# Bioinformatik

Sandra Held

Otto-von-Guericke Universität Magdeburg
sandra.held@st.ovgu.de

**Zusammenfassung.** Die Bioinformatik hat sich als interdisziplinäre Wissenschaft, welche Mathematik, Informatik und Molekularbiologie miteinander verknüpft, fest etabliert. Unter Verwendung von mathematischen und informatischen Techniken werden biologische Daten wie DNA- und Proteinsequenzen organisiert und analysiert. Datenbanken bilden die Grundlage der Forschung im Bereich Bioinformatik und der Datenbestand wächst exponentiell. Des Weiteren bedingen die speziellen Anforderungen der Bioinformatik Algorithmen zur Analyse von biologischen Daten, bspw. um Gene zu identifizieren und Vorhersagen bezüglich Struktur und Funktion von Proteinen bis hin zu Genomen treffen zu können.

## 1 Einleitung

Die Bioinformatik ist eine interdisziplinäre Wissenschaft, die mit Hilfe von Mathematik und Informatik molekularbiologische Fragen zu lösen versucht. Von besonderem Interesse sind evolutionäre Verwandtschaften von Organismen, die evolutionäre Strukturentwicklung von Proteinen sowie das Erkennen von Zusammenhängen zwischen Sequenzen, deren Struktur und ihren Funktionen. Außerdem versucht man die Regulation von Genverbänden und Genomen zu verstehen [11]. Es ist eine recht junge Wissenschaft, deren Anfänge in den 1960er Jahren liegen. In dieser Zeit wurden die Sammlungen von bekannten Aminosäuresequenzen rapide größer. Man gewann zunehmend die Überzeugung, dass Makromoleküle Informationen tragen und schnelle Computer wurden im Bereich der biologischen Forschung weithin verfügbar [4]. 1977 wurde der Begriff Bioinformatik dann erstmalig von der dänischen Forscherin Paulien Hogeweg verwendet [2]. Durch Verbesserung von Technik und Steigerung des Verständnis molekularbiologischen Strukturen wuchs der Datenbestand exponentiell. In den 1960er Jahren war es noch möglich alle bekannte Proteinsequenzen in Fachzeitschriften abzudrucken. Nur ein Jahrzehnt später war daran nicht mehr zu denken.
Zunächst wird auf die Datengrundlage eingegangen, anschließend Analysemethoden sowie Herausforderungen der Bioinformatik aufgezeigt und ein Ausblick auf zukünftige Fragestellungen gegeben.

## 2 Datenbanken

Datenbanken als Sammlung und Ordnung biologischer und biochemischer Daten bilden die Grundlage der biologischen Forschung. Die Datenmenge wuchs und

wächst exponentiell, so dass 1983 die erste Onlinedatenbank für Proteinsequenzen (PIR)[1] eingeführt wurde [8]. Heute gibt es eine Vielzahl von Datenbanken für Sequenzen, Struktur und Information, Funktion sowie Metadatenbanken. Eine Übersicht über einige Datenbanken und ihre Zusammenhänge liefert Abbildung 1.

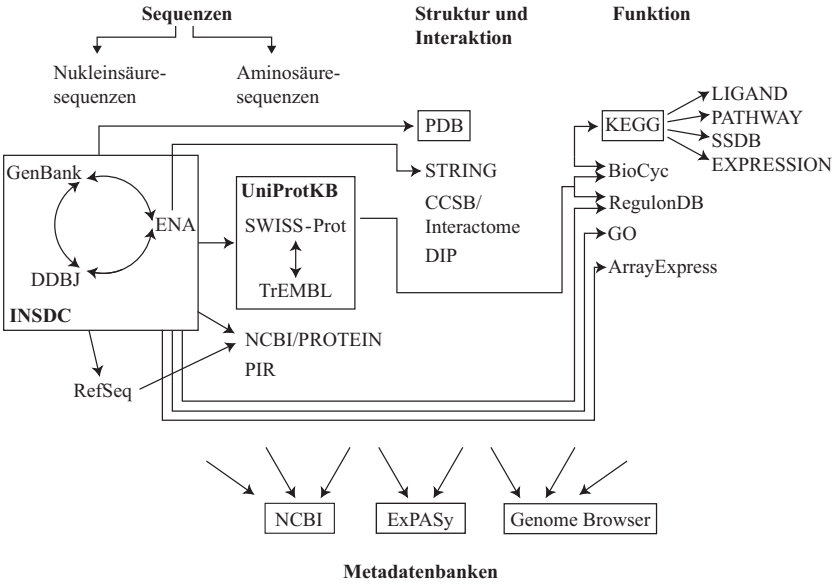


**Abb. 1.** Übersicht über existierende Datenbanken sortiert nach Sequenzen, Struktur und Information, Funktion sowie Metadatenbanken. Abbildung entnommen aus [6].

