Spektrum DER WISSENSCHAFT EXTRA

KOSMOLOGIE
Das simulierte Universum

BIOMECHANIK
Biomoleküle auf der virtuellen Streckbank

COMPUTERLINGUISTIK
Wissen für die künstliche Intelligenz

Datengetriebene Wissenschaft

Spektrum DER WISSENSCHAFT und Heidelberger Institut für Theoretische Studien
Eine Publikation von
HITS

Spektrum DER WISSENSCHAFT EXTRA 2/11

Reinhard Breuer

Das Unmögliche demnächst – nur Wunder dauern etwas länger

Als ich in die Wissenschaft einstieg, war man schon froh, wenn man mit dem Rechner eine Kurve »plotten« konnte – gerne auch mal in Abhängigkeit von zwei Variablen, also echt dreidimensional! Später gelang es vereinzelt sogar, dynamische Probleme mit Hilfe von Differenzialgleichungen für einige Zeitschritte zu verfolgen. Und das vom »Terminal« aus, ganz ohne Lochstreifen oder -karten – ein absolutes Highlight in der Computersteinzeit.

Der Rückblick sei gestattet, um die Dimensionen des Fortschritts zu begreifen: Welche Probleme sich heute mit Superrechnern behandeln lassen und welch riesige Datenmengen dabei produziert und gezielt ausgewertet werden, überstieg noch vor wenigen Jahren fast die Vorstellungskraft oder galt schlicht als unmöglich. Darum erstaunt es mich immer wieder, was in der Wissenschaft inzwischen alles machbar ist – und was insbesondere auch an dem vom Wissenschaftsmäzen Klaus Tschira gegründeten Heidelberger Institut für Theoretische Studien passiert, mit dem sich dieser Sonderteil beschäftigt.

Die dort tätigen Forscher stellen ausgewählte Projekte ihrer Arbeit auf den folgenden Seiten selbst vor, betreut und unterstützt von der Redaktion von »Spektrum der Wissenschaft«. Einige haben vorher schon in unserem Magazin geschrieben oder wurden darin porträtiert: so der Computerlinguist Michael Strube (»Wikipedia: Wissen für die Künstliche Intelligenz«, 12/2010, S. 94) und der Astrophysiker Volker Springel (»Vielleicht laufen wir einem Phantom nach«, 11/2010, S. 34).

Wer hätte sich träumen lassen, was Forscher heutzutage mit Hilfe von Simulationen – so nennt man die Berechnungen inzwischen – alles ergründen können: neben der Entstehung von Galaxien (S. 10) und der automatischen Erkennung natürlicher Sprachen (S. 30) auch die Stammesgeschichte von Organismen (S. 22) oder die Wechselwirkung von Proteinen (S. 14). Man möchte an numerische Zauberei glauben, so schnell gerät das (einst) Unmögliche in Reichweite – nur Wunder dauern immer noch etwas länger.

Mit den exponentiell anwachsenden Datenmengen und Publikationen wächst aber zugleich das Problem, sich darin noch zurechtzufinden. Entsprechend arbeiten auch Gruppen am HITS über Datenbankmanagement, beispielsweise um für Forscher Informationen über Stoffwechselprozesse bereitzustellen (S. 26). Denn die Simulation von Problemen mit niemals völlig zutreffenden, aber oft nützlichen Modellen (wie Klaus Tschira in seinem Editorial auf der nächsten Seite vermerkt) ist nur ein Aspekt jener »datengetriebenen Wissenschaft«, die mit dem Siegeszug der Höchstleistungsrechner immer mehr an Bedeutung gewinnt. Auf die immensen Herausforderungen, vor die sie alle Forschungsgebiete stellt, weist HITS-Chef Andreas Reuter in seinem Beitrag ab S. 6 hin.
Ob diese Herausforderungen schon überall verstanden sind, lässt sich bezweifeln. Wohin jedoch die abenteuerliche Reise vermutlich geht, können Sie bei der Lektüre der folgenden Artikel erahnen.

Reinhard Breuer
Editor-at-Large
Spektrum der Wissenschaft

TIM WEGNER, © KLAUS TSCHIRA STIFTUNG

Klaus Tschira

Alle Modelle sind falsch, aber einige immerhin nützlich

Was tun Wissenschaftler, die in der Grundlagenforschung arbeiten, die also versuchen, bestimmte Teilaspekte der uns umgebenden Welt zu verstehen? Manche von ihnen machen das, was die meisten Menschen von Wissenschaftlern erwarten: Sie beobachten, zählen, messen, registrieren, katalogisieren. Das sind die Empiriker. Sie streben danach, möglichst genaue Informationen darüber zu erhalten, wie Vorgänge in der Natur ablaufen.

Aber das ist nur der eine Teil des wissenschaftlichen Geschäfts. Für den anderen Teil sind die Theoretiker zuständig, die versuchen, in den Beobachtungen der Experimentatoren Gesetzmäßigkeiten zu erkennen und diese so zu formulieren, dass sie nicht nur mit den vorhandenen Beobachtungen übereinstimmen, sondern auch das Ergebnis von Experimenten voraussagen können, die noch gar nicht durchgeführt worden sind. Solche Gesetzmäßigkeiten können unterschiedliche Gestalt annehmen: Formeln, Diagramme, Computerprogramme und so weiter.

Jede Theorie verkörpert ein Modell des betrachteten Ausschnitts der Wirklichkeit und ist insofern stets eine Abstraktion oder Idealisierung: Sie beschreibt die Realität niemals absolut genau, sondern erfasst bestimmte relevante Aspekte »hinreichend gut« – unter Vernachlässigung anderer, für die Fragestellung irrelevanter Details. So gesehen sind alle Modelle falsch, wie der Statistiker George Box von der University of Wisconsin in Madison provokant formulierte. Sie können gleichwohl nützlich sein, sofern sie die – zumindest näherungsweise – Berechnung von Effekten erlauben, über die noch keine Messungen vorliegen. Die Wettervorhersage etwa beruht auf vielen Vereinfachungen und trifft nicht immer zu – aber sie ist, zumindest gelegentlich, sehr nützlich.

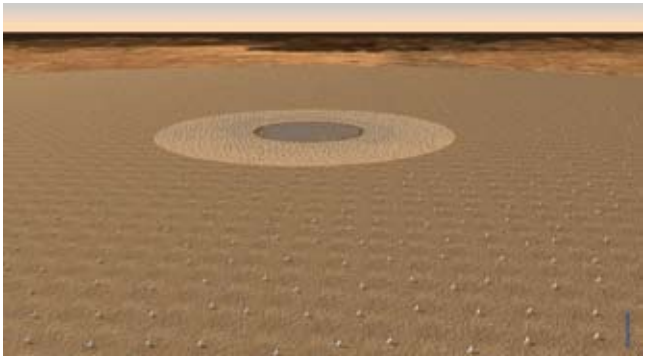
Die zwei genannten Arbeitsweisen ergänzen sich auf fruchtbare Weise. So nutzte Johannes Kepler das umfangreiche Beobachtungsmaterial Tycho Brahes zur Formulierung seiner Planetengesetze – ein klassisches Beispiel dafür, wie Messergebnisse durch Theoriebildung zu neuen Erkenntnissen führen. Manchmal ist das theoretische Modell auch zuerst da. Dann dienen Messungen dazu, es durch Vergleich mit seinen Voraussagen nachträglich zu bestätigen oder zu widerlegen. In diese Kategorie fällt Einsteins allgemeine Relativitätstheorie, die erst Jahre später experimentell untermauert wurde.

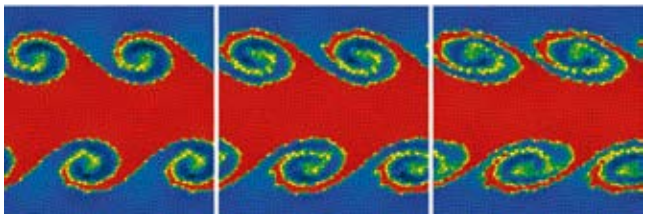
Natürlich befruchten Theorien auch die empirische Seite der Wissenschaft. So ermöglichen sie neue experimentelle Fragestellungen oder innovative Messverfahren. Weder Experimente noch Theorien allein verhelfen also zu grundlegenden neuen Einsichten. Nur ihr Wechselspiel bringt die Wissenschaft voran.

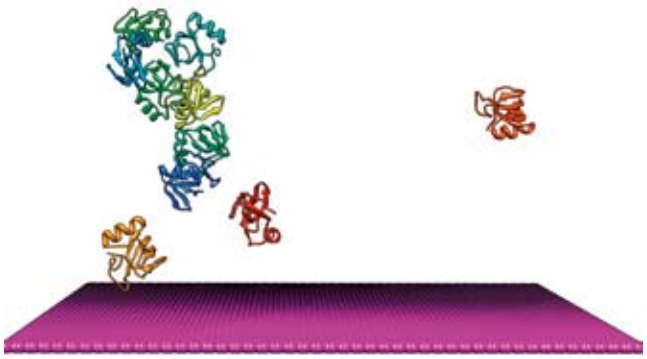
Seit etwa 20 Jahren verschiebt sich jedoch die Balance zwischen Experiment und Theorie in einem Maß, das teils schon eine Entkopplung befürchten lässt. Der Hauptgrund dafür ist, dass erheblich mehr Fördermittel in Experimentiereinrichtungen wie Beschleuniger, Teleskope, Sequenzierer oder Computer geflossen sind als in die Theoriebildung. In Verbindung mit dem rasanten Leistungszuwachs in der Halbleitertechnik kam es so zur Ansammlung gigantischer Datenmengen, die kein Mensch mehr allein durch Sichten und Nachdenken verarbeiten kann. Die Frage, wie solche Datenfluten jemals zu Theorien verdichtet, zu Erkenntnis veredelt werden können, geriet völlig in den Hintergrund. Das war für mich der Impuls zur Gründung des gemeinnützigen Heidelberger Instituts für Theoretische Studien (HITS).

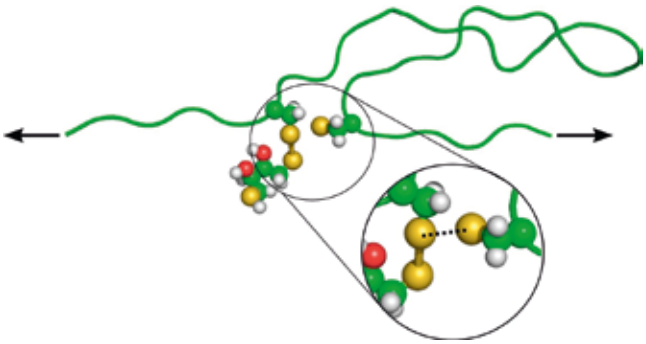
Daten gibt es, wie gesagt, in Hülle und Fülle, und zwar auf allen Gebieten der Naturwissenschaften und darüber hinaus. Die Forschungsgruppen des HITS sollen technisch und organisatorisch die Möglichkeit bekommen, Methoden zu entwickeln und zusammen mit experimentell arbeitenden Forschern zu erproben, die es erlauben, diese Datenmengen effektiv zu verwalten und zur Gewinnung neuer Einsichten nutzbar zu machen. Wenn dabei gelegentlich regelrechte Forschungs-Hits entstehen, ist das ganz im Sinne des Erfinders.

Klaus Tschira
Geschäftsführer HITS gGmbH

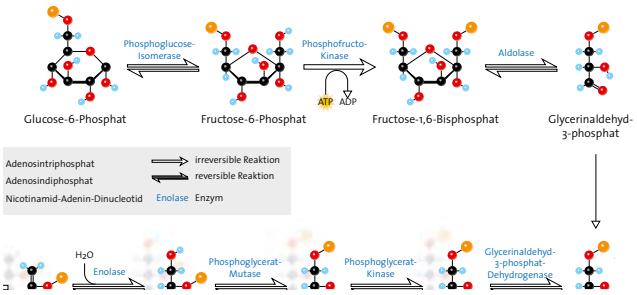


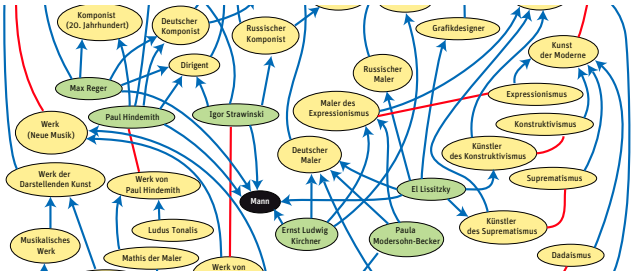






MIT FREUNDLICHER GENEHMIGUNG VON ANNE ASHLEY UND GERARD MANNING, SALK INSTITUTE





Datengetriebene Forschung – Herausforderung für die Informatik

Auf allen Gebieten der Natur- und Ingenieurwissenschaften gewinnt eine Arbeitsmethode an Bedeutung, bei der die Analyse von sehr großen Mengen an Daten zu neuen Erkenntnissen führt. Mit welchen technischen und organisatorischen Maßnahmen kann die Informatik eine solche datengetriebene Forschung unterstützen?

Von Andreas Reuter

Derzeit vollzieht sich ein grundlegender Wandel in den Natur- und Ingenieurwissenschaften – und das gleich auf mehreren Ebenen. So ändern sich etwa als Folge des Internets und der darauf aufbauenden Dienste die Kommunikationsstrukturen innerhalb und zwischen den Fachgebieten ganz erheblich. Außerdem entwickeln sich neue Organisationsformen für wissenschaftliche Einrichtungen, die der rasch wachsenden Komplexität der Forschungsvorhaben Rechnung tragen. Das wiederum erfordert neue Finanzierungsmodelle für wissenschaftliche (Groß-)Vorhaben. Ferner wandelt sich die Position der Wissenschaft in der Gesellschaft: Ihr wird – zumindest in den westlichen Ländern – sehr viel mehr Transparenz und Rechtfertigung der Ziele und Methoden abverlangt als früher.

Aber auch die wissenschaftliche Methodik selbst befindet sich im Umbruch. In der öffentlichen Wahrnehmung, die sich in Filmen oder Reportagen widerspiegelt, erscheint Wissenschaft immer noch als Tätigkeit, der einzelne (vorzugsweise geniale) Forscher in der Abgeschiedenheit eines Labors nachgehen. Dort kommen sie nach zähem Ringen und diversen Geistesblitzen zu bahnbrechenden Erkenntnissen – oder aber erfinden eine fürchterliche Waffe, je nachdem, ob sie auf der Seite der Guten oder der Schurken stehen.

Dieses romantisierende Bild hat mit moderner Forschung wenig zu tun. Zum einen wird Wissenschaft in immer größeren, komplexeren Projekten und Projektverbünden organisiert – man denke etwa an den Large Hadron Collider (LHC) bei der Europäischen Organisation für Kernforschung CERN in Genf. Solche Vorhaben lassen sich ohne industrielle Methoden und eine hochgradig arbeitsteilige Vorgehensweise nicht bewältigen. Zum anderen stehen Wissenschaftler immer häufiger nur noch in sehr indirektem Kontakt mit den Gegenständen ihrer Untersuchung, seien es Zellen oder Galaxien.

Wenn wir die Entwicklung im Methodenvorrat der Naturwissenschaften einmal Revue passieren lassen (unterer Kasten auf S. 8), so gab es ganz am Anfang die empirische Beschreibung, die sich gelegentlich zu – gleichfalls empirisch abgeleiteten – Handlungsregeln verdichtete. Parallel dazu, aber doch mit einer merklichen Verzögerung, entwickelte sich die Theoriebildung. Ihr Ziel war, formalisierte Modelle der beobachteten Phänomene zu erstellen. Diese sind aber nur um den Preis idealisierender Annahmen möglich – beispielsweise durch Vernachlässigung der Reibung bei der Beschreibung von Bewegungsgesetzen.

Rasch anschwellende Datenflut

Der nächste Schritt bestand darin, den Grad der Idealisierung zu verringern, um auch komplexe Vorgänge wie etwa die Verformung eines Autos beim Aufprall oder die Flugeigenschaften eines Flugzeugs so realistisch beschreiben zu können, dass sich die Ergebnisse der Modellanalyse auf das tatsächliche System übertragen lassen. Dies führte zu sehr komplizierten Modellen, deren Gleichungen nicht mehr direkt lösbar waren. In solchen Fällen bleibt nur die Möglichkeit, mit Methoden der numerischen Mathematik Näherungslösungen zu bestimmen, was bei großen Problemen wie etwa der Crashsimulation im Automobilbau Gleichungssysteme mit Hunderttausenden oder Millionen von Unbekannten ergibt. Deren Handhabung ist überhaupt nur noch mit dem Computer möglich.

Heute gilt die Simulation vielfach als dritte Säule der Wissenschaft – neben Experiment und Theorie. Manche sprechen ihr zwar diesen Rang ab und betrachten sie nur als eine von mehreren möglichen Arten, theoretische Modelle auszuwerten. Tatsache aber ist, dass ohne Simulation viele Modelle »steril« bleiben würden, da es nicht möglich wäre, Ergebnisse daraus abzuleiten.

Von der Simulation führt der Weg schließlich zur datengetriebenen Wissenschaft. Auf den ersten Blick scheint sie nichts grundsätzlich Neues zu bieten; schließlich geht es nur um die Zusammenführung von Experiment (Messung), Theoriebildung und Simulation zu einem kohärenten Methodenvorrat. Das eigentlich Interessante ist jedoch der Grund, der diese Zusammenführung notwendig macht: die rasch wachsende Menge von Daten, die von Messgeräten (wie Satelliten, Teleskopen, Sequenziermaschinen und Microarrays) oder aus Simulationen (etwa Klimavorhersagen und Szenarienanalysen) stammen. Denn die Menge neu erzeugter und gespeicherter Daten verdoppelt sich jedes Jahr, oder anders ausgedrückt: In jedem einzelnen Jahr fallen mehr experimentelle oder Simulationsdaten an als in allen Vorjahren zusammen. Am Anfang sieht eine solche exponentielle Wachstumskurve noch relativ harmlos aus, und tatsächlich konnten Forscher immerhin bis ins 20. Jahrhundert hinein Messergebnisse durch Sichten und Darübernachdenken analysieren.

Mit zunehmender Automatisierung der Messgeräte und dem breiteren Einsatz von Simulationsmethoden sind die Daten in vielen Projekten jedoch schon längst auf einen Umfang angewachsen, der es völlig unmöglich macht, sie im herkömmlichen Sinn direkt in Augenschein zu nehmen. Hierzu nur zwei

Beispiele: Das LHC-Experiment des CERN wird im Vollbetrieb 15 Petabytes (PB) pro Jahr erzeugen, und beim Square Kilometer Array (einem für 2024 geplanten System von Radioteleskopen) soll es sogar 1 PB pro Tag sein. Liegen die Milliarden der aktuellen Finanzkrise schon jenseits der menschlichen Vorstellungskraft, so verhält es sich mit den Petabytes noch eine Million Mal schlimmer (oberer Kasten auf S. 8). Wenn ein Mensch 80 Jahre lang ohne Unterbrechung nichts anderes täte, als sich 1 PB an Ergebnissen »anzusehen«, müsste er pro Sekunde 320 000 Buchstaben (ein Taschenbuch) lesen, um ganz durchzukommen.

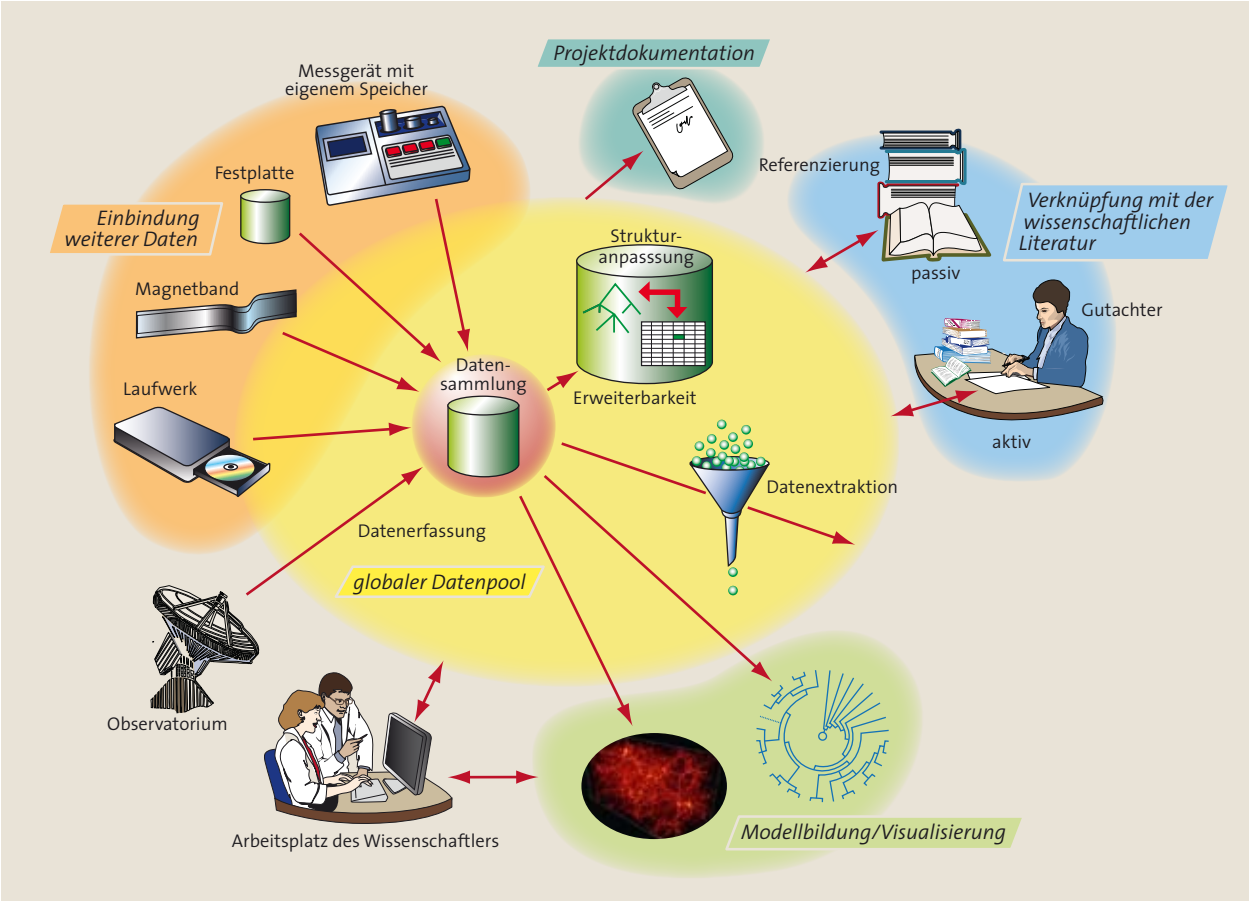
Es bleibt also nichts anderes übrig, als die experimentellen Daten zunächst von Software unterschiedlichster Art aufbereiten zu lassen. Das Datenvolumen muss durch Verdichtung, Selektion, statistische Analyse, Visualisierung und andere Verfahren so weit reduziert werden, dass das Ergebnis für den Menschen wieder aufnehmbar ist. Vor 50 Jahren haben Wissenschaftler noch unmittelbar durch die Teleskope oder Mikroskope geschaut, selbst die Messgeräte abgelesen und die Vorgänge im Reagenzglas beobachtet. Heute kommen sie mit den Experimenten oft erst durch das in Berührung, was auf dem Bildschirm ihres PCs erscheint, nachdem es über viele Stufen hinweg gefiltert, komprimiert und visualisiert worden ist.

Forschung im heutigen Sinn besteht also großteils in der durch Computer, Datenbanken und viele andere Softwarewerkzeuge unterstützten Verarbeitung sehr großer Mengen von Daten, die aus einer Vielzahl von Quellen stammen. Und damit stellt sich die Frage, was die Informationstechnik (IT) dazu beitragen kann und muss. Die offensichtliche Antwort lautet: Ihre Aufgabe ist es, Hardware- und Softwaresysteme zur Verfügung zu stellen, die es den Wissenschaftlern ermöglichen, alle für ihre Fragestellung erforderlichen Auswertungen effizient durchzuführen, ohne sich dabei um IT-spezifische Aspekte kümmern zu müssen. Ein Biologe will schließlich Biologie betreiben und nicht programmieren. Aber was heißt das konkret? In den folgenden Abschnitten skizziere ich die wichtigsten Forderungen an die IT (Übersicht im Kasten unten).

Zunächst müssen die von den Experimenten oder Simulationen kommenden Daten zuverlässig gespeichert werden – und dies unter Umständen mit enormer Geschwindigkeit, wenn man an die oben zitierten Beispiele denkt. Es darf keine Unterbrechungen geben, weil viele Versuche nicht wiederholbar sind. Ferner gilt es, die Daten schon beim Erfassen zu prüfen, zu filtern und für die langfristige

Computergestützter Umgang mit riesigen Datenmengen

Zum Anlegen, Verwalten und Nutzbarmachen eines globalen Datenpools braucht es Software, die vielerlei Anforderungen erfüllen muss. Einige wesentliche sind hier in der Grafik veranschaulicht.



SPEKTRUM DER WISSENSCHAFT / BUSKE-GRAFIK, NACH ANDREAS REUTER

Speicherung aufzubereiten, was weitere hohe Anforderungen an die Leistungsfähigkeit der Hard- und Software stellt.

Das Abspeichern hat dabei so zu erfolgen, dass die Bestände wachsen können, unter Umständen um mehrere Größenordnungen. Außerdem muss jederzeit eine Erweiterung um neue Informationskategorien und Datenstrukturen möglich sein.

Verwandt damit ist die Forderung, Daten aus verschiedenen Projekten und Disziplinen miteinander verknüpfen zu können, um übergreifende Fragen zu untersuchen. Zum Beispiel müssen in der Klimaforschung meteorologische, ozeanografische, geografische, statistische und etliche weitere Datensammlungen zueinander in Beziehung gesetzt werden. Das scheitert heute oft an ihrem unterschiedlichen Aufbau. So verwenden die einzelnen Disziplinen häufig andere Begriffe und Einheiten oder nicht einmal dasselbe Koordinatensystem. Da jedes Fachgebiet zudem seine eigenen Modellierungsmethoden einsetzt, muss es möglich sein, die Daten flexibel in der dafür erforderlichen Struktur bereitzustellen.

Zur Verarbeitung der Rohdaten gehört auch, sie zu verdichten; denn nur in komprimierter Form kann der Forscher die enthaltene Information aufnehmen. Die Software sollte möglichst verschiedene Arten der Verdichtung erlauben, so dass sich im Einzelfall diejenige Methode auswählen lässt, die am besten zu den jeweiligen Daten und Modellen passt. Von besonderer Bedeutung ist dabei die visuelle Darstellung.

Meist müssen Datenbestände für verschiedene Auswertungen immer wieder durchsucht und verarbeitet werden. Wenn sie sehr groß sind, beansprucht das viel Zeit. Die Geschwindigkeit des Zugriffs auf gespeicherte Daten beträgt heute bestenfalls 10^{12} Bytes (1 Terabyte) pro Sekunde; 10 PB zu durchsuchen, dauert somit rund drei Stunden. Um übermäßige Wartezeiten zu vermeiden, sollte man deshalb den Daten Indexstrukturen überstülpen können, die es erlauben, jederzeit gezielt relevante Teilmengen auszuwählen.

Wenn Forschungsarbeiten auf der Auswertung verschiedener Datensammlungen beruhen, ist es zudem unabdingbar, dass die entsprechenden Publikationen eindeutig auf die zu Grunde liegenden Datenbestände verweisen. Dabei müssen Bestände und Software zur Auswertung auch für die Gutachter und andere Leser der Artikel zugänglich sein, weil eine Beurteilung solcher Veröffentlichungen anders nicht möglich ist.

Schließlich ist zu berücksichtigen, dass wissenschaftliche Projekte immer öfter gemeinsam von mehreren Instituten und Arbeitsgruppen durchgeführt werden. Jede Einrichtung erzeugt oder verarbeitet in diesem Fall einen Teil der Daten, wobei andere Kooperationspartner eventuell auf ihre Ergebnisse zugreifen. Da auch Urheberrechte und Fragen der wissenschaftlichen Priorität eine Rolle spielen, muss gewährleistet sein, dass keine Gruppe Daten einer anderen sehen kann, die diese nicht zur gemeinsamen Nutzung freigegeben hat. Eng damit verwandt ist die Forderung, dass alle Interaktionen der Wissenschaftler mit den Datenbeständen – wie Modelldefinitionen, Auswertungen, Veröffentlichungen und so weiter – automatisch zu einer Projektdokumentation zusammengeführt werden.

Allerdings sollen die Schutzvorkehrungen die Zusammenarbeit nicht behindern. Tat-

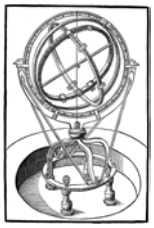
Größenvergleich

1 Petabyte = 10^{15} Bytes = 1 000 000 000 000 000 Bytes
Buch mit 330 Seiten: 1 Million = 10^6 Buchstaben (1 Buchstabe entspricht 1 Byte)
Library of Congress: Rund 31 Millionen Bücher (ohne Handschriften, Fotos und so weiter); 1 PB entspricht also dem Umfang von 10 Millionen Kongressbibliotheken.
Schnelle DSL-Leitung: 50 Mbit/Sekunde ≈ 8×10^6 Bytes/Sekunde
Transfer von 1 PB über diese Leitung: $1{,}25 \times 10^8$ Sekunden ≈ 1448 Tage ≈ 4 Jahre

Entwicklung der wissenschaftlichen Vorgehensweise

Bis vor rund 300 Jahren: Empirie

Wissenschaft beschränkt sich auf die empirische Beschreibung der Naturphänomene. Gelegentlich werden auch (empirisch abgeleitete) Rechenregeln entwickelt, etwa zum Erstellen von Kalendern.


AUS: TYCHO BRAHE, MECHANICA, 1602

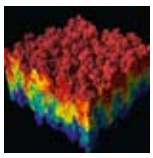
Seit 300 Jahren: Theorie

Forscher gehen dazu über, Naturphänomene zu generalisieren und in Form von (mathematischen) Modellen theoretisch erklärbar zu machen.

$$\left(\frac{\dot{a}}{a}\right) = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

Seit etwa 50 Jahren: Simulation

Naturphänomene wachsender Komplexität lassen sich mit zunehmender Genauigkeit auf Computern simulieren – oft unter Rückgriff auf mathematische Modelle.

