. Die Abbildung 1.2 illustriert den Projektaufbau in einer Übersicht.

Der erste Teilbereich betrachtet die Integration relevanter Informationsressourcen. Dazu wird eine Anzahl von Datenquellen, die von den Projektpartnern entwickelt wurden, für die speziellen Anforderungen des Projektes angepaßt und zusammen mit weiteren, externen Quellen teilweise integriert. Die Entwicklung einer formalen Beschreibung regulatorischer und metabolischer Netzwerke bildet den zweiten Teilbereich des Verbundprojektes. Neben der Beschreibung der Architektur von Netzwerken zur Regulation der Signaltransduktion und Transkription werden auch metabolische Netzwerke formal beschrieben. Im dritten Teilbereich werden die entwickelten Vorgehensweisen und Modelle

am Anwendungsbeispiel MODY getestet. Hier werden anhand einer Beispielkrankheit (Diabetes mellitus, Typ MODY) die Ergebnisse des Projektes überprüft und bei Bedarf Modelle oder Verfahren angepaßt. Dazu werden beispielsweise regulatorische Komponenten in Promotorsequenzen identifiziert und mit den Datenquellen der beteiligten Projektpartner auf ihre Relevanz untersucht.

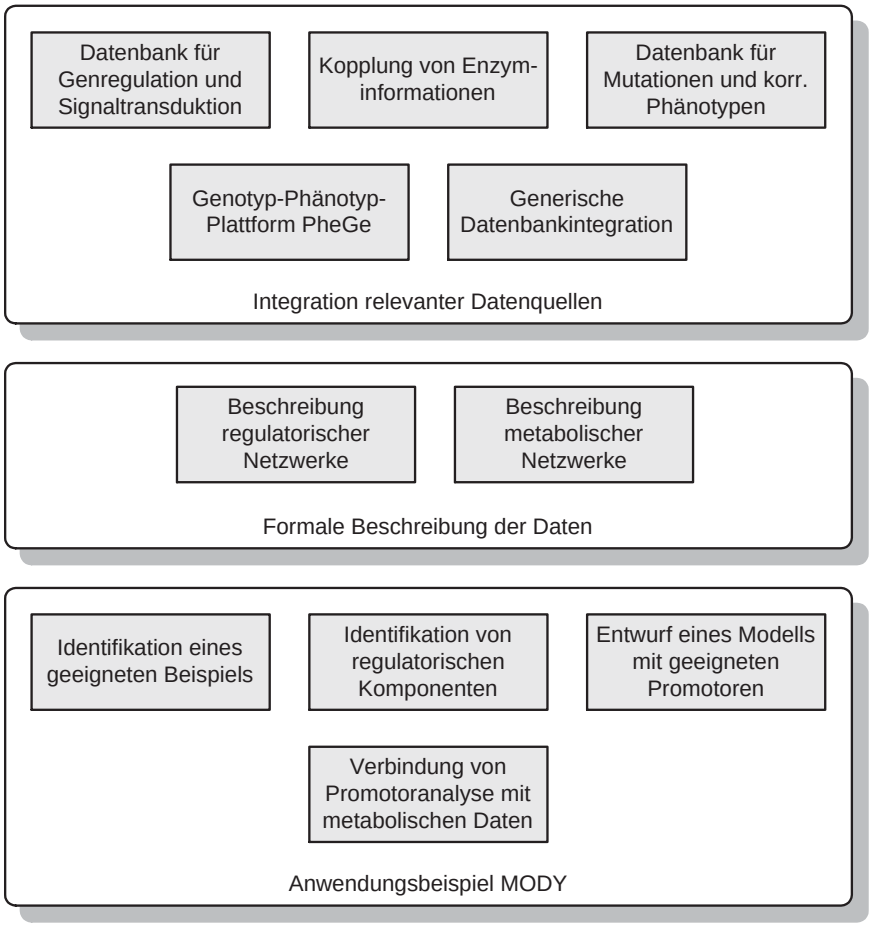

Abbildung 1.2: Übersicht zum Aufbau des DHGP–Projektes „Modellierung genregulatorischer Netzwerke “

Die vorliegende Arbeit liefert Beiträge zum ersten Teilbereich des Projektes und bildet die Grundlage für eine Reihe prototypischer Softwareentwicklungen, die auch in diesem Konsortium Anwendung fanden. Dies schließt die Entwicklung einer Datenbank für Mutationen und assoziierte Phänotypen sowie die Integration verschiedener medizinischer und molekularbiologischer Datenquellen ein. Auf der Basis der integrierten Daten konnten dann Möglichkeiten untersucht werden, um die Suche nach möglichen Zusammenhängen zwischen Genotyp und Phänotyp, den Genotyp–Phänotyp–Korrelationen, zu unterstützen.

Der Begriff der *Genotyp–Phänotyp–Korrelation* beschreibt Zusammenhänge zwischen der molekulargenetischen Ebene, dem einzelnen Gen oder dem gesamten Genom, das

durch eine wohldefinierte Nukleotidsequenz beschrieben wird, und der klinischen Ebene, die sich als Menge von direkt oder indirekt beobachtbaren Merkmalen des Organismus manifestiert. Zwischen der DNS–Sequenz des Genotyps und dem letztendlichen Erscheinungsbild des Individuums, dem Phänotyp, liegen jedoch eine Reihe von unterschiedlichen Zwischenschritten, u.a. Proteinsynthese, Genregulation, beteiligte Stoffwechselwege und entsprechende Umwelteinflüsse, die ein Wirknetz bilden, das eine enorme Komplexität entwickelt.

Zur Untersuchung von Beziehungen zwischen Genotyp und Phänotyp werden in dieser Arbeit zwei Ansätze verfolgt. Die entsprechenden unterschiedlichen Vorgehensweisen sind in der Abbildung 1.3 dargestellt. Für eine Genotyp–Phänotyp–Korrelation im Kleinen wird ein Informationssystem für Muationen und Phänotypen entworfen, implementiert und mit Fallberichten zu angeborenen Stoffwechselerkrankungen angereichert. Mit diesem Vorgehen wird durch die Bereitstellung einer geeigneten elektronischen Infrastruktur die Erfassung relevanter Datensätze, die in der klinischen Arbeit durch medizinische Fachexperten ermittelt werden, und die direkte Verbindung zwischen den klinischen Daten und den entsprechenden molekulargenetischen Untersuchungsergebnissen ermöglicht. Als alternative Möglichkeit wird die Genotyp–Phänotyp–Korrelation im Großen vorgestellt, mit der durch Datenintegration eine Reihe von Life–Science–Datenquellen angesprochen werden kann und Beziehungen zwischen den integrierten Daten hergestellt werden kann.

Im Rahmen dieser Arbeit wird somit ein Beitrag zur Nutzung von Methoden der Informatik im Forschungsbereich der Biologie geleistet. Aufbauend auf der Sammlung relevanter Daten über Mutationen und ihre korrespondierenden Phänotypen und der Analyse weiterer verfügbaren Datenquellen in den verschiedenen molekularbiologischen und medizinischen Bereichen wird ein Gesamtschema entwickelt, das ausgewählte Daten über einen Integrationsdienst verbindet und einem Nutzer homogen und nicht–redundant bereitstellt. Auf der Grundlage gesammelter und integrierter Daten werden verschiedene Verfahren angewandt, um die Suche nach Genotyp–Phänotyp–Korrelationen zu unterstützen.

1.2 Inhalt und Struktur

Die Konzeption eines Architekturvorschlages für eine integrierte Analyseumgebung zur Unterstützung der Suche von Genotyp–Phänotyp–Korrelationen und die Überprüfung der vorgeschlagenen Vorgehensweise werden als Ergebnisse der vorliegenden Arbeit in den letzten Kapiteln vorgestellt werden. Bevor jedoch diese präsentiert werden, sind Vorbetrachtungen und Analyseschritte notwendig, die in ihrem Ablauf und ihrer Abfolge grob den nachfolgenden Kapiteln entsprechen. Eine Orientierung an der vorliegenden Struktur der Arbeit führt den Leser also ausgehend von der Zielstellung über die durchgeführten Vorarbeiten bis zur Präsentation und Diskussion der Ergebnisse.

Die molekularbiologischen und informationstechnischen Grundlagen werden im nachfolgenden 2. Kapitel dargestellt. Aus Sicht der Biologie wird dabei auf die DNS als Träger

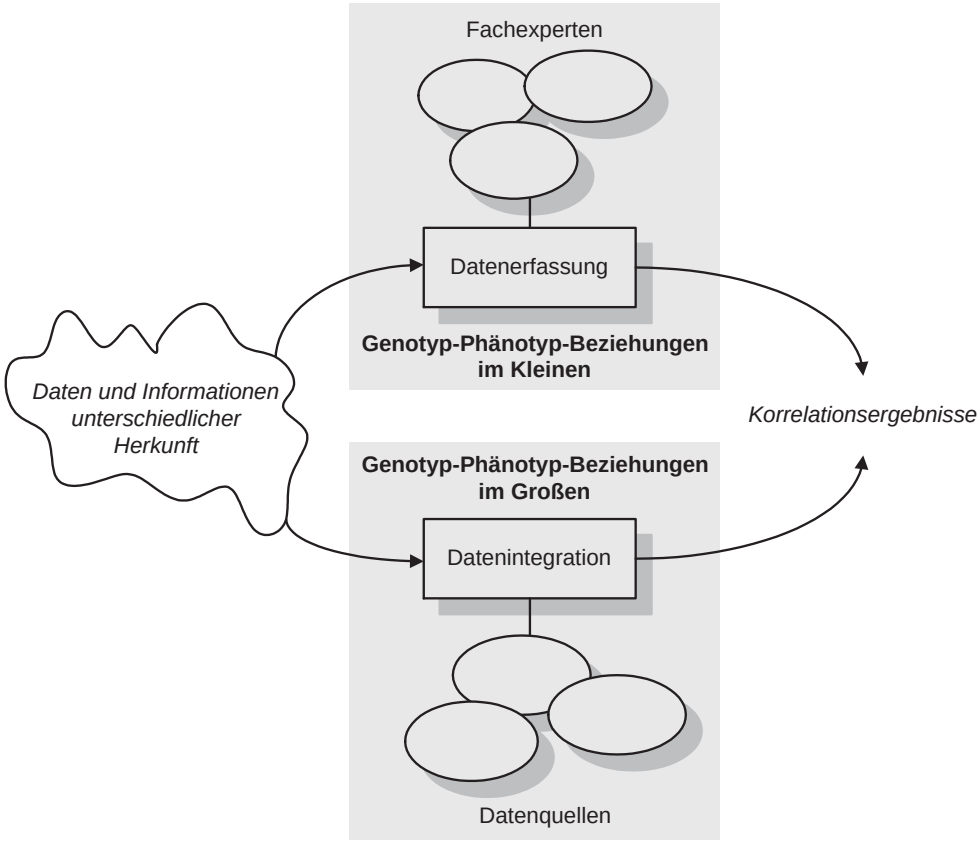

Abbildung 1.3: Untersuchung von Genotyp–Phänotyp–Korrelationen durch zwei unterschiedliche Ansätze

der genetischen Information, die Enzyme und die Stoffwechselwege fokussiert. Für die im Rahmen dieser Arbeit besonders interessanten Stoffwechselerkrankungen werden Entstehung und Charakteristika vorgestellt. Außerdem soll der Begriff des Wirkstoffpointings näher erläutert werden. Der zweite Teil dieses Kapitels, die Grundlagen aus dem Blickwinkel der Informatik, skizziert die Merkmale eines Informationssystems und relationaler Datenbanksysteme. Abschließend erfolgen einführende Bemerkungen zum Prinzip des fallbasierten Schließens (*Case–based Reasoning, CBR*).

Der erste Teil des Kapitals 3 soll eine Auswahl bereits bestehender und via WWW zugreifbarer Datenquellen vorstellen und analysieren. Dabei wird besonders Wert auf die Modellierung und Darstellung der Verbindung zu anderen Systemen gelegt (z.B. MIM–Nummern als Schlüssel in OMIM, EC in BRENDA u.a.). Diese Ergebnisse fließen später in die Modellierung des Gesamtsystemes ein. Als Entscheidungshilfe für die Auswahl der zu integrierenden medizinischen und molekularbiologischen Datenquellen werden die Resultate der vorangehenden Sichtung in einer zusammenfassenden Tabelle strukturiert. Nachdem nun die verschiedenen, bereits verfügbaren Datenbanken und Informationssysteme vorgestellt wurden, widmet sich der nachfolgende Teil des 3. Kapitels

der Datenintegration und ihrer Realisierung. Dazu wird die Motivation zur Integration unter Berücksichtigung von Notwendigkeit und Nutzen der Zusammenführung der unterschiedlichen Daten ausgeführt. Anschließend werden verschiedene Integrationsarchitekturen vorgestellt. Aufbauend auf den vorgestellten Architekturen werden bestehende Integrationsansätze anhand einer Reihe von Merkmalen gegenübergestellt und bewertet. Einige dieser Ansätze werden anschließend detaillierter vorgestellt. Dabei werden Vor- und Nachteile diskutiert und als Ergebnis der Untersuchung die Nutzung des FRIDAQ–Frameworks nach [Sch02] mit der Realisierung im BioDataServer durch [FHL+02] begründet.

Die Motivation und der Entwurf einer Mutations– und Phänotyp–Datenbank unter dem Namen *Ra*re *Me*tabolic *Di*seases Database Ramedis zur Sammlung klinischer Daten, wie Laborparameter und Symptomatik einzelner Patienten, sind Inhalt des Kapitels 4. Dazu werden prinzipielle Anforderungen an eine Datenbank für Mutationen und assoziierte Phänotypen formuliert. Dabei werden verschiedene Aspekte, beispielsweise Erweiterbarkeit, Vergleichbarkeit, Referenzierbarkeit, Datenschutz und Datensicherheit beachtet. Anschließend werden bereits vorhandene Systeme als aktueller Stand der Technik untersucht und unter Berücksichtigung der formulierten Anforderungen zusammenfassend verglichen. Aus diesen vorbereitenden Arbeiten resultiert ein Architekturvorschlag, der sich in verschiedene Komponenten gliedert und als Prototyp realisiert wurde. Durch eine enge Zusammenarbeit mit der Universität Tübingen und dem Klinikum Reutlingen bei der Entwicklung des Systemes wurde sichergestellt, daß die Anwendung den Bedürfnissen der zukünftigen Anwender weitgehend entspricht und sich das System beim Einsatz in der Praxis bewährt hat.

Für die Auswertung des gesammelten Datenbestandes werden Anfragemöglichkeiten bereitgestellt, die sich an den Prinzipien des fallbasierten Suchens, einem Teilbereich des CBR, orientieren. Über die Frage nach den Eigenschaften, die einen Suchanfrage charakterisieren, werden bereits gespeicherte, ähnliche Fälle untersucht. Dieses Vorgehen ermöglicht beispielsweise eine Unterstützung bei der Differentialdiagnostik, die auf die Abgrenzung und Identifikation einer bestimmten Erkrankung innerhalb einer Menge von symptomatisch ähnlichen Krankheiten ausgerichtet ist. Durch eine paralle Sammlung von molekulargenetischen Untersuchungen, die krankheitsrelevante Mutationen feststellen, werden bereits spezifische Korrelationen von Genotyp und Phänotyp für einzelne Fälle verfügbar. Der Datenbestand von Ramedis wird später in das Gesamtsystem integriert und liefert für das untersuchte Szenario die klinischen Daten zum Phänotyp.

Das Kapitel 5 widmet sich zwei Wegen zur Auswertung und Analyse der integrierten molekularbiologischen und medizinischen Daten. Dazu werden beispielhaft molekularbiologische und medizinische Daten der Domänen klinischer Phänotyp, biochemische Reaktionen und Reaktionsketten sowie genomische Sequenzen untersucht. Das umfaßt die Vorstellung und Anwendung verschiedener Ansätze zur Berechnung von Ähnlichkeiten innerhalb von Datenbeständen dieser Domänen. Eine anschließende zusammenfassende Gegenüberstellung der Ansätze zeigt ihre Eignung für eine Anwendung im Rahmen des Gesamtsystemes. Ein zweiter Abschnitt des Kapitels beleuchtet die Untersuchung von

möglichen Korrelationen zwischen Genotyp und Phänotyp innerhalb des integrierten Datenbestandes auf Basis einer Graphenstruktur. Somit wird, basierend auf den verschiedenen verteilten Datenquellen, den klinischen Daten aus Ramedis und mit Hilfe des Integrationsdienstes eine Möglichkeit geschaffen, die Anfragen nach Beziehungen zwischen den angeschlossenen Datenquellen zuläßt.

Im 6. Kapitel werden nun die Ergebnisse der vorangehenden Kapitel zusammengeführt. Dabei wird ein integriertes Schema entwickelt, daß eine einfache Genotyp–Phänotyp–Korrelation erlaubt. Hier wird unter anderem erkennbar, wie sich die verschiedenen Datenquellen verbinden lassen und welche Daten zur Analyse herangezogen werden sollen. Der integrierte Zugriff auf die verschiedenen Datenquellen in Verbindung mit geeigneten Auswertungsmethoden und einer einheitlichen Präsentation erlaubt nun im Rahmen eines Prototypen, der auf Basis der vorgeschlagenen Architektur entwickelt wurde, eine einfache und effiziente Nutzung der vorhandenen Ressourcen. Am Beispiel der Erkrankung Diabetes mellitus MODY 1 wird außerdem die Reichhaltigkeit des integrierten Datenbestandes vorgestellt.

Das abschließende Kapitel 7 wird die Ergebnisse der vorliegenden Arbeit zusammenfassend darstellen. Dazu wird ausgehend von der formulierten Zielstellung eine Übersicht der erreichten Ergebnisse gegeben. Die gesammelten Resultate werden dabei kritisch diskutiert und ein Ausblick auf erforderliche, weiterführende Arbeiten gegeben.

Im Anhang werden ergänzende Informationen bereitgestellt, die das Verständnis der Arbeit oder bestimmter Teile erleichtern sollen. Dazu werden im Anhang A die URLs der in dieser Arbeit vorgestellten und weiterer Life–Science–Datenquellen in einer Übersicht dargestellt. Da neben der Klassifikation von Architekturen und bestehenden Ansätzen nicht vertiefend auf den Integrationsschritt eingegangen wurde, werden im Anhang B einige vertiefende, technische Informationen zu diesem Vorgang bereitgestellt. So wird am Beispiel der Datenbank EMBL die Struktur der Flatfile–Daten vorgestellt und das entsprechende Adapterschema für den Integrationsdienst BioDataServer zugeordnet. Verschiedene Fachbegriffe aus den Bereichen Biologie und Informatik stellt abschließend das Kapitel D vor.

2 Molekularbiologische und informationstechnische Grundlagen

Dieses Kapitel soll die molekularbiologischen und informationstechnischen Grundlagen für die vorliegende Arbeit vermitteln. Da sowohl Informatiker als auch Biologen und Mediziner angesprochen werden und das Verständnis für die Eigenschaften der verschiedenen Fächer in der interdisziplinären Forschung von besonderer Bedeutung ist, wird ein Überblick der wichtigsten Grundlagen präsentiert.

Dazu werden im ersten Abschnitt einige Bemerkungen zu grundlegenden Begriffen aus der Molekularbiologie gemacht, die für ein Verständnis der späteren Betrachtung verschiedener Life–Science–Datenquellen notwendig erscheinen. Anschließend werden ausgewählte Definitionen und Methoden der Informatik vorgestellt, die im weiteren Verlauf der Arbeit angewendet werden. Zur weiteren Vertiefung sei auf die einschlägige Literatur verwiesen.

2.1 Molekularbiologische Grundlagen

Durch die Schlagworte Humangenomprojekt, Gentherapie und Klonen ist die DNS als Bauplan der Organismen seit einigen Jahren vermehrt in das Interesse der Öffentlichkeit gerückt. Trotz enormer Fortschritte auf diesem Gebiet erweisen sich die biochemischen inter– und intrazellulären Vorgänge als sehr komplex und sind somit auch nicht vollständig in elektronischer Form erfaßt.

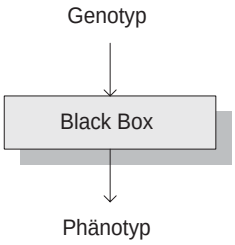

Abbildung 2.1: Vom Genotyp zum Phänotyp

Ohne das vollständige Verständnis des Weges von der genetischen Information in der

DNS (*Genotyp*) zur Manifestation eines Merkmals (*Phänotyp*) werden diese Vorgänge wie in Abbildung 2.1 eine Blackbox bleiben. Die Untersuchung des Zusammenhanges zwischen Sequenz, Struktur und Funktion innerhalb der Zelle und des Organismus entwickelt sich zu einer der wichtigsten Bereiche innerhalb der „post–genomischen“ Phase.

2.1.1 Die DNS als Träger der genetischen Information

Als Träger der genetischen Informationen in den Lebewesen fungieren die Nukleinsäuren. Mit der *Desoxyribonukleinsäure* (DNS) und der *Ribonukleinsäure* (RNS) sind zwei Arten von Nukleinsäuren in den Zellen der Organismen zu finden. Diese Nukleinsäuren sind Ketten aus Nukleotiden. Dabei ist jedes Nukleotid aus drei Bausteinen aufgebaut: der Phosphorsäure, einem Zucker und einem stickstoffhaltigen Ring – der Base. In der DNS treten als Basen *Adenin*, *Cytosin*, *Guanin* und *Thymin* auf; in der RNS wird Thymin durch die Base *Uracil* ersetzt. Die Nukleotide selbst werden über die Phosphorsäure zu Ketten verknüpft, sie verbindet stets die Zuckerbausteine der benachbarten Nukleotide.

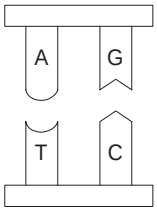

Abbildung 2.2: Basenpaarung

Ein Modell der DNS–Struktur wurde 1953 von WATSON und CRICK [WC53] entwickelt. Sie verknüpften zwei Polynukleotidketten zu einer schraubenartig gedrehten DNS–Doppelhelix, wobei sich die Basen der Nukleotide strickleiterartig paaren. Die vier Basen der DNS ordnen sich gegenüber an, wobei nur Guanin mit Cytosin und nur Adenin mit Thymin miteinander Wasserstoffbrücken ausbilden. Diese Basenpaarung zeigt die Abbildung 2.2. Dadurch sind die beiden Stränge der Helix nicht identisch, sondern komplementär aufgebaut, da durch jede Base des einen Stranges der zugehörige Partner auf dem anderen Strang festgelegt wird. Durch dieses Modell wurde erstmals teilweise verständlich, wie die DNS die genetische Information trägt. Die Abfolge der Nukleotide innerhalb dieser Kette entspricht einem Code, da durch drei aufeinanderfolgende Basen eine bestimmte Aminosäure festgelegt wird. Mit der Veränderung der Sequenz der Nukleotidkette wird dann ebenfalls der genetische Code verändert. Bei der Zuordnung von biochemischen Funktionen zu DNS–Sequenzen wird der Begriff des Genes nun durch *Cistron* ersetzt. Das Cistron bezeichnet eine Nukleotidsequenz, die eine biochemische Funktionseinheit kodiert.

Die Bestimmung der Reihenfolge der Nukleotide der DNS wird für immer längere DNS–Stücke durchgeführt. Das Ergebnis ist eine Sequenz von Basen. Jedoch kann es bei der Sequenzierung zu Mehrdeutigkeiten kommen, die bei der Speicherung gekennzeichnet wer-

Adenin	A
Cytosin	C
Guanin	G
Thymin	T
Uracil	U

Tabelle 2.1: Nukleotidbasen und ihre Ein–Buchstaben-Abkürzungen

den müssen. Die sequenzierten Basen werden über Buchstaben kodiert. Die Abkürzungen der Basen sind in Tabelle 2.1 zu finden.

Veränderungen der genetischen Information werden als Mutationen bezeichnet. Sie entstehen durch Umwelteinflüsse und die Labilität der Bausteine der DNS. Diese Veränderungen des Genoms sind selten und können sich in einigen Fällen als Krankheiten manifestieren, da durch ein verändertes Gen bei der Proteinsynthese fehlerhafte oder funktionslose Proteine gebildet werden können. Nachfolgend werden Arten von Mutationen erläutert; die Abbildung 2.3 illustriert dabei eine Teilmenge, die Chromosomenmutationen. Diese und weitergehende Informationen zur Molekulargenetik geben [Kni97] und [Bro99].

Genom–Mutationen
: Drastische Veränderungen des gesamten Genoms, z.B. Veränderung der Chromosomenzahl

Chromosomen–Mutationen
: Veränderungen der Form und Struktur von Chromosomen

Translokation
: Verlagerung eines Chromosomenstückes von seinem ursprünglichen Ort auf ein anderes Chromosom oder an eine andere Stelle des gleichen Chromosoms

Deletion
: Verlust von Abschnitten eines Chromosoms

Insertion
: Einbau eines DNS–Stückes in ein Chromosom

Inversion
: Verdrehung eines Chromosomenabschnittes um 180 Grad

Gen–Mutationen
: Veränderung der Nukleotidsequenz innerhalb eines Genes

Nukleotid–Austausch
: Veränderung der genetischen Information durch den Austausch eines normalen Nukleotids gegen ein anderes

Leseraster–Mutation
: Veränderung der Nukleotidsequenz in einem Gen durch Addition (Insertion) oder Verlust (Deletion) von Nukleotiden

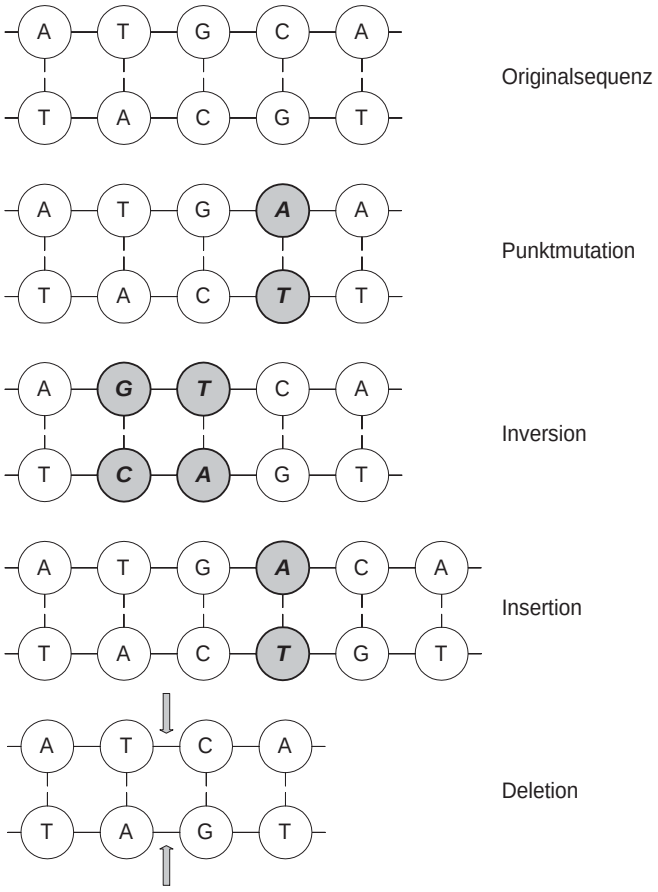

Abbildung 2.3: Grafische Darstellung von Chromosomenmutationen (nach [Bro99])

2.1.2 Von der DNS zum Protein

Alle Stoffwechselvorgänge im Organismus werden durch Enzyme gesteuert. Die meisten Enzyme sind Proteine, die als Makromoleküle aus vielen Einzelbausteinen bestehen – aus den Aminosäuren. Um diese Aminosäuren in der korrekten Weise zusammenbauen zu können, wird die in der DNS gespeicherte genetische Information als Bauplan für die Proteine genutzt. Dazu wird die erforderliche Nukleotidsequenz von der DNS abgeschrieben und auf einen RNS–Strang kopiert. Dieser Vorgang wird als *Transkription* bezeichnet. Entsprechend der RNS–Struktur werden anschließend während der *Translation* die Aminosäuren als Kette zu einem Protein zusammengesetzt.

2.1.2.1 Transkription

Die genetische Information auf der DNS befindet sich im Zellkern, die Synthese der Proteine aber erfolgt an den Ribosomen im Zytoplasma. Daher muß die Information vom Kern in das Zytoplasma übertragen werden. Dazu wird eine Abschrift der Nukleotid–Folge eines Gen–Abschnittes von der DNS auf die *Boten–RNS*[1] (mRNA) angefertigt. Die für diesen Vorgang notwendigen Enzyme heißen *DNS–abhängige*[2] *RNS–Polymerasen*. Die Herstellung eines komplementären RNS–Moleküles zur einem auf dem DNS–Strang befindlichen Sequenzabschnitt wird als Transkription bezeichnet und ist die erste Phase der Genexpression.

Die RNS–Polymerase bindet dabei bevorzugt an Stellen auf der DNS, die vor einem Genanfang liegen. Ebenfalls kommt sie am Ende des Transkriptionsabschnittes zum Stillstand. Somit muß in der Basensequenz der DNS die Information zum Starten und Stoppen der Transkription verschlüsselt sein. Die Erkennungs– oder Bindungsstelle am Anfang des Gens wird als *Promotor* bezeichnet, das Ende als *Terminator*. Eine Transkriptionseinheit ist die Menge der in einem bestimmten RNA–Molekül transkribierten DNA–Sequenzen. Sie beginnt somit am Promotor und endet am Terminator.

2.1.2.2 Translation

Die auf der mRNA vorliegende Basensequenz muß nun in eine Abfolge von Aminosäuren übersetzt werden, die entsprechend ihrer Verknüpfung ein bestimmtes Protein bilden. Dabei wird jeweils eine Sequenz von drei Basen einer Aminosäure zugeordnet. Die Tabelle 2.2 enthält die zwanzig in Proteinen vorkommenden Aminosäuren und ihre Drei– und Ein–Buchstaben–Abkürzungen. Diese Dreierfolge von Nukleotidbasen wird als *Basen–Triplett* oder *Codon*[3] bezeichnet. Bei dieser Kombination ergeben sich jedoch 64 Möglichkeiten zur Bestimmung einer Aminosäure. Da jedoch nur 20 Aminosäuren im Organismus existieren, kann auch eine Aminosäure durch mehrere verschiedene Codons bezeichnet werden. Die Tabelle 2.3 zeigt die Codons und die ihnen zugeordneten Aminosäuren.

Die Aminosäuren werden im Zytoplasma an eine *Transfer–RNS* (tRNA) gebunden. Dazu besitzt die tRNS einen bestimmten Aufbau, so daß nur eine spezifische Aminosäure binden kann. Außerdem tritt an diesem RNS–Typ ein Basen–Triplett auf, das komplementär zum Codon auf der mRNA ist, das die gebundene Aminosäure codiert. Entsprechend wird diese Dreiergruppe auch als *Anticodon* bezeichnet.

Um die auf der mRNA befindliche Basensequenz zu übersetzen, bewegt sich das Ribosom entlang des RNS–Stranges. Dabei werden nun die von der tRNA transportieren Aminosäuren entsprechend dem Codon auf der mRNA und dem Anticodon auf der tRNA zu

[1] Der Ursprung liegt im englischen Begriff 'messenger-RNA'.

[2] Diese genauere Bezeichnung ist nützlich, da auch Polymerasen existieren, die RNS–Sequenzen auf RNS übertragen. So beispielsweise bei manchen Viren, die ihre genetische Information in RNS speichern.

[3] Parallel zu Codon bei der mRNA werden Basen–Tripletts auf der DNS, die die Aminosäuren codieren, als *Codogene* bezeichnet.

Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Asparaginsäure	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutaminsäure	Glu	E
Glycin	Gly	G
Histidin	His	H
Isoleucin	Ile	I
Leucin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptophan	Trp	W
Tyrosin	Thy	Y
Valin	Val	V

Tabelle 2.2: Aminosäuren und ihre Drei– und Ein–Buchstaben–Abkürzungen

einem Protein verknüpft. Der Beginn und das Ende der Translation wird ebenfalls durch *Start–* und *Stop–Codons* gekennzeichnet.

2.1.2.3 Proteine und Enzyme

In vielen biologischen Prozessen spielen Proteine eine wichtige Rolle. Sie wirken als Katalysatoren chemischer Reaktionen und übernehmen im Organismus eine Reihe weiterer Funktionen vom Transport spezifischer Moleküle bis hin zur Immunabwehr. Proteine sind Makromoleküle, die die Fähigkeit besitzen, auf unterschiedlichste Moleküle spezifisch zu reagieren. Nachfolgende Liste soll die verschiedenen Funktionen darstellen.

Enzymatische Katalyse
: Die Enzyme – zu denen auch spezielle Proteine gehören – katalysieren chemische Reaktionen in biologischen Systemen meist durch die millionenfache Erhöhung der Reaktionsgeschwindigkeit. Dabei arbeiten nicht nur Proteine als Enzyme, es existieren auch katalytisch aktive RNS–Moleküle.

Transport und Speicherung
: Durch spezielle Proteine, die Transportproteine, wird der Transport von Molekülen

erste Position	zweite Position				dritte Position
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	*Stop*	*Stop*	A
	Leu	Ser	*Stop*	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lyr	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Tabelle 2.3: Die Codons aus Nukleotidbasen und ihre zugehörigen Aminosäuren

und Ionen realisiert. So dient beispielsweise Hämoglobin als Träger des Sauerstoffs in den Erythrozyten. Speicherproteine speichern Aminosäuren und andere Substanzen für den zukünftigen Gebrauch durch den Organismus, beispielsweise Ferritin, das Eisen in der Leber speichert.

Koordinierte Bewegung

Die Kontraktion des Muskelgewebes durch eine gleitende Bewegung zweier Arten von Proteinfilamenten (Kontraktile Proteine: Aktin und Myosin in Muskeln) ermöglichen es den Organismen, sich zu bewegen.

Mechanische Stützfunktion

Die Strukturproteine sind Teil des stützenden Gerüstes der Organismen, z.B. Kollagen in Sehnen, Knochen und Knorpel der Wirbeltiere. Sie gewährleisten die Zugfestigkeit dieser Gewebe.

Immunabwehr

Antikörper sind ebenfalls spezifische Proteine, die Fremdsubstanzen erkennen und binden. Diese Schutzproteine schützen gegen Krankheitserreger und bei Verletzungen.

Erzeugung und Übertragung von Nervenimpulsen

Bei der Übermittlung der neurotransmittervermittelten Antwort von Nervenzellen sind Rezeptorproteine beteiligt.

Kontrolle von Wachstum und Differenzierung

Bestimmte Proteine greifen regulierend in Wachstums– und Differenzierungsprozesse ein. Dabei spielt die kontrollierte, zeitlich abgestimmte Expression der genetischen Information eine wichtige Rolle.

Diese vielfältige Funktionalität ergibt sich aus der Ausbildung der Proteine in unterschiedlichen dreidimensionalen Strukturen, die jeweils verschiedene Moleküle binden können. Die einfache Reihenfolge oder Sequenz der Aminosäuren eines Proteins wird als *Primärstruktur* bezeichnet. Wie bereits im vorangehenden Abschnitt dargelegt, sind diese Aminosäuren die elementaren Struktureinheiten der Proteine. Dabei spezifiziert die Sequenz der Nukleotide in der DNS eine komplementäre Sequenz von Nukleotiden der RNS, die wiederum die Aminosäuresequenz des Proteins bestimmt. Jedes Protein verfügt also über eine einzigartige, wohldefinierte Abfolge von Aminosäuren, die genetisch festgelegt ist.

Durch die Verknüpfung der existierenden Aminosäuren durch Peptidbindungen entstehen Polypeptide. Ihre Länge liegt meistens zwischen 100 und 800 Bausteinen. Sequenzen mit weniger als 20 Aminosäure–Bausteinen heißen *Peptide*. Für die Funktion eines Proteins ist jedoch nicht nur die Reihenfolge der Aminosäuren verantwortlich. Jeder Protein faltet und dreht sich außerdem in einer charakteristischen Form. Diese Phänomene werden als Ausbildung der *Sekundär–* und *Tertiärstruktur* bezeichnet.

Die Enzyme sind meistens Proteine, die als Katalysatoren in biologischen Systemen wirken. Sie sind in ihrer Funktion spezifisch und besitzen in ihrer Wirkung eine hohe katalytische Aktivität. Damit können sie die Geschwindigkeit einer biochemische Reaktion um den Faktor 10^6 erhöhen.

2.1.2.4 Genregulation

Zur Anpassung des Organismus an verschiedene Umweltbedingungen und Einflüsse wird die Menge aller Gene nicht ständig exprimiert. Vielmehr sind viele Gene mit speziellen Aufgaben inaktiv und werden erst bei Bedarf angeschaltet. Außerdem wird die Geschwindigkeit der Genexpression zur Kontrolle der Menge der Genprodukte in der Zelle reguliert, um etwa auf die Veränderung von Nährstoffen zu reagieren. Die Gene jedoch, die ständig zur Aufrechterhaltung der Zellfunktion benötigt werden, nennt man *konstitutive Gene*.

Die Kontrolle der Transkription ist ein wesentliches Element der Regulierung der Genaktivität. Diese Regulation eines Genes erfolgt mittels kurzer, regulatorischer Bereiche auf der DNS als Bindungsstellen für eine bestimmte Klasse von Proteinen — den Transkriptionsfaktoren. Durch die Wechselwirkung zwischen diesen Proteinen und den Transkriptionsfaktorbindungsstellen als definierte Nukleotidsequenzen, besteht die Möglichkeit, die Aktivität der Expression eines Genes zu fördern oder zu hemmen. Dennoch wird bisher die Genregulation auf der Ebene der Transkription nur teilweise verstanden, insbesondere die Kinetik dieses Vorganges.

2.1.3 Metabolismus und Stoffwechselwege

Zellen gewinnen Energie aus ihrer Umgebung und wandeln Nährstoffe durch viele miteinander verbundene chemische Reaktionen in Zellkomponenten um. Die Menge dieser biochemischen Prozesse in der Zelle wird als Stoffwechsel oder *Metabolismus* bezeichnet. Dazu gehört im wesentlichen die Proteinsynthese, die *Biosynthese* und die Zellkommunikation. Der Begriff der Proteinsynthese wurde bereits im Abschnitt 2.1.2 erläutert – er bezeichnet den Übersetzungsprozeß von der genetischen Information zum Protein. Unter Biosynthese werden alle enzymatisch gesteuerten biochemischen Reaktionen zusammengefaßt. Der Stoffwechsel wird durch viele Mechanismen reguliert. Beispielsweise können die Mengen einiger entscheidender Enzyme durch Regulation ihrer Biosynthese– und Abbauraten gesteuert werden.

Die meisten zentralen Moleküle des Stoffwechsels sind für alle Lebensformen identisch. Zudem sind viele Stoffwechselmuster in Bakterien, Pflanzen und Tiere weitgehend gleich. Diese in Wechselwirkung stehenden biochemischen Reaktionen der Biosynthese werden in der Literatur mit dem Begriff *Metabolic Pathways* (Stoffwechselwege) bezeichnet [Mav90, Mic99]. Diese Vorgänge werden durch die in der Proteinsynthese hergestellten Enzyme gesteuert. Dabei werden Zwischenprodukte produziert und konsumiert, die Synthese von Enzymen gefördert und gehemmt.

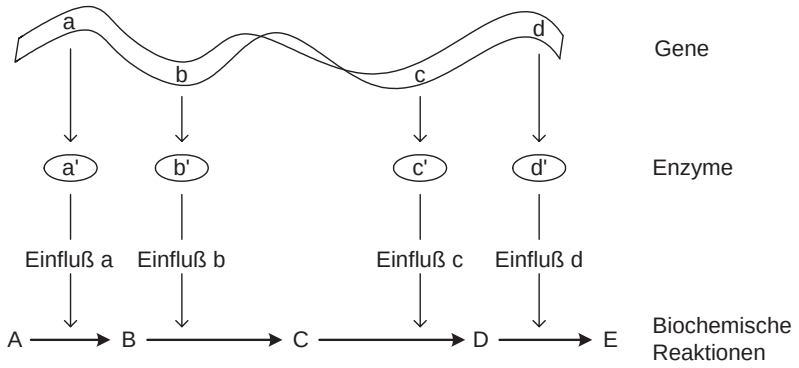

Abbildung 2.4: Steuerung einer Reaktionskette im Stoffwechsel (nach [Ste87])

Die Abbildung 2.4 soll diese Abläufe vereinfacht verdeutlichen. Die in den Genen a, .., d gespeicherten Informationen werden während der Transkription kopiert und durch die Translation übersetzt, so daß die entsprechenden Proteine synthetisiert werden können. Hier wirken sie als Enzyme a', .., d' auf biochemische Reaktionen, die als Reaktionskette, beispielsweise durch Konsumption von Reaktionsprodukten vorangehender Reaktionen, verbunden sind. Die Substanzen in dieser Folge von biochemische Reaktionen wurden mit den Großbuchstaben A, B, .., E bezeichnet, wobei A für den Ausgangsstoff, B, C und D für Zwischenprodukte und E für das Reaktionsprodukt diese Abfolge steht.

2.1.4 Erbkrankheiten und Stoffwechseldefekte

Die in den vorangestellten Abschnitten vorgestellten Mechanismen der Übersetzung der genetischen Information auf der DNS bis hin zur Manifestation eines Merkmals im Phänotyp sind für alle Lebewesen essentiell. Durch eine funktionelle Störung in den Zellen, Geweben und Organen des Körpers aufgrund veränderter biochemischer Reaktionen der unterschiedlichsten Art kann eine Erkrankung hervorgerufen werden. Sie entsteht dabei durch verschiedene Einflüsse: äußere Faktoren (*Exposition*), zeitweilige und unterschiedliche Anfälligkeiten bzw. Empfänglichkeiten (*Disposition*) oder durch eine spezielle, oftmals vererbte Gesamtveranlagung (*Konstitution*) [Ste87].

Als spezielle Gruppe von Krankheiten werden die Erbkrankheiten betrachtet. Sie werden durch krankhafte Veränderungen im menschlichen Genom verursacht. Diese treten dann wiederholt entsprechend bestimmten Regeln bei Vorfahren und Nachkommen des Erkrankten auf. Dabei wird die falsche genetische Information als defekte Erbanlage an die Kinder weitergegeben. Erkrankungen, die auf den Einfluß der Gene zurückgeführt werden können, wurden in der nachfolgenden Übersicht nach ihrer Ursache unterteilt:

Genbedingte Erkrankung
: wird durch die Wirkung spezifischer Erbanlagen oder Gene verursacht,

Chromosomenbedingte Erkrankung
: deren Ursache liegt in einer strukturellen oder zahlenmäßigen Anomalie der Chromosomen, wie z.B. beim Down–Syndrom,

Geninkompatibilität
: durch die Unverträglichkeit bestimmter phänotypischer Merkmale von an sich normalen Erbanlagen bei Mutter und Kind (z.B. fetale Erythroblastose durch Blutgruppenunverträglichkeit).

Vererbte Erkrankungen sind jedoch nicht mit angeborenen Krankheiten gleichzusetzen. Während Vererbung die Weitergabe eines genetischen Defektes von Generation zu Generation beinhaltet, umfaßt der Begriff 'angeboren' lediglich den Nachweis eines bestimmten Merkmales bei der Geburt. Eine Aussage über die Ursache oder den Zeitpunkt der Entstehung wird nicht getroffen. Beispielhaft seien folgende vererbbare Entwicklungsstörungen und Anomalien genannt:

- Fehlbildungen und Erkrankungen des Skeletts,
- Erkrankungen des Blutes, des Herzens und der Gefäße,
- Erkrankungen der Verdauungsorgane, der Atmungsorgane und der Ausscheidungsorgane,
- Erkrankungen der Sinnesorgane,

- Nerven– und Muskelerkrankungen,
- Stoffwechselkrankheiten.

Eine der häufigsten angeborenen Stoffwechselerkrankungen ist der Ornithin–Transcarbamylase–Mangel. Diese Erkrankung tritt auf, wenn eine Mutation des Genes besteht, das den Bauplan für das Enzym Ornithin–Transcarbamylase (OTC) darstellt. Dadurch kann das Enzym OTC nicht mehr korrekt synthetisiert werden, so daß es zu einer Fehlfunktion in der Biosynthese kommt. Dieser Mangel manifestiert sich im Harnstoffzyklus, in dem das ausgefallene Enzym die Reaktion von Ornithin zu Citrullin katalysiert. Somit kann im Falle eines Gendefektes diese Reaktion nicht mehr oder nur in einem beschränktem Umfang ausgeführt werden, da eine abgestufte klinische Ausprägung möglich ist.

Die Patienten mit OTC–Mangel werden häufig mit einer speziellen Diät behandelt oder müssen sich der Dialyse unterziehen, um die Anhäufung nicht abgebauter Stoffwechselprodukte im Organismus zu verhindern. Eine prinzipiell mögliche Behandlungsmethode wäre aber auch die Aktivierung eines alternativen Stoffwechselweges (*alternativer Metabolic Pathway*). Eine solche Alternative könnte die blockierte oder reduzierte Reaktion überbrücken und den Abbau und die Ausscheidung der Stoffwechselprodukte unterstützen.

2.1.5 Wirkstoffpointing

Um neue Therapien für Stoffwechselerkrankungen zu entwickeln und um verschiedene Therapieformen gegeneinander abwägen zu können, ist es wichtig zu wissen, wo und wie Medikamente im Organismus wirken. Dieses Wissen über die Angriffspunkte und die Wirkungsweise pharmakologischer Substanzen wird als *Wirkstoffpointing* (*drug pointing*) bezeichnet [HMPS99].

Medikamente sind im allgemeinen als biologisch aktive Substanzen oder Substanzmischungen anzusehen, die im menschlichen Organismus bestimmte Wirkungen haben. Zu einem Wirkstoff (*Agent*) gehört immer ein entsprechendes Zielmolekül (*Target*), auf das der Wirkstoff einwirkt. Diese Wechselwirkung von Agent und Target beruht auf dem Schlüssel–Schloß–Prinzip. Ein Wirkstoff kann dabei im Idealfall mit nur einem Zielmolekül wechselwirken, weil dieser wie ein Schlüssel ins Schloß paßt. Das Target wird nun durch den Einfluß des Agenten entweder aktiviert oder passiviert. Dadurch können weitere Moleküle beeinflußt werden, die dann im Endeffekt eine physiologischen Reaktion bewirken können. Diesen prinzipiellen Vorgang zeigt die Abbildung 2.5, wobei Wirkstoffe z.B. auf Gene oder andere Moleküle wirken können. Jedoch gibt es eine Reihe von Abweichungen von diesem Prinzip, da Fragen der Pharmakogenetik und Pharmakokinetik hier vernachlässigt wurden.

Zu den Targets für Wirkstoffe gehören vor allem Proteine, darunter insbesondere Enzyme, Ionenkanäle, Transportermoleküle und Rezeptoren. Die meisten der altbekannten Wirkstoffe sind hier einzuordnen, beispielsweise die Acetylsalicylsäure als Enzyminhibitor,

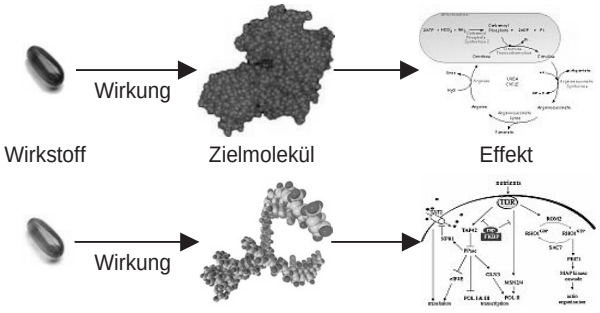

Abbildung 2.5: Pharmakologisches Prinzip des Wirkstoffpointings

die auch als Bestandteil des Medikamentes Aspirin bekannt ist. Eine umfassende Darstellung der Wirkungsweisen von Medikamenten, des Wirkstoffdesigns und weiterer Themen der Pharmakologie geben [BKK96], [Lül99] und [Mut96].

2.2 Informationstechnische Grundlagen

Dieser Abschnitt wird einige Begriffe im Bereich der Datenbank– und Informationssysteme definieren. Außerdem wird ein Einblick in die Prinzipien des fallbasierten Schließens gegeben. In diesem Zusammenhang werden ebenfalls einfache Verfahren zur Ähnlichkeitsbewertung vorgestellt.

