Einleitung Modul 2

„Big Data“ heißt: Daten zum Sprechen bringen

Große Datensätze setzen sich oft aus Millionen einzelner Informationen zusammen, die aus ganz unterschiedlichen Quellen stammen können und in einen neuen Kontext gebracht werden müssen. Forschungsdaten aus automatisierten, multiparallelen Hochdurchsatzanalysen, aber auch komplexe Umwelt- oder Klimadaten können hierfür die Grundlage sein. Da diese Daten zunächst jedoch ohne große Aussagekraft sind, besteht die Herausforderung der Wissenschaft darin, die entscheidenden Fragen zu stellen und die Daten so zum Sprechen zu bringen.

Algorithmen steuern das „Handeln“ von Computern

Zum besseren Verständnis können Daten als einzelne Buchstaben oder Wörter betrachtet werden, die erst durch Zusammenfügen zu verständlichen Sätzen und durch Kombinieren zu sinnvollen Texten werden. Diese Aufgabe übernehmen spezielle Computeralgorithmen, die in der Lage sind, die Datensätze automatisiert zu strukturieren und auszuwerten. Algorithmen sind Abfolgen von präzisen Handlungsanweisungen, die festlegen, wie der Computer mit den Daten umgehen soll. Wenn also Daten Wörter und Buchstaben darstellen, handelt es sich bei Algorithmen um die grammatikalischen Regeln, nach denen ein verständlicher Text entsteht.

Alles beginnt mit den richtigen Fragen

Um Forschungsdaten analysieren und auswerten zu können, müssen diese zunächst im Experiment erhoben werden. In der Pflanzenforschung können solche Daten beispielsweise genetische Informationen oder Merkmale einzelner Pflanzen, Populationen oder Gruppen unterschiedlicher Arten sein.

Um eine Hypothesen zu überprüfen und um biologische Prozesse besser zu verstehen, konzentriert sich die Wissenschaft sowohl auf Auffälligkeiten als auch auf bestimmte wiederkehrende Muster und Gemeinsamkeiten. Mit Hilfe von multivariaten Analysemethoden können zum Beispiel mehrere Variablen bzw. Merkmale auf einmal untersucht und anschließend mit Hilfe der Hauptkomponenten-Analyse strukturiert abgebildet werden.

Die Speicherung und Archivierung von Daten spielt eine wichtige Rolle

Forschungsdaten benötigen wie Bild-, Musik- und Textdateien Speicherplatz. So wie Urlaubsfotos auf der Festplatte eines Computers gespeichert werden, werden auch wissenschaftliche Daten auf Festplatten abgelegt. Diese Festplatten befinden sich häufig in verschiedenen Computern, die über das Internet miteinander verbunden sind. Für die Forschung bedeutet das, dass sie Zugang zu den Forschungsdaten anderer Arbeitsgruppen erhalten kann. Ein Forscher aus Tokyo kann beispielsweise auf die Daten einer amerikanischen Kollegin zugreifen. Diese Form der Speicherung und Archivierung ist eine tragende Säule von „Big Data“. Der freie Zugang zu Literatur und Materialien aus der Wissenschaft, zu denen auch wissenschaftliche Rohdaten zählen, wird häufig unter dem Ausdruck „Open Access“ zusammengefasst.

„Big Data“ erfordert einheitliche Standards

Damit der Austausch von Forschungsdatensätze sinnvoll ablaufen kann, ist es entscheidend, einheitliche Regeln in der Klassifizierung dieser Daten, sogenannte Standards, zu etablieren. Da nicht immer unmittelbar erkennbar ist, welche Informationen ein Datensatz enthält, muss er mit nachvollziehbaren Zusatzinformationen beschriftet werden. Diese Zusatzinformationen werden als „Meta-Daten“ bezeichnet. Sie sind mit der Deklaration von Lebensmittelverpackungen vergleichbar. Neben einer „Zutatenliste“ kann den Meta-Daten beispielsweise entnommen werden, unter welchen Umständen die Daten gewonnen wurden.

