de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

VON DER DATENANALYSE ZUM VERSTEHEN KOMPLEXER BIOLOGISCHER SYSTEME

Highlights aus dem Deutschen Netzwerk für Bioinformatik-Infrastruktur

www.denbi.de

LIEBE LESERINNEN UND LESER

Prof. Dr. Andreas Tauch (links) , Prof. Dr. Alfred Pühler (rechts)

Die Erzeugung von Big Data gehört heutzutage zu den Charakteristika der Lebenswissenschaften. Da die Analyse großer Datenmengen zunehmend ein vertieftes Verständnis der Bioinformatik erfordert, wurde vor fünf Jahren das Deutsche Netzwerk für Bioinformatik-Infrastruktur (de.NBI) ins Leben gerufen, das durch Angebote von Service, Training und Rechenkapazitäten zur Analyse solcher Datenmengen beiträgt.

Das vom BMBF geförderte de.NBI-Netzwerk setzt sich aus einer großen Anzahl von Einzelprojekten zusammen, die thematisch in acht Servicezentren aufgeteilt sind. Im März 2020 feiert das Netzwerk nun sein fünfjähriges Bestehen. Aus diesem Anlass wurde die vorliegende Jubiläumsbroschüre auf den Weg gebracht, die über die Themen des Netzwerks informiert. Insbesondere wurde Wert darauf gelegt, anwendungsorientierte Aspekte aus den Bereichen Pflanze, Mikrobe und Medizin in den Vordergrund zu stellen. Lassen Sie sich von der Vielfalt unserer Themen überraschen.

Neben der Vorstellung des Netzwerks, wurde ein Interview mit dem de.NBI-Koordinator und dem de.NBI-Geschäftsstellenleiter aufgenommen. Dieses Interview beschäftigt sich sowohl mit Struktur und Organisation des Netzwerks als auch mit den vielen in der Zwischenzeit angestoßenen Aktivitäten. Schließlich wird noch über die verschiedenen Aufgabenfelder des de.NBI-Netzwerks berichtet. Neben Service- und Trainingsaspekten erfolgt eine Darstellung der de.NBI-Cloud und des Industrieforums.

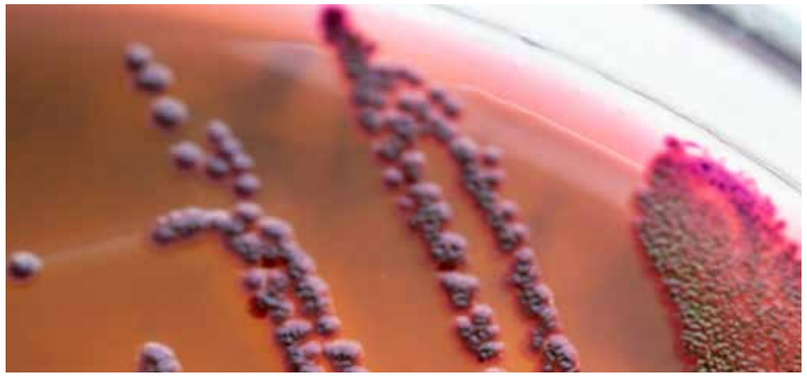
Die Broschüre soll dazu beitragen, die im Netzwerk angesiedelten Themen einem größeren Publikum näherzubringen.

Wir wünschen nun allen Leserinnen und Lesern eine interessante Lektüre.

Alfred Pühler
de.NBI-Koordinator

Andreas Tauch
de.NBI-Geschäftsstellenleiter

INHALT



VON DER DATENANALYSE ZUM VERSTEHEN KOMPLEXER BIOLOGISCHER SYSTEME
INHALT

PFLANZEN-BIOINFORMATIK – MODERNE PFLANZENFORSCHUNG UND PFLANZENZÜCHTUNG VORAN BRINGEN

Die Erzeugung großer Datenmengen hat in der Pflanzenforschung und Pflanzenzüchtung Einzug gehalten. Daten alleine bedingen aber noch keinen wissenschaftlichen Fortschritt. Durch die bioinformatische Analyse von Sequenz- sowie von Transkriptom-, Proteom- oder Metabolomdaten können jedoch detaillierte Informationen über wichtige genetische und physiolo gische Vorgänge in Kulturpflanzen erfasst und so deren züchterische Potenziale besser genutzt werden. Die Zukunft der Pflanzenforschung und Pflanzenzüchtung ist ohne Bioinformatik deshalb nicht mehr denkbar.

GRÜNE BIOINFORMATIK – DIE ENTSCHLÜSSELUNG DER WURZELN DER ZIVILISATION

Pflanzen sind unsere Begleiter: als Gewürze, Dekoration oder natürlich als Grundlage unserer Ernährung und sogar als Ursprung unserer Zivilisation. Heutige Sorten sind Resultat jahrtausendelanger Züchtung. Dieser Prozess hält an und neue Hochdurchsatzmethoden liefern Daten zur steten Verbesserung unserer Sorten. de.NBI trägt dazu bei, dies für die Forschung nutzbar zu machen und somit zur nachhaltigen Lebensmittelproduktion und -versorgung beizutragen.

GRÜNE BIOINFORMATIK – DIE ENTSCHLÜSSELUNG DER WURZELN DER ZIVILISATION.
PFLANZENBIOINFORMATIK

PFLANZEN- UND TIERZUCHT SIND GRUNDLAGE UNSERER ZIVILISATION

Vor etwa 20.000 Jahren begann im sogenannten fruchtbaren Halbmond zwischen östlichem Mittelmeer und dem Zweistromland (dem heutigen Irak) der Übergang zum sesshaften, bäuerlichen Leben. Eine Triebkraft dieser Umstellung auf den Anbau und die Züchtung von für den Menschen vorteilhaften Nutzpflanzen war eine durch klimatische Veränderungen angetriebene Entwicklung. Die damals beginnende Warmzeit erforderte einen Ausgleich zu schwindenden Nahrungsangeboten an Wildtieren. Im Zuge dieser neolithischen Revolution wurden Nutzpflanzen und Tiere erstmals domestiziert. Angebaut wurden zunächst Getreide (Abbildung 1) und Hülsenfrüchte. Die ersten domestizierten Tiere waren Ziegen, Schafe und Rinder. Dieses wird als Initialzündung und wesentlicher Wegbereiter unserer heutigen Zivilisation und Kultur in Form erster Hochkulturen in Mesopotamien und Ägypten angesehen. Denn nur durch die planbare und verlässliche Verfügbarkeit von Nahrung wurde die Grundlage für Kultur und Stabilität einer rapide wachsenden Bevölkerung gelegt. Die Züchtung und Auswahl für den Menschen vorteilhafter Pflanzen und Tiere prägt auch noch unsere heutige Kultur. Unsere Landschaften und Anbaugebiete sind geprägt von Organismen (Pflanzen), die nicht durch Evolution, sondern durch gezielte Kultivierung, Selektion und klassische Züchtung durch den Menschen entstanden sind. Alte Triebkräfte bleiben auch heute aktueller denn je. Herausforderungen wie ein rapider Klimawandel, eine dramatisch wachsende Erdbevölkerung und versalzene oder anderweitig minderwertige Böden stellen uns heute vor ähnlich drängende Herausforderungen wie vor 20.000 Jahren.

ENTSCHLÜSSELUNG VON GENOMEN HILFT, AKTUELLE HERAUSFORDERUNGEN DER PFLANZENZÜCHTUNG ZU BEWÄLTIGEN

Zum Glück sind in den Pflanzen durch Millionen von Jahren Evolution bereits Antworten zu sich ständig ändernden Umweltbedingungen in ihrem Bauplan, dem Genom, zusammengetragen. Neben

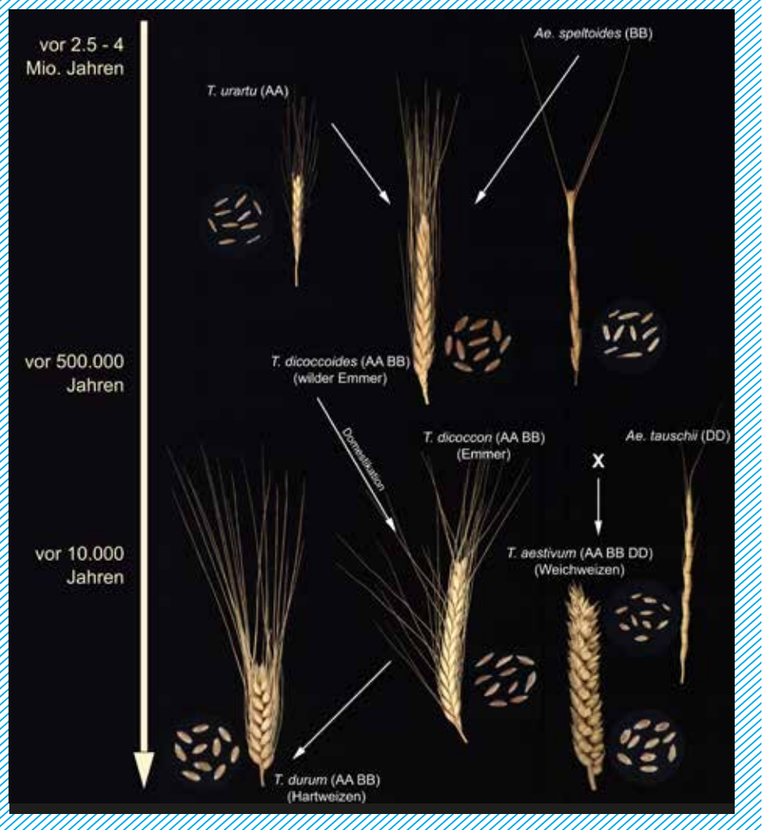
universellen Bauteilen hat jede Spezies oder auch Unterart ihre eigenen, teils speziellen Bauteile bzw. Gene hervorgebracht, die in ihrer Gesamtheit der sogenannte Genotyp sind. Durch das Verständnis dieser in Form von DNA-Molekülen in jeder Zelle vorliegenden Gene können wir uns wie auch schon vor 20.000 Jahren das genetische Lösungsrepertoire der Natur zunutze machen und versuchen, günstige Eigenschaften in kultivierte Sorten einzubringen. Dies kann entweder durch klassische Kreuzung und Auswahl auf dem Feld oder durch gezielte molekulare Analyse von Genomen über Pflanzengenbanken erfolgen. Jedoch sind die Genome vieler Nutzpflanzen – zum Beispiel Mais und Getreide wie Weizen und Gerste – in Größe und Aufbau erschreckend komplex und übertreffen in Teilen die Komplexität des menschlichen Genoms bei Weitem (Abbildung 2). Die heutigen Möglichkeiten moderner biologischer und genomischer Forschung sind jedoch eine weitaus günstigere Ausgangslage als vor 20.000 Jahren. Lösungen zur gezielten Ermittlung aller Gene, Genvarianten, Genomstrukturen und weiterer molekularer Eigenschaften konnten erst vor relativ Kurzem erarbeitet werden und bisher können nur spezialisierte Labors oder gar ganze Konsortien diese Art der Analytik durchführen. Die umfangreichen neuen und tiefergehenden Daten erlauben nun jedoch, völlig neue Fragen zu stellen und mögliche neue Zusammenhänge zu untersuchen. Daneben ist auch in der (pflanzen-)biologischen genomischen Forschung eine breite Demokratisierung und, damit verknüpft, eine massive Digitalisierung ehemals stark klassisch experimentell geprägter Bereiche zu beobachten. Um diesen Prozess zu begleiten und zu unterstützen, werden spezialisierte analytische Softwareprogramme und Vorhersagemodelle von den erfahrenen Expertengruppen im Pflanzenservicezentrum GCBN (German Crop Bio-Greenformatics Network) schrittweise auch in einer dafür eigens installierten und maßgeschneiderten analytischen Cloud zur Verfügung gestellt. Dies soll die Breitenanwendung von ehemals auf Spezialistengruppen beschränkten analytischen Prozessen unterstützen und letztlich eine breite Emanzipation *in silico* basierter pflanzengenomischer Forschung schaffen.

> Teilweise übertreffen Genome vieler Nutzpflanzen in Größe und Aufbau die Komplexität des menschlichen Genoms bei Weitem.

Breites Interesse und die Anwendung nicht nur in der Forschung, sondern auch in der angewandten Züchtungsforschung haben sich bereits über die letzten zwei Jahrzehnte entwickelt. Dabei sind die Verzahnung und der Austausch von ehemals als Grundlagenforschung betrachteten Bereichen mit anwendungsorientierter Forschung sowie unternehmensgebundener Entwicklung sehr eng geworden. Ein Beispiel dafür ist, dass durch die Züchtung auch unerwünschte Eigenschaften selektiert wurden, nämlich wenn es um die Anreicherung von Schadstoffen geht: Enthält der Boden Cadmium, reichert sich das Schwermetall im modernen Hartweizen an, aber nicht im originalen wilden Emmer. Das verantwortliche Gen hat im Hartweizen seine Funktion verloren. Die Züchtung fokussiert nun darauf, das funktionsfähige Gen wieder einzukreuzen [1]. Ähnliche Aspekte werden in Bezug auf die verbreiteten Weizensorten und Glutensensitivität untersucht und werden auch dort Eingang in die Züchtung finden [2, 3].

Als wichtigster nächster Schritt wird das Verstehen der Wechselwirkung von Genotyp, also der genetischen Information, und den Pflanzeneigenschaften, dem Phänotyp, sowie der Interaktionen mit der Umwelt angesehen. Drängende Umwelt- und Klimaproblematiken, Klimaänderungen, Hungersnöte, Unruhen und Migrationsströme sind eng mit dieser nur vordergründig rein wissenschaftlichen Problemstellung verknüpft. Lösungen für diese Aufgaben sind jedoch die Grundlage, um zumindest einige dieser großen Herausforderungen zu meistern. Zur Bearbeitung sind beispielsweise der standardisierte Zugriff und die strukturierte Bereitstellung einer breiten Palette sogenannter Omics-Daten mit modernen Verfahren der Computertechnik notwendig. Da dies nicht in einzelnen Laboren zu bewerkstelligen ist, ist unser Ziel, die geballte bioinformatische Expertise und die verfügbaren enormen Nutzpflanzen-basierten Datenmengen einer breiteren Anwendergemeinschaft – vom pflanzlichen Molekularbiologen bis hin zum Züchter – strukturiert und leicht zugänglich verfügbar zu machen und entsprechende Software zur laborübergreifenden Analyse und Anwendung anzubieten [4].

Neben wiederverwendbarer Software liegt ein Hauptaugenmerk auch auf dem ökonomischen Umgang mit den erzeugten Daten. Alle Daten sollen auffindbar (findable), zugreifbar (accessible), mit anderen Daten zusammenfügbar (interoperable) und wiederverwendbar (reusable) gespeichert und verwaltet werden (Abbildung 3). Diese Eigenschaften werden unter dem Begriff FAIR zusammengefasst und bilden ein wesentliches Ziel der Arbeiten im Pflanzenservicezentrum GCBN.



ABBILDNG 1: Entwicklungsgeschichte unseres heutigen Weizens. Vor etwa 500.000 Jahren bildete sich der wilde Emmer (*Triticum dicoccoides*) durch eine Verschmelzung von zwei diploiden Wildgräsern, wilder Einkorn, *T. uartu* (AA) und einem Ziegengras, *Ae. speltoides* (BB) zu einem tetraploiden AABB Genom. Mit der Sesshaftwerdung des Menschen begann vor ca. 10.000 Jahren ein Selektionsprozess, in dessen Verlauf zunächst der kultivierte Emmer (*Triticum dicoccon*) und daraus wiederum der Pastaweizen (= Hartweizen, *Triticum durum*) entwickelt wurde. Der hexaploide Brotweizen (= Weichweizen, *Triticum aestivum*, AABBDD) entstand etwa zeitgleich durch die Verschmelzung von tetraploidem Emmer mit einem weiteren, eher unscheinbaren Ziegengras (*Aegilops tauschii*) und wurde als beliebte Nahrungsquelle ebenfalls weitergezüchtet. (Bild: Gudrun Schütze, IPK Gatersleben)

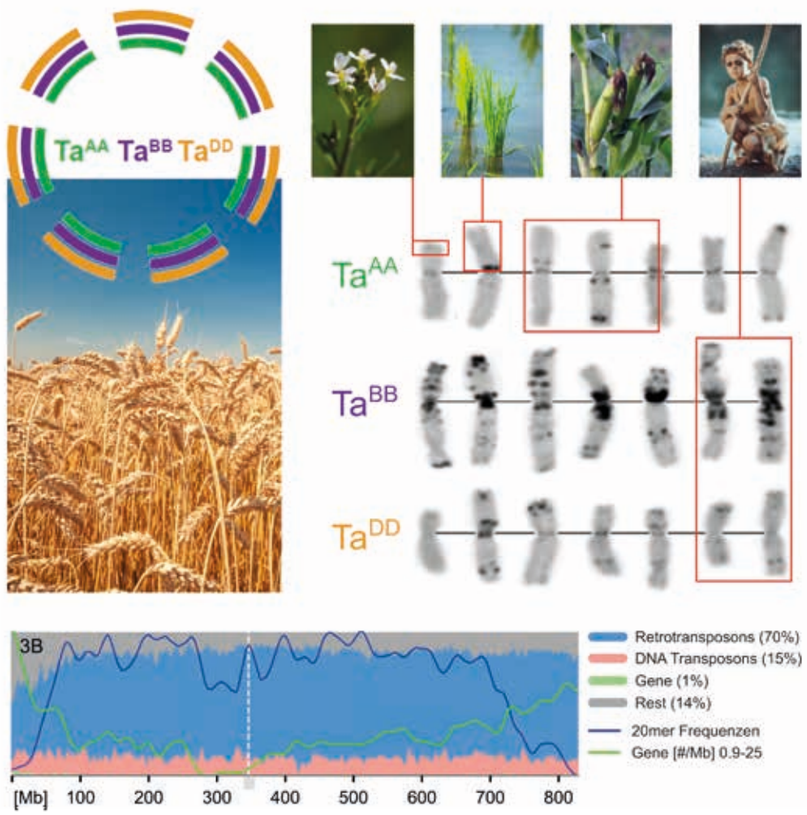


ABBILDUNG 2: Die komplexe Struktur des Brotweizengenoms. Mit einer Größe von 16 Gbp ist das Weizengenom fünfmal größer als das menschliche Genom. Brotweizen ist hexaploid und besteht aus drei sehr ähnlichen Subgenomen, genannt A,B und D, mit je sieben Chromosomen. Die roten Kästchen markieren die Genomgrößen von Acker-Schmalwand (0.13 Gbp) – dem ersten seit dem Jahr 2000 verfügbaren Pflanzengenom –, Reis (0.4 Gbp), Mais (2.3 Gbp) und Mensch (2.2 Gbp) in Relation zum Weizengenom. Der untere Teil zeigt die Architektur eines typischen Getreidechromosoms am Beispiel von Weizen (3B) als gestapeltes Balkendiagramm (0-100 %). Die Genomlandschaft wird von Transposons, überwiegend LTR-Retrotransposons dominiert, die durch ihre hohe Repetitivität (blaue Linie) die Assemblierung solcher Genome massiv erschweren. Die Gene als Hauptakteure der Merkmale sind wie Nadeln im Heuhaufen, sie stellen nur 1 % der gesamten DNA-Sequenz dar und sind an den Enden der Chromosomen stark angereichtert (grüne Linie). (Bilder von rechts nach links: Bild Weizen © vovan/AdobeStock) Bild Blume © lehic/AdobeStock, Bild Reis © comzeal/AdobeStock, Bild Mais © orestligetka/AdobeStock, Bild Kind EmotionPhoto/AdobeStock)

ABBILDUNG 3: FAIRe Daten für Forschung und Pflanzenzüchtung. Die linke Abbildung ist eine typische Übersichtsdarstellung zu den wesentlichen Eigenschaften eines (Nutz-)Pflanzengenoms aus der Sicht des Bioinformatikers am Beispiel des tetraploiden Pastaweizengenoms. Die von den einzelnen Genomprojekten erzeugten Daten werden zurzeit im Rahmen von Pilotprojekten mit Phänotypdaten verbunden, um die biochemischen Grundlagen von züchterisch relevanten Merkmalen besser zu verstehen. Damit die erarbeiteten wertvollen Datenressourcen auch weiterhin und in anderen Kontexten genutzt werden können, werden die Daten nach dem FAIR-Prinzip strukturiert, verschlagwortet und archiviert. Die Karte zeigt den weltweiten Datenzugriff auf das e!DAL Archiv System vom IPK (Plant Genomics & Phenomics Research Data Repository, http://edal-pgp.ipk-gatersleben.de). Bild links oben: © ktsdesign/AdobeStock, Bild links unten: © sdecoret/AdobeStock, Bild rechts oben: dppn.plant-phenotyping-network.de)

REFERENZEN [1] Nat Genet 2019;51(5):885-895. DOI: 10.1038/s41588-019-0381-3. **[2]** Science 2018;361(6403). DOI: 10.1126/science.aar7191. **[3]** Sci Adv 2018;4(8):eaar8602. DOI: 10.1126/sciadv.aar8602. **[4]** Genome Biology 2020. DOI: 10.1186/s13059-019-1899-5.

AUTOREN Heidrun Gundlach[1], Matthias Lange[2], Marie Bolger[3], Björn Usadel[3], Uwe Scholz[2], Klaus F. X. Mayer[1]
[1] *Plant Genome and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg,*
[2] *Bioinformatik und Informationstechnologie, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) Gatersleben, Corrensstraße 3, 06466 Seeland*
[3] *BG-2 Pflanzenwissenschaften, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52428 Jülich*

DIE CHEMISCHE DIVERSITÄT IN DER PFLANZENWELT

Die Rolle der Artenvielfalt auf unserem Planeten war lange Zeit nur wenig beachtet. Das hat sich in den letzten Jahren sowohl in der Wissenschaft als auch in der öffentlichen Wahrnehmung geändert. Neben der Biodiversität zählt auch die Untersuchung der Vielfalt an Inhaltsstoffen, die Chemodiversität, zu diesem Forschungsgebiet.

DIE CHEMISCHE DIVERSITÄT IN DER PFLANZENWELT
PFLANZENBIOINFORMATIK

Die Analyse der natürlichen Inhaltsstoffe von Pflanzen dient vielerlei Zwecken. So konnten zum Beispiel bereits zahlreiche chemische Substanzen aus Pflanzen als Heilmittel beim Menschen Verwendung finden. Sekundäre Stoffwechselprodukte steuern zudem eine Vielzahl von Interaktionsprozessen sowohl innerhalb der Pflanze als auch zwischen verschiedenen Pflanzen und den Mikroorganismen in ihrer Umgebung. Chemische Substanzen geben daher Aufschluss über eine Vielzahl an wichtigen biologischen Prozessen. Doch über viele dieser Naturstoffe ist bislang nichts bekannt – weder über ihre chemische Struktur noch ihre biologische oder ökologische Funktion. Daher beschäftigt sich das Forschungsgebiet der chemischen Ökologie mit solchen Fragestellungen und darüber hinaus mit der Bedeutung von chemischer Diversität.

Die technische Analyse der natürlichen Inhaltsstoffe von Pflanzen erfolgt oftmals mit einem Massenspektrometer. Dazu werden Proben der Pflanzen gesammelt, die Inhaltsstoffe im Labor beispielsweise in Wasser und Methanol extrahiert und anschließend mit einer Kombination aus Chromatografie und Massenspektrometrie analysiert (Abbildung 1).

Dabei fallen eine Unmenge an komplexen Rohdaten an, die Aufschluss über das Masse-zu-Ladung-Verhältnis und die chromatografische Retentionszeit der Substanzen geben. Diese Rohdaten lassen sich als Fingerprint der Pflanzen interpretieren und erlauben bereits, die Proben mit statistischen Methoden zu untersuchen, um biologische und ökologische Fragestellungen zu bearbeiten.

Die Abbildungen in diesem Beitrag zeigen einige Beispiele für Untersuchungen aus dem Bereich Eco-Metabolomics, an denen das Center for Integrative Bioinformatics (CIBI) beteiligt ist.

OHNE MOOS NICHTS LOS

Moose sind die ältesten Landpflanzen der Erde und kommen in fast allen Ökosystemen vor. Sie gelten als außerordentlich gute Bioindikatoren, die Veränderungen in der Umwelt anzeigen, wie etwa Verunreinigungen oder Schadstoffe in der Luft, die zu Schadbildungen oder beeinträchtigtem Wachstum der Moose führen können. Bislang wurden solche Veränderungen hauptsächlich im Wachstum und den morphologischen Eigenschaften betrachtet, nicht jedoch auf Ebene der biochemischen Zusammensetzung. Am Leibniz-Institut für Pflanzenbiochemie (IPB) wurden daher die biochemischen Veränderungen in verschiedenen Moosarten über die Jahreszeiten hinweg und im Hinblick auf unterschiedliche Lebensbedingungen und ihre Verwandtschaft zueinander (Phylogenie) mit Massenspektrometrie analysiert und anschließend mit Methoden der Bioinformatik ausgewertet.

Die Studie [1] gibt Hinweise, welche Zusammenhänge zwischen verschiedenen Lebensweisen und Selektionsstrategien der Moose und ihrer biochemischen Anpassung an veränderte Lebens- und Umweltbedingungen bestehen. Dieser ungerichtete Eco-Metabolomicsansatz liefert damit wertvolle biochemische Erkenntnisse, die unser Verständnis grundlegender ökologischer Strategien verbessern und als Grundlage zukünftiger Forschung (Hypothesengenerierung) dienen können. Wir haben zudem einen repräsentativen Datensatz und einen bioinformatischen Workflow erstellt, der in zukünftigen Metabolomics-Studien wiederverwendet werden kann.

MACBESST AM IDIV – DER PFLANZLICHE FINGERABDRUCK ALS WEGWEISER

Bei MacBeSSt geht es nicht um (klassische) Literatur, sondern um das Projekt „Metabolite Changes in Biodiversity Levels and Seasonal Shifts" am Deutschen Zentrum für integrative Biodiversitätsforschung (iDiv) Halle-Jena-Leipzig, das sich ebenfalls mit der (chemischen) Diversität in der Pflanzenwelt beschäftigt.

Im Gegensatz zu medizinisch relevanten Pflanzen, wie zum Beispiel Salbei oder Johanniskraut, ist über die sekundären Inhaltsstoffe (Metabolite) von Grasslandarten bisher wenig bekannt. Um den Metabolit-Fingerabdruck dieser Arten zu erforschen, untersuchen wir Pflanzen, die im Jena-Experiment [2] zusammen mit anderen Pflanzenarten gewachsen sind. Da auch veränderte Tageslängen, wärmere Temperaturen und die Wasserversorgung eine große Rolle in der Pflanzenentwicklung spielen, haben wir die

Proben von 13 Arten an vier Zeitpunkten zwischen Mai und Oktober genommen, um saisonale Unterschiede im metabolischen Fingerabdruck wiederzufinden.

Für die Analyse der Fingerabdrücke ist vor allem die Zusammensetzung dieser Artengemeinschaften von Bedeutung, da eine veränderte Nachbarschaft auch einen veränderten Fingerabdruck bedeuten könnte. Um diese Einflüsse genau unter die Lupe zu nehmen, haben wir Artengemeinschaften beprobt, die sich aus einer einzelnen Art (Monokultur) sowie aus zwei, vier oder acht verschiedenen Arten zusammensetzten. Mithilfe des Massenspektrometers, das an einen Flüssigkeitschromatografen gekoppelt ist, werden die Pflanzenextrakte gemessen. Die aufgenommenen Daten können anschließend statistisch ausgewertet und auf Zusammenhänge untersucht werden.

Die untersuchten äußeren Einflüsse, Artengemeinschaft und Saison, finden sich in Form veränderter Mengen der pflanzlichen Inhaltsstoffe wieder – weisen also den Weg, den die Pflanze bisher genommen hat. Hierbei verändert sich die Dimension des Fingerabdrucks aber nicht, weshalb es möglich ist, alle untersuchten Arten anhand ihres einzigartigen Musters über das ganze Jahr hinweg zu identifizieren. Durch das experimentelle Design kann das Projekt die Beziehungen zwischen Pflanzenarten, Artengemeinschaften, Jahreszeiten und der Umwelt untersuchen und damit eine Brücke zwischen den Forschungsgebieten Ökologie, Biochemie und Bioinformatik schlagen.

METABOLITENIDENTIFIKATION

Die Aufgaben der Bioinformatiker enden allerdings nicht bei der Analyse der Fingerprints, denn für eine biologische (oder ökologische) Interpretation wird die Annotation der chemischen Struktur benötigt. Dazu gibt es zwei Ansätze, für die im de.NBI-Netzwerk entsprechende Services angeboten werden.

Zum Beispiel können die Spektren aus dem Massenspektrometer mit den Einträgen einer Referenzdatenbank von bekannten Substanzen verglichen werden. Die MassBank [3] enthält mehr als 50.000 Einträge zu mehr als 13.000 Substanzen. Das CIBI entwickelt die Software und hilft, neue Daten aus der Nutzergemeinschaft zu integrieren. Aber nicht immer sind Referenzdaten verfügbar, denn oft sind die Reinsubstanzen nicht erhältlich. In solchen Fällen helfen In-silico-Vorhersagen mit Methoden der Bioinformatik (Computational Metabolomics).

Das am Leibniz-Institut für Pflanzenbiochemie entwickelte MetFrag [4] lässt sich sowohl online als auch in der de.NBI-Cloud nutzen. Im Rahmen unserer Moos-Studie (siehe oben) analysieren wir auch Substanzklassen und haben MassBank mit bislang unbekannten Spektren von Moosen erweitert.

Insbesondere in der Metabolomik steigt mit der großen Anzahl von Proben und Merkmalen in den experimentellen Ergebnissen der Bedarf an automatisierter Datenverarbeitung. Workflow- oder Pipeline-Tools sind hierbei visuelle Programmiersprachen, die es Biologen und Forschenden in der Biomedizin ermöglichen, modernste Algorithmen und Datenanalysen auf große Datensätze anzuwenden. Sie sind bereits im kommerziellen Data-Mining, aber auch in wissenschaftlichen Bereichen wie der Pharmaforschung oder der Genomik weit verbreitet. Zeitaufwendige Aufgaben können in leistungsfähige Cloud-Infrastrukturen ausgelagert werden. Mit der Einrichtung der de.NBI-Cloud wird es also einfacher, Metabolomics-Workflows zu entwickeln und zu betreiben. Die Cloud macht's!

WISSEN IST DAS EINZIGE GUT, DAS SICH VERMEHRT, WENN MAN ES TEILT

Zur biologischen bzw. ökologischen Forschung gehört auch, die Daten für die Nachwelt zur Verfügung zu stellen. Dafür ist das Metabolomics-Datenrepositorium MetaboLights am EMBL-EBI prädestiniert. Das de.NBI-Netzwerk und das Servicezentrum CIBI unterstützen vor allem die deutsche Nutzergemeinde, hochwertige Metabolomicsdaten FAIR zu veröffentlichen. Das bedeutet: Sie sind findable (also auffindbar) durch aussagekräftige Metadaten und entsprechende Suchmaschinen; es ist geregelt, wie sie accessible (also zugänglich) sind; sie sind interoperable, können also mit weiteren Daten kombiniert werden, und sind reusable (also wiederverwendbar), zum Beispiel in späteren Forschungsprojekten.

Die Daten zu den oben beschriebenen Beispielen finden sich als Studien MTBLS520, MTBLS709 und MTBLS679 in der Forschungsdatenbank MetaboLights.

Um diese Themen auch den kommenden Generationen von Forschenden und der interessierten Öffentlichkeit nahezubringen, gibt es unterschiedliche Bildungs- und Trainingsangebote. Um möglichst früh zu beginnen, lernen interessierte Schülerinnen und Schüler der Oberstufe im Rahmen der Sommerschule BioByte an der Martin-Luther Universität Halle-Wittenberg, wie man Naturstoffe extrahiert und die resultierenden Daten anschließend auswertet. Die tiefer gehenden de.NBI-Trainingsangebote richten sich an Wissenschaftler und Wissenschaftlerinnen aus verschiedenen Fachrichtungen, vom Master bis zum PostDoc. Hier gibt es sowohl kurze Workshops als auch längere Angebote wie die einwöchige Metabolomics Winterschool.

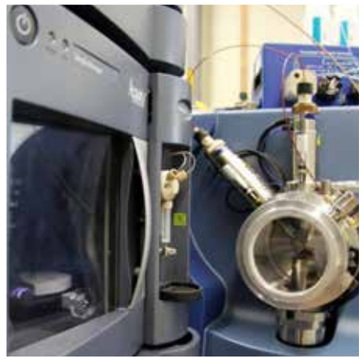


ABBILDUNG 1: Modernes Massenspektrometer im Labor aus [5].

ABBILDUNG 2: Verschiedene Moose im botanischen Garten der Martin-Luther Universität Halle-Wittenberg aus [6].

FAZIT

Viele der hier beschriebenen Herausforderungen für die (Eco-) Metabolomik gelten auch für andere, auf den ersten Blick nicht naheliegende Disziplinen. Eine Aufgabe zum Beispiel in der Umweltforschung ist die Überwachung der Wasserqualität, die den Vergleich von Proben über Standorte, Zeit oder nach erfolgter Wasseraufbereitung erfordert. Auch bei der Kontrolle von Lebensmitteln werden Proben auf ihre biochemische Zusammensetzung hin untersucht und können von Bioinformatik profitieren.

Das de.NBI-Netzwerk deckt in mehreren seiner Servicezentren verschiedene Aspekte der Metabolomik ab, darunter das Center for Integrative Bioinformatics (CIBI). Mit dem Aufkommen der de.NBI-Cloud kann die Datenverwaltung und -verarbeitung auch großer Studien mit vielen Proben übernommen werden. Bioinformatiker sind damit ein integraler Teil in interdisziplinären Teams mit Molekularbiologen, Biochemikern und Ökologen und helfen bei der Aufklärung und Konservierung der Diversität in der Pflanzenwelt unseres Planeten.

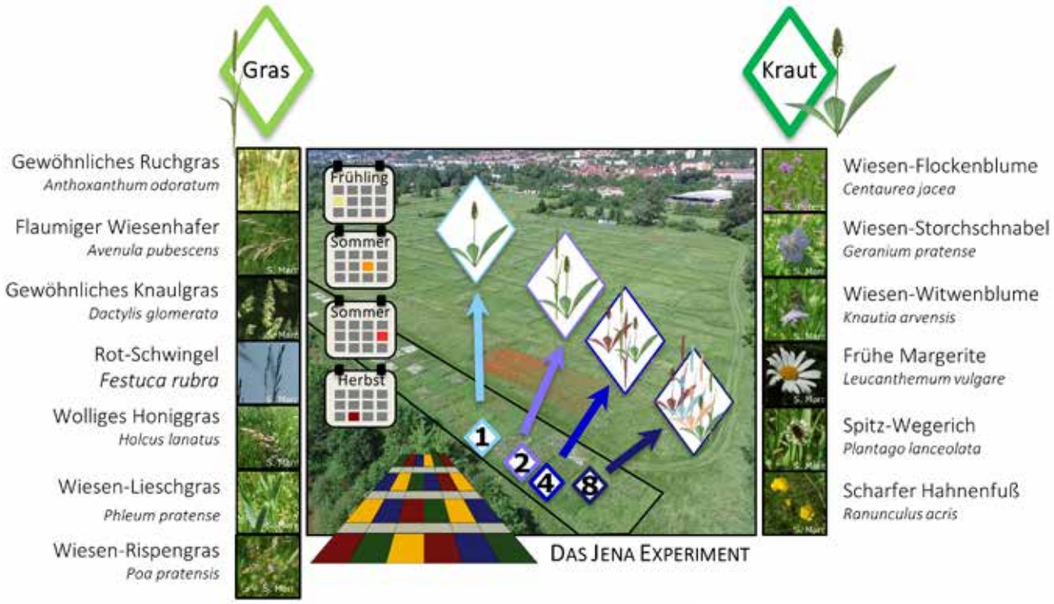


ABBILDUNG 3: Metabolite Changes in Biodiversity Levels and Seasonal Shifts (MacBeSSt) im Jena-Experiment.

THE JENA EXPERIMENT

REFERENZEN [1] Metabolites 2019, 9(10), 222. DOI:org/10.3390/metabo9100222 **[2]**] http://www.the-jena-experiment.de/Video.html **[3]** https://massbank.eu/ **[4]** https://msbi.ipb-halle.de/Metfrag **[5]** https://www.ipb-halle.de/forschung/technologie-plattformen/metabolomics/ **[6]** Präsentation K. Peters zu https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4361

AUTOREN Kristian Peters[1], Susanne Marr[1,2,3] und Steffen Neumann[1,3]
[1] *Leibniz-Institut für Pflanzenbiochemie (IPB), Weinberg 3, 06120 Halle (Saale)*
[2] *Martin-Luther-Universität Halle-Wittenberg, Universitätsplatz 10, 06108 Halle (Saale)*
[3] *Deutsches Zentrum für integrative Biodiversitätsforschung (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig*

MIKROBIELLE BIOINFORMATIK – ANALYSE DER VIELFALT AUF UNSEREM PLANETEN

Das Leben auf unserem Planeten wird in nahezu allen Bereichen durch mikroskopisch kleine Lebewesen, den Mikroorganismen, mitbestimmt. Die Erforschung ihrer Lebensvorgänge gelingt heutzutage in faszinierender Detailtiefe, wobei Omics-Daten und deren bioinformatische Analyse eine Schlüsselstellung einnehmen.

MIKROORGANISMEN – DIE UNSICHTBARE MEHRHEIT IN UNSEREN OZEANEN

Mensch und Meer leben schon immer in einer engen Beziehung. Ozeane bedecken ca. 70 Prozent der Erdoberfläche und rund die Hälfte der Weltbevölkerung lebt in den Küstengebieten. Durch die Fischerei bietet das Meer Nahrung für Millionen von Menschen und ist seit Tausenden von Jahren einer der wichtigsten Handelswege. In den letzten Jahrzehnten prägt der Tourismus als bedeutender Wirtschaftsfaktor zunehmend die Küstenregionen. Die Ozeane sind ebenfalls die Heimat von Millionen von Tier- und Pflanzenarten und Milliarden von Mikroorganismen. Als die unsichtbare Mehrheit bilden diese die Grundlage des Nahrungsnetzes im Meer und sind für das Recycling praktisch aller Nährstoffe global verantwortlich. Sie zu erforschen, gelingt nur durch das geschickte Zusammenspiel molekularer Techniken und deren bioinformatischer Analyse basierend auf Biodiversitäts-, Funktions- und Umweltdatenbanken.

DIE BEDEUTUNG DER MARINEN MIKROORGANISMEN

Marine Mikroorganismen sind mikroskopisch kleine, einzellige Organismen, zu denen Bakterien, Viren, kleine Algen und Archaeen zählen. Sie sind zwar winzig, existieren aber in sehr großer Anzahl überall in den Ozeanen, von den tiefsten Stellen am und im Meeresboden bis hin zur sonnendurchfluteten Wasseroberfläche. Ein Milliliter Meerwasser, das heißt ein Tausendstelliter, enthält bis zu eine Million Mikroorganismen (Abbildung 1). Es gibt in einem Liter Meerwasser somit mehr Mikroorganismen als Menschen auf der gesamten Erde. Verantwortlich für den globalen Nähr- und Energiestoffwechsel, sind sie unentbehrlich für die Funktionsweise der Ozeane [1].

Marine Mikroorganismen wirken sich auf unser tägliches Leben sowie unser Wohlbefinden aus. Das gilt unabhängig davon, ob man an der Küste oder im Inland lebt. Neben dem Ab- und Umbau von Nährstoffen ist die Photosynthese eine wichtige Aufgabe. Wie die Pflanzen können auch einige Meeresmikroben, etwa die Cyanobakterien, mithilfe der Lichtenergie der Sonne Kohlendioxid (CO_2) und Wasser in Zucker umwandeln. Während dieses Vorgangs wird Sauerstoff (O_2) produziert und an die Umwelt abgegeben. Wissenschaftler gehen davon aus, dass rund die Hälfte der weltweiten Sauerstoffproduktion aus den Ozeanen stammt, während die andere Hälfte aus anderen Lebensräumen wie Wäldern oder Böden kommt. Meeresmikroben produzieren somit den Sauerstoff für jeden zweiten unserer Atemzüge.

MIKROORGANISMEN – DIE UNSICHTBARE MEHRHEIT IN UNSEREN OZEANEN
MIKROBIELLE BIOINFORMATIK

ABBILDUNG 1: Das Bild zeigt Mikroorganismen auf einer Alge. Die Mikroben wurden mit Hilfe eines fluoreszierenden Farbstoffes sichtbar gemacht. (Foto: © Max-Planck-Institut für Marine Mikrobiologie / P. Gomez-Perreira / B. Fuchs)

Ein weiteres Beispiel für ihre Bedeutung ist die Fähigkeit einiger Mikroben, Öl abzubauen. Einige Arten ernähren sich davon und können deshalb helfen, Ölteppiche nach Tankerunglücken zu beseitigen. Kürzlich hat man sogar Mikroorganismen gefunden, die bestimmte Formen von Plastik abbauen können. Leider dauert dies Jahrzehnte und ist somit kein wirksamer Schutz gegen die Vermüllung der Meere [2].

Im Bereich der medizinischen und biotechnologischen Anwendung setzt man ebenfalls große Hoffnungen auf die Möglichkeiten mariner Mikroorganismen. Antibiotika sind Stoffwechselprodukte von Bakterien oder auch Pilzen, die die Eigenschaft haben, anderen Mikroorganismen zu schaden, indem sie ihr Wachstum hemmen oder sie töten. Der häufige Einsatz von Antibiotika hat dazu geführt, dass viele Mikroorganismen nicht mehr darauf ansprechen, das heißt, resistent sind. Die Wissenschaft hofft, im Meer bislang unbekannte antibiotisch wirkende Stoffe finden zu können. Dass dies möglich ist, zeigt ein kürzlich abgeschlossenes Forschungsprojekt, in dem ein antibiotisch wirkendes Produkt entdeckt wurde, das aus einem bis dahin unbekannten Bakterium stammt. Das neue Antibiotikum wird allerdings zunächst nur bei der Fischzucht in Aquakulturen eingesetzt, mit dem Ziel, die Tiere vor krank machenden Erregern zu schützen. Die Zulassung als Medikament bedarf ausgiebiger Versuchsreihen, die meist über zehn Jahre dauern.

In der Biotechnologie bedarf es biochemischer Reaktionen, um die Umwandlung organischer Stoffe zu katalysieren. Diese Aufgabe übernehmen Enzyme. Bei Enzymen handelt es sich um Eiweißstoffe, die von lebenden Zellen gebildet werden. Sie besitzen die Fähigkeit, die Reaktionsgeschwindigkeit von biochemischen Vorgängen zu erhöhen. Zellulose, Hauptbestandteil pflanzlicher Zellwände, wird als Rohstoff für die Papierherstellung genutzt. Enzyme, die in der Lage sind, Zellulose abzubauen, werden als Zellulasen bezeichnet und diese helfen dabei, das Material geschmeidig zu machen. Solche Enzyme stammen unter anderem aus Bakterien, die in der Tiefsee oder in antarktischen Gewässern leben. Auch mit Blick auf die Waschmittelindustrie ruht die Hoffnung auf den kalten Gewässern der Ozeane. Früher war es üblich, weiße Textilien bei sehr hohen Temperaturen zu waschen, um Verunreinigungen durch Hitze zu entfernen. Hohe Temperaturen sind jedoch gleichbedeutend mit einem hohen Energiebedarf. Mit dem verstärkten Einsatz von fett- und eiweißabbauenden Enzymen in Waschmitteln wird Wäsche auch bei vergleichsweise niedrigen Temperaturen energiesparend sauber.

WIE WERDEN MIKROORGANISMEN ERFORSCHT?

Bis vor Kurzem benötigten Wissenschaftlerinnen und Wissenschaftler eine Mikrobenreinkultur, um scheinbar einfachen Fragen wie „Welche Arten von Mikroorganismen gibt es überhaupt?“, „Was können sie?“ und „Wie interagieren sie mit ihrer Umwelt?“ beantworten zu können. Reinkultur bedeutet, dass einzelne Mikroorganismen im Labor ohne ihre natürliche Umgebung und die Gesellschaft anderer Mikroorganismen heranwachsen müssen. Da sich diese Laborbedingungen sehr von den Bedingungen in den Meeren unterscheiden, ist es extrem schwierig, Meeresmikroben zu züchten. Schätzungsweise können nur ein bis zehn Prozent der marinen Mikroorganismen im Labor kultiviert werden. Glücklicherweise wurden in den letzten Jahren neue molekulare Techniken entwickelt, die eine Untersuchung von Meeresmikroben ermöglichen, ohne dass dazu eine Reinkultur im Labor nötig ist (Abbildung 2).

> Die menschliche DNA enthält 25.000 bis 35.000 Gene.

