
Vorwort

Blättert man durch die aktuellen Ausgaben der Journale *Nature* und *Science*, so findet man kaum ein biologisches Forschungspapier ohne erhebliche Verwendung bioinformatischer Methoden: Sequenzalignment, die Konstruktion phylogenetischer Bäume, der Rückgriff auf biologische Datenbanken mithilfe bioinformatischer Werkzeuge, die Verwendung statistischer Analysemethoden, die aus der bioinformatischen Forschung heraus entstanden sind.

Zum Zeitpunkt des Schreibens an diesem Vorwort ist in der Zeitschrift *Nature* zum Beispiel ein bemerkenswertes Papier¹ zu den Galapagos-Finken erschienen, die Charles Darwin zu einem Schlüsselbeispiel evolutionärer Abläufe gemacht hat. Durch Rückgriff auf eine Vielzahl genetischer und bioinformatischer Methoden (vor allem phylogenetische Rekonstruktionen und statistische Analysen) können die Autoren eine Verbindung zwischen der genetischen Variation und der Schnabelform herstellen, die für Darwin eine zentrale Komponente der Spezies-*fitness* war.

In einer 2012 erschienenen Arbeit in *Science*² untersuchten die Autoren die Algenblüte mithilfe eines Metagenomikansatzes. Sie hatten sich zum Ziel gesetzt, ein Paradoxon der Meeresbiologie zu lösen: Wie können ohne starke räumliche Trennung so viele Mikroorganismen koexistieren? Mit hoher zeitlicher Auflösung wurde der DNA-Inhalt von Proben sequenziert, analysiert und daraus die Spezieszusammensetzung extrahiert. Die Arbeit brachte ein neues Ordnungsprinzip dieses wichtigen biologischen Systems zutage: Die Bakterienpopulationen im Wasser während des Verlaufs der Algenblüte sind hoch optimiert für die Nahrungsverfügbarkeit der verschiedenen zeitlichen Phasen. So entstehen zeitlich getrennte Nischen. Diese Arbeit ist aus vielen Blickwinkeln äußerst charakteristisch für den Alltag der modernen Bioinformatik: Herausragende Erkenntnisse werden erzielt, wenn Bioinformatik eingebettet wird in große Konsortien aus experimentell, mathematisch und informatisch

¹ Lamichhaney S et al. (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518:371–375.

² Teeling H et al. (2012) Substrate controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* 336:608–611.

arbeitenden Wissenschaftlern. Ohne die bioinformatischen Methoden zum Klassifizieren der Sequenzsegmente und zur metabolischen Interpretation der bakteriellen Enzyminventare hätten sich die durch diese Arbeit gewonnenen Erkenntnisse nicht erzielen lassen.

Es gibt viele weitere Beispiele für die zentrale Bedeutung der Bioinformatik in den großen biologischen Forschungskonsortien. Das ENCODE-Projekt³, bei dem aus genomischer Variation evolutionär konservierte Bereiche jenseits der Gene sichtbar gemacht wurden, hat in den letzten Jahren einen enormen und sehr fundamentalen Beitrag für unser Verständnis genomischer Information und der Regulation biologischer Prozesse geleistet. Die Vielzahl von Erkenntnissen aus diesem Projekt könnte zudem große Anwendungen in dem sich gerade entwickelnden Gebiet der personalisierten Medizin haben. Hier zeigen sich Bioinformatik und Systembiologie aufs Engste verzahnt mit biologischen und medizinischen Fragestellungen.

Mit der stark erweiterten Neuauflage unseres Lehrbuchs versuchen wir, einigen dieser neueren Entwicklungen der Bioinformatik einen angemessenen Raum zu geben, ohne die Grundlagen, die 2006 den Schwerpunkt der ersten Auflage bildeten, zu vernachlässigen. So wurde die Diskussion des *next-generation sequencing* ergänzt und der wichtige Aspekt der Datenzusammenführung oder Datenintegration stärker hervorgehoben. Besonders die Kap. 5.2 und 5.3 führen in eine kleine Auswahl aktueller Themen ein, mit denen die Anwendungserfolge der Bioinformatik der letzten Jahre illustriert werden sollen. Die wesentlichen Merkmale des Buchs sind auch in der zweiten Auflage erhalten geblieben: Ein Verständnis bioinformatischer Methoden wird anhand der großen grundlegenden Algorithmen entwickelt; das Computeralgebra-Programm *Mathematica* stellt die Programmierumgebung für praktische Anwendungen dar. Dabei wurden alle Beispiele an die neueste *Mathematica*-Version (10.0) angepasst. Wichtige Entwicklungen der letzten Jahre betreffen ebenso Softwarewerkzeuge für eine Vielzahl konkreter Anwendungen wie den großen statistischen Rahmen der Bioinformatik (speziell Fragen der Inferenz und der Schätzung unter unvollständiger Information). Auch die Verzahnung und der Austausch mit Gebieten der Algorithmik und der angewandten Informatik haben spürbaren Einfluss auf Aspekte der Bioinformatik.

