# Aktuelles Schlagwort
# “Datenbanken in der Bioinformatik”

François Bry und Peer Kröger

Institut für Informatik, Ludwig-Maximilians-Universität München
http://www.pms.informatik.uni-muenchen.de

Die Bioinformatik [2] kann als Anwendung von Informatikmethoden zur Untersuchung von Problemen der Molekularbiologie definiert werden, die auf sehr große Datenmengen beruhen und einer umfangreichen Datenanalyse bedürfen. Folglich spielen Datenbanken in der Bioinformatik eine zentrale Rolle. Die Bioinformatik-Datenbanken haben interessante Merkmale, die in herkömmlichen Datenbanken selten vorkommen. Insbesondere zeigen Datenbanken mit unterschiedlichen Molekularbiologie-Daten einheitliche Merkmale.

## Fragestellungen, Daten und Verwendung von Datenbanken

**Sequenz-Datenbanken zur DNS-Analyse und Sequenzierung.** Proteine, die “Bausteine des Lebens”, sind aus Aminosäuren nach einem Bauplan zusammengesetzt, der in der DNS kodiert ist. Die DNS ist ein lineares Polymer, Sequenz genannt, das aus vier Nukleinsäuren aufgebaut ist. Innerhalb der DNS gibt es kodierende sowie nicht-kodierende Abschnitte, deren Grenzen schwierig zu erkennen sind. Ziel der Sequenzierung ist es, die kodierenden sowie nicht-kodierenden Bereiche in der DNS-Sequenz eines Organismus zu ermitteln. DNS-Analyse-Methoden werden eingesetzt, um kodierende sowie weitere Bereiche zu erkennen. Die Daten werden in (meist sehr großen) Sequenz-Datenbanken verwaltet: GenBank [3] enthält z.B. ca. $2 \cdot 10^7$ Nukleotidsequenzen und referenziert ca. $2 \cdot 10^{10}$ Vorkommen von Nukleotiden. Das Wachstum der meisten Sequenz-Datenbanken ist exponentiell.

**Proteinsequenz- und Proteinstruktur-Datenbanken zur Strukturvorhersage.** Die Funktion eines Proteins hängt von seiner 3D-Struktur ab, so dass die Vorhersage der Struktur von Protein-Molekülen aus ihrer Sequenz, die sogenannte Protein-Faltung, ein Hauptziel der Biologie ist. Bioinformatik-Methoden werden eingesetzt, um Annäherungen an die tatsächliche Struktur von Proteinen zu berechnen und damit Laboruntersuchungen einzuschränken. Homologie-basierte Ansätze vergleichen dabei die Aminosäuresequenzen bereits bekannter Proteine mit der Sequenz des unbekannten Proteins. Dafür werden Proteinsequenz- und Proteinstruktur-Datenbanken aufgebaut und Ähnlichkeitsanfragen an solche Datenbanken gestellt.

**Biochemische Pfade.** Ein biochemischer Pfad ist eine abstrakte Modellierung von aufeinanderfolgenden chemischen Reaktionen in einer Zelle. Besondere Beachtung finden Metabolische Pfade (Reaktionswege im Stoffwechsel) und Regulatorische Pfade (Kontrollmechanismen in der Genexpression). Zum Auffinden biochemischer Pfade werden unter anderem Sequenz-Datenbanken verwendet.

Die so gewonnenen Daten werden in speziellen Datenbanken verwaltet.

**Sequenz-Datenbanken zur Ermittlung phylogenetischer Bäume.** Die Evolution verändert über die Jahre hinweg die Kodierung der Proteine in der DNS. Modelle dieser Veränderung werden auf die Daten von Sequenz-Datenbanken angewandt, um die Stammbäume, phylogenetische Bäume genannt, einzelner Organsimen zu ermitteln.

**Genexpressionanalyse.** Ein Gen wird meist als DNS-Abschnitt definiert, der ein Protein kodiert. Zellen besitzen Mechanismen für die sogenannte Genexpression, d.h. um ein spezielles Gen in der DNS zu lesen und daraus das kodierte Protein zu synthetisieren. Mit sogenannten DNS-Chips kann das Expressionsniveau (intuitiv: die "Konzentration") mehrerer tausend Gene gemessen werden, die eine Zelle zu einem Zeitpunkt exprimiert. Datenbanktechniken zum Data-Mining (z.B. Clustering-Methoden) werden dazu benutzt, Gene mit ähnlichen Expressionsmustern in Gruppen zusammen zu fassen.

