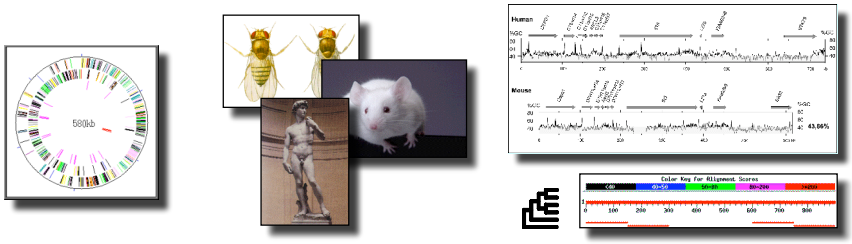
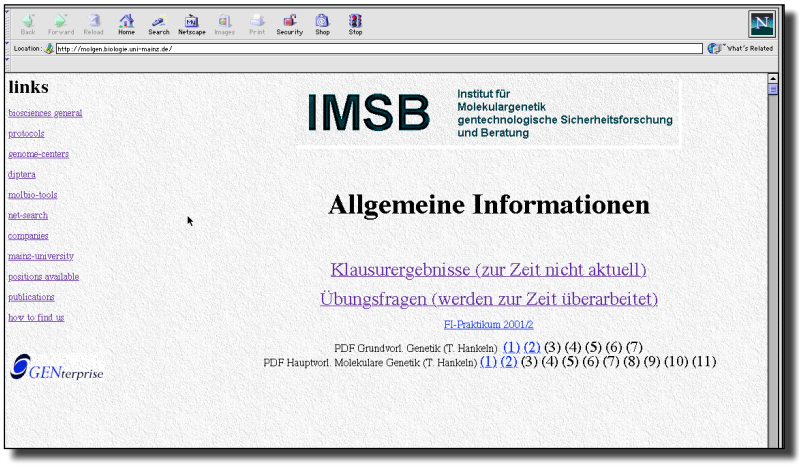
SS 2010

Bioinformatik-Was ist das?

Thomas Hankeln
Institut für Molekulargenetik



pdfs http://molgen.biologie.uni-mainz.de



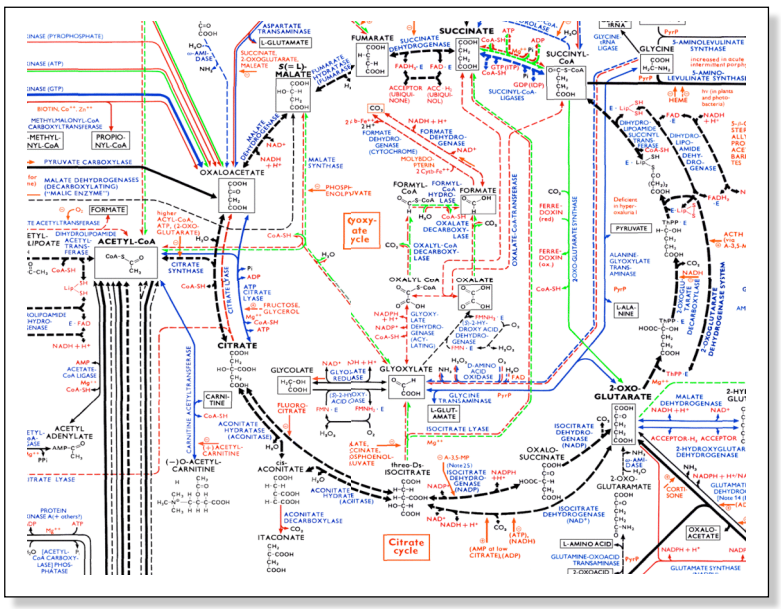
Bioinformatik /computational biology

„Anwendung mathematischer, statistischer und Computer-Methoden zur Analyse biologischer, biophysischer und biochemischer Daten“ (Georgia Inst. Technol.)

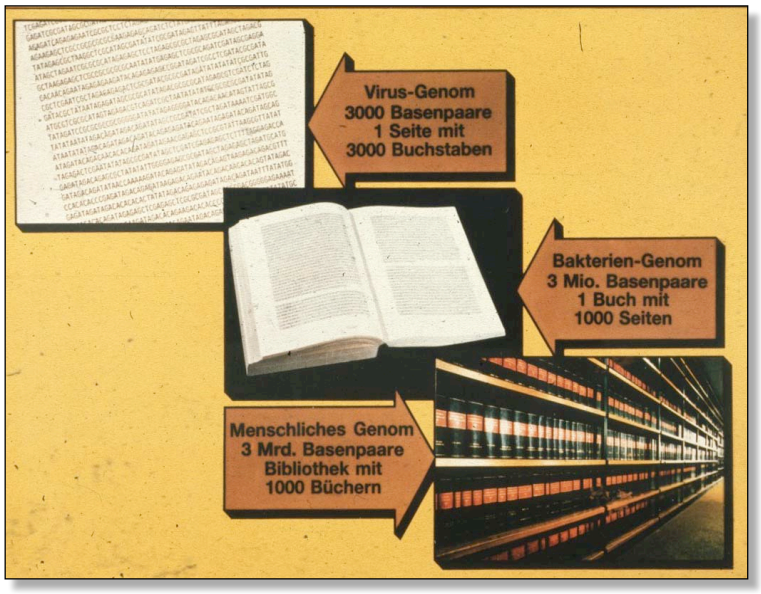
„Entwicklung von Datenbanken und Algorithmen für die biologische Forschung“ (whatis.com)

„Kombination von Computerwissenschaften, Informations-Technologie und Genetik zur Analyse der genetischen Information“ (BitsJournal.com)

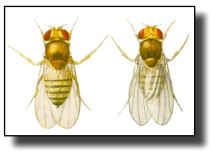
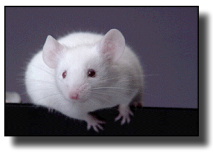
Warum Informatik in der Biologie?



Warum Informatik in der Biologie?



Warum Informatik in der Biologie?

• Bäcker-Hefe	12 069 kb	6 200 Gene
• Fadenwurm	97 000 kb	20 000 Gene
• Drosophila melanogaster	137 000 kb	14 000 Gene
• Homo sapiens	3 000 000 kb	<25 000 Gene
• Reis	400 000 kb	>50 000 Gene !
• Ackerschmalwand	125 000 kb	>25 500 Gene

Genom-Projekte bei Modellorganismen der biologischen Forschung lassen die Datenmengen rasch anwachsen

BioInformatik

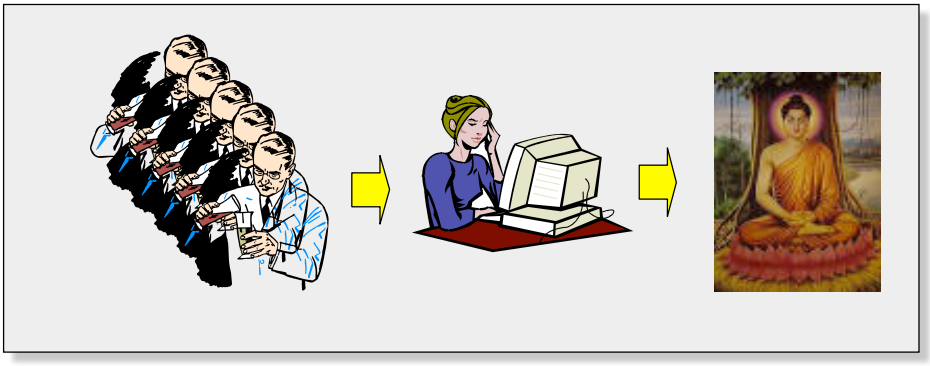
Genetik
Biochemie
Physiologie

Algorithmen*
Datenbanken
Visualisierung
Simulation

> Verständnis biologischer Zusammenhänge
> Kenntnis informatischer Methoden

*eine Menge eindeutiger Anweisungen zur Lösung eines Problems

Die Vision...



www.systemsbiology.org

Muss ich programmieren können?


Architekt & Maurer

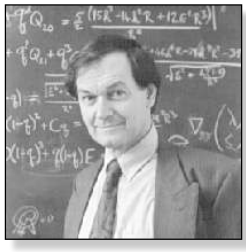
Nützlich sind: > Web sites basteln
> PERL als Programmiersprache
> UNIX/Linux als Betriebssysteme
> SQL als Datenbanksprache

Penrose's Gesetz

"Jede Formel in einem Buch halbiert die Anzahl der Leser".

$$L' = L \times 0{,}5^{F}$$

- L': Anzahl der tatsächlichen Leser
- L: Anzahl der potentiellen Leser
- F: Anzahl der Formeln



Bioinformatik

- Wie sie nicht sein sollte -

Literaturauswahl

Mount, D.M. *Bioinformatics*. Cold Spring Harbor Press 2004
(für den -zukünftigen- Profi, z. T. kompliziert)

Hansen, A. *Bioinformatik. Ein Leitfaden für Naturwissenschaftler.* Birkhäuser 2004

Graur, D, Li W-H *Fundamentals of Molecular Evolution*. Sinauer 2000 (Super, aber nur Phylogenie/Evolution)

Das Szenario ...ein neues tödliches Virus!

Severe Acute Respiratory Syndrome

- Symptome: ähnlich Lungenentzündung
- 114 Tage-Epidemie (2002/2003)
- 8098 Erkrankungen, 774 Tote
- 29 Länder betroffen
- eine zeitweise paralysierte asiatische Volkswirtschaft…

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und „Sequenzierung“

⇩

- Computer: Erkennen der Virusgene und -proteine (Genvorhersage)

Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)

⇩

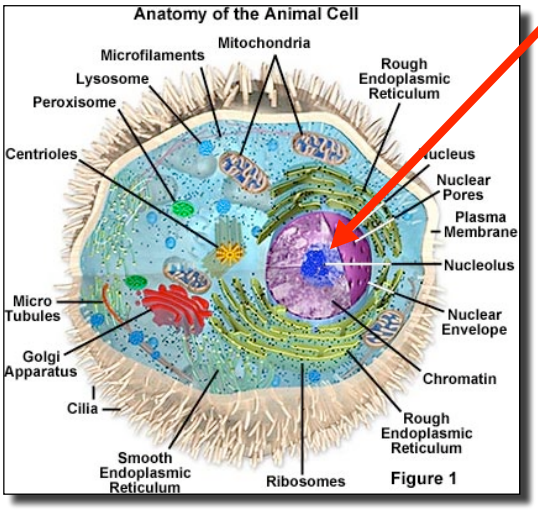
Verwandtschaft? (Phylogenetische Rekonstruktion)

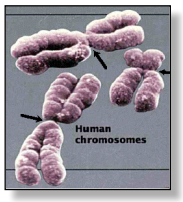
Struktur der Proteine? (Struktur-Vorhersage, -Modellierung)

Wirkstoff-Design

- Labor: Wirkstoff-Test

Jede Zelle enthält den Zellkern mit der genetischen Information, der DNA

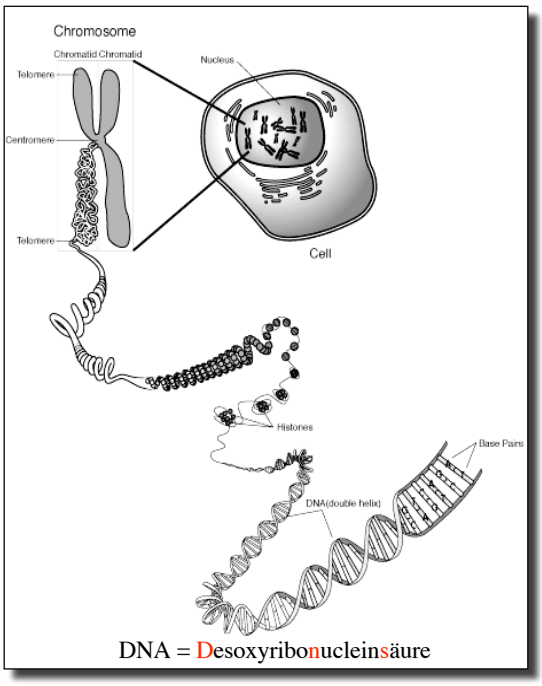




Gene sind Abschnitte auf einem langen, fädigen Molekül, der **DNA**.

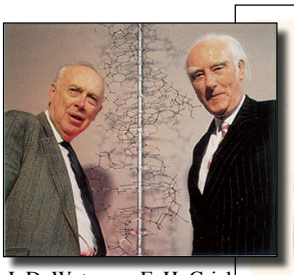
Die DNA ist auf **Chromosomen** aufgeteilt.

Das **Genom** ist die Gesamtzahl aller Gene einer Zelle.

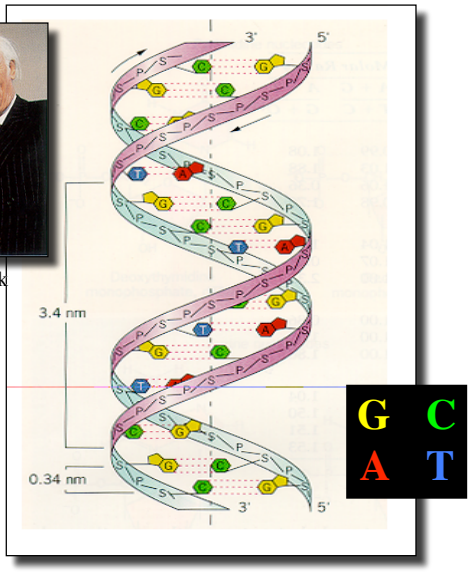


DNA = Desoxyribonucleinsäure

Die DNA besteht aus einer Abfolge (Sequenz) von 4 verschiedenen Bausteinen !



J. D. Watson F. H. Crick



Schreiben einer DNA-Sequenz...

- immer von links (5‘ Ende) nach rechts (3‘ Ende)
- meist nur ein Strang („Watson“ oder „Crick“)

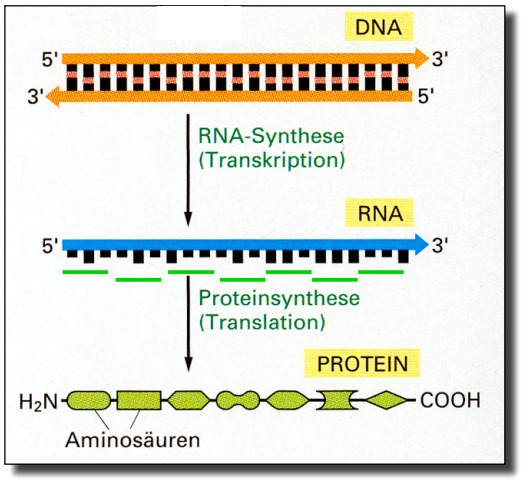
Beispiel:

5‘-GAGGGCTACTGCA-3‘

oder

5‘-TGCAGTAGCCCTC-3‘

Die Abfolge der 4 „Basen" der DNA enthält die Bauanleitung des Lebens !

