

Von Wechselgeld und Butterblumen

Eine kurze Reise durch die Bioinformatik

Sebastian Böcker

Lehrstuhl für Bioinformatik

Institut für Informatik

Friedrich-Schiller-Universität Jena



I think the biggest innovations of the twenty-first century will be the intersection of biology and technology. A new era is beginning, just like the digital age.

(Steve Jobs, biography by Walter Isaacson)



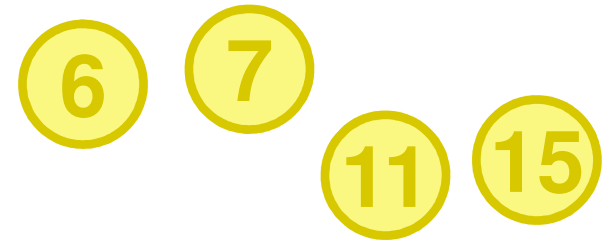
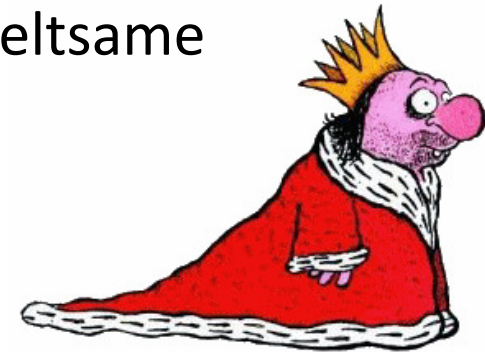
Es war einmal...

Vor gar nicht allzu langer Zeit lebte in einem fernen, fernen Land ein König, und der hatte eine gar seltsame Idee:

In seinem Land sollten nur noch Münzen geprägt werden, die die Werte 6, 7, 11 und 15 zeigten.

Das Volk fügte sich mit Freude, aber die Weisen des Landes fragten sich:

Welche Beträge können dann in den königlichen Geschäften noch als Wechselgeld heraus gegeben werden?





Wechselgeld-Problem

- Gegeben die Münzwerte:



- Welche Werte m sind über 6,7,11,15 zerlegbar?

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	...
✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓

Diagram illustrating the coin change problem. A sequence of numbers from 0 to 18 is shown, with green checkmarks indicating which values can be formed using the given coin denominations (6, 7, 11, 15). Red crosses indicate values that cannot be formed. Green arrows show the progression from 7 to 13 (+6) and from 13 to 19 (+6), illustrating the concept of adding a coin of value 6 to a valid sum.

- Falls wir m zerlegen können, dann können wir auch $m + 1 \cdot 6$, $m + 2 \cdot 6$, $m + 3 \cdot 6$, ... zerlegen
- n_r sei kleinste zerlegbare Zahl in Restklasse r (modulo 6)



Restklassen-Tabelle

$$a_1 = 6$$

r	n_r
0	0
1	7
2	14
3	15
4	22
5	11



- Gibt es Wechselgeld für $m = 20$?

$$20 \geq 14 = n_2 \quad \text{Ja!}$$

- Gibt es Wechselgeld für $m = 16$?

$$16 < 22 = n_4 \quad \text{Nein!}$$

- n_r sei kleinste zerlegbare Zahl in Restklasse r (modulo 6)



Round Robin Algorithmus

r	6			
0	0			
1	∞			
2	∞			
3	∞			
4	∞			
5	∞			

Initialisierung

Bei einer Münze $a_1=6$ sind $n=0, 6, 12, \dots$ zerlegbar



Round Robin Algorithmus

r	⑥	⑦		
0	0	0		
1	∞	7		
2	∞	14		
3	∞	21		
4	∞	28		
5	∞	35		

Iteration

Gegeben die Restklassen-Tabelle für zwei Münzen ⑥, ⑦.



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7		
2	∞	14		
3	∞	21		
4	∞	28		
5	∞	35		



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7		
2	∞	14		
3	∞	21		
4	∞	28		
5	∞	35	11	

$$0 + 11 = 11, 11 \bmod 6 = 5, \min \{n_5, 11\} = 11$$



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7		
2	∞	14		
3	∞	21		
4	∞	28	22	
5	∞	35	11	

$$11 + 11 = 22, 22 \bmod 6 = 4, \min \{n_4, 22\} = 22$$



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7		
2	∞	14		
3	∞	21	21	
4	∞	28	22	
5	∞	35	11	

$$22 + 11 = 33, 33 \bmod 6 = 3, \min \{n_3, 33\} = 21$$



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7		
2	∞	14	14	
3	∞	21	21	
4	∞	28	22	
5	∞	35	11	

$$21 + 11 = 32, 32 \bmod 6 = 2, \min \{n_2, 32\} = 14$$



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7	7	
2	∞	14	14	
3	∞	21	21	
4	∞	28	22	
5	∞	35	11	

$$14 + 11 = 25, 25 \bmod 6 = 1, \min \{n_1, 25\} = 7$$



Round Robin Algorithmus

r	6	7	11	
0	0	0	0	
1	∞	7	7	
2	∞	14	14	
3	∞	21	21	
4	∞	28	22	
5	∞	35	11	

$$7 + 11 = 18, 18 \bmod 6 = 0$$



Round Robin Algorithmus

r	6	7	11	15
0	0	0	0	0
1	∞	7	7	7
2	∞	14	14	14
3	∞	21	21	15
4	∞	28	22	22
5	∞	35	11	11



Round Robin Algorithmus

r	6	7	11	15
0	0	0	0	0
1	∞	7	7	7
2	∞	14	14	14
3	∞	21	21	15
4	∞	28	22	22
5	∞	35	11	11

Letzte Spalte ist die Restklassen-Tabelle!



Alle Arten, Wechselgeld zu geben

r	6	7	11	15
0	0	0	0	0
1	∞	7	7	7
2	∞	14	14	14
3	∞	21	21	15
4	∞	28	22	22
5	∞	35	11	11

$$28 = \begin{cases} 0 \cdot 15 + \begin{cases} 0 \cdot 11 + 4 \cdot 7 + 0 \cdot 6 \\ 2 \cdot 11 + 0 \cdot 7 + 1 \cdot 6 \end{cases} \\ 1 \cdot 15 + 0 \cdot 11 + 1 \cdot 7 + 1 \cdot 6 \end{cases}$$

Zwischenfrage

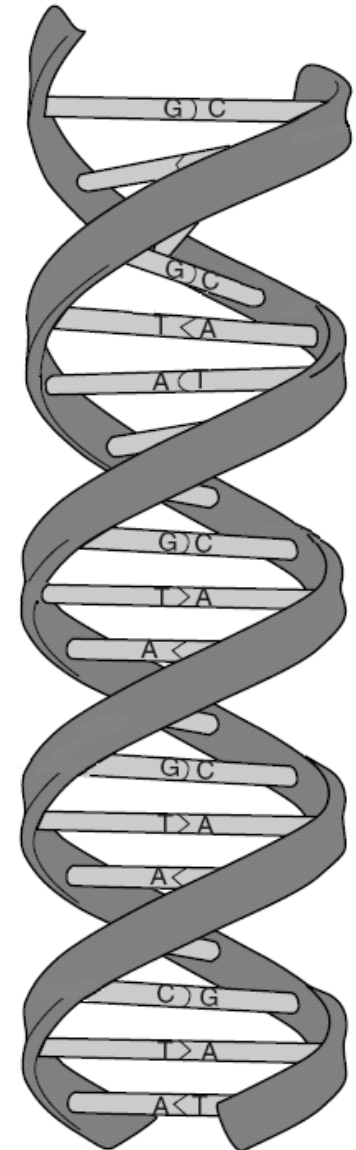


...und was hat das alles mit
Bioinformatik zu tun?