Eine Vorstellung der Grundlagen ist erforderlich, da diese Verfahren genutzt werden, um die Aufgabenstellung zu bearbeiten. Dabei werden beispielsweise Datenbank– und Informationssysteme zur Gewinnung, Speicherung, Umformung, zum Transport und zur Darstellung der vorhandenen medizinischen und molekularbiologischen Daten genutzt. Bestimmte Verfahren werden im weiteren Verlauf der Arbeit angewendet.

2.2.1 Informationssysteme und Relationale Datenbanksysteme

Die Grundlage für die Untersuchungen von Beziehungen zwischen Genotyp und Phänotyp bilden die zu erfassenden oder zu integrierenden molekularbiologischen und medizinischen Daten, die die erforderlichen Informationen von der Nukleotidsequenz bis zu den klinischen Merkmalen eines Patienten zur Verfügung stellen. Bevor jedoch einige dieser Datenquellen im Kapitel 3 vorgestellt werden, muß eine Klärung des Begriffes Datenquelle erfolgen. Außerdem sollen grundlegende Begriffe im Bereich der Datenbank– und Informationssysteme geklärt werden.

Definition 2.1 (Datenquelle) *Eine Datenquelle besteht aus mindestens einem Computer (rechentechnische Einheit), auf dem Daten in einem definierten Format gespeichert sind und auf die über bestimmte Schnittstellen zugegriffen werden kann. (nach [Sch02])*

Ein Teil der verfügbaren Datenquellen sind die sogenannten Flat–Files, die in der molekularbiologischen Praxis heute noch benutzt werden und zu einer Zeit entstanden, als das Datenvolumen noch minimal war. Flat–Files sind strukturierte Dateien im ASCII–Format, in denen die Datensätze sequentiell abgelegt und deren Datenfelder durch Schlüsselworte voreinander getrennt werden. Die meisten Datenquellen in diesem Anwendungsgebiet sind jedoch Datenbanksysteme und Informationssysteme, die angelegt wurden, um molekularbiologische Daten zu speichern und für eine Weiterverwendung rechentechnisch verfügbar zu machen.

Definition 2.2 (Datenbank) *Eine Datenbank ist eine strukturierte Sammlung von Daten, welche Fakten über spezielle Anwendungen eines modellierten Ausschnittes der Realwelt repräsentiert, die persistent und weitgehend redundanzfrei gespeichert werden. (nach [Sch02])*

Die Bezeichnung *Informationssystem* (IS) wird heute häufig für eine Vielzahl von Anwendungen benutzt. Dabei ist jedoch die Struktur, der Umfang und die Funktionalität der bezeichneten Systeme vielfach sehr unterschiedlich. Auch werden in der biologischen und medizinischen Praxis die Begriffe Datenbank und Informationssystem häufig synonym verwendet. Um einen Eindruck von der Komplexität eines IS zu geben, soll dieser Begriff im folgenden vorgestellt und einige Eigenschaften eines Informationssystems umrissen werden.

Definition 2.3 (Informationssystem) *Ein Informationssystem bezeichnet ein komplexes, zusammengesetztes Softwaresystem mit aufeinander bezogenen informationsverarbeitenden Operationen. Diese können in Gewinnung, Speicherung, Umformung, Transport und Darstellung gegliedert werden. (nach [Saa93])*

Ein Informationssystem besitzt nach [Saa93] weiterhin meist folgende Eigenschaften:

- Das IS realisiert eine dauerhafte (persistente) Speicherung von Daten. Dabei ist eine Datenbank oder ein Datenbanksystem als Teilmenge des IS anzusehen. Die resultierenden Informationen werden durch die Verknüpfung der Daten aus der Datenbank mit geeigneten Methoden und Interpretationen gewonnen. Diese Resultate können durch eine Wiederholung von Anfragen über dem gleichen Datenbestand beliebig oft wiedergewonnen werden.
- Das IS wertet die gespeicherten Daten anwendungsspezifisch aus.
- Ein IS ist durch Anpassungen und Erweiterungen dynamisch. Der Zustand des Systems kann durch Änderungen an der Datenbank, Regeln und Metadaten angepaßt und verändert werden.
- Das IS integriert weitere (externe) Informationsquellen. Das können Datenbanken oder andere Quellen (Funkzeit, GPS–Daten, Sensorendaten) sein.

Diese Eigenschaften zeigen, daß ein Informationssystem durch sehr viele, zum Teil unabhängige Aspekte charakterisiert wird. Es wird typischerweise für eine Nutzung über einen langen Zeitraum entworfen und unterliegt auch während der Nutzung ständigen Wachstums– und Aktualisierungsprozessen.

2.2.2 Fallbasiertes Schließen

In den meisten Expertensystemen wird das gespeicherte Wissen durch Regeln, Frames oder Klauseln formalisiert. Dadurch wird ein schwieriger und lang andauernder Prozeß der Wissensakquisition notwendig, da das zu erfassende Expertenwissen kaum in der entsprechenden formalisierten Form vorhanden ist. Vielmehr gewinnt ein Fachexperte nach [Goo96] seine fachspezifischen Erfahrungen durch den langjährigen Umgang mit ähnlichen Problemstellungen. Dabei merkt er sich die resultierenden Erkenntnisse, die anzuwendenden Verfahren und Methoden insbesondere im Kontext gelöster Aufgabenstellungen. Somit entsteht zwangsläufig während der notwendige Strukturierung dieses Wissens und seiner Formalisierung im Rahmen der vorgegebenen Wissensrepräsentation ein Verlust dieses episodischen Erfahrungswissens durch seine Transformation.

Durch die Wiederverwendung des Erfahrungswissens von Fachexperten zur Bearbeitung zukünftiger Probleme und die Ergänzung durch das Hinzufügen damit gelöster neuer Problemstellungen soll beim fallbasierten Schließen (*Case–based Reasoning*, CBR) die rechnerunterstützte Nutzung von episodischem Erfahrungswissen ermöglicht werden. Erste Wurzeln diese Ansatzes sind bei [Sch82] zu finden. Aus der Kognitionspsychologie wurde eine Theorie zum Verstehen, Erinnern und Lernen gemachter Erfahrungen entwickelt. Unter der Bezeichnung *Dynamic Memory* wurde die Anlehnung der Methode an das menschliche Gedächtnis beschrieben. Dabei wird eine dynamische Angleichnung der internen Struktur an neue Verhältnisse und leichtes Lernen aus neuer Erfahrungen betont. In [Kol83] wurde dann mit dem CYRUS–System eine Architektur vorgestellt, die ein episodische Gedächtnis mit der entsprechenden Abrufstrategie verband. Der besondere Vorteil dieser Anwendung wurde durch die Vereinfachnung des zeitintensiven Wissensaquisitionsschrittes erreicht, da häufig bereits große Mengen an Falldaten, beispielsweise in Krankenakten, vorhanden sind, die direkt in das System einfließen können.

Die Nutzung des CBR–Ansatzes verspricht nach [Wat95] durch die Suche und Nutzung existierender Problemlösungen folgende Vorteile:

- Das CBR–System benötigt kein explizites Domänenmodell und der bis dahin notwendige Prozeß zur Wissensaquisition reduziert sich auf die Sammlung von Fallberichten. Existierende Datenbestände, die beispielsweise elektronisch als Patientenakten gespeichert sind, können mit geringem Aufwand direkt in das CBR–System integriert werden.

- Der Implementations–Prozeß eines CBR–Systems beschränkt sich weitgehend auf die Identifikation signifikanter Merkmale, die einen Fall beschreiben.

- Die Anwendung von Datenbanktechniken erlaubt es einem CBR-System, große Mengen von Daten zu halten und zu verwalten.
- Die Wartung und Pflege eines CBR–Systemes vereinfacht sich, da neues Wissen in Form von neuen Fallberichten zum existierenden Datenbestand hinzugefügt wird.

Wie bereits beschrieben, basiert der CBR–Ansatz auf gespeicherten Fällen, die bereits gelöste Problemsituationen beschreiben. Neue Problemstellungen werden nun durch die Suche nach ähnlichen Situationen in der Fallbasis und ihrer Adaption auf das neue Problem gelöst. Dieser Prozeß wurde durch [AP94] als Zyklus mit vier Schritten beschrieben. Bevor jedoch näher auf dieses Vorgehensmodell eingegangen wird, sollen einige Begriffe und Definitionen erläutert werden, die für das Verständnis des CBR–Ansatzes notwendig sind. Dabei wurde sich an der Terminologie von [Goo96] orientiert.

Definition 2.4 (Problemstellung) *Eine Problemstellung besteht aus einer Menge von 2–Tupeln aus Merkmal und Merkmalausprägung. Alle nicht in dieser Menge enthaltenen Merkmale sind in der Problemstellung nicht erfaßt.*

Die Problemstellung ist dabei Teil des zu untersuchenden Klassifikationsproblemes. Sie wird in diesem Zusammenhang durch die Problemlösung ergänzt. Ein Fall (*case*) wird somit durch zwei endliche, disjunkte Mengen von Problemmerkmalen und –lösungen beschrieben. Eine neue Problemsituation wird als neuer Fall (*new case, unsolved case*) bezeichnet. Innerhalb einer Problemstellung werden beobachtete Eigenschaften, Symptome oder Fragen gesammelt. Diese werden als Merkmale bezeichnet.

Definition 2.5 (Merkmal) *Ein Merkmal beschreibt einen erfaßbaren Wert eines zu untersuchenden Systems. Der zu einem Merkmal erfaßte Wert heißt Ausprägung des Merkmals.*

Die Ausprägung eines Merkmales ist natürlich abhängig von der Art der Beantwortung — dem Datentyp der Antwort. Sie kann beispielsweise ein boolscher oder numerischer Wert sein. Jedoch müssen nicht alle Merkmale abgefragt werden, da sogenannte Merkmalsinterpretationen auch innerhalb eines Schrittes zur Vorverarbeitung der verfügbaren Daten aus anderen, erfaßten Merkmalen hergeleitet werden können. Die Lösung eines Problemes kann im medizinischen Sprachgebrauch auch durch den Begriff Diagnose umschrieben werden.

Definition 2.6 (Lösung) *Eine Lösung wird durch ihre Bezeichnung beschrieben. Sobald sie im Rahmen einer aktuellen Problembeschreibung zur Kandidatenmenge der Problemlösungen gehört, wird sie als 'etabliert' bezeichnet.*

Eine bekannte Problemstellung und ihre korrespondierende Lösung werden in der Fallbasis (*case base*) des CBR–Systemes gespeichert. Diese Fallbasis (Falldatenbank) besteht aus einer typischerweise sehr großen Menge von Fällen.

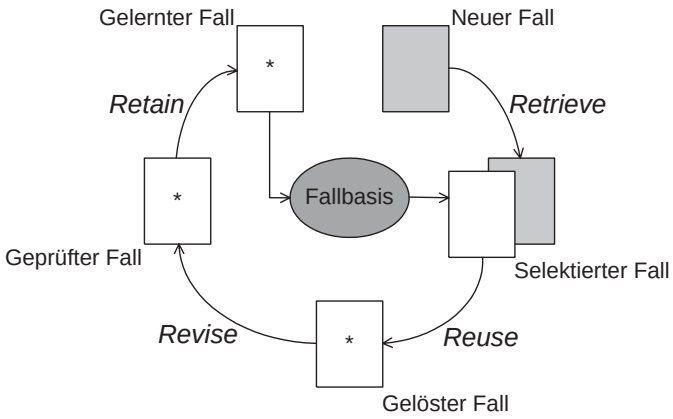

Abbildung 2.6: Prozeßmodell des CBR–Zyklus (nach [AP94])

In der Abbildung 2.6 ist der CBR–Zyklus nach [AP94] dargestellt. Er besteht aus den nachfolgenden vier Teilschritten.

1. **Retrieve** (Abrufen)
   Die neue Problemsituation wird durch einen *neuen Fall* beschrieben und in das CBR–System importiert. Durch Zugriff auf die Fallbasis wird eine Menge von bereits gelösten Fällen selektiert, indem sie durch ein geeignetes Ähnlichkeitsmaß bewertet werden.

2. **Reuse** (Wiederverwenden)
   Durch die Kombination der ähnlichen Fälle mit dem neuen Fall entsteht ein *gelöster Fall*, der einen Lösungsvorschlag für die aktuelle Problemstellung enthält.

3. **Revise** (Überprüfen)
   Die vorgeschlagene Lösung innerhalb des gelösten Falles wird auf ihre Eignung im Bezug auf die aktuelle Problemsituation geprüft. Daraufhin wird gegebenenfalls die vorgeschlagene Lösung noch korrigiert. Resultat dieses Schrittes ist ein *getesteter Fall*.

4. **Retain** (Speichern)
   Damit die gesammelten Erfahrungen für eine Nutzung in der Zukunft zur Verfügung stehen, kann der getestete Fall als *gelernter Fall* in die Falldatenbank eingearbeitet werden oder exisitierende Fälle werden modifiziert.

In der Praxis werden jedoch nicht alle Schritte des CBR–Zyklus durchlaufen und auf eine Problemsituation angewendet. Viele Anwendungen unterstützen den Nutzer beispielsweise nur beim Case Retrieval und Case Reuse. Diese abgestufte Verwendung der CBR–Schritte wird grundsätzlich in zwei Ansätze unterteilt.

Fallvergleichendes fallbasiertes Schließen (Case–Match CBR)
: Suche nach dem ähnlichsten Fall und Entscheidung, ob die Lösung dieses Falles auf die aktuelle Situation übertragen werden kann

Falladaptierendes fallbasiertes Schließen (Case–Adaption CBR)
: Angleichung der teilweise passenden Lösung des ähnlichsten Falles mit zusätzlichem Domänenwissen an die aktuelle Problemstellung

Dieser vierstufige CBR–Zyklus wurde durch verschiedene Zwischenschritte verfeinert. Dabei können sich einzelne Schritte innerhalb des Prozeßmodelles überlappen.

1. Vorauswahl einer Auswahl von geeigneten Fällen (potentielle Lösungskandidaten) aus der Falldatenbank als Grundlage für die weiteren Schritte
2. Ähnlichkeitsbewertung der vorausgewählten Fälle
3. Lösungstransfer
4. Testen und Kritisieren der gefundenen Lösung
5. Evaluation der Ergebnisse
6. Lernen

Die unterschiedlichen Schritte des CBR–Zyklus werden durch spezielle Algorithmen und Methoden unterstützt. So kann beispielsweise die Retrieve–Phase durch eine geeignete Indexierung der gespeicherten Fälle innerhalb der Fallbasis verbessert werden. Von besonderer Bedeutung sind außerdem Methoden zur Auswahl des ähnlichsten Falles innerhalb einer angemessener Zeitspanne. Während des Lösungstransfers ist eine allgemeine Vorgehensweise zur Interpretation, Adaption, Kombination der Lösung des oder der ähnlichsten Fälle wichtig.

2.2.3 Ähnlichkeitsbewertung

Häufig reicht die Untersuchung auf das Vorliegen von Identität oder Gleichheit von zwei Objekten nicht aus, um eine entsprechende Fragestellung adäquat zu beantworten. Deshalb wird auf Verfahren zurückgegriffen, die nach bestimmten Gesichtspunkten die Ähnlichkeit dieser beiden zu untersuchenden Objekte berechnen. Dabei haben die verschiedenen verwendeten Ähnlichkeitskonzepte Einfluß auf die möglichen Formalisierungen und Auswirkungen auf die Datenanalyse. Unabhängig von dem derzeitigen Verständnis des Ähnlichkeitsbegriffes und der Veränderung seiner Betrachtung innerhalb zurückliegender Epochen, wurde nach [Len96] die Ähnlichkeit als ein erklärendes und ordnendes Prinzip für die Betrachtung der Welt angesehen.

Die Definition eines allgemeinen Ähnlichkeitsmaßes ist sehr schwierig. Vielmehr ist anhand von Wissen über die zu untersuchende Fachdomäne nun eine analytische Methode zu definieren, die zur Untersuchung der Ähnlichkeit von Objekten aus dieser Domäne geeignet ist. So wird in der Definition der Ähnlichkeit von Dreiecken in der Geometrie entweder die Übereinstimmung aller Winkelgrößen oder aller entsprechenden Streckenlängenverhältnisse ähnlicher Figuren gefordert. Eine weitergehende mathematische Formalisierung hingegen betrachtet die zu untersuchenden Objekte in einer Repräsentation von Punkten in einem n–dimensionalen Merkmalsraum, in dem jede Dimension durch ein Merkmal aufgespannt wird. Auf diesem wird dann ein geeignetes Maß definiert, das den Abstand der Punkte als Unähnlichkeit der durch sie repräsentierten Objekte wiedergibt. Ein Beispiel dafür ist der Abstand nach Euklid in der Definition 2.7.

Definition 2.7 (Abstand nach Euklid)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Bei der Defintion von Ähnlichkeitsmaßen hat sich nach [Goo96] gezeigt, daß zur Steigerung der Ausdrucksstärke möglichst viel Anwendungswissen modellierbar sein soll. Dies führt jedoch gleichermaßen zu einer entsprechend höheren Komplexität und Aufwendigkeit der Vergleiche. Eine Verfeinerung der Berechnung des euklidschen Abstandes durch Gewichte für die einzelnen Merkmale führt zur Definition 2.8.

Definition 2.8 (Gewichteter Abstand nach Euklid)

$$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i(x_i - y_i)^2} \qquad mit \quad \sum_{i=1}^{n} w_i = 1, w_i \geq 0$$

Ein weiteres Maß für die Ähnlichkeit von Merkmalen, die durch zwei Mengen A und B charakterisiert sind, ist zum Beispiel der Jaccard-Koeffizient. Dieser nimmt seinen niedrigsten Wert dann an, wenn A und B disjunkt sind und erreicht seinen höchsten Wert wenn A und B die gleichen Elemente enthalten. Dieses Verfahren wird häufig in der Dokumentenverarbeitung eingesetzt.

2.3 Zusammenfassung

In diesem Kapitel wurden verschiedene Grundlagen aus zwei Fachgebieten vorgestellt — der Biologie und der Informatik. Dies war einerseits notwendig, weil die im späteren Verlauf der Arbeit benutzten Life–Science–Datenquellen in großem Umfang molekularbiologische und medizinische Daten von hoher Spezifität liefern. Um diese für eine Datenintegration vorzubereiten und die Zusammenhänge innerhalb der Datenbestände und

zwischen den verteilten Datenbeständen zu durchschauen, ist ein grundlegendes Verständnis der biologischen Zusammenhänge unverzichtbar. Andererseits werden natürlich Methoden der Informatik genutzt, um diese Datenbestände beispielweise zu speichern, zu verarbeiten und dem Nutzer zugänglich zu machen.

Dazu wurden grundlegende Bemerkungen zur DNS als Träger der genetischen Information in den Organismen und ihre Nutzung zur Synthese von Proteinen gemacht. Die Bedeutung von speziellen Proteinen, den Enzymen, innerhalb des Metabolismus und der biochemischen Reaktionen verdeutlichte die Erläuterung der Entstehung von Erbkrankheiten und Stoffwechseldefekten. Im Bereich der Informatik wurde die elementare Bedeutung von Datenbanksystemen für die Speicherung von Daten hervorgehoben, die dann die Grundlage für Informationssysteme bilden. Zur Analyse der integrierten Daten wurde der Ansatz des fallbasierten Schließens vorgestellt, der bereits in vielen medizinischen Expertensystemen angewendet wird und durch allgemeine Betrachtungen zur Ähnlichkeitsbewertung von Objekten ergänzt wird.

3 Analyse von Datenquellen und Integrationsansätzen

Im Anschluß an die Einführung in die Grundlagen aus Sicht von Biologie und Informatik werden nun die erforderlichen Vorbetrachtungen zur Sichtung und Integration relevanter Daten angestellt. Diese Daten sind in einer Vielzahl von verteilten Datenquellen gespeichert, die sich neben den spezifischen Aspekten der betrachteten biologischen Systeme auch durch unterschiedliche interne Repräsentationen, verwendete Systeme und Plattformen, Zugriffsmöglichkeiten und Komplexität voneinander unterscheiden.

Um eine Integration von medizinischen und molekularbiologischen Datenquellen durchzuführen, ist somit in einem ersten Schritt die Menge der verfügbaren Datenquellen zu strukturieren und zu bewerten, so daß durch geeignete Bewertungsmaßstäbe in einer zusammenfassenden Gegenüberstellung eine Auswahl ermöglicht wird. Dabei werden jedoch nur Datenquellen angesprochen, die für die Zielstellung dieser Arbeit eine besondere Bedeutung haben.

Im Anschluß werden verschiedene Architekturen zur Integration von Daten vorgestellt. Dabei werden auch die Besonderheiten dieser speziellen Datenquellen betrachtet, da Life–Science–Datenquellen beispielsweise besonders komplexe Schemata besitzen, die sich auch häufig ändern können. Außerdem erweitert sich der Datenbestand dieser Quellen kontinuierlich. Bereits vorhandene Integrationsansätze werden den vorgestellten Architekturen zugeordnet und anhand von bestimmten Kriterien bewertet, um ihre Eignung abzuschätzen.

3.1 Verfügbare Datenquellen

In diesem Kapitel sollen diejenigen Life–Science–Datenquellen vorgestellt werden, die für die vorliegende Fragestellung besonders interessant sind. Dazu werden sie zuerst einzeln vorgestellt und anschließend in einer zusammenfassenden Gegenüberstellung dargestellt. Eine weitergehende Vertiefung kann im Rahmen dieser Arbeit nicht gegeben werden, da die Menge der verfügbaren Datenbanken und Informationssysteme ständig steigt und für fast jede Spezialisierung der Biologie, Molekularbiologie und Biochemie spezielle Datenquellen angelegt werden. Um diesen Sachverhalt zu verdeutlichen, zeigt die Abbildung 3.1 einen Ausschnitt von wichtigen Informationen, die allein mit einem Gen assoziiert werden können.

Seit durch neue biotechnologische Methoden große Datenmengen innerhalb kurzer Zeiträume erzeugt werden können, ist dieses Datenaufkommen nur noch durch eine automatisierte, informationstechnische Aufbereitung, Speicherung und Analyse effizient nutzbar. Der Einsatz der Rechentechnik beschränkt sich dabei nicht nur auf die Verwendung von Datenbankmanagementsystemen. Viele weitere Forschungsgebiete im Bereich der Molekularbiologie, Biotechnologie, Pharmazie und Medizin sind ohne den Einsatz von Informationstechnik nicht denkbar.

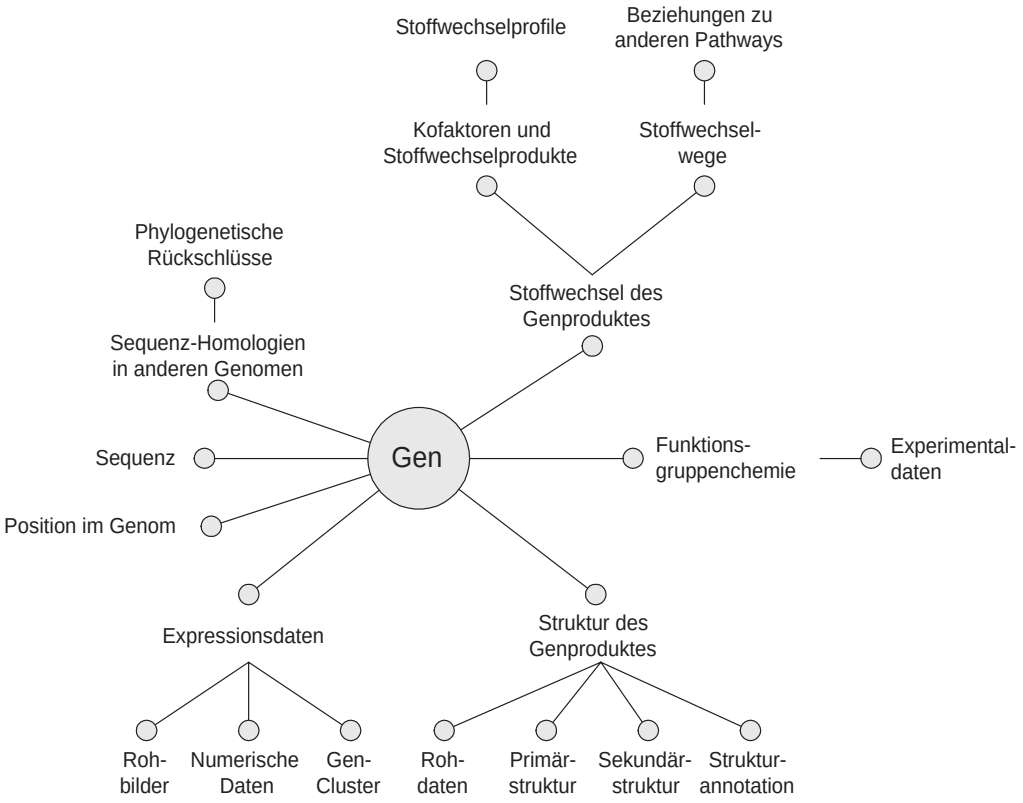

Abbildung 3.1: Wissenswerte Informationen zu einem Gen (nach [GJ02])

Ein jährlich erscheinendes Sonderheft der *Nucleic Acids Research* gibt einen sehr guten Überblick über die wichtigsten Vertreter der verschiedenen Datenquellen [Bax02, Bax03, Gal04]. So präsentierten sich in der nunmehr 11. Ausgabe vom Januar 2004 über 140 unterschiedliche molekularbiologische Quellen. Überdies hat der Autor eine Zusammenstellung von 548 Datenquellen und ihren Links in einer Liste[1] bereitgestellt. Eine Sichtung verschiedener genomischer Datenquellen führten 1998 BORSANI [BBB98] und seine Kollegen durch. Als deutschsprachiges Werk beschäftigt sich das Bioinformatik-Lehrbuch von RAUHUT [Rau01] mit den derzeit gebräuchlichsten Sequenz- und Proteindatenbanken, Werkzeugen und experimentellen Ansätzen. Ein ähnlich anwendungsorientiertes, englischsprachiges Werk von BAXEVANIS und QUELLETTE [BQ01] ist 2001

[1] Diese Liste ist über http://nar.oupjournals.org erreichbar.

bereits in der zweiten Auflage erschienen.

Die URLs der vorgestellten Life–Science–Datenquellen sind im Anhang A verzeichnet. Im Anhang B wird außerdem ein vertiefender Einblick in die Struktur, Bedeutung und Zusammenhang der verfügbaren Daten der Sequenzdatenquelle EMBL gegeben, um den Integrationsvorgang näher zu beleuchten. Sollte sich jedoch die Anzahl oder der Umfang der zu integrierenden Datenquellen in Zukunft erweitern, so ist die Vorgehensweise für diese Quellen natürlich ähnlich zu den folgenden analysierten Datenquellen.

3.1.1 Medizinische Datenquellen

Unter dem Begriff 'Medizinische Datenquellen' werden in diesem Abschnitt Systeme präsentiert, die aus klinischer Sicht allgemeine und spezifische Daten über Gendefekte (Mutationen) halten, beispielsweise korrespondierende Phänotypen, Symptome, klinische Diagnostik und Therapie.

So ist OMIM ein in der klinischen Praxis häufig genutztes System, das vielfältige Informationen zu einer großen Anzahl verschiedener Erkrankungen bereitstellt. Neben weiteren allgemeinen Datenquellen, die wie OMIM und Metagene [FT98] ein möglichst breites Spektrum von Defekten beschreiben, existieren aber auch Systeme, deren Ziel es ist, viele Daten zu einer spezifischen Erkrankung oder Erkrankungsfamilie zu sammeln, so zum Beispiel PAHdb über die Hyperphenylalaninämien, insbesondere Phenylketonurie (PKU).

Online Mendelian Inheritance in Man (OMIM)

Die *Mendelian Inheritance in Man* wurde 1998 bereits zum zwölften Mal [McK98] als gedruckte Ausgabe veröffentlicht. Sie ist eine umfassende Zusammenstellung von Informationen zu Genen und genetischen Störungen aus der medizinischen und molekularbiologischen Fachliteratur zur Unterstützung von Forschung und Lehre in der Humangenetik und der klinischen Praxis. Eine elektronische Version dieses Nachschlagewerkes ist unter dem Namen *Online Mendelian Inheritance in Man* (OMIM) mit über 13000 Einträgen[2] im WWW frei verfügbar und wird täglich aktualisiert. Dieses System wird durch das US–amerikanische Humangenomprojekt mit öffentlichen Mitteln finanziert [HSA+02].

Jeder OMIM–Eintrag wird durch einen sechsstelligen, numerischen Identifikator bezeichnet, wobei die erste Ziffer eine Unterteilung der Einträge in die Vererbungsweise[3] eines Merkmales (Erbgang) durchführt. Diese sogenannte MIM–Nummer hat sich mittlerweise auch in weiteren Systemen als Identifikator durchgesetzt und wird vielfach genutzt, um einen Verweis beispielsweise als HTML–Link auf OMIM herzustellen. Als Beispiel sei an dieser Stelle die *Frequency of Inherited Disorders Database* [AHK+01] genannt, in der

[2] Stand Oktober 2001

[3] z.B. autosomal–dominant, autosomal–rezessiv, X–chromosomal, Y–chromosomal

Inzidenz[4] und Prävalenz[5] von genetischen Erkrankungen aus der Literatur registriert werden. Mittlerweile wurde auch eine OMIM–ähnliche Datenbank für genetische Störungen bei Tieren unter dem Namen OMIA [Nic03] ins Leben gerufen.

PAHdb

In [NBPS98, SWS+00] wird PAHdb als eine relationale, ortsspezifische Datenbank für Mutationen des menschlichen Phenylalaninhydroxylase–Genes (PAH) und assoziierte Phänotypen auf der Ebene der Proteine, Metaboliten und des Organismus beschrieben. Dieses System wurde als Prototyp für weitere locusspezifische Datenbanken entwickelt. Die gesammelten Daten werden von Medizinern übermittelt oder der Fachliteratur entnommen. Alle Einträge werden vor der Veröffentlichung datiert und verifiziert. Die verschiedenen Allelvarianten werden nach einer einheitlichen Nomenklatur [DA01] bezeichnet, die sich mittlerweile auch in anderen Veröffentlichungen durchgesetzt hat und als Quasi–Standard angesehen werden kann. Außerdem wurden in diesem System bereits erste Ansätze zur Verbindung von Genotyp und Phänotyp sowie Struktur und Funktion umgesetzt [WPNS98].

Ein Teil der beschriebenen Gendefekte verursachen die Stoffwechselerkrankung Phenylketonurie (PKU, MIM–Nummer: 261600). Das Enzym Phenylalaninhydroxylase ist im menschlichen Organismus für die biochemische Reaktion notwendig, die Phenylalanin zu Thyrosin metabolisiert. Durch den Defekt bei der Synthetisierung des Genproduktes reichert sich nun Phenylalanin im Stoffwechsel des Patienten an und führt bei nicht adäquat oder nicht rechtzeitig therapierten Erkrankten im Kindesalter zu schwerwiegenden Schädigungen bei der Entwicklung des Gehirnes. Bei rechtzeitiger Erkennung des Defektes, z.B. durch das Neugeborenenscreening, wird als therapeutische Maßnahme typischerweise eine phenylalaninreduzierte Diät durchgeführt.

HGMD

Die *Human Gene Mutation Database* sammelt nach [CBK98, KBF+00] veröffentlichte Genmutationen, die Erkrankungen des Menschen verursachen. HGMD–Einträge umfassen die entsprechende Literaturreferenz, die assoziierte Erkrankung, den Gennamen und den Ort der Mutation auf dem Chromosom. Sie werden durch eine Kombination von manueller und automatisierter Suche aus 250 regelmäßig erscheinenden Zeitschriften gewonnen.

Neben den verschiedenen Mutationsarten werden auch die entsprechenden Veränderungen der Nukleotidsequenz im Rahmen des Gens oder Chromosoms dargestellt. So wird beispielsweise bei einem Nukleotidaustausch das betroffene Triplet angegeben und die

[4] Anzahl neuer Erkrankungsfälle in der Zeiteinheit

[5] Häufigkeit aller Fälle einer bestimmten Krankheit in einer Population zum Zeitpunkt der Untersuchung

mutierte Base gekennzeichnet. Im Januar 2003 waren etwa 32000 Mutationen in 1300 Genen verzeichnet.

3.1.2 Genomische Sequenzdatenquellen

Durch zahlreiche laufende Genomprojekte werden in großem Umfang DNS–Sequenzdatenbanken erstellt. In der post–genomischen Ära bilden diese Daten den Ausgangspunkt für eine rechnerunterstützte Analyse und Weiterverarbeitung im Rahmen der Genomforschung, wie beispielsweise zur Untersuchung des Zusammenhanges von Sequenz und Funktion.

Die im Rahmen des öffentlich geförderten Internationalen Humangenomprojektes (*International Human Genom Sequencing Consortium*) gewonnen genomischen Sequenzdaten werden in den folgenden drei wichtigsten internationalen Datenbanken dieser Domäne hinterlegt.

DNA Database of Japan (DDBJ)
: am Center for Information Biology des National Institute of Genetics in Mishima, Japan [MSGT03, MSI+04]

EMBL Nucleotide Database
: des European Molecular Biology Laboratory (EMBL) am European Bioinformatics Institute (EBI) in Hinxton, UK [KAA+04]

GenBank
: des National Center for Biotechnology (NCBI) der National Institutes of Health (NIH) in Bethesda, USA [BKL+04]

Diese drei Datenbanken bilden die *International Nucleotide Sequence Database Collaboration*. Sie dienen als Eingabepunkte für neue Sequenzen. Ihre Inhalte werden täglich miteinander abgeglichen und werden somit deckungsgleich gehalten. Die Formate der Datenbankeinträge und Zugriffsschnittstellen unterscheiden sich jedoch in einigen Punkten. Der Identifikator für die Sequenzen, die Accession–Number, und die ihr zugeordneten Informationen sind aber in den drei Systemen identisch. Nachfolgend sollen nun exemplarisch die Eigenschaften von EMBL und GenBank vorgestellt werden.

EMBL

Die europäische DNS–Sequenzdatenbank ist die EMBL Nucleotide Sequence Database [SBB+03, KAA+04]. Die Dateneingabe erfolgt für einzelne Einträge mittels WebIn, während ein automatisiertes Verfahren die Sequenzierungsdaten von Genomsequenzierungszentren und dem Europäischen Patentamt übermittelt. Quartalsweise werden neue Datenbankversionen freigegeben. Durch die Nutzung des am EBI entwickelten *Sequence*

Retrieval System (SRS) ist es möglich, umfangreiche Daten aus weiteren, angebundenen Datenquellen zu erhalten. Weiterhin werden eine Reihe von Werkzeugen angeboten, die die Suche nach ähnlichen Nukleotidsequenzen innerhalb des aktuellen Datenbestandes ermöglichen, beispielsweise Blitz, Fasta und BLAST.

Von besonderer Bedeutung für die spätere Nutzbarkeit der Daten ist eine aufmerksame Bearbeitung der übermittelten Einzelsequenzen. Die Sequenzen waren häufig Bestandteil von molekulargenetischen Experimenten, die ihre Eigenschaften und Funktion aufklären sollten. Die Daten aus Übermittlungen von Genomsequenzierungsprojekten hingegen besitzen meist nur vorläufige Annotationen, die automatisch erstellt wurden. Da diese Sequenzannotationen von größter Bedeutung sind, werden übermittelte Einzelsequenzen zurückgewiesen, die keine Annotationen aufweisen.

GenBank

GenBank [BKL+03, BKL+04] ist eine umfassende genomische Sequenzdatenbank, die der Öffentlichkeit DNA–Sequenzen von über 119000 verschiedenen Organismen bereitstellt, die hauptsächtlich aus Einzellaboren oder großen Sequenzierungsprojekten übermittelt werden. Im August 2002 enthielt GenBank mehr als 23 Millarden Nukleotide in 18 Millionen Sequenzen. Dieser Datenbestand verdoppelte sich in der zurückliegenden Zeit regelmäßig innerhalb von etwa 15 Monaten. Deshalb wird aller zwei Monate ein kompletter Datenbestand und täglich eine Aktualisierung veröffentlicht, die via FTP verfügbar sind.

Jeder GenBank–Eintrag enthält die Accession–Number als Schlüssel, eine kurze Beschreibung der Sequenz, den wissenschaftlichen Namen, die Taxonomie des Ursprungsorganismus, Literaturreferenzen und eine Tabelle von Merkmalen (*features*), wie das translatierte Protein und die eigentliche Nukleotidsequenz.

3.1.3 Proteinsequenz– und Proteinstrukturdatenquellen

Die beiden bekanntesten Proteinsequenzdatenbanken sind SWISS–PROT und PIR, die ihre Aminosäuresequenzen aus Nukleotiddatenbanken, wie EMBL, beziehen. Der eigentliche Eintrag in diesen Datenbanken ist eine Folge von Aminosäuren, die aus sequenzierten DNA–Abschnitten abgeleitet wurde. Proteinsequenzen selber werden nur selten direkt sequenziert, so daß sich der Nutzer auf die Korrektheit der DNA–Sequenz und ihrer Ableitung verlassen muß. Die Aktivitäten der Forschungsgruppen um SWISS–PROT und PIR werden außerdem im UniProt–System konzentriert.

Die Strukturen von Proteinen werden in steigender Quantität durch Röntgenstrukturanalyse und Kernresonanzspektroskopie ermittelt. In der ersten öffentlichen Datenquelle der Molekularbiologie, der PDB, wurden seit 1971 die 3D–Strukturen von biologischen Makromolekülen gespeichert. Als gebräuchliche Strukturdatenquellen seien außerdem

SCOP (*Structural Classification of Proteins*) und MMDB (*Molecular Modeling Database*) genannt.

SWISS–PROT

SWISS–PROT ist eine kommentierte Proteinsequenzdatenbank, bei der in hohem Maße auf umfangreiche Bemerkungen[6] (*Annotationen*) zu den Proteinsequenzen, geringe Redundanz der Daten und die Verbindung zu anderen Datenquellen in Form von Verweisen Wert gelegt wurde [BA00, BBA+03]. Sie ist 1986 an der Universität Genf, Schweiz entwickelt und in Zusammenarbeit mit dem EMBL (*European Molecular Biology Laboratory*) in Heidelberg gepflegt worden. Nunmehr wird dieses System partnerschaftlich vom EMBL und dem SIB (*Swiss Institute of Bioinformatics*) betrieben, wobei die Aktivitäten des EMBL jetzt beim EBI (*European Bioinformatics Institute*) in Hinxton, UK angesiedelt sind.

Die aktuelle Version von SWISS–PROT enthält in 115000 Einzeleinträgen[7] neben der Aminosäurensequenz, den Annotationen und Verweisen auch Synonyme, Literaturreferenzen und Schlüsselworte für eine schnelle und komfortable Sichtung der Daten. Das Format der Daten orientiert sich an der EMBL–Nukleotidsequenzdatenbank.

Aufgrund des steigenden Volumens der Daten, die vom Genomprojekt geliefert werden und nun aber auf einem hohen Qualitätsniveau annotiert werden sollen, bildeten sich Engpässe und somit zeitliche Verzögerungen während des Annotationsverfahrens. Um dennoch die sequenzierten Daten schnellstmöglich elektronisch bereitzustellen, wurde 1996 TrEMBL (*Translation of EMBL nucleotide sequence database*) entwickelt. TrEMBL enthält 682000 Rechner–annotierte Einträge[8], die durch die Translation aller kodierenden Sequenzen der EMBL–Datenbank, außer den bereits in SWISS–PROT vorhandenen CDS (Coding Sequences), gewonnen werden.

PIR–PSD

Die *Protein Information Resource* (PIR) erstellt in Zusammenarbeit mit dem MIPS (*Munich Information Center for Protein Sequences*) und der japanischen *Protein Information Database* (JIPID) die *PIR-International Protein Sequence Database* (PIR–PSD), die nach eigenen Angaben am umfassensten und sachkundigsten annotierte Proteinsequenzdatenbank [BGH+01, WYH+03]. Dabei sind die einzelnen Einträge in Proteinfamilien klassifiziert.

Die PIR–PSD wird mit standardisierter Fachterminologie annotiert, um die Qualität und Interoperabilität der Daten zu sichern. Dazu werden u.a. Originalliteratur und Zusammen-

[6] z.B. Funktion des Proteins, Struktur, post–translationale Modifikationen, Ähnlichkeiten zu anderen Proteinen, assoziierte Erkrankungen

[7] Stand Oktober 2002

[8] Stand Oktober 2002

fassungen aus MEDLINE verwendet und referenziert. Außerdem sind die Einzeleinträge mit Statusinformationen versehen, die Auskunft über die Bedingungen geben, unter denen die Aussagen zu den Proteinen und ihren Eigenschaften getroffen wurde. Durch die Nutzung von regelbasierten Methoden und durch Klassifikationsverfahren werden ähnliche Sequenzen mit Informationen angereichert und die Integrität der Einträge überprüft.

UniProt

Die Datenquellen SWISS–PROT, TrEMBL und PIR wurden außerdem unter dem Namen *Universal Protein Knowledgebase* (UniProt) zusammengefaßt und bieten dort eine Sammlung der einzelnen Sequenzinformationen und funktionalen Daten an [ABW+04]. Sie unterteilt sich in drei Abschnitte: das UniProt Archive (UniParc), die zentrale UniProt Knowledgebase (UniProt) und die UniProt NREF Datenbank (UniRef).

UniParc bietet eine nicht–redundante Sammlung von Proteinsequenzen aus verschiedenen Quellen, dazu gehören SWISS–PROT, TrEMBL, PIR–PSD, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBAse, WormBase und europäische, amerikanische und japanische Patentämter. Während eine Sequenz in unterschiedlichen Quellen mehrfach verzeichnet sein kann, wird in UniParc jede einzigartige Sequenz nur einmal gespeichert und mit einem eindeutigen Schlüssel versehen. Weiterhin wird dieser Eintrag mit Referenzen zu den Quelldatenbanken versehen. Veränderungen der Originalsequenz können anhand von Versionsnummern verfolgt werden.

Für die zentrale UniProt–Datenbank wurden Proteinsequenzen, Annotation und funktionale Informationen aus SWISS–PROT und TrEMBL um Einträge aus PIR–PSD ergänzt. Wechselseitige Referenzen erleichtern die Suche nach spezifischen Daten. Aus der Struktur der Quellen resultiert eine Unterscheidung der Informationen in manuell annotierte Sequenzen, die durch eine Literaturrecherche und evaluierte Computeranalysen entstanden sind, und rechnerannotierte Sequenzen, die bisher noch nicht manuell überprüft, sondern durch automatisierte Verfahren angereichert wurden.

UniRef bietet dem Nutzer nicht–redundante Verknüpfungen verfügbarer Sequenzen eines oder mehrerer Organismen, die mit Links zu den beteiligten Sequenzen, der Taxonomie, den Literaturreferenzen und der resultierenden Sequenz versehen sind. Der Wert der Identität der verknüpften Sequenzen unterteilt dabei die Resultate in drei Gruppen (NREF100, NREF90, NREF50).

PDB

In der *Protein Data Bank* (PDB) werden Strukturdaten biologischer Makromoleküle gesammelt [BWF+00]. Seit 1971 konnte sich PDB als zentrale, weltweite Datenbank für Proteinstrukturen etablieren und wird seit 1998 durch ein Konsortium von Forschungseinrichtungen, das *Research Collaboraty for Structural Bioinformatics* (RCSB), betreut. Sie erfaßt und archiviert experimentell ermittelte dreidimensionale Strukturen. Vor der

Aufnahme einer neuen Struktur werden die Daten überprüft und mit einem Identifikator versehen. Da die Veröffentlichung kristallografischer Ergebnisse in Fachzeitschriften häufig an eine Eintragung in PDB gebunden ist, sind die meisten von öffentlichen Forschungseinrichtungen ermittelten Proteinstrukturen in dieser Datenbank zu finden [GJ02].

Die Struktur eines Moleküles wird in einem dreidimensionalen Raumgitter angeordnet und durch die jeweilige Position der Atome in diesem Gitter beschrieben. Neben den Positionen werden Informationen zu Atomcharakter, Bindungen und Wechselwirkungen gehalten. Die in einer Strukturdatei gespeicherten Daten können durch geeignete Softwarewerkzeuge visualisiert werden und können so beispielsweise ein frei drehbares dreidimensionales Bild repräsentieren.

3.1.4 Metabolische und regulatorische Datenquellen

Durch die Nutzung metabolischer und regulatorischer Datenquellen werden Informationen über die Zusammenhänge innerhalb von Stoffwechselnetzwerken (metabolische Netze) und genregulatorischen Netzwerken bereitgestellt. Während metabolische Netze aus einer Reihe von enzymkatalysierten Einzelreaktionen bestehen, steuern genregulatorischen Netzwerke die Aktivierung von Genen zur Synthese der Genprodukte (Genexpression).

BRENDA

In BRENDA (*Braunschweig Enzyme Database*) werden Enzymdaten und metabolische Informationen aus der Primärliteratur manuell gesammelt [SCS02, SCE+04]. Die Datenbank stellt biochemische und molekulare Daten zu etwa 83000 Enzymen[9] von 9800 verschiedenen Organismen zur Verfügung, die durch 4200 EC–Nummern klassifiziert werden. Zu diesen Daten zählen Informationen über katalysierte Reaktionen, Vorkommen, Sequenz, Kinetik, Substrate und Produkte, Inhibitoren, Kofaktoren, Aktivatoren, Struktur und Stabilität.

Für eine grobe Einschätzung vorhandener Literaturreferenzen wird ein Verfahren zur automatischen Extraktion von relevanten Daten entwickelt. Verschiedene Werkzeuge ermöglichen einen komfortablen Zugang zum Datenbestand. Neben den typischen Suchformularen werden Browser angeboten, die auf der Baumstruktur der EC–Klassifikation (ECTree) oder der Taxonomie von Organismen (TaxTree) basieren.

KEGG

Die *Kyoto Encyclopedia of Genes and Genomes* wurde im Rahmen des japanischen Humangenomprojektes angelegt und setzt sich aus drei Hauptkomponenten zusammen: der

[9] Stand Januar 2004

Pathway–Datenbank mit Daten über Netzwerke der molekularen Interaktion, der Genes–Datenbank mit Daten zu Genen und Proteinen, die aus Sequenzierungsprojekten gewonnen werden und der Ligand–Datenbank, die Daten über Moleküle und chemische Reaktionen beinhaltet, die in Zellprozessen eine Rolle spielen [KGKN02, KGK+04]. Die Datenstruktur repräsentiert Graphen, die mit verschiedenen Algorithmen untersucht werden können, um typische Grapheigenschaften zu finden, die mit biologischen Zusammenhängen assoziiert werden können.

Die Pathway–Datenbank von KEGG bietet in über 201 Referenzpathways[10] — allgemeinen, speziesübergreifenden Netzwerken — die beteiligten Einzelreaktionen, die dann beispielsweise den Zugriff auf bereits bekannte 3D-Strukturen von Enzymen, korrespondierende genetische Erkrankungen und Gene ermöglichen.

TRANSPATH und TRANSFAC

Die deutsche Firma Biobase bietet TRANSPATH [KVC+03] als Datenbank genregulatorischer Netzwerke an, die umfangreiche Informationen zur Signaltransduktion mit Werkzeugen zur Visualisierung und Analyse verbindet. Durch die Integration von TRANSFAC können umfassende Netzwerke vom Ligand zu den Zielgenen und ihrer Genprodukten bereitgestellt werden.

Die TRANSFAC–Datenbank [WCF+01, MFG+03] wurde eingerichtet, um die Interaktion von Transkriptionsfaktoren und ihren DNS–Bindungsstellen zu modellieren und ihren Einfluß auf die Genexpression darzustellen. Den Kern dieses Systems bilden also die Transkriptionsfaktoren (FACTOR), ihre Bindungsstellen (SITE) und die regulierten Gene (GENE).

3.1.5 Wirkstoffdatenquellen

Durch das bessere Verständnis der grundlegenden Mechanismen, die die Entstehung der vielen verschiedenen Erkrankungen bewirken, steigen natürlich auch die Ansatzpunkte, um therapeutisch in die Krankheitsentstehung und den –verlauf einzugreifen. Besonders das Zusammenspiel zwischen Informationstechnik und Molekularbiologie scheint geeignet, um potentielle Wirkstoffe für neue Medikamente zu entwickeln. Fortschritte, beispielsweise in der Untersuchung von Protein–Interaktionen [The02a], sind jedoch nur durch eine geeignete Unterstützung mit entsprechender informationstechnischer Infrastruktur möglich. Im Bereich der Datenbanken für Wirkstoffe existieren jedoch nur einige Datenquellen. Firmen, die im Bereich der pharmakologischen Forschung aktiv sind, besitzen ihre eigenen Informationssysteme, die aber nach außen abgeschirmt sind.