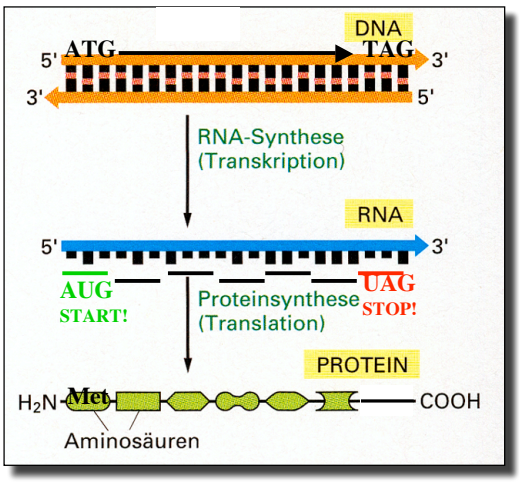


Informationsspeicher

Informationsabschrift

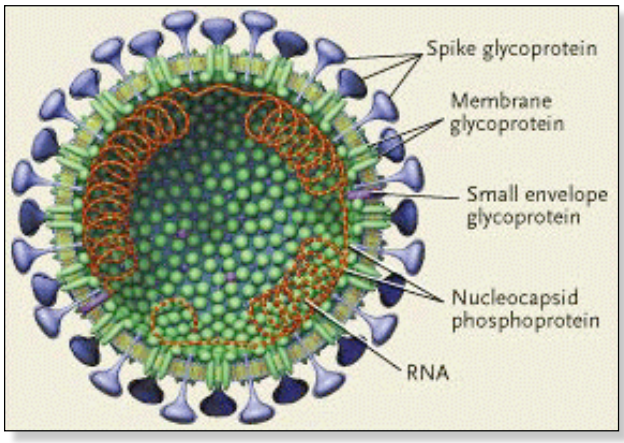
Produkt

Wie erkenne ich ein proteinkodierendes Gen?

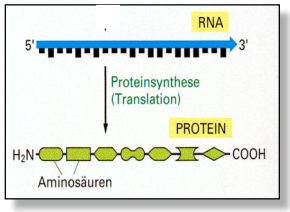


ORF
= offener Leserahmen

Viren haben eigene Erbinformation (manchmal aus RNA)



SARS Coronavirus



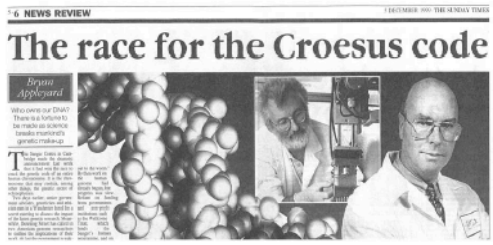
Die Ära der Genomforschung

„Even the smallest functional DNA varieties seen, those occurring in small phages, must have something like 5000 nucleotides in a row. We may, therefore, leave the task of reading the complete nucleotide sequence of a DNA for the next century, which will, however, have other worries.

Progress in Nucleic Acid Research and Molecular Biology, 1968

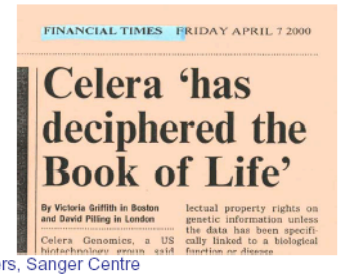
Phi-X 174 sequenced, Nature 1977

HGP: das Rennen!


6 NEWS REVIEW

The race for the Croesus code

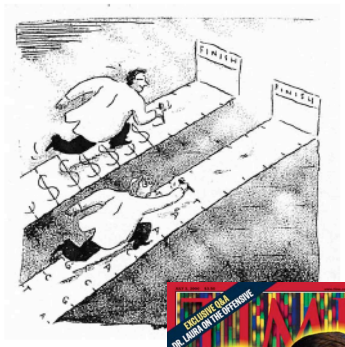
Brian Appleyard

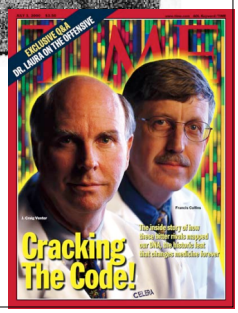

FINANCIAL TIMES FRIDAY APRIL 7 2000

Celera 'has deciphered the Book of Life'

By Victoria Griffith in Boston and David Pilling in London

Celera Genomics, a US biotechnology group, said ... lectual property rights on genetic information unless the data has been specifically linked to a biological function or disease





Source: Jane Rogers, Sanger Centre

Methoden der DNA-Sequenzierung

1977

- chemische Sequenzierung (**Maxam & Gilbert**)
- enzymatische Sequenzierung (**Sanger**)

synonym: > Kettenabbruch-Sequenzierung
> Didesoxy-Sequenzierung

Das Sanger-Verfahren

Sequenz bekannt Sequenz unbekannt

3‘-GATCCTGACATGAGGATCTAGATCCGTA……-5‘ **DNA-Matrize**

5‘-CTAGGACTGTAC-3‘ **>>>DNA-Synthese>>>** **Primer**

5‘-CTAGGACTGTAC T^{Stop}

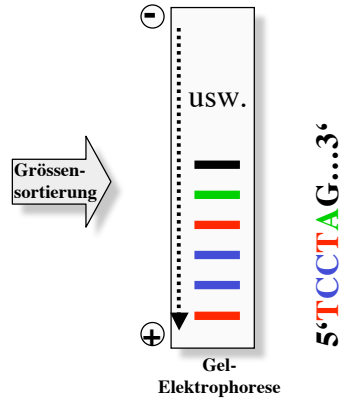
5‘-CTAGGACTGTAC TC^{Stop}

5‘-CTAGGACTGTAC TCC^{Stop}

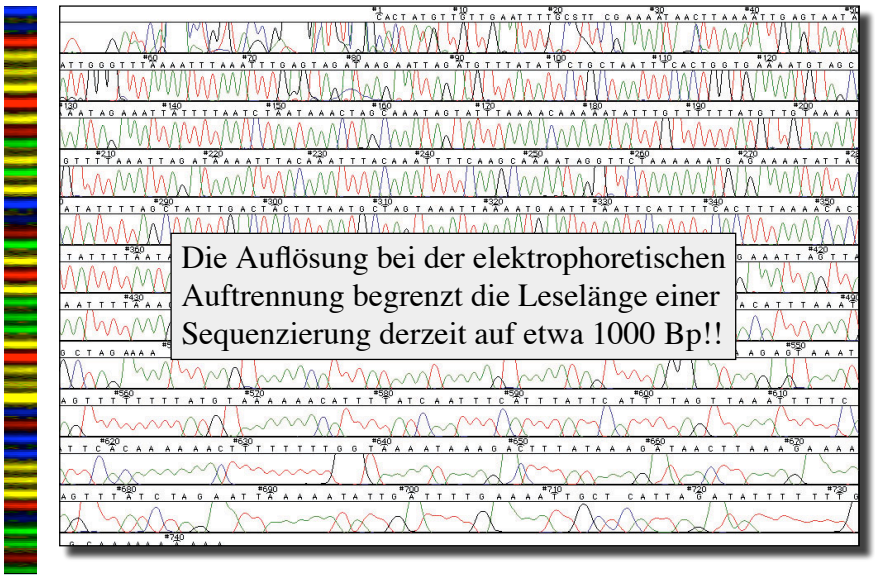
5‘-CTAGGACTGTAC TCCT^{Stop}

5‘-CTAGGACTGTAC TCCTA^{Stop}

5‘-CTAGGACTGTAC TCCTAG^{Stop}



Sequenzdaten- Chromatogramm

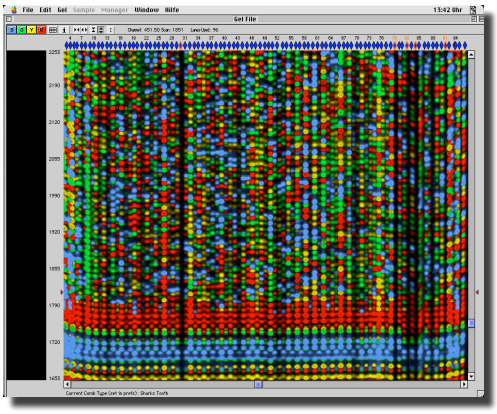


Die Auflösung bei der elektrophoretischen Auftrennung begrenzt die Leselänge einer Sequenzierung derzeit auf etwa 1000 Bp!!

Hochdurchsatz-DNA-Sequenzierung

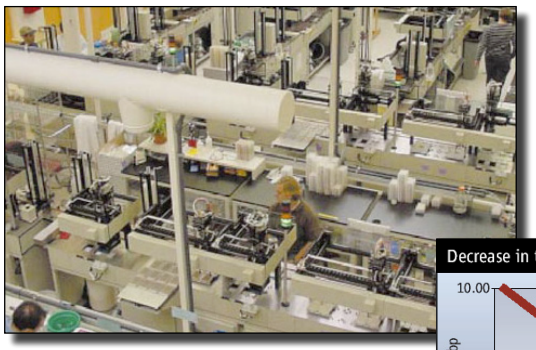
ABI 3730 Sequencer

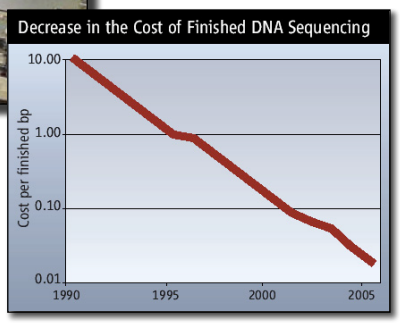
Kapillaren



96 Spuren x 1000 Basen = ca. 100 000 Basen in ca. 2 Std

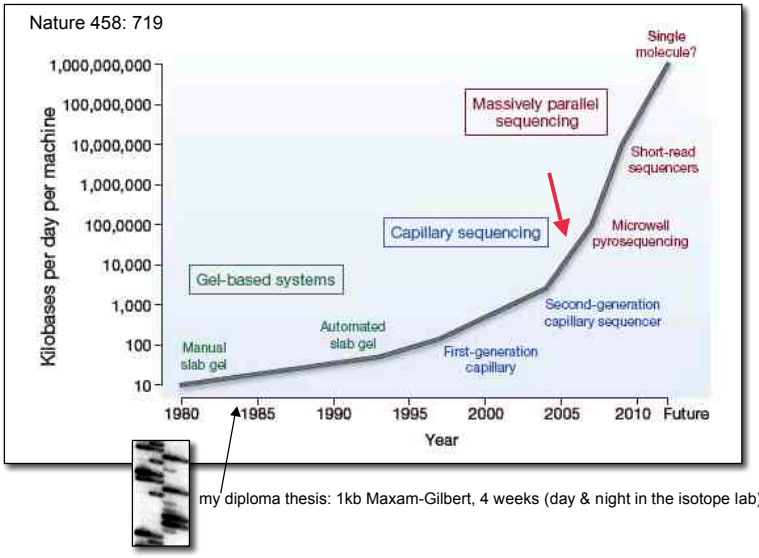
Sequenzierzentren arbeiten industriell...





Sequencing technology: A million-fold improvement!

Nature 458: 719



my diploma thesis: 1kb Maxam-Gilbert, 4 weeks (day & night in the isotope lab)

NGS technology: How to...

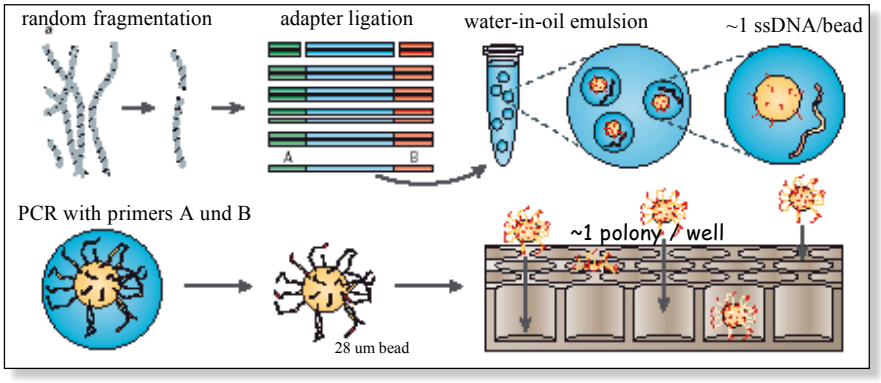
tedious cloning
high chemical costs
slow electrophoresis

PCR or even single molecules
extreme miniaturisation
massively-parallel read-out

Schritt 1:

Vermehrung der DNA

e. g. emulsion PCR

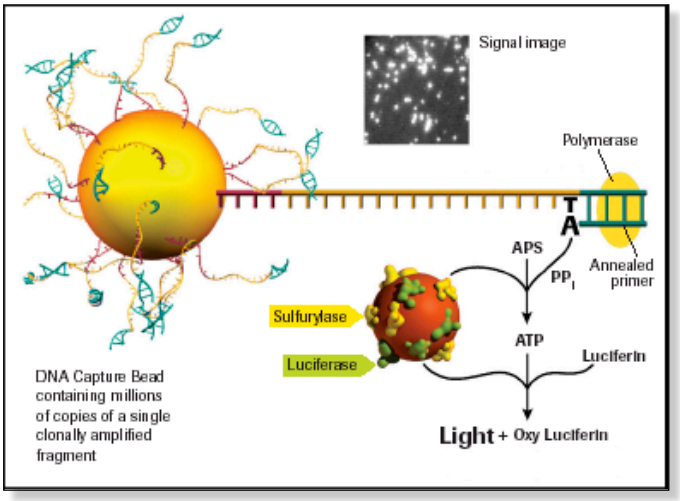


(2 mio wells in 454/FLX technology)

2.Schritt:

Sequenzierung ohne Separation

Beispiel: Pyro-Sequencing (Ronaghi et al. 1996, 1998)



Erstes von den 4 dNTPs wird zugegeben. Nur bei Einbau wird PPi frei.

Sulfurylase synthetisiert aus PPi und Adenosin-5-Phosphosulfat (APS) ein ATP.

ATP wird von Luciferase für Lichtemission benutzt.
Licht~ATP~PPi~Nt-Einbau

Apyrase spaltet restliches dNTP und ATP.

