Silke Eckstein

Informations-management in der Systembiologie

Datenbanken, Integration, Modellierung

Springer

Informationsmanagement in der Systembiologie

Silke Eckstein

Informationsmanagement in der Systembiologie

Datenbanken, Integration, Modellierung

Springer

Dr. Silke Eckstein
Institut für Informationssysteme
TU Braunschweig
Deutschland
s.eckstein@tu-bs.de

ISBN 978-3-642-18233-4 e-ISBN 978-3-642-18234-1
DOI 10.1007/978-3-642-18234-1
Springer Heidelberg Dordrecht London New York

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Einbandentwurf: deblik, Berlin

Gedruckt auf säurefreiem Papier

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Vorwort

Die Systembiologie wäre in ihrer heutigen Form ohne Beiträge aus der Informatik nicht denkbar. Die Aufgaben, die die Informatik – in Kooperation mit anderen Disziplinen – übernehmen kann, umfassen die Strukturierung, Speicherung und Bereitstellung der verschiedensten Arten von Daten sowie die Entwicklung von Austauschformaten und Integrationsansätzen. Die Entwicklung von Analyse- und Simulationsmethoden und deren effiziente Implementierung gehören ebenso dazu wie die Erstellung von Visualisierungswerkzeugen und Modelldatenbanken.

Das vorliegende Buch bietet eine Informatikperspektive auf die Systembiologie. Der Fokus liegt dabei auf Verwaltung, Austausch und Integration der anfallenden Daten sowie auf der Modellbildung mit unterschiedlichen, insbesondere algorithmischen Methoden. Dabei wird der Leser schrittweise von den Daten über die zur Verfügung stehenden Datenbanken und deren Intergrationsmöglichkeiten hin zu verschiedenen Modellierungsansätzen mit unterschiedlichen Analysemöglichkeiten geführt. Ein Kapitel über die biologischen Grundlagen ermöglicht Nicht-Biologen einen raschen Einstieg in das Thema.

Damit eignet sich das Buch zum einen für Informatiker, die sich in Richtung Systembiologie orientieren wollen, aber auch für (System-)Biologen, die eine Informatikperspektive auf ihr Fach kennen lernen möchten, sowie für Personen anderer Disziplinen, die in dem interdisziplinären Gebiet der Systembiologie arbeiten.

Das Buch wendet sich somit an alle, die einen breiten, fundierten Überblick über das Gebiet gewinnen wollen und gibt zahlreiche Hinweise auf vertiefende Literatur zu den einzelnen Themen. Es ist sowohl zum Selbststudium als auch als Grundlage für eine entsprechende Vorlesung geeignet.

Dieses Buch ist im Rahmen meiner Tätigkeit als Leiterin der Bioinformatikgruppe am Institut für Informationssysteme der TU Braunschweig entstanden. Diese habe ich ab 2002 im Rahmen des Braunschweiger Bioinformatik-Kompetenzzentrums „Intergenomics“ mit aufgebaut. Unsere direkten Kooperationspartner im Rahmen des Kompetenzzentrums waren die Arbeitsgruppen für Mikrobiologie von Prof. Dieter Jahn und für Pflanzengenetik von Prof. Reinhard Hehl der TU Braunschweig, mit denen wir unsere ersten fachübergreifenden Diskussionen führten. Später kamen Kooperationen mit der Biotechnologiefirma Biobase aus Braunschweig-Wolfenbüttel, der Bioinformatikgruppe der TU Braunschweig von

Prof. Dietmar Schomburg sowie der Arbeitsgruppe für zelluläre Proteomforschung des Helmholtzzentrums für Infektionsforschung in Braunschweig von Dr. Lothar Jänsch hinzu.

Allen Kooperationspartnern möchte ich für ihre Diskussionsbereitschaft und ihre Anregungen danken, die direkt oder indirekt zur Entstehung dieses Buchs beigetragen haben. Daher geht mein Dank für fachübergreifende Diskussionen an Lorenz Bülow, Reinhard Hehl, Karsten Hiller, Lothar Jänsch, Dieter Jahn, Thorsten Johl, Frank Klawonn, Mathias Krull, Maren Lang, Richard Münch, Claudia Pommerenke, Susanne Quester, Ida Retter, Alexander Riemer, Maurice Scheer und Thomas Ulas.

Prof. Hans-Dieter Ehrich danke ich ganz herzlich dafür, dass er es mir ermöglicht hat, die Bioinformatikgruppe an seinem Institut aufzubauen, und Prof. Wolf-Tilo Balke dafür, dass er mich darin unterstützt, die Bioinformatikgruppe an seinem Institut weiterzuführen.

Ohne die ehemaligen Mitglieder meiner Gruppe wäre dieses Buchprojekt sicher nicht entstanden. Sie haben sich mit Ideen, Einsatz, Kritik, Unterstützung und Diskussionsbeiträgen in die Gruppe eingebracht und sie durch unterschiedliche Charaktereigenschaften und Ansichten lebendig werden lassen. Ein herzliches Dankeschön geht daher an Andreas Kupfer, Brigitte Mathiak und Claudia Täubner. Unterstützt wurden wir durch vorübergehende „Importe“ aus der Biologie: Britta Reis, Caroline Rio-Bartulos und Sophie von Elsner waren immer zur Stelle, um alle möglichen biologischen Fragen zu beantworten und uns auch sonst auf vielfältige Weise zu unterstützen.

Auch meinen früheren und aktuellen Kolleginnen und Kollegen möchte ich für Unterstützung und Anregungen ganz unterschiedlicher Art danken: Peter Ahlbrecht, Regine Dalkıran, Patrick Hennig, Silviu Homoceanu, Benjamin Köhncke, Maik Kollmann, Andreas Kupfer, Christoph Lofi, Thomas Mack, Brigitte Mathiak, Karl Neumann, Olivera Pavlovic, Ralf Pinger, Joachim Selke, Claudia Täubner, Sarah Tauscher, Sascha Tönnies und Jörg Weimar.

Bei dem Team vom Springer-Verlag – insbesondere bei Dorothea Glaunsinger – bedanke ich mich für die professionelle und engagierte Unterstützung.

Mein ganz spezieller Dank gilt meiner Familie und insbesondere Rainer und Bjarne für ihre Liebe, ihre Geduld mit mir und ihre Unterstützung, ohne die dies Alles nicht möglich gewesen wäre. Danke!

Braunschweig Silke Eckstein
Januar 2011

Inhaltsverzeichnis

Kapitel 1
Einleitung

Wie interagieren die Komponenten einer Zelle, um die Struktur und das spezifische Verhalten dieser Zelle hervorzurufen?

Wie interagieren Zellen, um die Struktur und das Verhalten von Organen und Organismen hervorzurufen?

Diese beiden Fragen nach den intra- und den interzellulären Abläufen sind nach [Wol07] die Kernfragen der Systembiologie. Beantworten lassen sie sich nur durch ein iteratives Vorgehen, bei dem experimentell erzeugte Daten integriert und in Modelle umgesetzt werden, deren Analyse und Simulation das Verständnis dieser zellulären Abläufe erweitern und wiederum Anregungen zu neuen Experimenten geben. Dabei ist eine interdisziplinäre Zusammenarbeit zwischen Biologie, Mathematik und Informatik notwendig, um die benötigten Daten, Methoden und Werkzeuge zur Verfügung zu stellen.

Das vorliegende Buch bietet eine Informatikperspektive auf die Systembiologie mit einem Fokus auf Verwaltung, Austausch und Integration der anfallenden Daten sowie auf die Modellbildung mit informatischen Methoden.

1.1 Systembiologie

Was genau ist mit der Erforschung der intra- und interzellulären Abläufe gemeint? Laut [Kit00] ist es das Hauptanliegen der Systembiologie, biologische Organismen in ihrer Gesamtheit zu verstehen. Es soll ein integriertes Bild aller ablaufenden Prozesse über alle Ebenen, vom Genom über das Proteom bis hin zum Verhalten und zur Biomechanik des Gesamtorganismus, gewonnen werden.