Eine sehr lesenswerte Übersicht über die bewegte Geschichte der Bioinformatik bietet Ouzounis (2012).⁴ Wie sehr selektives Zitieren und die selektive Interpretation statistischer Befunde gerade die Bioinformatik gefährden, ist von Boulesteix (2010) beschrieben worden.⁵

Die Bioinformatik hat sich in den letzten sieben Jahren, seit der ersten Auflage dieses Lehrbuchs, deutlich in Richtung der Informatik verschoben. Es ist aus unserer

³ The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 488:57–74.

⁴ Ouzounis CA (2012) Rise and demise of bioinformatics? Promise and progress. *PLoS Computational Biology* 8:e1002487.

⁵ Boulesteix AL (2010) Over-optimism in bioinformatics research. *Bioinformatics* 26:437–439.

Sicht zwingend, diese wichtige und sich aus ihrer Anwendung heraus definierende Disziplin nah an ihrem biologischen Gegenstand auszugestalten.

Wir haben es – gerade vor diesem Hintergrund – als besonders wichtig empfunden, den Ton der ersten Auflage nicht durch übermäßige Formalisierung oder ein zu starkes Gewicht auf den statistischen Fragestellungen zu verändern. Nach wie vor ist unsere ideale Leserschaft gegeben durch Studierende der Bioinformatik, Biologie, Biotechnologie oder Medizin, die – durchaus auch ohne größere Programmierkenntnisse – einen Überblick über die Methodenvielfalt der Bioinformatik gewinnen möchten oder auch sich konkrete Werkzeuge zum Analysieren und Interpretieren ihrer eigenen Daten erarbeiten wollen.

Während der Arbeit an dieser zweiten Auflage sind wir von vielen Seiten intensiv unterstützt worden.

Den Mitgliedern der Arbeitsgruppe *Computational Systems Biology* an der Jacobs University in Bremen danken wir für das ‘Testlesen’ einiger zentraler Kapitel, besonders Christoph Fretter, Sergio Grimbs und Jens Christian Claussen.

Viele Inhalte sind durch intensive Diskussionen mit Kollegen motiviert. Gedankt sei an dieser Stelle besonders Frank Oliver Glöckner (Bremen), Ivo Grosse (Halle), Christian Hammann (Bremen), Werner E. Helm (Darmstadt) und Nicole Radde (Stuttgart). Nikolaus Sonnenschein (Kopenhagen) danken wir für die Diskussionen zur MASSTOOLBOX.

Andreas Held danken wir für das kompetente Lektorat des Manuskripts. Herrn Rainer Plaumann danken wir für die sorgfältige Umsetzung verschiedener Korrekturdurchgänge in LaTeX.

Stefanie Wolf vom Springer-Verlag danken wir erneut sehr herzlich für die bewährt gute und aufmerksame Betreuung des Projektes.

Bremen und Weihenstephan im Februar 2015,

Marc-Thorsten Hütt
Manuel Dehnert

Vorwort zur ersten Auflage

Bioinformatische Methoden bilden eine unverzichtbare Grundlage nahezu aller Aspekte der Molekularbiologie und Genetik. Der methodische Apparat zeigt sich dem Anwender vor allem in der Gestalt mehr oder weniger gut bedienbarer Softwarewerkzeuge. Doch die entscheidenden Fragen beginnen dahinter:

- Was bedeuten die Voreinstellungen der Parameter?
- Wie aussagekräftig ist das Resultat?
- Ist das Resultat optimal oder würde eine verfeinerte bioinformatische Methode auf ein besseres Verständnis des biologischen Systems führen?
- Wie kann man eigene Ergebnisse mit anderen vergleichen, die mit einem etwas anderen bioinformatischen Analyseverfahren gewonnen wurden?

