

Universität Stuttgart  
Institut für Biochemie und Technische Biochemie  
Allmandring 31, 70569 Stuttgart  
<http://www.itb.uni-stuttgart.de/>

Prof. Dr. Jürgen Pleiss  
Tel. (+49)711-685 63191  
Fax (+49)711-685 63196  
E-mail [Juergen.Pleiss@itb.uni-stuttgart.de](mailto:Juergen.Pleiss@itb.uni-stuttgart.de)

---

**Wintersemester 2023/2024**

# **Modul "Wissenschaftliche Methodik I"**

**(MSc Technische Biologie)**

## **Datenbanken**

**24.10.2023**

**Jan Range: Übungen "Datenbanken", V57.04, 8.11.2023, 8<sup>00</sup> - 9<sup>30</sup>**

# Datenbanken

## 1. Warum Datenbanken? Biologische Daten = "Big Data"

- Große, schnell wachsende Datenmengen
- Hohe Komplexität: Verknüpfung zwischen unterschiedlichen Daten und verteilten Datenquellen
- Schwach strukturierte Daten
- Fehlerhafte und inkonsistente Inhalte
- Fehlende Daten

## 2. Was sind Datenbanken?

- Datenbank: System zur elektronischen Datenverwaltung
- Strukturierung von komplexen Daten: relationales Datenmodell
- Datenbankmanagementsystem (DBMS): Datenbankverwaltung und Zugangskontrolle

# 1. Biologische Daten = "Big Data"

## 1. Große, schnell wachsende Datenmengen

### Gründe:

**Fallende Kosten** und größere Geschwindigkeit der Erzeugung neuer Daten

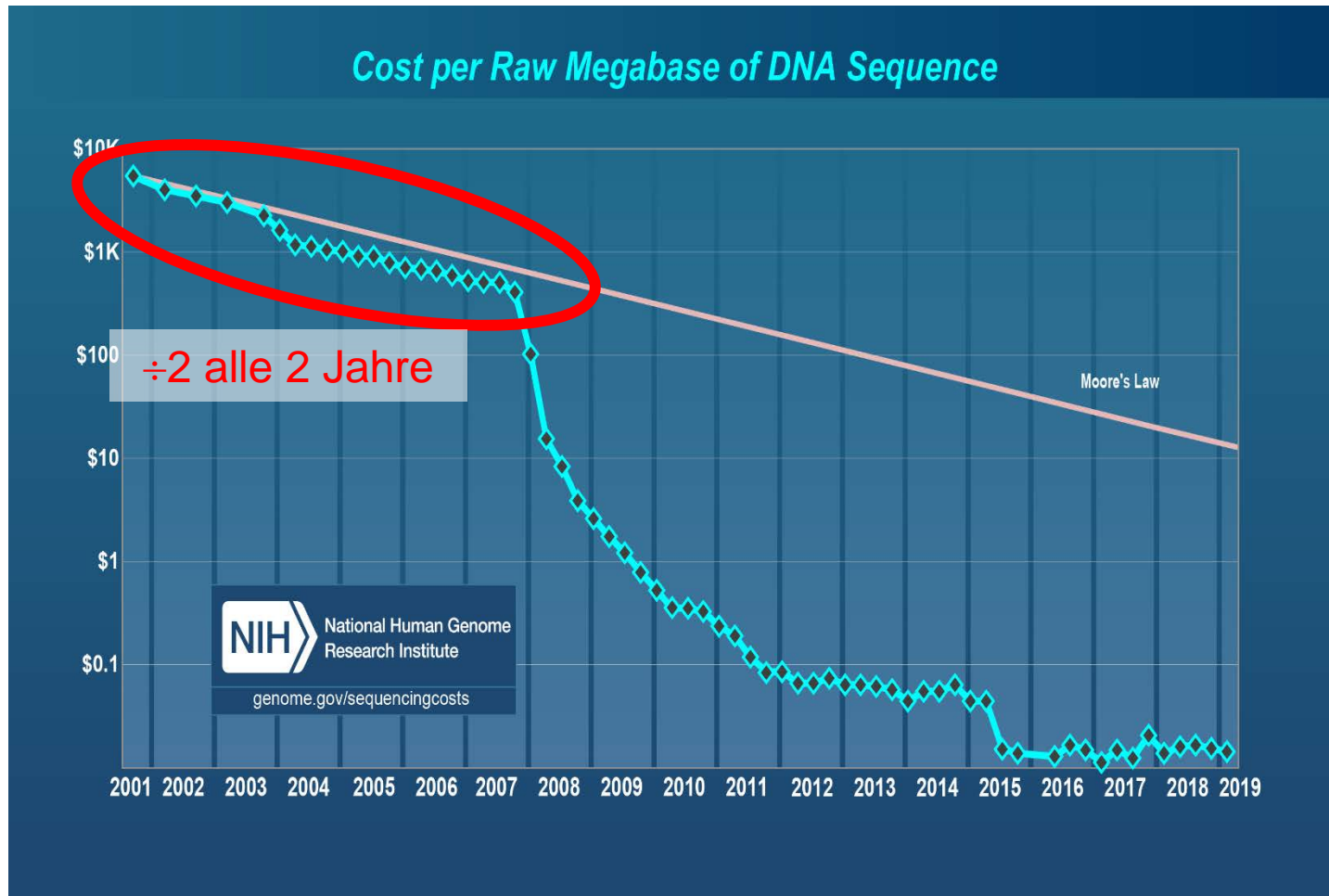
- DNA-Sequenzierung
- Hochdurchsatzverfahren für Screening: microfluidics, fluorescence activated cell sorting (FACS), Pipettierroboter → schnell, geringere Volumina

**Neue Methoden** und Messverfahren

- NMR Analytik ermöglicht die gleichzeitige Messung des Zeitverlaufs vieler Metaboliten
- Sequenzierung des Metagenoms ganzer Habitate
- Proteomik: Identifikation und Quantifizierung aller Proteine einer Zelle

# Fallende Kosten der DNA-Sequenzierung

## Kosten der DNA-Sequenzierung

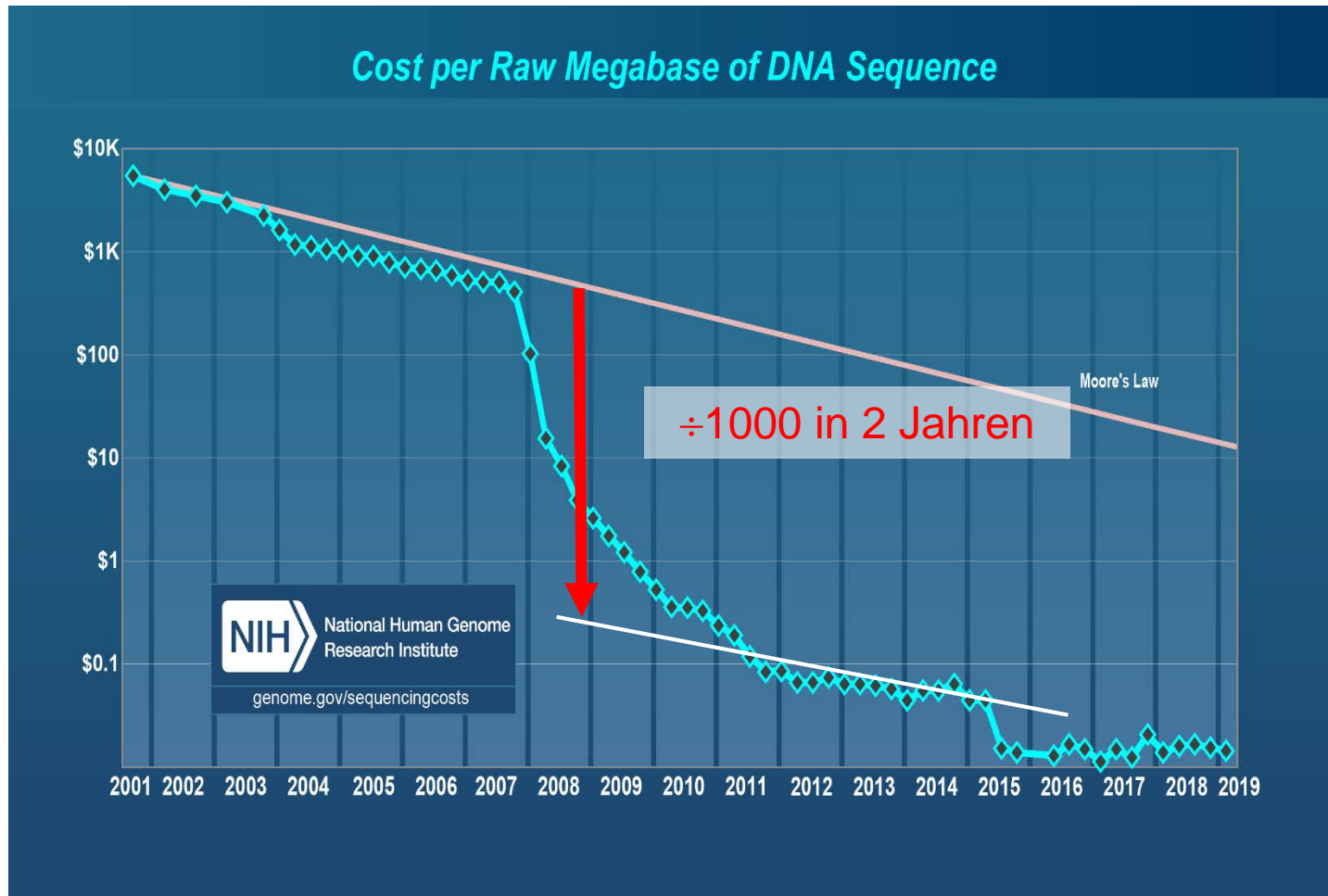


2001-2007 :  
Sanger-Sequenzierung  
(first generation)



[https://www.genome.gov/images/content/costpermb\\_2017.jpg](https://www.genome.gov/images/content/costpermb_2017.jpg)