Oder etwas anders ausgedrückt: Die Systembiologie untersucht, wie die Komponenten einer Zelle oder eines Organismus Interaktionsnetzwerke bilden und wie diese Netzwerke die Funktionen der Zelle hervorrufen, die dem beobachtbaren Erscheinungsbild – dem Phänotyp – entsprechen [Pal06].

In [IGH01] wird darüber hinaus betont, dass die zu untersuchenden biologischen Systeme systematisch pertubiert und die Antworten auf Gen-, Protein- und Interaktions-Ebene beobachtet werden. Diese Daten werden integriert und bilden die Basis, um Modelle aufzustellen, die die Struktur des biologischen Systems und seine Antworten auf die verschiedenen Pertubationen beschreiben.

S. Eckstein, *Informationsmanagement in der Systembiologie*,
DOI 10.1007/978-3-642-18234-1_1,

Möglich geworden ist die Bearbeitung solcher Fragestellungen durch immense Fortschritte bei der Sequenzierung von Genomen sowie durch die Entwicklung von Hochdurchsatzexperimenten zum Beispiel zur Gen- und zur Proteinexpressionsanalyse, die darüber Auskunft geben, unter welchen Bedingungen welche Gene aktiv sind bzw. welche Proteine exprimiert werden. Ein Anliegen der Systembiologie ist es also, solche Daten so zu integrieren und in Modelle zu fassen, dass sie die biologischen Systeme möglichst genau beschreiben und dadurch helfen, ihre Funktionsweise zu verstehen.

Um dieses Ziel zu erreichen, ist ein interdisziplinäres Vorgehen nötig. Die Durchführung von Experimenten und Messungen, die Interpretation der Daten und die Aufstellung von Hypothesen ist Aufgabe der beteiligten Biologen. Die Unterstützung bei der Modellierung, die Simulation der Modelle und die Analyse der Ergebnisse ordnet [SJW06] der Systemtheorie zu und die Datenverwaltung, die Visualisierung sowie die Erstellung von Softwarewerkzeugen der Informationstechnik.

In [KHK+05] wird auf eine genaue Festlegung, wer (welche Disziplin) was zu tun hat, verzichtet. Dafür halten die Autoren die Beteiligung aus Biologie, Chemie, Physik, Mathematik, Ingenieurswissenschaften, Regelungstechnik und Informatik für angezeigt.

1.2 Systembiologie aus einer Informatikperspektive

Die Beiträge, die die Informatik – in Kooperation mit den anderen Disziplinen – zur Systembiologie leisten kann, sind vielfältig. Sie umfassen zum Beispiel die Strukturierung, Speicherung und Bereitstellung der verschiedensten Arten von Daten sowie die Entwicklung von Austauschformaten und Integrationsansätzen. Die Entwicklung von Analyse- und Simulationsmethoden und deren effiziente Implementierung gehören ebenso dazu wie die Erstellung von Visualisierungswerkzeugen und Modelldatenbanken.

Mit der fast schon klassischen Bioinformatik, die Hütt und Dehnert als „die Entwicklung und das Betreiben von Datenbanken, Software und mathematischen Werkzeugen zur Analyse, Organisation und Interpretation biologischer Daten“ definieren [HD06], gibt es starke Überschneidungen. In typischen Bioinformatik-Lehrbüchern stehen aber neben den biologischen Datenbanken Algorithmen zum Sequenzvergleich und zur Berechnung von phylogenetischen Stammbäumen sowie zur Strukturvorhersage von Proteinen im Vordergrund. Diese Methoden werden in der Systembiologie selbstverständlich verwendet, aber ihre Entwicklung ist nicht Kernaufgabe der Systembiologie. Hier erhält dagegen etwa die Zusammenführung von Daten aus verschiedenen Datenquellen und die Umsetzung in Modelle ein viel größeres Gewicht.

Bei der Modellbildung kann die Informatik Beiträge leisten, die über das Bereitstellen effizienter Werkzeuge hinausgehen: In [Pri09] prägt Priami den Begriff der „algorithmischen Systembiologie“. Er versteht darunter die Modellierung biologischer Systeme mit algorithmischen Ansätzen und operationaler Semantik im Gegensatz zur klassischen mathematischen Modellierung in Form von Gleichungen und

mit einer denotationalen Semantik. Beide Ansätze unterscheiden sich grundsätzlich und erlauben unterschiedliche Sichtweisen auf die zu modellierenden Systeme.

In Anbetracht der vielfältigen Aufgaben, die die Informatik in der Systembiologie wahrnehmen kann, könnte die Perspektive dieses Buches auch als „datenorientierte Bioinformatik und algorithmische Systembiologie“ bezeichnet werden. Nach einer Einführung in die biologischen Grundlagen im nächsten Kapitel widmen wir uns den verschiedenen Arten von Daten und Datenbanken in diesem Gebiet, der Integration der Daten sowie der Modellierung biologischer Netzwerke. Alle drei Themen wollen wir im Folgenden kurz motivieren.

1.3 Datenbanken

Die Systembiologie wäre nicht möglich ohne eine ausreichend große und breite Datenbasis. Will man die intra- und interzellulären Vorgänge erforschen, so muss man die Bestandteile der Zelle kennen. Damit ist zum einen der prinzipielle Aufbau aus Zellmembran, Zytoplasma, Zellkern, DNA etc. gemeint. Zum anderen müssen aber auch die miteinander interagierenden Moleküle bekannt sein, die die verschiedenen Arten von Interaktionsnetzwerken in der Zelle aufspannen. Man braucht Informationen über die Gensequenz des Organismus, die vorhandenen Gene und deren Funktionsweise. Man muss die Struktur der Proteine kennen, um ihre Reaktionsmöglichkeiten zu bestimmen und vieles mehr.

Für alle diese und noch viele weitere Arten von Informationen wurden molekularbiologische Datenbanken entwickelt, deren Anzahl und Umfang seit Jahren rapide ansteigt.

Um eine Orientierung in diesem Gebiet zu ermöglichen, geben wir einen Überblick über die Datenbanken, der sich an den biologischen Grundlagen orientiert, die wir im Kapitel vorher eingeführt haben. Außerdem stellen wir exemplarisch die wichtigsten Datenbanken jeder Art kurz vor. Dadurch bekommt der Leser einen Eindruck von den verschiedenen Arten molekularbiologischer Datenbanken. Er kann solche, von denen er z. B. in einem Projektkontext zum ersten Mal hört, mit wenigen Nachfragen inhaltlich einordnen und kennt zudem die großen Datenbanken in dem jeweiligen Gebiet, deren Einsatz auch in Betracht gezogen werden könnte.

Zunehmend finden Austauschformate für die verschiedensten Arten von molekularbiologischen Daten Verbreitung. Da einer der Schwerpunkte dieses Buches auf der Modellierung biologischer Netzwerke liegt, konzentrieren wir uns auf Austauschformate in diesem Bereich. Wir stellen mehrere XML-Formate zum Austausch von Interaktionsdaten und Netzwerken vor und vergleichen diese miteinander. Der Leser wird dadurch in die Lage versetzt, zu entscheiden, welches Austauschformat für ein bestimmtes Projekt am besten geeignet ist.

Die in den molekularbiologischen Datenbanken vorhandenen Daten werden oft nicht direkt von ihren Erzeugern an diese Datenbanken übermittelt, sondern in Publikationen veröffentlicht. Diese Publikationen werden dann wiederum von den Datenbankbetreibern auf relevante Daten hin untersucht, die sie in ihre Datenbanken

aufnehmen wollen. Wir geben einen Überblick über die relevanten digitalen Bibliotheken im Gebiet der Systembiologie sowie über Information-Retrieval und Text-Mining-Methoden zur automatischen Textanalyse. Der Leser weiß anschließend, welche digitalen Bibliotheken Veröffentlichungen zu welchen Themen bereitstellen und er bekommt einen Einstieg in die automatische Textanalyse.

