



Univ.-Prof. Mag. Dr. Thomas Rattei

Bioinformatik als Schlüsseldisziplin im Zeitalter der molekularen und systemorientierten Biologie

Bioinformatik ist eine noch recht junge Forschungsrichtung, hinter der sich ein breitgefächertes, anwendungsorientiertes Tätigkeitsfeld verbirgt. In seiner Antrittsvorlesung am Mittwoch, 30. Juni 2010, gab Thomas Rattei, Professor für "In Silico Genomics" und Leiter des neuen Departments für Computational Systems Biology an der Fakultät für Lebenswissenschaften, einen Einblick in die Möglichkeiten und Perspektiven des Forschungsbereichs. Seine Antrittsvorlesung zum Thema "Bioinformatik als Schlüsseldisziplin im Zeitalter der molekularen und systemorientierten Biologie" hat er in einem Gastbeitrag zusammengefasst.

Die Entwicklung von Methoden zur Sequenzierung von DNA hat eine neue Ära in der Biologie eingeleitet, in welcher sowohl Genome als auch die darin kodierten Biomoleküle umfassend auf molekularer Ebene untersucht werden können. Die heutige Verfügbarkeit zahlreicher kompletter Genomsequenzen unterstützt eine Vielzahl biologischer Forschungsdisziplinen. In den letzten Jahren wurden Methoden zur "Metagenomik" etabliert, welche die Sequenzierung von DNA selbst aus nicht aufgereinigten Umweltproben erlauben wie z.B. kompletter mikrobieller Gemeinschaften.

Die Prozesse der Gewinnung genomischer und metagenomischer DNA sowie deren Interpretation sind in hohem Maß auf Computermethoden angewiesen. Die Bioinformatik ist daher in den letzten Jahren zu einer Schlüsseldisziplin der molekularen und systemorientierten Biologie geworden. Als neue Forschungsrichtung ist sie ganz besonders auf eine enge Verbindung zwischen Lehre und Forschung angewiesen. Dieser Entwicklung Rechnung tragend, wurden im März 2010 die Professur für "In silico Genomik" und das neue "Department für Computational Systems Biology" gegründet. In dieser Antrittsvorlesung soll dieses neue Forschungsgebiet vorgestellt und anhand ausgewählter Beispiele und Fragestellungen erläutert werden.

Information in biologischen Sequenzen

Alle Lebewesen bestehen aus kleinen Einheiten, den Zellen. Diese sind entweder eigenständig überlebensfähig oder bilden größere Organismen. Wenn wir ein Mikroskop auf Zellen richten, wird bei hoher Vergrößerung die innere Struktur der Zellen erkennbar. Sowohl in Zellen mit Zellkern (z.B. menschliche Zellen), als auch ohne diesen (z.B. Bakterien-Zellen), finden wir lange Ketten eines besonderen Molekültyps. Die Desoxy-Ribonukleinsäureketten (DNS) bestehen aus vier Bausteinen (genannt A, C, G, T), deren Abfolge in sog. Sequenzen die Erbinformation der Zellen darstellt. Dieses Erbgut wird von den Zellen bei deren Teilung verdoppelt und an die Tochterzellen weitergegeben. Während des Bestehens einer Zelle dient es als beständige "Kopiervorlage" für die Abschrift der funktionellen Abschnitte, die wir als Gene bezeichnen.



Komplexe Gebilde aus verschiedenen Eiweißen und weiteren Molekülen realisieren diese Kopiervorgänge mit hoher Präzision. Die bestehenden Gen-Kopien werden als Transkripte bezeichnet und entweder direkt in der Zelle wirksam oder benötigen eine weitere Verarbeitung. Diese besteht meist in der Übersetzung der Gene in Eiweiße (Proteine), wobei jeweils 3 DNS-Bausteine, die Information für einen der 20 möglichen Eiweißbausteine (die Aminosäuren) liefern. Dieser Übersetzungsvorgang erfordert ebenfalls eine komplexe molekulare Maschine, die selbst aus zahlreichen Eiweißen und anderen Molekülen besteht und als Ribosom bezeichnet wird. Nach der Übersetzung der Gene in Eiweiße falten sich diese und nehmen die verschiedensten Aufgaben in der Zelle wahr, von der Ausbildung stabiler Raumstrukturen bis zum Stoffwechsel.