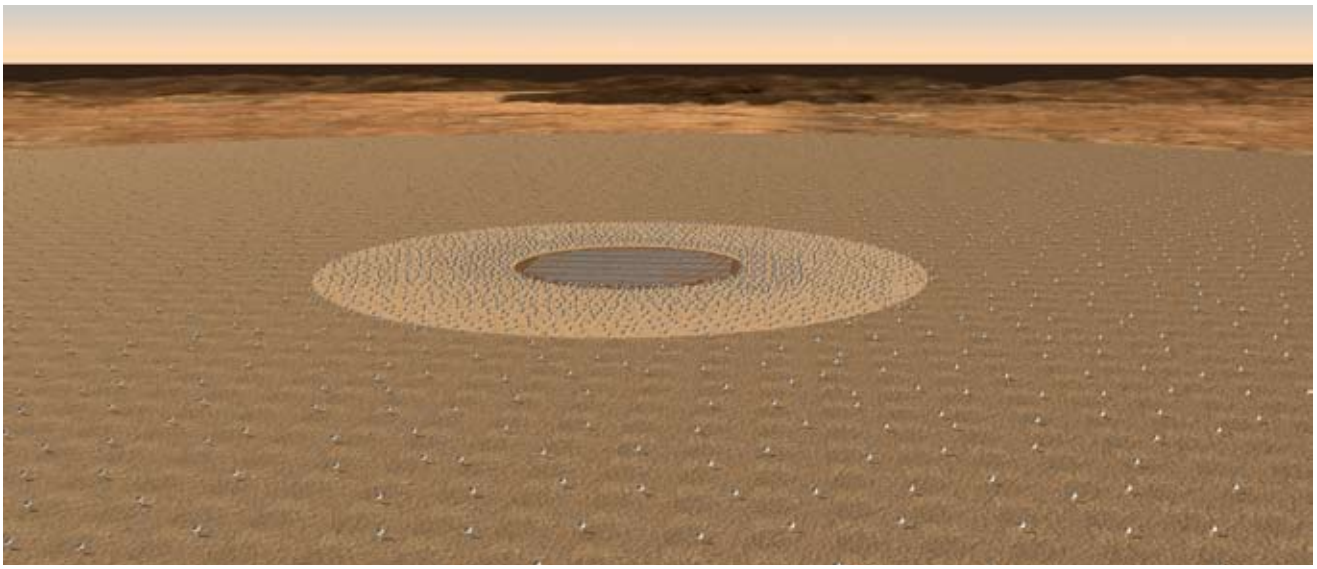

LLNL.GOV

Heute: Datengetriebene Wissenschaft

Experiment, Theoriebildung und Simulation werden zusammengeführt:

- Geräte und Simulationen erzeugen sehr große Mengen von Daten.
- Diese Daten werden durch Software aufbereitet.
- Die Daten und die daraus abgeleiteten Informationen werden in Computern gespeichert.
- Die Wissenschaftler analysieren die Datensammlungen mit Hilfe von Suchverfahren, statistischen Methoden, Visualisierungsverfahren und so weiter.


KLEINES FOTO: ESO, STÉPHANE GUISARD; RECHTS: BESSELFUNCTIONS, CC-BY-2.5



SKA / KILO STUDIOS

Das Square Kilometer Array, ein für 2024 geplantes System von Radioteleskopen, wird 1 Petabyte (PB) an Daten pro Tag liefern. Solch riesige Datenmengen lassen sich nicht mehr ohne äußerst leistungsfähige Computer und ausgefeilte Software auswerten.

sächlich scheuen viele Wissenschaftler immer noch davor zurück, ihre Ergebnisse in eine gemeinsam mit anderen genutzte Datenbank zu stellen, auch wenn es strikte Zugriffskontrollen gibt. Oft schicken dieselben Forscher ihre Daten freilich bedenkenlos per E-Mail an Kollegen, obwohl im Prinzip jeder deren Inhalt während der Übertragung mitlesen kann.

Schließlich muss sichergestellt sein, dass relevante Daten nicht durch Hardwareausfälle oder Bedienfehler verloren gehen können. Viele Fördereinrichtungen für Forschungsprojekte verlangen Mindestaufbewahrungsfristen für alle projektbezogenen Daten und Ergebnisse.

Das Ende pragmatischer Schnellschüsse

Heute werden die genannten Probleme oft in jedem Institut oder für jedes Projekt durch Rückgriff auf etwas halbwegs Brauchbares immer wieder von Neuem gelöst. Diese Ad-hoc-Lösungen sind in der Regel aber so spezifisch, dass sie schon für das nächste Projekt nicht mehr taugen (jedenfalls nicht vollständig). Außerdem legt jedes Labor und jede Projektgruppe eigene Regeln und Konventionen fest. Das macht die Übertragbarkeit der Daten oft schwierig bis unmöglich. Statt pragmatischer Schnellschüsse müssen in Zukunft also generische Lösungen her, die sich für eine große Klasse von Problemen und für unterschiedliche Auswertungsbedürfnisse eignen. Auf sie hinzuwirken, ist auch eine Aufgabe der nationalen und supranationalen Fördereinrichtungen. Anderenfalls wäre eine datenzentrierte Kooperation über verschiedene Disziplinen hinweg zum Scheitern verurteilt.

Im Zusammenhang mit der computergestützten Wissenschaft sind aber nicht nur methodische und Informatikprobleme zu lösen. So erfordert etwa die Möglichkeit zur Integration von Datenbeständen über die Grenzen von Projekten und Disziplinen hinweg die Definition von Standards möglichst großer Reichweite. Außerdem können Zentren zum Verwalten umfangreicher Datenbestände sowie die Hochleistungsrechner zu deren Bearbeitung nicht an jedem Institut oder auch nur an jeder Universität eingerichtet werden – das wäre viel zu teuer. Sinnvoll ist eine hierarchische Organisation mit wenigen Supercomputerzentren an der Spitze, einigen »großen« Zentren darunter und vielen Institutsservern auf der dritten Stufe.

Der Aufbau solcher nationalen oder besser noch internationalen Kooperationsstrukturen ist naturgemäß auch ein politisches Thema, in das Standortpräferenzen und Prestigefragen hineinspielen. Immerhin laufen bereits die erforderlichen Abstimmungsprozesse in Deutschland, Europa, den USA, Australien oder China. Das nächste Ziel für die Spitze der Hierarchie ist jedenfalls schon definiert: ein Rechner, der rund 1000-mal so schnell arbeitet wie der heutige Rekordhalter, also eine Leistung im Bereich von Exaflops (10^{18} Rechenoperationen pro Sekunde) erbringt.

Die Informationstechnologie hat somit eine ganze Reihe von Problemen zu lösen, um der modernen, datengetriebenen Wissenschaft gerecht zu werden –, und eines der schwierigsten, die Parallelverarbeitung auf Millionen von Rechenknoten, habe ich nicht einmal angesprochen. Wichtig ist, dass die Werkzeugentwicklung auf Seiten der Informatik Hand in Hand mit Methodenentwicklung auf Seiten der Wissenschaft geht. Denn nur so funktioniert jenes Wechselspiel, das seit jeher Triebfeder des wissenschaftlichen Fortschritts war: Neue Methoden stellen neue Anforderungen, und neue technische Möglichkeiten eröffnen den Weg zu neuen Methoden.

DER AUTOR

Andreas Reuter ist Professor für Informatik an der Universität Heidelberg und Geschäftsführer des Heidelberger Instituts für Theoretische Studien (HITS).

QUELLEN

Bell, G. et al.: Petascale Computational Systems: Balanced Cyber-Infrastructure in a Data-Centric World. Letter to NSF Cyberinfrastructure Directorate. In: IEEE Computer 39, S. 110–112, 2006

Hey, T. et al.: The Fourth Paradigm – Data-Intensive Scientific Discovery. Microsoft Corporation, 2009

Der Kosmos im Computer

Die Arbeitsgruppe »Theoretische Astrophysik« schlägt eine Brücke vom Universum kurz nach dem Urknall bis zur Gegenwart. In den fortgeschrittensten Supercomputersimulationen untersuchen die Forscher, wie sich aus der einst homogen verteilten Materie die heutige Vielfalt von Galaxien entwickeln konnte.

Von Volker Springel

Astronomie und Astrophysik beschäftigen sich mit dem wohl größten aller denkbaren Forschungsgegenstände: dem Universum als Ganzem. Tatsächlich sprengen die Dimensionen der Zahlen in diesen Disziplinen die menschliche Vorstellungskraft und Erfahrungswelt. Welche physikalische Größe man auch betrachtet – ob Temperatur, Dichte, Druck oder Magnetfeldstärke –, im Universum finden wir dafür fast durchweg Zahlenwerte, die um viele Größenordnungen über allem liegen, was wir auf der Erde und in unseren Laboratorien je werden messen können.

Schon grundlegende Tatsachen über den Kosmos übersteigen unseren Erfahrungshorizont. Wir wissen heute, dass das Universum etwa 13,6 Milliarden Jahre alt ist, dass dieses Raumzeitgebilde expandiert und dass sich die Expansion sogar immer weiter beschleunigt. Wir wissen, dass Sterne viele hundert Millionen Jahre lang leben – aber nicht ewig –, dass Planeten um andere Sterne eher die Regel als die Ausnahme sind und dass große Galaxien gewaltige Schwarze Löcher beherbergen.

Vielleicht die größte Zumutung, welche die moderne Kosmologie für unseren Verstand bereithält, ist aber die Entdeckung, dass das Universum vor allem so genannte Dunkle Materie und Dunkle Energie enthält. Erstere besteht aus einer bislang noch nicht nachgewiesenen Teilchenart, die sich vor allem durch ihre Schwerkraftwirkung verrät. Die Dunkle Energie ist noch rätselhafter. Forscher machen sie für die beschleunigte Ausdehnung des Kosmos verantwortlich.

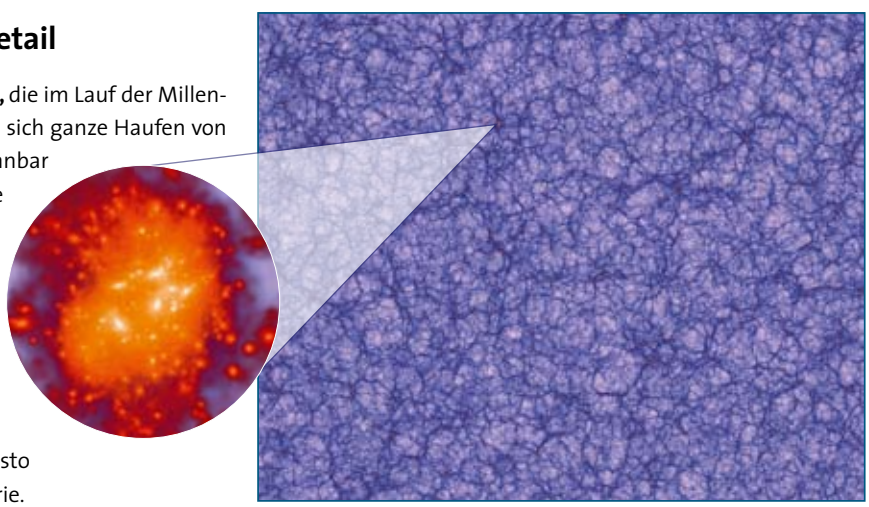
Im Universum dominieren also keineswegs die Atome der »normalen«, so genannten baryonischen Materie. Vielmehr repräsentiert der Stoff, aus dem wir selbst ebenso wie Sterne und Galaxien bestehen, gerade einmal vier Prozent der kosmischen Energiedichte.

Diese Erkenntnis verdanken wir dem Lambda-CDM-Modell *(Lambda Cold Dark Matter)*, das als Standardmodell der Kosmologie gilt. Als umfassende Theorie des Universums erklärt es eine Vielzahl astronomischer Daten und macht auch genaue Voraussagen darüber, wie das All unmittelbar nach dem heißen Urknall vor 13,6 Milliarden Jahren ausgesehen hat. Zu jener Zeit waren Materie und Strahlung fast perfekt gleichmäßig verteilt, abgesehen von winzigen Abweichungen, den Folgen von Quantenfluktuationen in einer frühen Phase des Urknalls. Diese lassen sich noch heute messen, denn sie sind dem extrem gleichmäßigen »Hintergrund« aus Mikrowellenstrahlung aufgeprägt, der das All erfüllt. Die Astronomen vermuten, dass die Schwankungen gleichsam die Saatkörner für alle späteren von der Schwerkraft geformten Materiestrukturen im Universum darstellen.

Um die Entstehung dieser Strukturen zu untersuchen, sind wir mittlerweile nicht mehr allein auf Beobachtungen angewiesen. Vielmehr haben sich Computersimulationen als außerordentlich wichtiges neues Forschungsinstrument etabliert. Dank ihrer Hilfe lassen sich komplexe physikalische Gleichungssysteme lösen, ohne dass wir auf Vereinfachungen zurückgreifen müssen, welche die Ergebnisse verfälschen. Auch virtuelle astrophysikalische Experimente sind nun möglich. Im Computer können wir beispielsweise zwei Galaxien kollidieren und miteinander verschmelzen

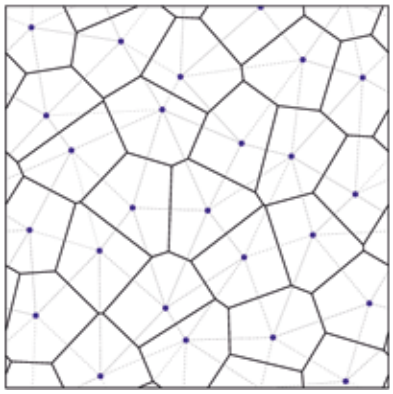
Vom großen Ganzen zum Detail

In den Filamenten aus Dunkler Materie, die im Lauf der Millennium-XXL-Simulation entstehen, bilden sich ganze Haufen von Galaxien, im Bildausschnitt rechts erkennbar als kleine, helle Flecken. Die Kantenlänge dieses zweidimensionalen Ausschnitts aus der Simulation beträgt mehrere Milliarden Lichtjahre. Zoomt man in sie hinein (kreisförmiger Bildausschnitt, Durchmesser rund 20 Millionen Lichtjahre), sieht man die Materieansammlungen in höherer Auflösung. Je heller hier die Bildpunkte, desto größer ist die Dichte der Dunklen Materie.



ALLE ABBILDUNGEN DIESES ARTIKELS: VOLKER SPRINGEL

Dem Geschehen dynamisch auf der Spur



Um Gase oder Flüssigkeiten in einem Volumen zu untersuchen, kann man den Raum in statische Zellen unterteilen. Besser ist jedoch ein Voronoi-Gitter (Grafik), wie es der Simulationskode AREPO für die Berechnung von strömenden Gasen verwendet. Jede Zelle umfasst den Raumbereich, der dem zugehörigen Punkt am nächsten liegt. Die Wände zwischen den Zellen sind die Ebenen, welche die Verbindungsstrecken (gepunktet) benachbarter Punkte in der Mitte senkrecht durchschneiden. Verschiebt man die Punkte mit der lokalen Gasgeschwindigkeit, verändert sich das Gitter dynamisch. Die räumliche Auflösung des Verfahrens ist dadurch gerade dort besonders hoch, wo viel geschieht.

lassen. Solche Verschmelzungsprozesse spielten eine entscheidende Rolle beim Aufbau immer größerer Galaxien. Während wir sie in der Natur nie beobachten können – schließlich benötigen sie Jahrmilliarden –, lassen sie sich nun am Rechner simulieren.

Genau solchen Experimenten widmet sich meine Arbeitsgruppe »Theoretische Astrophysik« am Heidelberger Institut für Theoretische Studien (HITS). Mit ihnen wollen wir eine Brücke vom Universum kurz nach dem Urknall, als es sich durch nur wenige Parameter vollständig beschreiben ließ, bis zu seinem heutigen komplexen Zustand schlagen. Vor allem haben wir uns zum Ziel gesetzt, das Phänomen der Galaxienbildung über die gesamte Zeit seit dem Urknall aufzuklären.

Bei der Entstehung von Galaxien ist ein außerordentlich breites Spektrum an physikalischen Prozessen im Spiel. Es reicht von der Dynamik der Dunklen Materie und der Dunklen Energie über Vorgänge bei der Sternentstehung bis hin zur Entwicklung superschwerer Schwarzer Löcher, zu elektromagnetischen Strahlungsprozessen und zur Magnetohydrodynamik. Es sind vor allem Computersimulationen, welche diese Komplexität berechenbar machen.

Ein aktuelles Beispiel dafür ist die Millennium-XXL-Simulation, die wir unlängst mit Kollegen des internationalen Virgo-Konsortiums auf dem JuRoPa-Supercomputer am Forschungszentrum Jülich durchgeführt haben. In diesem Modell verfolgen wir die Entstehung kosmischer Strukturen wie Galaxien und Galaxienhaufen. Es enthält 303 Milliarden (6720^3) Dunkle-Materie-Bausteine, die eine würfelförmige Raumregion mit einer Kantenlänge von weit mehr als zehn Milliarden Lichtjahren erfüllen. Die Dunkle-Materie-Bausteine unserer Simulation sind dabei nicht als Elementarteilchen zu verstehen. Vielmehr ist jeder einzelne von ihnen ein fiktives Makropartikel mit einer Milliarde Sonnenmassen.

Die weltgrößte kosmologische Simulation

Ihre Auflösung und ihr Volumen machen Millennium-XXL, die ihren Vorgänger darin um den Faktor 30 übertrifft, zur weltweit größten kosmologischen Simulation überhaupt. Sie liefert unerreicht genaue statistische Daten über die großräumige Struktur des Kosmos und die Entstehungsgeschichte von etwa 500 Millionen Galaxien.

Diese Daten sind unerlässlich, um zukünftige Beobachtungsprogramme, welche die zeitliche Entwicklung der Dunklen Energie im Universum und ihre physikalische Natur ergründen sollen, zu kalibrieren und systematische Fehlerquellen auszuschließen. Der Grundgedanke besteht darin, dass die beobachtbare Galaxienverteilung Rückschlüsse auf die tatsächliche Materieverteilung im Universum zulässt. Welcher Art diese Beziehung ist, die vom Galaxientyp und auch von der Zeit abhängt, können wir dank der Simulationsdaten genau untersuchen. In der Materieverteilung finden wir wiederum so genannte baryonische akustische Oszillationen, die ihrerseits eine Signatur der Expansionsgeschichte des Alls darstellen und damit wichtige Hinweise auf die Eigenschaften der Dunklen Energie geben.

Seit den frühen 1980er Jahren hat sich die Zahl der Teilchen in den jeweils größten kosmologischen Simulationen etwa alle eineinhalb Jahre verdoppelt. Diesem langjährigen Trend zufolge hätte die Millennium-XXL-Simulation erst im Jahr 2015 möglich sein sollen. Dass sie schon heute realisiert wurde, ist unseren neuen Strategien im Umgang mit extrem großen Datenmengen und den darauf aufbauenden Galaxienmodellen zu verdanken. Sie fanden ihren Niederschlag zum einen in einer speziell angepassten Version unserer Simulationssoftware GADGET3. Zum anderen reizten wir die Möglichkeiten des JuRoPa-Supercomputers am Forschungszentrum Jülich voll aus.

Dort schufteten alles in allem 12 288 Prozessoren gemeinsam an der Rechnung. In insgesamt fast drei Millionen Arbeitsstunden führten sie 86 Trilliarden Kraftberechnungen aus. Jede einzelne davon ermittelt die gravitative Wechselwirkung eines einzelnen Dunkle-Materie-Bausteins mit allen anderen Komponenten der Simulation. Dank der Parallelisierung der Berechnungen erhielten wir das Ergebnis schon nach 9,3 Tagen. Ein gewöhnlicher Computerprozessor, der eine Rechnung nach der anderen ausführt, hätte dazu gut 300 Jahre benötigt.

Einer der wichtigsten Faktoren, welche die Größe solcher Simulationen beschränken, ist der Speicherbedarf. Für unseren neuen Kode entwickelten wir daher auch besonders speichereffiziente und schnelle Berechnungsverfahren. Am Ende benötigte die Rechnung für die 303 Milliarden Teilchen dennoch fast 30 Terabyte oder 30 000 Gigabyte Hauptspeicher, womit wir den uns zugeteilten Speicher des Superrechners vollständig ausnutzten.

Das riesige Volumen der Millennium-XXL-Simulation erlaubt es, auch extrem seltene Ereignisse und Objekte aufzuspüren, beispielsweise sehr massereiche Galaxienhaufen. Das Lambda-CDM-Modell sagt voraus, dass die Masse von Galaxienhaufen eine recht scharf definierte Obergrenze im Bereich von einigen 10^{15} Sonnenmassen besitzt. In jüngster Zeit wurden tatsächlich einige Exemplare entdeckt, die recht nahe an dieser Grenze liegen. Manche Forscher behaupten sogar, sie lägen bereits darüber. In der Millennium-

XXL-Simulation bilden sich tatsächlich auch Galaxienhaufen, die ein wenig mehr Masse besitzen. Noch besteht daher kein offensichtlicher Grund zur Besorgnis: Alle Galaxienhaufen, die je beobachtet wurden, lassen sich weiterhin mit dem kosmologischen Standardmodell erklären. Doch schon die Entdeckung eines einzigen Haufens, dessen Masse diese Grenze deutlich überschreitet, könnte es widerlegen.

Eher ein Gas als eine Flüssigkeit

Trotz ihrer beeindruckenden Größe besitzt die Millennium-XXL-Simulation einen Nachteil: Über kleinräumige Strukturen und Vorgänge in einzelnen Galaxien trifft sie nur wenige Aussagen. Schließlich ist selbst ein Objekt von der Größe der Milchstraße durch gerade einmal 1000 Bausteine repräsentiert. Hinzu kommt: Unsere Simulation behandelt die normale baryonische Materie der Einfachheit halber als stoßfreies Fluid; als einzige Wechselwirkung ist also die Schwerkraft berücksichtigt. Tatsächlich unterliegt die Materie aber Druckkräften und verhält sich damit eher wie ein ideales Gas. Außerdem kann sie thermische Energie verlieren, indem sie Strahlung abgibt. Unter der Wirkung der Schwerkraft kann sie also, weil sie von Hitze weniger stark auseinandergetrieben wird, noch viel stärker verklumpen als Dunkle Materie.

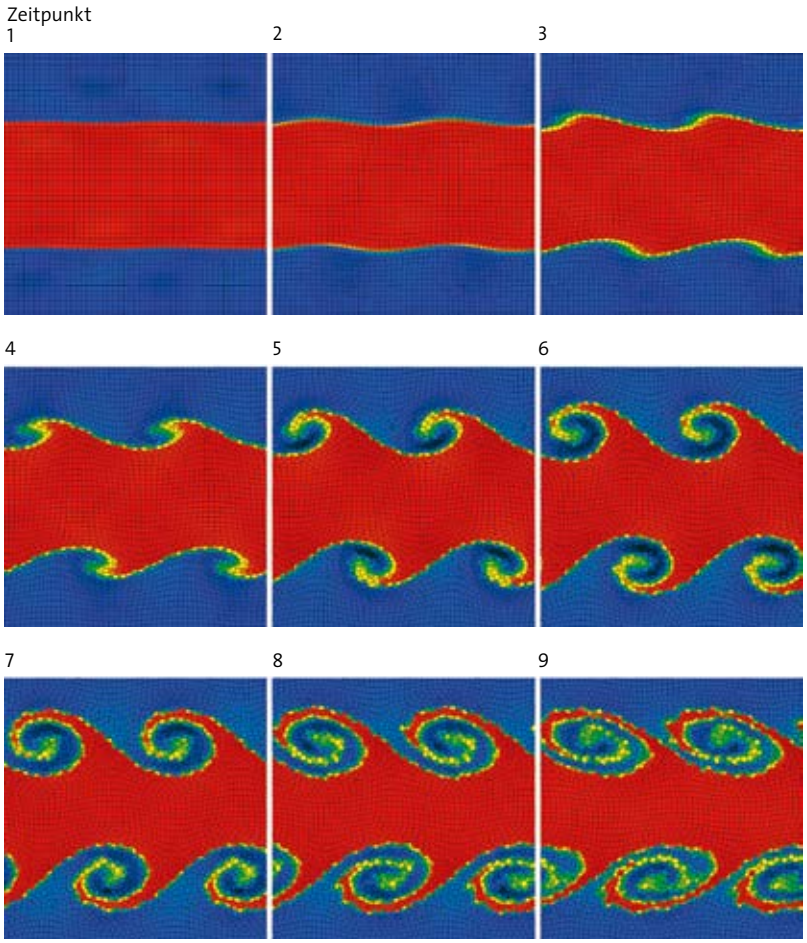
Diese Unterschiede von baryonischer und Dunkler Materie werden auf kleinen Skalen wichtig. Wir müssen also die baryonischen Prozesse korrekt simulieren, wenn unser Modell auch über die inneren Regionen von Galaxien Aussagen treffen soll. Die Berechnung des hydrodynamischen Verhaltens normaler Materie erweist sich allerdings als ausgesprochen anspruchsvoll. Die typische Dichte des Wasserstoff- und Heliumgases, das sich zu sternbildenden Galaxien verdichtet, ist sehr niedrig. Ein solches ideales Gas, in dem praktisch keine innere Reibung stattfindet, neigt über einen sehr weiten Skalenbereich hinweg stark zu Turbulenzen. Zudem führen große Unterschiede in Temperatur, Dichte und Geschwindigkeit zu gewaltigen Überschallströmungen. Und schließlich »spürt« auch jedes Teilchen im Gas die Schwerkraft aller anderen Gaspartikel. Während diese so genannte Eigengravitation bei strömungsmechanischen Problemen auf der Erde völlig vernachlässigbar ist, gewinnt sie in der Astrophysik entscheidende Bedeutung. So kontrahiert etwa eine Gaswolke nur deshalb allmählich zu einem Stern, weil sich die Teilchen gegenseitig anziehen.

Astrophysiker müssen also neue Wege gehen, um geeignete numerische Verfahren für die Kosmologie zu entwickeln. Die zentrale Idee des Ansatzes zur Simulation baryonischer Gase, den unsere Gruppe entwickelt hat, ist der Einsatz eines unstrukturierten Gitters, das im Unterschied zu herkömmlichen Verfahren nicht stationär ist, sondern sich mit dem Gas mitbewegen kann. Dadurch lässt sich genau dort, wo die relevanten Prozesse stattfinden, eine hohe Auflösung erzielen. Bei der neuen Methode gehen wir von einem Satz von Punkten im Raum aus, die jeweils von einer so genannten Voronoi-Zelle umgeben sind. Diese besteht einfach aus derjenigen Raumregion, die näher an diesem Punkt liegt als an irgendeinem anderen. Gemeinsam bilden die Voronoi-Zellen dann ein Voronoi-Gitter, das den Raum gewissermaßen pflastert. Die Wände zwischen den Zellen sind die Ebenen, welche die Verbindungsstrecken benachbarter Punkte in der Mitte senkrecht durchschneiden (siehe Abbildung S. 11). Nun kann man, während sich Gestalt und Topologie des Gitters kontinuierlich ändern, die Bewegung der einzelnen Punkte der lokalen Bewegung des Gases anpassen.

Darüber hinaus gelang es uns, ein so genanntes Godunov-Verfahren höherer Ord-

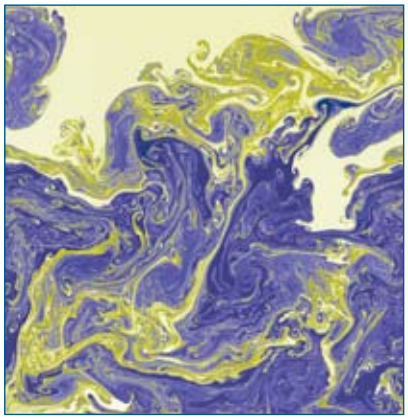
Klügere Algorithmen, weniger Artefakte

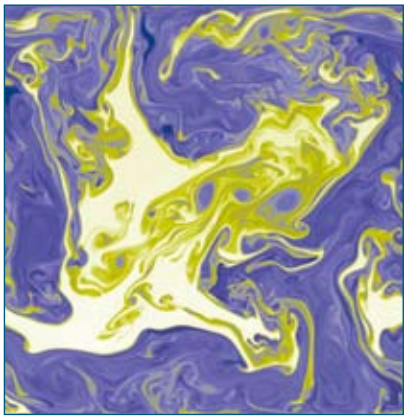
Bewegen sich zwei Phasen eines Gases aneinander vorbei – im Beispiel fließt eine dichte Phase (rot) nach rechts, eine weniger dichte (blau) nach links –, entsteht eine so genannte Scherströmung, die zu typischen Kevin-Helmholtz-Wirbeln führt (unterste Zeile). Ein dynamisch mitbewegtes Voronoi-Gitter (schwarz umrandete Gitterzellen) erlaubt es, sie korrekt und ohne Artefakte darzustellen.



Näher an der Realität

Rayleigh-Taylor-Instabilitäten führen dazu, dass sich zwei Phasen eines Fluids turbulent miteinander vermischen. Das Bild links zeigt das Simulationsergebnis bei mitbewegtem Gitter, rechts kam ein traditionelles festes Gitter zum Einsatz. Letzteres führt zu größeren Advektionsfehlern, so dass sich die simulierten Fluide lokal viel stärker als in der Realität vermischen. Auch die feine Schichtung der Phasen geht früher verloren.





nung auf dem bewegten Gitter zu implementieren. Mit seiner Hilfe können wir mit analytischen Methoden bestimmen, wie viel Masse, Energie und Impuls eine Zelle nach jedem Zeitschritt enthält.

Mitfließende Gitter

Der wesentliche Vorteil ist dabei der lagrangesche Charakter der Methode. Wenn irgendwo im Universum eine neue Galaxie entsteht und sich die Gasdichte in dieser Region millionenfach erhöht, dann fließt das Gitter automatisch mit. Es erlaubt also genau dort eine stark erhöhte räumliche Auflösung, wo die Galaxie entsteht. Daneben erweisen sich die Zahlenwerte der Ergebnisse, anders als in traditionellen Gittermethoden, als vollständig unabhängig vom verwendeten Bezugssystem.

Das fließende Gitter verringert zudem Advektionsfehler. Zu diesem Typ von Berechnungsfehler kommt es, wenn ein Masseteilchen mit der Strömung mitgeführt wird und dabei nicht vollständig, sondern nur teilweise von einer Zelle in die nächste übertritt, so dass es zu einem unerwünschten Ausschmieren der Strömung kommt. Wegen der diskreten Struktur des Gitters lässt sich dieser Vorgang mathematisch nicht exakt darstellen. In einem bewegten Gitter kann die Zelle hingegen passend mitbewegt werden, so dass sich viele Advektionsfehler von vornherein vermeiden lassen und ein künstliches Mischen in hohem Maß verhindert wird.

