



Grafik: GBF  
Foto: van den Heuvel

## Bioinformatik

# Werkzeug und Schlüssel zur Genomanalyse

Hans-Werner Mewes

Ohne leistungsfähige Rechenanlagen wäre die biochemische Forschung heute längst nicht mehr denkbar. Computersysteme helfen, experimentelle Daten auszuwerten: Bei der Sequenzierung von Nukleinsäuren und Proteinen können aus Tausenden von Fragmenten die richtigen Abfolgen der Bausteine errechnet werden, Gen-Datenbanken erlauben wie beim Human-Genom-Projekt den schnellen, weltweiten Zugriff auf Informationen. Der Artikel zeigt die Möglichkeiten und Grenzen dieser Disziplin, der Bioinformatik, auf.

**D**ie DNA höherer Organismen besteht aus vielen Millionen oder auch einigen Milliarden von Basenpaaren, die kontrolliert abgelesen werden. Damit ergibt sich ein Informationsgehalt, der für das menschliche Auffassungsvermögen weder begreifbar noch erfassbar ist. Aufgabe der Bioinformatik ist es, Struktur in die Datenflut zu bringen und die gewonnenen Informationen zu interpretieren.

Die Meinung darüber, was Bioinformatik bedeutet, geht selbst unter den Fachleuten auseinander. Eine Definition der Bioinformatik, die ihre pragmatischen Aspekte in den Vordergrund stellt, bezeichnet die Bioinformatik als Disziplin, die mit den Methoden der Informatik, Statistik und Mathematik versucht, Fragen der molekularen Biologie zu beantworten. Insbesondere die zunehmende Entschlüsselung der DNA-Sequenzen verschiedenster Organismen und das Erforschen der Genfunktionen führen zu einer Flut an Daten, die sinnvoll geordnet, analysiert und gespeichert werden müssen. Dabei dient die Bioinformatik zum einen als Werkzeug, um aus Daten Informationen zu machen. Zum anderen haben Bioinformatiker die Aufgabe, die Methoden zu entwickeln, die für die Bearbeitung und Interpretation dieser Daten notwendig sind.

## Ohne leistungsfähige Rechner geht nichts

Einer der großen Zufälle in der Geschichte der Naturwissenschaften ist die Parallelität der Entwicklung der Molekularbiologie und der Computertechnologie. Ohne leistungsfähige Rechner ist eine systematische Analyse der Genominformationen nicht möglich. Allein der Vergleich von tausend Proteinsequenzen untereinander erfordert die Lösung einer halben Million von Optimierungsaufgaben und deren statistische Bewertung. Derzeit gibt es bereits über eine Million Proteinsequenzen in den Datenbanken, dennoch ist die Berechnung aller Vergleiche mit den verfügbaren Rechnern durchführbar.

Die Interpretation dieser Datenflut ist vergleichbar mit der (Re-) Konstruktion eines Autos aus der – auch noch mit schweren Fehlern behafteten – ungeordneten Liste seiner Einzelteile. Um das Fahrzeug beziehungsweise das Genom verstehen zu können, müssen wir zusätzliche Informationen über die Struktur der Teile, ihre Interaktionen und ihre funktionellen Eigenschaften besitzen. In lebenden Systemen wird dies noch zusätzlich verkompliziert durch Veränderungen während des zellulären Lebens-

zyklus. Im Gegensatz zum Auto, das zwar altert, aber die Summe seiner Einzelteile und ihre Funktion normalerweise konstant hält, unterliegt die Zelle einem ständigen Wandel. Die Summe der mit der zellulären Information verbundenen bekannten Informationen können wir als „Biologisches Wissen“ bezeichnen. Die Darstellung des biologischen Wissens in systematischer, strukturierter Form ist eine der wichtigsten Herausforderungen in der Bioinformatik.

## Dienstleistung und Kreativität gefragt

Der Erfolg der Bioinformatik ist auch durch den hohen praktischen Nutzen bedingt, den sie für den experimentell arbeitenden Biologen hat. Die Bioinformatik stellt Datenbanken bereit und verknüpft diese mit Anwendungen, die auf die Forschungsprojekte der einzelnen Biologen zugeschnitten sind.

Die in den molekularbiologischen Datenbanken verfügbaren strukturierten Informationen werden täglich millionenfach genutzt. Die weltweit am häufigsten genutzte Ressource ist das National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, das nicht nur Gensequenzen, sondern auch die gesamte Literatur der Biowissenschaften verwaltet und über die Datenbank „Medline“ zugänglich macht.

Zu dieser notwendigen, oft ungeliebten Infrastrukturfunktion kommt die wissenschaftliche, kreative Aufgabe der Bioinformatik dazu. Erkenntnisse durch die systematische Anwendung der Algorithmen auf große Datenmengen zu gewinnen, oder auch neue Algorithmen zu entwickeln, ist selbstverständlich die Lieblingsbeschäftigung der Bioinformatiker. Im Englischen wird diese Forschung auch gern als „computational biology“ im Gegensatz zur experimentellen Biologie bezeichnet.