Sequenzdatenbanken enthalten u.a. DNA-, RNA- und Proteinsequenzen. Drei große Datenbanken aus diesem Bereich sind GenBank[2], European Nucleotide Archive (ENA)[3] und DNA Data Bank of Japan (DDBJ)[4]. Inzwischen sind sie vernetzt und tauschen regelmäßig Informationen untereinander aus. Allein die ENA Datenbank enthält 746.3 Millionen Sequenzen und 1,782.6 Milliarden Basen[5].

[1] `http://pir.georgetown.edu/`
[2] `http://www.ncbi.nlm.nih.gov/genbank/`
[3] `https://www.ebi.ac.uk/ena`
[4] `http://www.ddbj.nig.ac.jp/`
[5] `http://www.ebi.ac.uk/ena/about/statistics`, Stand 16.06.16

In die Bereiche Struktur und Information fallen Sammlungen komplett sequenzierter Genome sowie 3D-Struktur-Datenbanken. Als Beispiel ist die Datenbank PDB[6] zu nennen, die Strukturen von Proteinen, Nukleinsäuren und Proteinkomplexen enthält [8]. Eine neuere Datenbank ist Pfam[7]. Sie enthält Proteinfamilien, für deren Bestimmung 3D-Struktur-Datenbanken die Grundlage bilden.
Datenbanken für Funktionen enthalten Informationen zu Stoffwechselwegen. KEGG[8] enthält bspw. von allen zellulären Prozessen, u.a. dem Membrantransport, graphische Präsentationen [5].

## 3 Analyse biologischer Daten

Die unterschiedlichen Datentypen haben jeweils eigene Fragestellungen, auf die getrennt eingegangen wird. Grundlage der Algorithmen, die auf molekularbiologischen Daten arbeiten, ist die Umwandlung der Sequenzen in Zeichenketten und Arrays.

### 3.1 Rohsequenzen von DNA

DNA Rohsequenzen werden untersucht, um Genvorhersagen treffen zu können und Gene von Pseudogenen zu unterscheiden [7]. Dafür ist es entscheidend zu wissen, welche Regionen der Sequenz codiert bzw. nicht codiert sind. Die codierten Bereiche nennt man Exons. Sie werden in die fertige Proteinsequenz übersetzt und ergeben somit echte Gene. Introns sind nicht codierte Abschnitte, die vor der Translation aus der RNA-Sequenz entfernt werden. Somit sind sie nicht in der Proteinsequenz zu finden und werden daher als Pseudogene bezeichnet. Ein wichtiger Algorithmus zur Identifikation von Genen durch Strukturvorhersage ist Genscan [6].

### 3.2 Proteinsequenzen

Der Fokus in diesem Bereich liegt auf der Bestimmung von Sequenzähnlichkeiten. Bewertet wird die statistisch signifikante Ähnlichkeit zweier Sequenzen bis hin zur evolutionären Verwandtschaft.
Eine Methode zur Bewertung von Ähnlichkeiten ist die Bestimmung von Alignments. Dabei werden übereinstimmende Sequenzelemente (Basen) beider Sequenzen gesucht und deren Position bestimmt. Die einfachste Art die Übereinstimmung von Elementen zu bestimmen ist beide als Zeichenketten untereinander zu setzen und ähnliche Elemente zu verbinden. Dies ist in Abbildung 2 veranschaulicht. Das Ähnlichkeitsmaß wird dabei über eine Matrix festgesetzt. Außerdem können in Alignments Lücken auftreten, sogenannte Gaps. Sie entstehen durch Einfügen bzw. Löschen von Positionen innerhalb einer Sequenz auf DNA-Ebene, was durch Mutationen verursacht wird [6].