Ein Beispiel zeigen die Bilder links. Hier strömen unterschiedlich dichte Gase aneinander vorbei. Dabei wachsen kleine Störungen an der Grenzfläche schnell zu wellenartigen Wirbeln heran, welche die beiden Phasen schließlich turbulent miteinander vermischen. Will man diese so genannten Kelvin-Helmholtz-Instabilitäten numerisch beschreiben, führen Advektionsfehler in der Regel dazu, dass sich die Phasen im Modell früher vermischen als in der Realität. Indem wir diese Fehler stark reduzieren, können wir Überschallströmungen und Turbulenzen mit größerer Präzision darstellen (Bilder links und oben). Deshalb wollen wir das neue Verfahren auch in unserer Simulationssoftware AREPO einsetzen. An ersten Rechnungen dieser Art arbeiten wir bereits intensiv, sowohl mit Kollegen am Harvard Center for Astrophysics in Cambridge (Massachusetts) als auch im Virgo-Konsortium.

Außerdem wollen wir in der nächsten Zeit endlich die Entstehung von Spiralgalaxien besser verstehen lernen. Sternsysteme dieses Typs sind zwar die häufigsten im Universum, doch in bisherigen Simulationen bildeten sich fast ausschließlich elliptische Galaxien. Wir vermuten die Gründe dafür in einem unzureichenden Verständnis der Regulation der Sternentstehung durch bestimmte astrophysikalische Prozesse wie etwa die Explosion von Sternen als Supernovae. Auch die mangelnde Genauigkeit der bisher eingesetzten numerischen Methoden spielt eine Rolle. Zumindest dieses zweite Problem wird unser neuer AREPO-Kode möglicherweise lösen können.

Die vielleicht größte Aufgabe der Kosmologen besteht in diesen Jahren aber darin, die Rätsel um die Dunkle Seite des Kosmos aufzuklären. Mit unseren Simulationen versuchen wir, sie dabei zu unterstützen – indem wir physikalische Modelle überprüfen helfen, die eines Tages unser gesamtes Universum beschreiben könnten.

DER AUTOR

Volker Springel hat in Tübingen und an der University of California in Berkeley Physik studiert und im Jahr 2000 an der Ludwig-Maximilians-Universität München promoviert. Als Postdoc war er an der Harvard University in Cambridge (Massachusetts) und am Max-Planck-Institut für Astrophysik in Garching, wo er anschließend bis 2010 eine Forschungsgruppe zur numerischen Kosmologie leitete. Seither ist er Professor für Theoretische Astrophysik an der Universität Heidelberg. Hier forscht er am Heidelberger Institut für Theoretische Studien (HITS) und am Astronomischen Recheninstitut des Zentrums für Astronomie.

QUELLEN

Springel, V.: E pur si muove: Galilean-invariant Cosmological Hydrodynamical Simulations on a Moving Mesh. In: Monthly Notices of the Royal Astronomical Society 401, S. 791–851, 2010. Vorab publiziert auf **http://arxiv.org/abs/0901.4107**

Vogelsberger, M. et al.: Moving Mesh Cosmology: Numerical Techniques and Global Statistics. Eingereicht. Vorab publiziert auf **http://arxiv.org/abs/1109.1281**

WEBLINK

www.h-its.org/tap
Details zu Millennium-Simulationen und weiteren Simulationsprojekten der HITS-Arbeitsgruppe Theoretische Astrophysik

Das biomolekulare Erkennungspuzzle

Proteine sind die Funktionsträger des Lebens. Ihre Wechselwirkungen miteinander und mit anderen Biomolekülen sorgen dafür, dass Zellen ihre Aufgabe im Organismus erfüllen. Um diese Wechselwirkungen besser zu verstehen, setzen Forscher zunehmend rechnergestützte Methoden ein. Computersimulationen von Proteininteraktionen leisten auch einen immer wichtigeren Beitrag zum Design von Wirkstoffen gegen Krankheiten und in der Biotechnologie.

Von Rebecca C. Wade

In einer Zelle wimmelt es geradezu von großen und kleinen Molekülen, die ständig in Bewegung sind. Wie finden und erkennen sie in diesem Gewirr ihre jeweiligen Bindungspartner? Wie können sie mit mehreren anderen Molekülen zusammen Komplexe bilden? Und wie kommt es, dass manche dieser Vorgänge schnell und andere langsam ablaufen? Bei der Suche nach Lösungsstrategien für das Puzzle der biomolekularen Erkennung helfen neben ausgeklügelten Experimenten und biochemischen Untersuchungen vermehrt Berechnungen und Simulationen am Computer. Mit ihnen befassen wir uns in der Arbeitsgruppe »Molekulare und zelluläre Modellierung« am Heidelberger Institut für Theoretische Studien.

Betrachten Sie zum Beispiel ein Puzzle aus 2000 Teilen, das ein Schloss in einer schönen Landschaft zeigt. Einige Plättchen lassen sich ganz einfach platzieren: Flaggen, Turmspitzen oder auch Mauerkanten. Bei anderen hilft nur geduldiges Probieren. Das gilt etwa für grünliche oder bräunliche Teile, die zu den Bäumen im Wald gehören, oder für solche in den verschiedenen Blautönen des Himmels.

Bei der Bindung zwischen Biomolekülen spielt wie im Puzzle die Passform eine wesentliche Rolle. Dies erkannte vor über einem Jahrhundert bereits Emil Fischer, der die Wechselwirkungen zwischen Enzymen und Substraten mit dem Bild von Schlüssel und Schloss beschrieb. Doch wie bei den Puzzleteilen reicht die Gestalt nicht aus, um alle möglichen Wechselwirkungen eindeutig zu beschreiben. Einige Moleküle tragen gut definierte »Flaggen«, die ihre Position in der Zelle oder ihre Beziehung zu anderen Stoffen bestimmen. So legen etwa bei manchen Proteinen kurze lineare Sequenzmotive fest, an welcher Stelle sich eine andere Substanz anlagern kann. In vielen Fällen jedoch ist weniger offensichtlich, woran Moleküle einander erkennen.

Wie stark und selektiv sich zwei Substanzen aneinander binden, hängt von der freien Energie der betreffenden Bindung ab. Diese wiederum setzt sich aus verschiedenen Komponenten zusammen. Das Problem ist, dass diese oft groß sind und teils entgegengesetzte Wirkungen haben. Aus diesem Grund bedarf es sehr genauer Berechnungen, um aus den Einzelkomponenten die (häufig sehr kleine) Summe korrekt zu ermitteln.

Eine weitere Schwierigkeit liegt darin, dass die relative Bedeutung der Komponenten von Fall zu Fall variiert, was es schwer macht, ein allgemein gültiges Computermodell für ein solches Problem zu entwickeln. So dominieren bei einer Bindung zwischen Proteinen manchmal weit reichende elektrostatische Kräfte, dank deren sich Moleküle auch über größere Distanzen hin finden. Manchmal spielen sie dagegen kaum eine Rolle. In solchen Fällen leisten zum Beispiel anziehende Kräfte zwischen hydrophoben (Wasser meidenden) Gruppen, die nur eine geringe Reichweite haben, den größten Beitrag zur Bindungsstärke. Das Problem der genauen Beschreibung der physikochemischen Wechselwirkungen zwischen Molekülen – sei es mit einer auf physikalischen Gesetzmäßigkeiten basierenden Energiefunktion oder einer rein empirisch aufgestellten Funktion – wird gewöhnlich als Scoring-Problem bezeichnet.

Eine weitere Herausforderung ist das so genannte Sampling-Problem. Schon beim Puzzle gibt es unzählige denkbare Kombinationen der einzelnen Plättchen – und der Spieler bemüht sich, die Möglichkeiten einzugrenzen, um die Anzahl der vergeblichen Versuche beim Einpassen eines Teils zu verringern. Ein Puzzle ist jedoch nur ein zweidimensionales Objekt. Das Durchprobieren

Suche nach Enzymhemmern am Computermodell

Auf dem Strukturbild eines Enzyms namens LmPTR1, das im *Leishmania*-Parasiten vorkommt und sich deshalb als Angriffspunkt für Medikamente gegen die Leishmaniose eignet, ist die Oberfläche der vier identischen Untereinheiten in verschiedenen Farben dargestellt (links). An einem der aktiven Zentren haftet sein gewöhnliches Substrat, ein Molekül namens Pteridin (dunkelviolett), zusammen mit dem Kofaktor (NADPH, türkis). Die Ausschnittvergrößerung (rechts) zeigt die Bindungstasche des Enzyms (graue Moleküloberfläche) mit zwei daran angelagerten potenziellen Hemmstoffen. Farbig hervorgehoben sind Proteinregionen, die laut Berechnung die Bindung Wasser abweisender (gelb) oder Wasser liebender funktioneller Gruppen (blau) begünstigen. Die Wirkstoffkandidaten (gelb, hellblau) lagern sich zwischen dem Kofaktor (grau) und einer aromatischen Aminosäure des Proteins ein (alle drei als Stäbchenmodell dargestellt).

möglicher Konstellationen in einem dreidimensionalen biomolekularen System mit ungleich mehr »Teilen« stellt noch viel höhere Anforderungen.

So hat jedes Teilchen im Raum drei Freiheitsgrade für die Translation und drei für die Rotation. Hinzu kommt, dass die Moleküle nicht starr wie Puzzleteile sind, sondern auf Grund thermischer Bewegungen ständig ihre Gestalt ändern. Auch können sie, wenn sie eine Bindung eingehen, ihre Form aneinander anpassen. Biomolekulare Systeme haben also extrem viele Freiheitsgrade, was die Entwicklung detaillierter Modelle erschwert. Diese müssen schließlich alle für die molekulare Erkennung relevanten Variablen genau genug berücksichtigen, ohne dabei die Möglichkeiten des Computers zu überschreiten.

Viele Wege führen zum Modell

Es gibt verschiedene Ansätze zur Konstruktion von Modellen, mit denen sich die Erkennung zwischen Biomolekülen simulieren und vorhersagen lässt. Hier möchte ich auf die zwei gebräuchlichsten näher eingehen, die sich auch miteinander kombinieren lassen.

Die erste Strategie folgt dem bioinformatischen Ansatz. Die Grundlage sind hier experimentelle Ergebnisse, die in eigens dafür angelegten Datenbanken gesammelt werden. Dabei handelt es sich etwa um Molekülstrukturen oder um die Abfolge der Aminosäuren von Proteinen oder die Basensequenz von Genen.

Die Datenbanken werden nun nach Übereinstimmungen beziehungsweise Unterschieden zwischen den Einträgen durchsucht. Findet man etwa Ähnlichkeiten in der Sequenz von Genen oder Proteinen, so deutet das darauf hin, dass die betreffenden Bereiche auf Grund ihrer Funktion während der Evolution weit gehend erhalten geblieben sind. Anhand solcher Sequenzmotive sowie der räumlichen Anordnung der Atome im Molekül gelingt es in einigen Fällen, Bindungsstellen zu identifizieren und die Position der Bindungspartner im Komplex vorherzusagen. Mit Hilfe der dreidimensionalen Molekülstrukturen und der wissensbasierten Analysemethoden können die Forscher dann die Bindungsaffinitäten zwischen Molekülen abschätzen.

Die wachsende Menge an genetischen und strukturellen Daten macht diese Strategie zwar zusehends leistungsfähiger, aber die Qualität ihrer Ergebnisse variiert stark mit den verwendeten Daten und hängt zudem davon ab, inwieweit es gelingt, die jeweils relevanten Informationen aus Datenbanken herauszufiltern.

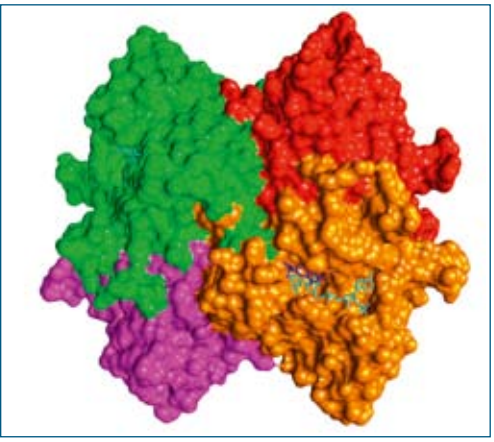
Die zweite Strategie nutzt physikalisch-chemische Prinzipien zur Modellierung biomolekularer Interaktionen. Dabei erstellen die Forscher mathematische Energiefunktionen, in die physikalische Bindungsfaktoren wie die Van-der-Waals-Wechselwirkungen oder elektrostatische Kräfte eingehen. Nur in wenigen Fällen lohnt es sich hierbei, auf die genauen, aber auch sehr rechenintensiven Methoden der Quantenmechanik zurückzugreifen.

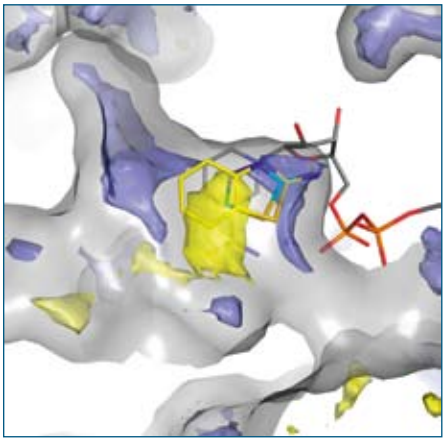
Üblicherweise beschränkt man sich auf den Einsatz molekularmechanischer Modelle, bei denen jedes Atom durch eine passend gewählte Kugel repräsentiert wird. Die Rolle der Bindungen zwischen den Atomen übernehmen Federn mit empirisch bestimmten Eigenschaften wie der Rückstellkraft.

Für die Simulation großer Systeme mit sehr vielen Atomen reicht oft eine weniger detaillierte Darstellung. In solchen so genannten *Coarse-Grain*-Modellen werden mehrere Atome, zum Beispiel Seitenketten von Proteinen oder sogar ganze Proteine, zu größeren Partikeln zusammengefasst und mit parametrisierten Interaktionsprofilen versehen.

Durch Verwendung geeigneter Energiefunktionen, wie sie bei Moleküldynamik- oder Monte-Carlo-Verfahren zum Einsatz kommen, lassen sich zudem thermische Bewegungen und Verformungen der Biomoleküle simulieren. Das erlaubt nicht nur die Vorhersage von Bindungsstellen und der Struktur von Molekülkomplexen, sondern auch eine Abschätzung der Bindungsstärke und -kinetik. Letztere beschreibt die Geschwindigkeit, mit der sich die Bindung bildet beziehungsweise auflöst.

Das Problem, wie sich Biomoleküle erkennen, gehört zwar zunächst in die Grundlagenforschung, ist aber auch beim gezielten Entwurf von Medikamenten und der Entwicklung künstlich modifizierter Proteine von großer Bedeutung. So werden die rechnergestützten Methoden zur Lösung dieses Problems besonders in der pharmazeutischen, agrochemischen und biotechnologischen Industrie eingesetzt. Dort leisten sie gute Dienste vor allem bei der Suche nach biomolekularen Interaktionen in der Wirkstoffentwicklung und bei der Vorhersage, wie sich Mutationen auf die Struktur und Eigenschaften von Proteinen auswirken. Hier möchte ich den Einsatz dieser computergestützten Methoden anhand unserer eigenen Arbeiten beschreiben. Diese reichen von der Medikamentenentwicklung bis zu Untersuchungen der DNA-Anordnung, der Oligomerisierungszustände von Proteinen und der Oberflächenaktivierung.





REBECCA C. WADE UND DOMANTAS MOTIEJUNAS

Wirkstoffe gegen Parasiten

Die Leishmaniose ist eine schwere Erkrankung, an der weltweit rund zwölf Millionen Menschen leiden. Sie tritt hauptsächlich in ärmeren Ländern der warmen Klimazonen auf. Auslöser sind einzellige Parasiten aus der Familie der Trypanosomatidae, die durch Bisse von Sandmücken übertragen werden. Heutige Medikamente sind nur bedingt wirksam und haben viele Nebenwirkungen. Außerdem ist der Erreger gegen viele von ihnen schon mehr oder weniger resistent.

Als aussichtsreicher Angriffspunkt für neuartige Arzneimittel gegen den Parasiten *Leishmania major* ließ sich eine Pteridinreduktase

namens LmPTR1 ausmachen. Sie gehört gemeinsam mit der Dihydrofolatreduktase (DHFR) zum Folatstoffwechselweg und ist wichtig für die DNA-Synthese. Wird sie zusammen mit DHFR gehemmt, kann der Parasit keine neue Erbsubstanz synthetisieren und sich folglich auch nicht vermehren.

Beiden Enzymen ist gemeinsam, dass sie sowohl den Kofaktor NADPH als auch das Substrat Folsäure (oder Abwandlungen davon) bei ihrer enzymatischen Aktivität verwenden. Im Gegensatz zur Dihydrofolatreduktase, die bei den Parasiten wie auch beim Menschen vorkommt, findet man die Pteridinreduktase jedoch nur beim Parasiten. Gelingt es nun, Verbindungen zu finden, die nicht aus der Stoffklasse der Folsäuren stammen, sich aber dennoch spezifisch an das parasitäre Enzym LmPTR1 heften, minimiert man das Risiko von Nebenwirkungen beim Menschen.

Die Kristallstruktur des Enzyms LmPTR1 war schon bekannt. Wir konnten sie also benutzen, um bei einem virtuellen Screening eine große Substanzbibliothek nach geeigneten Verbindungen zu durchsuchen, die gut in das aktive Zentrum des Enzyms passen und keine Ähnlichkeiten zu Folsäurederivaten aufweisen. Um die angestrebte Hemmwirkung zu erzielen, war es wichtig, dass die Ringstrukturen der Wirkstoffkandidaten zwischen denen des Kofaktors und den aromatischen Seitenketten des Proteins zu liegen kommen (siehe Kasten auf S. 14/15).

Wie Kollegen in Italien und Belgien anhand von Laborexperimenten zeigen konnten, hemmen einige der von uns identifizierten potenziellen Wirkstoffe tatsächlich die Enzymaktivität von LmPTR1. Um diese Substanzen zu optimieren, untersuchten wir in weiteren Simulationen, wie sich durch Austausch einzelner Atome oder Atomgruppen die Bindung an das aktive Zentrum des Enzyms verstärken lässt.

Zwei rechnerbasierte Entwicklungsdurchgänge und eine anschließende experimentelle Prüfung am isolierten Enzym lieferten so 18 spezifisch wirksame LmPTR1-Inhibitoren. Sechs davon hemmten nicht nur die Aktivität des isolierten Enzyms, sondern auch das Wachstum der Parasiten in Zellkultur. Eine dieser Substanzen entspricht sogar dem Wirkstoff eines Medikaments, das bereits zur Behandlung von Erkrankungen des Zentralnervensystems zugelassen ist. Möglicherweise lässt sich dessen Anwendungsbereich auf die Therapie parasitärer Erkrankungen ausweiten.

Dieses Projekt macht deutlich, wie sich unsere rechnergestützten Proteinsimulationen und die von unseren Kollegen in Italien und Belgien durchgeführten Laborexperimente erfolgreich ergänzen. Auch wenn solche computerbasierten Ansätze in der pharmazeutischen Industrie weit verbreitet sind, darf das nicht darüber hinwegtäuschen, dass Standardverfahren häufig Einschränkungen unterliegen und an das zu untersuchende Zielprotein speziell angepasst werden müssen. Beim LmPTR1 war es etwa entscheidend, dass wir vier Wassermoleküle im aktiven Zentrum des Proteins berücksichtigten. Dadurch gelang es, die für die Wirkstoffentwicklung wichtige korrekte Orientierung der Wirkstoffkandidaten zu ermitteln, auch wenn wir die Enzymaktivität beziehungsweise Bindungsstärke nicht zuverlässig vorhersagen konnten.

Raffinierte Packung der DNA

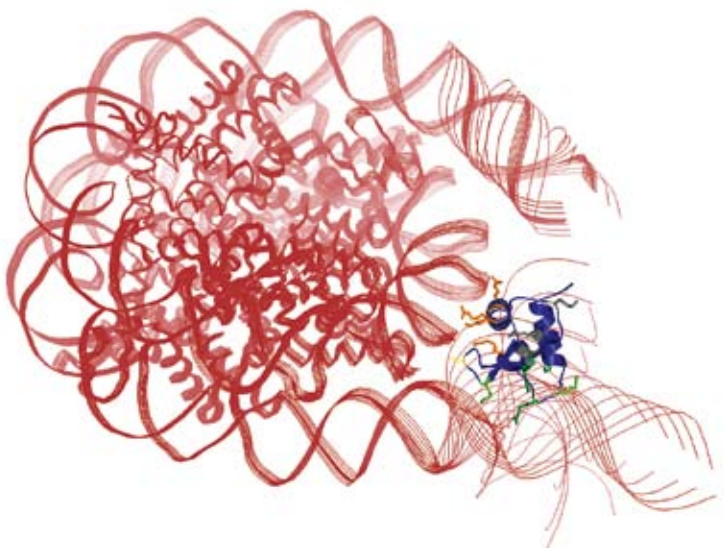
Der Kern einer eukaryotischen Zelle enthält fadenförmige DNA mit einem Durchmesser von etwa 10 bis 20 Mikrometern und einer Gesamtlänge von zwei Metern. Damit die Erbsubstanz überhaupt in die Zelle passt, muss sie zu einer kompakten Struktur, dem so genannten Chromatin, aufgewickelt werden. Um dies zu bewerkstelligen, benutzt die Zelle Histone: positiv geladene Proteine, die sich an die negativ geladenen Nukleinsäuren binden, aus denen die Erbsubstanz besteht.

Den Grundbaustein des Chromatins bilden die Nukleosomen, um deren Proteinkern sich die DNA spulenförmig wickelt. Zwischen ihnen erstrecken sich zunächst noch freiliegende Abschnitte des DNA-Fadens, die als Linker-DNA bezeichnet werden. An die Nukleosomen heften sich die so genannten Linker-Histone. Diese kleinen Proteine sorgen dafür, dass sich die perlschnurartige Nukleosomenkette zickzackförmig zusammenlagert oder wie eine Wendeltreppe windet und so die kompakten Chromatinfasern bildet. Ferner tragen sie dazu bei, das Abschreiben und Vervielfältigen der DNA zu regulieren.

Anders als die Ladungsunterschiede zwischen den Histonen und der DNA vermuten lassen, beruht die Bindung nicht nur auf elektrostatischen Wechselwirkungen. Wir wollten daher genauer wissen, wie sich die Linker-Histone an die Nukleosomen anlagern. Zu diesem Zweck untersuchten wir die Wanderung der kleinen Proteine zum Nukleosom, indem wir ihre brownsche Molekularbewe-

Erbfaden am Wickel

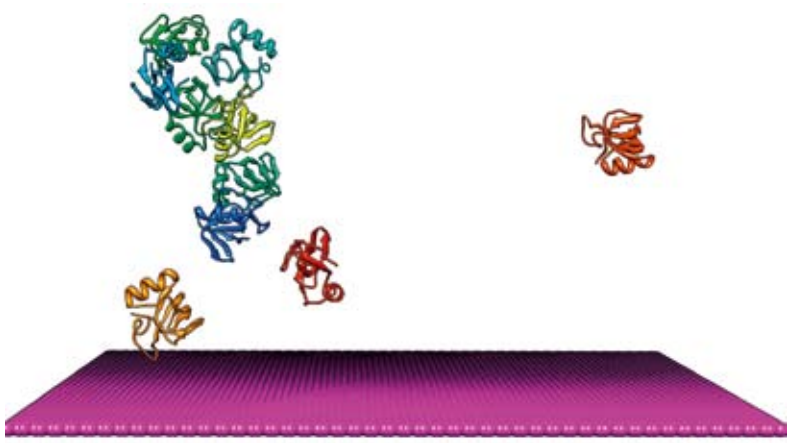
Simulationen ergaben, wie sich das Linker-Histon (blau) an das Nukleosom (braun) bindet und so zur Packung der DNA beiträgt. Unter Berücksichtigung der Flexibilität der beiden Enden des DNA-Stücks gibt es eine Reihe möglicher Anordnungen, von denen 13 als Überlagerung dargestellt sind. Die an die DNA gebundenen Aminosäurereste sind in Orange (nukleosomale DNA) und Grün (Linker-DNA) gezeigt.



GEORGI PACHOV UND REBECCA C. WADE

Der Haftkraft von Pilzsporen auf der Spur

Die Simulation der Diffusion von Hydrophobinmolekülen in wässriger Lösung in Gegenwart einer Graphitoberfläche half, ihre hohe Haftfähigkeit zu ergründen. Die einzelnen Proteinmoleküle sind in verschiedenen Farben dargestellt.



PAOLO MEREGHETTI

gung simulierten. Dabei konnten wir beobachten, wie elektrostatische Interaktionen das Linker-Histon anziehen und es so lenken, dass es sich in einer bestimmten Orientierung an das Nukleosom bindet. Außerdem sahen wir, wie die räumliche Gestalt des Nukleosoms den Vorgang beeinflusst.

In unseren Simulationen ließen wir das Linker-Histon an verschiedene Konformationen des Nukleosoms beziehungsweise seiner flankierenden Linker-DNA-Stücke anlagern und konnten so den vorherrschenden Bindungsmodus herausfinden. Dieser entspricht den experimentell ermittelten Daten. Es zeigte sich, dass das Linker-Histon asymmetrisch am Übergang zwischen der nukleosomalen DNA und einem der beiden Linker-DNA-Stücke haftet (siehe Kasten links). War die Konformation des Nukleosoms weniger kompakt, band es sich an einer stärker beengten Stelle an die Linker-DNA. Die Entdeckung zweier unterschiedlicher Bindungsarten deutet darauf hin, dass das Linker-Histon mit über die Struktur des Chromatins bestimmt. Indem es die Nukleosomen konformationsabhängig erkennt und stabilisiert, fördert es die Bildung enger Zickzack- oder lockerer Schraubenwindungen.

Damit das Durchprobieren vieler verschiedener Konformationen und Bindungsstellen die Kapazität unserer Computer nicht überstieg, nahmen wir für die Simulationen Vereinfachungen vor. Ausgehend von den errechneten diffusionsgetriebenen »Begegnungskomplexen« lassen sich nun mit einem verfeinerten Modell, das die Flexibilität der Makromoleküle vollständig berücksichtigt, weitere Details untersuchen. So kann jetzt beispielsweise bestimmt werden, wie die Partnermoleküle ihre Gestalt während des Bindungsvorgangs aneinander anpassen und welchen Einfluss dabei die nur auf kurze Distanzen wirksamen Wasserstoffbrückenbindungen und hydrophoben Wechselwirkungen haben.

Das Geheimnis extremer Oberflächenaktivität

Hydrophobine sind kleine Proteine mit der höchsten Oberflächenaktivität aller bekannten Eiweißstoffe. Sie kommen in der Hülle von Pilzsporen vor und haften auch an äußerst glatten Oberflächen. Wegen dieser Eigenschaft sind Hydrophobine für biotechnologische Anwendungen wie die Herstellung von Biosensoren oder die Immobilisierung von Enzymen von großem Interesse. Wir wollten wissen, wie sie sich in Lösung verhalten und worauf ihre extreme Haftfähigkeit an Oberflächen beruht. Dazu simulierten wir die brownsche Molekularbewegung von hunderten Hydrophobinmolekülen in wässriger Lösung in Gegenwart einer Graphitoberfläche.

Ausgehend von einer experimentell ermittelten Proteinstruktur des Klasse-II-Hydrophobins (HFBI) aus dem Schimmelpilz *Trichoderma reesei* verwendeten wir ein Modell, bei dem die Proteine mit atomarer Auflösung dargestellt waren, aber nicht ihre Konformation verändern konnten. Die Simulationen zeigten, wie Dipol-Dipol-Wechselwirkungen das Erkennen der Moleküle untereinander fördern. Zudem konnten wir sehen, wie die Proteine Oligomere bilden: Die in der Simulation gefundene Zusammenlagerung von jeweils vier Hydrophobinen entspricht den kristallografisch nachgewiesenen Homotetrameren. An der Graphitoberfläche neigen diese Oligomere dazu, sich aufzulösen und über ihre hydrophobe Außenseite mit dem Festkörper in Kontakt zu treten (Kasten links).

Die biomolekulare Erkennung ist ein anspruchsvolles Problem, das modernste rechnergestützte Methoden aus den verschiedensten Fachgebieten erfordert. Die drei hier beschriebenen Anwendungen sind gute Beispiele dafür, wie Computersimulationen dabei helfen können, dieses äußerst komplexe Problem zu lösen.

DIE AUTORIN

Rebecca C. Wade studierte Physik an der University of Oxford (B. A. hons. 1985) und promovierte in molekularer Biophysik (D. Phil. 1988). Danach forschte sie an den Universitäten Houston und Illinois. 1992 bis 2001 war sie Gruppenleiterin am European Molecular Biology Laboratory (EMBL) in Heidelberg. Seit 2001 leitet sie die Gruppe »Molecular and Cellular Modeling« (MCM), zunächst bei der EML Research gGmbH und seit 2010 am Heidelberger Institut für Theoretische Studien (HITS).

QUELLEN

Ferrari, S. et al.: Virtual Screening Identification of Nonfolate Compounds, Including a CNS Drug, as Antiparasitic Agents Inhibiting Pteridine Reductase. In: Journal of Medical Chemistry 54, S. 211–221, 2011

Mereghetti, P. et al.: Brownian Dynamics Simulation of Protein Solutions: Structural and Dynamical Properties. Biophysical Journal 99, S. 782–791, 2010

Mereghetti, P., Wade, R. C.: Diffusion of Hydrophobin Proteins in Solution and Interactions with a Graphite Surface. In: BMC Biophysics 4, Artikel 9, 2011, doi:10.1186/2046-1682-4-9

Pachov, G. et al.: On the Structure and Dynamics of the Complex of the Nucleosome and the Linker Histone. In: Nucleic Acid Research 2011, doi: 10.1093/nar/gkr101

Zerren an Biomolekülen im Computer

Mechanische Kräfte sind lebenswichtig – im großen wie im kleinen Maßstab. Eine Forschungsgruppe am Heidelberger Institut für Theoretische Studien untersucht ihre Wirkung auf der kleinsten Ebene: vom Protein bis hin zur einzelnen chemischen Bindung.