Das Buch des Lebens ist in gedruckter Form nicht lesbar, es verschließt sich der Interpretation. Zuallererst stellt sich ein quantitatives Problem: Während der Text des zweiten Teils des „Faust“ in 300 Kilobyte abgespeichert werden kann, ist das Hefegenom 40 mal, das Humangenom etwa 9000 mal so umfangreich.

Foto: van den Heuvel

# Werkzeug und Schlüssel zur Genomanalyse

## Experiment und Modell in Einklang bringen

Um Einblick in biologische Zusammenhänge zu gewinnen, werden die Werkzeuge der Bioinformatik auf experimentelle Daten

### Algorithmus

Ein Algorithmus ist eine Handlungsanweisung, die bei genauer Anwendung nach einer endlichen Anzahl von Schritten mit Sicherheit zur Lösung einer Aufgabe führt. Meist werden Algorithmen in Form eines Computerprogramms angewendet und dienen zur Lösung mathematischer Probleme.

angewendet. Bei der Sequenzierung von Genomen beispielsweise können immer nur Fragmente mit einer Länge von etwa 500 Basen analysiert werden. Dadurch erhält man einzelne Teile eines eindimensionalen Puzzles, die in der richtigen Reihenfolge miteinander verknüpft werden müssen, um letztlich die Basenabfolge des vollständigen Genoms zu rekonstruieren. Dann gilt es, die einzelnen

**Um eine möglichst große Zahl von Proben in möglichst kurzer Zeit zu analysieren, werden viele Arbeitsschritte heute automatisiert. Dabei können mehrere Proben synchron bearbeitet werden.**

Fotos: GSF/GAZ



Gene voneinander abzugrenzen und regulatorische Bereiche zu identifizieren. Dabei helfen Computerprogramme wie zum Beispiel MatInspector, mit dem die Bindungsstellen von Transkriptionsfaktoren erkannt werden können. Transkriptionsfaktoren sind bestimmte Proteine, deren spezifische Eigenschaften die Aktivität des Gens durch die Kontrolle seiner Synthese regulieren.

Ein weiteres aktuelles Beispiel für die große Bedeutung der Bioinformatik zur Interpretation experimenteller Daten ist die Expressionsanalyse. Im Idealfall kann durch die Expressionsanalyse für jedes Gen ermittelt werden, wie aktiv es ist, indem die individuellen mRNAs quantifiziert werden. Vor allem die zeitliche Abfolge der mRNA-Synthese oder der Vergleich von gesundem und krankem Gewebe sind interessant, denn sie erlauben wichtige Aufschlüsse über dynamische Abläufe in der Zelle. Da viele Gene gleichzeitig betrachtet werden, können insbesondere die Beziehungen zwischen Gruppen von Genen besser erfasst werden. In der Regel liefert eine Expressionsanalyse mehrere tausend Messpunkte. Ohne die Methoden der Bioinformatik wäre es unmöglich, Ordnung in diese gewaltige Datenflut zu bringen und die Expressionsprofile zu analysieren und zu interpretieren.

Experimentelle Daten dienen als Grundlage zur Entwicklung neuer Algorithmen. Die Genvorhersage in eukaryotischen Genomen liefert dafür ein Beispiel: Liegt die Sequenz eines Gens vor, müssen zunächst die funktionellen Abschnitte der DNA-Sequenz identifiziert werden. Die Sequenz allein reicht zur Entschlüsselung der biologischen Zusammenhänge nicht aus, nicht ein-



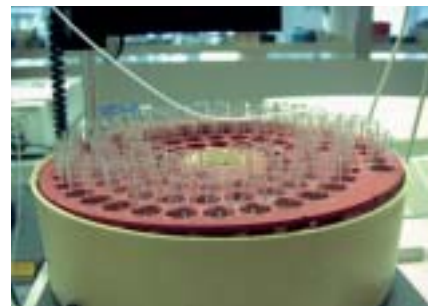
mal die Genvorhersage ist perfekt, denn höhere Eukaryoten verfügen über komplexe Mechanismen, die primär transkribierten RNA-Sequenzen weiter zu verarbeiten und dabei bestimmte Abschnitte der RNA herauszuschneiden. Die Struktur des von dieser Gensequenz letztlich kodierten Proteins kann nur dann korrekt vorhergesagt werden, wenn es gelingt, diesen Prozess detailgenau zu modellieren. Der Bioinformatiker hat die Aufgabe, den besten Algorithmus für eine möglichst zuverlässige Vorhersage zu entwickeln.

