Springer-Lehrbuch

Marc-Thorsten Hütt Manuel Dehnert

Methoden der Bioinformatik

Eine Einführung

205 Abbildungen

Springer

Prof. Dr. Marc-Thorsten Hütt
Manuel Dehnert
Technische Universität Darmstadt
Institut für Botanik
Schnittspahnstraße 3–5
64287 Darmstadt

E-mail: huett@bio.tu-darmstadt.de
dehnert@bio.tu-darmstadt.de

Bibliografische Information der Deutschen Bibliothek

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

ISBN-10 3-540-25687-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-25687-8 Springer Berlin Heidelberg New York

Springer ist ein Unternehmen von Springer Science+Business Media
springer.de

Planung: Dr. Dieter Czeschlik, Heidelberg
Redaktion: Stefanie Wolf, Heidelberg
Satz: Druckfertige Vorlagen der Autoren
Herstellung: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Einbandgestaltung: deblik, Berlin
Titelbild: Visualisierung des Protein-Interaktionsnetzwerks von *Helicobacter pylori* (vgl. Abbildung 5.38)

Gedruckt auf säurefreiem Papier SPIN 11419518 29/3150YL - 5 4 3 2 1 0

Meiner Tochter Milena für ihre unbestechlichen Einschätzungen.
– MTH

Für Vinciane, mit der ich das Leben genieße.
– MD

Vorwort

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 Software-Werkzeuge. 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, zum Beispiel 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 Softwarepakete 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. Differentialgleichungen 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, so dass *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* gehört auch eine Form von Begriffs- und Gedankengeschichte der Bioinformatik und eine kurze Einführung der biologischen Schlüsselbegriffe dieser Disziplin. Am Ende des Kapitels steht 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 Grundprinzipen der Parameterschätzung zu besprechen. Solche Werkzeuge erlauben zum Beispiel, 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 Kapitel 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 *Kapitel 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 zwischen 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 (zum Beispiel 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-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) Bibliographie 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 professionelles* als Mathematiker und theoretischer Physiker nicht verbergen kann.

Darmstadt im Januar 2006, *Marc-Thorsten Hütt*
Manuel Dehnert

Inhaltsverzeichnis