Von Ilona Baldus und Frauke Gräter

Ob Pflanze oder Säugetier, kein Lebewesen kann ohne Einwirkung mechanischer Kräfte überleben. Ein beeindruckendes Beispiel dafür kommt aus der Raumfahrt: Während eines mehrwöchigen Aufenthalts im All würde ein Astronaut ohne spezielles Krafttraining einen erheblichen Teil seiner Knochenmasse verlieren. Woran liegt das? Der menschliche Körper erneuert ständig sein Knochengewebe und baut es dafür kontinuierlich ab. Der gleichzeitige erneute Aufbau hängt allerdings davon ab, wie stark der Knochen benutzt wird – das heißt, in welchem Maß Kräfte durch Stehen, Gehen und Laufen darauf einwirken. Im Weltall ist die Gravitation um ein Vielfaches geringer als am Erdboden, was die Belastung der Knochen stark reduziert und ihren Wiederaufbau verzögert. Nur das Krafttraining im All verhindert also, dass ein Raumfahrer mit stark geschwächtem Skelett auf die Erde zurückkehrt.

Der große Einfluss mechanischer Kräfte auf das Leben zeigt sich selbst auf der Ebene einzelner Zellen. Auch sie reagieren in einer ungewohnten Umgebung manchmal anders als normal, wie beispielsweise André E. X. Brown und Dennis E. Discher von der University of Pennsylvania in Philadelphia 2009 festgestellt haben. Demnach wachsen Nervenzellen auf dem harten Boden der im Labor verwendeten Petrischalen weitaus schlechter als auf einer weichen, elastischen Oberfläche, an der sie fester haften. Das wirft natürlich die Frage auf, wie lebende Organismen oder gar einzelne Zellen die auf sie einwirkende mechanische Kraft eigentlich spüren. Verfügen sie über spezielle Sensoren, die einen Zug oder Druck registrieren und mit einem Signal darauf antworten, das ein biochemisches Programm in Gang setzt?

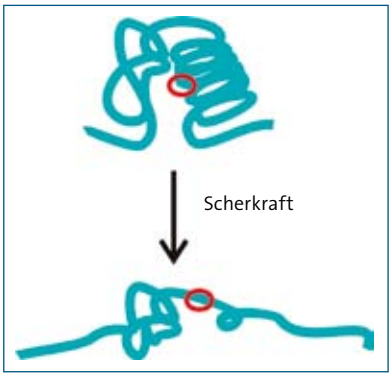
Der Antwort auf diese Frage sind Forscher in den letzten Jahren ein gutes Stück näher gekommen. Offenbar gibt es tatsächlich Kraftsensoren, und vereinzelt wurden sie auch schon identifiziert. Wie sie genau funktionieren, lässt sich experimentell aber nur schwer und oft ausschließlich indirekt beobachten; denn es handelt sich meistens um Proteine, also Eiweißstoffe, die typischerweise nicht mehr als wenige Nanometer (milliardstel Meter) messen.

In der Gruppe für »Molekulare Biomechanik« am Heidelberger Institut für Theoretische Studien (HITS) benutzen wir deshalb leistungsstarke Computer und physikalische Modelle, um den Einfluss mechanischer Kräfte auf einzelne Proteinmoleküle zu ergründen. Wir möchten die Funktionsweise solcher winzigen molekularen Kraftsensoren im Detail verstehen. Manche Krankheiten beruhen darauf, dass das Messen und Verarbeiten der mechanischen Kraft in bestimmten Geweben gestört ist. Mit unseren Untersuchungen verfolgen wir das Ziel, in Zusammenarbeit mit Medizinern die molekularen Mechanismen hinter solchen Störungen aufzudecken.

Scherkräfte im Blut

Der so genannte Von-Willebrand-Faktor (VWF) bietet ein anschauliches Beispiel für den Einfluss mechanischer Kräfte auf Vorgänge in Lebewesen. Es handelt sich um ein Protein im Blut, das die Blutgerinnung einleitet. Fehlt der VWF oder wirkt er nur unzureichend, kommen Blutungen nicht zum Stillstand, was tödlich sein kann. In diesem Fall sprechen Mediziner auch vom Von-Willebrand-Syndrom. Umgekehrt kann eine zu starke Wirkung des VWF, also eine über-

Kontrolle der Blutgerinnung



ILONA BALDUS UND FRAUKE GRÄTER

Der Von-Willebrand-Faktor spielt eine wichtige Rolle bei der Blutgerinnung. Scherspannungen beim Austreten von Blut aus einer Wunde strecken das verknäuelte fadenförmige Molekül. Dadurch wird es klebrig und verbindet sich mit Blutplättchen zu engmaschigen Netzwerken. Wie Computersimulationen ergaben, legt die Entfaltung allerdings auch eine Stelle frei, an der Enzyme das Molekül zerschneiden können (roter Kreis). Das verhindert ein Überschießen der Gerinnungsreaktion und die Bildung von Thromben. In der Schemazeichnung ist nur der relevante Teil des in Wahrheit viel größeren Proteins gezeigt.

mäßige Blutgerinnung, die Bildung von Pfropfen – Thrombosen – verursachen. Mechanische Kräfte kontrollieren das Gleichgewicht zwischen Blutfluss und -gerinnung in den Adern und besonders im Umkreis einer Wunde.

Generell gilt: Wann immer eine Flüssigkeit durch ein Rohr strömt, entsteht eine Scherkraft, weil die Strömung in der Rohrmitte schneller ist als am Rand. Ein mitschwimmender Faden wird dadurch gestreckt. Der VWF ist ein solcher Faden, allerdings so winzig klein und dünn, dass er sich nur im Mikroskop erkennen lässt. Er geht im Scherfluss, der bei einer Verletzung besonders hoch ist, von einem verknäuelten in den langgestreckten Zustand über (Kasten links). Als Folge davon wird er besonders klebrig und verbindet sich mit anderen solchen Fäden und mit Blutplättchen zu engmaschigen Netzwerken. Sie bilden das erste Gerüst für Blutgerinnsel, die das Ausfließen von Blut verhindern.

Interessanterweise gibt es im langen Fadenmolekül des VWF eine Stelle, an der ihn ein anderer Blutbestandteil, eine Protease, zerschneiden kann. Im verknäuelten Zustand ist diese Schnittstelle tief im Inneren – in der so genannten A2-Domäne – verborgen und damit für die molekulare Schere schlecht zugänglich. Wir vermuteten jedoch, dass sie frei gelegt wird, wenn sich der Faden durch die Scherkraft streckt.

Um Klarheit zu gewinnen, untersuchten wir den Vorgang in Computersimulationen. Hierzu befestigten wir an den Enden des verknäuelten Proteins virtuelle Federn und bewegten diese voneinander weg. Das klingt viel einfacher, als es ist. Tatsächlich erforderte es sehr aufwändige Rechnungen, denen Modelle der klassischen newtonschen Physik zu Grunde lagen. Am Ende aber konnten wir so ermitteln, wie sich der VWF unter Zugspannung entfaltet. In der Tat geben bestimmte Teile des Fadenmoleküls sukzessive der Kraft nach und lösen sich voneinander. Dabei wird schließlich auch die Spaltstelle frei gelegt, an der die Schneideenzyme ansetzen können.

Das ist für die biologische Rolle des VWF sehr wichtig. Die Netzwerke, zu denen sich die von der Scherkraft gestreckten Fäden zusammenlagern, sind zwar für die Blutgerinnung notwendig, doch ein unbegrenztes Wachstum würde das Gefäß für alle Zeit komplett verstopfen. Besonders hohe Zugkräfte entfalten den VWF deshalb so weit, dass die Protease Zutritt zur Schnittstelle erhält. So kann sich ein Gleichgewicht zwischen Blutgerinnung und Auflösung der Blutpfropfen einstellen. Der Kraftsensor dafür ist der VWF. Er übersetzt ein rein mechanisches in ein biochemisches Signal – ein lebenswichtiger Vorgang.

Bindungsbruch unter Spannung

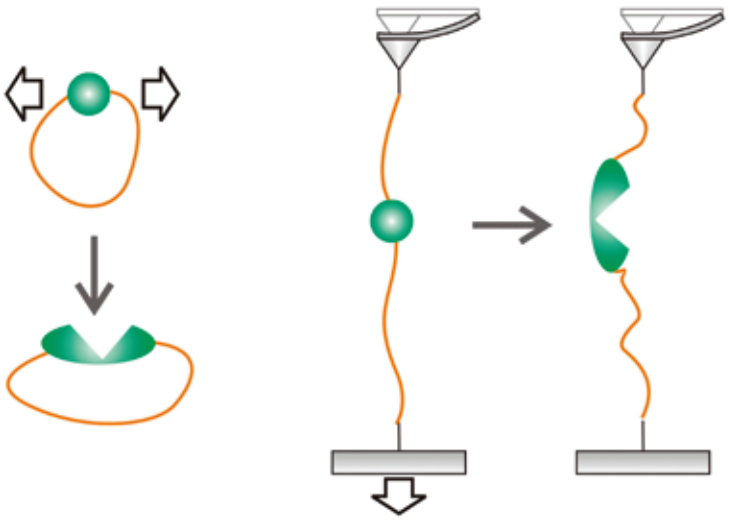
Spielen mechanische Kräfte auch in noch kleineren Dimensionen eine Rolle? Jegliche feste Materie besteht aus Atomen, die durch chemische Bindungen zusammengehalten werden. Diese ähneln Klebstoffen unterschiedlicher Haftkraft. Je nach Funktion des Moleküls sind sie sehr stark oder leicht zu lösen. Manche wirken als Schalter, der nach Bedarf geöffnet wird, andere sollen einer molekularen Struktur dauerhafte Stabilität verleihen. Das gilt zum Beispiel für die Bindungen, die das Rückgrat eines Proteins aufbauen.

Angesichts der großen Bedeutung mechanischer Signale in Lebewesen liegt die Vermutung nahe, dass Zugspannungen auch chemische Bindungen in Molekülen verstärken oder schwächen können. Wann und wie passiert das? Dieser Frage sind wir nachgegangen. Wir wollten wissen, wie leicht sich eine Bindung lösen lässt, wenn man von beiden Seiten daran zieht wie an einem Seil. Allerdings ist rohe Gewalt nicht immer das beste Mittel. Man denke nur an eine klemmende Tür. Meist gibt sie zwar umso eher nach, je stärker man dagegendrückt. Trotzdem ist es nicht immer angebracht, sich mit voller Wucht dagegenzuwerfen. Besser versucht man vielleicht zunächst, die Tür vorsichtig mit der Klinke zu öffnen. Auch bei der chemischen Bindung hängt die ideale Öffnungsmethode vom Einzelfall ab.

In Proteinen gibt es verschiedenste Wechselwirkungen zwischen den Atomen. Zwei davon wollen wir hier betrachten: Wasserstoffbrücken und kovalente Bindungen. Erstere ähneln Klettverschlüssen. Einzeln lassen sie sich leicht lösen, aber im Verbund sind sie sehr stabil. Wasserstoffbrücken halten Strukturelemente wie das Beta-Faltblatt und die Alpha-Helix zusammen. Im letzteren Fall sind die Proteinbausteine, die Aminosäuren, wie

Auf Biegen und Brechen

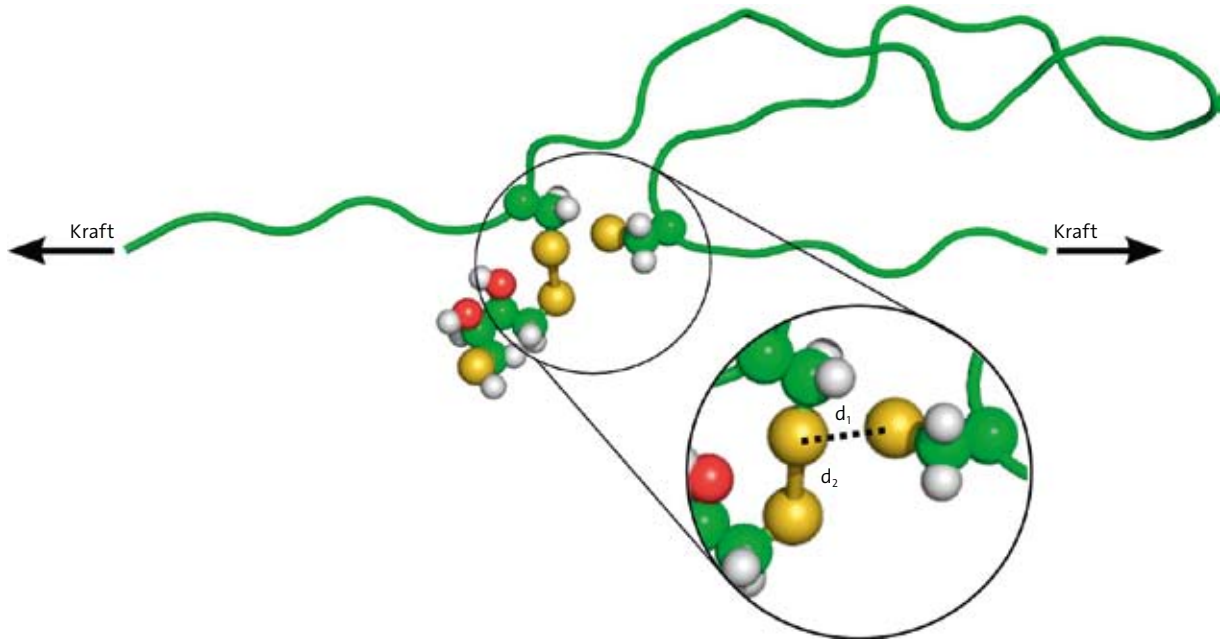
Inwieweit mechanische Kräfte das Öffnen einer Bindung erleichtern, lässt sich experimentell ermitteln. So herrscht in Ringmolekülen je nach ihrer Größe eine unterschiedlich starke Spannung (links). Man kann nun prüfen, ob sich dieser Umstand auf die Geschwindigkeit des Bindungsbruchs auswirkt. Eine andere Möglichkeit ist, das Molekül in ein Kraftmikroskop einzuspannen und durch Verbiegen des Tastarms (Cantilevers) einen Zug darauf auszuüben (rechts).



ILONA BALDUS UND FRAUKE GRÄTER

Molekulare Brückensprengung

Das kleine Molekül DTT (Dithiothreitol) zerstört die Disulfidbrücke d_1 in einem Protein (Titin) unter Bildung einer neuen Disulfidbrücke d_2. Wie Computersimulationen ergaben, erleichtert eine angelegte Zugspannung den Bindungsbruch vor allem dadurch, dass das Schwefelatom von DTT schon aus größerer Entfernung die neue Disulfidbindung eingehen und die alte dabei lösen kann.



ILONA BALDUS UND FRAUKE GRÄTER

in einer Wendeltreppe angeordnet: Das Rückgrat bildet das Gerüst und die Wasserstoffbrücken das Geländer. In einem Beta-Faltblatt verlaufen zwei Abschnitte des Proteinrückgrats parallel zueinander. Wasserstoffbrücken verbinden diese Stränge durch elektrostatische Kräfte miteinander. Beta-Faltblätter geben Proteinen zwar große Stabilität, lassen sich aber bei genügend Zugkraft auftrennen.

Kovalente Bindungen sind wesentlich fester. Dabei teilen sich zwei Atome ein Elektronenpaar. Ein biologisch wichtiges Beispiel sind Schwefel-Schwefel-Bindungen oder, wie Chemiker sagen, Disulfidbrücken. Sie bilden sich etwa zwischen zwei Molekülen der Aminosäure Cystein. Solche Bindungen haben meist die Aufgabe, die Struktur des Proteins zu stabilisieren – auch gegen von außen einwirkende Zugkräfte.

Neuerdings lässt sich im Labor beobachten, wie Disulfidbrücken oder andere kovalente Bindungen unter Zugspannung aufbrechen (Kasten auf S. 19). Dazu befestigt man ein einzelnes Molekül in einem Kraftmikroskop mit einem Ende an der Unterlage und mit dem anderen an der Spitze des Tastarms. Dieser besteht aus einer Blattfeder, mit der sich eine mechanische Kraft auf das eingespannte Molekül ausüben lässt. So kann man direkt verfolgen, wie leicht sich die Bindung bei welcher Zugkraft öffnet.

Doch nackte Gewalt führt dabei nicht zum Ziel. Wie in der Natur geht es darum, die Bindung so sanft wie möglich zu lösen. Das gelingt durch Zugabe von kleinen Hilfsmolekülen, so genannten Reduktionsmitteln. Diese enthalten bei Disulfidbrücken ein Schwefelatom, das sich mit dem einen Teil der Disulfidbrücke verbindet und so den anderen daraus verdrängt.

Derartige Messungen im Labor ergaben, dass sich Disulfidbindungen in Proteinen unter Mitwirkung eines Reduktionsmittels sehr leicht aufbrechen lassen. Zugspannungen von wenigen hundert Pikonewton reichen bereits aus. Das entspricht in etwa der Kraft, die ein einzelner Mensch aufwenden müsste, um mit der gesamten Weltbevölkerung zusammen ein 1-Euro-Stück hochzuhalten. Dabei gilt: Je größer die Zugspannung, desto instabiler wird die Bindung und desto schneller löst sie sich. Dies ist ganz ähnlich wie bei einem Gummiband: Je stärker man daran zieht, desto eher reißt es.

Man könnte meinen, die Mechanochemie einer solchen Reaktion damit verstanden zu haben – gerade weil man sich den Effekt der Zugkraft intuitiv vorstellen kann. Aber wie so oft sind die Zusammenhänge komplexer, als sie auf den ersten Blick erscheinen. So gibt es Reaktionspartner, bei denen die mechanische Kraft das Öffnen der chemischen Bindung erschwert! In einem anderen Fall, den eine Gruppe um Roman Boulatov von der University of Illinois in Urbana-Champaign 2009 entdeckte, löst sich die Disulfidbrücke unabhängig von der an ihr angreifenden Zugspannung immer gleich schnell. Für dieses Experiment bauten die Wissenschaftler die Schwefel-Schwefel-Bindung in kleine ringförmige Moleküle ein. Über die Größe des Rings konnten sie die darin herrschende Spannung gezielt verändern (siehe Kasten auf S. 19).

Wie beeinflusst eine mechanische Kraft also eine chemische Bindung? Warum er-

leichtert sie in bestimmten Fällen deren Öffnung? Wird die Bindung durch die Zugspannung vorgedehnt oder auf andere Weise geschwächt? Oder ist sie einfach nur leichter zugänglich für das Reduktionsmittel, weil das gesamte Molekül dabei auseinandergezogen wird?

In unserer Forschungsgruppe am HITS in Heidelberg suchen wir auf numerischem Weg nach Antworten auf diese Fragen. Deshalb haben wir den Bindungsbruch am Computer simuliert. Dabei zeigte sich in Einklang mit den experimentellen Befunden, dass sich Bindungen normalerweise mit steigender Kraft schneller lösen.

Wir können allerdings auch gewissermaßen genauer hinschauen, was im Einzelnen passiert. So erhalten wir Einblicke in Abläufe, die experimentell nur sehr schwer und mit großem Aufwand zugänglich wären. Zum Beispiel können wir die Reaktion in Einzelschritte zerlegen. Das Öffnen der Disulfidbrücke beginnt damit, dass sich das Schwefelatom des Reduktionsmittels der Bindung nähert, die unter Spannung steht. Es nimmt mit einem der beiden Brückenschwefelatome Kontakt auf und bildet mit ihm eine neue Disulfidbindung. Dabei wird das andere Schwefelatom verdrängt und die ehemalige Disulfidbrücke gesprengt (Kasten links).

Paradoxe Wirkung einer äußeren Zugkraft

Diesen Vorgang bezeichnen Chemiker als bimolekulare nukleophile Substitutionsreaktion (S_N2). Am Computer haben wir die einzelnen Schritte unter die Lupe genommen. Dabei interessierten wir uns für zwei Messgrößen: den Abstand zwischen den Schwefelatomen in der aufbrechenden (d_1) und in der neu entstehenden Disulfidbrücke (d_2).

Das Ergebnis war überraschend. Zwar hatten wir erwartet, dass sich beide Bindungslängen, also d_1 und d_2, während des Reaktionsprozesses ändern und die ursprüngliche Schwefel-Schwefel-Bindung von der Zugkraft verlängert wird. Allerdings fiel die Dehnung nur sehr gering aus. Wirklich unerwartet war hingegen, dass auch d_2 von der externen Kraft beeinflusst wird, obwohl diese nur auf d_1 wirkt. Wie wir feststellten, muss sich unter Zugspannung das Schwefelatom des Reduktionsmittels der Bindung nicht mehr so weit nähern, um sie zu öffnen. Sie zerbricht schon in größerem Abstand, weil sie von der äußeren Kraft geschwächt ist. Das beschleunigt die Reaktion.

Wie man sieht, sind der Ablauf des Bindungsbruchs und der Einfluss der Kraft darauf komplexe Angelegenheiten. Das macht die Computersimulationen und ihre Interpretation äußerst aufwändig. Es gibt jedoch eine dazu komplementäre Methode, die direkter und dadurch einfacher ist: die Betrachtung der Energielandschaft einer Reaktion. Daraus lässt sich unmittelbar ersehen, wie leicht eine Umsetzung abläuft.

Energielandschaften gleichen Gebirgen. Am wohlsten fühlen sich die Stoffe im Tal. Je tiefer es ist, desto besser. Der Weg von einem Tal ins andere führt über einen Berg oder Pass. Im Falle der Disulfidbrücke ist das zu erreichende Tal die offene Bindung.

Auch die Rolle der mechanischen Kraft lässt sich mit der Energielandschaft veranschaulichen. Sie hebt das betreffende Molekül ein Stück weit aus seinem Tal heraus, was den Weg über den Berg bereits deutlich erleichtert. Außerdem senkt sie das zu erreichende Tal ab und erniedrigt zugleich den Pass dorthin.

Wir haben auch solche Energielandschaften berechnet. Dabei bestätigte sich, dass mit steigender Zugkraft, die auf eine Disulfidbrücke wirkt, das Tal für die offene Bindung immer weiter absinkt. Das macht das Lösen der Verknüpfung energetisch vorteilhafter. Im Einklang mit den Ergebnissen der Computersimulation dehnt die Kraft also nicht einfach nur die Disulfidbrücke, sondern wirkt sich auf das ganze Molekül aus. So ändert sie Winkel und verdreht Strukturelemente, was die Schwefel-Schwefel-Bindung zusätzlich destabilisiert. Sobald die Brücke bricht, können Winkel und verzerrte Strukturelemente ihre ursprüngliche Position wieder einnehmen. Bei diesem Entspannen wird sehr viel Energie frei. Auch hier bietet sich der Vergleich mit dem Gummiband an: Spannt man es stark und zerschneidet es, so kehrt es mit einem kräftigen Schnalzen in seinen ungespannten Zustand zurück.

Wie man sieht, sind lebendige Systeme auf verschiedenste Weise mechanischen Kräften ausgesetzt. Biologische Strukturen, von kleinen Eiweißmolekülen bis zu Zellen und Geweben, haben im Verlauf der Evolution die Fähigkeit erlangt, gezielt darauf zu reagieren. So kann der Organismus mechanische Reize als lebenswichtige Informationen direkt in biochemische Signale umwandeln und verarbeiten. Computersimulationen, wie wir sie in der Gruppe »Molekulare Biomechanik« am HITS in Heidelberg durchführen, helfen Schlüsselprozesse aufzudecken, die sich diese Kraftsensoren zu Nutze machen. Die Erkenntnisse, die wir dabei gewinnen, schaffen letztendlich die Voraussetzung dafür, korrigierend in Störungen der Signalkaskade bei Krankheiten einzugreifen oder die natürlichen Vorbilder im Labor für andere Zwecke nachzuahmen. Wir sind gespannt!

DIE AUTORINNEN

Ilona Baldus (oben) hat an der Universität Heidelberg Chemie studiert. Sie ist Doktorandin bei Frauke Gräter und untersucht den Einfluss mechanischer Kräfte auf Redoxpotenziale von Proteinen.

Frauke Gräter ist seit 2009 Leiterin der Forschungsgruppe »Molekulare Biomechanik« am Heidelberger Institut für Theoretische Studien (HITS). Zuvor leitete sie eine Nachwuchsforschergruppe, die an der Chinese Academy of Sciences in Schanghai, einem Partnerinstitut der Max-Planck-Gesellschaft, und an der Universität Heidelberg angesiedelt war. Nach ihrer Promotion an der Universität Göttingen war die Chemikerin bis 2007 am Max-Planck-Institut für Biophysikalische Chemie in Göttingen und an der Columbia University in New York tätig.

QUELLEN

Baldauf, C. et al.: Shear-Induced Unfolding Activates von Willebrand Factor A2 Domain for Proteolysis. In: Journal of Thrombosis and Haemostasis 7, S. 2096–2105, 2009

Brown, A. E. X., Discher, D. E.: Conformational Changes and Signaling in Cell and Matrix Physics. In: Current Biology 19, S. R781–R789, 2009

Kucharski, T. J. et al.: Kinetics of Thiol/Disulfide Exchange Correlate Weakly with the Restoring Force in the Disulfide Moiety. In: Angewandte Chemie 121, S. 7174–7177, 2009

Li, W., Gräter, F.: Atomistic Evidence of how Force Dynamically Regulates Thiol/Disulfide Exchange. In: Journal of the American Chemical Society 132, S. 16790–16795, 2010

Wiita, A. P. et al.: Probing the Chemistry of Thioredoxin Catalysis with Force. In: Nature 450, S. 124–127, 2007

Hochleistungsrechner und der Stammbaum des Lebens

Eine wahre Flut von DNA-Daten ermöglicht inzwischen immer präzisere Rekonstruktionen von Stammbäumen – im Prinzip jedenfalls. In der Praxis überfordert die Suche nach der optimalen Lösung auch die leistungsfähigsten Computer. Die Herausforderung heißt deshalb, die Effizienz der Programme für Näherungslösungen zu steigern.

Von Alexandros Stamatakis

Die computergestützte Berechnung von Stammbäumen, welche die Verwandtschaftsverhältnisse zwischen Organismen wiedergeben, ist eine verhältnismäßig junge Disziplin. Doch reichen ihre Anfänge immerhin bis in die 1960er Jahre zurück. Für jeden Organismus beziehungsweise jede Spezies, deren Position im Stammbaum ermittelt werden soll, liegen typischerweise DNA-Daten oder Angaben zu morphologischen Merkmalen vor – etwa über die Knochenform. Bei Bakterien kann es sich auch um chemische Eigenschaften handeln, die für die jeweilige Spezies charakteristisch sind.

Das Ziel besteht darin, anhand geeigneter Modelle denjenigen Stammbaum zu rekonstruieren, der am besten zu den vorliegenden Daten passt. Mathematisch gesehen, handelt es sich also um ein Optimierungsproblem. Dahinter steckt die stillschweigende Annahme oder Hoffnung, dass der »optimale« Stammbaum auch der wahre ist. An seinen Blättern befinden sich die Organismen, für welche DNA-Daten vorliegen. Die inneren Knoten – sprich: Verzweigungen – repräsentieren hypothetische gemeinsame Vorfahren.

Von diesen existieren in der Regel keine DNA-Daten, weil sich normalerweise nur aus lebenden Organismen Erbsubstanz gewinnen lässt. Allerdings gab es in letzter Zeit bedeutende Fortschritte bei der Sequenzierung alter DNA; dadurch ist es insbesondere der Gruppe um Svante Pääbo vom Max-Planck-Institut für evolutionäre Anthropologie in Leipzig gelungen, das Neandertalergenom zu entziffern.

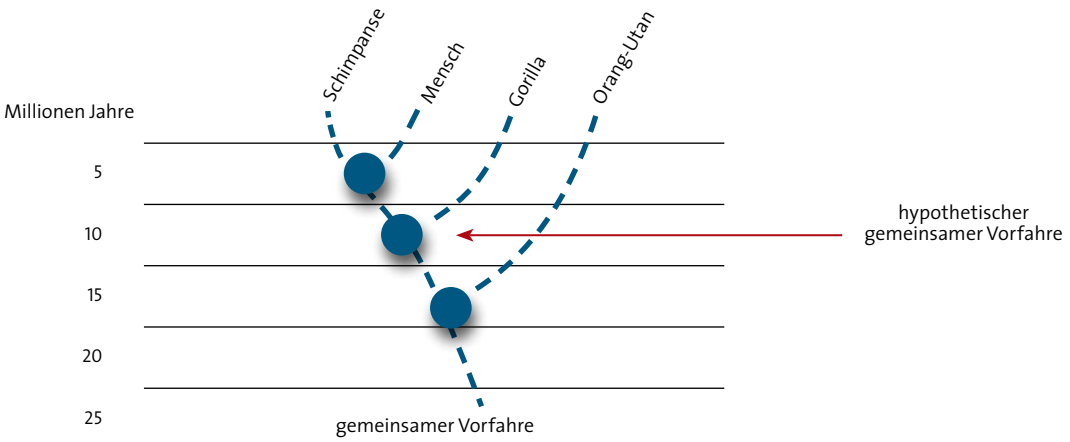
Betrachten wir ein klassisches Beispiel: den Stammbaum von Mensch, Schimpanse, Gorilla und Orang-Utan. Der auf DNA-Sequenzen beruhende Eingabedatensatz könnte, grob vereinfacht, dann so aussehen:

Mensch	AAACCCCGTTTTT
Gorilla	AAACTTTAAGGGT
Schimpanse	AAGATTCGTTTTT
Orang-Utan	AGAATCCGTTTGT

Dabei stehen die Buchstaben für die Basen Adenin, Thymin, Cytosin und Guanin, die das genetische Alphabet ausmachen. Ein möglicher Stammbaum für diese Daten ist im Kasten unten gezeigt. Dabei bleibt offen, wo der gemeinsame Vorfahr aller Menschenaffen, das heißt die Wurzel des Baums, liegt. Diese wird zur Vereinfachung der mathematischen Modelle üblicherweise weggelassen.

