Aktuelles Schlagwort Bioinformatik

Erschienen in *Informatik Spektrum,* Okt. 1999

Rolf Backofen[1], François Bry[1], Peter Clote[1], Hans-Peter Kriegel[1], Thomas Seidl[1] und Klaus Schulz[2]

[1]Institut für Informatik und [2]Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians-Universität München

Die Bioinformatik tangiert einerseits die Molekularbiologie, die Biochemie und die Genetik, andererseits die Theoretische und Praktische Informatik und die Computerlinguistik. Sie verfügt über einen homogenen und breiten Bestand an offenen Problemen. Sie gewinnt immer mehr an Bedeutung in Biologie und Genetik und wird schon industriell eingesetzt.

Ziele der Bioinformatik. Die Proteine sind komplexe Moleküle, die Bausteine aller Lebensformen sind. Die Vielfalt an Proteinen ist gewaltig: Z.B. kommen im menschlichen Körper über eine Million von verschiedenenen Proteinen vor. Proteine setzen sich aus Aminosäuren in einer Weise zusammen, die in der DNA (Deoxyribose Nucleic Acid oder Desoxyribonukleinsäure) kodiert ist. Die DNA ist ein lineares Polymer, das aus 4 Nukleotiden aufgebaut ist. Eine Teilsequenz aus 3 Nukleotiden heißt Codon. Jede der 20 Aminosäure wird durch 1 bis 6 der $4^3 = 64$ Codons kodiert, die meisten in mehrfacher Weise. Ein Ziel der **Sequenzanalyse** ist es, die Bereiche der DNA zu finden, die ein Protein kodieren. Dies ist sehr komplex: Es gibt viele nicht-kodierende Bereiche; die Anfänge von kodierenden Bereichen und nicht-kodierenden Bereichen sind schwierig zu erkennen; die kodierenden Bereiche sind nicht notwendigerweise verbunden. Zur Sequenzanalyse werden derzeit Informatikmethoden mit biochemischen Laboruntersuchungen kombiniert.

Die DNA-Analyse und die Verarbeitung natürlicher Sprachen haben vieles gemeinsam: „Kodierungen`` werden aus Sequenzen erkannt; riesige Mengen von empirisch gewonnen Daten stehen zur Verfügung; Gesetzmäßigkeiten sind erkennbar, die bisher nur teilweise verstanden wurden; die systematische Erfassung von Ähnlichkeiten ist notwendig; die Grenze zwischen wohlgeformten und nicht-wohlgeformten Zeichenfolgen ist fließend. In beiden Bereichen werden stochastische Methoden wie **stochastische Grammatiken** und **Hidden-Markov-Modelle** eingesetzt.

Die **Vorhersage der Proteinstruktur** ist ein Hauptziel der Biowissenschaften, weil die Funktion eines Proteins von seiner 3-dimensionalen Gestalt abhängt. Ihre vollständige Lösung käme in Medizin und Pharmazie einer wissenschaftlichen und wirtschaftlichen Revolution gleich. Um schwierige Laboruntersuchungen zu vermeiden, werden Informatikmethoden zur **Proteinfaltung** eingesetzt, die die Struktur eines Proteins aus seiner Aminosäuresequenz ermitteln.

Man unterscheidet die Primär-, Sekundär, Tertiär- und Quartär- (oder Quaternär-) Strukturen eines Proteins. Die **Primärstruktur** ist die Aminosäuresequenz. Die **Sekundärstruktur** ist eine Abstraktion der 3-dimensionalen Gestalt in Form von lokalen Faltungsmustern namens α -**Helix**, β - **Faltblatt** und **Turn**. Die **Tertiärstruktur** ist die 3-dimensionale Gestalt von Proteinuntereinheiten. Die **Quartärstruktur** gibt wieder, wie sich die verschiedenen Untereinheiten eines Proteins räumlich zusammenlagern. Bisher sind die Tertiärstrukturen von nur ca. 9 000 Sequenzen bekannt. Die **homologie-basierte Vorhersage** der