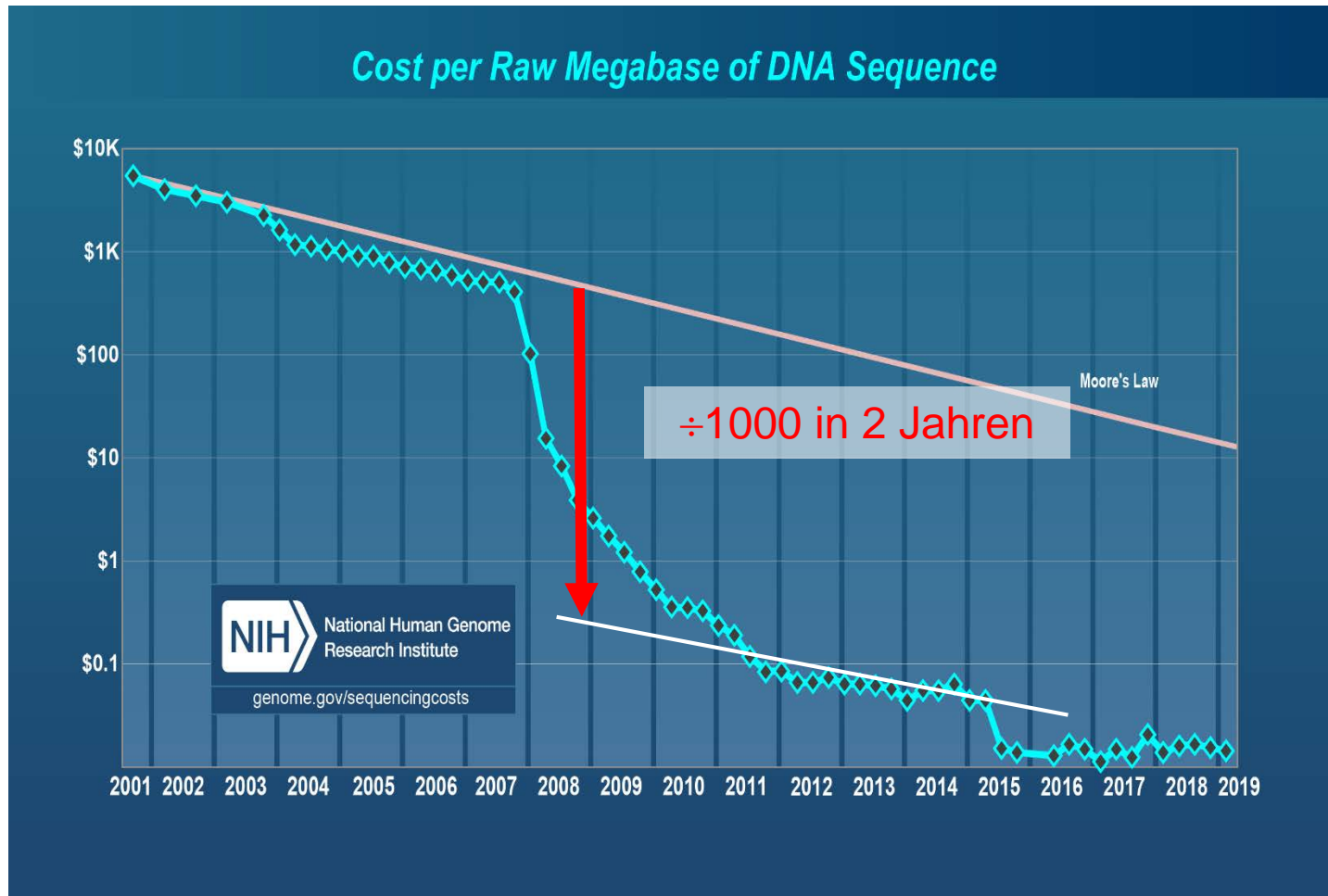
# Fallende Kosten der DNA-Sequenzierung



2008-2014 :  
454, Illumina, SOLiD  
(*second generation,*  
*next generation*)



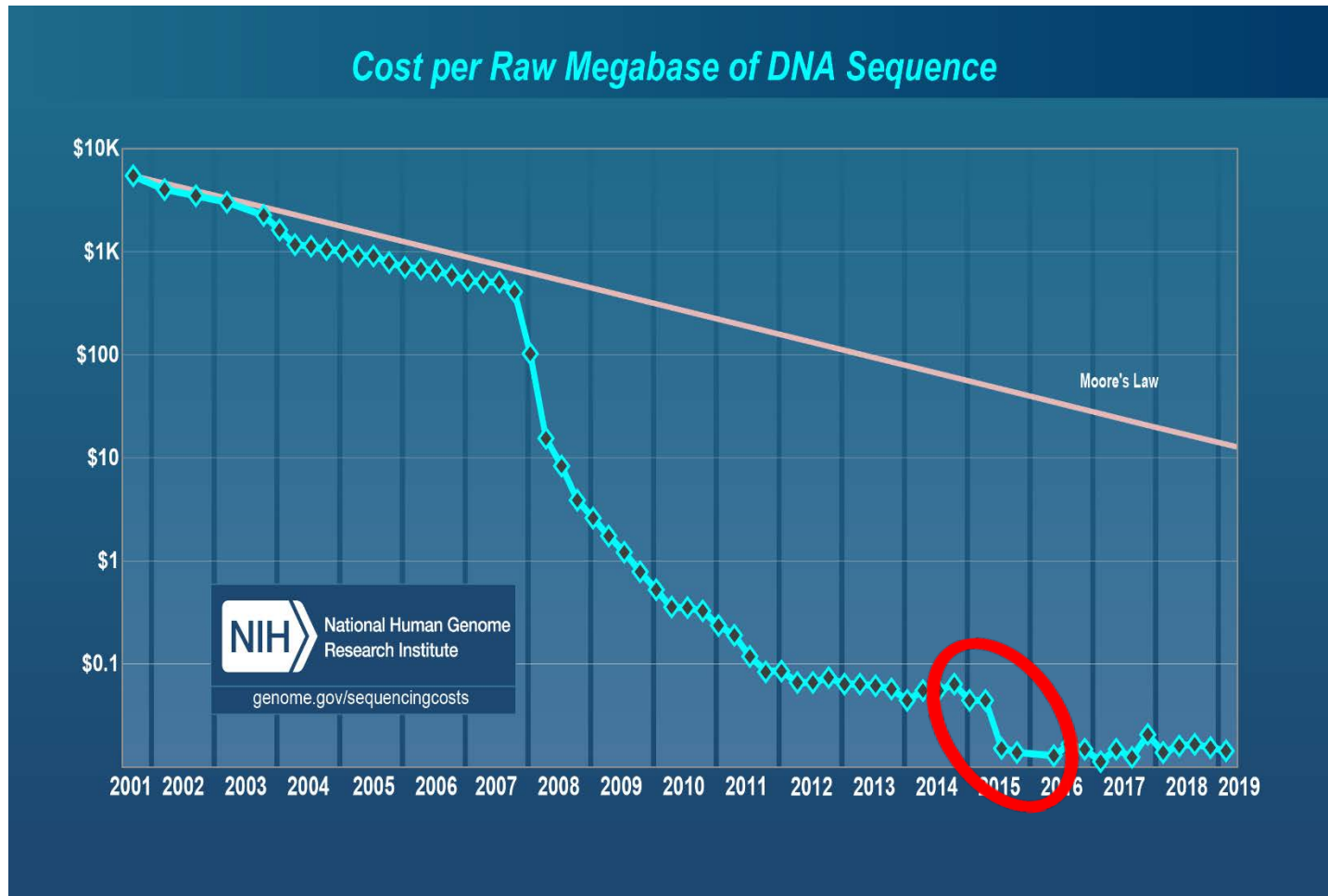
# Fallende Kosten der DNA-Sequenzierung



2008-2014 :  
454, Illumina, SOLiD  
(*second generation,  
next generation*)



# Fallende Kosten der DNA-Sequenzierung



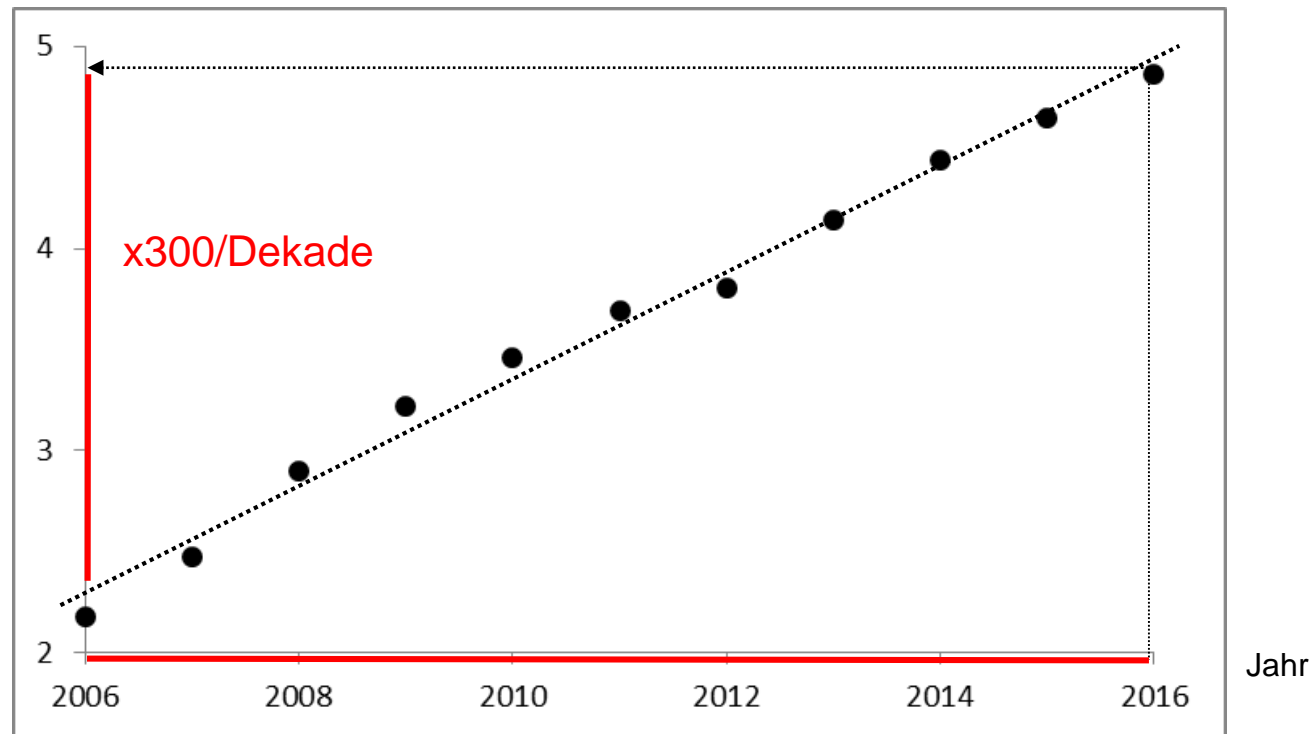
seit 2015 :  
single-molecule real-  
time, FRET, nanopore  
(*third generation*)