1.4 Integration

Damit die ambitionierten Ziele der Systembiologie erreicht werden können, ist eine breite Datenbasis zwar unumgänglich, die verschiedensten Arten von Daten müssen aber zur Erstellung aussagekräftiger Modelle auch zusammengeführt werden. Hier sind die im letzten Abschnitt erwähnten Austauschformate hilfreich und auch notwendig aber nicht ausreichend. Benötigt werden vielmehr Integrationsansätze, die auch die Semantik der Daten berücksichtigen.

Daher stellen wir zunächst die verschiedenen Dimensionen der Informationsintegration vor und betrachten die Ansätze etwas genauer, die bei der Integration biologischer Daten häufig eingesetzt werden. Das Hauptthema ist aber die semantische Integration von Daten mit Hilfe von Ontologien.

Die zur Zeit am weitesten verbreitete Ontologiesprache ist die Web Ontology Language (OWL), die wir ausführlich vorstellen. Als Grundlage dafür führen wir zunächst Ontologien als solche ein. Als formale Basis für viele existierende Ontologiesprachen werden Beschreibungslogiken verwendet. Es handelt sich dabei um entscheidbare Teile der Prädikatenlogik, die je nach unterstützten Konzepten unterschiedlich ausdrucksstark sind und sich entsprechend auch in ihrer Berechnungskomplexität unterscheiden. Neben den Beschreibungslogiken bilden RDF, das Resource Description Framework, und RDF Schema die Grundlagen von OWL, welche wir ebenfalls einführen.

In der Molekularbiologie gibt es einige weit verbreitete Ontologien, von denen wir die in unserem Kontext wichtigsten vorstellen: die Gene Ontology (GO), die Open Biomedical Ontologies (OBO) sowie BioPAX (Biological Pathway eXchange) für den Austausch von Pathwaydaten. Die Grundideen der semantischen Integration von Daten werden anhand eines Ansatzes zur Synchronisation von Datenbanken und Ontologien vorgestellt, der systemübergreifende Anfragen unterstützt.

Der Leser wird in die Lage versetzt, Integrationsansätze nach bestimmten Kriterien klassifizieren zu können. Er kennt die häufigsten zur Integration biologischer Datenbanken eingesetzten Verfahren und kann ihre Vor- und Nachteile abschätzen. Er hat eine ausführliche Einführung in OWL, RDF und RDF Schema bekommen, sodass er nun in der Lage ist, entsprechende Ontologien zu verstehen und selbst zu erstellen. Außerdem hat er deren formale Grundlagen kennen gelernt. Er hat die wichtigsten Informationen zu Gene Ontology, OBO und BioPAX bekommen und anhand eines ganz konkreten Ansatzes gesehen, wie semantische Integration funktionieren kann.

1.5 Modellierung

Die Modellierung, Simulation und Analyse biologischer Netzwerke bildet die Grundlage dafür, die intra- und interzellulären Prozesse zu verstehen und ein integriertes Bild der Abläufe in einem Organismus zu entwickeln. Dabei erlauben es die Gensequenzierung und die funktionale Analyse von Genomen erstmals, biologische Netzwerke in einem großem Stil zu rekonstruieren. Am Beispiel von metabolischen Netzwerken zeigen wir, wie eine solche Rekonstruktion durchgeführt wird. Darauf aufbauend können verschiedenste Arten der Modellierung und der Analyse zum Einsatz kommen:

- Graphentheoretische Ansätze, mit denen die topologischen Eigenschaften der biologischen Netzwerke untersucht werden, mit dem Ziel, daraus Rückschlüsse auf ihre funktionellen Eigenschaften zu ziehen.
- Stöchiometrische Analysen der Netzwerke mit dem Ziel, die wahrscheinlichsten Stoffflüsse zu finden.
- Verschiedenste Arten von mathematischer und algorithmischer Modellierung zur Beantwortung unterschiedlicher Fragestellungen.

Während sich Fachbücher häufig auf bestimmte Modellierungsansätze oder -richtungen konzentrieren, werden hier ganz unterschiedliche Ansätze nebeneinander gestellt. Dies geschieht zum einen, um einen breiten Einstieg in das Gebiet zu geben und den Blick dafür zu öffnen, welche Methoden in einem konkreten Projekt die vielversprechendsten sind. Die andere Motivation ist die, dass es für das Hauptanliegen der Systembiologie, biologische Organismen in ihrer Gesamtheit zu verstehen, unbedingt notwendig ist, diese unter ganz unterschiedlichen Gesichtspunkten zu betrachten und die so gewonnenen Erkenntnisse wiederum zusammen zu bringen.

Da Modelle immer Abstraktionen von der Wirklichkeit sind, lassen sie sich auch nach der Art der Abstraktionen, die sie vornehmen, klassifizieren. Wir diskutieren verschiedene Modellierungsdimensionen und ordnen unterschiedliche Ansätze darin ein.

Nach der breiten Einführung in die Modellierung fokussieren wir auf algorithmische Modellierungsansätze und geben einen Überblick über die verschiedenen Sprachen, die hier zum Einsatz kommen können. Anschließend greifen wir uns einen Ansatz heraus – die Petri-Netze – den wir ausführlich präsentieren. Die Entscheidung ist dabei auf die Petri-Netze gefallen, da sie eine graphische Repräsentation genauso mitbringen wie eine rigorose mathematische Fundierung. Verschiedene Petri-Netz-Erweiterungen greifen unterschiedliche Modellierungsansätze auf und die zugehörigen Werkzeuge stellen vielfältige Analysemöglichkeiten zur Verfügung.

Der Leser erhält einen breiten Überblick über die verschiedenen Modellierungsrichtungen und Analysemöglichkeiten. Außerdem bekommt er die Gelegenheit, einen konkreten Ansatz zu vertiefen, was wiederum die Anwendung anderer Modellierungsansätze erleichtert.

Kapitel 2
Biologische Grundlagen

Ein Organismus wie zum Beispiel der menschliche Körper besteht aus Billionen von Zellen, die jeweils einen Zellkern enthalten. Jeder dieser Zellkerne wiederum enthält einen Chromosomensatz in doppelter Ausführung, der als Genom bezeichnet wird. Das menschliche Genom besteht aus 23 Chromosomenpaaren. Jedes Chromosom ist ein langes DNA-Molekül, das die Form einer Doppelhelix hat und funktionale Regionen enthält, die Gene.

Auf den zwei Chromosomen eines Chromosomenpaars befinden sich dieselben Gene an denselben Stellen, aber meistens mit unterschiedlicher Ausprägung. Jedes Gen kann in verschiedenen Ausprägungen existieren, die auch als Allele bezeichnet werden. Jedes Allel eines bestimmten Gens kodiert für eine andere Version einer bestimmten Eigenschaft, zum Beispiel grüne versus blaue Augenfarbe.

Die Gene wiederum beeinflussen bzw. steuern den gesamten Organismus: Sie kodieren für bestimmte Proteine, d. h., dass durch die Aktivierung bestimmter Gene bestimmte Proteine hergestellt werden (vgl. Abschn. 2.5). Proteine nehmen in jedem Organismus eine zentrale Rolle ein, da sie verschiedenste Funktionen haben: vom Transport über Stoffwechsel und Strukturaufbau bis hin zur Signalweiterleitung.

Proteine interagieren miteinander in Netzwerken, die neben Signalweiterleitung und Stoffwechselfunktionen wiederum auch genregulatorische Aufgaben übernehmen können. Das bedeutet, dass Proteine Gene aktivieren können, was wiederum zur Synthese anderer Proteine führt. Solche Interaktionen von Proteinen sind es, die einen Organismus am Leben erhalten und sein Agieren ermöglichen. Diese grundlegenden Zusammenhänge werden wir in den nachfolgenden Abschnitten ausführlicher erörtern.