[10] Stand September 2001

MDDrugDB

In der Arbeitsgruppe Bioinformatik/Medizinische Informatik der Universität Magdeburg wurde der Prototyp eines Informationssystemes für Wirkstoffe und Medikamente entwickelt. Das Ziel dieses Systemes mit dem Namen MDDrugDB ist die Unterscheidung der verschiedenen Wirkmechanismen von Medikamenten [KTSH01]. Es soll dargestellt werden, innerhalb welcher biochemischer Reaktion ein Wirkstoff in einen Stoffwechselweg eingreift. Dadurch kann beispielsweise die Möglichkeit zur Aktivierung alternativer Stoffwechselwege zur Umgehung defekter Einzelreaktionen innerhalb einer Reaktionskette untersucht werden.

In diesem System werden jedoch nicht nur klassische Medikamente erfaßt, sondern auch neuere Therapieverfahren, die sich teilweise noch im Versuchsstadium befinden, wie beispielsweise die Anwendung gentechnisch hergestellter Proteine. Bei der überwiegenden Anzahl von Medikamenten wirkt ein kleineres Molekül als Ligand auf ein makromolekulares Targetmolekül, wie ein Enzym, einen Rezeptor oder Transporter. In zunehmendem Maße werden jedoch auch Medikamente entwickelt, die auf andere Ziele wirken.

Die Wirkstoffdatenbank basiert auf einem Oracle-Datenbanksystem, wobei die Nutzerschnittstelle über dynamisch generierte HTML-Seiten realisiert wurde. Das Datenbankmodell basiert auf dem Prinzip der Medikamentenwirkung. Diese werden, wie in der Abbildung 2.5 dargestellt, vereinfacht als Wirkstoff–Target–Effekt–Relation aufgefaßt. Als Hauptdatenquelle dient die medizinische Literaturdatenbank MEDLINE.

Im Internet ist eine Vielzahl von Datenquellen verfügbar, die die hier angesprochenen Wirkstoffe und Medikamente adressieren. Allerdings geben diese keine Auskunft über die zugrundeliegenden molekularen Wirkmechanismen. Außerdem sind sie, im Gegensatz zu einem Großteil der molekularbiologischen Datenbanken mit Informationen über Gene, Enzyme und Stoffwechselwege, nicht frei zugänglich.

PharmGKB

Die PharmGKB (*Pharmacogenetics and Pharmacogenomics Knowledge Base*) ist eine Datenbank zur Bereitstellung von Informationen über die Beziehung zwischen der Wirkung von pharmazeutischen Stoffen und genetischen Einflüssen [HOR+02]. Dies wird durch die Sammlung von genotypischen, phänotypischen und klinischen Daten aus der Forschung erreicht.

Zur Herstellung einer Verbindung zwischen genetischen und phänotypischen Variationen können eine Vielzahl von Experimenten durchgeführt werden. Es ist jedoch problematisch, diese Resultate in einem Datenbanksystem abzulegen und analysierbar zu machen. Dazu werden nun die experimentellen Ergebnisse einer Kategorie zugeordnet, die ihre Bedeutung im Rahmen der pharmakologisch–genomischen Untersuchungen anzeigt. Die Daten werden dabei in einem XML–Format eingesandt, um anschließend überprüft und eingearbeitet zu werden.

3.1.6 Zusammenfassende Gegenüberstellung

Nachdem in den vorangegangenen Abschnitten einzelne Datenquellen entsprechend ihrer Klassifikation vorgestellt wurden, sollen sie nun in einer zusammenfassenden Tabelle 3.1 gegenübergestellt werden. Dazu werden verschiedene Merkmale herangezogen, die im Rahmen dieser Arbeit die Möglichkeit geben, eine Entscheidung über eine Integration der zur Verfügung stehenden Datenbestände zu treffen.

Domäne
: Die bestehenden Datenquellen werden anhand der bereitgestellten medizinischen oder molekularbiologischen Daten einer Domäne zugeordnet.

Format
: Für die Anbindung der verteilten Datenquellen an einen Integrationsdienst sind Kenntnisse über das bereitgestellte Format der zu integrierenden Daten notwendig. Die verschiedenen Formen, in denen die Daten angeboten werden, können Flatfiles, HTML–Dokumente oder XML–Dateien sein. Ein Teil der Datenquellen bietet auch einen direkten Zugriff über eine Programmierschnittstelle an, so daß Datenbankanfragesprachen, beispielsweise SQL oder OQL, genutzt werden können. Ebenfalls sind teilweise prototypische Implementierungen für Anfragen durch CORBA oder Web Services und SOAP verfügbar oder geplant.

Lizenzfrei
: Der Zugriff und die Nutzung von Daten aus den Datenquellen kann aufgrund unterschiedlicher Lizenzmodelle bestimmten Einschränkungen unterliegen. Während Inhalte aus öffentlich geförderten Projekten meist für eine akademische Nutzung im Rahmen von Forschung und Lehre kostenfrei zur Verfügung gestellt werden, ist eine kommerzielle Verwertung durch Firmen häufig mit erheblichen Kosten verbunden. Der in dieser Arbeit vorgestellte Prototyp wurde aber im Rahmen eines öffentlichen Forschungsprojektes erstellt und kann somit als rein akademische Anwendung betrachtet werden. Somit bezeichnet das Merkmal der Lizenz hier die Verfügbarkeit eines kostenfreien Zugriffes zur akademischen Nutzung. Dazu werden für dieses Merkmal die Ausprägungen *ja*, *teilweise*, *nein* und *unbekannt* genutzt.

3.2 Integrationsarchitekturen und bestehende Ansätze

Die entstehenden Probleme der wachsenden Spezialisierung von Datenquellen und ihrer unterschiedlichen Formate wurden im Laufe der ersten Phase des Humangenomprojektes besonders deutlich. Natürlich wurden auch schon vorher Überlegungen zur Zusammenführung der verschiedenen Datenbestände diskutiert, um Gesetzmäßigkeiten und

Datenquelle	Merkmale		
	Domäne	Format	Lizenzfrei
BRENDA	Metabolismus	HTML	ja
DDBJ	gen. Sequenzen	Flatfile, HTML, XML	ja
EMBL	gen. Sequenzen	Flatfile, HTML, XML	ja
GenBank	gen. Sequenzen	Flatfile, HTML, XML	ja
HGMD	Mutationen	HTML	ja
KEGG	Metabolismus	HTML, XML	ja
MDDrugDB	Wirkstoffe	HTML	ja
OMIM	Krankheiten	Flatfile, HTML	ja
PAHdb	Mutationen	HTML	ja
PDB	Proteinstruktur	HTML	ja
PharmGKB	Wirkstoffe	HTML	teilweise
PIR–PSD	Proteinsequenzen	HTML, XML, SQL	ja
Ramedis	Mutationen	HTML, SQL	ja
SWISS–PROT	Proteinsequenzen	Flatfile, HTML	ja
TRANSFAC/TRANSPATH	Genregulation	Flatfile, HTML	teilweise
UniProt	Proteinsequnzen	Flatfile, HTML, XML	ja

Tabelle 3.1: Gegenüberstellung verschiedener medizinischer und molekularbiologischer Datenquellen anhand von drei Merkmalen

Zusammenhänge auf einer höheren Ebene untersuchen zu können. Durch die rapide ansteigenden Datenmengen wurde dieses Problem jedoch weiter in den Blickpunkt der Verantwortlichen gerückt.

Diese weitgehend voneinander unabhängig entwickelten Life–Science–Datenquellen unterscheiden sich in einer Vielzahl von Aspekten und sind bestenfalls miteinander durch gegenseitige Hyperlinks verbunden. Wird ein professionelles Datenbankmanagementsystem genutzt, so werden typischerweise relationale oder objektorientierte Ansätze verfolgt. Häufig sind jedoch auch noch Flatfiles in eigenen Formaten vorhanden. Daneben existieren die Schemakonflikte, die sich durch unterschiedliche Namensgebung, Relationenschemata und Datentypen ergeben. Darüber hinaus ist auch der Zugriff auf die Datenquellen nicht einheitlich. Typischerweise werden Suchmasken als HTML–Dokumente bereitgestellt. Modernere Systeme unterstützen bereits den direkten Zugriff über Datenbankanfragesprachen (SQL, OQL) und Programmierschnittstellen (CORBA, JDBC).

So wurde 1995 in [DOB95] die Bedeutung der Integration von molekularbiologischen Datenquellen für die Forschung herausgestellt und dabei die technischen Herausforderungen diskutiert, die verschiedenen Ansätze klassifiziert und verfügbare Methoden und Werkzeuge untersucht. Die Bewertung von Integrationsansätzen wurde mit Hilfe von zwei Dimensionen durchgeführt: dem Grad der Integration und dem Grad der Kopplung. Diese beiden Merkmale werden auch in [Sch02] genutzt und um weitere ergänzt.

Bei [Kar95, LR03] werden mit Blick auf die aktuellen Forschungen Anforderungen und Voraussetzungen für die Architektur einer anfragebasierten Infrastruktur formuliert. Insbesondere werden dabei die folgenden Eckpunkte und Besonderheiten molekularbiologischer Datenbestände formuliert.

- Die verschiedenen medizinischen und molekularbiologischen Datenquellen besitzen Heterogenität auf unterschiedlichen Ebenen. Zur Nutzung von Daten aus verschiedenen verteilten und heterogenen Quellen sind geeignete, flexible Mechanismen bereitzustellen, die anwendungsspezifische Integrationen unterstützen. (Forderung nach Datenintegration)

- Molekularbiologische Datenquellen haben komplexe Schemata, die sich häufig verändern. Dadurch wird für eine anwendungsabhängige Datenintegration häufig manuelle Arbeit von Fachexperten benötigt, die das Schema des integrierten Systemes für spezifische Projekte semantisch korrekt anlegen und seine Elemente eindeutig definieren. (Forderung nach semantischer Korrektheit und Redundanzfreiheit)

- Nutzer mit biologischem Fachwissen müssen nicht über Wissen zu Herkunft, Zugriffsmechanismen und Schema der einzelnen Datenquellen verfügen, sollten jedoch mit Informationen über die erwartete Qualität der integrierten Daten einer Quelle in Bezug auf Fehlerfreiheit, Vollständigkeit und Aktualität versorgt werden. (Forderung nach Transparenz)

- Die Möglichkeit zur Verarbeitung einer komplexen Anfrage über unterschiedlichen Datenquellen erlaubt es Anwendern Fragen zu beantworten, die bei der Verwendung von Hypertext–Ansätzen nicht möglich gewesen wären.

- Viele Nutzer können aus den hunderten von verfügbaren biologischen Datenquellen nicht die kleine Untermenge auswählen, die sie für ihre spezifische Anfrage benötigen. Die Daten der angeschlossenen Datenquellen sollen jedoch durch den Integrationsdienst uneingeschränkt zugreifbar sein. (Forderung nach Spezifität und Vollständigkeit)

- Die Anwender benötigen rechtzeitig Zugriff auf die aktuellsten Datenbestände, der auch im Hinblick auf die Performance des Systemes den Ansprüchen des Anwenders genügt. (Forderung nach Aktualität und Performance)

Ausgehend von diesen Besonderheiten molekularbiologischer Datenbestände wird eine Architektur benötigt, die Zugriff auf verteilte, heterogene Datenquellen bietet. Dabei müssen Schemata und Daten der angeschlossenen Quellen auf möglichst aktuellem Stand gehalten werden. Die erforderliche Anbindung der Datenquellen muß durch einen Fachexperten übernommen werden und sollte nicht dem Nutzer überlassen werden. In den folgenden Abschnitten werden Architekturen und bestehende Ansätze zur Integration von Datenquellen vorgestellt und diskutiert. Dabei soll untersucht werden, ob bereits ein

geeigneter Integrationsdienste existiert, der sich für die Nutzung im Rahmen der vorliegenden Arbeit anbietet.

3.2.1 Architekturen zur Integration von Datenquellen

Zur Integration verteilter, heterogener Datenquellen existieren eine Reihe von Architekturen, die sich voneinander in vier prinzipielle Klassen abgrenzen lassen. Diese wurden in verschiedenen Veröffentlichungen beispielsweise von [Kar95, DOB95, FHL+02] gegenübergestellt und diskutiert. Dennoch hat sich bisher keine dieser Architekturen in der Breite durchgesetzt, so daß eine Vielzahl von konkurrierenden Ansätzen existiert. Diese vier Architekturklassen faßt die nachfolgende Aufstellung zusammen, bevor sie anschließend detailliert erläutert werden.

- Hypertext–Navigation
- Multidatenbanken
- Föderierte Datenbanken
- Data Warehouse

Die Verbindung verteilter Datenquellen durch Hypertext–Links im HTML–Dokument ist derzeit am gebräuchlichsten, da sie einfach zu realisieren ist und für Anwender durch die Nutzung eines HTML–Browsers problemlos zur Verfügung gestellt werden kann. Durch diese Art der Integration wird jedoch keine Rücksicht auf die Heterogenität der verlinkten Datenquellen genommen, so daß in diesem Fall häufig von einer unechten Integration gesprochen wird, da Daten für eine Weiterverarbeitung trotzdem zusammengesucht und verbunden werden müssen. Außerdem besteht die Gefahr, daß die verwendeten URLs veraltet oder ungültig sind. Die häufige Nutzung dieser Verbindung zwischen verteilten Datenbanken wurde bei [BK03] betont, der beispielsweise 111 zufällig ausgewählte molekularbiologische Datenquellen untersuchte, von denen 87 Prozent Hypertext–Links zu externen Quellen aufwiesen.

Eine Koppelung von verschiedenen Datenquellen zu Komponenten eines verbundenen Systemes wird als Multidatenbank bezeichnet. Die nachfolgend vorgestellte und in Abbildung 3.2 dargestellte Unterscheidung von Multidatenbanksystemen wurde in [SL90] vorgeschlagen und beinhaltet föderierte Datenbanken als Teilmenge. Es existieren jedoch in der Literatur weitere, davon abweichende Möglichkeiten zur Klassifikation. Die folgenden Abschnitte werden allgemeine und speziellere Architekturen für eine Integration von verteilten, heterogenen Datenquellen kurz vorstellen und in die gewählte Klassifikationshierarchie einordnen.

Zur Unterstützung einer Integration werden häufig ontologiebasierte Konzepte angewendet, die durch die Definition von bestimmten Konzepten und Beziehungen zwischen die-

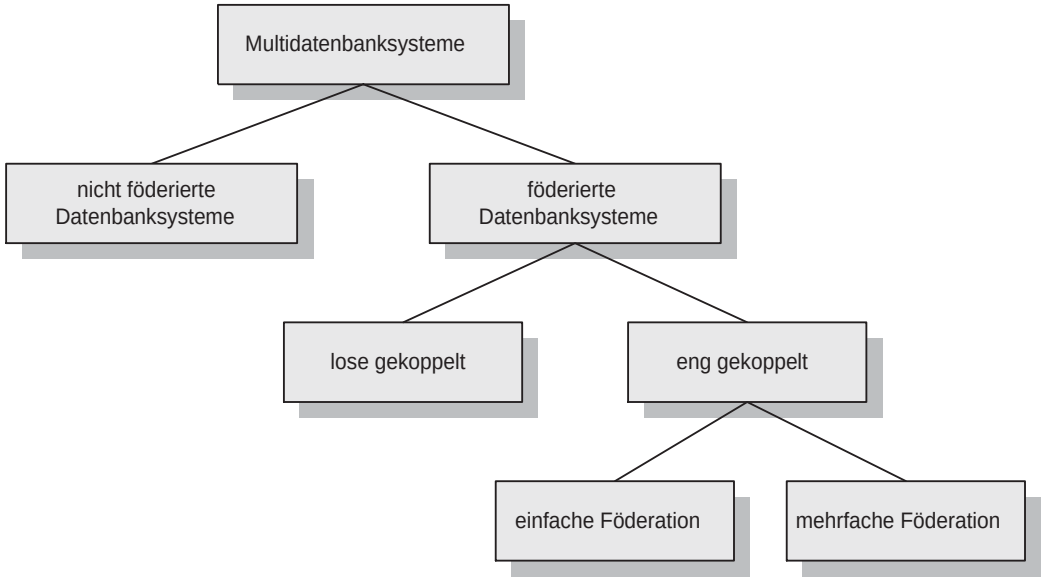

Abbildung 3.2: Klassifikation für Multidatenbanksysteme (nach [SL90])

sen Konzepten die Semantik von Datenbankinhalten beschreiben. So könnten zwei Konzepte *Enzym* und *Amylase* repräsentieren. Durch eine Relation *ist ein* wäre die Beziehung *Amylase ist ein Enzym* hergestellt.

Bei [Köh03] wird eine Datenbank mit dem Namen SEMEDA (*Semantic Meta Database*) vorgestellt, die dem Nutzer durch Bereitstellung von semantischen Informationen in Form einer Ontologie Anfragen an integrierte Datenbanken erlaubt. Dazu benötigt der Nutzer keine Informationen über das zugrundeliegende Schema der Quelle oder andere technische Details der angefragten Datenquellen. Diese Vereinfachung wird durch die Speicherung von Metadaten über die Datenstruktur der Datenquellen in SEMEDA erreicht.

Multidatenbanksysteme

Nach [Con97] beruht die Multidatenbanken–Architektur auf gekoppelten Datenbanksystemen als Komponenten innerhalb eines Verbundes. Die Verwaltung des gesamten Datenbestandes muß nicht von einem Gesamtsystem übernommen werden, vielmehr wird der Datenbestand meist in unabhängigen Partitionen entworfen und verwaltet. Dabei wird dem Nutzer durch eine geeignete Anfragesprache der Zugriff auf die verschiedenen Datenquellen ermöglicht. Die bei dieser Integration auftretenden Konflikte, z.B. in mehreren Datenquellen redundant gespeicherte Daten, strukturelle Unterschiede zwischen den Datenquellen, Unterschiede in der Bezeichnungsweise und Wertinkonsistenzen zwischen den Daten der einzelnen Datenquellen müssen von der zur Verfügung gestellten Sprache behandelt werden können.

Behalten die Komponentendatenquellen einen gewissen Grad an Autonomie, so werden

sie als föderiertes Datenbanksystem bezeichnet. Anderenfalls, bei der Übernahme der Verwaltung des gesamten Datenbestandes durch ein zentrale Instanz ist das System nicht mehr föderiert. Bei welchem Autonomiegrad jedoch die Grenze zwischen föderierten und nicht föderierten System verläuft, ist unterschiedlich. Die Abbildung 3.3 zeigt die bei [LMR90] vorgeschlagene Referenzarchitektur, die aus den folgenden Schemata besteht.

- Das *Physische Schema* beschreibt die physische (interne) Struktur der von der Komponentendatenquelle verwalteten Daten.
- Das *interne logische Schema* stellt das konzeptionelle Schema (implementierungsunabhängig) der Komponente dar.
- Das *konzeptionelle Schema* stellt dem Nutzer der Multidatenbank die Gesamtheit oder aber spezielle Sichten des internen logischen Schemas der einzelnen Komponentendatenquellen zur Verfügung.
- Das *externe Schema* wird vom Nutzer auf den konzeptionellen Schemata der einzelnen Komponenten durch Verwendung der bereitgestellten Sprache angelegt.
- Das *Abhängigkeitsschema* beschreibt Abhängigkeiten zwischen den Daten der einzelnen Komponentendatenquellen, sogenannte Interdatenbankabhängigkeiten.

Diese Schemata sind in drei Ebenen: der internen Ebene, der konzeptionellen Ebene und der externen Ebene angeordnet. Der Nutzer führt die Erstellung der spezifischen integrierten Sicht auf die von ihm benötigten Daten selbständig durch. Eine Anfrage über einer Reihe von einzelnen Datenquellen wird mit der Multidatenbankanfragesprache spezifiziert. In einer zentralen Einheit wird diese Anfrage zerlegt und den entsprechenden Komponentendatenquellen zugeleitet. Die jeweiligen Resultate werden nun von der Verarbeitungseinheit zusammengefügt und an den anfragenden Nutzer zurückgegeben.

Föderierte Datenbanksysteme

Der Begriff des föderierten Datenbanksystems wurde 1985 in [HM85] eingeführt und bei [SL90] präzisiert. Demnach besteht ein föderiertes Datenbanksystem aus einer Menge kooperierender aber autonomer Datenquellen und einer übergeordneten Ebene, dem Föderierungsdienst oder föderierten Datenbankmanagementsystem, das Manipulationen auf den Komponenten kontrolliert und koordiniert, somit also globalen Anwendungen den Zugriff auf die einzelnen Komponentensysteme erlaubt. Die Betonung liegt hierbei jedoch auf einer bestimmten Autonomie der an der Föderation beteiligten Komponenten, d.h. das Gesamtsystem wird nicht vollständig über diese zentrale Instanz verwaltet. Eine allgemeine Architektur föderierter Datenbanksysteme ist in der Abbildung 3.4 dargestellt.

Abhängig vom Grad der Kopplung werden föderierter Datenbanksysteme in eng und lose gekoppelte Systeme unterschieden. Die lose Kopplung der einzelnen Datenquellen wird in

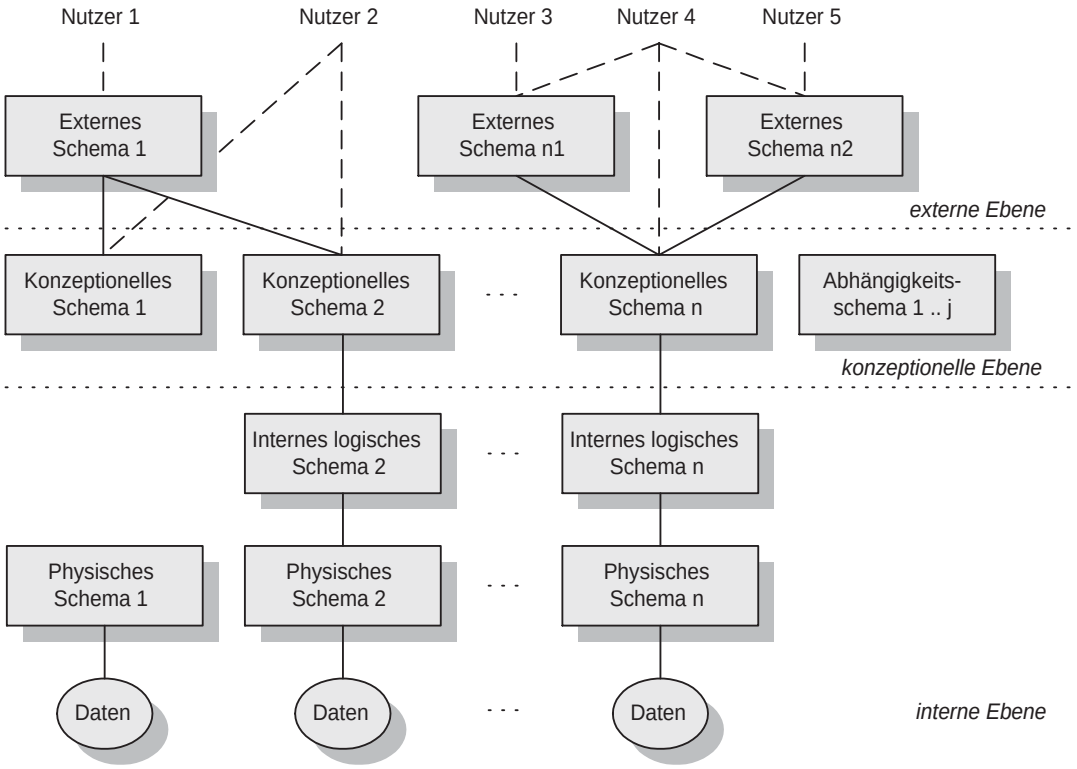

Abbildung 3.3: Referenzarchitektur für Multidatenbanken (nach [LMR90])

erster Linie durch die Verantwortung des Benutzers für die Zusammenführung der teilnehmenden Komponentendatenquellen, den Zugriff auf die verschiedenen Systeme und die geeignete Kopplung der Daten gekennzeichnet. Dieses Vorgehen ermöglicht dem Benutzer durch die Möglichkeit der nutzerspezifischen Entwicklung verschiedener, voneinander unabhängiger Föderierungsschemata ein hohes Maß an Flexibilität. Dazu im Gegensatz steht die enge Kopplung, bei der ein Schema festgelegt wird, auf dem die einzelnen Komponentenschemata abgebildet werden. Der Nutzer wird auf diesem Wege zwar von einem Teil des Aufwandes befreit, er wird jedoch Kompromisse bei der individuellen Auswahl der Daten eingehen müssen.

Für eng gekoppelte, föderierte Datenbanksysteme existiert nun eine weitere Unterscheidungsmöglichkeit. Stellt der den Komponentendatenquellen übergeordnete Föderierungsdienst nur ein einziges globales Schema zur Verfügung, so wird von einer einfachen Föderation gesprochen. Existieren hingegen mehrere föderierte Schemata nebeneinander, in die unterschiedliche Beschreibungen für die Bereitstellung der zu integrierenden Daten einfließen, wird dies als mehrfache Föderation bezeichnet.

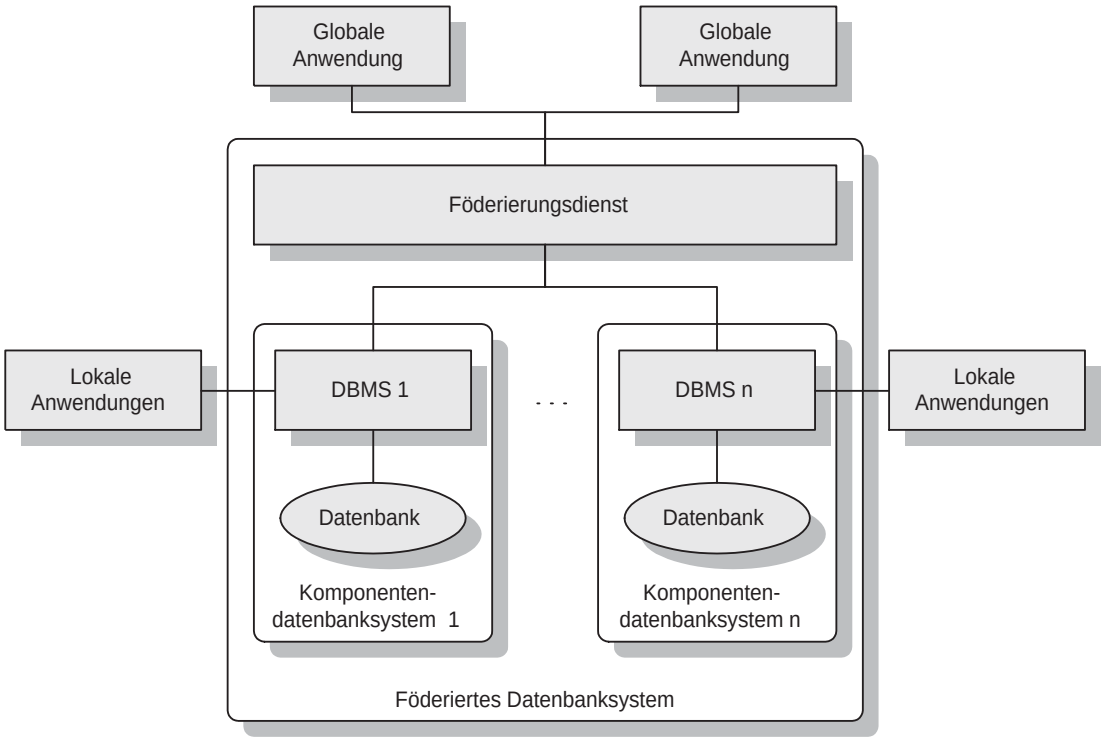

Abbildung 3.4: Allgemeine Architektur föderierter Datenbanksysteme (nach [Con97])

Data Warehouses

Vorwiegend im Bereich der Wirtschaftsinformatik wurde durch die Notwendigkeit, zur Steuerung und Kontrolle von Unternehmensaktivitäten verteilte Datenbestände zu integrieren und auszuwerten, das Konzept des *Data Warehouses* entwickelt. Durch die Integration und Bereinigung von Daten aus unternehmensweit verteilten oder weiteren externen Quellen können Informationen mit höherer Qualität generiert werden. Für Entscheidungsträger sollen so betriebliche Daten in unterschiedlichen Dimensionen zusammengeführt, bereinigt, weiterverarbeitet und visualisiert werden und so Grundlage für die Entscheidung zwischen verschiedenen Handlungsalternativen sein.

Von besonderer Bedeutung ist bei dieser mehrdimensionalen Sicht die zeitliche Entwicklung anderer Dimensionen. Ein Nutzer ist damit in der Lage, Entwicklungen über einen definierten Zeitraum hinweg zu beobachten und Tendenzen für die Zukunft abzuleiten. Typischerweise wird ein Data Warehouse nach [DG98] als eine themenorientierte, integrierte, zeitbezogene und dauerhafte Sammlung von entscheidungsunterstützenden Informationen betrachtet.

In den letzten Jahre wurde der Begriff des *virtuellen Data Warehouses* geprägt, das als föderiertes System eine Menge von Komponentensytemen (Data Marts) enthält und für diese ein globales Schema anbietet, ohne eine vollständige Materialisierung der Daten

durch Replikation durchzuführen.

3.2.2 Bewertung von Integrationsansätzen

Die Arbeit von SCHOLZ [Sch02] analysiert verschiedene molekularbiologische Integrationsansätze und stellt diese einander anhand von zehn Merkmalen gegenüber, von denen bereits zwei in [Kar95] diskutiert wurden. Die Tabelle 3.2 zeigt die Ergebnisse dieser Untersuchung und dient als Vorbereitung für die Auswahl eines geeigneten Integrationsansatzes. Außerdem wurden aktuelle Arbeiten ebenfalls eingeordnet und bewertet. Die einzelnen Merkmale in dieser Tabelle wurden mit Abkürzungen versehen. Nachfolgend werden diese Merkmale und ihre Ausprägungen erläutert.

Grad der Integration G

Der Grad der Integration bezieht sich auf die Existenz eines integrierten Schemas. Im Fall der *engen Kopplung* (+) werden die Schemata der zu integrierenden Datenquellen zu einem gemeinsamen Modell zusammengefaßt, auf dessen Basis anschließend ein neues, integriertes globales Schema entwickelt wird. Dies wird in föderierten Datenbanksystemen durchgeführt. Im Kontrast dazu wird zwar bei der *losen Kopplung* (-) auch ein globales Schema angelegt — es ist jedoch nur die Menge der Originalschemata der Komponentendatenquellen.

Materialisierung der Daten M

Der Grad der Materialisierung der integrierten Daten ist bei der Nutzung von *Views* (+) am geringsten, da hier ein direkter Zugriff über eine Netzwerkverbindung auf die Originaldatenquelle genutzt wird. Dieses Vorgehen sichert zwar einen ständig aktuellen Datenbestand, kann jedoch durch Verfügbarkeit und Qualität der Verbindung negativ beeinflußt werden. Bei der *vollständigen Materialisierung* (-) hingegen werden Kopien der Komponentendatenquellen im Integrationssystem angelegt, so daß ein performanter Zugriff sichergestellt wird. Hier müssen natürlich Abstriche bei der Aktualität der Daten hingenommen werden.

Realisierungsstand R

Beim Realisierungsstand wird der tatsächliche Status der Umsetzung betrachtet. Dazu werden *konkrete Implementierungen* (+) und *theoretische Ansätze* (-) unterschieden.

Plattformunabhängigkeit P

Das Ziel der meisten Softwareentwicklungen ist eine weitgehende *Unabhängigkeit von der verwendeten Plattform* (+). Dennoch ist der Einsatz von Software auf unterschiedlichen Systemen nicht immer ohne größeren Aufwand zu realisieren, so daß eine *Plattformabhängigkeit* (-) heute vielfach akzeptiert werden muß.

Internetfähigkeit I

Dieses Merkmal unterscheidet zwischen der Möglichkeit des Zugriffes auf das In-

Systeme	Merkmale									
	G	M	R	P	I	SA	SP	SF	F	U
BioKleisli	-	+	+	+	+	+	-	-	+	-
Biology Workbench	+	-	+	+	+	-	-	-	-	-
DiscoveryLink	+	+	+	+	+	+	+	+	-	+
Entrez	-	+	+	+	+	-	-	+	-	-
FRIDAQ	+	+	+	+	+	+	+	+	+	+
GeneCards	+	-	+	+	+	-	-	-	-	-
HUSAR	-	-	+	+	+	-	+	+	-	-
IGD	+	-	-	-	+	+	-	+	-	-
ISYS	+	+	+	+	-	+	+	+	+	-
LIMBO	-	-	-	+	+	+	-	+	-	-
Moby Dick	+	-	-	-	-	+	+	+	+	+
PEDANT	+	-	+	+	+	+	-	+	+	+
SRS	-	-	+	+	+	-	+	+	+	-
TAMBIS	-	+	+	+	+	+	-	-	-	-

Tabelle 3.2: Gegenüberstellung verschiedener molekularbiologischer Integrationsansätze anhand von zehn Merkmalen (nach [Sch02])

tegrationssystem über das *Inter- oder Intranet* (+) und einer erforderlichen *lokalen Installation* (-).

Schnittstellen SA, SP, SF

Die drei Schnittstellen–Merkmale erfassen die Unterstützung von Standardanfragesprachen (SA), z.B. SQL oder OQL, die Möglichkeit zur Anbindung mit Programmiersprachen (SP), z.B. über JDBC oder ODBC und die Bereitstellung verschiedener Ausgabeformate, wie HTML oder XML.

Flexibilität F

Ein Maß für die Flexibilität einer Integrationslösung ist die Fähigkeit des Systems, auf unbekannte Nutzeranforderungen zu reagieren oder adäquat angepaßt zu werden. Eine *fixe Lösung* (-) besteht somit aus einem abgeschlossenem System, das keine Möglichkeit zu Anpassung oder Veränderung der Konfiguration bereitstellt. Dem gegenüber steht die *variable Lösung* (+), die nach der Festlegung einer Nutzerkonfiguration eine Generierung des Systemes oder von Systemteilen durchführt, um spezifischen Anforderungen an die Anwendung nachzukommen.

Unterstützung der Informationsfusion U

Dieses Merkmal charakterisiert die Fähigkeit eines Systems zur Kombination, Verdichtung und Interpretation von Daten aus verschiedenen, heterogenen Datenquellen zur Ableitung von Informationen einer neuen Qualität.

Die vorgestellten Merkmale und die Bewertung vorhandener Ansätze in einer Tabelle dient der Vorauswahl für die nachfolgende detailliertere Betrachtung. Ausgehend von den aufgeführten Merkmalen, die zur Gegenüberstellung existierender molekularbiologischer Integrationsansätze in der Tabelle 3.2 herangezogen wurden, werden bei dieser näheren Betrachtung sechs Ansätze (DiscoveryLink, FRIDAQ, GUS, K2/Kleisli, SRS, TAMBIS) untersucht. Ihre Auswahl erfolgte durch die besondere Berücksichtigung der Merkmale Materialisierung der Daten, Realisierungsstand des Systemes und Flexibilität der Lösung.

3.2.3 Vorstellung ausgewählter Integrationsansätze

Aufbauend auf den vorgestellten Integrationsarchitekturen wurden in der Literatur eine Vielzahl von Ansätzen und Prototypen beschrieben, die sowohl rein akademischen Charakter hatten, als auch in kommerziell verwertete Anwendungen eingebunden wurden. Einige der im vorangehenden Abschnitt bewerteten Systeme sollen nun weitergehend untersucht und präsentiert werden, um eine Entscheidung über die Verwendung eines dieser Systeme zur Integration zu treffen.

DiscoveryLink

Mit dem DiscoveryLink–System [HSK+01] vervollständigte die Firma IBM ihr Angebot im Produktbereich Life Science. Dabei wird dem Nutzer eine Middleware zur Verfügung gestellt, die komplexe SQL–Anfragen auf unterschiedliche, verteilte Datenquellen erlaubt. Der Anwender greift dafür nur auf eine virtuelle Datenbank zu. Aufbauend auf Garlic, dem Prototypen eines föderierten Datenbanksystemes, beruht DiscoveryLink auf zwei Schlüsseltechnologien: der Wrapper–Architektur und der Anfrageoptimierung.

Diese Architektur wird in der Abbildung 3.5 verdeutlicht. So verbinden sich Anwendungen unter Verwendung verschiedener, angebotener Schnittstellen zum DiscoveryLink–Server und setzen ihre Anfragen in SQL[11] ab. Die Daten, die notwendig sind, um diese Anfrage zu beantworten, kommen meist aus verschiedenen Datenquellen. Während eines Registrierungs–Prozesses werden dazu die Wrapper über SQL–DDL beim System angemeldt und die erforderlichen Datenbestände als Relationen in den Quellen identifiziert. Diese Datenquellen können sowohl einfache Flat–Files sein, aber auch kommerzielle Datenbanksysteme. Nach der Identifikation der notwendigen Datenquellen wird ein Ausführungsplan für die spezielle Anfrage aufgestellt, der die Originalanfrage in eine Reihe von Fragmenten zerlegt, die auf den einzelnen Datenquellen auszuführen sind.

Die Wrapper agieren dabei als Vermittler zwischen den einzelnen Datenquellen und dem DiscoveryLink–Server. Sie übernehmen folgende Aufgaben:

- Abbildung der Daten aus der Quelle auf das relationale Modell von DiscoveryLink,

[11] Derzeit werden die Statements INSERT, UPDATE und DELETE nicht unterstützt.

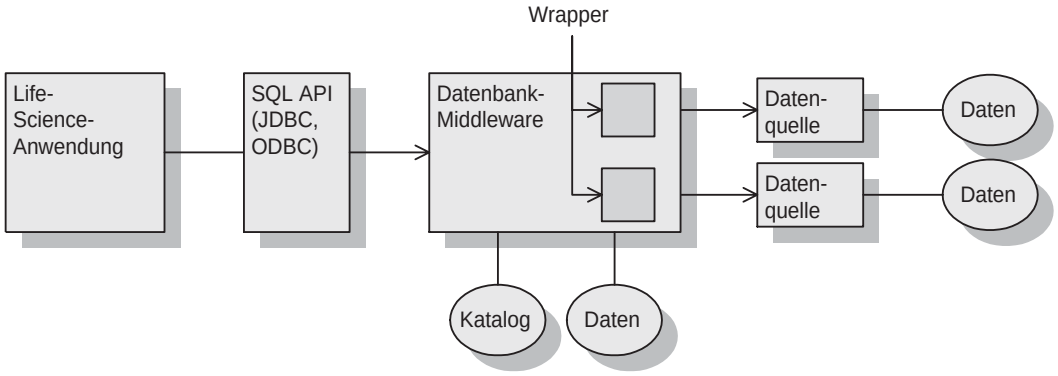

Abbildung 3.5: DiscoveryLink–Architektur (nach [HSK+01])

- Vorhaltung von Informationen über die Fähigkeiten der Quelle bei der Anfrageverarbeitung,
- Abbildung der vom Wrapper übermittelten Anfragefragmente in Anfragen, die von der Quelle verarbeitet werden können,
- Weitergabe der quellenspezifischen Anfrage und Rückgabe der Ergebnisse.

DiscoveryLink ist somit ein eng gekoppeltes föderiertes Datenbanksystem. Dabei wird jede Anfrage auf den Originaldatenquellen ausgeführt und verschiedene Schnittstellen ermöglichen einen lesenden SQL–Zugriff. Für zusätzliche Datenquellen müssen jedoch spezifische Wrapper implementiert werden.

FRIDAQ

Als Framework zur Integration molekularbiologischer Datenquellen wird FRIDAQ bei [Sch02] vorgestellt. Es repräsentiert einen Architekturvorschlag für die Erzeugung eines modularen, konfigurierbaren Systems, das einen universellen und standardisierten Zugriff auf einen den jeweiligen Anforderungen entsprechenden integrierten Datenbestand realisiert. Die Abbildung 3.6 illustriert den FRIDAQ–Architekturvorschlag in UML–Notation.

Dieses Framework besteht im wesentlichen aus drei hierarchisch angeordneten Komponenten: der *Datenintegrationskomponente*, einer *lokalen Speicherungskomponente* und der *Browserkomponente*. Die *existierenden Datenquellen* werden über die *Datenintegrationskomponente* angebunden. Sie stellt die Verbindung zu den weltweit verteilten Datenquellen her, indem über entsprechende Zugriffsmodule die angeforderten Daten der angebundenen Datenquelle homogenisiert werden. Für die verschiedenen angebundenen Datenquellen wird ein gemeinsames, intergriertes Schema genutzt. Die *lokale Speicherungskomponente* hält die Integrationsergebnisse lokal für eine Wiederverwendung be-

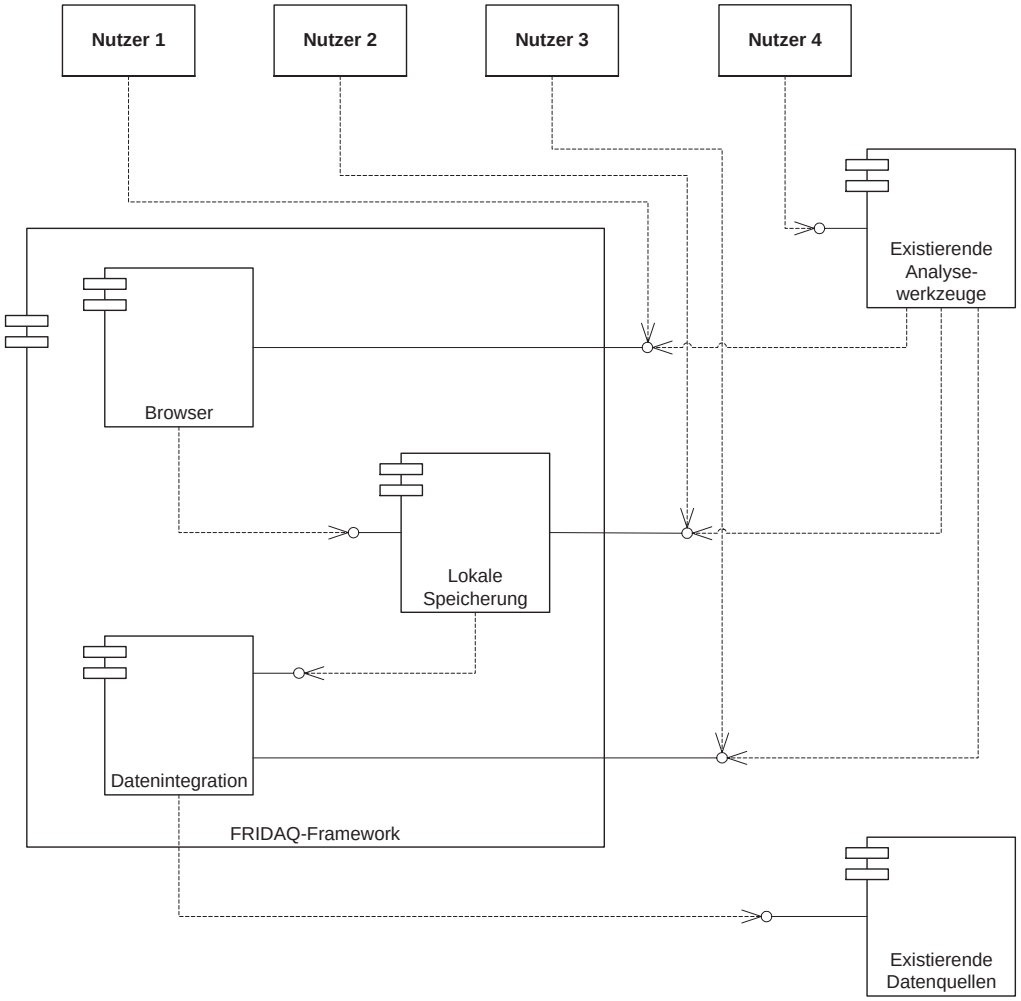

Abbildung 3.6: Architektur des FRIDAQ–Frameworks als UML–Klassendiagramm (nach [Sch02])

reit. Dabei kann festgelegt werden, welche Daten persistent gespeichert werden sollen. Mit der *Browserkomponente* wird ein Werkzeug bereitgestellt, um selbständig und unabhängig auf dem Datenbestand der *lokalen Speicherungskomponente* zu arbeiten und die Prozesse innerhalb der einzelnen Komponenten zu überwachen und zu regeln. Die *Nutzer* und *existierende Analysewerkzeuge* können dabei über wohldefinierte Schnittstellen auf jede der drei Komponenten des Frameworks zugreifen.

Ein mögliches Vorgehen zur Erzeugung eines auf diesem Framework basierenden Systems zeigt die Abbildung 3.7 als UML–Aktivitätsdiagramm. Dabei wird in einem ersten Schritt untersucht, in welchem Maße die Anforderungen des Nutzers durch die Datenintegrationskomponente erfüllt werden. Abhängig vom Ergebnis dieser Analyse, muß die Datenintegrationskomponente womöglich angepaßt werden. Sind die Anforderungen erfüllt oder wurde die Komponente angepaßt, so wird auf Basis eines Nutzerschemas eine nut-

zerspezifische Datenbank in der Sicherungskomponente erzeugt. Abschließend wird noch ein Browser erzeugt, durch den beispielsweise Anfragen an die integrierten Datenquellen formuliert werden.

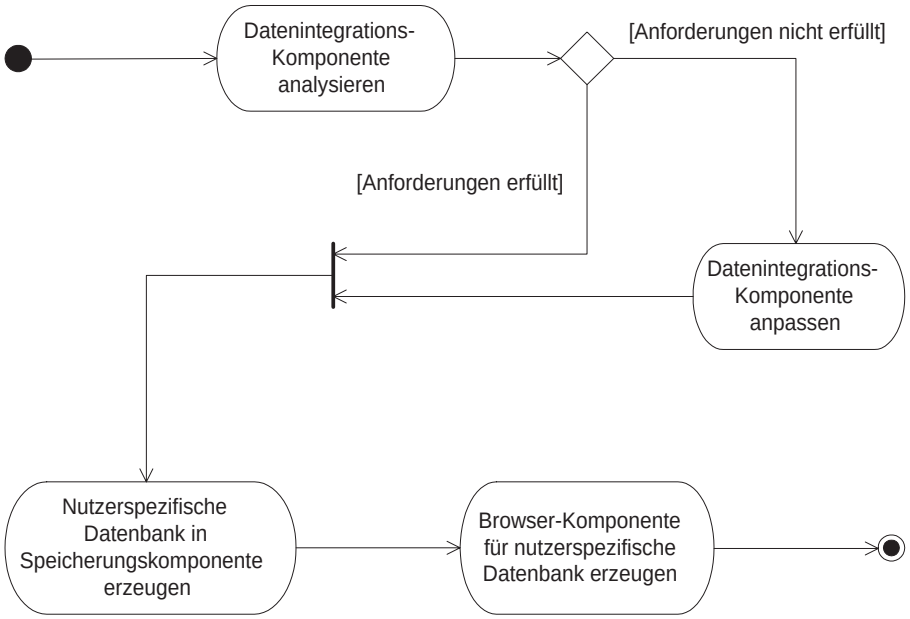

Abbildung 3.7: Vorgehensmodell zur Anwendung des FRIDAQ–Frameworks als UML–Aktivitätsdiagramm (nach [Sch02])

Aufbauend auf diesem Framework wurde in [FHL+02] mit dem Namen BioDataServer der Prototyp eines Integrationssytemes entwickelt, welches, basierend auf Mediatoren, SQL–Anfragen auf integrierte Datenquellen unterstützt. Die Architektur dieses Prototypen ist in der Abbildung 3.8 zu sehen und wurde folgendermaßen beschrieben. Durch eine *Kommunikationsschnittstelle* können die Nutzer und Anwendungen über das TCP/IP auf den Prototypen zugreifen und somit Daten senden und empfangen. Die Steuerung des Prozessmanagements, Änderungen der globalen Integrationsschemata, die Kontrolle der Datenbankadapter und der Abruf von Informationen über den Status der Integration erfolgt durch ein *Administrationsmodul*, bei dem sich authorisierte Benutzer anmelden können.