# Genom-Sequenzen

Anzahl der sequenzierten Genome

$\log_{10}$  (Zahl der sequenzierten Genome)



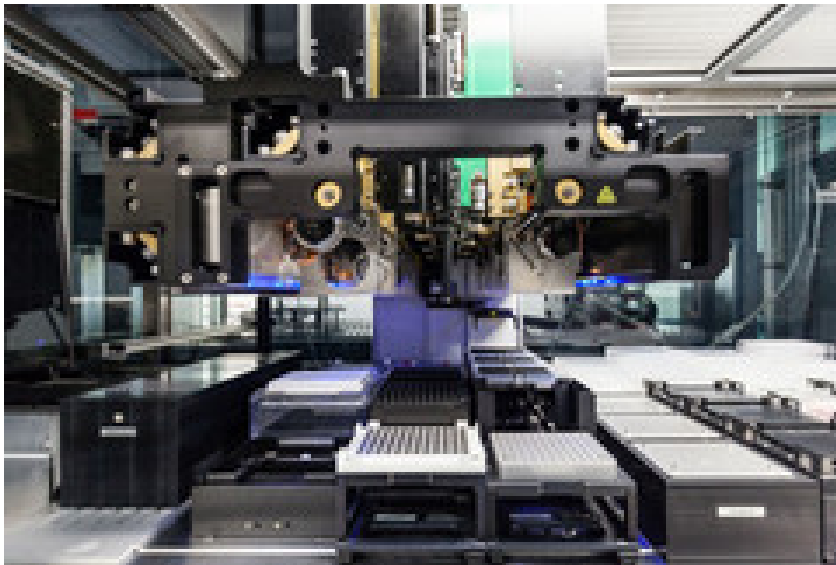
Genomes OnLine Database: <https://gold.jgi.doe.gov/statistics>



## Biologische Daten = Big Data

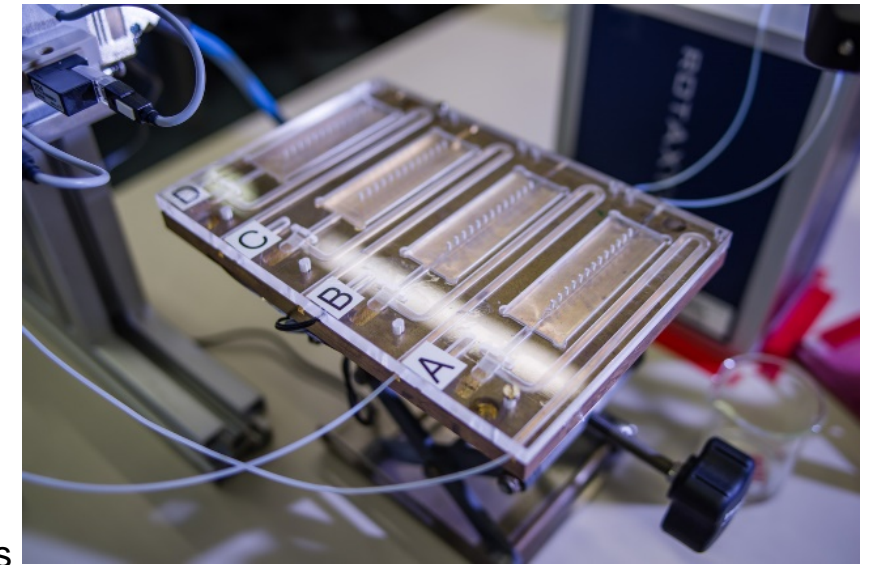
**High throughput Sequenzierung** (3.Generation):      Länge der reads: 100 kbp  
Kosten : 100 Mbp/€  
Produktivität: 10 Gbp/Tag ( **$10^6$**  Gene/Tag)

**High throughput assay Methoden:**      microtiter plate (96 wells à 100 µl) :  $10^3$  Experimente/Tag  
microdroplet (pl – nl) :  **$10^8$  -  $10^9$**  Experimente/Tag



Pipetting robot

<https://www.kiwi.tu-berlin.de/>



Microfluidics

<https://www.niemeyer-lab.de/applications/>

# Big Data in der Biologie

## 2. Hohe Komplexität: Verknüpfung zwischen unterschiedlichen Daten und (verteilten) Datenquellen

"-omics Daten" : unterschiedliche Datentypen, Datenmodelle, Datenbanken:

- Genom: Genomische Information einer Zelle oder eines Virus
- Metagenom: Gesamtheit der genomischen Information der Mikroorganismen eines Habitats
- Epigenom: Gesamtheit von epigenetischen Zuständen (Veränderung von DNA oder Histone)
- Transkriptom: Gesamtheit aller RNA-Moleküle einer Zelle
- Proteom: Gesamtheit aller Proteine in einer Zelle oder Gruppe von Zellen
- Metabolom: Gesamtheit aller Metaboliten einer Zelle
- Lipidom: Gesamtheit der zellulären Lipide
- Glycom: Gesamtheit der zellulären Zucker und Kohlenhydrate
- Fluxom: Gesamtheit der metabolischen Flüsse in einer Zelle
- Connectom: Netzwerk der Neuronen und Gehirnzellen

# Big Data in der Biologie

## 2. Hohe Komplexität: Verknüpfung zwischen unterschiedlichen Daten und (verteilten) Datenquellen

Beispiel: Verknüpfung von Informationen über ein Protein:

- Aminosäuresequenz → DNA-Sequenz, Lokalisation im Genom,...
- Proteinstruktur → Interaktion mit anderen Proteinen, mit Liganden,...
- Lokalisation in der Zelle → verschiedene Zustände der Zelle,...
- Bindungsaffinitäten und Funktion → Reaktionsbedingungen, Substrate,...

Beispiel eines UniProt-Eintrags mit " high-quality annotation": **P14779 (CPXB\_BACMB)**

Protein | Bifunctional cytochrome P450/NADPH--P450 reductase

Gene | cyp102A1


UniProtKB - P14779  
(CPXB\_BACMB)

### Organism


*Bacillus megaterium* (strain ATCC 14581 / DSM 32 / JCM 2506 / NBRC 15308 / NCIMB 9376 / NCTC 10342 / VKM B-512)

Status |  Reviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>


## Function<sup>i</sup>

Functions as a fatty acid monooxygenase (PubMed:3106359, PubMed:1727637, PubMed:16566047, PubMed:7578081, PubMed:11695892, PubMed:14653735, PubMed:16403573, PubMed:18004886, PubMed:17077084, PubMed:17868686, PubMed:18298086, PubMed:18619466, PubMed:18721129, PubMed:19492389, PubMed:20180779, PubMed:21110374, PubMed:21875028). Catalyzes hydroxylation of fatty acids at omega-1, omega-2 and omega-3 positions (PubMed:1727637, PubMed:21875028). Shows activity toward medium and long-chain fatty acids, with optimum chain lengths of 12, 14 and 16 carbons (lauric, myristic, and palmitic acids). Able to metabolize some of these primary metabolites to secondary and tertiary products (PubMed:1727637). Marginal activity towards short chain lengths of 8-10 carbons (PubMed:1727637, PubMed:18619466). Hydroxylates highly branched fatty acids, which play an essential role in membrane fluidity regulation (PubMed:16566047). Also displays a NADPH-dependent reductase activity in the C-terminal domain, which allows electron transfer from NADPH to the heme iron of the cytochrome P450 N-terminal domain (PubMed:3106359, PubMed:1727637, PubMed:16566047, PubMed:7578081, PubMed:11695892, PubMed:14653735, PubMed:16403573, PubMed:18004886, PubMed:17077084, PubMed:17868686, PubMed:18298086, PubMed:18619466, PubMed:18721129, PubMed:19492389, PubMed:20180779, PubMed:21110374, PubMed:21875028). Involved in inactivation of quorum sensing signals of other competing bacteria by oxidizing efficiently acyl homoserine lactones (AHLs), molecules involved in quorum sensing signaling pathways, and their lactonolysis products acyl homoserines (AHs) (PubMed:18020460).  18 Publications ▼

### Catalytic activity<sup>i</sup>


$\text{NADPH} + n \text{ oxidized hemoprotein} = \text{NADP}^+ + n \text{ reduced hemoprotein}$ .  19 Publications ▼


$\text{RH} + [\text{reduced NADPH--hemoprotein reductase}] + \text{O}_2 = \text{ROH} + [\text{oxidized NADPH--hemoprotein reductase}] + \text{H}_2\text{O}$ .


 19 Publications ▼

## Cofactor<sup>i</sup>


Protein has several cofactor binding sites:

FAD  1 Publication ▼


FMN  1 Publication ▼


heme  5 Publications ▼


## Enzyme regulation<sup>i</sup>


Inhibited by N-(12-imidazolyl-dodecanoyl)-L-leucine.  1 Publication ▼


## Kinetics<sup>i</sup>


kcat is 84.1 s<sup>-1</sup> for lauric acid (PubMed:16403573). kcat is 1480 min<sup>-1</sup> for palmitic acid. kcat is 1880 min<sup>-1</sup> for N-palmitoylglycine. kcat is 1690 min<sup>-1</sup> for N-palmitoyl-L-methionine. kcat is 610 min<sup>-1</sup> for N-palmitoyl-L-glutamine. kcat is 485 min<sup>-1</sup> for N-palmitoyl-L-glutamic acid. kcat is 1160 min<sup>-1</sup> for N-palmitoyl-L-leucine (PubMed:18004886). kcat is 28 s<sup>-1</sup> for lauric acid (PubMed:17868686). kcat is 2770 min<sup>-1</sup> for laurate/dodecanoate (PubMed:18721129). kcat is 77 for lauric acid (PubMed:19492389). kcat is 2770 min<sup>-1</sup> for laurate/dodecanoate (PubMed:20180779). kcat is 16400 min<sup>-1</sup> for arachidonate (PubMed:20180779). kcat is 91.4 for palmitic acid (PubMed:21110374).  7 Publications ▼


K<sub>M</sub>=250 μM for lauric acid at pH 7.4 at room temperature  1 Publication ▼