Durch die Sequenzierung von Genomen hat die Forschung in diesem Gebiet in den letzten zwei Jahrzehnten erhebliche Fortschritte gemacht. Wir beginnen daher mit einem Blick auf das Humangenomprojekt.

2.1 Das Humangenomprojekt

Das Humangenomprojekt startete 1990 in den USA als öffentlich finanziertes Projekt mit dem Ziel, bis 2010 das menschliche Genom sequenziert zu haben. 1995 schloss sich Deutschland diesem Ziel mit dem deutschen Humangenomprojekt

S. Eckstein, *Informationsmanagement in der Systembiologie*,
DOI 10.1007/978-3-642-18234-1_2,

(DHGP) an. Insgesamt arbeiteten mehr als 1.000 Wissenschaftler in über 40 Ländern im Rahmen dieses Projekts zusammen. Konkurrenz bekam das Projekt 1998 durch die private Firma Celera von Graig Venter, die mit 100 Wissenschaftlern und 50 Technikern an den Start ging.

2001 verkündeten beide Gruppen die Sequenzierung des menschlichen Erbguts in einer Draft-Version, das öffentliche Projekt publizierte in Nature [Int01] und das private in Science [VAM+01]. Der Begriff Draft-Version ist in diesem Zusammenhang so zu verstehen, dass es beiden Gruppen gelang, den euchromatischen Anteil des menschlichen Genoms nahezu vollständig zu sequenzieren, also die Bereiche auf der DNA, die genetisch aktiv sind, d. h. aus denen Proteine exprimiert werden können. Im Oktober 2004 veröffentlichte das öffentlich finanzierte Projekt eine vorläufig endgültige Sequenz [Int04].

Wie muss man sich die Sequenzierung eines Genoms vorstellen? Das menschliche Genom besteht aus 23 Chromosomenpaaren, die die Erbinformationen enthalten. Das sind 46 lange DNA-Moleküle, die sich im Zellkern befinden. Jedes dieser Chromosomen besteht aus einem DNA-Doppelstrang, der sich wiederum aus Abfolgen von 4 verschiedenen Nukleotiden zusammensetzt. Diese 4 Nukleotide werden mit den Buchstaben A,C,G und T bezeichnet. Das menschliche Genom lässt sich daher als eine lange Zeichenkette über dem Alphabet A, C, G, T beschreiben. Ziel der Sequenzierung ist es somit, diese Zeichenkette zu entziffern. Die Sequenziermaschinen sind heutzutage aber noch nicht in der Lage, die langen DNA-Moleküle in einem Schritt zu lesen. Das Vorgehen ist daher wie folgt:

1. Zunächst wird die DNA repliziert, um mehrere identische Kopien zu erhalten.
2. Anschließend werden diese Kopien in Stücke zerlegt, was zum Beispiel mit Ultraschall geschehen kann. Dabei muss dafür gesorgt werden, dass jede der Kopien in unterschiedliche Stücke zerlegt wird.
3. Diese Einzelstücke können von den Sequenziermaschinen verarbeitet werden, das heißt ihre Buchstabenfolge kann abgelesen werden.
4. Durch die überlappenden Teilstücke der DNA-Kopien kann auf die Gesamtabfolge zurückgeschlossen werden. Dazu werden Einzelstücke gesucht, deren Anfangsstück mit dem Endstück eines anderen Einzelstücks überlappt.
5. Diese überlappenden Einzelstücke werden dann in der richtigen Reihenfolge zusammengesetzt. Diesen Vorgang nennt man Assemblierung. Dabei muss eine gute Fehlerbehandlung erfolgen, da mit mehrfach vorkommenden Teilsequenzen und Lesefehlern umgegangen werden muss. Die Assemblierungsalgorithmen müssen sehr effizient sein, da die Anzahl der Einzelstücke und somit auch die Anzahl der durchzuführenden paarweisen Vergleiche auf mögliche Überlappungen sehr hoch ist.

In den Medien wurde in diesem Zusammenhang häufig von der Entschlüsselung des menschlichen Erbguts gesprochen. Dies ist nicht korrekt, da das Wort Entschlüsselung impliziert, dass die Funktionsweise des Erbguts bekannt sei. Das ist aber nicht oder nur in kleinen Teilbereichen der Fall. Was seit 2001 bekannt ist, ist die Sequenz des menschlichen Erbguts, also die Abfolge der Basen auf der DNA. An der tatsächlichen Entschlüsselung wird man noch Jahre oder Jahrzehnte arbeiten.

2.2 Zellen und Organismen

Alle lebenden Organismen, so unterschiedlich sie auch sein mögen, bestehen aus Zellen, kleinen Kompartimenten, die im Großen und Ganzen dieselben Bestandteile besitzen. Man geht sogar davon aus, dass es vor über 3 Milliarden Jahren eine Urzelle gab, von der alle heutigen Zellen abstammen. Folgende Eigenschaften sind allen Zellen gemein [ABH+05]:

- Sie wachsen und vermehren sich, sie wandeln verschiedene Energieformen ineinander um, sie nehmen ihre Umgebung wahr und reagieren auf Reize.
- Das Innere einer Zelle wird durch eine Plasmamembran von der Umgebung abgetrennt.
- In allen Zellen sind Erbinformationen in Form von DNA enthalten. Mit Hilfe der Gene auf der DNA können die Zellen Proteine synthetisieren.
- Obwohl alle Zellen in vielzelligen Organismen die gleiche DNA besitzen, können sie sich sehr unterschiedlich verhalten. Das bedeutet, dass sie ihre biochemischen Aktivitäten entsprechend der Reize steuern, die sie aus ihrer Umgebung empfangen.
- 96,5% der Masse von lebenden Zellen bestehen aus den vier Elementen Kohlenstoff (C), Wasserstoff (H), Stickstoff (N) und Sauerstoff (O).

Es gibt zwei grundsätzlich unterschiedliche Arten von Zellen: Als Eukaryoten werden solche bezeichnet, die einen Zellkern und eine Zellmembran besitzen, und als Prokaryoten diejenigen ohne Zellkern. Die Bezeichnung Eukaryot kommt aus dem Griechischen und setzt sich aus den zwei Bestandteilen *eu*, echt und *karyos*, Kern zusammen: mit einem echten Kern ausgestattet. Zu den Eukaryoten gehören beispielsweise alle Tiere und Pflanzen sowie Hefearten. Die DNA der Eukaryoten befindet sich, verteilt auf mehrere Chromosomen, im Zellkern. Die restlichen Bestandteile der Zelle – außer dem Zellkern – bilden das Cytoplasma.

Die DNA der Prokaryoten ist nicht auf mehrere Chromosomen aufgeteilt und wird nicht durch einen speziellen Zellkern vom Rest der Zelle abgetrennnt. Zu den Prokaryoten gehören z. B. alle Bakterien. Alle Prokaryoten sind Einzeller aber nicht alle Einzeller sind auch Prokaryoten. Backhefe (baker's yeast, *Saccharomyces cerevisiae*) ist ein Beispiel für einen eukaryotischen Einzeller.

Zellen bestehen aus Molekülen, die – abgesehen vom Wasser – fast alle organischer Natur sind. Als organische Moleküle werden solche bezeichnet, die aus Kohlenstoffverbindungen bestehen, alle anderen werden anorganisch genannt. Darüberhinaus lassen sich die in Zellen vorkommenden Moleküle in vier grundlegende Gruppen einteilen: kleine Moleküle („small molecules"), Proteine, DNA und RNA. Die letzteren drei werden auch als Makromoleküle bezeichnet und in den nächsten Abschnitten ausführlicher besprochen.