Genomik (DNA)



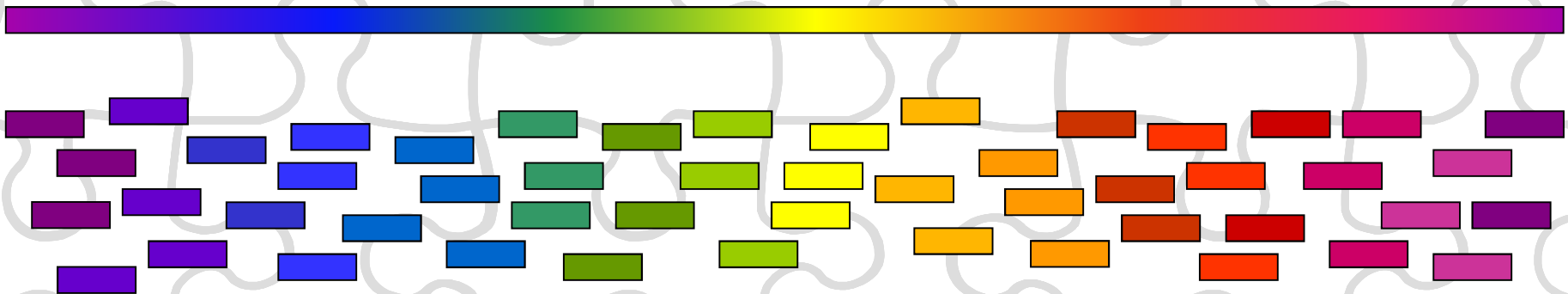
- die **Information** jeder lebenden Zelle ist gespeichert in ihrem Genom (Erbgut)
- codiert ist diese Information als DNA (Desoxyribonukleinsäure)
- Text mit **4 Buchstaben**: A,C,G,T
- jede Zelle des Körpers hat **dieselbe Information**
- Gene kodieren Proteine (Aminosäure-Ketten)
- Genom liefert **statistisches Bild** der Zelle





Genom-Assemblierung

- Genome werden heutzutage in der Regel mit der **Schrotschuss-Methode** sequenziert:



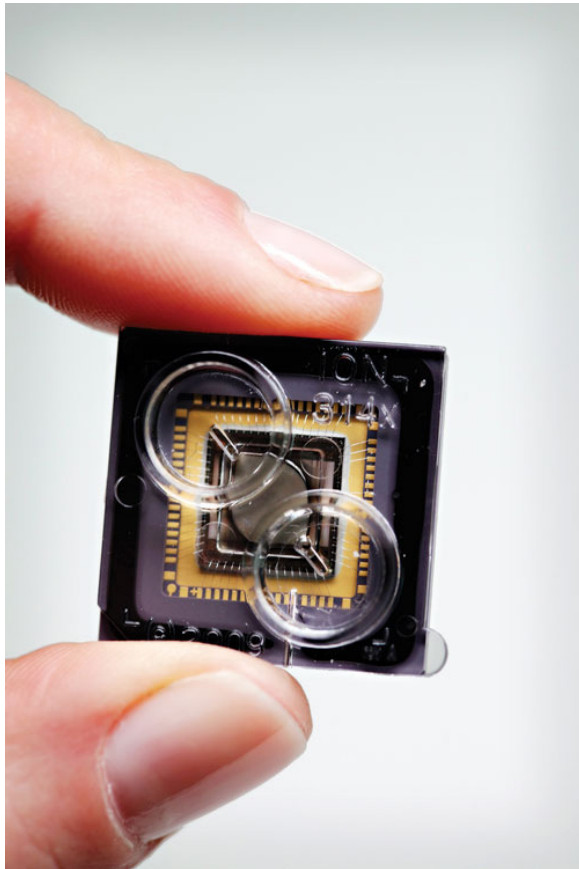
- dabei werden die Sequenzen von **Millionen von Bruchstücken** bestimmt, je 600–800 Buchstaben lang
- Wie puzzelt man die **überlappenden Bruchstücke** wieder zu einem kontinuierlichen Genom zusammen?
- unmöglich für einen Menschen, **extrem schwer** für einen Computer...

Sequenzsuche



- die Bibel hat 2.3 Millionen Buchstaben
- menschliches Genom hat 3 Milliarden Buchstaben
- Am 22. August 2005 waren alles in allem 100 Milliarden Buchstaben sequenziert: Fruchtfliege, Hefepilz, Maus, Ratte, Hund, ...
- Was bedeutet dieser Text? Welche Teile kodieren Proteine, welche Teile steuern die Expression?
Problem der Genom-Annotierung
- Mustersuche in Texten: Ähnliche Sequenz lässt auf ähnliche Funktion schließen (**BLAST**)

Next Generation Sequencing



- bis zu einem **Terrabyte** (1000 Gigabyte, 1000000 Megabyte) an DNA oder RNA können in einer Woche gelesen werden

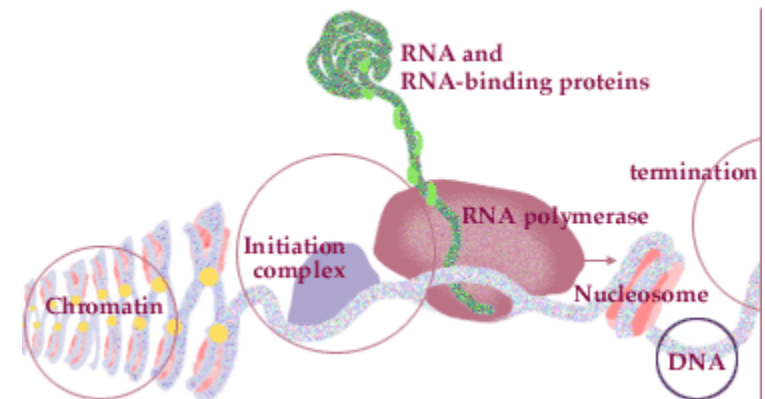


Transkriptomik (Boten-RNA)

- bei höheren Lebewesen ist die DNA im Zellkern eingeschlossen
- zuerst wird eine **Kopie** der DNA-Sequenz angelegt: Boten-RNA, Ribonukleinsäure (**transkribiert**)
- die Boten-RNA wird außerhalb des Zellkerns in das Protein „übersetzt“
- liefert **dynamisches Bild** der Zelle



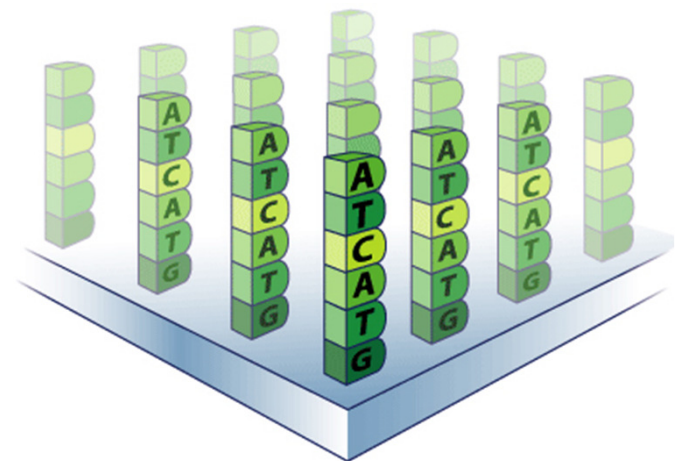
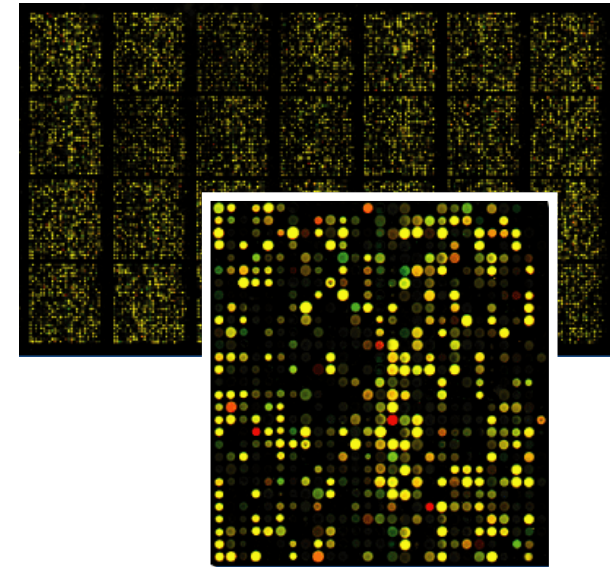
Nobelpreis Chemie 2006
für Roger D. Kornberg