Bioinformatik 1 **Infobox**

- *Bioinformatik ist ein interdisziplinäreres Studium, das biologische und biochemische Kenntnisse mit Informatik verbindet. Für die Pflanzenforschung wird die Bioinformatik zunehmend wichtiger, um die großen Datenmengen der Forschung zu katalogisieren, auszuwerten und zu interpretieren.*
- *Die Bioinformatik ist auch ein Ort innovativer Lernkonzepte, wie etwa die Online-Plattform Rosalind zeigt. Rosalind ist eine Lernplattform der Bioinformatik und des problemorientierten Programmierens. http://rosalind.info/problems/locations*
- *Eine der größten deutschen Investitionen in die Bioinformatik der letzten drei Jahre ist die BMBF-Initiative „Deutsches Netzwerk für Bioinformatik-Infrastruktur“ vom September 2013.*

Einleitung Modul 2

Moderne Technologien für die Biologie

Durch moderne Technologien sind in den vergangenen Jahren völlig neue Teilgebiete der Biologie entstanden, die „Omics" genannt werden. Ziel dieser Forschungsrichtungen ist es, biologische Systeme in ihrer Gesamtzeit zu verstehen. Statt sich also auf Einzelelemente, wie ein Gen oder ein Stoffwechselprodukt (Metabolit) zu konzentrieren, wird die Dynamik und Vernetzung aller Elemente des gesamten Genoms oder Metaboloms untersucht. Die zugehörigen Teilgebiete heißen „Genomik", „Epigenomik", „Transkriptomik",„Proteomik" und„Metabolomik".

Die Voraussetzung für diese neuen Forschungszweige bilden moderne Hochleistungstechnologien. Hochdurchsatzverfahren ermöglichen es, ein Genom schnell und kostengünstig zu sequenzieren, d.h. die Reihenfolge der Grundbausteine der Erbinformationen auf den Chromosomen zu bestimmen. Moderne Sequenzierungsmethoden erlauben es heute, das menschliche Genom in wenigen Wochen und für wenige tausend Dollar zu sequenzieren. Die Sequenzierung des ersten menschlichen Genoms, die im Rahmen des internationalen „Humangenomprojekts" 1990 startete, dauerte 13 Jahre und kostete rund 3 Milliarden Dollar. Über 1.000 Personen waren weltweit an diesem Projekt beteiligt.

Was kann „Big Data" leisten?

Das bessere Verständnis der Gesamtheit biologischer Systeme ermöglicht es, diese auch im Wechselspiel mit ihrer Umwelt zu betrachten. Die Metabolomik versucht beispielsweise Stoffwechselprodukte von Pflanzen zu identifizieren, die eine gesundheitsfördernde Wirkung besitzen. Hängt der Gehalt dieser Stoffe mit bestimmten Umweltfaktoren zusammen, und kann die Regulation der Synthese in der Pflanze verstanden werden, kann gezielt daran gearbeitet werden, die Produktion dieser Stoffwechselprodukte durch die Pflanze zu steuern.

Eine andere vielversprechende Herausforderung besteht darin „stärkere" Pflanzen zu züchten. So fand ein Forschungsteam unter Beteiligung von Professor Björn Usadel (im Interview auf Seite 22) einen Weg, Tomaten unempfindlicher gegenüber Hitze und Trockenheit zu machen. Als Vorbild diente dabei eine Wildtomatensorte, deren Erbgut mit dem unserer Kulturtomaten verglichen wurde. Dabei wurde ein Gen identifiziert, das die Bildung einer schützenden Wachsschicht steuert, die die Wildtomate vor Hitze und Trockenheit schützt.