K<sub>M</sub>=34 μM for N-beta-oxolauroyl-DL-homoserine lactone  1 Publication ▼


K<sub>M</sub>=210 μM for N-beta-oxolauroyl-DL-homoserine  1 Publication ▼


K<sub>M</sub>=140 μM for N-lauroyl-DL-homoserine  1 Publication ▼


K<sub>M</sub>=322 μM for lauric acid at pH 7.5 and 15 degrees Celsius  1 Publication ▼

K<sub>M</sub>=265 μM for lauric acid  1 Publication ▼


K<sub>M</sub>=16 mM for indole  1 Publication ▼

K<sub>M</sub>=87.4 μM for laurate/dodecanoate at pH 7.0 and 25 degrees Celsius  1 Publication ▼

K<sub>M</sub>=230 μM for lauric acid at pH 7.4  1 Publication ▼

K<sub>M</sub>=87.4 μM for laurate/dodecanoate at 25 degrees Celsius  1 Publication ▼

K<sub>M</sub>=5.1 μM for arachidonate at 25 degrees Celsius  1 Publication ▼

K<sub>M</sub>=42.4 μM for palmitic acid at pH 7.4 and 30 degrees Celsius  1 Publication ▼

## Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Binding site <sup>i</sup>	264	Fatty acid  Combined sources   Curated 1 Publication			1
Site <sup>i</sup>	269	Important for catalytic activity  2 Publications			1
Metal binding <sup>i</sup>	401	Iron (heme axial ligand)  Combined sources 22 Publications			1
Binding site <sup>i</sup>	438	Fatty acid  Combined sources   Curated 1 Publication			1

## Regions

Feature key	Position(s)	Description	Actions	Graphical view	Length
Nucleotide binding <sup>i</sup>	489 – 494	FMN  Combined sources   1 Publication			6
Nucleotide binding <sup>i</sup>	536 – 539	FMN  Combined sources   1 Publication			4
Nucleotide binding <sup>i</sup>	570 – 572	FMN  Combined sources   1 Publication			3
Nucleotide binding <sup>i</sup>	578 – 580	FMN  Combined sources   1 Publication			3

## GO - Molecular function<sup>i</sup>

- aromatase activity Source: UniProtKB
- FMN binding Source: InterPro
- heme binding Source: InterPro
- identical protein binding Source: IntAct
- iron ion binding Source: UniProtKB
- NADPH-hemoprotein reductase activity Source: UniProtKB
- oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen Source: UniProtKB

[Complete GO annotation...](#)












## Keywords<sup>i</sup>

Molecular function	Monooxygenase, Oxidoreductase
Biological process	Electron transport, Transport
Ligand	FAD, Flavoprotein, FMN, Heme, Iron, Metal-binding, NADP

## Enzyme and pathway databases

BioCyc <sup>i</sup>	MetaCyc:MONOMER-17698.
BRENDA <sup>i</sup>	1.14.14.1. 656. 1.6.2.4. 656.

## Names & Taxonomy<sup>i</sup>


Protein names <sup>i</sup>	<p><i>Recommended name:</i></p> <p><b>Bifunctional cytochrome P450/NADPH--P450 reductase</b>  Curated</p> <p><i>Alternative name(s):</i></p> <ul style="list-style-type: none"><li>• Cytochrome P450(BM-3)  1 Publication ▼</li><li>• Cytochrome P450BM-3  1 Publication ▼  Imported ▼</li><li>• Fatty acid monooxygenase  1 Publication ▼</li><li>• Flavocytochrome P450 BM3  2 Publications ▼</li></ul> <p><u>Including the following 2 domains:</u></p> <ul style="list-style-type: none"><li>• Cytochrome P450 102A1 (EC:1.14.14.1  7 Publications ▼)</li><li>• NADPH--cytochrome P450 reductase (EC:1.6.2.4  7 Publications ▼)</li></ul>
Gene names <sup>i</sup>	<p>Name: <b>cyp102A1</b>  Imported ▼</p> <p>Synonyms: cyp102</p> <p>ORF Names: BG04_163  Imported ▼</p>
Organism <sup>i</sup>	Bacillus megaterium (strain ATCC 14581 / DSM 32 / JCM 2506 / NBRC 15308 / NCIMB 9376 / NCTC 10342 / VKM B-512)
Taxonomic identifier <sup>i</sup>	1348623 [NCBI]
Taxonomic lineage <sup>i</sup>	Bacteria > Firmicutes > Bacilli > Bacillales > Bacillaceae > Bacillus > 
Proteomes <sup>i</sup>	UP000031829 Component <sup>i</sup> : Chromosome



## Subcellular location<sup>i</sup>

- Cytoplasm  By similarity

### GO - Cellular component<sup>i</sup>

- cytoplasm  Source: UniProtKB-SubCell




[Complete GO annotation...](#)

### Keywords - Cellular component<sup>i</sup>










Cytoplasm

## Pathology & Biotech<sup>i</sup>

### Biotechnological use<sup>i</sup>

This protein is a target of protein engineering. Its selectivity-directing and activity-enhancing mutations have been extensively studied and the designed mutations allow this enzyme to act on non-native substrates and/or in order to enhance production of synthetically desirable end-products.  Curated  3 Publications 

### Mutagenesis

Feature key	Position(s)	Description	Actions	Graphical view	Length
Mutagenesis <sup>i</sup>	48	R → Q or S: 2-3-fold decrease in binding affinity for N-myristoyl-L-methionine as substrate.  1 Publication 			1
Mutagenesis <sup>i</sup>	75	A → G: Higher activity in the hydroxylation of highly branched fatty acids; when associated with V-88 and Q-189.  1 Publication 			1
Mutagenesis <sup>i</sup>	83	A → F: 800-fold binding affinity for laurate as substrate. High coupling of NADPH consumption to laurate formation. Very much more effective in indole hydroxylation. Favors omega-2 hydroxylation. Significantly higher rates of NADPH consumption in the absence of substrate. No temperature-dependent shifts to low-spin in complex with palmitate.  1 Publication 			1



## Chemistry databases

DrugBank <sup>i</sup>	<a href="#">DB08086</a> . N-[12-(1H-imidazol-1-yl)dodecanoyl]-L-leucine. <a href="#">DB03440</a> . N-Hexadecanoylglycine. <a href="#">DB04257</a> . Palmitoleic Acid.
-----------------------	---

## PTM / Processing<sup>i</sup>

### Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Chain <sup>i</sup> (PRO_0000052205)	1 – 1049	Bifunctional cytochrome P450/NADPH--P450 reductase <a href="#">Add</a> <a href="#">BLAST</a>			1049

## Expression<sup>i</sup>

### Induction<sup>i</sup>

By pentobarbital (PubMed:[1544926](#), PubMed:[3106359](#)). Expression is negatively regulated by repressor bm3R1 at the transcriptional level (PubMed:[1544926](#)). [2 Publications](#) ▼

## Interaction<sup>i</sup>

### Binary interactions<sup>i</sup>

With	Entry	#Exp.	IntAct	Notes
itself		2	<a href="#">EBI-7701704</a> , <a href="#">EBI-7701704</a>	

### GO - Molecular function<sup>i</sup>

- [identical protein binding](#) [Source: IntAct](#) ▼

[Complete GO annotation...](#)

### Protein-protein interaction databases

MINT <sup>i</sup>	<a href="#">MINT-8313368</a> .
-------------------	--------------------------------

## Chemistry databases

BindingDB<sup>i</sup> P14779.

## Structure<sup>i</sup>

### Secondary structure



Legend: ■ Helix ■ Turn ■ Beta strand ■ PDB Structure known for this area

[Show more details](#)

### 3D structure databases























Select the link destinations: <input checked="" type="radio"/> PDB <sup>i</sup> <input type="radio"/> RCSB PDB <sup>i</sup> <input type="radio"/> PDBj <sup>i</sup>	PDB entry	Method	Resolution (Å)	Chain	Positions	PDBsum
	<a href="#">1BU7</a>	X-ray	1.65	A/B	<a href="#">2-456</a>	<a href="#">[»]</a>
	<a href="#">1BVY</a>	X-ray	2.03	A/B	<a href="#">2-459</a>	<a href="#">[»]</a>
				F	<a href="#">460-650</a>	<a href="#">[»]</a>
	<a href="#">1FAG</a>	X-ray	2.70	A/B/C/D	<a href="#">2-472</a>	<a href="#">[»]</a>
	<a href="#">1FAH</a>	X-ray	2.30	A/B	<a href="#">2-472</a>	<a href="#">[»]</a>

# Family & Domains<sup>i</sup>


## Domains and Repeats

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain <sup>i</sup>	483 – 622	Flavodoxin-like <div>PROSITE-ProRule annotation</div>	<div>Add</div> <div>BLAST</div>	<div></div>	140
Domain <sup>i</sup>	660 – 892	FAD-binding FR-type <div>PROSITE-ProRule annotation</div>	<div>Add</div> <div>BLAST</div>	<div></div>	233

## Region

Feature key	Position(s)	Description	Actions	Graphical view	Length
Region <sup>i</sup>	2 – 472	Cytochrome P450	 Add  BLAST	<div><div></div></div>	471
Region <sup>i</sup>	21 – 30	Fatty acid binding  Combined sources  Curated  1 Publication 		<div><div></div></div>	10
Region <sup>i</sup>	76 – 88	Fatty acid binding  Combined sources  Curated  1 Publication 	 Add  BLAST	<div><div></div></div>	13
Region <sup>i</sup>	182 – 189	Fatty acid binding  Combined sources  Curated  1 Publication 		<div><div></div></div>	8
Region <sup>i</sup>	329 – 331	Fatty acid binding  Combined sources  Curated  1 Publication 		<div><div></div></div>	3
Region <sup>i</sup>	473 – 1049	NADPH--P450 reductase	 Add  BLAST	<div><div></div></div>	577

## Sequence similarities<sup>i</sup>

In the N-terminal section; belongs to the [cytochrome P450 family](#).  Curated

## Phylogenomic databases

eggNOG <sup>i</sup>	<a href="#">ENOG4107EER</a> . Bacteria. <a href="#">COG0369</a> . LUCA.
KO <sup>i</sup>	<a href="#">K14338</a> .