Gen-Chips

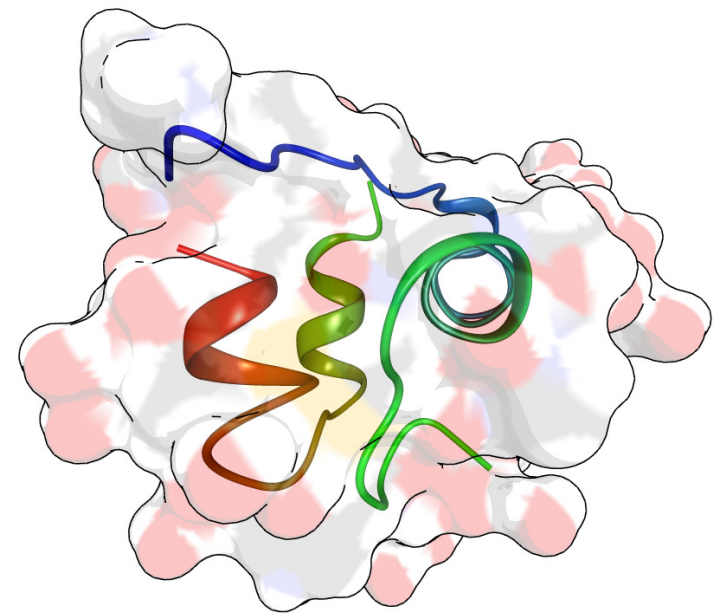
- für **alle Gene des Menschen** wird parallel überprüft, ob sie aktiv sind
- über die meisten untersuchten Gene ist wenig oder **nichts bekannt**
- Experiment wird unter verschiedenen Bedingungen (Hunger) oder als Zeitreihe wiederholt
- Gene, die sich **ähnlich verhalten**, sind möglicherweise an den **selben biochemischen Prozessen** beteiligt
- Bildverarbeitung, **Clustering**





Proteomik (Proteine)

- Proteine sind „molekularen Maschinen“, die Stoffe transportieren, Ionen pumpen, Signalstoffe erkennen...
- Protein als Kette aus 20 verschiedenen Aminosäuren aufgebaut, Länge der Kette bis zu 30000 dieser A.
- Proteine falten sich zu komplexen 3-dimensionalen Gebilden
- Menschen haben ungefähr 500000 verschiedene Proteine
- Welche Proteine sind in einer Zelle zu finden? Unterscheiden sich gesunde von Krebszellen?

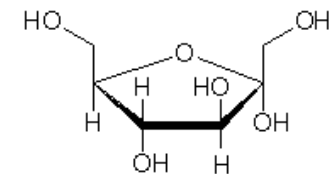


Insulin

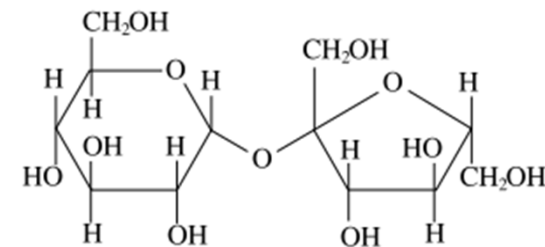


Metabolomik (Zucker, Fette, ...)

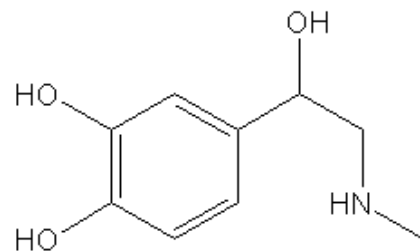
- die „kleinen“ Moleküle in der Zelle heißen Metaboliten
- eigentlich alles außer DNA, RNA und Proteinen:
 - Einfachzucker, Mehrfachzucker, Lipide, ...
 - Adrenalin, Coffein, Nikotin, ...
 - Antibiotika



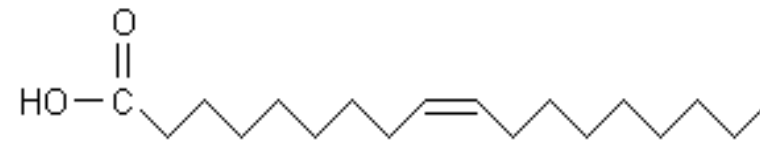
Fructofuranose



Saccharose



Adrenalin

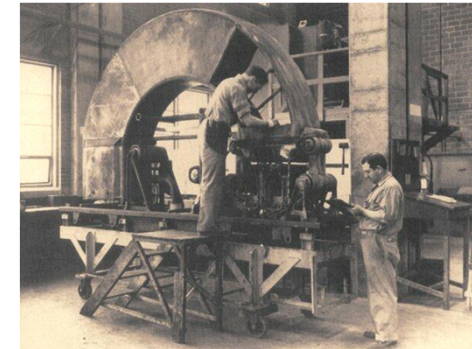


Ölsäure



Massenspektrometrie

- Erfindung der Massenspektrometrie **Anfang des 20. Jahrhunderts** durch Sir J.J. Thomson
- Anwendungen: Studium u. Auftrennung von Isotopen, Erdölindustrie, ...
- 1966 Anfänge der **Peptid-Analyse**
- 1985/1989 Entwicklung „weicher“ Ionisationsmethoden (MALDI/ESI)
- Heute **die Standardtechnik** zur Analyse von Proteinen und Metaboliten

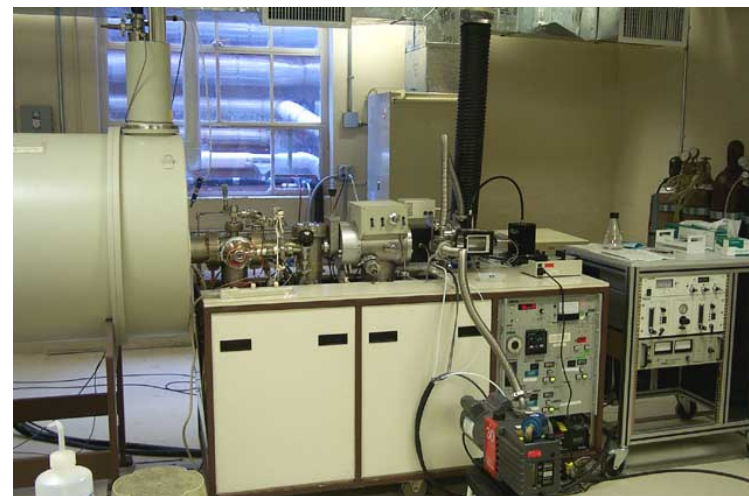




Hochauflösende Massenspektrometrie

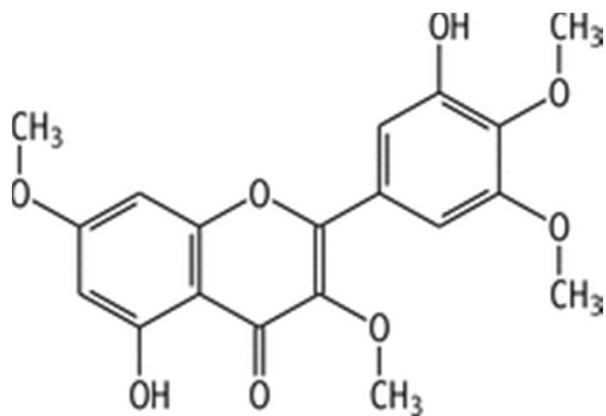
- Fourier Transform MS kann die Masse eines Moleküls mit sehr hoher Genauigkeit bestimmen:

Messungengenauigkeit ist ein Tausendstel der Masse eines einzelnen Neutrons!