Next Generation Sequencing

	454 Roche	Illumina	ABI SOLiD	good ol' Sanger
DNA matrix	emulsion PCR, (28 µm beads)	bridge PCR, isothermal (10^6 /cm^2)	emulsion PCR, (1 µm beads)	plasmid clones
sequencing method	seq-by-synthesis: Pyrosequencing	seq-by-synthesis: ‚reversible' Dye-Terminators	sequencing-by-ligation	Dye-terminator 96 capillaries
read length	400 bp (up to 1000?)	2 x 75 bp (up to 2x100?)	35 bp or 2x25 (up to 100?)	up to 1000 bp
reads	up to 1.5 Mio	up to 270 Mio	up to 320 Mio	96 per run
data	up to 600 Mbp	up to 27 Gbp	up to 32 Gbp	0.1 Mbp
runtime	10 hrs	9 days	10 days	2 hrs

Celebrity genomics without the Y chromosome: Glenn Close has her genome sequenced

Category: **complete genomics** • **illumina** • **next-generation sequencing** • **solexa** • **whole-genome sequencing**

Posted on: March 11, 2010 9:30 AM, by **Daniel MacArthur**

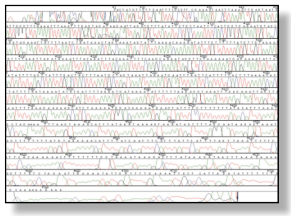
Zoe McDougall from Oxford Nanopore points me to a press release from Illumina announcing a new era of celebrity genomics:

> Illumina, Inc. (NASDAQ:ILMN) today announced that it has sequenced the DNA of American actress Glenn Close, the first publicly named female to have her DNA sequenced to full coverage. The service was completed in Illumina's CLIA certified and CAP accredited laboratory utilizing Illumina's Genome Analyzer technology and following the established process shown at http://www.everygenome.com/. Ms. Close's DNA was sequenced to an average depth greater than 30 fold, providing information on SNP variation and allowing for the analysis of other structural characteristics of the genome such as insertions, deletions and rearrangements. Specifically, over 95% of the known genome was reported, including over 12 million genotype calls on previously documented SNPs. In addition, 379,000 SNPs previously not reported in any public database were found.

While there's nothing new about celebrity genomics, previous examples have largely been "scientific celebrities" (such as Jim Watson and Craig Venter) - so Close is the first genome with broader celebrity status, and also the first named individual without a Y chromosome to rack up her 6 billion base pairs. That's of negligible interest scientifically, but there's no doubt this will dramatically increase the public profile of whole genome sequencing.

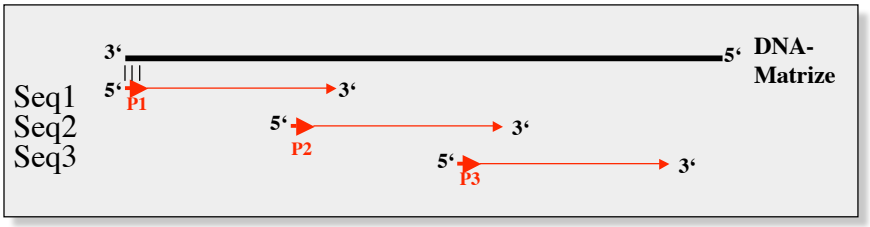
Sequenzierungsstrategien sind erforderlich!

Die Auflösung bei der elektrophoretischen Auftrennung begrenzt die Leselänge einer Sanger Sequenzierung auf etwa 1000 Bp!!



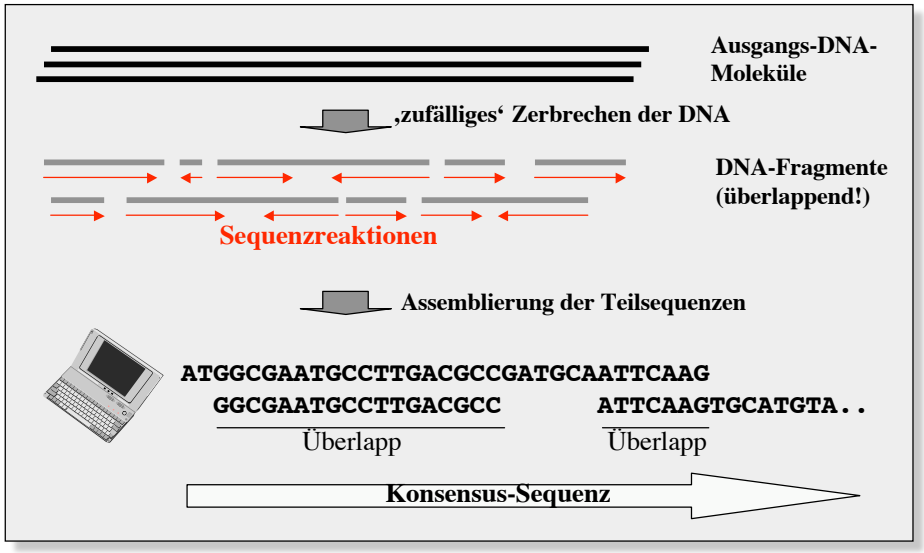
Längere DNA-Moleküle (z. B. ganze Genome) müssen schrittweise (in kleinen Stücken) sequenziert werden. Diese DNA-Sequenzstücke müssen dann zum Genom zusammen-Gesetz werden („Assemblierung“).

Die ‚Primer Walking'-Strategie



- sequentieller Ablauf > langsam
- geordnete Strategie > übersichtlich
- vergleichsweise teuer (Primer kosten Geld)

Die ‚shotgun'-Strategie



Alignment: die Schlüssel-Technik der Bioinformatik!

```
Query: 1   tctacggggccgtagtgcaggccatgagtcgaggctgggatggcgagtaagag 53
           ||||||| || || ||||| |||||||| |||||||||||| || |||||||||
Sbjct: 616 tctacggagctgtggtgcaagccatgagccgaggctgggacggggagtaagag 668
```

Nt-Substitution

As-Austausch

Gap bzw. InDel

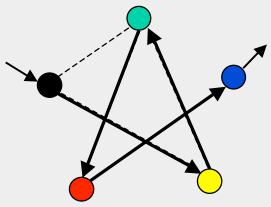
```
Query: 5   EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQY--NCRQFSSPEDCLSSPEFL 62
           + ELIR SW ++ ++ + HG +LF+RLF L+P+LL LF Y  NC     S +DCLSSPEFL
Sbjct: 8   DKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYTTNC---GSTQDCLSSPEFL 64
```

ähnliche As

identische As

Assemblierung von Sequenzen:
Das ‚shortest common superstring' Problem

•Sequenz-Reads als ‚Knoten'
•Überlappung als ‚Kanten'



Hamilton-Pfad
…passiert jeden Knoten nur einmal!

Konsensus-String aus Hamilton-Pfad
ergibt die gesuchte Gesamtsequenz

Assemblierung der Gesamtsequenz aus Einzel-Reads

Reads = { **TTACTAC, TTTTATG, GCATGCC,**
TAAGGTT, ACCCCAG, GCATGCA }

5' **AACCTTACTACTGGGGTTTTATGCATGCATGCC** 3' Watson
3' **TTGGAATGATGACCCCAAAATACGTACGTACGG** 5' Crick

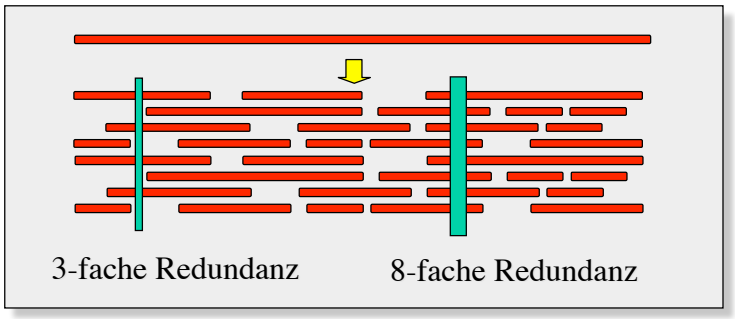
Der Assembly-Algorithmus vergleicht automatisch die Reads und ihre „Reverse Complements".
Er schreibt dann allerdings nur einen Strang auf....

5'**AACCTTACTACTGGGGTTTTATGCATGCATGCC**3'

Assemblierung der Gesamtsequenz aus Einzel-Reads

- Die Einzelsequenzen enthalten üblicherweise Fehler.
- Der Algorithmus muss also nach definierten Kriterien eine Konsensus-Sequenz erstellen.

Die Abdeckung der Gesamtsequenz erfordert eine „Redundanz"

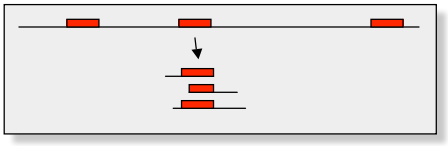


Ideal zur Absicherung der Sequenz an jeder Position ist eine Redundanz von 10x!

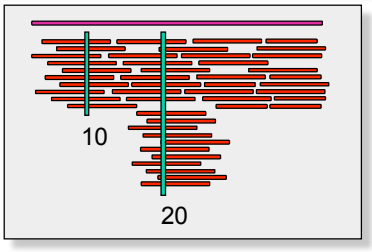
<u>Problem</u>: Aufwand, Kosten!

Probleme beim Assembly

- „Repeats“: besonders problematisch, wenn
  - > repeats länger als Leseweite sind
  - > repeats fast identisch sind

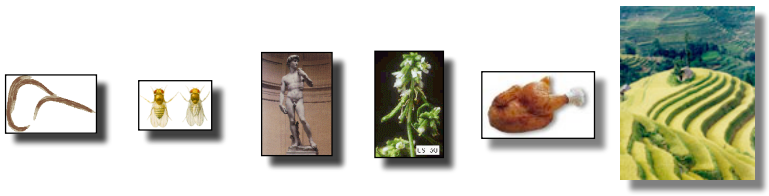


falsches „alignment“
aufgrund starker Ähnlichkeit
repetitiver Sequenzkopien



Überproportionale Redundanz im Alignment
zeigt problematische Stellen mit Repeats an

Genomgrößen im Vergleich



• Hefe	12 069 kb	6 607 Gene
• Fadenwurm	97 000 kb	20 178 Gene
• Fliege	137 000 kb	13 601 Gene
• Homo sapiens	>3 000 000 kb	19 042 Gene?
• Reis	400 000 kb	>50 000 Gene !
• Ackerschmalwand	125 000 kb	>25 500 Gene
• Huhn	1 000 000 kb	<23 000 Gene

Das Genom des SARS-Virus

- 1 Monat nach Virus-Identifikation 2 Genome sequenziert!
- Länge : 29 740 Bp (RNA)
- nach 3 Monaten > 20 Virus-Isolate sequenziert

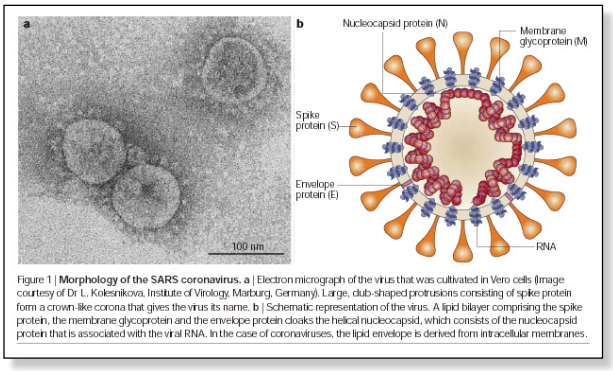


Figure 1 | **Morphology of the SARS coronavirus.** a | Electron micrograph of the virus that was cultivated in Vero cells (Image courtesy of Dr L. Kolesnikova, Institute of Virology, Marburg, Germany). Large, club-shaped protrusions consisting of spike protein form a crown-like corona that gives the virus its name. b | Schematic representation of the virus. A lipid bilayer comprising the spike protein, the membrane glycoprotein and the envelope protein cloaks the helical nucleocapsid, which consists of the nucleocapsid protein that is associated with the viral RNA. In the case of coronaviruses, the lipid envelope is derived from intracellular membranes.

Review: Stadler et al. (2003) Nature Reviews Micobiology 1, 209-218

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und „Sequenzierung“
- Computer: Erkennen der Virusgene und -proteine (Genvorhersage)

  Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)

  Verwandtschaft? (Phylogenetische Rekonstruktion)

  Struktur der Proteine? (Struktur-Vorhersage, -Modellierung)

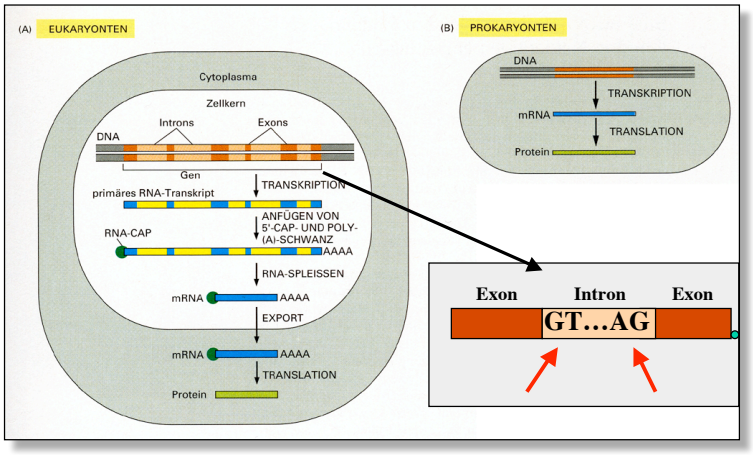
  Wirkstoff-Design
- Labor: Wirkstoff-Test

Genvorhersage und Genomannotation

Wo steckt denn nun die genetische Information?

```
  1 ccgaacgctt atagagagct atagagtgaa agctgagaag aaccaaaacg gagcataaac
 61 atgaacagcg atgaggtgca actgatcaag aagacctggg aaatccccgt ggcaacacca
121 acagattctg gagcggcgat actgacgcag tttttcaacc gctttccgtc caacttggag
181 aagttcccct tccgcgatgt tcctttggag gagctaagtg tgagttgtac cttacacata
241 ggtcttcaat taactcaaga ttaacttgat ctgttttctt tcagggaaat gctcgcttcc
301 gagcacatgc cggcagaatc ataagggtct ttgacgagtc catccaggtc ctgggccagg
361 atggcgatct ggagaagctg gacgagatct ggaccaaaat tgccgttagt cacattccgc
421 ggaccgtttc caaggagtct tacaacgtaa gttgaacact gcagtcgagc tctcgacttt
481 gagatacctg ttggtcagat agtggaagtt gaaagctata tgacatttaa aaattcaatt
541 gcatttaaaa catcatttta tttttttag caactgaaag gagttatcct ggatgtgctg
601 acagctgcct gcagtctgga cgagagtcaa gcggccacgt gggccaagct ggtggaccat
661 gtctacgcaa tcatcttcaa ggcgatcgac gacgacggca acgccaagta gatgaggcag
721 ctggaggtgg agatgcaacc gaatccgcgg a
```

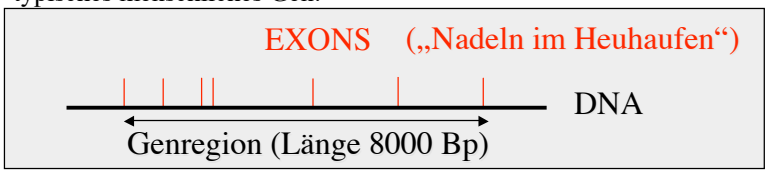
Bei Eukaryoten-Genomen ist Generkennung besonders schwierig



Die Gene bestehen aus proteinkodierenden Abschnitten („Exons“) und nicht-kodierenden „Introns“, die durch Spleißen aus der mRNA entfernt werden.