Die Bioinformatik-Datenbanken sind sehr unterschiedlich bezüglich ihres Inhaltes. Manche Datenbanken speichern die Daten aus einem einzigen, möglicherweise schon abgeschlossenen Forschungsprojekt. Andere Datenbanken sind das Ergebnis einer weltweiten und andauernden Zusammenarbeit zwischen Forschungsteams. Manche Datenbanken beinhalten Daten über einen einzigen Organismus, andere über alle Vorkommen eines Proteins in allen möglichen Organismen. In manchen Datenbanken werden neue Daten erst nach Korrektheit- und Konsistenzüberprüfungen aufgenommen. In anderen Datenbanken finden solche Überprüfungen nicht statt. Dennoch haben die Bioinformatik-Datenbanken erstaunlich viele Gemeinsamkeiten, was Datenmodellierung und -management angeht. [3] gibt einen Überblick über die verbreitesten Datenbanken der Bioinformatik.

## Datenmodellierung und -management

Bemerkenswert ist, dass Datenmodellierung und -management kaum von der Art der Molekularbiologie-Daten abhängen.

**Datenmodelle.** Bioinformatik-Datenbanken verwenden vier Formen der Datenmodellierung:

- ASCII-Texte, Flat-Files genannt;
- Datenmodelle, die für herkömmliche Datenbanken entwickelt wurden;
- das Object-Protocol Model (OPM);
- das ACEDB-Datenmodell.

*Flat-Files.* Die Datensätze aus Bioinformatik-Datenbanken, die als Sammlungen von Flat-Files implementiert sind, sind entweder unstrukturiert (Abbildung 1) oder mittels textueller Bezeichner, line type genannt, stukturiert (Abbildung 2). Es gibt keinen einheitlichen Satz an solchen Bezeichnern, der in den meisten (oder vielen) Bioinformatik-Datenbanken verwendet wird. Sowohl die Kodierung von Begriffen mit "line types" wie auch die Begriffe selbst können sich zwischen Bioinformatik-Datenbanken erheblich unterscheiden. Sehr viele Bioinformatik-Datenbanken sind immer noch als

Sammlungen von Flat-Files implementiert: ca. 40% der in [3] untersuchten Bioinformatik-Datenbanken sind Flat-File-Sammlungen. Zudem sind Flat-Files der *de facto* Standard zum Datenaustausch in der Bioinformatik.

```
>H3PM3|H3PM3 histone H3 - garden pea. [ Pisum sativum ]|peaH3
ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRFRPGTVALREIRKYQKSTEL
LIRKLPFQRLVREIAQDFKTDLRFQSSAVSALQEAAEAYLVGLFEDTNLCAIHAKRVTIM
PKDIQLARRIRGERA
```

Abbildung 1: Auszug aus der Bioinformatik-Datenbank HDB

```
ID   PILI_PSEAE     STANDARD;      PRT;   178 AA.
AC   P43502;

DR   EMBL; L22036; AAA25951.1; -.
DR   INTERPRO; IPR002545; -.
DR   PFAM; PF01584; CheW; 1.
KW   Sensory transduction.
SQ   SEQUENCE   178 AA;  19934 MW;  634A1A4B135A7E77 CRC64;
     MSDVQTPFQL LVDIDQRCRR LAAGLPAQQE AVQSWSGIGF RMGGRFFVAP MGEVGEVLME
     PRYTQLPGVK TWVKGVANVR GRLLPIMDLC GFLGTELSPL RKQRRVLVVE MLIVFAGLIV
     DEVFGMQMFP VDTFSEQLPP LEAALQPFIM GVFMREQPWL VFSPMALAQM QGFLDVAV
```

Abbildung 2: Auszug aus der Bioinformatik-Datenbank SWISS-PROT

*Relationale und Objekt-Datenmodelle.* Etwa 35% der in [3] untersuchten Bioinformatik-Datenbanken werden von einem herkömmlichen (relationalen, objekt-orientierten oder objekt-relationalen) Datenbanksystem verwaltet. Ihre Daten werden folglich unter Verwendung des relationalen oder des Objekt-Datenmodels repräsentiert. Mit dem Objekt-Datenmodel werden die (meist weitgehend strukturierten) Molekularbiologie-Daten gut repräsentiert. Mit dem relationalen Modell werden die Daten meist unübersichtlich und wenig intuitiv repräsentiert.