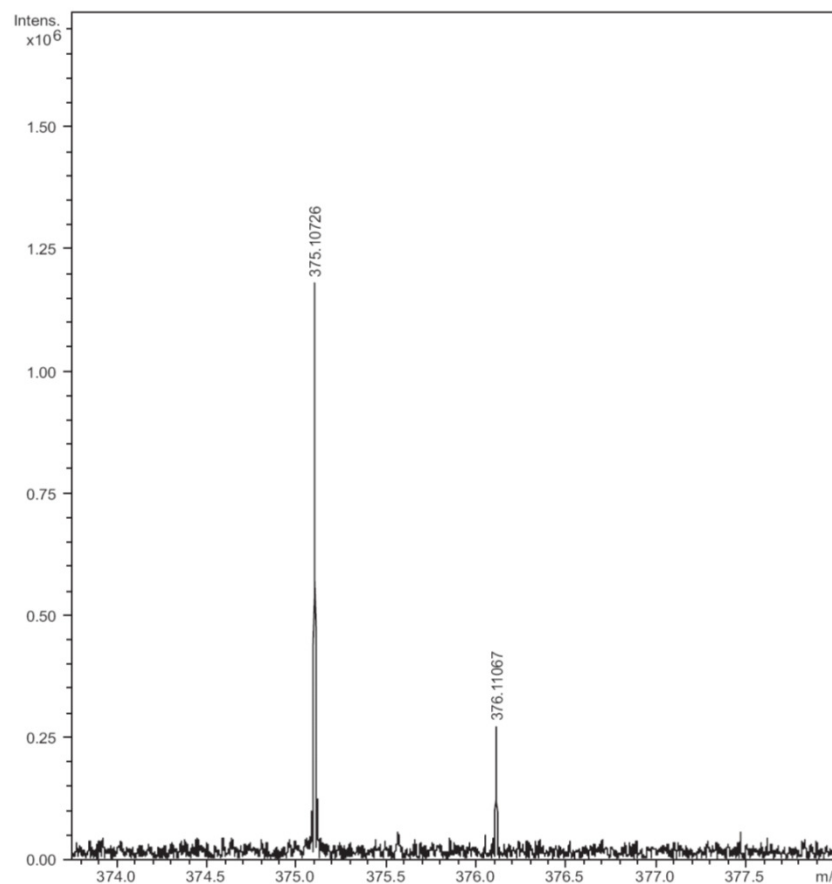
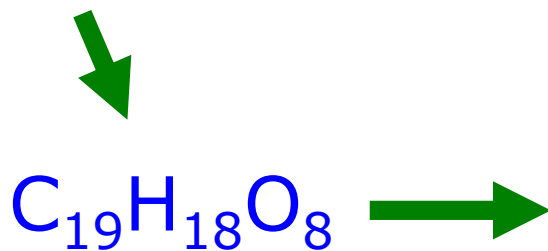


Spektrum eines Metaboliten

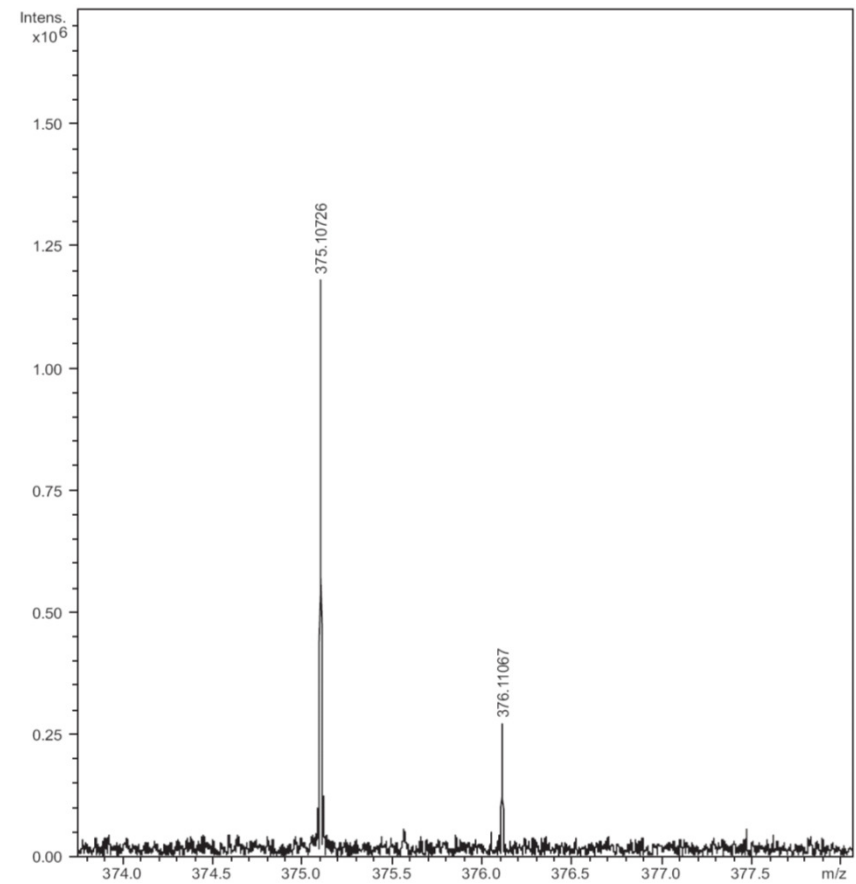


Flavon

[5,3'-Dihydroxy-3,7,4',5'-
tetramethoxyflavon]