Wasser ist ein Beispiel für ein anorganisches kleines Molekül. Organische kleine Moleküle zeichnen sich dadurch aus, dass sie aus bis zu 30 Kohlenstoffverbindungen bestehen. Beispiele sind Fettsäuren, aus denen die Zellmembranen aufgebaut sind, Zucker, Nukleotide und Aminosäuren. Nukleotide bestehen, wie wir weiter

unten noch genauer sehen werden, aus jeweils einem Zucker, einem Phospahtrest und einer Base. Sie sind die Grundbausteine für DNA- und RNA-Moleküle. Aminosäuren werden zu Proteinen zusammengesetzt.

2.3 Genome, Chromosomen und DNA

Die Gesamtheit der genetischen Information eines Organismus wird als Genom bezeichnet und ist in DNA-Molekülen gespeichert. Eukaryoten besitzen mehrere DNA-Moleküle, die in Chromosomen strukturiert sind. Das menschliche Genom besteht aus 23 Chromosomenpaaren mit etwa 3 Milliarden Basenpaaren und ca. 25.000 Genen, wobei allerdings über die genaue Anzahl der Gene noch Unsicherheit besteht. Damit besitzt der Mensch nur doppelt so viele Gene wie eine Fliege und fünfmal so viele wie das Bakterium *E. coli*. Eine lebensfähige Zelle benötigt vermutlich weniger als 400 Gene.

Die Abkürzung DNA steht für *deoxyribonucleic acid* (oder auf deutsch *Desoxyribonukleinsäure*) und DNA-Moleküle bestehen aus zwei langen Molekülketten, die zu einer Doppelhelix zusammengesetzt sind. Das heißt die Molekülketten winden sich schraubenförmig um eine gemeinsame, fiktive Achse. Dabei verlaufen die Stränge in entgegengesetzter Richtung, sind also antiparallel. Die Doppelhelix verläuft von oben her betrachtet im Uhrzeigersinn, ist also rechtsherum gedreht. Diese dreidimensionale Struktur der DNA (vgl. Abb. 2.1) entdeckten James D. Watson und Francis H.C. Crick 1953 [WC53a, WC53b].

Am spezifischen Aufbau der DNA sind zwei Arten chemischer Bindungen beteiligt, kovalente Bindungen und Wasserstoffbrückenbindungen.

2.3.1 Chemische Bindungen

Es gibt unterschiedlich starke chemische Bindungen, also Bindungen zwischen den kleinsten Teilchen in chemischen Stoffen (z. B. zwischen Atomen, Anionen, Kationen oder Molekülen). Starke Bindungen sind z. B. kovalente Bindungen, die auch als Atom- oder Elektronenpaarbindungen bezeichnet werden. Charakteristisch für diese Bindungen ist, dass sich zwei Atome die Elektronen in ihrer äußeren Schale teilen (vgl. Abb. 2.2, linke Seite), wodurch sie ihre äußeren Schalen auffüllen und somit eine stabilere Elektronenanordnung erreichen. Kovalente Bindungen kommen in Zellen nur mit Hilfe von Enzymen als Katalysatoren zustande und können auch nur mit Hilfe von Enzymen wieder gelöst werden [ABH+05]. Sie sorgen somit für den festen Zusammenhalt von Atomen in Verbindungen. Beispiele für solche Verbindungen sind Moleküle.

Insbesondere Kohlenstoff (C) ist aufgrund seines Aufbaus dazu in der Lage, mit Hilfe kovalenter Bindungen große Moleküle zu bilden. Das liegt daran, dass die recht kleinen Kohlenstoffatome vier Elektronen und vier freie Plätze in ihrer äußeren Schale besitzen und somit vier kovalente Bindungen zu anderen Atomen

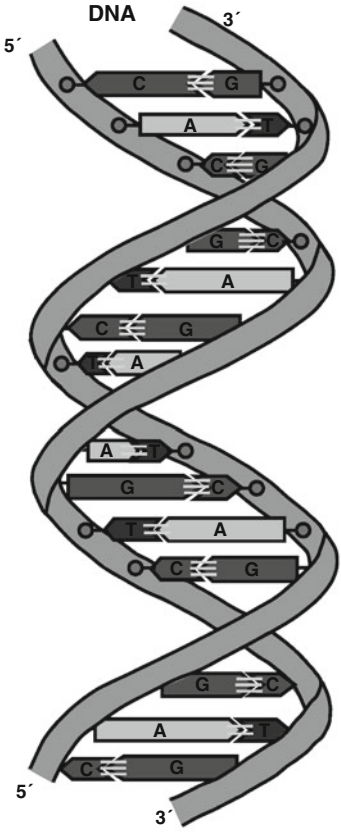

Abb. 2.1 Doppelhelixstruktur der DNA

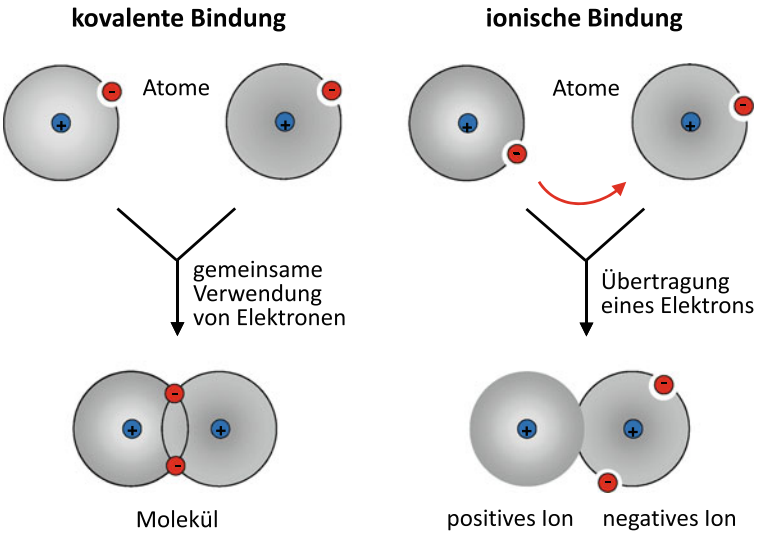

Abb. 2.2 Atom- und Ionenbindungen (angelehnt an [ABH+05])

eingehen können. Besonders stabil sind dabei kovalente C-C-Bindungen, die die Form von Ketten, von verzweigten Strukturen oder auch von Ringen annehmen können [ABH+05].

Ionenbindungen, die dadurch entstehen, dass ein Atom Elektronen an ein anderes Atom abgibt (vgl. Abb. 2.2, rechte Seite), sind ein Beispiel für nichtkovalente Bindungen. Ein anderes Beispiel sind Wasserstoffbrückenbindungen, welche elektrostatischer Natur sind. Sie kommen dadurch zustande, dass polare kovalente Bindungen existieren. Das bedeutet, dass bei einer kovalenten Bindung ein Atom (aufgrund der relativen Größe seines positiven Atomkerns) die Elektronen etwas stärker zu sich herüberzieht als das andere Atom. Dadurch entstehen unterschiedliche elektrische Ladungen an den verschiedenen Enden eines Moleküls. Dies führt wiederum dazu, dass zwischen dem positiv geladenen Ende eines und dem negativ geladenen Ende eines anderen Moleküls Anziehungskräfte entstehen. Die darauf basierenden Bindungen nennt man Wasserstoffbrückenbindungen, da typischerweise ein Wasserstoffatom das positive Ende eines Moleküls bildet und mit Sauerstoff oder Stickstoff eine solche Bindung eingeht. Die Bindungen sind sehr schwach und können z. B. durch Erhitzen gelöst werden.

Ebenfalls sehr schwache Bindungen entstehen durch Van-der-Waals-Anziehungen. Es handelt sich dabei um elektrostatische Wechselwirkungen, welche durch flukturierende elektrische Ladungen zwischen Atomen entstehen, wenn diese sich lange genug nahe kommen.

