Arbeitsmaterial **Modul 2**

Die Biologie erzeugt massenweise Daten, die Bioinformatik hebt die darin verborgenen Schätze. Mit Algorithmen und mathematischen Modellen wird am Computer neues Wissen über Pflanzen geschaffen.

Mit Rechenkraft voraus
Datenmengen mit Potential für die Naturwissenschaften

Die moderne Biologie produziert Unmengen an Daten, aber Daten sind noch keine Ergebnisse. Erst richtig ausgewertet, analysiert, verstanden und angewandt werden aus Zahlenkolonnen Erkenntnisse. Die Zukunft gehört mehr und mehr der Bioinformatik.

Foto: © M. Arlt

Im Jahr 2000 wurden die ersten Pflanzengenome sequenziert. Diese gehörten zur Modellpflanze *Arabidopsis thaliana* und zur Weltnahrungspflanze Reis. Zwei Jahre später folgten detailliertere Informationen von Rundkorn- und Langkornreis. In den folgenden Jahren ging es Schlag auf Schlag. Inzwischen sind die Genome von zahlreichen Pflanzen wie Tomate, Kartoffel, Mais, aber auch Kautschuk- und Kakaobaum in Datenbanken abgelegt. Auch Informationen über Transkripte, Proteine und Metaboliten findet man dort.

Die Bioinformatik nutzt diesen Schatz und macht bahnbrechende Entdeckungen, ohne dass ein Labor betreten werden muss. Statt in der Erde zu buddeln, werden Gensequenzen und Transkript-, Protein- oder Metabolitenprofile ausgewertet, die von der experimentellen Biologie *en masse* produziert werden. In riesigen Tabellen wird nach Auffälligkeiten in Zahlenkolonnen und Ausreißern aus der Statistik gesucht. „Auch wir haben Hypothesen, die wir anhand von Daten überprüfen. Der einzige Unterschied ist, dass wir die Daten nicht selber messen", erklärt Dr. Dirk Walther, der am Max-Planck-Institut für Molekulare Pflanzenphysiologie die Arbeitsgruppe Bioinformatik leitet.

Erst 2012 machte das Hallenser Forschungsteam um Ivo Grosse Schlagzeilen in *Nature*, als sie herausfanden, dass die pflanzliche Embryogenese, also die Entwicklung der befruchteten Eizelle zum Embryo, genau wie die tierische nach dem Sanduhr-Modell abläuft. Die Daten, auf denen ihre Arbeit beruht, stammen aus dem Internet. Keinen einzigen Wert hatte das Forschungsteam selbst gemessen, das hatten andere für sie erledigt.

Das einzige Limit ist manchmal die Leistung der Computer. Deswegen setzt die Forschung auch hier auf die Kraft der Masse. Es werden Freiwillige rekrutiert, die ihre ungenutzte Rechenpower für die Suche nach Leben im All (Seti@home) oder der richtigen Faltung von Proteinen zur Verfügung stellen (Fold@home). Manchmal werden komplizierte Probleme auch in Computerspiele verpackt, damit der Spaßfaktor gesteigert wird und sich möglichst viele an der Lösung beteiligen. Ein Beispiel dafür ist *foldit*, wo jeder selbst daran arbeiten kann, Proteine in ihre richtige Form zu bringen. Obwohl Arbeit hier vielleicht das falsche Wort ist.

Überhaupt findet Wissenschaft längst nicht mehr nur im Labor statt. Die „Cititzen Science"-Bewegung wächst trotz ihres angelsächsischen Namens nicht nur in den Vereinigten Staaten. Auch hierzulande sammeln Interessierte in ihrer Freizeit Daten über Zugvögel oder Insektenpopulationen, sie kartieren Pflanzen und die Ausbreitung von invasiven Spezies. Fans der „do it yourself"-Biologie gehen noch einen Schritt weiter und versuchen sich in Laboren an komplizierteren Unterfangen wie der Analyse der eigenen DNA.

