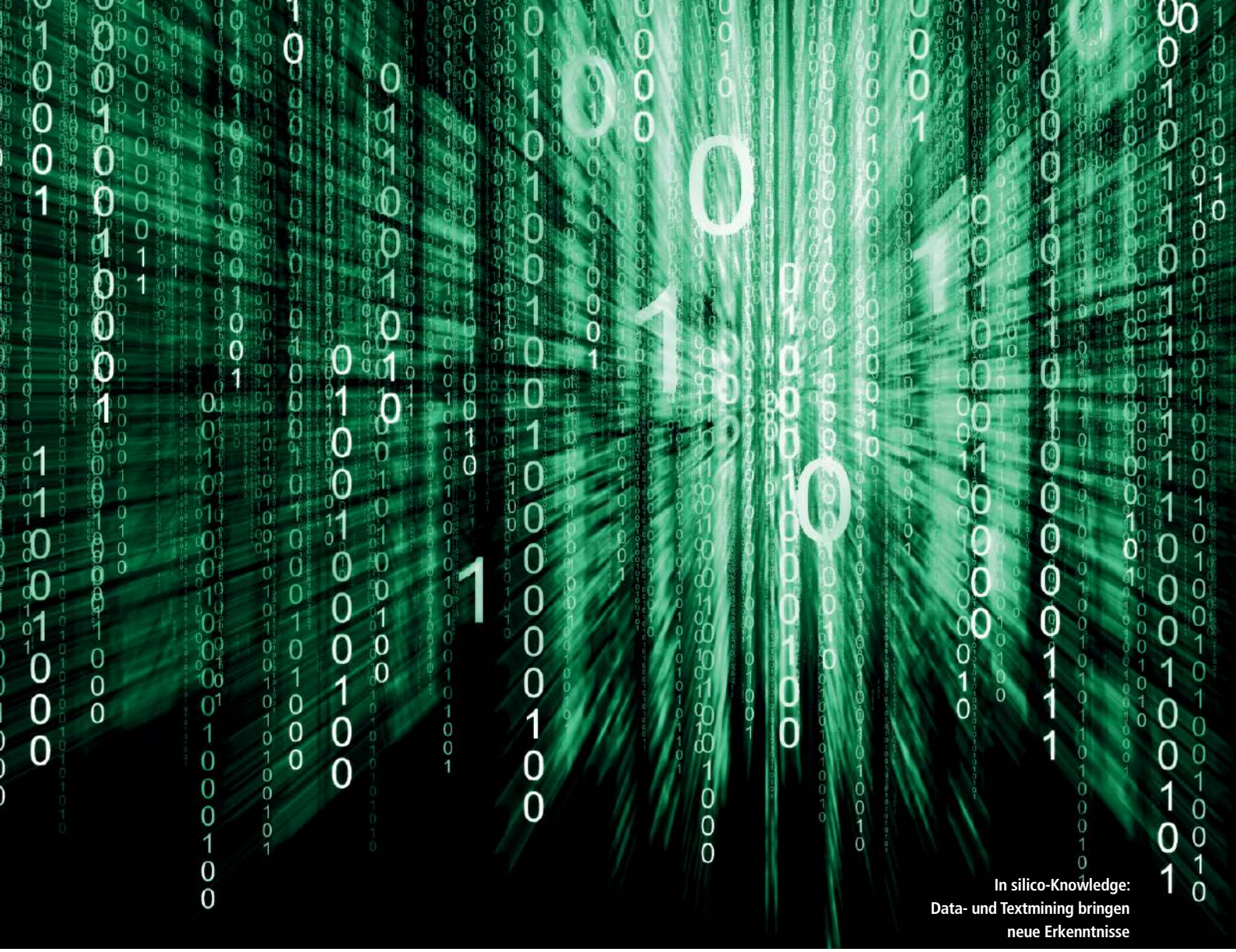


In silico-Knowledge:
Data- und Textmining bringen
neue Erkenntnisse

Goldgrube Big Scientific Data

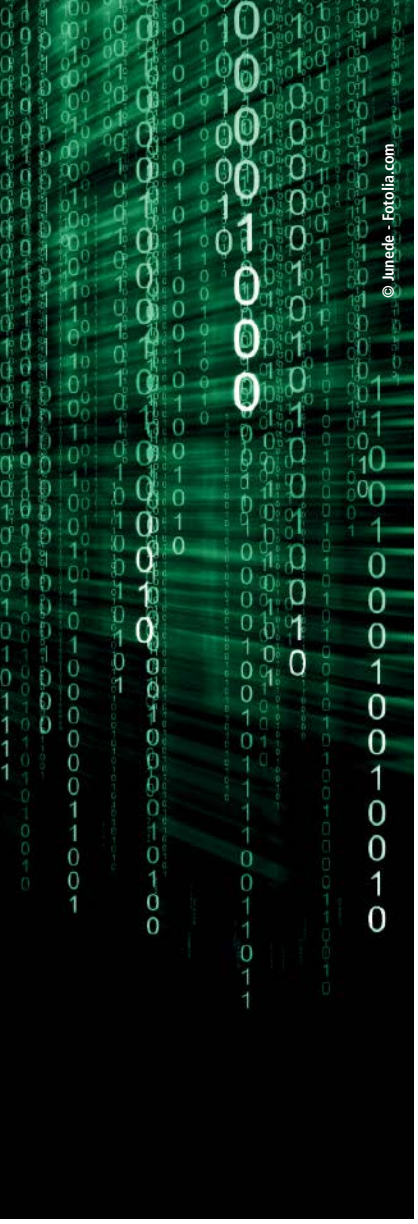
Große Datenmengen strukturieren, analysieren und nutzen

Die Masse an Datenberge aus der modernen Biologie nimmt explosionsartig zu. Die Ansammlung von massiven Rohdaten aus den neu entwickelten Omics-Technologien erfordert neue Strategien. Datamining und Textmining sind die schnellen computergestützten Lösungen, um große Datenmengen aufzuarbeiten. Für viele ist Textmining eine Goldgrube geworden. Aus sämtlichen wissenschaftlichen Arbeiten werden rasch neue Erkenntnisse und neue Zusammenhänge für das Gesundheitswesen ausgegraben.

Datenberge aus den Omics-Forschungsbereichen

In Wikipedia sind mehr als 35 Omics-Bereiche verzeichnet. Es handelt sich dabei um verschiedene Gebiete der modernen Biologie und beinhaltet die Erfassung von großen Datenmengen. Die Wortneubildung -omics hat sich inzwischen auch auf andere Teilgebiete ausgeweitet. Mittlerweile wurden eine Vielzahl der eingeführten Begriffsschöpfungen mit der Nachsilbe -omics auch in die deutsche Sprache übernommen [1]

Zu den bekannten Forschungsgebieten wie Genomik (Genetik), Proteomik (Eiweiße) und Metabolomik (Stoffwechsel) sind durch die modernen Methoden der Molekularbiologie viele andere klassische Fachgebiete erweitert worden. Ein besonders gutes Beispiel dafür ist die Erweiterung