2.3.2 Aufbau und Funktion der DNA

Die beiden zu einer Doppelhelix zusammengesetzten DNA-Stränge sind Polynukleotidketten, die über Wasserstoffbrückenbindungen miteinander verbunden sind. Dabei sind Polynukleotide lineare, aperiodische, aus Nukleotiden zusammengesetzte chemische Verbindungen, die auch Polymere genannt werden. Die Bezeichnung Polymer steht für „aus vielen gleichen Teilen bestehend“. Die „vielen gleichen Teile“ sind in der DNA vier verschiedene Nukleotide, die jeweils aus dem Zucker Desoxyribose, einer Phosphatgruppe und einer stickstoffhaltigen Base bestehen. Die in der DNA vorkommenden Basen sind Adenin (A), Cytosin (C), Guanin (G) und Thymin (T), deren abkürzende Bezeichnungen den typischen 4-Buchstaben-Code ergeben, mit dem sich DNA-Moleküle charakterisieren lassen.

Die Basen kodieren also die genetische Information.

Chemisch gesehen sind Zucker Moleküle mit der Summenformel $C_n H_{2n} O_n$, weshalb sie oft auch als Kohlenhydrate bezeichnet werden. In unserem Kontext sind die beiden Zucker Ribose und Desoxyribose interessant, wobei letzterer eine Ausnahme von der obigen Summenformel darstellt (vgl. Abb. 2.3). Der Zucker Desoxyribose ist auch der Namensgeber für die DNA (Desoxyribonukleinsäure).

Nukleotide entstehen also durch kovalente Bindungen zwischen dem Zucker Desoxyribose, einer Phosphatgruppe und einer Base, wie es in Abb. 2.4 zu sehen ist. Dabei wird bei der Bindung des Zuckers mit dem Phosphat das Wasserstoffatom abgespalten und bei der Bindung des Zuckers mit der Base die OH-Gruppe. Die

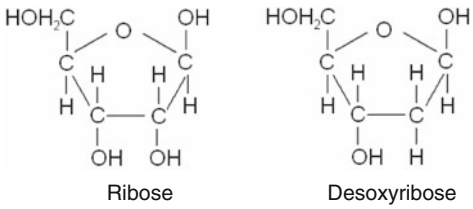

Abb. 2.3 Strukturformeln für die Zucker Ribose und Desoxyribose

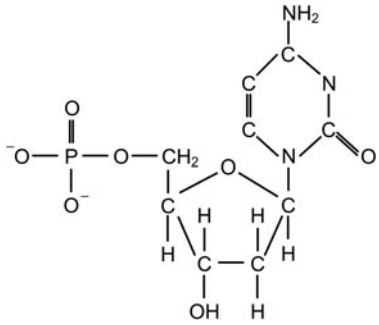

Abb. 2.4 Nukleotid

Basen Thymin und Cytosin bestehen aus einem Ring und werden auch als Pyrimidine bezeichnet. Adenin und Guanin dagegen bestehen aus zwei Ringen und werden als Purine bezeichnet (vgl. auch Abb. 2.5).

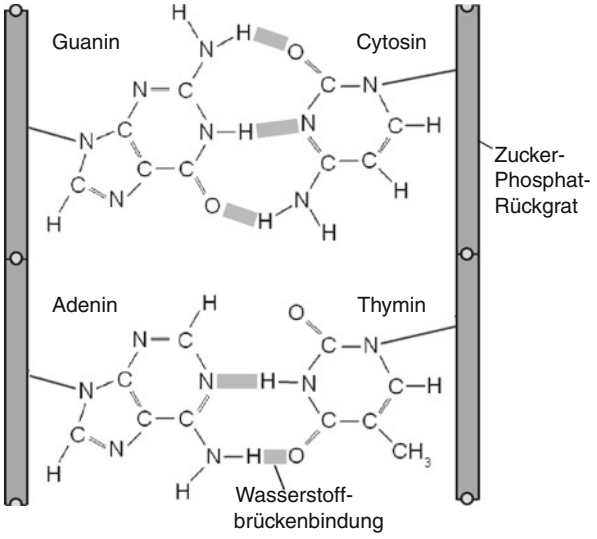

Abb. 2.5 Wasserstoffbrückenbindungen zwischen den Basen A und T bzw. C und G (angelehnt an [ABH+05])

Nukleotide werden durch Bindungen zwischen dem 3′ und dem 5′-Kohlenstoffatom zu Nukleinsäuren verknüpft. Die DNA-Moleküle haben daher ein 5′- und ein 3′-Ende und werden in dieser Richtung gelesen.

Wir hatten ja bereits gesehen, dass DNA-Moleküle als Doppelhelices aufgebaut sind. Die beiden Stränge enthalten komplementäre Informationen. Das heißt, dass, wenn auf der einen Seite eine bestimmte Base auftritt, auf der anderen Seite immer eine bestimmte andere Base vorhanden ist. Die Basenpaare, die sich dabei zusammenfinden, sind aufgrund ihrer chemischen und räumlichen Struktur A und T sowie C und G. Bei den dabei auftretenden Bindungen handelt es sich um Wasserstoffbrückenbindungen, die im Gegensatz zu kovalenten oder Ionenbindungen eher schwach sind. In Abb. 2.5 sind die Wasserstoffbrückenbindungen zwischen A und T bzw. C und G dargestellt.

Dadurch dass die Doppelhelixstruktur der DNA aus komplementären Basensträngen besteht, reicht also zur Beschreibung eines DNA-Moleküls die Angabe eines der beiden Basenstränge aus. Der Aufbau der DNA als Doppelhelix mit komplementären Basensträngen sorgt dafür, dass sich die genetische Information vererben lässt: Bei der Zellteilung muss jede Kindzelle eine Kopie des elterlichen Genoms erhalten, d. h. die DNA muss repliziert werden. Dies geschieht dadurch, dass die Doppelhelix in 2 einzelne Stränge aufgespalten wird, für die dann jeweils neue komplementäre Stränge gebildet werden.

Die Funktionen der DNA sind das Speichern, Abrufen und Übersetzen von genetischen Anweisungen zur Erzeugung und zum „Betrieb“ eines Organismus. Dabei enthalten nur bestimmte Abschnitte auf der DNA genetische Informationen – die Gene. Sie sind für die Proteinsynthese zuständig. Das heißt, die Gene enthalten die Information, wie bestimmte Proteine hergestellt werden.

2.4 Proteine

In diesem Abschnitt betrachten wir, wie Proteine aufgebaut sind und welche Funktionen sie haben, bevor wir uns im nächsten Abschnitt der Genexpression, also der Erstellung von Proteinen aus der DNA zuwenden.

Proteine – oder umgangssprachlich Eiweiße – haben vielfältige Aufgaben innerhalb von Zellen, die von Signalweiterleitungen über enzymatische Funktionen bis zur Erzeugung von Bewegung reichen. Entsprechend groß ist die Anzahl an unterschiedlichen Proteinen. Sie werden auch als Bausteine der Zelle bezeichnet und machen den größten Teil ihres Trockengewichts aus [ABH+05]. Chemisch gesehen sind Proteine Makromoleküle, die aus Aminosäuren bestehen, welche mit Peptidbindungen zu langen unverzweigten Ketten verknüpft sind. Die Proteingröße variiert von unter 100 bis hin zu 3.000 Aminosäuren. Durch die Aminosäuresequenz wird sowohl die dreidimensionale Struktur als auch die Funktion des Proteins bestimmt. Wir betrachten daher im Folgenden den Aufbau von Aminosäuren.

2.4.1 Aminosäuren

Aminosäuren bestehen aus einem zentralen Kohlenstoffatom (C), einem Wasserstoffatom (H), einer Carboxylgruppe (einer Säure, COOH), einer Aminogruppe (H_2N) und einem Rest, der auch als Seitengruppe bezeichnet wird (vgl. Abb. 2.6). Es gibt 20 verschiedene Aminosäuren, die sich jeweils durch ihre Seitengruppe unterscheiden. Einen Überblick gibt Tabelle 2.1. Dort sind die Aminosäuren zusammen mit ihren gebräuchlichsten Abkürzungen dargestellt, dem Drei-Buchstaben-Code (Three-Letter-Code, 3LC) und dem Ein-Buchstaben-Code (One-Letter-Code, 1LC).