Die gesamte Information über einen Organismus existiert in seinem genetischen Code, der sogenannten DNA, die deshalb als Bauplan des Lebens bezeichnet wird. Dieser informiert die Zelle darüber, was wann zu tun ist. Die DNA kann in kleine Unterabschnitte unterteilt werden, die man als Gene bezeichnet. In der DNA eines Lebewesens gibt es Tausende Gene und jedes Gen hat eine spezifische Funktion. So enthält die menschliche DNA beispielsweise 25.000 bis 35.000 Gene, aber nur sehr wenige davon sind für einzelne Merkmale wie für die Augen- oder Haarfarbe verantwortlich. Mithilfe der sogenannten Next Generation Sequencing (NGS) – Technologie kann man mit relativ wenig technischem und finanziellem Aufwand die DNA einer kompletten Mikroorganismengemeinschaft lesen [3]. Dieser Ansatz wird auch als Metagenom-Sequenzierung bezeichnet und liefert eine Liste der Gene aller Mikroorganismen, die in einem bestimmten Gebiet vorkommen.

BIOINFORMATISCHE ANALYSE

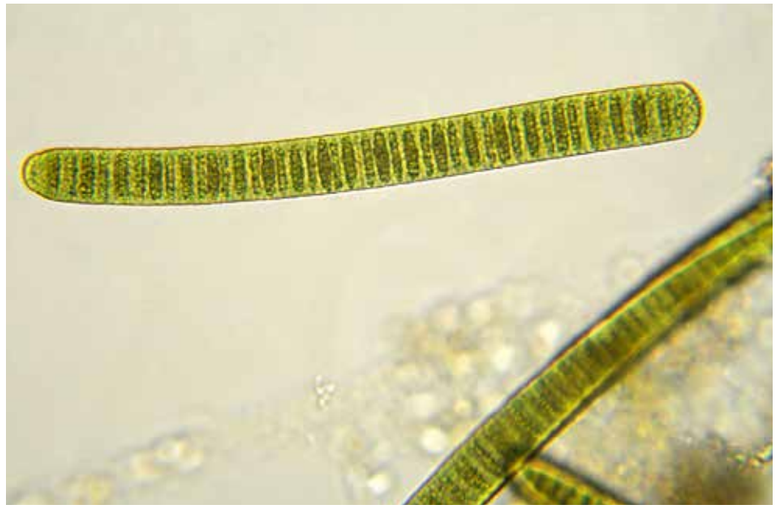
Einige Gene kommen in allen Organismen auf der Erde vor und weisen dabei geringe, aber dennoch bedeutende Unterschiede zwischen den Organismen auf. Ein Beispiel für ein solches Gen ist die ribosomale RNA (rDNA). Da dieses Gen für jede Spezies einzigartig ist, kann man es wie eine Art Fingerabdruck für einen Mikroorganismus nutzen, ähnlich wie bei der Analyse von menschlichen Fingerabdrücken. Strafverfolgungsbehörden speichern alle Fingerabdrücke in riesigen Datenbanken, um sie mit anderen Fingerabdrücken vergleichen zu können, die zum Beispiel an Tatorten erfasst wurden. So können mögliche Täter identifiziert werden. Das gleiche Prinzip wird auch in der Molekularbiologie angewandt. Man bestimmt die Gensequenz der rDNA und vergleicht diese mit der bereits vorhandenen Information, welche in Referenzdatenbanken gespeichert ist. Die im BioData-Leistungszentrum beheimatete SILVA-Datenbank [4] ist eine der beiden weltweit führenden rDNA-Referenzdatenbanken. Mit fast zehn Millionen Einträgen, die den gesamten Stammbaum des Lebens umfassen, stellt sie derzeit das umfassendste Repositorium für qualitätsgeprüfte rDNA-Sequenzen dar. Aufgrund ihrer systemrelevanten Bedeutung für die gesamte Wissenschaft wurde SILVA kürzlich zur ELIXIR Core Data Resource ernannt. So können Mikroorganismen identifiziert und Antworten auf die Frage „Welche Arten von Meeresmikroben gibt es in meiner Probe?“ gefunden werden.

Analog dazu wird die Frage „Was können sie?“ durch eine Ähnlichkeitssuche aller gefundenen Gene gegen die BRENDA-Referenzdatenbank für Enzymfunktionen beantwortet. So ist es möglich, ein Modell der zum Zeitpunkt der Probennahme vorhandenen enzymatischen Funktionen und möglicher Stoffwechselwege zu erstellen. Neben einem besseren Verständnis des Ökosystems als Ganzes sind Gensequenzen, die Anwendungen im Bereich der Medizin oder Biotechnologie finden könnten, von besonderem Interesse.

WIE INTERAGIEREN SIE MIT IHRER UMWELT?

Um die Funktionen und die Stabilität eines Ökosystems zu verstehen, reicht die Information über Diversität und Funktion der Mikroorganismen nicht aus. Dafür ist es notwendig, ihren Lebensraum selbst zu beschreiben. Dieser wird sowohl durch die Interaktion zwischen den Lebewesen als auch durch ihre Umweltbedingungen (zum Beispiel Nährstoffe, Temperatur, Salzgehalt, Wassertiefe/Druck) charak-

ABBILDUNG 2: Beispiele für Sterivex-Filter zur Konzentration von Mikroorganismen für die metagenomische Analyse. Im Rahmen des Ocean Sampling Days 2014 wurden rund 200 Proben von marinen Forschern rund um den Globus genommen. (Foto: Anna Kopf)



terisiert. Zum Teil können diese Faktoren zeitgleich mit der Probennahme der Mikroorganismen bestimmt werden. Allerdings sind für eine genaue Charakterisierung der Umwelt häufig komplexe Analysen von Wasser- und Meeresbodenproben im Labor notwendig.

Nur wenn alle diese Informationen zusammenspielen, ist man in der Lage, die komplexen Wechselbeziehungen von Organismen mit ihrer Umwelt zu verstehen, um genauere Vorhersagen zu treffen, wie sich globale Veränderungen, wie zum Beispiel die Erwärmung der Meere als Teil des Klimawandels, auswirken. Dabei sind Einzelmessungen oft unzureichende Momentaufnahmen. Die technischen Weiterentwicklungen der letzten Jahrzehnte ermöglichen es heute jedoch, eine Vielzahl von Umweltfaktoren kontinuierlich, automatisiert zu messen. Dafür werden sowohl stationäre als auch mobile Messsysteme genutzt. Ein Beispiel dafür sind die 3.800 weltweit treibenden Argo Floats. In regelmäßigen Intervallen bestimmen diese Systeme automatisch in den oberen 2.000 Metern der Ozeane die Temperatur und den Salzgehalt. Unter Verwendung von Satellitenverbindungen stehen diese Daten mit einer geringen Zeitverzögerung Wissenschaft und Öffentlichkeit zur Verfügung [5].

Die dauerhafte Bereitstellung einer Vielzahl von Daten ist essenziell für die Erforschung von globalen Entwicklungen wie dem Klimawandel und dem Artensterben. Dies kann nur durch die Speicherung in Datenarchiven sichergestellt werden. Eines der weltweit führenden Systeme dafür ist der Data Publisher for Earth und Environmental Science-PANGAEA. Das zertifizierte World – Data Center [6] wird gemeinschaftlich vom Zentrum für Marine Umweltwissenschaften MARUM an der Universität Bremen und dem Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung betrieben. Die de.NBI-Datenbank PANGAEA stellt mit über 16 Milliarden Datenpunkten eine umfangreiche Sammlung an wissenschaftlichen Daten einer großen Nutzergemeinde zur Verfügung [7]. Diese umfasst Daten aus Erde und Umwelt, aber auch Vorkommen und Verteilung sowohl von Lebewesen als auch von biochemischen Molekülen. Auf die Daten kann über die Webseite [8] zugegriffen werden, aber für Fachleute gibt es auch die Möglichkeit, die Daten über maschinelle Schnittstellen abzurufen, um sie für weitere Analysen zur Verfügung zu haben.

Die Wechselbeziehungen zwischen Mikroorganismen und den größeren Lebewesen auf unserem Planeten unterliegen einem empfindlichen Gleichgewicht und werden durch Umweltverschmutzung und sich wandelnde Klimabedingungen gefährdet. Um die Umwelt zu schützen, benötigen wir ein fundiertes Wissen über die Mikroorganismen, die im Meer leben, ihre Funktionen und wie sie untereinander und mit der Umwelt interagieren. Das Leistungszentrum BioData bietet dazu weltweit anerkannte Datenbanken für Umwelt- und Biodiversitätsforschung sowie medizinische und biotechnologische Anwendungen.

REFERENZEN [1] Nat Rev Microbiol 200;5(10):759-69. DOI: 10.1038/nrmicro1749. **[2]** Appl Microbiol Biotechnol 2018; 102:7669-7678. DOI: 10.1007/s00253-018-9195-y. **[3]** Nat Rev Genet 2016;17(6):333-51. DOI: 10.1038/nrg.2016.49. **[4]** Nucleic Acids Res 2013; 41(Database issue): D590–D596. DOI: 10.1093/nar/gks1219. **[5]** http://www.argo.ucsd.edu/ **[6]** http://www.icsu-wds.org/ **[7]** J Biotechnol 2017;261:177-186. DOI: 10.1016/j.jbiotec.2017.07.016. **[8]** https://www.pangaea.de/

AUTOREN Janine Felden[1], und Frank Oliver Glöckner[1,2]
[1] MARUM - Center for Marine Environmental Sciences University of Bremen and Alfred-Wegener-Institute, Helmholtz Center for Polar- and Marine Research, Bremerhaven
[2] Jacobs University Bremen, Bremen

ERFORSCHUNG DER TIEFSEE MIT BIOINFORMATISCHER BILDANALYSE

Die Erforschung und Überwachung der Tiefsee und der Einflüsse des Menschen stellen eine große interdisziplinäre wissenschaftliche Herausforderung dar. Zur Auswertung großer Mengen Unterwasserbilder werden neue effiziente bioinformatische Lösungsansätze gesucht. Das neue BIIGLE 2.0-System hat sich in kurzer Zeit zu einem wertvollen und international sehr anerkannten Werkzeug zur Verwaltung, Visualisierung, Annotation und algorithmischen Analyse von Unterwasserbilddaten entwickelt.

In der Meeresforschung werden verstärkt Bild- und Videodaten aufgezeichnet, um den Zustand und die Entwicklung von Ökosystemen zu erfassen. Die Menge der anfallenden Daten erfordert eine softwaregestützte Auswertung. Für diesen Forschungsbereich wurde im Jahr 2009 das Bio-Image Indexing and Graphical Labelling Environment (BIIGLE) als erstes Online-Annotationssystem für Bilddaten aus der Meeresforschung vorgestellt und erfährt seitdem eine stetig steigende Akzeptanz in den Meereswissenschaften.

Die Erforschung und Beobachtung der Ozeane hat neben dem üblichen Forschungsdrang der Menschheit in diesem Jahrtausend noch zusätzlich an Bedeutung gewonnen. Einerseits müssen die Einflüsse des Klimawandels auf die marinen Ökosysteme erfasst werden. Andererseits müssen auch weitere, ganz direkte Einflüsse des Menschen auf die Weltmeere (zum Beispiel durch Überfischung, Rohstoffgewinnung oder Tourismus) erfasst, studiert und beurteilt werden. In den letzten zehn Jahren haben Technologien wie die hochauflösende Digitalfotografie zu signifikanten Fortschritten in der ingenieurtechnischen Entwicklung von mobilen oder stationären Unterwasser-Trägersystemen geführt. So konnten mit modernen Systemen wie ROV (Remotely Operated Vehicle), AUV (Autonomous Underwater Vehicle), OFOS (Ocean Floor Observation System) oder FUO (Fixed Underwater Observatory) Methoden entwickelt werden, um große Areale des Meeresbodens mit Fotos oder Videos in hoher Qualität zu erfassen oder kleine Areale über einen langen Zeitraum in Fotosequenzen zu beobachten [1]. Die digitalen Bilddaten enthalten vielfältige Informationen über die taxonomische Zusammensetzung sowie die morphologischen Eigenschaften der größeren Fauna. Die Auswertung der rapide wachsenden Menge an Bilddaten benötigt allerdings dringend die Unterstützung durch geeignete Algorithmen und spezialisierte Softwaresysteme.

METHODEN DER BILDAUSWERTUNG

Ziele der Auswertung der Bilddaten sind in den meisten Fällen die Bestimmung und Markierung einer Region im Bild (Schritt 1) und die semantische Annotation dieser Bildregion (Schritt 2). Schritt 1 besteht beispielsweise aus der Wahl eines Bildpunktes, einer Kreis- bzw. Rechteckform oder eines individuell gezeichneten

Polygons an einem definierten Ort im Bild. In Schritt 2 wird eine semantische Kategorie entweder frei formuliert oder aus einem Katalog gewählt und mit der Bildregion verknüpft. Hierbei kann es sich um vordefinierte taxonomische Kataloge aus der Biologie (zum Beispiel aus der WoRMS-Datenbank) oder andere Kataloge handeln, die etwa verschiedene Sorten nicht biologischer Objekte (zum Beispiel Müll) beschreiben. Aufgrund der relativ hohen Diversität einerseits und der teilweise sehr niedrigen Dichte pro Spezies andererseits ist eine Vollautomatisierung der beiden Schritte zur Auswertung der Bilddaten auf absehbare Zeit nicht realistisch. Allein wegen der genannten Gegebenheiten fehlt in der Regel eine ausreichende Menge an semantisch annotierten Bildausschnitten, um moderne maschinelle Lernalgorithmen (sogenanntes Deep Learning) zur automatischen Detektion und/oder Klassifikation der Objekte in den Bild- und Videodaten anzuwenden.

TECHNISCHE UMSETZUNG

Die Arbeitsgruppe Biodata Mining an der Universität Bielefeld hat 2009 ein erstes Online-Annotationssystem für Bilddaten vorgestellt [2]. Dieses System namens BIIGLE ermöglichte es den Meeresbiologinnen und Meeresbiologen zum ersten Mal, mithilfe einer Internetverbindung ihre Bilddaten abzurufen, zu sichten und nachhaltig auszuwerten. Weiterhin erlaubte das System, mit einem sehr einfach und effizient zu bedienenden grafischen Werkzeug die sogenannten Objects of Interest in den Bildern zu markieren und mit vordefinierten semantischen Kategorien zu verknüpfen. Obwohl die vorrangige Motivation hinter BIIGLE die Sammlung von Trainingsdaten für maschinelles Lernen war, wurde das System sehr schnell in den Forschungsbereichen der Meeresbiologie und -geologie populär und wurde in die Arbeitsprozesse zur Auswertung der Bilddaten integriert.

BIIGLE 2.0

Im Jahr 2017 wurde das BIIGLE-System komplett neu implementiert, um weitere Funktionen hinzuzufügen und den gewachsenen Ansprüchen zu entsprechen, die durch eine gestiegene Anzahl von Nutzenden mit diversen Forschungskontexten entstanden sind [3]. Zu den wichtigsten neuen Funktionen zählen neue grafische Annotationswerkzeuge (zum Beispiel magic wand, Polygone; Abbildung 1a), Werkzeuge zur Qualitätssicherung der Annotationen (Abbildung 1b), ein Werkzeug zur Videoannotation, dynamisch und interaktiv durch die Nutzenden gestaltbare hierarchische Kataloge aus semantischen Kategorien (Abbildung 1c), sowie automatische Laserpunktdetektion, neue Geovisualisierungen und ein automatisches Werkzeug zur Objektdetektion auf Basis von maschinellen Lernverfahren.

Seit der Veröffentlichung von BIIGLE 2.0 im Jahr 2017 ist die Zahl der Nutzenden und die Zahl der Annotationen in BIIGLE stetig angestiegen (Abbildung 2). Zu den Nutzenden gehören Meeresforschungsinstitute wie das GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, das Senckenberg Forschungsinstitut in Wilhelmshaven, das französische Ifremer, das englische National Oceanography Centre sowie zahlreiche Universitäten und Forschungsgruppen aus der ganzen Welt. Die meereswissenschaftlichen Themenfelder und Bildtypen nehmen stetig in ihrer Anzahl und Diversität zu. So werden inzwischen neben Bildern und Videos aus mobilen oder stationären Trägersystemen auch Bilder aus der Hellfeldmikroskopie zur Klassifikation von Plankton oder erkranktem Zellgewebe, sowie Luftbildaufnahmen von Flugdrohnen im BIIGLE 2.0-System ausgewertet.

Form von Grafikprozessoren zum wissenschaftlichen Rechnen dar. Diese ermöglichten erstmals die Implementierung von modernen Methoden des maschinellen Lernens in BIIGLE. Ein Beispiel dafür ist die Methode Machine Learning Assisted Image Annotation [4], die seit Anfang 2019 allen Nutzern von BIIGLE zur Verfügung steht.

In Zukunft soll die Nutzung der durch die de.NBI Cloud verfügbaren Ressourcen in BIIGLE weiter ausgedehnt werden. Zum einen sollen weitere Methoden des maschinellen Lernens unter Verwendung von Grafikprozessoren verfügbar gemacht werden. Zum anderen soll das System auf eine bessere Skalierbarkeit durch den Einsatz mehrerer virtueller Maschinen in OpenStack vorbereitet werden, um der steigenden Popularität und Zahl der Nutzenden gerecht zu werden.

HOHE AKZEPTANZ IN DER COMMUNITY

BIIGLE wird seit Februar 2018 vollständig in der OpenStack-Cloud betrieben, die von de.NBI in Bielefeld bereitgestellt wird. Die Umstellung auf OpenStack stellte einen wesentlichen Fortschritt für den Betrieb und die Weiterentwicklung von BIIGLE dar. So konnte durch die Nutzung modernerer Hardware und Software die Geschwindigkeit des Systems mehr als verdoppelt werden. Weiterhin konnte durch die Nutzung von mehreren separaten virtuellen Maschinen in OpenStack die Ausfallsicherheit und Wartbarkeit verbessert werden. Der OpenStack-Dienst zum Speichern großer Datenmengen wurde sukzessive in BIIGLE eingebunden. Neben Bild- und Videodaten verwaltet BIIGLE über diesen Dienst inzwischen auch mehrere Millionen dynamisch generierte Einzeldateien. Einen weiteren Fortschritt stellte die Verfügbarkeit von leistungsfähiger Spezial-Hardware in

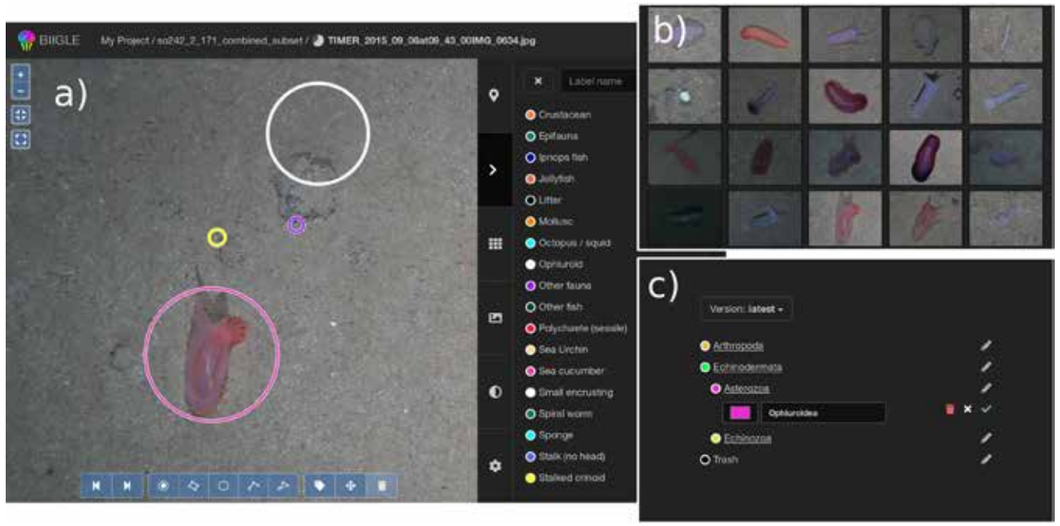


ABBILDUNG 1 (oben): Teile der BIIGLE-Benutzeroberfläche. a) Das Annotationswerkzeug mit Kreisannotationen in der Hauptansicht und der verfügbare Katalog an semantischen Kategorien in der Seitenleiste. b) Übersicht über vorhandene Annotationen zur Qualitätssicherung in dem Label Review Grid Overview-Werkzeug. c) Ansicht zum Bearbeiten eines hierarchischen Katalogs von semantischen Kategorien.

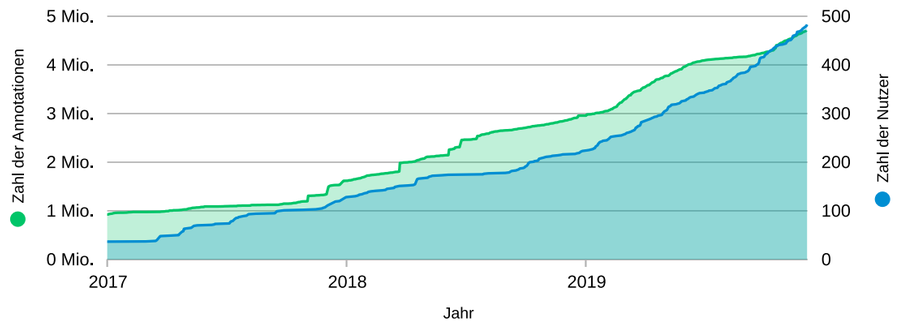


ABBILDUNG 2 (unten): Die Zahl der Annotationen (grün, linke Achse) und die Zahl der Nutzer (blau, rechte Achse) in BIIGLE 2.0 seit der Veröffentlichung in 2017. Die Anfangswerte entstammen der Übernahme von Daten aus der Vorläuferversion von BIIGLE 2.0.

REFERENZEN [1] Oceanography and Marine Biology, 216, pp 9-80. DOI: 10.1201/9781315368597. **[2]** OCEANS 2009-EUROPE. DOI: 10.1109/OCEANSE.2009.5278332. **[3]** Front. Mar. Sci., 28 March 2017 DOI: 10.3389/fmars.2017.00083. **[4]** PLoS One. 2018; 13(11): e0207498. DOI: 10.1371/journal.pone.0207498.

AUTOREN Martin Zurowietz[1], Tim W. Nattkemper[1]
[1]AG Biodata Mining, Technische Fakultät, Universität Bielefeld, Universitätsstraße 25, 33615 Bielefeld

NICHT KULTIVIERBARE BAKTERIEN

Der Zugang zum größten genetischen Schatz der Erde

Aktuelle Schätzungen gehen davon aus, dass die 16.000 bisher kultivierten und beschriebenen Bakterienarten weniger als 0,1 % der in der Natur vorhandenen Arten abdecken. Für eine systematische Erschließung dieses weltweit größten Reservoirs an genetischen Informationen ist die Kultivierung der limitierende Faktor. Die notwendigen Parameter müssen bisher mühsam durch empirische Tests ermittelt werden.

16.000

Aktuelle Schätzungen ...

GEHEN DAVON AUS, DASS DIE 16.000 BISHER KULTIVIERTEN UND BESCHRIEBENEN SPEZIES NUR 0,001 % BIS 0,1 % DER IN DER NATUR VORHANDENEN ARTEN ABDECKEN.

93.000

Bisher wurden ...

FÜR CA. 93.000 ENZYME 150.000 LITERATURREFERENZEN VON WISSENSCHAFTLERN MANUELL AUSGEWERTET UND INSGESAMT 4,7 MILLIONEN DATEN EXTRAHIERT.

Mit der Erfindung des ersten Mikroskops wurden im Jahr 1676 durch Antonie van Leeuwenhoek bereits die ersten Bakterien entdeckt. Über viele Jahre beschränkte sich die Charakterisierung von Bakterien auf die Beobachtung der Morphologie. Erst Ende des 19., Anfang des 20. Jahrhunderts wurden zunehmend physiologische Tests entwickelt, die Unterschiede im Stoffwechsel, dem Aufbau der Zellwand sowie der Resistenz auf Antibiotika zeigten. Bis heute werden neue Bakterienarten mit bis zu 150 physiologischen Eigenschaften beschrieben, um zum einen die besonderen Fähigkeiten neu entdeckter Spezies zu ermitteln und zum anderen Unterschiede zu nahe verwandten Spezies festzustellen. Unterstützt werden diese phänotypischen Untersuchungen heutzutage durch Sequenzanalysen. Anhand der Sequenzen können Verwandtschaftsverhältnisse (die Phylogenie) zu bereits beschriebenen Spezies aufgeklärt werden. Insbesondere kann durch die Sequenzierung der vollständigen Genome aber auch das genetische Potenzial der neuen Spezies untersucht werden. Während neue Sequenzdaten zuverlässig in großen Repositorien hinterlegt werden, wo sie für die Wissenschaft zur Verfügung stehen, finden sich phänotypische Daten relativ verborgen in Laborbüchern oder im Fließtext von Veröffentlichungen. Um die Verfügbarkeit der phänotypischen Daten nachhaltig zu verbessern, werden in den Datenbanken BRENDA [1] und Bac*Dive* [2] manuell Daten aus Publikationen extrahiert, standardisiert und systematisch verfügbar gemacht.

ENZYMDATEN IN BRENDA

In der BRENDA-Datenbank werden seit 30 Jahren Enzyme mit all ihren Eigenschaften charakterisiert. BRENDA hat sich zu einem der weltweit wichtigsten und am stärksten verwendeten Informationssysteme in den Lebenswissenschaften entwickelt und gehört zu den ELIXIR Core Data Resources. In BRENDA werden Daten aus den unterschiedlichsten Quellen zusammengefasst, recherchierbar gemacht und für die Benutzer aufgearbeitet. Dabei ist die manuelle Textauswertung bei Weitem die aufwendigste, aber auf absehbare Zeit noch unverzichtbare Methode, um die in der Literatur sonst nicht zugängliche Information strukturiert für den Wissenschaftler zur Verfügung zu stellen. Für ca. 93.000 Enzyme wurden bisher 150.000 Literaturreferenzen von Wissenschaftlern manuell ausgewertet und insgesamt 4,7 Millionen Daten extrahiert. Um jedoch einen kompletten Überblick über die Literatur zu den klassifizierten Enzymen zu erhalten, werden zusätzlich Textmining-Verfahren eingesetzt. Die Informationen über das Auftreten von Enzymen in Organismen konnten hierdurch im Vergleich zur manuellen Auswertung vervierfacht werden. Insgesamt werden auf diese Weise 3,8 Millionen Literaturzitate erfasst. Darüber hinaus werden auch Daten aus anderen Datenbanken automatisch integriert, wie zum Beispiel Proteinsequenzen aus der UniProt-Sequenzdatenbank oder 3D-Strukturen aus der PDB–Proteinstruktur-Datenbank.

METADATEN ZU BAKTERIEN IN BAC*DIVE*

Seit 2012 wird am Leibniz-Institut DSMZ-Deutsche Sammlung für Mikroorganismen und Zellkulturen mit The Bacterial Diversity Metadatabase (Bac*Dive*) eine Datenbank entwickelt, die bislang nicht verfügbare mikrobiologische Forschungsdaten zugänglich macht. Die erste Datenbankversion beinhaltete Basisdaten wie die Taxonomie, Kultivierungsbedingungen und den Ursprungsort für mehr als 23.000 Bakterien und Archaeen. Über die letzten Jahre wurden die Nutzungsmöglichkeiten von Bac*Dive* erheblich weiterentwickelt. Es

900.000

Aktuell ist ...

BAC*DIVE* MIT ÜBER 900.000 DATENPUNKTEN ZU 80.584 STÄMMEN DIE WELTWEIT UMFASSENDSTE DATENBANK FÜR BAKTERIELLE METADATEN.

31.826

Es wurde...

HIERFÜR EIN DATENSATZ AUS TEMPERATURDATEN VON 31.826 ENZYMEN UND WACHSTUMSTEMPERATURWERTEN VON 21.498 MIKROORGANISMEN GENERIERT.

6,5 MIO.

Darüber hinaus ...

IST DAS MODELL IN DER LAGE, OPTIMALE AKTIVITÄTSTEMPERATUREN FÜR 6,5 MILLIONEN ENZYME VORAUSZUSAGEN.

ABBILDUNG 1: Erfolgreich kultivierte Bakterien auf Agarplatten ©DSMZ.

erfolgte die Mobilisierung neuer Datentypen aus den internen Datenbanken der Kultursammlungen, die bis dahin nicht öffentlich zugänglich waren. Darüber hinaus wurde 2015 damit begonnen, manuell Daten für bis zu 152 Datenfelder aus Speziesbeschreibungen in der Literatur zu extrahieren und in Bac*Dive* zu integrieren. So stehen bereits Daten aus über 6.000 Speziesbeschreibungen zur Verfügung. Mit dem Ziel, möglichst alle phänotypischen Informationen aus Speziesbeschreibungen in der reinen datenbasierten Form in Bac*Dive* verfügbar und durchsuchbar zu machen, wird diese Sammlung kontinuierlich erweitert. Aktuell ist Bac*Dive* mit über 900.000 Datenpunkten zu 80.584 Stämmen die weltweit umfassendste Datenbank für bakterielle Metadaten.

DIE SYNTHESE VON DATEN ERÖFFNET NEUE MÖGLICHKEITEN

Großes Potenzial steckt in der Kombination von Daten aus verschiedenen Quellen, wodurch sich vollständig neue Analysemöglichkeiten eröffnen. Dabei zu überwindende Hürden sind schlechte Auffindbarkeit, beschränkter Zugang, technische Inkompatibilität der Formate und unzureichende Standardisierung. Daher wurde mit der Veröffentlichung der FAIR-Prinzipien (findable, accessible, interoperable, reusable) ein Kulturwandel in der Wissenschaft gestartet, diese Hürden abzubauen und die Verfügbarkeit und Nachnutzung von wissenschaftlichen Daten zu verbessern. Ein passendes Beispiel dafür, welcher Mehrwert durch die Neukombination von Daten erreicht werden kann, wird im Folgenden beschrieben. Der Schwede Martin Engqvist verglich kürzlich in seiner Untersuchung [3] die Kultivierungstemperaturen von Bakterien aus Bac*Dive* mit den optimalen Temperaturdaten für die Aktivität der Enzyme aus BRENDA. Hierfür generierte er einen Datensatz aus Temperaturdaten von 31.826 Enzymen und Wachstumstemperaturwerten von 21.498 Mikroorganismen. Mit diesen Daten konnte er eine starke Korrelation zwischen Wachstumstemperatur und optimaler Enzymtemperatur nachweisen, die eine enge Verknüpfung dieser beiden Parameter zeigt. So kombiniert, bieten sich zahlreiche Möglichkeiten, systematisch die Enzymfunktionen abhängig von der Wachstumstemperatur zu untersuchen. Dabei ist dieser Datensatz nur der erste Schritt zu wesentlich weitreichenderen Studien zur Vorhersage von bisher unbekannten Parametern.

DIE VORHERSAGE VON KULTIVIERUNGSPARAMETERN FÜR BISHER NICHT KULTIVIERBARE BAKTERIEN

Flächendeckend verfügbare, standardisierte Informationen sind die Voraussetzung, um Vorhersagen für bisher nicht bekannte Parameter zu machen. In einer Folgestudie wurde von Forschern um Martin Engqvist auf der Basis des zuvor erzeugten Datensatzes ein Modell entwickelt, das mithilfe von Proteinsequenzdaten präzise die optimale Wachstumstemperatur für Bakterien vorhersagt. Darüber hinaus ist das Modell in der Lage, optimale Aktivitätstemperaturen für 6,5 Millionen Enzyme vorherszusagen.

Die optimale Wachstumstemperatur ist dabei nur einer von vielen Kultivierungsparametern, welche für eine erfolgreiche Kultivierung eines neuen Isolats notwendig sind. Aber andere Studien haben auch hierfür bereits eine Lösung. So haben Forscher um Alice McHardy die Software Traitar entwickelt, die auf Basis von Genomsequenzen von Bakterien bis zu 67 phänotypische Parameter mit unterschiedlicher Sicherheit vorhersagen kann [4]. Diese Parameter umfassen unter anderem Verwertung von Nährstoffen wie Zuckern und Aminosäuren, Salzkonzentration des Mediums, Morphologie und Sauerstoffabhängigkeit. Dies zeigt, dass es durch die Kombination von Daten aus verschiedenen Quellen und durch Kombination von Modellen und der Software verschiedener Entwickler möglich ist, bereits viele Voraussagen für Wachstumsbedingungen für bisher nicht kultivierbare Bakterien zu machen. Durch die konsequente Verbesserung der Datengrundlage können mithilfe dieser Modelle langwierige und kostenaufwendige Laborarbeiten in Zukunft reduziert und so die Effizienz und der Durchsatz in der Untersuchung neuer Bakterienarten deutlich gesteigert werden.

BEDEUTUNG DER VORHERSAGEN DURCH KÜNSTLICHE INTELLIGENZ FÜR DIE WISSENSCHAFT

Erst vor Kurzem konnte gezeigt werden, dass eine mit 100.000 Bildern trainierte Künstliche Intelligenz deutlich bessere Ergebnisse in der Vorhersage von schwarzem Hautkrebs erzielte, als selbst erfahrene Dermatologen [5]. Die Forscher verwendeten in dieser Studie ein künstliches neuronales Netz (Convolutional Neural Network), welches dann aus einem Testdatensatz von 100 Bildern 95 % der Melanome korrekt erkannte. Die Unterstützung durch Künstliche Intelligenz (KI) in der Datenauswertung und in der Vorhersage von bisher nicht bekannten Parametern eröffnet neue Möglichkeiten. Vor allem beim Erkennen von Zusammenhängen in großen Datenmengen kann ein gut trainierter KI-Algorithmus dem Menschen überlegen sein und Vorhersagen mit einer hohen Präzision tätigen. Diese Vorhersagen wiederum dienen dem Menschen als Ausgangspunkt für weitergehende Forschung. Allerdings reichen Vorhersagen alleine nicht aus. Um wissenschaftliche Hypothesen zu belegen, wird daher die Validierung von Vorhersagen im Labor immer ein essenzieller Teil in den Naturwissenschaften bleiben.

Um das große Potenzial der KI-unterstützten Analysen in Zukunft noch besser ausnutzen zu können, sind große Datensätze notwendig, die eine hohe Qualität und ein hohes Maß an Standardisierung aufweisen. Um das zu gewährleisten, sind Datenbanken wie Bac*Dive* und BRENDA, welche Forschungsdaten im großen Stil sammeln, standardisieren und den Wissenschaftlern in hocheffizienter Weise zur Verfügung stellen, essenziell.

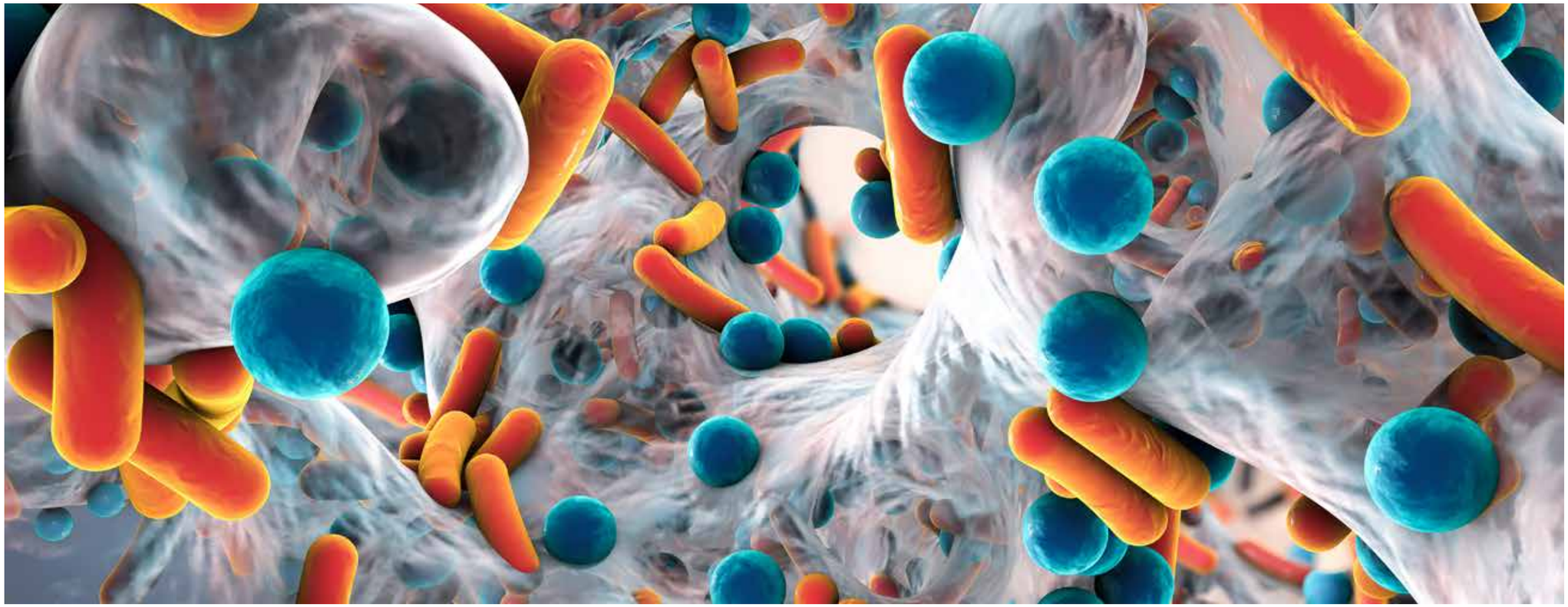
REFERENZEN [1] BMC Microbiol 2018;18(1):177. DOI: 10.1186/s12866-018-1320-7. **[2]** Ann Oncol 2018;29(8):1836-1842. DOI: 10.1093/annonc/mdy166. **[3]** Nucleic Acids Res 2019;47(D1):D542-D549. DOI: 10.1093/nar/gky1048. **[4]** Nucleic Acids Res 2019;47(D1):D631-D636. DoOI: 10.1093/nar/gky879. **[5]** MSystems 2016; 1(6): e00101-16. DOI: 10.1128/mSystems.00101-16.

AUTOREN Lorenz C. Reimer[1], Dietmar Schomburg[2], Jörg Overmann[1]
[1] *Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen, Inhoffenstr. 7B, 38124 Braunschweig*
[2] *Institut für Biochemie, Biotechnolgie & Bioinformatik, Technische Universität Braunschweig, Rebenring 56, 38106 Braunschweig*

IDENTIFIZIERUNG UND ANALYSE

resistenter Krankenhauskeime mithilfe der de.NBI-Cloud

Antibiotikaresistente Bakterien treten weltweit vermehrt in Kliniken, Nutztieren, Nahrungsmitteln sowie der Umwelt auf. Aufgrund zunehmender Resistenzen auch gegen Reserveantibiotika sind diese oft schwierig zu bekämpfen oder gar unbehandelbar. Die Software ASA³P ermöglicht die umfassende Analyse bakterieller Genome und bietet somit die Grundlage für die Entwicklung neuer Bekämpfungsstrategien.



IDENTIFIZIERUNG UND ANALYSE RESISTENTER KRANKENHAUSKEIME MITHILFE DER de.NBI-CLOUD
MIKROBIELLE BIOINFORMATIK

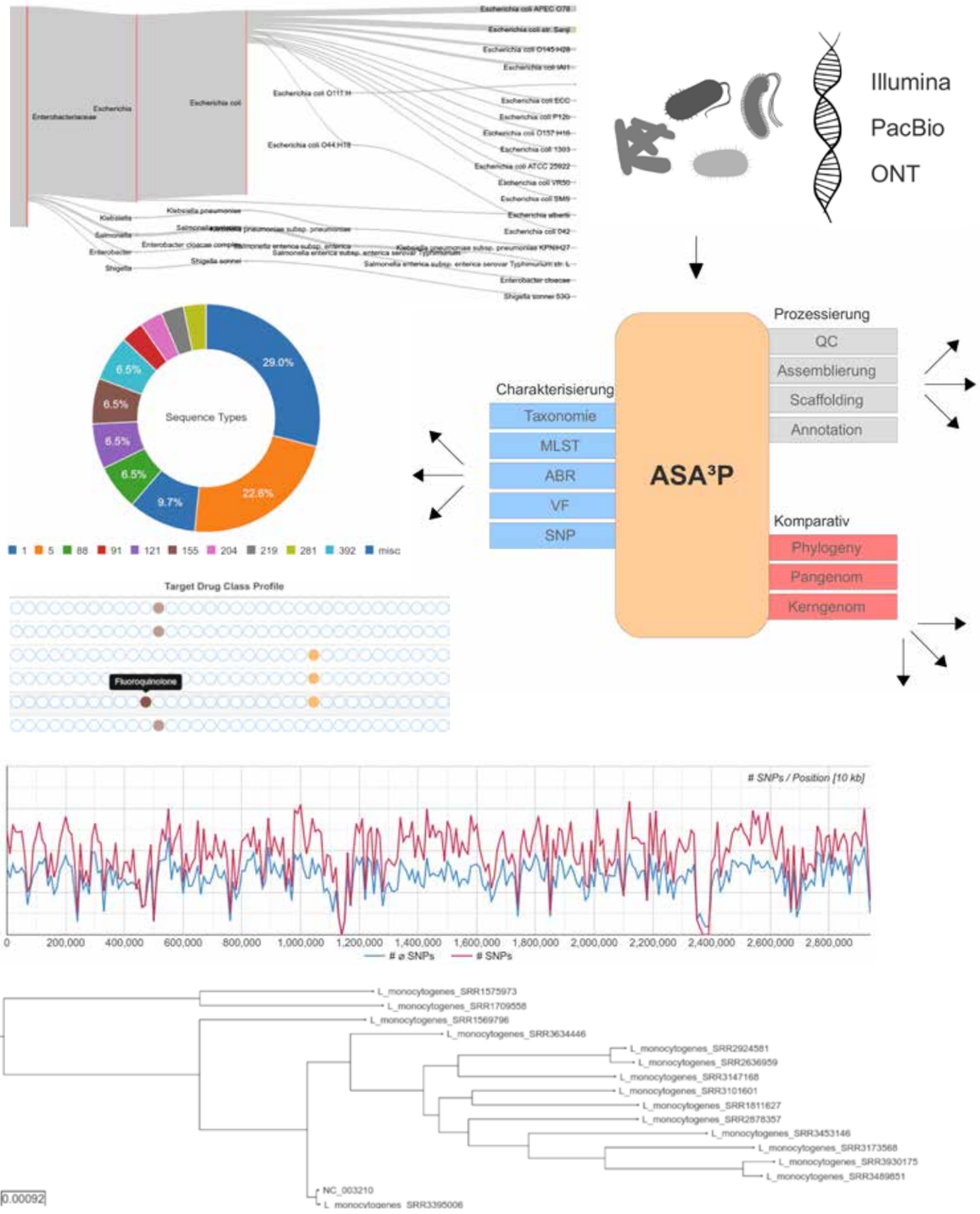
WELTWEITE BEDROHUNG DURCH ANTIBIOTIKARESISTENTE BAKTERIEN

Antibiotikaresistente Bakterien wurden im Jahr 2015 in der EU und im europäischen Wirtschaftsraum für etwa 670.000 Infektionen und 33.110 Todesfälle verantwortlich gemacht. Bis zum Jahr 2050 könnten Antibiotikaresistente Bakterien global betrachtet zum Tod von bis zu zehn Millionen Menschen führen und Kosten in Höhe von 94 Billionen Euro verursachen [1]. Die zunehmende Verbreitung von Antibiotikaresistenzen ist jedoch nicht nur ein Problem für die Krankenhäuser. Antibiotikaresistente pathogene Bakterien wurden in vielen weiteren Bereichen, wie zum Beispiel in Nutztieren, Nahrungsmitteln und der Umwelt, nachgewiesen. Die Weltgesundheitsorganisation (WHO) veröffentlichte 2018 eine Prioritätenliste zur Entwicklung neuer Antibiotika gegen pathogene Bakterien. Carbapenem-resistente, Gram-negative Bakterien (*Enterobacterales, P. aeruginosa, A. baumannii,* sogenannte ESKAPE-Organismen*)* standen auf dieser Liste ganz weit oben [2]. Gerade diese multiresistenten Bakterien tauchen in letzter Zeit immer häufiger auf. Die Sorge, eine postantibiotische Ära zu erreichen, in der bakterielle Infektionen kaum noch mit Antibiotika zu behandeln sind, steigt. Um dieser Gefahr beispielsweise durch die Entwicklung neuer Antibiotika entgegenzuwirken, ist eine genaue Kenntnis der Bakterien vonnöten. Hierzu müssen deren Eigenschaften möglichst genau und für möglichst viele Bakterien analysiert werden.