Dieses Buch führt ein in die thematischen, praktischen und mathematischen Grundlagen der Bioinformatik: Es stellt dar, wie man von mathematischen Kerngedanken zu einer konkreten bioinformatischen Methode gelangt. An anderen Stellen wird der umgekehrte Weg gegangen und Schritt für Schritt werden die mathematischen Verfahren und Strategien hinter einer gegebenen (als Bioinformatik-Software erhältlichen) Implementierung aufgezeigt. Damit bietet sich dem Anwender aus der Biologie die Möglichkeit, ohne erhebliche zusätzliche Mathematik- oder Programmierkenntnisse zu erwerben, die Verfahren hinter den etablierten bioinformatischen Methoden zu verstehen.

Die beschriebenen Methoden sind exemplarisch. Es war uns wichtiger, einzelne Verfahren auf allen Ebenen (mathematisches Konzept – algorithmische Realisierung – praktische Anwendung) darzustellen, als methodische Vollständigkeit anzustreben. Wir denken, dass diese Auswahl zu einem Grundverständnis bioinformatischen Arbeitens führt und vor allem eine Denkweise darstellt, die sich als nützlich für die alltäglichen Problemstellungen von mit Sequenzdaten konfrontierten Studierenden und wissenschaftlich Arbeitenden erweist.

Ein entscheidender Vorteil dieser Perspektive ist, dass viele Programmoptionen dann deutlich klarer werden. Wir werden sehen, dass die Beherrschung einiger weniger zentraler Algorithmen große Teile der Bioinformatik in Grundzügen zugänglich zu machen vermag. An einigen Stellen werden wir bis in die Tiefe der formalen algorithmischen und mathematischen Fundamente dieser Verfahren gehen, z. B. bei paarweisem Sequenzalignment, bei Hidden-Markov-Modellen und bei der Konstruktion phylogenetischer Bäume.

Eine wichtige Frage der Präsentation ist, auf welche Weise man Biologen und Medizinern die erforderlichen praktischen Kenntnisse und Fertigkeiten im Umgang mit Algorithmen und Programmierung vermitteln kann. Im Gegensatz zu den meisten solchen Versuchen, die sich auf die Anwendung fertig implementierter Softwarepa-

kete oder aber auf einfache Programmieraufgaben (z. B. in Perl) beschränken, sind wir den Weg über das Computeralgebra-Programm *Mathematica* gegangen. *Mathematica* erlaubt, elementare Beispiele bioinformatischer Anwendungen in transparenter Weise darzustellen, mathematische Hypothesen unmittelbar zu testen und die im Verlaufe des Buchs diskutierten Algorithmen schnell und effizient zu implementieren. In *Mathematica* stehen die Bausteine mathematischer Modelle (z. B. Differenzialgleichungen oder spezielle Funktionen) auf derselben Stufe wie Verfahren der Datenhandhabung, Datenanalyse und Visualisierung. Das gesamte Spektrum bioinformatischer Arbeitsweisen kann so in derselben Umgebung erfolgen. Unerwartete Unterstützung hat dieser Weg durch die aktuelle Entwicklung von *Mathematica* erfahren. Seit der Version 5.1 (im Oktober 2005) beschäftigen sich zentrale Neuerungen mit der Behandlung von Symbolsequenzen, sodass *Mathematica* sich zu einer wichtigen bioinformatischen Forschungs- und Arbeitsoberfläche weiterentwickelt.

Im ersten Kapitel wird der Rahmen bioinformatischer Beschäftigung dargestellt. Zu einer solchen *Skizze des Fachs* gehören auch eine Form von Begriffs- und Gedanken-geschichte der Bioinformatik und eine kurze Einführung der biologischen Schlüsselbegriffe dieser Disziplin. Am Ende des Kapitels stehen eine – sehr subjektive – Liste von Kernfragen und eine erste Einführung in *Mathematica*.

Einen ersten Blick auf das mathematische Fundament der Bioinformatik soll dann das Kapitel *Statistische Analyse von DNA-Sequenzen* erlauben. Dabei werden zuerst einige elementare Eigenschaften von Wahrscheinlichkeiten diskutiert und die mathematische Notation bereitgestellt, um Wahrscheinlichkeitsmodelle einzuführen und die Grundprinzipien der Parameterschätzung zu besprechen. Solche Werkzeuge erlauben z. B. die Frage zu diskutieren, wie sich DNA-Sequenzen von zufälligen Symbolabfolgen unterscheiden. Diese Frage bildet den Ausgangspunkt, den Zusammenhang von Sequenz, Struktur und Funktion zu diskutieren, und führt später schließlich auf wichtige statistische Eigenschaften eukaryotischer Genome.