der klassischen Ernährung zur Nutriomics. Um ein ganzheitliches Verständnis für die Interaktion von Nährstoffen und menschlichen Organismus zu bekommen, wurden zur klassischen Ernährung weitere Omics-Technologien hinzugefügt. So hat sich die moderne Ernährung der Nutriomics mit den Gebieten Nutrigenetics (Nahrungsgenetik eines Individuum), Nutrigenomics (Ernährungsgenetik der Bevölkerung) und Nutriepigenetics (Epigenetik) ergänzt. Die neu gewonnenen Zusammenhänge und Erkenntnisse sollen vor allem die Entwicklung von Ernährungsstrategien für häufig vorkommende Herz-Kreislauf-



Erkrankungen oder Diabetes vorantreiben.

Sicherlich hat jeder Omics-Bereich seine eigene Besonderheit und seine eigenen charakteristischen Eigenschaften. Eine Omics-übergreifende Vernetzung wird für die meisten Omics-Bereiche noch angestrebt. Eine größere Herausforderung ist sicherlich die Verknüpfung von Omics-Bereichen aus der roten Biotechnologie (medizinische Biotechnologie) und der grünen Biotechnologie (Agrargentechnik). Sie erfordert einen multidisziplinären Ansatz und die enge Zusammenarbeit von Spezialisten und Forschungsteams.

Human Genomprojekt: Genomics

In der Flut der exponentiellen Entwicklung von riesigen Datenmengen hat im Besonderen das humane Genomproject gezeigt, dass die zielgerechte Nutzung und Analyse von humanen Sequenzen, ein komplexes Management für Datenspeicherung und anschließende Datenverarbeitung erforderlich macht.