[6] `http://www.rcsb.org/pdb/home/home.do`
[7] `http://pfam.xfam.org/`
[8] `http://www.genome.jp/kegg/pathway.html`

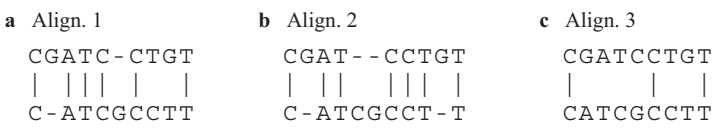


**Abb. 2.** Beispiele für paarweise Sequenzalignments. Übereinstimmungen zwischen den Paaren werden durch Verbindungslinien markiert, Gaps durch Bindestriche innerhalb der einzelnen Sequenz. Abbildung entnommen aus [6].

Ist die Ähnlichkeit zweier Sequenzen hoch genug, erlaubt dies zuverlässige Prognosen bezüglich der Struktur des Proteins sowie Indizien für seine Funktion. Treten Gaps auf, sind gegebenenfalls Rückschlüsse auf evolutionäre Beziehungen der verglichenen Sequenzen möglich, da diese durch Mutationen verursacht werden. Paarweise Alignments bilden zudem die Grundlage der Algorithmen BLAST [1] und FASTA [9], die Datenbanken nach ähnlichen Sequenzen durchsuchen.
Neben den paarweisen Alignments gibt es auch die Möglichkeit multiple Alignments zu bestimmen, welche genauere Informationen liefern, wie Aminosäuren an einzelnen Positionen verteilt sind.
Um letztendlich evolutionäre Verwandtschaften und Beziehungen von Organismen untersuchen zu können müssen phylogenetische Untersuchungen folgen. Dazu werden Bäume bzw. Netze aus multiplen Alginments berechnet. Vergleicht man bestimmte Eigenschaften, so kann man über diese Bäume bestimmen, ob gemeinsame Vorfahren vorliegen, also ob sie homolog sind.

### 3.3 Strukturdaten von Proteinen und anderen Makromolekülen

Der reine Vergleich von Sequenzdaten ist hilfreich, lässt aber kaum Rückschlüsse auf die Funktionsweise des Proteins zu. Daher wird zusätzlich der genau Aufbau von Sekundärstruktur und Tertiärstruktur von Proteinen ermittelt. Deren Untersuchung geht mit Proteingeometrie einher, bei der u.a. Winkelmessung sowie Berechnungen von Oberflächen und Volumen eine Rolle spielen. Auf diese Weise lassen sich Rückschlüsse über die nötige Energie zur Stabilisierung von Makromolekülen ziehen. Außerdem kann man mit Hilfe von Strukturdaten Interaktionen zwischen Proteinen und Untereinheiten, wie bspw. DNA oder RNA, erkennen [7].

### 3.4 Genome

Die Gesamtheit aller Gene bezeichnet man als Genom eines Organismus. Genome sind dementsprechend die Datenträger eines Organismus, die als Einheit die meisten Informationen am Stück enthalten. Aus informatischer Sicht ist ihre Analyse am komplexesten [10], da für die Untersuchung die Komplettsequenzierung der DNA grundlegend ist. Untersucht werden Genome wiederum auf ihre Struktur und biologische Funktionen von einzelnen Genomabschnitten, wobei es je nach

Organismus signifikant Unterschiede gibt [6]. Es gibt verschiedene Projekte, die sich mit der Genombestimmung beschäftigen, u.a. mit dem menschlichen. Dazu zählen das 1000 Genomes Project und ENCODE.
Das 1000 Genomes Project hat es sich zur Aufgabe gemacht Variationen der Sequenzen im menschlichen Genoms zu katalogisieren [12]. Gleichzeitig wird an der Entwicklung bioinformatischen Analysetechniken gearbeitet. Variationen zu bestimmen
Das zweites bedeutende Projekt in diesem Bereich ist ENCODE (Encyclopedia of DNA Elements) [13], dessen Ziel es ist alle funktionellen Elemente, die das menschliche Genom codiert, zu finden und zu bestimmen.