Solche Betrachtungen werden uns unmittelbar auf zwei Schlüsselkonzepte der statistischen Analyse von Symbolsequenzen führen, nämlich Markov-Modelle und Hidden-Markov-Modelle. Damit schließt sich die Kluft zwischen den elementaren mathematischen Begriffen und den etablierten Techniken, die – als Hidden-Markov-Modelle – in so unterschiedlichen Bereichen wie Genidentifikation, Aspekten der Proteinstruktur und Sequenzalignment Anwendung finden.

In Kap. 2 bilden die unbehandelten, elementaren Sequenzdaten den Ausgangspunkt und es wird versucht, Schritt für Schritt mit immer fortgeschritteneren mathematischen Methoden Informationen aus diesen Daten zu extrahieren. In Kap. 3 kehrt sich diese Blickrichtung um. Nun werden wir etablierte Analysemethoden und Betrachtungsweisen in den Vordergrund stellen, um dann hinter die Kulissen der gegebenen Softwareimplementierungen zu gelangen und die dort wirksamen mathematischen Verfahren zu entdecken und zu verstehen. Den Anfang bilden Verfahren des Sequenzvergleichs, gefolgt von phylogenetischen Analysen, die Unterschiede zwi-

schen ähnlichen Sequenzsegmenten in einen kausalen Zusammenhang bringen. Am Ende stehen einige Bemerkungen zu bioinformatischen Datenbanken.

Kapitel 4 beschäftigt sich mit der Sequenzanalyse auf der Skala vollständiger Genome mit Methoden der Informationstheorie. Dahinter verbirgt sich der Versuch zu quantifizieren, wie stark ein Symbol aus einer Sequenz mit einem anderen Symbol in einiger Entfernung korreliert ist, welche Systematiken diese Korrelationen aufweisen und wie Beiträge zu solchen Korrelationen mit biologischen Eigenschaften in Verbindung gebracht werden können. Die zentralen Begriffe der Informationstheorie, Information und Entropie, gewinnen in der Bioinformatik immer stärker an Bedeutung. Ein Grund ist dabei, neben lokalen Aspekten einer Sequenz (also etwa einzelne Gene) auch globale Eigenschaften einer Sequenz zu diskutieren (z. B. Fluktuationen der Dinukleotidhäufigkeiten oder die Verteilung repetitiver Elemente).

Aus unserer Sicht entwickelt sich gerade zur Zeit die Bioinformatik mit dem Aufkommen viele Einzelinformationen integrierender Datenbanken zu einem Startpunkt einer molekular verorteten Systembiologie. Um diesen Weg hin zu einer systemorientierten und modellierenden Bioinformatik weiter auszugestalten, sind Kenntnisse aus angrenzenden Themenfeldern erforderlich, etwa der Komplexitätsforschung und der Netzwerkbiologie. Kapitel 5 stellt diese Gedanken dar. Besonders bei Methoden der Graphentheorie steht hinter den formalen, abstrakten Begriffen ein interessanter vereinheitlichender Blick auf extrem unterschiedliche biologische Objekte: Mit denselben theoretischen Werkzeugen können so genetische Netzwerke, Protein-Interaktionsnetzwerke und metabolische Regulationsnetzwerke beschrieben, funktionell charakterisiert und verglichen werden.

Die Komplexitätstheorie ist eng verzahnt mit fraktaler Geometrie. Ein Fraktal ist ein selbstähnliches Objekt, d. h. dass eine vergrößerte Kopie eines Ausschnitts vom Original nicht prinzipiell zu unterscheiden ist. Ein solches Fehlen einer charakteristischen Längenskala (die bei Vergrößerungen oder Verkleinerungen eine Orientierung liefern könnte) gibt es auch bei DNA-Sequenzen. Wir werden fraktale Eigenschaften von DNA-Sequenzen sichtbar machen, um diese Parallele zwischen Fraktalen und DNA-Sequenzen schließlich auf ihre biologischen Ursachen zu befragen. Am Ende des Kapitels werden wir schließlich eine interessante, aber keineswegs triviale Parallele zwischen genetischen Netzwerken und fraktaler Geometrie verfolgen, die sich als sehr eleganter Zugang zur Systembiologie herausstellen wird.