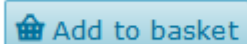
## Family and domain databases

Gene3D <sup>i</sup>	<a href="#">3.40.50.360</a> . 1 hit.
InterPro <sup>i</sup>	<a href="#">View protein in InterPro</a> <a href="#">IPR023206</a> . Bifunctional_P450_P450_red. <a href="#">IPR001128</a> . Cyt_P450. <a href="#">IPR017972</a> . Cyt_P450_CS. <a href="#">IPR003097</a> . FAD-binding_1. <a href="#">IPR017927</a> . Fd_Rdtase_FAD-bd. <a href="#">IPR001094</a> . Flavodoxin-like. <a href="#">IPR008254</a> . Flavodoxin/NO_synth. <a href="#">IPR001709</a> . Flavoprot_Pyr_Nucl_cyt_Rdtase. <a href="#">IPR029039</a> . Flavoprotein-like_dom. <a href="#">IPR001433</a> . OxRdtase_FAD/NAD-bd. <a href="#">IPR017938</a> . Riboflavin_synthase-like_b-brl.
Pfam <sup>i</sup>	<a href="#">View protein in Pfam</a> <a href="#">PF00667</a> . FAD_binding_1. 1 hit. <a href="#">PF00258</a> . Flavodoxin_1. 1 hit. <a href="#">PF00175</a> . NAD_binding_1. 1 hit. <a href="#">PF00067</a> . p450. 1 hit.
PIRSF <sup>i</sup>	<a href="#">PIRSF000209</a> . Bifunctional_P450_P450R. 1 hit.
PRINTS <sup>i</sup>	<a href="#">PR00369</a> . FLAVODOXIN. <a href="#">PR00371</a> . FPNCR.
SUPFAM <sup>i</sup>	<a href="#">SSF48264</a> . SSF48264. 1 hit. <a href="#">SSF52218</a> . SSF52218. 1 hit. <a href="#">SSF63380</a> . SSF63380. 1 hit.
PROSITE <sup>i</sup>	<a href="#">View protein in PROSITE</a> <a href="#">PS00086</a> . CYTOCHROME_P450. 1 hit. <a href="#">PS51384</a> . FAD_FR. 1 hit. <a href="#">PS50902</a> . FLAVODOXIN_LIKE. 1 hit.

# Sequence<sup>i</sup>

Sequence status<sup>i</sup>: Complete.

P14779-1 [UniParc]



« Hide

10	20	30	40	50
MTIKEMPQPK	TFGELKNLPL	LNTDKPVQAL	MKIADELGEI	FKFEAPGRVT
60	70	80	90	100
RYLSSQRLIK	EACDESRFDK	NLSQALKFVR	DFAGDGLFTS	WTHEKNWKKA
110	120	130	140	150
HNILLPSFSQ	QAMKGYHAMM	VDIAVQLVQK	WERLNADEHI	EVPEMTRLT

**Length:** 1,049

**Mass (Da):** 117,781

**Last modified:** January 23, 2007 - v2

**Checksum:**<sup>i</sup> B0BE61F8A2EE33D5

BLAST

GO

## Sequence databases

Select the link

destinations:

☒ EMBL<sup>i</sup>

☐ GenBank<sup>i</sup>

☐ DDBJ<sup>i</sup>

PIR<sup>i</sup>

RefSeq<sup>i</sup>

J04832 Genomic DNA. Translation: AAA87602.1.

CP009920 Genomic DNA. Translation: AJI21949.1.

S87512 Genomic DNA. Translation: AAK19020.1.

A34286.

WP\_034650526.1. NZ\_JJMH01000056.1.

## Genome annotation databases

EnsemblBacteria<sup>i</sup>

GeneID<sup>i</sup>




KEGG<sup>i</sup>

AJI21949; AJI21949; BG04\_163.

29911283.

bmeg:BG04\_163.

## Similar proteins<sup>i</sup>

90% Identity		50% Identity					
Entry	Cluster members	Organisms	Length	Cluster ID	Cluster name	Size	
P14779	UPI0000110A76 UPI000181D221 UPI0000110A75 UPI0000110989 UPI000252CACA UPI0000E69E3C UPI000533FF6F UPI0000111590 F2Q6X4 K7N7U0 +36	Bacillus megaterium (strain ATCC 14581 / DSM 32 / JCM 2506 / NBRC 15308 / NCIMB 9376 / NCTC 10342 / VKM B-512) Bacillus megaterium Bacillus megaterium (strain DSM 319) Bacillus megaterium Q3 Bacillus aryabhattai B8W22 Bacillus sp. FJAT-21351 Bacillus megaterium (strain WSH-002) Bacillus sp. Leaf75 Bacillus flexus Bacillus aryabhattai And more	1,049	UniRef90_P14779	Cluster: Bifunctional cytochrome P450/NADPH--P450 reductase	47	  

Full view

## Entry information<sup>i</sup>

Entry name <sup>i</sup>	CPXB_BACMB		
Accession <sup>i</sup>	Primary (citable) accession number: <b>P14779</b> Secondary accession number(s): A0A0B6AQ66, Q9AE23		
Entry history <sup>i</sup>	Integrated into UniProtKB/Swiss-Prot:	April 1, 1990	
	Last sequence update:	January 23, 2007	
	Last modified:	July 5, 2017	

# Big Data in der Biologie

## 3. Schwach strukturierte Daten

- Fehlen einer Ontologie (= formales Repräsentationssystem)

Beispiele: Name desselben Proteins: lipase oder triglyceride hydrolase oder esterase oder hydrolase oder...

Einheiten: katalytische Aktivität in  $\text{s}^{-1}$  oder in  $\text{min}^{-1}$  oder ...

Messmethoden: Proteinkonzentration über Bradford-Test oder über Absorption bei 280 nm oder ...

# Big Data in der Biologie

## 3. Schwach strukturierte Daten

- Fehlen einer Ontologie (= formales Repräsentationssystem)
- Überwiegend werden Daten als Freitext, Abbildung oder Tabelle publiziert

Beispiel: eine Tabelle oder eine Abbildung

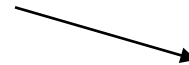


Table 3. Initial reaction rates, conversion after 1 h and electron coupling efficiency of wild type CYP153 A-CPR and RPIS variants. The highest values for each property are in bold. Average values and standard deviations were derived from triplicate experiments.						
	WT	L115 K	S120D	D153 K	K166Q	E425 L
Initial rate [ $\mu\text{mol}/\text{min} \cdot \mu\text{mol}$ ]	<b>23.2 <math>\pm</math> 0.7</b>	1.9 $\pm$ 0.1	5.3 $\pm$ 0.7	2.2 $\pm$ 0.2	10.6 $\pm$ 0.6	2.2 $\pm$ 0.6
Conversion after 1 h [%]	<b>80.1<sup>a</sup></b>	27.4 $\pm$ 1.7	44.4 $\pm$ 6.0	27.2 $\pm$ 0.4	52.8 $\pm$ 3.7	29.8 $\pm$ 6.1
Electron coupling efficiency [%]	53.3 $\pm$ 2.2	54.6 $\pm$ 1.6	45.9 $\pm$ 1.6	58.1 $\pm$ 2.0	<b>63.7 <math>\pm</math> 0.6</b>	54.0 $\pm$ 1.3

<sup>a</sup> one sample analyzed

ChemistrySelect 2016, 6, 1243–1251    Wiley Online Library    1247    © 2016 Wiley

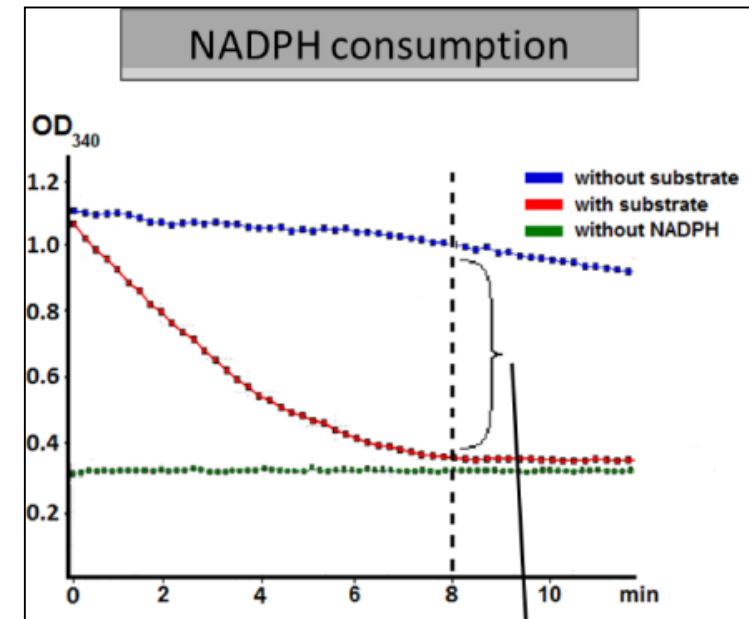
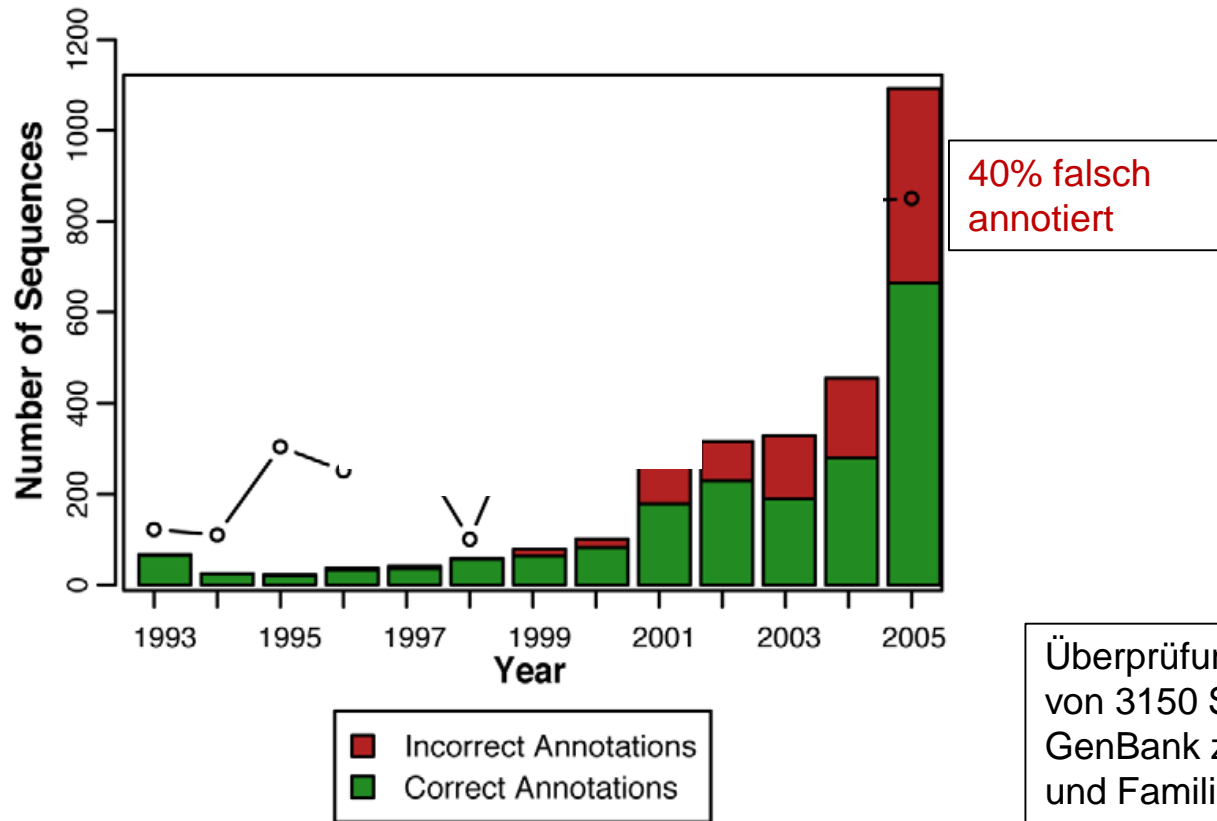