In Tabelle 2.1 ist eine 21. Aminosäure – Selenocystein – aufgeführt, die aber selbst nicht direkt in der DNA kodiert wird. Es wird immer zunächst Cystein gebildet, das dann durch weitere Prozesse in Selenocystein umgewandelt wird. Des Weiteren kommt diese Aminosäure extrem selten vor, sodass man im Allgemeinen von 20 in Proteinen vorkommenden Aminosäuren spricht.

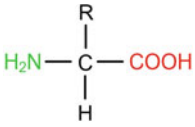

Abb. 2.6 Aminosäure

Tabelle 2.1 Die Aminosäuren mit ihren Abkürzungen

Aminosäure	3LC	1 LC
Alanin	Ala	A
Arginin	Arg	R
Asparagin	Asn	N
Asparaginsäure	Asp	D
Cystein	Cys	C
Glutamin	Gln	Q
Glutaminsäure	Glu	E
Glyzin	Gly	G
Histidin	His	H
Isoleuzin	Ile	I
Leuzin	Leu	L
Lysin	Lys	K
Methionin	Met	M
Phenylalanin	Phe	F
Prolin	Pro	P
Serin	Ser	S
Threonin	Thr	T
Tryptophan	Trp	W
Tyrosin	Tyr	Y
Valin	Val	V
Selenocystein	Sec	U
Asparginsäure oder Aspargin	Asx	B
Glutaminsäure oder Glutamin	Glx	Z
Beliebige Aminosäure	Xaa	X

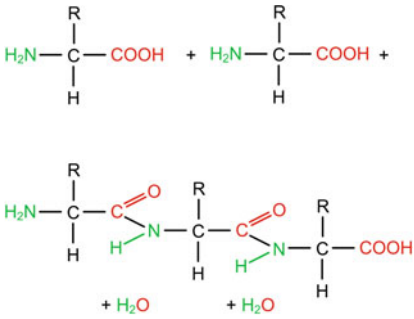

Abb. 2.7 Kurze Aminosäurenkette

Proteine sind Makromoleküle, die aus Aminosäuren bestehen, welche mit Peptidbindungen zu langen unverzweigten Ketten verknüpft sind. Kurze Aminosäureketten von weniger als 100 Aminosäuren werden Peptide genannt, Proteine auch Polypeptide, daher auch die Bezeichnung „Peptidbindung“ für die Bindung zwischen den Anminosäuren. Ein anderer häufig verwendeter Name für diese kovalente Bindung ist Amidbindung. Dabei bindet jeweils die Carboxylgruppe der einen Aminosäure unter Abspaltung von Wasser an die Aminogruppe der nächsten. Es handelt sich also um eine Kondensationsreaktion. Abbildung 2.7 zeigt eine kurze Kette von Aminosäuren. Auch Proteine haben eine Lesereihenfolge: Und zwar vom Amino- oder N-Terminus, das ist die freie Amino-Gruppe auf der linken Seite, zum Carboxyl- oder C-Terminus, der freien Carboxyl-Gruppe auf der rechten Seite. Typischerweise wird zur Darstellung von Peptiden oder Proteinen nicht die Strukturformel angegeben sondern ihre Aminosäuresequenz im Ein- oder Drei-Buchstaben-Code.

2.4.2 Struktur von Proteinen

Die Struktur von Proteinen wird auf vier verschiedenen Ebenen betrachtet:

- Als **Primärstruktur** wird die Aminosäuresequenz bezeichnet (vgl. Abschn. 2.4.1), also eine Zeichenreihe über einem 20-elementigen Alphabet.
- Die **Sekundärstruktur** beschreibt Regelmäßigkeiten in der lokalen Struktur, z. B. α-Helices und β-Faltblätter.
- Als **Tertiärstruktur** wird die 3-D-Struktur von Proteinen bezeichnet.
- Und eine **Quartärstruktur** besitzen Proteine, die aus mehreren Polypeptidketten bestehen.

Beispiele für die räumliche Struktur von Proteinen auf allen vier Ebenen sind in Abb. 2.8 zu sehen. Die Sekundärstruktur von Proteinen wird durch nichtkovalente Bindungen – Ionen-, Wasserstoffbrücken- und Van-der-Waals-Bindungen – gebildet, die zwischen der CO-Gruppe (Carboxyl-Gruppe) einer Peptidbindung und der NH-Gruppe (Amino-Gruppe) einer anderen Peptidbindung entstehen. Zur Bildung

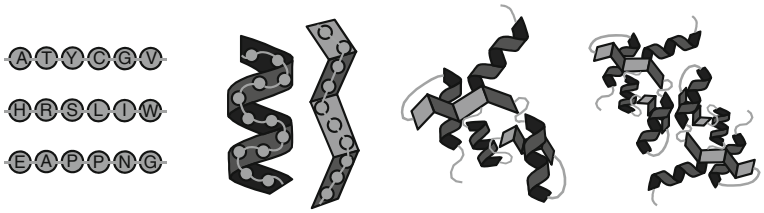

Abb. 2.8 Primär-, Sekundär-, Tertiär- und Quartärstruktur von Proteinen

der Tertiärstruktur tragen wiederum nichtkovalente Bindungen bei sowie spezifische räumliche Anordnungen. Wir wollen hier nicht weiter auf die Proteinstrukturen eingehen, diese sind allein schon ein umfangreiches Forschungsgebiet. Wichtig in unserem Kontext ist lediglich zu wissen, dass zum einen die dreidimensionale Struktur der Proteine für ihre Funktion ausschlaggebend ist und dass diese Struktur allein durch die Aminosäuresequenz festgelegt wird.

2.4.3 Arten und Funktionen von Proteinen

Es wurde bereits angedeutet, dass Proteine viele verschiedene Aufgaben in Zellen wahrnehmen. Die wichtigsten werden im Folgenden besprochen (Klassifikation und Beispiele nach [ABH+05]): **Enzyme** katalysieren Auf- oder Abbau kovalenter Bindungen. In Zellen sind tausende unterschiedlicher Enzyme enthalten, die jeweils als Katalysator für eine bestimmte Reaktion dienen. Beispiele sind das Enzym *Pepsin*, das im Magen Proteine aus der Nahrung abbaut, *Ribulosebisphosphat-Carboxylase*, das in Pflanzen bei der Umwandlung von Kohlendioxid zu Glucose beteiligt ist und die Gruppe der *Proteinkinasen*, die Phosphatgruppen in Proteinmoleküle einbauen.

Strukturproteine stützen Zellen und Gewebe mechanisch. *α-Keratin* etwa ist der Hauptbestandteil von Haar und Horn. Kleine Moleküle oder Ionen werden von **Transportproteinen** transportiert. Beispielsweise transportieren *Hämoglobin* Sauerstoff und *Transferrin* Eisen im Blutstrom. In Zellmembranen eingebettet sind viele Proteine, die kleine Moleküle oder Ionen durch die Membran transportieren. **Motorproteine** sind für die Bewegung in Zellen oder Geweben zuständig. In den Skelettmuskelzellen des Menschen liefert *Myosin* die Antriebskraft für Bewegungen.

Speicherproteine speichern kleine Moleküle oder Ionen. Zum Beispiel wird Eisen in der Leber an das kleine Protein *Ferritin* gebunden und dadurch gespeichert. Im Kontext dieses Buches sehr wichtig sind **Signalproteine**, die Signale von Zelle zu Zelle und innerhalb von Zellen übertragen. Ein Beispiel ist das kleine Protein *Insulin*, das den Glucosespiegel im Blut kontrolliert, ein anderes ist der *Epidermiswachstumsfaktor (EGF)*, der das Wachstum und die Teilung von Epithelzellen stimuliert. Epithel ist eine der vier Grundgewebearten; die anderen sind Binde-, Muskel- und Nervengewebe.