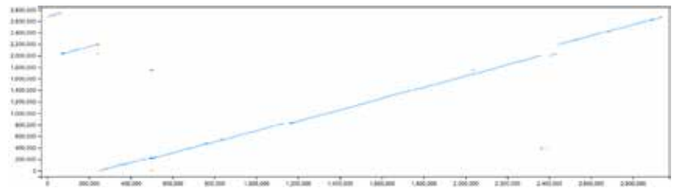
NUTZUNG DER GENOMSEQUENZIERUNG IN DER ANTIBIOTIKARESISTENZFORSCHUNG

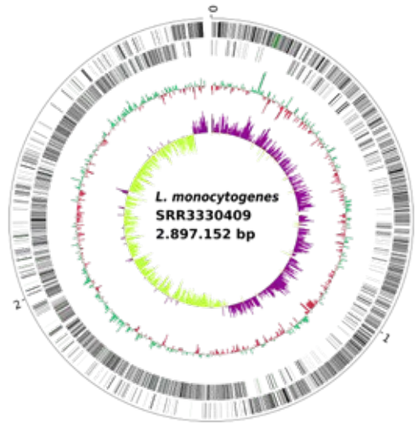
Die Methoden, mit denen Bakterien charakterisiert werden können, haben sich seit dem letzten Jahrhundert deutlich verändert. In den letzten zwanzig Jahren wurden bedeutende Fortschritte auf dem Gebiet der DNA-Sequenzierung erzielt, sodass heute das vollständige Erbgut von Bakterien innerhalb weniger Stunden entziffert werden kann. Dabei wird zunächst die Abfolge der einzelnen Nukleotide (Buchstaben) des bakteriellen Erbguts identifiziert, bevor mithilfe bioinformatischer Methoden die Funktion einzelner Sequenzabschnitte analysiert wird. Durch rapide gesunkene Kosten werden diese Methoden nun vermehrt im Hochdurchsatz für die Untersuchung Antibiotikaresistenter Bakterien eingesetzt. Dies hat zu einer starken Zunahme verfügbarer bakterieller Genomdaten geführt. So sind zum Beispiel bis heute 219.763 *Salmonella*- sowie 106.458 *E. coli*-Stämme sequenziert worden [3].

IDENTIFIZIERUNG UND ANALYSE RESISTENTER KRANKENHAUSKEIME MITHILFE DER de.NBI-CLOUD
MIKROBIELLE BIOINFORMATIK



# Contigs	N50 Cov	N50	Genome Size	GC
19	38	345.992	2.896.690	38
20	31	513.368	3.086.657	38
28	19	509.539	3.023.654	38
85	38	66.196	2.855.624	38
23	45	438.161	2.935.600	38
54	80	261.717	3.027.510	38
27	144	372.909	3.079.427	38
50	65	503.866	3.167.548	38
16	21	1.468.329	2.893.735	38





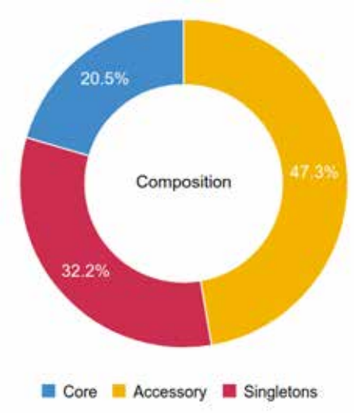


ABBILDUNG 1: Automatisierte Analyse bakterieller Genome mit ASA³P. Die bioinformatische Software ASA³P verarbeitet vollautomatisch die Rohdaten moderner Sequenziermaschinen und führt umfassende und hochspezialisierte Analysen durch. Die vielfältigen und komplexen Analyseergebnisse werden anschaulich visualisiert [2].

HOCHPARALLELE ANALYSE BAKTERIELLER GENOME MITTELS ASA³P

Die Nutzung von Genomsequenzdaten zur Charakterisierung Antibiotikaresistenter Bakterien bietet zwar eine Reihe von Vorteilen; bei deren Generierung und Verarbeitung in Hochdurchsatz-Verfahren sind jedoch verschiedene Herausforderungen zu bewältigen. Auf der einen Seite kann aus den Genomdaten eine Vielzahl an Informationen über diese Bakterien extrahiert werden, welche sonst mit anderen Methoden nicht so einfach bzw. kostengünstig zu generieren wären. Die inzwischen sehr exakten Sequenzdaten bieten die Möglichkeit, einen hochaufgelösten genetischen Fingerabdruck einzelner Bakterien zu erzeugen. So können Antibiotikaresistenzgene oder Pathogenitätsfaktoren identifiziert und die Verwandtschaft zu anderen Bakterien bestimmt werden. Diese genetischen Fingerabdrücke bilden zum einen die Basis für die Entwicklung neuer Strategien gegen Antibiotikaresistente Bakterien; zum anderen können sie in Form vereinfachter Berichte auch an Kliniken oder öffentliche Gesundheitseinrichtungen zurückgemeldet werden.

Auf der anderen Seite stoßen diese Verfahren schnell auf ein allgemeines Problem: Die genetischen Fingerabdrücke müssen aus einer großen Ausgangsdatenmenge extrahiert werden. Bei der Analyse weniger Bakterien ist dies noch manuell möglich, bei der gleichzeitigen Analyse von Dutzenden, Hunderten oder gar Tausenden Bakterien ist jedoch eine automatisierte und hochparallele Analysesoftware erforderlich, da die generierte Menge an Ausgangs-

daten immer weiter steigt und sich heute bereits in der Größenordnung einiger Terabytes bewegt. Zur gezielten und umfassenden Analyse der Genomsequenzdaten wurde daher im Zuge einer Kooperation mit dem Deutschen Zentrum für Infektionsforschung (DZIF, unter Prof. Dr. Trinad Chakraborty) sowie der am de.NBI-Standort Gießen ansässigen Arbeitsgruppe von Prof. Dr. Alexander Goesmann die Analysesoftware ASA³P (Automatic Bacterial Isolate Assembly, Annotation and Analyses Pipeline) entwickelt [2]. ASA³P wurde für die Verarbeitung von Sequenzdaten führender Sequenziertechnologien optimiert. Im ersten Schritt der Analysesoftware werden die Genomsequenzdaten einer Qualitätskontrolle unterzogen und fehlerhafte Daten aussortiert. Anschließend werden die verbleibenden Daten genutzt, um die genetische Information der einzelnen Bakterien abzuleiten (genetischer Fingerabdruck). Im letzten Schritt können die genetischen Fingerabdrücke mehrerer Bakterien miteinander verglichen werden. ASA³P kann innerhalb von Stunden hochauflösende genetische Fingerabdrücke von Hunderten Bakterien erstellen – eine Arbeit, die in Zeiten manueller Ansätze mehrere Wochen bis Monate erfordert hätte. Dies wurde durch spezielle technische Anpassungen ermöglicht, sodass bei Bedarf auch die enormen Rechenkapazitäten der de.NBI-Cloud-Infrastruktur optimal genutzt werden können. Es stehen Wissenschaftlerinnen und Wissenschaftlern verschiedener Disziplinen in der de.NBI-Cloud umfangreiche Rechenkapazitäten zur Verfügung, um sowohl wissenschaftlich spannende als auch gesellschaftlich dringende Probleme und Fragestellungen zu erforschen.

ANWENDUNGSBEISPIELE VON ASA³P

Die Software ASA³P wird im Rahmen von nationalen und internationalen Kooperationen verwendet. So konnten bereits weit über 5.500 bakterielle Krankheitserreger aus Deutschland, Europa und Afrika systematisch analysiert und neue Erkenntnisse zur Bekämpfung der Antibiotikaresistenz gewonnen werden. Zwei Anwendungsbeispiele von ASA³P sollen im Nachfolgenden vorgestellt werden.

HOCHRESISTENTE KEIME IN DEUTSCHLAND ENTDECKT

Im Rahmen der Zusammenarbeit mit dem DZIF wurden Antibiotikaresistente klinische Keime gesammelt, sequenziert und mittels ASA³P analysiert. Bei der Untersuchung der genetischen Fingerabdrücke stellte sich heraus, dass sich unter den analysierten Keimen bereits extrem resistente Bakterien befanden. Diese Bakterien zeigten Resistenzen gegen Antibiotika aus vielen unterschiedlichen Klassen, darunter auch die Reserveantibiotika Colistin und Carbapeneme [4].

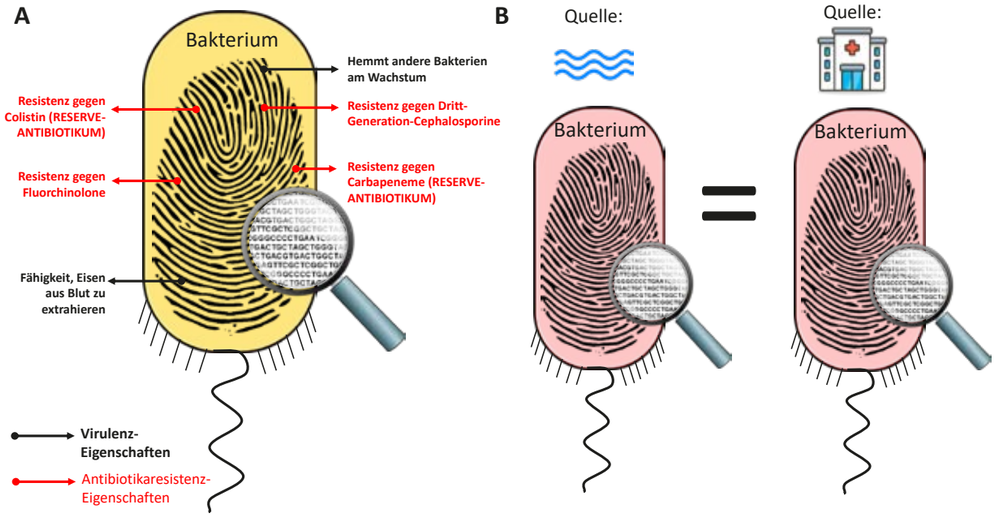


ABBILDUNG 2: Zwei Einsatzmöglichkeiten für die Analysesoftware ASA³P: a) Erstellung von genetischen Fingerabdrücken mit ausgewählten Beispielen für Virulenz- und Antibiotikaresistenz-Eigenschaften; b) Vergleich von Bakterien aus unterschiedlichen Quellen.

VERGLEICHENDE ANALYSE VON WASSERKEIMEN

Eine andere Studie, in der ASA³P zum Einsatz kam, wurde in Kooperation mit Reportern des NDR durchgeführt. Die initiale Fragestellung war, ob multiresistente Bakterien in Gewässern zu finden sind und, wenn ja, ob diese Bakterien bereits früher im klinischen Zusammenhang eine Rolle gespielt haben. Die Genombasierte vergleichende Analyse mittels ASA³P zeigte, dass im Wasser multiresistente Bakterien zu finden sind, die eine hohe Ähnlichkeit zu human-assoziierten Bakterien zeigen. Dies bedeutet zum einen, dass Wasser ein bislang wenig untersuchtes Reservoir für multiresistente Bakterien darstellt, zum anderen aber auch, dass von Gewässern eine potenzielle Gefahr für den Menschen ausgehen kann [5].

AUSBLICK

Die Anwendungsmöglichkeiten von ASA³P zur Analyse mikrobieller Genome sind nahezu unbegrenzt. Die generierten genetischen Fingerabdrücke können mit vielfältigen klinischen Daten kombiniert werden, um bakterielle Strategien zu Antibiotikaresistenzen zu verstehen und neue Ansätze zur Bekämpfung zu entwickeln. Die kombinierte Entwicklung von Genombasierten Ansätzen und leistungsfähigen Softwarelösungen ist ein aufstrebendes Feld in einem systembiologischen Ansatz, um neuartige Einblicke in die Antibiotikaresistenz bakterieller Erreger zu ermöglichen. Mittelfristig sollen diese Ansätze in diagnostische Werkzeuge überführt und zur Vorhersage zukünftiger Entwicklungen genutzt werden können.

Um diesem Anspruch gerecht zu werden, wurde das Microbial Genome Research Center (MGRC) [6] als eine neue Querschnittsplattform etabliert. Diese Plattform beinhaltet eine Datenbank-Komponente und eine Biobank-Komponente. Die Datenbank-Komponente kombiniert vielfältige Daten (genetische Fingerabdrücke, Antibiotikaresistenzdaten, präklinische und klinische Datensätze, Daten aus klassischen Kohorten- und epidemiologischen Studien). Die Biobank-Komponente gibt Wissenschaftlern und Akteuren aus der Industrie den Zugang zu gut charakterisierten aktuellen und historischen Isolaten, sodass neue Ansätze experimentell getestet werden können.

Das MGRC schließt somit die Lücke zwischen grundlegenden bioinformatischen Analysen und der Medizininformatik. Durch die integrierte Auswertung der verschiedenen Daten wird das MGRC dazu beitragen, die Antibiotikaresistenzlast zu bestimmen sowie das Infektionsmanagement und die Infektionskontrolle zu verbessern. Es sollen Daten für Frühwarnsysteme zur Detektion von Ausbrüchen bereitgestellt und Hochrisikoklone identifiziert werden. Dies soll schließlich dazu dienen, die Effektivität von Maßnahmen gegen Antibiotikaresistente Bakterien zu erhöhen und Übertragungen in Krankenhäusern einzudämmen.

REFERENZEN [1] https://www.ime.fraunhofer.de/de/presse/IMI_Project_GNA_NOW.html **[2]** PLOS Computational Biology. DOI: 10.1371/journal.pcbi.1007134. **[3]** Lancet Infect. Dis. 18, 318-327. DOI:10.1016/S1473-3099(17)30753-3. **[4]** https://www.dzif.de/de/wenn-antibiotika-versagen-neues-gen-fuer-antibiotika-resistenz-auch-deutschland-nachgewiesen **[5]** https://www.ndr.de/fernsehen/sendungen/panorama_die_reporter/Auf-der-Spur-der-Superkeime,panorama8258.html **[6]** http://enterobase.warwick.ac.uk/

AUTOREN Oliver Schwengers[1], Linda Falgenhauer[2], Karina Brinkrolf[1], Trinad Chakraborty[2], Alexander Goesmann[1]
[1] *Bioinformatik & Systembiologie, Justus-Liebig-Universität Gießen, 35392 Gießen*
[2] *Institut für Medizinische Mikrobiologie, Justus-Liebig-Universität Gießen, 35392 Gießen und Deutsches Zentrum für Infektionsforschung, Standort Gießen-Marburg-Langen, Justus-Liebig-Universität Gießen, 35392 Gießen*

PHYLOGENETISCHE ANALYSEN ALS WERKZEUG ZUR IDENTIFIZIERUNG VON KRANKHEITSERREGERN

In den letzten 15 Jahren hat die stete Weiterentwicklung der DNA-Sequenzierung es ermöglicht, ganze Gruppen von Bakterien auf Ähnlichkeiten und Unterschiede zu untersuchen. Eines der etabliertesten Werkzeuge in diesem Feld ist die EDGAR-Plattform, die im de.NBI-Netzwerk bereitgestellt und sowohl in der taxonomischen Grundlagenforschung als auch bei konkreten klinischen Fragestellungen weltweit genutzt wird.

8.079

Derzeit werden ...

FÜR 322 GATTUNGEN MIT INSGESAMT 8.079 GENOMEN PROJEKTE IN EDGAR BEREITGESTELLT.

4.400

Zusätzlich gibt es ...

226 PROJEKTE MIT 4.400 GENOMEN, IN DENEN DIE TYPSTÄMME VON TAXONOMISCHEN FAMILIEN ANALYSIERT WERDEN KÖNNEN.

Die Weiterentwicklung moderner Methoden zur DNA-Sequenzierung hat dazu geführt, dass heutzutage möglich ist, ganze Gruppen von Bakterien auf Ähnlichkeiten und Unterschiede zu untersuchen, ein Forschungsansatz, der sich komparative Genomik nennt. Wenn dabei die Abstammungsverhältnisse zwischen den verschiedenen bakteriellen Arten (Spezies) im Mittelpunkt stehen, spricht man von Phylogenomik. Eines der etabliertesten Werkzeuge in der komparativen Genomik und Phylogenomik ist die EDGAR-Plattform, die an der Justus-Liebig-Universität Gießen im Rahmen des Deutschen Netzwerks für Bioinformatik-Infrastruktur (de.NBI) vom Bielefeld-Gießen Resource Center for Microbial Bioinformatics (BiGi) entwickelt und bereitgestellt wird.

DIE EDGAR-PLATTFORM FÜR PHYLOGENOMIK

In den letzten zehn Jahren hat sich die EDGAR-Plattform [1] zu einem der Standardwerkzeuge in der komparativen Genomik entwickelt. EDGAR bietet zahlreiche Analyse- und Visualisierungsfunktionen wie die Berechnung der geteilten und individuellen Genausstattung innerhalb von Genomgruppen, Venn-Diagramme zur Darstellung der differenziellen Genverteilung, zirkuläre Genom-Plots oder multiple Syntenie-Plots. Ein spezieller Fokus der Software liegt auf der Phylogenomik. Die webbasierte Software bietet Zugang zu einer breiten Palette von Werkzeugen zur Analyse der Abstammungsverhältnissse und taxonomischen Einordnung bakterieller Spezies. Dabei werden insbesondere Methoden zur Berechnung von Abstammungsbäumen sowie zur Berechnung von Genom-zu-Genom-Distanzen bereitgestellt; bekannte Methoden sind die Average Nucleotide Identity (ANI) oder die Average Amino Acid Identity (AAI).

Die EDGAR-Datenbank umfasst 12.479 Genome.

Die in EDGAR implementierten Methoden stehen Wissenschaftlerinnen und Wissenschaftlern in vorberechneten Projekten für eine Vielzahl bakterieller Genome als öffentliche Datenbank frei zur Verfügung. Derzeit werden Projekte für 322 Gattungen mit insgesamt 8.079 Genomen bereitgestellt; zusätzlich gibt es noch 226 Projekte mit 4.400 Genomen, in denen die Typstämme von taxonomischen Familien analysiert werden können. Insgesamt umfasst die als de.NBI-Service bereitgestellte EDGAR-Datenbank also 12.479 Genome.

Über die öffentliche EDGAR-Datenbank hinaus bietet EDGAR auch die Möglichkeit, unveröffentlichte Daten im Rahmen wissenschaftlicher Kooperationen in passwortgeschützten Projekten zu analysieren. Eine sehr erfolgreiche Kooperation entwickelte sich in den letzten Jahren mit dem Landesbetrieb Hessisches Landeslabor (LHL), der Verbraucherschutzeinrichtung des Landes Hessen in den Bereichen Veterinärmedizin, Lebensmittelanalytik und Landwirtschaft. Einige wissenschaftliche Resultate aus der Kooperation des LHL mit dem de.NBI-Service EDGAR sollen im folgenden Abschnitt beispielhaft dargestellt werden.

DEN LETZTEN BEISSEN DIE … RATTEN!

Rattenbissfieber ist eine vergleichsweise selten diagnostizierte und unbekannte Zoonose, also eine Infektionskrankheit, die von Tieren auf den Menschen übertragen werden kann (Abbildung 2). *Streptobacillus (S.) moniliformis* ist der wichtigste ursächliche Erreger. Beim Menschen äußert sich Rattenbissfieber mit hohem Fieber, rötlichen Hautausschlägen und Gelenkentzündungen; bei schwerwiegenden Komplikationen (zum Beispiel Gehirnabszessen, Herzklappenentzündungen oder Blutstrominfektionen) kann die Krankheit tödlich verlaufen. Gelegentlich können auch andere Tiere erkranken, darunter Puten, verschiedene Nagerarten sowie Koalabären und nicht humane Primaten.

Obwohl Rattenbissfieber weltweit auftritt und die Kolonisierungsrate der meist nicht erkrankten Ratte bei über 90 % liegen kann, gilt die Infektion als unterdiagnostiziert und ist selbst unter medizinischem

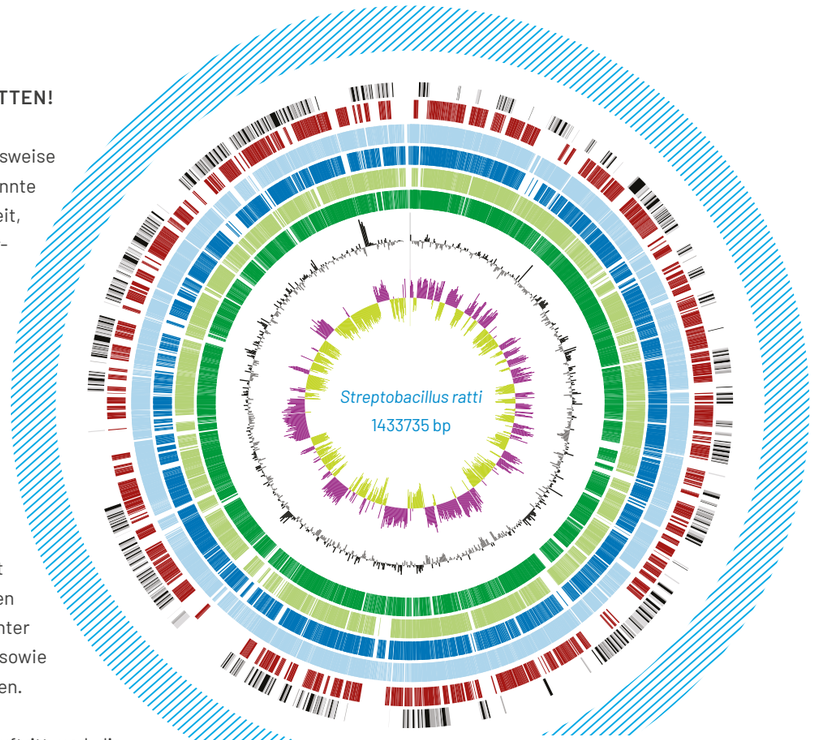


ABBILDUNG 1: Zirkuläre Darstellung des Genoms von *Streptobacillus rattii* OGS16T im Vergleich mit vier anderen Streptobacillus-Genomen. Der äußere schwarze Ring zeigt die Verteilung der Gene in *S. rattii*. Der rote Ring zeigt die Gene, die in den vier Vergleichsstämmen konserviert sind. Die grünen und blauen Ringe zeigen jeweils die Anordnung übereinstimmender Gene in den ausgewählten *Streptobacillus-Genome S. moniliformis, S. hongkongensis, S. felis* und *S. notomytis*.

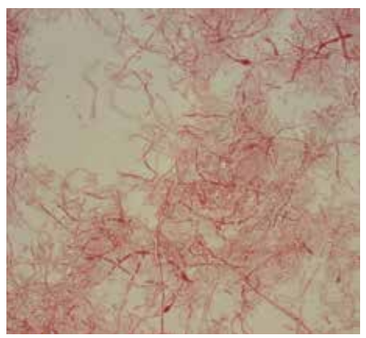


ABBILDUNG 2: Rattenbissfieber sowie eine Reihe weiterer Zoonosen können nicht selten selbst von als Haustiere gezüchteten Farbmorphen der Wanderratte übertragen werden. Der oft sehr unkritische Umgang mit diesen Haustieren führt gerade bei Kindern häufig zu Erkrankungen. (Foto links: https://pixabay.com/de/photos/ratte-m%C3%A4dchen-park-457984/, Foto rechts: Tobias Eisenberg)

Fachpersonal relativ unbekannt. Trotz intensivierter Forschung betrifft dies insbesondere die Variabilität des Erregers, dessen Pathogenese sowie seine Ausstattung mit Virulenzfaktoren [2]. Eigene Untersuchungen konnten zeigen, dass die für fast 90 Jahre nur aus *S. moniliformis* bestehende Gattung *Streptobacillus* tatsächlich artenreicher ist. Inzwischen konnte diese Gattung um vier Spezies (*S. hongkongensis*, *S. felis*, *S. notomytis* und *S. ratti*) erweitert werden (Abbildung 1), von welchen mindestens eine auch bereits im Zusammenhang mit menschlichem Rattenbissfieber erwähnt wurde. Oft sind diese neuen Erreger nur anhand einzelner oder weniger Stämme beschrieben worden und auch für *S. moniliformis* ließen sich trotz weltweiter Akquise lediglich rund 24 Isolate in einer Stammsammlung am LHL zusammenführen. Von ihnen wurde nachfolgend das Genom sequenziert. Die zeitliche und räumliche Bandbreite der Isolate war dabei gewaltig und erstreckte sich über 90 Jahre, nahezu alle Erdteile und unterschiedliche Wirtsspezies, von welchen *S. moniliformis* zuvor isoliert worden war.

Weil die Ähnlichkeit insbesondere des 16S-rRNA-Gens innerhalb dieser Verwandtschaftsgruppe sehr hoch ist und Unterscheidungen auf Speziesebene somit zweifelbehaftet bleiben, wurde versucht, höher diskriminierende Gensequenzen zu finden, um eine speziesspezifische Diagnostik vorantreiben zu können [3]. Mithilfe von EDGAR wurden aber auch phylogenetische Fragestellungen innerhalb der Gattung sowie in den nächstverwandten Taxagruppen untersucht. Auch bei Analysen zum Vorkommen von Virulenzgenen, Resistenzfaktoren und Phagen bei *Streptobacillus* kam die EDGAR-Plattform zum Einsatz. Somit kommt knapp ein Jahrhundert nach der Erstbeschreibung des Erregers erstmals Licht in wesentliche Aspekte dieser vernachlässigten Zoonose.

EDGAR UND DAS DRECKIGE DUTZEND

In Analogie zur Schreckenshitliste der Giftstoffe wurde von der US-amerikanischen Gesundheitsbehörde CDC auch eine solche Liste für potenziell waffenfähige biologische Agenzien zusammengestellt. Auf der Liste von einem Dutzend Bioterrorismus-Agenzien der zweithöchsten Kategorie findet sich die Bakteriengattung *Brucella*, weil sie beim Menschen schwere, mitunter tödliche, in jedem Fall aber monatelange Erkrankungen auslöst. Die Brucellose ist außerhalb dieser Betrachtungen gleichfalls eine Zoonose, die hierzulande nur sehr selten auftritt, in Endemiegebieten jedoch jährlich geschätzte 500.000 Neuinfektionen verursacht. Diese nehmen ihren Ausgang im Kontakt mit infizierten Tieren oder durch Verzehr von rohen tierischen Lebensmitteln. Bislang galten Brucellen als Krankheitserreger ausschließlich bei Säugetieren. Nachdem der Arbeitsgruppe am LHL 2012 erstmals der Nachweis von Brucellen bei Fröschen (Abbildung 3) und somit bei einer unerwarteten und relativ weit entfernt verwandten Tierklasse gelang [4], wurden diese in den Folgejahren bei weiteren Amphibien weltweit nachgewiesen. Einige Jahre später wurde erneut der Erstnachweis bei einer weiteren Tierklasse durch unsere Zusammenarbeit geführt und näher entschlüsselt: Dem Brucella-Nachweis bei einem tropischen Stachelrochen [5] folgte eine umfangreiche Genomcharakterisierung mit EDGAR, an welcher auch das Institut für Mikro-

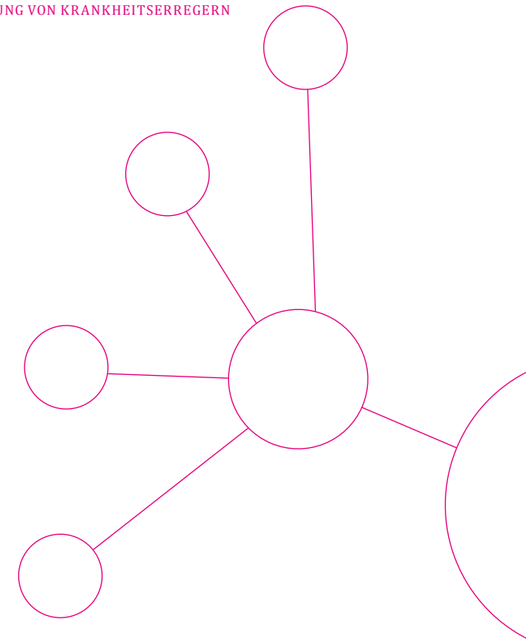
ABBILDUNG 3: Am LHL gelang erstmals der Nachweis, dass Brucellen auch Frösche befallen können. Im Bild oben ein Glasfrosch (*Sachatamia ilex*) aus Costa Rica. (Foto oben: Tobias Eisenberg, Foto unten: iStock)

biologie der Bundeswehr beteiligt war. Die Stämme von Fröschen und Rochen sind sehr nahe miteinander verwandt und verkörpern mittlerweile eine eigenständige phylogenetische Stellung innerhalb der Gattung *Brucella*. Derzeit ist noch nicht sehr viel darüber bekannt, ob diese Bakterien dieselben schwerwiegenden Erkrankungen beim Menschen auslösen können wie ihre bei landwirtschaftlichen Nutztieren vorkommende Verwandtschaft. Jedoch wurden bei schwer erkrankten Menschen auch bereits ähnliche Stämme isoliert, ohne dass man bis heute einen Bezug zu den in Rede stehenden wechselwarmen Wirtstieren hergestellt hatte. Möglicherweise liegt hier eine noch vergleichsweise basale Entwicklungsform vor, die einen Übergangszustand zwischen einem von abgestorbenen organischen Substanzen lebenden Bodenbewohner und einem hoch an Säugetiere und den Menschen angepassten Infektionserreger repräsentieren könnte. Mit EDGAR konnten in den Genomen der Stämme von Fischen und Fröschen nämlich sowohl Gene aus harmlosen Bodenbakterien als auch dieselben Virulenzgene klassischer Säuger-Brucellen identifiziert werden. Weitere Analysen werden zeigen, ob von diesen Stämmen eine ähnliche Gefahr ausgeht.

30.000

Die EDGAR-Plattform ...

GEHÖRT ZU DEN MEISTGENUTZTEN SERVICES DES de.NBI-NETZWERKS MIT NUTZERN AUS ÜBER 200 UNIVERSITÄTEN UND FORSCHUNGSINSTITUTEN WELTWEIT UND EINEM ANALYSEAUFKOMMEN VON FAST 30.000 BAKTERIELLEN GENOMEN PRO JAHR.



EDGAR AUF DEM WEG IN DIE ZUKUNFT

Die angeführten Beispiele demonstrieren die vielfältige Nutzbarkeit der EDGAR-Plattform zur Analyse von bakteriellen Genomen sowohl in der taxonomischen Grundlagenforschung als auch in Bezug auf konkrete klinische Fragestellungen. Dementsprechend gehört EDGAR zu den meistgenutzten Services des de.NBI-Netzwerks mit Nutzern aus über 200 Universitäten und Forschungsinstituten weltweit und einem Analyseaufkommen von annähernd 30.000 bakteriellen Genomen pro Jahr.

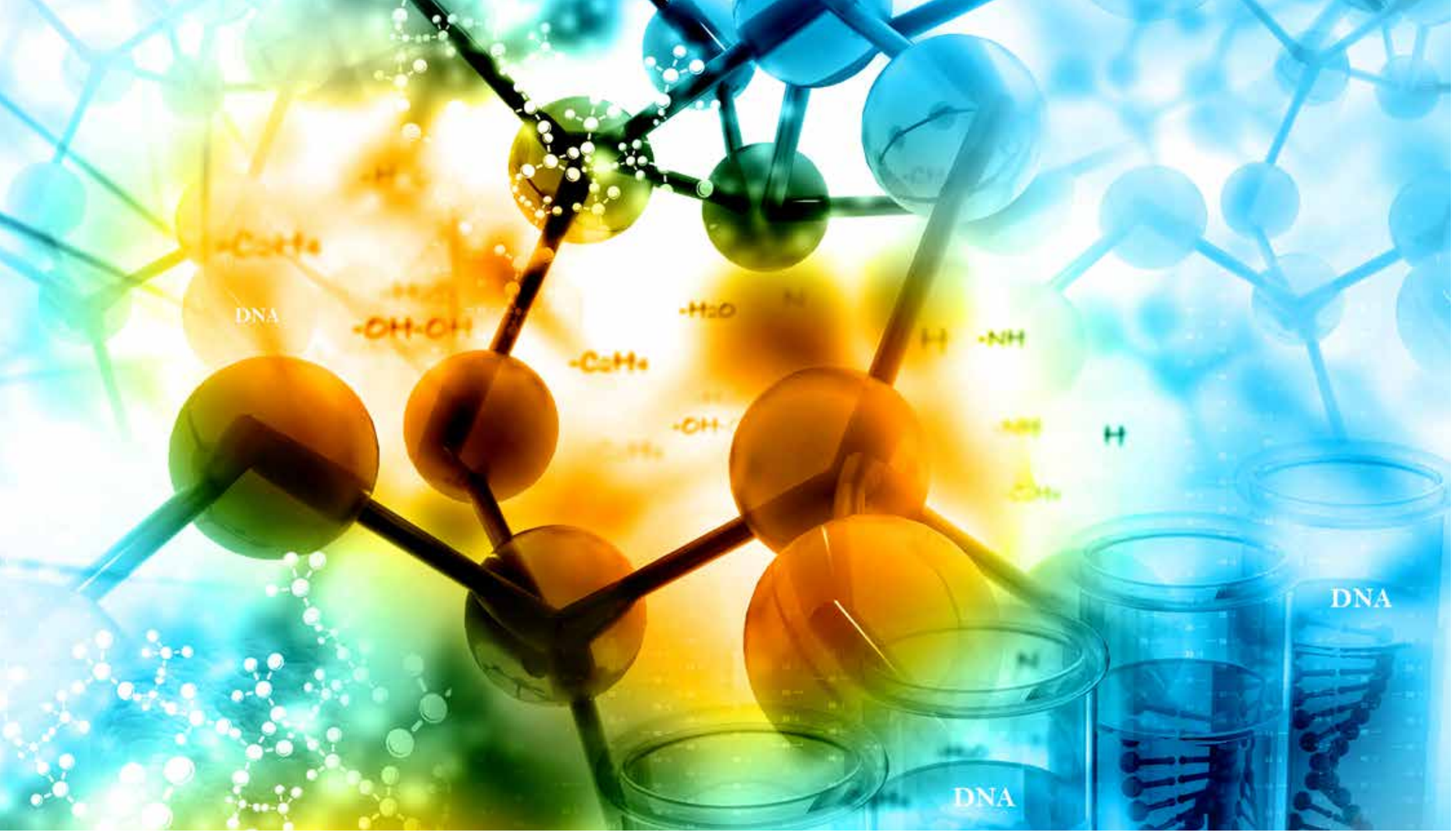
Da die Kapazität moderner Sequenziersysteme bei sinkenden Kosten weiter zunimmt, sind stetig technische Anpassungen an EDGAR nötig, um mit den Datenmengen Schritt zu halten. Daher ist ein kompletter Austausch der zugrunde liegenden Datenstruktur geplant, der es erlaubt, die EDGAR-Analysen entsprechend der Anzahl der untersuchten Genome auf skalierbare Weise mit den dafür benötigten Hardware-Ressourcen zu versorgen. Dabei sollen bei Bedarf auch die extrem umfangreichen Ressourcen der de.NBI-Cloud genutzt werden können. Zusammen mit den damit einhergehenden Änderungen im Datenmanagement soll die Funktionsfähigkeit von EDGAR auch für Großprojekte mit Hunderten bis hin zu Tausenden von Genomen sichergestellt werden. Ein weiterer Fokus wird auf der Integration neuer phylogenomischer Analysen liegen. So sind diverse schnelle Alternativen zu den etablierten ANI/AAI-Methoden verfügbar. Diese werden derzeit evaluiert und zukünftig in die EDGAR-Plattform integriert.

Durch die Integration von zeitgemäßen Markergen-basierten Ansätzen, wie der Verwendung des Universal Bacterial Core Genome (UBCG), ist EDGAR auf einem guten Weg, in Zukunft eine wichtige Rolle in der komparativen Genomik im Allgemeinen und der Phylogenomik im Speziellen zu spielen.

EDGAR-WEBSERVER

REFERENZEN **[1]** In Bergey's Manual of Systematics of Archaea and Bacteria. DOI:10.1002/9781118960608.bm00038. **[2]** VVB Laufersweiler Verlag 2018; URL: http://geb.uni-giessen.de/geb/volltexte/2018/13567/. **[3]** BMC Genomics 2016;17(1):864. DOI:org/10.1186/s12864-016-3206-0. **[4]** Appl Environ Microbiol. 2012;78(10):3753-5. DOI:10.1128/AEM.07509-11. **[5]** Antonie Van Leeuwenhoek. 2017;110(2):221-234. DOI:10.1007/s10482-016-0792-4.

AUTOREN Jochen Blom[1], Tobias Eisenberg[2], Alexander Goesmann[1]
[1] *Bioinformatik & Systembiologie, Justus-Liebig-Universität, 35392 Gießen*
[2] *Landesbetrieb Hessisches Landeslabor, Schubertstraße 60, 35392 Gießen*



BRENDA – EINE ESSENZIELLE RESSOURCE

für die Entwicklung von biotechnologischen Stoffproduktionswegen

Der Artikel beschreibt die hohe Bedeutung des BRENDA Enzym-Informationssystems für die Entwicklung neuartiger biotechnologischer Produktionsverfahren für komplexe Wirk- und Wertstoffe. Im Rahmen solcher Projekte wird BRENDA sowohl für das Design neuer metabolischer Pfade bei der Auswahl geeigneter Mikroorganismen als auch für das Training von KI-Software für die Experimentplanung eingesetzt.

Biotechnologische Stoffproduktion ist eines der am schnellsten wachsenden Anwendungsfelder in der Bioökonomie. Neben der Verwendung von Stoffwechselwegen, die in bekannten Organismen natürlicherweise vorkommen, wie zum Beispiel die Alkoholproduktion in Hefe oder die Produktion von Antibiotika in Pilzen, wird die Produktion neuartiger Produkte in letzter Zeit häufig durch die Kombination von Stoffwechselwegen aus verschiedenen Organismen konzipiert oder über das ab-initio-Design völlig neuer Stoffwechselwege durch die zielorientierte Kombination von Enzymen zugänglich. Essenziell hierfür ist die genaue Kenntnis der Enzymeigenschaften für die Auswahl geeigneter Enzyme im Hinblick auf Substratspezifität, Stoffumsatz, Stabilität, Temperatur usw.

HINTERGRUND

Die moderne biologische und biotechnologische Forschung ist schon seit vielen Jahren ohne den permanent notwendigen Zugang zu Faktendatenbanken nicht denkbar. Spätestens mit den ersten Sequenzierungen von Genen oder Proteinen sowie den ersten 3D-Strukturbestimmungen von Proteinen vor 50 Jahren wurde deutlich, dass hier in großem Umfang experimentelle Daten entstehen, die nicht mehr manuell, sondern nur über Algorithmen effizient ausgewertet werden können und für die Auswertung oder effiziente Planung von Experimenten unabdingbar sind.

Während aber Sequenz- und Strukturdaten in Repositorien hinterlegt werden, sind heute die allermeisten anderen Daten, wie zum Beispiel Funktionen und Eigenschaften von Proteinen, unzugänglich in Publikationen versteckt. Um diese in strukturierter Form zugänglich zu machen, müssen diese Publikationen manuell oder in begrenztem Umfang durch Textmining-Verfahren ausgewertet, strukturiert, standardisiert und schließlich effizient zugänglich gemacht werden, wie dies zum Beispiel für Enzyme in BRENDA [1] geschieht.

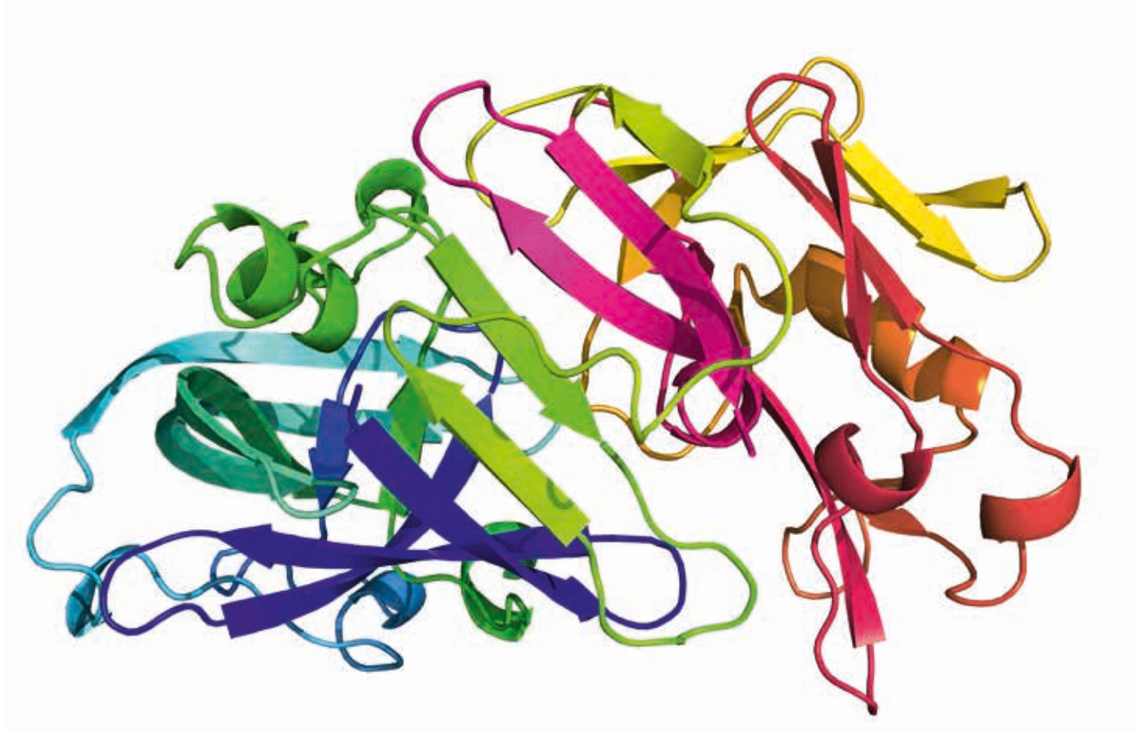
ENZYMDATEN AUS BRENDA

In der BRENDA-Datenbank werden seit 30 Jahren Enzyme mit all ihren Eigenschaften charakterisiert. BRENDA hat sich zu einem der weltweit wichtigsten und am stärksten verwendeten Informationssysteme in den Lebenswissenschaften entwickelt und gehört zu den ELIXIR Core Data Resources. Mehr als 80.000 Benutzer aus allen Ländern der Erde werden pro Monat gezählt. In BRENDA werden Daten aus den unterschiedlichsten Quellen zusammengefasst, recherchierbar gemacht und für die Benutzer aufgearbeitet. Dabei ist die manuelle Textauswertung bei Weitem die aufwendigste, aber auf absehbare Zeit noch unverzichtbare Methode, um die in der Literatur sonst nicht zugängliche Information strukturiert für den Wissenschaftler zur Verfügung zu stellen. Für ca. 93.000 Enzyme wurden bisher 150.000 Literaturreferenzen von Wissenschaftlern manuell ausgewertet und insgesamt 4,7 Millionen Daten extrahiert. Um jedoch einen kompletten Überblick über die Literatur zu den klassifizierten Enzymen zu erhalten, werden zusätzlich Textmining-Verfahren eingesetzt. Insbesondere zum Vorkommen von Enzymen, zum Zusammenhang von Enzym und Krankheit sowie zu bestimmten kinetischen Daten können Informationen durch automatische Textverarbeitungsverfahren aus Titel oder Abstrakt einer Publikation mit hoher Genauigkeit ermittelt und integriert werden. Die Informationen über das Auftreten von Enzymen in Organismen konnten hierdurch im Vergleich zur manuellen Auswertung vervierfacht werden. Insgesamt werden auf diese Weise 3,8 Millionen Literaturzitate erfasst.

Im Gegensatz zu den Pathway-Datenbanken werden in BRENDA alle bekannten Reaktionen und Substrate, das heißt auch in Organismen nicht vorkommende, aufgenommen.

BRENDA – eines der weltweit wichtigsten und am stärksten verwendeten Informationssysteme in den Lebenswissenschaften.

Daten aus anderen Datenbanken werden automatisch integriert, zum Beispiel Proteinsequenzen aus der UniProt-Sequenzdatenbank, 3D-Strukturen aus der PDB-Proteinstruktur-Datenbank, sequenzierte Genome, taxonomische Daten, Ontologien mit Bezug zu Enzymfunktionen und vieles mehr. Berechnete Daten vervollständigen



die Information. Hierzu gehören zum Beispiel die Vorhersagen von Enzym-Funktion (Genomannotation), Protein-Lokalisierung, Transmembranregionen oder Organismusklassen-spezifische statistische Verteilungen kinetischer Parameter.

BRENDA wird für eine Vielzahl von verschiedensten Projekten aus dem gesamten Bereich der Lebenswissenschaften verwendet, wie zum Beispiel aus Abbildung 1 hervorgeht, die die häufigsten wissenschaftlichen Begriffe aus den Titeln von ca. 1.600 Veröffentlichungen, die BRENDA zitieren, wiedergibt. Die Größe der Buchstaben repräsentiert die Frequenz des jeweiligen Begriffs.

BRENDA UND BIOTECHNOLOGISCHE ANWENDUNGEN

Eine Vielzahl von Publikationen beschreibt die Verwendung von BRENDA-Daten für das Design von Enzymen mit neuen Eigenschaften sowie für das Design von ganzen Stoffwechselwegen, die entweder in einem bestimmten Organismus gentechnologisch integriert werden oder für hocheffiziente In-vitro-Produktionssysteme verwendet werden.