Das Problem der Gen-Identifizierung

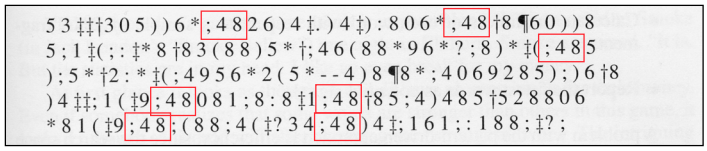
typisches menschliches Gen:



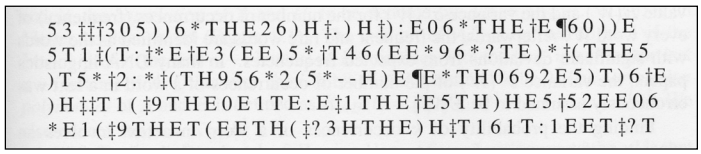
- Funktionelle Teile eines Gens sind als Schnipsel (Exons) verteilt (durchschnittliche Länge: nur 145 Basenpaare)

Alles geht! Oder:
Edgar Allen Poe und die DNA-Linguistik

Zum Schatz von Captain Kidd... („The Gold-Bug“)



- häufigstes engl. Wort? ;48 the



Die (vereinfachte) Aufgabe:

- gegeben sind uncharakterisierte Genom-DNA-Sequenzen

- FINDE...
  - Protein-kodierende Regionen
  - Exon/Intron-Grenzen
  - mögliche genregulatorische Abschnitte

Mache daraus ein Modell für die Struktur des Gens!

Warum „vereinfacht"?

- nicht alle Gene werden in Proteine übersetzt!
  (RNA-Gene)

- auch nicht alle Genregionen proteinkodierender Gene werden in Proteine übersetzt
  (5' und 3'-untranslatierte Exons)

- Gene werden alternativ gespleißt.
  Die ALT-mRNAs können unterschiedliche Proteine kodieren.

Welche „Signale" von Genen kennen wir?

- Repetitive DNA = keine Gene > wegfiltern

- Startkodons, Stopkodons > ORFS („open reading frames")
- Spleiß-Donor/Akzeptor-Stellen ("GT-intron-AG")

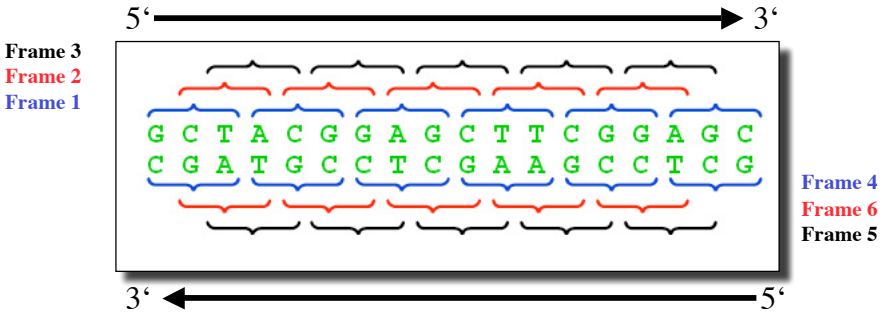
- Promoter: Bindemotive für Transkriptionsfaktoren („Boxen")
  Startpunkt der Transkription (+1, cap site)
  CpG-Inseln
- Polyadenylierungssignal (AATAAA) am Ende des Transkripts

Welchen besonderen „Inhalt" haben Gene?

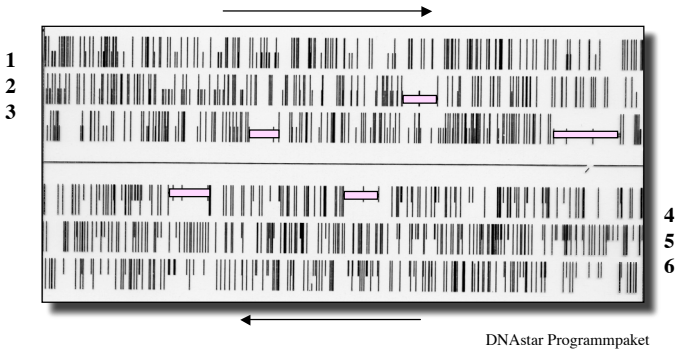
- „codon usage" innerhalb von ORFs

Proteinkodierende Gene haben einen „besonderen Inhalt“

- sie lassen sich als einen „offenen Leserahmen“ (ORF) lesen, d. h. in eine ununterbrochene Aminosäurefolge übersetzen



Suche nach ORFs



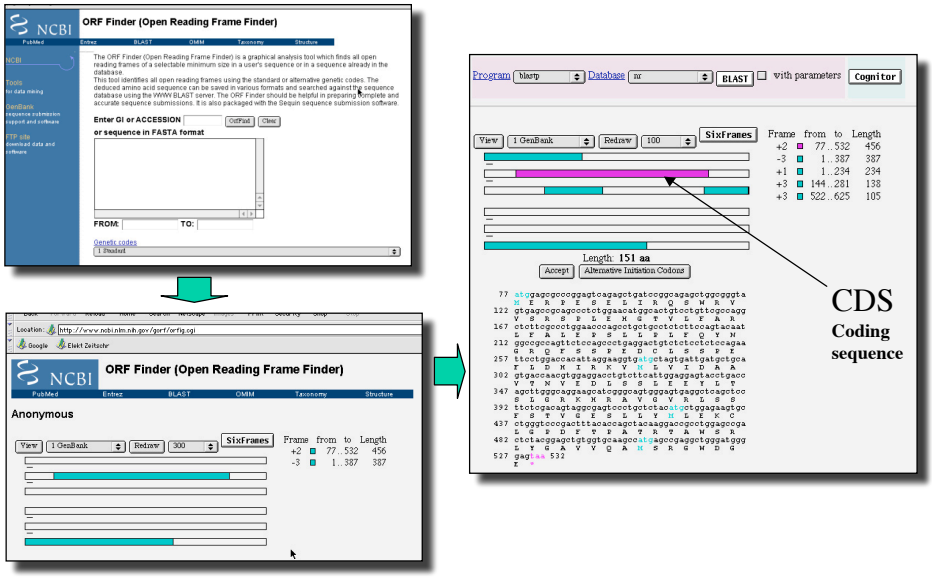
DNAstar Programmpaket

Start
Stop

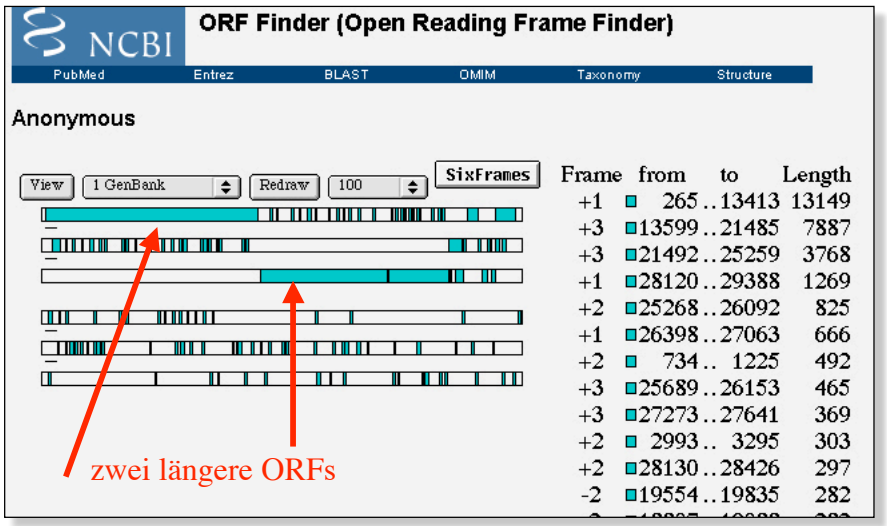
Potenzielle Gene

Der NCBI-ORFfinder

http://www.ncbi.nlm.nih.gov/gorf/gorf.html



NCBI-ORFfinder: SARS-Genom



SARS-Genom und seine Gene

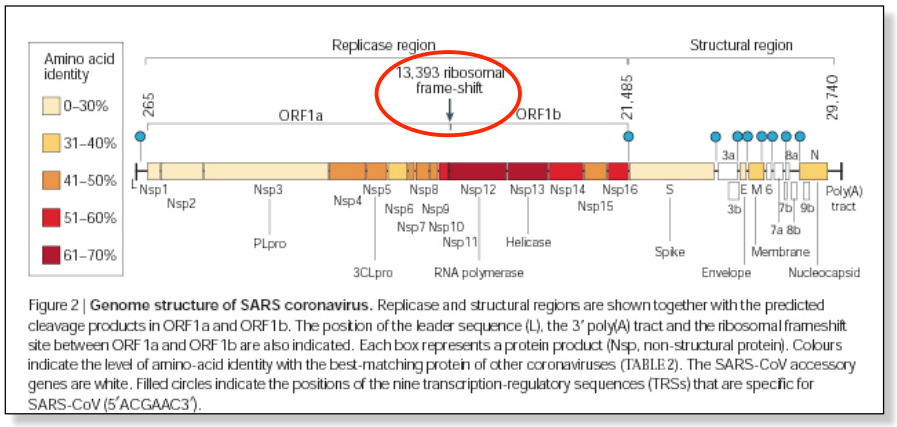


Figure 2 | **Genome structure of SARS coronavirus.** Replicase and structural regions are shown together with the predicted cleavage products in ORF1a and ORF1b. The position of the leader sequence (L), the 3′ poly(A) tract and the ribosomal frameshift site between ORF1a and ORF1b are also indicated. Each box represents a protein product (Nsp, non-structural protein). Colours indicate the level of amino-acid identity with the best-matching protein of other coronaviruses (TABLE 2). The SARS-CoV accessory genes are white. Filled circles indicate the positions of the nine transcription-regulatory sequences (TRSs) that are specific for SARS-CoV (5′ACGAAC3′).

ORF1a und ORF 1b werden zunächst in <u>ein</u> Protein übersetzt (ORF1a/b), (was dann in mehrere Teilproteine zerlegt wird)

SARS-Genom und seine Gene

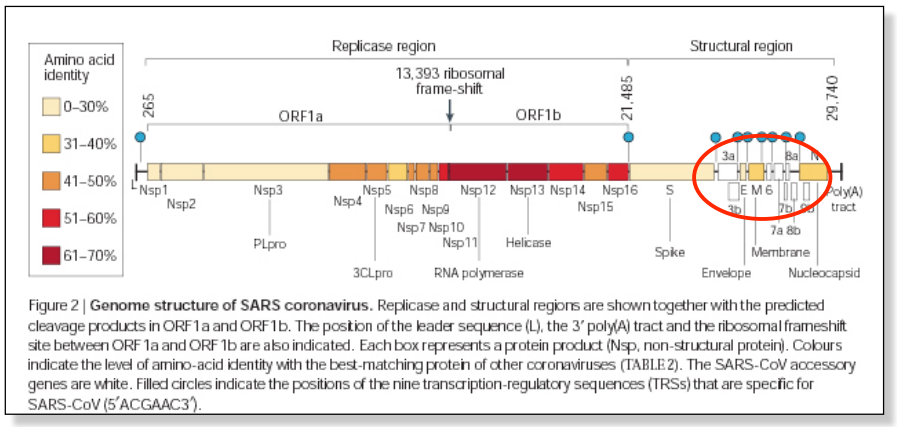


Figure 2 | **Genome structure of SARS coronavirus.** Replicase and structural regions are shown together with the predicted cleavage products in ORF1a and ORF1b. The position of the leader sequence (L), the 3′ poly(A) tract and the ribosomal frameshift site between ORF1a and ORF1b are also indicated. Each box represents a protein product (Nsp, non-structural protein). Colours indicate the level of amino-acid identity with the best-matching protein of other coronaviruses (TABLE 2). The SARS-CoV accessory genes are white. Filled circles indicate the positions of the nine transcription-regulatory sequences (TRSs) that are specific for SARS-CoV (5′ACGAAC3′).

Virengenome haben im Gegensatz zu Eukaryoten-Genomen häufig <u>überlappende Genbereiche!</u>

SARS-Gene und Proteine

Table 1 | **Predicted SARS-CoV proteins**

ORF	SARS-CoV proteins	Length (amino acids)	Position in the polyprotein	Functional and structural predictions
Replicase region				
ORF1a	Nsp1	180	1M–180G	?
	Nsp2	638	181A–818G	?
	Nsp3 (PLpro)	1922	819A–2740G	Papain-like cysteine protease–cleavage of Nsp1–Nsp4, adenosine diphosphate-ribose 1-phosphatase (ADRP), 2 TMD
	Nsp4	500	2741K–3240Q	3 TMD
	Nsp5 (3CLpro)	306	3241S–3546Q	3C-like cysteine protease–cleavage of Nsp4–Nsp16
	Nsp6	290	3547G–3836Q	5 TMD
	Nsp7	83	3837S–3919Q	?
	Nsp8	198	3920A–4117Q	?
	Nsp9	113	4118N–4230Q	?
	Nsp10	139	4231A–4369Q	Growth-factor-like domain
	Nsp11	13	4370S–4382V	?
ORF1b	Nsp12 (RdRp)	932	4370S–5301Q	RNA-dependent RNA polymerase
	Nsp13 (Helicase)	601	5302A–5902Q	Helicase, zinc-binding domain, NTPase
	Nsp14	527	5903A–6429Q	Exonuclease (ExoN homologue)
	Nsp15	346	6430S–6775Q	EndoRNAse (XendoU homologue)
	Nsp16	298	6776A–7073N	mRNA cap-1 methyltransferase
Structural region				
ORF2	Spike (S) protein	1255		1 TMD, ≥12 *N*-glycosylation sites
ORF3a	?	274		2 TMD, 1 *N*-glycosylation site, 10 *O*-glycosylation sites
ORF3b	?	154		?
ORF4	Envelope (E) protein	76		1 TMD, 2 *N*-glycosylation sites
ORF5	Membrane (M) protein	221		3 TMD, 1 *N*-glycosylation site
ORF6	?	63		1 TMD
ORF7a	?	122		1 TMD
ORF7b	?	44		1 TMD
ORF8a	?	39		Membrane-associated
ORF8b	?	84		1 *N*-glycosylation site
ORF9a	Nucleocapsid (N) protein	422		
ORF9b	?	98		1 *O*-glycosylation site

The analyses are based on the sequence of the SARS-CoV FRA isolate (GenBank accession number AY310120). Transmembrane domains (TMDs) were predicted using the program PSORT (threshold is less than –2); the glycosylation sites were predicted using the NetNGlyc server (see NetNGlyc in the Online links). Information on the functional predictions has been taken from REFS 20,33. Nsp, non-structural protein.