Über die Komponenten *SQL–Anfrageverarbeitung* und *Datenintegration* wird eine Untermenge der Standardanfragesprache SQL zur Verfügung gestellt. Außerdem werden die verschiedenen globalen Integrationsschemata verwaltet, die die Basis für den Integrationsvorgang bilden. Die Anfrage wird nun in Teilanfragen aufgelöst und in Integrationsoperatoren übersetzt. Nachfolgend wird eine Ausführungshierarchie angelegt und abgearbeitet. Die ermittelten Teilergebnisse werden dann wieder zusammengefügt. Über die *Datenzugriffssteuerung* werden die unterschiedlichen *Datenbankadapter* und der *Cacheadapter* angesprochen. Er lädt dazu zur Laufzeit die einzelnen, benötigten Adapter, verwaltet eine Liste der Adapter, leitet die Schemata der angekoppelten Datenquellen weiter

und verarbeitet eventuell auftretende Fehlermeldung der Adapter. Die *Datenbankadapter* realisieren den homogenen Zugriff auf die Datenquellen durch die Reproduktion einer relationalen Sicht auf den spezifischen Datenbestand und ermöglichen die Operationen zum Zugriff auf die Datenquelle. Wiederholte Anfragen werden durch die Nutzung eines Zwischenspeichers, den *Cacheadapter*, unterstützt.

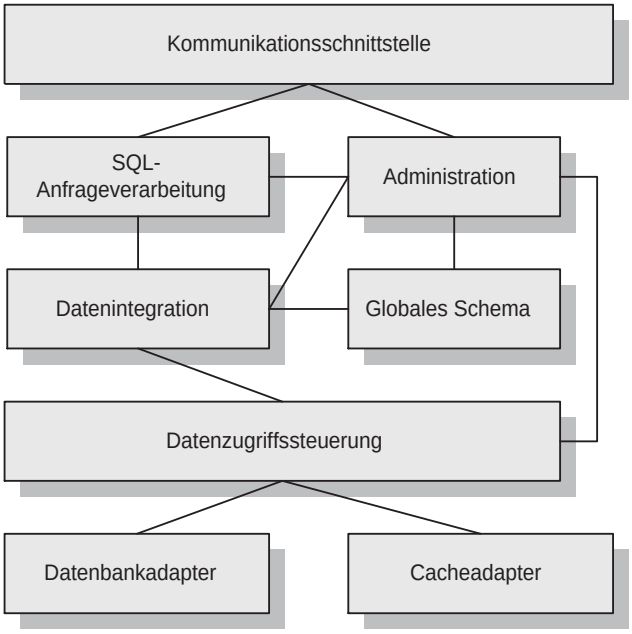

Abbildung 3.8: Architektur der Mediator–basierten Prototypen des BioDataServers (nach [FHL+02])

Der BioDataServer wurde entwickelt, um einer großen Anzahl von Anwendungen im Bereich der Bioinformatik den homogenisierten Zugriff auf nicht–materialisierte, integrierte Daten zu ermöglichen. Die Anzahl der dabei theoretisch integrierbaren Datenquellen wird bisher nur durch die eingeschränkte Anzahl der verfügbaren Adapter beschränkt. Der Vorgang der Adapterentwicklung wird jedoch teilweise schon durch Software–Werkzeuge unterstützt.

K2/Kleisli

Ein Vergleich in [DCB+01] stellt zwei unterschiedliche Ansätze gegenüber: K2, ein lose gekoppeltes föderiertes Datenbanksystem und GUS, ein Data Warehouse. Dementsprechend wird nun kurz K2 und nachfolgend GUS vorgestellt.

Als Nachfolger von BioKleisli [DOTW97] nutzt das K2–System ein umfangreiches Modell von Datentypen, wie Mengen, Multimengen und Listen, die mit Daten aus den Komponentendatenquellen gefüllt werden können. Als Anfragesprache wird OQL genutzt. Wie bei einigen anderen föderierten Systemen werden spezialisierte Komponenten, die

hier als Data Driver bezeichnet werden, eingesetzt, die den Zugriff auf die unterschiedlichen Datenquellen regeln und die entsprechenden Anfrageergebnisse in der speziellen K2–Datenstruktur zurückgeben.

Das K2–System ist als Prototyp verfügbar und kann durch die Verwendung mit Web–Browsern plattformunabhängig eingesetzt werden. Seine Architektur bildete den Ausgangspunkt für die Entwicklung von TAMBIS.

GUS

Im gleichen Projekt wie K2 wurde auch GUS (*Genomics Unified Schema*) als Data Warehouse entwickelt, bei dem auch eine Schemaintegration durchgeführt wurde. Es integriert Daten in einem objektorientierten Schema nach dem Dogma DNS – RNS – Protein. Diese Datenbestände zu annotierten Nukleotid– und Aminosäuresequenzen stammen in erster Linie aus GenBank und SWISS–PROT.

SRS

SRS (*Sequence Retrieval System*) wurde zu Beginn der 90er Jahren am EMBL durch [EA93, EUA96] als Plattform zur Datenintegration entwickelt. Als Basis wurden Datenquellen als Flat–Files genutzt und mit Indexen versehen, so daß ein performanter Zugriff ermöglicht wird. Nach dem Erwerb der Entwicklungs– und Vertriebslizenz durch die Lion Bioscience AG (Heidelberg) wurde SRS komplett überarbeitet und bezeichnet sich nun als derzeitigen Marktführer auf dem Gebiet der bioinformatischen Datenintegration.

Die Anbindung der Datenquellen erfolgt über den Export in ein Flat–File. Dazu muß das Schema der Datenquelle zuerst als Grammatik beschrieben werden. Anschließend werden die Daten als lokale Kopie im Integrationssystem angelegt. Durch die Interaktion mit einer Webschnittstelle werden die vom Nutzer gewünschten Datenquellen, Attribute und Verknüpfungen ausgewählt. Die Anfrageergebnisse werden dann entsprechend ihrer Quelle in Listenform präsentiert. Eine Integration der Ergebnisse wird nicht unterstützt.

Durch die Nutzung einer eigenen Sprache ist es möglich, Suchen über Indexen, boolsche Operationen und Links zwischen Datenquellen zu verknüpfen. Typischerweise wird die Generierung dieser Konstrukte jedoch durch ein Nutzerinterface in Form verschiedener HTML–Seiten übernommen. Die Anfragesprache ist mengenorientiert, so bestimmt beispielsweise die Anfrage

```
[embl-organism:human]
```

alle Einträge der Datenbank EMBL mit der Zeichenkette human im Datenfeld organism.

Das SRS stellt ein Data Warehouse ohne Schemaintegration dar, wobei Kopien der Originaldatenquellen als Flat–Files angelegt und indexiert werden. Mit Hilfe des Paketes SRS

Relational können die relationalen Datenbankmanagementsysteme Oracle und MySQL angebunden werden. Eine nicht–redundante Datenhaltung ist nicht realisiert. Die Vorteile von SRS liegen besonders auf den Gebieten Realisierungsstand, Internetfähigkeit und Plattformunabhängigkeit. Jedoch werden Standardanfragesprachen nicht unterstützt. Dafür werden Schnittstellen für Java, C++, Perl und CORBA, als Ausgabeformate XML und HTML angeboten. In [DOB95] wird SRS nicht als Werkzeug zur vollständigen Integration anerkannt, da weder ein globales Schema noch eine einheitliche Nutzerschnittstelle die Navigation über die angeschlossenen Datenquellen unterstützt.

TAMBIS

Die Bezeichnung TAMBIS [GSN+01] steht für Transparent Access to Multiple Bioinformatics Information Sources, ein Projekt, das ontologiebasiert Daten für Biologen integrieren soll. Durch die Nutzung von Wrappern wird der Zugriff auf die einzelnen Datenquellen realisiert, so daß ein Nutzer nur auf eine virtuelle Datenbank zugreift.

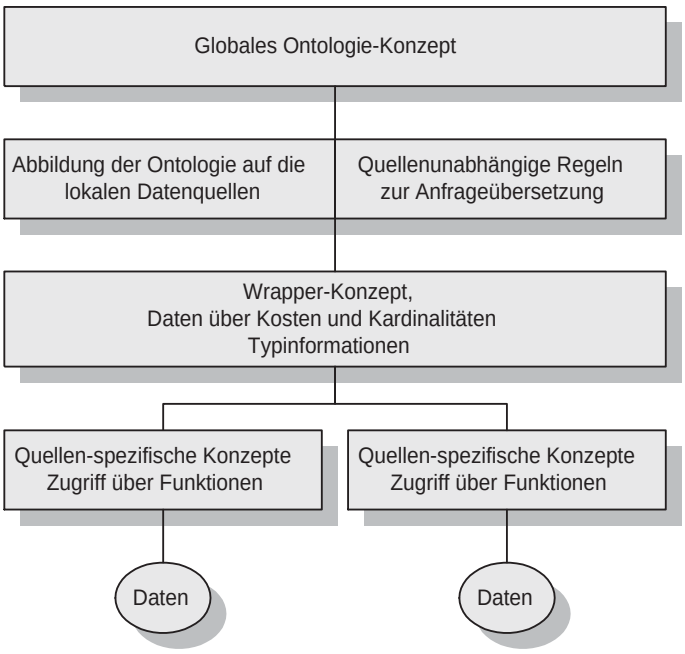

Abbildung 3.9: TAMBIS–Modell (nach [GSN+01])

Die Anfragen werden unabhängig von den betroffenen Datenquellen über eine grafische Nutzerschnittstelle formuliert, anschließend in eine Menge geordneter, quellenspezifischer Anfragen übersetzt und zur Ausführung weitergeleitet. Eine Ontologie, die etwa 1800 biologische Konzepte und ihre Beziehungen umfaßt, stellt dabei Informationen zur Generierung der Nutzerschnittstelle und für die Übersetzung der Anfragen bereit. Das System, wie in Abbildung 3.9 verdeutlicht, besteht aus folgenden Hauptkomponenten:

- der Ontologie mit biologischen Konzepten,
- der grafischen Schnittstelle zur Anfrageformulierung,
- dem Dienstmodell, das die biologische Ontologie mit den Schemata der Datenquellen verbindet,
- dem Modul zur Transformation der Anfragen,
- dem Wrapper–Dienst, der auf die externen Datenquellen zugreift.

Das TAMBIS–System ist den lose gekoppelten föderierten Datenbanksystemen zuzuordnen, da die Anfragesprache CPL verwendet wird und über eine Schemaintegration keine Aussage getroffen wird. Weiterhin wird auf die originalen Datenquellen zugegriffen. Ein Prototyp ist im Internet verfügbar. Eine Nutzung von Integrationsergebnissen mit externen Analysewerkzeugen ist nicht möglich, da Schnittstellen zur Anbindung von Anwendungen nicht existieren. Besonderer Wert wurde auf den ontologiebasierten Ansatz gelegt, eine Betrachtung zur effizienten Anfrageverarbeitung für verteilte, heterogene Datenquellen erfolgt nicht.

3.3 Zusammenfassung

Im ersten Teil des vorliegenden Kapitels wurde eine Auswahl der gebräuchlichsten molekularbiologischen und medizinischen Datenquellen vorgestellt. Dabei sind die molekularbiologischen Quellen in genomische Sequenzendatenquellen, Proteinsequenzedatenquellen, metabolische und regulatorische Datenquellen und Wirkstoffdatenquellen unterteilt worden. Außerdem wurden verschiedene medizinische Datenquellen betrachtet. Anhand ausgewählter Merkmale wurden die einzelnen Quellen gegenübergestellt und auf ihre Eignung und Anforderungen für eine Integration geprüft. Dabei wurde besonderer Wert auf die angebotenen Datenformate und Fragen der Nutzungslizenz gelegt.

Die aktuelle Bedeutung der Integration dieser Datenquellen und die besonderen Anforderungen an diesen Prozeß, die sich aus der komplizierten Struktur und großen Menge der Daten ergeben, wurden im zweiten Teil des Kapitels herausgearbeitet. Anschließend wurden verschiedene Architekturen zur Integration, beispielsweise föderierte Datenbanksysteme und Data Warehouses vorgestellt. Diese verschiedenen Architekturen wurden in einer Vielzahl von Ansätzen und prototypischen Systemen umgesetzt. Eine Auswahl von Systemen konnte mit Merkmalen, die sich in der Literatur bereits bewährt haben, verglichen werden. Dabei wurden auch neuere Arbeiten berücksichtigt.

Aufbauend auf dieser Analyse wurde mit FRIDAQ ein bestehender Architekturvorschlag ausgewählt, der einer Integration von Datenquellen, die im Rahmen der vorliegenden Fragestellung benötigt werden, zugrundeliegen soll. Im Rahmen der hier vorgestellten und bewerteten Ansätze ist dieses System besonders geeignet, um für eine Integration der bereits analysierten Datenquellen eingesetzt zu werden. Diese Entscheidung wird einerseits

durch die Gegenüberstellung der verschiedenen Ansätze in der Tabelle 3.2 bedingt. Andererseits sind die Systeme mit einem vergleichbaren Funktionsumfang, so zum Beispiel DiscoveryLink und SRS, kommerzielle Produkte und können somit nicht uneingeschränkt verändert und genutzt werden.

FRIDAQ wurde als Framework konzipiert, um die Erzeugung eines modularen und konfigurierbaren Systems zu ermöglichen, das einen universellen und standardisierten Zugriff auf integrierte Datenbestände realisiert. Einige der Module dieses Frameworks wurden bereits von Dritten prototypisch im BioDataServer umgesetzt und sollen für die Integration im Rahmen dieser Arbeit wiederverwendet werden.

4 Datenbank für Mutationen und assoziierte Phänotypen

Die Defekte der DNS–Sequenz eines Organismus in Form verschiedener Mutationen[1] bilden die Ursache für die unterschiedlichsten Erkrankungen, die sich im Laufe der Entwicklung als beobachtbare oder auch nicht beobachtbare Phänotypen präsentieren. Ein Ausgangspunkt für eine Genotyp–Phänotyp-Korrelation könnten somit die klinischen Daten eines Patienten sein. Dazu wurde in Kooperation mit der Universität Tübingen und der Klinik für Kinder- und Jugendmedizin des Kreiskrankenhauses Reutlingen unter der Leitung von Prof. K. F. Trefz ein Informationssystem mit dem Namen *Ramedis* (Rare Metabolic Diseases Database) zur Publikation, Sammlung und Analyse seltener, angeborener Stoffwechselerkrankungen entwickelt [MST+01, TSM+02].

Das Informationssystem Ramedis beinhaltet derzeit[2] etwa 700 Fälle, die u.a. durch insgesamt 3800 Symptome und 18500 Laborwerte charakterisiert werden. Im nachfolgenden Kapitel sollen Motivation und Anforderungen für den Entwurf und die Entwicklung dieses Systemes aufgezeigt werden. Die formulierten Forderungen an das System charakterisieren eine universell einsetzbare Datenbank für Mutationen und assoziierte Phänotypen angeborener Stoffwechselerkrankungen. Bereits bestehende Ansätze in diesem Bereich werden anhand der spezifizierten Anforderungen gegenübergestellt und verglichen. Weiterhin werden die besonderen Merkmale von Ramedis herausgestellt und Architektur sowie Protoyp präsentiert.

4.1 Motivation

In der medizinischen Fachliteratur, beispielsweise in Tagungsbänden, werden häufig Fallberichte seltener Erkrankungen veröffentlicht, die es den entsprechend spezialisierten Medizinern erlauben, aktuelle und geeignete therapeutische Maßnahmen zu ermitteln, zu diskutieren oder der medizinischen Fachwelt zu präsentieren. Mittlerweile ist in diesen Publikationsformen natürlich auch eine Online–Recherche durch das WWW über geeignete Schnittstellen möglich.

Leider gehen jedoch viele wertvolle Informationen zum besseren Verständnis derartiger Krankheiten oftmals in der Flut anderer Daten verloren oder sind zum Teil nur schwer

[1] vgl. dazu Abschnitt 2.1.1

[2] Stand August 2004, vgl. dazu Tabelle 4.2

aufzufinden. Außerdem nutzt natürlich jeder Autor eines solchen Fallberichtes seine persönliche Darstellung und Begriffswelt, die jedoch von Synonymen und der subjektiven Beschreibung eines Ausschnittes des Gesamtbildes eines Patienten geprägt ist. Eine informationstechnische Aufbereitung und Auswertung dieser Quellen beispielsweise mit Text–Mining–Methoden würde also sehr aufwendig und nur durch entsprechenden Einsatz von medizinischen Fachexperten möglich, die eine Einordnung der extrahierten Daten in ein geeignetes Modell unterstützen.

Neben publizierten Fallberichten existieren auch weitere Datenquellen, die Informationen über phänotypische Merkmale von Patienten mit Gendefekten enthalten. Einzelne Datenbanken im Internet wurden von Projektgruppen oder Interessengemeinschaften angelegt, um Diagnostik, Therapie und Forschung auf dem Gebiet der Gendefekte zu unterstützen. Drei ausgewählte System werden im nachfolgenden Abschnitt 4.3 vorgestellt und in ihrer Eignung für die vorliegende Zielstellung bewertet.

4.2 Anforderungen

Ausgehend von der Motivation und den allgemeinen Schlußfolgerungen, die nach der Betrachtung bereits zur Verfügung stehenden Datenquellen und ihrer Einschätzung hinsichtlich der Nutzbarkeit auftraten, werden nachfolgend die Schlüsselanforderungen an ein System formuliert, das Mutationen und ihre korrespondierenden Phänotypen in Form von Fallberichten speichert, verwaltet und aufbereitet. Dabei sollen die typischen, an eine Softwareentwicklung gestellten Ansprüche, wie beispielsweise Wartbarkeit, Qualität und weitere nicht aufgeführt werden. Durch Interviews und die Diskussion unterschiedlicher Fallbeispiele wurden die Erwartungen der späteren Nutzer herausgearbeitet und in der nachfolgenden Aufstellung berücksichtigt.

Anbindung weiterer Datenquellen

Ausgehend von einzelnen Fallberichten, die einen Erkrankungsverlauf beschreiben, können weitere Informationen, die jedoch nicht in der Datenbank gespeichert sind, angefordert werden. Dies kann beispielsweise durch geeignete Verbindung mit weiteren Datenquellen durch HTML–Links möglich werden. Dadurch entsteht für den Anwender ein Mehrwert, da ein Teil der aufwendigen und zeitraubenden Recherche in Fachliteratur und Internet erleichtert wird.

Datenbank–Schnittstelle

In der klinischen Forschung ist es häufig notwendig, die Daten einer Reihe von Fallberichten unter Berücksichtigung einer Auswahl von Parametern zu untersuchen. Dadurch ist es notwendig, eine Anbindung anderer Analysesoftware, beispielsweise für erweiterte statistische Auswertungen, zu ermöglichen. Diesem Anspruch könnte durch die Bereitstellung einer ODBC–Schnittstelle zur Datenbank oder die Generierung von Dateien mit standardisierten Datensätzen nachgekommen werden. Ein

Datenformat wie XML wäre aufgrund seiner Flexibilität für diese Funktion besonders geeignet, wird jedoch erst von wenigen kommerziell angebotenen Statistikanwendungen unterstützt.

Datenschutz

Bei der Arbeit mit Patientendaten ist in besonderem Maße auf die datenschutzrechtlichen Aspekte zur Wahrung der Persönlichkeitsrechte zu achten. Diese Forderung ist während des Entwurfsprozesses unbedingt zu beachten und bei der späteren Nutzung des Systemes durch begleitende Maßnahmen, wie die Einholung von Einverständniserklärungen der Patienten durch die entsprechenden Autoren und die Durchführung der Anonymisierung von Patientendaten, sicherzustellen. Entsprechende Regelungen und Durchführungsbestimmungen sind in den Datenschutzgesetzen von Bund und Ländern zu finden. Für Datensammlungen im Rahmen von Forschungsprojekten gelten in bestimmten Fällen vereinfachte Regelungen.

Datensicherheit

Der schreibende Zugriff auf gespeicherte Patientendaten ist nur dem publizierenden Autor gestattet. Den sonstigen, üblichen Forderungen nach Datensicherheit in Bezug auf die Datenhaltung ist durch die Auswahl eines geeigneten Datenbankmanagementsystemes Rechnung zu tragen. Während der Datenübertragung zwischen Nutzer und Datenbank ist ebenso auf eine geeignete Sicherheitsarchitektur zu achten. Diese Forderung sollte in Bezug auf die persönlichen Daten von Patienten, die über das Netz geschickt werden, unbedingt beachtet werden, obwohl diese Daten bereits anonymisiert sind. Jedoch ermöglicht es den Einsatz für klinische Studien und erhöht das allgemeine Vertrauen der Nutzer in die Anwendung, was insbesondere im medizinischen Bereich nicht unterschätzt werden darf.

Dateneingabe und –auswertung

Zur Unterstützung des Anwenders müssen geeignete Software–Werzeuge bereitgestellt werden, die die Eingabe und Auswertung der Daten ermöglichen. Diese Anwendungen müssen natürlich ihrerseits die hier beschriebenen Anforderungen unterstützen. Bei der Entwicklung der Anwendung ist insbesondere auf eine leichte und intuitive Installation sowie Bedienbarkeit zu achten.

Datensammlung durch Fachexperten

Die Eingabe und Pflege der Daten über Mutationen und begleitende Parameter erfolgt primär durch die behandelnden Fachexperten. Dazu müssen diese natürlich durch eine entsprechende Nutzerschnittstelle und den Zugriff auf weitergehende Dienste motiviert werden. Eine manuelle Extraktion von Daten aus ausgewählten, publizierten Fallberichten ist nur ergänzend vorgesehen. Da die Dateneingabe für allgemeine Publikationen oder klinische Studien aus Zeitgründen parallel zur Untersuchung eines Patienten erfolgen sollte, ist eine einfache und schnelle Eingabeprozedur notwendig. Interessierte Fachexperten müssen die Möglichkeit haben, sich für eine Nutzung des Systemes zu registrieren. Eine Überprüfung der Anmeldung sollte dann möglichst zeitnah geschehen.

Erkrankungsunabhängige Einsetzbarkeit
Das System muß so entworfen werden, daß sich eine Anwendung für möglichst viele unterschiedliche Erkrankungen eignet, die durch Gendefekte verursacht werden. Dazu ist es erforderlich, dynamische Datenstrukturen vorzusehen.

Erweiterbarkeit
Das System muß geeignete Möglichkeiten bereitstellen, die eine Erweiterbarkeit der Datenbank um spezifische Charakteristika eines Falles, beispielsweise zur Aufnahme neuer Untersuchungsparameter, Diagnosen oder Therapiemöglichkeiten ohne eine Veränderung des Datenmodells zulassen, die bei der Implementierung noch nicht berücksichtigt wurden. Diese Verwaltung der Parameter sollte nur für Nutzer mit erweiterten Rechten zugänglich sein.

Hohe Vergleichbarkeit gespeicherter Daten
Bei der Eingabe von Daten durch die Nutzer ist sicherzustellen, daß keine synonymen oder homonymen Begriffe für medizinische Sachverhalte verwendet werden, die eine spätere Analyse der gespeicherten Daten erschweren. Wenn für die Eingabe spezifischer Untersuchungsparameter die Angabe von Einheiten erforderlich ist, so sollte dem Nutzer keine freie Texteingabe ermöglicht werden. Vielmehr sind standardisierte, internationale Einheiten, wie die SI–Einheiten für den Anwender zur Auswahl bereitzustellen.

Plattformunabhängigkeit, weltweiter Zugriff, intuitive Bedienung
Zur Erschließung eines möglichst großen Nutzerkreises ist das System plattformunabhängig zu entwerfen und muß einen möglichst weltweiten Zugriff auf die angebotenen Dienste bereitstellen. Da die erwünschten Daten nur durch eine Interaktion mit dem Nutzer gesammelt werden können, muß eine entsprechend komfortable und intuitiv bedienbare grafische Nutzerschnittstelle entworfen werden, die den Zugriff auf das System weitgehend vereinfacht und unterstützt. Ein WWW–Portal kann dazu die erforderlichen Instruktionen, Programme und Web–Seiten bereitstellen.

Referenzierbarkeit
Eine eindeutige Identifizierung der verfügbaren Fallberichte ist für den Nutzer unbedingt erforderlich. Außerdem sollten die einzelnen Fallberichte durch die Konstruktion einer URL referenziert und somit über den Browser leicht wiederauffindbar sein. Damit können dann einzelne Fallberichte auch in anderen Medien verbreitet werden und sind dennoch auf einem aktuellen Stand diskutierbar.

Umfassende Fallberichte
Die durch das System zu verwaltenden Daten sind in Form von umfassenden Fallberichten zu organisieren. Die dazu notwendigen Parameter zur Beschreibung des Falles, seiner Therapie und seines Krankheitsverlauf sind bereitzustellen.

Mit diesen Anforderungen werden die Umrisse eines Systemes skizziert, das sich universell zur Verwaltung und Speicherung von Daten zu Mutationen und Phänotypen einsetzen

läßt, sich aber trotzdem durch seine Ausrichtung von Krankenhausinformationssystemen absetzt. Bei der Zusammenstellung der Anforderungen wurden insbesondere die Wünsche der späteren Nutzer beachtet. Außerdem ist es damit möglich, verwandte Arbeiten anhand einheitlicher Maßstäbe zu beurteilen.

4.3 Diskussion vorhandener Mutationsdatenbanken

Die Idee der Sammlung von Mutationen und assoziierten Phänotypen ist in der medizinischen und molekulargenetischen Welt schon häufig verfolgt worden. Leider beschränken sich die derzeit in der Literatur beschriebenen und im WWW verfügbaren Datenbanken bisher auf die Beschreibung von einzelnen Krankheiten oder Erkrankungsklassen mit festgelegten klinischen Parametern. Dieser Zustand ist hauptsächlich darauf zurückzuführen, daß bei der Entwicklung dieser Werkzeuge für Forschungsprojekte oder spezielle klinische Studien immer nur die Beschreibung eines Ausschnittes innerhalb der Menge der Gendefekte betrachtet werden sollte und somit keinerlei Veranlassung bestand, einen allgemeinen Ansatz zu verfolgen. Die folgenden drei Datenbanken sollen beispielhaft den derzeitigen Stand der Technik auf diesem Gebiet illustrieren.

4.3.1 PAH–Mutationsdatenbank

Im Abschnitt 3.1.1 wurde PAHdb bereits als eine relationale Datenbank für Mutationen des Phenylalaninhydroxylase–Genes (PAH) und die damit assoziierten Phänotypen vorgestellt. Dabei wird ein Ansatz zur Datensammlung verfolgt, der die Nutzer des Systemes einbezieht. Die interessanten Informationen werden über eine Schnittstelle von engagierten Medizinern eingegeben und so der Fachwelt zur Verfügung gestellt.

Bei der Eingabeprozedur werden jedoch nur wenige vordefinierte Auswahlfelder genutzt, so daß viele Freitexteingaben ermöglicht werden, die eine Vergleichbarbeit der Daten erschweren. In einer neueren Veröffentlichung [SHK+03] werden einige Veränderungen der grafischen Nutzeroberfläche beschrieben, die diesen kritischen Prozeß erleichtern sollen und zu weiteren Eingaben motiviert. Diese Dateneingabe wird in mehrere Bereiche (Autorendaten, Referenz, Mutation, Zuordnung, Genotyp–Phänotyp–Beschreibung, In–vitro–Analyse) gruppiert und beschränkt sich auf festgelegte Parameter, die vom Nutzer in ihrer Ausprägung zu bewerten sind. Somit ist eine Erweiterbarkeit des Systemes oder eine erkrankungsunabhängige Einsetzbarkeit nicht mit vertretbarem Aufwand zu erreichen.

Dem Nutzer werden außerdem keine Informationen über die verwendete Softwarearchitektur oder ein direkter Datenbankzugriff zugänglich gemacht. Nur mit Hilfe eines Suchformulares als HTML–Dokument sind Nutzeranfragen über den vorgegebenen Parametern möglich. In [KTJ+97] wurde PAHdb bereits für Untersuchungen zur Genotyp–Pänotyp–Korrelation genutzt. Bis Oktober 2003 wurden nach Angaben der Verantwortli-

chen 462 Mutationen gesammelt.

4.3.2 Mutationsdatenbank für Tetrahydrobiopterin–Mangel (BIODEF und BIOMDB)

Dieses System beleuchtet nach [Bla96] den Tetrahydrobiopterin–Mangel, eine Stoffwechselstörung, die ähnlich wie die PKU zu erhöhten Phenylalanin–Konzentrationen führt, deren Symptomatik sich jedoch nicht durch eine Phenylalanin–arme Diät beeinflussen läßt. Das Gesamtsystem teilt sich in zwei Datenquellen, BIODEF und BIOMDB, die durch Hyperlinks miteinander verbunden sind. Dabei fokussiert BIOMDB auf die Mutation und die sie beschreibenden Informationen. BIODEF hingegen hält ergänzende Patientendaten, wie Geburtsjahr, Geschlecht, Herkunft, einige allgemeine klinische Daten, Laborparameter und korrespondierende Literaturreferenzen.

Parallel zu PAHdb ist jedoch hier ebenfalls nur eine feste Anzahl von Parametern vorgegeben, die über ein Web-Formular ausgefüllt werden und dann vom Autor an die Datenbank zu übertragen sind. Bei der Eingabe können keine vordefinierten Werte eingetragen werden, so daß nur freie Texteingaben möglich sind. Allgemeine Patientendaten, molekulargenetische Untersuchungsergebnisse und Laborparameter werden über eine Suchfunktion zur Auswertung bereitgestellt. Die verfügbaren Patientendaten können durch eine laufende Numerierung identifiziert werden, ein direkter Zugriff auf einen Eintrag über die ID ist jedoch nicht möglich.

4.3.3 ARPKD–Mutationsdatenbank

Eine kompakte Datenbank für die autosomal rezessive polyzystische Nierenerkrankung (ARPKD) bietet die Humangenetikabteilung der RWTH Aachen über das WWW an. Neben Informationen aus medizinischen Veröffentlichungen steht ein Formular zur Verfügung, über das Mutationsdaten an die Datenbank übertragen werden können. Dabei wird sich jedoch auf eine festgelegte Menge von Parametern beschränkt. Dazu zählen allgemeine Patientendaten (Geschlecht, Herkunft), genotypische Beschreibung der Mutation und Informationen über den klinischen Status des Patienten (letzter Untersuchungszeitpunkt, ausgewählte Laborparameter und Symptome). Der vorhandene Datenbestand wird dem Nutzer in Form einer Tabelle präsentiert. Weitergehende Such– oder Analysemöglichkeiten sind nicht vorhanden.

4.3.4 Zusammenfassende Gegenüberstellung

Bei allen hier vorgestellten und in der Literatur beschriebenen Mutationsdatenbanken ist festzustellen, daß sie zwar auf ihrem Spezialgebiet meist eine akzeptierte Lösung zur Sammlung von Mutationen und Phänotypen bilden, jedoch zur Nutzung als allgemeiner

Ansatz nicht geeignet sind. Eine Fokussierung auf die alleinige Nutzung von Daten aus einigen ausgewählten Mutationsdatenquellen durch Integration würde somit die Möglichkeiten des Gesamtsystemes erheblich beschränken, da so nur Anfragen und Untersuchungen zu den integrierten Phänotypen möglich wären.

Der Entwurf einer allgemeineren Architektur und die Implementierung eines entsprechenden Prototypen ist zwar mit einem erheblichen Aufwand verbunden, verspricht jedoch auch einen erweiterten Anwendungsbereich und somit höhere Nutzerakzeptanz. Bei keiner der betrachteten Datenquellen ist die Möglichkeit vorgesehen, eine Anbindung zu weiteren medizinischen oder molekularbiologischen Datenquellen herzustellen. Außerdem ist ein Zugriff auf die vorhandenen Daten über weitere Schnittstellen, beispielsweise mittels Programmier– oder Datenbankanfragesprachen nicht möglich.

Die Tabelle 4.1 ermöglicht anhand des dargestellten Anforderungskataloges eine Gegenüberstellung und Einschätzung verwandter Arbeiten. Dazu wurde die Erfüllung der Anforderung mit den Attributen *ja*, *nein* und *teilweise* versehen. Für den Fall, daß eine qualifizierte Einschätzung eines Merkmales nicht möglich war, wurde das Attribut *unbekannt* verwendet.

Anforderung	PAHdb	BIODEF	AKPKD
Datenbank–Schnittstelle	nein	nein	nein
Anbindung weiterer Datenquellen	nein	nein	nein
Datenschutz	ja	ja	ja
Datensicherheit	unbekannt	unbekannt	unbekannt
Dateingabe und –auswertung	ja	ja	ja
Datensammlung durch Fachexperten	ja	ja	ja
Erkrankungsunabhängige Einsetzbarkeit	nein	nein	nein
Erweiterbarkeit	nein	nein	nein
Vergleichbarkeit der Daten	teilweise	nein	teilweise
Plattformunabhängigkeit, usw.	ja	ja	ja
Referenzierbarkeit	nein	teilweise	teilweise
Umfassende Fallberichte	nein	teilweise	teilweise

Tabelle 4.1: Gegenüberstellung existierender Mutationsdatenbanken anhand ausgewählter Anforderungen

Aus den in der Tabelle gegenübergestellten Beispielen für Mutationsdatenbanken geht deutlich hervor, daß ein solches universell einsetzbares System derzeit noch nicht verwendet wird. Durch die beobachtete Beschränkung der Anwendungen auf die aktuell untersuchten Fragestellungen bleibt eine weitergehende, unkomplizierte Nutzung der gesammelten Daten und der zugrundeliegenden Architektur unerreichbar.

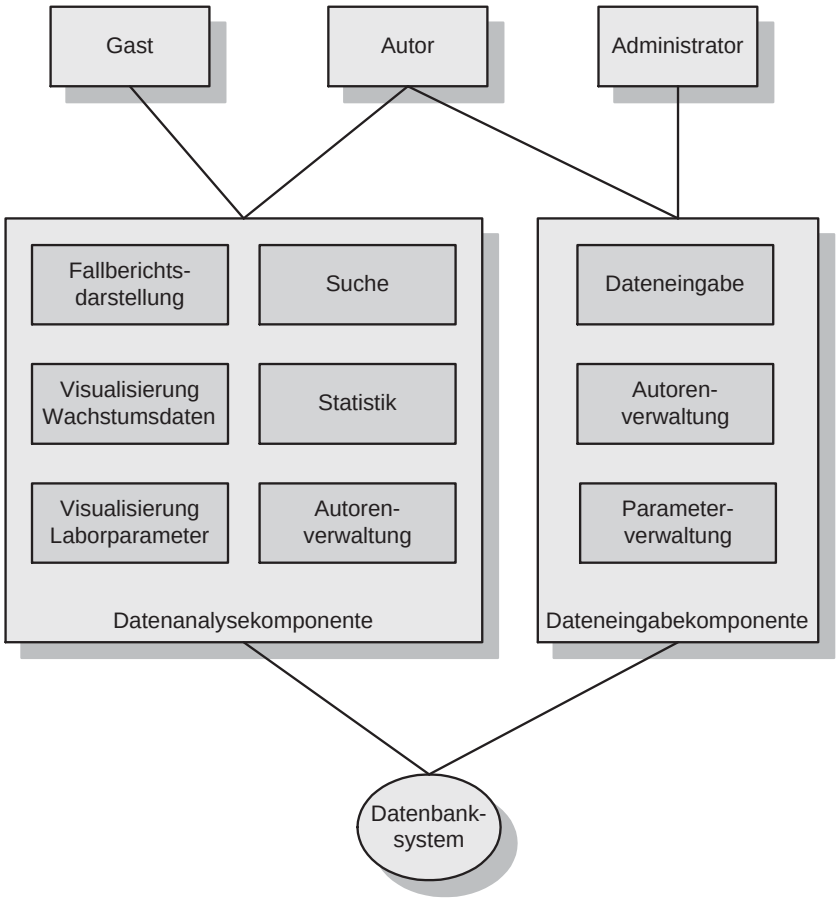

Abbildung 4.1: Architektur des Ramedis–Systemes mit unterschiedlichen Nutzern, Analyse– und Eingabekomponente, Datenbanksystem

4.4 Architekturvorschlag

Nachdem in den vorangehenden Abschnitten der aktuelle Stand der Entwicklungen im Bereich der Datenbanken für Mutationen und assoziierte Phänotypen im allgemeinen und an drei Beispielen aufgezeigt und grundlegende Anforderungen an eine Architektur für einen allgemeinen Ansatz beschrieben wurden, soll nun ein konkreter Architekturvorschlag präsentiert werden. Dieser wurde auf Basis der dargestellten Anforderungen entwickelt und soll dem Nutzer die Möglichkeit bieten, für eine große Anzahl von Erkrankungen, die durch Gendefekte verursacht werden, Informationen im Stil ausführlicher Fallberichte zu aggregieren und zu analysieren.

Der in Abbildung 4.1 als Komponentendiagramm dargestellte Architekturvorschlag beinhaltet vier Hauptbereiche. Diese werden in der folgenden Übersicht erläutert. Dabei werden auch die jeweils untergeordneten Teilkomponenten vorgestellt, die die erforderlichen Funktionalitäten zur Erfüllung der ausgeführten Anforderungen realisieren.

Nutzer

Die verschiedenen Nutzer, die parallel auf Dienste zugreifen können, erhalten unterschiedliche Rechte, die davon abhängig den Zugriff auf die einzelnen Software–Module ermöglichen oder verhindern. So besitzt ein *Gast* nur die Möglichkeit, lesend auf Fallberichte zuzugreifen. Veränderungen am Datenbestand sind für ihn nicht möglich. Das Anlegen und Ergänzen von Fallberichten ist den jeweiligen *Autoren* vorbehalten. Diese müssen sich registrieren lassen und erhalten ein aktives Nutzerkonto nur nach einer Prüfung ihrer persönlichen Anmeldungsdaten, so daß die Qualität der eingegebenen Fallberichte sichergestellt wird. Zur Erweiterung der vorhandenen Untersuchungsparameter und beispielsweise der Einheiten sind nur *Administratoren* berechtigt. Der *Administrator* hat auch die Möglichkeit, Nutzerkonten von *Autoren* zu aktivieren oder zu deaktivieren.

Dateneingabekomponente

Eine Softwarekomponente zur Dateneingabe realisiert mit drei verschiedenen Modulen die Funktionen, die zur Pflege des Datenbestandes notwendig sind. Über die *Autorenverwaltung* werden von einem Administrator die Zugangsdaten angelegt und verwaltet, so daß authorisierten Autoren der Zugang zum System ermöglicht wird. Diese nutzen nun das *Dateneingabemodul*, um Daten über Mutationen und Phänotypen in Form von Fallberichten zu speichern. Der Autor hat hier die Möglichkeit, einen neuen Fallbericht anzulegen oder einen bereits vorhanden zu editieren. In dieser Anwendung besitzt der Autor natürlich nur den Zugriff auf die von ihm eingerichteten Fälle. Zur Eingabe von Untersuchungsergebnissen werden vordefinierte Untersuchungsparameter zur Auswahl angeboten, so daß Probleme mit Synonymen weitgehend vermieden werden können. Mit Hilfe der *Parameterverwaltung* können von den Administratoren zusätzliche Parameter angelegt werden, die zur Beschreibung eines Falles notwendig werden können.

Datenanalysekomponente

Zur Analyse der gespeicherten Fallberichte steht eine Datenanalysekomponente zur Verfügung, die den Nutzern den lesenden Zugriff auf die gespeicherten Daten erlaubt. Dazu wird ein *Suchmodul* angeboten, das neben der einfachen Suche nach einzelnen Parametern der Fallberichte auch eine kombinierte Anfrage im Sinne des fallbasierten Suchens unterstützt. Durch Interaktion mit dem Nutzer wird anschließend die gewünschte *Fallberichtsdarstellung* geladen, die durch zwei Visualisierungskomponenten für Wachstumsdaten und Laborparameter vervollständigt wird. Diese Darstellungen haben sich in der klinischen Praxis zur Analyse und Präsentation von Entwicklungstendenzen bereits als sehr hilfreich erwiesen. Über die *Autorenverwaltung* können sich Mediziner zur Dateneingabe anmelden. Ein einfaches *Statistikmodul* gibt allgemeine Informationen zum aktuellen Datenbestand, beispielsweise die Anzahl der Fallberichte und der zugehörigen Parameter.

Datenbanksystem

Ein Datenbanksystem, das die grundlegenden Anforderungen an nicht–redundante Datenhaltung, Datensicherheit, usw. erfüllt, wird genutzt, um die anfallenden Daten

zu verwalten und auf Anforderung bereitzustellen.

An dieser Stelle soll nun ebenfalls der konzeptionelle Entwurf der Datenbank vorgestellt werden. Dazu wurde in der Abbildung 4.2 eine vereinfachte, formale Beschreibung der benötigten Informationsstrukturen durchgeführt. Mit besonderer Beachtung der Forderung nach Erweiterbarkeit und hoher Vergleichbarkeit der Daten wurde eine Vorauswahl von Untersuchungsparametern in speziellen Relationen angelegt, beispielsweise für Laboruntersuchungen und ihre Parametereinheiten. Diese wurden in Zusammenarbeit mit Medizinern erstellt und bilden somit die Grundlage für den anzureichernden Datenbestand. Der Vorteil dieser vordefinierten Listen liegt außerdem in ihrer leichten Ergänzbarkeit. Somit wird dem Nutzer eine umfangreiche Auswahl von Untersuchungsparametern präsentiert, die von Fachexperten geprüft wurde und dennoch so dynamisch ist, daß sie bei Bedarf erweitert werden kann.

Im Gegensatz zu herkömmlichen Veröffentlichungen wird zur Eingabe von Daten in Ramedis eine Vorgehensweise angewandt, die die spätere Vergleichbarkeit und Verwertbarkeit der gesammelten Daten erhöht. Durch den Einsatz der großen Anzahl von vorgegebenen Werten, die zur Charakterisierung der Fallberichte ausgewählt werden können, werden Konflikte durch Synonyme und Homonyme vermieden. Um eine möglichst vollständige Sicht auf einen Fall zu ermöglichen, wird neben allgemeinen Angaben die Eingabe einer Vielzahl von Eigenschaften ermöglicht, z.B. klinische Symptome, Laboruntersuchungen, molekulargenetische Untersuchungen, Diät und Medikation.

4.5 Realisierung

Aufbauend auf den in den vorangehenden Abschnitten formulierten Anforderungen an eine universell einsetzbare Mutationsdatenbank und den Ergebnissen der Betrachtung verwandter Datenbanken wurde ein Architekturvorschlag entwickelt, für den in diesem Abschnitt die Realisierung eines Prototypen ausgeführt werden soll. Dazu werden kurz die angewendeten Methoden und Resultate dargestellt. Das entwickelte System läßt sich analog zur vorgeschlagenen Architektur in Abbildung 4.1 in zwei Komponenten unterteilen, durch die der Nutzer mit dem System interagiert: eine Eingabe- und eine Auswertungskomponente. Der Zugriff auf beide Anwendungsbereiche wird primär über die HTML–Dokumente des WWW–Portales *www.ramedis.de* realisiert. Den Umfang der Daten, mit denen ein spezifischer Fallbericht angereichert werden kann, wird an einem Beispiel im Anhang C ausgeführt.

4.5.1 Dateneingabekomponente

Die Eingabekomponente ist eine Java–Applikation, die zur Eingabe und Veränderung fallspezifischer Daten verwendet wird. Dabei muß sich der Nutzer als Autor identifizieren

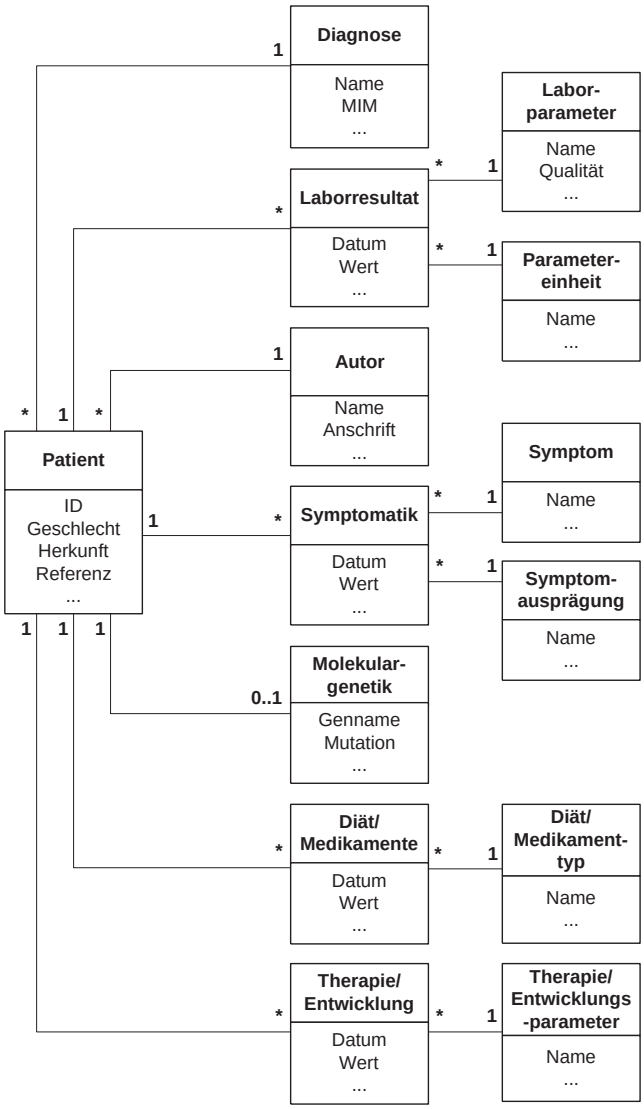

Abbildung 4.2: Vereinfachtes Relationenschema der Ramedis–Datenbank

und kann nur auf die von ihm angelegten Datensätze zugreifen. Durch Verwendung der Java–Web–Start–Technologie läßt sich diese Anwendung auf dem Computer des Nutzers plattformunabhängig über das WWW starten, dabei werden auch automatische Software–Aktualisierungen unterstützt.

Die Abbildung 4.3 zeigt eine Maske des Eingabewerkzeuges mit allgemeinen Falldaten. Die Werte zur Diagnose, ethnischen Herkunft der Eltern und Nationalität des beschriebenen Patienten sind aus den vordefinierten Listen auszuwählen. Die Datumsangaben zum Geburtstag und Diagnosestellung werden nur zur internen Berechnung benötigt, da aus Datenschutzgründen diese persönlichen Informationen nicht der Öffentlichkeit über die Auswertungskomponente zugänglich gemacht werden können. Aus diesem Grunde wird für zeitabhängige Parameter nur das Alter berechnet und angezeigt. In der Abbildung 4.3 der Dateneingabekomponente sind deshalb diese Felder auch gelöscht worden. Der Zugriff auf weitere Eingabeparamter erfolgt über die Auswahl *Symptoms, Laboratory, Diet/Drugs, Therapy/Development, Molecular Genetics* und *Pictures*. Nach der Speicherung eines neuen Falles wird dem Autor ein Fallschlüssel angezeigt, über den eine spätere Identifikation des Falles ermöglicht wird. Mit Hilfe dieses Fallschlüssels können vom Autor dann beispielsweise Daten einer Nachuntersuchung hinzugefügt werden.