Besonders hervorgehoben wird von den Autoren häufig die Tatsache, dass die BRENDA-Daten manuell erstellt und somit von höherer Qualität als automatisch erzeugte Daten seien. Hier werden von den ca. 50 Datenfeldern in BRENDA insbesondere die breite Beschreibung der von jedem Enzym katalysierten Umsetzungen unter Einschluss der Umsetzung von synthetischen Verbindungen, die kinetischen Daten, die Informationen über Aktivatoren und Inhibitoren sowie die Informationen über die Stabilität der Enzyme bei bestimmten Temperaturen, pH-Werten und gegenüber Sauerstoff und organischen Lösungsmitteln verwendet. Die Information in BRENDA über das Vorkommen in bestimmten Organismen ebenso wie über die Tatsache, dass das Enzym in bestimmten Organismen nicht beobachtet wird, sowie die Information, ob es sich um ein Membran-gebundenes Enzym handelt, spielen ebenfalls eine Rolle wie auch die Zusammenstellung der Sequenzen homologer Enzyme in Relation zur Substratspezifität bzw. Stabilität.

Aus der hohen Zahl der Anwendungen sollen hier nur fünf verschiedenartige instruktive Beispiele aus kürzlich erschienenen Publikationen kurz beschrieben werden. In einem Enzym-Design-Projekt verwendeten die Autoren den in BRENDA angegebenen Bereich von kinetischen Daten für verschiedene 3,4-Dihydroxyphenylacetaldehydsynthasen zum Training eines Algorithmus (M-Path), der ihnen die Konstruktion eines bifunktionellen Enzyms erlaubt, das alternativ als Aldehydsynthase und Decarboxylase arbeitet und zur Produktion von zum Beispiel Dopamin verwendet werden kann [2].

Im Hinblick auf konkrete Anwendungen bei der Konstruktion von ganzen Stoffwechselwegen sei hier nur die Konstruktion einer synthetischen Biochemie-Plattform zur zellfreien Produktion von Monoterpenen aus Glucose erwähnt [3]. Die Autoren verwenden kinetische

BRENDA – EINE ESSENZIELLE RESSOURCE FÜR DIE ENTWICKLUNG
VON BIOTECHNOLOGISCHEN STOFFPRODUKTIONSWEGEN
MIKROBIELLE BIOINFORMATIK

Werte aus BRENDA zur Konstruktion eines Modells zur Planung eines Systems aus 27 Enzymen, das stabil ohne Zugabe von ATP oder NADH zum Beispiel Limonen, Pinen und Sabinen mit >95 % Ausbeute und Titern von >5 g/l mit einer einzigen Glucose-Zugabe produziert. Die mit dem System erzielten Konzentrationen der Produkte liegen eine Größenordnung über der Konzentration, die zelltoxisch wäre und mit zum Beispiel bakteriellen Systemen erreicht werden könnte.

In einem zweiten Projekt beschreiben die Autoren die Produktion von Glucarsäure mit 75 % Ausbeute aus Sucrose durch in-vitro-metabolisches Engineering [4]. Glucarsäure findet Anwendung in der Lebensmittel-, Kosmetik- und pharmazeutischen Industrie.

In einem Review-Artikel beschreiben die Autoren Ansätze, die in BRENDA integrierten „Nebenaktivitäten" von Enzymen gegenüber natürlichen und künstlichen Substraten zum Verständnis dazu zu verwenden, wie in der Evolution neue metabolische Pfade entstehen und wie sie für die Entwicklung von neuartigen biotechnologischen Prozessen benutzt werden können [5]. Diese „Nebenaktivitäten" haben oft eine um Größenordnungen geringere katalytische Effizienz und werden in den typischen *Pathway*-Datenbanken nicht erwähnt.

Schon vor einigen Jahren wurde diese ausschließlich in BRENDA gespeicherte Information unter anderem zum Training eines Algorithmus verwendet, der in der Lage ist, in Organismen alternative metabolische Pfade zwischen zwei Metaboliten zu entdecken und diese Information zu nutzen [6].

ABBILDUNG 1: Die BRENDA Word Map – häufigste Begriffe aus den Titeln von 1.600 Publikationen, die BRENDA zitieren.

REFERENZEN [1] Nucleic Acids Res 2019;47(D1):D542-D549. DOI: 10.1093/nar/gky1048. **[2]** Nat Commun 2019;10(1):2015. DOI: 10.1038/s41467-019-09610-2. **[3]** Nat Commun 2017;8:15526. DOI: 10.1038/ncomms15526. **[4]** ChemSusChem 2019;12(10):2278-2285. DOI: 10.1002/cssc.201900185. **[5]** Curr Opin Biotechnol 2018;49:108-114. DOI: 10.1016/j.copbio.2017.07.015. **[6]** Bioinformatics 2009;25(22):2975-82. DOI: 10.1093/bioinformatics/btp507.

AUTOREN Dietmar Schomburg[1], Ida Schomburg[1], Lisa Jeske[1], Antje Chang[1], Sandra Placzek[1]
[1] *Institut für Biochemie, Biotechnologie & Bioinformatik, BRICS, Technische Universität Braunschweig, Rebenring 56, 38106 Braunschweig*

HUMANE BIOINFORMATIK – DER NUTZEN FÜR DIE MEDIZIN

In der modernen Humanmedizin werden in Zukunft individuelle Sequenzdaten, aber auch Omics-Daten eine wichtige Rolle spielen. Die Nutzung dieser Daten bietet neue Perspektiven für die Erforschung von Krankheiten und deren Entstehung – bis hin zu einer frühzeitigen und individualisierten Behandlung.

VON PROTEINSTRUKTUREN ZU NEUEN MEDIKAMENTEN

Welche Proteine spielen bei einer bestimmten Erkrankung eine Rolle und was ist über sie bekannt? Welche Eigenschaften muss ein Wirkstoff haben, der diese Proteine beeinflusst? Gibt es Forschungsdaten dazu und wie ist deren Qualität? Proteins*Plus* und BRENDA bieten Antworten auf Fragen, die bereits im Rahmen des rationalen Wirkstoffentwurfs vor teuren Laboruntersuchungen gestellt werden können.

VON PROTEINSTRUKTUREN ZU NEUEN MEDIKAMENTEN
HUMANE BIOINFORMATIK

Konventionelle Medikamentenentwicklung ist ein kosten- und zeitintensiver Prozess. Für gewöhnlich dauert der Entwicklungszyklus eines Medikaments 14 Jahre und kostet über 800 Millionen US-Dollar. Um Kosten und Zeit zu sparen, wird der rationale Wirkstoffentwurf eingesetzt, in dem bereits vor intensiven Laboruntersuchungen Computermodelle genutzt werden. Der Webservice Proteins*Plus* in Verbindung mit dem Enzym-Informationssystem BRENDA stellt wichtige Komponenten für diesen Prozess zur Verfügung. Wie genau, demonstrieren wir anhand des Proteins Aldosereduktase, welches bei einer Diabeteserkrankung für schwerwiegende Folgeerkrankungen mitverantwortlich ist.

Beim rationalen Wirkstoffentwurf wird zunächst der Wirkort des zu entwickelnden Medikaments bestimmt, häufig ein Enzym. Ein Enzym ist ein Protein, das eine spezifische chemische Reaktion ermöglicht oder beschleunigt. Dabei tritt es mit anderen Proteinen oder kleineren Molekülen, den sogenannten Liganden, in Kontakt und bildet mit diesen Wechselwirkungen aus. Liganden können kleine, natürlich vorkommende Moleküle in der Zelle oder Wirkstoffe von Medikamenten sein. Durch Liganden, die an bestimmte Zielproteine binden, wird deren Funktion beeinflusst, was in der Medikamentenentwicklung genutzt wird. Um neue Wirkstoffe für Medikamente entwickeln zu können, ist es wichtig, neben der Funktionsweise auch die räumliche Struktur des Proteins zu kennen. Der Fokus liegt dabei auf der Region, an die der Wirkstoff binden soll, dem sogenannten aktiven Zentrum. Auf Basis struktureller Informationen über Protein und Ligand können Computermodelle Vorhersagen über die Wechselwirkungen zwischen diesen beiden treffen. Dieses Wissen ist hilfreich für die Auswahl von kleinen Molekülen, die als Startstrukturen für die Entwicklung neuer Wirkstoffe in Medikamenten dienen können.

Krankheiten, für die erfolgreich Medikamente mithilfe des rationalen Designs entwickelt wurden, sind HIV, Tuberkulose, Krebs, Diabetes, Rheuma und viele andere [1].

ANWENDUNGSBEISPIEL ALDOSEREDUKTASE – VERMINDERUNG VON DIABETES-FOLGEERKRANKUNGEN

Die Aldosereduktase ist ein Enzym (EC: 1.1.1.21); unter anderem wandelt es Glucose in Sorbit um und reduziert Aldehyde, die in unterschiedlichen Stoffwechselwegen entstehen. Eine Diabeteserkrankung führt häufig zu zeitweise stark erhöhten Glucosewerten im Blut. Die Umsetzung von Glucose zu Sorbit führt zu einer Ansammlung von Sorbit im Körper, da dieses nur sehr langsam abgebaut werden kann. Hohe Sorbitwerte wiederum sind vor allem für Nieren, Nerven und Augen sehr schädlich. Um Folgeschäden von Diabetes zu lindern, versucht man daher, Medikamente zu entwickeln, die die Aldosereduktase hemmen und somit die Umsetzung von Glucose zu Sorbit reduzieren.

2.000

BRENDA bietet ...

DEM FORSCHER FÜR DIE ALDOSEREDUKTASE UMFASSENDE INFORMATIONEN MIT MEHR ALS 2.000 DATEN AUS 89 PUBLIKATIONEN.

DATEN ZUR ALDOSEREDUKTASE AUS DEM ENZYM-INFORMATIONSSYSTEM BRENDA

Das Enzym-Informationssystem BRENDA [2] hat sich zu einem der weltweit wichtigsten und am stärksten verwendeten Informationssysteme in den Lebenswissenschaften entwickelt und gehört zu den ELIXIR Core Data Resources.

In BRENDA werden Daten aus den unterschiedlichsten Quellen zusammengefasst, recherchierbar gemacht und für die Benutzer aufgearbeitet.

Für ca. 93.000 Enzyme wurden bisher 150.000 Literaturreferenzen von Wissenschaftlern manuell ausgewertet und insgesamt 4,7 Millionen Daten extrahiert. In Kombination mit Text-Mining und Datenintegrationsverfahren werden insgesamt Daten aus 3,8 Millionen Literaturzitaten erfasst.

Zu der hier in Frage stehenden humanen Aldosereduktase bietet BRENDA dem Forscher umfassende Informationen mit mehr als 2.000 Daten aus 89 Publikationen. Für das Design von Wirkstoffen sind insbesondere die ca. 500 bekannten Inhibitoren, die in BRENDA mit wichtigen Daten wie Inhibitionskonstanten, Verweisen auf Proteinstrukturdaten und wissenschaftlichen Publikationen verknüpft sind, von hoher Bedeutung. Insgesamt 26 Verweise führen zu Publikationen, in denen die medizinische Relevanz des Enzyms für die Entwicklung von Wirkstoffen bei Diabetes diskutiert wird. Auf diese Weise ist für Entwicklerinnen und Entwickler neuer Wirkstoffe eine schnelle und effiziente Erfassung des wissenschaftlichen Hintergrundes gegeben.

UNTERSUCHUNG DER ALDOSEREDUKTASE MIT DEM WEBSERVICE PROTEINS*PLUS*

Proteins*Plus* [3] ist ein Webservice [4], der am Zentrum für Bioinformatik Hamburg entwickelte Softwarewerkzeuge zum rationalen Wirkstoffentwurf öffentlich verfügbar macht. Somit können Wissenschaftlerinnen und Wissenschaftler online Proteinstrukturen für ihre Forschung auswählen und analysieren.

Von zentraler Bedeutung für das Arbeiten mit Proteinstrukturen ist deren Visualisierung, die beim Proteins*Plus*-Webserver durch den integrierten NGL viewer [5] realisiert wird.

Wenn bereits Liganden in der Proteinstruktur vorhanden sind, werden diese als Strukturdiagramme visualisiert und verfügbar gemacht. Mithilfe der in Proteins*Plus* verfügbaren Werkzeuge können wichtige Informationen über die Eigenschaften des aktiven Zentrums der Aldosereduktase gesammelt und aufbereitet werden.

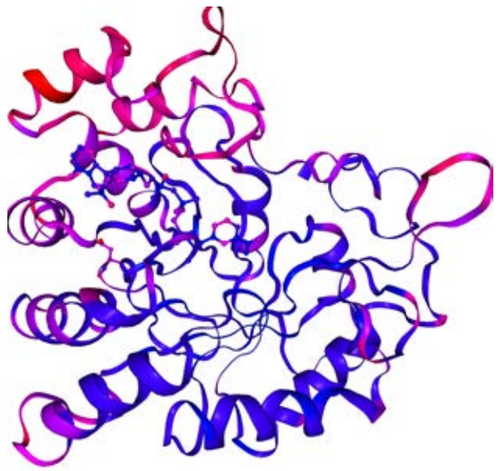


ABBILDUNG 1: Aldosereduktasestruktur mit EDIA-Färbung: Rot = schlechte Qualität, blau = gute Qualität.

160.000

Es werden ...

IN DER PDB CA. 160.000 3D-STRUKTURDATEN VON GROSSEN BIOLOGISCHEN MOLEKÜLEN BEREITGESTELLT.

WELCHE ALDOSEREDUKTASE-STRUKTUREN GIBT ES?

Die Grundlage aller Berechnungen im Proteins*Plus*-Webservice stellen Proteinstrukturen dar. In der öffentlich zugänglichen Protein-Datenbank PDB [6] werden ca. 160.000 3D-Strukturdaten von großen biologischen Molekülen bereitgestellt. Proteinstrukturen werden über einen vierstelligen alphanumerischen Code archiviert. Dieser muss dem Wissenschaftler nicht bekannt sein; Services wie Proteins*Plus* bieten moderne Textsuchfunktionen, wie wir sie von Internetsuchmaschinen kennen. Die Textsuche nach dem Beispielprotein „aldose reductase" ergibt 187 Treffer, die nach verschiedensten Kriterien weiter reduziert werden können. Wir entscheiden uns für eine sogenannte Holostruktur mit dem Code 1ah4, welche neben dem Protein einen für die Funktion wichtigen Kofaktor enthält.

QUALITATIVE ANALYSE VON STRUKTURMODELLEN

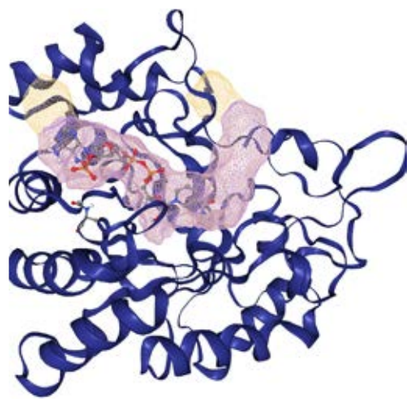
Zu Beginn der Medikamentenentwicklung ist es notwendig, die Qualität der Daten, auf deren Basis gearbeitet werden soll, zu überprüfen. Dafür stellt der Proteins*Plus*-Webserver zwei Softwarewerkzeuge zur Verfügung.

Eines dieser Werkzeuge ist EDIA, ein Programm zur Überprüfung eines dreidimensionalen Strukturmodells mit der zugrunde liegenden Elektronendichte. Die Elektronendichte ist das primäre Ergebnis, welches bei der Strukturaufklärung entsteht. Basierend auf der Elektronendichte wird im Anschluss die 3D-Struktur modelliert. Experimentelle Daten wie Elektronendichtekarten enthalten Varianzen und Ungenauigkeiten, die für die weitere Verwendung von Bedeutung sind. EDIA dient der Berechnung und Darstellung der Genauigkeit des Modells. Für die von uns ausgewählte Aldosereduktase-Holostruktur wird mithilfe von EDIA sichtbar, welche Bereiche der Struktur weniger gut aufgelöst sind (Abbildung 1). Diese liegen in unserem Fall allerdings außerhalb des aktiven Zentrums, welches für die weiteren Analysen von Interesse ist und eine hohe Genauigkeit aufweist.

IDENTIFIKATION DES AKTIVEN ZENTRUMS

Im Anschluss an die qualitative Untersuchung der Proteinstruktur wird die konkrete Ausdehnung des aktiven Zentrums bestimmt. Da die Holostruktur der Aldosereduktase noch keinen Wirkstoff enthält, wird mithilfe des DoGSiteScorers die potenzielle Bindetasche bestimmt. Basierend auf topologischen und chemischen

a



b

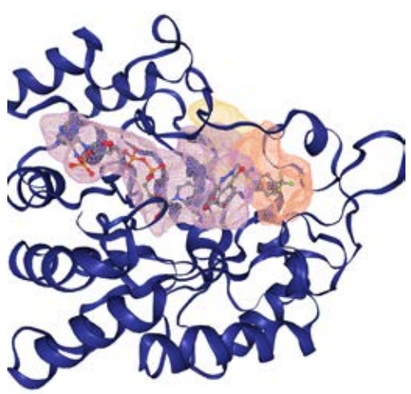


ABBILDUNG 2: Aldosereduktasestrukturen mit durch den DoGSiteScorer bestimmten potenziellen Bindetaschen: a) Holostruktur 1ah4, lila = Subtasche mit Kofaktor; b) Aldosereduktasestruktur mit Kofaktor und Zopolrestat 1dfr, orange = Öffnung einer weiteren Region der Bindetasche, hervorgerufen durch das Medikament Zopolrestat.

Eigenschaften untersucht der DoGSite-Scorer die Proteinstruktur, listet mögliche Bindetaschen auf und berechnet für die Bindetaschen die Wahrscheinlichkeit, dass diese mit einem Wirkstoff wechselwirken können. Für die Aldosereduktase werden acht Taschen gefunden, von denen die Tasche P_0 den Kofaktor enthält und zusätzlich genug Platz für ein kleines Molekül bietet (Abbildung 2a).

BEKANNTE LIGANDEN UND DEREN STRUKTURELLE AUSWIRKUNGEN

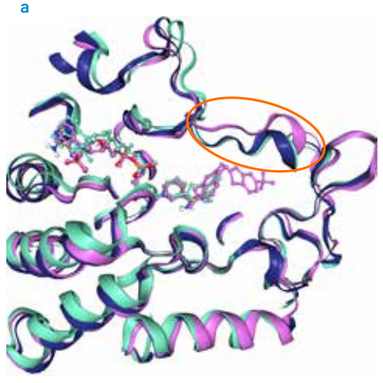
Im Anschluss an die Definition der Bindetasche stellt sich die Frage, wie flexibel diese ist und ob es bereits bekannte Liganden, sowohl natürliche Substrate als auch Medikamente, gibt. Dafür wird mit SIENA in der PDB nach identischen Bindetaschen gesucht. Basierend auf der durch den DoGSiteScorer definierten Bindetasche sucht SIENA nach Bindetaschen mit nahezu identischer Aminosäurensequenz, die sich allerdings in ihrer räumlichen Struktur unterscheiden können. Die Suche mit SIENA ergibt 164 Treffer, die nun visuell weiter untersucht werden können. Zwei Strukturen – mit den PDB-Codes 1ah0 und 1frb – beinhalten Inhibitoren der Aldosereduktase: 1ah0 enthält den recht kompakten Wirkstoff Sorbinil und 1frb enthält den deutlich größeren Wirkstoff Zopolrestat, der langgestreckt in der Bindetasche liegt (Abbildung 3). Hier fällt beim Vergleich mit dem Holoprotein auf, dass sich die 3D-Anordnung des Proteins verändert hat und das große Molekül Zopolrestat in einen Bereich ragt, der in der ursprünglichen Holostruktur nicht zugänglich war. Durch den größeren Liganden Zopolrestat hat sich somit die Bindetasche der Aldosereduktase weiter geöffnet (Abbildung 2b).

SCHLUSSFOLGERUNGEN AUS DER STRUKTURELLEN ANALYSE DER ALDOSEREDUKTASE

Die strukturelle Untersuchung der Aldosereduktase zeigt, dass die Bindetasche, die das aktive Zentrum des Proteins bildet, sehr flexibel ist und ein Ligand zu strukturellen Anpassungen führen kann. Dieses Phänomen, in der Fachsprache als Induced Fit bezeichnet, verdeutlicht die Relevanz der genauen strukturellen Analyse im Wirkstoffentwurf.

Für die Medikamentenentwicklung bedeutet dies, dass möglichst unterschiedliche Strukturen verwendet werden sollten, um ein breites Spektrum an Strukturvariationen zu integrieren. Proteins*Plus* unterstützt auch diesen Prozess und ermöglicht so die effektive Nutzung von Strukturdaten für die computergestützte Suche nach neuen Wirkstoffen.

a



b

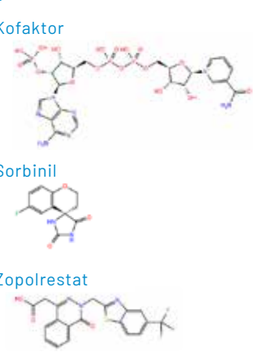


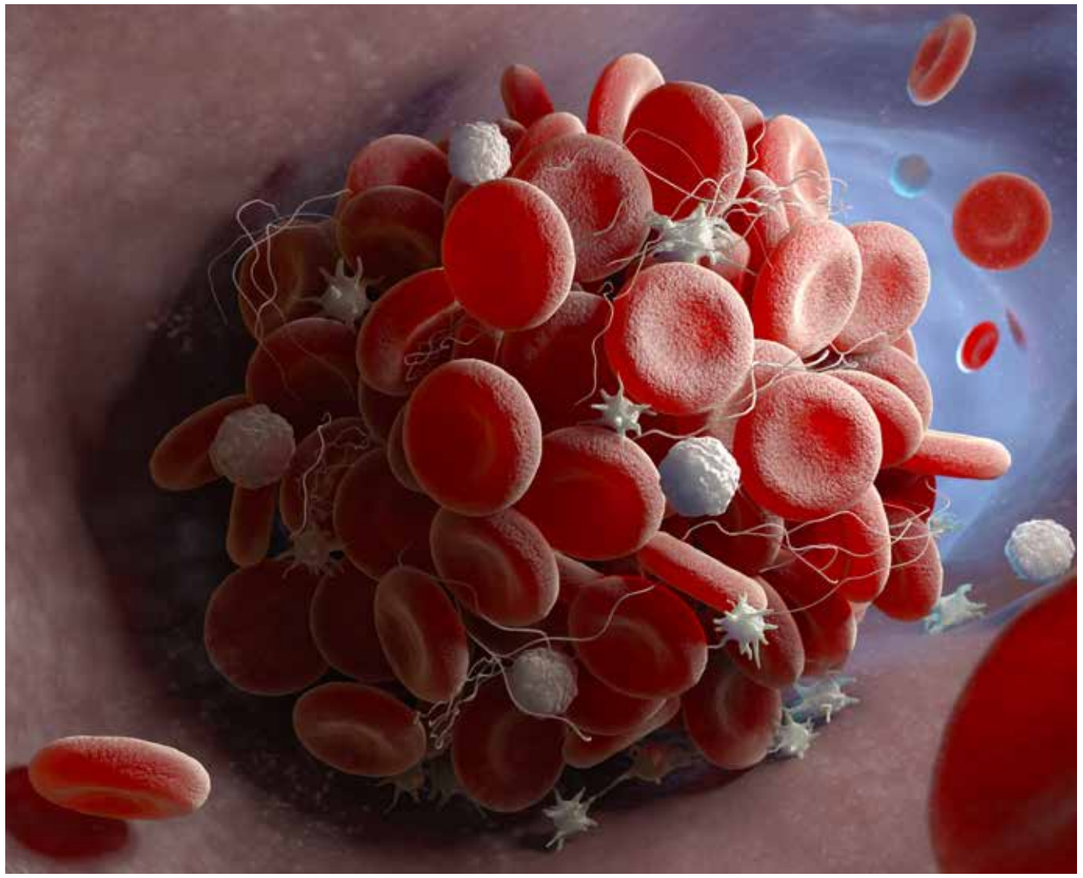
ABBILDUNG 3: a) Überlagerung von Aldosereduktasestrukturen (blau = Holostruktur, grün = Sorbinil, pink = Zopolrestat) mithilfe von SIENA-Färbung, Zopolrestat führt zu einer Öffnung der Bindetasche (orange eingekreist); b) 2D-Struktur der gebundenen Liganden.

REFERENZEN [1] Int. J. Mol. Sci. 2019, 20 (11). DOI: 10.3390/ijms20112783. **[2]** Nucleic Acids Res. 2019, 47: D542–D549. DOI: 10.1093/nar/gky1048. **[3]** Nucleic Acids Res. 2017, 45 (W1), W337–W343. DOI: 10.1093/nar/gkx333. **[4]** https://proteins.plus/ **[5]** Bioinformatics 2018, 34 (21), 3755–3758. DOI: 10.1093/bioinformatics/bty419. **[6]** https://www.rcsb.org/

AUTOREN Katrin Schöning-Stierand[1], Eva Nittinger[1], Dietmar Schomburg[2], Ida Schomburg[2], Matthias Rarey[1]
[1] *Universität Hamburg, ZBH – Zentrum für Bioinformatik, Bundesstraße 43, 20146 Hamburg, http://uhh.de/zbh*
[2] *Technische Universität Braunschweig, BRICS, Rebenring 56, 38106 Braunschweig*

LIPIDOMIK – WIE LIPIDE DIE BLUT-GERINNUNG STEUERN

Lipide, abgeleitet vom griechischen Wort für Fett, sind neben Proteinen und Kohlehydraten die häufigsten Biomoleküle jeder Zelle und zuständig für verschiedene Funktionen wie Schutz, Energiespeicher und Signaltransduktion. Hier präsentieren wir am Beispiel der Blutplättchen, wie mittels bioinformatorischer Techniken das Lipidom, die Gesamtheit aller Lipide, analysiert und wichtige Einblicke in die Blutgerinnung mit medizinischer Implikation gewonnen werden können.



LIPIDOMIK – WIE LIPIDE DIE BLUTGERINNUNG STEUERN
HUMANE BIOINFORMATIK

WAS IST DIE LIPIDOMIK?

Die Lipidomik ist eine noch recht junge Forschungsrichtung, in der mit modernen massenspektrometrischen und anderen chemisch-analytischen Hochdurchsatzmethoden die Struktur, die Zusammensetzung und die genaue Menge von Lipiden in biologischen Proben bestimmt werden. Dies ermöglicht den Vergleich von Lipidkonzentrationen zwischen gesunden und erkrankten Patienten und die Bestimmung von Lipiden als Biomarker für die Diagnose von Krankheiten und deren Verlauf sowie zur Kontrolle des Behandlungsfortschritts. Allerdings reicht eine isolierte Betrachtung von Lipiden allein nicht aus. Daher forscht die Lipidomik auch interdisziplinär an der Verknüpfung von Informationen über Gene, Proteine, Regulation und Exposition (zum Beispiel Umweltgifte), um ein besseres systemisches Gesamtbild als Grundlage der personalisierten Medizin zu ermöglichen.

BIOINFORMATISCHE ANWENDUNGEN FÜR DIE LIPIDOMIK

de.NBI unterstützt die bioinformatische Seite der Lipidomik in Deutschland im Rahmen des Teilprojekts „Lipidomics Informatics in the Life Sciences" (LIFS) [1] mit den Forschungspartnern Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V. in Dortmund (AG Lipidomics, Robert Ahrends), dem Forschungszentrum Borstel, Leibniz-Lungenzentrum (AG Bioanalytische Chemie, Dominik Schwudke) sowie dem Max-Planck-Institut für molekulare Zellbiologie und Genetik in Dresden (AG Biologische Massenspektrometrie, Andrej Shevchenko). Die Partner entwickeln und pflegen hierfür Programme zur Durchführung und Auswertung massenspektrometrischer Messungen wie LipidXplorer [2] und LipidCreator & Skyline [3] zur Bestimmung von Lipididentität und -konzentration sowie zum Vergleich von Lipidprofilen aus verschiedenen Messungen mit LUX Score [4] und Clover (Abbildung 1). Am Standort Dortmund wird die Referenzdatenbank LipidCompass für Lipidkonzentrationen in verschiedenen Geweben aufgebaut. Hierfür werden zunächst Proben der Modellsysteme Blutplättchen und -plasma als Referenz mit quantifizierten Lipiden von verschiedenen nationalen wie internationalen Kooperationspartnern hinterlegt.

ANWENDUNGSFALL: STEUERUNG DER BLUTGERINNUNG DURCH LIPIDE IN BLUTPLÄTTCHEN

Blutplättchen (Thrombozyten) spielen eine wichtige Rolle bei der Blutgerinnung nach Verletzungen der Blutgefäße. Durch die Aktivierung infolge einer solchen Verletzung verändern diese ihre Form und vernetzen sich mithilfe von Fibrin mit ihren Nachbarn. Dies führt zur Bildung eines Blutgerinnsels (Thrombus), welches die verletzte Stelle verstopft und so einen weiteren Blutverlust verhindert. Leider werden Blutplättchen auch durch andere Faktoren aktiviert, was zu Thromben in ansonsten nicht verletzten, aber eventuell durch Vorerkrankungen geschädigten Blutgefäßen führt. Dies hat unerwünschte Nebenwirkungen, da durch die Verstopfung wichtiger Blutgefäße die Nährstoff- und Sauerstoffversorgung von Organen oder Körperteilen teilweise oder ganz unterbrochen wird. Typische akute Folgen sind hier zum Beispiel der Herzinfarkt sowie Embolien. Diese führen Jahr für Jahr weltweit zu zahlreichen Todesfällen und sehr häufig zu starken gesundheitlichen Einschränkungen der betroffenen Patienten. Allerdings gibt es auch (zumeist erbliche) Krankheiten, bei denen die Blutgerinnung gestört ist, wodurch innere sowie äußere Verletzungen zu massivem Blutverlust der betroffenen Personen führen können, da dort die Bildung eines stabilen Thrombus zum Wundverschluss unterbleibt.

Unser Anwendungsfall [5] liefert eine erste Bestandsaufnahme der von Blutplättchen in der Maus gebildeten Lipide und von deren Konzentrationen im ruhenden Zustand sowie nach deren Aktivierung und validiert diese an menschlichen Blutplättchen. Ein besonderer Schwerpunkt lag auf dem besseren Verständnis des Stoffwechselmechanismus der Niemann-Pick-Typ-A/B-Krankheit in Blutplättchen. Diese erbliche Lipidspeicher-Erkrankung führt unter anderem zu einer erheblich reduzierten Lebenserwartung der betroffenen Personen sowie als Folge des beeinträchtigten Lipidstoffwechsels zu einer stark reduzierten Blutgerinnungsfähigkeit.

Hierbei wurde festgestellt, dass ein bestimmtes Lipid (die Spezies PI 18:0-20:4, Abbildung 2) während der Aktivierung der Blutplättchen hauptsächlich als Vorstufe weiterer, für den Gerinnungsmechanismus wichtiger Lipide dient. Bei Niemann-Pick-Typ-A/B-Betroffenen fehlt ein bestimmtes Protein, welches die Vorstufe von Lysoshpingomyelin (SPC) nun nicht mehr zu Ceramiden umwandelt, sondern zu einer Anreicherung von SPC im Laufe der Plättchenaktivierung führt. SPC wiederum stört die Bildung von Blutgerinnseln, was anhand von gesunden menschlichen Blutplättchen nach Aktivierung belegt wurde. Allerdings hat die Studie auch Hinweise auf weitere Mechanismen ergeben, die in zukünftigen Arbeiten genauer untersucht werden sollen.

LIFS bietet zahlreiche Veranstaltungen, Schulungen und Workshops zur Lipidbioinformatik

LIFS.ISAS.DE

ZUKÜNFTIGE ANWENDUNGSMÖGLICHKEITEN

Ein genaues Verständnis der biochemischen Mechanismen hinter der Thrombenbildung unterstützt Mediziner und Pharmakologen zukünftig dabei, gezielte Medikamente und Behandlungen zu entwickeln, die dabei helfen können, Infarkte, Thrombosen und Embolien zu verhindern bzw. die Blutgerinnung besser steuern zu können. Weiterhin erlauben aus den Lipidprofilen abgeleitete diagnostische Biomarker die frühzeitige Entdeckung von Thrombosen sowie die Überwachung des Behandlungsfortschritts. Diese Arbeit wird daher in Zukunft dazu beitragen, dass vielen Patienten mit akuten und chronischen Störungen der Blutgerinnung schneller und zielgenauer geholfen werden kann, um Todesfälle und negative, infarkt- und thrombosebedingte gesundheitliche Langzeitfolgen besser verhindern zu können.

Für die automatisierte Analyse und Auswertung der massenspektrometrischen Messungen wurden die Programme LipidXplorer und LipidCreator mit Skyline verwendet. Die Datenintegration und die statistischen Vergleiche wurden mit Python und R und die Abbildung auf metabolische Netzwerke mit Cytoscape [6] umgesetzt. Quantitative Visualisierungen der Lipidkonzentrationen zwischen Mensch und Maus wurden auf Basis einer R/Shiny-Anwendung implementiert. Diese Werkzeuge halfen maßgeblich dabei, die Vielzahl an Messungen und Lipidkonzentrationen überhaupt erst sinnvoll kombinieren, vergleichen sowie interpretieren zu können. Das LIFS-Projekt stellt diese selbst entwickelten Anwendungen unentgeltlich zur Verfügung, sodass auch andere Forscher im Bereich der Lipidomik diese für ihre eigenen Arbeiten nutzen können.

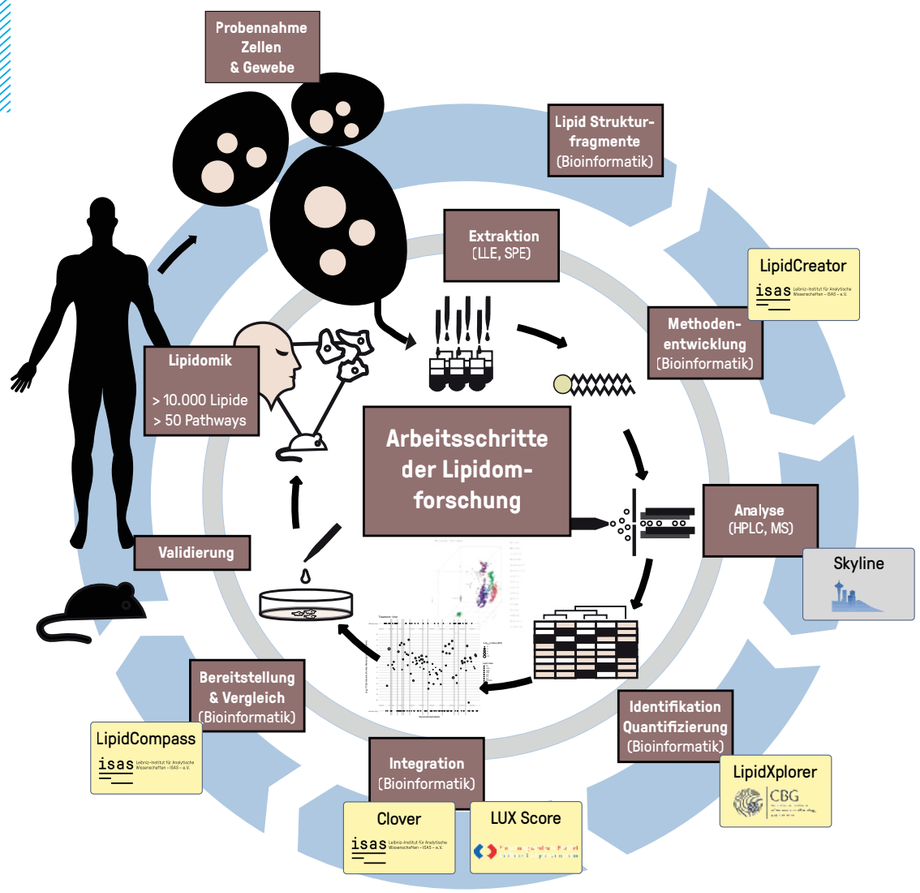
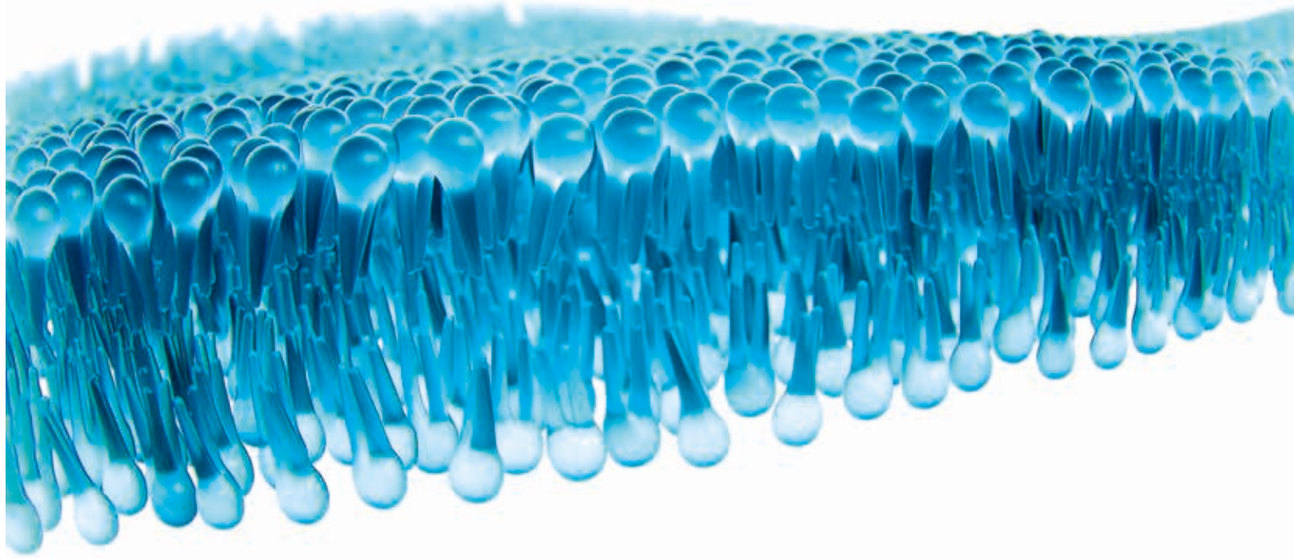


ABBILDUNG 1: Arbeitsschritte der Lipidomforschung. Die LIFS-Partner entwickeln für die einzelnen Schritte passend zugeschnittene Programme, wie z. B. LipidCreator, LipidXplorer, LUX Score, Clover und LipidCompass (gelb hinterlegt), um von der Probennahme bis zur Auswertung, Datenintegration und -visualisierung sowie der Bereitstellung der Daten einen reibungslosen Informationsfluss zu ermöglichen. Hierzu werden auch bereits etablierte Programme, wie z. B. Skyline (grau hinterlegt), spezifisch erweitert, hier z. B. durch das Skyline-Plugin LipidCreator, um diese in den Arbeitsablauf zu integrieren.



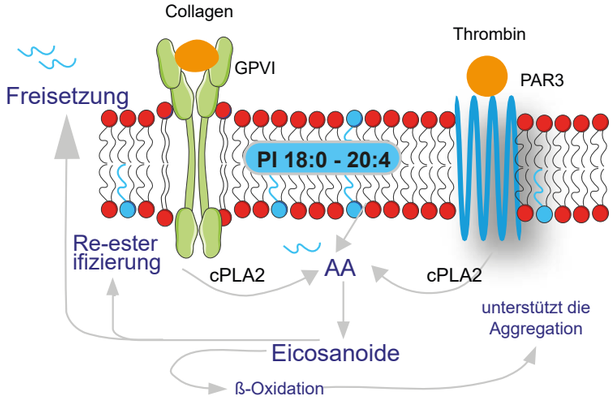


ABBILDUNG 2: Collagen- und Thrombininduzierte Generierung von Arachidonsäure während der Thrombozytenaktivierung. Die Abbildung zeigt zwei Wege auf, wie Lipidmediatoren aus dem Phospholipid PI 18:0 - 20:4 als Vorläufermolekül gebildet werden und anschließend in Lipidmediatoren umgewandelt oder verstoffwechselt werden. Abbildung angepasst aus [4].

REFERENZEN [1] Journal of Biotechnology 261, 131–136 (2017). DOI: 10.1016/j.jbiotec.2017.08.010. **[2]** PLoS ONE 7, e29851 (2012). DOI: 10.1371/journal.pone.0029851. **[3]** Bioinformatics 26, 966–968 (2010). DOI: 10.1093/bioinformatics/btq054. **[4]** PLOS Computational Biology 11, e1004511 (2015). DOI: 10.1371/journal.pcbi.1004511. **[5]** Blood, blood-2017-12-822890 (2018). DOI: 10.1182/blood-2017-12-822890. **[6]** Genome Res. 13, 2498–2504 (2003). DOI: 10.1101/gr.1239303.

AUTOREN Nils Hoffmann[1,5], Dominik Kopczynski[2,5], Fadi Al Machot[3], Dominik Schwudke[3], Jacobo Miranda Ackerman[4], Andrej Shevchenko[4], Robert Ahrends[1,5]
(weitere de.NBI-externe Mitarbeiter: Bing Peng[5], Cristina Coman[5], Canan Has[5])
[1] LIFS 1, [2]BioInfra.Prot 2, [3]LIFS 2, Forschungszentrum Borstel, Leibniz-Lungenzentrum, Borstel, [4]LIFS 3, Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Dresden, [5]Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund

DIE ERFORSCHUNG DES MENSCHLICHEN MIKROBIOMS

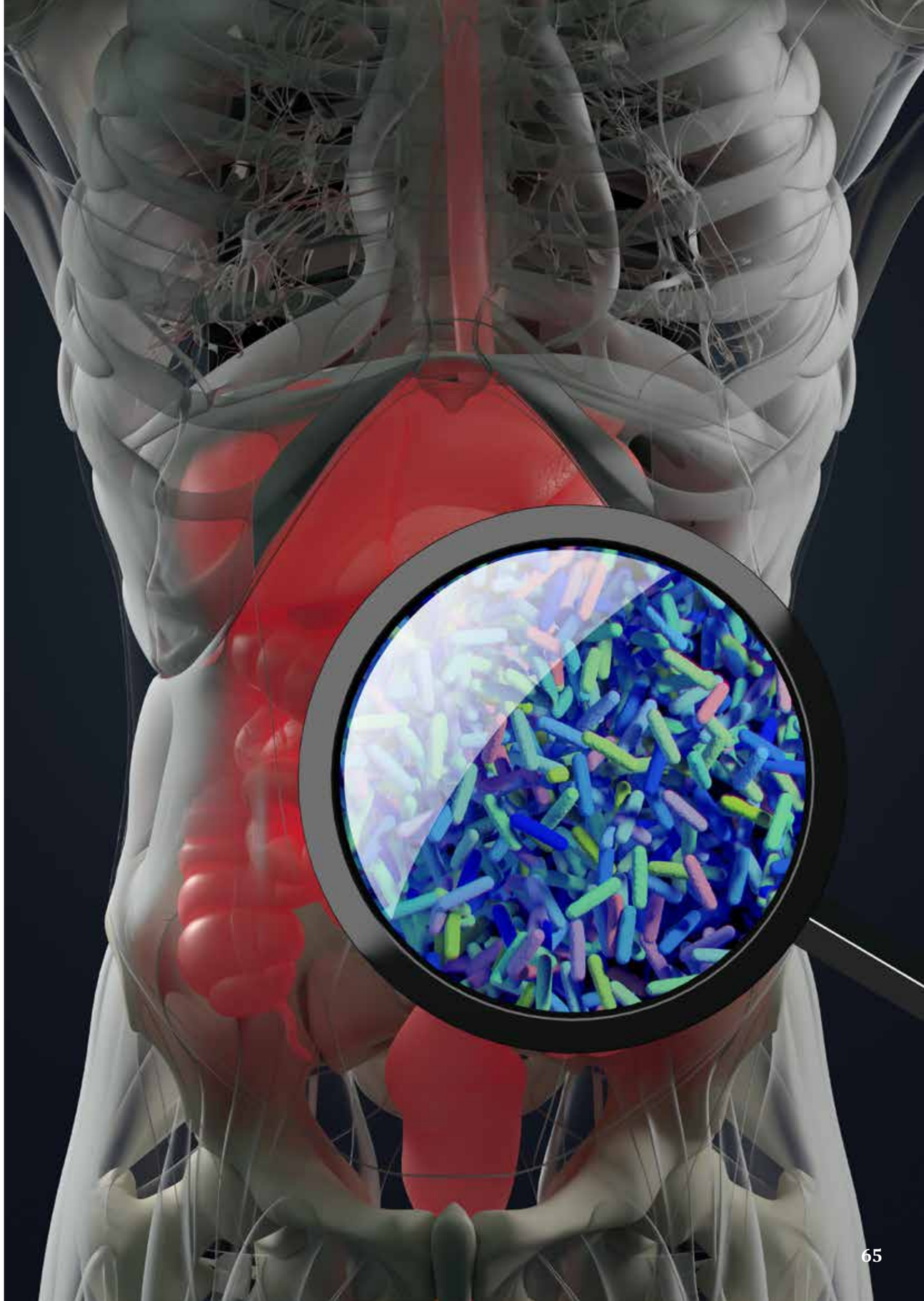
gibt Aufschluss über die Entstehung von Krankheiten und eröffnet neue Behandlungswege

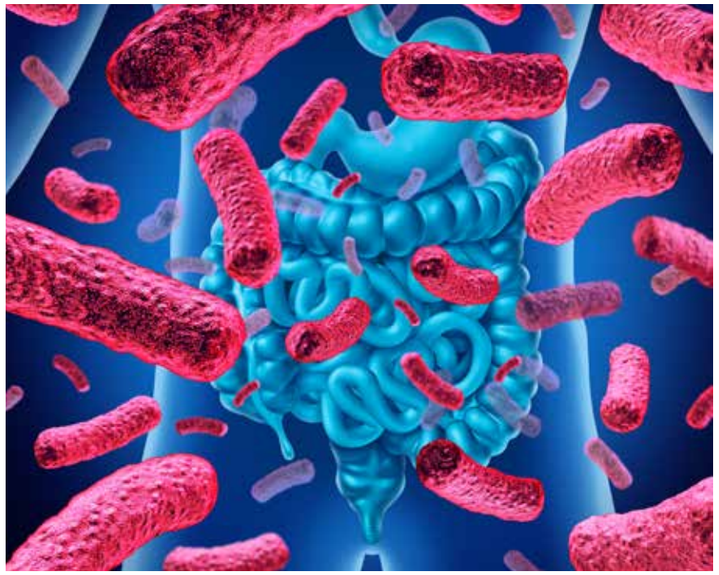
Die Mikrobiomforschung erkundet unsere mikrobiellen Mitbewohner und ihren Einfluss auf unsere Gesundheit. Dabei spielt die bioinformatische Datenanalyse eine entscheidende Rolle, um die Wechselwirkungen zwischen Mensch und Mikrobiom besser zu verstehen und daraus Biomarker – zum Beispiel für die Darmkrebsfrüherkennung – abzuleiten. Damit wird sie zukünftig zur Prävention von Krankheiten und zur Entwicklung neuer Therapiemöglichkeiten beitragen.

