



# Molekulare Bioinformatik

Ralf Hofestädt, Universität Bielefeld,  
Hans-Peter Lenhof, Universität des Saarlandes

Die Sequenzierung des menschlichen Genoms, die von den Medien als die Mondlandung der Biologie gefeiert wurde, war das herausragende wissenschaftliche Ereignis der letzten Jahre. Der sequenzierte genetische Bauplan des Menschen ist heute die Quelle für Visionen und Hoffnungen im Bereich der Medizin und Pharmazie. Diese visionären Gedanken repräsentieren die wesentliche Motivation für den massiven Aufbau, den die Bioinformatik in Deutschland in extrem kurzer Zeit vollzogen hat. In den vergangenen 60 Jahren hat sich nicht nur die elektronische Datenverarbeitung in einem atemberaubenden Tempo entwickelt. Zeitlich parallel dazu hat sich die *Molekulare Biologie* in den Mittelpunkt der aktuellen wissenschaftlichen Aktivitäten und Förderprogramme katapultiert. In den 40er-Jahren konnte die Desoxyribonukleinsäure (DNS) als genetischer Informationsträger identifiziert werden. In den 50er-Jahren wurde der Aufbau und die Struktur (Doppelhelix) der DNS durch James Watson und Francis Crick aufgeklärt. In den 60er-Jahren gelang die Entzifferung des genetischen Codes, der für die Übersetzung der genetischen Information (Triplett = lineare Abfolge von drei Basen {Adenin, Guanin, Thymin und Cytosin}) in die Sprache der Peptide (Aminosäuren) zuständig ist. Die grundlegenden Methoden der DNS-Sequenzierung wurden in den 70er-Jahren entwickelt. Diese Techniken konnten in den 80er-Jahren zu einer tragfähigen Technologie ausgebaut werden. In den 90er-Jahren stand

die Biotechnologie unter Anderem durch das weltweit angelegte Human Genome Projekt im Blickpunkt der Öffentlichkeit. Die Entwicklung der Biotechnologie hatte und hat eine Welle von Firmengründungen zur Folge. In Deutschland wurden diese Aktivitäten ganz wesentlich durch das BMBF Förderkonzept *BioRegio* angestoßen. Neben den fünf Zentren, die vom BMBF im Rahmen dieses Förderkonzeptes unterstützt wurden, konnte in einer Reihe von anderen Regionen eine rasante Entwicklung im Bereich der Biotechnologie beobachtet werden.

In der Initialisierungsphase des Humangenomprojekts hatte James Watson bereits deutlich darauf hingewiesen, dass die anfallenden Massen an Daten (Wörter über dem Alphabet A, T, G, C sowie die Hintergrundinformationen) adäquat zu speichern und zu analysieren seien. Die Komplexität dieser Aufgabe wurde aber zu diesem Zeitpunkt gewaltig unterschätzt. Somit wurden auch nicht direkt begleitende Fördermittel verfügbar gemacht. Verschiedene Aktivitäten aus der Informatik, Biomathematik, Biophysik und der Theoretischen Biologie flossen langsam zusammen und formten die Bioinformatik in den verschiedenen Ländern. So wurde 1992 eine Fachgruppe in der *Gesellschaft für Informatik* gegründet, die sich die Unterstützung der Molekularen Biologie zur Aufgabe machte. 1993 richtete die Fachgruppe die erste nationale Fachtagung in Bonn aus. 1996 wurde diese jährlich stattfindende Konferenz erstmals international ausgetragen (GCB 96 in

Leipzig). Zu diesem Zeitpunkt gehörte die GCB zu den wenigen internationalen Konferenzen auf dem Gebiet der Bioinformatik.

Die erste Phase der Bioinformatik ist dadurch gekennzeichnet, dass Datenbanken und Informationssysteme – seit Mitte der 90er-Jahre im Internet – spezifisch entwickelt und implementiert werden/wurden, um die molekularen Daten systematisch zu erfassen und verfügbar zu machen. Heute sind circa 500 Informationsquellen verfügbar, die spezielle Daten, molekulare Mechanismen und Komponenten repräsentieren. Neben den sequenzierten Genomen und den bereits annotierten Genen sind hier Enzyme, Proteine, biochemische Reaktionen, Signalpathways usw. zu nennen. Neben den Datenmanagementaufgaben hat sich die Bioinformatik in dieser Phase natürlich auch schon mit der Entwicklung und Implementierung von Analysetools beschäftigt, die aus den gespeicherten Daten neues molekularbiologisches Wissen extrahieren. Ein klassisches Beispiel für Analysetools sind Sequenzalignment-Verfahren, die den Grad der Ähnlichkeit von Sequenzen bestimmen. Weitere Schwerpunkte der Forschung waren Sequenzassemblierung (Genomsequenzierung), Gen-Annotation (Suche nach Genen), Funktions- und Strukturvorhersagen, Protein-Folding, Protein-Docking usw. Einige dieser Fragestellungen sind auch heute noch nicht wirklich gelöst und sie gehören weiterhin zu den großen wissenschaftlichen Herausforderungen (*Grand-Challenges*) der Bioinformatik.



Nachdem die Sequenzierung des menschlichen Genoms erfolgreich abgeschlossen werden konnte, rücken andere Fragestellungen aus den Bereichen Molekularbiologie und Genetik, sowie deren potenzielle Anwendungen in der Medizin und Pharmazie in den Mittelpunkt des Interesses. Die zentrale Fragestellung hierbei ist die Entschlüsselung des menschlichen Genoms, d. h. ausgehend von der Sequenz des menschlichen Genoms möchte man ein tiefergehendes Verständnis der molekularen Maschine Mensch entwickeln. Das Ziel dieser Forschung ist es, alle molekularen Bausteine unseres Körpers zu studieren und die Funktion, Struktur und Dynamik (Wechselwirkungen) dieser Bausteine zu verstehen.