Tertiärstruktur beruht auf einem Vergleich der Sequenz, deren Struktur ermittelt werden soll, mit Sequenzen, deren Tertiärstrukturen schon bekannt sind. Bei der **Ab-Initio-Strukturvorhersage** werden Tertiärstrukturen mit minimaler freier Energie durch globale Optimierungsmethoden ermittelt. Die Frage ob und wie ein Protein mit anderen Molekülen einen stabilen Komplex bilden kann, heißt **Protein-Docking-Problem**. Verfahren zum **1:1-Docking** einzelner Proteinpaare liefern eine relative Positionierung der Moleküle zueinander. Beim **1:n-Docking** werden in einer Moleküldatenbank Docking-Partner für ein gegebenes Protein gesucht. Zum 1:n-Docking sowie zur homologie-basierten Strukturvorhersage werden Verfahren der Molekulardynamik, diskrete Techniken, genetische Algorithmen, geometrische Algorithmen sowie Data-Mining- und Knowledge-Discovery-Methoden eingesetzt.

Derzeit gibt es mehr als hundert verschiedene Datenbanken in der Molekularbiologie: u.a. DDBJ, EMBL, GenBank, PIR und SwissProt. Viele dieser Datenbanken sind sehr groß. GenBank enthält z.B. ca. 4×10^6 Nukleotidsequenzen, die insgesamt aus ca. 3×10^{12} Vorkommen von Nukleotiden bestehen. Den Biodatenbanken liegt kein einheitliches Schema zu Grunde. Die Verknüpfung **heterogener Biodatenbanken** und die **Schemaintegration für Biodatenbanken** sind weitgehend noch ungelöste Probleme von großer wirtschaftlicher Bedeutung.

Die Evolution verändert mit der Zeit die in der DNA kodierten Proteine. Es ist möglich durch computergestütze Sequenzanalysen und Klassifizierung diese Veränderungen und daraus die Stammbäume, **phylogenetische Bäume** genannt, von verwandten Spezies zu ermitteln. Der Ansatz wird u.a. in der evolutionären Paläontologie eingesetzt.

In letzter Zeit gewinnt die computergestützte Ermittlung von **metabolischen** und **regulatorischen Pfaden** an Wichtigkeit. Ein metabolischer Pfad ist eine abstrakte Darstellung eines Stoffwechselprozesses, der die daran beteiligten Proteine und Moleküle auflistet. Ein regulatorischer Pfad stellt die Informationsflüße in einem Zelltyp dar, deren Missverhalten die Grundlage für viele Krankheiten - wie etwa Krebs - bildet. Um die Ähnlichkeit von metabolischen bzw. regulatorischen Pfaden in unterschiedlichen Organismen zu untersuchen, werden Methoden der Mustererkennung, der Ähnlichkeitssuche in Datenbanken und der Sequenzanalyse eingesetzt.

Gene sind DNA-Bereiche, die Proteine kodieren und dadurch Erbeigenschaften bestimmen. Ein Gen kann in Zellen eines bestimmten Typs ,,exprimiert``, d.h. zur Proteinentwicklung ,,abgelesen``, werden. Man spricht von **Genexpression**. Mit einem einzelnen **DNA-Chip** können die Konzentrationen oder **Expressionsniveaus** von tausenden bis hunderttausenden Genen gemessen werden, die in einem bestimmten Zelltyp exprimiert werden. In **differentiellen Displays** lassen sich die Unterschiede zwischen den Expressionsniveaus in gesunden und kranken Zellen desselben Zelltyps feststellen. Die sehr umfangreichen Daten, die in dieser Weise erhalten werden, sind der Ausgangspunkt für neue Ansätze zur Diagnose und Therapie, die Informatikmethoden und biochemischen Laboruntersuchungen kombinieren.

Etwas abseits der Bioinformatik, jedoch im Bezug dazu ist die Grundlagenforschung zum Thema **Biocomputing**. Damit wird der Einsatz von molekularbiologischen Methoden wie **Gelelektrophorese** und **Polymerase-Kettenreaktion** zur Berechnung komplexer mathematischer Probleme wie etwa kryptographischer Entschlüsselung bezeichnet. Die Durchsetzungschancen des Biocomputings sind noch völlig offen.