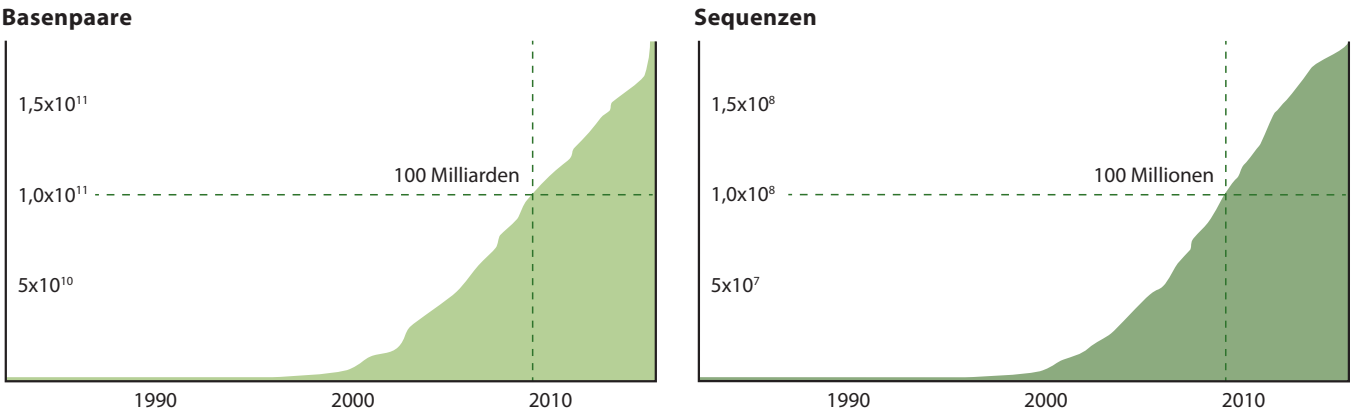
Die stärkere Verfügbarkeit wissenschaftlicher Daten, wie zum Beispiel Sequenzinformationen, schafft ein völlig neues Betätigungsfeld für interessierte Laien.

Bioinformatik 2 **Infobox**

- ***GenBank ist eine der wichtigsten molekularbiologischen Datenbanken. Aufgrund der immer neuen Technologien zur Sequenzierung weist der gesamte Sequenzinhalt von GenBank seit fast drei Jahrzehnten durchgehend ein exponentielles Wachstum auf (Abbildung S.20). www.ncbi.nlm.nih.gov/genbank/statistics***
- ***Wie die Bioinformatik unterliegt auch die Systembiologie als junge Disziplin einem erheblichen Wandel. Schon aus diesem Grund ist eine Grenzziehung zwischen den beiden Gebieten äußerst schwierig. Bioinformatische Methoden setzen typischerweise auf der Ebene der Sequenzinformation an (DNA-, RNA- oder Proteinsequenzen), während im Zentrum der Systembiologie eher die systemische Kontextualisierung von Beobachtungen steht (zum Beispiel das Abbilden von Daten auf Signalpfade oder verschiedene intrazelluläre Netzwerke), durchaus auch gestützt durch quantitative mathematische Modelle der zugrundeliegenden biologischen Systeme.***
- ***Netzwerke sind in den letzten 15 Jahren zu einer wichtigen Datenstruktur der Bioinformatik und der Systembiologie geworden. Daten (zum Beispiel zu Transkriptionsfaktoren und ihren Bindestellen oder zu Protein-Protein-Wechselwirkungen) lassen sich effizient in Form eines Netzwerks zusammenfassen. Zugleich bieten die vielfältigen Analysemethoden komplexer Netzwerke einen neuen Blick auf die in dieser Form repräsentierte biologische Information.***

Quelle: Hütt, Dehnert, Methoden der Bioinformatik, 1. Auflage, Springer-Verlag 2006

Arbeitsmaterial Modul 2



GenBank ist eine große DNA-Sequenzdatenbanken des National Center of Biotechnology Information (NCBI) am National Institutes of Health (NIH). Aktuell sind bei GenBank rund 182 Milliarden Basenpaare und rund 178 Millionen Sequenzen gespeichert (Stand: Oktober 2014; Daten entnommen unter http://www.ncbi.nlm.nih.gov/genbank/statistics).

Viele Daten sind frei zugänglich und jeder kann sich an ihnen ausprobieren

Die teuren Analysegeräte die heute in jeder Universität und jedem Forschungsinstitut zur Standardausstattung gehören, können in einer einzigen Probe den Gehalt von tausenden Transkripten und Proteinen und immerhin noch hunderten Metaboliten gleichzeitig messen. Es werden mehr Daten produziert, als mit der heutigen Technologie analysiert werden können.