Grundlage für die Optimierung ist eine abstrakte Funktion f, eine Rechenvorschrift, die zu einem gegebenen Stammbaum und zu gegebenen DNA-Daten einen Zahlenwert liefert: die »Plausibilität« *(likelihood)*. Je höher dieser Wert, desto besser ist der Stammbaum mit den Daten vereinbar. Wenn man also drei

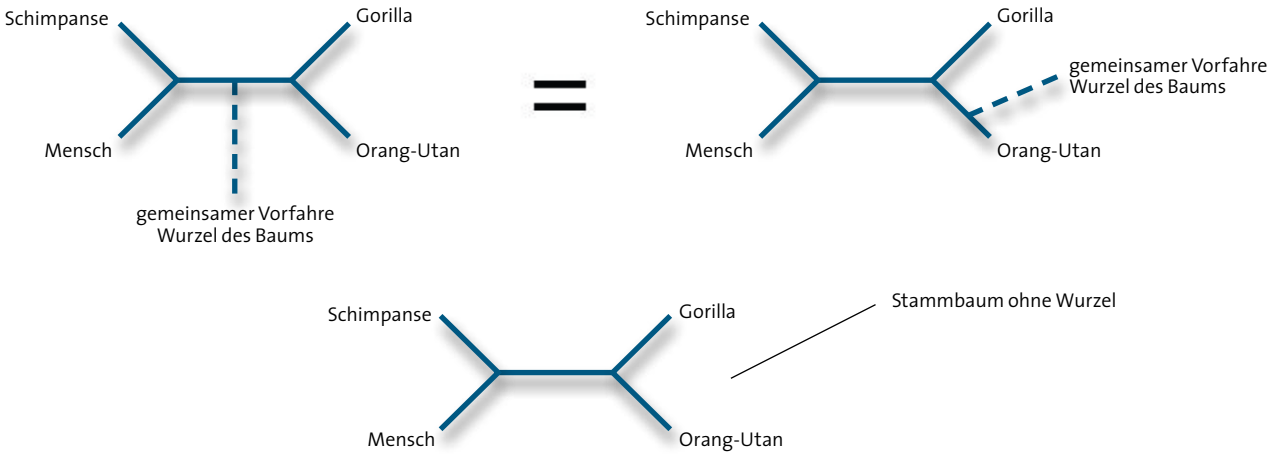
Der DNA-Stammbaum der Menschenaffen



ALLE ABBILDUNGEN DIESES ARTIKELS: ALEXANDROS STAMATAKIS

Stammbäume ohne Wurzel

Anhand von DNA-Daten der Menschenaffen allein lässt sich keine Aussage über die Wurzel des Stammbaums machen. Sie kann an den verschiedensten Stellen liegen (oben). Dem trägt die Darstellung ohne Wurzel Rechnung (unten).



Stammbäume in Betracht zieht, muss man die Funktion für alle drei berechnen. Der optimale Baum ist dann derjenige, für den der größte Wert herauskommt.

In unserem Beispiel mit den Menschenaffen lässt sich dieses Problem leicht lösen, weil für vier Organismen nur drei unterschiedliche wurzellose Bäume existieren (Kasten oben). Dabei erweist sich derjenige, bei dem der Mensch mit dem Schimpansen näher verwandt ist als beide mit dem Gorilla und dem Orang-Utan, als plausibelste Lösung. Doch wie sieht die Funktion f aus? In der Praxis benutzt man dafür statistische Modelle, die auf Schätzungen beruhen, wie wahrscheinlich Mutationen sind, bei denen eine der vier Basen durch eine andere ersetzt wird.

Das grundsätzliche Problem bei diesem Verfahren ist, dass die Anzahl der möglichen Bäume extrem stark mit der Anzahl der enthaltenen Spezies zunimmt. So beläuft sie sich bei 50 Arten, was heutzutage noch eine relativ kleine Zahl ist, bereits auf $2{,}84 \cdot 10^{76}$ Kandidaten. Für jeden von ihnen müsste der Wert der Funktion f berechnet werden, denn es gibt keinen Trick, einen Großteil davon von vornherein auszuschließen. Unter der optimistischen Annahme, dass diese Berechnung für einen Baum mit 50 Organismen eine Sekunde Rechenzeit benötigt, würde die Evaluierung aller Bäume auf einem einzelnen Prozessor $9 \cdot 10^{68}$ Jahre dauern. Selbst mit der gesamten Rechenkraft auf der Erde wäre diese Aufgabe vermutlich nicht innerhalb eines vernünftigen Zeitraums zu schaffen.

Optimierungsprobleme, für die der Bedarf an Rechenzeit derart schnell zunimmt, kommen in vielen Bereichen der Informatik vor und heißen NP-vollständig. Peter Gritzmann und René Brandenberg beschreiben sie in ihrem populärwissenschaftlichen Buch »Das Geheimnis des kürzesten Weges« auf für Laien gut verständliche Art und Weise.

Automatische Suchverfahren

Da das Problem nicht exakt lösbar ist, behilft man sich mit so genannten heuristischen Suchverfahren, die zwar nicht die beste, aber zumindest eine ziemlich gute Lösung liefern. Leider gibt es bei der Berechnung von Stammbäumen keine Möglichkeit, mit Sicherheit zu sagen, wie weit das Ergebnis einer solchen approximativen Suche vom Optimum entfernt ist. Deshalb ist es unerlässlich, dass Biologen den gefundenen Baum anhand ihres Wissens auf Plausibilität prüfen.

Man kann das Suchverfahren auch an sehr schnell evolvierenden Organismen wie etwa Viren testen, deren Stammbaum über die letzten Jahre bis Jahrzehnte bekannt ist. Auch im Erfolgsfall bietet das jedoch keine Gewähr dafür, dass die Methode bei Lebewesen, die sich im Verlauf von Jahrmillionen entwickelt haben, genauso gut funktioniert.

Ein weiterer Unsicherheitsfaktor ist die Programmverifikation. Selbst wenn die Evolutionsmodelle und Rekonstruktionsverfahren perfekt sind, heißt das nicht, dass sie auch korrekt auf dem Computer umgesetzt wurden. Durch die starke Zunahme rechnerbasierter Datenanalysen in der Biologie haben Fehler in Veröffentlichungen, die auf Programmierfehlern beruhen, in jüngster Zeit stark zugenommen. Zusammen mit meinem Doktoranden Fernando Izquierdo-Carrasco habe ich die Probleme der Verifikation von Stammbäumen und von Programmen zu deren Berechnung kürzlich ausführlich dargelegt *(Briefings in Bioinformatics 12, S. 270)*.

Trotz solcher Schwierigkeiten und Unsicherheiten kommen Verfahren zur Rekonstruktion von Stammbäumen in der medizinischen und biologischen Forschung heute routinemäßig zum Einsatz. So dienen sie etwa dazu, den Ursprung von Virusepidemien zu ermitteln oder die bakterielle Zusammensetzung der Darmflora zu analysieren. Um das berühmte Zitat des russischen Genetikers Theodosius Dobzhansky (1900 – 1975) zu bemühen: »In der Biologie macht nichts Sinn, außer im Licht der Evolution.«

Was sind die aktuellen Entwicklungen und Herausforderungen auf dem Gebiet der Stammbaumberechnung? Zuallererst ist die Revolution bei der DNA-Sequenzierung zu nennen. Die Analyse des Erbguts wurde durch bahnbrechende Fortschritte in den letzten fünf bis sechs Jahren wesentlich vereinfacht und beschleunigt, so dass zugleich die Kosten

dramatisch gesunken sind. Dadurch lassen sich inzwischen auch komplette Genome einer Spezies sehr viel leichter entziffern. Während vor zehn Jahren die Sequenzierung des menschlichen Erbguts noch Schlagzeilen machte, nehmen heute selbst Biologen eher gelangweilt zur Kenntnis, dass schon wieder irgendein Genom entschlüsselt wurde.

Die Herausforderung verlagert sich daher zunehmend vom Labor zur Datenverarbeitung. Das Hauptproblem besteht darin, dass die Menge der DNA-Daten wesentlich schneller zunimmt als die Rechengeschwindigkeit der Computer oder Prozessoren zu ihrer Analyse. Das betrifft sowohl die Bioinformatik als auch ihre Teildisziplin, die rechnergestützte Ermittlung von Stammbäumen. Die Computerwissenschaftler stehen deshalb vor der schwierigen Aufgabe, immer effizientere Programme und Methoden zur Datenspeicherung und -analyse bereitzustellen.

Ohne Hoch- und Höchstleistungsrechner, in denen mehrere Einzelrechner (Prozessoren) gleichzeitig an einem Problem arbeiten, lässt sich die Datenflut vielfach nicht mehr bewältigen. Zur Rekonstruktion von Stammbäumen standen noch vor zehn Jahren lediglich die Sequenzen von ein oder zwei Genen zur Verfügung, die jeweils etwa 1000 Basenpaare umfassten. Inzwischen liegen immer öfter die weitaus umfangreicheren kompletten Genome vor. So besteht das Erbgut des Menschen aus etwa 20 000 bis 25 000 Genen; nach einigen Schätzungen sind es sogar bis zu 75 000.

Diese Datenflut stellt die Informatiker vor enorme Probleme. Das gilt insbesondere für den Speicherplatzbedarf der Programme zur Stammbaumrekonstruktion, da zur Berechnung der Bewertungsfunktion f zunehmend komplette Genome für 50 oder 100 Spezies im Arbeitsspeicher gehalten werden müssen.

Ziel: Effiziente Bewertung der Güte eines Stammbaums

Solche Programme verbringen bis zu 99 Prozent ihrer Gesamtlaufzeit damit, die Funktion f für verschiedene denkbare Bäume auszuwerten (Kasten unten). Deshalb besteht eines der Hauptziele der von mir geleiteten Scientific Computing Group am Heidelberger Institut für Theoretische Studien darin, die Zeit und den Speicherplatzbedarf für diese Aufgabe so weit wie möglich zu reduzieren.

Über die vergangenen zehn Jahre haben wir das frei verfügbare Programm RAxML *(Randomized Accelerated Maximum Likelihood)* entwickelt. Statt die Menge aller Stammbäume erschöpfend abzuarbeiten – was aussichtslos wäre –, konstruiert das Programm zu Beginn eine Anzahl von Bäumen, indem es Blatt für Blatt in zufälliger Reihenfolge an jeweils optimaler Stelle einfügt. Es versucht diese Bäume zu verbessern, indem es ganze Äste abschneidet und an anderer Stelle wieder einsetzt, das Ganze im Rahmen eines kombinatorischen Optimierungsverfahrens namens *simulated annealing*. RAxML gehört zu den fünf bis sechs weltweit am meisten benutzten Programmen zur Stammbaumrekonstruktion.

Der Webserver http://phylobench.vital-it.ch/raxml-bb/ bietet auch interessierten Laien die Möglichkeit, es auszuprobieren; ein kleiner Testdatensatz findet sich unter www.exelixis-lab.org/dna.phy.

Zur Beschleunigung der Rechnung verfolgen wir verschiedene Ansätze. So sind wir auf der Suche nach Tricks, um redundante Berechnungen zu vermeiden und Speicherplatz zu sparen. Ausgangspunkt hierfür ist die mathematische Beschreibung der Wahrscheinlichkeitsberechnungen: Wir bemühen uns, die Funktion f so zu transformieren, dass sie bei geringerem Speicherbedarf und weniger Rechenoperationen genau das gleiche Ergebnis liefert. Von großer Bedeutung ist auch, das Programm an moderne Rechnerarchitekturen anzupassen. Dadurch lassen sich die Ressourcen der eingesetzten Prozessoren besser nutzen. Das ermöglicht einen höheren Datendurchsatz und steigert so die Anzahl der evaluierten Bäume pro Sekunde.

Wir gehen allerdings auch den umgekehrten Weg und fragen uns, wie die ideale Rechnerarchitektur für unser Programm aussehen würde. In diesem Teilprojekt entwerfen wir optimale Schaltkreise zur Berechnung der Wahrscheinlichkeitsfunktion f. Zum Testen und Verifizieren unserer Architekturen benutzen wir so genannte Field Programmable Gate Arrays, bei denen es sich um eine Art programmierbare Hardware handelt. Sie bestehen aus vielen elektronischen Grundbausteinen (»Gattern«), die sich mittels einer Hardware-Beschreibungssprache dynamisch miteinander verbinden lassen, um die vorgegebene Schaltung nachzubilden.

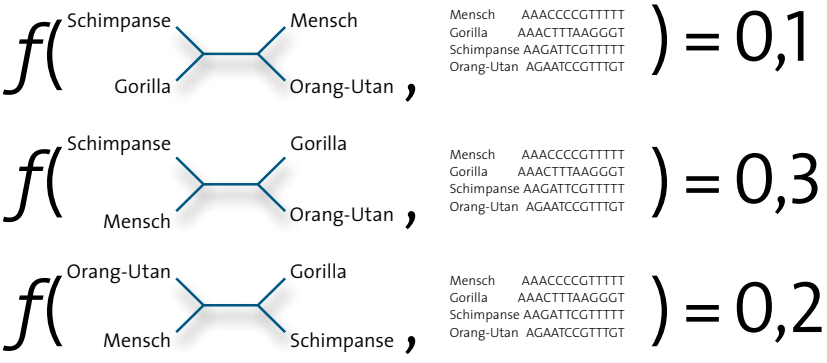
Bei all diesen Versuchen achten wir darauf, dass unsere Ergebnisse nicht nur auf RAxML anwendbar sind, sondern auch auf alle anderen likelihood-basierten Programme zur Stammbaumberechnung. Deren Geschwindigkeit hängt ja gleichfalls entscheidend davon ab, wie effizient die Funktion f auf dem Rechner umgesetzt ist.

Wie erwähnt, lassen sich sehr umfangreiche, speicherintensive Datensätze inzwischen nur noch mit Hochleistungsrechnern verarbeiten. Am HITS steht uns solch ein großer Parallelrechner zur Verfügung. Das System besteht aus 42 Rechenknoten mit je 48 Prozessoren, die durch ein leistungsfähiges Netzwerk miteinander verbunden sind.

Idealerweise gilt es, diese insgesamt 2016 Prozessoren alle gleichzeitig zu beschäftigen.

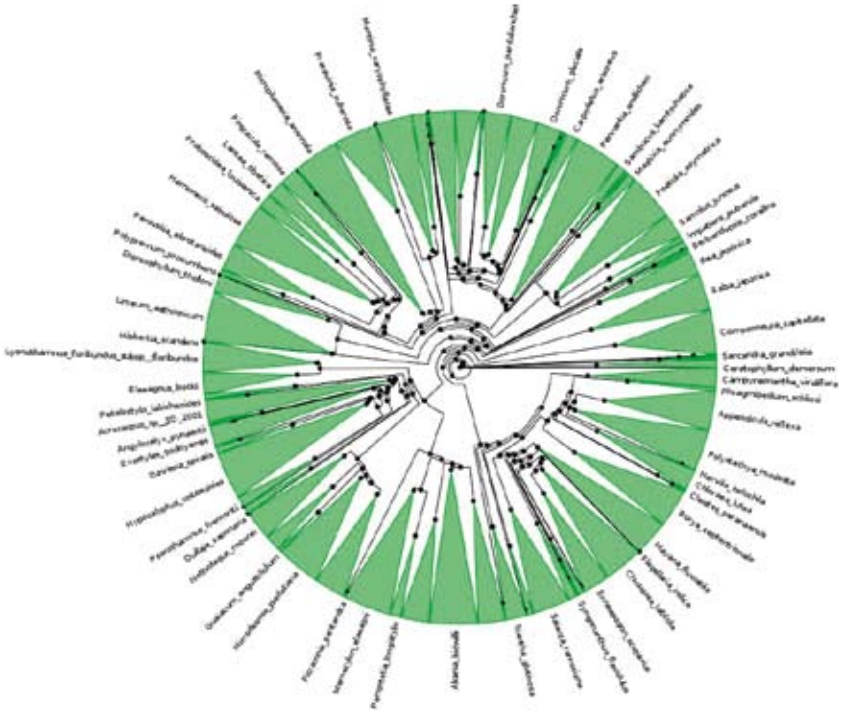
Berechnung des Verwandtschaftsgrads

Für vier Spezies existieren nur drei unterschiedliche wurzellose Stammbäume. Die Funktion f berechnet die Wahrscheinlichkeit, dass der betreffende Baum zu den DNA-Daten passt. Ihre Werte zeigen, dass Mensch und Schimpanse enger miteinander verwandt sind als mit Gorilla und Orang-Utan.



Beispiel eines ausgedehnten DNA-Stammbaums

Diesen Stammbaum für 56 000 Pflanzen errechnete die Gruppe des Autors kürzlich in Zusammenarbeit mit Forschern von der Yale University und der Brown University in den USA.



Am besten wäre es, wenn jeder von ihnen einen anderen Stammbaum evaluieren würde. Dazu müsste der einzelne Prozessor jedoch das komplette Datenmaterial im eigenen Arbeitsspeicher verfügbar haben – wozu dieser möglicherweise nicht ausreicht. Da liegt es nahe, die Aufgabe in Teilaufgaben zu zerlegen, die jede für sich nur eine relativ kleine Teilmenge aller Daten erfordern, und diese entsprechend auf die Prozessoren zu verteilen. Allerdings darf die einzelne Teilaufgabe auch nicht zu klein sein; sonst nimmt der Austausch von Daten, der vor und nach der Erledigung jeder Teilaufgabe erforderlich ist, einen zu großen Teil der Rechenzeit in Anspruch. Die Analyse und Identifizierung solcher Teilaufgaben ist nicht einfach und bildet einen der Schwerpunkte im Teilgebiet der Informatik, das sich mit der parallelen Programmierung beschäftigt.

Auch hier gilt, dass die angewandten Parallelisierungsstrategien auf alle likelihoodbasierten Programme übertragbar sein sollten und es auch sind. Mit RAxML wurden schon bis zu 1024 Prozessoren simultan zur Berechnung der Funktion f für einen einzigen Stammbaum eingesetzt, wobei diese Zahl kein Limit darstellt. Das Programm nutzt auch die Fähigkeit zur Parallelverarbeitung bei Mehrkernprozessoren, wie sie in allen neueren Laptops und Desktops zu finden sind.

Abgesehen von unseren Bemühungen, die Effizienz der Programme zur Stammbaumberechnung zu steigern, beschäftigen wir uns aber auch mit der Analyse sehr großer biologischer Datensätze. Diese interdisziplinären Projekte verbessern unser Verständnis der Biologie und helfen uns, aktuelle rechnerische oder methodische Herausforderungen zu erkennen. Beispielhaft sei hier das »plant tree of life grand challenge project« genannt, das von der Deutschen Forschungsgemeinschaft (DFG) und der National Science Foundation in den USA gefördert wird. Sein Hauptziel besteht darin, einen umfassenden Stammbaum der Pflanzen mit etwa 500 000 Spezies zu berechnen und online zur Verfügung zu stellen, so dass Biologen ihn für weiterführende Analysen nutzen können. Das ist eine Herkulesaufgabe, zumal die benötigten Daten keineswegs komplett vorliegen. Noch nie konnte ein Stammbaum dieser Größenordnung berechnet werden. Allerdings lassen sich auf dem Hochleistungsrechner des HITS mit Daten von etwa 20 Genen schon Bäume für 120 000 Spezies berechnen. In Zusammenarbeit mit Kollegen an der Yale University und der Brown University in den USA haben wir vor Kurzem einen Stammbaum der Pflanzen mit etwa 56 000 Spezies rekonstruiert und publiziert – den größten seiner Art bisher (Kasten links).

Obwohl es noch ein weiter Weg ist, kommen wir unserem Endziel, der Berechnung des Stammbaums aller Lebewesen, allmählich näher. Die stetige Verbesserung der Sequenzierverfahren und Rechnerarchitekturen lässt uns hoffen, dass wir dieses Ziel eines Tages auch erreichen werden.

DER AUTOR

Alexandros Stamatakis leitet am Heidelberger Institut für Theoretische Studien die Scientific Computing Group. Er hat an der Technischen Universität München Informatik studiert und dort im Jahr 2004 in der Informatik promoviert. Nach Postdoc-Stationen auf Kreta und an der ETH Lausanne (Schweiz) war er von 2008 bis 2010 als Nachwuchsgruppenleiter an der Ludwig-Maximilians-Universität und später an der TU München (Emmy-Noether-Programm der DFG) tätig, bevor er im Oktober 2010 ans HITS kam.

QUELLEN

Alachiotis, N. et al.: A Reconfigurable Architecture for the Phylogenetic Likelihood Function. Konferenzbeitrag, FPL Prag 2009. Online unter: http://sco.h-its.org/exelixis/nikos/publications.html

Gritzmann, P., Brandenberg, R.: Das Geheimnis des kürzesten Weges: ein mathematisches Abenteuer. Springer, Berlin/Heidelberg 2004

Ott, M. et al.: Large-Scale Maximum Likelihood-Based Phylogenetic Analysis on the IBM BlueGene/L. In: Proceedings of IEEE/ACM Supercomputing (SC2007) Conference, Reno, Nevada, November 2007

Stamatakis, A., Izquierdo-Carrasco, F.: Result Verification, Code Verification and Computation of Support Values in Phylogenetics. In: Briefings in Bioinformatics 12, S. 270–279, 2011

Stamatakis, A., Alachiotis, N.: Time and Memory Efficient Likelihood-Based Tree Searches on Gappy Phylogenomic Alignments. In: Bioinformatics 26, S. i132–i139, 2010

Pfade im Informationsdschungel

Wer die verschlungenen Wege des Stoffwechsels erforscht, benötigt Orientierungshilfe. Die Datenbank SABIO-RK hilft mit allerlei Finessen der Informatik, benötigte Daten in der Flut an Publikationen zu finden.

Von Wolfgang Müller

Allein vor dem Rechner sitzend, versunken in einer abstrakten Welt aus Bits und Bytes – das ist das Bild, das sich viele von der Arbeit des Informatikers machen. Tatsächlich sieht die Realität oft anders aus. So unterstützt die HITS-Gruppe »Scientific Databases and Visualization« (SDBV) Systembiologen durch die Einrichtung und Pflege spezieller Datenbanken. Das erfordert interdisziplinäre Zusammenarbeit und regen Austausch mit den Nutzern.

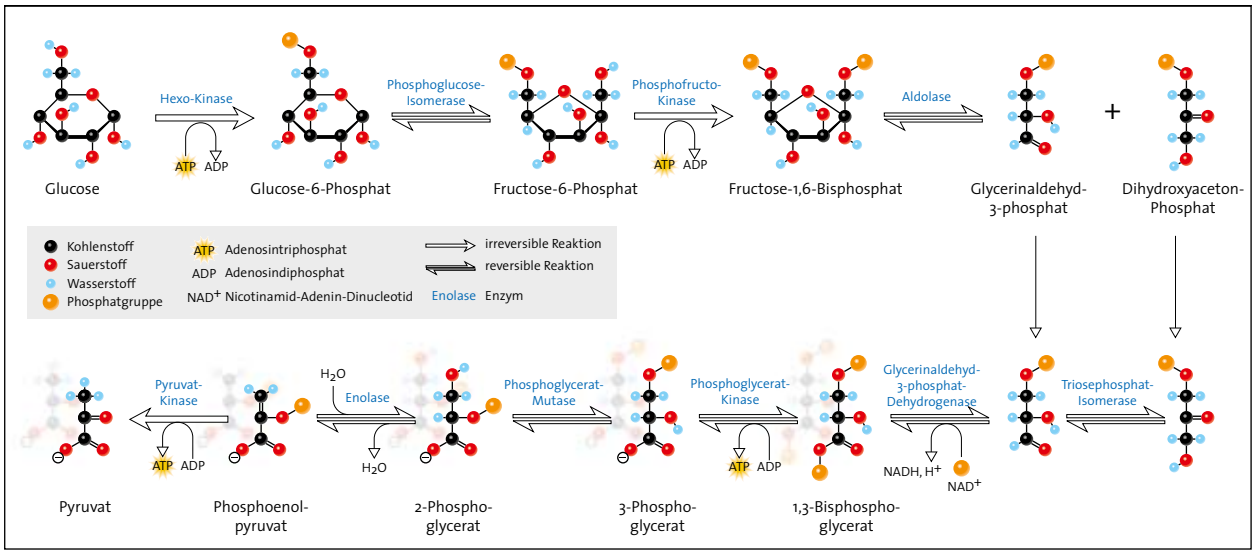
Systembiologen betrachten Vorgänge in lebenden Organismen nicht isoliert, sondern in größeren Zusammenhängen. Da sich hierbei schnell zu viele Informationen für einen einzigen Kopf anhäufen, arbeiten diese Forscher disziplinübergreifend zusammen. Während Experimentatoren sich zum Beispiel intensiv mit der Messung von Vorgängen innerhalb der Zelle befassen, haben Theoretiker etwa Stoffwechselketten und deren Kombinationen im Blick. Sie versuchen die zu Grunde liegenden biochemischen Prozesse in mathematischen Modellen zu formulieren, um nicht allein das »Wer reagiert mit wem?« zu beantworten, sondern auch Fragen wie »Wie schnell läuft die Reaktionskette bei den gegebenen äußeren Bedingungen ab?«. Solche kinetischen Modelle sind Differenzialgleichungen, die beispielsweise die zeitliche Veränderung der Glukosekonzentration und der durch den Abbau des Moleküls entstehenden Produkte widerspiegeln (unter anderem ATP und ADP, die als Energieträger im Körper fungieren). Über Koeffizienten lassen sich diese Gleichungen an die Temperatur, den pH-Wert und andere Parameter anpassen.

Wie überall in der Wissenschaft folgt der Erkenntnisgewinn dem immer gleichen Schema: Auf der Basis bereits publizierter Forschungsergebnisse entsteht eine Hypothese, die experimentell überprüft wird; die Analyse der Messergebnisse begründet dann ein Modell dessen, was im Experiment passiert ist. Alle gewonnenen Informationen werden schließlich publiziert und speisen wiederum neue Theorien und Experimente.

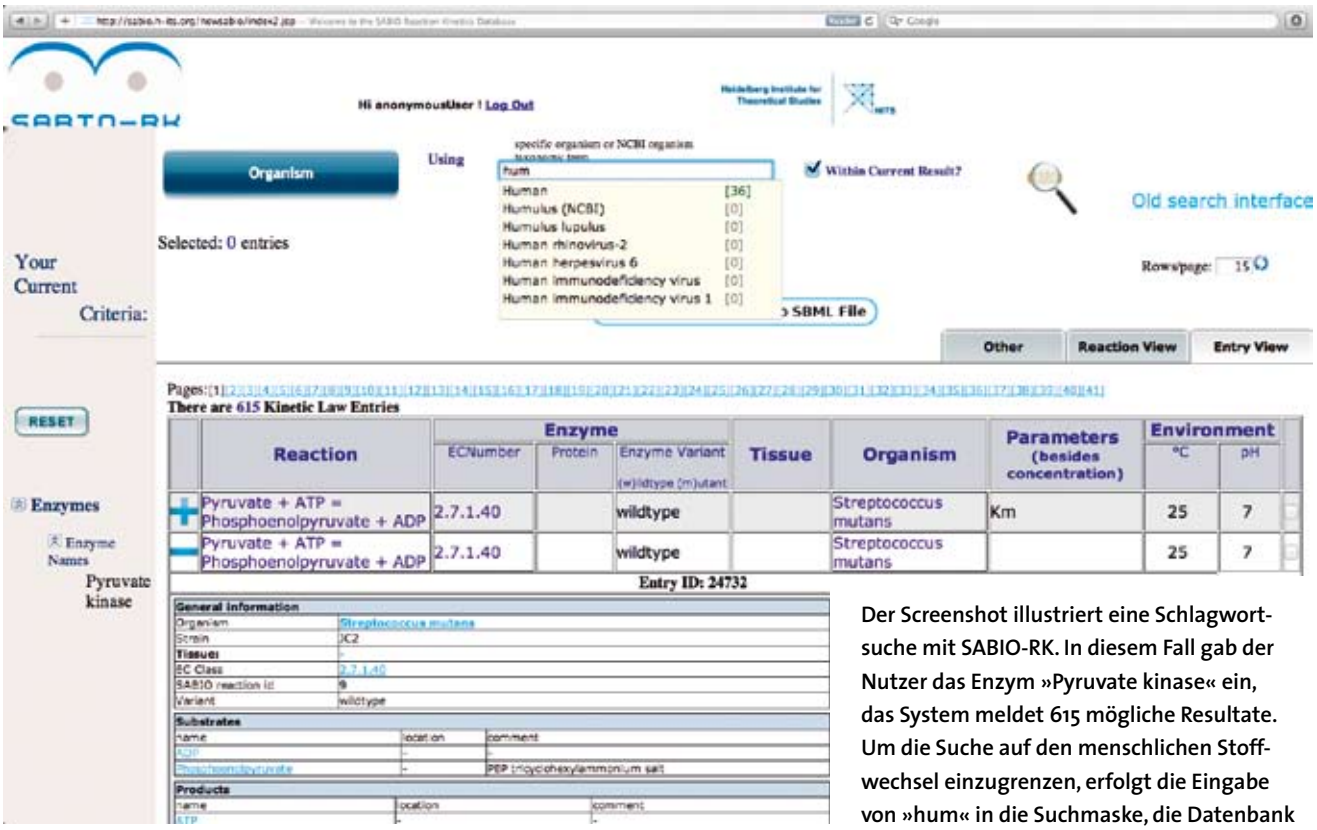
Und gerade an dieser Stelle helfen Datenbanken. Denn der Austausch über gedruckte Journale ist nicht nur langsam, es fällt Wissen-

Reaktionsketten im Visier der Forscher

Beim Glukosestoffwechsel wird das Zuckermolekül in einer Reaktionskette zu Pyruvat umgesetzt; es entstehen außerdem die Energieträger Adenosindiphosphat und Adenosintriphosphat (ADP und ATP). Immer wieder greifen dabei Enzyme wie die Hexokinase ein und katalysieren einen Zwischenschritt. Wie schnell aber laufen die Reaktionen ab, welchen Einfluss haben die Stoffkonzentrationen und die Umgebungsbedingungen? Solche Zusatzinformationen zu Stoffwechselwegen enthält die Datenbank SABIO-RK.



SPEKTRUM DER WISSENSCHAFT / ART FOR SCIENCE



Der Screenshot illustriert eine Schlagwortsuche mit SABIO-RK. In diesem Fall gab der Nutzer das Enzym »Pyruvate kinase« ein, das System meldet 615 mögliche Resultate. Um die Suche auf den menschlichen Stoffwechsel einzugrenzen, erfolgt die Eingabe von »hum« in die Suchmaske, die Datenbank liefert dazu eine Reihe von Vorschlägen.

schaftlern auch zunehmend schwerer, aus der gesamten Flut an Informationen nur die das jeweilige Thema betreffenden herauszufiltern. Aktuell verzeichnet PubMed, eine der wichtigsten Publikationsverzeichnisse für die Medizin, allein zur Leber – dem zentralen Organ des Stoffwechsels – mehr als 700 000 Veröffentlichungen. Um die jeweils relevanten zu ermitteln und daraus die für eine bestimmte Fragestellung wichtigen Daten zu entnehmen, benötigt ein Forscher die Unterstützung der elektronischen Medien.

Hierzu hat unsere Gruppe die Datenbank SABIO-RK *(System for the Analysis of Biochemical Pathways – Reaction Kinetics)* entwickelt. Wie es der Name andeutet, enthält sie von uns aufbereitete Angaben zu Stoffwechselwegen. So genannte Biokuratoren wählen zunächst potenziell nützliche Artikel anhand der Zusammenfassungen in PubMed aus. Hilfskräfte lesen diese Publikationen und geben die daraus entnommenen Daten zunächst in eine nichtöffentliche Version der Datenbank ein. Nun kommen wieder die Biokuratoren zum Zuge, die zum einen darauf achten, dass Gleiches gleich gespeichert wird. Dies betrifft sowohl die formale Struktur der Informationen als auch die Detailtiefe eventueller zusätzlicher Kommentare. Zum anderen sollte Gleiches auch gleich bezeichnet sein.