Die Bioinformatik in der Industrie. Unter den bereits etablierten industriellen Anwendungen der Bioinformatik finden sich die **Arzneimittelentwicklung**, die **Gentherapie**, die **Erkennung von genetischen Risikofaktoren** wie die Geisteskrankheit Fragile-X-Syndrom und die Huntingtonskrankheit, die nicht unumstrittene **Genveränderung** in der Agrar- und Züchtungswirtschaft und die **biometrischen Unterschriften**. Die für die pharmazeutische Industrie hochrelevante **Genexpression** trägt viel zum gegenwärtigen industriellen Interesse an der Bioinformatik bei. Führende Konzerne wie *BASF*, *Hoechst*, *Hoffmann-La Roche* (mit *Boehringer Mannheim*), *Merck*, *Millenium Pharmaceutical*, *Rhône Poulenc* und *SmithKline Beecham* und Unternehmen wie *Affymetrix*, *Artemis*, *Incyte* und *LION AG* betreiben in erheblichem Umfang Forschung und Entwicklung in der Bioinformatik. Unternehmen wie *Medigenomix*, *GPC*, *Epidaurus* und *Switch Biotech*, die in der **Genomanalytik** tätig sind, nutzen die Bioinformatik intensiv.

Methoden und Perspektiven der Bioinformatik. Die Bioinformatik setzt Methoden aus verschiedenen Gebieten der Informatik ein: u.a. Kombinatorische Optimierung, Integer Linear Programming, Constraint-Programmierung, Algorithmik und Formale Sprachen, Genetische Algorithmen, Geometrische Algorithmen, Stochastische Algorithmen, Neuronale Netze, Mustererkennung, Maschinelles Lernen, Inductive Logic Programming, Knowledge Discovery und Data Mining, Computerlinguistik. Wegen einerseits der gewaltigen Datenmengen, die behandelt werden müssen, und andererseits den allgegenwärtigen Ausnahmen zu erkennbaren Gesetzmäßigkeiten bieten die Biowissenschaften ein herausragendes Anwendungsfeld für die moderne Informatik.

Die potentielle Einsatzmöglichkeiten der Informatik in Biowissenschaften gehen weit über ihre derzeitigen Anwendungen hinaus. Die Rolle, die die Informatik bei den Biowissenschaften nun spielt, ähnelt der Rolle der Mathematik in der Physik: Erst der Einsatz von Informatikmethoden ermöglicht es, in Biowissenschaften Modelle zu bilden und mit ihnen zu rechnen statt im Reagenzglas zu experimentiere.

Der Wichtigkeit der Bioinformatik sowohl für die Biowissenschaften als auch für die Informatik wird derzeit in der Lehre noch nicht ausreichend Rechnung getragen. Hochschulabsolventen, die eine fundierte Informatik- und Biologieausbildung vorweisen können, sind derzeit „Mangelware``. Nur an wenigen Universitäten werden Studiengänge in Bioinformatik angeboten oder eingerichtet.

In der Bioindustrie wird die Bioinformatik als Schlüsseltechnologie angesehen. Nicht zuletzt junge Biotechnologieunternehmen wie diejenigen der „BioTech-Regionen`` (s. http://www.bioregio.com) sind auf eine starke universitäre Bioinformatik angewiesen.

Die Autoren danken den anonymen Gutachtern für ihre konstruktiven Hinweise.

Literatur

- [1] Baldi, P. und Brunak, S. Bioinformatics - The machine learning approach, MIT Press, 1998
- [2] Baxevanis, A. D. und Ouellette, F. Bioinformatics - A practical guide to the analysis of genes and proteins, Wiley, 1998
- [3] Durbin, R. et al. Biological sequence analysis, Cambridge University Press, 1998
- [4] Giegerich, R. et al. Empfehlung zur Einrichtung von Studiengängen im Fach Bioinformatik, Fachgruppe 4.0.2 der GI, 1997 - http://wwwiti.cs.uni-magdeburg.de/iti_bm/fb4/gi/bioinfocurric.ps

- [5] Gusfield, G. Algorithms on strings, trees, and sequences: Computer science and computational biology, Cambridge University Press, 1997
- [6] Hunter, L. ed. Artificial intelligence and molecular biology, MIT Press 1993
- [7] Lengauer, T. et al. Bioinformatik - Diagnose von Krankheiten und Entwicklung von Wirkstoffen mit Hilfe des Computers, In Software, Spektrum-der-Wissenschaft-Dossier 2/1999, 38-43
- [8] Searls, D. B. The computational linguistics of biological sequences in [6]
- [9] Suhai, S. ed. Theoretical and computational methods in genome research, Plenum Press, 1997
- [10] Waterman, M. Introduction to computational biology, Chapmann & Hall, 1995

François Bry

1st October 1999