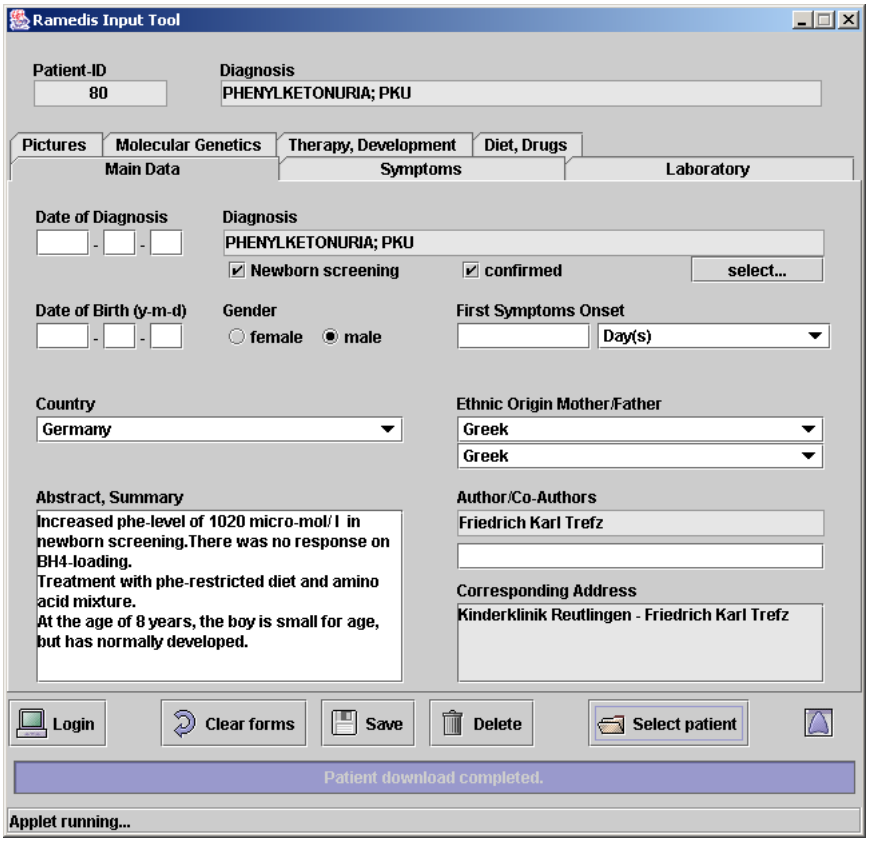

Abbildung 4.3: Allgemeine Daten zu einem Fallbericht in der Dateneingabekomponente von Ramedis

4.5.2 Datenauswertungskomponente

Die Auswertung erfolgt über eine Reihe von dynamisch generierte HTML–Seiten, die die Falldaten anonymisiert zusammenfassen und wie in der Abbildung 4.4 darstellen. Hier findet sich ein Teil der allgemeinen Informationen aus der Dateneingabekomponente wieder. Außerdem sind die Ergebnisse der molekulargenetischen Untersuchung zu erkennen. Im rechten Fenster befindet sich das Resultat der Visualisierung des Wachstumsparameters Länge zusammen mit Perzentilen aus statistischen Angaben im Kindes– und Jugendalter aus verschiedenen deutschen Stichproben nach [KWK+01]. In dem vorliegenden Fall wurde die Länge des Patienten in einem Alter von drei bis zehn Jahren dargestellt und die Vergleichsgraphen seiner entsprechenden Vergleichsgruppe eingezeichnet.

Zur Unterstützung der Navigation innerhalb des umfangreichen Datenbestandes wird eine Schnittstelle für Suchanfragen bereitgestellt. Dabei ist eine Suche über verschiedenen, einzelnen Attributen, z.B. Autor oder Diagnose möglich, die durch die Nutzung der vordefinierten Listen erheblich vereinfacht wird. Zur Unterstützung der Suche nach Kombinationen von Untersuchungsparametern wurde außerdem eine fallbasierte Suche implementiert, die dann eine nach Ähnlichkeit sortierte Menge von Fällen zurückliefert [Hin00]. Dazu kann eine Auswahl von qualitativen Laborparametern, Symptomen und ethnischer Herkunft getroffen werden. Diese Anfrageschnittstelle wird trotz ihrer Zugehörigkeit zur Datenauswertungskomponente im nachfolgenden, eigenen Abschnitt erläutert, da sie im System eine besonders wichtige Aufgabe erfüllt.

4.5.3 Fallbasierte Anfrageschnittstelle

Auf einen fallbasierten Ansatz wurde zurückgegriffen, da in herkömmlichen Expertensystemen das Wissen durch Regeln, Frames, Klausel, usw. formalisiert werden muß. Resultat dieses Vorgehens ist somit häufig ein langwieriger Wissensaquisitionsprozeß, da das Expertenwissen oft nicht in entsprechend formalisierter Form vorliegt. Der Experte gewinnt sein Wissen aus langjährigen Erfahrungen mit ähnlichen Problemstellungen, das er sich im Kontext gelöster Probleme merkt. Dieses Erfahrungswissens muß nun strukturiert, um Faktenwissen ergänzt und entsprechend der Wissensrepräsentation formalisiert werden. Das fallbasierte Schließen (Case–based Reasoning, CBR) verfolgt nun das Ziel, Erfahrungswissen zum Lösen zukünftiger Probleme heranzuziehen und durch Hinzufügung neuer gelöster Probleme dieses Wissen zu ergänzen. Weitergehende Grundlagen zu diesem Verfahren sind im Abschnitt 2.2.2 zu finden.

Ausgangspunkt für die Anwendung fallbasierter Anfragen innerhalb des vorgestellten Informationssystemes ist somit die geeignete Strukturierung und Bereitstellung des für die Problemlösung zur Verfügung stehenden Wissens. Um dieses Wissen zu formulieren, wird mit der Wissensrepräsentation ein geeigneter Formalismus angeboten. Dazu werden im allgemeinen die folgenden Zielsetzungen beachtet.

- Definition der Objekte der untersuchten Domäne

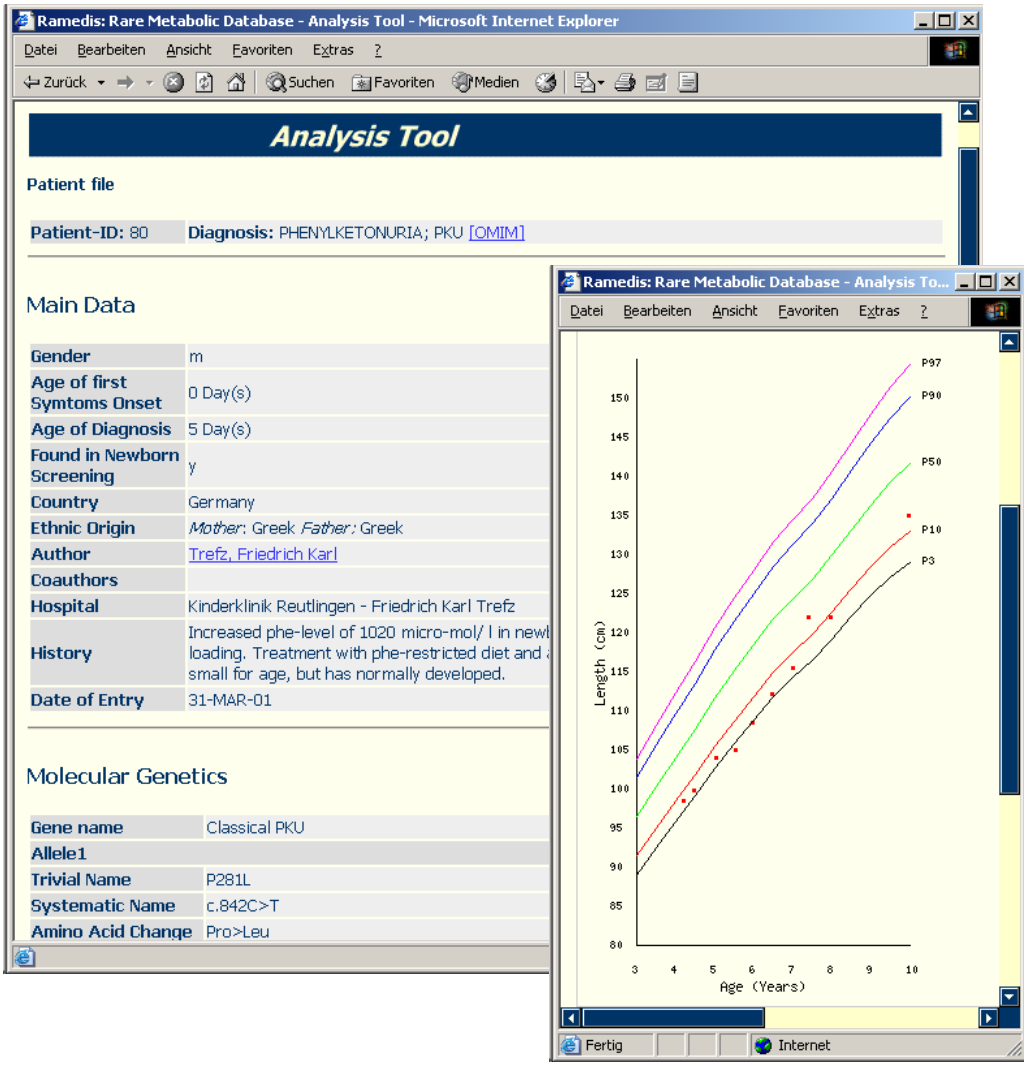

Abbildung 4.4: Darstellung des Ausschnittes eines Fallberichtes und der Visualisierung des Wachstumsparameters Länge in perzentiler Darstellung in der Auswertungskomponente von Ramedis

- Beschreibung der Abhängigkeiten und Beziehungen zwischen den Objekten der Domäne
- Erhaltung der Domänenstruktur
- Vermeidung der Repräsentation von redundanten Informationen
- Unterstützung der Ähnlichkeitsbewertung und der Lösungsadaption

In der Literatur werden unterschiedliche Auffassungen über die formale Beschreibung, Repräsentation und Verwaltung von Fällen und Falldaten vertreten. In dieser Arbeit wird letztlich die Implementierung der Wissensrepräsentation primär durch die Nutzung eines relationalen Datenbankmodelles beeinflußt. Dazu werden die relevanten Falldaten in eine Problembeschreibung und die entsprechende Lösung unterteilt. Dabei besteht die Problembeschreibung aus einer Menge von Merkmalen in den Bereichen allgemeine Patientendaten (z.B. Geschlecht, ethnische Herkunft), Laboruntersuchungsergebnisse, Symptome sowie Therapie– und Entwicklungsinformationen. Die Problemlösung des Falles ist eine Diagnose, die mit einem boolschen Attribut versehen ist, das den Grad der Sicherheit dieser Diagnose angibt. Für die Ausführung von fallbasierten Anfragen ist typischerweise die Angabe von Problembeschreibung und Problemlösung im Rahmen der Falldaten ausreichend. Für den praktischen Einsatz und zur Erhöhung des Nutzwertes wird auch eine Reihe von Zusatzinformationen hinterlegt, beispielsweise das Erfassungsdatum des Falles, beliebige Textdaten in Form eines Kommentares oder Abbildungen. Ein Beispielfall wird dazu im Anhang C dargestellt.

In der Abbildung 4.2 ist ein vereinfachtes Relationenschema der Datenbank abgebildet. Ausgehend von einer zentralen Relation, in der für jeden Fall oder Patienten ein Eintrag angelegt wird, und dem zugehörigen Autor, der für diesen Fallbericht verantwortlich ist, werden alle weiteren relevanten Merkmale in gesonderten Relationen abgelegt, die unterschiedliche Merkmalsdomänen abdecken. Der Entwurf dieser Relationenstruktur soll einen effizienten Zugriff auf die abgelegten Fälle in der Fallbasis ermöglichen. Dabei wurde die Unterstützung der angewendeten Strategie zur Vorauswahl von Fällen, des relationalen Retievals, durch die Anlage geeigneter Indexe über den entsprechenden Relationen ermöglicht.

Das im Abschnitt 2.2.2 vorgestellte Prozeßmodell muß für einige Anwendungen des CBR nicht vollständig realisert werden. Unter der Bezeichnung fallvergleichendes CBR kann für bestimmte Anwendungen auf eine Adaption der gefundenen Lösung verzichtet werden. Zu dieses Anwendungsszenarien, wie Klassifikation, Entscheidungsunterstützung und diagnostische Anwendung gehört auch die in der vorliegenden Arbeit untersuchte Problemstellung.

Entsprechend der Retrieval–Phase des CBR–Zyklus nach [AP94] in der Abbildung 2.6 sind die ähnlichsten Fälle zu ermitteln. Dieser Schritt kann bei sehr großen Fallbasen einen großen Berechnungsaufwand nach sich ziehen, so daß eine geeignete Vorauswahl der Menge der potentiell relevanten Fälle notwendig ist, um die zu untersuchen-

de Fallmenge zu reduzieren. In [Wes96] werden dazu drei verschiedene Verfahren zur Durchführung einer geeigneten Vorauswahl präsentiert.

Sequentielles Retrieval
: Auf jeden Fall der Fallbasis wird das Verfahren zur Ähnlichkeitsbestimmung anwendet. Dadurch wird die Vollständigkeit und Korrektheit des Vorgehens sichergestellt und ein entsprechender Algorithmus ist einfach zu implementieren. Jedoch wird damit ein hoher Berechnungsaufwand in Kauf genommen.

Indexorientierten Retrieval
: Eine Indexstruktur wird für die Vorauswahl relevanter Fälle genutzt. Dabei muß jedoch der Aufwand für die Implementierung des Verfahrens und die Indexerstellung beachtet werden.

Relationales Retieval
: Hier wird die Fallbasis um die Fälle reduziert, die keinen Beitrag zu Lösung liefern können. Durch die Abfrage eines passenden Ausschnittes der Fallbasis, den möglichen Lösungskandidaten, wird die Bearbeitung auf diesen Ausschnitt begrenzt. Die dabei gefundenen Lösungskandidaten sind nun in einer sequentiellen Phase anhand des vorgegebenen Ähnlichkeitsmaßes anzuordnen.

Innerhalb der vorgestellten, fallbasierten Anfrageschnittstelle wird das relationale Retrieval auf der Basis des MAC/FAC–Modelles [FGL95], wie in der Abbildung 4.5 darbestellt, durchgeführt. Wie bereits angedeutet, wird dazu in der MAC–Phase (*many are called*) eine Vorauswahl möglicher Lösungskandidaten aus der Fallbasis getroffen. Die dabei typischerweise genutzten Indexierungsmechanismen und Anfragetypen lehnen sich an die Strukturen und Methoden an, die im Rahmen relationaler Datenbanksysteme genutzt werden [Wes96]. Anschließend wird in der FAC–Phase (*few are choosen*) die Ähnlichkeit zwischen der aktuellen Problemstellung und den ausgewählten Lösungskandidaten ermittelt und der ähnlichste Fall ausgewählt.

Für die Durchführung einer Vorauswahl in der Fallbasis im Rahmen der MAC–Phase können prinzipiell vier Ansätze unterschieden werden, die sich durch die Charakterisierung, wann ein Fallbeispiel als potentieller Lösungskandidat ausgenommen wird, differenzieren lassen.

Gleichheit
: Ein Fall der Fallbasis gehört dann zur Menge der potentiellen Lösungskandidaten, wenn alle Merkmale des Falles mit den entsprechenden Merkmalen der aktuellen Problemsituation übereinstimmen.

Partielle Gleichheit
: Ein Fall der Fallbasis gehört dann zur Menge der potentiellen Lösungskandidaten, wenn mindestens ein Merkmal des Falles mit den Merkmalen der aktuellen Problemsituation übereinstimmt.

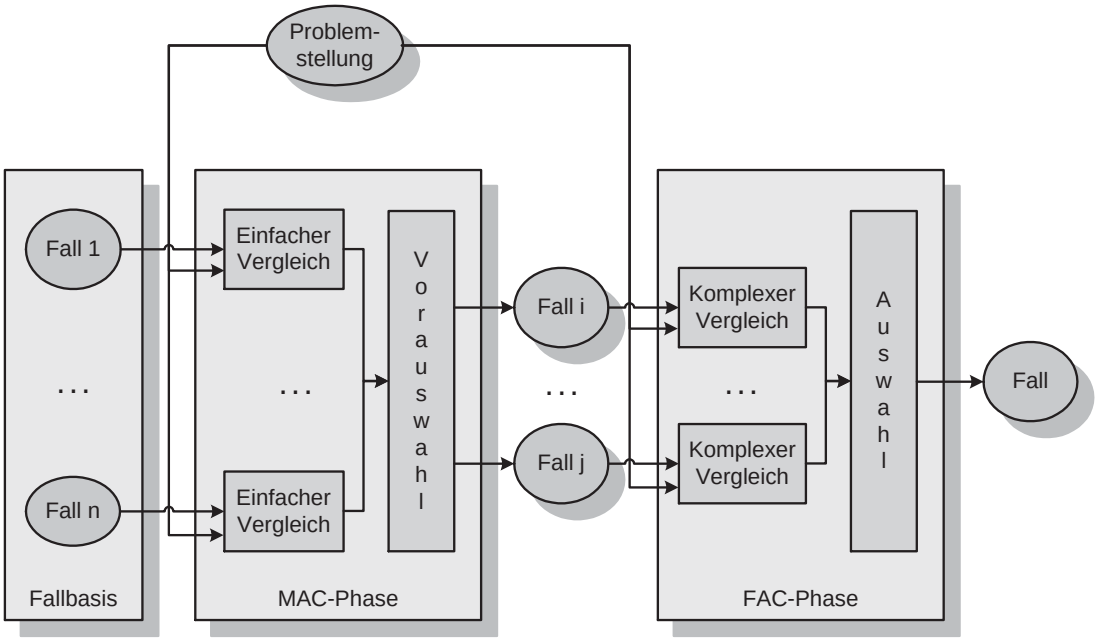

Abbildung 4.5: Darstellung des MAC/FAC–Modell (nach [FGL95])

Lokale Ähnlichkeit
: Ein Fall der Fallbasis gehört dann zur Menge der potentiellen Lösungskandidaten, wenn alle Merkmale des Falles sehr ähnlich zu den entsprechenden Merkmalen der aktuellen Problemsituation sind.

Partielle lokale Ähnlichkeit
: Ein Fall der Fallbasis gehört dann zur Menge der potentiellen Lösungskandidaten, wenn mindestens ein Merkmal des Falles sehr ähnlich zu den Merkmalen der aktuellen Problemsituation ist.

Auf Basis dieser Ansätze zur Vorauswahl lassen sich direkt Anfrageprädikate an ein relationales Datenbanksystem formulieren. Im Rahmen der hier vorgestellten Anfrageschnittstelle wurde die partielle Gleichheit verwendet, um mit geringem Berechnungsaufwand eine akzeptable Vorauswahl zu ermöglichen. Dazu werden für die über eine grafische Nutzerschnittstelle eingegebenen Merkmale der aktuellen Problemstellung alle in der Fallbasis hinterlegten Fälle selektiert, für die zumindest ein Merkmal übereinstimmt.

Bei Verfahren zur Einschränkung der Fallbasis auf eine Menge potentieller Lösungskandidaten können jedoch zwei Gruppen von Retrievalfehlern auftreten, die unterschiedlichen Einfluß auf die Vollständigkeit der Ergebnismenge haben. Dabei wird zwischen α- und β–Fehlern unterschieden.

α–Fehler
: In der Vorauswahl wird ein Fall der Fallbasis nicht berücksichtigt, der jedoch aus-

reichend ähnlich zur gegebenen Problemsituation ist. Dies führt zu unvollständigen Lösungsmengen, da Fälle bei der Ähnlichkeitsberechnung unberücksichtigt bleiben.

β–Fehler
Ein Fall, der zur gegebenen Problemsituation nicht ausreichend ähnlich ist, wird dennoch in der Vorauswahl berücksichtigt. Dieser Fehler erhöht den Aufwand für die Ähnlichkeitsberechnung, da Fälle untersucht werden, für die dieses Vorgehen unnötig ist. Die Qualität des Retrievalergebnissen wird jedoch nicht beeinflußt.

Neben der Vorauswahl potentieller Lösungskandidaten ist beim Abrufen (Retrieve) der gespeicherten Fälle die Ähnlichkeit zum Suchfall zu berechnen. Dazu ist in einem ersten Schritt die lokale Ähnlichkeit der einzelnen Merkmale zu bestimmen, um anschließend mit Hilfe eines globalen Ähnlichkeitsmaßes aus den Teilähnlichkeiten die Gesamtähnlichkeit des Falles zu ermitteln. Als allgemeinen Ansatz für die Berechnung der globalen Ähnlichkeit $sim(F)$ für einem Fall F wurde die folgende Gleichung genutzt, die eine Wichtung der Einzelähnlichkeiten über die Festlegung der Gewichte $w_i \in [0, 1]$ erlaubt.

$$sim(F) = \sum_{i=1}^{n} w_i \cdot sim_i \quad \text{mit} \quad \sum_{i=1}^{n} w_i = 1$$

Im vorliegenden Fall werden nun die drei Merkmale Symptom, Laboruntersuchung und ethnische Herkunft betrachtet, so daß sich eine konkretisierte Formel zur Berechnung der globalen Ähnlichkeit für diese drei Merkmalsmengen $Symptome_F$, $Laborparameter_F$, $Herkunft_F$ ergibt.

$$\begin{aligned} sim(F) &= w_1 \cdot sim(Symptome_F) \\ &+ w_2 \cdot sim(Laborparameter_F) \\ &+ w_3 \cdot sim(Herkunft_F) \end{aligned}$$

Bei [Goo96] wird betont, daß bei der Bestimmung der Ähnlichkeit zweier Mengen von Merkmalen, einzelne Merkmale teilweise generell oder aber bei einer bestimmten Lösung unterschiedlich relevant sind. So ist beispielsweise das Symptom Haarverlust nur für wenige Stoffwechselerkrankungen typisch, während im Gegensatz dazu das Symptom Fieber häufig bei verschiedenen Erkrankungen auftritt. Es ist also von Vorteil, diese Informationen über die unterschiedliche Bedeutung verschiedener Merkmale innerhalb des Datenbestandes zu hinterlegen und in die Ähnlichkeitsberechnung einfließen lassen zu können.

Eine derartige Gewichtung von Symptomen und Laboruntersuchungen wird durch einen Ansatz erreicht, der Merkmalen, die in wenigen Erkrankungen relevant sind, ein höheres

Gewicht gibt, während ein Merkmal, das häufig in Fällen auftritt, ein geringes Gewicht zugewiesen bekommt. Die Festlegung der Gewichte in Abhängigkeit des entsprechenden Merkmales wurde durch die beteiligten medizinischen Fachexperten durchgeführt, kann jedoch bei einer Überprüfung der Werte noch verändert werden. Insgesamt wurden mehrere Wichtungsgruppen gebildet, die eine Bedeutung des Merkmals von unwichtig bis extrem wichtig ermöglichen.

Am Beispiel Merkmalsmengen der Symptome $Symptome_F$ wird nun die Zusammenführung der einzelnen lokalen Ähnlichkeiten für die betrachteten Merkmale unter Beachtung der gewählten Gewichtung dargestellt. Dazu wird in einem ersten Schritt der lokalen Ähnlichkeitsberechnung anhand der vorgegebenen Symptome das maximal mögliche Gewicht $SympGew_{max}$ ermittelt. Ausgehend vom aktuell selektierten Fall F der Fallbasis wird die Summe der korrespondierenden Gewichte $SympGew_F$ berechnet. Daraus ergibt sich die Ähnlichkeit der Symptome des Falles zur vorgegebenen Problemsituation.

$$sim(Symptome_F) = \frac{SympGew_F \cdot 100}{SympGew_{max}}$$

Basierend auf dem Prozeßmodell des CBR–Zyklus und dem MAC/FAC–Modell wurde ein Algorithmus entwickelt, der zu einer Suchanfrage ähnliche Fälle innerhalb der gesammelten Patientendaten in der Mutations– und Phänotypendatenbank Ramedis sucht. Als einfaches Beispiel für die Nutzung des CBR–Ansatzes auf bestehenden Daten wurde eine grafische Nutzerschnittstelle implementiert, die Anfragen auf dem Datenbestand über den Parametern Symptom, Laboruntersuchung und ethnische Herkunft ermöglicht. Die Auswahl der einzelnen Werte beschränkt sich dabei auf bereits im Datenbestand genutzte Ausprägungen der Merkmale. Zusätzlich wird die Gewichtung der Einzelmerkmale bei der Berechnung der Fallähnlichkeit spezifiziert. Nach der Beschreibung der Problemsituation als Anfrage über die Suchparameter werden in einem Zwischenschritt potentiell ähnliche Fälle innerhalb der Datenbank untersucht und mit den entsprechenden Ähnlichkeitsbewertungen zwischengespeichert. Dieses Vorgehen wird vereinfacht in der Abbildung 4.6 verdeutlicht.

Resultat dieser Verarbeitungsschritte ist eine Liste von Fällen, die eine bestimmte Ähnlichkeit zu der in der Anfrage beschriebenen Problemsituation aufweisen. Zusätzlich zum korrespondierenden Ähnlichkeitswert werden die zur Berechnung herangezogenen Merkmale und die Falldiagnose aufgeführt. Eine beispielhafte Anfrage könnte entsprechend mit dem Symptom Fieber in Verbindung mit der Untersuchung des Laborwertes Phenylalanin konstruiert werden.

Diese fallbasierten Suchanfragen können zur Unterstützung der Differentialdiagnostik genutzt werden, die auf die Abgrenzung und Identifikation einer bestimmten Erkrankung innerhalb einer Menge von symptomatisch ähnlichen Krankheiten ausgerichtet ist. Durch die paralle Sammlung von molekulargenetischen Untersuchungen in Ramedis, die krankheitsrelevante Mutationen feststellen, werden bereits spezifische Korrelationen von Genotyp und Phänotyp für einzelne Fälle verfügbar.

```
Algorithmus FallAehnlichkeit (Suchparameter, Fallbasis) {
  Initialisiere eindeutige Session;
  Berechne maximale Gewichte;
  Waehle potentiell aehnliche Faelle aus der Fallbasis;
  Fuer jeden vorselektierten Fall {
    Berechne Gewicht der Symptome;
    Berechne Gewicht der Laboruntersuchungen;
    Berechne Gewicht der ethnischen Herkunft;
    Berechne die resultierende Fallaehnlichkeit;
    Speichere Fallinformationen mit Session und Aehnlichkeit;
  }
  Praesentiere berechnete Aehnlichkeiten fuer jeden Fall;
}
```

Abbildung 4.6: Algorithmus für die Suche von ähnlichen Fällen in Ramedis anhand der Parameter Symptom, Laboruntersuchung und ethnische Herkunft

4.5.4 Datenbanksystem

Zur Speicherung der Falldaten wird ein relationales Oracle–Datenbankmanagementsystem eingesetzt, in dem die in Abbildung 4.2 dargestellte, vereinfachte Datenbankstruktur mit insgesamt 33 Relationen implementiert wurde. Dabei finden sich die verschiedenen klinischen Parameter, wie Symptomatik, Laborresultat, Molekulargenetik, Diät/Medikamente und Entwicklung/Therapie im Schema wieder.

Die Tabelle 4.2 zeigt eine entsprechende Übersicht der in Ramedis gehaltenen Daten. Im oberen Teil der Tabelle sind Informationen über die registrierten Nutzer und gespeicherte Fallberichte verzeichnet. Im mittleren Teil der Tabelle ist die Anzahl der vordefinierten Untersuchungsparameter zu finden. Die Anzahl der für die Pflege der Fallberichte eingetragenen Parameter ist im unteren Teil der Tabelle zu sehen. Diese wurden von den Autoren ausgewählt und als Patientendaten verzeichnet. Diese Übersicht zeigt, daß das entwickelte Informationssystem eine große Menge von vordefinierten Parametern bereitstellt und bereits viele Fallberichte im System dokumentiert wurden.

4.6 Zusammenfassung

Als Basis für die geplante Untersuchung von Genotyp–Phänotyp–Korrelationen wurden in diesem Kapitel Motivation und verwandte Ansätze zur Umsetzung einer Datenbank für Mutationen und assoziierte klinische Phänotypen beschrieben. Dabei zeigte sich, daß die bisher angelegten, derartigen digitale Datenbestände auf ausgewählte Krankheiten und Krankheitsfamilien fokussiert wurden und sich somit nicht für die vorliegende Arbeit

Autoren	56
Fallberichte	689
Vordefinierte Parameter	
Diagnosen	364
Symptome	633
Laboruntersuchungen	1355
Therapie und Entwicklung	83
Diät und Medikamente	144
Übermittelte Parameter	
Symptome	3791
Symptome pro Fall	5.5
Laboruntersuchungen	18391
Laboruntersuchungen pro Fall	26.69
Therapie und Entwicklung	3297
Diät und Medikamente	1598
Abbildungen	83
Molekulargenetische Untersuchungen	295

Tabelle 4.2: Übersicht zum aktuellen Datenvolumen in Ramedis (Stand August 2004)

eignen. Dieser Sachverhalt wurde außerdem beispielhaft an drei Datenbanken illustriert.

Anhand der verwendeten Arbeiten und in Zusammenarbeit mit Projektpartnern aus dem medizinischen Umfeld wurden allgemeine Anforderungen an eine universell einsetzbare Architektur für eine solche Datenbank formuliert. Aufbauend auf diesen Anforderungen wurden die drei betrachteten, verwandten Datenquellen gegenübergestellt und auf die Existenz der erwünschten Merkmale hin untersucht. Von diesem Anforderungskatalog ausgehend konnte außerdem ein Architekturvorschlag präsentiert werden, der durch eine enge Kooperation mit den zukünftigen Anwendern und die Nutzung aktueller Technologien besonders geeignet ist, um weltweit umfassend Daten zu genetisch verursachten Erkrankungen in Form von Fallberichten zu sammeln.

Dazu wurden unter dem Oberbegriff *Ramedis: Rare Metabolic Diseases Database* verschiedene Software–Werkzeuge mit grafischen Nutzeroberflächen implementiert, die es erlauben, Daten innerhalb der Fallberichte anzulegen, zu editieren und auszuwerten. Durch die Unterstützung der Visualisierung von Laboruntersuchungesergebnissen und Wachstumsparametern werden Fachexperten motiviert, die von ihnen untersuchten Fälle in das System einzutragen. Die so gesammelten Daten bieten die Möglichkeit, auf hohem qualitativen Niveau phänotypische Daten für eine Untersuchung von Genotyp–Phänotyp–Korrelationen zu nutzen.

Im Laufe dieses Jahres ist der Anschluß weiterer Stoffwechselzentren in Deutschland geplant. Außerdem soll ein Review–Komitee eingerichtet werden, das die Qualität der eingegebenen Daten überwacht und bewertet. Durch Erweiterungen der Software–Module

des Ramedis–Systemes ist außerdem die Unterstützung nationaler und internationaler klinischer Studien vorgesehen.

5 Ähnlichkeiten und Beziehungen in Life–Science–Datenbeständen

In den vorausgegangenen Kapiteln wurden verschiedene Life–Science–Datenquellen vorgestellt und analysiert. Sie bilden die Datenbasis, auf der eine Integration durchgeführt wird, um zielorientiert und nutzerzentriert ausgewählte Daten aus den verschiedenen Quellen in eine Integrationsdatenbank zu überführen, auf der dann Analysewerkzeuge zum Einsatz kommen können. Dieser integrierte Datenbestand wird dabei Informationen über unterschiedliche Aspekte eines biologischen Systemes beinhalten. Bei der Untersuchung der vielfältigen Zusammenhänge innerhalb und zwischen den einzelnen Komponenten des Systemes ist es erforderlich, für Fragestellungen des Nutzers die Möglichkeit der Suche nach ähnlichen Resultaten zu bieten, da eindeutige Ergebnisse oft nicht zu finden sind.

Für die Berechnung von Ähnlichkeiten lassen sich in unterschiedlichen Fachgebieten der Informatik, z.B. der Bioinformatik, Dokumentenverarbeitung und bei der Suche in Multimedia–Datenbanken, Beispiele finden. So besitzen molekulare Sequenzen, wie DNS, RNS oder Aminosäuresequenzen bei hoher Sequenzähnlichkeit in der Regel ebenfalls eine hohe funktionale oder strukturelle Ähnlichkeit. Das heißt, durch den Vergleich von zwei Sequenzen kann über die Struktur auf die zu erwartende Funktion geschlossen werden.

Auch die Berechnung von Ähnlichkeiten auf der Grundlage des Inhaltes von Textinformationen ist nun bereits seit langem Gegenstand der wissenschaftlichen Forschung und wird auch im Rahmen kommerzieller Produkte eingesetzt. Mit der immer weiteren Verbreitung multimedialer Datenbanken gewinnt ebenfalls das Problem der effizienten Suche in derartigen Datenbeständen derzeit stark an Bedeutung und wird in unterschiedlichen Anwendungen bereits unterstützt. Dabei werden entweder die Parameter eines Bildes, wie Schlüsselworte, Datum und Größe ausgewertet oder sichtbare Merkmale der Bildes, wie Formen und Texturen untersucht. Darauf basierend ist dann die Suche nach Ähnlichkeiten zu einem vorgegebenen Musterbild oder bestimmten Merkmalen möglich.

Der Begriff der Ähnlichkeit und verschiedene gebräuchliche Ähnlichkeitsmaße wurden bereits im Abschnitt 2.2.3 angesprochen. Dabei wurde herausgestellt, daß unterschiedliche Ansätze existieren, um ein Maß für die Ähnlichkeit zu definieren. Dabei hängt die Definition des entsprechenden Maßes natürlich von der vorliegenden und zu untersuchenden Domäne ab. Eine allgemeine Berechnungsvorschrift zu finden, ist deshalb kaum möglich. Außerdem führt der Versuch, das Maß möglichst ausdrucksstark zu modellieren, häufig

zu aufwendigen und komplexen Vergleichen. Deshalb werden auch Verfahren genutzt, die nur anhand bestimmter Merkmale von untersuchten Objekten einen Anhaltspunkt für eine mögliche Ähnlichkeit finden sollen. Da sie jedoch meist nur Ausschnitte der Realwelt betrachten, können ihre Ergebnisse nur als Näherungswerte betrachtet werden, so daß der Nutzer die Qualität gefundener Lösungen aus seiner Sicht beurteilen muß.

In diesem Kapitel sollen nun Möglichkeiten vorgestellt und diskutiert werden, innerhalb spezifischer Domänen im molekularbiologischen Anwendungsfeld Ähnlichkeiten zu berechnen. Dazu wurden vier ausgewählte Domänen näher betrachtet, die durch die Integration molekularbiologischer Datenbestände und die Anforderungen des laufenden Projektes entsprechend Abschnitt 1.1 bestimmt wurden. Neben der Betrachtung der domänenspezifischen Ähnlichkeiten wird auch die geeignete Zusammenführung dieser Einzelergebnisse der Domänen zu einer Ähnlichkeit auf Szenarioebene untersucht. Das heißt, ausgehend von einem Modell, in dem über einem integrierten Datenbestand nutzerspezifisch Domänen definiert wurden, wird ebenfalls ein Ähnlichkeitsmaß definiert, das verschiedene Pfade vom Genotyp zum Phänotyp auf dieser Basis gegenüberstellt.

5.1 Ähnlichkeit auf Domänenebene

Bevor die Untersuchung von Ähnlichkeiten auf der Ebene des globalen Schemas der Integrationsdatenbank vorgenommen werden kann, sollen ähnlich dem Bottom–up–Ansatz Teilaufgaben definiert und bearbeitet werden. Dazu wurde eine Strukturierung des Datenbestandes auf semantischer Ebene in Informationsdomänen vorgenommen. Diese ergeben sich meist zwangsläufig aus der vorliegenden medizinischen oder molekularbiologischen Fragestellung und den integrierten Life–Science–Datenquellen. Eine Abgrenzung des Begriffes der Informationsdomäne wird in der Definition 5.1 vorgenommen. In den weiteren Ausführungen wird jedoch häufig auch die Bezeichnung Domäne synonym für Informationsdomäne angewendet.

Definition 5.1 (Informationsdomäne) *Eine Informationsdomäne ist eine strukturierte Zusammenfassung von Daten aus einem oder mehreren Life–Science–Datenbeständen, die im Rahmen einer Anwendung durch einen Fachexperten festgelegt wird.*

Bei der Vorstellung des Ähnlichkeitsbegriffes und seiner Anwendung in den unterschiedlichen Zusammenhängen wurde deutlich, daß die verschiedenen Verfahren zur Berechnung von Ähnlichkeiten zwischen Objekten auch zu Ergebnissen führt, die nur im Kontext der jeweiligen Fragestellung sinnvoll analysiert werden können. Um aber diese Verfahren innerhalb einer gemeinsamen Analyseumgebung zusammenzuführen, ist es notwendig, verbindliche Anforderungen an ein Ähnlichkeitsmaß festzulegen. Diese Anforderungen werden in der folgenden Definition 5.2 umrissen.

Definition 5.2 (Ähnlichkeitsmaß) *Sei M eine Menge von Merkmalen oder Problembeschreibungen. Eine Funktion $sim : M \rightarrow [0, 1]$ heißt Ähnlichkeitsmaß auf M, wenn gilt:*
$sim(x, x) = 1$ *für alle $x \in M$ (reflexiv) und*
$sim(x, y) = sim(y, x)$ *für alle $x, y \in M$ (symmetrisch).*

In den folgenden Abschnitten werden nun verschiedene Verfahren vorgestellt, die eine Bewertung von Ähnlichlichkeiten innerhalb bestimmter Domänen medizinischer und molekularbiologischer Datenbestände ermöglichen. Dazu wurden vorhandene Arbeiten aus der Literatur herangezogen. Es sollen jedoch auch eigene Ansätze diskutiert werden. Eine Gegenüberstellung der untersuchten Verfahren in tabellarischer Form wird dann diesen Abschnitt zusammenfassen.

5.1.1 Domäne der klinischen Phänotypen

Aus Sicht der unterschiedlichen Fachbereiche der Biologie oder Medizin kann der Begriff des Phänotypen natürlich auch unterschiedliche Bedeutungen besitzen. Für einen Molekularbiologen, der sich mit der Proteinsynthese beschäftigt, repräsentiert sich der Phänotyp eines bestimmten Gendefektes vielfach schon in einem veränderten Genprodukt. Demgegenüber steht der Mediziner in der Klinik, dessen Hauptaugenmerk sich nun vielmehr auf die Symptomatik oder die veränderten Laborparameter eines Patienten richten wird. In diesem Abschnitt sollen Möglichkeiten untersucht und vorgestellt werden, klinische Phänotypen miteinander zu vergleichen. Dazu werden als Grundlage in erster Linie die Daten des Ramedis–Systemes, einer Datenbank für Mutationen und assoziierte Phänotypen, die im Kapitel 4 präsentiert wurde, benutzt.

Die diesem Informationssystem zugrundeliegende Datenbank hält Daten über durchgeführte Untersuchungen, beispielsweise Symptome und Laborwerte, aber auch therapeutische Maßnahmen, wie die Verabreichung von Medikamenten oder die Verordnung bestimmter Diäten. Diese standardisierten ein– oder mehrdimensionalen Merkmale eignen sich zur Untersuchung der Ähnlichkeit zwischen zwei Fallberichten, jedoch darf die große Abhängigkeit des klinischen Phänotypes von Umwelteinflüssen nicht vernachläßigt werden. Eine Übersicht von Merkmalen, die für einen Vergleich zwischen Fallberichten herangezogen werden können, zeigt die Tabelle 5.1. Außerdem wurden in dieser Aufstellung die Antworttypen aufgeführt und eine Einschätzung, ob eine Belegung dieses Merkmales für einen Fallbericht obligatorisch ist.

Zur Verbindung der verschiedenen Merkmale eines Falles kann beispielsweise der gewichtete euklidsche Abstand genutzt werden, der in der Definiton 2.8 vorgestellt wurde. In der Tabelle 5.2 wird beispielhaft die Ähnlichkeit von zwei Fällen zu einem Referenzfall anhand von ausgewählten Merkmalen bestimmt. Dabei wurden die gewichteten Einzelergebnisse summiert, und können so die Werte $0 \leq sim(x, y) \leq 1$ annehmen. Dieses Beispiel zeigt eine Auswahl von Untersuchungsparametern aus dem Ramedis–System, die für eine Suche nach einem PKU–Patienten denkbar wären. Dazu wurden als allgemeine Merkmale die Teilnahme am Neugeborenenscreeningprogramm und die Bestäti-

Merkmal	Antworttypen	obligatorisch
Alter	numerisch	ja
Geschlecht	One–Choice	ja
Ethnische Herkunft Mutter/Vater	One–Choice	nein
Neugeborenenscreening	Ja/Nein	nein
Diagnose	One–Choice	ja
Diagnose bestätigt	Ja/Nein	ja
Laborwert (qualitativ)	Multiple–Choice	nein
Laborwert (quantitativ)	numerisch	nein
Symptom	Multiple–Choice	nein
Therapie	Multiple–Choice	nein

Tabelle 5.1: Übersicht der Merkmale für den klinischen Phänotyp innerhalb der Ramedis–Datenbank mit Antworttypen nach [Goo96]

gung der gestellten Diagnose gewählt. Neben dem quantitativen Laborwert Phenylalanin im Blut wurden außerdem die Auffälligkeit im neonatalen Screening und das Vorliegen einer Tetrahydrobiopterin–Sensitivität herangezogen.

Merkmal	Gewicht	Referenzfall	Fallbeispiel 1	Fallbeispiel 2
Teilnahme am Neugeborenenscreening	0.1	ja	ja	ja
Diagnose bestätigt	0.1	ja	nein	ja
Laborwert (quantitativ) Phenylalanin in mg/dl	0.3	13	5	14
Symptom Auffälligkeit im neonatalen Screening	0.25	ja	ja	ja
Symptom Tetrahydrobiopterin–Sensitivität	0.25	ja	nein	ja
Ähnlichkeit			0.46	0.98

Tabelle 5.2: Beispiel für den Vergleich von klinischen Phänotypen

Die Tabelle zeigt auf der linken Seite die verwendeten Merkmale, die benutzten Gewichte und die Merkmalsausprägungen des Referenzfalles. Die rechte Hälfte der Aufstellung verdeutlicht die Unterschiede der einzelnen Merkmalsausprägungen für zwei Beispielfälle, die in der Vorauswahl berücksichtigt wurden. Im unteren Teil der Tabelle befindet sich die für die Beispielfälle berechnete Ähnlichkeit zum Referenzfall für die ausgewählten Merkmale. Die in diesem Abschnitt vorgestellten, verschiedenen Merkmale eines Falles zur Bewertung seiner Ähnlichkeit zu einem Referenzfall gehen bisher über die im

Ramedis–System implementierten Ergebnisse hinaus.

5.1.2 Domäne der biochemischen Reaktionen und Reaktionsketten

Die Informationen über die am Stoffwechsel beteiligten Metabolite und biochemischen Prozesse werden in vielfältiger Form und Präsentation in unterschiedlichen Datenquellen gehalten und dem interessierten Nutzer zur Verfügung gestellt. Von besonderer Bedeutung sind dabei die Daten über die biochemischen Reaktionen, die in ihrem Zusammenspiel die Basis für die Komplexität des Metabolismus legen. Die an den Reaktionen beteiligten Metabolite sind ebenso wie Informationen über die Verbindung von mehreren biochemischen Einzelreaktionen zu komplexen Reaktionsketten in den unter Abschnitt 3.1 betrachteten Datenquellen verfügbar.

Jedoch ist es im typischen Anwendungsfall häufig nicht möglich, bei der Suche nach spezifischen Einzelreaktionen und ihren Vorgängern oder Nachfolgern in Reaktionsketten alle beteiligten Metabolite manuell zu spezifizieren, um den konkreten Datenbankeintrag zielgenau zu finden. In manchen Fällen muß außerdem davon ausgegangen werden, daß die entsprechenden Datenbankinhalte unvollständig oder gar unrichtig sind. Diese Fehler im Datenbestand können aus unterschiedlichen Gründen auftreten. So ist es denkbar, daß bei integrierten Datenbeständen Inkonsistenzen während des Integrationsprozessen entstanden sind. Weiterhin sind manche biochemischen Prozesse noch nicht in ihrem gesamten Umfang beschrieben.

Aus diesen Gründen ist es nun notwendig, bei der Suche von Elementen in der Domäne der biochemischen Reaktionen und Reaktionsketten in vorhandenen Datenbeständen eine Möglichkeit zu finden, die einserseits eine unscharfe Suche erlaubt, jedoch die wahrscheinlich enorme und unübersichtliche Anzahl von Ergebnissen geeignet bewertet, so daß der Nutzer in seinem Bemühen, schnell und präzise auf die gewünschten Daten zuzugreifen, unterstützt wird. Die Gewichtung der Resultate einer unscharfen Suche innerhalb eines Datenbestandes nach spezifischen Eigenschaften, wie chemischer Funktion oder räumlicher Struktur, ist somit wenig sinnvoll. Vielmehr sollten dabei aus der Sicht der Informatik die beteiligten Metabolite in ihrer Anordnung innerhalb der chemischen Reaktion betrachtet werden. Eine endgültige Abschätzung und Bewertung des Ergebnisses der Datenbankanfrage bleibt jedoch immer dem entsprechenden Fachexperten, hier beispielsweise dem Biologen oder Biochemiker, überlassen.

Die Formalisierung von biochemischen Reaktionen und aus diesen zusammengesetzten Reaktionsketten oder Stoffwechselwegen wird in der Literatur typischerweise mit der Speicherung von Datensätzen in Datenbanken oder der Vorhersage, Modellierung und Simulation von komplexen, metabolischen Netzwerken verbunden. Diese hat sich nach [PPW+03] von der Betrachtung von Einzelreaktionen zu komplexen Netzwerken weiterentwickelt. So wurden beispielsweise zuerst durch Experimente stoichiometrische Daten zu einzelnen chemischer Reaktionen gewonnen. Durch die fortschreitende Katalogisierung vieler Einzelreaktionen wurde dann die Beschreibung traditioneller Stoffwech-

selwege ermöglicht. Der derzeitige Stand dieser Entwicklung ist die Fähigkeit zur mathematischen Beschreibung von komplexen Stoffwechselwegen durch Netzwerke.

Als Beispiel für die Anwendung stoichiometrischer und thermodynamischer Daten kann das METATOOL von [PSN+99] zur Untersuchung metabolischer Netzwerke genannt werden. Die Untersuchung von Metabolic Pathways wird außerdem durch verschiedene Methoden unterstützt, die in anderen Gebieten der Informatik weit verbreitet sind. So wird zur Identifikation von Reaktionsketten aus biochemischen Reaktionen bei [SLP+01] ein graphentheoretischer Ansatz verfolgt, während [KZL00] Petri–Netze zur Analyse von metabolischen Netzwerken aus verschiedenen Datenbanken nutzt. Die Berechnung von Stoffwechselwegen auf der Basis von Regeln, die beispielsweise aus KEGG oder anderen Datenquellen gewonnen werden können, wird in [Hof96] und [OGFK98] vorgestellt.

Um nun innerhalb der Domäne der biochemischen Reaktionen die Ähnlichkeit zwischen vorliegenden Reaktionen zu untersuchen, muß vorher natürlich eine Formalisierung des ablaufenden Prozesses und seiner begleitenden Umstände vorgenommen werden. Die wohl trivialste Möglichkeit eine biochemische Reaktion zu formalisieren, beschränkt sich auf die Betrachtung von Substraten und Produkten der jeweiligen Reaktion. Diese Daten zu den Reaktionsbeteiligten sind auch in vielen Datenquellen verfügbar. Das metabolische Gemisch wird somit in der Definition 5.3 vorläufig auf eine Menge von Vor- und Nachbedingungen reduziert.

Definition 5.3 (Biochemische Reaktion als Substrat–Produkt–Beziehung) *Eine biochemische Reaktion r kann als Substrat–Produkt–Beziehung mit einem 2–Tupel der Form $r = (V, N)$ beschrieben werden, wobei V die Menge von Vorbedingungen und N die Menge der Nachbedingungen bezeichnet.*

Im Reaktionsschema wird wie im Beispiel 5.1 durch den Reaktionspfeil ($\rightarrow$) die entsprechende Reaktionsrichtung angegeben. Die jeweilige Reaktionsrichtung schließt jedoch nach [Bud89] die Reversibilität der betrachteten chemischen Reaktion nicht aus. Nur wenn der reversible Charakter der Reaktion besonders betont werden soll, werden Substrate und Produkte durch zwei in entgegensetzte Richtungen weisende Pfeile ($\leftrightarrow$) verbunden. In den folgenden Betrachtungen soll jedoch davon ausgegangen werden, daß durch den Reaktionspfeil die Reaktionsrichtung festgelegt wird, so daß sich die Mengen der Vor- und Nachbedingungen bestimmen lassen. Meistens hat jedoch eine Vertauschung von Vor- und Nachbedingungen keinen Einfluß auf das letztliche Ergebnis der Berechnung.