ORFs mit z.T. unbekannter Identität und Funktion

ORF-Suche ist nicht ausreichend, um Genmodelle vorherzusagen!

Moderne integrierte Genvorhersage-Programme verbinden Suche nach Signalen mit neueren statistischen Methoden...

...Hidden Markov Models (HMM)

Markov WER??

- Andrei Andreyevich Markov (1856-1922)
- Markov-Kette:

Eine *Markovkette* ist ein stochastischer Prozess, der nacheinander eine Reihe von Zuständen mit einer gewissen Wahrscheinlichkeit durchläuft. Dabei hängt die Wahrscheinlichkeit für den jeweils nächsten Zustand nur vom aktuellen Zustand ab:

$P(t_{i+1}|t_i, t_{i-1},\ldots,t_j) = P(t_{i+1}|t_i)$

Pfeile geben Übergangswahrscheinlichkeiten an

Hidden Markov Models

- verwende **statistische Informationen,** um Abfolgen (z. B. Sequenzen) zu klassifizieren
- Analogie:

  „Automatische Erkennung der Sprache eines Textes“

  In einem typischen deutschen Text macht der Buchstabe ‚e‘ ca. 16,55% aller Buchstaben aus, in einem schwedischen nur ca. 9.77%.

  ⇒ zähle die e‘s im Text, um zu berechnen mit welcher Wahrscheinlichkeit es sich um einen deutschen Text handelt

Hidden Markov Models

Was ist denn da „hidden“??

- wir ***sehen*** nur die **„e‘s“** — **„emission“**
- dahinter ***versteckt*** sich die Information:

  „dies ist ein deutscher Text“ — **„state“**

Hidden Markov Models

- Anwendungsgebiete in der Bioinformatik:

  > Vorhersage der Genstruktur (Exons/Introns)

  > Vorhersage von Promoterbereichen

  > Erstellung von Modellen für Proteinfamilien zum Suchen nach entfernt verwandten Proteinen in DB („profile HMMs“)

Von der reinen Textsuche zum HMM

1 ACA---ATG
2 TCAACTATC
3 ACAC--AGC
4 AGA---ATC
5 ACCG--ATC

Bsp.: Fünf Sequenzen, die ein funktionell wichtiges Signal definieren

Textsuche würde erfolgen nach:

`(AT)(GC)(AC)(ACGT)* A(TG)(GC)`

Kann bei Suche nicht unterscheiden zwischen...

...einer plausiblen Sequenz (zB der Konsensus-S.)

`ACAC--ATC`

...und einer höchst unwahrscheinlichen Sequenz

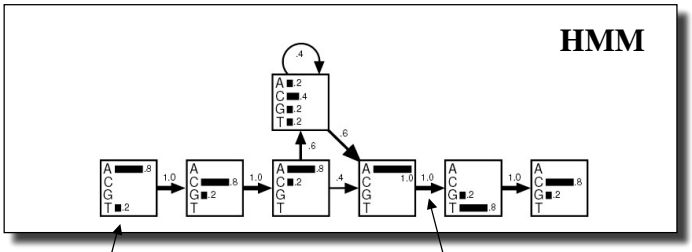
`TGCT--AGG`

Von der reinen Textsuche zum HMM

Also besser:

Bewerten, ob Sequenzabfolge „gut“ in das Alignment passt...

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



Emissions-
wahrscheinlichkeit

Übergangs-
wahrscheinlichkeit

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und „Sequenzierung“
- Computer: Erkennen der Virusgene und -proteine (Genvorhersage)

Ähnlichkeit zu bekannten Genen oder Proteinen? (Datenbanksuchen)

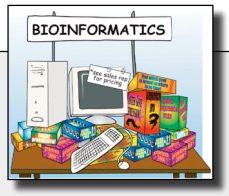
Verwandtschaft? (Phylogenetische Rekonstruktion)

Struktur der Proteine? (Struktur-Vorhersage, -Modellierung)

Wirkstoff-Design

- Labor: Wirkstoff-Test

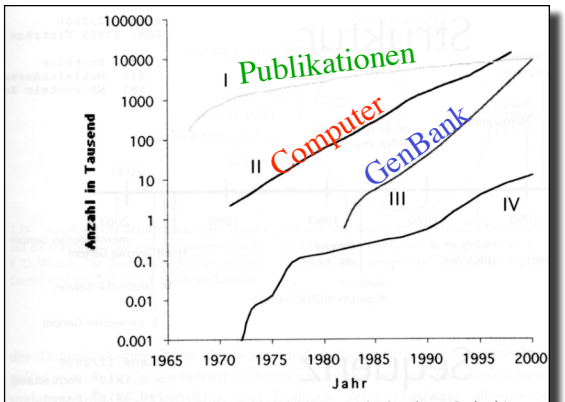
Datenbanken in der Molekularbiologie



- **Literaturdatenbanken**
- **Sequenzdatenbanken**
  - primäre DB: annotierte DNA- u. Proteinsequenzen
  - abgeleitete DB: interpretierte Sequenzdaten

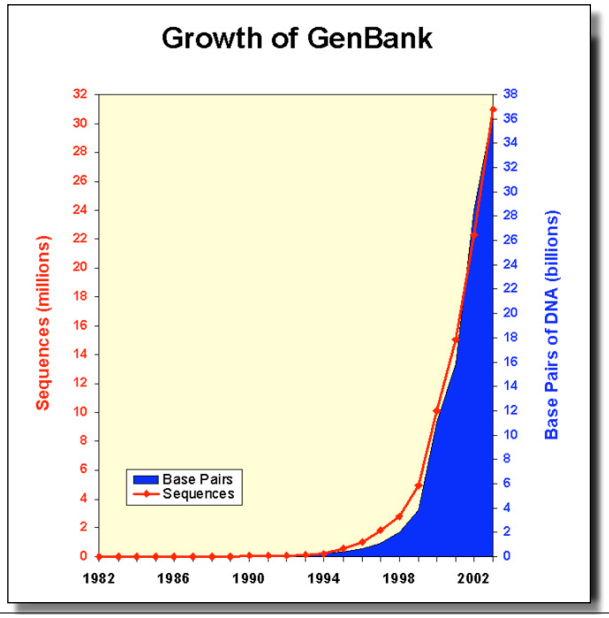
    (z.B. Proteindomänen oder Stoffwechselwege)

Datenbanken-Wachstum



1.10 Kurve I zeigt den Anstieg der in MEDLINE enthaltenen Literaturreferenzen. (Quelle: NLM Annual Reports).
Kurve II beschreibt die Entwicklung der Leistungsfähigkeit von Computerhardware (Moore's Law: Anzahl von Transistoren/Chip), [http://www.physics.udel.edu/wwwusers/watson/scen103/intel.html].
Kurve III beschreibt die exponentielle Zunahme der GenBank Einträge [http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html].
Kurve IV beschreibt das Wachstum der 3D Strukturdatenbank PDB [http://www.rcsb.org/pdb/holdings_table.html].
Es ist offensichtlich, daß sich im Verhältnis der Sequenzzahlen zur Rechnerleistung und zur Anzahl der Veröffentlichungen zur Zeit eine ungünstige Schere öffnet.

Datenbank-Wachstum



22,617,000,000 bases in 18,197,000 sequence records (August 2002)

35,599,621,471 bases in 29,819,397 sequence records (Oct 2003)

43,194,602,655 bases in 38,941,263 sequence records (Oct 2004)

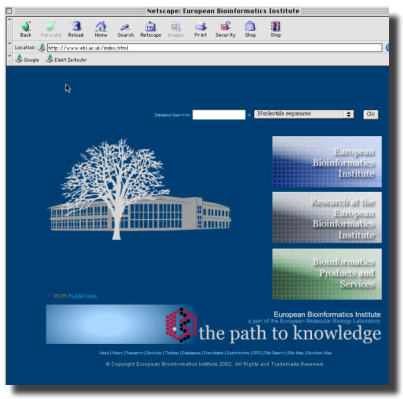
Datenbanken in der Molekularbiologie

http://www.ncbi.nlm.nih.gov/

National Center for Biotechnology Information, Am NIH, Bethesda, Maryland, USA

http://www.ebi.ac.uk

European Bioinformatics Institute, Sanger Campus, Hinxton, GB

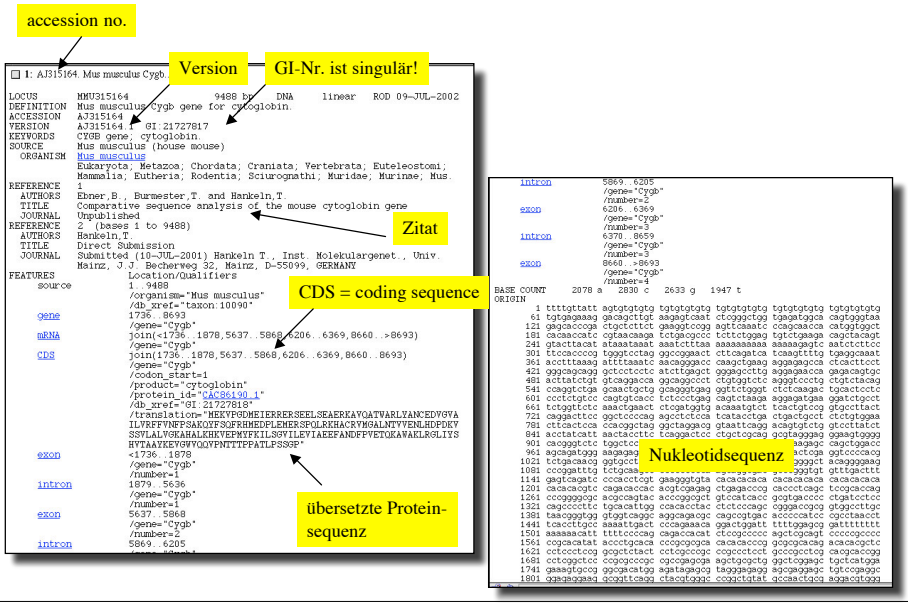


Sequenz-Datenbanken

NCBI	> GenBank (1979)
EBI	> EMBL database (1980)
Genome-Net	> DDBJ = DNA database of Japan (1984)

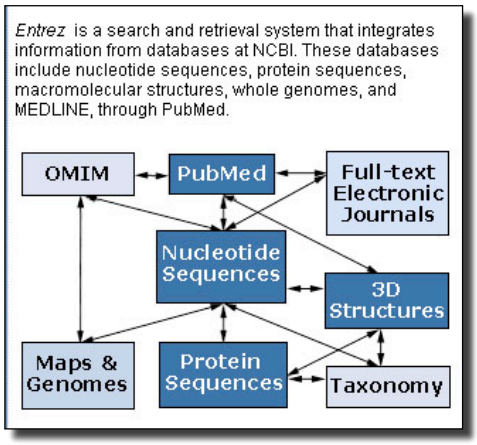
- Täglicher Abgleich erfolgt zwischen allen drei Datenbanken
- dennoch Unterschiede in der Redundanz und Annotations-Präzision
- jeder darf Einträge vornehmen!

Ein GenBank-Eintrag



Integrierte Such-Werkzeuge!

- Entrez /NCBI
- SRS sequence retrieval system /EBI



www.ncbi.nlm.nih.gov/Entrez/

Suche in Sequenzdatenbanken

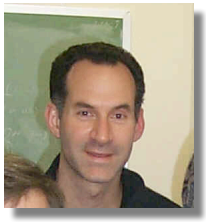
Eine bekannte verwandte Sequenz in den Datenbanken ermöglicht einen ersten schnellen Hinweis auf die Identität und Funktion einer unbekannten Sequenz.