*Object-Protocol Model (OPM).* OPM [6] wurde zur Repräsentation von wissenschaftlichen (Labor-) Experimenten entwickelt. Es eignet sich besonders gut zur Repräsentation der zeitlichen Bedingungen und des Datenflußes zwischen Teilexperimenten. Folglich eignet sich OPM gut zur Repräsentation von Phenotyp-Daten, d.h. Daten über die Dynamik von biologischen Prozessen. OPM weist viele Merkmale von SDM [7] und dem Datenmodell von $O_2$ [5] auf.

*ACEDB.* ACEDB (mit "E") [1] ist ein Datenbank-Managementsystem mit einem eigenen, speziellen Datenmodell, welches ursprünglich für die Bioinformatik-Datenbank ACeDB (mit "e") entwickelt wurde (ACeDB ist das Kürzel von "A *C. elegans* Database"). ACEDB findet in den Bioinformatik-Datenbanken breite Anwendung zur Repräsentation von genetischen Daten. Grund dafür ist die Flexibilität des Datenmodells, welches viele Aspekte des semistrukturierten Ansatzes zur Datenmodellierung [4] besitzt.

**Datenbankmanagement.** Viele (über 30% der in [3] untersuchten) Bioinformatik-Datenbanken werden mit einem relationalen Datenbank-Management-System (DBMS) verwaltet, obwohl das relationale Datenmodell zur Repräsentation von molekularbiologischen Daten wenig geeignet ist. Nur

wenige (ca. 9% der in [3] untersuchten) Bioinformatik-Datenbanken werden mit einem objekt-orientierten DBMS verwaltet, obwohl die objekt-orientierte Modellierung von molekularbiologischen Daten sehr passend ist. Diese Lage ist sicherlich auch auf die rasche Entwicklung der Bioinformatik, auf das extrem schnelle Wachstum der Bioinformatik-Datenbanken sowie auf den beschränkten Erfolg der objekt-orientierten DBMS zurückzuführen. Das speziell für die Bioinformatik entwickelte DBMS ACDB ([3]) wird zur Verwaltung von noch wenigen (ca. 4% der in [3] untersuchten) Bioinformatik-Datenbanken eingesetzt.

**Querverweise.** Oft verweisen Datensätze von Bioinformatik-Datenbanken auf Beschreibungen der Experimente, durch die die Daten gewonnen wurden und/oder auf ähnliche Daten in derselben Datenbank oder in anderen Bioinformatik-Datenbanken. Meist werden diese Verweise mittels (künstlicher) Primärschlüssel realisiert und als Hypertext-Links implementiert. Die Hypertext-Verlinkung ist ein besonders auffälliges Merkmal der Bioinformatik-Datenbanken.

**Anfragen und Crawling.** Die meisten Bioinformatik-Datenbanken bieten (oft hierarchisch organisierte) Web-Formulare, womit Anfragen an die Datenbank gestellt werden können. Solche Schnittstellen sind leicht zu benutzen, aber ermöglichen meist nur begrenzte Anfragen. Anfragesprachen im herkömmlichen Sinn werden selten zur Verfügung gestellt, u.a. weil ihre Verwendung Fachkenntnisse erfordern, über die nur wenige Benutzer von Bioinformatik-Datenbanken verfügen. Zusätzlich stellen fast alle Bioinformatik-Datenbanken ihre Daten in den unterschiedlichsten Formaten als Flat-Files zum Herunterladen zur Verfügung. Das System SRS zur Integration von Bioinformatik-Datenbanken bietet eine originelle Anfragesprache. Mit dieser Anfragesprache kann eine Navigation durch Datenbanken und Datensätzen spezifiziert werden. Auffällig originell sind die Crawling-Konstrukte der SRS-Anfragesprache zur Verfolgung von Hypertext-Links. Interessanterweise bietet keine Anfragesprache für XML [4] solche Crawling-Konstrukte.