Die Forschung der letzten Jahre enthüllt zunehmend, in welchem Umfang die mikrobielle Besiedlung des Menschen unsere Gesundheit beeinflusst; charakteristische Veränderungen des Mikrobioms sind für eine Vielzahl von Krankheiten erkennbar. Beispielsweise wurden mikrobielle Biomarker zur Darmkrebsfrüherkennung identifiziert, die derzeit klinisch erprobt werden. Zudem haben Forscher, unter anderem des de.NBI-Netzwerks, begonnen, Wechselwirkungen zwischen Medikamenten und Darmbakterien systematisch zu untersuchen. Die Voraussetzungen dafür wurden durch Fortschritte in der Hochdurchsatz-Sequenzierung des mikrobiellen Erbguts (DNA), aber auch durch bioinformatische Entwicklungen zur Auswertung dieser Sequenzdaten geschaffen. Insbesondere den entscheidenden Beitrag der Bioinformatik möchten wir hier am Beispiel kürzlich veröffentlichter Studien verdeutlichen.

DAS MENSCHLICHE MIKROBIOM IST ARTENREICH UND SO INDIVIDUELL WIE EIN FINGERABDRUCK

Die mikrobiellen Mitbewohner des Menschen, ihre Gene und nicht zuletzt ihre Stoffwechselprodukte, die das Umweltmilieu entscheidend prägen, werden insgesamt als Mikrobiom bezeichnet. Neben anderen Mikroorganismen wie Hefen und Viren können mehr als 1.000 verschiedene Bakterien- und Archaeen-Arten unseren Darm besiedeln. Die Zusammensetzung dieser artenreichen mikrobiellen Gemeinschaft – früher auch als Darmflora bezeichnet – variiert von Mensch zu Mensch; selbst eineiige Zwillinge haben ein unterschiedliches Darmmikrobiom [1]. Noch gewaltiger ist die genetische Vielfalt des Mikrobioms – das Metagenom, welches die Gesamtheit aller mikrobiellen Gene bezeichnet, umfasst allein im Falle des Darmmikrobioms eine etwa 100-fach größere Zahl an Genen als das menschliche Genom.





DIE MIKROBIOMFORSCHUNG UNTERSUCHT EINZELNE BAKTERIELLE GENE SOWIE DAS ERBGUT GANZER MIKROBIELLER ÖKOSYSTEME – DAS METAGENOM

Diese Erkenntnisse verdanken wir vor allem den weitreichenden Entwicklungen neuer Sequenziertechnologien, die heute die Entschlüsselung der genetischen Information mit enormen Durchsatzraten ermöglichen. Mit der sogenannten Shotgun-Metagenomik können sogar alle Gene in allen Organismen, die in einer bestimmten Probe vorhanden sind, gleichzeitig ausgelesen werden. Dass dabei die Mikroorganismen nicht kultiviert werden müssen, ist ein entscheidender Vorteil, da viele Mikroorganismen unter Laborbedingungen nicht wachsen.

Allerdings stellt die Analyse von Metagenom-Sequenzdaten eine enorme bioinformatische Herausforderung dar. So sind beispielsweise die Kategorisierung der bakteriellen Vielfalt einer Probe (Abbildung 1) und die Bestimmung der Häufigkeiten einzelner Bakterienarten darin (taxonomische Bestimmung und Quantifikation) ein zentraler bioinformatischer Analyseschritt. Um dabei höchste Genauigkeit zu erreichen, haben Forscher am EMBL das mOTUs Computerprogramm entwickelt. Dessen Kernstück bildet eine umfassende Datenbank, die Gene aller Bakterien enthält, die bisher kultiviert und deren Genom entschlüsselt ist, aber zusätzlich auch solche Gene, die direkt aus Metagenomdaten gewonnen wurden. Deren genaue Einordnung in den bakteriellen Stammbaum erlaubt es der mOTUs-Software, auch die Häufigkeit von bisher nicht kultivierten Bakterien in Metagenomen zu bestimmen, was die Genauigkeit der bakteriellen Biodiversitätsanalysen gegenüber allen anderen bisher verfügbaren Analysewerkzeugen entscheidend verbessert.

Zusätzlich zu solchen Biodiversitätsanalysen können Forscher ein Metagenom auch daraufhin untersuchen, welche Stoffwechselwege den Mikroben zur Verfügung stehen, welche biochemischen Produkte sich daraus ergeben und welche Bedeutung diese möglicherweise für den Gesundheitszustand des menschlichen Organismus haben. Dass der mikrobielle Stoffwechsel in seiner enormen Vielfalt bisher nur sehr lückenhaft erforscht ist, erschwert solche Analysen und macht vielfach statistische Inferenzen und Extrapolationen nötig. Dabei dienen ebenfalls umfassende Datenbanken, die die bisher bekannte evolutionäre und funktionelle Vielfalt mikrobieller Gene und Stoffwechselwege kartieren, als Grundlage. Seit mehr als zehn Jahren pflegen und erweitern Forscher am EMBL eine solche Datenbank, EggNog genannt. Insbesondere die Unsicherheiten bezüglich der Richtigkeit und Vollständigkeit der Informationen in diesen Datenbanken wurden dabei gründlich untersucht und mit anderen Datenbanken verglichen. Auf dieser Grundlage wird mit großem zeitlichem und finanziellem Aufwand die Qualität der Datenbank durch manuelles Kuratieren stetig gesteigert.

DAS MENSCHLICHE MIKROBIOM TRÄGT ENTSCHEIDEND ZUR GESUNDHEIT BEI

Eingehende Analysen des mikrobiellen Stoffwechsels im menschlichen Darm haben dazu beigetragen, die Auffassung zu revidieren, dass Bakterien grundsätzlich Krankheiten verursachen. Ganz im Gegenteil zeigen etliche Studien, dass ein gesundes Darmmikrobiom zu unserem Wohlbefinden beiträgt. Solche gesundheitsfördernden Bakterien trainieren das Immunsystem, sie bilden einen hochwirksamen Schutz gegen das unkontrollierte Wachstum von Krankheitserregern und ihr Stoffwechsel liefert uns viele wichtige – teilweise essenzielle – Vitamine und Nährstoffe. Der mikrobielle Stoffwechsel ist so eng mit unserem körpereigenen verwoben, dass er selbst neurale Steuerungsprozesse und die zelluläre Regeneration beeinflusst [2] (Abbildung 2).

DIE ERFORSCHUNG DES MENSCHLICHEN MIKROBIOMS GIBT AUFSCHLUSS ÜBER
DIE ENTSTEHUNG VON KRANKHEITEN UND ERÖFFNET NEUE BEHANDLUNGSWEGE
HUMANE BIOINFORMATIK

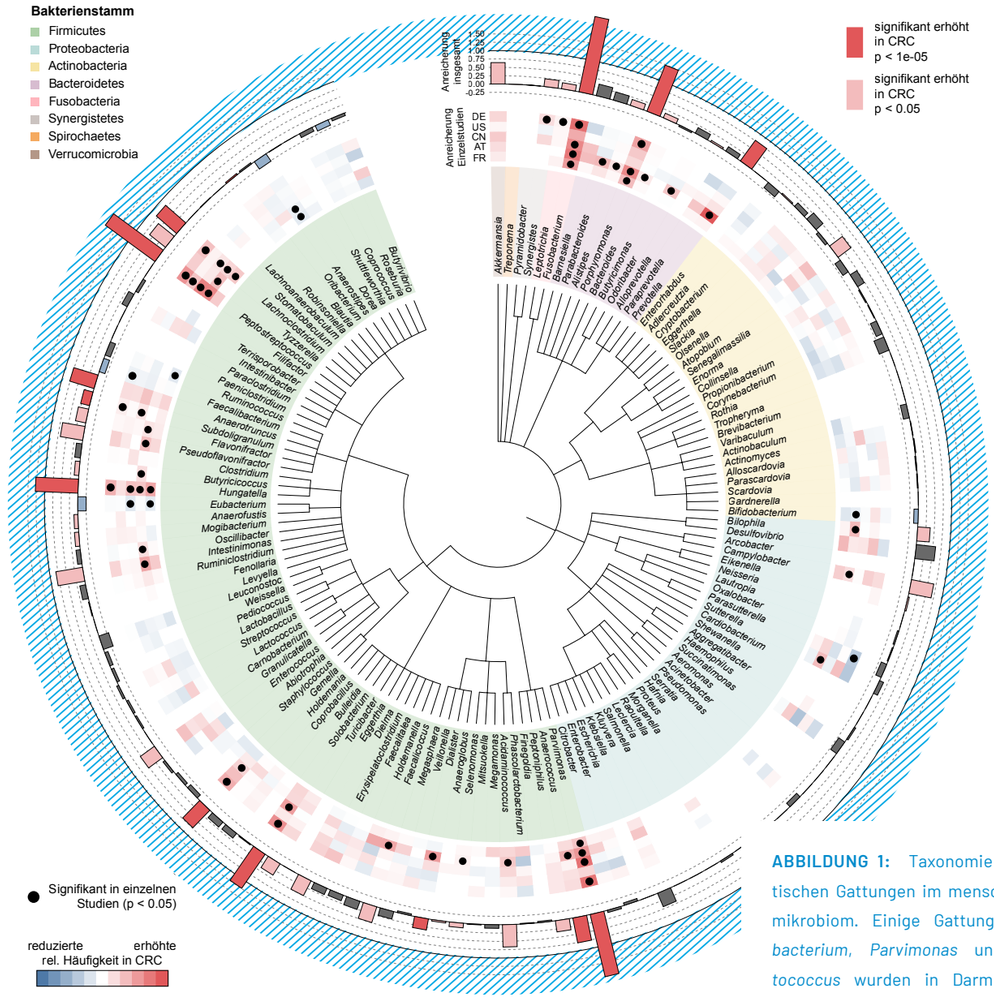


ABBILDUNG 1: Taxonomie der prokaryotischen Gattungen im menschlichen Darmmikrobiom. Einige Gattungen wie *Fusobacterium*, *Parvimonas* und *Peptostreptococcus* wurden in Darmkrebspatienten (*Colorectal cancer, CRC*) stark vermehrt vorgefunden und das relativ konsistent in Daten von mehreren Studien.

VERÄNDERUNGEN IM DARMMIKROBIOM GEHEN MIT VIELEN KRANKHEITEN EINHER

Trotz des vielfachen Nachweises der positiven Effekte des Darmmikrobioms auf die menschliche Gesundheit ist es bisher nicht gelungen zu definieren, was genau ein gesundes intestinales Mikrobiom ausmacht. Zu diesem Zweck wurden auch vergleichende Analysen, sogenannte Assoziationsstudien, durchgeführt, in denen das Mikrobiom von Patientengruppen mit gesunden Probanden verglichen wird. So lassen sich systematisch mit der untersuchten Krankheit assoziierte Mikrobiom-Veränderungen erkennen. Tatsächlich wurden unzählige Mikrobiom-Assoziationsstudien mit einer Vielzahl von Krankheiten in den letzten Jahren publiziert. Da diese sich oft auf geringe Patientenzahlen stützen, ist nicht in allen Fällen gesichert, dass die Ergebnisse auch reproduzierbar sind. Der statistischen Methodik für solche Studien kommt daher eine Schlüsselrolle zu. Ein ganzes Instrumentarium an solchen statistischen Werkzeugen, speziell für die Mikrobiomanalyse, haben Forscher des EMBL als Paket auf der R/Bioconductor-Plattform zur Verfügung gestellt. Dieses SIAMCAT genannte Softwarepaket ermöglicht die genaue statistische Auswertung von Mikrobiom-Assoziationsstudien unter Berücksichtigung möglicher anderer Einflüsse (technischer Art, aber auch bei unterschiedlicher Herkunft, Ernährungsweise etc. der Probanden), die ansonsten zu fälschlichen Krankheitsassoziationen führen könnten.

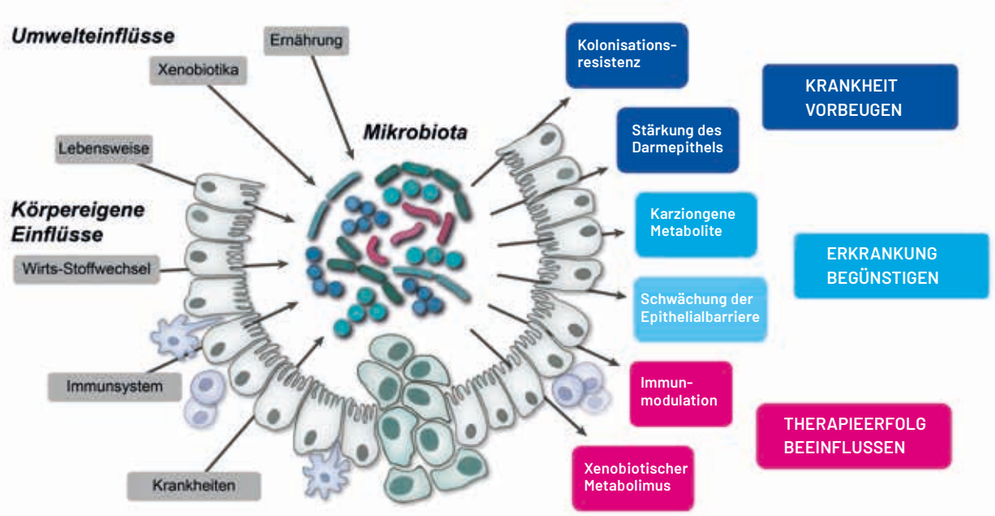


ABBILDUNG 2: Das Darmmikrobiom wird durch zahlreiche Umwelt- und Wirtsfaktoren beeinflusst. Diese Einflüsse können Veränderungen im Mikrobiom bewirken, die krankheitsfördernde Effekte oder Auswirkungen auf den Behandlungserfolg durch Medikamente haben können. Aufgrund seiner Individualität stellt das Darmmikrobiom also einen Individuum-spezifischen Risikofaktor in der Krankheitsentstehung und für therapeutische Komplikationen dar.

Die statistische Auswertung vieler Mikrobiom-Assoziationsstudien und weitergehender Untersuchungen an Tiermodellen hat bisher vor allem eines klargemacht: Es kommt auf die genaue Zusammensetzung des Mikrobioms an. Denn während die mikrobielle Artenvielfalt meist positiv ist, können einzelne Mikroben Krankheitsverläufe beschleunigen und die Medikamentenwirkung sowie Nebenwirkungen beeinflussen.

MIKROBIELLE BIOMARKERFORSCHUNG

Im Fall von Darmkrebs haben Forscher am EMBL in mehreren Veröffentlichungen gezeigt, dass sich anhand der Zusammensetzung des Darmmikrobioms Tumorpatienten von krebsfreien Probanden unterscheiden lassen. Ein kürzlich in der renommierten Fachzeitschrift „Nature Medicine" veröffentlichter Artikel veranschaulicht, wie die oben beschriebenen Konzepte und bioinformatischen Werkzeuge kombiniert werden können, um Veränderungen des Darmmikrobioms bei Darmkrebspatienten im Detail aufzuklären. Die Wissenschaftler des EMBL um Georg Zeller und ihre internationalen Forschungspartner beschreiben dort in einem studienübergreifenden Vergleich (Meta-Analyse) stark erhöhte Vorkommen von 29 Bakterienarten in den Darmkrebspatienten der acht untersuchten Studien [3] (Abbildung 1). Ihre Ergebnisse verdeutlichen, dass die Variabilität in der Zusammensetzung des menschlichen Darmmikrobioms nicht ausschließlich von externen Faktoren wie Ernährung und Lebensweise abhängt, sondern bestimmte Arten von Bakterien in Darmkrebspatienten grundsätzlich in größeren Mengen als in der gesunden Bevölkerung vorzufinden sind; diese sind also prinzipiell als mikrobielle Krebs-Biomarker global einsetzbar. Ein entsprechendes diagnostisches Verfahren zur Krebsfrüherkennung (nicht-invasives Darmkrebs-Screening) befindet sich aktuell in der klinischen Erprobung.

Ferner gibt eine detaillierte Analyse der mikrobiellen Genfunktionen in Darmkrebs-Metagenomdaten Aufschluss darüber, welche Stoffwechselprodukte in Krebspatienten angereichert sind. Dabei fanden die Forscher am EMBL heraus, dass die Stoffwechselwege zum Abbau von fett- und fleischhaltigen Nahrungsmitteln und zur Synthese krebserregender sekundärer Gallensalze in den Patientenproben in deutlich höheren und gleichzeitig die zum Abbau von pflanzlichen Kohlenhydraten aus Ballaststoffen in geringeren Mengen vorzufinden sind als in Proben von gesunden Menschen.

Diese Erkenntnisse über das Darmmikrobiom sind mit epidemiologischen Studien zu Ernährungsrisiken für die Darmkrebsentwicklung konsistent und könnten zukünftig zu verbesserten Ansätzen in der personalisierten Krebsprävention weiterentwickelt werden.

Entwicklung von nicht-invasiven und genauen Methoden zur Früherkennung von Darmkrebs.

AUSBLICK

Obwohl die Mikrobiomforschung erst am Anfang steht, macht sie große Hoffnungen auf die Verbesserung von Gesundheit und Wohlbefinden.

Das menschliche Mikrobiom ist ein hochaktuelles Thema, dem sich weltweit zahlreiche Forscher widmen. In den letzten Jahren wurde immer deutlicher, welch vielfältige Einflüsse die mikrobiellen Mitbewohner auf unseren Körper haben. Sie lenken das Immunsystem, wandeln Medikamente chemisch um oder steuern das Sättigungsgefühl. Um Nichtwissenschaftler an diesem Erkenntnisprozess teilhaben zu lassen, wurde von EMBL-Forschern um Peer Bork die Studie my.microbes initiiert, die durch eine große Probandenzahl aus der allgemeinen Bevölkerung zu einem besseren Verständnis der Wechselwirkung zwischen dem Menschen und seinem Mikrobiom beitragen soll [4].

Langfristig erhofft man sich, die Erkenntnisse über das Darmmikrobiom gezielt zur Krankheitsprävention und personalisierten Therapie einsetzen zu können. So wurde die Zusammensetzung des Darmmikrobioms bereits als wichtiger Faktor für den Erfolg von Immuntherapien bei Krebspatienten erkannt. Obwohl noch wenig bekannt ist, über welche molekularen Mechanismen die Mikrobiota eine Immunaktivierung erreicht, laufen bereits klinische Studien, die darauf abzielen, das Darmmikrobiom zu verändern, um Immuntherapien wirksamer zu machen [5].

Als weiterer Meilenstein in der Mikrobiomforschung gilt die Erkenntnis, dass nicht nur Antibiotika die nützliche mikrobielle Gemeinschaft in unserem Darm aus dem Gleichgewicht bringen können; andere Medikamente haben einen ähnlichen Effekt, wie eine Studie von EMBL-Wissenschaftlern um Peer Bork zeigt [6]. Demnach hemmt jedes vierte der mehr als 1.000 untersuchten Medikamente aus sämtlichen nicht antibiotischen Wirkstoffklassen das Wachstum unserer Darmbakterien – vom Entzündungshemmer bis zum Antipsychotikum.

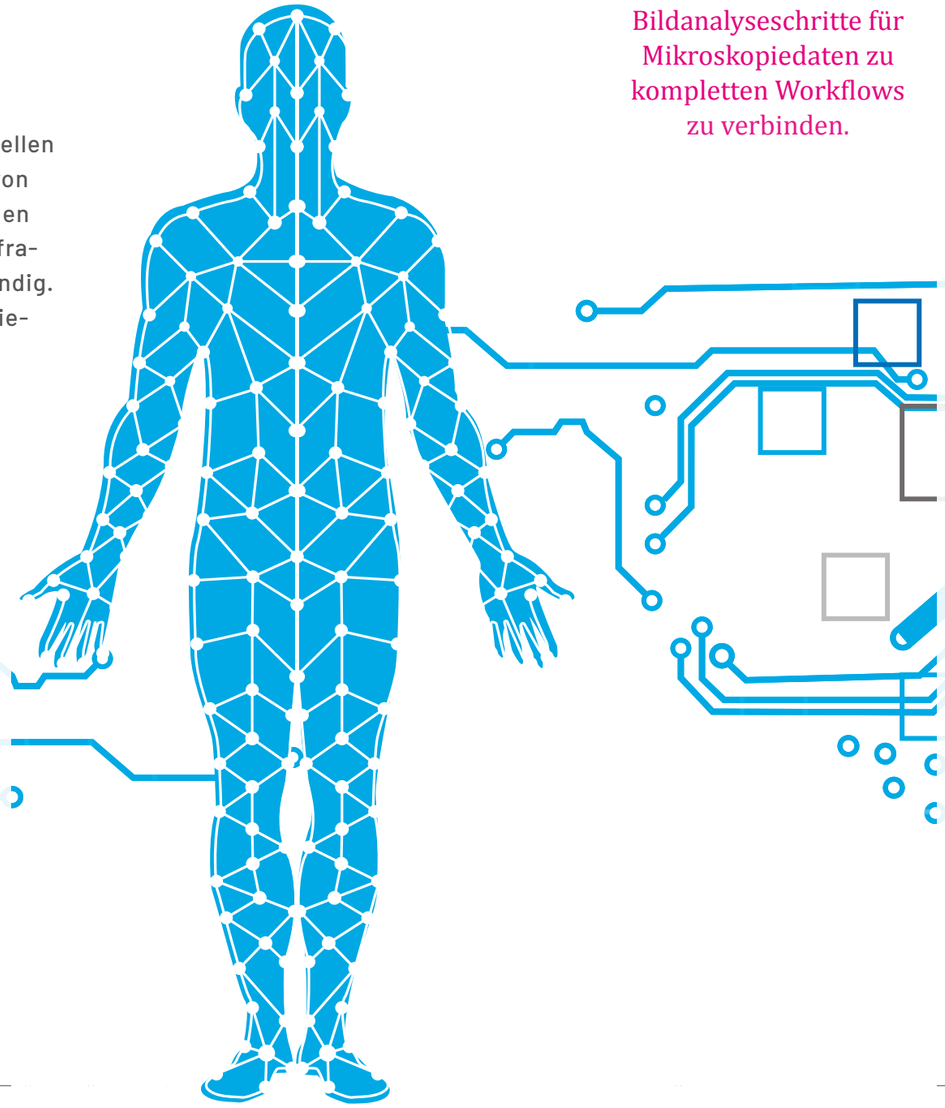
Mikrobiomforschung ist ein junges – und hochgradig interdisziplinäres – Forschungsgebiet, in dem die Bioinformatik eine Schlüsselrolle einnimmt. Das schnell und stetig wachsende Volumen und die Komplexität der Forschungsdaten erfordern immer leistungsfähigere bioinformatische Algorithmen und Softwareprogramme, die zunehmend Entwicklungen im Bereich der Künstlichen Intelligenz und des maschinellen Lernens aufgreifen. Dieses Potenzial für intelligente Analysen auszuschöpfen, verspricht weitere rapide Fortschritte bei der Entschlüsselung der komplexen Wechselwirkungen zwischen dem menschlichen Organismus und seinen mikrobiellen Bewohnern.

REFERENZEN [1] Dtsch. Med. Wochenschr. 142:267-274. DOI: 10.1055/s-0043-124940. **[2]** N Engl J Med. 2016 Dec 15;375(24):2369-2379. DOI: 10.1056/NEJMra1600266. **[3]** Nat Med. 2019 Apr;25(4):679-689. DOI: 10.1038/s41591-019-0406-6. **[4]** http://my.microbes.eu **[5]** Nat Med. 2019 Mar;25(3):377-388. DOI: 10.1038/s41591-019-0377-7. **[6]** Nature. 2018 Mar 29;555(7698):623-628.DOI: 10.1038/nature25979.

AUTOREN Ulrike Trojahn[1], Jakob Wirbel[1], Peer Bork[1], Georg Zeller[1]
[1] *European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg*

WAS UNS DIE EIGENSCHAFTEN MENSCHLICHER ZELLEN ÜBER KREBSERKRANKUNGEN VERRATEN

Das Erscheinungsbild – auch Phänotyp genannt – menschlicher Zellen gibt Aufschluss über deren Funktionsweise und die Entstehung von Krankheiten. Um die enormen Mengen an digitalen Daten von Zellen zuverlässig auswerten zu können, sind eine gut ausgebaute IT-Infrastruktur und fortschrittliche bioinformatische Werkzeuge notwendig. Dabei spielen insbesondere Hochdurchsatzverfahren, Mikroskopie-Bildgebung, computerbasierte Bildanalyse und biologische Datenbanken eine wichtige Rolle.



Der menschliche Körper setzt sich aus Billionen von Zellen zusammen, die unterschiedliche Gewebe im Körper bilden. Wenn Zellen von ihrer vorgegebenen Funktion abweichen, können Krankheiten entstehen. Ein Ziel der Forschung in den Lebenswissenschaften ist es, die Funktionsweise von Zellen zu verstehen, um mögliche Ansatzpunkte für die Behandlung von Krankheiten zu finden, die aufgrund von Fehlfunktionen von Zellen und deren Gene entstehen. Hierzu untersuchen Wissenschaftler die Zellen im Gewebe beispielsweise mittels Mikroskopie-Verfahren. Andere Ansätze nutzen molekularbiologische Technologien wie RNA-Interferenz oder die Genschere CRISPR/Cas9 im großen Maßstab, um durch die gezielte Hemmung von Genen deren Funktion zu entschlüsseln. Heutige Hochdurchsatz-Screening-Verfahren ermöglichen die Untersuchung einer Vielzahl an Zellen in kürzester Zeit. Dabei entstehen enorme Mengen an digitalen Daten (Big Data), die gespeichert, ausgewertet, in einen biologischen Kontext gebracht und langfristig zugänglich gemacht werden müssen. Dies stellt hohe Ansprüche an die IT-Infrastruktur und bioinformatische Werkzeuge (Abbildung 1).

VON DER AUTOMATISIERTEN MIKROSKOPIE BIS ZUR ERFASSUNG DER PHÄNOTYPEN

Für die Analyse großer Mengen an Bilddaten, wie sie etwa in der Hochdurchsatz-Bildgebung von humanen Zellen mittels automatisierter Mikroskopie erzeugt werden, stehen mittlerweile verschiedene bioinformatische Werkzeuge zur Verfügung, wie beispielsweise CellProfiler oder Bibliotheken für verschiedene Programmiersprachen wie R, Python oder Matlab. Zur effizienten Nutzung vieler dieser Programme sind spezielle bioinformatische Kenntnisse erforderlich. Die Software „Konstanz Information Miner" KNIME (www.knime.org) bietet dagegen einen einfachen und intuitiven Ansatz und wird in der Arbeitsgruppe von Dr. Holger Erfle an der Universität Heidelberg genutzt. Individuelle Prozessierungs- und Analyseschritte lassen sich damit grafisch zu kompletten Arbeitsabläufen, sogenannten Workflows, verbinden.

Die Software-Plattformen KNIME und Galaxy ermöglichen individuelle Bildanalyseschritte für Mikroskopiedaten zu kompletten Workflows zu verbinden.

Hierbei werden zunächst die Bilddaten aufgearbeitet und ihnen dann die dem jeweiligen Experiment zugehörigen übergeordneten Daten (Metadaten), wie beispielsweise Koordinaten, und die experimentspezifische Behandlung der Zellen zugeordnet. In den Bildern werden dann die einzelnen Zellen identifiziert und individuelle Eigenschaften wie etwa ihre Helligkeit, ihre Form oder ihre Struktur erfasst. Anhand dieser Werte erfolgt die Einteilung der Zellen in Kategorien entsprechend ihrer Erscheinungsbilder, den sogenannten Phänotypen. Das Auftreten bestimmter Phänotypen oder deren Änderungen lassen Rückschlüsse auf die Reaktion der Zellen auf die Auswirkung verschiedener Behandlungen zu, beispielsweise die Hemmung oder Hochregulierung einzelner Gene.

Der Vorteil dieser Workflows liegt darin, dass diese wiederverwendet und weitergegeben sowie durch Anpassung der Parameter auf unterschiedliche Bilddaten angewandt werden können. Weiterhin ist es möglich, einzelne Teile des Workflows zu gruppieren und diese hierdurch modular miteinander zu vernetzen. Zudem ergeben sich weitere vielfältige Möglichkeiten, wenn die automatische Bildanalyse in die mikroskopische Bilderfassung integriert wird, sodass im Sinne eines Rückkopplungsmechanismus repräsentative Zellen oder seltene Phänotypen gezielt aufgenommen werden oder die Auflösung von ausgewählten Bereichen erhöht werden kann. Dadurch werden zugleich der Zeitaufwand und das Datenvolumen im Vergleich zu Standard-Hochdurchsatzverfahren reduziert, bei denen zunächst alle Daten aufgenommen und danach ausgewertet werden.

Diese Bilderfassung mit mehreren Auflösungsstufen wurde beispielsweise zur Untersuchung von Telomeren, den Enden der Chromosomen, in Prostatakrebsgewebe eingesetzt. Telomere verkürzen sich bei jeder Zellteilung; Tumorzellen müssen diese also aktiv wieder verlängern können, um sich ungehindert weiter teilen und vermehren zu können. In einem Ansatz zur Untersuchung von Gewebeproben mehrerer Patienten nahm das Mikroskop zunächst eine Übersicht der Proben auf einem Objektträger auf (sogenanntes Gewebe-Mikroarray; englisch: Tissue Microarray). In diesen Aufnahmen wurden automatisch Zellkerne identifiziert, für die dann die Telomere mittels hochauflösender 3D-Mikroskopie aufgenommen und analysiert wurden (Abbildung 2). Hierdurch konnten die Forscher spezifische Informationen über die Verteilung und Größe der Telomere und damit über die Mechanismen der Telomer-Verlängerung gewinnen [1].

CLOUD-TECHNOLOGIEN ZUR WEBBASIERTEN ANALYSE VON MIKROSKOPIE-BILDERN

Mit den stetig wachsenden Datenmengen in der biomedizinischen Forschung gewinnt die automatische Analyse immer mehr an Bedeutung. Besonders große Datenmengen entstehen bei der Aufnahme von Mikroskopie-Bildern, deren Anzahl sehr hoch sein kann und deren

Größe mehrere Gigapixel umfassen kann. Dadurch steigen die Herausforderungen an die Rechenkapazitäten und die Methoden zur computerbasierten Auswertung der Bilddaten. Cloud-basierte Lösungen ermöglichen die Nutzung zentraler schneller Recheninfrastruktur. Durch Cloud-Technologien kann komplexe Recheninfrastruktur transparent zur Verfügung gestellt werden und Bilddaten müssen nicht mehr auf einzelnen Rechnern von Wissenschaftlern kopiert werden. Die effiziente und zuverlässige automatische Auswertung von Mikroskopie-Bilddaten hat das Potenzial die Identifikation von krankheitsrelevanten Biomarkern zu verbessern.

Um große Datenmengen von Mikroskopie-Bildern automatisch in der Cloud auszuwerten, hat die Arbeitsgruppe von PD Dr. Karl Rohr an der Universität Heidelberg die webbasierte Plattform Galaxy erweitert und das System Galaxy Image Analysis entwickelt [2]. Der Einsatz einer webbasierten Schnittstelle für die Cloud ermöglicht die Durchführung automatischer Analysen in der Cloud mittels eines Standard-Webbrowsers. Dies hat den Vorteil, dass Nutzer auf ihrem eigenen Computer keine Software installieren müssen. Zudem können zentral neue Bildanalysemethoden von Informatikern effizient an Biologen und Mediziner über die Plattform bereitgestellt werden. Dies beinhaltet beispielsweise Methoden der Bildsegmentierung und Bildregistrierung. Bildsegmentierung ist wichtig, um die Umrandung von wichtigen Objekten wie Zellen oder Gewebe zu erkennen. Bildregistrierung wird benötigt, um Objekte, die aus verschiedenen Blickwinkeln oder durch verschiedene Bildmodalitäten aufgenommen wurden, in Relation zu setzen (Abbildung 3). Gerade durch den Einsatz von maschinellen Lernverfahren bzw. Methoden der Künstlichen Intelligenz wie Deep Learning werden besonders gute Ergebnisse erzielt. Bei Deep Learning werden tiefe neuronale Netze, das heißt Netze aus künstlichen Neuronen mit einer großen Anzahl an Netzschichten, anhand von Beispielen trainiert. Das Training benötigt eine hohe Rechenkapazität der IT-Infrastruktur, die durch das verwendete Cloud-basierte System bereitgestellt wird. Galaxy wurde ursprünglich für die Analyse von Genomdaten entwickelt. Durch unsere Erweiterung können jetzt auch Bilddaten sowie Genomdaten und Bilddaten zusammen analysiert werden.

In einer interdisziplinären Kooperation wird Galaxy Image Analysis beispielsweise eingesetzt, um histologische Mikroskopie-Bilder und Massenspektrometrie-Daten (MALDI) kombiniert auszuwerten (Arbeitsgruppen O. Schilling, Universitätsklinikum Freiburg; B. Grüning, Universität Freiburg; K. Rohr/T. Wollmann, Universität Heidelberg). MALDI bietet die Möglichkeit, vergleichsweise effizient ein ortsaufgelöstes Massenspektrogramm von Gewebe aufzunehmen (Abbildung 3). Dadurch können präzisere Krebsdiagnosen routinemäßig durchgeführt werden. Für die automatisierte Auswertung der Bilder wurde eine computergestützte Methode (Workflow) entwickelt [3]. Dabei werden neue Bildsegmentierungs- und Bildregistrierungsverfahren kombiniert. Diese können von Biologen und Medizinern in ihre eigenen Workflows über die Galaxy-Plattform integriert werden. Galaxy Image Analysis wird für unterschiedliche Anwendungen bereitgestellt, insbesondere über die Galaxy Europe Platform (ELIXIR) und die de.NBI-Cloud.

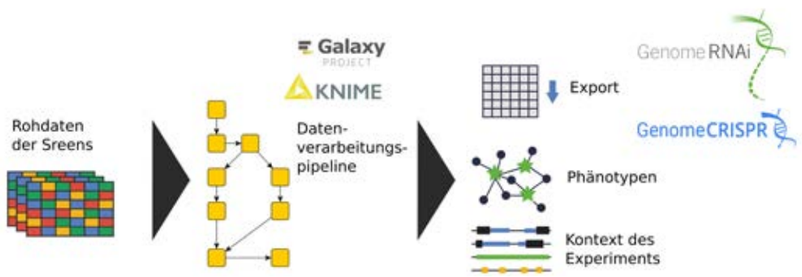


ABBILDUNG 1: Beispiel von Verarbeitungsschritten bei Hochdurchsatzverfahren von der Gewinnung der originalen Daten über die automatische Auswertung zur Identifikation von Phänotypen (Zelländerungen) und der langfristigen Speicherung bis hin zur Einordnung in den biologischen Kontext.

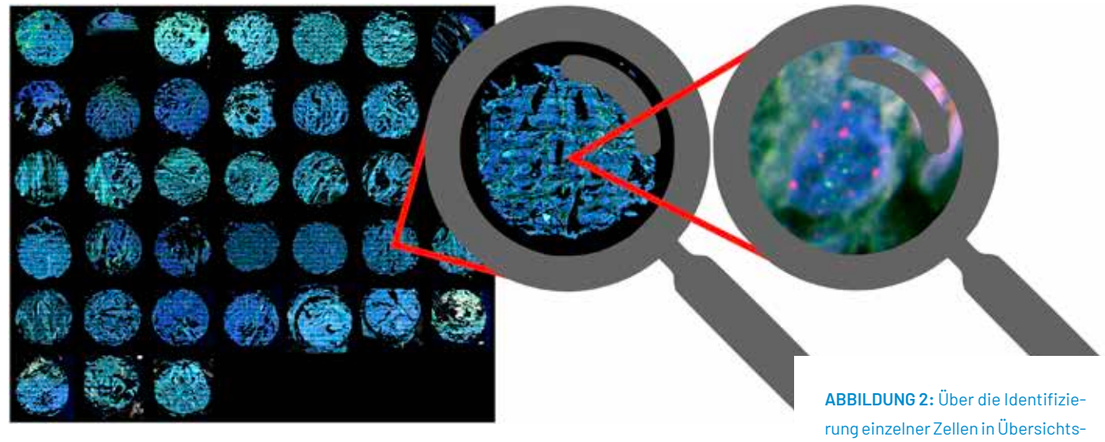
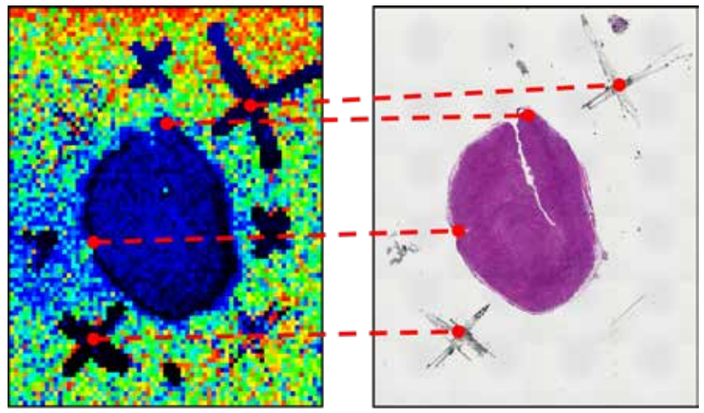


ABBILDUNG 2: Über die Identifizierung einzelner Zellen in Übersichtsbildern eines Präparates können verschiedene Vergrößerungsstufen und dadurch eine Multiskalen-Aufnahme realisiert werden.



MALDI-Bild Histologisches Bild

ABBILDUNG 3: MALDI- und histologisches Beispielbild mit eingezeichneten korrespondierenden Landmarken, die mittels Registrierung genutzt werden, um komplementäre Bildinformationen in Beziehung zu setzen.

Die beiden Datenbanken GenomeRNAi und GenomeCRISPR machen Daten aus großangelegten Hochdurchsatzexperimenten mit Millionen von Messungen strukturiert verfügbar.

MIT HILFE VON DATENBANKEN SCHALTPLÄNE VON GENEN ERSTELLEN

Mit Hilfe von Experimenten im Hochdurchsatzverfahren und Datenanalyse-Workflows können systematisch Messungen an Milliarden von Zellen durchgeführt werden. Um diese Datenmengen effizient auswerten und interpretieren zu können, ist eine leistungsfähige Dateninfrastruktur erforderlich. Um Daten aus diesen groß angelegten Hochdurchsatzexperimenten, bei denen Millionen von Messungen zugleich durchgeführt werden, strukturiert verfügbar zu machen, müssen diese in speziell konzipierten Datenbanken systematisch abgelegt werden. Zu diesem Zweck betreibt die Arbeitsgruppe von Prof. Michael Boutros am Deutschen Krebsforschungszentrum (DKFZ) und der Universität Heidelberg die beiden Datenbanken GenomeRNAi und GenomeCRISPR [4]. Diese Datenbanken beinhalten Ergebnisse aus Hunderten von Hochdurchsatzexperimenten, in denen mittels molekularbiologischer Methoden wie RNA-Interferenz (RNAi) oder CRISPR/Cas9 die Funktion von Genen gezielt beeinflusst wurde. Forscher aus Deutschland und der ganzen Welt können gezielt auf diese Daten zugreifen und diese zur Beantwortung biomedizinischer Fragestellungen heranziehen.

Die GenomeCRISPR-Datenbank beinhaltet beispielsweise Daten aus Experimenten, in denen mithilfe der Genschere CRISPR/Cas9 einzelne Gene systematisch in vielen verschiedenen Krebsarten ausgeschaltet wurden, woraufhin die Auswirkung des Genverlusts auf das Tumorwachstum gemessen wurde. Krebszellen sind für ihr Wachstum auf mutierte Gene angewiesen, die sich in gesunden Körperzellen nicht finden. Diese veränderten Gene ermöglichen dem Krebs, zu wachsen und sich auszubreiten. Da die Erkrankung auf diese Veränderungen angewiesen ist, nicht aber gesunde Zellen, stellen diese mutierten Gene interessante Angriffspunkte für neue Therapien dar. Oft können jedoch gerade diese aus technischen Gründen nicht angegriffen werden. Die GenomeCRISPR-Datenbank hilft den Wissenschaftlern beim Umgehen des Problems, indem sie aus den umfassenden Datensätzen Genschaltpläne erstellen und daraus weitere Angriffspunkte identifizieren können. Zum Beispiel reagieren Krebszellen oft empfindlich auf den Verlust von Genen, die sich in diesen Schaltplänen in der Nähe von den im Krebs veränderten Genen befinden.

Beispielsweise nutzten die Wissenschaftler in der Arbeitsgruppe von Prof. Boutros in einer aktuellen Untersuchung die Daten in GenomeCRISPR, um eine umfassende Karte für genetische Schaltpläne von Krebszellen zu erstellen (Abbildung 4). Dabei entdeckten sie, dass die beiden Gene GANAB und PRKCSH die Ausschüttung von sogenannten Wnt-Liganden kontrollieren [5]. Mithilfe dieser Signalmoleküle können sich benachbarte Krebszellen gegenseitig zum Wachstum anregen – ein Prozess, der vor allem bei Bauchspeicheldrüsen-, Darm- und Leberkrebs eine wichtige Rolle spielt. Diese Arbeiten zeigen, wie eine Vielzahl an genetischen Screens integriert und dadurch neue Erkenntnisse durch bioinformatische Analysen gewonnen werden können. Dabei handelt es sich um ein Verfahren, das dynamisch mit der Anzahl an Daten mitwachsen und weitere Aufschlüsse über die Funktion von Zellen geben wird.

REFERENZEN [1] Methods. 2017 Feb 1;114:60-73. DOI: 10.1016/j.ymeth.2016.09.014. **[2]** J Biotechnol. 2017 Nov 10;261:70-75. DOI: 10.1016/j.jbiotec.2017.07.019. **[3]** bioRxiv. 628719. DOI: https://doi.org/10.1101/628719. **[4]** Nucleic Acids Res. 2017 Jan 4;45(D1):D679-D686. DOI: 10.1093/nar/gkw997. **[5]** Mol Syst Biol. 2018 Feb 21;14(2):e7656. DOI: 10.15252/msb.20177656.