Damit kommt man schnell zum Dilemma und zur Herausforderung der Bioinformatik: Solange der Prozess nicht perfekt modelliert werden kann, ist es erforderlich, die Ergebnisse der Modellierung experimentell zu überprüfen und die verwendeten Algorithmen manuell zu verbessern. Häufig ist eine perfekte Modellierung gar nicht möglich, wenn es sich um Systeme mit einer großen Zahl von Parametern handelt. Die Korrekturen können nun wieder dem Theoretiker Anhaltspunkte für die Verbesserung der Algorithmen geben.

## Grenzen der Bioinformatik

Die Identifizierung der genetischen Elemente auf der DNA gelingt heutzutage längst nicht perfekt, besonders der komplexe Aufbau regulatorischer Bereiche bereitet große Schwierigkeiten. Die Suche nach verbesserten Identifizierungsmethoden ist eine Suche nach den richtigen Algorithmen, die benötigt werden, um die einzelnen genetischen Elemente zu erkennen.

Noch wesentlich anspruchsvoller wird es, wenn es um die Aufklärung der Funktion korrekt





vorhergesagter Elemente geht. Bewährte funktionelle und strukturelle Eigenschaften blieben während der Evolution oft erhalten, deshalb bedingen ähnliche Sequenzen oft auch ähnliche Funktionen. Die Analyse verwandter Sequenzen gibt zwar Hinweise auf konservierte und damit funktionell wichtige Segmente in DNA und Proteinen, beantwortet aber noch lange nicht die Frage nach den funktionellen Mechanismen. Hier ist die Bioinformatik wesentlich auf assoziative Informationen angewiesen, also der Zuweisung von funktionellen oder strukturellen Eigenschaften der Gene über experimentelle Daten.

Die komplexen Auswirkungen eines defekten Gens auf den Organismus lassen sich nur schwer vorhersagen. Gleiches gilt für Wechselwirkungen zwischen Genen durch Transkriptionskontrolle, für regulatorische Netzwerke und komplexe genetische Erkrankungen. Hier gilt: Die Eigenschaften von Genen, Proteinen oder sogar der Funktion einzelner Aminosäuren können durch die Auswertung experimenteller Daten rekonstruiert werden, der umgekehrte Weg ist jedoch noch verschlossen, das heißt allein aus der Sequenz lassen sich die Genfunktionen in der Regel nicht vorhersagen. Die Vorhersage funktioneller Eigenschaften ist nur dann zuverlässig, wenn evolutionär eng verwandte Moleküle verglichen werden, von denen eines experimentell bewiesene Eigenschaften aufweist.

Man darf wegen der Schwierigkeiten bei der Vorhersage funktioneller Eigenschaften das Potential der Bioinformatik allerdings nicht abwerten, im Gegenteil. Die Tatsache, dass aufgrund der assoziati-

ven Zuweisung, also der Anwendung von Algorithmen und der Nutzung bereits bekannter Sequenz/Struktur/Funktionsbeziehungen zwischen 35 und 70 Prozent aller Gene eines Organismus funktionelle Eigenschaften zugewiesen werden können, beweist die Leistungsfähigkeit der Methoden.

### Unterschiedlichste Disziplinen vernetzen

Mit dem Paradigmenwechsel von der deskriptiven zur systematischen Biologie ist die Bioinformatik zu einem wichtigen Element der Genomforschung geworden. Ihr Erfolg hängt von der Fähigkeit ab, traditionell und methodisch weit auseinander liegende Disziplinen zu verknüpfen. Das durch das Nationale Genomforschungsnetz (NGFN) geförderte Projekt zur Bioinformatik der Genomanalyse von Säugergenomen (BFAM) versucht, die Brücke zu schlagen zwischen der experimentellen Genomforschung, der angewandten Bioinformatik und der durch Infor-

**Ziel der Bioinformatik ist die Vernetzung der Biowissenschaften mit der Datenverarbeitung. Leistungsfähige Rechner steuern Laborroboter und erfassen die Datenflut, wie sie bei der Genomanalyse anfällt.**

*Foto: GSF/GAZ*

matiker und Mathematiker vertretenen Theorie.

Die Bioinformatik wird in Zukunft weiter an Bedeutung gewinnen. Weniger die zentralen Einrichtungen, sondern vielmehr die kooperativen Interaktionen, die Netzwerke, werden die Bioinformatik prägen. Theorie, Methodenentwicklung, technische Implementierung in Form integrierender Werkzeuge werden die experimentellen Methoden der Biologen ergänzen. Der Herausforderung, interdisziplinär denken zu müssen, erhält den Bioinformatikern genauso wie den beteiligten Wissenschaftlern anderer Disziplinen die Freude an ihrer Arbeit.

#### Internettipp:

Münchener Informationszentrum für Proteinsequenzen: <http://mips.gsf.de>



Bei der Analyse von genomischer DNA können immer nur kleine Fragmente mit überlappenden Randbereichen sequenziert werden. Der Computer setzt diese „Puzzle-Teile“ dann zusammen.

*Foto: GBF*