Populärstes (und schnellstes) Werkzeug:

BLAST (Altschul et al. 1991, 1997)

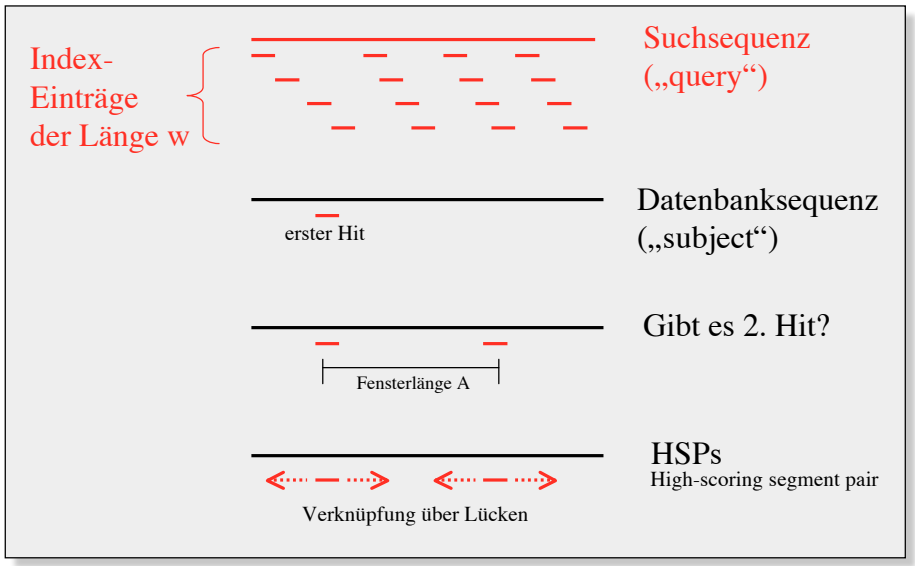
„Basic Local Alignment Search Tool"

Stephen Altschul

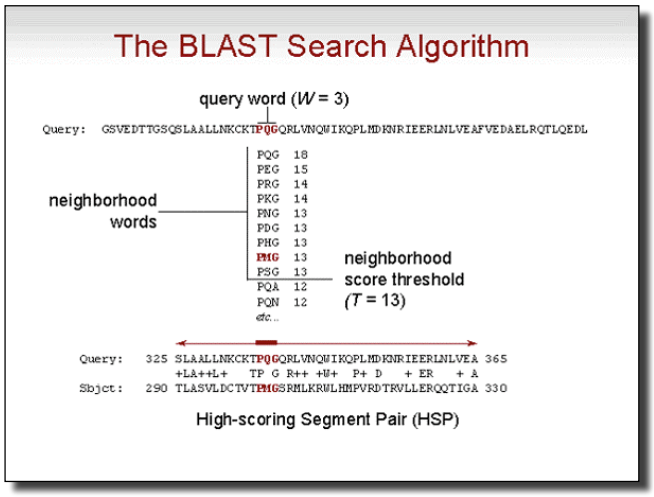


David Lipman

BLAST



BLAST



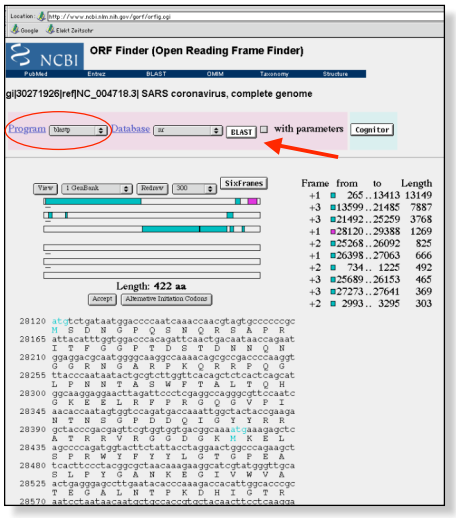
(word size W = 11 bei DNA)

- zunächst wird nach kurzen lokal passenden Abschnitten („words“) gesucht
- dabei werden auch **ähnliche** word-hits akzeptiert
- dann versucht BLAST, die Bereiche neben den „matching words“ unter Einbeziehung von Lücken zu optimieren

BLAST

1. Suchsequenz wird in ‚words‘ der Länge w „zerbrochen“
2. mit Index dieser ‚words‘ wird Datenbank durchsucht
3. ein „word hit“ liegt vor, wenn das ‚word‘ exakt oder in ähnlicher Form (threshhold-Score >T) erkannt wird
   > word size kann hoch bleiben (speed) ohne Sensitivitätsverlust
   > erhöhe T : weniger ‚background words‘, schneller
   > erniedrige T : entfernte Verwandschaften zu finden
4. ausgehend von ‚word hit‘ wird lokales optimales alignment verlängert, bis Score S durch mismatches stark abfällt (= HSP, high-scoring segment pair)
   > dabei können kleine Lücken toleriert werden

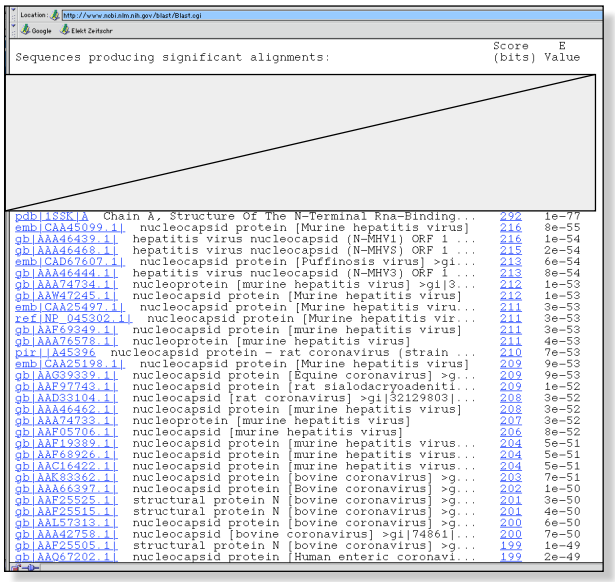
BLAST-Suche der ORFs des SARS-CoV



Programm BLASTP
d.h. Suche auf Proteinsequenzebene

Datenbank nr (non-redundant)

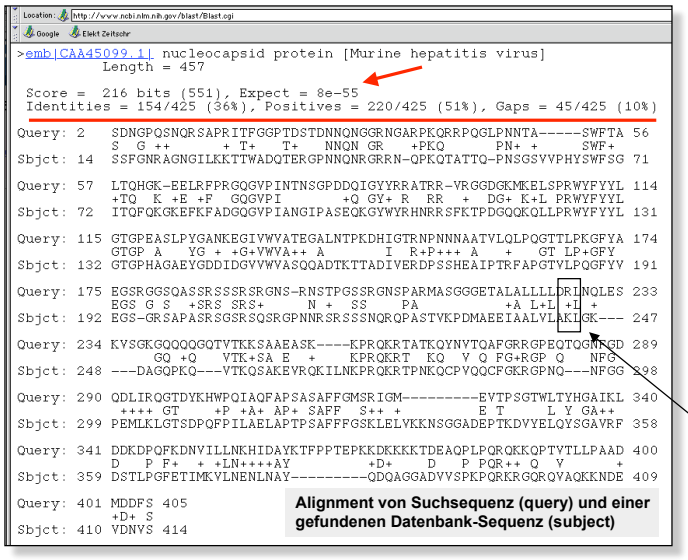
BLAST-Suche: die Ergebnisliste



..es gibt natürlich schon einige SARS-Datenbankeinträge

…unser ORF ist ähnlich zu **Nukleo-capsid-Protein** anderer **Coronaviren!**

BLAST-Suche: ein Alignment!



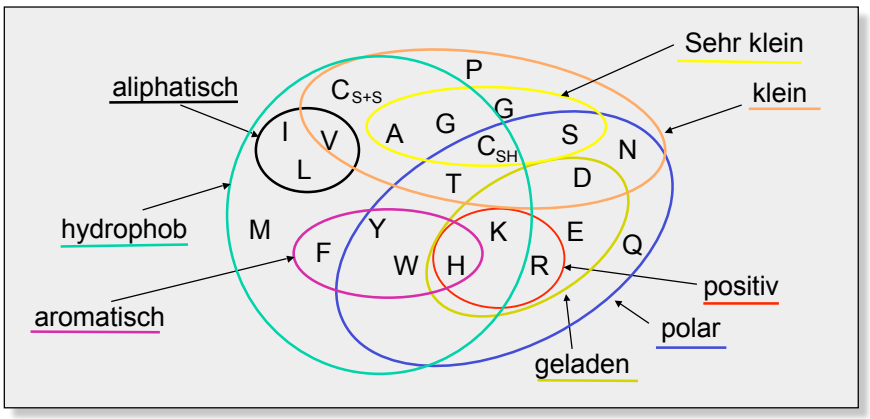
```
>emb|CAA45099.1| nucleocapsid protein [Murine hepatitis virus]
          Length = 457

 Score =  216 bits (551), Expect = 8e-55
 Identities = 154/425 (36%), Positives = 220/425 (51%), Gaps = 45/425 (10%)

Query: 2   SDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTA-----SWFTA 56
           S  G ++      + T+   T+   NNQN GR   +PKQ      PN+ +     SWF+
Sbjct: 14  SSFGNRAGNGILKKTTWADQTERGPNNQNRGRRN-QPKQTATTQ-PNSGSVVPHYSWFSG 71

Query: 57  LTQHGK-EELRFPRGQGVPINTNSGPDDQIGYYRRATRR-VRGGDGKMKELSPRWYFYYL 114
           +TQ  K +E +F  GQGVPI        +Q GY+ R  RR  +   DG+ K+L PRWYFYYL
Sbjct: 72  ITQFQKGKEFKFADGQGVPIANGIPASEQKGYWYRHNRRSFKTPDGQQKQLLPRWYFYYL 131

Query: 115 GTGPEASLPYGANKEGIVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFYA 174
           GTGP A    YG + +G+VWVA++ A       I  R+P+++ A    +   GT LP+GFY
Sbjct: 132 GTGPHAGAEYGDDIDGVVWVASQQADTKTTADIVERDPSSHEAIPTRFAPGTVLPQGFYV 191

Query: 175 EGSRGGSQASSRSSSRSRGNS-RNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLES 233
           EGS G S  +SRS SRS+     N +  SS    PA           +A L+L +L +
Sbjct: 192 EGS-GRSAPASRSGSRSQSRGPNNRSRSSSNQRQPASTVKPDMAEEIAALVLAKLGK--- 247

Query: 234 KVSGKGQQQQGQTVTKKSAAEASK----KPRQKRTATKQYNVTQAFGRRGPEQTQGNFGD 289
           GQ +Q   VTK+SA E  +    KPRQERT  KQ  V Q FG+RGP Q   NFG
Sbjct: 248 ---DAGQPKQ---VTKQSAKEVRQKILNKPRQKRTPNKQCPYQQCFGKRGPNQ---NFGG 298

Query: 290 QDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGM---------EVTPSGTWLTYHGAIKL 340
            ++++ GT    +P +A+ AP+ SAFF   S++ +          E T     L Y GA++
Sbjct: 299 PEMLKLGTSDPQFPILAELAPTPSAFFPGSKLELVKKNSGGADEPTKDVYELQYSGAVRF 358

Query: 341 DDKDPQFKDNVILLNKHIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAAD 400
           D   P F+  + +LN++++AY        +D+    D   P PQR++ Q  V      +
Sbjct: 359 DSTLPGFETIMKVLNENLNAY---------QDQAGGADVVSPKPQRKRGQRQVAQKKNDE 409

Query: 401 MDDFS 405
           +D+ S
Sbjct: 410 VDNYS 414
```

Alignment von Suchsequenz (query) und einer gefundenen Datenbank-Sequenz (subject)

Identität 36%
Similarität 51%*
Lücken 10%

*bezieht chemische Ähnlichkeit von Aminosäuren ein

Exkurs: Protein-Similarität



Je mehr Linien von einer zur anderen Aminosäure zu überqueren sind, desto chemisch unähnlicher sind die beiden As.

BLAST bewertet die Signifikanz eines Alignments!

Sequences producing significant alignments:	Score (bits)	E Value
gb|AAR23257.1| nucleocapsid protein [SARS coronavirus Sino3...	862	0.0
gb|AAR87518.1| putative nucleocapsid protein N [SARS corona...	862	0.0
gb|AAT76155.1| nucleocapsid protein [SARS coronavirus TJF]	860	0.0
gb|AAP50495.1| nucleocapsid protein [SARS coronavirus FRA] ...	860	0.0
gb|AAS48456.1| nucleocapsid protein [SARS coronavirus BJ01]	859	0.0
gb|AAR12990.1| nucleocapsid protein [SARS coronavirus HB]	859	0.0
gb|AAP30714.1| putative nucleocapsid protein [SARS coronavi...	858	0.0
gb|AAP82974.1| nucleocapsid protein [SARS coronavirus Shanh...	858	0.0
gb|AAS01074.1| nucleocapsid protein [SARS coronavirus CUHK-L2]	469	e-131
gb|AAS48575.1| nucleocapsid protein [SARS coronavirus xw002]	446	e-124
gb|AAS48576.1| nucleocapsid protein [SARS coronavirus cw049]	444	e-123
gb|AAS48577.1| nucleocapsid protein [SARS coronavirus cw037]	437	e-121
pdb|1SSK|A Chain A, Structure Of The N-Terminal Rna-Binding...	292	1e-77
emb|CAA45099.1| nucleocapsid protein [Murine hepatitis virus]	216	8e-55
gb|AAA46439.1| hepatitis virus nucleocapsid (N-MHV1) ORF 1 ...	216	1e-54
gb|AAA46468.1| hepatitis virus nucleocapsid (N-MHVS) ORF 1 ...	215	2e-54

Der E-Wert gibt die Wahrscheinlichkeit an, mit der der für den Match gefundene Score-Wert aus Zufall beim Durchsuchen einer Datenbank der verwendeten Größe auftritt.

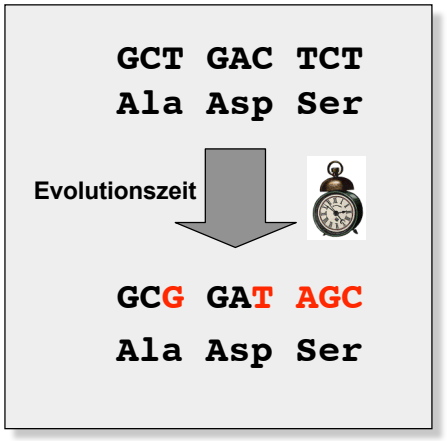
Table 1 | **Predicted SARS-CoV proteins**

ORF	SARS-CoV proteins	Length (amino acids)	Position in the polyprotein	Functional and structural predictions
Replicase region				
ORF1a	Nsp1	180	1M–180G	?
	Nsp2	638	181A–818G	?
	Nsp3 (PLpro)	1922	819A–2740G	Papain-like cysteine protease-cleavage of Nsp1–Nsp4, adenosine diphosphate-ribose 1-phosphatase (ADRP), 2 TMD
	Nsp4	500	2741K–3240Q	3 TMD
	Nsp5 (3CLpro)	306	3241S–3546Q	3C-like cysteine protease-cleavage of Nsp4–Nsp16
	Nsp6	290	3547G–3836Q	5 TMD
	Nsp7	83	3837S–3919Q	?
	Nsp8	198	3920A–4117Q	?
	Nsp9	113	4118N–4230Q	?
	Nsp10	139	4231A–4369Q	Growth-factor-like domain
	Nsp11	13	4370S–4382V	?
ORF1b	Nsp12 (RdRp)	932	4370S–5301Q	RNA-dependent RNA polymerase
	Nsp13 (Helicase)	601	5302A–5902Q	Helicase, zinc-binding domain, NTPase
	Nsp14	527	5903A–6429Q	Exonuclease (ExoN homologue)
	Nsp15	346	6430S–6775Q	EndoRNAse (XendoU homologue)
	Nsp16	298	6776A–7073N	mRNA cap-1 methyltransferase
Structural region				
ORF2	Spike (S) protein	1255		1 TMD, ≥12 *N*-glycosylation sites
ORF3a	?	274		2 TMD, 1 *N*-glycosylation site, 10 *O*-glycosylation sites
ORF3b	?	154		?
ORF4	Envelope (E) protein	76		1 TMD, 2 *N*-glycosylation sites
ORF5	Membrane (M) protein	221		3 TMD, 1 *N*-glycosylation site
ORF6	?	63		1 TMD
ORF7a	?	122		1 TMD
ORF7b	?	44		1 TMD
ORF8a	?	39		Membrane-associated
ORF8b	?	84		1 *N*-glycosylation site
ORF9a	Nucleocapsid (N) protein	422		
ORF9b	?	98		1 *O*-glycosylation site

The analyses are based on the sequence of the SARS-CoV FRA isolate (GenBank accession number AY310120). Transmembrane domains (TMDs) were predicted using the program PSORT (threshold is less than –2); the glycosylation sites were predicted using the NetNGlyc server (see NetNGlyc in the Online links). Information on the functional predictions has been taken from REFS 20,33. Nsp, non-structural protein.