In den Anwendungsbeispielen der Methoden beschränken wir uns sehr oft auf Eukaryoten. Vor allem steht an vielen Stellen das menschliche Genom im Vordergrund. Entsprechend sind auch viele Kapitel zum biologischen Hintergrund (etwa zum Genaufbau oder zur Genomorganisation) sehr stark aus einer eukaryotischen Perspektive abgefasst.

Die angegebenen Literaturhinweise zu jedem Kapitel spiegeln stark unsere persönlichen Vorlieben wider. Diese Empfehlungen sind keinesfalls als systematische (oder gar vollständige) Bibliografie zu verstehen. Zu vielen Fachbegriffen geben wir im

Text die englische Übersetzung an, um die Suche nach Forschungsartikeln zu diesem Thema zu erleichtern.

Natürlich ist dieses Buch auch ein Produkt intensiver Dialoge.

Wir danken Werner E. Helm für viele Diskussionen über die Verbindung von Statistik und Biologie und für unsere gemeinsamen bioinformatischen Forschungsarbeiten, die explizit (vor allem in Kapitel 2 und 4) oder implizit in die hier diskutierten Anwendungen eingeflossen sind.

Danken möchten wir auch den Teilnehmerinnen und Teilnehmern unserer Darmstädter Bioinformatik-Lehrveranstaltungen, die mit Fragen, kritischen Rückmeldungen und Diskussionen unsere Darstellung mit geprägt haben.

Christiane Hilgardt danken wir für die Mitarbeit bei der Erstellung der elektronischen Fassung und eine gründliche Lektüre der biologischen Fallbeispiele. Eine Vielzahl von Skizzen wurde von Doris Schäfer in sehr schöne digitale Abbildungen verwandelt. Teile der elektronischen Fassung wurden zudem von Carsten Marr, Stefan Schmelz, Rainer Plaumann und Philipp Weil bearbeitet.

Gerhard Thiel und Brigitte Hertel danken wir für ihre Hilfestellungen bei Kaliumkanälen und anderen Beispielen zur Verbindung von Sequenz und Struktur von Proteinen.

Arnulf Kletzin hat große Teile des Manuskripts, die Aspekte der praktischen Bioinformatik betreffen, kritisch durchgesehen.

Erich Bohl hat einen großen Einfluss auf die begriffliche und mathematische Feinstruktur der Kapitel 1 und 2 gehabt.

Für eine sehr umfassende Durchsicht des Textes und des Layouts der Endfassung danken wir Stefan Christ und Heike Hameister.

Kapitel 4.2 hat sehr profitiert von einem gemeinsam mit Markus Porto (Fachbereich Physik der TU Darmstadt) veranstalteten Seminar ‘Biophysikalische Prinzipien der Genomorganisation’ im Sommersemester 2005.

Für intensive Korrekturen von Teilen des Manuskripts, für Diskussionen und Änderungsvorschläge danken wir zudem Erich Bohl, Stefan Bornholdt, Markus Domschke, Christoph Fretter, Werner E. Helm, Christiane Hilgardt, Franz-Josef Meyer-Almes, Dirk Plendl, Markus Porto, Stefanie Sammet, Gerhard Thiel und Katrin Wolff.

Dem Team vom Springer-Verlag, vor allem Frau Stefanie Wolf, danken wir für die engagierte und professionelle Betreuung dieses Buchprojekts.

Es ist klar, dass der gedruckte Text nicht das ideale Medium für die Programmzeilen unserer *Mathematica*-Exkurse darstellt. Wir haben daher die elektronischen Fassungen auf einer Internetseite www.bioinformatik-mathematica.de zur Verfügung gestellt.

Das Buch ist geprägt von unserer persönlichen Perspektive, die – trotz aller intensiven Auseinandersetzung mit der Biologie und der jahrelangen Arbeit in diesem Fach – unsere *formations professionnelles* als Mathematiker und theoretischer Physiker nicht verbergen kann.

Darmstadt im Januar 2006,

Marc-Thorsten Hütt
Manuel Dehnert

Methoden der Bioinformatik

Eine Einführung zur Anwendung in Biologie und Medizin

Hütt, M.-T.; Dehnert, M.

2016, XVII, 406 S. 217 Abb., Softcover

ISBN: 978-3-662-46149-5