AUTOREN Manuel Gunkel[1], Thomas Wollmann[1], Benedikt Rauscher[1,2], Holger Erfle[1], Michael Boutros[1,2], Karl Rohr[1]
[1] *Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg*
[2] *Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg*

WAS UNS DIE EIGENSCHAFTEN MENSCHLICHER ZELLEN ÜBER KREBSERKRANKUNGEN VERRATEN
HUMANE BIOINFORMATIK

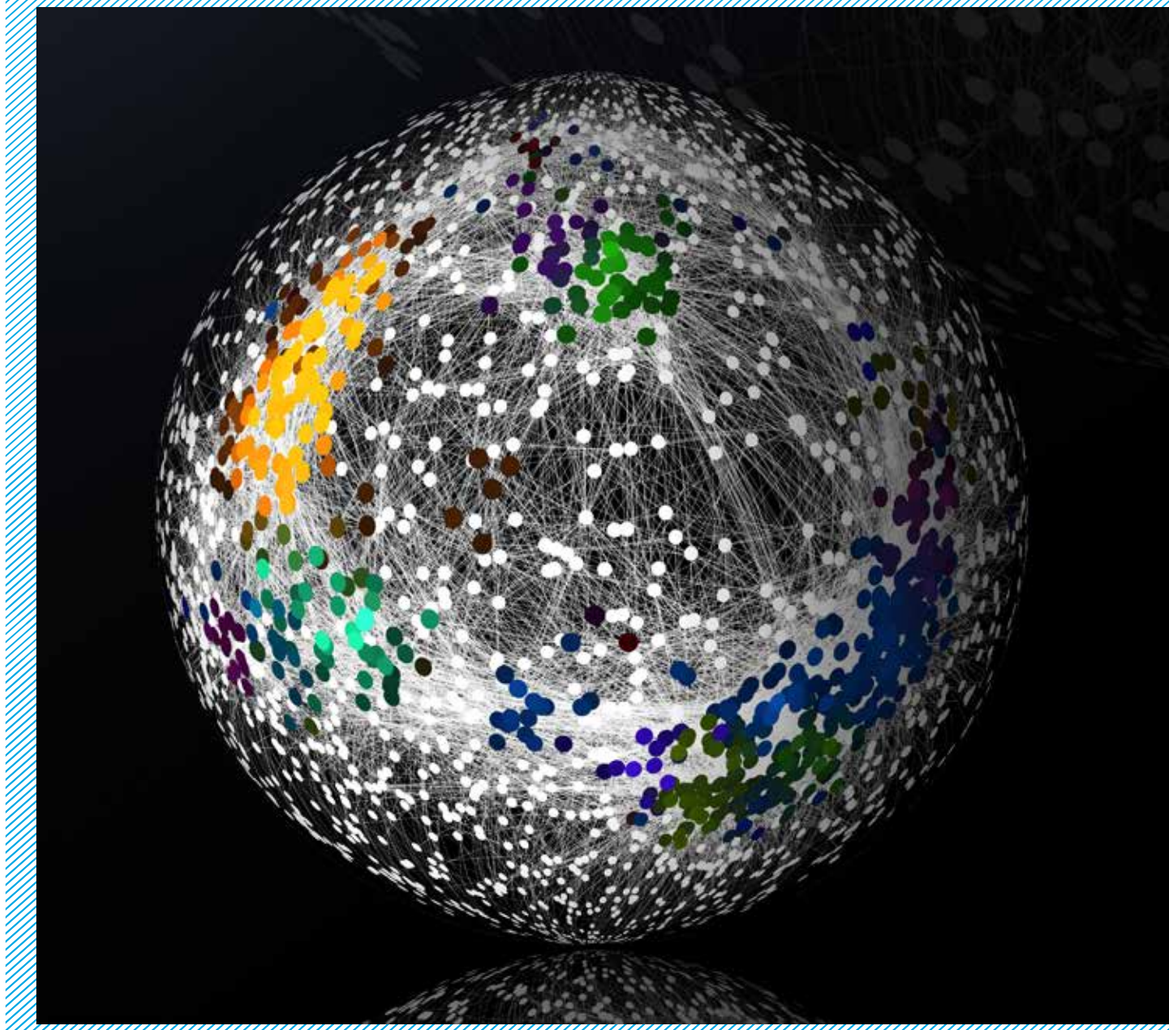
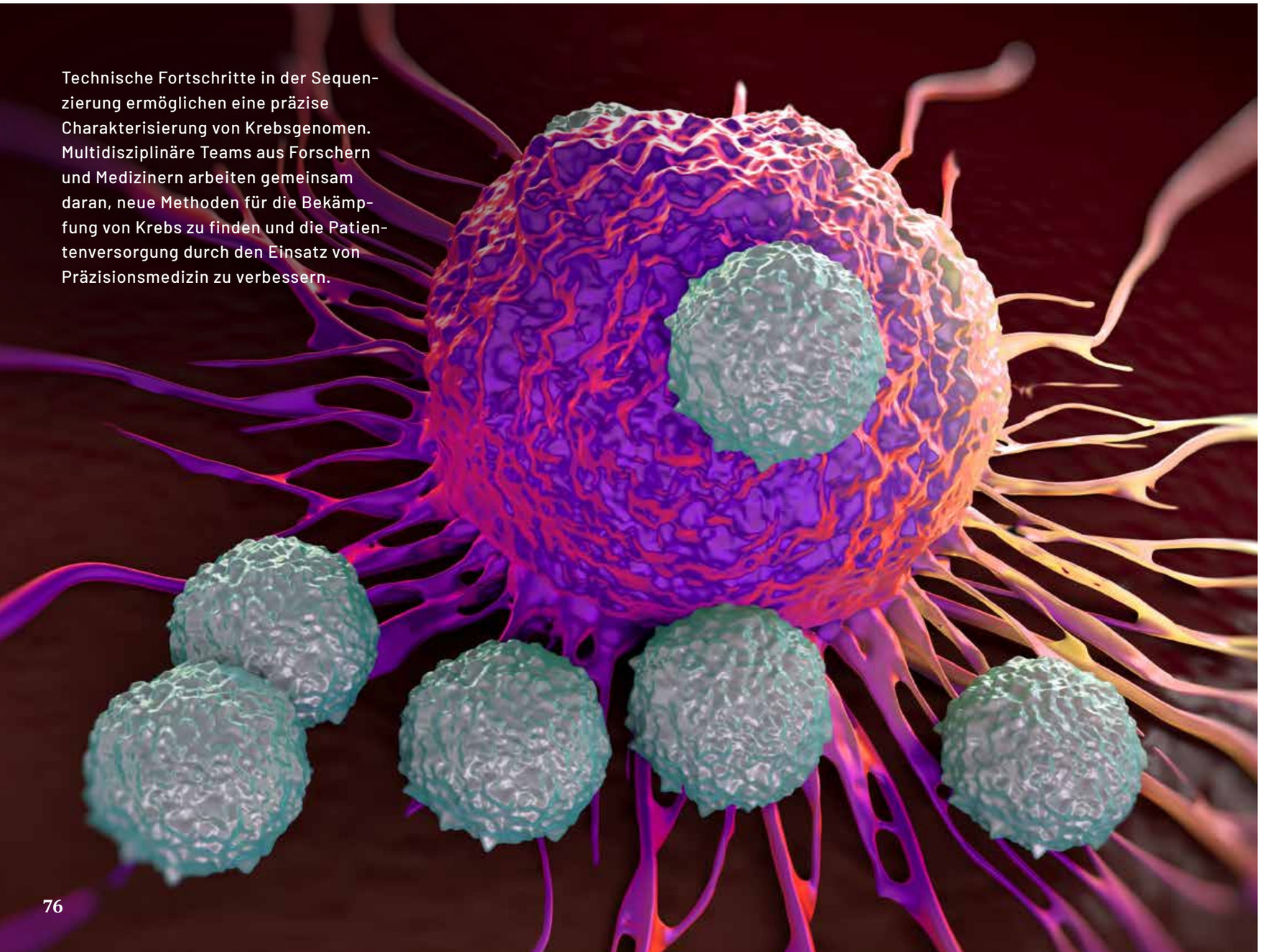


ABBILDUNG 4: Ein genetischer Schaltplan basierend auf der GenomeCRISPR-Datenbank. Jeder Punkt steht für ein Gen. Diese sind im Schaltplan miteinander verbunden, wenn sie funktionell miteinander verwandt sind – das heißt wenn sie für die Regulation derselben biologischen Prozesse verantwortlich sind. Einzelne Prozesse, die bei Krebserkrankungen eine besonders wichtige Rolle spielen, sind farblich hervorgehoben. So kann festgestellt werden, ob und inwieweit bisher unerforschte Gene bei der Krebserkrankung eine Rolle spielen.

PERSONALISIERTE MEDIZIN ZUR CHARAKTERISIERUNG VON TUMORERKRANKUNGEN



Technische Fortschritte in der Sequenzierung ermöglichen eine präzise Charakterisierung von Krebsgenomen. Multidisziplinäre Teams aus Forschern und Medizinern arbeiten gemeinsam daran, neue Methoden für die Bekämpfung von Krebs zu finden und die Patientenversorgung durch den Einsatz von Präzisionsmedizin zu verbessern.

PERSONALISIERTE MEDIZIN ZUR CHARAKTERISIERUNG VON TUMORERKRANKUNGEN
HUMANE BIOINFORMATIK

DIE DNA IST DER BAUPLAN DES MENSCHLICHEN LEBENS

A C G T – so heißen die vier DNA-Basen, die die Bausteine des Lebens, so wie wir es kennen, bilden. Seit über vier Milliarden Jahren sind es diese DNA-Sequenzen, die genetische Information kodieren, welche an unsere Nachkommen weitergegeben wird; sie erfüllen mehrere Grundfunktionen des Lebens: Wachstum und Reproduktion. Wir Menschen haben 3,2 Milliarden DNA-Basen, die auf 23 Chromosomenpaare verteilt sind. Hieraus bildet sich unser Genom. Unser Genom definiert die Vorlage von mehr als 20.000 Genen. Das Produkt jedes dieser Gene hat eine präzise Funktion, arbeitet in einem komplexen Netzwerk mit anderen Genprodukten zusammen und gemeinsam steuern sie jeden biologischen Prozess in unserem Körper. Unterschiede in unserer DNA führen zu der Vielfalt an Phänotypen, die wir um uns herum sehen. Wenn jedoch Teile unserer DNA beschädigt oder mutiert sind, kann dies negative Folgen haben und zu Krankheiten führen.

Als die Menschheit die Wichtigkeit der Entschlüsselung der menschlichen DNA-Sequenz erkannte, wurde das Humangenomprojekt ins Leben gerufen: Dabei handelte es sich um eine internationale Zusammenarbeit zwischen 20 verschiedenen Instituten, in denen über 13 Jahre hinweg (von 1990 bis 2003) das menschliche Genom sequenziert und zusammengesetzt wurde. Die Kosten für dieses Unterfangen waren immens – 2,7 Milliarden US-Dollar. Die Entschlüsselung der menschlichen Genomsequenz hat uns nicht nur geholfen, die menschliche Biologie, sondern auch die Entstehung vieler Krankheiten einschließlich Krebs besser zu verstehen.

DER ANFANG: BIG DATA AUS DER GENOMIK GIBT AUFSCHLUSS ÜBER KREBSTREIBER

In den letzten Jahren wurden neue Generationen an DNA-Sequenzierungstechnologien entwickelt, welche immer schneller, kostengünstiger und zugänglicher wurden. Mit den neuesten Technologien können wir ein menschliches Genom für weniger als 1.000 Euro in weniger als einer Woche sequenzieren. Diese bemerkenswerte Kosten- und Zeitersparnis bei der Sequenzierung eines menschlichen Genoms hat es den Forschern ermöglicht, die Ursachen für eine Vielzahl von Krankheiten zu erforschen, von denen die Krebsgenomik in den letzten Jahren ein Schwerpunkt war. Dies führte zu internationalen Bemühungen, welche das Ziel hatten zu verstehen, wie DNA-Mutationen die Entstehung von Krebs bei verschiedenen Krebsarten beeinflussen.

Die größten Konsortien in dieser Hinsicht sind das International Cancer Genome Consortium (ICGC) und The Cancer Genome Atlas (TCGA), die gemeinsam über 23.000 Patienten für mehr als 30 verschiedene Krebsarten sequenziert haben.

> Das ist Big Data! Es wurden über 23.000 Patienten für mehr als 30 verschiedene Krebsarten sequenziert.

Darüber hinaus wurde die Gründung des Heidelberger Instituts für Personalisierte Onkologie (HIPO) als Gemeinschaftsprojekt des Deutschen Krebsforschungszentrums (DKFZ), des Nationalen Centrums für Tumorerkrankungen (NCT) und der Universität Heidelberg unterstützt. HIPO hat bis heute fast 100 Projekte initiiert und über 3.000 Patientenproben analysiert. Diese Konsortien haben Kliniker und Forscher zusammengeführt, um die medizinischen und technischen Herausforderungen bei der Analyse dieser Daten anzugehen. Die wichtigsten Durchbrüche stammen hierbei von Big Data-Analysemethoden und multidisziplinären Teams, welche versuchen, diese Daten zu verstehen.

ABBILDUNG 1: Heidelberg NCT MASTER Ablauf. Die sechs Hauptschritte im Workflow sind: (1) Patienteneinwilligung und Registrierung, (2) Probenbewertung, Lagerung und Verarbeitung, (3) molekulare Profilerstellung und bioinformatische Analyse, (4) klinische Interpretation und Validierung molekularer Daten, (5) Diskussion der Ergebnisse in einem Molekularen Tumorboard (MTB) und (6) Behandlungs- und Therapieempfehlungen. Angelehnt an [6].

DIE AKTION: JEDER TUMOR IST ANDERS UND MUSS ENTSPRECHEND BEHANDELT WERDEN

Von Beginn der Krebsgenomik an wurde großer Wert darauf gelegt, dass die gewonnenen Erkenntnisse schnell Patienten, die an verschiedenen Tumorarten erkrankt sind, im Rahmen von translationalen Forschungsprojekten zunutze gemacht werden können. Die Genomik hat daher entscheidend zur Entwicklung der sogenannten Präzisionsmedizin bzw. Präzisionsonkologie beigetragen. Zusätzlich zu umfangreichen Sequenzierprogrammen wurden molekulare Tumorboards, in denen spezialisierte Ärzte verschiedener Fachrichtungen zusammen mit Bioinformatikern und anderen Naturwissenschaftlern über einzelne Fälle beraten, aufgebaut [2, 3]. Patienten, die bestimmte Einschlusskriterien erfüllen (außerordentlich junge Patienten; außerordentlich seltener Tumor; alle etablierten Therapien durchlaufen, ohne eine Heilung erzielt zu haben), kann diese moderne, aber sehr umfangreiche Diagnostik zur Verfügung gestellt werden. Die Erkenntnis, dass Treiber nicht nur spezifisch für bestimmte Tumorarten sind, hat zur Etablierung von personalisierten Sequenzierungsprogrammen wie NCT MASTER und INFORM in Heidelberg geführt [2, 3] (Abbildung 1). Diese Programme vereinen Logistik, Probenverarbeitung, Sequenzierung, Analyse und klinische Auswertung mit dem Ziel, innerhalb von vier bis sechs Wochen nach der Biopsie eine Therapieempfehlung zu erhalten. Biologen, Pathologen, Bioinformatiker und Ärzte koordinieren, analysieren und interpretieren gemeinsam die Genomsequenzierung von Krebspatienten, die auf die Standardtherapie nicht angesprochen haben. Die Ergebnisse werden in einem molekularen Tumorboard mit Experten aus verschiedenen Fachrichtungen diskutiert und sodann eine Therapieempfehlung ausgesprochen (Abbildung 2). Durch diesen personalisierten Ansatz wird in 75 % der Fälle mindestens eine Mutation identifiziert, die zur Steuerung der weiteren Therapie herangezogen werden kann. Zwei Drittel davon werden durch klinische Evidenz gestützt und in mehr als 35 % der Fälle wird die empfohlene Therapie umgesetzt.

Der Erfolg der Programme zeigt sich in zahlreichen individuellen Heilversuchen, dem Einsatz von Medikamenten, die für andere Tumoren zugelassen sind, und der Durchführung von neuen Krebstherapien wie der Immuntherapie. Zusammenfassend lässt sich sagen, dass die Nutzung von Hochdurchsatzverfahren in Kombination mit einem Spezialistenteam einen wesentlichen diagnostischen, therapeutischen und prognostischen Zusatznutzen für die Patienten bringt.

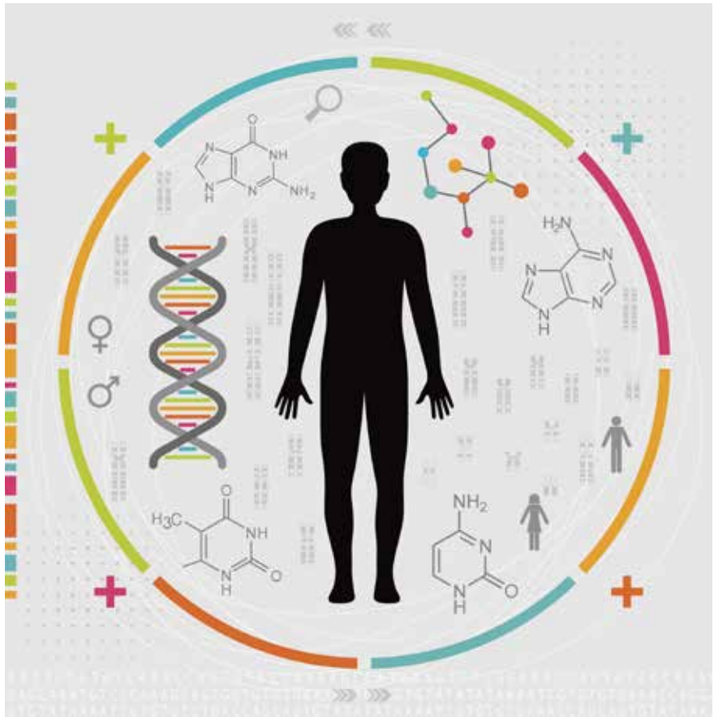
DIE ERKENNTNIS: MOLEKULARE ERKENNTNISSE FÜHREN ZU NEUEN THERAPIEMÖGLICHKEITEN – „BIOMARKERN“

Unabhängig vom Ursprungsgewebe des sequenzierten Tumors wird bei dieser Form der Diagnostik nach bestimmten behandelbaren Konstellationen gesucht. Die Behandelbarkeit kann entweder aus dem Vorliegen bestimmter Mutationen in bestimmten Genen folgen (targetable lesions) oder aus allgemeineren kombinierten Merkmalen. Ein diagnostisches Merkmal, das zu einer therapierelevanten Konsequenz führt, wird „Biomarker“ genannt.

Eine sehr häufige Gruppe von targetable lesions in Tumoren verschiedener Entitäten sind aktivierende Mutationen in Tyrosinkinase-Rezeptoren oder ihnen nachgeschalteten Signalkaskaden. Diese Rezeptoren regeln im gesunden Gewebe die Kommunikation einer Zelle mit ihrer Umgebung und die Reaktion der Zelle auf äußere Stimuli. Eine konstitutive Überaktivierung eines Signalwegs, der beispielsweise die Zellproliferation anregt, kann zur Entstehung eines Tumors führen.

Ein Beispiel für einen kombinierten Biomarker stellt die Detektion von bestimmten DNA-Reparaturdefekten dar. Jede Zelle hat mehrere molekulare Mechanismen, um Mutationen zu erkennen und zu reparieren. Ein sehr effizienter Mechanismus zur fehlerfreien Korrektur von Mutationen ist die homologe Rekombination. Fällt diese in einer Tumorzelle aus, weil zum Beispiel eines der an diesem Reparaturweg beteiligten Gene selber mutiert ist, so akkumuliert die betroffene Zelle über die Zeit immer mehr Mutationen. Diese können zumindest teilweise durch andere, noch intakte DNA-Reparaturmechanismen korrigiert werden. Gibt man dem betroffenen Patienten nun jedoch ein Medikament, das einen weiteren Reparaturmechanismus hemmt, so ist die Gesamtreparaturkapazität der Krebszelle unter Umständen erschöpft, diejenige der gesunden Zellen desselben Patienten hingegen nicht, denn in diesen Zellen funktioniert die homologe Rekombination ja noch. Eine solche Konstellation, bestehend aus einer Mutation oder einem Biomarker und der Wirksamkeit eines Medikaments, nennt man synthetische Letalität. Um diese ausnutzen zu können, bedarf es einer präzisen Erkennung des zugrunde liegenden Merkmals, beispielsweise des Defekts in der homologen Rekombination. Allerdings gibt es Konstellationen, in denen die ursächliche Mutation für den Defekt nicht gefunden wird. Dann kann es wichtig sein, mittels Mustererkennung den Abdruck dieses DNA-Reparaturdefekts auf das Genom zu erkennen. Im Falle der homologen Rekombination wurden dazu verschiedene Methoden und Maßzahlen entwickelt (Mutationssignaturen, HRD-Score oder auch kombinierte Maße wie der der TOP-ART-Studie zugrunde liegende Score [4]).

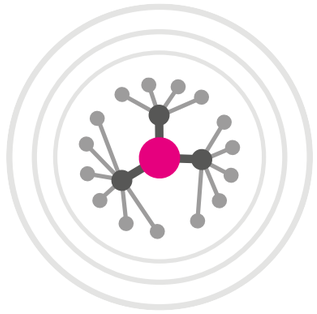
Ein anderes Beispiel für einen kombinierten Biomarker ist, die Gesamtzahl der Mutationen in einer Probe zu erfassen, insbesondere die Gesamtzahl in kodierenden Bereichen des Genoms (nur 2 % des menschlichen Genoms kodieren direkt Gene). Jede nicht synonyme Mutation kann eine Veränderung nicht nur im entsprechenden Protein hervorrufen, sondern auch in zerkleinerten Stücken dieses Proteins (Peptiden), die von den Zellen zur Erkennung durch das Immunsystem auf ihrer Oberfläche präsentiert werden (Neo-Epitope). Präsentiert eine Zelle ein Peptid, welches eine Mutation enthält, so kann sie vom Immunsystem als Tumorzelle erkannt werden und unter Umständen von T-Zellen getötet und abgeräumt werden. Dies ist die körpereigene Tumorabwehr, die sehr effizient funktioniert und jeden Tag in jedem Menschen ca. 6.000 neu maligne entartete Zellen vernichtet. Manche Tumoren besitzen allerdings die Fähigkeit, T-Zellen in ihrer Umgebung durch bestimmte Signale schwach und träge zu machen. Eine neue Klasse von Medikamenten, die sogenannten Immune Checkpoint Inhibitors (ICI), hemmen diese schwächende Signalkaskade und führen damit zu einer Reaktivierung der zytotoxischen T-Zellen. Die oben genannte Gesamtzahl der Mutationen im kodierenden Bereich, die die Anzahl der Neo-Epitope bestimmt, ist ein prädiktiver Wert für die Wirksamkeit einer Therapie mit ICI.



DIE VORAUSSETZUNG: GROSSE DATENMENGEN BENÖTIGEN EINE GROSSE INFRASTRUKTUR

Die Auswertung dieser Daten benötigt nicht nur eine große Expertise, sondern auch eine große, spezialisierte Infrastruktur. So werden allein zum Speichern der Daten eines Krebspatienten, bei dem der Tumor und das Blut sequenziert wurden, 500 Gigabyte Speicher zur Verfügung gestellt. Bei mehreren Tausend Patienten sind das schnell mehrere Petabyte (1 Petabyte = 1.000.000 Gigabyte). Diese Daten müssen zudem in komplexen und rechenintensiven Schritten analysiert werden. Um auf den Daten Forschung zu betreiben, müssen diese häufig mit großen Datensätzen wie denen des ICGC oder TCGA verglichen werden.

Leider haben nur wenige Institute die Kapazität, um mehrere Petabyte an Daten herunterzuladen, zu speichern und zu analysieren. Um dies zu umgehen, wird zunehmend versucht, die Datensätze in Clouds zu speichern. In diesen Clouds können Wissenschaftler, nachdem sie nachgewiesen haben, dass sie berechtigt sind, auf den Daten zu arbeiten, die Daten analysieren. In de.NBI wird zum Beispiel derzeit ein Spiegel der ICGC-Daten etabliert. Neben einer erhöhten Effizienz erlaubt diese Form der Datenanalyse auch allen Wissenschaftlern aller Institute, an den großen Datensätzen zu arbeiten. Das Teilen von Ressourcen setzt die in diesem Bereich der Wissenschaft dringend benötigten Kräfte frei.



DAS FAZIT: DIE ZUKUNFT DER PERSONALISIERTEN ONKOLOGIE

Getrieben vom Erfolg der ersten Präzisionsonkologieprogramme, nimmt deren Zahl immer weiter zu. Zum einen starten immer mehr Universitätskliniken und Zentren eigene Programme, zum anderen werden vorhandene Programme ausgeweitet, zum Beispiel durch die Gründung eines zweiten NCT in Dresden. Um an allen Standorten die gleiche Qualität zu gewährleisten, bedarf es standardisierter, leicht zu teilender Analyseverfahren. Es hat sich gezeigt, dass dafür das Installieren gemeinsamer Software in Clouds am effizientesten ist.

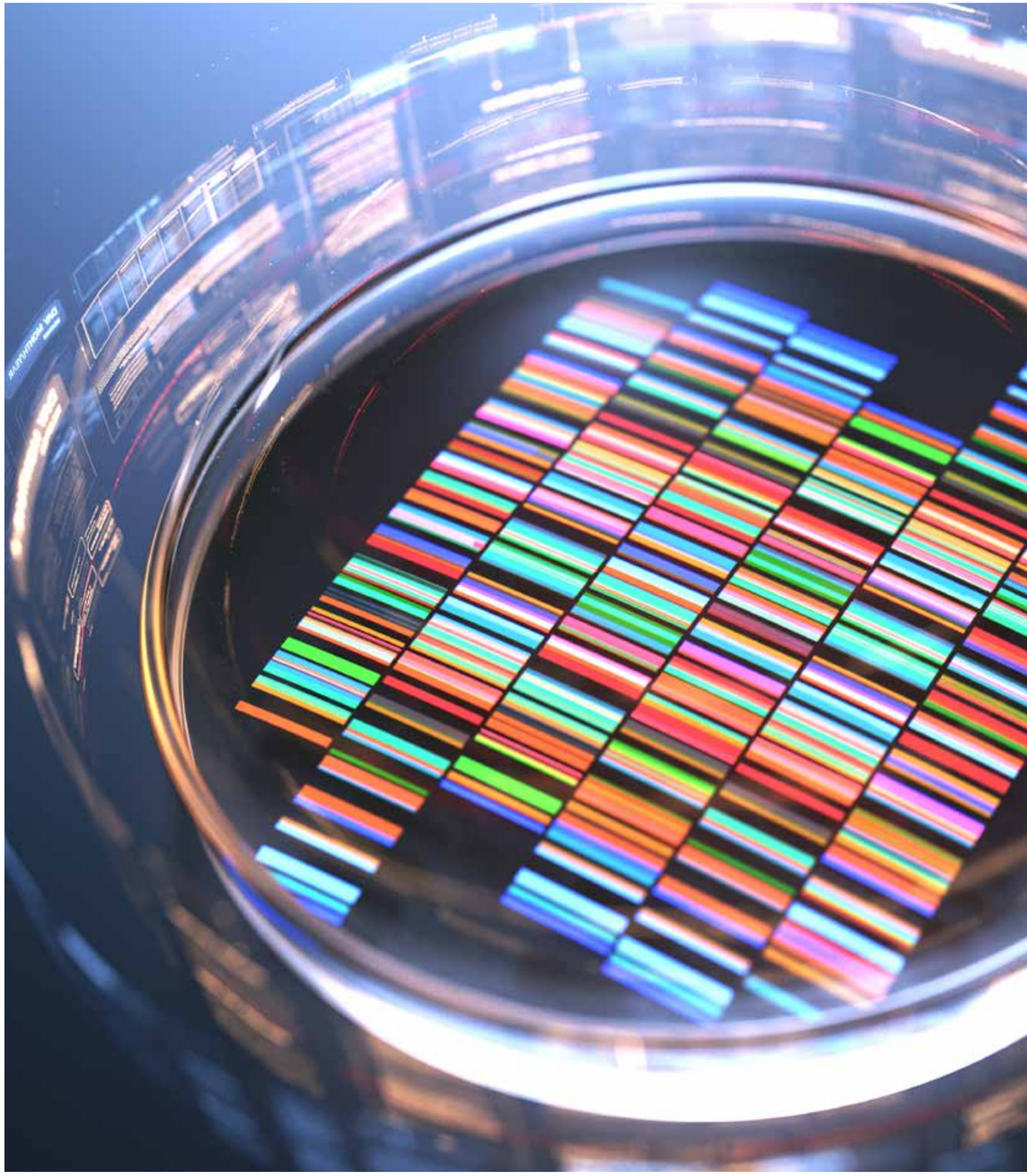
Obwohl die Präzisionsonkologie bereits viel erreicht und viele neue Erkenntnisse auch für seltene Krebsarten gewonnen hat, macht sie schnell große Fortschritte. Weltweit arbeiten unzählige Ärzte und Wissenschaftler daran, um das Motto des DKFZ Realität werden zu lassen: *Forschen für ein Leben ohne Krebs.*

REFERENZEN [1] Nat Com 2019;10(1):368. DOI:10.1038/s41467-018-08069-x. **[2]** Int J Cancer 2017;141(5):877-886. DOI: 10.1002/ijc.30828. **[3]** Eur J Cancer 2016;65:91-101. DOI: 10.1016/j.ejca.2016.06.009. **[4]** https://www.nct-heidelberg.de/das-nct/newsroom/aktuelles/details/top-art-studie-den-krebszellen-gezielt-das-reparaturwerkzeug-wegnehmen.html **[5]** Nature 2018;555: 469–474. DOI: 10.1038/nature26000. **[6]** https://www.nct-heidelberg.de/fileadmin/media/nct-heidelberg/forschung/nct%20master/nct_HD_master_k6.pdf

AUTOREN *Naveed Ishaque [1], Ivo Buchhalter [2], Daniel Hübschmann [2,3,5], Barbara Hutter [2], Franziska Mueller [1], Matthias Bieg [1], Nina Haberman [4], Jan Korbel [4], Benedikt Brors [2], Stefan Fröhling [2,3], Roland Eils [1,6]*

[1] Berliner Institut für Gesundheitsforschung (BIH) und Charité – Universitätsmedizin Berlin
[2] Deutsches Krebsforschungszentrum (DKFZ), Heidelberg
[3] Nationales Centrum für Tumorerkrankungen (NCT), Heidelberg
[4] The European Molecular Biology Laboratory (EMBL), Heidelberg
[5] Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM), Heidelberg
[6] Medizinischen Fakultät und Universitätsklinikum Heidelberg, Heidelberg

PERSONALISIERTE MEDIZIN ZUR CHARAKTERISIERUNG VON TUMORERKRANKUNGEN
HUMANE BIOINFORMATIK



ANALYSE DER GENREGULATION MENSCHLICHER ZELLEN MITTELS MASCHINELLEM LERNEN

Maschinelle Lernverfahren und insbesondere Deep Learning haben sich in den letzten Jahren von enormer Bedeutung erwiesen, um neues Wissen über genregulatorische Mechanismen zu erlangen. Wir stellen ein neues Softwarepaket namens Janggu zur Verfügung, welches die Etablierung von Deep-Learning-Anwendungen mit genomischen Daten unterstützt. Janggu reduziert den softwareentwicklerischen Aufwand und ermöglicht, biologische Fragen effizienter beantworten zu können.

VON GENEN ZU ZELLEN

Jede Zelle in unserem Körper enthält das gesamte Erbgut in Form der DNA-Sequenz, welche in verschiedene Abschnitte unterteilt ist. Die wohl bekanntesten dieser Abschnitte werden als Gene bezeichnet. Die meisten Gene beschreiben Bauanleitungen für Proteine, welche wiederum als molekulare Werkzeuge für die Umsetzung der biochemischen Prozesse in unserem Körper notwendig sind.

Zwar beinhaltet jede Zelle die gesamte DNA-Sequenz, aber Leberzellen haben andere Aufgaben zu erfüllen als beispielsweise Muskel- oder Nervenzellen. Dies funktioniert dadurch, dass in unterschiedlichen Zelltypen unterschiedliche Gene aktiv, also exprimiert sind. So gibt es etwa leber- oder muskelspezifische Gene, welche ausschließlich in den jeweiligen Zellen aktiv abgelesen werden und in die entsprechenden Proteine umgewandelt werden. Dieser Prozess erfordert einen hohen Grad an Koordination, welche als Genregulation bezeichnet wird. Obwohl uns die menschliche DNA-Sequenz bereits seit Beginn der Jahrtausendwende in ihrer gesamten Länge bekannt ist, sind die Regulation vieler Gene und die damit verbundenen Prozesse noch längst nicht bis ins Detail verstanden. Das liegt unter anderem daran, dass Genregulation ein koordinierter und hochkomplexer Prozess ist, der durch die Verpackung der DNA, epigenetische Modifikationen und das Binden von Proteinen an die DNA-Sequenz gesteuert wird.

BIOTECHNOLOGISCHE MESSVERFAHREN HELFEN DIE REGULATION VON GENEN ZU VERSTEHEN

In den letzten Jahren haben biotechnologische Fortschritte, insbesondere die Hochdurchsatzsequenzierung, zu neuen Erkenntnissen über die Genregulation beigetragen. Diese Verfahren erlauben es, Millionen kurzer DNA- oder RNA-Sequenzstücke zu erfassen, welche direkt oder indirekt das Resultat genregulatorischer Aktivitäten sind und dadurch Rückschlüsse auf die Genregulation ermöglichen. Beispiele solcher Hochdurchsatzprotokolle sind etwa: ChIP-seq, womit von Proteinen gebundene Regionen in der DNA identifiziert oder epigenetische Modifikationen detektiert werden; RNA-seq, womit die Genexpression quantifiziert wird; und ATAC-seq, womit offen zugängliche von fest verpackten DNA-Regionen unterschieden werden. Die Messung dieser Vorgänge hat allerdings zu einer Explosion des Datenvolumens geführt. Datenmengen von Hunderten von Gigabyte als Resultat eines einzelnen Experimentes sind heute keine Seltenheit mehr. Die Durchführung solcher Experimente unter unterschiedlichen Bedingungen, wie etwa für unterschiedliche Zelltypen, Spezies oder Krankheiten, trägt überdies multiplikativ zum Datenwachstum in der Genomik bei.

Seit wenigen Jahren können solche Messungen sogar in einzelnen Zellen erfasst werden. Diese ermöglichen eine bislang unübertroffene Auflösung der zellbiologischen und entwicklungsbiologischen Vorgänge. In der sogenannten Einzelzell-RNA-Sequenzierung wurden so beispielsweise bereits Genexpressionsprofile für über zwei Millionen Zellen in einer einzelnen Studie berichtet [1].

MASCHINELLES LERNEN ALS WERKZEUG ZUR ANALYSE VON GENOMIK-DATEN

Die großen Datenmengen aus der Genomik können von Hand nicht mehr analysiert und interpretiert werden. Dies erfordert die Neu- oder Weiterentwicklung von Datenanalyseverfahren, welche stetig an die neuen biotechnologischen Methoden angepasst werden müssen. Von enormer Bedeutung haben sich Methoden des maschinellen Lernens erwiesen, welche es ermöglichen, komplexe Zusammenhänge aus den großen Datenmengen zu extrahieren. Diese Methoden sind nicht nur in der Biologie weit verbreitet, sondern sind in praktisch allen Domänen mit hohem Datenvolumen im Einsatz, beispielsweise in der Bildverarbeitung oder der Sprachanalyse und Spracherkennung. Methoden des maschinellen Lernens halten auch vermehrt in der Medizin Einzug. So haben diese Methoden in einigen Bereichen der Pathologie bereits das Niveau von Fachärzten erreicht.

Den meisten konventionellen Lernverfahren geht ein sogenannter Merkmalsextraktionsschritt voran. Solche Merkmale werden von menschlichen Experten der Domäne bereitgestellt und dienen als Basis für die Vorhersagen. Wenn beispielsweise ein automatisches Lernverfahren verwendet werden soll, um automatisch die Namen von Personen in einem Text zu erkennen (im Unterschied zu Verben, Nomen etc.), dann geht das Programm Wort für Wort über den Text und achtet dabei auf relevante Merkmale zur Vorhersage. Hier könnte etwa eine Titelbezeichnung oder Ansprache im vorangehenden Wort (zum Beispiel Dr., Prof., Frau oder Herr) nützlich zur Namenserkennung sein. Weitere Merkmale könnten in diesem Fall von Linguisten bereitgestellt werden. Der maschinelle Lernalgorithmus hat dann die Aufgabe, diese Merkmale je nach Wichtigkeit oder Genauigkeit zu gewichten, um zur erfolgreichen Namensvorhersage zu gelangen. Das Angewiesensein auf qualitativ hochwertiges Expertenwissen ist oft ein Nachteil von konventionellen maschinellen Lernverfahren, da dieses teuer ist und nicht in allen Domänen zur Verfügung steht.

In den letzten Jahren hat die Anwendung tiefer neuronaler Netze (deep neural networks), welche eine Unterkategorie des maschinellen Lernens darstellt, zu großen Erfolgen geführt. Die von der Neurobiologie inspirierten neuronalen Netze bestehen aus einfachen parametrisierbaren Funktionen, den sogenannten Neuronen, die hierarchisch in Schichten aneinandergefügt werden. Die Anordnung in vielen, oft hierarchisch organisierten Schichten wird als Deep Learning bezeichnet. Solche Modelle können relevante Merkmale automatisch erlernen und komplexe Zusammenhänge widerspiegeln, um das eigentliche Vorhersageproblem zu lösen. Als Konsequenz sind neuronale Netze besonders schnell und flexibel für unterschiedlichste Probleme anwendbar, da die Notwendigkeit von domänenspezifischer Expertise auf ein Minimum reduziert wird.

Durch ihre Ausdrucksstärke haben neuronale Netze nicht nur in vielen Fällen konventionelle maschinelle Lernverfahren weit übertroffen, sondern sich zum Teil sogar als dem Menschen überlegen gezeigt. Diese Eigenschaften sowie die Fähigkeit, aus Millionen von Datenpunkten Wissen zu schöpfen, haben tiefen neuronalen Netzen auch in der Genomik zu großer Popularität verholfen. Mit ihrer Einführung in die Genomik ist eine Lawine von weiteren Studien gefolgt, in denen unterschiedlichste genomische Datentypen kombiniert wurden, um neue Einsichten in die epigenetischen und genregulatorischen Aspekte der Zellbiologie zu erlangen [2].

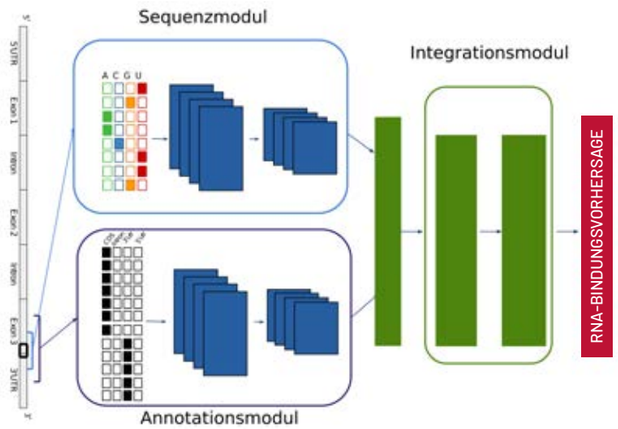


ABBILDUNG 1: Schematische Darstellung eines tiefen neuronalen Netzes für die Genomik. Das Netzwerk extrahiert Merkmale aus der DNA-Sequenz sowie der Genannotation, welche im Integrationsmodul in hierarchischer Form zu übergeordneten Merkmalen kombiniert werden. Die letzte Schicht dient zur Vorhersage von RNA-Proteinbindungen (Adaptiert von Ghanbari, 2019 [3]).

Ein wichtiger Aspekt in biomedizinischen Anwendungen ist es dabei, nicht nur ein Vorhersageproblem zu lösen, sondern auch zu verstehen, welche Informationen die relevantesten sind, um Einsichten in die zugrunde liegenden biochemischen Zusammenhänge zu erlangen. So werden tiefe neuronale Netze etwa dazu verwendet, um aus RNA-Sequenzen und Gen-Annotationen auf RNA–Protein-Interaktionen zu schließen (Abbildung 1). Dabei kann man mithilfe von Gradienten-basierten Methoden anschließend einen Einblick in die Entscheidungsvorgänge des Netzwerks erlangen und so die Plausibilität der Vorhersagen prüfen sowie Sequenzvarianten bewerten, die die Interaktionen stören können [3].

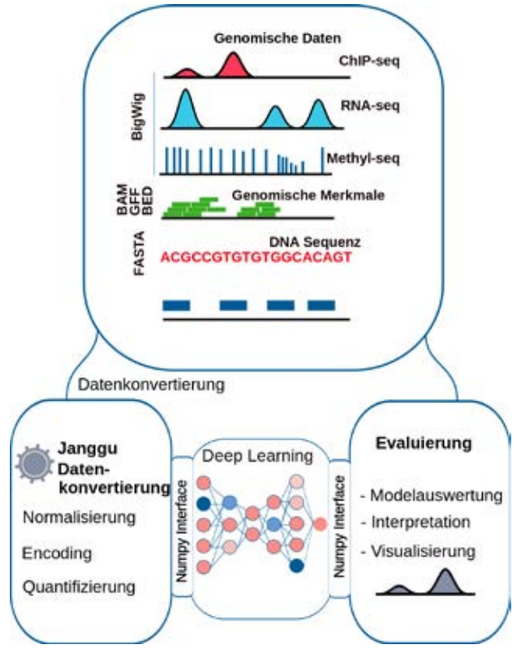


ABBILDUNG 2 Janggu unterstützt die Entwicklung von Deep-Learning-Anwendungen in der Genomik. Dies wird einerseits durch Module erzielt, welche die rohen Daten automatisch in die für die neuronalen Netzwerke benötigten Formate transformieren. Andererseits stellt Janggu eine Reihe von Methoden zur Auswertung der Ergebnisse bereit, etwa zur Messung der Vorhersagequalität oder zur Visualisierung im genomischen Kontext (Urheber: [4]).

WIE JANGGU DEEP LEARNING IN DER GENOMIK UNTERSTÜTZT

In wenigen Jahren sind bereits eine Vielzahl an Deep-Learning-Anwendungen in der Genomik entwickelt worden. Allerdings sind die meisten dieser Anwendungen zur Beantwortung spezieller Fragestellungen konzipiert: Sie erfordern fest vorgegebene Daten oder verwenden ein festgelegtes Netzwerkmodell. Durch laufend neu publizierte Datensätze oder durch neue Messprotokolle ist es allerdings immer notwendig, neuen biologischen Fragestellungen mit speziell angepassten Deep-Learning-Anwendungen auf den Grund zu gehen. Häufig sind bestehende Anwendungen allerdings nur unter enormem Entwicklungsaufwand adaptierbar, was dazu führt, dass Bioinformatiker einen wesentlichen Teil ihrer Zeit mit der Bearbeitung technischer Details verbringen anstatt mit der Beantwortung des eigentlichen biomedizinischen Problems.

Aus diesem Grund haben wir ein Softwarepaket namens Janggu entworfen, welches Bioinformatikern qualitätsgeprüfte Softwarelösungen für die häufigsten Schritte der Softwareentwicklung anbietet und es damit erleichtert, Deep-Learning-Anwendungen zu entwickeln (Abbildung 2) [4]. Eine wesentliche Hürde ist es etwa, die gemessenen Genomikdaten so zu transformieren, dass sie unmittelbar mit vorhandenen Deep-Learning-Softwaremodulen kompatibel sind. Dies war zuvor nur unter erheblichem redundantem Programmieraufwand möglich und wird durch Janggu für einen großen Teil der gängigen Datenformate einheitlich gelöst. Des Weiteren bietet Janggu eine Reihe an Validierungsmethoden an, um die Plausibilität und Qualität der Vorhersagen zu kontrollieren. So können Vorhersagen im genomischen Kontext visualisiert werden.

Die Flexibilität von Janggu wurde anhand mehrerer prototypischer Anwendungen dargestellt, die von der Vorhersage der Interaktionen von Transkriptionsfaktor-Proteinen mit DNA-Sequenzen bis zur Vorhersage der Genexpression anhand von epigenetischen Daten reichen [4]. Dabei werden technische Aspekte der Softwareentwicklung auf ein Minimum reduziert und ein hoher Durchsatz zur Beantwortung neuer Fragen erzielt.

REFERENZEN [1] Nature 2019;566:496–502. DOI: 10.1038/s41586-019-0969-x. **[2]** Nat Rev Gen 2019;20:389–403. DOI: 10.1038/s41576-019-0122-6. **[3]** Biorxiv 2019. DOI: https://doi.org/10.1101/518191. **[4]** Biorxiv 2019. DOI: https://doi.org/10.1101/700450.

AUTOREN Wolfgang Kopp[1], Philipp Boß[1], Altuna Akalin[1], Uwe Ohler[1]
[1] *Berlin Institute for Medical Systems Biology, Max-Delbrück-Centrum für Molekulare Medizin, Berlin*

RNA

in der Medizinischen Diagnostik

In der Zellfunktion spielen RNA-Moleküle eine wesentlich wichtigere Rolle als bisher angenommen. Nicht-kodierende RNA haben eine Funktion in der Zellregulation und bieten völlig neue Therapieoptionen. Über die RNA kann sogar die Expression von Genen in einzelnen Zellen gemessen werden. Damit können aus einem Tumor die wenigen Zellen identifiziert werden, die Metastasen bilden können. Eine Hauptaufgabe ist hierbei die hochwertige Analyse der Daten.