Das menschliche Genom

Wieviel Information steckt nun im Erbgut (dem Genom) einer Zelle? Verdeutlichen wir uns dies am Beispiel einer menschlichen Zelle, z.B. aus unserer Haut. Die DNS-Moleküle in den verschiedenen Chromosomen enthalten ca. 3 Milliarden Paare von DNS-Bausteinen. Würden wir diese Paare jeweils mit einem Buchstaben bezeichnen (A, C, G oder T) und in typischer Schriftgröße auf Buchseiten drucken, so benötigten wir dafür 1,5 Millionen Seiten. Diese würden ca. 50 große Bucherschranke füllen. Wohlgermerkt, dies ist die Information die das Erbgut einer einzigen menschlichen Zelle repräsentiert. Nicht alle Teile dieser Information sind jedoch einzigartig - es gibt zahlreiche Wiederholungen im menschlichen Genom. Es wäre also möglich, diese wiederholten Abschnitte zu katalogisieren und im Text durch Verweise zu ersetzen. Dies würde die Anzahl benötigter Bücherregale erheblich reduzieren.

Der menschliche Organismus enthält ca. 10.000 Milliarden Zellen, welche alle nahezu identische Erbinformation besitzen. Darüber hinaus ist der menschliche Körper jedoch von ungefähr der gleichen Menge an Bakterienzellen besiedelt. Normalerweise helfen uns diese Bakterien, gesund zu bleiben und unsere Nahrung gut verdauen zu können. Nur selten gerät unsere Bakterienflora aus dem Gleichgewicht und verursacht Erkrankungen.

Die Untersuchung der Bakterien in unserem Organismus steckt zwar noch in den Anfängen, aber aus der Analyse des Erbguts dieser Bakterien wissen wir, dass es sich um zahlreiche verschiedene Arten handelt. Die Information dieser Genome ist viel kompakter als diejenige der menschlichen Zellen und würde beim o.g. Druck nur wenige Bücher benötigen. Aufgrund der Bakterienvielfalt stellt deren Genominformation jedoch eine erhebliche Erweiterung der menschlichen Genomdaten dar. Es liegt auf der Hand, dass diese gewaltigen Informationsmengen nur mit Hilfe von Computern analysiert werden können.

Analyse von Genomsequenzen

Die Ermittlung der Abfolge der DNS-Sequenzen in Genomen bildet den wichtigsten Schritt bei der wissenschaftlichen Analyse der Erbinformation. Sie wird als Sequenzierung bezeichnet und hat sich in den letzten Jahrzehnten revolutionär weiter entwickelt. Typisch für alle derzeit verfügbaren Sequenzierungsmethoden ist es, dass nur jeweils kurze Abschnitte der Genome sequenziert werden können. Diese kurzen Sequenzfragmente müssen nachfolgend mit Hilfe von Computerprogrammen



einander zugeordnet und zusammengefügt werden. Dies ist typischerweise die erste Aufgabe für die Bioinformatik in einer Genomanalyse, der dann viele weitere Arbeitsschritte am Computer folgen. Für die Sequenzierung des vollständigen Erbguts einer menschlichen Zelle sind ca. eine Woche Zeit und ca. 50.000 EUR erforderlich, während das Erbgut eines unbekanntes Bakteriums bereits an einen Tag und um ca. 10.000 EUR sequenziert werden kann.

Die weitere Analyse von Genomsequenzen erfolgt mit Hilfe spezieller Computerprogramme, die durch Bioinformatiker entwickelt werden. Viele dieser Programme sind so komplex, dass auch deren Anwendung fundiertes bioinformatisches Wissen erfordert. Beginnend von einer Genomsequenz, die nur aus einer langen Abfolge von vier Buchstaben (A, C, G und T) besteht, werden Vorhersagen über die Lage von Genen im Genom, deren Übersetzung in Eiweiße, deren Funktion und Struktur berechnet.