Über eine Suchmaske mit geeigneten Filtern kann ein Nutzer auf die Datenbank zugreifen – etwa nach Reaktionen suchen, an denen bestimmte Moleküle beteiligt sind. Die Informationen werden zudem auf Wunsch als SBML-Dateien ausgegeben, also in der Systems Biology Markup Language, einem international standardisierten Dateiformat der systembiologischen Modellierung. Ferner gibt es Verknüpfungen zu anderen Datensammlungen: So kann man sich mit einem Klick bei ChEBI *(Chemical Entities of Biological Interest)*, einer Datenbank, die am European Bioinformatics Institute in Hinxton (England) entwickelt und gepflegt wird, weitere Informationen zu einem Reaktionspartner holen.

Problematische Vielfalt der Namen

Für diese Arbeit benötigen wir mehr als die Expertise in der Informatik. Es genügt nicht zu wissen, wie Nutzer in einer Datenbank suchen und wie man sie dabei optimal unterstützen kann. Wir müssen auch verstehen, wie Systembiologen Daten gewinnen und einsetzen. Zudem ist SABIO-RK zwar einerseits eine Webanwendung, die wie ein Ingenieurserzeugnis geplant und gebaut werden muss. Darum herum ranken sich aber andererseits auch interessante Forschungsthemen.

So sind die Namen der reagierenden Stoffe oft nicht eindeutig, was die Forderung, Gleiches gleich zu benennen, zu einer anspruchsvollen Aufgabe macht. Beispielsweise bezeichnen das deutsche »Wasser« und die chemische Formel H_2O die gleiche Substanz. Für das englische *water* listet die Datenbank ChEBI nicht weniger als 14 Synonyme auf.

Auch die IUPAC, eine internationale Organisation, die regelt, wie chemische Verbindungen zu bezeichnen sind, lässt hier viel Spielraum. Ein Beispiel aus dem Glukosestoffwechsel: Glyceraldehyd-3-Phosphat, das korrekt auch als 3-Phosphoglyceraldehyd geschrieben werden kann, denn die standardisierte Nomenklatur erlaubt die Umstellung von Namensteilen.

Eine Vereinheitlichung ist bereits Teil der Kuratierung. So darf es schon bei der Eingabe nur entweder Glucose oder Glukose geben. Genauer gesagt, speichern wir nicht einen Textnamen, sondern die standardisierten Be-

zeichner der ChEBI: Der Glukose entspricht dort der Identifikator ChEBI:17234, der eindeutig und sprachunabhängig ist. Um eine derartige Umsetzung in einen standardisierten Bezeichner schon bei der Eingabe von Suchbegriffen durch die Nutzer zu unterstützen, lassen sich gängige Verfahren der Sprachverarbeitung wie Stemming-Algorithmen leider nicht einsetzen. Diese bilden Worte auf einen gemeinsamen Wortstamm ab, könnten beispielsweise für »gehst« und »geht« die Basis »geh« finden.

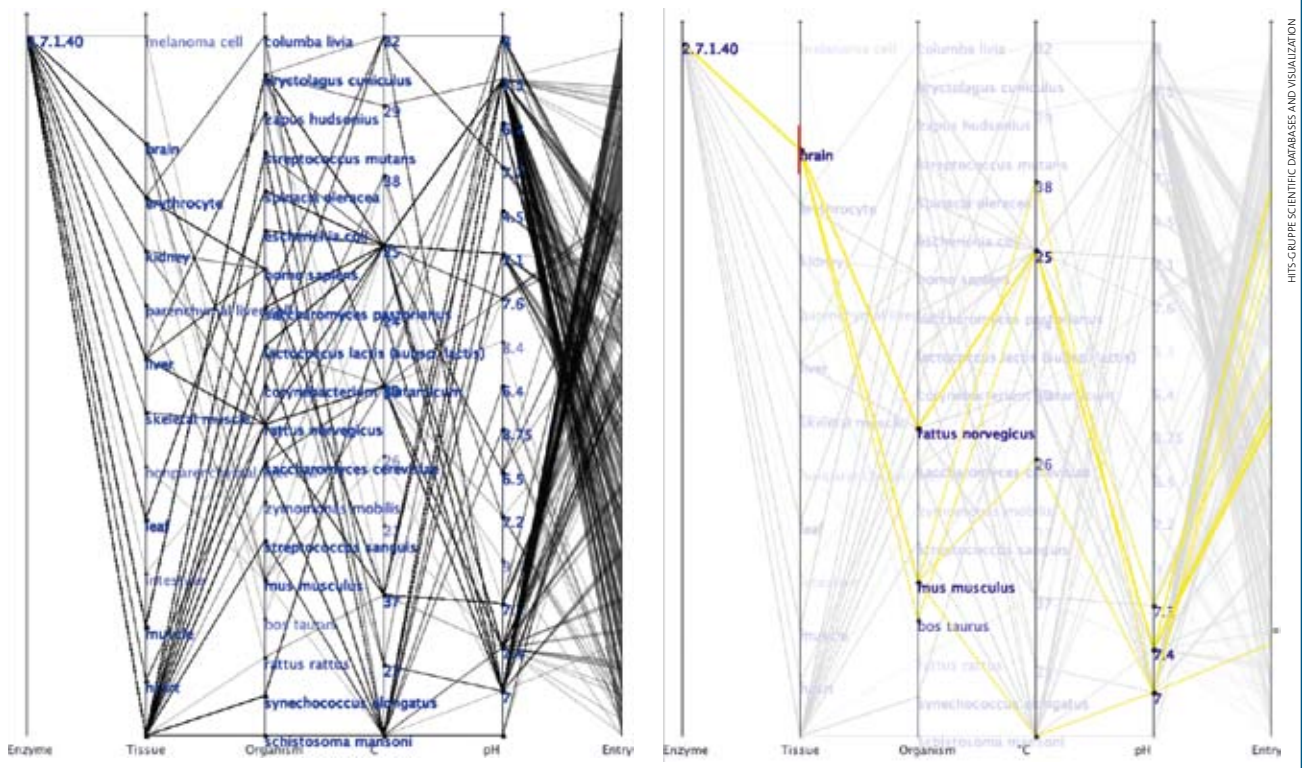
In langjähriger Zusammenarbeit mit der Gruppe von Uwe Reyle an der Universität Stuttgart entstanden zwei neue Verfahren zur Namen-Normalisierung. Das eine folgt einem morphologischen Ansatz, untersucht also die Form des Wortes. Dazu müssen wir jeden Namen in Wortbestandteile zerlegen. Diese werden sortiert, manche durch andere ersetzt. Die einzelnen Schritte sind jeweils so gewählt, dass Wörter gleichen Sinns auf gleiche künstliche Wörter abgebildet werden. Beispielsweise entfernt dieses Verfahren in den IUPAC-konformen englischen Bezeichnungen *1-butanol* und *butan-1-ol* die Bindestriche, sortiert die Wortbestandteile und kommt in beiden Fällen zu dem identischen Ergebnis *1butanol*.

Der zweite Ansatz hingegen beschäftigt sich mit dem Sinn der Wörter, ist also semantischer Natur. Dieser Algorithmus übersetzt Molekülbezeichnungen in chemische Strukturformeln und käme damit im Beispielfall ebenfalls zu dem Ergebnis, dass die zwei verschiedenen Wörter identische chemische Strukturen bezeichnen. Zwar wird aus dem Problem, korrekte Wort-Transformationsregeln zu suchen, nun die Aufgabe, Molekülnamen korrekt in Strukturen umzusetzen. Doch kann der semantische Ansatz viel mehr, vermag sogar mit Überbegriffen umzugehen: Sucht man etwa nach einer Reaktion eines Alkohols mit einem anderen Molekül, wäre der semantische Ansatz der Namen-Normalisierung im Vorteil, einerlei um welchen Alkohol es sich handelt; ein morphologischer Ansatz müsste hierzu stark erweitert werden.

Wir verfolgen deshalb beide Verfahren parallel. Die morphologische Methode steht kurz vor dem Einsatz, die semantische ist davon noch weiter entfernt. In der aktuellen Implementierung arbeitet sie auch deutlich langsamer. Als Anbieter einer Dienstleistung müssen wir uns fragen, mit welchem Aufwand wir

Stöbern in der Datenbank mit parallelen Koordinaten

Die Datenbank SABIO-RK entspricht einem vieldimensionalen Raum. So genannte parallele Koordinaten ermöglichen dennoch eine intuitive Herangehensweise. Das Beispiel beschränkt die Suche auf sechs Dimensionen: Enzym, Gewebetyp, Organismus, Umgebungstemperatur, pH-Wert und den Eintrag in der Datenbank (Entry-ID). Fragt man nach Reaktionen, die durch eine Pyruvatkinase katalysiert werden, ergibt sich das linke Bild. Offenbar wären Daten für verschiedene Zelltypen wie Melanome oder Erythrozyten abrufbar (linke Grafik), im Fokus der Suche stehen aber nur Informationen zu Gehirnzellen. Markieren des Kreuzungspunkts »Brain Tissue« lässt nicht relevante Linien verblassen. Auf einen Blick sieht der Nutzer nun beispielsweise, dass er Messergebnisse für Experimente an Ratten *(Rattus norwegicus)* abrufen kann (rechte Grafik). Sofern er sich aber für die Kinetik des Enzyms bei 26 Grad Celsius interessiert, müsste er auf Messungen an Mäusen *(Mus musculus)* zurückgreifen.



wie viel Resultat im Sinne unserer Nutzer bekommen. Gibt es Problemstellungen, bei denen er gern mehrere Sekunden wartet, bis ein Ergebnis vorliegt? Die Diskussion ist noch offen. Klar ist aber, dass wir mit beiden Ansätzen nicht immer richtigliegen. Ihr einziger Zweck ist es, den vom Nutzer gewählten Bezeichner einmalig in einen standardisierten umzusetzen. Intern wird dann nur noch dieser verwendet.

Wer in SABIO-RK nach bestimmten Reaktanten sucht, profitiert von dieser Namen-Normalisierung, denn er muss sich keine Gedanken darum machen, wie die Substanz oder das Enzym in den für ihn wichtigen Publikationen bezeichnet wurde. Viele Nutzer bringen aber weniger Vorwissen über unsere Datensammlung mit. Prinzipbedingt enthält sie nur einen sehr kleinen, aber gut gewählten und relevanten Teil der veröffentlichten reaktionskinetischen Daten.

Das Mantra der Datenbanksuche

Wir wollen einem Biologen ermöglichen, schnell herauszufinden, ob die für ihn wichtigen Daten in unserer Sammlung vorhanden sind. Das lässt sich vielleicht mit einem Kunden vergleichen, der ein Kaufhaus betritt, um eine ihm passende Hose zu finden. Vielleicht ist schon die Größe klar, aber andere Merkmale kommen erst bei der Suche selbst in den Sinn (Jeans, blau, elastischer Stoff).

Die meisten der heutzutage gebräuchlichen Such-Interfaces ermöglichen das, was die Computerwissenschaftler Ben Shneiderman und Catherine Plaisant von der University of Maryland *fact finding* und *extended fact finding* nennen. Hier geht es im Wesentlichen darum, bereits vorhandenes Wissen zu ergänzen wie: Es gibt eine Naturkonstante c, die Lichtgeschwindigkeit im Vakuum. Wie groß ist diese? Oder im Kontext der Systembiologie: Wie stark bindet ein bestimmtes Enzym bei einem pH-Wert von fünf an sein Substrat?

Dem Kunden im Kaufhaus wäre damit nicht geholfen, da er nur sehr vage Vorgaben machen kann. Dennoch wird er das gewünschte Produkt finden, da Einkaufszentren Anhaltspunkte zur Navigation geben. So wird er die Parfümerie- oder Süßwarenabteilung ignorieren, die richtige Etage für die Herrenoberbekleidung ansteuern, dort zu den Hosenständern gelangen, eine Grobauswahl anhand der Größen treffen und so fort.

Diesem Vorgehen entsprechen so genannte explorative Suchansätze. Sie sollen einem Nutzer sozusagen ein Gefühl für die Datensammlung vermitteln. Shneiderman hat dazu 1996 sein Visual Information Seeking Mantra für die Datenbankprogrammierung formuliert: *»Overview first, zoom and filter, details on demand«* – zunächst gilt es, einen Überblick zu vermitteln, dann immer näher heranzugehen und Daten herauszufiltern, schließlich Details bei Bedarf anzuzeigen.

Das bekannte Webprogramm Google Maps ist ein schönes Beispiel für eine gelungene Realisierung dieses Mantras. Nehmen wir an, Sie möchten einen abgelegenen Campingplatz in Südfrankreich finden, dann würden Sie von der Weltkugel ausgehend nach Südfrankreich zoomen, dort besonders grüne Regionen und darin wiederum nach Campingplätzen suchen. Erst dann kommen Details: Wie heißt der Ort, wie der Campingplatz, wie haben Nutzer ihn bewertet? Und all das geht so schnell vonstatten, dass kaum jemand bemerkt, wie die Satellitenkarte mit jedem Zoom nachgeladen wird.

Dass diese Technik auch für die Wissenschaft taugt, beweist SubtiPathways, eine Entwicklung der Universität Göttingen. Anstatt der Landkarten präsentiert es Stoffwechselwege; an manchen Orten darauf sind weitere Informationen hinterlegt. Dieser Ansatz eignet sich aber nur für Datensätze mit maximal fünf Dimensionen. Volker Springels Sternsimulationen (siehe den Beitrag S. 10) sind dafür ein beeindruckendes Beispiel: Verschiedene Energiedichten im dreidimensionalen Raum werden durch Helligkeiten als vierte Dimension dargestellt, deren zeitliche Veränderung erweitert die Simulation um eine fünfte Achse.

Stoffwechselpfade sind aber vielschichtiger: Es ist wichtig, an welchem Organismus, welchem Gewebe und welchem Zelltyp eine Messung durchgeführt wurde; oft sind vier bis fünf verschiedene Moleküle an einer Reaktion beteiligt, die sich je nach den äußeren Bedingungen unterschiedlich verhalten. Um einen solchen Prozess auf unseren dreidimensionalen Anschauungsraum abzubilden – und dann Techniken wie in Google Maps einzusetzen –, müssten wir Dimensionen weglassen.

Für deren Auswahl gibt es zwar durchaus Verfahren, wünschenswert im Sinne der explorativen Suche wäre aber nach wie vor, die Gesamtheit intuitiv erfassen zu können. Die von uns favorisierte Lösung sind so genannte parallele Koordinaten. Hier werden Punkte in hochdimensionalen Räumen nicht auf den dreidimensionalen Anschauungsraum projiziert, sondern als miteinander verknüpfte Linienzüge dargestellt (siehe Kasten linke Seite). Dabei entspricht jeder dieser Züge einer Dimension, daher die Bezeichnung des Verfahrens: Die Koordinatenachsen werden parallel zueinander gestellt, ein Punkt im vieldimensionalen Raum auf einen Linienzug in einer Fläche abgerollt.

Das Grundverfahren wurde bereits im ausgehenden 19. Jahrhundert entwickelt und seitdem auf verschiedene Problemstellungen angepasst. Seine Anwendung in Suchszenarien ist dennoch keineswegs trivial, da viele Fragen zu beantworten sind: Wie sollte man die Achsen anordnen, wie Kreuzungen besser kenntlich machen, wie dem Nutzer die Auswahl der ihn interessierenden Bereiche erleichtern? Für all diese Fragestellungen gibt es generelle und auch für das Problem angepasste Antworten.

Die beschriebenen Techniken – die Namen-Normalisierung ebenso wie die explorative Suche – haben zum Ziel, dem Nutzer einen Überblick zu geben und in ihm Erwartungen an Suchresultate zu wecken, die das System dann auch erfüllen kann. Um dies gut machen zu können, müssen wir erfahren, wie die Nutzer arbeiten, sowie ihre Wünsche und ihre Prioritäten kennen. Wir müssen als interdisziplinär arbeitende Gruppe in der Lage sein, Vorschläge zu machen. Dabei sind unsere Kuratoren wichtig, die aus der Biologie und Biochemie kommen und die Rolle der Nutzer übernehmen können, gleichzeitig aber auch informatisches Verständnis haben. Für viele andere Fragestellungen hingegen ist die direkte Zusammenarbeit mit Systembiologen außerhalb der Gruppe unerlässlich.

DER AUTOR

Wolfgang Müller studierte Experimentalphysik in Konstanz und parallel dazu Informatik an der Fernuniversität Hagen. Er habilitierte sich an der Universität Bamberg mit einer Arbeit zur Suche in selbstorganisierten verteilten Systemen. Seit 2009 leitet er die SDBV-Gruppe, die 1999 am EML Research, dem Vorgänger des HITS, von Isabel Rojas gegründet wurde.

Kreativ durch Analogien

Gleiche Strukturen erkennen bei Dingen, die auf den ersten Blick nichts miteinander gemein haben: Das ist das Arbeitsprinzip, mit dem die interdisziplinäre Computerlinguistik ihre Erfolge erzielt.

Von Michael Strube

Die Computerlinguistik vereinigt Elemente von Informatik und Linguistik; sie verwendet darüber hinaus Methoden aus weiteren Gebieten wie Mathematik, Psychologie, Statistik und künstliche Intelligenz. Der Reiz und die Herausforderung einer solchen interdisziplinären Wissenschaft liegen darin, Analogien zwischen Konzepten aus weit entfernten Teilgebieten zu erkennen und zu nutzen.

Paradebeispiel dafür ist einer der entscheidenden Durchbrüche, welche die Computerlinguistik prägten. Es geht um das »Parsing«: Ein Computerprogramm, genauer gesagt ein Compiler, nimmt Zeichen für Zeichen den Input des Benutzers entgegen, der in diesem Fall seinerseits aus dem Text eines Computerprogramms besteht, und ermittelt dessen Struktur. Im Prinzip dasselbe tut ein Mensch, der einen gesprochenen Satz hört und versteht.

Diese Analogie ist noch nicht besonders bemerkenswert, weil die Entwickler der Programmiersprachen und der zugehörigen Parserprogramme von Anfang an stark von der Linguistik beeinflusst waren; da verwundert es nicht, dass sie deren Denkstrukturen übernommen haben. Aber die Analogie funktioniert auch in Gegenrichtung. Erst als die Informatiker Methoden aus dem Kompilieren formaler Sprachen – insbesondere Programmiersprachen – auf natürliche Sprache übertrugen, wurde das Parsing von gewöhnlichen Sätzen überhaupt effektiv berechenbar. Erst dann konnten sie also Programme schreiben, die einen normalen, gesprochenen Satz hören und in akzeptabler Zeit zumindest seine grammatische Struktur erkennen.

Mehr noch: Ein solches Programm soll vor dem eigentlichen Parsing kontinuierliche Sprache erkennen, das heißt im pausenlosen Strom der gesprochenen Laute einzelne Wörter und damit auch die Grenzen zwischen den Wörtern ausfindig machen, und das unabhängig von der Person des Sprechers und mit großem Wortschatz. Diese Aufgabe in ausreichender Qualität zu lösen, gelang erst mit Hilfe einer weiteren Analogie. Man interpretiert das Sprachsignal als verrauschte, das heißt durch zufällige Störungen verunreinigte Version einer Zeichenkette, die dekodiert werden muss. Dank der neuen Betrachtungsweise lassen sich nun statistische Methoden aus der Informationstheorie anwenden.

Koreferenzresolution mit annotierten Paaren

As we know, Putin has kept putting off this visit to Japan since last year, like back then when Yeltsin repeatedly postponed his trip to Japan.
That is to say, Japan asked for too high a price.
That is, it asked the Russian president to come to Japan to make concessions on territorial issues.
Well, well, the Russian president was still unwilling, was unwilling to make concessions.

Im Text oben sind als koreferent erkannte Erwähnungen farbig unterlegt und durch gleichfarbige Striche miteinander verbunden. Hier kommt es nicht nur darauf an zu verstehen, dass »his« sich auf »Yeltsin« bezieht und »it« auf Japan, sondern auch darauf, dass mit »the Russian president« »Putin« gemeint ist. Letzteres erfordert sogar Weltwissen, nämlich dass zu der Zeit, als dieser Text geäußert wurde, nicht mehr Boris Jelzin, sondern Wladimir Putin russischer Präsident war.

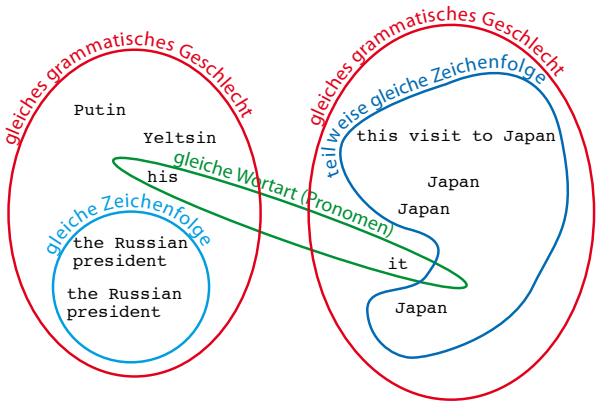
Als Trainingsmaterial für Lernprogramme dienen Listen (»Annotationen«) aus Paaren von Erwähnungen, zum Beispiel aus obigem Text, mit der – von menschlichen Bearbeitern hinzugefügten – Angabe, ob diese Erwähnungen sich auf denselben Gegenstand beziehen (koreferent sind, blauer Strich) oder nicht (roter Strich).

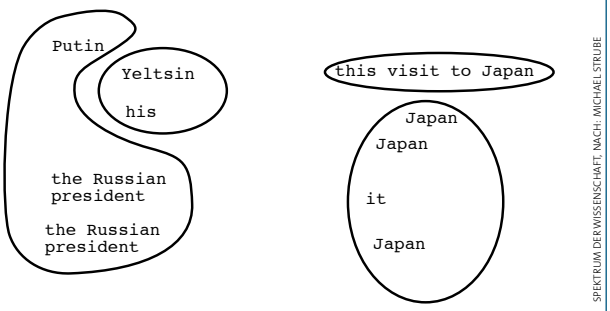
Putin	this visit to Japan	rot
Putin	Japan	rot
this visit to Japan	Japan	rot
Japan	Yeltsin	rot
this visit to Japan	Yeltsin	rot
Putin	Yeltsin	rot
Yeltsin	his	blau
his	Japan	rot
Yeltsin	Japan	rot
Japan	Japan	blau
Japan	it	blau
it	the Russian president	rot
Japan	the Russian president	rot
Japan	the Russian president	rot
his	the Russian president	rot
Yeltsin	the Russian president	rot
Japan	the Russian president	rot
this visit to Japan	the Russian president	rot
Putin	the Russian president	blau
...	...	

SPEKTRUM DER WISSENSCHAFT

Koreferenzresolution mit Hypergraphen

Das Programm definiert zunächst mit Hilfe einzelner Merkmale Teilmengen aller Erwähnungen (Hypergraphen) als Kandidaten für Koreferenzketten. Die Teilmengen sind im linken Bild durch farbige Umrandungen dargestellt. Sie werden dann mit Hilfe von Algorithmen der linearen Algebra verrechnet; das Ergebnis ist die korrekte Zerlegung (rechts).





SPEKTRUM DER WISSENSCHAFT, NACH: MICHAEL STRUBE

Von den so entwickelten Methoden profitierte schließlich immens die maschinelle Übersetzung. Hier trägt dieselbe Analogie: Die Ausgangssprache wird als verrauschte Version der Zielsprache angesehen. Obwohl die automatische Übersetzung auf den ersten Blick nichts mit der Spracherkennung gemein hat, erkannten Computerlinguisten eine Strukturähnlichkeit und übertrugen den Lösungsansatz von der Spracherkennung auf die automatische Übersetzung.

Ist »er« Putin oder Jelzin?

Hier wird ein Muster deutlich: Man löst ein computerlinguistisches Problem, indem man eine Analogie zu einem scheinbar entfernten Gebiet erkennt – natürliche Sprachen und Programmiersprachen, Spracherkennung und Informationstheorie, maschinelle Übersetzung und Spracherkennung. Zwei Studien aus meiner Arbeitsgruppe zeigen im Folgenden, wie eine solche Übertragung im Einzelfall geleistet werden kann.

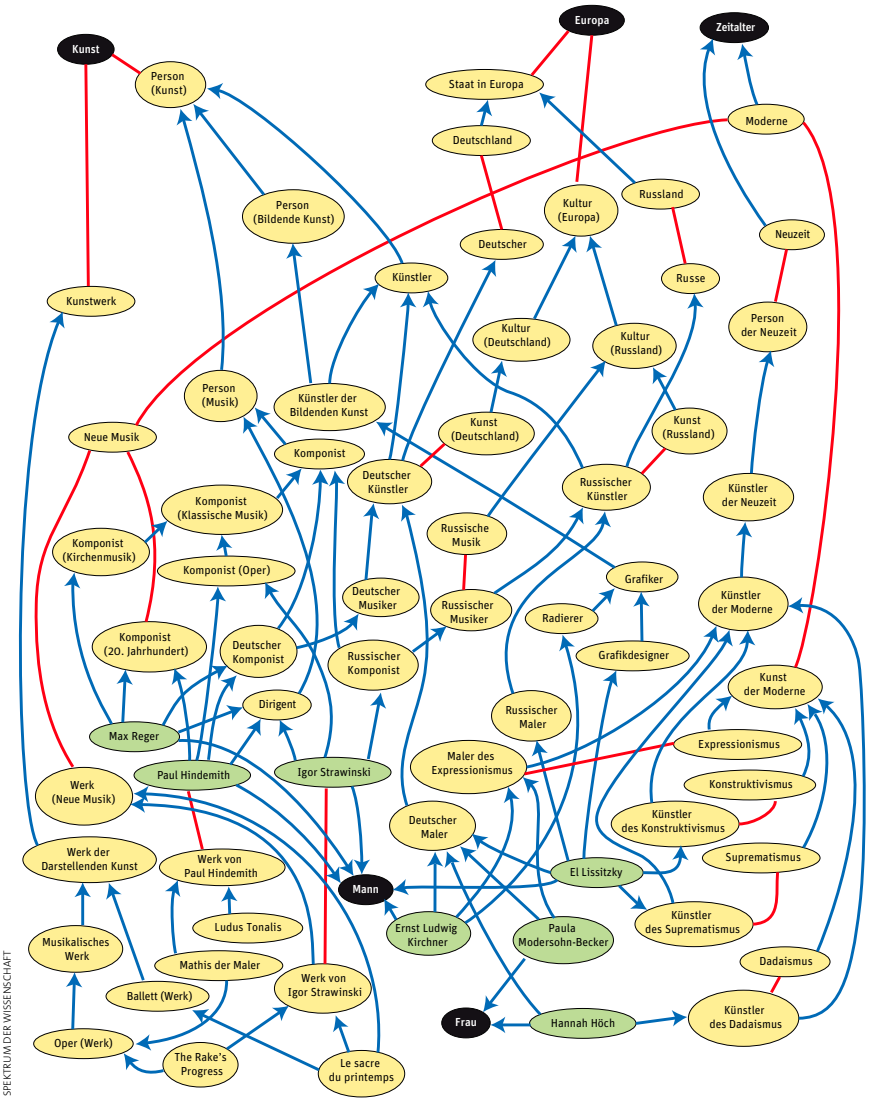
Eine wichtige Aufgabe beim automatischen Verstehen von Texten ist die so genannte Koreferenzresolution: zu erkennen, dass sich mehrere Ausdrücke im Text (»Erwähnungen«) auf denselben Gegenstand beziehen (»koreferieren«). Eine Erwähnung kann zum Beispiel ein Eigenname in unterschiedlichen Varianten, ein Pronomen oder auch eine zusammengesetzte Nominalphrase sein. In dem Text im Kasten links sind die Erwähnungen »Putin« und »the Russian president« koreferent, ebenso »Yeltsin« und »his« sowie »Japan« und »it«. Formal gesprochen kommt es darauf an, alle Erwähnungen in Teilmengen aufzuteilen, deren Elemente zueinander koreferent sind; und natürlich darf eine Erwähnung nicht zwei verschiedenen Teilmengen angehören. Diese Mengen heißen auch »Koreferenzketten«, weil sie häufig, wie im Kasten, durch verbindende Striche dargestellt werden.

Frühe Arbeiten in der Computerlinguistik griffen Erkenntnisse aus der Linguistik auf und stellten komplexe Regeln für die Koreferenzresolution auf, die eine vollständige syntaktische und häufig auch semantische Analyse des Textes voraussetzten. Da dieser Ansatz nicht robust genug für eine Anwendung im größeren Stil war, wurden seit den späten 1990er Jahren zunehmend Verfahren des maschinellen Lernens eingesetzt: Ein Programm leitet automatisch aus von Menschen vorgegebenen korrekten Lösungen, die als Trainings- und Testdaten dienen (»Annotationen«), Regeln oder statistische Zusammenhänge ab.

Damit die Standardverfahren des maschinellen Lernens angewendet werden können, arbeitet man mit Paaren von Erwähnungen. Eine Annotation besteht aus einer Liste solcher Paare mitsamt der Angabe, ob die beiden Erwähnungen eines Paars koreferent sind oder nicht (Kasten links, rechte Grafik). Das Programm lernt nicht nur danach, es gibt zu einem neu vorgelegten Text Listen von Paaren aus. Diese »paarweise Klassifikation« hat den Vorteil, dass sie bekannten und gut verstandenen Methoden des maschinellen Lernens zugänglich ist. Nachteil ist, dass Wissen um den Kontext verloren geht. So kann es einem solchen Programm durchaus passieren, dass es »Putin« und »Yeltsin« durch

Glossar

- **Syntax** ist die grammatische Struktur eines Textes, Semantik seine Bedeutung.
- **Parsing:** Einen Eingabetext Zeichen für Zeichen entgegennehmen, dabei Grenzen zwischen bedeutungstragenden Elementen (»Wörtern«) und in gewissen Grenzen die Struktur des Texts erkennen.
- Zwei Ausdrücke im Text (»Erwähnungen«) **koreferieren**, wenn sie denselben Gegenstand bezeichnen.
- **Koreferenzresolution** ist die Identifizierung koreferenter Erwähnungen.
- **Annotation** ist ein von einem menschlichen Bearbeiter mit Zusatzinformationen versehenes Textbeispiel für das maschinelle Lernen.
- Ein **Synset** ist eine Menge annähernd synonymer Ausdrücke in der Datenbank WordNet.