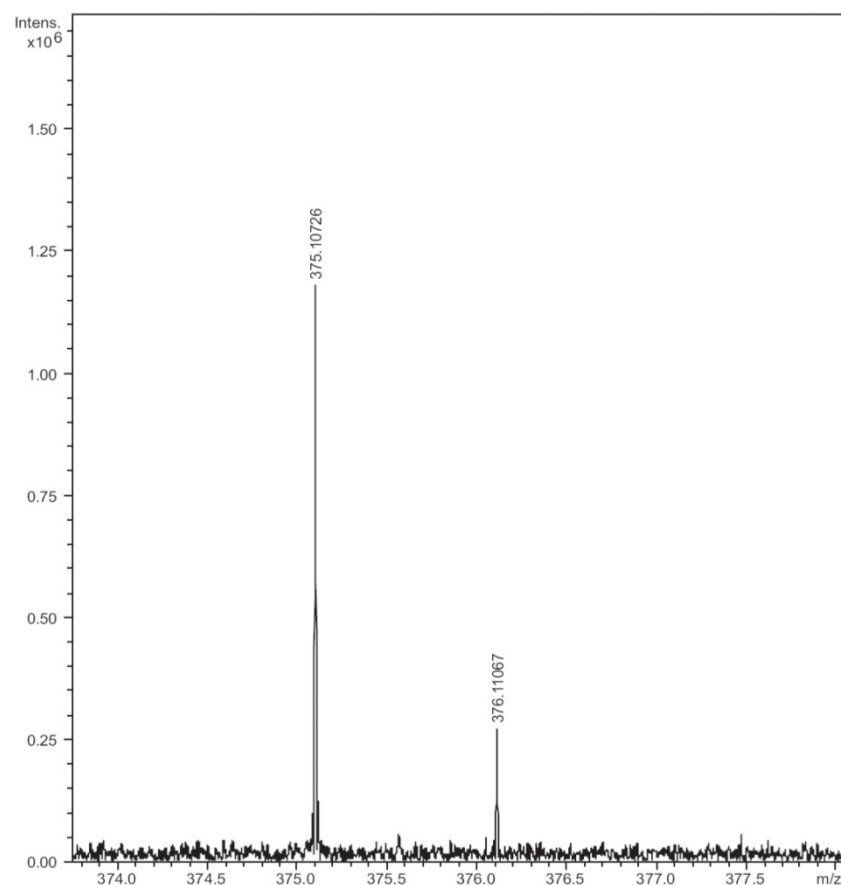
Spektrum eines Metaboliten



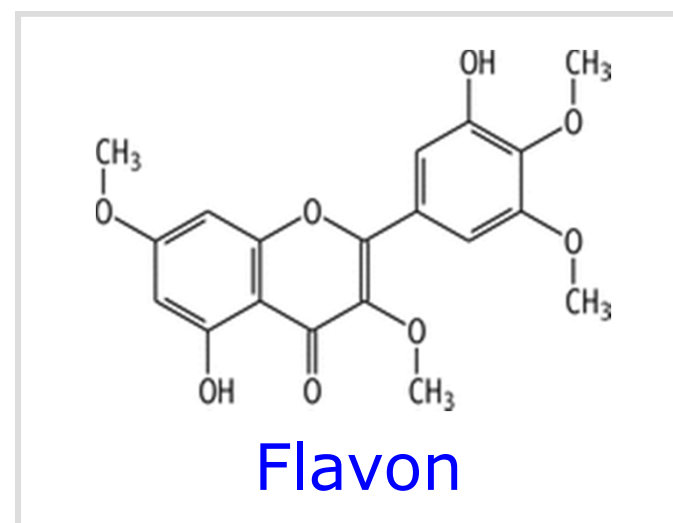
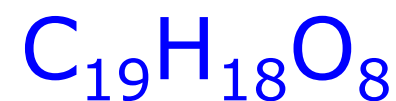


Unsere Aufgabe

Eingabe: Hoch-
aufgelöstes Spektrum



Ausgabe:
Summenformel





Zerlegung der Molekül-Masse

Element	Symbol	Protonen	Neutronen	Masse (Da)
Wasserstoff	^1H	1	0	1.007825
Kohlenstoff	^{12}C	6	6	12.0
Sauerstoff	^{16}O	8	8	15.994915

- **Nominal-Masse** ist Anz. Protonen plus Anz. Neutronen
- die Masse eines Atoms ist **ungleich** der **Summe der Massen** seiner Protonen, Neutronen und Elektronen!

6 Protonen plus 6 Neutronen plus 6 Elektronen
haben in der Summe eine Masse von 12.09596 Da



Zerlegung der Molekül-Masse

Element	Masse (Da)
Wasserstoff	1.007825
Kohlenstoff	12.0
Sauerstoff	15.994915

1008

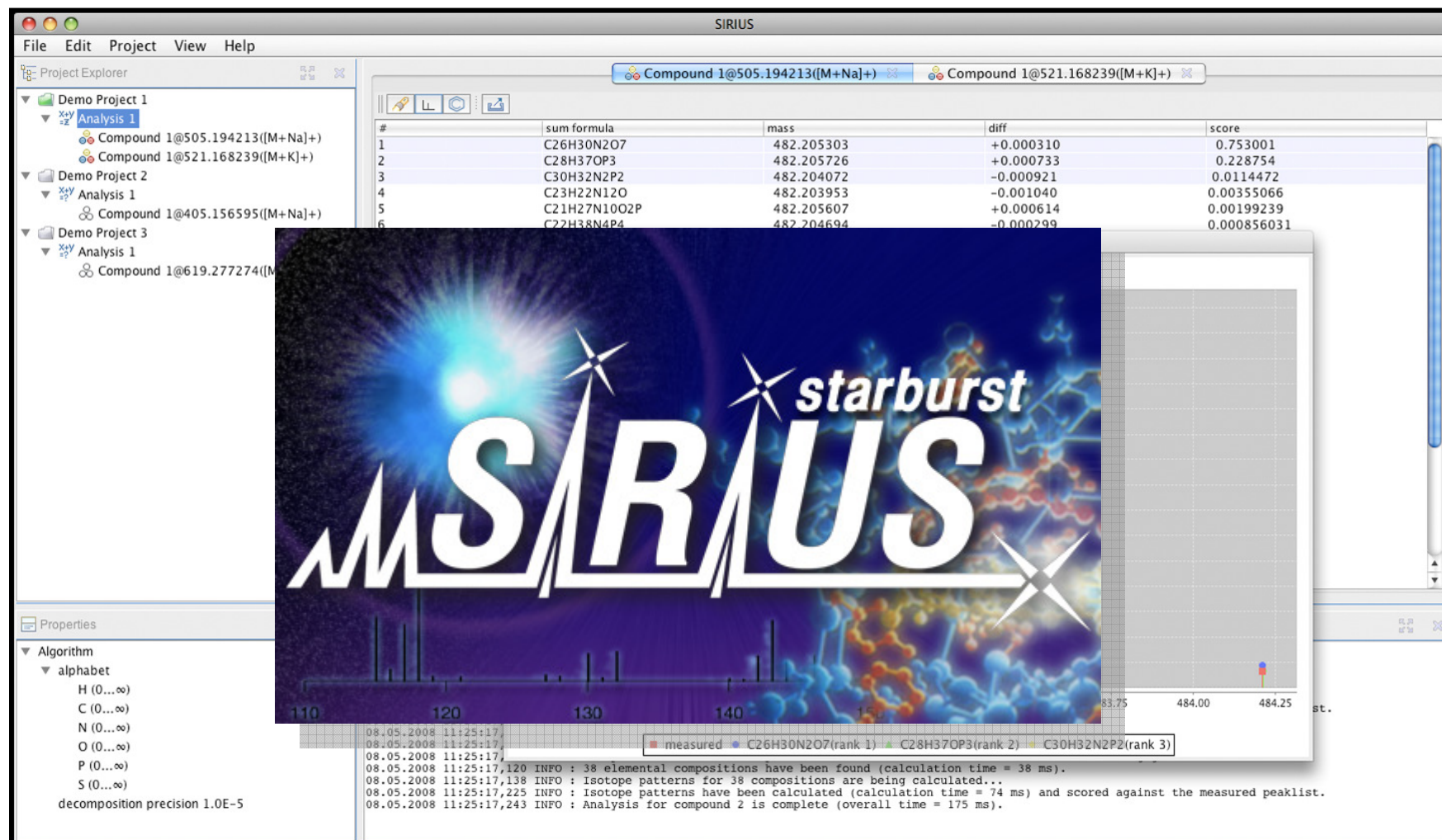
12000

15995

- Unsere Aufgabe:
Finde alle Moleküle mit Masse ungefähr 375.1072 Da
- Masse der Elemente mal 1000, runden gibt Münzwerte
- Generiere Wechselgeld für die Summe 375107