Werkzeuge und Aufgaben der Bioinformatik

Diese komplexe Beschreibung von Genomdaten bildet die Basis für die weitere biologische Forschung an den untersuchten Organismen und hilft, Experimente zielgerichtet zu gestalten. Die benötigten Werkzeuge der Bioinformatik reichen dabei von Arbeitsplatzcomputern über Server und Datenspeicher, Spezialsoftware, Datenbanken aller bekannten biologischen Sequenzen und ihrer Funktion und Vernetzung bis hin zum Internet, mit dessen Hilfe der essenzielle Informationsaustausch und sogar komplexe Berechnungen möglich sind.

In den Jahren 2002-2010 sind in unserer Arbeitsgruppe die Genome zahlreicher Organismen untersucht worden. Dies waren zumeist Bakterien, wie z.B. menschliche Krankheitserreger. Aber auch die Genome von Pflanzen und Tieren wurden durch uns analysiert. In den letzten Jahren verschiebt sich der Arbeitsschwerpunkt zunehmend auf das komplexe Erbgut ganzer Lebensgemeinschaften, die aus zahlreichen Einzelorganismen bestehen. Diese wachsen oft nur in der Gemeinschaft und müssen somit in diesem Kontext untersucht werden.

Die zusätzliche Aufgabe der Bioinformatik ist es dann, die sequenzierten Genomfragmente den in der Gemeinschaft lebenden Organismen zuzuordnen. Trotz der stetig anwachsenden Datenmenge aus Genomsequenzierungen ist deren grundlegende bioinformatische Analyse mit Hilfe modernster Technik und Software eine lösbare Aufgabe. Die schwierigste wissenschaftliche Herausforderung bildet vielmehr das Verständnis des komplexen Zusammenwirkens der Gene, wie im folgenden Abschnitt genauer erläutert wird.

Biologische Netzwerke

Biomoleküle, wie z.B. die Eiweiße, können oft nur funktionieren, wenn sie miteinander wechselwirken. Alle diese Wechselwirkungen formen komplexe Netzwerke mit einer so großen Anzahl von Verbindungen, dass die umfassende Analyse dieser Informationen nur mit Computerunterstützung möglich ist. Besonders aufschlussreich ist dabei die Verknüpfung verschiedener Informationstypen, wie z.B. die Wechselwirkung der Eiweiße untereinander und die Auswirkung des Fehlens der Proteine auf das Wachstum des Organismus. Bei der gut untersuchten Hefe zeigt



sich, dass vielfach (aber nicht immer) diejenigen Eiweiße mit besonders vielen Interaktionspartnern für das Wachstum der Hefezellen zwingend erforderlich sind.

Konkrete Beispiele für biologische Netzwerke sind Protein-Protein-Interaktionsnetzwerke und Stoffwechselnetzwerke. Wechselwirkungen zwischen Proteinen lassen sich z.B. bei der Signalweiterleitung in sog. G-Protein gekoppelten Rezeptoren beobachten. Derartige Rezeptoren sind z.B. wichtig für unser Sehen, Hören und Riechen. Hier wird durch einen von außen ankommenden Reiz eine Aktivierung des Rezeptors verursacht, infolge dessen die Bindung eines an den Rezeptor gekoppelten Eiweißes gelöst wird. So kann der Reiz sehr schnell ins Innere der Zelle geleitet werden und dort weitere Reaktionen auslösen. Stoffwechselnetzwerke hingegen beschreiben die Stoffumwandlungen in einem Organismus, die zumeist unter dem Einfluss von Enzymen, also Eiweißen, ablaufen.

Ein einfaches Beispiel für die Bedeutung der Stoffwechselnetzwerke ist die Erbkrankheit Phenylketonurie. Diese tritt bei ca. jedem 8000. Neugeborenen auf und erfordert eine strenge Diät, um Schäden bei der Entwicklung des Nervensystems zu vermeiden. Verursacht wird diese Erkrankung meist durch einen Defekt im Gen PAH, das für das Enzym Phenylalaninhydroxylase verantwortlich ist. Dieses Gen liegt auf dem langen Arm des Chromosoms 12 in unseren Zellen.