Das humane Genomprojekt wurde in 2002 durch die Entschlüsselung der kompletten humanen DNA-Sequenz (Genomics), die in öffentlichen Datenbanken erhältlich ist, abgeschlossen. Mit der Bereitstellung von mehr als 3 Milliarden DNA-Basenpaaren war einer der ersten Schritte getan, um das gesamte Genom des menschlichen Organismus auf der Ebene der Desoxyribonukleinsäure (DNA) systematisch zu untersuchen. Obwohl der Code der DNA aus nur den 4 Basen, Adenin (A), Thymin (T), Cytosin (C) und Guanin (G) besteht, enthält sie die gesamte Information um den menschlichen Körper zu entwickeln.

Exom-Sequenzierung

Ziel war es, die wichtigen codierenden und funktionellen Bereiche der DNA genaustes zu identifizieren und relevante Bereiche für die biomedizinische Anwendung zu erkennen. Da nur ca. 5 % des gesamten Genoms für Proteine kodierend sind, mussten die funktionellen Bereiche der Sequenzen, den sogenannten exprimierten Exonbereiche, gefiltert werden. Daraus entwickelten sich die verschiedensten Hochdurchsatzsequenzierungs-Techniken, die unter dem Begriff Next Generation Sequencing Technologien (NGS) zusammengefasst werden. [2]

Das humane Genom Projekt wurde in 2008 auf das 1000-Genome-Projekct erweitert um eine Vielzahl an Sequenzen und deren Varianten miteinander vergleichen zu können [3]. Diese Erweiterung war natürlich eine neue Dimension in unserem Genomzeitalter, die nicht nur die medizinische und pharmazeutische Anwendungsbereiche vorantreibt, sondern auch die elektronische Datenspeicherung und Verarbeitung herausfordert. Die Speicherkapazitäten von Computern wurden für Rohdaten von Genomsequenzen im Umfang von sechs Petabyte (6 Millionen Gigabytes) erweitert.

Computergestützte Lösungen für BIG Scientific Data

Um Datenanalysen und Auswertungen im großen Maßstab durchzuführen und möglichst viele Eigenschaften parallel zu messen, benötigt es an der Entwicklung von geeigneten computergestützten Hilfsmitteln und ausgefeilten Analyseprogrammen. Sicherlich ist hier bei der Entwicklung von solchen Analyseprogrammen, die einen Organismus als ganzes System betrachten und auf allen Ebenen miteinander vergleichen, Expertenwissen von Spezialisten gefragt [4]. Intelligente Werkzeuge und computergesteuerte Lösungen zur Analyse, Klassifizierung und Mustererkennung von unstrukturierten Rohdaten sind mittlerweile als Datamining-Technologien frei zugänglich zu herhalten. Die Welt der DNA-Moleküle kann mit klinischen Daten und Bilderverarbeitungen und auch epidemischen Fragestellungen daher auf viel bessere Weise vernetzt und verknüpfen werden.

Interessant ist dabei auch die finanzielle Seite: es wurde prognostiziert, dass durch das Herausfiltern von relevante Daten für das Gesundheitswesen, das sogenannte Datamining, in den Vereinigten Staaten eine Kosteneinsparung von 450 Billionen Dollar einbringen würde [4].

Textmining zum Erkenntnisgewinn

In der Flut der exponentiellen Entwicklung von Rohdaten aus den Omics-Technologien und deren Analysen

hat auch eine Explosion der Anzahl von wissenschaftlichen Veröffentlichungen zur Folge. Wir sind längst in der Ära angekommen, wo große Datenmengen sogenannte BIG Science Data bzw. Massendaten-Wissenschaft um die ganze Welt gehen. Es ist daher ratsam für jeden der sich mit innovativen Techniken und wissenschaftlichen Projekten beschäftigt, vor einem Beginn eines Projektes die verfügbaren wissenschaftliche Arbeiten und Datenbestände systematisch auf Relevanzen hin zu untersuchen. Programme und Computerplattformen wurden entwickelt, die es ermöglichen im großen Maßstab wissenschaftliche Text zu filtern. Ein sogenannter Textmining-Prozess bietet sogar die Möglichkeit, die Gesamtheit der wissenschaftlichen Arbeiten, miteinbegriffen Abstracts, Papers, Posters und auch Patente, systematisch nach relevanten Fragestellungen zu untersuchen. Einige Firmen nutzen solche Textmining-Strategien mit dem Ziel neue Querverbindungen und neue Wirkstoffe zu entdecken. Für viele ist Textmining eine Goldgrube geworden. Aus allen aktuellen Publikationen werden rasch neue Erkenntnisse und neue Zusammenhänge ausgegraben.