Die Zelle als Funktionseinheit wird von drei Molekülgruppen dominiert: DNA, RNA und Proteinen. Die DNA ist der Träger der Erbinformation, welche die Gene unter Verwendung von vier Basen, A, C, G und T, kodiert. Diese Gensequenz wird wiederum ausgelesen und über den Zwischenschritt der RNA-Kopie, bestehend aus A, C, G und U, in Proteine übersetzt. Proteine führen dann bestimmte Funktionen aus. Dabei werden nicht in jedem Zelltyp alle Gene ausgelesen, sondern die verwendeten Gene sowie deren Menge bestimmen, ob eine Zelle eine Leberzelle, Herzzelle oder einer der ca. 300 anderen Zelltypen des Körpers ist. Man kann die DNA sozusagen mit der Legislative vergleichen, die bestimmt, welche Funktionen möglich sind, während die Proteine die Funktion ausführen, also die Exekutive darstellen. Was sind aber dann die RNA-Moleküle? Lange Zeit ging man davon aus, dass sie nur die Eigenschaft haben, Vorlagen für Proteine zu sein. Allerdings wurde in den letzten Jahren deutlich, dass die RNA eine wesentlich wichtigere Rolle hat als bisher angenommen. So gibt es eine Vielzahl von sogenannten nicht-kodierenden RNAs, also RNAs, die nicht in Proteine übersetzt werden, die aber eine wesentliche Rolle bei der Regulation der Zellfunktion ausüben. Man kann die RNA daher vielleicht am besten mit der Judikative vergleichen, da sie die Funktion der Proteine (also der Exekutive) über Regulationsmechanismen kontrolliert. Allerdings kontrollieren die Proteine auch die Funktion der RNA, was zu einem komplizierten Regelkreis führt. Ist dieser Regelkreis gestört, führt das zu kranken Zellen.

RNA-MOLEKÜLE ALS MEDIKAMENT

Dieser Regelkreis bietet eine völlig neue Therapieoption, die bisher nur in geringem Maß genutzt wird. Existierende Medikamente haben meistens bestimmte krank machende Proteine als Angriffspunkt. Insbesondere bei genetischen Erkrankungen kann dies aber teilweise nur im eingeschränkten Maße zum Erfolg führen. Ein Beispiel ist die Spinale Muskelatrophie, die eine der meistverbreiteten genetisch bedingten Todesursachen bei Säuglingen ist [1]. Sie wird durch eine Mutation in einem Gen (SMN1) verursacht, was dazu führt, dass nicht genügend Proteine aus diesem Gen gebildet werden, um die richtige Funktion von Muskelzellen zu gewährleisten. Allerdings gibt es von dem Gen eine

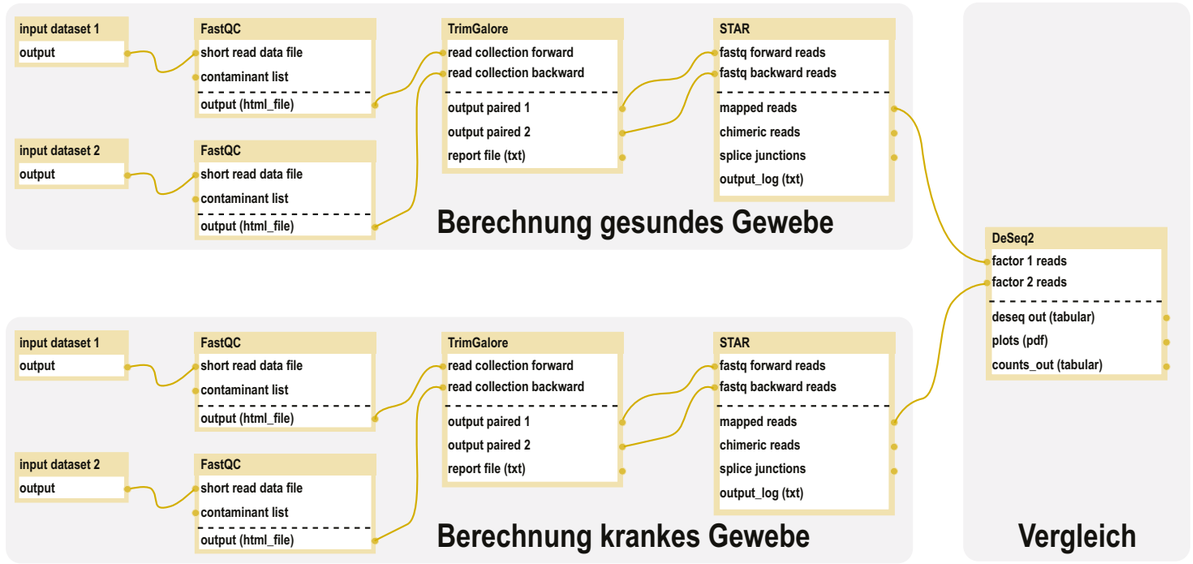


ABBILDUNG 1: Workflow für den Vergleich von gesunden und kranken Geweben durch RNA-Sequenzierung. Dabei werden die Sequenzdaten, die in zwei Dateien für den Vorwärts- und Rückwärtsstrang der DNA vorliegen, zunächst mit dem Programm FastQC auf Qualität überprüft. Fehler in der Sequenzierung werden dadurch entdeckt. Als Nächstes werden, wie beschrieben, die Adapter entfernt (TrimGalore) und die Sequenzen (genannt Reads) den bekannten Genen zugeordnet. Dies erfolgt durch das sogenannte Mappen auf das Referenzgenom. Das heißt, jedem Read wird durch das Tool STAR seine genaue Position auf dem Genom zugewiesen und dann pro Gen die Anzahl der Reads bestimmt. Dies ist dann das Expressionsprofil des gesunden Gewebes. Gleiches wird mit dem kranken Gewebe gemacht und das Programm DeSeq2 führt den Vergleich durch. Es bestimmt, welche Gene sehr unterschiedlich in beiden Geweben sind. Dies sind dann Kandidatengene für eine Krankheit.

leicht veränderte Kopie (SMN2) in unserer DNA, die häufig keinen Gendefekt aufweist. Ein neues Medikament mit dem Wirkstoff Nusinersen, das 2017 in der EU zugelassen wurde, verwendet eine RNA und ihre regulierende Wirkung, um das Gen SMN2 in höherer Kopienzahl zu produzieren, was wiederum die Krankheitsfolgen mildert.

RNA-BIOINFORMATIK ALS DETEKTIV ODER: WER IST DIE BÖSE ZELLE?

Der Zelltyp ist also nicht durch das Genom, sondern die verwendeten Gene bzw. genauer durch die Anzahl der RNA-Kopien pro ausgelesenem Gen definiert. Dies nennt man die Expression einer Zelle. Der Typ einer Zelle ist damit durch ihr Expressionsprofil definiert. Die RNA-Sequenzierung erlaubt es, das Expressionsprofil einer Menge von Zellen (bzw. der Zellen eines Gewebes) zu bestimmen. Kranke Gewebe können damit durch abnormale Expressionsprofile im Vergleich zu gesundem Gewebe bestimmt werden.

Diese Definition gilt aber nicht nur für gesunde Zellen, sondern auch für kranke Zellen. Auch Krebszellen sind ursprünglich nichts anderes als veränderte Körperzellen, die den größten Teil der Erbinformationen mit normalen Körperzellen teilen. Allerdings besteht ein Krebsgeschwür nicht nur aus Tumorzellen, sondern sie benötigen die Unterstützung von benachbarten Zellen (Stroma), um das Krebsgeschwür aufrechtzuerhalten [2]. Um es sehr vereinfacht auszudrücken: Jemand muss den Einkauf durchführen (Blutgefäße und die dazugehörigen Zellen) oder das Haus zusammenhalten (Bindegewebszellen), damit das Krebsgeschwür weiter durch ungezügelte Teilung von Tumorzellen wachsen kann. Auch bei den Tumorzellen gibt es große Unterschiede. Viele sind einfach Stubenhocker, die sich zwar unkontrolliert teilen, aber nicht aus dem Geschwür ausbrechen. Der viel schlimmere Kandidat sind jedoch Krebszellen, die aus dem Geschwür ausbrechen und neue Regionen im Körper suchen, somit den Krebs ausstrahlen. Allerdings sind das oft nur sehr wenige Zellen, die damit auch leicht übersehen werden können, insbesondere in frühen Stadien.

Hier setzt die Einzelzell-Sequenzierung an. Dabei werden typischerweise mehrere Tausend Zellen eines Tumors vereinzelt sequenziert und deren RNA-Profile bestimmt. Zelltypen werden durch den Vergleich ihrer Expressionsprofile bestimmt. Mit dieser Methode können dann besonders bösartige Tumorzellen identifiziert werden [3].

So einfach dies in der Theorie klingt, so komplex sind die technische Realisierung und die Anforderung an die digitale Datenverarbeitung. Um dies etwas klarer zu machen, muss man sich den Prozess der Sequenzierung näher ansehen. Sequenziermaschinen hängen an jedes RNA-Molekül, das abgelesen und digitalisiert wird, eine bestimmte, charakteristische RNA-Sequenz an, die aus technischen Gründen benötigt wird, den sogenannten Adapter. Der einfache, aber geniale Trick der Einzelzell-Sequenzierung ist, diese Adapter um ein kurzes Stück zu erweitern und damit die einzelnen Zellen zu identifizieren. Nur so ist es möglich, genügend RNA aus allen Zellen zu sammeln, damit man sie sequenzieren kann. Um ein kleines Beispiel zu nennen, nehmen wir mal der Einfachheit halber an, dass der vom Sequenziergerät benötigte Adapter aus einer Folge von fünf Gs besteht, also genau GGGGG. Nun wird in jeder Zelle an jedes RNA-Molekül (eine Folge von A, C, G und U) dieser Adapter und eine Folge von drei weiteren Nukleotiden angehängt, mit der die Zelle identifiziert wird. Diese Dreiersequenzen werden dann als Zahlen interpretiert, also AAA = 1, AAC = 2, AAG = 3, AAU = 4, ACA = 5 und so weiter. Damit wird an alle RNAs der Zelle 1 die Sequenz GGGGGAAA angehängt, an alle RNAs der Zelle 2 die Sequenz GGGGGAAC etc. Durch diesen Trick kann man dann 4 hoch 3 bzw. 64 Zellen eindeutig identifizieren. In Wirklichkeit sind diese Sequenzen länger, damit man in der Lage ist, mehrere Tausend Zellen zu identifizieren.

100 MIO.

Sequenzen ...

UMFASST EIN REALER DATENSATZ.

UND WIE SOLL EIN MENSCH DIESE DATEN INTERPRETIEREN?

Dieses Vorgehen ist an sich schon kompliziert, wird aber noch dadurch erschwert, dass man nicht nur fünf oder zehn RNA-Moleküle aus den Zellen sequenzieren muss, sondern viel mehr. Ein wirklicher Datensatz besteht dann aus 100 Millionen Sequenzen, die zum Beispiel von der Form GGGGGAAUUUUAGACCCCAUCAAA sowie hundert weiteren Basen sind. Wie soll man das als Mensch verstehen und herausfinden, dass es sich um ein RNA-Molekül eines Gens mit der DNA-Sequenz TTTTAGACCCATCAAAC ... in der fünften Zelle handelt? Und wie soll ein Mensch diese Information Tausenden von Zellen zuordnen und daraus dann die bösen Tumorzellen entdecken?

Die Antwort ist einfach: Gar nicht. Dies muss durch Computerprogramme bewerkstelligt werden. Es gibt hierfür eine Vielzahl von Programmen, die vom Zentrum für RNA-Bioinformatik verwaltet, angepasst und kostenlos einer breiten Forschergemeinschaft zur Verfügung gestellt werden. Hierunter sind auch Programme, die Aptamere entfernen, sie den Zellen zuordnen, die angehängte RNA-Sequenz bestimmten Genen zuordnen (sogenannte Mapper) und daraus Expressionsprofile für die einzelnen Zellen erstellen (Abbildung 1). Dabei werden typischerweise ca. 2.000-3.000 Gene und deren Expression erfasst. Aber selbst danach würde der Vergleich, ob die Expression der 2.000-3.000 Gene in der Zelle X ähnlich zu dem Profil der Zelle Y ist, jeden Menschen überfordern. Daher gibt es Programme, die diesen Vergleich durch-

führen und dem Mediziner letztendlich Gruppen von Zellen und deren Häufigkeit aufzeigen. Anhand dieser Visualisierung der digitalen Daten kann dann ein Mediziner oder Lebenswissenschaftler seine Schlüsse ziehen. Für ein Analyseproblem werden sogenannte Workflows erstellt, die typischerweise ein bis mehrere Dutzend Programme in einen sinnvollen Ablauf bringen, damit diese Visualisierung gelingt und neue Erkenntnisse gewonnen werden können (Abbildung 2).

Das RNA-Bioinformatik-Zentrum des Deutschen Netzwerks für Bioinformatik-Infrastruktur besteht aus sieben Partnern aus ganz Deutschland, die sich zur Aufgabe gemacht haben, die notwendigen Tools, Workflows und Visualisierungen zu entwickeln und einem jeden zugänglich zu machen. Auf unserem Galaxy-Server stellen wir zum Beispiel mehr als 2.000 verschiedene Tools zur Verfügung, die beliebig zur Analyse von hochkomplexen Daten verknüpft werden können.

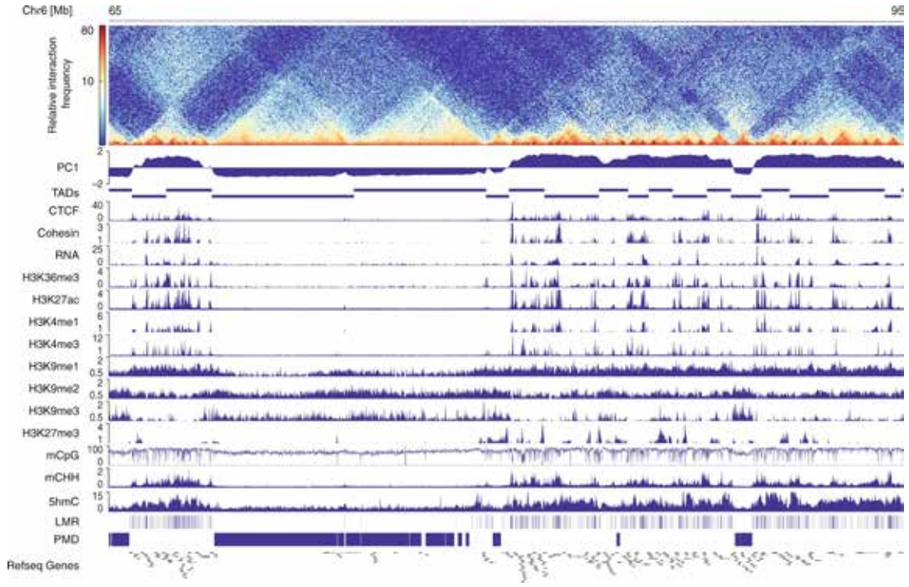
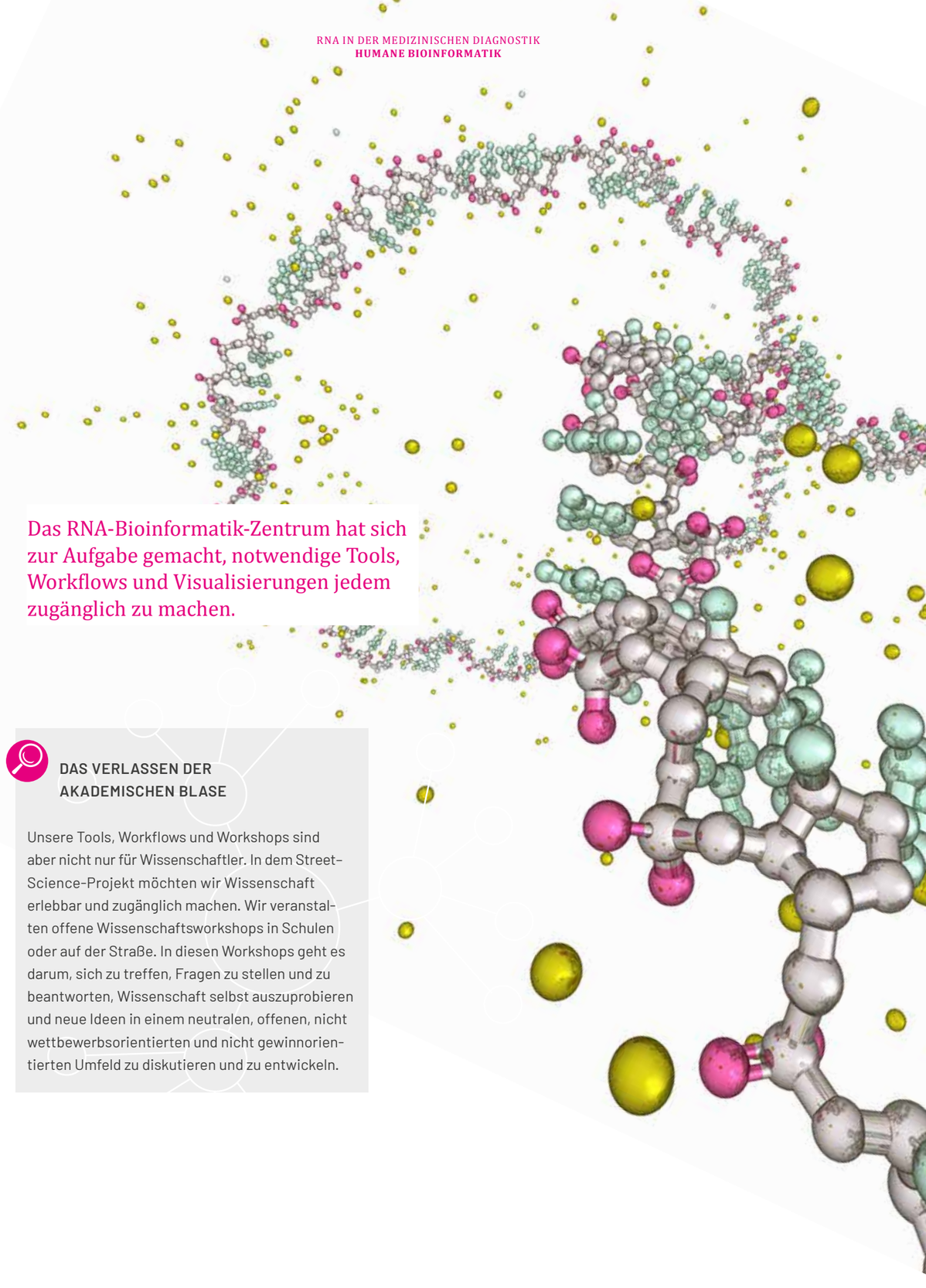


ABBILDUNG 2: Visualisierung mehrerer Sequenzierungen, die unterschiedliche Eigenschaften des Genoms untersuchen. Jede Reihe (also PC1, TADs usw.) entspricht dabei einem Sequenzierungs-Experiment, das z. B. die Struktur der DNA (TADs) oder ihre epigenetischen Veränderungen (H3K36me3 usw.) bestimmt. In der letzten Reihe sind die Reads einer normalen RNA-Sequenzierung dargestellt. Diese kompakte Darstellung erlaubt es dem Lebenswissenschaftler, die Ergebnisse der verschiedenen Experimente korrekt zu interpretieren. (Bild aus: Nothjunge, S., Nührenberg, T.G., Grüning, B.A. et al. DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. Nat Commun 8, 1667 (2017) doi:10.1038/s41467-017-01724-9)

REFERENZEN [1] https://www.presseportal.de/pm/102449/3651447 **[2]** https://www.uniklinikum-leipzig.de/einrichtungen/dermatologie/Seiten/forschung-prof-simon-tumor-stroma-interaktionen-.aspx **[3]** https://science.sciencemag.org/content/352/6282/189

AUTOREN Rolf Backofen[1] und Björn Grüning[1]
[1] *Albert-Ludwigs-Universität Freiburg, Institut für Informatik, Georges-Köhler-Allee 106, 79110 Freiburg*

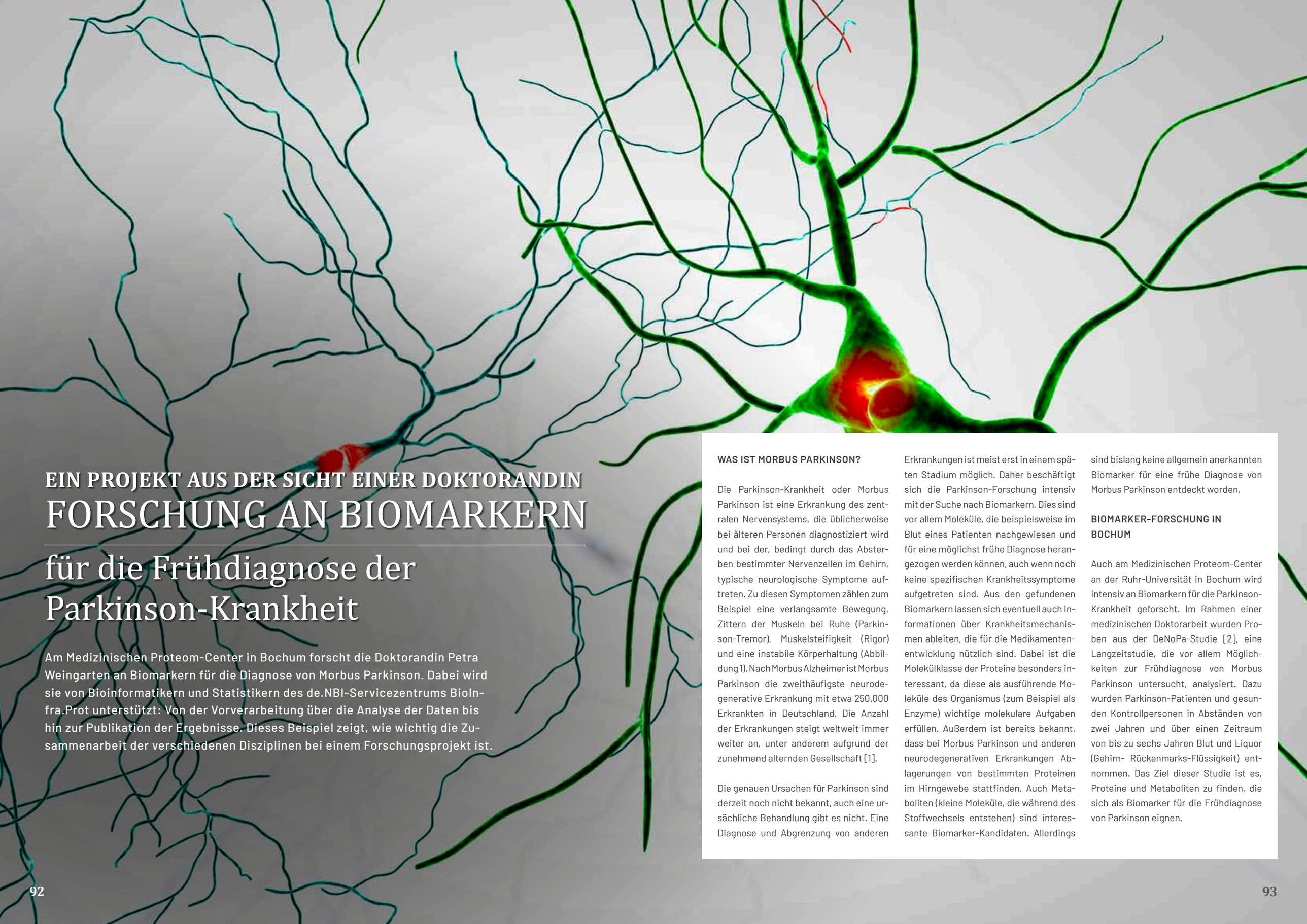
RNA IN DER MEDIZINISCHEN DIAGNOSTIK
HUMANE BIOINFORMATIK



Das RNA-Bioinformatik-Zentrum hat sich zur Aufgabe gemacht, notwendige Tools, Workflows und Visualisierungen jedem zugänglich zu machen.

DAS VERLASSEN DER AKADEMISCHEN BLASE

Unsere Tools, Workflows und Workshops sind aber nicht nur für Wissenschaftler. In dem Street-Science-Projekt möchten wir Wissenschaft erlebbar und zugänglich machen. Wir veranstalten offene Wissenschaftsworkshops in Schulen oder auf der Straße. In diesen Workshops geht es darum, sich zu treffen, Fragen zu stellen und zu beantworten, Wissenschaft selbst auszuprobieren und neue Ideen in einem neutralen, offenen, nicht wettbewerbsorientierten und nicht gewinnorientierten Umfeld zu diskutieren und zu entwickeln.



EIN PROJEKT AUS DER SICHT EINER DOKTORANDIN

FORSCHUNG AN BIOMARKERN

für die Frühdiagnose der Parkinson-Krankheit

Am Medizinischen Proteom-Center in Bochum forscht die Doktorandin Petra Weingarten an Biomarkern für die Diagnose von Morbus Parkinson. Dabei wird sie von Bioinformatikern und Statistikern des de.NBI-Servicezentrums BioInfra.Prot unterstützt: Von der Vorverarbeitung über die Analyse der Daten bis hin zur Publikation der Ergebnisse. Dieses Beispiel zeigt, wie wichtig die Zusammenarbeit der verschiedenen Disziplinen bei einem Forschungsprojekt ist.

WAS IST MORBUS PARKINSON?

Die Parkinson-Krankheit oder Morbus Parkinson ist eine Erkrankung des zentralen Nervensystems, die üblicherweise bei älteren Personen diagnostiziert wird und bei der, bedingt durch das Absterben bestimmter Nervenzellen im Gehirn, typische neurologische Symptome auftreten. Zu diesen Symptomen zählen zum Beispiel eine verlangsamte Bewegung, Zittern der Muskeln bei Ruhe (Parkinson-Tremor), Muskelsteifigkeit (Rigor) und eine instabile Körperhaltung (Abbildung 1). Nach Morbus Alzheimer ist Morbus Parkinson die zweithäufigste neurodegenerative Erkrankung mit etwa 250.000 Erkrankten in Deutschland. Die Anzahl der Erkrankungen steigt weltweit immer weiter an, unter anderem aufgrund der zunehmend alternden Gesellschaft [1].

Die genauen Ursachen für Parkinson sind derzeit noch nicht bekannt, auch eine ursächliche Behandlung gibt es nicht. Eine Diagnose und Abgrenzung von anderen Erkrankungen ist meist erst in einem späten Stadium möglich. Daher beschäftigt sich die Parkinson-Forschung intensiv mit der Suche nach Biomarkern. Dies sind vor allem Moleküle, die beispielsweise im Blut eines Patienten nachgewiesen und für eine möglichst frühe Diagnose herangezogen werden können, auch wenn noch keine spezifischen Krankheitssymptome aufgetreten sind. Aus den gefundenen Biomarkern lassen sich eventuell auch Informationen über Krankheitsmechanismen ableiten, die für die Medikamentenentwicklung nützlich sind. Dabei ist die Molekülklasse der Proteine besonders interessant, da diese als ausführende Moleküle des Organismus (zum Beispiel als Enzyme) wichtige molekulare Aufgaben erfüllen. Außerdem ist bereits bekannt, dass bei Morbus Parkinson und anderen neurodegenerativen Erkrankungen Ablagerungen von bestimmten Proteinen im Hirngewebe stattfinden. Auch Metaboliten (kleine Moleküle, die während des Stoffwechsels entstehen) sind interessante Biomarker-Kandidaten. Allerdings sind bislang keine allgemein anerkannten Biomarker für eine frühe Diagnose von Morbus Parkinson entdeckt worden.

BIOMARKER-FORSCHUNG IN BOCHUM

Auch am Medizinischen Proteom-Center an der Ruhr-Universität in Bochum wird intensiv an Biomarkern für die Parkinson-Krankheit geforscht. Im Rahmen einer medizinischen Doktorarbeit wurden Proben aus der DeNoPa-Studie [2], eine Langzeitstudie, die vor allem Möglichkeiten zur Frühdiagnose von Morbus Parkinson untersucht, analysiert. Dazu wurden Parkinson-Patienten und gesunden Kontrollpersonen in Abständen von zwei Jahren und über einen Zeitraum von bis zu sechs Jahren Blut und Liquor (Gehirn- Rückenmarks-Flüssigkeit) entnommen. Das Ziel dieser Studie ist es, Proteine und Metaboliten zu finden, die sich als Biomarker für die Frühdiagnose von Parkinson eignen.

Aufgrund des komplexen Studiendesigns (Abbildung 2) und der vielfältigen Analysemöglichkeiten wurde die Doktorandin Petra Weingarten von der Statistikerin Karin Schork und dem Bioinformatiker Michael Turewicz unterstützt, die im Rahmen des de.NBI-Servicezentrums BioInfra.Prot Beratung und Analysen im Bereich der Bioinformatik und Statistik für Proteomik-Daten anbieten. In einem ersten Vorgespräch wurden die Ziele des Projektes und die bisherigen Vorarbeiten besprochen. Laut Karin Schork sollte eine erste Besprechung möglichst früh im Projekt erfolgen: *„Man denkt immer, die statistische Auswertung kommt bei so einem Projekt erst ganz am Schluss. Dabei ist auch das Studiendesign ganz am Anfang ein entscheidendes Element. Es ist wichtig, bereits vor der Messung der Daten Kontakt zu einem Statistiker oder Bioinformatiker aufzunehmen und über das geplante Projekt zu sprechen. So kann man früh mögliche Herausforderungen erkennen und auch einigen Problemen bei der Analyse später vorbeugen."*

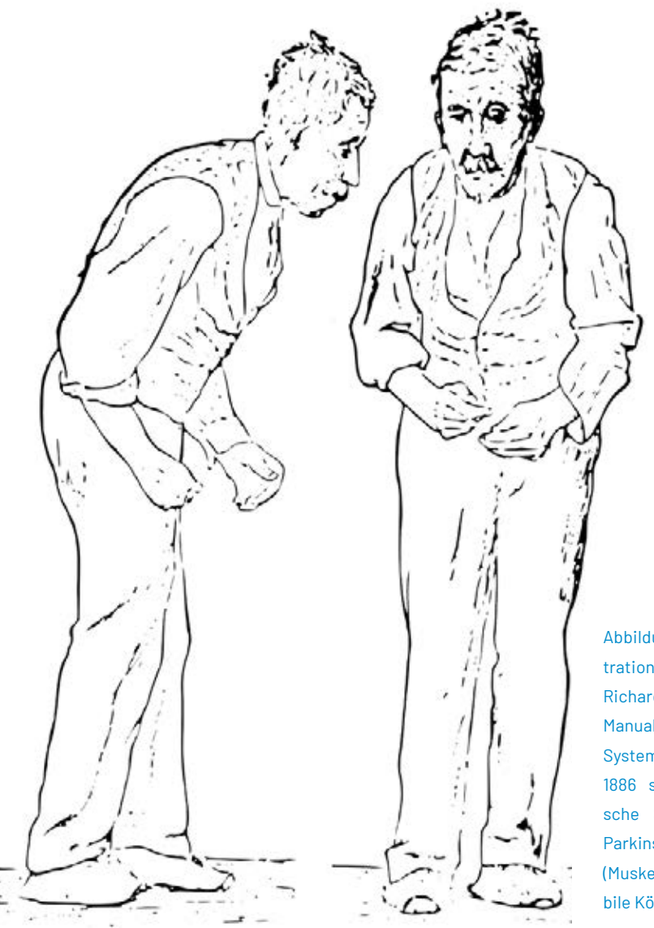


Abbildung 1: Diese Illustration von Sir William Richard Gowers aus „A Manual of the Nervous System" aus dem Jahr 1886 stellt einige typische Symptome eines Parkinson-Patienten dar (Muskelsteifigkeit, instabile Körperhaltung) [4].

HERAUSFORDERUNG: METABOLITEN

Bei diesem ersten Gespräch stellte sich heraus, dass es bei diesem Projekt eine große Herausforderung geben würde: die Analyse und Verarbeitung von Metabolitdaten. Im Gegensatz zu der Analyse von Proteindaten gab es diesbezüglich, sowohl vonseiten des Labors als auch der Datenanalyse bisher kaum Erfahrungen. Die Messung der Metabolitdaten erfolgte mit einem kommerziellen Kit, das eine Etablierung dieser Technik im Labor erforderte. Bei der Datenanalyse warf vor allem die Vorverarbeitung der Daten Fragen auf: Welche Vorverarbeitungsschritte müssen noch erfolgen und welche wurden bereits durch die mitgelieferte Software abgedeckt? Können die Metabolitdaten eigentlich bei der statistischen Analyse genauso behandelt werden wie Proteindaten? Nach vielen Besprechungen, Überprüfungen verschiedener Methoden und einer Telefonkonferenz mit der Herstellerfirma des Kits konnte eine Strategie zur Vorverarbeitung der Daten gefunden und die Daten für die statistische Auswertung vorbereitet werden.

de.NBI-TRAINING UNTERSTÜTZT DIE BERATUNG

Für die statistische Auswertung der Metabolit- und Proteindaten wurden Skripte in der Programmiersprache R vorbereitet, die sowohl Vergleiche der Daten über die verschiedenen Zeitpunkte als auch zwischen Patienten sowie Kontrollen ermöglichen und dazu passende Grafiken erstellen. Michael Turewicz und Karin Schork bieten als de.NBI-Training regelmäßig einmal im Jahr einen Einführungskurs zu R an, in dem der grundlegende Umgang mit dieser Programmiersprache eingeübt wird. Auch Petra Weingarten hat diesen Einführungskurs besucht und ist dadurch in der Lage, die bereitgestellten Skripte selbst an neue Daten anzupassen und kleinere Änderungen vorzunehmen: *„Der R-Kurs hat mir bei meiner Arbeit sehr geholfen. Im Kurs wurde ich langsam und verständlich an die relevanten Funktionen herangeführt und konnte dadurch die mir zur Verfügung gestellten R-Skripte lesen und auch anhand der Skripte manche Analyseschritte nachvollziehen."*

FORSCHUNG AN BIOMARKERN FÜR DIE FRÜHDIAGNOSE DER PARKINSON-KRANKHEIT
HUMANE BIOINFORMATIK

VIELVERSPRECHENDE ERGEBNISSE

Während eines der vielen Beratungsgespräche wurden die Ergebnisse der statistischen und bioinformatischen Analyse vorgestellt. Es wurden einige vielversprechende Biomarker-Kandidaten gefunden, die in noch folgenden Experimenten validiert werden sollen. Neben den Analysen einzelner Proteine und Metaboliten wurden auch sogenannte Biomarker-Panels untersucht. *„Vor dem Hintergrund der natürlichen biologischen Variabilität, von individuell unterschiedlichen Krankheitsverläufen und von diversen Krankheitssubtypen ist davon auszugehen, dass eine kleine Menge von verschiedenen Biomolekülen in Kombination besser als diagnostischer Biomarker geeignet ist als einzelne Proteine oder Metaboliten. Diese werden mit Methoden des maschinellen Lernens gesucht und sie können die komplexen molekularen Muster, anhand derer Morbus Parkinson erkannt werden kann, besser abbilden. Eine solche Menge von Biomolekülen bezeichnet man als Biomarker-Panel"*, erklärt Michael Turewicz. Die abschließende Analyse hierzu steht derzeit noch aus.

AUF DEM WEG ZUR PUBLIKATION

Die Ergebnisse der Studie sollen außerdem in einem wissenschaftlichen Artikel zusammengefasst und publiziert werden. In der Proteomik ist es üblich und von vielen wissenschaftlichen Zeitschriften bei einer Publikation verlangt, dass auch die zur Studie gehörenden Rohdaten der Messungen veröffentlicht werden. Dafür gibt es das PRIDE-Archiv [3], in das die Daten hochgeladen werden können und so nach einer erfolgreichen Publikation öffentlich für Re-Analysen zur Verfügung stehen. Da das Hochladen der großen Datenmengen oftmals nicht ganz unproblematisch ist, bietet BioInfra.Prot einen Upload-Service an, der bei Fragen und Problemen hilft. Außerdem wurde ein Tool entwickelt, das die Daten in Standardformate konvertiert, die von PRIDE akzeptiert werden.

ERFOLGREICHER PROJEKTVERLAUF

Insgesamt sind beide Seiten sehr zufrieden mit dem bisherigen Verlauf des Projekts; einige weitere Analysen und die Erstellung einer Publikation werden in den nächsten Wochen folgen. Petra Weingarten schreibt derzeit an den letzten Kapiteln ihrer Doktorarbeit, für die die Kooperation mit BioInfra.Prot ein großer Gewinn war: *„Die enge Zusammenarbeit mit der Bioinformatik und Statistik hat mir ein sicheres Gefühl bei dem Umgang mit den Daten gegeben. Die Möglichkeit, Fragen direkt mit Fachleuten gut verständlich klären zu können, war unersetzbar und ich weiß nicht, wie ich es ohne die Hilfe effektiv geschafft hätte."*

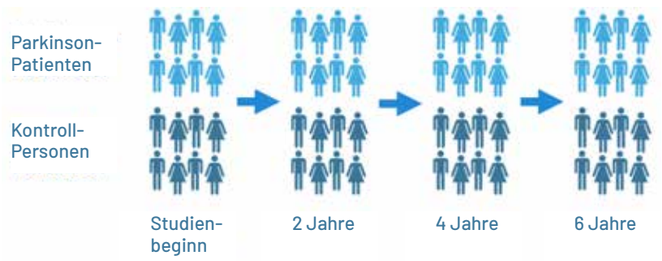


ABBILDUNG 2: Studiendesign: Blutplasma- und Liquor-Proben werden von Parkinson-Patienten und gesunden Kontrollpersonen zu Studienbeginn und nach zwei, vier und sechs Jahren entnommen und auf Metaboliten und Proteine hin untersucht. (Bild: Karin Schork, Petra Weingarten)

REFERENZEN [1] https://www.parkinson-gesellschaft.de/aktuelles/36-von-der-forschung-in-die-klinik-diedeutsche-parkinson-gesellschaft-mit-neuer-praesenz-im-web.html **[2]** https://www.denopa.de/ **[3]** https://www.ebi.ac.uk/pride/archive/ **[4]** https://de.wikipedia.org/wiki/Parkinson-Krankheit#/media/Datei:Sir_William_Richard_Gowers_Parkinson_Disease_sketch_1886.svg

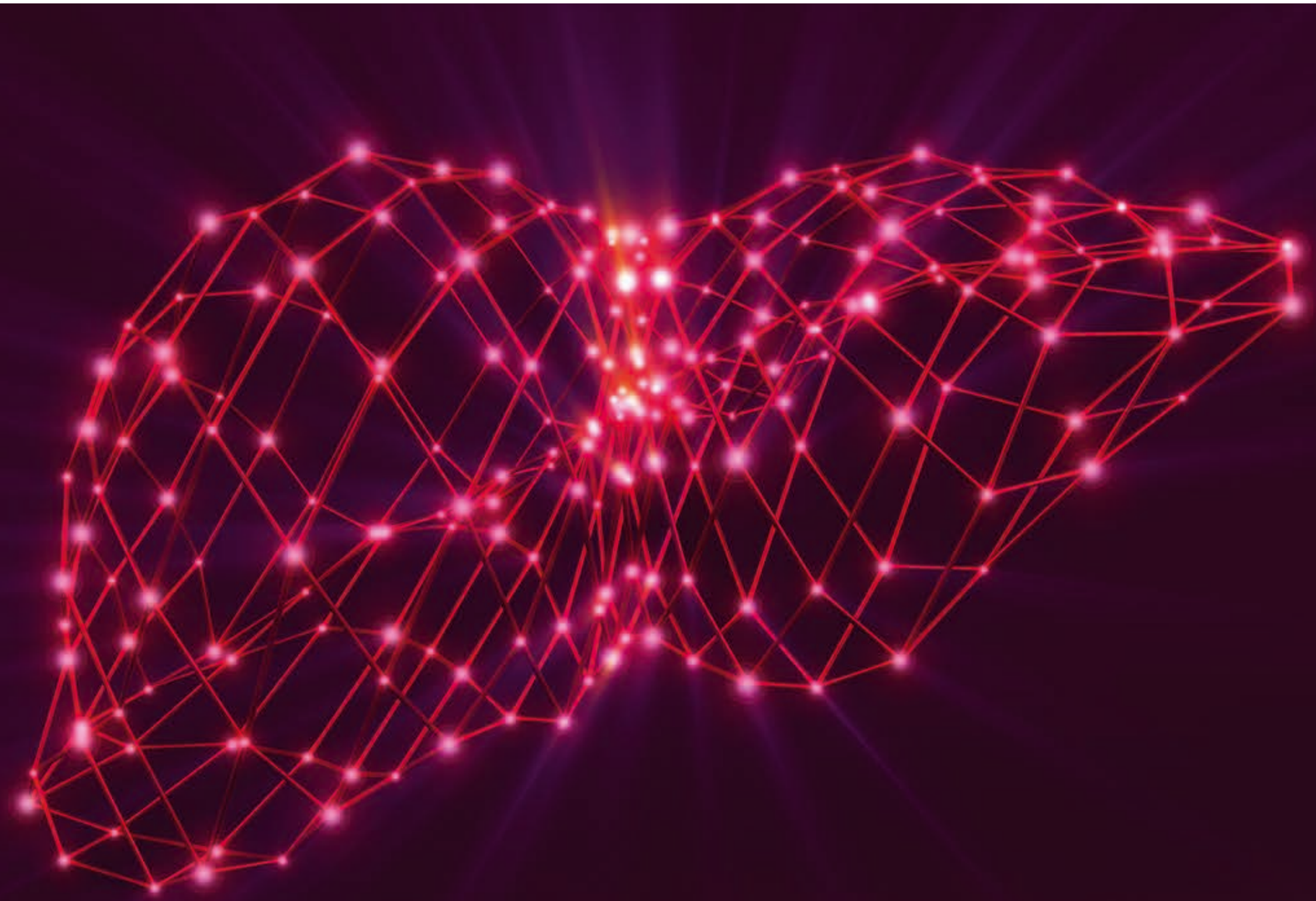
AUTOREN Karin Schork[1], Petra Weingarten[1], Martin Eisenacher[1] und Michael Turewicz[1]
[1] *Ruhr-Universität Bochum, Medizinische Fakultät, Medizinisches Proteom-Center, Gesundheitscampus 4, 44801 Bochum*

SYSTEMMEDIZIN DER LEBER – HERAUSFORDERUNG FÜR DAS DATENMANAGEMENT

Eine der Besonderheiten der Leber ist ihre Heilungsfähigkeit. Klinisch relevante Erkenntnisse zu gewinnen, wie langfristige Herausforderungen die Leber trotzdem schädigen und zu fortschreitenden Lebererkrankungen führen können, ist Gegenstand des LiSyM-Projekts. In diesem arbeiten Wissenschaftlerinnen und Wissenschaftler aus Labor, Theorie und Klinik zusammen. Der vorliegende Artikel beschreibt, wie dieses vielschichtige Projekt im Datenmanagement abgebildet wird.

Die Leber ist ein außergewöhnliches Organ mit vielen verschiedenen Aufgaben in unserem Körper und erstaunlicher Leistungsfähigkeit. Schon die alten Griechen wussten, dass Teile der Leber funktionsfähig immer wieder neu gebildet werden können. Der Sage nach wurde der Titan Prometheus als Strafe dafür, den Menschen die Kunst des Feuermachens verraten zu haben, von den Göttern auf grausame Weise bestraft. Er wurde an einen Stein im Kaukasus angekettet und ein Adler fraß täglich Teile seiner Leber, die sich dann immer wieder neu bildete, bis der Adler am nächsten Tag wiederkam.

Die Belastungen einer heutigen Leber sind profaner, aber deutlich vielfältiger. Trotz der Fähigkeit, sich neu zu bilden, kann eine Leber langfristig, schleichend verletzt werden. Diese langfristigen Leberschädigungen beginnen zumeist mit der Bildung einer sogenannten Fettleber. Zirka 20 % der westlichen Bevölkerung leiden an einer sogenannten nicht alkoholischen Fettleber, also einer Leber, die Fetttröpfchen eingelagert hat, geschädigt ist, aber wahrscheinlich nicht durch Alkohol geschädigt wurde. Einige nicht alkoholische Fettlebern entzünden sich und entwickeln eine nichtalkoholische Steatohepatitis, eine Form der Leberentzündung. Andere Patienten bleiben bis auf die Verfettung gesund. Worauf beruht dieser Unterschied? Was führt nun zu einem Fortschreiten der Krankheit? Was schützt die Leber davor? Diese und weitere Fragen stehen im Zentrum des Forschungsnetzwerks Systemmedizin der Leber (LiSyM), gefördert vom deutschen Bundesministerium für Bildung und Forschung. LiSyM verfolgt einen systemmedizinischen Ansatz: Mit verschiedenen Ansätzen wird versucht, biologische Systeme mithilfe von simulierbaren Computermodellen zu verstehen und dieses Verständnis dann in der Klinik anzuwenden bzw. durch die Entwicklung neuer Therapieansätze einer klinischen Anwendung näher zu bringen.