Normalerweise wird die in eiweißreicher Nahrung enthaltene Aminosäure Phenylalanin durch dieses Enzym zu Tyrosin, einer anderen Aminosäure, abgebaut. Beim Defekt des Enzyms PAH funktioniert dieser Abbau nicht mehr. Jedoch nehmen wir auch Tyrosin mit der Nahrung auf, so dass nicht etwa der Mangel an Tyrosin die Krankheit der Phenylketonurie auslöst. Diese ist vielmehr durch das Stoffwechselnetzwerk um Phenylalanin verständlich. Daraus geht hervor, dass bei einem Defekt des PAH Enzyms das gesamte Phenylalanin über andere Stoffwechselwege abgebaut werden muss. Die wichtigste Alternative ist der Abbau mittels des Enzyms L-Aminosäure-Oxidase zu Phenylpyruvat, das bei zu hoher Konzentration zu erheblichen Schädigungen des Nervensystems führt.

Biologische Netzwerke sind somit der Schlüssel zum Verständnis der biologischen Funktion von Gen- und Genomsequenzen. Sie sind derart komplex, dass ein Computereinsatz bei ihrer Analyse unerlässlich ist. Die Bioinformatik organisiert, analysiert und vergleicht biologische Netzwerke. So lassen sich z.B. Datensätze verschiedener experimenteller Methoden für die Messung von Interaktionen im Computer miteinander kombinieren, um ein umfassendes Gesamtnetzwerk zu erstellen. Die Computer-Modellierung derartiger biologischer Netzwerke im Computer beschreibt deren Struktur und Verhalten in Abhängigkeit von wesentlichen Rahmenbedingungen und sagt Eigenschaften gesamter biologischer Systeme voraus, die dann experimentell überprüft werden können. Dies ist derzeit Gegenstand intensiver Forschungen in der Bioinformatik und wird helfen, die Komplexität biologischer Organismen und Lebensgemeinschaften zu verstehen.

In der Arbeit unserer Arbeitsgruppe sind biologische Netzwerke in verschiedenen Projekten vertreten. So studieren wir Strukturprinzipien in Protein-Netzwerken aus Tieren, Pflanzen, Pilzen und Bakterien, um Module von Proteinen zu identifizieren welche zu den gleichen Aufgaben in der Zelle beitragen. Die Kenntnis dieser Module hilft, die Funktion unerforschter Proteine vorherzusagen und die Prinzipien der



zellulären Funktionen besser zu verstehen. In anderen Projekten analysieren wir Stoffwechsel-Netzwerke von Bakterien, um anhand der vorhergesagten Transportprozesse und chemischen Reaktionen das Wachstum der Bakterien in Medien mit verschiedener Zusammenstellung von Nährstoffen vorhersagen zu können. Diese Arbeit hilft bei der Kultivierung der Bakterien im Labor und trägt zum besseren Verständnis ihrer individuellen Lebensweisen bei.

Zusammenfassung

Die moderne biologische Forschung erzeugt enorme Datenmengen, die nur mit Hilfe von Computern bearbeitet werden können. Die Bioinformatik organisiert, verknüpft und analysiert diese Daten. Mit Hilfe von Computermodellen kann das Verhalten biologischer Systeme beschrieben und vorhergesagt werden. Dabei ist eine enge Zusammenarbeit von Theorie und Experiment erforderlich, die in idealer Form einen ständigen Kreislauf des Messens, Auswertens, Modellierens und Modifizierens bildet. Aufgrund der großen Bedeutung der Biologie für unser tägliches Leben, z.B. als Teil der modernen Medizin, in der Biotechnologie, Landwirtschaft für Nahrungsmittel- und Energieproduktion oder in der Abwasserreinigung, wird die biologische Forschung und mit ihr die Bioinformatik entscheidend zur Entwicklung unserer Gesellschaft und zur Lösung der zahlreichen anstehenden globalen Probleme beitragen können. Dazu wird das neu gegründete Department für Computational Systems Biology an der Universität Wien tatkräftig beitragen.

Univ.-Prof. Mag. Dr. Thomas Rattei ist seit März 2010 Professor für "In Silico Genomics" und Leiter des neu gegründeten Departments für Computational Systems Biology an der Fakultät für Lebenswissenschaften.