Annotation der SARS-Proteine/ Gene

Warum wohl SARS-BLAST-Suche auf Proteinebene?



Während der Evolution wird die DNA durch ‚stille' Mutationen stark verändert, während die Selektion die Veränderung auf Aminosäureebene weitgehend verhindert:

- Suche auf DNA-Ebene funktioniert nur zwischen nahe verwandten Taxa/ Genen
- Suche auf Aminosäureebene kann noch Ähnlichkeiten von entfernt verwandten Sequenzen detektieren

Gene identifizieren durch Datenbanksuchen

Ein passender ‚Match' mit einem bekannten Gen (auf Nukleotidebene) oder Protein (Aminosäureebene) ist der direkteste Beweis, dass in der Suchsequenz ein Gen liegt.

Vorzugsweise wird zuerst nach Datenbankeinträgen desselben oder näher verwandter Organismen gesucht (auf DNA-Ebene), dann auf Proteinebene nach Ähnlichkeiten in entfernten Organismen (oder entfernt verwandten Proteinen).

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und „Sequenzierung“
- Computer: Erkennen der Virusgene und -proteine (Genvorhersage)

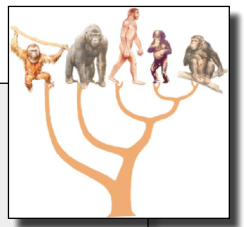
  Ähnlichkeit zu bekannten Genen oder Proteinen? (Datenbanksuchen)

  Verwandtschaft? (Phylogenetische Rekonstruktion)

  Struktur der Proteine? (Struktur-Vorhersage, -Modellierung)

  Wirkstoff-Design
- Labor: Wirkstoff-Test

Molekulare Phylogenie



- **Verwandtschaft von Organismen** (molekulare Systematik, Forensik)
- **Verwandschaft zwischen Genen/Proteinen** (Genomevolution, Gen/Proteinfunktion)
- **Wie haben sich Lebewesen ausgebreitet** (Anthropologie, Ökologie, Epidemiologie)

Molekulare Daten und Phylogenie

- Sequenzen sind direkt vererbt; keine Umwelteinflüsse
- Sequenzdaten sind in großer Menge relativ kostengünstig und schnell zu erhalten (Dank sei der PCR!!!)
- weitgehend frei von Interpretationseinflüssen („reduziert“, „etwas abgeflacht“ etc.)
- Sequenzen evolvieren insgesamt gleichförmiger als morphologische oder physiologische Charaktere
- Sequenzen erlauben Vergleiche über große Distanzen (Tiere, Pilze, Pflanzen)
- „sophisticated“ Modelle zur mathematisch/statistischen Behandlung der Sequenzevolution existieren

Dennoch: auch molekulare Daten können zu falschen Stammbäumen führen

Phylogenie-Rekonstruktion ist kein triviales Problem

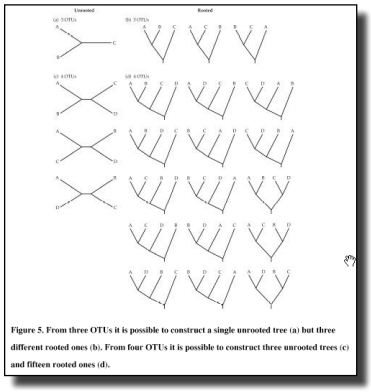


Figure 5. From three OTUs it is possible to construct a single unrooted tree (a) but three different rooted ones (b). From four OTUs it is possible to construct three unrooted trees (c) and fifteen rooted ones (d).

- es ist viel leichter und sicherer, einen unverwurzelten Baum zu erstellen:
  d. h. nur dann „rooten“, wenn die Outgroup klar definiert ist

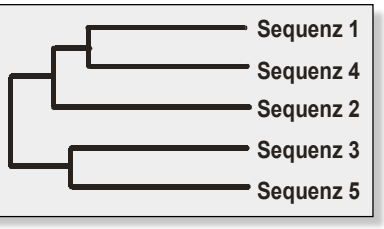
TABLE 5.1 Possible numbers of rooted and unrooted trees up to 10 OTUs

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

From Felsenstein (1978).

Die allgemeine Vorgehensweise…

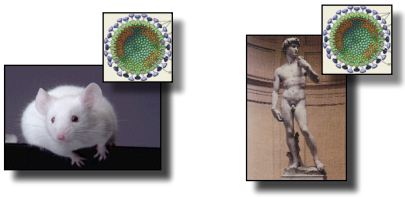
```
Sequenz 1:   KIADKNFTYRHHNQLV
Sequenz 2:   KVAEKNMTFRRFNDII
Sequenz 3:   KIADKDFTYRHW-QLV
Sequenz 4:   KVADKNFSYRHHNNVV
Sequenz 5:   KLADKQFTFRHH-QLV
```



- Multiples Sequenzalignment erstellen (DNA oder Protein)
- Sequenzen vergleichen > Ähnlichkeit bestimmen
- Aus Ähnlichkeitsmaß die Verwandschaft rekonstruieren (Baum)

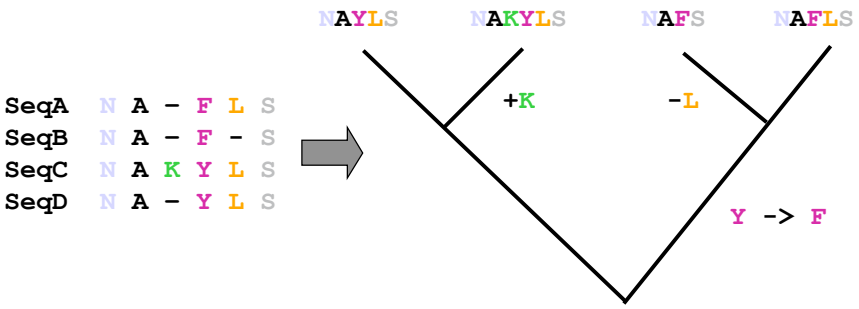
Wann DNA? Wann Protein?

Eng verwandte SARS-Varianten
in der Population



Corona-Virus-Gruppen
aus verschiedenen Spezies

Multiples Alignment ist eine Hypothese zur Sequenzevolution



Warum ist es problematisch, das „richtige" Alignment zu konstruieren?

- 2 x 300 Bp = 10^{88} mögliche Alignments!!!
- Computer-Algorithmen erforderlich, die ohne ausführliche Suche auskommen.

Warum ist problematisch, das „richtige" Alignment zu konstruieren?

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

Keine mismatches

Keine internen Lücken

„Von beidem Etwas"

Was ist richtig?

Wir treffen damit Annahme über den Ablauf der Evolution!!!!

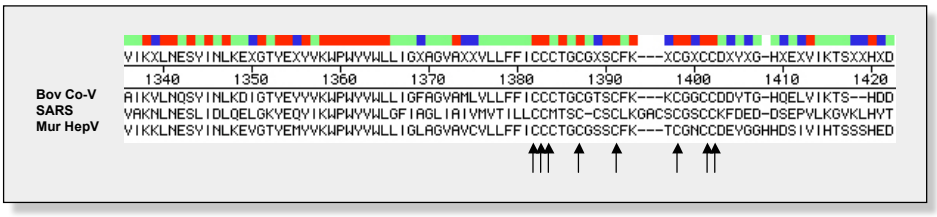
Jede Sequenz lässt sich mit einer jeden anderen Sequenz alignen!

Aber macht das Alignment auch Sinn?

Also: Haben wir die richtigen Annahmen über den Verlauf der Evolution getroffen??

Wir brauchen **„evolutionäre Modelle"**, um ein möglichst richtiges Alignment zu erstellen!

Wie erstellt man ein möglichst „richtiges" Alignment ?



‚Evolutionsmodell': Die Aminosäure Cystein ist für die Proteinstruktur äußerst wichtig!

- Cysteine sind daher **konserviert** während der Evolution!
- Cysteine können daher beim Alignment zweier Proteinsequenzen als **Ankerpunkte** dienen
- ein Alignment mit übereinanderstehenden Cysteinen würde danach mit Pluspunkten **‚belohnt'**

Exkurs:

SARS: konservierte Cysteine im Alignment des spike-Proteins

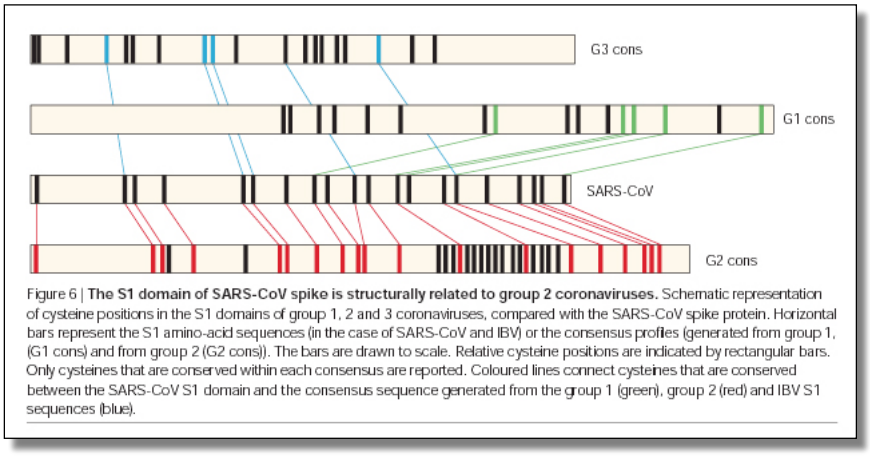


Figure 6 | **The S1 domain of SARS-CoV spike is structurally related to group 2 coronaviruses.** Schematic representation of cysteine positions in the S1 domains of group 1, 2 and 3 coronaviruses, compared with the SARS-CoV spike protein. Horizontal bars represent the S1 amino-acid sequences (in the case of SARS-CoV and IBV) or the consensus profiles (generated from group 1, (G1 cons) and from group 2 (G2 cons)). The bars are drawn to scale. Relative cysteine positions are indicated by rectangular bars. Only cysteines that are conserved within each consensus are reported. Coloured lines connect cysteines that are conserved between the SARS-CoV S1 domain and the consensus sequence generated from the group 1 (green), group 2 (red) and IBV S1 sequences (blue).

Verwandschaft von SARS zu Gruppe 2-Coronaviren?

Ein einfacher Score-Wert zur Bewertung eines Alignments

$$S = Y - \sum W_k \times Z_k$$

S = Similarity-Score (‚Belohnungspunkte')
Y = Anzahl an Matches

Zk = Anzahl der gaps mit Länge k
Wk = gap penalty für gaps der Länge k

Mit Setzen der gap penalty trifft man Annahmen über die relative Häufigkeit von indel-Mutationen während der Evolution!

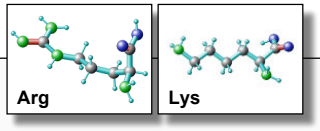
Eine einfache Identitätsmatrix bei Nukleotidsequenzen

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

- alle Richtungen von Nt-Austauschen sind gleich wahrscheinlich

- bei jedem „match" beider Sequenzen gibt es 1 Belohnungspunkt für den Übereinstimmungs-Score

Substitutions-Matrizen für Proteine



- **bei Proteinen gibt es 20 As!**
- chemisch-funktionelle Ähnlichkeit bestimmt Wahrscheinlichkeit eines Austauschs während der Evolution. Daher...
- ...sind die „Kosten“ für bestimmte Austausche (bzw. die Belohnung für gleiche As) unterschiedlich hoch!
- Definition der Kosten erfolgt über Matrizen:

  z. B. **PAM-Matrizen** (Dayhoff 1978)

PAM-Matrizen

		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cysteine	C	12																			
Hydrophilic	S	0	2																		
	T	−2	1	3																	
	P	−3	1	0	6																
	A	−2	1	1	1	2															
	G	−3	1	0	−1	1	5														
Acid-amide	N	−4	1	0	−1	0	0	2													
	D	−5	0	0	−1	0	1	2	4												
	E	−5	0	0	−1	0	0	1	3	4											
	Q	−5	−1	−1	0	0	−1	1	2	2	4										
Basic	H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
	R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
	K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
Hydro-phobic	M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
	I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
	L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
	V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
Aromatic	F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
	Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	2	7	10	
	W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	3	4	−5	−2	−6	0	0	17

Fig. 5.7 The PAM 250 matrix. For each pair of amino acids (see Table 3.1, p. 41, for key to the one-letter codes for amino acids) the matrix gives the ratio of the frequency at which the pair is observed in pairwise comparisons of proteins to that are expected due to chance alone, expressed as a 'log odd'. Amino acids that regularly replace each other have a positive score, amino acids that rarely replace each other have negative scores. Note that replacements more often occur among chemically related amino acids (indicated on the left). From Dayhoff (1978: Fig. 84).

...definieren ‚Belohnungswerte‘ für zwei Aminosäuren, die sich in einem Alignment gegenüberstehen:

- positiver Wert = Aminosäuren, die sich häufig in Alignments gegenüberstehen und somit ‚funktionell konserviert‘ sind

z.B. W-W 17
C -C 12

aber W-V - 6

Wir haben also Kriterien (Substitutionsmatrizen, gap penalties), um Alignments zu bewerten.

Aber wie werden Alignments erstellt?

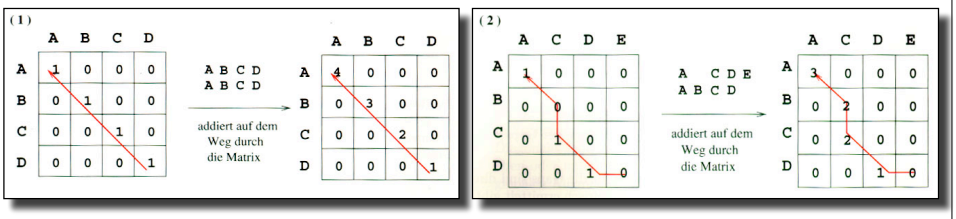
Needleman-Wunsch (N-W) 1970

- Bei Erstellung des Alignments werden zunächst kleine Problem-Schritte gelöst. Dann wird aus den Teillösungen das Gesamtalignment rekonstruiert

- Algorithmus: „dynamic programming“

Needleman-Wunsch

- es wird zunächst eine zweidimensionale Matrix mit den beiden zu vergleichenden Sequenzen erstellt
- in die Zellen der Matrix wird der Alignment-Score für die jeweils verglichenen Sequenzpositionen hineingeschrieben. Die Berechnung des Score-Werts erfolgt natürlich anhand einer Substitutionsmatrize.
- das Alignment ergibt sich als Pfad durch die Matrix. Der Pfad mit der höchsten Endsumme gewinnt...