Doch die Bioinformatik holt auf. „Fortschritte im Hochleistungsrechnen sowie numerische Algorithmen haben computerintensive Simulationen von komplexen mathematischen Modellen sowie die Bearbeitung von Milliarden oder sogar Billionen Datenpunkten ermöglicht und sogar zur Routine gemacht", schreiben die Biologen Steve Long und Mark Stitt in der Fachzeitschrift *Plant, Cell and Environment*. Aber Algorithmen schreiben sich nicht von alleine und hinter jedem mathematischen Modell sitzt ein schlauer Kopf, der es erdacht hat. Die Bioinformatik sucht deshalb händeringend Nachwuchs.

Ein wenig Statistik im Grundstudium, das reicht heute nicht mehr

In Deutschland bieten inzwischen viele Hochschulen den Studiengang Bioinformatik an, daneben gibt es zahlreiche verwandte Angebote wie „Bioinformatik und Genomforschung" oder „Bioinformatik und Systembiologie". Einer der Vorreiter war die Freie Universität Berlin, die im November 2000 als einen Bachelor- und Masterstudiengang Bioinformatik eingerichtet hat. „Die Freie Universität hat damit auf den zunehmenden Bedarf an interdisziplinär arbeitenden Kapazitäten für Biotechnologie reagiert", heißt es in der Pressemitteilung dazu. „Für die akademische Forschung ist zurzeit genug Nachwuchs da", so Dr. Dirk Walther, „aber wenn sich die personalisierte Medizin weiter so rasant entwickelt, dann werden zukünftig im medizinischen Bereich vermehrt bioinformatische Fachkräfte gebraucht werden."

Diese müssen sich besonders für Informatik und Mathematik interessieren. Doch während diese Fähigkeiten in der Physik schon lange als wichtiges Grundlagenwissen gelten, wurden sie in der Biologie oft nur wenig beachtet. Etwas Mathematik im Grundstudium, gerade genug um statistische Signifikanz zu berechnen, das reicht heute nicht mehr. Wer in der Biologie mithalten will, der muss sich in Zukunft auch mit mathematischer Modellierung auskennen. „Wissenschaftler mit dieser Ausbildung werden vermutlich einige der wichtigsten zukünftigen Entdeckungen machen", schreiben Long und Stitt in ihrem Artikel.

Dazu braucht man nicht einmal teure Spezialausrüstung, keine millionenschweren Messgeräte und keine energiehungrigen Gewächshäuser. Ein paar Computer mit Internetanschluss und Menschen mit Ideen, das ist alles. „In dieser Disziplin kann man relativ schnell zur Weltspitze aufsteigen", so Dr. Dirk Walter. Genau das macht die Bioinformatik so interessant für Entwicklungs- und Schwellenländer.

Wer weiß, vielleicht geht auch der Nobelpreis für Medizin eines Tages an Bioinformatiker. Die Chemie hat es vorgemacht, die Preisträger von 2013 wurden für ihr Computermodell zur Simulation chemischer Reaktionen geehrt.

Diesen Artikel online lesen ***http://bit.ly/1sG7V68***

Arbeitsaufträge

1. ***Informieren Sie sich über die Genomgrößen folgender Pflanzen. Legen Sie eine Tabelle mit der Zahl der Basenpaare (Mb = Megabasenpaare = 10^6 Basenpaare) an und sortieren Sie diese nach der Größe: Ackerschmalwand (Arabidopsis thaliana), Eukalyptus (Eucalyptus grandis), Gerste (Hordeum vulgare), Gurke (Cucumis sativus), Kartoffel (Solanum tuberosum), Kulturapfel (Malus domestica), Mais (Zea mays), Orange (Citrus sinensis), Paprika/ Spanischer Pfeffer (Capsicum annuum), Reis (Oryza sativa), Tomate (Solanum lycopersicum), Virginischer Tabak (Nicotiana tabacum), Walderdbeere (Fragaria vesca), Weichweizen (Triticum aestivum), Westliche Balsam-Pappel (Populus trichocarpa), Zuckermelone (Cucumis melo), Zuckerrübe (Beta vulgaris).***
2. ***Was versteht man unter der „Citizen Science"? Stellen Sie Beispielprojekte vor, die Sie persönlich interessant finden.***