Fig. 2 Illustration of the measurement for the coupling efficiency.



# Big Data in der Biologie

## 4. Fehlerhafte und inkonsistente Inhalte



Überprüfung der Zuordnung von 3150 Sequenzen der GenBank zu Superfamilie und Familie

Superfamily	Family	E.C. No.
Enolase <sup>1</sup>	Enolase <sup>1</sup>	4.2.1.11
	Galactonate dehydratase	4.2.1.6
	Mandelate racemase	5.1.2.2
	Glucarate dehydratase	4.2.1.40
	Methylaspartate ammonia-lyase	4.3.1.2
	ortho-succinyl benzoate synthase	4.2.1.113
	Dipeptide epimerase	—
	Chloromuconate cycloisomerase	5.5.1.7
	Muconate cycloisomerase	5.5.1.1
	L-fuconate dehydratase	4.2.1.68
Crotonase	Dodecenoyl-CoA delta-isomerase (mitochondrial)	5.3.3.8
	Delta(3,5)-delta(2,4)-dienoyl-CoA isomerase	—
	Methylmalonyl-CoA decarboxylase	4.1.1.41
	3-Hydroxyisobutyryl-CoA hydrolase	3.1.2.4
	4-Chlorobenzoate dehalogenase	3.8.1.7
Vicinal Oxygen Chelate (VOC)	1,4-Dihydroxy-2-naphthoyl-CoA synthase	—
	Methylmalonyl-CoA epimerase	5.1.99.1
	4-Hydroxyphenylpyruvate dioxygenase	1.13.11.27
	FosA	2.5.1.18
Terpene Cyclase	Glyoxalase I	4.4.1.5
	5-Epi-aristolochene synthase	—
	Bornyl diphosphate synthase	5.5.1.8
	Pentalenene synthase	4.2.3.7
	Squalene-hopene synthase	5.4.99.17
	Trichodiene synthase	4.2.3.6
Haloacid Dehalogenase (HAD)	Aristolochene synthase	4.2.3.9
	Deoxy-D-mannose-octulosonate 8-phosphate phosphatase	3.1.3.45
	Phosphonoacetaldehyde hydrolase	3.11.1.1
	2-Haloacid dehalogenase	3.8.1.2
Amidohydrolase (AH)	Beta-phosphoglucomutase	5.4.2.6
	Cytosine deaminase	3.5.4.1
	Adenosine deaminase	3.5.4.4
	N-acyl-D-amino-acid deacylase	3.5.1.81
	L-hydantoinase	3.5.2.2
	D-hydantoinase	3.5.2.2
	Urease	3.5.1.5
	Isoaspartyl dipeptidase	—

# Big Data in der Biologie

## 5. Fehlende Daten: *data-poor domains* / *data-rich domains*

<1% der  $90 \cdot 10^6$  Sequenzen in der UniProtKB Datenbank haben "high-quality annotation"

data-rich: DNA-Sequenz

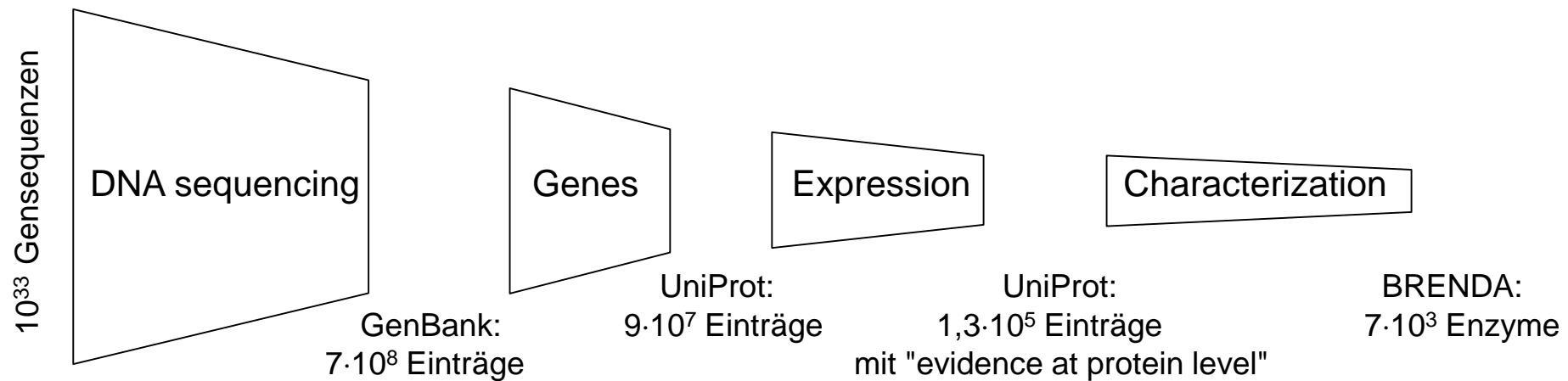
data-poor: Funktion und biochemische Eigenschaften

Grund: Kosten für Biochemie-Experimente >> Kosten einer DNA-Sequenzierung

# Big Data in der Biologie

## 5. Fehlende Daten: *data-poor domains* / *data-rich domains*

<1% der  $90 \cdot 10^6$  Sequenzen in der UniProtKB Datenbank haben "high-quality annotation"



## Zusammenfassung: Big Data in der Biologie

1. Große, schnell wachsende Datenmengen
2. Hohe Komplexität: Verknüpfung zwischen unterschiedlichen Daten und (verteilten) Datenquellen
3. Schwach strukturierte Daten
4. Fehlerhafte und inkonsistente Inhalte
5. Fehlende Daten: *data-poor domains* / *data-rich domains*

## 2. Datenbanken

Datenbanken dienen der strukturierten Speicherung großer Datenmengen

### **Daten** (*data*):

*"reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing; data can be processed by humans or by automatic means"* (ISO/IEC 2382:2015)

### **Datenbank** (*databank*):

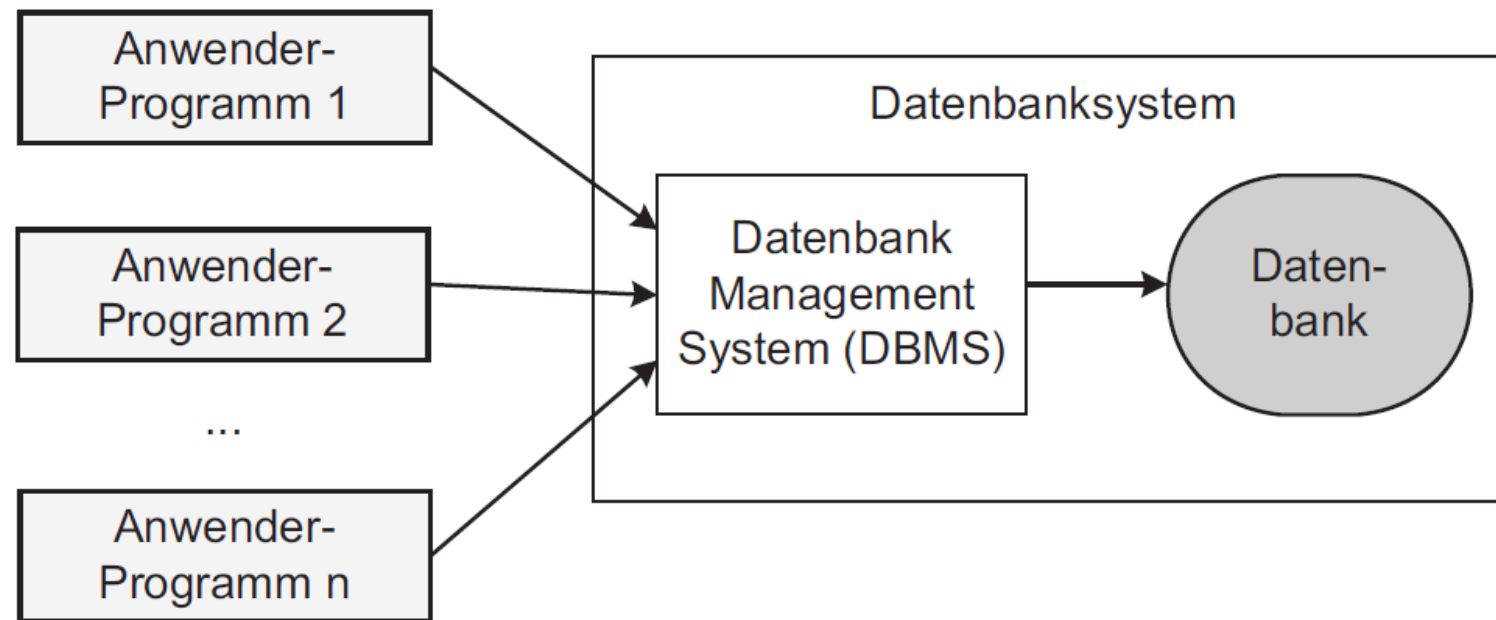
*"set of data related to a given subject and organized in such a way that it can be consulted by subscribers"* (ISO/IEC 2382:2015)

- logisch zusammengehöriger Datenbestand
- System zur Datenverwaltung

# Datenbanken

Eine Datenbank (Datenbanksystem) besteht aus

1. einem Datenbankmanagementsystem zur Datenbankverwaltung und Zugangskontrolle
2. den (strukturierten) Daten → meist: relationales Datenmodell
3. einer Datenbanksprache: Abfrage und Verwaltung → z.B. SQL ("Structured Query Language")



# Datenbankmanagementsystem

**Ziel:** Daten effizient, konsistent und dauerhaft speichern und bereitstellen

## **Datenbankverwaltungssystem** (*Database Management System, DBMS*)

Software, die den Aufbau, die Verwaltung und die Verwendung von Datenbanken in einem Rechensystem ermöglicht:

- Beschreibung der Daten (*data description*)
- Änderung der Daten (*data manipulation*)
- Schnittstellen zwischen Datenbank und Umgebung (Rechner, Netzwerk)
- Instandhaltung (*maintenance*)
- Gewährleistung der Konsistenz (*data integrity*)
- Verwalten von Transaktionen (*transaction management*)

# Datenbankmanagementsystem

## Beispiele für relationale Datenbankverwaltungssysteme

*(Relational Database Management Systems):*

Firebird (open source)

MySQL (open source)

Oracle Database (Oracle)

Microsoft Access (Microsoft)

...



# Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)

**Beispiel:** Erstellung einer Datenbank, die in verschiedenen Publikationen berichtete Viskosität  $\eta$  (in mPa·s) und Dichte  $\rho$  (in g·cm<sup>-3</sup>) für verschiedene wässrige Mischungen bei einem Wasseranteil  $\chi_w$  und einer Temperatur  $T$  (in K) beschreibt.

Codd, E. F.: *A relational model of data for large shared data banks*. Communications of the ACM, 13(6):377-387, 1970.

Codd, E. F.: *Relational database: a practical foundation for productivity*. Communications of the ACM, 25(2):109-117, 1982.

Ernst, H. *et al.*: *Grundkurs Informatik*, Springer Fachmedien Wiesbaden, 5. Auflage, 2015, S. 341ff.

## Relationales Datenmodell


- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)

**Beispiel:** Erstellung einer Datenbank, die in verschiedenen Publikationen berichtete Viskosität  $\eta$  (in mPa·s) und Dichte  $\rho$  (in g·cm<sup>-3</sup>) für **verschiedene wässrige Mischungen** bei einem Wasseranteil  $\chi_w$  und einer Temperatur **T** (in K) beschreibt.

# Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)
- Die Tupel beschreiben einzelne Einträge (Datensätze)

<b><u>Tabelle: Mischung</u></b>	
<b>ID</b>	<b>Name</b>
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	



# Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)
- Die Tupel beschreiben einzelne Einträge (Datensätze)
- Die **Attribute** beschreiben verschiedene Eigenschaften

<u>Tabelle: Mischung</u>	
ID	Name
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	

Attribute

# Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)
- Die Tupel beschreiben einzelne Einträge (Datensätze)
- Die Attribute beschreiben verschiedene Eigenschaften
- Jede Tabelle enthält als Attribut einen eindeutigen Primärschlüssel (**primary key**)

<u>Tabelle: Mischung</u>	
ID	Name
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
↑	...

Primärschlüssel (primary key )  
eindeutig (= kommt jeweils nur einmal vor)

## Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)
- Die Tupel beschreiben einzelne Einträge (Datensätze)
- Die Attribute beschreiben verschiedene Eigenschaften
- Jede Tabelle enthält als Attribut einen eindeutigen Primärschlüssel (primary key)
- Eine Tabelle beschreibt damit eine bestimmte Klasse von Objekten durch ihre Eigenschaften

<b><u>Tabelle: Mischung</u></b>	
<b>ID</b>	<b>Name</b>
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	...

<b><u>Tabelle: Viskosität</u></b>				
<b>ID</b>	<b><math>\eta</math></b> (in mPa·s)	<b><math>\chi_w</math></b>	<b>T</b> (in K)	<b>Literatur</b> (DOI)
1	0.566	0.0856	308.15	10.1021/je60011a015
2	...	...	...	...
...				

# Relationales Datenmodell

- Eine relationale Datenbank besteht aus mehreren miteinander verknüpften Tabellen (Edgar F. Codd, 1970)
- Eine Tabelle besteht aus Zeilen (Tupel) und Spalten (Attribute)
- Die Tupel beschreiben einzelne Einträge (Datensätze)
- Die Attribute beschreiben verschiedene Eigenschaften
- Jede Tabelle enthält als Attribut einen eindeutigen Primärschlüssel (primary key)
- Eine Tabelle beschreibt damit eine bestimmte Klasse von Objekten durch ihre Eigenschaften
- Beziehung (Relation) zwischen den Tabellen durch Fremdschlüssel (**foreign key**)

<u>Tabelle: Mischung</u>	
ID	Name
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	...

<u>Tabelle: Viskosität</u>					
ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur (DOI)	Mischung_ID
1	0.566	0.0856	308.15	10.1021/je60011a015	3
2	...	...	...	...	
...					

Foreign key:

kann mehrfach vorkommen

# Relationales Datenmodell

Relation:

- 1:n** jede Mischung ist mit einer oder mehreren Viskositätswerten verknüpft  
jeder Viskositätswert ist mit genau einer Mischung verknüpft

<u>Tabelle: Mischung</u>	
ID	Name
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	...

<u>Tabelle: Viskosität</u>					
ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur (DOI)	Mischung_ID
1	0.566	0.0856	308.15	10.1021/je60011a015	3
2	0.729...	0.43	308.15	10.1021/je60011a015	3
...					



# Relationales Datenmodell

Relation:

- 1:n** jede Mischung ist mit einer oder mehreren Dichtewerten verknüpft  
jeder Dichtewert ist mit genau einer Mischung verknüpft

<u>Tabelle: Mischung</u>	
ID	Name
1	Glycerin-Wasser
2	Ethylenglycol-Wasser
3	Methanol-Wasser
...	...

<u>Tabelle: Viskosität</u>					
ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur (DOI)	Mischung_ID
1	0.566	0.0856	308.15	10.1021/je60011a015	3
2	0.729...	0.43	308.15	10.1021/je60011a015	3
...					

<u>Tabelle: Dichte</u>					
ID	$\rho$ (g·cm <sup>-3</sup> )	$\chi_w$	T (in K)	Literatur (DOI)	Mischung_ID
1	0.9971	1	278.15	10.1007/BF00508889	2
2	...	...	...	...	3
...					

# Relationales Datenmodell

Designprinzip: Vermeidung von **Redundanzen** (mehrfach vorkommender Attributwerte) :

- Vermeidung von Eingabefehlern
- Verringerung der Speicherplatzbedarfs
- Beschleunigung der Abfragen

**Tabelle: Viskosität**

ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur (DOI)	Mischung_ID
1	0.566	0.0856	308.15	10.1021/je60011a015	3
2	0.729...	0.43	308.15	10.1021/je60011a015	3
...					

## Relationales Datenmodell

Daher: separate Tabelle für Literatur, da dasselbe Paper mehrfach vorkommt

**Tabelle: Literatur**

ID	DOI	Journal	Jahr
1	10.1021/je60011a015	J.Chem.Eng.Data	1961
2	10.1038/s41598-020-78101-y	Sci.Rep.	2020
3	10.1007/BF00508889	Int.J.Thermophys.	1985

**Tabelle: Viskosität**

ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur_ID	Mischung_ID
1	0.566	0.0856	308.15	1	3
2	0.729...	0.43	308.15	1	3
...					

# Relationales Datenmodell

Relation:

**1:n** in einer Publikationen können mehrere Viskositätswerte vorkommen  
jeder Viskositätswert wird in genau einer Publikationen berichtet

<u>Tabelle: Literatur</u>			
ID	DOI	Journal	Jahr
1	10.1021/je60011a015	J.Chem.Eng.Data	1961
2	10.1038/s41598-020-78101-y	Sci.Rep.	2020
3	10.1007/BF00508889	Int.J.Thermophys.	1985

<u>Tabelle: Viskosität</u>					
ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur_ID	Mischung_ID
1	0.566	0.0856	308.15	1	3
2	0.729...	0.43	308.15	1	3
...					

# Relationales Datenmodell

Relation:

**1:n** in einer Publikationen können mehrere Dichtewerte vorkommen  
jeder Dichtewert wird in genau einer Publikationen berichtet

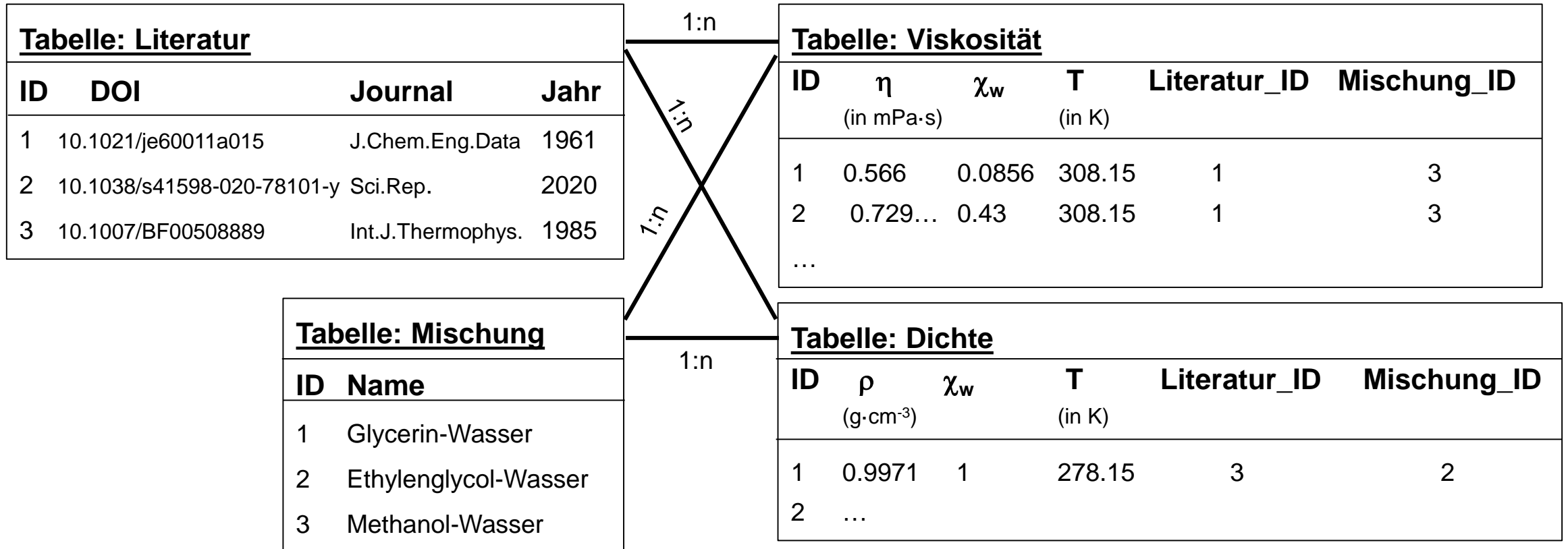
<u>Tabelle: Literatur</u>			
ID	DOI	Journal	Jahr
1	10.1021/je60011a015	J.Chem.Eng.Data	1961
2	10.1038/s41598-020-78101-y	Sci.Rep.	2020
3	10.1007/BF00508889	Int.J.Thermophys.	1985