Tabelle 1: Datenbanken zur automatischen Analyse von wissenschaftlichen Texten: Textmining

Name	Angaben URL
ALIBABA	http://alibaba.informatik.hu-berlin.de
AnneO'Tate	http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi
AskMedline	http://askmedline.nlm.nih.gov/ask/ask.php
Chillibot	http://www.chilibot.net
Copub	http://services.nbic.nl/copub/portal
Coremine medical	www.coremine.com
DPWP GenList	http://dpwebpage.nia.nih.gov
eTBLAST	http://bioinformatics.ca/links_directory/database/10750/etblast
Eurotos: Brain	http://www.euretos.com/brain
Facta+	http://www.nactem.ac.uk/facta/
Ferret	http://omictools.com/ferret-s10159.html
GeneValorization	http://www.bioguide-project.net
GoPubMed	http://www.gopubmed.com/web/gopubmed/www/GoPubMed/Search/index.html
HubMed	http://git.macropus.org/hubmed
Interact	http://www.ebi.ac.uk/intact
Liger-Cat	http://ligercat.ubio.org
MEDIE	http://www.nactem.ac.uk/medie
MedlineRanker	http://omictools.com/medlineranker-s5075.html
MedMiner	http://discover.nci.nih.gov/host/1999_medminer_abstract.jsp
Medstory	http://www.medstory.com
MedSum	http://webtools.mf.uni-lj.si/public/medsum.html
MeshPubMed	http://www.nlm.nih.gov/bsd/disted/meshtutorial
MiSearch	http://portal.ncibi.org/gateway/misearch.html
Novoseek	https://www.crunchbase.com/product/novoseek
PICO	http://pubmedhh.nlm.nih.gov/nlmd/pico/piconew.php
PMinstant	http://pminstant.com/
ProteinCorral	http://ubio.bioinfo.cnio.es/biotools/textmining/node/136
PubAnatomy	http://pubanatomy.org/
PubCrawler	http://pubcrawler.gen.tcd.ie
PubFlow	http://www.pubflow.uni-kiel.de
PubFocus	http://www.pubfocus.com
PubGet	http://pubget.com
Pubmatrix	http://pubmatrix.grc.nia.nih.gov
PubMedReminer	http://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi
PubNet	https://www.pubnet.org
PubReminer	http://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi
PubViz	https://connect.umms.med.umich.edu
Quertle	http://www.quertle.com
ReleMed	http://pages.citebite.com
Twease	http://omictools.com/twease-s5981.html
XplorMed	http://xplormed.ogic.ca

Die Liste der frei erhältlichen Software Programme zur automatischen Analyse von wissenschaftlichen Texten ist lang (Tabelle 1) Sicherlich bieten diese Programme oder Plattformen ein hinreichendes Hilfsmitteln um auf dem neusten Stand zu bleiben und Wissenslücken zu schließen [5]. Inzwischen hat sich dazu auch der neue Begriff Bibliomics etabliert und beinhaltet die komplette Ansammlung von biologischen Veröffentlichung und der damit assoziierten Information. Dazu entsprechend wurde auch der Begriff Bibliome geschaffen, und umfasst die Gesamtheit der biologischen Literatur.

Zusammenfassend zeigt sich, dass –omics und Data- und Textmining auf vielen Ebenen neue Möglichkeiten für die Wissenschaft eröffnet, die in den kommenden Jahren sicherlich noch an Bedeutung gewinnen wird. Während es früher vor allem eine Herausforderung war, Daten zu gewinnen, ist es in diesen Bereichen heute vor allem eine Herausforderung die großen Datenmengen zu strukturieren, zu analysieren und übergeordnete Muster zu erkennen und zu nutzen.

Referenzen

[1] Wikipedia. URL: https://de.wikipedia.org/wiki/-omik

[2] Neveling, K. und Hoischen, A.: Einführung in die Grundlagen der Hochdurchsatzsequenzierung. medizinische genetik, 26(2), 231-238 (2014)

[3] 1000 Genomes. URL: http://www.1000genomes.org

[4] Herland, M. *et al.*: A review of data mining using big data in health informatics. Journal of Big Data, 1(1), 1-35 (2014)

[5] Wikipedia.URL: https://en.wikipedia.org/wiki/List_of_text_mining_software

Kontakt

Dr. Jutta Wirth
Bioseminars & Wageningen University
Niederlande
jutta.wirth@wur.nl
www.biobioseminars.com