Vom Alignment zu einem einfachen Baum-Rekonstruktionsverfahren…

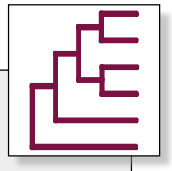
Aus dem Alignment ergibt sich zunächst, wie ähnlich oder unähnlich die Sequenzen zueinander sind.

Meist wird eine Distanzmatrix erstellt:

	A	B	C	D
OTU* A	0	6	10	18
OTU B		0	12	20
OTU C			0	19
OTU D				0

*** OTU = operational taxonomic unit: z. B. Spezies, Gen, Protein**

Vom Alignment zu einem einfachen Baum-Rekonstruktionsverfahren…



Sokal and Michener 1967!

UPGMA

=

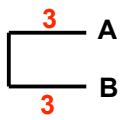
Unweighted Pair-Group Method using Arithmetric Means

…eine Methode, die auf der Berechnung von Unähnlichkeiten (Distanzen) der alignierten Sequenzen beruht („Distanz-Methode")

UPGMA

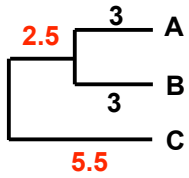
1. **Ausgangs-Distanz-Matrix**

	A	B	C	D
OTU A	0	6	10	18
OTU B		0	12	20
OTU C			0	19
OTU D				0

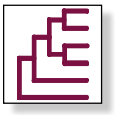


2. **Neu berechnete Distanz-Matrix**

	A/B	C	D
OTU A/B	0	11	19
OTU C		0	19
OTU D			0

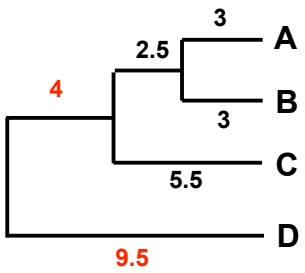


UPGMA

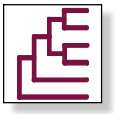


3. **Neu berechnete Distanz-Matrix**

		A/B/C	D
Sequenz	A/B/C	0	19
Sequenz	D		0



UPGMA

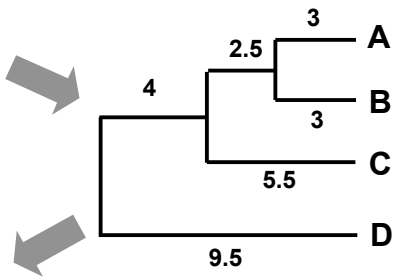


Ausgangsmatrix

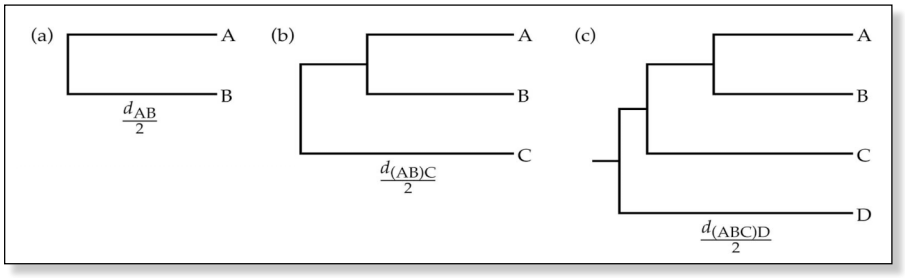
	A	B	C	D
OTU A	0	6	10	18
OTU B		0	12	20
OTU C			0	19
OTU D				0

rekonstruierte Matrix

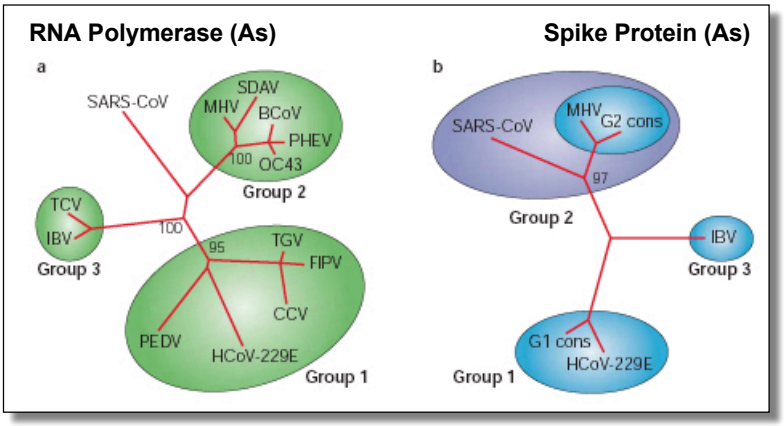
	A	B	C	D
OTU A	0	6	11	19
OTU B		0	11	19
OTU C			0	19
OTU D				0



UPGMA



SARS-Phylogenie



Unterschiedliche Datensätze und Rekonstruktionsmethoden können leicht unterschiedliche Baum-Topologien ergeben!!

Aber: SARS Co-V ist alter Verwandter der Gruppe 2 Coronaviren

SARS-Phylogenie

Table 2 | **Protein homologies between SARS-CoV and other coronaviruses**

	Group 1			Group 2		Group 3
SARS-CoV proteins	HCoV-229E	TGV	PEDV	MHV	BCoV	IBV
Nsp1	<20.0	<20.0	<20.0	27.0	<20.0	<20.0
Nsp2	<20.0	**23.0**	23.0	<20.0	20.0	<20.0
Nsp3 (PLpro)	25.1	26.6*	*24.1**	26.2	**26.8**	23.3
Nsp4	26.8	26.0	28.4	**43.1**	42.4	28.5
Nsp5 (3CLpro)	*40.4*	43.8	44.6	**50.0**	48.4	41.0
Nsp6	30.0	*27.0*	29.4	34.2	**35.5**	28.5
Nsp7	38.6	42.2	39.8	**47.5**	46.1	*37.3*
Nsp8	**48.2**	42.9	43.9	46.8	47.3	*38.7*
Nsp9	45.1	38.9	45.1	45.1	**46.9**	39.8
Nsp10	53.8	54.5	56.1	56.2	55.4	**58.3**
Nsp11	–	–	–	–	–	–
Nsp12 (RdRp)	59.8	59.6	60.0	**67.3**	66.9	62.4
Nsp13 (Helicase)	60.7	62.0	62.3	67.2	**68.6**	*58.9*
Nsp14	52.3	53.7	52.3	**57.6**	**57.6**	52.0
Nsp15	43.1	43.0	45.4	**45.9**	45.0	*40.2*
Nsp16 (Methyltransferase)	56.4	54.4	55.3	63.0	**65.0**	*53.4*
Spike (S) protein	*28.8*	31.0*	30.3	**31.1**	31.0	32.7*
Envelope (E) protein	33.0*	**27.9**	*20.0*	23.0	26.5	23.2
Membrane (M) glycoprotein	*30.6*	32.5	34.8	40.8	**41.9**	32.5
Nucleocapsid protein (N)	*26.9*	30.1	29.5	37.3	**37.4**	31.5

Numbers indicate the amino-acid identity between the predicted SARS-CoV proteins and the corresponding gene products of other coronaviruses (as a percentage). More conserved pairs are in bold; more variable pairs are in italic. The program FASTA was used for sequence comparison. Asterisks indicate that the alignment was obtained using only a fragment of the whole protein. Nsp, non-structural protein.

SARS-Phylogenie

DNA (Komplettgenom)

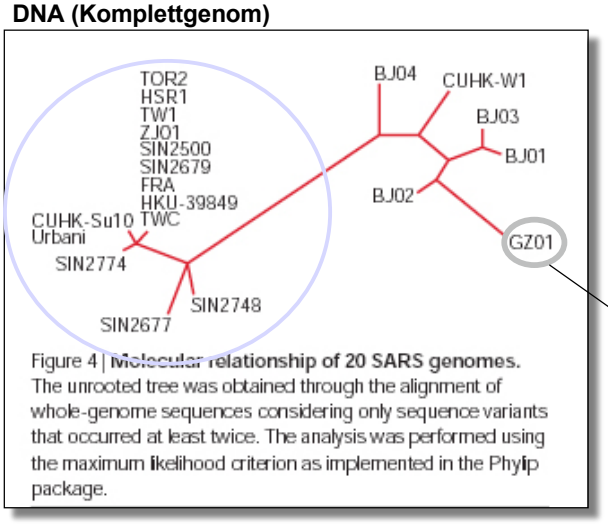


Figure 4 | Molecular relationship of 20 SARS genomes. The unrooted tree was obtained through the alignment of whole-genome sequences considering only sequence variants that occurred at least twice. The analysis was performed using the maximum likelihood criterion as implemented in the Phylip package.

- **Varianten sind >99% identisch. Dennoch ist eine geographische Zuordnung möglich.**

Sequenz zeigt Besonderheit:

Sein Spike-Gen hat 29 Bp zusätzlich, die sonst nur in tierischen SARS-Verwandten gefunden worden sind!

SARS-Phylogenie

Larvenroller - palm civet
(*Paguma larvata*)

Marderhund - Raccoon dog
(*Nyctereutes procyonoides*)

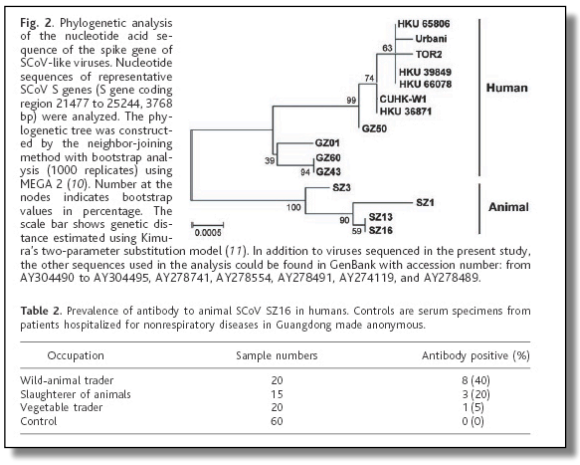
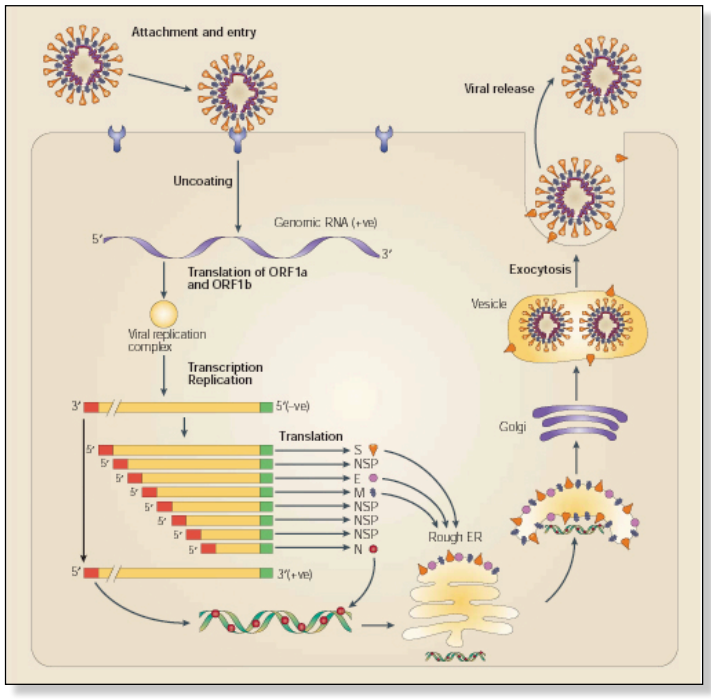


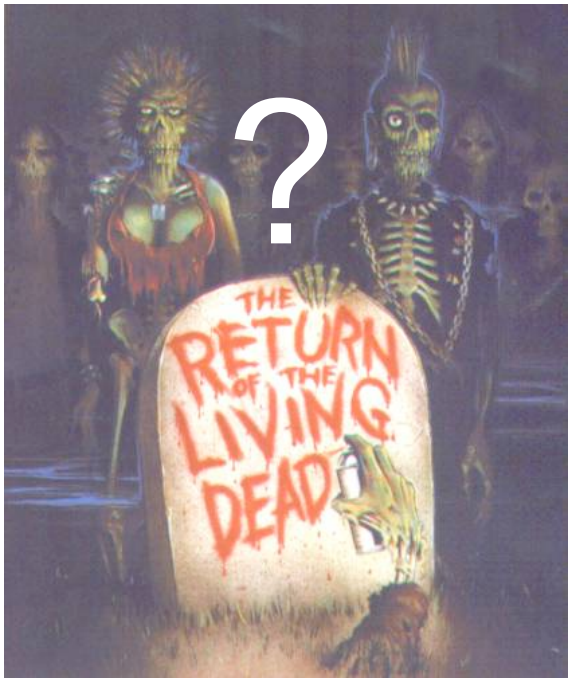
Fig. 2. Phylogenetic analysis of the nucleotide acid sequence of the spike gene of SCoV-like viruses. Nucleotide sequences of representative SCoV S genes (S gene coding region 21477 to 25244, 3768 bp) were analyzed. The phylogenetic tree was constructed by the neighbor-joining method with bootstrap analysis (1000 replicates) using MEGA 2 (*10*). Number at the nodes indicates bootstrap values in percentage. The scale bar shows genetic distance estimated using Kimura's two-parameter substitution model (*11*). In addition to viruses sequenced in the present study, the other sequences used in the analysis could be found in GenBank with accession number: from AY304490 to AY304495, AY278741, AY278554, AY278491, AY274119, and AY278489.

Table 2. Prevalence of antibody to animal SCoV SZ16 in humans. Controls are serum specimens from patients hospitalized for nonrespiratory diseases in Guangdong made anonymous.

Occupation	Sample numbers	Antibody positive (%)
Wild-animal trader	20	8 (40)
Slaughterer of animals	15	3 (20)
Vegetable trader	20	1 (5)
Control	60	0 (0)

SARS-Lebens-Zyklus




?
THE
RETURN
OF THE
LIVING
DEAD