<u>Tabelle: Viskosität</u>					
ID	$\eta$ (in mPa·s)	$\chi_w$	T (in K)	Literatur_ID	Mischung_ID
1	0.566	0.0856	308.15	1	3
2	0.729...	0.43	308.15	1	3
...					

<u>Tabelle: Dichte</u>					
ID	$\rho$ (g·cm <sup>-3</sup> )	$\chi_w$	T (in K)	Literatur_ID	Mischung_ID
1	0.9971	1	278.15	3	2
2	...				

# Relationales Datenmodell

Relationales Datenmodell aus vier Tabellen mit 1:n Beziehungen, das die in verschiedenen Publikationen berichtete Viskosität  $\eta$  und Dichte  $\rho$  für verschiedene wässrige Mischungen bei einem Wasseranteil  $\chi_w$  und einer Temperatur  $T$  beschreibt.



# Relationales Datenmodell

Weitere Tabelle mit den Autoren der Publikationen aus der Tabelle "Literatur"

Relation:

**n:m** in einer Publikationen können mehrere Autoren genannt werden  
jeder Wissenschaftler kann Autor mehrerer Publikationen sein

<u>Tabelle: Literatur</u>			
ID	DOI	Journal	Jahr
1	10.1021/je60011a015	J.Chem.Eng.Data	1961
2	10.1038/s41598-020-78101-y	Sci.Rep.	2020
3	10.1007/BF00508889	Int.J.Thermophys.	1985
4	10.1351/pac198557081083	Pure Appl. Chem.	1985

n:m

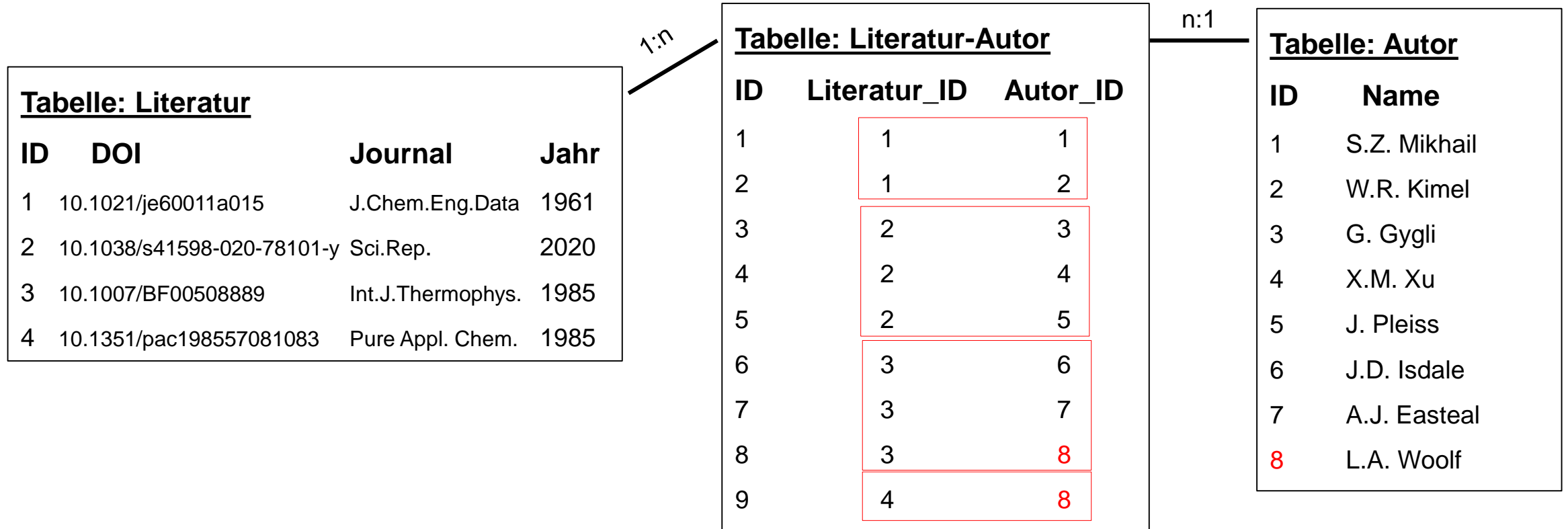
<u>Tabelle: Autor</u>	
ID	Name
1	S.Z. Mikhail
2	W.R. Kimel
3	G. Gygli
4	X.M. Xu
5	J. Pleiss
6	J.D. Isdale
7	A.J. Easteal
8	L.A. Woolf

# Relationales Datenmodell

Weitere Tabelle mit den Autoren der Publikationen aus der Tabelle "Literatur"

Verknüpfung mit Tabelle "Literatur" über eine n:m Beziehung

**n:m** in einer Publikationen können mehrere Autoren genannt werden  
jeder Wissenschaftler kann Autor mehrerer Publikationen sein

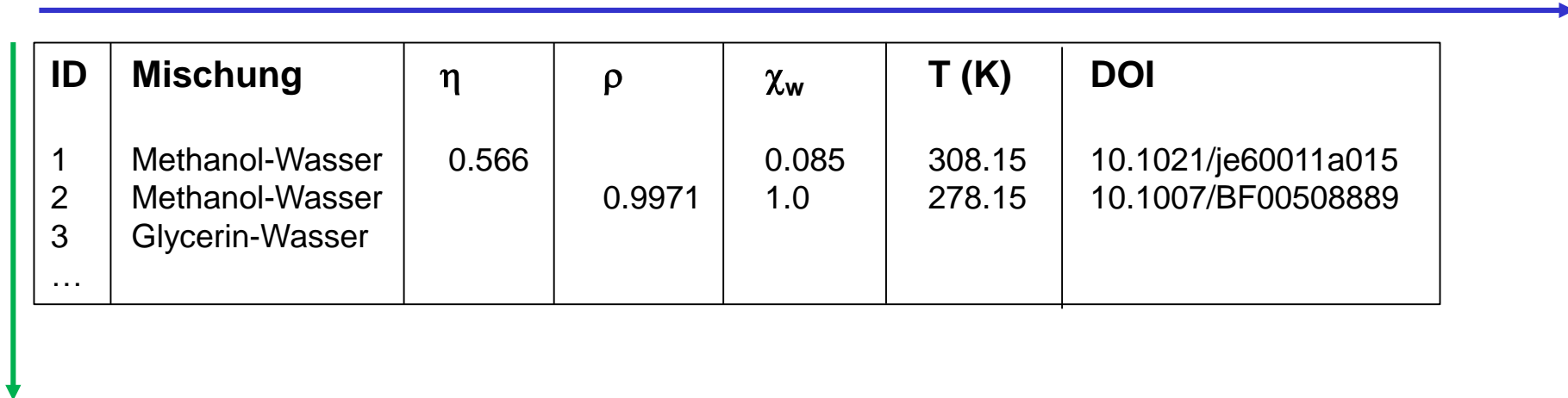




# Relationales Datenmodell

Warum mehrere verknüpfte Tabellen und nicht nur eine Tabelle?

- **schnelles Wachstum der Datenmenge:** 100 → 100000 → 100 Mio → ... Einträge
- **schnell wachsende Komplexität der Daten** durch weitere Attribute: 10 → 100 → 1000 → ... Attribute



ID	Mischung	$\eta$	$\rho$	$\chi_w$	T (K)	DOI
1	Methanol-Wasser	0.566	0.9971	0.085	308.15	10.1021/je60011a015
2	Methanol-Wasser			1.0	278.15	10.1007/BF00508889
3	Glycerin-Wasser					
...						

## Vorteile:

- **Vermeidung von Redundanz** (z.B. Mehrfachnennung von "Methanol-Wasser") durch separate Tabelle "Mischung"
- **Konsistenz:** Vermeidung von Widersprüchen, z.B. ID 3 Mischung = "Glycerin-Wasser", in ID 4 = "Wasser-Glycerin"
- **Effizienz:** schneller Zugriff auch bei großen Datenbanken und komplexen Abfragen; Speichereffizienz ("Methanol-Wasser" wird nur einmal in der Tabelle "Mischung" beschrieben)

# Datenbanksprache

**SQL** (*Structured Query Language*):

Datenbanksprache zum Abfragen und Manipulieren der Daten

Beispiel:

```
SELECT eta, chiW  
FROM Viskosität  
WHERE Mischung_ID = 3
```

# Praktische Übungen

## Auslesen einer Datenbank via SQL und R

- Datenbank-Schema der Vorlesung wird verwendet
- Konditionelles Querying mit dem SQLDF Paket von R

## Datenauswertung via R oder Python

- Anwendung von Hypothesentests
- Korrelations-Analyse

→ SQLDF Tutorial: <https://dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/sql.html>

→ Python Tutorial: <https://wiki.python.org/moin/BeginnersGuide/NonProgrammers>

## Zusammenfassung: Datenbanken

Eine relationale Datenbank besteht aus

1. einem Datenbankmanagementsystem (DBMS)

2. dem relationalen Datenmodell

Tabellen → Tupel, Attribute

Verknüpfung der Tabellen durch Schlüssel (*keys*)

3. einer Datenbanksprache → SQL