SPEKTRUM DER WISSENSCHAFT

Dieser kleine Ausschnitt aus dem Kategoriennetz der deutschen Wikipedia konzentriert sich auf die nähere Umgebung der Einträge zu einigen Künstlern vom Beginn des 20. Jahrhunderts. Blaue Pfeile haben die Bedeutung »ist ein« (»ein russischer Komponist ist ein Komponist«, »Igor Strawinski war ein Dirigent«), rote Linien kennzeichnen das Wissen, dass eine solche Beziehung nicht besteht.

eine lange Koreferenzkette verbindet – und damit in einen Topf wirft –, weil an irgendeiner Stelle das Pronomen »his« zum einen wie zum anderen passt und schon die Feststellung, dass »his« nur einen der beiden Herren meinen kann, über die Betrachtung einzelner Paare hinausgeht.

Dies war – stark vereinfacht – der Stand der Forschung, als vor drei Jahren Jie Cai als Doktorandin in meiner Arbeitsgruppe anfing. Wir fragten uns, wie man das Problem der Koreferenz angemessener repräsentieren und insbesondere Wissen über den Kontext in die Entscheidung mit einbeziehen kann. Dabei legten wir das oben beschriebene Konzept zu Grunde, dass Koreferenzketten eigentlich Mengen sind und es darum geht, jede Erwähnung genau einer Menge zuzuweisen. In der Informatik fanden wir ein geeignetes Analogon zu dieser Aufgabe: die Clusteranalyse. Man ordne Datenpunkte Mengen (»Clustern«) zu, und zwar so, dass eng benachbarte Datenpunkte in der Regel in ein und denselben Cluster geraten. Nur kann man zwar zu zwei (durch Koordinaten gegebenen) Datenpunkten in einfacher Weise deren Entfernung definieren; aber das funktioniert für Erwähnungen nicht. Allenfalls sind Erwähnungen Punkte (»Knoten«) in einem Graphen, die genau dann durch eine Kante verbunden sind, wenn sie koreferent sind; aber das ist ja erst das Ergebnis der Analyse und nicht der Ausgangspunkt. Diese Kanten wiederum drücken nichts weiter aus als eine paarweise Klassifikation und bieten daher keinen Fortschritt.

Weiter kommt man mit einem neuen Konzept. Ein Hypergraph ist ein verallgemeinerter Graph, bei dem eine Kante mehr als zwei Knoten miteinander verbinden kann. Damit ist er die graphentheoretische Entsprechung einer Menge, und wir haben eine angemessene Darstellung des Koreferenzproblems gefunden: Erwähnungen sind Knoten im Hypergraphen, und jeder Gegenstand ist eine Hyperkante, die alle seine koreferenten Erwähnungen umfasst. Das Problem der Koreferenzresolution kann dann als Clusteranalyse für Hypergraphen aufgefasst werden.

Mit diesem neuen theoretischen Rahmen ist unser Programm zur Koreferenzresolution nicht mehr ausschließlich auf die Beispielpaare der paarweisen Klassifikation angewiesen. Vielmehr zieht es eine Vielzahl von »Merkmalen« *(features)* heran. Ein Merkmal ist ein Indiz dafür, dass zwei Erwähnungen im Prinzip koreferent sein können. Eines von ihnen zeigt an, ob Erwähnungen der gleichen semantischen Klasse angehören, also zum Beispiel beide eine Person, einen Ort oder ein Fahrzeug bezeichnen. Ein anderes Merkmal stellt dar, ob Erwähnungen die gleiche Zeichenkette enthalten (»Präsident Putin«, »Wladimir Putin«, »Putin«, …). Weitere Merkmale enthalten Wissen über grammatische Eigenschaften einer Erwähnung wie Genus, Numerus und Person, über ihre syntaktische Rolle (Subjekt, Objekt, …) oder bestimmte syntaktische Beziehungen zwischen zwei Erwähnungen, etwa dass eine Erwähnung Apposition einer anderen ist (»Wladimir Putin, der russische Präsident, …«). Auch der Abstand im Text zwischen zwei Erwähnungen, gezählt in Wörtern oder Sätzen, wird als Merkmal ausgedrückt. Insgesamt arbeiten wir mit etwa 20 unterschiedlichen Merkmalen.

Unser Programm erstellt im ersten Schritt zu jedem Merkmal einen Satz von Hyperkanten. Diese sind manchmal gewöhnliche Kanten, verbinden also nur zwei Erwähnungen, zum Beispiel bei dem Merkmal für den Abstand im Text. Die meisten aber umfassen mehr als zwei Erwähnungen; sie machen die Stärke des Verfahrens aus. Allen Hyperkanten werden mit Hilfe von annotierten Trainingsdaten Gewichte zugewiesen; das sind Zahlen,

die bezeichnen, wie stark das mit dem Merkmal ausgedrückte Indiz für Koreferenz ist. Da das Verfahren robust ist gegenüber kleinen Abweichungen bei den Gewichten, kommt es mit fünf Prozent der Trainingsdaten aus, die für die paarweise Klassifikation erforderlich sind. Das ist von entscheidender Bedeutung, da Annotationen für jedes Sachgebiet neu erstellt werden müssen, viele Stunden menschlicher Arbeit erfordern und daher teuer sind.

Die mit Gewichten versehenen Hyperkanten lassen sich in Matrizen umwandeln. Die wiederum kann man mit Standardmethoden aus der linearen Algebra so transformieren, dass am Ende eine korrekte Zerlegung in Mengen koreferenter Erwähnungen steht (Kasten S. 31 oben).

Wikipedia als lexikalische Datenbank

In Experimenten mit Standarddatensätzen konnten Jie Cai und ich zeigen, dass unsere Methode trotz deutlich geringeren Bedarfs an Lernstoff wesentlich bessere Ergebnisse bei der Koreferenzresolution erzielt als die üblichen Verfahren, und das in etwa einem Viertel der Rechenzeit. Wegen des geringen Trainingsaufwands ist es uns auch gelungen, unser Verfahren ohne größere Mühe auf eine neues Sachgebiet zu übertragen: Inzwischen analysiert es nicht nur Nachrichtentexte, sondern auch Arztberichte.

Aufgaben wie die Koreferenzresolution benötigen über das linguistische Wissen (»Ist ›Putin‹ ein Substantiv oder ein Verb?«) hinaus auch Wissen über Objekte in der Welt und ihre Beziehungen zueinander (»Ist ›Putin‹ ein Mensch oder ein Ort?«). Koreferenzrelationen bestehen häufig zwischen einem Unter- und einem Oberbegriff, etwa »Wladimir Putin« und »der russische Präsident«, »der russische Politiker«. Im Oktober 2005 stellte sich meinem damaligen Doktoranden Simone Paolo Ponzetto und mir die Frage, wie wir unserem Koreferenzresolutionssystem dieses Wissen zur Verfügung stellen können.

Die in der Computerlinguistik populärste Ressource für derartiges Wissen ist »WordNet«, eine lexikalische Datenbank, die Wörter so genannten »Synsets« zuordnet, die jeweils eine Menge (annähernd) synonymer Ausdrücke enthalten. Die Synsets sind in einer Taxonomie angeordnet und durch viele weitere semantische Relationen miteinander verknüpft, so dass sich ein reichhaltiges semantisches Netzwerk ergibt. WordNet enthält aber nur wenig Wissen über durch mit Eigennamen bezeichnete Konzepte. So gibt es in der aktuellen Version (Stand 30. Mai 2011) zwar einen Eintrag über »Vladimir Putin«; »Boris Yeltsin« hat allerdings nie Eingang in die Datenbank gefunden. Wir waren also auf der Suche nach einer Wissensquelle, die mehr Informationen über durch Eigennamen bezeichnete Konzepte enthält und dennoch so gut strukturiert ist wie WordNet.

Ein Blick auf die im Oktober 2005 noch recht kleine »Wikipedia« zeigte uns, dass diese Online-Enzyklopädie die erste Bedingung erfüllt. Die zweite Bedingung erforderte einen erneuten, unbefangenen Blick. Im Gegensatz zu gewöhnlichen, unstrukturierten Webseiten enthält Wikipedia neben dem ebenfalls unstrukturierten Text einige Strukturelemente, die unserer Aufgabe dienlich waren. So findet man am Ende jedes Artikels die Liste der Kategorien, denen er angehört. Die Kategorien selbst sind ebenfalls kategorisiert, so dass man mit ihrer Hilfe von einem Artikel zu einem anderen gelangen kann, der mit dem ersten semantisch verwandt ist.

Damit war klar: Wenn es gelingt, aus den Wikipedia-Kategorien ein semantisches Netz zu extrahieren, dann verfügt man über eine Ressource, die WordNet zumindest bei den durch Eigennamen bezeichneten Konzepten überlegen ist. In der Folge haben Ponzetto und ich (später stieß Vivi Nastase als Postdoc zum Team) mehrere Verfahren entwickelt, die Wikipedia zuerst in ein semantisches Netzwerk umwandeln, dann in eine Taxonomie und schließlich in ein Netzwerk mit reichhaltigen semantischen Relationen (Spektrum der Wissenschaft 12/2010, S. 94; Bild S. 33). Die Anwendung auf mehrere computerlinguistische Probleme belegte die Richtigkeit unserer Grundannahme.

Die beiden hier beschriebenen Projekte weisen eine Gemeinsamkeit auf. Beim Problem der Koreferenzresolution kam es darauf an, auf einer abstrakten Ebene die Strukturgleichheit zwischen dem linguistischen Phänomen der Koreferenz, dem mathematischen Konzept der Menge und dem graphentheoretischen Konstrukt des Hypergraphen zu sehen. Bei der Wissensextraktion aus Wikipedia ging es darum, das Kategoriensystem in Wikipedia als Netzwerk zu erkennen, dessen Kanten semantische Nähe ausdrücken und dessen Knoten – Wikipedia-Artikel und -Kategorien – den »Synsets« aus WordNet entsprechen. Hat man diese Strukturgleichheit erst einmal gefunden, ist es relativ leicht, sie zu nutzen – in diesem Fall Wikipedia in ein semantisches Netzwerk umzuformen und darauf weitere Strukturen aufzubauen.

In beiden Beispielen war es entscheidend, Analogien zwischen auf den ersten Blick nicht zusammenhängenden Gebieten zu erkennen. In einem interdisziplinären Gebiet wie der Computerlinguistik gilt dies auch eine Abstraktionsstufe höher: Es kommt darauf an, Analogien zwischen Analogien zu sehen. »Good mathematicians see analogies between theorems or theories. The very best ones see analogies between analogies«, so der bedeutende Mathematiker Stanislaw Ulam (1909–1984) in seinem Werk »Analogies between analogies«.

Die wissenschaftliche Umgebung bei HITS stellt in dieser Beziehung eine einmalige Chance dar, da die Interdisziplinarität zu den Voraussetzungen seiner Existenz zählt. Vielleicht werde ich eines Tages sogar Methoden aus der Biomechanik oder der theoretischen Astrophysik auf computerlinguistische Probleme anwenden!

DER AUTOR

Michael Strube, Jahrgang 1965, wurde 1996 an der Universität Freiburg mit einer Dissertation in Computerlinguistik promoviert. Nach einer Postdoc-Zeit an der University of Pennsylvania in Philadelphia kam er 2000 als wissenschaftlicher Mitarbeiter zur EML Research gGmbH in Heidelberg. Ein Jahr später wurde er Leiter der Natural Language Processing Group des Instituts, das mittlerweile Heidelberger Institut für Theoretische Studien heißt. Er ist Honorarprofessor an der Universität Heidelberg im Fach Computerlinguistik.

QUELLEN

Cai, J., Strube, M.: End-to-End Coreference Resolution via Hypergraph Partitioning. In: Proceedings of the 23rd International Conference on Computational Linguistics, Peking, 23.–27. August 2010, S. 143–151. Download über www.aclweb.org/anthology/C/C10/

Ponzetto, S. P., Strube, M.: Taxonomy Induction Based on a Collaboratively Built Knowledge Repository. In: Artificial Intelligence 175, S. 1737–1756, 2011

Virtuelle Forschungsumgebungen für morgen

Um Wissenschaftlern die Infrastruktur bieten zu können, die sie für ihre Arbeit in der Zukunft brauchen, müssen Hochschulen und außeruniversitäre Institutionen ihre Kräfte bündeln und neue Wege beschreiten.

Von Uwe Schwiegelshohn

Nur dort, wo der Boden und das Angebot an Wasser und Licht ihren Bedürfnissen genügen, werden Pflanzen gedeihen und Frucht tragen. Genauso verhält es sich auch mit der Wissenschaft: Ein Forscher benötigt eine seinem Thema angemessene Umgebung, um herausragende Ergebnisse zu erzielen. Das war schon in der Antike so, wobei sich die erforderliche Infrastruktur im Lauf der Jahrhunderte freilich beträchtlich erweitert hat. Doch auch wenn wir heute von »virtuellen Forschungsumgebungen« sprechen, sind die Grundbedürfnisse doch erstaunlich gleich geblieben. Gelehrte brauchen vor allen Dingen eines: die Möglichkeit, sich mit anderen Experten ihres Fachs auszutauschen.

Weil diese beiden Grundpfeiler jeder Forschung Ende des 4. Jahrhunderts v. Chr. am Museion Alexandrias gegeben waren, wurde es zur zentralen Stätte antiker Gelehrsamkeit. Nirgends sonst beherbergte eine Bibliothek eine solche Vielzahl an Schriften – hunderttausende sollen es gewesen sein. Nicht anders als heute ermöglichten diese frühen Publikationen eine indirekte Kommunikation zwischen Forschern über Generationen hinweg.

Auf Grund seiner Bedeutung wurde das Museion oft von den Großen der Zeit geleitet, etwa von Eratosthenes, der den Erdumfang und die Schiefe der Ekliptik vermaß, oder von dem frühen Sprachwissenschaftler Aristophanes. Selbst längere Reisen und die damit verbundenen Gefahren schreckten Wissen Suchende nicht ab. Seine einzigartige Ausstattung verdankte das Museion dem Engagement des ptolemäischen Herrschergeschlechts. Obwohl die Wirtschaftsmacht ihres Landes noch nicht davon abhing, wissenschaftliche Erkenntnisse in technische Innovationen umzumünzen, legten diese Könige großen Wert darauf, den »Forschungsstandort« Alexandria attraktiv zu machen und so in der Antike den Wettbewerb um die klügsten Köpfe zu gewinnen.

Während es damals nur wenige solcher Stätten der Gelehrsamkeit gab, änderte sich die Situation im Spätmittelalter deutlich. Mit dem Untergang des Römischen Reichs im 5. Jahrhundert war eine Phase weit gehender wissenschaftlicher Stagnation angebrochen. Nun aber wurden die antiken naturphilosophischen Erkenntnisse wiederentdeckt, und die Mächtigen ihrer Zeit gründeten Universitäten als neue Form, Studium und Forschung eine Heimat zu geben. Das Modell erwies sich als erfolgreich. Um 1230 gab es bereits etwa 20 solcher Einrichtungen in Europa, 1789 waren es schon 142. Im deutschen Sprachraum vollzog sich diese Entwicklung etwas langsamer. Bis 1400 gab es erst drei Universitäten, bis 1500 wuchs ihre Zahl aber auf zehn. Heute sind es in Deutschland allein ungefähr 100. Hinzu kommen noch andere Arten von Hochschulen und außeruniversitäre Forschungseinrichtungen.

Niedergang der Bibliotheken

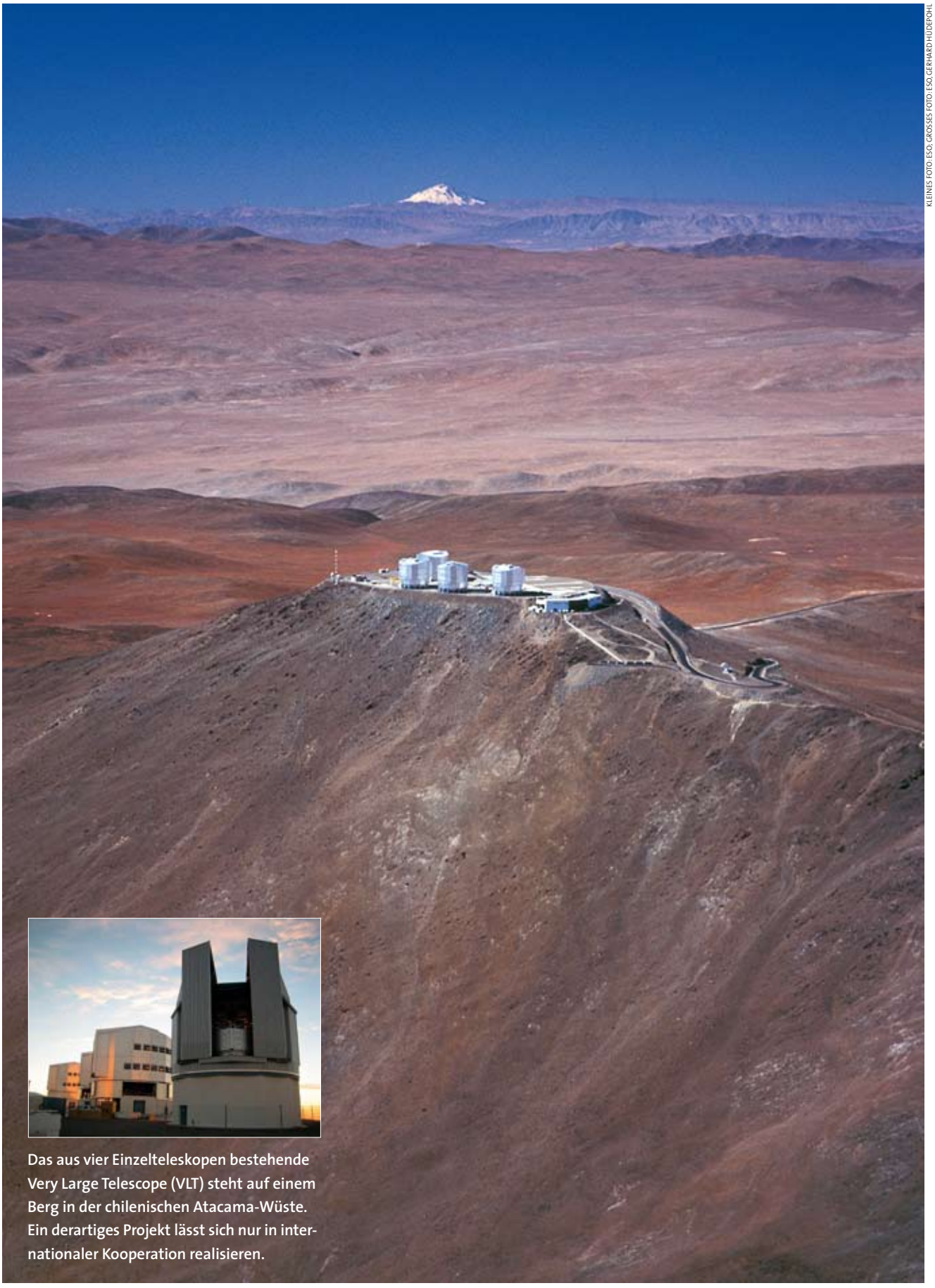
Auch in den mittelalterlichen Universitäten spielten die Bibliotheken eine tragende Rolle, und daran hat sich bis in die Gegenwart nichts geändert. Nach wie vor ist die Publikation das primäre Mittel, Forschungsergebnisse in der Fachwelt zu verbreiten. Seit der Erfindung des Buchdrucks durch Johannes Gutenberg Mitte des 15. Jahrhunderts lassen sie sich leicht vervielfältigen – und dank des Aufkommens der Zeitungen und schließlich des Wissenschaftsjournalismus auch einer breiten Öffentlichkeit vermitteln.

Dennoch gab es seit den Zeiten Galileo Galileis (1564–1642) eine strukturelle Veränderung: Schriften – und damit die Bibliotheken – verloren mit der Einführung von Experiment und Beobachtung ihre Alleinstellung im Erkenntnisprozess. Dieser Trend setzt sich bis in die Gegenwart fort, in der neben den Bibliotheken als weitere Querschnittsfunktion die Rechenzentren zur Verarbeitung von Forschungsdaten aufkamen. Inzwischen wendet eine typische technische Universität durchschnittlich weniger als zwei Prozent ihres jährlichen Etats für die Ausstattung ihrer Bibliothek auf, hingegen über fünf Prozent für Laborräume und technische Einrichtungen. Dies war und ist die Konsequenz einer veränderten Forschungslandschaft, in der sich die Natur- von den reinen Geisteswissenschaften lösten und größeren Raum einnahmen.

Da Experimente disziplinspezifisch sind, erfordern sie unterschiedliche Forschungsumgebungen. Angesichts einer wachsenden Zahl von Teildisziplinen wird es für eine Universität immer aufwändiger, das ganze Spektrum der Wissenschaften abzubilden, auch wenn sich Forschungsumgebungen bei vergleichbaren Fragestellungen durchaus ähneln.

Mitunter benötigen wissenschaftliche Instrumente spezifische Einsatzorte. Dazu gehören astronomische Teleskope, die einen dunstfreien Himmel erfordern (siehe Foto rechts), oder die polaren Beobachtungsstationen und Forschungsschiffe der Klimaforscher. Die effiziente Nutzung dieser weit entfernten Beobachtungsstandorte verlangt, große Datenvolumina von dort schnell zu den jeweiligen Wissenschaftlern an ihren Heimatuniversitäten zu übermitteln.

Während Gelehrte noch Anfang des 20. Jahrhunderts in Briefwechseln Informationen austauschten und Theorien diskutierten, wollen Forscher heute mit anderen ohne Verzögerung und unabhängig vom Aufenthaltsort in Verbindung treten können. Das leisten die modernen, globalen Kommunikationssysteme, darunter vor allem das Internet. Hierfür



KLEINES FOTO: ESO, GROSSES FOTO: ESO, GERHARD HÜDEPOHL

Das aus vier Einzelteleskopen bestehende Very Large Telescope (VLT) steht auf einem Berg in der chilenischen Atacama-Wüste. Ein derartiges Projekt lässt sich nur in internationaler Kooperation realisieren.

die technische Infrastruktur bereitzustellen, ist ebenfalls eine Kernkompetenz der Rechenzentren. Sie versetzen nicht nur Wissenschaftler – vorwiegend Naturwissenschaftler – aus unterschiedlichen Einrichtungen und Ländern in die Lage, gemeinsam zu forschen und zu veröffentlichen, sie unterstützen auch den zunehmend interdisziplinären Charakter der Wissenschaft. Fragt man beispielsweise, welche Auswirkungen die globale Erwärmung und die mit ihr einhergehenden Veränderungen von Lebensräumen auf die Verbreitung von Krankheiten haben werden, sind Spezialisten verschiedener Fachrichtungen gefragt.

Von modernen Forschungsumgebungen wird erwartet, dass sie eine solche Vernetzung unterstützen. Das Vorhandensein der genannten Kommunikationssysteme allein genügt dafür nicht mehr. Ebenso wichtig wird die Kompatibilität zwischen früher isoliert funktionierenden Laboren. Das erfordert die Einrichtung und Pflege möglichst standardisierter Schnittstellen, was ebenfalls in die Kompetenz der Rechenzentren fällt und ihre Bedeutung noch steigert.

Die wachsende Anzahl der Universitäten und die immer aufwändigere technische Ausstattung lässt freilich die Kosten steigen. So entstand der Begriff der Forschungsinfrastruktur, die neben der Qualität der darin enthaltenen Forschungsumgebungen auch organisatorische Aspekte wie die Kosteneffizienz berücksichtigt. Eines der ersten Beispiele dafür waren die in den 1970er Jahren entstandenen deutschen Bibliotheksverbünde. Zunächst wurden Zentralkataloge geschaffen, in denen ein Titel nur einmal aufgeführt ist, was die Katalogpflege vereinfacht. Später kam die Onlinefernleihe hinzu, die einzelne Büchereien entlastete, weil nun keine mehr ein komplettes Literaturangebot vorhalten musste. Dank der modernen Kommunikationstechnik verlängert sich die Wartezeit nur geringfügig.

AKG BERLIN / BIBLIOTHÈQUE DE TROYES, MS 129, FOL. 32

Das Studium von Schriften bildete im Mittelalter die Basis aller Gelehrsamkeit (oben: ein Hörsaal der Pariser Sorbonne im 15. Jahrhundert). Mit dem Aufkommen des Experiments im 16. Jahrhundert verloren Bibliotheken an Bedeutung. Manche Forschungsfragen

Eine weitere Neuentwicklung ist die virtuelle Forschungsumgebung, die annähernd die gleiche Funktionalität wie eine ideal ausgestattete lokale aufweist, obwohl nicht mehr alle Komponenten am Standort existieren. Voraussetzung ist eine gesteuerte Kooperation zwischen den Trägern der lokalen Umgebungen und die Vernetzung der einzelnen Forscher untereinander. Somit ist die typische Forschungsumgebung von heute immer eine virtuelle, da sie unterschiedliche Standorte verbindet.

Das betrifft insbesondere wissenschaftliche Experimente, die aus Kostengründen nur an wenigen oder gar nur an einem Ort durchgeführt werden können. Man denke etwa an die Projekte aus dem Bereich der Teilchenphysik am Large Hadron Collider (LHC) am CERN (Europäische Organisation für Kernforschung) in der Nähe von Genf. Solche Organisationen wurden gegründet, um Großexperimente durchzuführen. Sie sind in der Regel durch öffentliche Mittel finanziert und bilden eine wesentliche Säule von Forschungsinfrastrukturen. Umfangreiche und den angeschlossenen Wissenschaftlern zugänglich zu machende Datenvolumina entstehen aber auch durch die Digitalisierung von Literatur und Kultur oder im Zuge einer Vielzahl klei-

FELIX MÜLLER, BERN

Im 3. Jahrhundert v. Chr. avancierte die Universität Alexandrias zu einem Zentrum des Wissens. Von christlichen Fanatikern zerstört, wurde sie im 4. Jahrhundert n. Chr. andernorts neu gebaut. Polnische Archäologen glauben, einige der Hörsäle entdeckt zu haben.

CERN

erfordern die Konzentration von Experimentiereinrichtungen an einem Ort (im Bild oben der LHC am CERN). Von dort müssen die Ergebnisse über Kommunikationsnetzwerke zu den über die ganze Welt verteilten Wissenschaftlern weitergeleitet werden.

nerer Studien mit aufwändigen Bilddaten wie in der Medizin. Diese Informationen werden in der Regel in großen Archiven gesammelt und wiederum anderen Forschern zur Verfügung gestellt.

Sowohl aus dem Interesse der Beteiligten als auch aus Effizienzgründen sollten all diese Daten möglichst vielen Gruppen zugänglich sein. Damit entsteht ein Bedarf an virtuellen Forschungsumgebungen, die institutionsübergreifend aufgebaut sind. Im Gegensatz zum erwähnten Bibliotheksverbund oder zum Rechenzentrum wären diese zwar disziplinspezifisch, durch Synergieeffekte würden aber die Kosten reduziert.

Rechnen in der Wolke

Virtuelle Forschungsumgebungen sollen vor allem notwendige Dienste für die beteiligten Wissenschaftler anbieten, angefangen von der Verbindung zu anderen Forschern, wissenschaftlichen Geräten oder Datenspeichern an weit entfernten Orten bis hin zur Bereitstellung und Pflege benötigter Software für die Auswertung von Messergebnissen. Es wäre ineffizient, wenn solche Software von jedem Wissenschaftler selbst erstellt werden müsste, wie dies in der Vergangenheit oft der Fall war. Nachdem aber jetzt die Institutionsgrenzen einmal aufgebrochen sind, bietet es sich an, sie auch hier zu überschreiten. Bei den LHC-Experimenten ist das durch das so genannte Worldwide LHC Computing Grid (WLCG) – ein aus miteinander kommunizierenden Rechnern auf der ganzen Welt bestehendes Netzwerk – bereits in Ansätzen geschehen.

Das erfordert eine dienstleistungsorientierte Softwaretechnologie. Ein Beispiel dafür ist das so genannte Cloud Computing. Eine solche »Rechnerwolke« besteht aus einem Netzwerk von Computern, aus dem ein Anbieter die nachgefragten Ressourcen dynamisch zuweist. Letzterer weiß also nicht mehr, wo konkret jene Maschinen stehen, die seine Daten oder eine bestimmte Software vorhalten – all das bleibt ihm wie hinter einer Wolke verborgen.

Hier bietet sich eine weitere Chance für die Hochschulen, Kosten zu sparen und gleichzeitig ein Mehr an Infrastruktur zu bieten. Gegenwärtig versorgen ihre Rechenzentren noch die vor Ort arbeitenden Forscher. Angesichts der Fragmentierung der Hochschullandschaft in viele Disziplinen mit jeweils nur einer kleinen Zahl von Wissenschaftlern pro Universität lässt sich auf diese Weise kaum die nötige hohe Auslastung erreichen. Vernünftiger wären Rechenzentrumsverbünde analog den Bibliotheksverbünden; erste Schritte in diese Richtung wurden sowohl innerhalb von Bundesländern als auch über Bundesländergrenzen hinweg schon unternommen.

Im Extremfall könnte ein Rechenzentrum die gesamte Infrastruktur stellen – beziehungsweise als eigenständiges Unternehmen ausgegliedert werden; man spricht von Infrastructure-as-a-Service. Dieser Ansatz ist vor allem in solchen Fächern sinnvoll, die die verfügbare Technologie möglichst optimal ausnutzen wollen. Geht es dagegen nur um die Ausführung von bestimmten Programmen, etwa zur statistischen Auswertung, ist das Konzept Software-as-a-Service interessanter. Ein Forscher könnte dann eine speziell für seine Aufgabenstellung entwickelte Software verwenden, ohne sich um deren Implementierung kümmern oder selbst über die notwendige Hardware verfügen zu müssen.

Das würde auch die lokalen Rechenzentren entlasten, da sie in Zukunft kaum in der Lage sein werden, die Vielzahl unterschiedlicher Anwendungssoftware für die jeweils wenigen Nutzer bereitzustellen und zu pflegen. Zudem sinkt das Risiko, dass ein in einem Projekt entwickeltes Verfahren vergessen und in einem anderen neu entwickelt wird.