Beispiel 5.1 (Substrate und Produkte einer biochemischen Reaktion)
Typischerweise wird in der Chemie eine Reaktionsgleichung in der Form

$$S \rightarrow P$$

angegeben. Dabei bezeichnet S die Menge der Substrate der Reaktion und P die Menge der Produkte. In der Praxis entsteht beispielsweiser folgender Ausdruck.

$$2H_2 + O_2 \rightarrow 2H_2O$$

Unter Verwendung der Definition 5.4 kann bereits eine einfache Ähnlichkeit zwischen zwei biochemischen Reaktionen berechnet werden. Bei den nachfolgenden Untersuchungen wird natürlich von $V_1 \cup V_2 \neq \emptyset$ und $N_1 \cup N_2 \neq \emptyset$ ausgegangen.

Definition 5.4 (Ähnlichkeit in Substrat–Produkt–Beziehungen) *Die Ähnlichkeit zwischen zwei gegebenen biochemischen Reaktionen $r_1 = (V_1, N_1)$ und $r_2 = (V_2, N_2)$ berechnet sich nach*

$$sim(r_1, r_2) = \frac{|V_1 \cap V_2|}{|V_1| + |V_2|} + \frac{|N_1 \cap N_2|}{|N_1| + |N_2|}.$$

Beispiel 5.2 (Ähnlichkeit von Substrat–Produkt–Beziehungen) *Entsprechend der vorangehenden Berechnungsvorschrift wird für den Vergleich der Reaktion r_1*

$$\textit{L-Arginin} + H_2O \rightarrow \textit{L-Ornithin} + \textit{Harnstoff}$$

mit folgender Reaktion r_2

$$\textit{L-Arginin} + \textit{4-Aminobutanoat} \rightarrow \textit{L-Ornithin} + \textit{4-Guanidinobutanoat}$$

eine Ähnlichkeit berechnet werden. Für diese beiden Reaktionen ergeben sich aus den Substraten und Produkten nun die entsprechenden Mengen der Vor– und Nachbedingungen.

$$\begin{aligned} V_1 &= \{\textit{L-Arginin}, H_2O\}, \\ N_1 &= \{\textit{L-Ornithin}, \textit{Harnstoff}\}, \\ V_2 &= \{\textit{L-Arginin}, \textit{4-Aminobutanoat}\}, \\ N_2 &= \{\textit{L-Ornithin}, \textit{4-Guanidinobutanoat}\} \end{aligned}$$

Nach der Definition 5.4 wird folgende Ähnlichkeit ermittelt.

$$sim(r_1, r_2) = \frac{1}{2}$$

Zur genaueren Modellierung einer biochemischen Reaktion werden nach [Hof96] jedoch vier spezifische Mengen von Metaboliten benötigt. Betrachtet man ein metabolisches Gemisch, so wird die Menge der Metaboliten, die die Ausgangssituation beschreibt, als *Vorhergemisch* bezeichnet. Ebenso wird mit dem *Nachhergemisch* die Endsituation der biochemischen Reaktion skizziert. Außerdem werden diese Reaktionen dadurch gekennzeichnet, daß biochemische Strukturen durch enzymatisch gesteuerte Prozesse modifiziert und in eine neue Struktur überführt werden. Neben diesen enzymatischen Vorgängen existieren auch Metaboliten, die die Reaktionsgeschwindigkeit der biochemischen Reaktion beeinflussen. Diese Metabolitengemische, die einen bestimmten positiven oder negativen Einfluß auf den Ablauf der Reaktion haben, werden als *Fördergemisch* und *Hemmgemisch* bezeichnet. Eine Formalisierung der biochemischen Reaktion unter Verwendung der vier unterschiedlichen Mengen führt zur Definition 5.5.

Definition 5.5 (Biochemische Reaktion) *Eine biochemische Reaktion r wird im folgenden durch ein 4–Tupel (V, N, F, H) dargestellt, wobei mit V die Menge von Vorbedingungen, mit N die Menge der Nachbedingungen, mit F die Menge der Fördersubstanzen und mit H die Menge der Hemmsubstanzen bezeichnet wird.*

Ausgehend von dieser neuen, konkretisierten Formalisierung einer biochemischen Reaktion muß nun die Vorschrift zur Berechnung der Ähnlichkeit angepaßt werden. Dabei werden Faktoren eingeführt, die eine Gewichtung der an der Reaktion beteiligten Metaboliten ermöglichen. In der Definition 5.4 wurden diese Gewichte nicht angelegt, sondern bereits mit $w_1 = w_2 = 0.5$ angesetzt und gekürzt. In den nachfolgenden Berechnungsvorschriften der Definition 5.6 können die einzelnen Gewichte eingesetzt werden, wobei diese jedoch typischerweise für alle Reaktionsbeteiligten in der gleichen Höhe festgelegt werden. Um den Einfluß der Förder– und Hemmsubstanzen auf das Gesamtergebnis zu verringern, können dann die Gewichte entsprechend kleiner gewählt werden. Somit ist wie im Beispiel 5.3 eine Regelung der Bedeutung von Reaktionsbeteiligten bei einer spezifischen Suchanfrage möglich.

Definition 5.6 (Ähnlichkeit biochemischer Reaktionen) *Die Ähnlichkeit zweier gegebenen biochemischen Reaktionen $r_1 = (V_1, N_1, F_1, H_1)$ und $r_2 = (V_2, N_2, F_2, H_2)$ berechnet sich nach*

$$sim(r_1, r_2) = w_1 \frac{2|V_1 \cap V_2|}{|V_1| + |V_2|} + w_2 \frac{2|N_1 \cap N_2|}{|N_1| + |N_2|} + w_3 \frac{2|F_1 \cap F_2|}{|F_1| + |F_2|} + w_4 \frac{2|H_1 \cap H_2|}{|H_1| + |H_2|},$$

wobei die einzelnen beteiligten Metabolitmengen mit Hilfe der Faktoren w_i gewichtet werden können, wenn gilt

$$\sum_{i=1}^{4} w_i = 1, w_i \geq 0 \quad \textit{und} \quad F_1 \cup F_2 \neq \emptyset, \quad H_1 \cup H_2 \neq \emptyset.$$

Beispiel 5.3 (Vergleich zweier biochemischer Reaktionen) *Der Vergleich zweier biochemischer Reaktionen r_1*

L-Arginin $+ H_2O +$ *Adenosin* $+$ *Citrullin* $\rightarrow$ *L-Ornithin* $+$ *Harnstoff* $+$ *Adenosin* $+$ *Citrullin*

und r_2

L-Arginin + 4-Aminobutanoat → L-Ornithin + 4-Guanidinobutanoat

soll auf Basis der Berechnungsvorschrift 5.6 durchgeführt werden. Für diese beiden Reaktionen ergeben sich aus den Reaktionsgleichungen nun die entsprechenden Mengen der Vor– und Nachbedingungen sowie der Förder– und Hemmsubstanzen.

$$\begin{aligned} V_1 &= \{\textit{L-Arginin}, H_2O\}, \\ N_1 &= \{\textit{L-Ornithin}, \textit{Harnstoff}\}, \\ F_1 &= \{\textit{Adenosin}, \textit{Citrullin}\}, \\ H_1 &= \{\} \end{aligned}$$

und

$$\begin{aligned} V_2 &= \{\textit{L-Arginin}, \textit{4-Aminobutanoat}\}, \\ N_2 &= \{\textit{L-Ornithin}, \textit{4-Guanidinobutanoat}\}, \\ F_2 &= \{\}, \\ H_2 &= \{\} \end{aligned}$$

Für die Berechnung werden die nachfolgenden Gewichte festgelegt.

$$\begin{aligned} w_1 &= 0.45, \\ w_2 &= 0.45, \\ w_3 &= 0.1, \\ w_4 &= 0, da \quad H_1 \cup H_2 = \emptyset \end{aligned}$$

Nach der Definition 5.6 wird folgende Ähnlichkeit ermittelt.

$$sim(r_1, r_2) = 0.45\frac{1}{2} + 0.45\frac{1}{2} + 0 + 0 = 0.45$$

Pathway Alignments

Die unterschiedlichen biochemischen Reaktionen finden im Organismus in der Regel nicht unabhängig voneinander statt. Vielmehr bilden sie ein System aus aufeinander beruhenden Teilen, bei dem Reaktionsprodukte wieder in anderen Reaktionen Ausgangsstoffe darstellen. Deshalb werden in der theoretischen Betrachtung diese einzelnen Reaktionen ebenfalls zu komplexen Netzwerken, den Stoffwechselwegen oder Metabolic Pathways, gruppiert und auf gegenseitige Ähnlichkeiten untersucht. Unter dem Begriff *Pathway Alignment* werden diese Verfahren zusammengefaßt. Dabei werden jedoch nicht nur biochemische Reaktionsketten untersucht, sondern auch Signaltransduktionskaskaden, genregulatorische Systeme oder Abfolgen von Protein–Interaktionen.

Unter dem Namen *PathBlast* wird von [KSK+03] ein System vorgestellt, das auf der Basis von zwei Protein–Interaktions–Netzwerken die gemeinsamen, konservierten Elemente berechnet. Dabei wird ein ähnliches Konzept wie beim Blast–Algorithmus [AGM+90] angewandt, um ein Alignment zu ermitteln, wobei zwischen zwei untersuchten Pfaden auftretende Variationen ebenfalls durch „Gaps“ und „Mismatches“ eingeordnet und bewertet werden.

Der Vergleich von Metabolic Pathways auf der Basis der Teilreaktionen durch Gegenüberstellung der beteiligten Enzyme wird bei [TMH00] vorgeschlagen. Dazu werden die am Pathway beteiligten Enzyme mit einem Alignment angeordnet und dann über eine Ähnlichkeitsfunktion bewertet. Diese Funktion stützt sich auf die Strukturierung der Enzyme durch die EC–Nomenklatur, durch die eine Einteilung in Gruppen auf insgesamt vier Ebenen ermöglicht wird. Über eine Zahlenkombination mit vier Positionen, die durch einen Punkt getrennt werden, sind die Enzyme so den entsprechenden Klassen zugeordnet. Die

erste Ebene bezeichnet den Reaktionstyp des Enzymes, die zweite Ebene seine gruppenspezifische Wirkung, in der dritten Ebene die Substratspezifität und in der vierten Ebene werden die einzelnen Enzyme dann aufgelistet. Das nachfolgende Beispiel 5.4 zeigt diese Strukturierung für ein Enzym. Dieses Vorgehen wurde in [MTM02] auf Stoffwechselwege von E. coli angewandt und lieferte erste Ergebnisse.

Beispiel 5.4 (Eingruppierung nach der EC–Nomenklatur) *Das Enzyme Alkoholdehydrogenase wird nach der EC–Nomenklatur eingeordnet. Dabei nimmt die Klassifikation mit steigender Tiefe der Gruppierung an Spezifität zu.*

```
1.x.x.x  Oxydoreductases
1.1.x.x  Acting on the CH-OH group of donors
1.1.1.x  With NAD+ or NADP+ as acceptor
1.1.1.1  Alcohol dehydrogenase; Aldehyde reductase
```

Ein ähnlicher Ansatz zur Gegenüberstellung von biochemischen Reaktionen verfolgt [Che02] mit dem *PathAligner*, der biochemische Reaktionsketten ebenfalls anhand der beteiligten Enzyme vergleicht und nachfolgend näher vorgestellt werden soll. Die Ähnlichkeitsfunktion bewertet wie im Beispiel 5.5 dann zwei Enzyme anhand ihrer EC–Eingruppierung. Ihr Wert ist 1, wenn die EC–Nummern identisch sind, 0.75, wenn sie bis auf die letzte Stelle identisch sind, usw. und 0, wenn die beiden Enzyme keine gemeinsame Hauptklasse haben.

Beispiel 5.5 (Ähnlichkeit zweier Stoffwechselwege nach [Che02]) *Ausgehend von den zwei Stoffwechselwegen* $E_1 = \{4.3.2.1, 6.3.4.5, 2.1.3.3\}$ *und* $E_2 = \{6.3.4.16, 2.1.3.3\}$ *wird ein Alignment gebildet.*

```
{4.3.2.1, 6.3.4.5,  2.1.3.3}
{   -   , 6.3.4.16, 2.1.3.3}
```

Die Ähnlichkeit der Stoffwechselwege E_1' *und* E_2' *berechnet sich anschließend nach*

$$\frac{1}{3}(0 + 0.75 + 1) = 0.58$$

Die Abbildung 5.1 zeigt das Ergebnis der Beispielanfrage bei der Nutzung des PathAligner über das WWW. Dabei wurden die korrespondierenden Enzyme für die Bewertung der Ähnlichkeit durch unterschiedliche Farben hervorgehoben.

Die Fokussierung auf die beteiligten Enzyme wurde in ähnlicher Weise bereits bei [DSS+99] vorgenommen, aber auch durch weitere Verfahren ergänzt. Dabei wurden drei alternative Möglichkeiten kombiniert: Analyse und Vergleich biochemischer Daten, Analyse der Stoffwechselwege und gegenüberstellende Analyse von Genomsequenzen.

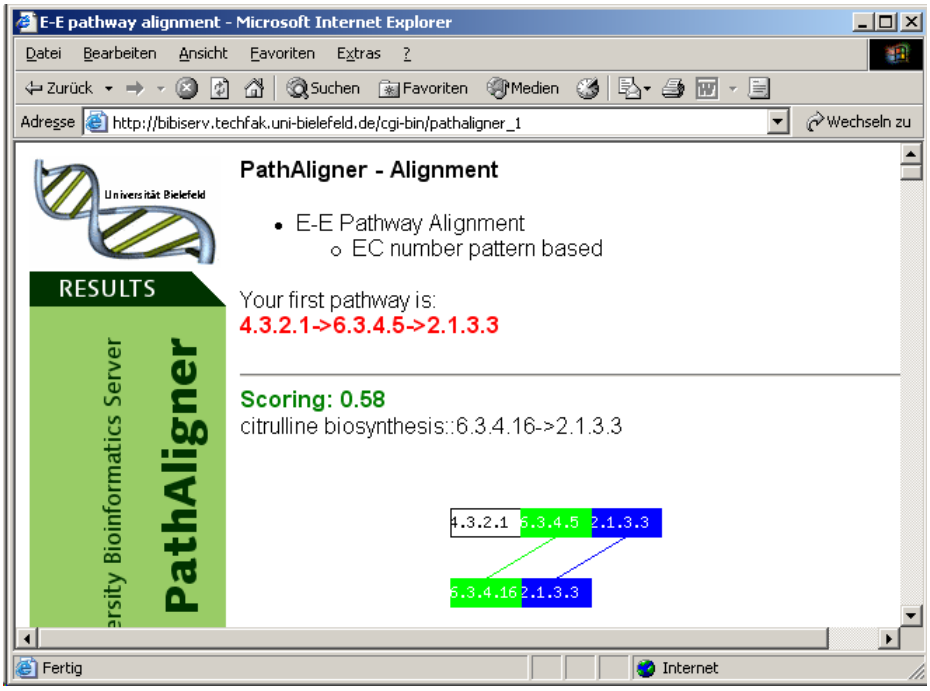

Abbildung 5.1: Darstellung der Beispielanfrage nach Nutzung des PathAligner im WWW

5.1.3 Domäne der genomischen Sequenzen

Sobald auf der Ebene der genomischen Sequenzen nach Ähnlichkeiten gesucht wird, werden im Bereich der Molekularbiologie Verfahren zum Vergleich von Sequenzen oder Zeichenfolgen eingesetzt — die *Sequenzalignments*. Dabei wird die Abfolge der Basen innerhalb der DNS bzw. RNS als Wort eines Alphabetes verstanden und miteinander verglichen. Diese Sequenzalignments bieten die Möglichkeit, verschiedene Sequenzen qualitativ und quantitativ zu beurteilen. Die unterschiedlichen Möglichkeiten, wie Sequenzen durch Mutationen aus anderen hervorgehen können, sind in der Abbildung 2.3 illustriert worden. Eine weitere Möglichkeit, die Ähnlichkeit von Organismen auf der genomischen Ebene zu berechnen, wird im Bereich der Züchtungsforschung angewandt und zieht dazu die *verwandtschaftlichen Verhältnisse* der untersuchten Lebewesen heran.

Sequenzalignments

Im Rahmen der Untersuchung von Primärsequenzen, also Nukleotid- oder Aminosäureketten, soll dem zu untersuchenden Gen gewöhnlich eine bestimmte biochemische Funktion oder ein Phänotyp zugeordnet werden. Dazu müssen nun in der Gensequenz Zeichen für die mögliche Funktion des korrespondierenden Proteins gefunden werden. Dazu wird diese nun anderen Sequenzen mit bereits bekannter Funktion gegenübergestellt. Die Bedeutung dieser grundlegenden Analysen wird in [Rau01] formuliert: „Der Vergleich

von Sequenzen unbekannter Biomoleküle mit Sequenzen solcher Biomoleküle bekannter Funktion (und eventuell auch bekannter Struktur) ist ein zentrales Anliegen der Bioinformatik."

Die Schlußfolgerung ähnlicher Funktion bei verwandten Sequenzen resultiert aus der Feststellung, daß sich bei der Entstehung neuer Arten das zugrunde liegende biochemische System des Organismus nicht vollkommen neu entwickelt. Vielmehr stellt die Natur meist durch kleine Modifikationen des Genoms neuartige Funktionen bereit. An dieser Stelle sind nun die verschiedenen, zur Beschreibung von Sequenzvergleichen notwendigen Bezeichnungen nach [Rau01] voneinander abzugrenzen. Häufig werden diese trotz ihrer unterschiedlichen Bedeutung synonym verwendet. Eine totale Übereinstimmung der untersuchten Sequenzen an jeder Position wird als die *Sequenzidentität* bezeichnet. Wie bereits erwähnt, entwickeln sich die biochemischen Abläufe bei neuen Organismen nicht vollkommen neu. Werden nun deren Proteine verglichen, so kann festgestellt werden, daß Aminosäuren mit ähnlichen Eigenschaften eher durch einander substituiert werden als durch unähnliche Aminosäuren. Diese Neigung zu bestimmten Substitutionen wird im Rahmen der *Sequenzähnlichkeit* bewertet. Als allgemeiner Ausdruck für eine evolutionär bedingte Verwandschaftsbeziehung zwischen Sequenzen wird *Sequenzhomologie* gebraucht. Zwei homologe Sequenzen sind somit aufgrund ihrer gemeinsamen Herkunft ähnlich, während ähnliche Sequenzen zueinander nicht homolog sein müssen.

Ein Verfahren, das den Sequenzvergleich durchführt, stellt nun zwei Sequenzen in voller Länge oder nur in bestimmten Abschnitten gegenüber. Anhand eines festgelegten Bewertungsschemas wird dann die prozentuale Übereinstimmung der gewählten Gegenüberstellung ermittelt. Um eine bessere Bewertung zu erreichen, können die untersuchten Sequenzen dann gegeneinander verschoben werden. Die entsprechenden Bewertungsparameter zielen darauf, die Übereinstimmung zwischen den untersuchten Sequenzen darzustellen, wobei beispielsweise auch Lücken (*Gaps*) in der gewählten Paarung beachtet werden. Um jedoch herauszufinden, ob das letztendlich berechnete Alignment auch von entsprechender statistischen Signifikanz ist oder ob die ermittelte Übereinstimmung auch bei der Gegenüberstellung zufällig ausgewählter Sequenzen erreicht werden konnte, muß anschließend die Qualität des Alignments überprüft werden.

Als allgemeiner Algorithmus zur Suche von Ähnlichkeiten in Aminosäuresequenzen von zwei Proteinen wurde 1970 in [NW70] von NEEDLEMAN und WUNSCH ein Verfahren vorgestellt, das die Bewertung von Homologie zwischen diesen Proteinen erlaubt. Dieser Ansatz wurde vielfach als Grundlage für weitere, variierte und verbesserte Algorithmen genutzt, bei denen nun beispielsweise auch Lücken (*Gaps*) in den zu vergleichenden Sequenzen beachtet wurden oder statistische Angaben über den Austausch von Aminosäuren in Substitutionsmatrizen gehalten werden. Mit BLAST (*Basic Local Alignment Search Tool*) wurde dann in [AGM$^+$90] eine mittlerweile weit verbreitete Software publiziert, die eine unbekannte Sequenz mit schon bekannten, in öffentlich zugänglichen Datenbeständen wie GenBank gespeicherten Sequenzen vergleicht. Eine ältere, aber ebenso gebräuchliche Methode ist die Suche mit FASTA. Weitergehende Informationen zur praktischen Analyse von Sequenzen und zu den benutzten Algorithmen und Programmen

geben [BQ01, GJ02, Rau01].

Verwandschaftsbeziehungen

Die Untersuchung der genetischen Ähnlichkeit verschiedener Organismen anhand des Verwandtschaftsverhältnisse hat sich bereits mit dem Beginn der modernen Züchtungsforschung zu einem Gebiet besonderer Bedeutung entwickelt und wird im Rahmen der genetischen Statistik ausgewertet. Dabei werden die Begriffe Ähnlichkeit und Distanz ebenfalls wechselweise verwendet, so daß sich der Wert der Distanz durch die Berechnung der Differenz von 1 und dem Ähnlichkeitswert ergibt.

In diesem Zusammenhang soll hier die Ermittlung des *Coefficient of parentage* (Elternschaftskoeffizient) zwischen zwei Elementen vorgestellt werden. In [Kem69] wird anhand der jeweiligen Stammbäume dieser Organismen ein Vorgehen beschrieben, das einen numerischen Wert zur Bewertung der Verwandtschaftsverhältnisse nach der Definition 5.7 liefert. Dieses wurde eingeführt, um während eines Zuchtprogrammes von Tieren oder Pflanzen anhand eines qualitativen Maßsystems Inzucht zwischen eng verwandten Individuen zu vermeiden. So sind beispielsweise Sohn und Vater enger verwandt als Großvater und Enkel.

Definition 5.7 (Coefficient of parentage (Elternschaftskoeffizient)) *Der Coefficient of parentage r_{XY} bezeichnet die Wahrscheinlichkeit, daß ein beliebiges Gen des Individuums X durch Abstammung identisch mit einem beliebigen Gen des Individuums Y ist.*

Zur Illustration des Vorgehens zur Ermittlung des Elternschaftkoeffizienten wird im nachfolgenden Beispiel eine mögliche Konstellation von miteinander verwandten Individuen vorgestellt.

Beispiel 5.6 (Verwandtschaftliche Beziehungen zwischen Organismen) *In der Abbildung 5.2 werden die möglichen verwandtschaftlichen Beziehungen zwischen bestimmten Individuen einer Art dargestellt. Dabei sind A und B nicht verwandt und besitzen somit einen Koeffizienten von $r_{AB} = 0$. Die direkten Nachkommen von A und B sind C und D, während E aus der Paarung $A \times D$ und F aus $B \times E$ hervorgeht.*

Die daraus resultierenden Elternschaftskoeffizienten werden in der Tabelle 5.3 dargestellt. Dabei berechnet sich beispielsweise der Koeffizient r_{AC}, also die Verwandschaft zwischen den Individuen A und C aus

$$r_{AC} = \frac{1}{2}(r_{AA} + r_{AB}) = \frac{1}{4}$$

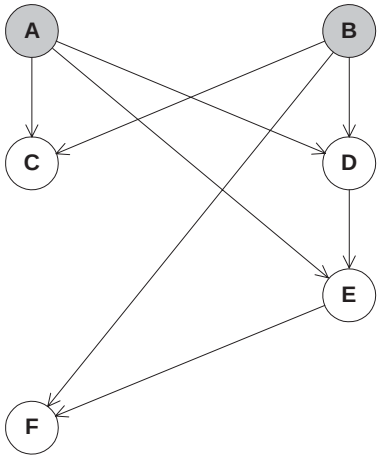

Abbildung 5.2: Darstellung verwandtschaftlicher Beziehungen zwischen Organismen (nach [Kem69])

	A	B	C	D	E	F
A	$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{16}$
B	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{5}{16}$
C	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
D	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{5}{16}$
E	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{3}{8}$
F	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{5}{16}$	$\frac{3}{8}$	$\frac{9}{16}$

Tabelle 5.3: Berechnung des Elternschaftskoeffizienten für ein Beispiel (nach [Kem69])

5.1.4 Zusammenfassende Gegenüberstellung

Die in den vorangegangenen Abschnitten vorgestellten Verfahren zur Untersuchung von Ähnlichkeiten in molekularbiologischen und medizinischen Datenbeständen sollen in diesem Abschnitt anhand verschiedener Merkmal verglichen und eingeordnet werden. Dazu wurden die folgenden fünf Merkmale herangezogen.

Ähnlichkeitsmaß

Bezugnehmend auf die Definition 5.2, die Anforderungen an ein Ähnlichkeitsmaß festlegt, wird untersucht, ob diese Anforderungen durch das betrachtete Verfahren erfüllt werden.

Algorithmus

Dieses Merkmal beschreibt, ob für das untersuchte Verfahren ein Algorithmus beschrieben wurde. Diese Beschreibung kann beispielsweise in Pseudocode vorhanden sein und erleichtert eine eventuelle Umsetzung des Verfahrens.

Anwendung
Der tatsächliche Status der Umsetzung dieses Verfahrens wird durch dieses Merkmal gekennzeichnet. Der Realisierungsstand kann beispielsweise durch eine im WWW verfügbare prototypische Implementierung nachgewiesen werden.

Bibliothek
Zur Verwendung der untersuchten Verfahren in der Praxis werden Bibliotheken für verschiedene Programmiersprachen bereitgestellt, so daß eine Anbindung und Nutzung in einem Prototypen ohne großen Aufwand ermöglicht wird. Die vorgestellten Algorithmen sind dann bereits beispielsweise in anwendbaren Methoden oder Funktionen implementiert.

Domäne
Durch die Integration verschiedener molekularbiologischer Datenquellen sind unterschiedliche Aspekte des betrachteten biologischen Systemes in der Integrationsdatenbank vertreten. Zur Strukturierung des Datenbestandes werden Domänen entsprechend der Defintion 5.1 zugeordnet.

Die Tabelle 5.4 zeigt die Verfahren und ihre Merkmale. Die Bezeichnung einiger Merkmale mußte dabei abgekürzt werden. Die Ausprägung der Merkmale außer der Domäne wird durch *ja*, *nein* und *unbekannt* dargestellt.

Verfahren	Merkmal				
	Domäne	Ähnlich.	Algo.	Anwend.	Biblio.
Fallbasiert	klin. Phänotyp	ja	ja	ja	nein
Reaktionsbeteiligte	biochem. Reaktion	ja	ja	nein	nein
PathAligner	Stoffwechselwege	ja	ja	ja	ja
PathBlast	Stoffwechselwege	nein	ja	ja	nein
Sequenzalignment	gen. Sequenzen	nein	ja	ja	ja
Verwandtschaft	gen. Sequenzen	ja	ja	nein	nein

Tabelle 5.4: Gegenüberstellung von Verfahren zur Ähnlichkeitsbewertung anhand verschiedener Merkmale

5.2 Genotyp–Phänotyp–Korrelation auf Szenarioebene

Derzeit werden nach [KST02] zur Untersuchung von Genotyp–Phänotyp–Korrelation in der Biologie drei verschiedene Methoden angewandt: „quantitative genetics“, „molecular genetics“und „animal models“. Beim quantitativen Ansatz werden dazu genetische und Umweltfaktoren untersucht, die unterschiedliche individuelle Merkmale bestimmter Merkmale verursachen. Diese Merkmale können physischer Art (Körpergröße, Gewicht),

persönliche Merkmale (Agressivität, Selbstlosigkeit) aber auch kognitive Merkmale (Intelligenz, Gedächtnis) sein. Existiert für ein bestimmtes Merkmal eine interessante Variation innerhalb der untersuchten Population, so wird nach einer entsprechenden genetischen Veränderung innerhalb der Studie gesucht. Dabei darf jedoch der starke Einfluß von Umweltfaktoren nicht außer Acht gelassen werden. Der molekulargenetische Ansatz hingegen verfolgt die Identifikation spezifischer Gene aufgrund von veränderten Proteinfunktionen, die zurückverfolgt werden. Die selektive Zucht von natürlichen Merkmalen und die Erschaffung von transgenen Pflanzen und Tieren im dritten Ansatz erlaubt die Untersuchung der Auswirkungen genetischer Veränderungen in unterschiedlichen Entwicklungsstadien.

Der Ausgangspunkt für die Suche nach Beziehungen zwischen Genotypen und Phänotypen in der vorliegenden Arbeit sind die Daten, die durch den Integrationsprozeß in einer Integrationsdatenbank zusammengeführt wurden. Diese gesammelten Daten wurden durch ihre Zuordnung zu spezifischen Informationsdomänen durch den Anwender entsprechend der Definition 5.1 für die weitere Nutzung strukturiert. Durch unterschiedliche Fremdschlüsselbeziehungen sind die Relationen der Datenbank innerhalb der Domänen und zwischen den einzelnen Domänen miteinander verbunden. Die integrierten medizinischen und molekularbiologischen Daten bilden die Grundlage für den nachfolgend vorgestellten Ansatz.

Zur Entwicklung eines allgemeinen Ansatzes zur Identifikation von Genotyp–Phänotyp–Korrelationen in einem derart strukturierten Datenbestand ist es erforderlich, die Suche unabhängig von der Semantik der integrierten Datenbestände zu realisieren. Dieses Vorgehen ist notwendig, da in den unterschiedlichen Disziplinen der Medizin, Biologie oder Molekulargenetik verschiedene Auffassungen darüber vorhanden sind, wodurch beispielsweise der Phänotyp eines Organismus beschrieben wird. Aus den Blickwinkeln der Anwender könnte der Phänotyp im Zusammenhang mit einer Erkrankung die Manifestation bestimmter klinischer Symptome, ein defektes Enzym oder auch die blockierte Abfolge biochemischer Reaktionen sein. Deshalb wird in der Definition 5.8 eine Formalisierung des Begriffes Genotyp–Phänotyp–Korrelation vorgenommen.

Definition 5.8 (Genotyp–Phänotyp–Korrelation) *Als Genotyp–Phänotyp–Korrelation K wird im folgenden ein 4–Tupel K = (G, P, ϕ, ψ) bezeichnet mit*
der Menge genotypischer Merkmale g_i (Genotyp) $G = \{g_1, g_2, .., g_n\}$, $n \in N$,
der Menge phänotypischer Merkmale p_i (Phänotyp) $P = \{p_1, p_2, .., p_m\}$, $m \in N$,
der Abbildung $\phi : G \rightarrow P$, die genotypische Merkmale in phänotypische Merkmale überführt,
der Abbildung $\psi : P \rightarrow G$, die phänotypische Merkmale in genotypische Merkmale überführt.

Die dargestellte einfache Formalisierung einer Genotyp–Phänotyp–Korrelation läßt sich nun durch eine Übertragung auf einen Graphen weiter ausbauen. Dabei werden die integrierten Datensätze als Knoten und die Beziehungen zwischen den einzelnen Datensätzen

als Kanten repräsentiert. Die Tabelle 5.5 verdeutlicht dazu die Gegenüberstellung von Objekten innerhalb der Problemstellung und den äquivalenten Elementen der Graphentheorie.

Problemstellung	Graphentheorie
Datenbestand	Graph
Datensätze	Knoten
Beziehungen	Kanten
Informationsdomänen	Subgraphen
Ähnlichkeiten/Distanzen	Gewichte

Tabelle 5.5: Gegenüberstellung von Objekten innerhalb der Problemstellung und den äquivalenten Elementen der Graphentheorie

Zur Verdeutlichung des Vorgehens in den folgenden Abschnitten wird eine allgemeine Fragestellung formuliert, die im Verlauf entsprechend den vorgenommenen Abstraktions– und Analyseschritten angepaßt wird.

Allgemeine Fragestellung

Welche genotypischen und phänotypischen Merkmale können in Life–Science–Datenbeständen miteinander in Beziehung gebracht werden?

Ausgehend von dieser Fragestellung soll eine weitergehende Formalisierung der Zusammenhänge und Beziehungen innerhalb des integrierten Datenbestandes durchgeführt werden. Dazu erfolgt die Übertragung der Problemsituation in die Graphentheorie. Eine Untersuchung von Life–Science–Datenbeständen durch Nutzung von Graphen wird zum Beispiel bei [LNRV03, LMNR04] durch die Trennung dieser Daten in eine logische und eine physische Ebene realisiert. Auf der physischen Ebene befinden sich die tatsächlichen Datenquellen und die Links, die sich zwischen ihnen befinden. Die logische Ebene hingegen enthält Objektklassen, beispielsweise von Konzepten oder Ontologien. Diese Objektklassen können durch eine oder mehrere physische Datenquellen implementiert werden. Zum Beispiel ist eine Umsetzung der logischen Klasse „Citation“ durch die Datenquelle PubMed möglich. Die in diesem Ansatz genutzten logischen Klassen werden in der vorliegenden Arbeit durch die Informationsdomänen repräsentiert.

Die Betrachtung von Genotyp–Phänotyp–Korrelationen auf der Basis von Graphen ermöglicht die Anwendung von bekannten Algorithmen auf der Datenstruktur des Graphen, die bereits in ihrer Komplexität untersucht wurden. Die verschiedenen genotypischen und phänotypischen Merkmale in den Mengen G und P der Definition 5.8 finden sich als Knoten des Graphen wieder, so daß eine Korrelation zwischen diesen nun durch die Suche geeigneter Kantenzüge identifizierbar wird.

Der dadurch entstehende Graph $X = (V, E)$ mit den endlichen Mengen V an Knoten und E an Kanten ist schlicht. Das heißt, er ist endlich, ungerichtet und besitzt keine Schlingen und Mehrfachkanten. Dieser Graph wird durch die Zuordnung der Knoten, die in

der Integrationsdatenbank als Datensätze bestehen, zu $i \in N$ Informationsdomänen in Subgraphen $X_i = (V_i, E_i)$ geteilt, wobei $\bigcup_{i=1}^{n} V_i = V$ gilt. Unter Berücksichtigung der vorliegenden Fragestellung können Kanten zwischen Knoten eines Subgraphen vorerst vernachläßigt werden, so daß gilt $E_i = \emptyset$. Das Beispiel 5.7 verdeutlicht diese Zusammenhänge und stellt einen Graphen in einer Abbildung zeichnerisch dar.

Beispiel 5.7 (Graph mit Subgraphen) *Für einen Graphen* $X = (V, E)$ *mit*
$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$ *und*
$E = \{[1, 5], [1, 6], [2, 6], [3, 6], [3, 7], [4, 7], [6, 8], [6, 9], [7, 10], [7, 11], [7, 12], [8, 13], [9, 13], [9, 15], [11, 15], [12, 16]\}$
werden mit $n = 4$ *Informationsdomänen Subgraphen* $X_i = (V_i, E_i)$ *festgelegt. Dabei ist*
$V_1 = \{1, 2, 3, 4\}$,
$V_2 = \{5, 6, 7\}$,
$V_3 = \{8, 9, 10, 11, 12\}$ *und*
$V_4 = \{13, 14, 15, 16\}$.
Eine mögliche zeichnerische Darstellung dieses Graphen zeigt die Abbildung 5.3. Dabei wurde entsprechend der Unterteilung in Subgraphen eine hierarchische Struktur gewählt, die jeweils alle Knoten eines Subgraphen horizontal anordnet.

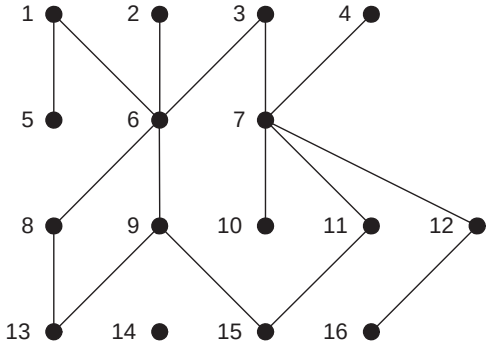

Abbildung 5.3: Darstellung der Graphenstruktur für einen Beispieldatenbestand

Mit der vorangehenden Formalisierung der vorliegenden Struktur des Datenbestandes ist die Grundlage für eine Untersuchung der integrierten Daten gegeben. Die im Rahmen dieser Arbeit verfolgte Fragestellung der Suche nach Beziehungen zwischen genotypischen und phänotypischen Merkmalen wird nun zuerst mit der Sicht auf die Datenebene beschrieben. Die einzelnen Merkmale des Genotypes und des Phänotypes sind als Datensätze im Datenbestand integriert worden. Wenn zwischen diesen Merkmalen bereits Zusammenhänge, die von unterschiedlicher Natur sein können, beobachtet wurden, so sind die Merkmale im Datenbestand typischerweise durch ihre Schlüssel miteinander in Verbindung gesetzt. Somit läßt sich die nachfolgende Fragestellung formulieren.

Fragestellung auf Ebene der integrierten Daten

Welche Datensätze einer ausgewählten Informationsdomäne lassen sich bei der

Verfolgung von Beziehungen zwischen den Elementen des integrierten Datenbestandes ausgehend von einem bestimmten Ausgangspunkt erreichen?

Zur Untersuchung des aus dem integrierten Datenbestand konstruierten Graphen sind Algorithmen notwendig, die abhängig von bestimmten Parametern den Graphen durchsuchen und einen oder mehrere Knoten als Ergebnis liefern. Dazu wird die untersuchte Fragestellung von der Datenebene auf die Graphenebene angepaßt. Als Parameter werden der Graph $X = (V, E)$, der Subgraph $X_j \subseteq X$, in dem die Zielknoten erwartet werden, und der Startknoten $u \in E$ übergeben.

Fragestellung auf Ebene des Graphen

Welche Knoten eines ausgewählten Subgraphen X_j lassen sich bei der Verfolgung von Kantenfolgen eines Graphen $X = (V, E)$ ausgehend von einem bestimmten Knoten u erreichen?

In einem ersten Vorschlag wird zur algorithmischen Lösung das fundamentale Prinzip der Breitensuche angewendet, auf dem viele bekannte Graphalgorithmen basieren. Grundlage der Breitensuche sind dabei zwei Listen zur Speicherung von zu untersuchenden und bereits besuchten Knoten. In der Abbildung 5.4 wird dieser Algorithmus skizziert.

```
Algorithmus Breitensuche(Graph X, Subgraph J, Knoten u) {
  Initialisiere Knotenliste K mit dem Knoten u;
  Initialisiere Knotenliste B als leer;
  Initialisiere Ergebnisliste E als leer;
  Solange die Liste K nicht leer ist {
    Entnimm der Liste K das erste Element k;
    Fuege den Knoten k der Liste B hinzu;
    Ist der Knoten k im Subgraph J {
      Fuege den Knoten k zur Liste E hinzu;
    } sonst {
      Fuer alle benachbarten Knoten i von k {
        Ist der Knoten i nicht in der Liste B {
          Fuege den Knoten i der Liste K hinzu;
        }
      }
    }
  }
}
```

Abbildung 5.4: Algorithmus für die Breitensuche im Graphen X bei gegebenem Subgraphen J und dem Startknoten u

Ausgehend von einem gegebenen Startknoten u werden sukzessiv die Sphären für u berechnet. Die dabei gefundenen Knoten werden auf ihre Zugehörigkeit zum Subgraphen $J = X_j$ überprüft und bei positivem Resultat der Ergebnisliste hinzugefügt.

In einem Beispieldatenbestand, wie in der Abbildung 5.5 vereinfachend dargestellt, mit den Informationsdomänen *Sequence*, *Protein*, *Disease* und *Patient* können nun Anfragen formuliert und ausgeführt werden. Die Festlegung auf den *Patient*–Datensatz mit der Identifikation 715 als Startknoten und den Subgraphen *Sequence* als Zieldomäne würde die Datensätze K03020, U49897, S61296 als resultierende Knoten der Suche liefern. Über einem integrierten Datenbestand sind so Beziehungen zwischen Datensätzen über die Grenzen der ursprünglichen, einzelnen Datenquellen hinweg möglich.

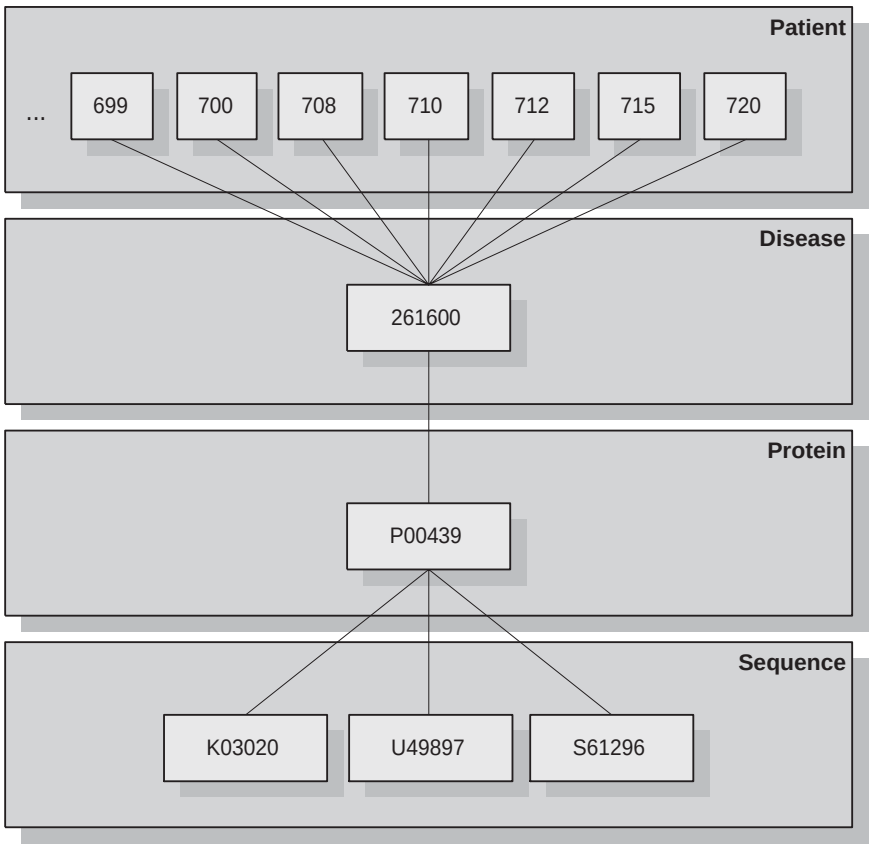

Abbildung 5.5: Mögliche Genotyp–Phänotyp–Korrelationen als Graph mit der Kennzeichnung der untersuchten Domänen mit ihren englischen Bezeichnungen und den Identifikatoren der beteiligten Datensätze

Im Abschnitt 5.1 wurden Verfahren untersucht, die Ähnlichkeitsbewertungen innerhalb bestimmter medizinischer und molekularbiologischer Domänen erlauben. Nachdem nun bereits durch die Nutzung von Graphen die Suche nach Beziehungen zwischen genotypischen und phänotypischen Merkmalen vorgestellt wurde, soll als Ausblick die Frage gestellt werden, wie in diesem Szenario eine Ausweitung auf ähnliche Datensätze

ermöglicht werden kann. Durch die Nutzung von Graphen bietet sich dazu die Erweiterung der Graphenstruktur um gewichtete Kanten an, so daß Ähnlichkeitswerte als Gewichte innerhalb eines Algorithmus berücksichtigt werden können und eine Aussage über die zu erwartende Relevanz der Ergebnisses für die aktuelle Problemsituation getroffen werden kann.

5.3 Zusammenfassung

Aufbauend auf der Analyse der integrierten Daten und Datenquellen wurden im Rahmen eines laufenden DHGP–Projektes in Zusammenarbeit mit den Kooperationspartnern verschiedene Datenquellen für eine Integration ausgewählt. Dabei decken diese einzelnen Datenquellen unterschiedliche Aspekte biologischer Systeme ab und stellen durch explizite Verweise oder Referenzen Beziehungen zwischen den enthaltenen Informationsdomänen her. In diesem Kapitel wurden neben Arbeiten aus der Literatur eigene Ansätze zur Untersuchung von Ähnlichkeiten und Beziehungen innerhalb und zwischen den integrierten Datenbeständen vorgestellt. Sie dienen als weitergehende Überlegungen für eine Analyse und Aufbereitung eines integrierten Life–Science–Datenbestandes.

Dazu wurden im ersten Teil dieses Kapitels für die Domänen der klinischen Phänotypen, der biochemischen Reaktionen und Reaktionsketten und der genomischen Sequenzen unterschiedliche Ansätze präsentiert, die eine Gegenüberstellung von ähnlichen Objekten ermöglichen. So wurden beispielsweise für klinische Phänotypen die verfügbaren Merkmale in der Ramedis–Datenbank untersucht und klassifiziert, so daß bestimmte Merkmalskombinationen für eine Analyse ausgewählt und ähnliche Fälle gesucht werden können. Ein Prototyp dieser Ähnlichkeitssuche innerhalb der Domäne der klinischen Phänotypen wurde bereits prototypisch in Ramedis implementiert.

Die eigentliche Unterstützung der Suche nach Korrelationen von Genotyp und Phänotyp innerhalb eines integrierten Datenbestandes wird im zweiten Abschnitt diskutiert. Dazu werden die typischen Vorgehensweisen in Molekulargenetik und Medizin vorgestellt. Durch die Zusammenführung der unterschiedlichen Datenquellen mit Hilfe der Datenintegration wird nun jedoch die Untersuchung von Beziehungen auf der Ebene der elektronischen Datenbestände möglich. Auf der Basis von Graphen wird ein Algorithmus zur Breitensuche angepaßt, so daß von einem Startknoten aus, der beispielsweise ein phänotypisches Merkmal repräsentert, erreichbare Knoten innerhalb einer vorher festgelegten Zieldomäne ermittelt werden, die bestimmte genotypische Merkmale widerspiegeln.

6 Vorstellung des Prototypen des Gesamtsystemes

Im Rahmen dieser Arbeit wurden bisher unterschiedliche methodische Grundlagen vorgestellt und analytische Schritte unternommen, um zwei Ansätze zur Untersuchung von Genotyp–Phänotyp–Korrelationen auf der Basis von gesammelten klinischen Daten sowie innerhalb eines integrierten molekularbiologischen und medizinischen Datenbestandes zu entwickeln. Aufbauend auf den Ergebnissen der Analyse verschiedener Integrationsansätze, der Untersuchung von Ähnlichkeitsmaßen auf verschiedenen Life–Science–Domänen und der Entwicklung eines Algorithmus zur Suche nach potentiellen Korrelationen von Genotyp und Phänotyp soll auf den folgenden Seiten ein Architekturvorschlag für eine integrierte Analyseumgebung aufgezeigt und erläutert werden. Durch die Auswahl und Integration einiger molekularbiologischer und medizinischer Datenquellen wird außerdem an einem Beispielszenario die Nutzung des Prototypen und das Vorgehen innerhalb der Umgebung verdeutlicht.

Die Einordnung in den Projektrahmen wird an dieser Stelle insbesondere durch die Zusammenarbeit mit den Projektpartnern und die Nutzung der entwickelten Prototypen durch die Teilprojekte charakterisiert. In diesem Zusammenhang wurde die Datenbank für Mutationen und korrespondierende Phänotypen (Ramedis) besonders intensiv an der Klinik für Kinder– und Jugendmedizin der Kreiskliniken Reutlingen eingesetzt und mit Fallberichten angereichert. Der vorangehende Softwareentwurfsprozeß konnte außerdem durch regelmäßige Konsultationen zielgerichtet auf medizinische Fachexperten als die späteren Anwender ausgerichtet werden. Die auf diesem Weg zusammengetragenen medizinischen und molekulargenetischen Patientendaten wurden anonymisiert als Fallberichte gespeichert und gehen außerdem teilweise als phänotypische Informationen in die Datenintegration ein.

Weitere prototypische Implementierungen wurden im Bereich der Datenintegration erstellt und in Kooperationen mit verschiedenen Partnern eingesetzt. So konnten einerseits Informationsressourcen durch Integration unterschiedlicher Datenquellen erschlossen werden, die von den Teilprojekten auf den Gebieten transkriptionale Regulation, Signalwege und Stoffwechselwege bereitgestellt wurden. Zusammen mit den klinischen Daten aus den gesammelten Fallberichten ermöglicht der Zugriff auf den integrierten Datenbestand die Suche nach potentiellen Kandidaten für Proteine, die im untersuchten Netzwerk der Beispielerkrankung Diabetes mellitus MODY interessant für weitergehende Untersuchungen sein können. Aufbauend auf diesen Daten wird so die Vorhersage regulatorischer Regionen und die Analyse der beteiligten Signalwege unterstützt.

6.1 Architektur und Komponenten

In diesem Abschnitt wird nun die Architektur eines integrierten Systemes zur Untersuchung von Genotyp–Phänotyp–Korrelationen auf der Basis von molekularbiologischen Daten skizziert. Dabei wird zuerst ein Überblick für das Gesamtsystem gegeben. Anschließend werden die einzelnen Komponenten detaillierter vorgestellt. Zur Verdeutlichung des Aufbaues und des Vorgehens werden entsprechende Abbildung eingesetzt.