## Datenanalyse und -integration

Auch bei der Datenanalyse gibt es überwiegend Gemeinsamkeiten zwischen den Bioinformatik-Datenbanken, unabhängig von der Art der gespeicherten Daten. Viele Werkzeuge zur Datenanalyse sind zwar tendenziell eher einer gewissen Art von Daten zu zuordnen, dennoch werden sie oft auch für Daten anderer Art angeboten und benutzt.

**Datenbanken, Datenanalyse und Data Mining.** Die meisten Bioinformatik-Datenbanken stellen Bioinformatik-Software für die Datenanalyse zur Verfügung. Diese Software sind entweder Implementierungen von verbreiteten Bioinformatik-Algorithmen (z.B. dem Smith-Watermann-Algorithmus) oder Werkzeuge (z.B. BLAST), die auf bekannten und/oder weniger bekannten Algorithmen und Verfahren beruhen. Einige dieser Werkzeuge (z.B. 3Dee) beziehen sich auf Datenbanken, so dass die Unterscheidung zwischen einer Methode zur Datenanalyse, die eine bestimmte Datenbank verwendet, und einer Datenbank, die eine bestimmte Methode zur Datenanalyse anbietet, schwierig sein kann. Viele Bioinformatik-Datenbanken bieten auch elementare Computer-Linguistik-Software zur Schlagwortsuche und Übersetzungen zwischen den geläufigsten Datenformaten von Bioinformatik-Datenbanken. [3] gibt einen Überblick über die verbreitesten Methoden und Werkzeuge zur Datenanalyse, die mit Bioinformatik-Datenbanken verwendet werden. Mit dem raschen und stetigen Wachstum

der Bioinformatik-Datenbanken sind Verfahren zum Knowledge Discovery und Data Mining unabdingbar geworden. Sie sind Gegenstand aktueller Forschung.

**Datenintegration.** Die Integration von Daten unterschiedlicher Art und/oder aus unterschiedlichen Bioinformatik-Datenbanken ist ein akutes Problem. Es treten semantische Konflikte auf: z.B. werden grundlegende Begriffe wie "Gen" in verschiedenen Datenbanken unterschiedlich ausgelegt. Frühe Systeme zur Datenintegration in der Bioinformatik (wie BioKleisli und SRS) berücksichtigen semantische Konflikte nicht. Neuere Ansätze (wie Tambis) versuchen semantische Konflikte meist mit Ontologien zu lösen. Das Problem, dass Daten unterschiedlicher (insbesondere auch sehr schlechter) Qualität integriert werden, ist bisher von keinem Integrationsansatz zufriedenstellend behandelt worden.

# Literatur


[1] ACEDB Dokumentationssammlung: http://genome.cornell.edu/acedocs/

[2] R. Backofen, F. Bry, P. Clote, H.-P. Kriegel, T. Seidl, K. Schulz. *Aktuelles Schlagwort Bioinformatik*. Informatik Spektrum, Vol. 22, No. 9, pp. 376-378, October 1999. Auch im Informatik-Lexikon der GI: http://www.gi-ev.de/informatik/lexikon/inf-lex-bioinformatik.shtml

[3] F. Bry, P. Kröger. *A Computational Biology Database Digest: Data, Data Analysis, and Data Management*. Research Report PMS-FB-2002-8, Institut für Informatik, Universität München, http://www.pms.informatik.uni-muenchen.de/publikationen/#PMS-FB-2002-8

[4] F. Bry, M. Kraus, D. Olteanu, S. Schaffert. *Aktuelles Schlagwort Semistrukturierte Daten*. Informatik Spektrum, Vol. 24, No. 4, pp. 230-233, August 2001. Auch im Informatik-Lexikon der GI: http://www.gi-ev.de/informatik/lexikon/inf-lex-semistrukturierte-daten.shtml

[5] F. Bancilhon, C. Delobel, P. Kanellakis. *Building an Object-Oriented Database System: The Story of $O_2$*. Morgan Kaufmann, 1992.
ISBN 1-55860-169-4.

[6] I.-M. Chen, V. Markowitz. *An Overview of the Object Protocol Model (OPM) and the OPM Data Management Tools*. Information Systems, 1995, Vol. 20, No. 5, pp. 393-418.

[7] M. Hammer, D. McLeod. *Database Description with SDM: A Semantic Database Model*. ACM Transactions on Database Systems, Vol. 6, No. 3, 1981.