„Big Data" führt zu neuen Arbeitsmethoden in der Wissenschaft

Im Jahr 2012 gelangte das Forschungsteam unter der Leitung von Ivo Grosse, Professor für Bioinformatik an der Martin-Luther-Universität Halle-Wittenberg, auf das Titelblatt des renommierten Fachmagazins *Nature*. Das Team fand heraus, dass es in der Embryonalentwicklung der Pflanzen eine Phase gibt, in der sich Embryonen artübergreifend zum Verwechseln ähneln. Dieses Phänomen, auch „Sanduhrmodell" genannt, war bisher nur aus der Tierwelt bekannt. Außergewöhnlich an der Vorgehensweise war, dass auf bereits existierende Daten zugriffen wurde, und diese unter neuer Fragestellung und mit Hilfe computerbasierter Methoden analysiert wurden.

Ein weiterer Gewinn durch „Big Data" besteht darin, verschiedenste Szenarien am Computer simulieren zu können. Auf diese Weise lässt sich etwa überprüfen, wie sich eine Pflanze mit einer bestimmten genetischen Ausstattung verhalten könnte, wenn sich einzelne Parameter wie Licht, Temperatur oder Niederschlagsmenge ändern. Diese rationalen Ansätze helfen Zeit und Kosten zu sparen. Die Ergebnisse der Simulationen können im Anschluss unter realen Bedingungen auf dem Feld überprüft werden.

Landwirtschaft 3.0

In der landwirtschaftlichen Praxis arbeiten heute bereits Prototypen von Feldrobotern, die voll automatisiert punktgenau Dünger und Pflanzenschutzmittel auf die Äcker auszubringen können. Grundlage hierfür ist die schnelle und präzise Interpretation großer Datenmengen in Kombination mit GPS- und spektroskopischen Daten. Unterstützung erhalten die Feldroboter dabei von Satelliten oder Agrardrohnen. Bodenstationen empfangen deren Daten und berechnen in Sekundenbruchteilen die optimale Menge an Dünge- oder Pflanzenschutzmitteln, die auf bestimmte Stellen des Ackers ausgebracht werden muss. Die benötigte Düngermenge orientiert sich maßgeblich am Zustand der Pflanzen oder dem Reifegrad ihrer Früchte. Diese Methoden kommen nicht nur dem Landwirt, sondern auch der Umwelt zugute.

Berufsfeld mit Zukunft!

Das Konzept von „Big Data" bringt in vielen Bereichen der Wissenschaft Kapazitäten aus unterschiedlichsten Disziplinen an einen Tisch. Im Kontext der Pflanzenforschung kombiniert die Bioinformatik das Wissen der Lebenswissenschaften aus Biologie, Genetik, Biochemie und Biotechnologie mit dem der Informatik, Mathematik und Statistik. Auf diese Weise ist in den letzten Jahren ein ganz neues Fachgebiet entstanden, das sich mittlerweile als ein eigener Wissenschaftszweig etabliert hat: die Bioinformatik. Ein Studium auf diesem Gebiet ist nicht nur theoretisch und praktisch, sondern auch technisch ausgerichtet und vermittelt neben den naturwissenschaftlichen Grundlagen, Spezialwissen aus den Bereichen der Informatik und Datenverarbeitung. Die Bioinformatik ist ein Berufsfeld mit Zukunft.

Arbeitsaufträge

1. ***Lesen Sie den Einleitungstext, und definieren Sie die folgenden Begriffe mit Ihren eigenen Worten: Algorithmus, Big Data, Datenbank, Meta-Daten und Bioinformatik.***
2. ***Erstellen Sie eine Tabelle mit den als „Omics-Technologien" bezeichneten Forschungszweigen der Biologie, den Systemen mit denen sich diese beschäftigen und den zugehörigen Einzelelementen aus denen sich diese Systeme zusammensetzen (z.B. Teilgebiet: Genomik, System: Genom, Einzelelement: Gen).***
3. ***Was ist mit „Landwirtschaft 3.0" gemeint? Diskutieren Sie im Klassenverband die Vor- und Nachteile von Landwirtschaft 3.0. Bilden Sie sich anhand dessen eine persönliche Meinung.***