## 4 Herausforderungen der Bioinformatik

Molekularbiologische Daten werden experimentell gewonnen und sind damit nicht exakt, da eine Reihe Fehlerquellen während der Gewinnung auftreten können [3]. Bei der Verarbeitung der Daten muss dieser Umstand eingerechnet werden.
Die Natur bringt durch Evolution zuverlässige Optimierungen hervor. Momentan erreichen die meisten in der Bioinformatik eingesetzten Algorithmen 80% echt positive Vorhersagen [8]. Die Optimierung zur Vorhersagekraft von bioinformatischen Algorithmen gestaltet sich jedoch schwierig. Denn Optimierung durch Evolution funktioniert durch gleichzeitige Anpassung mehrerer Parameter. Die Komplexität dieser Vorgänge bedingt, dass es schwierig ist die Rahmenbedingungen und Optimierungskriterien zu einer eindeutigen Zielfunktion zusammenzufassen, die algorithmisch umsetzbar ist.
Ebenfalls problematisch ist, dass in der Natur Eigenschaften nur bis zu dem Punkt optimiert werden, der gerade ausreicht, um das Überleben zu sichern. Es setzt sich in der Biologie also nicht zwangsläufig eine optimale Lösung durch. Aus informatischer Sicht bedeutet Optimierung jedoch immer das Finden einer optimalen Lösung. Jede algorithmische Lösung ist also prinzipiell eine Hypothese, deren biologische Relevanz zwingend durch Experimente verifiziert werden muss [3].

## 5 Ausblick

Die Bioinformatik sieht sich mit einer enormen, stetig wachsenden Datenmenge konfrontiert. Allein das Sequenzierzentrum Beijing Genomics Institute (BGI) generiert täglich 6 Tbyte Daten. Einerseits ermöglicht diese Entwicklung eine Vielzahl neuer und deutlich komplexerer Forschungsergebnisse, anderseits birgt sie auch viele Herausforderungen. Es wird eine enorme Speicherkapazität gebraucht und die Rechenleistung muss entsprechend ausgebaut sein. Algorithmen, die auf diesen großen Datensätzen arbeiten, müssen zwingend eine geringe Zeitkomplexität aufweisen [8].
Zwar herrscht eine große Fülle an Daten, jedoch sind diese über verschiedenste Ressourcen verteilt, die zumeist keine gemeinsame Schnittstelle haben. Mit Hilfe

integrierter Systeme sollen diese zukünftig vernetzt werden und über eine einheitliche Bedienoberfläche erreichbar sein [6].
Aus biologischer Sicht wird es in Zukunft vor allem darum gehen komplexe Zusammenhänge, zum Beispiel die Funktionsweise einer Zelle, zu verstehen.

## Literatur


1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403 – 410 (1990)
2. Attwood, T., Gisel, A., Bongcam-Rudloff, E., Eriksson, N.: Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective. INTECH Open Access Publisher (2011)
3. Böckenhauer, H.J., Bongartz, D.: Algorithmische Grundlagen der Bioinformatik: Modelle, Methoden und Komplexität. Springer-Verlag (2013)
4. Hagen, J.B.: The origins of bioinformatics. Nat Rev Genet 1(3), 231–236 (12 2000)
5. Hansen, A.: Bioinformatik: Ein Leitfaden für Naturwissenschaftler. Springer-Verlag (2013)
6. Hütt, M.T., Dehnert, M.: Methoden der Bioinformatik: Eine Einführung zur Anwendung in Biologie und Medizin. Springer-Verlag (2015)
7. Luscombe, N.M., Greenbaum, D., Gerstein, M., et al.: What is bioinformatics? a proposed definition and overview of the field. Methods of information in medicine 40(4), 346–358 (2001)
8. Merkl, R.: Bioinformatik: Grundlagen, Algorithmen, Anwendungen. John Wiley & Sons (2015)
9. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences 85(8), 2444–2448 (1988)
10. Polanski, A., Kimmel, M.: Bioinformatics. Springer Berlin Heidelberg (2007)
11. Rauhut, R.: Bioinformatik. John Wiley & Sons (2001)
12. The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature 467(7319), 1061–1073 (10 2010)
13. The ENCODE Project Consortium: A user's guide to the encyclopedia of dna elements (encode). PLoS Biol 9(4), 1–21 (04 2011)