Ausgangspunkt für den Entwurf des Gesamtsystemes sind Komponenten des im Abschnitt 3.2.3 unter dem Namen FRIDAQ [Sch02] vorgestellte Frameworkes zur Integration molekularbiologischer Datenbestände. Die Komponenten des Frameworkes zur Datenintegration und lokalen Speicherung werden als Module in der hier vorgestellten Architektur genutzt.

Einer der Vorteile dieses Frameworks ist die permanente Möglichkeit des Nutzers auf die aktuellen Originaldaten der integrierten Datenquellen zuzugreifen. Damit wird für die Anwendung eine höchstmögliche Aktualität der Daten garantiert. In der Praxis erweist sich diese Forderung jedoch als Quelle einer Vielzahl von Problemen. Bei komplexen Anfragen über dem integrierten Schema treten in der Regel Antwortzeiten auf, die eine effiziente Nutzung deutlich erschweren. Aus diesem Grund wurde zwar im Rahmen des FRIDAQ–Frameworkes die Anwendung eines Caches diskutiert, jedoch wird bei großen Datenbeständen und unterschiedlichen Anfragen dieser das Problem kaum ausreichend beheben können. Deshalb wurde in der nachfolgend vorgestellten Architektur eine Komponente zur Replikationssteuerung aufgenommen, die eine redundante Speicherung der Inhalte integrierter Datenquellen ermöglicht und steuert. Aus diesem Vorgehen, diese Daten als lokale Kopien bereitzuhalten, resultiert nun eine erhöhte Verfügbarkeit der Daten und eine Steigerung der Effizienz von Zugriffen.

6.1.1 Architektur im Überblick

In der Abbildung 6.1 ist die Architektur des Prototypen dargestellt. Sie besteht im wesentlichen aus drei Komponenten und verschiedenen *Nutzern*, die über Schnittstellen auf Systemfunktionen zugreifen. Das eigentliche System bildet dabei die Komponente *Integrierte Analyseumgebung*, die sich aus den Teilkomponenten *Genotyp–Phänotyp–Analyse*, *Domänendatenverwaltung*, *Lokale Speicherung*, *Replikationssteuerung* und *Datenintegration* zusammensetzt. Außerhalb dieser Komponenten befinden sich neben den *Nutzern* auch noch die Komponenten *Existierende Analysewerkzeuge* und *Existierende Datenquellen*.

Das wichtigste Element der vorgestellten Architektur bildet die *Integrierte Analyseumgebung*, die Zugriff auf die *Existierenden Datenquellen* nimmt und entsprechend der Systemkonfiguration molekularbiologische Daten integriert. Diese Daten werden innerhalb der *lokalen Speicherungskomponente* gehalten, die somit die Funktion einer Integrationsdatenbank übernimmt. Die zeitliche Kontrolle und Regelung der Integration wird durch

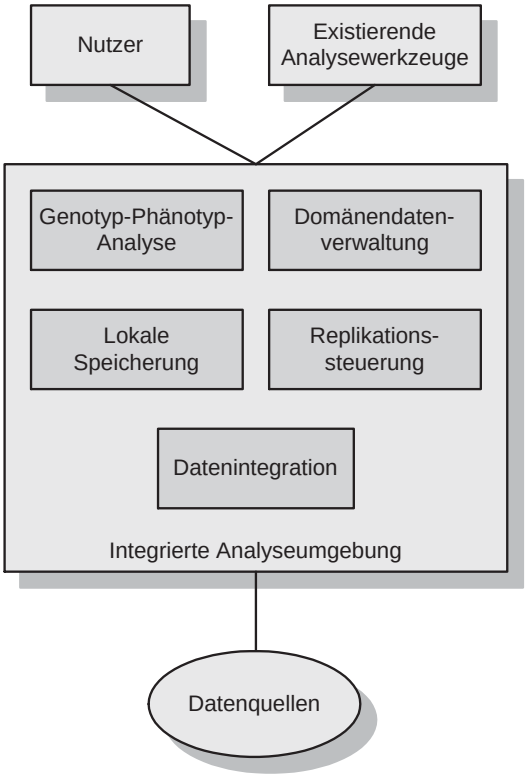

Abbildung 6.1: Architektur des Prototypen mit externen Analysewerkzeugen und den integrierten, externen Datenquellen

eine *Replikationssteuerung* übernommen. Über verschiedene Schnittstellen wird ein Zugriff auf die Daten innerhalb der *Lokalen Speicherung* ermöglicht, so daß sowohl *Existierende Analysewerkzeuge* als auch *Nutzer* Zugang zum integrierten Datenbestand erhalten.

Zur Untersuchung von Genotyp–Phänotyp–Korrelationen steht die Teilkomponente *Genotyp–Phänotyp–Analyse* bereit, die ebenfalls auf den Datenbestand der *Lokalen Speicherung* zugreift und über eine Schnittstelle für *Nutzer* Dienste anbietet. Die Teilkomponenten *Genotyp–Phänotyp–Analyse*, *Domänendatenverwaltung* und *Replikationssteuerung* innerhalb der *Integrierten Analyseumgebung* werden in den folgenden Teilabschnitten vertiefend erläutert und im anschließenden Beispielszenario in Nutzung und Zusammenspiel vorgestellt. Demgegenüber wird für nähere Informationen zur *Datenintegration* und *Lokalen Speicherung* auf Abschnitt 3.2.3 und die entsprechenden Referenzen verwiesen.

6.1.2 Replikationssteuerung

Einleitend wurden bereits die Notwendigkeit und die damit verbundenen Konsequenzen der Etablierung einer *Replikationssteuerung* innerhalb der *Integrierten Analyseumgebung*

skizziert. An dieser Stelle soll nun vertiefend auf diese Teilkomponenten eingegangen werden. Dazu werden in einem ersten Schritt die grundlegenden Anforderungen an eine Anwendung formuliert, die die mehrfache Speicherung von Datenobjekten steuern soll. Anschließend werden kurz die Vor– und Nachteile einer solchen Applikation vorgestellt.

Im Bereich der Betriebssystementwicklung sind die Probleme von Perfomancedifferenzen zwischen verschiedenen Zugriffsarten seit langem bekannt. So sind Festplattenzugriffe viel langsamer als Hauptspeicherzugriffe. Wegen dieser Zeitunterschiede wird die Arbeit mit dem Dateisystem durch Caching–Strategien [Tan02] optimiert. Dabei werden Bereiche des Festplattenspeichers im Hauptspeicher gehalten, um Zugriffsanfragen schneller beantworten zu können. Um diesen Cache zu verwalten, werden verschiedene Algorithmen verwendet, die anhand unterschiedlicher Präferenzen die Ersetzung von zwischengespeicherten Speicherbereichen und das Laden neuer Bereiche in den Cache optimieren. Diese Strategien eignen sich jedoch zur Anwendung im vorliegenden Kontext nicht, da eine bloße Zwischenspeicherung von bereits ermittelten Anfrageergebnissen oder eine vorausschauende Speicherung von noch nicht gestellten Anfragen nicht ausreicht oder kaum möglich ist, da unterschiedliche Nutzer Anfragen über einem heterogenen — aus verschiedenen Datenquellen stammenden — Datenbestand formulieren.

Der Begriff der Replikation wird typischerweise im Zusammenhang mit dem Einsatz verteilter oder mobiler Datenbanken verwendet. Dabei motiviert sich die mehrfache Speicherung eines Datenobjektes durch eine bessere Erreichbarkeit der Daten, eine Performancesteigerung durch höhere Zugriffsgeschwindigkeiten und eine erhöhte Ausfallsicherheit [BD96]. Jedoch entsteht dabei ein Konflikt zwischen der angestrebten Erhöhung von Verfügbarkeit und Performance einerseits und der Korrektheit des Gesamtsystemes andererseits. Denn durch die Speicherung von Repliken wird die in Datenbanksystemen typische Forderung nach Redundanzfreiheit verletzt. Außerdem steigt natürlich der Aufwand zur Aktualisierung des Gesamtsystemes mit den Änderungstransaktionen auf dem Datenbestand. Ebenso sinkt die Aktualität der Daten mit einer höheren Anzahl von Kopien, was jedoch in dieser Betrachtung nicht von Bedeutung ist, da jeweils nur eine Kopie angelegt wird.

Zusammenfassend läßt sich also feststellen, daß ein Kompromiß zwischen Vor– und Nachteilen der Replikation gefunden werden muß. Die verschiedenen Anforderungen sind im Überblick in der Abbildung 6.2 illustriert. Die Nutzung von Replikationsmechanismen innerhalb der hier vorgestellten *Integrierten Analyseumgebung* beschränkt sich jedoch auf einen lesenden Zugriff, so daß Änderungsoperationen nicht notwendig sind und der dabei anfallende Aufwand bei der Abwägung vernachlässigt werden kann.

Um mit erhöhter Verfügbarkeit und Performance von der Existenz einer lokalen Kopie des integrierten Datenbestandes zu profitieren, müssen geeignete Replikationsverfahren zur Sicherstellung der Konsistenz von originären und replizierten Daten angewendet werden. Die Nutzung dieser Vefahren vereinfacht sich im Gegensatz zu ihrer Anwendung im Rahmen verteilter und mobiler Datenbanken, da in der hier vorgeschlagenen Architektur nur eine einzelne Kopie erstellt werden soll, auf der auch keine Änderungstransaktionen durchgeführt werden sollen. Diese lokale Kopie — die Integrationsdatenbank — ist

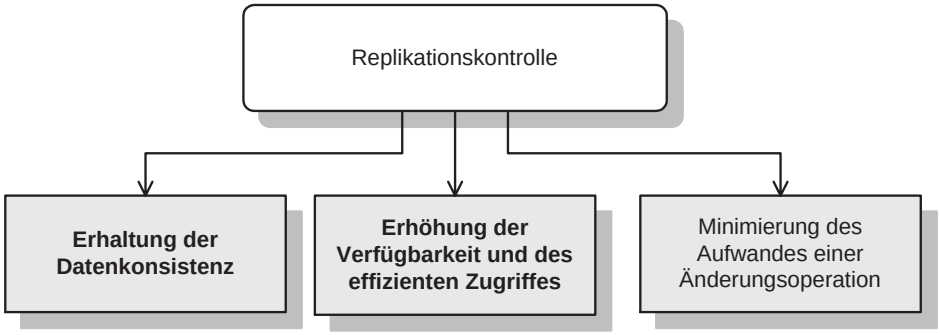

Abbildung 6.2: Darstellung der Zielkonflikte der Replikationskontrolle (nach [BD96])

jedoch in einem geeigneten Maße aktuell zu halten, ohne daß jede Anfrage auf den Originaldatenbestand weitergeleitet wird. Nachteil der Anwendung eines Replikationsverfahrens ist natürlich der dazu notwendige erhöhte Speicherplatzbedarf und der steigende Aufwand zur Aktualisierung des Datenbestandes.

6.1.3 Domänendatenverwaltung

Die Teilkomponente *Domänendatenverwaltung* innerhalb der *Integrierten Analyseumgebung* ermöglicht es dem Nutzer, individuelle Anforderungen an den Inhalt und die Präsentation des integrierten Datenbestandes zu definieren. Außerdem bildet sie die Grundlage für eine Untersuchung von Genotyp–Phänotyp–Korrelationen in der Teilkomponente *Genotyp–Phänotyp–Analyse* und stellt für diese Aufgabe zusätzlich benötigte semantische Daten über den integrierten Datenbestand bereit. Die Nutzung des Begriffes Domäne innerhalb dieser Komponente entspricht der Definition 5.1.

Im Rahmen des Integrationsprozesses wird aus einer Menge von existierenden molekularbiologischen Datenquellen eine Integrationsdatenbank angelegt, die unterschiedliche Aspekte der Betrachtung eines biologischen Systemes abbildet. Diese verschiedenen Sichtweisen können innerhalb einer vereinheitlichten, hierarchielosen Darstellungsweise nicht präsentiert werden. Deshalb ist es für den Nutzer von großer Bedeutung, über den integrierten Daten verschiedene nutzerspezifische Domänen in Form von Sichten anzulegen. Bestandteil einer solchen Sicht ist die Bezeichnung, die Definition verschiedener Attribute für Suchoperationen und die Festlegung der Repräsentation über die grafische Nutzerschnittstelle. Die Beziehungen zwischen Sichten können außerdem durch abstrakte Konzepte, wie Generalisierung oder Spezialisierung beschrieben werden.

Das typische Vorgehen bei der Definition der nutzerspezifischen Domänen auf dem inte-

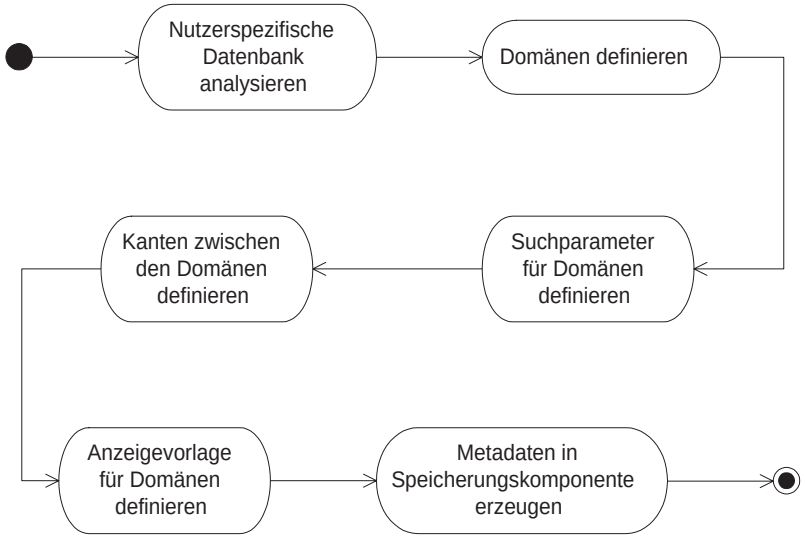

Abbildung 6.3: Vorgehen zur Erzeugung der Informationen über nutzerspezifische Domänen innerhalb der Domänendatenverwaltung als UML–Aktivitätsdiagramm

grierten Datenbestand wird in der Abbildung 6.3 illustriert. Nachdem durch die Nutzung der *Datenintegrations*–Komponente eine Integrationsdatenbank aus den angeschlossenen Datenquellen erzeugt wurde, kann der Anwender die für ihn relevante Sicht auf den Datenbestand festlegen. Das nachfolgend beschriebene Vorgehen ist durch eine geeignete grafische Nutzerschnittstelle zu unterstützen. Dieser Vorgang erzeugt Metadaten mit semantischen Informationen über den integrierten Datenbestand, die ebenfalls in der Datenbank gehalten werden.

Dazu ist von ihm natürlich ein ausreichendes Wissen über die integrierten Daten zu erwarten. Durch die Definition einer Domänenbeschreibung werden semantische Informationen durch den Nutzer bereitgestellt. Dazu gehören auch mögliche Suchparameter über dem integrierten Datenbestand, die dem Nutzer eine Erschließung über die grafische Nutzeroberfläche erleichtern. Anschließend werden die Beziehungen zwischen den Domänen durch Kanten definiert. Diese werden explizit durch SQL–Ausdrücke beschrieben und ermöglichen die Suche von potentiellen Beziehungen zwischen Genotypen und Phänotypen.

Zur Aufbereitung der gespeicherten Daten innerhalb der grafische Nutzerschnittstelle werden nun für jede Domäne spezifische Anzeigevorlagen erzeugt, die das Fachwissen des Nutzers über die bearbeitete Domäne wiedergeben. In einem abschließenden Schritt werden nun noch die erforderlichen Metadaten mit den Domäneninformationen innerhalb der Datenbank erzeugt.

6.1.4 Genotyp–Phänotyp–Analyse

Ziel der Teilkomponente zur Genotyp–Phänotyp–Analyse ist die Unterstützung des Nutzers bei der Suche und Verwaltung von Zusammenhängen auf Basis einer Genotyp–Phänotyp–Korrelation innerhalb des integrierten Datenbestandes. Dabei werden dem Anwender über eine grafische Nutzerschnittstelle geeignete Werkzeuge bereitgestellt, die eine Navigation innerhalb und zwischen den Informationsdomänen erlauben.

Entsprechend den im Abschnitt 5.2 vorgestellten Verfahren werden in dieser Teilkomponente die folgenden Funktionen unterstützt.

- Bereitstellung einer grafischen Nutzerschnittstelle
- Aufbereitung der Metadaten der Teilkomponente *Domänendatenverwaltung* zur Formulierung von Suchanfragen
- Anbindung der Integrationsdatenbank (Teilkomponente *Lokale Speicherung*) über Datenbankzugriffsfunktionen
- Übersichtsartige Darstellung von Ergebnissen der Suchanfragen
- Temporäre Speicherung von Zwischenergebnissen und Daten über Beziehungen zwischen Domänen
- Detaillierte Darstellung von Ergebnissen entsprechend den definierten Anzeigevorlagen der Teilkomponente *Domänendatenverwaltung*
- Übersichtsartige Darstellung von Genotyp–Phänotyp–Korrelation über dem integrierten Datenbestand

Die Navigation innerhalb der Integrationsdatenbank und die Suche nach spezifischen Datensätzen wird durch die Darstellung vorhandener Fremdschlüsselbeziehungen und bekannter Interaktionen zwischen den Einzelrelationen des Datenbestandes unterstützt. Durch sie können relevante Daten über Domänengrenzen hinweg zusammengeführt werden. Einige Screenshots der webbasierten, grafischen Nutzerschnittstelle sind in der Abbildung 6.6 dargestellt. Eine weitergehende, automatisierte Unterstützung der Suche nach Korrelationen zwischen Genotypen und Phänotypen ist durch Einbindung des im Abschnitt 5.2 vorgestellten Graphalgorithmus denkbar.

6.2 Vorgehen und Anwendung am Beispiel

In den vorangehenden Abschnitten wurde die Architektur eines Systemes zur Unterstützung der Suche nach Genotyp–Phänotyp–Korrelationen im Überblick vorgestellt. Weiterhin wurden einzelne Teilkomponenten erläutert. Dieser Abschnitt soll nun anhand

eines Beispielszenarios die Beantwortung spezifischer Fragestellungen illustrieren. Dabei wird jedoch nicht vertiefend auf die Anbindung neuer Datenquellen oder detailiert auf den Integrationsprozeß eingegangen.

Das hier vorgestellte Beispielszenario orientiert sich an der klinischen Sicht auf Stoffwechselerkrankungen, die durch angeborene Gendefekte verursacht werden. Zur Erläuterung der entsprechenden Zusammenhänge sei auf das 2. Kapitel verwiesen. Der Ausgangspunkt für dieses Fallbeispiel sei ein Mediziner, der basierend auf der Erkrankung eines Patienten weitergehende Informationen zu ähnlichen Erkrankungsfällen und den damit assoziierten Mutationen erhalten möchte.

Schritt 1: *Analyse der Anforderung an den zu integrierenden Datenbestand*

In einem ersten Schritt sind die möglichen Fragestellungen an das System abzuklären. In der Softwaretechnik würde dieses Vorgehen der Analysephase entsprechen. Dabei muß in enger Kooperation mit den späteren Nutzern der Umfang und der Inhalt der Fragestellungen diskutiert werden, für die das System konzipiert und der Datenbestand integriert werden soll. Für das Beispielszenario wären die folgende Anfragen denkbar.

- Welche anderen Fallberichte sind der aktuellen Problemstellung im Bezug auf Laborwerte oder Symptome ähnlich?
- Welche Diagnosen oder Therapiemöglichkeiten wurden innerhalb ähnlicher Fallberichte angewendet?
- Welches Enzym ist defekt, so daß eine bestimmte biochemische Reaktion innerhalb eines Stoffwechselweges nicht durchgeführt werden kann?
- Welche Erkrankungen werden durch den Ausfall eines bestimmten Proteins bedingt?
- Welche Nukleotid– oder Aminosäuresequenz ist einem bestimmten Protein zugeordnet?

Schritt 2: *Auswahl der zu integrierenden Datenbestände aus verschiedenen molekularbiologischen Datenquellen*

Auf Grundlage des vorangehenden Analyseschrittes werden nun unter den verschiedenen molekularbiologischen Datenquellen die zu integrierenden Datenbestände ausgewählt. Eine Vorstellung und Untersuchung einer Auswahl der gebräuchlichsten molekularbiologischen Datenquellen ist im Abschnitt 3.1 zu finden. Für das Beispielszenario werden die Datenquellen und ihr Integrationsbeitrag in der Abbildung 6.4 als Quader illustriert. Dabei sind die zu integrierenden Bestandteile der Quellen in einer dunkleren Farbe dargestellt.

Zur übersichtsweisen Darstellung der verfügbaren Daten auf dem Weg vom Phänotyp zum Genotyp wird das klinische Erscheinungsbild einer Erkrankung als Ausgangspunkt genutzt. Dazu wurden Datenbestände aus der Mutationsdatenbank *Ramedis* zu klinischen Phänotypen, wie beispielsweise Laborwerte und Symptome, und den entsprechenden Mutationen integriert.

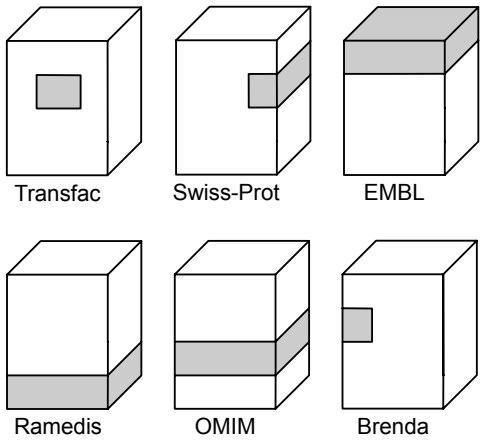

Abbildung 6.4: Verschiedene molekularbiologische Datenquellen und ihre zu integrierenden Datenbestände

Als allgemeine Datenquelle für Informationen über Erkrankungen, ihre Diagnostik und Therapie dient *OMIM*. Die an der Entstehung von Stoffwechselerkrankungen beteiligten Enzyme und die beeinflußten biochemischen Reaktionen werden aus der Datenbank *Brenda* gewonnen. Informationen über die an der Genregulation beteiligten Transkriptionsfaktoren wurden aus *Transfac* integriert. Die mit *Transfac* assoziierten Datenquellen *Transpath* und *PathoDB* liefern Wissen über Signalwege und die pathologisch relevanten mutierten Formen der Transkriptionsfaktoren und ihrer Bindungsstellen. Ergänzend hinzugefügt sind allgemeine Proteininformationen über *Swiss–Prot*. Die Originalsequenz (nicht mutiert) liefert die genomische Sequenzdatenbank *EMBL*.

Schritt 3: *Anlage einer Integrationsdatenbank mit den ausgewählten Inhalten und Definition von Domänen über den integrierten Daten*

Nachdem die verfügbaren Datenquellen analysiert wurden und der Umfang der zu integrierenden Daten festgelegt wurde, ist es nun notwendig, den Integrationsprozeß durchzuführen. Die für die Integration notwendigen Methoden und Teilschritte sind im Abschnitt 3.2 näher beschrieben. In dem vorliegend beschriebenen Vorgehen wird der BioDataServer nach [FHL+02] als Integrationsdienst genutzt, der die erforderlichen Daten aus den Quellen anfordert und entsprechend einem globalen Schema in der Integrationsdatenbank abgelegt.

In der Abbildung 6.5 a sind die ausgewählten Teile der einzelnen Datenquellen abgebildet, die in den integrierten Datenbestand einfließen sollen. Diese Ausschnitte der Datenquellen aus der Abbildung 6.4 werden, wie im Abbildungsteil b dargestellt, zusammengeführt und dann den entsprechenden Informationsdomänen zugeordnet. Diese Domänen Patient, Krankheit, Protein und Sequenz sind im Abbildungsteil c bezeichnet und werden außerdem innerhalb der Integrationsdatenbank durch Fremdschlüsselbeziehungen verbunden, so daß Verbindungen zwischen zu-

geordneten Informationen möglich sind.

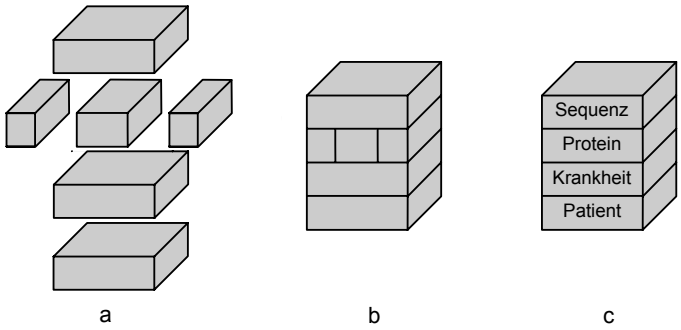

Abbildung 6.5: Nutzerspezifische Integrationsdatenbank mit den ausgewählten Inhalten und den definierten Domänen über den integrierten Daten

Schritt 4: *Anwendung der grafischen Nutzerschnittstelle für Anfragen*

Im Rahmen des abschließenden Schrittes ist es nun möglich, über eine grafische Nutzerschnittstelle, die in einem aktuellen HTML–Browser benutzbar ist, sowohl allgemeine Anfragen auf dem integrierten Datenbestand zu formulieren, als auch spezielle Pfade innerhalb der integrierten Daten als Genotyp–Phänotyp–Korrelationen zu untersuchen. Außerdem können die integrierten Daten übersichtsartig aufbereitet werden oder als Tupel in der Originalrelation betrachtet werden.

In der Abbildung 6.6 sind verschiedene Screenshots dieser webbasierten Nutzerschnittstelle dargestellt. Auf der linken oberen Seite dieser Abbildung (a) befindet sich die Suchmaske zur Formulierung von allgemeinen Anfragen über dem integrierten Datenbestand. Diese wird zur Ausführungszeit aus den im dritten Schritt angelegten Domänendaten generiert. Das rechte obere Fenster dieser Abbildung (b) zeigt eine mögliche Verbindung vom Genotyp (Sequence) zum Phänotyp (Patient) dieses Beispielszenarios. Zu den einzelnen Datensätzen lassen sich die zugeordneten Informationen in einer Übersicht anzeigen.

Der Zugriff auf die Einzelrelationen wird im linken unteren Screenshot (c) abgebildet. Die hier dargestellte Relation besitzt zwei Attribute, deren Werte in Tabellenform dargestellt werden, wobei Bedingungen über den einzelnen Attributwerten formuliert werden können, um die Ergebnismenge einzuschränken. Im rechten unteren Fenster (d) sind unterschiedliche Fremdschlüsselbeziehungen innerhalb des integrierten Datenbestandes aufgeführt, so daß die Navigation zwischen Relationen über gemeinsamen Identifikatoren möglich ist.

Die hier vorgestellten vier Schritte bilden natürlich nur eine grobe Übersicht des erforderlichen Vorgehens. Von besonderer Bedeutung ist dabei die detaillierte Ermittlung der Anforderungen der zukünftigen Nutzer an einen integrierten Datenbestand. Vielfach überdeckt bei der Vorstellung und Nutzung eines solchen integrierten Systemes die subjektive Wahrnehmung der Fachexperten den objektiven Nutzen der Anwendung, weil Daten

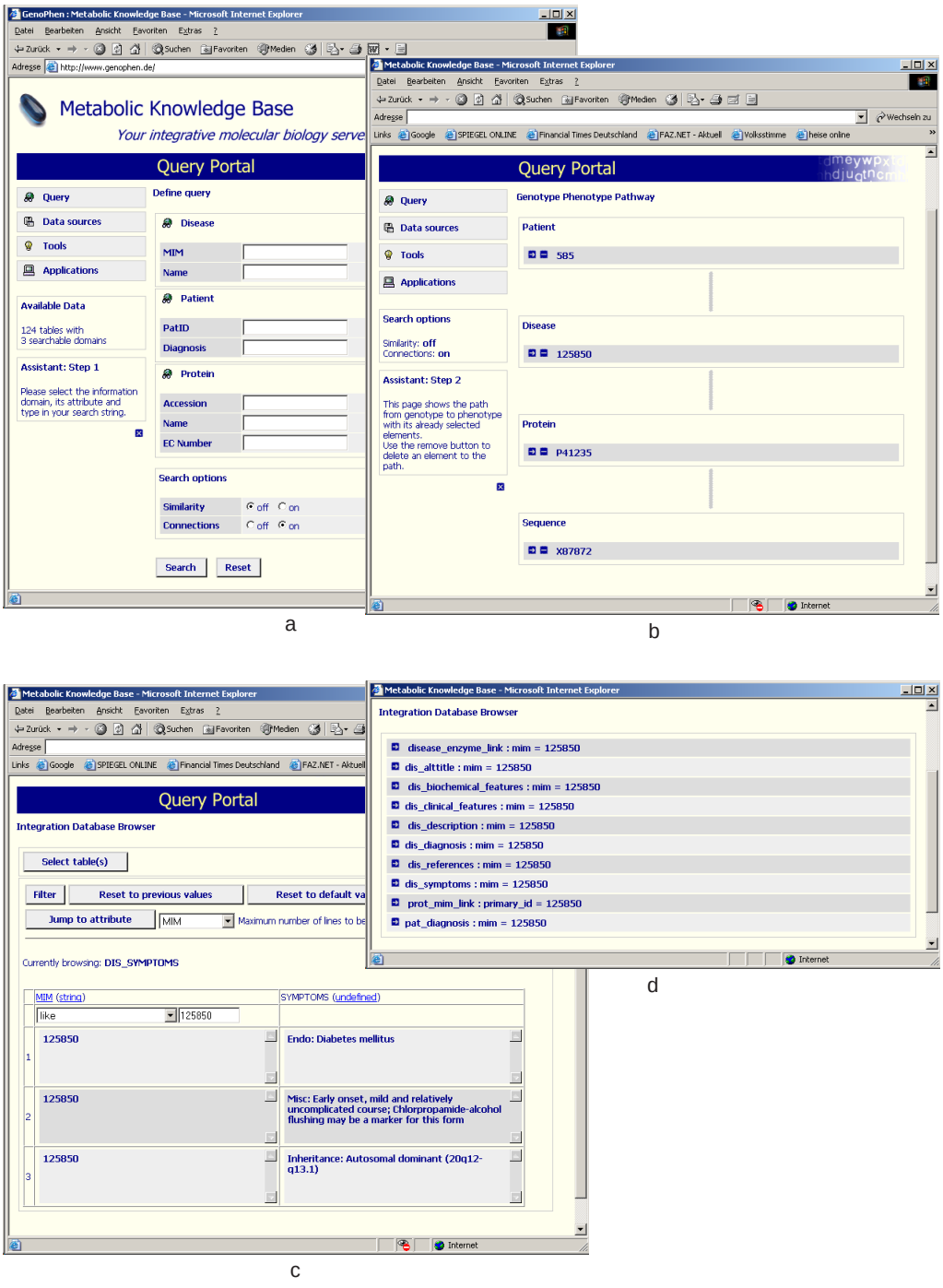

Abbildung 6.6: Screenshots der webbasierten, grafischen Nutzerschnittstelle mit (a) Anfragemaske mit vordefinierten Informationsdomänen, (b) Beispiel eines vom Nutzer ausgewählten Pfades vom Genotyp (Sequence) zum Phänotyp (Patient), (c) Darstellung von Datensätzen der Originalrelation in Tabellenform, (d) Auswahlmöglichkeit zwischen verschiedenen Fremdschlüsselbeziehungen innerhalb des integrierten Datenbestandes ausgehend von einem bestimmten Attribut

aus unterschiedlichen Quellen integriert wurden, ohne beispielsweise für den Biologen oder Mediziner einen Mehrwert zu schaffen. Solche prototypischen Integrationen sind jedoch für den Informatiker notwendig, um die Funktionsfähigkeit des Gesamtsystemes zu präsentieren. Daher ist insbesondere die Bedeutung der Auswahl der zu integrierenden Daten aus öffentlichen und forschungsgruppeninternen Quellen zu beachten. Erst eine weitergehende Verwendung der Integrationsumgebung in enger interdisziplinärer Kooperation mit Fachexperten erbringt anwenderorientierte und effizient einsetzbare Softwarewerkzeuge.

Nachfolgend wird an einem konkreten Beispiel die Reichhaltigkeit der in dem zuvor integrierten Datenbestand verfügbaren Informationen illustriert. Dazu sind im linken Teil (a) der Abbildung 6.7 die über dem integrierten Datenbestand angelegten Informationsdomänen *Patient*, *Erkrankung*, *Protein*, *Transkriptionsfaktor* und *Sequenz* dargestellt. Die einzelnen Domänen enthalten dabei unterschiedliche Informationen über die betrachteten Objekte, so können beispielsweise für einen bestimmten Patienten, der durch Datensätze innerhalb der Domäne Patient dargestellt wird, neben allgemeinen Daten (Geschlecht, Herkunft) auch Aussagen über Symptome, Laborparameter, Molekulargenetik, Therapie/Entwicklung und Diät/Medikamente gefunden werden.

Gegenstand des dargestellten Beispiels ist die heute weit verbreitete Erkrankung Diabetes mellitus. Sie besitzt verschiedene Untergruppen, von denen sich MODY (*Maturity-onset Diabetes of the Young*) besonders zur Demonstration der erzielten Integrationsergebnisse eignet, da die korrespondierenden Mutationen sechs verschiedenen Genen zugeordnet werden können und jeweils unterschiedliche Krankheitsbilder verursachen [FSG04]. Dabei kodieren fünf dieser Gene Transkriptionsfaktoren und ein Gen die Glukokinase, so daß eine molekulargenetische Bestätigung der Verdachtsdiagnose Diabetes MODY möglich ist.

Für den spezifischen MODY–Typ 1 wurden im rechten Teil (b) der Abbildung 6.7 ausgehend von der MIM–Nummer 125850 die Daten aufgeführt, die diesem Eintrag von anderen Domänen aus zugeordnet werden können. Dabei sind jedoch nur Ausschnitte der verfügbaren Informationen aufgeführt. So konnte eine Reihe von Patienten gefunden werden, die mit dieser Diagnose gespeichert sind. Für den Fall 585 sind allgemeine Daten, wie das Geschlecht, das Alter, in dem die Diagnose erstellt wurde, und die entsprechende Literaturreferenz in der Abbildung aufgeführt. Die für MODY 1 verantwortliche Mutation wurde im Gen HNF4A lokalisiert und der Domäne Protein zugeordnet. Das Genprodukt HNF-4alpha entfaltet seine Wirkung als Transkriptionsfaktor. Die entsprechenden genomischen Sequenzen für die beteiligten Gene sind nur als Links zur Datenquelle EMBL enthalten, da der Umfang von EMBL für eine prototypische Integration zu mächtig war und die Aminosäuresequenz bereits aus Swiss–Prot unter der Domäne Protein integriert wurde.

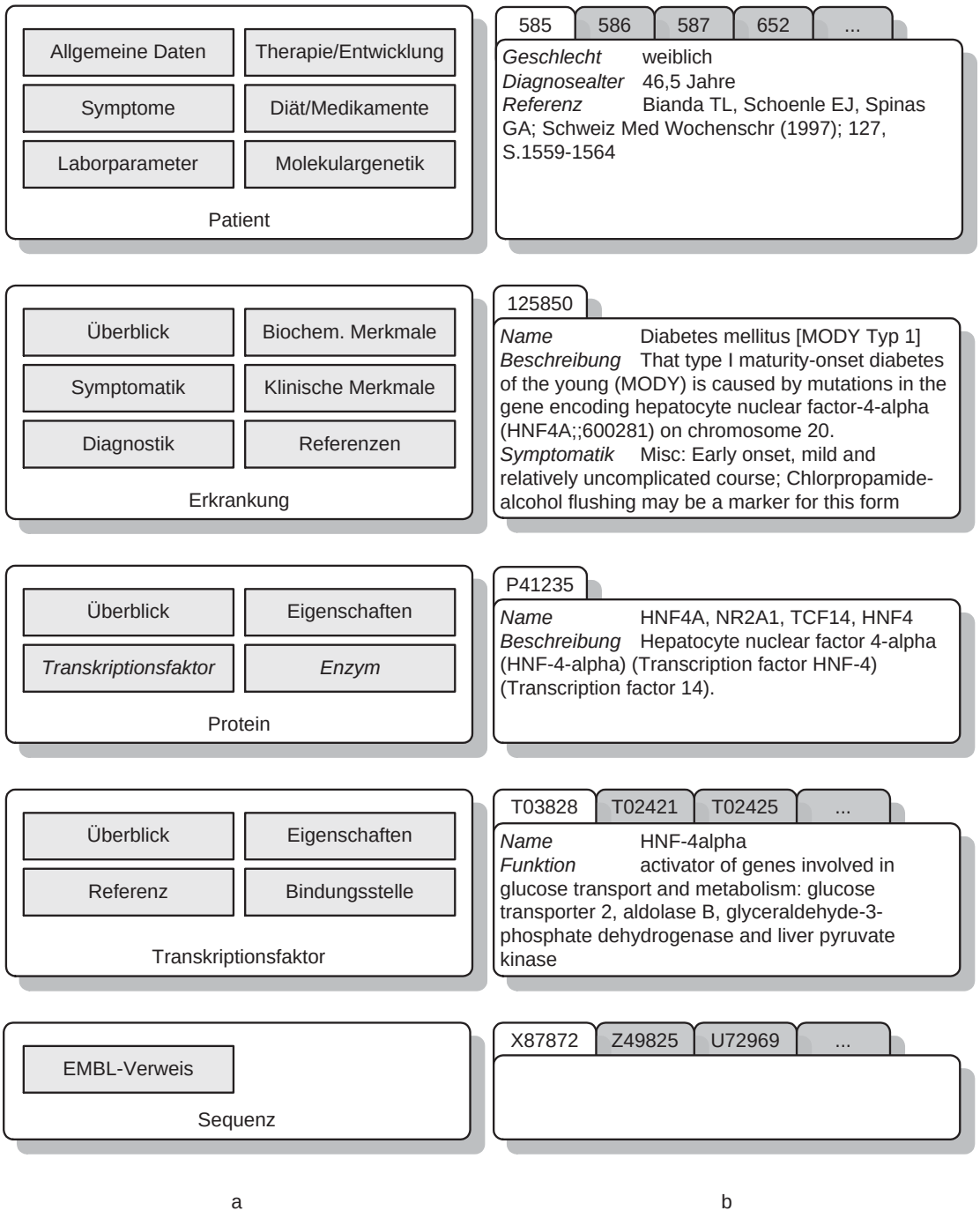

Abbildung 6.7: Darstellung des Datenumfanges der Integrationsdatenbank am Beispiel Diabetes mellitus MODY 1 mit (a) einem Überblick der über dem Datenbestand angelegten Informationsdomänen und den enthaltenen Daten sowie (b) der Darstellung einer Auswahl von einzelnen Datensätzen zur Beispielerkrankung, die aus verschiedenen Quellen integriert wurden

6.3 Zusammenfassung

In diesem Kapitel wurde die Architektur und der Prototyp eines Systemes vorgestellt, das auf der Grundlage von integrierten molekularbiologischen und medizinischen Daten die Untersuchung von Genotyp–Phänotyp–Korrelationen unterstützt. Diese Architektur des Gesamtsystemes wurde zu Beginn ausführlich dargestellt und die Zusammenarbeit der einzelnen Komponenten *Integrierte Analyseumgebung*, *Externe Analysewerkzeuge* und *Existierende Datenquellen* untereinander und mit den verschiedenen *Nutzern* erläutert.

Anschließend wurden zwei Teilkomponenten der *Integrierte Analyseumgebung*, die *Genotyp–Phänotyp–Analyse* und die *Replikationssteuerung* detailliert in Aufbau und Funktion beschrieben. Dabei wird insbesondere die Motivation zur Nutzung von Replikationsverfahren dargestellt und Vor– und Nachteile dieses Ansatzes diskutiert.

Den Abschluß des Kapitels bildet ein Beispielszenario, das anhand einer typischen Anfrage das Vorgehen zur Entwicklung eines integrierten Datenbestanden aufzeigt. Im Rahmen dieses Beispieles wurden außerdem die verschiedenen Möglichkeiten der *Genotyp–Phänotyp–Analyse* innerhalb der implementierten Anwendung präsentiert und durch ausgewählte Screenshots illustriert. Am Beispiel der Erkrankung Diabetes mellitus MODY 1 ist außerdem der Umfang der integrierten Daten aufgezeigt worden.

7 Zusammenfassung und Ausblick

In der biomedizinischen Forschung und Entwicklung kommt im allgemeinen eine Vielzahl von Datenbanken und Informationssystemen zum Einsatz, die über unterschiedliche Aspekte biologischer Systeme Auskunft geben, beginnend mit den genomischen Sequenzen über genregulatorische und metabolische Netzwerke bis hin zu klinischen Phänotypen. Für eine umfassende und effiziente Nutzung dieser wertvollen Datenbestände müssen die verteilten Daten integriert und dem Nutzer für Analysen bereitgestellt werden. Neben der offensichtlich erforderlichen Datenintegration muß der Anwender aber auch bei der Navigation innerhalb des zusammengeführten Datenbestandes durch geeignete Werkzeuge unterstützt werden, die eine zielgerichtete und effiziente Weiterverarbeitung der Daten ermöglichen.

Die Unterstützung des Nutzers bei der Navigation orientiert sich an *Genotyp–Phänotyp–Korrelationen*, die Zusammenhänge zwischen der molekulargenetischen Ebene, dem einzelnen Gen oder dem gesamten Genom und der klinischen Ebene, die sich als Menge von direkt oder indirekt beobachtbaren Merkmalen des Organismus manifestiert, beschreiben. Zwischen der DNS–Sequenz des Genotyps und dem letztendlichen Erscheinungsbild des Individuums, dem Phänotyp, liegen jedoch eine Reihe von unterschiedlichen Zwischenschritten, u.a. Proteinsynthese, Genregulation, beteiligte Stoffwechselwege und entsprechende Umwelteinflüsse, die ein Wirknetz bilden, das eine enorme Komplexität entwickelt.

Das sich permanent beschleunigende Wachstum der verfügbaren Menge an digitalisierten Life–Science–Daten hat zu etwa 600 gebräuchlichen, öffentlich nutzbaren Datenbanken und Informationssystemen [Gal04] geführt, wobei eine durchschnittliche Forschungsgruppe auf bis zu 40 Datenquellen parallel zugreift [LNRV03]. Eine einheitliche Verwaltung, aller von einer Anwendung aus unterschiedlichen verteilten Quellen benötigten Daten, wird jedoch typischerweise durch die Heterogenität dieser Quellen auf mehreren Ebenen erschwert.

Neben den klassischen Forderungen bei der Integration nach Transparenz, Vollständigkeit, semantischer Korrektheit und Redundanzfreiheit, gewinnen durch die besonderen Aspekte von Life–Science–Daten weitere Anforderungen an Bedeutung [DOB95, LR03]. So ist beispielsweise durch die ständig wachsende Datenmenge, die komplexen Schemata und ihrer permanenten Anpassung in den meisten medizinischen und molekularbiologischen Datenquellen neben der Aktualität des integrierten Datenbestandes auch ein

effizienter Zugriff sicherzustellen. Diese besonderen Anforderungen bei der Integration molekularbiologischer und medizinischer Datenquellen werden durch eine Reihe von Ansätzen widergespiegelt, die sich speziell auf die Datenintegration in dieser Anwendungsdomäne beziehen. Neben föderierten Datenbanksystemen ist die Nutzung von Data Warehouses häufig vorgeschlagen worden. Bisher konnte sich jedoch keine der verschiedenen akademischen und kommerziellen Architekturen, wie SRS [EA93], DiscoveryLink [HSK+01], BioKleisli [DOTW97] oder TAMBIS [GSN+01], in der Breite durchsetzen.

Die verfügbaren Life–Science–Datenquellen und bestehende Ansätze auf dem Gebiet der Datenintegration sowie offene Fragen aus der Sicht der Entwicklung einer integrierten Analyseumgebung zur Genotyp–Phänotyp–Korrelation bilden Ausgangspunkt und Motivation der vorliegenden Arbeit. Ausgehend von der Analyse der verbreiteten medizinischen und molekularbiologischen Datenquellen wurde anhand einer zusammenfassenden Gegenüberstellung eine Menge dieser Systeme zur Integration ausgewählt. Durch die Kooperation mit den Projektpartnern eines im Rahmen des Deutschen Humangenomprojektes geförderten Konsortiums wurde dabei der inhaltliche Rahmen für die Integration festgelegt und auch Zugriff auf Datenbestände von BRENDA [SCE+04] der Universtität zu Köln, sowie TRANSFAC [MFG+03] und TRANSPATH [KVC+03] der BioBase GmbH erlangt. Weitere Quellen für die Datenintegration bildeten OMIM [HSA+02], SwissProt [BBA+03] und EMBL [KAA+04]. Unter Berücksichtigung der spezifischen Eigenschaften und ihrem Vergleich an zehn Merkmalen wurde aus den bestehenden Integrationsansätzen der Integrationsdienst BioDataServer [FHL+02] aus der FRIDAQ–Architektur [Sch02] herangezogen, um als Komponente der Gesamtarchitektur zu dienen. Sie unterstützt dabei mediatorbasiert SQL–Anfragen auf den einzelnen zu integrierenden Datenquellen.

Im Rahmen der Analyse der vorhandenen Datenquellen auf ihren Informationsumfang und ihre Eignung zur Integration wurde festgestellt, daß zur Abbildung von klinischen Daten im Form von Mutationen und ihren assoziierten Phänotypen, beispielsweise durch molekulargenetische Untersuchungsergebnisse, Symptome und Laborparameter von Patienten, keine adäquate Datenquelle vorhanden ist. Diese Feststellung wurde getroffen, da sich bei der Untersuchung dieser bereits bestehenden Systeme, zum Beispiel PAHdb [SWS+00] und BIODEF [Bla96], heraustellte, daß sie mit einer Reihe von Forderungen, wie erkrankungsunabhängige Einsetzbarkeit, Erweiterbarkeit, Vergleichbarkeit und Referenzierbarkeit, nicht in Einklang stehen und somit für eine Nutzung im Rahmen dieser Arbeit nicht geeignet sind.

Auf der Basis dieser Untersuchungsergebnisse wurde ein universell einsetzbarer Architekturvorschlag für ein webbasiertes Informationssystem entwickelt, das Daten im Stil ausführlicher Fallberichte aggregiert und analysiert. In Zusammenarbeit mit der Universität Tübingen und der Kinderklinik Reutlingen als Referenznutzer wurde die vorgeschlagene Architektur unter dem Namen Ramedis [MST+01] prototypisch implementiert. Die ausgezeichnete Nutzbarkeit dieser Anwendung konnte durch eine hohe Anzahl von etwa 700 bereits gesammelten Fällen nachgewiesen werden. Eine Auswertung des

gespeicherten Datenbestandes wird durch fallbasierte Suchanfragen unterstützt, die sich auf Prinzipien des Case–based Reasoning stützen und ähnliche Fälle zur einer gegebenen Problemsituation auffinden können. Dieses Vorgehen ermöglicht beispielsweise eine Unterstützung des Nutzers bei der Differentialdiagnostik, die auf die Abgrenzung und Identifikation einer bestimmten Erkrankung innerhalb einer Menge von symptomatisch ähnlicher Krankheiten ausgerichtet ist. Außerdem wird durch die gleichzeitige Speicherung von molekulargenetischen Untersuchungsergebnissen und klinischen Parametern bereits eine Genotyp–Phänotyp–Korrelation im Kleinen realisiert.

Durch die Analyse der bestehenden Integrationsansätze und der Festlegung auf den BioDataServer als Teil des Anwendungssystemes wurde die Zusammenführung der ebenfalls bereits untersuchten molekularbiologischen und medizinischen Datenbestände möglich. Das parallel entwickelte Informationssystem Ramedis wurde auch als Quelle für die Datenintegration genutzt, so daß im Rahmen des laufenden BMBF–Projektes ein umfangreicher Datenbestand, beginnend mit Informationen zu DNA–Sequenzen, Proteinen, Enzymen, Transkriptionsfaktoren, biochemischen Reaktion und klinischen Phänotypen, zur Verfügung gestellt werden konnte.