SIRIUS Software

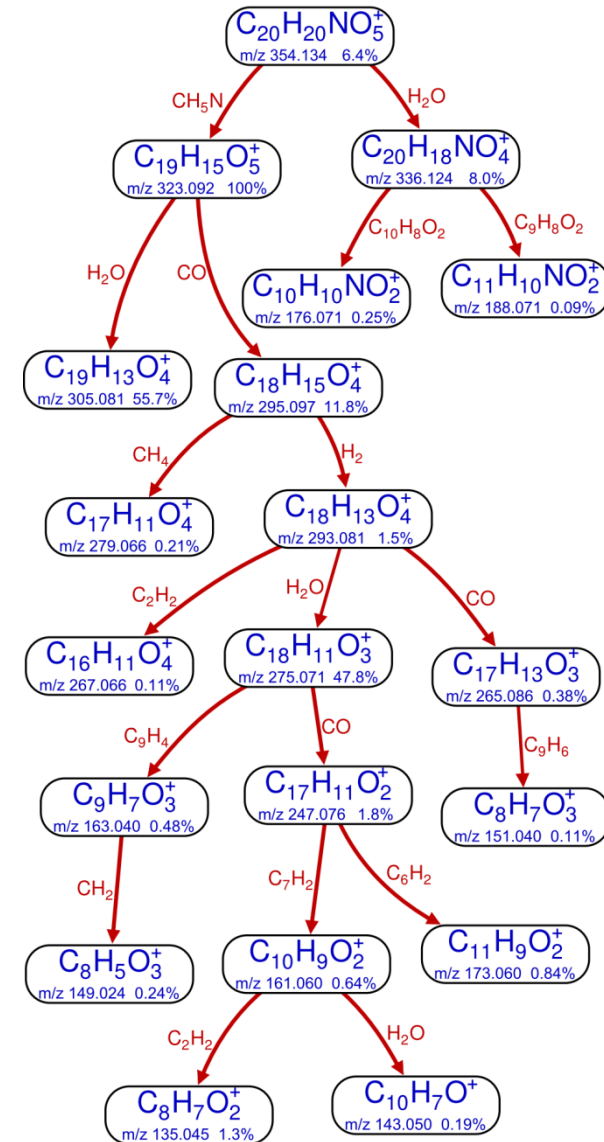
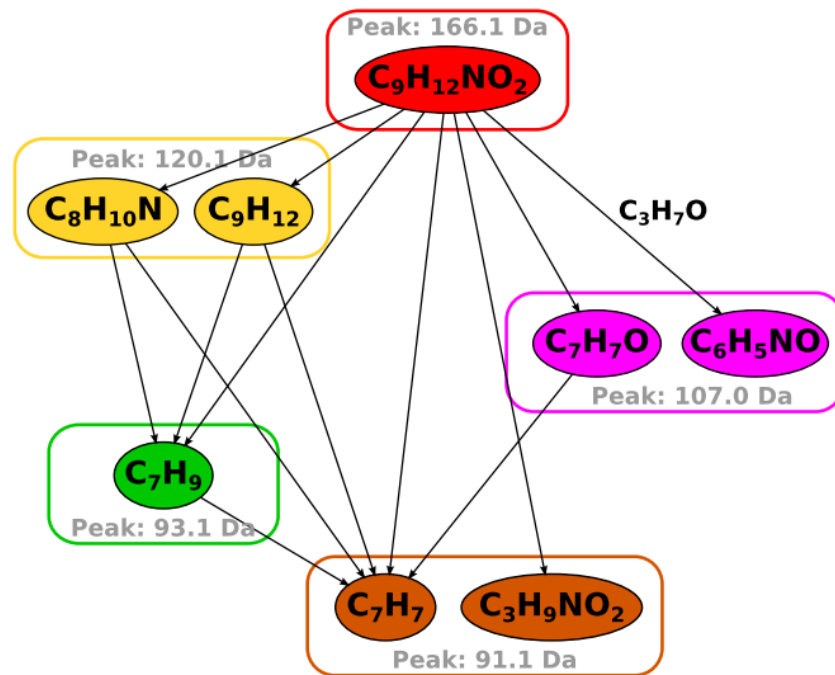
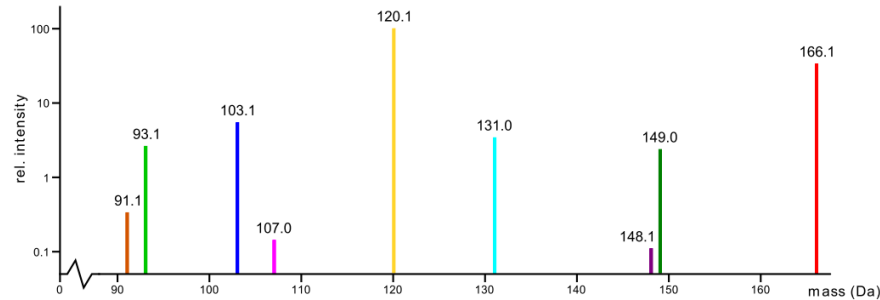


<http://bio.informatik.uni-jena.de/sirius>

Tandem Massenspektrometrie

- die Summenformel reicht uns nicht...
- Fragmentiere (zerhacke) den Metaboliten
- Identifiziere ihn aus den Massen der Fragmente







**ein Drittel der Menschheit ist latent
mit *M. tuberculosis* infiziert**

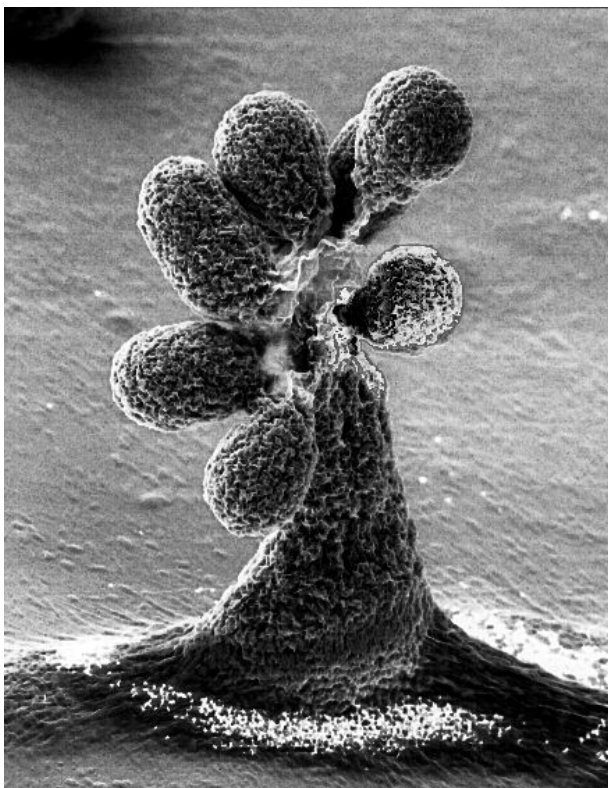
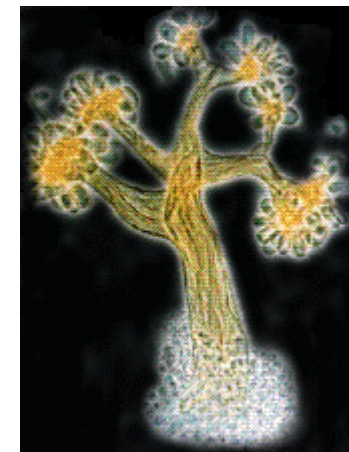
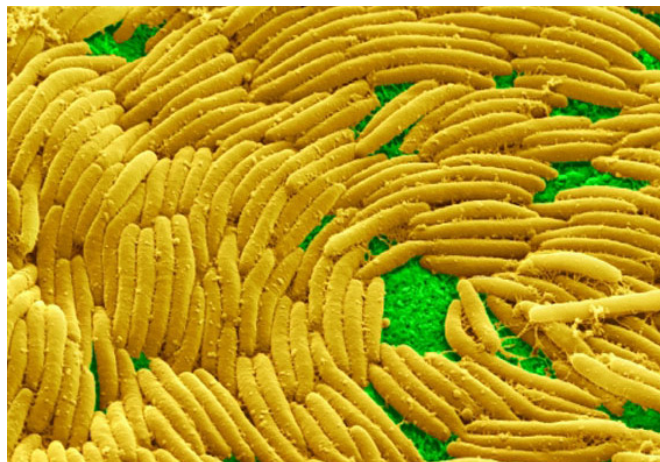
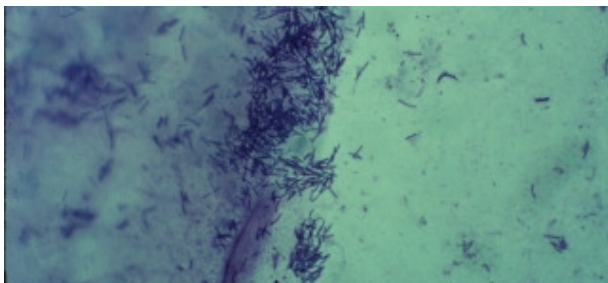
World Health Organisation (WHO)