Von **Rezeptorproteinen** werden Signale erkannt und ins Innere der Zelle weitergeleitet. Von Nervenenden ausgesendete chemische Signale werden in der Membran

von Muskelzellen durch *Acetylcholinrezeptoren* empfangen. Eine Leberzelle erhält über den *Insulinrezeptor* das Signal, mit Glucoseaufnahme auf das Hormon Insulin zu reagieren. **Genregulatorproteine** binden DNA, um Gene an- oder abzuschalten. So stellt zum Beispiel der *Lactoserepressor* in Bakterien die Gene für die Synthese von Enzymen zum Lactoseabbau (Milchzuckerabbau) ruhig. Es gibt viele verschiedene Proteine, die als genetische Schalter wirken, um die Entwicklung in vielzelligen Organismen zu kontrollieren. Daneben gibt es noch jede Menge Proteine mit speziellen Aufgaben, die sich nur schwer in Klassen zusammenfassen lassen.

Proteine haben in Organismen also vielfältige Funktionen und sie reagieren miteinander auf ebenso vielfältige Art und Weise. Diese Interaktionen von Proteinen werden mit Hilfe von sogenannten Pathways oder Netzwerken beschrieben, die häufig auch graphisch in semi-formaler Weise repräsentiert werden. Man unterscheidet zwischen verschiedenen Arten von Pathways und Netzwerken (vgl. Abschn. 2.7). Metabolische Pathways beschreiben Stoffwechselvorgänge, also die Umsetzung von Stoffklassen in andere. Regulatorische Pathways hingegen regeln die Antworten von Zellen auf externe Stimuli, indem sie zum Beispiel die Synthese oder den Abbau anderer Moleküle bewirken. Hier steht nicht die Stoffumwandlung im Vordergrund sondern die Weiterleitung von Signalen. Man spricht daher auch von Signaltransduktionswegen.

Ein solcher Signaltransduktionsweg ist der TLR4-Pathway. Um einen ersten Eindruck von Pathway-Darstellungen zu vermitteln, ist er in Abb. 2.9 in der Form dargestellt, wie man ihn in der TRANSPATH-Datenbank findet [KPV+06]. Man sieht

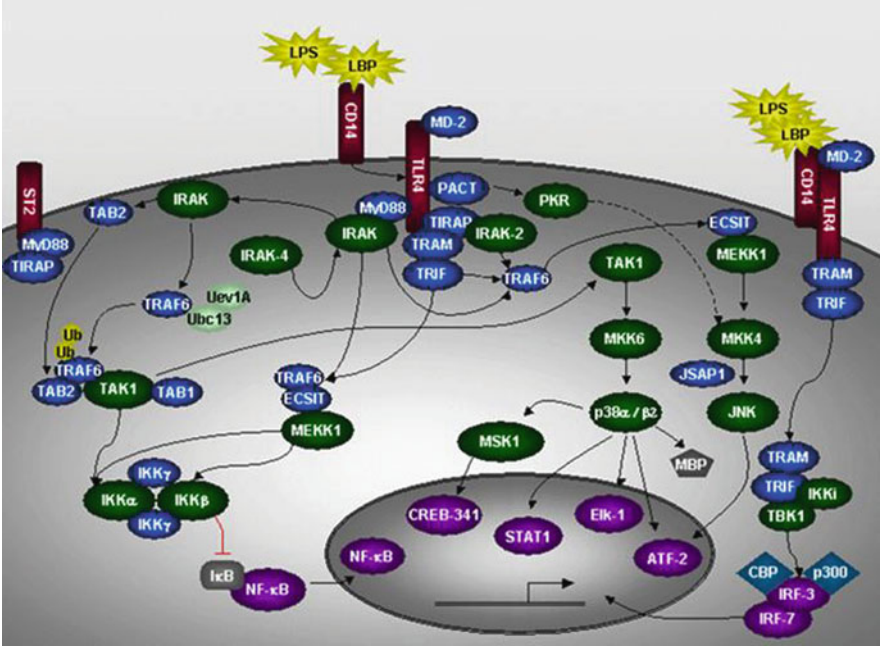

Abb. 2.9 Der TLR4-Pathway aus der TRANSPATH-Datenbank

unter anderem rechteckig dargestellte Rezeptoren, an die Signalmoleküle binden können, verschiedene andere Moleküle und Molekülkomplexe sowie Interaktionen zwischen diesen.

Bevor wir uns aber ausführlicher mit den verschiedenen Arten biologischer Netzwerke beschäftigen, schauen wir uns zunächst einmal den Zusammenhang zwischen Genen und Proteinen an, die Genexpression.

2.5 Genexpression

Gene sind bestimmte Teilbereiche von Genomen, die Proteine kodieren. Proteine werden, wie wir in Abschn. 2.4 gesehen haben, durch ihre Aminosäuresequenz charakterisiert. Diese wird in Genen durch Basen-Triplets – sogenannte Codons – kodiert. Das bedeutet, dass jede Aminosäure durch drei Basen beschrieben wird. Diesen Code nennt man auch den genetischen Code.

Allerdings wird die Information auf der DNA nicht direkt in Proteine umgesetzt sondern das entsprechende Segment der DNA wird zunächst in ein ähnliches Molekül abgebildet, die RNA. Dieser Schritt, der tatsächlich selbst noch aus diversen Einzelschritten besteht, wird auch als Transkription bezeichnet. Anschließend werden aus der RNA Proteine gebildet. Man spricht hier von Translation. Die Erstellung von Proteinen aufgrund der in der DNA kodierten Information wird Genexpression genannt, die also aus den Schritten Transkription und Translation besteht. Dieser Zusammenhang zwischen DNA, RNA und Proteinen ist die Basis der Molekularbiologie und wird oft auch als „zentrales Dogma“ bezeichnet.

Zwischen DNA und RNA bestehen sowohl chemische als auch strukturelle Unterschiede. Die Abkürzung RNA steht für Ribonukleinsäure, was darauf hindeutet, dass ihre Nukleotide mit einem anderen Zucker, nämlich Ribose, aufgebaut sind. Die Strukturformel dieses Zuckers wurde schon weiter oben in Abb. 2.3 in Abschn. 2.3 gezeigt. Der zweite chemische Unterschied besteht darin, dass anstelle der Base Thymin (T) in der RNA Uracil (U) zum Einsatz kommt.

Strukturell unterscheidet sich RNA von DNA zum einen dadurch, dass RNA einzelsträngig vorliegt statt wie DNA eine doppelsträngige Helix auszubilden. Des weiteren kann sich RNA intramolekular falten, dass heißt sie bildet ähnlich wie Proteine Sekundärstrukturen aus (vgl. Abb. 2.10).

Ganz abstrakt kann man Transkription und Translation als Abbildungen von einem Alphabet in ein anderes betrachten. Diese Sicht wird in Abb. 2.11 dargestellt. Bei der Transkription werden alle Ts in Us umgewandelt. Bei der Translation werden Basentripletts oder Codons in Aminosäuren umgesetzt. Da die RNA ein lineares Polymer aus 4 verschiedenen Nukleotiden ist, gibt es $4*4*4 = 64$ mögliche Kombinationen aus drei Buchstaben. Da nur 20 verschiedene Aminosäuren kodiert werden und auch alle Tripletts tatsächlich vorkommen, ist der Code redundant, d. h. die meisten Aminosäuren werden durch mehrere verschiedene Tripletts kodiert. Tabelle 2.2 zeigt den genetischen Code. Dabei befindet sich die erste Base eines Codons in der Spalte ganz links, die zweite Base in der obersten Zeile und die dritte Base in der Spalte ganz rechts.