Die entstehende Forschungsinfrastruktur besitzt dann zwei Komponenten, deren Zusammenspiel noch nicht geklärt ist: Auf der einen Seite übernimmt die Universität disziplinübergreifend die Strukturierung der Forschungsumgebung vor Ort, auf der anderen arbeitet der Träger einer virtuellen disziplinspezifischen Forschungsumgebung über die Grenzen der Institutionen hinweg. Offen sind bis jetzt die Mechanismen der Zusammenarbeit und die Finanzierung solcher Infrastrukturen. Um den Forschungsstandort Deutschland auch für die Zukunft gut zu positionieren, sollten diese Fragen so schnell wie möglich gelöst werden.

DER AUTOR

Uwe Schwiegelshohn leitet das Institut für Roboterforschung der Technischen Universität Dortmund, wo er sich vor allem auf die Gebiete Grid Computing und autonome mobile Roboter konzentriert. Er ist zudem Prorektor für den Geschäftsbereich Finanzen der Hochschule. In diesem Rahmen befasst er sich auch mit fakultätsübergreifenden Fragen der Strukturentwicklung.

Wissenschaft braucht Vernetzung

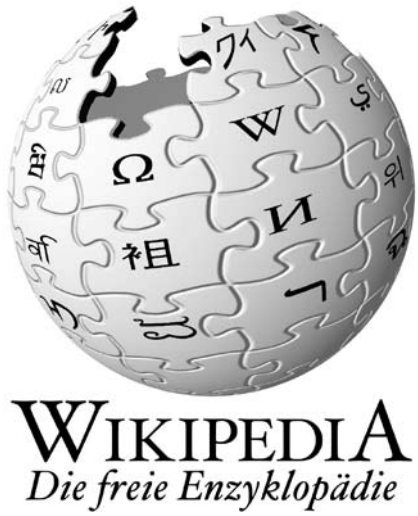
Forscher können der rapide anwachsenden Datenmengen nur Herr werden und sie zum rascheren Erkenntnisgewinn nutzen, wenn sie ihre Rolle als Mitglieder eines großen Netzwerks verstehen und akzeptieren. Dies erfordert neue Formen des Umgangs mit urheberrechtlichen Fragen und neue Modalitäten der Zusammenarbeit.

Von John Wilbanks

Die Gewinnung neuer Erkenntnisse durch die Analyse großer Datensammlungen wird oft als »viertes Paradigma« wissenschaftlichen Arbeitens bezeichnet. Unabhängig davon, ob man dem zustimmt, ist es sinnvoll, die ursprüngliche Bedeutung des Begriffs Paradigmenwechsel in Thomas Kuhns »Structure of Scientific Revolutions« noch einmal zu reflektieren.

Kuhns Modell beschreibt eine Welt der Wissenschaft, in der ein System von Ideen die Vorherrschaft erringt, sich etabliert und so eine Sicht der Welt hervorbringt (das »Paradigma«), die für sich selbst Macht und Einfluss gewinnt. Dieses System von Ideen bezieht seine Geltung daraus, dass es eine plausible Erklärung für beobachtbare Phänomene liefert. Auf diese Weise haben wir zum Beispiel den Äther als Träger des Lichts bekommen sowie die Miasmen-Theorie für Infektionskrankheiten und die Vorstellung, dass die Sonne um die Erde kreist. Das System von Ideen, die Sicht der Welt, das Paradigma verfestigt sich durch schrittweise Erweiterung. Jeder einzelne Wissenschaftler arbeitet in der Regel so, dass er das Paradigma Stück für Stück ergänzt. Wem es gelingt, ein großes Stück hinzuzufügen, der erlangt Autorität, Forschungsaufträge, Preise und Auszeichnungen – und Direktorenposten.

Alle Beteiligten profitieren von ihren Investitionen in ein System von Ideen, das über die Ideen selbst hinausreicht. Firmen und Regierungen (und die Leute, die für sie arbeiten) gründen Geschäftspläne und politische Vorgaben auf eine solche Sicht der Welt. Das führt zum Aufbau eines Schutzwalls – einer Art Immunsystem –, der das Weltbild gegen Angriffe abschirmt. Zweifler werden ins Abseits gedrängt. Neue Ideen fallen nicht auf fruchtbaren Boden, bekommen kein Geld und kein Personal. Furcht, Unsicherheit und Skepsis bestimmen die Reaktion auf originelle Vorstellungen, Methoden, Modelle und Ansätze, die dem herrschenden Paradigma zuwiderlaufen.



Das Onlinelexikon Wikipedia ist das bekannteste Beispiel einer für alle frei zugänglichen Website, welche die Gemeinschaft der Internetnutzer weltweit unentgeltlich aufgebaut hat, stetig erweitert, pflegt und aktualisiert.

Doch Weltanschauungen gehen unter und Paradigmen stürzen, wenn sie die Beobachtungen nicht mehr erklären können oder wenn ein Experiment zweifelsfrei nachweist, dass sie falsch sind. Der Äther hat sich nach Hunderten von Jahren stetiger Verfeinerung als Schimäre erwiesen, und so erging es dem Miasma und dem Geozentrismus. Die Zeit für einen Wechsel ist dann gekommen, wenn die alten Erklärungsmuster den neuen Realitäten nicht mehr gewachsen sind.

Dies scheint mir die Idee hinter Jim Grays Begründung eines vierten Paradigmas und dem Bild von der »Datenflut« zu sein: dass unsere Fähigkeit, Daten zu messen, zu speichern, zu analysieren und zu visualisieren, die neue Realität ist, der sich die Wissenschaft stellen muss. Daten sind der Kern dieses neuen Paradigmas, und es steht auf einer Stufe mit dem, was wir für den wissenschaftlichen Methodenvorrat halten: der experimentellen Beobachtung, der Theoriebildung und der Simulation.

Müssen wir die ersten drei Paradigmen also begraben? Keineswegs, vielmehr will ich sie feiern. Mit der experimentellen Beobachtung und Theoriebildung sind wir weit gekommen – von einem Weltbild, in dem die Sonne um die Erde kreist, bis zur Quantenphysik. Simulation ist das Herzstück vieler aktueller Forschungsaktivitäten, von der Rekonstruktion des antiken Rom bis hin zur Wettervorhersage. Die Genauigkeit von Simulationen und Prognosen steht im Zentrum heißer politischer Debatten um die Wirtschaftsentwicklung und den Klimawandel. Und natürlich gilt, dass Beobachtung und Theorie unabdingbar sind für gute Simulationen. Ich kann auf meinem Bildschirm sehr schön etwas simulieren, in dem die Gravitation nicht vorkommt, aber wenn ich mit meinem Auto über einen Klippenrand fahre, wird mich die Schwerkraft gnadenlos wieder einholen.

So gesehen handelt es sich also nicht um einen Paradigmenwechsel im kuhnschen Sinne. Daten werden nicht die gute alte Realität beiseiteschieben. Stattdessen stellen sie eine Reihe von Anforderungen an die Methoden und Konventionen, mit denen wir über Beob-

achtungen und Theorien kommunizieren, aber auch an die Robustheit und Komplexität unserer Simulationen und schließlich an die Art, wie wir unser Wissen darlegen, weitergeben und integrieren.

Was sich ändern muss, ist das Paradigma von uns selbst als Wissenschaftlern, nicht die alten Paradigmen des Erkenntnisgewinns. Als wir anfingen zu begreifen, dass alles Stoffliche aus Atomen besteht, dass wir das Produkt unserer Gene sind und dass die Erde um die Sonne kreist – da vollzogen sich Paradigmenwechsel im kuhnschen Sinne. Was wir hier diskutieren, geht quer durch all diese Umbrüche. Datengetriebene Forschung, richtig verstanden, wird mehr Paradigmenwechsel bei wissenschaftlichen Theorien in kürzerer Zeit hervorbringen, weil wir unser jeweiliges Weltbild sofort mit der »objektiven Realität« vergleichen können, die sich so effektiv messen lässt.

Netzwerke beschreiben den Umbruch durch die Datenflut vielleicht besser als die kuhnsche Dynamik. Ihre Skalierbarkeit kommt der Beherrschung von Unmassen an Daten entgegen – Netzwerke können gewaltige Mengen von Informationen in etwas Nützliches verwandeln, so dass die Überfülle an Informationen nicht länger ein »Problem« ist, das »gelöst« werden muss. Und beim Umgang mit der Datenflut können wir vom Entwurf der Netzwerke lernen: Wenn wir ihrer Herr werden wollen, müssen wir eine offene Stra-

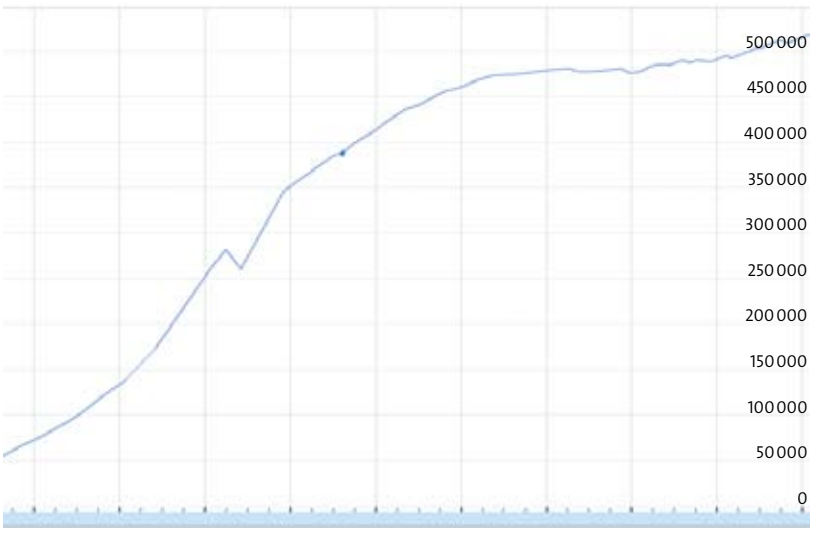
Der Apache-Webserver, der populärste Webserver im Internet, ist ein Paradebeispiel für Open-Source-software, die Enthusiasten weltweit unentgeltlich programmieren und der Allgemeinheit zur Verfügung stellen.

HTTP://WWW2.JUSCH.CH/DOKUS/DEBIANANWENDERHANDBUCH.DE/BILDER/WEBMIN/SERVERAPACHE.PNG

Freie Inhalte im Internet

Das Open Directory Project (ODP) gilt als größtes von Menschen gepflegtes Webverzeichnis des World Wide Web. Seine Inhalte sind für jeden kostenlos zugänglich und werden von freiwilligen Redakteuren unentgeltlich bearbeitet und aktualisiert. Die Grafik zeigt die Entwicklung der Einträge im deutschsprachigen Zweig des ODP.

QUELLE: HTTP://DE.WIKIPEDIA.ORG/W/INDEX.PHP?TITLE=DATEI:ODP_SITECOUNT_WORLD_DEUTSCH.PNG&FILETIMESTAMP=20100211081729



tegie verfolgen, die auf den Erfahrungen mit Netzwerken beruht.

Damit meine ich die Rechner- und Kommunikationsnetzwerke, die lediglich auf einem Satz von Protokollen aufgebaut sind, Schicht für Schicht, von Endpunkt zu Endpunkt. Das Internet und das Web sind anhand von Dokumenten realisiert worden, die standardisierte Methoden dafür definieren, wie Daten übertragen und dargestellt werden, und die zugleich ein Namensschema für Rechner und Dateien vorgeben. Da wir uns alle dieser Methoden bedienen und jeder sie nutzen kann, ohne um Erlaubnis zu fragen, entwickelt sich das Netzwerk ganz von selbst und wächst immer weiter.

So gesehen sind Daten nicht ein »viertes Paradigma«, sondern eine »vierte Netzwerkschicht« (auf dem Ethernet, TCP/IP und dem World Wide Web), die, von oben nach unten, mit den anderen Schichten kompatibel ist und zusammenwirkt oder »interoperiert«, wie Computerwissenschaftler sagen. Ich glaube, diese Sichtweise wird dem Wesen wissenschaftlicher Methodik eher gerecht als das Konzept eines Paradigmenwechsels mit seinem destruktiven Ansatz. Daten sind das Ergebnis allmählicher Fortschritte bei den Mess- und Beobachtungsverfahren. Sie untermauern die Theorie, sie treiben und validieren die Simulation, und sie werden am besten in standardisierter wechselseitiger Kommunikation mit den genannten Schichten des Wissensnetzwerks ausgetauscht.

Vorzüge des Prinzips der Offenheit

Krass gesagt ist das Paradigma, das zerstört werden muss, die Idee, dass wir Wissenschaftler als unvernetzte Individuen forschen. Wenn denn diese Metapher akzeptabel erscheint, hält sie für uns, die wir über den Entwurf eines Netzwerks für die wissenschaftliche Kommunikation nachdenken, zwei Lektionen bereit.

Die erste Lektion, frei nach David Isenberg, ist die, dass das Internet seine Durchschlagskraft einer ganz speziellen Eigenschaft verdankt: Es ist öffentlich. Das gilt gleich in mehrfacher Hinsicht. Die Definitionen der Standards, auf denen das Internet beruht, sind offen und frei zugänglich – frei zum Lesen, zum Herunterladen, zum Kopieren, zum Verwenden. Sie sind frei im urheberrechtlichen Sinn. Die Spezifikationen können von jedem herangezogen werden, der sie verbessern und

Eine Publikation von *Spektrum der Wissenschaft* und dem *Heidelberger Institut für Theoretische Studien*

Spektrum der Wissenschaft

Chefredakteur: Dr. Carsten Könneker
Editor-at-Large: Dr. Reinhard Breuer (v.i.S.d.P.)
Redaktionsleiter: Dr. Hartwig Hanser (Monatshefte), Dr. Gerhard Trageser (Sonderhefte)
Redaktion: Thilo Körkel (Online-Koordinator), Dr. Klaus-Dieter Linsmeier, Dr. Christoph Pöppe
Art Direction: Karsten Kramarczik
Layout: Sibylle Franz, Claus Schäfer
Schlussredaktion: Christina Meyberg (Ltg.), Sigrid Spies, Katharina Werle
Bildredaktion: Alice Krüßmann (Ltg.), Anke Lingg, Gabriela Rabe
Redaktionsassistenz: Anja Albat-Nollau, Britta Feuerstein
Verlag: Spektrum der Wissenschaft Verlagsgesellschaft mbH, Postfach 10 48 40, 69038 Heidelberg, Tel.: 06221 9126-600, Fax: 06221 9126-751 Amtsgericht Mannheim, HRB 338114
Verlagsleiter: Richard Zinken
Geschäftsleitung: Markus Bossle, Thomas Bleck

Heidelberger Institut für Theoretische Studien – HITS

Geschäftsführer: Dr. Klaus Tschira, Prof. Dr. Andreas Reuter
Presse- und Öffentlichkeitsarbeit: Dr. Peter Saueressig
Anschrift: HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Tel.: 06221 533-245, Fax: 06221 533-198

www.h-its.org

Spektrum der Wissenschaft CUSTOM PUBLISHING

Leitung: Dr. Joachim Schüring
Anschrift: Spektrum der Wissenschaft – Custom Publishing, Postfach 10 48 40, 69038 Heidelberg; Hausanschrift: Slevogtstraße 3–5, 69126 Heidelberg, Tel.: 06221 9126-612, Fax: 06221 9126-5612

www.spektrum.com/cp

Gesamtherstellung: L.N. Schaffrath Druckmedien GmbH & Co. KG, Marktweg 42–50, 47608 Geldern

Erscheinungstermin: Spektrum der Wissenschaft 12/2011

erweitern möchte, aber ihr Wert beruht nicht auf Optimierungen durch Einzelne, sondern darauf, dass sehr viele Menschen sie benutzen. Wie Isenberg anmerkt, bringt dies eine Reihe von »Wundern« hervor: Das Netzwerk wächst ohne zentrale Kontrollinstanz, es lässt uns Dinge verbessern, ohne um Erlaubnis zu fragen, es erschließt und fördert neue Märkte (denken Sie an E-Mail, Instant Messaging, soziale Netze – oder Pornografie). Versuche, die offene Struktur des Internets zu verändern, würden es in seiner Existenz gefährden. Das muss denjenigen unter uns, die in einer Welt der wirtschaftlichen Rivalitäten und der klassischen ökonomischen Theorien aufgewachsen sind, unbegreiflich erscheinen. Von ihrer Warte aus ist es widersinnig, dass Wikipedia existiert und noch dazu der Encyclopedia Britannica den Rang streitig macht.

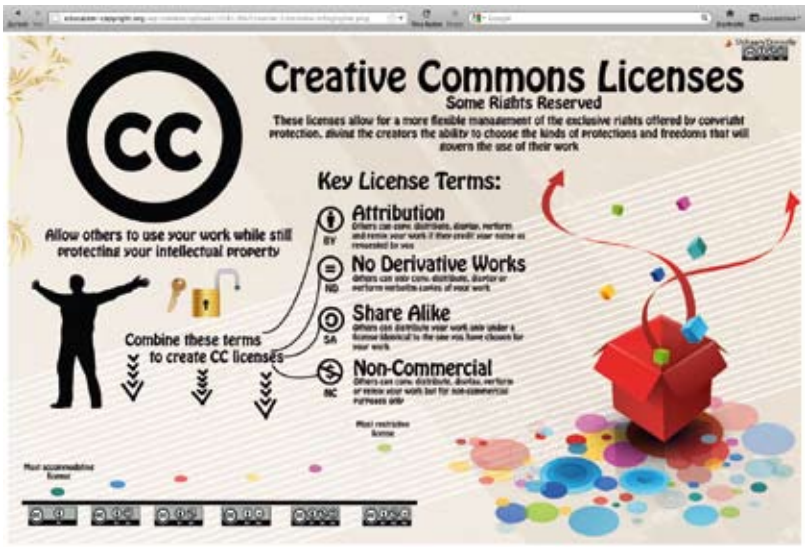
Aber, mit Galilei gesprochen: »Sie bewegt sich doch.« Wikipedia existiert, und das Netz – eine einvernehmliche Halluzination, die auf einer Sammlung technischer Standards beruht – transportiert Skype-Video-Anrufe zwischen mir und meiner Familie in Brasilien – und zwar umsonst. Es ist eine Innovationsmaschine wie keine je zuvor. Das Netz lehrt uns, dass neue Netzwerkschichten für den Umgang mit Daten die Idee der Offenheit beherzigen sollten – der Nutzung von Standards, die uns allen erlauben, frei zusammenzuarbeiten und die Segnungen des Netzes, die wir von der riesigen Dokumentensammlung des World Wide Web kennen, für die gigantischen Datensammlungen nutzbar zu machen, die wir so leicht zusammentragen können.

Die zweite Lektion kommt aus einer anderen offenen Welt, derjenigen der Open-Source-Software. Die Erstellung von Software nach dem Modell verteilter kleiner Einzelbeiträge, zusammengeführt durch technische und rechtliche Standardisierung, war auch so eine theoretische Unmöglichkeit, die durch die Realität des Internets einen wahrhaft kuhnschen Paradigmenwechsel erfuhr. Die Möglichkeit der jederzeitigen Kommunikation, verbunden mit günstigem Zugang zu Programmierwerkzeugen, und die weitsichtige Anwendung öffentlicher Urheberrechtslizenzen hatten einen seltsamen Effekt: Sie brachten Software hervor, die funktionierte und mit der Zeit immer umfangreicher und leistungsfähiger wurde. Die wichtige Erkenntnis ist, dass wir Millionen von Gehirnen anzapfen können, wenn wir standardisieren, und die so entstehenden Produkte in vielen Fällen besser sind als die in traditionellen, zentralisierten Umgebungen erzeugten. (Ein gutes Beispiel ist der Apache-Webserver, der seit 1996 der populärste Webserver im Internet ist.)

Creative Commons hat diese Lektionen auf die Lizenzierung angewendet und einen Satz von Standardlizenzen für digitale Medienprodukte entwickelt. Diese haben sich explosionsartig verbreitet und schützen mittlerweile mehrere hundert Millionen digitaler Objekte im Netz. Es zeigt sich, dass offene Lizenzen bemerkenswerte Vorteile haben: Sie ermöglichen (bei vernachlässigbaren Transaktionskosten) für digitale Objekte wie Musik oder Fotografien – und für wissenschaftliche Information – denselben Grad gemeinschaftlicher Nutzbarkeit, den wir von technischen Netzwerken kennen.

Urheberrechtsschutz für online publizierte Werke

Creative Commons bietet die Möglichkeit, geistiges Eigentum im Internet unter verschieden strikten Lizenzbedingungen zu veröffentlichen.



HTTP://EDUCATION-COPYRIGHT.ORG/WP-CONTENT/UPLOADS/2011/06/CREATIVE-COMMONS-INFOGRAPHIC.PNG

In diesem Lokal in Spanien ist nur Musik mit Creative-Commons-Lizenz aus dem Internet zu hören.

KLAUS GRAF / CC-BY-SA-2.5 (HTTP://CREATIVECOMMONS.ORG/LICENSES/BY-SA/2.5/DEED.DE)

Ich dachte,
ich spüre einen Paradigmenwechsel,
aber mir war nur die Unterhose hochgerutscht.

Scheinbar fehlende Anreize sind bei alldem der Punkt, der klassischen ökonomischen Theorien zuwiderläuft. Das ist ein anderes Beispiel für einen wahrhaft kuhnschen Paradigmenwechsel – die alte Theorie konnte keine Welt beschreiben, in der Menschen umsonst arbeiten, doch die neue Realität zeigt, dass genau dies passiert.

Forscher als Knoten im Netzwerk

Es gibt im Netz durchaus Widerstand gegen eine datenintensive Schicht. Doch der beruht längst nicht im gleichen Maß auf Urheberrechtsbedenken, wie das bei Software der Fall war (gleichwohl ist das Beharrungsvermögen des Urheberrechts groß, wenn es um die Anpassung der Fachgutachter-Kultur bei wissenschaftlichen Veröffentlichungen geht, was die »Webrevolution« in der wissenschaftlichen Literatur de facto verhindert). Zwar existieren im Zusammenhang mit Daten Urheberrechtsprobleme, aber Widerstand kommt noch von vielen anderen Seiten: Es ist schwierig, Daten mit Anmerkungen zu versehen und sie dann erneut zu benutzen, es ist schwierig, große Datenmengen zu übermitteln, es ist schwierig, Daten miteinander zu kombinieren, die nicht von vornherein dafür ausgelegt wurden, und so weiter. Dadurch haben Daten für alle außer denen, die sie erzeugen, eine sehr kurze Halbwertszeit. Dieser Widerstand hat seinen Ursprung im Paradigma von uns selbst als individuellen Wissenschaftlern, nicht in den Paradigmen der experimentellen Beobachtung, der Theoriebildung oder der Simulation.

Um ihn zu überwinden, müssen wir in Annotation und Qualitätssicherung investieren, in Hardware zur Speicherung und Wiedergabe von Daten sowie in die Grundlagen zu ihrer gemeinsamen Visualisierung und Analyse. Wir brauchen offene Standards, die es erlauben, Daten allen zugänglich zu machen und im Verbund zu nutzen. Wir brauchen eine verbindliche Definition für die Datenschicht. Und vor allem müssen wir Wissenschaftler aus allen Gebieten darin unterweisen, auf dieser neuen Datenschicht zu arbeiten. Solange unsere Ausbildungskultur von den Prinzipien der gildenartigen Mikrospezialisierung geprägt ist, wird der Wissenschaftsbetrieb der Datenschicht weiter erheblichen Widerstand entgegensetzen.

Wir sollten uns selbst als vernetzte Knoten sehen, die Daten weitergeben, Theorien testen und die Simulationen anderer Wissenschaftler benutzen. Angesichts der Tatsache, dass jede Kurve zur Beschreibung der Kapazitäten für das Sammeln von Daten exponentiell ansteigt, müssen wir unsere eigenen Kapazitäten zur Nutzung dieser Daten entsprechend steigern – und das schnellstmöglich. Wir müssen uns und unser Wissen vernetzen. Nichts, was die Menschen bislang hervorgebracht haben, wächst so schnell wie offene Netze.

Wie alle Vergleiche hat natürlich auch die Netzmetapher ihre Grenzen. Wissen ist schwieriger zu vernetzen als Dokumente. Ein kooperativer Arbeitsstil kann sich bei der Softwareentwicklung leichter von selbst herausbilden, weil die Werkzeuge billig und überall zugänglich sind – das trifft auf die Teilchenphysik oder Molekularbiologie nicht zu. Einige der großartigen Dinge im Web eignen sich nicht so gut für Wissenschaft und Forschung, weil das Prinzip der auf Konsens basierenden Einschätzungen nur die langweiligen Dinge zu Tage fördert, denen jeder zustimmt, aber nicht das Abgelegene, das oft viel interessanter ist.

Dennoch gibt es herzlich wenige Alternativen zum Netzwerkansatz. Die Datenflut ist da, und sie ebbt nicht ab. Wir können mehr und schneller messen als jemals zuvor. Und wir können Messungen in enormer Zahl gleichzeitig nebeneinander durchführen. Unsere Gehirnkapazität bleibt dagegen für alle Zeit auf ein Gehirn pro Person beschränkt. Wir müssen also zusammenarbeiten, wenn wir Schritt halten wollen, und Netzwerke sind die besten Kooperationswerkzeuge, die unsere Kultur hervorgebracht hat. Das aber bedeutet, dass wir unseren Umgang mit Daten genauso offen gestalten müssen wie die Protokolle, die Rechner und Dokumente miteinander verbinden. Es ist der einzige Weg, auf dem wir die erforderliche Leistungsstufe erreichen können. ~

DER AUTOR


SCIENCE COMMONS / CC-BY-3.0

John Wilbanks ist Executive Director of Science Commons bei der Organisation Creative Commons. Er hat die Bioinformatikfirma Intellico gegründet, die semantische Graphennetzwerke für die pharmazeutische Forschung entwickelt, und gehört dem Beirat der U.S. National Library of Medicine's PubMed Central an.

QUELLEN

Bell, G. et al.: Beyond the Data Deluge. In: Science 323, S. 1297–1298, 2009, doi: 10.1126/science.1170411

Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago 1996

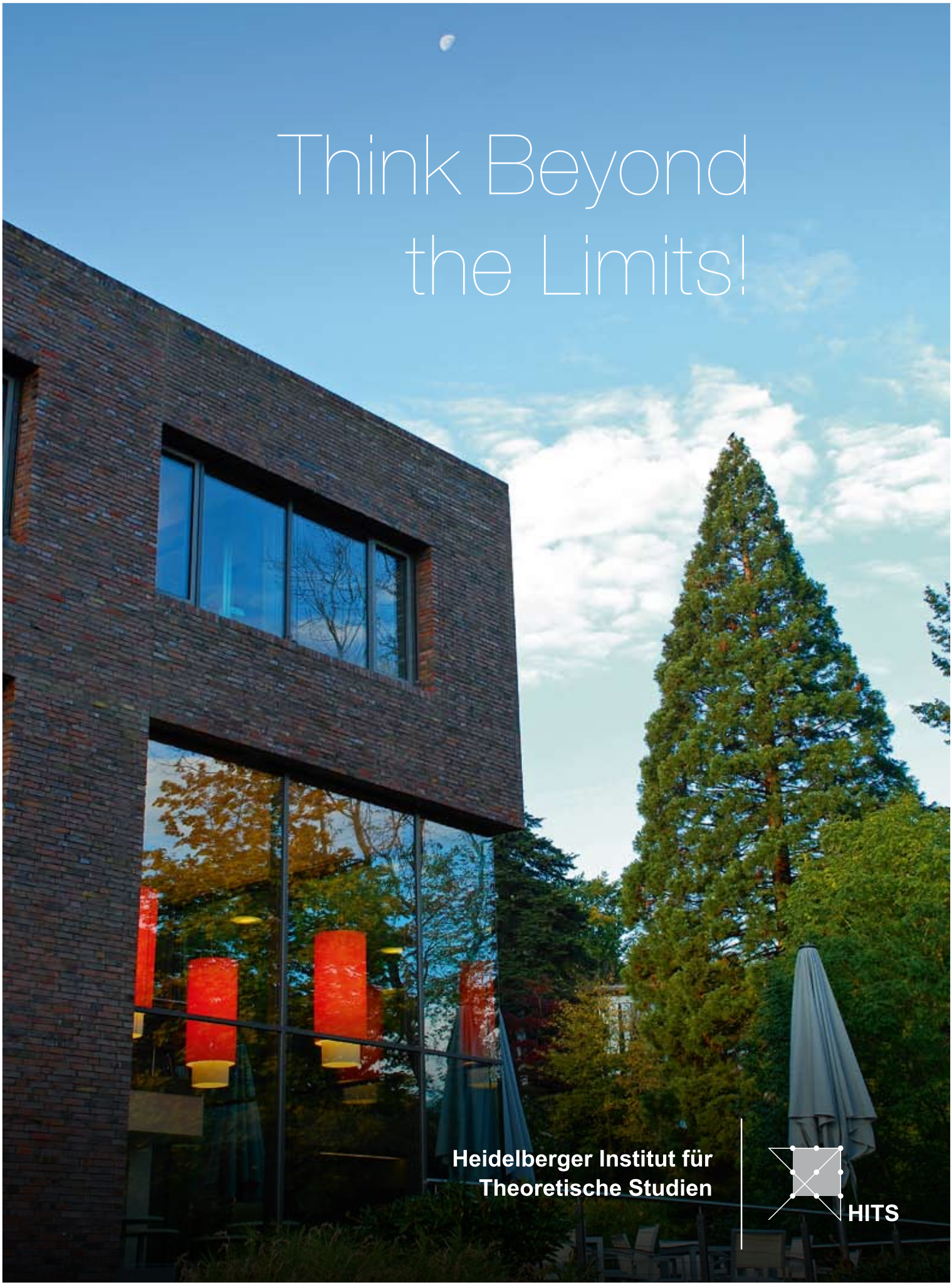
Science Commons Protocol on Open Access Data: http://sciencecommons.org/projects/publishing/open-access-data-protocol

Gekürzte Übersetzung des Kapitels »I Have Seen the Paradigm Shift, and It Is Us« aus »The Fourth Paradigm – Data-Intensive Scientific Discovery«. Herausgegeben von Tony Hey, Stuart Tensley und Kristin Tolle. Microsoft 2009

Das **Heidelberger Institut für Theoretische Studien (HITS)** ist das Forschungsinstitut der gemeinnützigen Klaus Tschira Stiftung. Der methodische Schwerpunkt liegt auf der Theorie- und Modellbildung. Dabei spielen rechnergestützte Simulationen und Datenerschließung eine zentrale Rolle. Derzeit arbeiten rund achtzig Forscher aus fünfzehn Ländern in den sechs Arbeitsgruppen, darunter zahlreiche Doktoranden und junge Gastwissenschaftler.

Heidelberger Institut für Theoretische Studien

HITS


Think Beyond
the Limits!
Heidelberger Institut für
Theoretische Studien
HITS