**mehr als 50% der meistverkauften
Medikamente, und 80% der Antibiotika
sind von Naturstoffen abgeleitet**

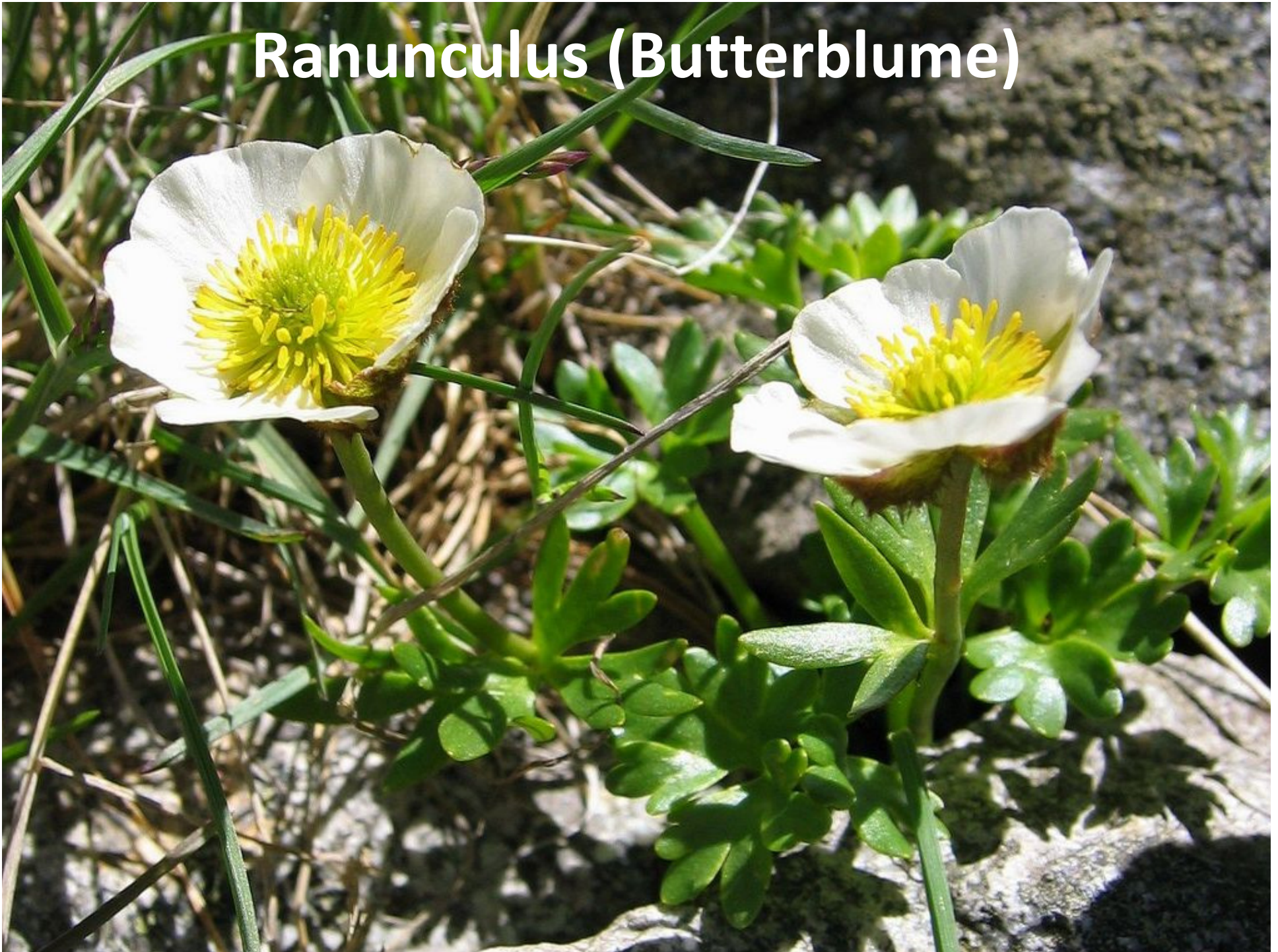
Newman *et al.*, 2007



Myxobakterien



Ranunculus (Butterblume)





Butterblumen in Neuseeland

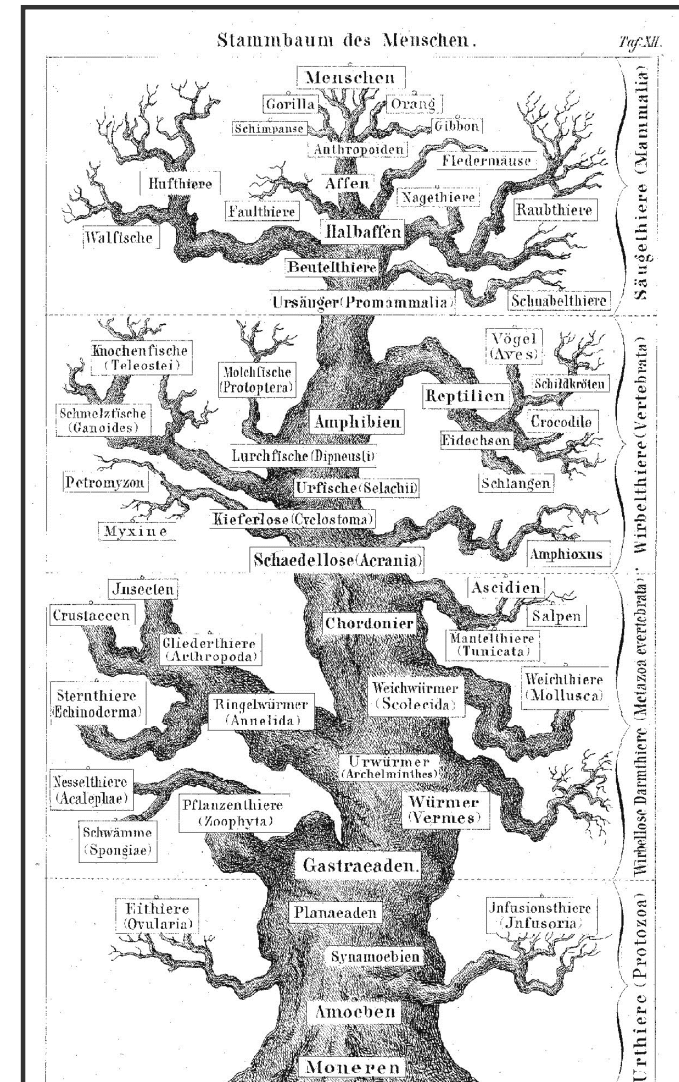
- gemeinsamer Vorfahre aller alpine Butterblumen in Neuseeland „landete“ vor **5 Millionen Jahren**
- diverse Morphologie, unterschiedlichste Lebensräume
- „schnelle“ geologische und klimatische Änderungen in Neuseeland bedingten **„schnelle“ Evolution**
- phylogenetisch interessant, aber **schwer zu analysieren**