Datenmanagement ist auch heute noch ein Schwerpunkt der Bioinformatik. Allerdings steht heute die benutzerspezifische Integration der Datenbestände und Analyse-Tools im Mittelpunkt des Interesses. Die Integrationsphase wird auch bei allen heute verfügbaren Informationsquellen direkt sichtbar. Die Datenbanken werden verlinkt, verschiedene Datenbanken häufig reorganisiert und dann zu einer neuen Datenbank integriert. Integrations-Tools spielen auch als Software-Produkte in diesem Bereich eine wichtige Rolle. So ist das Integrationstool SRS heute ein Schlüsselprodukt der Firma *Lion Bioscience* (Heidelberg). Die Firma *Biobase* (Wolfenbüttel) geht den klassischen Weg der Integration, indem die Produkte selbst zu einem neuen Produkt integriert werden. Auch *Biomax* (München) stellt eine Integrationssoftware zur Verfügung, um den Benutzeransprüchen gerecht werden zu können. Bei all diesen Aktivitäten ist aber nicht zu übersehen, dass sich die Bioinformatik derzeit in den Anfängen dieser zweiten Phase befindet, die auch mit dem Begriff *Integrative Bioinformatik* beschrieben werden kann. Es geht letztlich darum, die elektronische Infrastruktur zu gestalten, die es dann erlaubt, auf der Grundlage der

molekularen Datenbestände sowie geeigneter Modelle und Simulatoren Komponenten einer virtuellen Zelle zu implementieren. Die Implementierung der virtuellen Zelle ist letztlich die treibende Vision. Auch hier hat das BMBF Akzente setzen können und durch die Förderinitiative *Systembiologie* die Leberzelle in Deutschland als Objekt der weiteren Aktivitäten in den Blickpunkt gerückt.

Von der Realisierung dieser Vision sind wir heute weit entfernt, da viele grundlegende molekulare Mechanismen und Systeme (zum Beispiel die Genregulation) noch nicht tiefgehend verstanden werden. Die Bioinformatik wird auch einen wesentlichen Beitrag zum besseren Verständnis dieser komplexen Mechanismen leisten.

Dieses Themenheft charakterisiert die zweite und aktuelle Phase der Bioinformatik. Eine komplexe Datenlandschaft ist bereits gegeben. Modelle zur Analyse der mit diesen Daten in Verbindung stehenden metabolischen Prozesse sind vorhanden. Die Integration dieser Daten und Modelle ist aktueller Forschungsgegenstand und erlaubt die Analyse komplexer metabolischer Netzwerke. Integrierter Datenzugriff unter dem Blickwinkel der Modellierung und Simulation der biochemischen Mechanismen bildet die Grundlage für alle weiteren Aktivitäten in Richtung Virtuelle Zelle. Der Beitrag von *Helge Bertram* und *Thomas Dandekar* präsentiert einen Überblick über diese Zusammenhänge. Der Weg vom Genom zum metabolischen Netzwerk gehört zu den aktuellen Kernaufgaben der Forschung in den Bereichen der Molekularen Biologie und Bioinformatik. Dieser Artikel versucht, neben der Einführung in die Thematik auch die aktuellen Herausforderungen zu umschreiben.

Eine neue Dimension im Bereich der molekularen Forschung soll die Proteomic-Forschung eröffnen. Hier werden Proteinprofile von Robotern erfasst. Diese Ma-

schinen basieren auf verschiedenen Verfahren und Techniken, die in dem Artikel von *Oliver Kohlbacher* und *Knut Reinert* vorgestellt werden. Die Besonderheit dieser Daten liegt darin, dass erstmals neben dem Proteinprofil räumliche Daten und analytische Werte in vivo bestimmt werden. Proteomics aus der Sicht der Datenhaltung und Datenanalyse steht im Zentrum des Artikels von *Manuela Priuß*. Die automatisierte Genexpressionsanalyse steht im Gegensatz zu den Proteomic-Verfahren bereits seit einigen Jahren der Praxis und damit auch der Molekularen Medizin zur Verfügung. Mit Hilfe der Genexpressionsanalyse erhofft man, dass bestimmte Krankheiten anhand von spezifischen Expressionsmustern diagnostiziert werden können. Der Artikel von *Anja von Hydebreck* et al. versucht, genau in diesem Bereich der Anwendung Möglichkeiten und Perspektiven aufzuzeichnen. Die Integration von molekularen Daten und Methoden der Modellierung metabolischer Netzwerke ist das Thema des letzten Artikels dieses Themenheftes. Die Arbeitsgruppe von *Ernst Gilles* gehört dabei zu den Verfechtern der analytischen Methode.

Abschließend möchten wir uns bei allen Personen bedanken, die dazu beigetragen haben, dass dieses *it-Sonderheft* überhaupt entstehen konnte. Einen herzlichen Dank möchten wir den Autoren übermitteln, die mit uns den Versuch unternommen haben, ihr Arbeitsgebiet und in einer verständlichen Sprache zu vermitteln.

**Prof. Dr. Ralf Hofestädt**

Adresse: Universität Bielefeld, Technische Fakultät, Abt. Bioinformatik, Universitätsstraße 25, 33615 Bielefeld,  
E-Mail: hofestae@techfak.uni-bielefeld.de

**Prof. Dr. Hans-Peter Lenhof**

Adresse: Lehrstuhl für Bioinformatik, Universität des Saarlandes, Postfach 151150, 66041 Saarbrücken, E-Mail: len@bioinf.uni-sb.de