SYSTEMMEDIZIN DER LEBER-HERAUSFORDERUNG FÜR DAS DATENMANAGEMENT
HUMANE BIOINFORMATIK

In LiSyM arbeiten 37 Forschungsgruppen aus 23 verschiedenen Forschungszentren und Organisationen zusammen. Natürlich funktioniert dies nicht spontan, sondern benötigt eine sinnvolle Struktur und Organisation, die von einem zentralen Datenmanagement unterstützt wird. Weitere wichtige Gründe für Datenmanagement sind, eine Nachvollziehbarkeit und Nachnutzung von Daten zu ermöglichen. Hier spricht man von FAIRen-Daten: findable (auffindbar), accessible (zugreifbar) sowie interoperable (interoperabel, also kombinierbar) und reusable (wiederverwendbar). FAIR ist keine genaue Vorgabe zur Strukturierung und Formatierung von Daten, sondern eher ein ganzes Spektrum an sehr einfachen und grundlegenden Regeln, wie Daten fair gemacht werden können. Das Ziel ist ein nützlicher Kompromiss, der die FAIRness der Daten mit geringstem Aufwand erreicht.

Im Rahmen des LiSyM-Netzwerks wird auch Personal für Datenmanagement-Experten zur Weiterentwicklung der ver-

wendeten Softwareplattform sowie zum Sammeln von Anforderungen, welche unterschiedliche Nutzer an das Datenmanagement stellen, und zum Community-Management (inklusive Nutzertraining) gefördert. Diese projektgeförderten Experten arbeiten eng mit Personal aus dem de.NBI-SysBio-Team zusammen und nutzen neben eigenen Entwicklungen auch Resultate der Entwicklungsarbeit aus de.NBI-SysBio. Auch gemeinsame, projektübergreifende Konzeptions- und Entwicklungsarbeit für das Datenmanagement findet statt, wodurch Synergien entstehen, von denen alle beteiligten Projekte profitieren.

LiSyM ist in vier thematischen Säulen organisiert.

LiSyM ist in vier thematischen Säulen organisiert, in denen jeweils eine Frage untersucht wird (von Tier-Experiment bis Klinik). Jede Säule hat Partnerinnen bzw. Partner aus experimenteller Forschung, Modellierung und Klinik.

▶ **Säule 1:** *Early Metabolic Injury* befasst sich damit, wie die Fettleber in eine Leberentzündung übergeht.

▶ **Säule 2:** *Chronic Liver Disease Progression* befasst sich mit dem Übergang von einer Entzündung zur Zirrhose.

▶ **Säule 3:** *Regeneration and Repair in Acute-on-Chronic Liver Failure* befasst sich damit, wie man Leberheilung im Falle eines akuten Versagens einer chronisch erkrankten Leber begünstigen kann.

▶ **Säule 4:** *Liver Function Diagnostics*. Ziel dieser Säule ist die nicht invasive Diagnose von Leberschäden.

Komplettiert wird das Großprojekt durch das koordinierende Programmdirektorat unter der Leitung von Prof. Peter Janssen. Hier ist auch das Datenmanagement angesiedelt.

Eine Herausforderung für das Datenmanagement liegt im Vorhandensein von sehr unterschiedlichen Daten, die für die Generierung und Simulation der in LiSyM entwickelten Computermodelle zusammengeführt, also integriert werden müssen. Die Daten unterscheiden sich einerseits in ihrer Modalität (Bilddaten, Messdaten, Gen- bzw. Protein-Sequenzdaten, klinisch erhobene Daten usw.), andererseits auch in den genutzten Formaten für Daten und Metadaten (Daten, welche die eigentlichen Daten beschreiben und in Zusammenhang bringen) sowie in den Ansprüchen an die Privatsphäre und Datensicherheit. Mäuse genießen selbst keinen besonderen Schutz der Privatsphäre, wohl aber Menschen. Als Konsequenz sind Daten über Menschen anders zu behandeln als Daten, die aus Tieren oder Zelllinien im Labor gewonnen wurden.

Weitere Herausforderungen sind die Unterschiedlichkeit der gleichzeitig verwendeten Computeranwendungen sowie das räumlich verteilte Vorliegen von Daten aus dem Netzwerk. Patientenbezogene Daten dürfen meist die Organisation nicht verlassen, an der sie gewonnen wurden. Die Projektpartner können also nicht alle Daten gemeinsam an einem Ort zentral speichern, müssen diese aber trotzdem in Zusammenhang bringen können, auch über die Grenzen der Organisationen hinweg. Ferner haben manche Nutzer aus anderen Gründen lokale Datenspeicher oder weitere Werkzeuge, die sie einsetzen. Mit diesen muss das zentrale Datenmanagement-System eines solchen Forschungsnetzwerks kommunizieren können.

DIES FÜHRT UNS ZUR UNTENSTEHENDEN LISYM-DATENMANAGEMENT-ARCHITEKTUR:

Das Zentrum der Architektur ist LiSyM SEEK, eine Installation der SEEK-Software. Sie wird seit zehn Jahren gemeinsam von der Universität Manchester, Heidelberger Institut für Theoretische Studien (HITS) und anderen Partnern der FAIRDOM-Initiative [1] entwickelt und in verschiedenen europäischen und nationalen Forschungskonsortien eingesetzt. SEEK ist mit dem Wissen entwickelt, dass Datenmanagement in interdisziplinären Forschungsprojekten Daten aus den verschiedensten Quellen katalogisieren können muss. Es ist in der Lage, Daten zentral zu speichern, aber auch auf verteilt vorliegende Daten zu verweisen und diese miteinander zu vernetzen – also genau das, was im Fall LiSyM benötigt wird.

Wir nutzen für das LiSyM-Netzwerk eine eigene SEEK-Instanz aus Gründen der Flexibilität und der erhöhten Datensicherheit: Nicht jede Veränderung des Systems, die für LiSyM erforderlich oder nützlich ist, ist auch für andere SEEK-Instanzen brauchbar oder erforderlich, zum Beispiel für die von mehreren unterschiedlichen Forschungsprojekten und Konsortien gemeinsam parallel benutzte Instanz FAIRDOMHub. Diese SEEK-Installation, welche – wie auch LiSyM-SEEK – auf einem eigenen Server bei HITS betrieben wird, ist die meistbenutzte Instanz, die mehr als 100 Projekte verschiedener Größe mit Datenmanagement versorgt. Hierhin können LiSyM-Nutzer ihre Daten transferieren, sofern dies gewünscht wird. Dies erleichtert insbesondere das Teilen von einzelnen sowie von vernetzten Datensätzen mit anderen Projekten (also außerhalb von LiSyM) in der gemeinsam genutzten Plattform FAIRDOMHub. Sehr fein für jeden Datensatz einstellbare Zugriffsrechte ermöglichen dabei sowohl den Austausch vertraulicher Daten mit einzelnen anderen Nutzern oder Nutzerkreisen innerhalb eines Projekts als auch den Austausch über die Projektgrenzen hinweg.

Die SEEK-Software, also auch FAIRDOM-Hub und LiSyM SEEK, erlaubt es ihren Nutzern, Daten zu speichern, zu katalogisieren, mit Metadaten zu versehen, mit anderen Daten zu verknüpfen und schließlich mit anderen Nutzern zu teilen. Die Nutzer können dabei für jeden einzelnen Datensatz bestimmen, wer genau die Daten sehen und herunterladen darf, wer lediglich einige grundlegende Metadaten sieht und die eigentlichen Daten lediglich auf Anfrage an den Besitzer der Daten erhält und wer gar keinen Zugriff bekommt.

Dies darf auch über den Lebenszyklus der Daten hinweg verändert werden. Beispielsweise werden Daten neuer Laborexperimente vom Experimentator eventuell gespeichert, aber zunächst mit nur wenigen engen Kooperationspartnern geteilt. Zu einem späteren Zeitpunkt sollen diese Daten dann mit anderen Daten des Projekts in Verbindung gebracht werden, um zum Beispiel ein Computermodell daraus zu entwickeln. Die meisten SEEK-Nutzenden teilen ihre Daten bis zur Publikation zunächst nur innerhalb ihres Projekts. Darüber hinaus können Daten auch mittels geheimer Links, die auch mit einem Verfallsdatum versehen werden können, im Wesentlichen versteckt gehalten, aber mit Gutachtern geteilt werden: Nutzer mit dem geheimen Link bekommen Zugriff auf die Datei – auch ohne eigenes Benutzerkonto. Schließlich können Daten mit der Welt geteilt, also publiziert werden sowie mit stabilen, dauerhaft zitierbaren *Digital Object Identifiers* versehen werden, beispielsweise zum Verweis auf *Supporting* Material aus einer Publikation.

WIE KOMMEN DIE DATEN MIT IHREN METADATEN NUN IN DAS LISYM SEEK?

Neben dem klassischen manuellen Upload via Web-Interface durch den Nutzer bieten sich die folgenden Möglichkeiten (Abbildung 1).

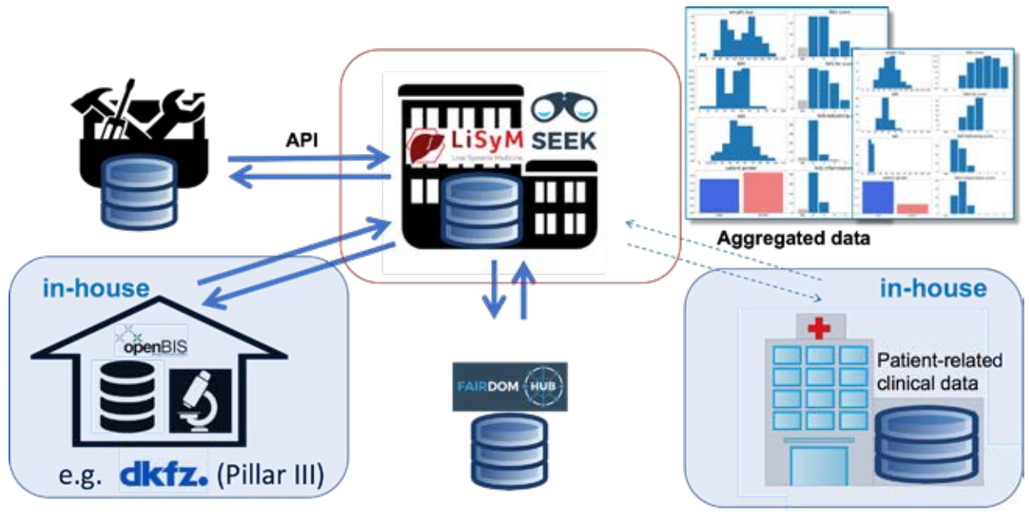


ABBILDUNG 1: Möglichkeiten des Datentransfers in LiSyM SEEK.

API zum Datentransfer: Es gibt eine webbasierte Programmierschnittstelle, die es ermöglicht, Daten direkt und programmgesteuert in SEEK hochzuladen. Dies kann beispielsweise mittels Python-Programmen geschehen. Wir bieten hierzu konkrete Beispielcodes und sind bei der Anpassung dieses Beispielcodes behilflich.

Link zu openBIS [2]: Das von unseren Kooperationspartnern an der ETH Zürich in der Schweiz entwickelte Labor-Informations- und Management-System (LIMS) openBIS [2] wird in Laboratorien zum Datenmanagement eingesetzt. Neuere Versionen dieses Systems bieten ein konfigurierbares Interface zu SEEK, das gemeinsam mit Partnern aus Manchester und Edinburgh entwickelt wurde. Hier ist es möglich, gezielt Experimente und Gruppen von Experimenten innerhalb von SEEK sichtbar zu machen. Sie sind dann in beiden Systemen sichtbar und können geteilt werden. Diese Möglichkeit wird beispielsweise bei unseren Partnern am DKFZ in Heidelberg eingesetzt.

Upload per Datenträger: Schließlich gibt es für große Daten auch die Möglichkeit, einen Upload über den Austausch von Datenträgern durchzuführen. Dies ist insbesondere für große Datenmengen wie Stapel zusammengehöriger, hochauflösender mikroskopischer Aufnahmen interessant.

AUSTAUSCH ÜBER KLINISCHE DATEN (ALS AGGREGIERTE ANONYMISIERTE DATEN)

Klinische Daten, welche in Forschungsprojekten verwendet werden sollen, sind für das wissenschaftliche Datenmanagement in kooperativ arbeitenden Forschungsverbünden eine große Herausforderung, da sie nicht ohne Weiteres über Organisationsgrenzen verschickt werden dürfen. Sie dürfen also im Normalfall nicht im FAIRDOMHub oder LiSyM SEEK gespeichert werden, da dies in der Regel nicht im Einklang mit dem Datenschutz steht. Es ist jedoch möglich sich über diese Daten auszutauschen, indem die Daten einzelner Patienten lokal in der Klinik gruppiert (aggregiert) werden und lediglich die Daten über die Patientengruppen, nicht aber über einzelne Patienten, mit Kooperationspartnern außerhalb der Klinik geteilt werden: Welche Eigenschaften haben die Teilnehmer an einer Studie? Hat vielleicht ein Partner die ersehnten Daten über jugendliche Leberkranke, die die Daten über ältere Kranke komplementieren könnten? Welche Verteilung bestimmter Leberwerte der an einer Studie beteiligten Kliniken gibt es?

> Eine große Herausforderung für die Forschung: Patientendaten dürfen aus Datenschutzgründen nicht einfach verschickt werden. Erst eine Zusammenfassung darf legal in LiSyM SEEK gespeichert und zwischen den Partnern ausgetauscht werden.

Hierfür bietet es sich an, einen mobilen Code zu verteilen. Wir haben auf Basis von Anaconda und Jupyter demonstriert, wie dies durchgeführt werden kann: Die klinischen Partner haben sich auf Tabellenvorlagen geeinigt, also in Excel vorliegende Beispielstrukturen für klinische Daten. Wir haben nun in Python Code implementiert, der diese Daten vor Ort in der Klinik liest, direkt analysiert und aggregiert, ohne dass die Daten transportiert werden. Die Zusammenfassungen der Analysen identifizieren Patienten nicht; deshalb können sie legal in LiSyM SEEK gespeichert und zwischen den Partnern ausgetauscht werden. Der Vorteil dieser Lösung ist, dass sie für alle Beteiligten leicht handhabbar ist. Aufseiten der klinischen Partner ist relativ wenig und leicht zu administrierende Software zu installieren. Sie benötigt keine Administratorrechte auf den Rechnern, auf denen sie läuft. Auf der anderen Seite ist der Automatisierungsgrad klein, die Software wird von den Partnern selbst gestartet, die Daten selbst zusammengeführt. Dies benötigt mehr Handarbeit, ist jedoch leichter zu sichern.

Zusammenhängende Daten können in SEEK gruppiert werden zu Assays, Studien und Investigations, und in Zusammenhang gebracht werden mit Standard OperatingProcedure Protokollen (SOPs), Beschreibungen der verwendeten biologischen Proben, resultierenden Computermodellen und Publikationen. Sie alle können auch in SEEK beschrieben und miteinander vernetzt werden, sofern sie aufeinander aufbauen oder in Beziehung stehen. Diese Strukturierung basiert auf entsprechenden Metadaten, welche die Zusammenhänge der Daten beschreiben. Daraus ergibt sich ein FAIRes Abbild der Daten und Metadaten des LiSyM-Forschungsnetzwerks (Abbildung 2).

Damit sind SEEK und unser darauf aufbauender Datenmanagement-Service, welcher auch wichtige Aspekte des User- und Erwartungsmanagements abdeckt, bestens geeignet, die Anforderungen an ein Datenmanagement-Konzept für ein solch verteiltes, kooperatives und interdisziplinäres Forschungskonsortium wie LiSyM zu erfüllen. Das wird komplettiert durch Angebote wie Beteiligung der Nutzer an der Planung der Weiterentwicklung der SEEK-Software und Trainings für die verschiedenen Nutzerkreise aus Labor, Theorie und klinischer Praxis. Dieses Gesamtpaket bieten wir im Rahmen von de.NBI auch anderen Nutzern an.

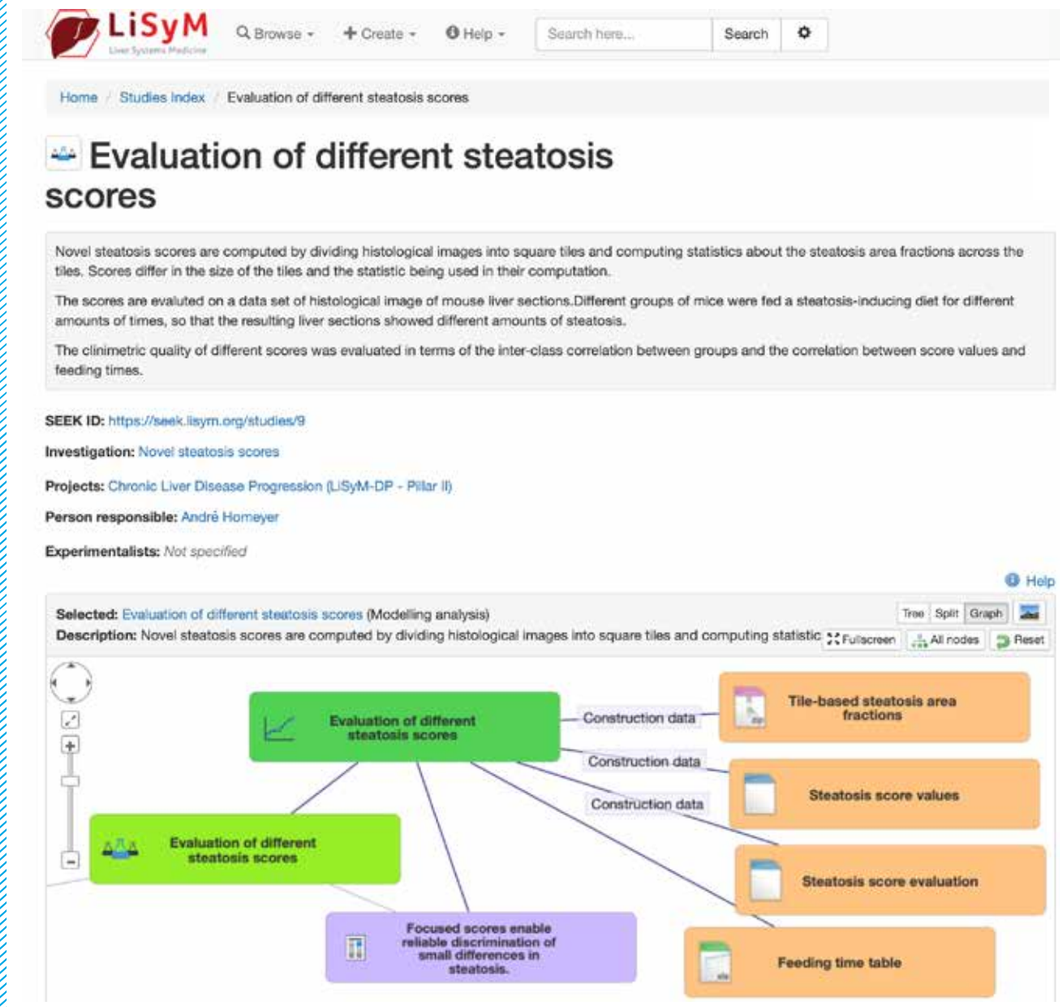


ABBILDUNG 2: Screenshot aus LiSyM SEEK zur Demonstration der Strukturierung zusammenhängender Datensätze (orange) und daraus resultierenden Publikationen (lila) zur Gruppierung in Assays und Studien (grün).

REFERENZEN [1] https://fair-dom.org **[2]** https://sis.id.ethz.ch/software/openbis.html

AUTOREN Martin Golebiewski[1] und Wolfgang Müller[1]
[1] *Heidelberger Institut für Theoretische Studien (HITS), Heidelberg*

DAS DEUTSCHE NETZWERK FÜR BIOINFORMATIK-INFRASTRUKTUR (de.NBI)

Das Deutsche Netzwerk für Bioinformatik-Infrastruktur (de.NBI) wird seit 2015 vom Bundesministerium für Bildung und Forschung (BMBF) gefördert, um Forschende in den Lebenswissenschaften bei der Auswertung großer Datenmengen in vielfältiger Weise zu unterstützen. Dazu bieten acht thematisch unterschiedlich aufgestellte Servicezentren ein breites Spektrum an bioinformatischen Tools und begleitenden Trainingskursen an. Außerdem steht die de.NBI-Cloud Rechenressourcen für die Analyse dieser Daten zur Verfügung.

DAS DEUTSCHE NETZWERK FÜR BIOINFORMATIK-INFRASTRUKTUR (de.NBI)

Im Deutschen Netzwerk für Bioinformatik-Infrastruktur (de.NBI) wird eine große Anzahl an Einzelprojekten, die an Universitäten und Forschungseinrichtungen angesiedelt sind, zu acht thematisch unterschiedlichen Servicezentren zusammengebunden. Die Durchführung der de.NBI- Aufgaben Service und Training liegt in der Verantwortung dieser Servicezentren. Darüberhinaus bietet de.NBI eine Cloud-Infrastruktur sowie ein Industrieforum an. Das de.NBI-Netzwerk wird von der Geschäftsstelle aus koordiniert.

32

Institutionen...

GEHÖREN ZUM NETZWERK.

42

Projekte ...

SIND IM NETZWERK EINGEBUNDEN.

250

Wissenschaftlerinnen und Wissenschaftler. ...

ARBEITEN IM NETZWERK.

DAS DEUTSCHE NETZWERK FÜR BIOINFORMATIK-INFRASTRUKTUR
de.NBI

THEMATISCHE SCHWERPUNKTE & SERVICEZENTREN:

- **HUMANE BIOINFORMATIK**
  HEIDELBERG CENTER FOR HUMAN BIOINFORMATICS (HD-HuB)
- **MIKROBIELLE BIOINFORMATIK**
  BIELEFELD-GIESSEN RESOURCE CENTER FOR MICROBIAL BIOINFORMATICS (BiGi)
- **PFLANZENBIOINFORMATIK**
  GERMAN CROP BIOGREENFORMATICS NETWORK (GCBN)
- **RNA-BIOINFORMATIK**
  RNA BIOINFORMATICS CENTER (RBC)
- **PROTEOMBIOINFORMATIK**
  BIOINFORMATICS FOR PROTEOMICS (BioInfra.Prot)
- **INTEGRATIVE BIOINFORMATIK**
  CENTER FOR INTEGRATIVE BIOINFORMATICS (CIBI)
- **BIODATENBANKEN**
  CENTER FOR BIOLOGICAL DATA (BioData)
- **DATENMANAGEMENT/SYSTEMBIOLOGIE**
  de.NBI SYSTEMS BIOLOGY SERVICE CENTER (de.NBI-SysBio)

- **STANDORTE DER SERVICEZENTREN-LEITUNGEN**
- **STANDORTE DER PARTNER**

DAS DEUTSCHE NETZWERK FÜR BIOINFORMATIK-INFRASTRUKTUR
de.NBI

BEITRAG DES de.NBI-NETZWERKS

zur Lösung des Datenproblems in den Lebenswissenschaften

Das de.NBI-Netzwerk besteht nun seit fünf Jahren. Es ist deshalb sinnvoll, die Aufgaben des Netzwerks zu beleuchten und nach den in der Zwischenzeit erzielten Ergebnissen zu fragen. Dazu führte Irena Maus, die in der de.NBI-Geschäftsstelle für Öffentlichkeitsarbeit verantwortlich ist, ein Interview mit dem de.NBI-Koordinator Alfred Pühler und dem de.NBI-Geschäftsstellenleiter Andreas Tauch.

IRENA MAUS: Das de.NBI-Netzwerk feiert sein fünfjähriges Jubiläum. Zu welchem Zweck wurde es eingerichtet?

ALFRED PÜHLER: Das de.NBI-Netzwerk wurde 2015 etabliert, um allen Forschenden in den Lebenswissenschaften eine Infrastruktur zur Verfügung zu stellen, die die Analyse umfangreicher Datenmengen mittels Bioinformatik ermöglicht. Diese Infrastruktur umfasste zunächst die Bereiche **Service** und **Training**. Im Servicebereich steht eine Vielzahl von Analyseprogrammen zur Verfügung, die zur Auswertung lebenswissenschaftlicher Daten genutzt werden kann. Neben diesem Dienstleistungsbereich ist der Schulungsbereich des de.NBI-Netzwerks von ausschlaggebender Bedeutung. Im Schulungsbereich wird der Umgang mit ausgewählten Programmen und den damit erzielten Ergebnissen vermittelt. Diese beiden Bereiche wurden in den letzten fünf Jahren konsequent ausgebaut. In der Zwischenzeit stehen mehr als 100 moderne Analyseprogramme zur Verfügung, deren Funktionsweise in über 280 Kursen vermittelt wurde. Insgesamt konnten bisher mehr als 5.000 Teilnehmende geschult werden.

IRENA MAUS: Wie ist dieses Netzwerk strukturiert und wie werden Entscheidungen getroffen?

ANDREAS TAUCH: Das de.NBI-Netzwerk hat sich bei seiner Etablierung thematisch ausgerichtet. Es besteht aus **acht Servicezentren**, die verschiedene Teildisziplinen in den Lebenswissenschaften abdecken, zum Beispiel Humane Bioinformatik, RNA-Bioinformatik oder Biodatenbanken. Das Netzwerk wird von einer **zentralen Koordinierungseinheit**, gelenkt, der der de.NBI-Koordinator und die Leiter der acht Servicezentren angehören. Dieses Gremium trifft sich vierteljährlich um anstehende Beschlüsse zu fassen. Es lässt sich dazu von **sieben Expertengruppen** aus dem Netzwerk beraten. Diese Vorgehensweise hat sich in den letzten Jahren bestens bewährt.

IRENA MAUS: Welche Entwicklung nahm das Netzwerk in den letzten fünf Jahren?

ALFRED PÜHLER: Neben den beiden Bereichen Dienstleistung und Schulung ergaben sich für das de.NBI-Netzwerk weitere Aufgaben. Hierzu zählte vor allem der Aufbau eines Rechnerbereichs, der de.NBI-Nutzern die Analyse von großen Datenmengen erlaubt. Das de.NBI-Netzwerk hat hier von Beginn an auf eine zukunftsorientierte Technik gesetzt und eine **de.NBI-Cloud** an mehreren Standorten errichtet. Eine weitere Aufgabe für das Netzwerk bestand im Aufbau von europäischen Kooperationen. Diese Aufgabe wurde durch den Beitritt Deutschlands zur **ELIXIR**-Organisation erleichtert. Schließlich wurde erfolgreich an der Etablierung eines industriellen Zweigs des de.NBI-Netzwerks gearbeitet. In den vergangenen Monaten wurde ein **de.NBI-Industrieforum** eingerichtet, das zurzeit 26 Firmen als Mitglieder zählt.

IRENA MAUS: Wie erfolgreich war die Etablierung der föderativen de.NBI-Cloud?

ANDREAS TAUCH: Die Etablierung einer eigenen Cloud wurde seit 2016 durch zusätzliche Fördermittel des BMBF möglich. Wir haben uns für den Aufbau einer föderativen Cloud an sechs deutschen Standorten entschieden. Das Projektmanagment dazu erfolgt zentral in der de.NBI-Geschäftsstelle. Das besondere Merkmal der **de.NBI-Cloud** ist, dass sie akademisch Forschenden als Bioinformatik-Infrastruktur kostenlos zur Verfügung gestellt wird. Der wissenschaftliche Erfolg der de.NBI-Cloud ist an ihren Zahlen abzulesen: über **700 angemeldete Forschende** mit über **200 laufenden Großprojekten!***

IRENA MAUS: Wie beteiligt sich de.NBI an der europäischen ELIXIR-Organisation?

ALFRED PÜHLER: ELIXIR ist ein europäisches Infrastruktur-Netzwerk, das sich die Aufgabe gestellt hat, alle Aspekte im Umgang mit lebenswissenschafltichen Daten in den Migliedsländern zu unterstützen. Damit verfolgt ELIXIR im europäischen Raum analoge Ziele wie das de.NBI-Netzwerk in Deutschland. Nach Beitritt Deutschands zu ELIXIR im Juli 2016 wurde das de.NBI-Netzwerk beauftragt, den **deutschen ELIXIR-Knoten** zu entwickeln. Dies wurde durch Beteiligung an ELIXIR-Aktivitäten umgesetzt. So wurden im Dienstleistungsbereich de.NBI-Auswerteprogramme über die ELIXIR-Organisation Nutzern europaweit zur Verfügung gestellt. Auch im Schulungsbereich erfolgte ein Abgleich mit ELIXIR-Partnern. Im Weiteren wurde die erfolgreich etablierte de.NBI-Cloud von einigen ELIXIR-Mitgliedsstaaten in Kooperationsprojekte eingebunden. Schließlich findet auch das de.NBI-Industrieforum auf ELIXIR-Ebene besondere Beachtung, da auch dort die Einbindung der europäischen Industrie in eine Bioinformatik-Infrastruktur vorangetrieben wird.

IRENA MAUS: Welche Aufgabe kommt dem de.NBI-Industrieforum zu?

ANDREAS TAUCH: Das **de.NBI-Industrieforum** stellt die jüngste Entwicklung des de.NBI-Netzwerks dar. Es ist ein **loser Verbund von derzeit 26 Firmen**, der sich im Laufe des Jahres 2019 gebildet hat. Im November haben sich Mitglieder des Forums erstmalig zu einer eintägigen Informationsveranstaltung in Berlin getroffen. Das Forum soll eine wissenschaftliche Zusammenarbeit zwischen de.NBI- und den Industriepartnern auf Projektebene ermöglichen, um das de.NBI-Knowhow der Auswertung großer Datenmengen in den Industriesektor zu übertragen. Die Mitgliedsfirmen haben wiederum Zugang zu den de.NBI-Trainingsaktivitäten und zu wissenschaftlichen de.NBI-Veranstaltungen und sie können selbst an der inhaltlichen Gestaltung des Forums mitwirken.

IRENA MAUS: Welche Anstrengungen wurden unternommen, um die Angebote und Dienstleistungen des de.NBI-Netzwerks langfristig zur Verfügung zu stellen?

ALFRED PÜHLER: In den letzten fünf Jahren konnte mit dem de.NBI-Projekt eine zukunftsorientierte Infrastruktur aufgebaut werden, deren Fortbestand durch einen **Verstetigungsschritt** gesichert werden sollte. Als de.NBI-Koordinator gehört dieser Verstetigungsauftrag zu meinen Hauptaufgaben. Intensive Recherchen haben gezeigt, dass eine Aufnahme des de.NBI-Netzwerks in die Leibniz-Gemeinschaft als Denkmöglichkeit in Frage kommt. Bis zu einer Übernahme durch Leibniz sind allerdings noch etliche Verhandlungen und Gespräche angesagt. Die angestrebte Verstetigung wird sich also nicht nahtlos an das zu Ende gehende de.NBI-Projekt anschließen. Erfreulicherweise hat das BMBF aber zugesagt, mit einer **Überbrückungsfinanzierung** das de.NBI-Netzwerk zunächst bis Ende des Jahres 2021 zu unterstützen. Die Mitglieder des de.NBI-Netzwerks sind für diese Lösung sehr dankbar und hoffen, dass in Zukunft die geplante Verstetigung ein langfristiges Bestehen des de.NBI-Netzwerks sichern wird.

Prof. Dr. Alfred Pühler
de.NBI-Koordinator (rechts)
puehler@cebitec.uni-bielefeld.de

Prof. Dr. Andreas Tauch
de.NBI-Geschäftsstellenleiter (links)
tauch@cebitec.uni-bielefeld.de

* Am Galaxy-Standort in Freiburg gibt es weitere ca. 10.000 Nutzende, die im wesentlichen Microjobs in der dortigen de.NBI-Cloud ausführen.

de.NBI-SERVICES
Tools, Workflows, Datenbanken, Consulting

Eine der Hauptaufgaben des de.NBI-Netzwerks liegt im Servicebereich. Hier wird ein vielfältiges Portfolio an Software, Webtools, Workflows und Datenbanken vorgehalten, das Forschenden aus den Lebenswissenschaften zur Analyse großer Datenmengen zur Verfügung steht. Neben statistischem Consulting gibt es auch Beratung zu den angebotenen Tools. Alle de.NBI-Tools sind Open Source.

100

de.NBI-SERVICES ...

STELLT MEHR ALS 100 TOOLS ZUR AUSWERTUNG VON GROSSEN DATENMENGEN IN DEN LEBENSWISSENSCHAFTEN ZUR VERFÜGUNG.

450.000

Nutzende ...

PRO MONAT.

Protein List Comparator EDGAR RNA-seq end-to-end workflow Excemplify Quality-standards Freiburg RNA tools webserver PIPmiR IPK-Blast-Server Github-repository-galaxytools BiBiServ tools INFO-RNA TPP Pan-Cancer-alignment-workflow microMUMMIE rightfield data-standardisation-and-conversion-service circBase iPATH KNIME Cellular phenotyping of microscope image data MORRE ReadXplorer SABIO RK services blockbuster GotohScan roddy CopraRNA tRNAdb PlabiPD TargetThermo specI SIACAT OTP Conveyor-workflows eggNOG NGS Pipelines CRISPR iTOL PIA Unique-peptide-finder GBIS galaxy rna workbench Patient-Searchtool SEEK SDA Hardware-Sharing PAA Vienna RNA package SABIO-RK PicTar mOTUs motifSearch pSILAC PLEXY RSVP SILVA IntaRNA MARNA DARIO MGX CARNA DEXSeq DELLY Bioinformatical consulting and statistical analysis of proteomics data IceLogo PIA memeris BRENDA e!DAL RNAsnoop ProMeTra EURISCO PlantsDB SNV-calling-pipeline Docker-images:-galaxy-stable PeptideShaker Enterotyping EBI-image-&-RBioFormats AntaRNA COMBINE-Archive-Toolkit CrossPlatformCommander GenomeRNAi doRiNA S-Peaker SILVAngs SpliceMap MeltDB segemehl ExpaRNA MITOS LocARNA BiVes Freiburg-Galaxy-Server RNAplex ProCon OpenMS BacDive snoStrip WaRSwap KNIME EMMA2 PANGAEA workflows-and-recipes Cloud/HPC IONiser Pan-Cancer-alignment-workflow

„Zur Auswertung biologischer Datensätze stellt de.NBI mehr als 100 bioinformatische Tools für Forschende zur Verfügung, inklusive Consulting durch Experten."

Rabeaa Alkhateeb
de.NBI-Servicekoordinatorin
contact@denbi.de
www.denbi.de/services

de.NBI-TRAINING
Workshops, Hackathons, Sommerschulen

Neben dem Service spielt vor allem der Trainingsbereich des de.NBI-Netzwerks eine bedeutsame Rolle. In zahlreichen Trainingskursen werden de.NBI-Nutzer im Umgang mit bioinformatischen Werkzeugen geschult und somit das Verständnis von erzielten Ergebnissen gestärkt. Aktuelle Entwicklungen im Bereich der Bioinformatik werden darüber hinaus in de.NBI-Symposien, speziellen Workshops und jährlich stattfindenden Sommerschulen aufgegriffen.

5.000

Teilnehmende ...

WURDEN IN de.NBI-KURSEN BISHER GESCHULT.

Advanced modeling with Copasi Analyzing metabolic networks with CellNetAnalyzer Applied Metaproteomics Workshop Bioimage Analysis Course Data Management For Plant Genomics & Phenomics Differential analysis of proteomic data using R Galaxy for linking bisulfite sequencing with RNA sequencing Galaxy workshop on HTS data analysis Genomics and Metagenomics training course Genomics training course Introduction to BRENDA and ProteinPlus Introduction to Python Programming Linux Command Line & Basic Scripting course Machine Learning in R Microscopy Image Analysis Course Nanopore Best Practice Workshop User Meeting de.NBI Cloud Proteomics and Metabolomics with OpenMS SILVA/BacDive Workshop: From Primer to Paper and Back Single-Cell Omics workshop Software Carpentry workshop Spring School "Computational Biology Starter" Statistical analysis & qualitative and quantitative comparison of lipidomics data Tool-Training for Proteomics Tools for Systems biology modeling and data exchange Training on microbial phylogeny and diversity analysis Metabolomics Data Clinic Data Interpretation of Whole-Genome and Exome Data in Cancer Research Statistics and Computing in Genome Data Science The Linux Command Line: From Basic Commands to Shell Scripting Phylogenetic reconstruction course DNA Methylation: Design to Discovery Eukaryote genome annotation workshop Big Data Training Course in Plant Genomics

280

Trainingskurse ...

WURDEN DURCHGEFÜHRT.

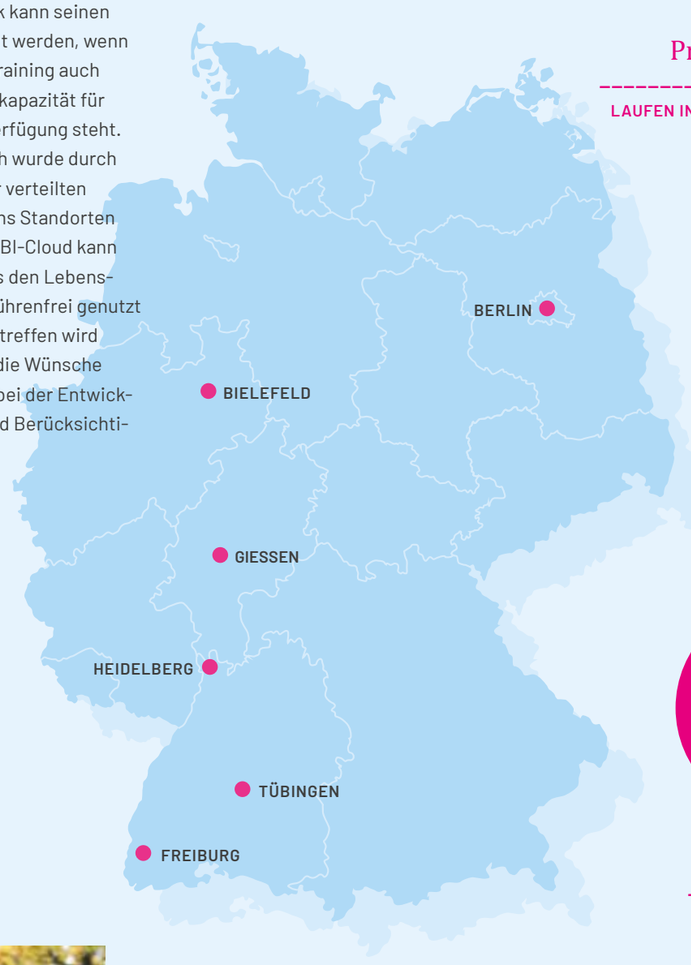
„Um unsere Tools optimal für die Datenauswertung zu nutzen, bieten wir verschiedenste Trainingskurse, Workshops, Hackathons und Sommerschulen an."

Daniel Wibberg
de.NBI-Trainingskoordinator
contact@denbi.de
www.denbi.de/training

de.NBI-CLOUD

Infrastruktur, Plattform und Software als Service

Das de.NBI-Netzwerk kann seinen Aufgaben nur gerecht werden, wenn neben Service und Training auch ausreichend Rechenkapazität für de.NBI-Nutzer zur Verfügung steht. Ein Compute-Bereich wurde durch die Etablierung einer verteilten de.NBI-Cloud an sechs Standorten geschaffen. Die de.NBI-Cloud kann von Forschenden aus den Lebenswissenschaften gebührenfrei genutzt werden. Über Nutzertreffen wird sichergestellt, dass die Wünsche von de.NBI-Nutzern bei der Entwicklung der de.NBI-Cloud Berücksichtigung finden.



250

Projekte ...

LAUFEN IN DER de.NBI-CLOUD.

38

Petabyte Speicherplatz

20.000

Rechenkerne

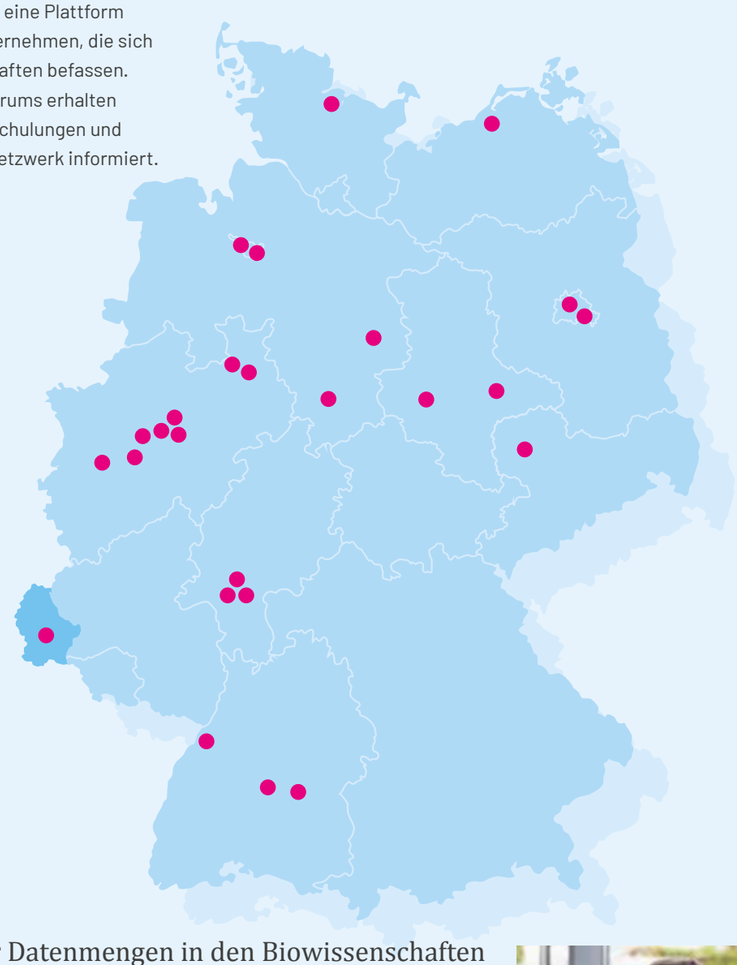
„Mit der Einrichtung der de.NBI-Cloud greifen wir in der Bioinformatik den internationalen Trend auf, skalierbare Ansätze zur Analyse großer Datenmengen zu entwickeln."

Peter Belmann
de.NBI-Cloud-Koordinator
cloud@denbi.de
www.denbi.de/cloud

de.NBI-INDUSTRIEFORUM

Software-Lösungen, Consulting, Networking

Das de.NBI-Industrieforum bietet eine Plattform zum Networking für Industrieunternehmen, die sich mit Big Data in den Biowissenschaften befassen. Mitglieder des de.NBI-Industrieforums erhalten Zugang zu de.NBI-Services und Schulungen und werden über Entwicklungen im Netzwerk informiert.



26

Mitglieder ...

IN DEUTSCHLAND UND LUXEMBURG.

„Die Analyse großer Datenmengen in den Biowissenschaften ist heutzutage auch für Industrieunternehmen äußerst relevant. Mit dem de.NBI-Industrieforum bieten wir einen Knowhow-Transfer zwischen Akademie und Industrie."

Manuel Wittchen
de.NBI-Industrieforum-Manager
contact@denbi.de
www.denbi.de/industrial-forum

Aktivitäten im de.NBI-Netzwerk

de.NBI-Vollversammlung 2018
_____
Berlin

Redaktionsteam der de.NBI-Geschäftsstelle
Von oben nach unten: Peter Belmann, Manuel Wittchen, Doris Jording, Daniel Wibberg, Andreas Tauch, Irena Maus, Tanja Dammann-Kalinowski, Alfred Pühler
_____
Bielefeld

DAS DEUTSCHE NETZWERK FÜR BIOINFORMATIK-INFRASTRUKTUR
de.NBI

de.NBI Training Course 2018: Introduction into targeted and untargeted metagenome analysis
_____
Gießen

de.NBI Summer School 2018: Riding the Data Life Cycle
_____
Braunschweig

Spring School 2019: Computational Biology Starter
_____
Gatersleben

de.NBI Training Course 2017: High-throughput genome analysis and comparative genomics
_____
Bielefeld

IMPRESSUM

Prof. Dr. Alfred Pühler
Deutsche Netzwerk für Bioinformatik-Infrastruktur (de.NBI)
de.NBI-Geschäftsstelle
Centrum für Biotechnologie
Universitätsstraße 27
33615 Bielefeld

Tel: +49 (0)521 106 8750
Fax: +49 (0)521 106 89046
E-Mail: contact@denbi.de

www. denbi.de
@denbiOffice
linkedin.com/company/de-nbi

Datum: Januar 2020

Bildnachweis:
iStockphoto, Pixabay, ROV-Team/GEOMAR (CC BY 4.0)

Design und Layout:
MEDIUM Werbeagentur GmbH, Bielefeld

Druck:
Bruns Druckwelt GmbH & Co. KG, Minden

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Fkz 031A532B
(de.NBI-Geschäftsstelle)

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

www.denbi.de