Häufig besteht bei der Formulierung von Fragestellungen über den unterschiedlichen Domänen des integrierten Datenbestandes die Notwendigkeit, nicht nur eindeutige Suchergebnisse zuzulassen, sondern auch ähnliche Resultate für eine Anfrage zu ermitteln. Diese Ähnlichkeitsanfragen sind in verschiedenen Bereichen bereits weit verbreitet, so beispielsweise in Multimedia– oder Geodatenbanksystemen. Die Untersuchung und Gegenüberstellung von Ansätzen aus der Literatur zur Bewertung von Ähnlichkeiten in den molekularbiologischen Domänen der biochemischen Reaktionsketten und der genomischen Sequenzen wurde durch die Entwicklung eigener Vorschläge zur Bewertung ähnlicher biochemischer Reaktionen und klinischer Phänotypen erweitert. Diese Verfahren bilden eine Ergänzung, mit der eine bessere Erschließung des integrierten Datenbestandes ermöglicht werden kann. Durch die Kombination von einzelnen Ähnlichkeitsberechnungen innerhalb der molekularbiologischen und medizinischen Domänen des integrierten Datenbestandes zur Betrachtung des Gesamtszenarios wurde eine Realisierung der Suche von Genotyp–Phänotyp–Korrelationen über Domänengrenzen hinweg im gesamten Datenbestand ermöglicht. Mit der Abbildung der Beziehungen zwischen den einzelnen Domänen auf einen Graphen konnte zur Analyse der typische Graphalgorithmus zur Breitensuche herangezogen und angepaßt werden, um die Präsentation eines formalen Ansatzes zu Genotyp–Phänotyp–Korrelation zu vervollständigen.

Nach der Untersuchung und Gegenüberstellung der Verfahren zur Bewertung von Ähnlichkeiten innerhalb des integrierten molekularbiologischen und medizinischen Datenbestandes wurden die zur Umsetzung eines Vorschlages zur Unterstützung der Identifikation von Genotyp–Phänotyp–Korrelationen notwendigen, einzelnen Softwarekomponenten in einer Gesamtarchitektur zusammengeführt. Diese Architektur stellt die folgenden Funktionen bereit: Datenintegration, lokale Speicherung (Integrationsdatenbank), Domänendatenverwaltung, Genotyp–Phänotyp–Analyse über eine Nutzerschnittstelle. Als Nachweis für die praktische Anwendbarkeit des entwickelten Ansatzes wurde ein webbasierter Pro-

totyp des Gesamtsystemes implementiert und zur Nutzung bereitgestellt.

Mit Hilfe der Verknüpfung dieser Werkzeuge ist beispielsweise eine nutzerspezifische Integration verschiedener heterogener Datenquellen, die Anreicherung des integrierten Datenbestandes um semantische Metadaten über die abgedeckten Informationsdomänen, die Unterstützung der Navigation im integrierten Datenbestand, die Verfolgung von Beziehungen zwischen Genotyp und Phänotyp sowie der Export relevanter Datensätze in verschiedenen Formaten möglich. Als besonderer Teil dieser Architektur ist das Informationssystem für Mutationen und assoziierte Phänotypen Ramedis hervorzuheben, das ebenfalls im Rahmen dieser Arbeit entwickelt wurde und dessen Datenbestand einen wichtigen Beitrag durch die Integration unter der Domäne der klinischen Phänotypen liefert.

Aufbauend auf dem vorgestellten Ansatz zur Integration von Life–Science–Datenquellen zur Unterstützung der Suche nach Beziehungen zwischen Genotypen und Phänotypen sind eine Reihe von Erweiterungsmöglichkeiten denkbar, die im folgenden kurz vorgestellt werden. Dabei wird zwischen den Erweiterungen des sich bereits in intensiver Nutzung befindlichen Ramedis–Informationssystemes und denen der prototypischen Analyseumgebung für Genotyp–Phänotyp–Beziehungen unterschieden.

Das bestehende Ramedis-System sollte erweitert werden, um eine weitergehende Sammlung und Nutzung von klinischen Parametern und molekulargenetischen Untersuchungsergebnissen zu ermöglichen. Dazu wäre die grafische Nutzeroberfläche um eine Schnittstelle und nachgelagerte Algorithmen zu ergänzen, die auf der Basis der gespeicherten Merkmale in der Datenbank eine umfangreichere Suche nach ähnlichen Fällen realisieren. Diese fallbasierte Anfrageschnittstelle orientiert sich dabei am Prinzip des fallvergleichenden Case-based Reasoning, bei dem eine Problemsituation durch eine Menge von Problemmerkmalen beschrieben wird. Durch Zugriff auf eine Fallbasis wird eine Reihe von bereits gelösten Fällen selektiert, die durch ein geeignetes Ähnlichkeitsmaß bewertet werden. Bisher wird dazu nur eine kleine Auswahl der verfügbaren Fallmerkmale herangezogen. Durch eine Ausweitung auf weitere, spezifische Parameter könnte eine vorgeschlagene Lösung besser auf ihre Eignung im Bezug auf die aktuelle Problemsituation geprüft werden. Ziel dieses Vorgehens ist weiterhin das Auffinden von bereits dokumentierten Fällen mit ähnlichen klinischen Parametern, die beispielsweise einen Vergleich mit Simulationsergebnissen oder eine Abgrenzung von Differentialdiagnosen ermöglichen könnten.

Neben der aktuell praktizierten manuellen Sammlung und Annotation von publizierten Fallberichten aus medizinischen Fachzeitschriften und der Eingabe von Fällen durch die behandelnden Ärzte ist die Anreicherung der Datenbank in Ramedis durch die Integration von Daten aus verwandten Systemen, wie PAHdb, wünschenswert. Durch diesen Schritt könnten im Rahmen von Kooperationsvereinbarungen qualitativ hochwertige Datensätze den bereits vorhandenen Datenbestand erweitern und somit einen umfangreicheren und vollständigeren Blick auf die untersuchten Erkrankungen geben.

Bei der Anwendung des Ramedis–Systems in der Praxis hat sich gelegentlich gezeigt, daß bei besonders aktuellen Krankheitsfällen, die durch den behandelnden Arzt in Fach-

zeitschriften publiziert werden sollen, Vorbehalte gegenüber einer Verfügbarkeit von spezifischen Parametern dieser Falles im WWW bestehen. Aus diesem Grunde wäre die Implementierung einer lokal anwendbaren Version von Ramedis wünschenswert, so daß die eingegebenen Daten nur dem Autor zu Verfügung stehen, aber nach der Publikation einfach in die öffentliche Version im WWW übertragen werden können. Die bei der Entwicklung von Ramedis angelegten Methoden und Schnittstellen bilden durch ihre Plattformunabhängigkeit einen guten Ausgangspunkt für eine zeitweilige Nutzung als nicht–öffentliche Fassung in Form einer lokalen Installation. Dabei ist jedoch zu beachten, daß die Autoren geeignet motiviert werden, die Daten weiterhin für die öffentliche Version zur Verfügung zu stellen.

Neben den verschiedenen Erweiterungen an Ramedis sind auch einige Weiterentwicklungen im Rahmen der vorgestellten Analyseumgebung für Genotyp–Phänotyp–Beziehungen auf Basis des integrierten Datenbestandes vorstellbar. Da bisher die Untersuchung von Beziehungen zwischen Genotypen und Phänotypen nur halbautomatisch durchgeführt wird, sollte die Implementierung der vorgeschlagenen Graphstruktur und die Ausführung des entsprechenden Suchalgorithmus auf dem integrierten Datenbestand mit höchster Priorität verfolgt werden. Außerdem wäre eine Abbildung von Ähnlichkeiten innerhalb einzelner Domänen als Gewichte von Kanten im Graphen wünschenswert. Dabei kann auf die in dieser Arbeit vorgestellten theoretischen Vorüberlegungen und Untersuchungen zurückgegriffen werden, so daß diese weitergehenden Analysemöglichkeiten durch prototypische Entwicklungen abgerundet werden.

Zusammenfassend kann festgestellt werden, daß im Rahmen dieser Arbeit ein Informationssystem für Mutationen und assoziierte Phänotypen mit dem Namen Ramedis geschaffen wurde, das bereits erfolgreich zur Sammlung und Gegenüberstellung von Fällen seltener Stoffwechselerkrankungen eingesetzt wird. Neben diesen klinischen Informationen wurden in einer Integrationsdatenbank Daten aus unterschiedlichen molekularbiologischen und medizinischen Quellen zusammengeführt. Durch die Unterstützung mit webbasierten Softwarewerkzeugen wird dabei dem Nutzer die effiziente Navigation und Suche nach Beziehungen zwischen Genotypen und Phänotypen auf der Basis des integrierten Datenbestandes ermöglicht.

A WWW-Adressen ausgewählter molekularbiologischer Datenquellen

Die nachfolgende Übersicht zeigt die URLs der in dieser Arbeit vorgestellten molekularbiologischen und medizinischen Datenquellen. Für weitere Informationen über die wichtigsten Quellen sei auf das jährlich erscheinende Sonderheft der *Nucleic Acids Research* [Bax02, Bax03, Gal04] verwiesen.

ARPKD/PKHD1 Mutation Database	http://www.humgen.rwth-aachen.de
Braunschweig Enzyme Database (BRENDA)	http://www.brenda.uni-koeln.de
DNA Database of Japan (DDBJ)	http://www.ddbj.nig.ac.jp
EMBL Nucleotide Database	http://www.ebi.ac.uk/embl/
Frequency of Inherited Disorders Database (FIDD)	http://www.uwcm.ac.uk/uwcm/mg/fidd/
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
Genomics Unified Schema (GUS)	http://www.gusdb.org
Human Gene Mutation Database (HGMD)	http://www.hgmd.org
Knowledgebase for Inborn Errors of Metabolism (METAGENE)	http://www.metagene.de
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg/
Molecular Modeling Database (MMDB)	http://www.ncbi.nlm.nih.gov/Structure/
Online Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/Omim/

Fortsetzung auf der nächsten Seite

Fortsetzung von der vorherigen Seite

PathBlast	http://www.pathblast.org
PharmGKB	http://pharmgkb.org
Phenylalanine Hydroxylase Locus Knowledgebase (PAHdb)	http://www.pahdb.mcgill.ca
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/
Rare Metabolic Diseases Database (Ramedis)	http://www.ramedis.de
Protein Data Bank (PDB)	http://www.rcsb.org/pdb/
Protein Information Resource (PIR)	http://pir.georgetown.edu
Semantic Meta Database (SEMEDA)	http://www-bm.ipk-gatersleben.de/semeda/
Structural Classification of Proteins (SCOP)	http://scop.mrc-lmb.cam.ac.uk/scop/
SWISS-PROT Protein Knowledgebase	http://www.expasy.org/sprot/
Tetrahydrobiopterin Deficiencies Database (BIODEF)	http://www.bh4.org
TRANSFAC	http://transfac.gbf.de/TRANSFAC/
TRANSPATH	http://www.biobase.de/pages/products/
Universal Protein Knowledgebase (UniProt)	http://www.uniprot.org

B Ausgewählte Datenquellen im Detail

Der folgende Abschnitt soll die Struktur der Daten und des Integrationsvorganges einer ausgewählten Quelle beleuchten. Dazu werden anhand von Ausschnitten der jeweiligen Darstellung der Daten Bemerkungen zu den unterschiedlichen Formaten und Bedeutungen gemacht. Anschließend werden die erforderlichen Schritte zur Integration dieser Datenquelle mit dem BioDataServer nach [FHL+02] erläutert.

B.1 EMBL

Die Datenquelle EMBL wurde als genomische Sequenzdatenbank bereits im Abschnitt 3.1.2 vorgestellt. Die Daten werden für den Nutzer als Flatfile, HTML–Dokument oder als XML–Datei bereitgestellt. In den beiden nachfolgenden Abschnitten ist für einen Datensatz ein Auszug des Flatfiles und des XML–Dokumentes dargestellt.

B.1.1 Originaldatensatz als Flatfile

```
ID   AF404777   standard; DNA; HUM; 171266 BP.
XX
AC   AF404777;
XX
SV   AF404777.1
XX
DT   23-FEB-2002 (Rel. 70, Created)
DT   23-FEB-2002 (Rel. 70, Last updated, Version 1)
XX
DE   Homo sapiens phenylalanine hydroxylase (PAH) gene, complete cds.
XX
KW   .
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-171266
RA   Konecki D.S., Lichter-Konecki U.;
RT   "Completion of the sequence of PAH, a model disease gene";
```

```
RL   Unpublished.
XX
RN   [2]
RP   1-171266
RA   Konecki D.S., Lichter-Konecki U.;
RT   ;
RL   Submitted (13-JUL-2001) to the EMBL/GenBank/DDBJ databases.
RL   Medical Genetics Branch, National Human Genome Research Institute, 10
RL   Center Drive, Bldg. 10 - 3D45, Bethesda, MD 20892, USA
XX
DR   GOA; Q8TEY0; Q8TEY0.
DR   SPTREMBL; Q8TEY0; Q8TEY0.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..171266
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
FT   mRNA            join(27306..27838,32011..32118,49993..50176,67364..67452,
FT                   78328..78395,89668..89864,92050..92185,93244..93313,
FT                   98051..98107,100571..100666,101223..101356,104487..104602,
FT                   105784..106676)
FT                   /gene="PAH"
FT                   /product="phenylalanine hydroxylase"
FT   CDS             join(27779..27838,32011..32118,49993..50176,67364..67452,
FT                   78328..78395,89668..89864,92050..92185,93244..93313,
FT                   98051..98107,100571..100666,101223..101356,104487..104602,
FT                   105784..105827)
FT                   /codon_start=1
FT                   /db_xref="GOA:Q8TEY0"
FT                   /db_xref="SPTREMBL:Q8TEY0"
FT                   /gene="PAH"
FT                   /product="phenylalanine hydroxylase"
FT                   /protein_id="AAL78816.1"
FT                   /translation="MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNGAISLIFSLKEEV
FT                   GALAKVLRLFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGA
FT                   TVHELSRDKKKDTVPWFPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRKQFAD
FT                   IAYNYRHGQPIPRVEYMEEEKKTWGTVFKTLKSLYKTHACYEYNHIFPLLEKYCGFHED
FT                   NIPQLEDVSQFLQTCTGFRLRPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPQP
FT                   DICHELLGHVPLFSDRSFAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKQGDSI
FT                   KAYGAGLLSSFGELQYCLSEKPKLLPLELEKTAIQNYTVTEFQPLYYVAESFNDAKEKV
FT                   RNFAATIPRPFSVRYDPYTQRIEVLDNTQQLKILADSINSEIGILCSALQKIK"
XX
SQ   Sequence 171266 BP; 50913 A; 34540 C; 35053 G; 50760 T; 0 other;
     ggaaatcctt tccaaaaagt actactacaa acaagcccag atggtgagaa ctatataaaa        60
     cactaactct tcaatgccca gacaccaatg aactagatgg tcaagtatca agatcatcta       120

     ...

     tataggtcaa gtgtttggtc caataaatgg ataagttgga actcaataat ccttattata    171240
     acttccagag tagagtgatg tttaaa                                         171266
//
```

B.1.2 Originaldatensatz im XML–Format

```
<?xml version="1.0" encoding="UTF-8" ?>
<?format DECIMAL="."?>
<!DOCTYPE Bsml (View Source for full doctype...)>
<!--  The BSML specification was created by Joseph H. Spitzner, Ph.D., LabBook, Inc.
 http://www.labbook.com -->
<Bsml>
 <Definitions>
```

```
<Sequences>
 <Sequence id="AF404777" ic-acckey="AF404777" title="AF404777" comment="Homo
  sapiens phenylalanine hydroxylase (PAH) gene, complete cds." length="171266"
  topology="linear" molecule="dna" representation="raw">
  <Attribute name="version" content="AF404777.1" />
  <Attribute name="organism-species" content="Homo sapiens (man)" />
  <Attribute name="organism-classification" content="Eukaryota; Metazoa; Chordata;
   Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini;
   Hominidae; Homo" />
  <Attribute name="source" content="Homo sapiens" />
  <Attribute name="date-created" content="23-FEB-2002" />
  <Attribute name="date-last-updated" content="23-FEB-2002" />
  <Attribute name="database-xref" content="SPTREMBL:Q8TEY0" />
  <Attribute name="database-xref" content="GOA:Q8TEY0" />
  <Attribute name="database-xref" content="EPD:EP74106" />
  <Feature-tables>
   <Feature-table>
    <Reference>
     <RefAuthors>Konecki D.S., Lichter-Konecki U.</RefAuthors>
     <RefTitle>Completion of the sequence of PAH, a model disease gene</RefTitle>
     <RefJournal>Unpublished Reference</RefJournal>
    </Reference>
    <Reference>
     <RefAuthors>Konecki D.S., Lichter-Konecki U.</RefAuthors>
     <RefJournal>Submitted (13-JUL-2001) to the EMBL/GenBank/DDBJ databases.
      Medical Genetics Branch, National Human Genome Research Institute, 10 Center
      Drive, Bldg. 10 - 3D45, Bethesda, MD 20892, USA</RefJournal>
    </Reference>
    <Feature id="FTR_AF404777.1_0" class="SOURCE" value-type="source"
     title="source" display-auto="1">
     <Qualifier value-type="organism" value="Homo sapiens" />
     <Qualifier value-type="db_xref" value="TAXONOMY:9606" />
     <Interval-loc startpos="1" endpos="171266" startopen="0" endopen="0"
      onepos="0" complement="0" />
    </Feature>
    <Feature id="FTR_AF404777.1_1" class="MRNA" value-type="mrna" title="PAH"
     display-auto="1">
     <Qualifier value-type="gene" value="PAH" />
     <Qualifier value-type="product" value="phenylalanine hydroxylase" />
     <Qualifier value-type="join" value="join(27306..27838,32011..32118,
      49993..50176,67364..67452,78328..78395,89668..89864,92050..92185,
      93244..93313,98051..98107,100571..100666,101223..101356,104487..104602,
      105784..106676)" />
     <Interval-loc startpos="27306" endpos="106676" startopen="0" endopen="0"
      onepos="0" complement="0" />
    </Feature>
    <Feature id="FTR_AF404777.1_2" class="CDS" value-type="cds" title="PAH"
     display-auto="1">
     <Qualifier value-type="gene" value="PAH" />
     <Qualifier value-type="product" value="phenylalanine hydroxylase" />
     <Qualifier value-type="codon_start" value="1" />
     <Qualifier value-type="translation" value="MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNG
      AISLIFSLKEEVGALAKVLRLFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGATVHE
      LSRDKKKDTVPWFPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRVEYMEEE
      KKTWGTVFKTLKSLYKTHACYEYNHIFPLLEKYCGFHEDNIPQLEDVSQFLQTCTGFRLRPVAGLLSSRDFLGGL
      AFRVFHCTQYIRHGSKPMYTPQPDICHELLGHVPLFSDRSFAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGL
      CKQGDSIKAYGAGLLSSFGELQYCLSEKPKLLPLELEKTAIQNYTVTEFQPLYYVAESFNDAKEKVRNFAATIPR
      PFSVRYDPYTQRIEVLDNTQQLKILADSINSEIGILCSALQKIK" />
     <Qualifier value-type="db_xref" value="GOA:Q8TEY0" />
     <Qualifier value-type="db_xref" value="SPTREMBL:Q8TEY0" />
     <Qualifier value-type="db_xref" value="PID:AAL78816.1" />
     <Qualifier value-type="join" value="join(27779..27838,32011..32118,
      49993..50176,67364..67452,78328..78395,89668..89864,92050..92185,
      93244..93313,98051..98107,100571..100666,101223..101356,104487..104602,
      105784..105827)" />
     <Interval-loc startpos="27779" endpos="105827" startopen="0" endopen="0"
      onepos="0" complement="0" />
```

```
      </Feature>
     </Feature-table>
    </Feature-tables>
    <Seq-data>ggaaatcctttccaaaaagtactactacaaacaagcccagatggtgagaa ctatataaaacactaact
    cttcaatgcccagacaccaatgaactagatgg tcaagtatcaagatcatctagaaaaacacgacctcaccaaatgaac
    taaa taagacaccaggaaccaatgctgcagagacagagatatgtgacctttcag acagggaattcagaatacctgct
    ttgaggaaaatcaacataattcaagat aacatagagaa
    ...
    aggatctccaatatcttcctccaatatcttcaggatagtaatctttatta ttatgaacttattgcagttaaaatgctc
    agtataggtcaagtgtttggtc caataaatggataagttggaactcaataatccttattataacttccagag tagag
    tgatgtttaaa</Seq-data>
   </Sequence>
  </Sequences>
 </Definitions>
</Bsml>
```

B.1.3 Adapterschema

```
XmlEMBLAdapter
class "main" {
name: "acc" type: "string" description: "" isKey: "true"
name: "seq_id" type: "string" description: "" isKey: "false"
name: "comment" type: "string" description: "" isKey: "false"
name: "length" type: "integer" description: "" isKey: "false"
name: "molecule" type: "integer" description: "" isKey: "false"
name: "version" type: "string" description: "" isKey: "false"
name: "organism" type: "string" description: "" isKey: "false"
name: "organism_classification" type: "string" description: "" isKey: "false"
name: "source" type: "string" description: "" isKey: "false"
name: "keywords" type: "string" description: "" isKey: "false"
name: "date_create" type: "string" description: "" isKey: "false"
name: "date_last_update" type: "string" description: "" isKey: "false"
name: "database_ref" type: "string" description: "" isKey: "false"
name: "sequence" type: "string" description: "" isKey: "false"
name: "embl_link" type: "string" description: "" isKey: "false"
name: "feature_acc_refs" type: "string" description: "" isKey: "false"
name: "mrna" type: "string" description: "" isKey: "false"
name: "gene" type: "string" description: "" isKey: "false"
name: "note" type: "string" description: "" isKey: "false"
name: "translation" type: "string" description: "" isKey: "false"
name: "function" type: "string" description: "" isKey: "false"
name: "product" type: "string" description: "" isKey: "false"
name: "ec" type: "string" description: "" isKey: "false"
name: "feature_link" type: "string" description: "" isKey: "false"
name: "startpos" type: "string" description: "" isKey: "false"
name: "endpos" type: "string" description: "" isKey: "false"
}
```

C Beispiel für einen Fallbericht

Die nachfolgende Übersicht zeigt beispielhaft den Umfang der Daten, die für einen Fallbericht in der Mutationsdatenbank Ramedis aufgenommen werden. Dabei wurden die originalen englischen Bezeichnungen der einzelnen Datensätze beibehalten. Wiederholt aufgenommene Daten, wie Laborparameter, sind mit dem Zeitpunkt der Untersuchung oder mit dem Beginn der Behandlung in Form des Alters versehen.

Allgemeine Daten

Patient-ID	81
Diagnosis	PHENYLKETONURIA; PKU
Gender	male
Age of first Symtoms Onset	0 Day(s)
Age of Diagnosis	5 Day(s)
Found in Newborn Screening	yes
Country	Germany
Ethnic Origin	Mother: German; Father: German
Author	Trefz, Friedrich Karl
Coauthors	M.K.
Hospital	Kinderklinik Reutlingen - Friedrich Karl Trefz
History	Mild hyperphenylalaninemia found in newborn screening program. Treatment with phe-restricted diet up to the age of 8 years. BH4-loading at the age of 8 years showed slight decrease of phe-level. Therapy with tetrahydrobiopterine was initiated since BH4-sensitivity had been confirmed. Normal psychomotoric development at the age of 8 years. IQ of 122 (03/2002)
Date of Entry	31-Mar-01

Fortsetzung auf der nächsten Seite

Fortsetzung von der vorherigen Seite

Molekulargenetik

Gene name	Non-PKU, HPA
Genotype	mut. / null
Allele 1	
Trivial Name	E390G
Systematic Name	c.1169A → G
Amino Acid Change	Glu → Gly
Nucleotide Change	GAG → GGG
Exon/Intron	Boundary/Promotor 11
Allele 2	
Trivial Name	R408W
Systematic Name	c.1222C → T
Amino Acid Change	Arg → Trp
Nucleotide Change	CGG → TGG
Exon/Intron	Boundary/Promotor 12

Symptome

tetrahydrobiopterin-sensitivity	yes, 8.74 Year(s)
no clinical signs or symptoms	yes, 8.41 Year(s)

Laborparameter

Phenylalanine	8.41 Year(s), 468 micro-mol/l, blood, Increased
Phenylalanine	8.47 Year(s), 336 micro-mol/l, blood, Increased
Phenylalanine	8.54 Year(s), 540 micro-mol/l, blood, Increased
Phenylalanine	8.7 Year(s), 360 micro-mol/l, blood, Increased
Tyrosine	8.41 Year(s), 93.5 micro-mol/l, blood, Normal
Tyrosine	8.47 Year(s), 77 micro-mol/l, blood, Normal
Tyrosine	8.54 Year(s), 132 micro-mol/l, blood, Normal
Tyrosine	8.64 Year(s), 82.5 micro-mol/l, blood, Normal

Therapie, Entwicklung

length	8.41 Year(s), 135 cm, development

Fortsetzung auf der nächsten Seite

Fortsetzung von der vorherigen Seite

length	10.47 Year(s), 146.5 cm, development
phenylalanine-restricted diet	10 Day(s)
weight	8.41 Year(s), 28.8 kg, development
weight	10.47 Year(s), 33.8 kg, development

Diät, Medikamente

artificial protein intake (specific amino acid mixture)	10.47 Year(s), 4.1 g/day oral
phenylalanine	8.73 Year(s), 1187 mg/day oral
protein (total)	8.73 Year(s), 12.3 g/day oral

D Glossar

Für eine Reihe von Begriffen, die in der vorliegenden Arbeit nicht durch gekennzeichnete Definitionen festgelegt wurden, soll in diesem Abschnitt zusammenfassend eine kurze Erläuterung gegeben werden

---

Base Bestandteil von Nukleotiden: die Purin–Basen Adenin und Guanin, sowie Pyrimidin–Basen Cytosin, Uracil und Thymin

Basenpaarung Bildung von Basenpaaren in doppelsträngiger DNS oder RNS

Chromosom Fadenförmiges Gebilde im Zellkern, das die Erbinformation in Form von Genen auf der DNS trägt; Weitergabe durch Replikation an die nächste Zellgeneration

Cistron Bezeichnung für eine Nukleotidsequenz, die eine biochemische Funktionseinheit codiert; identisch mit einem Gen

Codon Sequenz von drei aufeinanderfolgenden Basen, die als Bestandteil der Boten–RNS jeweils den Einbau einer Aminosäure, die Termination oder die Initiation signalisieren; wird auch als Basentripplet oder Code–Triplett bezeichnet

Deletion Chromosomen–Mutation durch den Verlust eines Abschnittes eines Chromosoms

DNS Träger der genetischen Information durch eine Doppelhelix, die aus zwei komplementären Strängen von verketteten Nukleotiden besteht; engl. DNA

Enzym Proteine, die an den meisten biochemischen Vorgängen in Organismen beteiligt sind, da sie die erforderliche Aktivierungsenergie herabsetzen und damit die Reaktionsgeschwindigkeit dieser Prozesse erhöhen

Genotyp Menge der genetischen Informationen eines Organismus

Insertion Chromosomen–Mutation durch Einbau eines DNS–Abschnittes in ein Chromosom

Life Science Lebenswissenschaften

Metabolic Pathway in Wechselwirkung stehende biochemische Reaktionen (Stoffwechselweg)

Metabolismus Menge der biochemischen Reaktionen in einer Zelle oder in einem Organismus

Mutation Veränderung des Erbgutes, die nicht durch Rekombination und Segregation entstehen

Nukleotid Chemische Verbindung aus Base, Pentose und Phosphorsäure, die als Bausteine in Nukleinsäuren verknüpft sind

RNS Ein- oder doppelstränge Nukleinsäure, enthält statt der Base Thymin Uracil; engl. RNA

Sequenz Lineare Anordnung (Kette) von Nukleotiden oder Aminosäuren

Transkription Biochemischer Prozeß zur Abschrift der Nukleotidsequenz eines Gen–Abschnittes von der DNS auf die Boten–RNS (mRNA)

Translation Biochemischer Prozeß, der in mehreren Teilschritten die Nukleotidsequenz der mRNA in die Aminosäuresequenz des Proteins übersetzt und dabei die verschiedenen Aminosäuren zu einem Protein verkettet

Translokation Verlagerung eines Chromosomenstückes von seinem ursprünglichen Ort auf ein anderes Chromosom oder an eine andere Stelle des gleichen Chromosoms

Phänotyp Erscheinungsbild eines Individuums aus direkt oder nicht direkt beobachtbaren Merkmalen

SI Internationales Einheitensystem (Systéme International d' Unités)

Literaturverzeichnis

[ABW+04] P. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi und L. L. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(1):115–119, 2004.

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers und D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[AHK+01] L. N. Al–Jadar, P. S. Harper, M. Krawczak, S. R. Palmer, B. N. Johansen und D. N. Cooper. The Frequency of Inherited Disorders Database. *Human Genetics*, 108(1):72–74, 2001.

[AP94] A. Aamodt und E. Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1):39–59, 1994.

[BA00] A. Bairoch und R. Apweiler. The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.

[Bax02] A. D. Baxevanis. The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Research*, 30(1):1–12, 2002.

[Bax03] A. D. Baxevanis. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Research*, 31(1):1–12, 2003.

[BBA+03] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout und M. Schneider. The SWISS–PROT protein sequence database and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[BBB98] G. Borsani, A. Ballabio und S. Banfi. A practical guide to orient yourself in the labyrinth of genome databases. *Human Molecular Genetics*, 7(10):1641–1648, 1998.

[BD96] T. Beuter und P. Dadam. Prinzipien der Replikationskontrolle in verteilten Systemen. *Informatik Forschung und Entwicklung*, 11(4):203–212, 1996.

[BGH+01] W. C. Barker, J. S. Garavelli, Z. Hou, H. Huang, R. S. Ledley, P. B. McGarvey, H.-W. Mewes, B. C. Orcutt, F. Pfeiffer, A. Tsugita, C. R. Vinayaka, C. Xiao, L. L. Yeh und C. Wu. Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Research*, 29(1):29–32, 2001.

[BK03] F. Bry und P. Kröger. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. *Distributed and Parallel Databases*, 13:7–42, 2003.

[BKK96] H.-J. Böhm, G. Klebe und H. Kubinyi. *Wirkstoffdesign*. Heidelberg; Berlin; Oxford: Spektrum Akademischer Verlag, 1996.

[BKL+03] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell und D. L. Wheeler. GenBank. *Nucleic Acids Research*, 31(1):23–27, 2003.

[BKL+04] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell und D. L. Wheeler. GenBank: update. *Nucleic Acids Research*, 32(1):23–26, 2004.

[Bla96] N. Blau. *The Hyperphenyalaninemias. A Differential Diagnosis and International Database of Tetrahydrobiopterin Deficiencies*. Marburg: Tectum Verlag, 1996.

[BQ01] A. D. Baxevanis und B. F. F. Quellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. New York: Wiley Interscience, 2001.

[Bro99] T. A. Brown. *Moderne Genetik*. Heidelberg; Berlin: Spektrum Akademischer Verlag, 1999. 2. Auflage.

[Bud89] E. Buddecke. *Grundriss der Biochemie*. Berlin; New York: de Gruyter, 1989.

[BWF+00] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov und P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[CBK98] N. D. Cooper, E. V. Ball und M: Krawczak. The human gene mutation database. *Nucleic Acids Research*, 26(1):285–287, 1998.

[Che02] M. Chen. Metabolic Pathway Alignment and Alternative Pathway Identification. In *Poster Abstracts Book of the European Conference on Computational Biology (ECCB) 2002, October 6–9, 2002, Saarbrücken, Germany*, S. 40–41, 2002.

[Con97] S. Conrad. *Föderierte Datenbanksysteme: Konzepte der Datenintegration*. Berlin; Heidelberg: Springer–Verlag, 1997.

[Con01] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[Cou02] Couzin, J. BREAKTHROUGH OF THE YEAR: Small RNAs Make Big Splash. *Science*, 298:2296–2297, 2002.

[DA01] J. T. den Dunnen und S. E. Antonarakis. Nomenclature for the description of human sequence variations. *Human Genetics*, 109:121–124, 2001.

[DCB+01] S. B. Davidson, J. Crabtree, B. P. Brunk, J. Schug, V. Tannen, G. C. Overton und Jr. C. J. Stoeckert. K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal*, 40(2):512–531, 2001.

[DEF+02] S. Döhr, F. Ehrentreich, G. Frauendienst-Egger, R. Hofestädt, O. Hofmann, M. Lange, U. Mischke, A. Potapov, D. Scheible, R. Schnee, U. Scholz, D. Schomburg, K. Seidl, T. Töpel, F.-K. Trefz, T. Werner und E. Wingender. Modeling of gene regulatory networks for genotype–phenotype information. *German Human Genome Project: Progress Report 1999 – 2002*, S. 70–71, 2002.

[DG98] G. Dodge und T. Gorman. *Oracle8 Data Warehousing*. New York: Wiley Computer Publishing, 1998.

[DOB95] S. B. Davidson, C. Overton und P. Buneman. Challenges in Integrating Biological Data Sources. *Journal of Computational Biology*, 2(4):557–572, 1995.

[DOTW97] S. B. Davidson, C. Overton, V. Tannen und L. Wong. BioKleisli: a digital library for biomedical researchers. *International Journal on Digital Libraries*, 1:36–53, 1997.

[DSS+99] T. Dandekar, S. Schuster, B. Snel, M. Huynen und P. Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochemical Journal*, 343:115–124, 1999.

[EA93] T. Etzold und P. Argos. SRS – an indexing and retrieval tools for flat-files data libraries. *CABIOS*, 9(1):49–57, 1993.

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

[FGL95] K. D. Forbus, D. Gentner und K. Law. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2):141–205, 1995.

[FHL+02] A. Freier, R. Hofestädt, M. Lange, U. Scholz und A. Stephanik. BioDataServer: A SQL–based service for the online integration of life science data. *In Silico Biology*, 2(0005), 2002. *Online Journal*: `http://www.bioinfo.de/isb/2002/02/0005/`.

[FSG04] H.-C. Fehmann, M. Z. Strowski und B. Göke. Diabetes mellitus mit monogen determinierter Störung der Beta–Zell–Funktion. *Deutsches Ärzteblatt*, 101(13):860–867, 2004.

[FT98] G. Frauendienst-Egger und F. K. Trefz. *METAGENE 3.0 Computersystem zur Diagnoseunterstützung angeborener Stoffwechselerkrankungen*. Stuttgart: Wissenschaftliche Verlagsgesellschaft mbH, 1998.

[Gal04] M. Y. Galperin. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Research*, 32(1):3–22, 2004.

[GJ02] C. Gibas und P. Jambeck. *Einführung in die Praktische Bioinformatik*. Köln: O'Reilly Verlag, 2002.

[Goo96] K. Goos. *Fallbasiertes Klassifizieren: Methoden, Integration und Evaluation*. St. Augustin: Infix, 1996.

[GSN+01] C. A. Goble, R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim und A. Brass. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40(2):532–551, 2001.

[Hin00] S. Hinze. Entwicklung einer Auswertungs- und Case-Based-Reasoning Komponente für die Patientendatenbank RAMEDIS. Diplomarbeit, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, Institut für Technische und Betriebliche Informationssysteme, Dezember 2000.

[HM85] D. Heimbigner und D. McLeod. A federated architecture for information management. *ACM Transactions on Information Systems (TOIS)*, 3(3):253–278, 1985.

[HMPS99] R. Hofestädt, U. Mischke, M. Prüß und U. Scholz. Metabolic Drug Pointing and Information Processing. *Medical Informatics Europe*, 68:12–15, 1999.

[Hof96] R. Hofestädt. *Theorie der regelbasierten Modellierung des Zellstoffwechsels*. Aachen: Shaker, 1996.

[HOR+02] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman und T. E. Klein. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Research*, 30(1):163–165, 2002.

[HS97] A. Heuer und G. Saake. *Datenbanken: Konzepte und Sprachen*. International Thomson Publishing, 1997.

[HSA+02] A. Hamosh, A. F. Scott, J. Amberger, , C. Bocchini, D. Valle und V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.

[HSK+01] L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice und W. C. Swope. DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, 40(2):489–511, 2001.

[KAA+04] T. Kulikova, P. Aldebert, N. Althorpe, W. Baker, K. Bates, P. Browne, A. van den Broek, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, M. Garcia–Pastor, N. Harte, C. Kanz, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, P. Stoehr, G. Stoesser, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu und R. Apweiler. The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, 32(1):27–30, 2004.

[Kar95] P. D. Karp. A Strategy for Database Interoperation. *Journal of Computational Biology*, 2(4):573–586, 1995.

[KBF+00] M. Krawczak, E. V. Ball, I. Fenton, P. D. Stenson, S. Abeysinghe, N. Thomas und D. N. Cooper. MDI Special Article – Human Gene Mutation Database – A Biomedical Information and Research Resource. *Human mutation*, 15(1):45–51, 2000.

[Kem69] O. Kempthorne. *An Introduction to Genetic Statistics*. Ames: The Iowa State University Press, 1969.

[KGK+04] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno und M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(1):277–280, 2004.

[KGKN02] M. Kanehisa, S. Goto, S. Kawashima und A. Nakaya. The KEGG database at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[Kni97] R. Knippers. *Molekulare Genetik*. Stuttgart, New York: Georg Thieme Verlag, 7. Auflage, 1997.

[Köh03] J. Köhler. *SEMEDA (Semantic Meta–Database): Ontology Based Semantic Integration of Biological Databases*. Dissertation, University Bielefeld, Technical Faculty, 2003.

[Kol83] J. L. Kolodner. Maintaining Organization in a Dynamic Long-Term Memory. *Cognitive Science*, 7(4):243–280, 1983.

[KSK+03] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell und T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 100(20):11394–11399, 2003.

[KST02] A. Karmiloff–Smith, G. Scerif und M. Thomas. Different Approaches to Relating Gernotype to Phenotype in Developmental Disorders. *Developmental psychobiology*, 40(3):311–322, 2002.

[KTJ+97] E. Kayaalp, E. Treacy, Waters. P. J., S. Byck, P. M. Nowacki und C. R. Scriver. Human Phenylalanine Hydroxylase Mutations and Hyperphenylalaninemia Phenotypes: A Metanalysis of Genotype-Phenotype-Correlations. *The American Journal of Human Genetics*, 61(1):1309–1317, 1997.

[KTSH01] R. Kauert, T. Töpel, U. Scholz und R. Hofestädt. Information System for the Support of Research, Diagnosis and Therapy of Inborn Metabolic Diseases. In *MedInfo 2001: Proceedings of the 10th World Congress on Health and Medical Informatics, London, September 2 – 5, 2001*, S. 353–356. Amsterdam: IOS Press, 2001.

[KVC+03] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov und E. Wingender. TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Research*, 31(1):97–100, 2003.

[KWK+01] K. Kromeyer–Hauschild, M. Wabitsch, D. Kunze, F. Geller, H. C. Geiß, V. Hesse, A. von Hippel, U. Jaeger, D. Johnsen, W. Korte, K. Menner, G. Müller, J. M. Müller, A. Niemann–Pilatus, T. Remer, F. Schaefer, H.–U. Wittchen, S. Zabransky, K. Zellner, A. Ziegler und J. Hebebrand. Perzentile fr den Body-mass-Index fr das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben. *Monatsschrift Kinderheilkunde*, 149(8):807–818, 2001.

[KZL00] R. Küffner, R. Zimmer und T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825–836, 2000.

[Len96] K. Lengnink. *Formalisierungen von Ähnlichkeit aus Sicht der Formalen Begriffsanalyse*. Aachen: Shaker, 1996.

[LMNR04] Z. Lacroix, H. Murthy, F. Naumann und L. Raschid. Links and paths through life science data sources. In E. Rahm, Herausgeber, *Data Integration in the Life Sciences, First International Workshop, DILS 2004, Leipzig, Germany, March 25–26, 2004, Proceedings*, *Lecture Notes in Computer Science*, Band 2994, S. 203–211. Springer, 2004.

[LMR90] W. Litwin, L. Mark und N. Roussopoulos. Interoperability of multiple autonomous databases. *ACM Computing Surveys (CSUR)*, 22(3):267–293, 1990.

[LNRV03] Z. Lacroix, F. Naumann, L. Raschid und E. M. Vidal. Exploring life science data sources. In *Proceedings of the IJCAI–03 Workshop on Information Integration on the Web (IIWeb–03), Acapulco, Mexico*, 2003.

[LR03] U. Leser und P. Rieger. Integration molekularbiologischer Daten. *Datenbankspektrum*, 3(6):56–66, 2003.

[Lül99] H. Lüllmann. *Pharmakologie und Toxikologie*. Stuttgart; New York: Thieme, 1999.

[Mav90] M. L. Mavrovouniotis. Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnology – Bioengineering*, 36:1070–1082, 1990.

[McK98] V. A. McKusick. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. Baltimore: John Hopkins University Press, 12. Auflage, 1998.

[MFG+03] V. Matys, E. Fricke, R. Geffers, E. Gšling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, H. Lewicki-Potapov, B. Michael, R. Mnch, S. Reuter, I. Rotert, H. Saxel, M. Scheer, S. Thiele und E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[Mic99] G. Michal. *Biochemical Pathways*. Heidelberg: Spektrum Akademischer Verlag, 1999.

[MR95] V. M. Markowitz und O. Ritter. Characterizing heterogeneous molecular biology database systems. *Journal of Computational Biology*, 2(4):547–556, 1995.

[MSGT03] S. Miyazaki, H. Sugawara, T. Gojobori und Y. Tateno. DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Research*, 31(1):13–16, 2003.

[MSI+04] S. Miyazaki, H. Sugawara, K. Ikeo, T. Gojobori und Y. Tateno. DDBJ in the stream of various biological data. *Nucleic Acids Research*, 32(1):31–34, 2004.

[MST+01] U. Mischke, U. Scholz, T. Töpel, D. Scheible, R. Hofestädt und F. K. Trefz. RAMEDIS – Rare Metabolic Diseases Publishing Tool for Genotype-Phenotype Correlation. In *MedInfo 2001: Proceedings of the 10th World Congress on Health and Medical Informatics, London, September 2 – 5, 2001*, S. 970–974. Amsterdam: IOS Press, 2001.

[MTM02] S. Miyake, Y. Tohsato und H. Matsuda. An Application of a Pathway Alignment Method to Comparative Analysis between Genome and Pathways. In *1st IEEE Computer Society Bioinformatics Conference (CSB 2002), 14–16 August 2002, Stanford, CA, USA*, S. 329. IEEE Computer Society, 2002.

[Mut96] E. Mutschler. *Arzneimittelwirkungen: Lehrbuch der Pharmakologie und Toxikologie*. Stuttgart: Wissenschaftliche Verlagsgesellschaft mbH, 1996.

[NBPS98] P. M. Nowacki, S. Byck, L. Prevost und C. R. Scriver. PAH Mutation Analysis Consortium Database: 1997. Prototype for relational locus–specific mutation databases. *Nucleic Acids Research*, 26(1):220–225, 1998.

[Nic03] F. W. Nicholas. Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. *Nucleic Acids Research*, 31(1):275–277, 2003.

[NW70] S. B. Needleman und C. D. Wunsch. A General Method Applicable to the Search of Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[OGFK98] H. Ogata, S. Goto, W. Fujibuchi und M. Kaneshisa. Computation with the KEGG pathway database. *BioSystems*, 47:119–128, 1998.

[PPW+03] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell und B. O. Palsson. Metabolic pathways in the post–genome era. *Trends in Biochemical Sciences*, 28(5):250–258, 2003.

[PSN+99] T. Pfeiffer, I. Sánchez–Valdenebro, J. C. Nuño, F. Montero und S. Schuster. METATOOL: for studying metabolic networks. *Bioinformatics*, 15(3):251–257, 1999.

[Rau01] R Rauhut. *Bioinformatik: Sequenz – Struktur – Funktion*. Weinheim et al.: Wiley-VCH, 2001.

[Saa93] G. Saake. *Objektorientierte Spezifikation von Informationssystemen*. Teubner, 1993.

[SBB+03] G. Stoesser, W. Baker, A. van den Broek, M. Garcia–Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara und R. Vaughan. The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Research*, 31(1):17–21, 2003.

[SCE+04] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn und D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(1):431–433, 2004.

[Sch82] R. C. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, 1982.

[Sch02] U. Scholz. *FRIDAQ – Ein Framework zur Integration molekularbiologischer Datenbestände*. Aachen: Shaker, 2002.

[SCS02] I. Schomburg, A. Chang und D. Schomburg. BRENDA; enzyme data and metabolic information. *Nucleic Acids Research*, 30(1):47–49, 2002.

[SHK+03] C.R. Scriver, M. Hurtubise, D. Konecki, M. Phommarinh, L. Prevost, H. Erlandsen, R. Stevens, P.J. Waters, S. Ryan, D. McDonald und C. Sarkissian. PAHdb 2003: What a locus–specific knowledgebase can do. *Human mutation*, 21(4):333–344, 2003.

[SL90] A. P. Sheth und J. A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR)*, 22(3):183–236, 1990.

[SLP+01] H. Seo, D.-Y. Lee, S. Park, L. T. Fan, S. Shafie, B. Bertók und F. Friedler. Graph–theoretical identification of pathways for biochemical reactions. *Biotechnology Letters*, 23(1):1551–1557, 2001.

[Ste87] H. Stengel. *Erbkrankheiten: Entstehung, Vererbung und Verhütung erbl. bedingter Entwicklungsstörungen, Anomalien und Krankheiten*. Stuttgart; New York: Schattauer, 1987.

[SWS+00] C. R. Scriver, P. J. Waters, C. Sarkissian, S. Ryan, L. Prevost, D. Cote, J. Novak, Saeed Teebi und P. Nowacki. PAHdb: A Locus–Specific Knowledgebase. *Human mutation*, 15(1):99–104, 2000.

[Tan02] A. S. Tanenbaum. *Moderne Betriebssysteme*. München: Pearson Studium, 2002.

[The02a] The News and Editorial Staffs. BREAKTHROUGH OF THE YEAR: Scorecard 2002. *Science*, 298:2299, 2002.

[The02b] The News and Editorial Staffs. BREAKTHROUGH OF THE YEAR: The Runners-Up. *Science*, 298:2297–2303, 2002.

[TMH00] Y. Tohsato, H. Matsuda und A. Hashimoto. A Multiple Alignment Algorithm for Metabolic Pathway Analysis using Enzyme Hierarchy. In P. E. Bourne, E. Gribskov, R. B. Altman, N. Jensen, D. Debra Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande und H. Helge Weissig, Herausgeber, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB), August 19–23, 2000, La Jolla / San Diego, CA, USA*, S. 376–383. AAAI, 2000.

[TSM+02] T. Töpel, U. Scholz, U. Mischke, D. Scheible, R. Hofestädt und F. Trefz. Supporting genotype-phenotype correlation with the rare metabolic diseases database Ramedis . *In Silico Biology*, 2(3):407–414, 2002.

[Ven01] Venter, J.C. et. al. The Sequence of the Human Genome. *Science*, 291:1304–1351, 2001.

[Wat95] I. Watson. An Introduction to Cased-Based Reasoning. In I. Watson, Herausgeber, *Progress in Case–Based Reasoning: First United Kingdom Workshop,*

Salford, UK, January 12, 1995; Proceedings, S. 3–16. Berlin; Heidelberg: Springer–Verlag, 1995.

[WC53] J. D. Watson und F. H. C. Crick. Genetical Implications of the Structure of Deoxyribonucleid Acid. *Nature*, 171(4361):964–967, 1953.

[WCF+01] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, M. Prüß, F. Schacherer, S. Thiele und S. Urbach. The TRANSFAC system on gene expression regulation. *Nucleic Acids Research*, 29(1):281–283, 2001.

[Wes96] S. Wess. *Fallbasiertes Problemlösen in wissensbasierten Systemen zur Entscheidungsunterstützung und Diagnostik: Grundlagen, Systeme und Entscheidungen*. Sankt Augustin: Infix, 1996.

[WPNS98] P. J. Waters, M. A. Parniak, P. Nowacki und C. R. Scriver. Mutation Update – In Vitro Expression Analysis of Mutations in Phenylalanine Hydroxylase: Linking Genotype to Phenotype and Structure to Function. *Human mutation*, 11(1):4–17, 1998.

[WYH+03] C. H. Wu, L. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang und W. C. Barker. The Protein Information Resource. *Nucleic Acids Research*, 31(1):345–347, 2003.