Super-Bäume



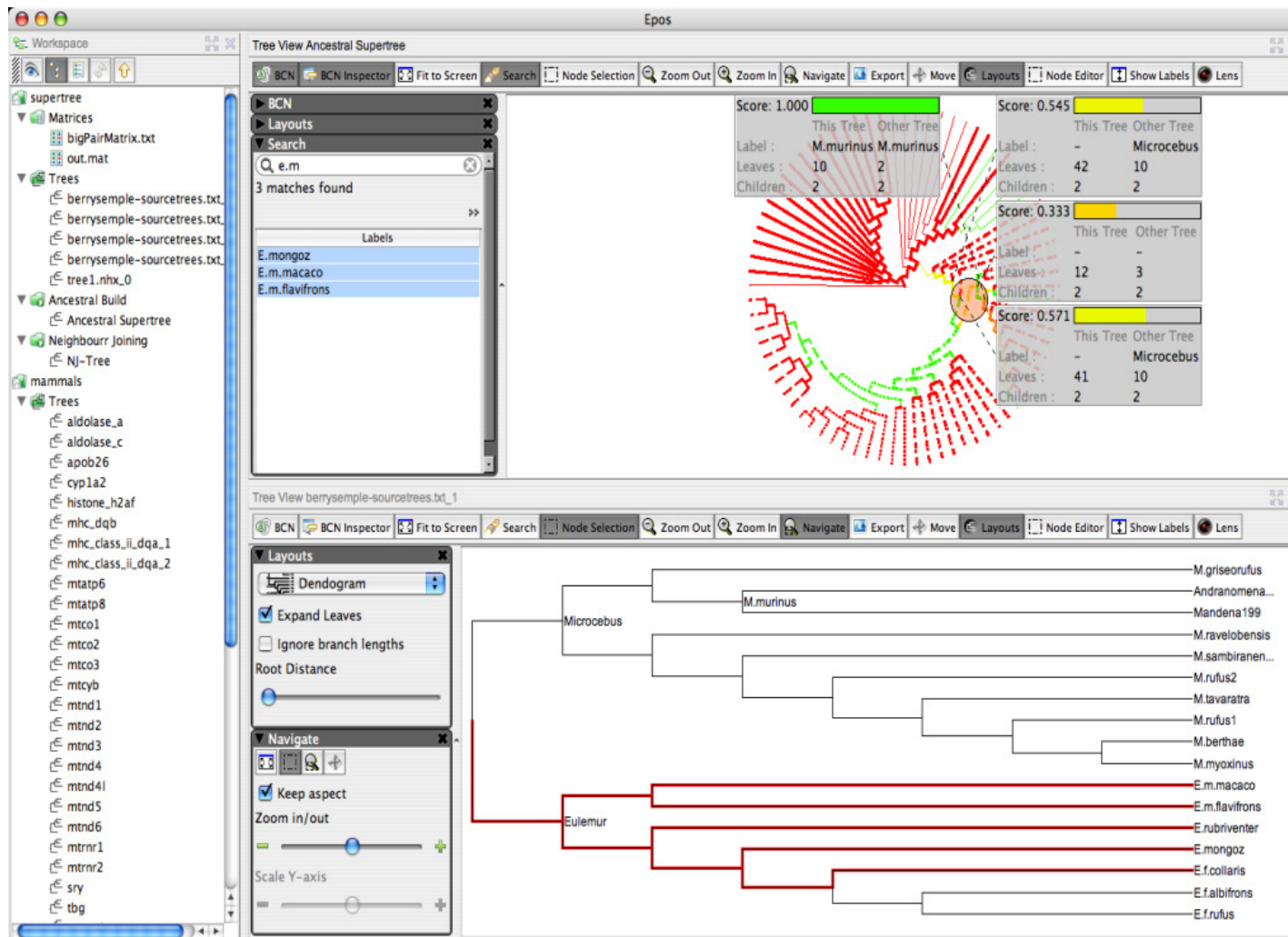
- Bioinformatische Methoden der Stammbaum-Rekonstruktion werden **langsam** für Stammbaum mit mehr als 100 Arten
- Ziel: **Baum des Lebens**
- Idee: **Supertree-Methoden**
- Unterbäume zu Superbäumen „**zusammenkleben**“
- **Effiziente Berechnung**
- Auflösen von **Widersprüchen**



(Ernst Haeckel, 1874)



EPoS: Estimating Phylogeny of Species



<http://bio.informatik.uni-jena.de/epos>

Feldforschung in der Phylogenetik



Lehrstuhl für Bioinformatik, FSU Jena

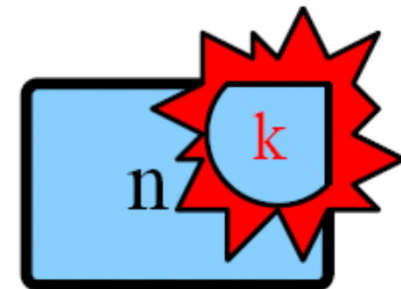
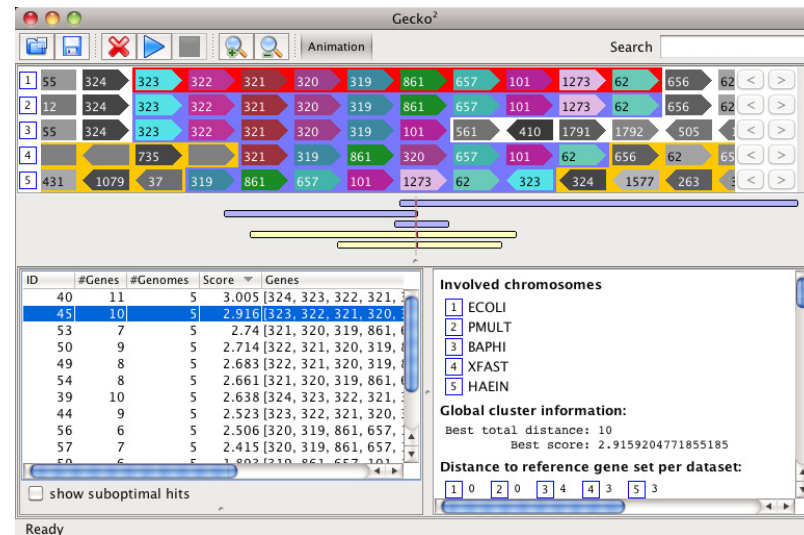


<http://www.bio.informatik.uni-jena.de/>



Forschung an meinem Lehrstuhl

- algorithmische Massenspektrometrie
- algorithmische Phylogenetik
- approximative Genclover
- schwere bioinformatische Probleme





Bioinformatik in Jena

- Prof. [Stefan Schuster](#), Lehrstuhl für Bioinformatik, Fakultät für Biologie und Pharmazie
- mein Lehrstuhl, Fakultät für Mathematik und Informatik
- Juniorprof. [Christoph Kaleta](#) (Fakultät Biologie & Pharmazie)
- Juniorprof. [Manja Marz](#) (Fakultät Mathematik & Informatik)
- Dr. [Reinhard Guthke](#), Leibniz-Institut für Naturstoff-Forschung
- Dr. [Jürgen Sühnel](#), Leibniz-Institut für Altersforschung
- Prof. Dr. [Marc Thilo Figge](#), Leibniz-Institut für Altersforschung
- [Jena Centre for Bioinformatics](#) (JCB)



Danke für eure Aufmerksamkeit!

<http://www.bio.informatik.uni-jena.de/>
<http://www.uni-jena.de/bioinformatik/>