JOHANNES GUTENBERG UNIVERSITÄT MAINZ

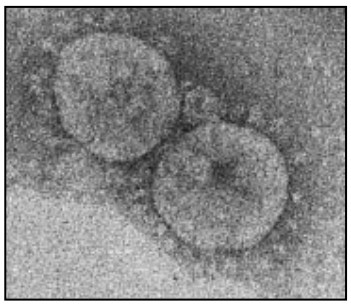
BSc Modul 4-2, Tage 1-3 SoSe 2017

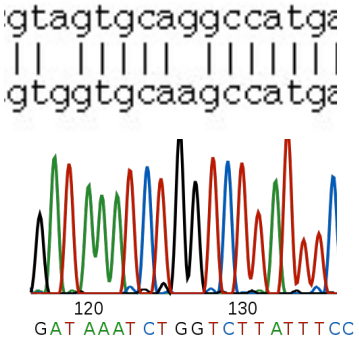
Von der DNA zum Medikament:

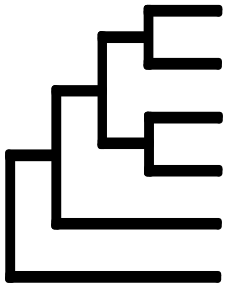
Bioinformatische Methoden in der Molekularbiologie

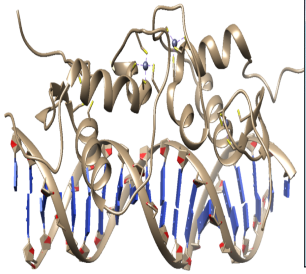
Thomas Hankeln
AG Molekulargenetik & Genomanalyse

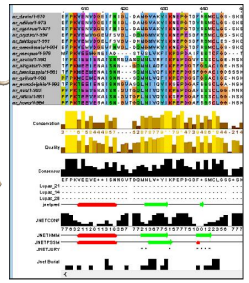
Miguel Andrade & Elmar Jaenicke
AG Bioinformatik

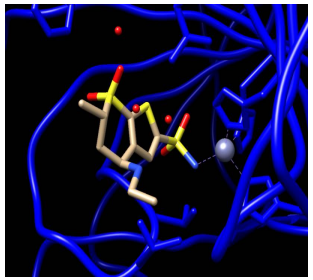




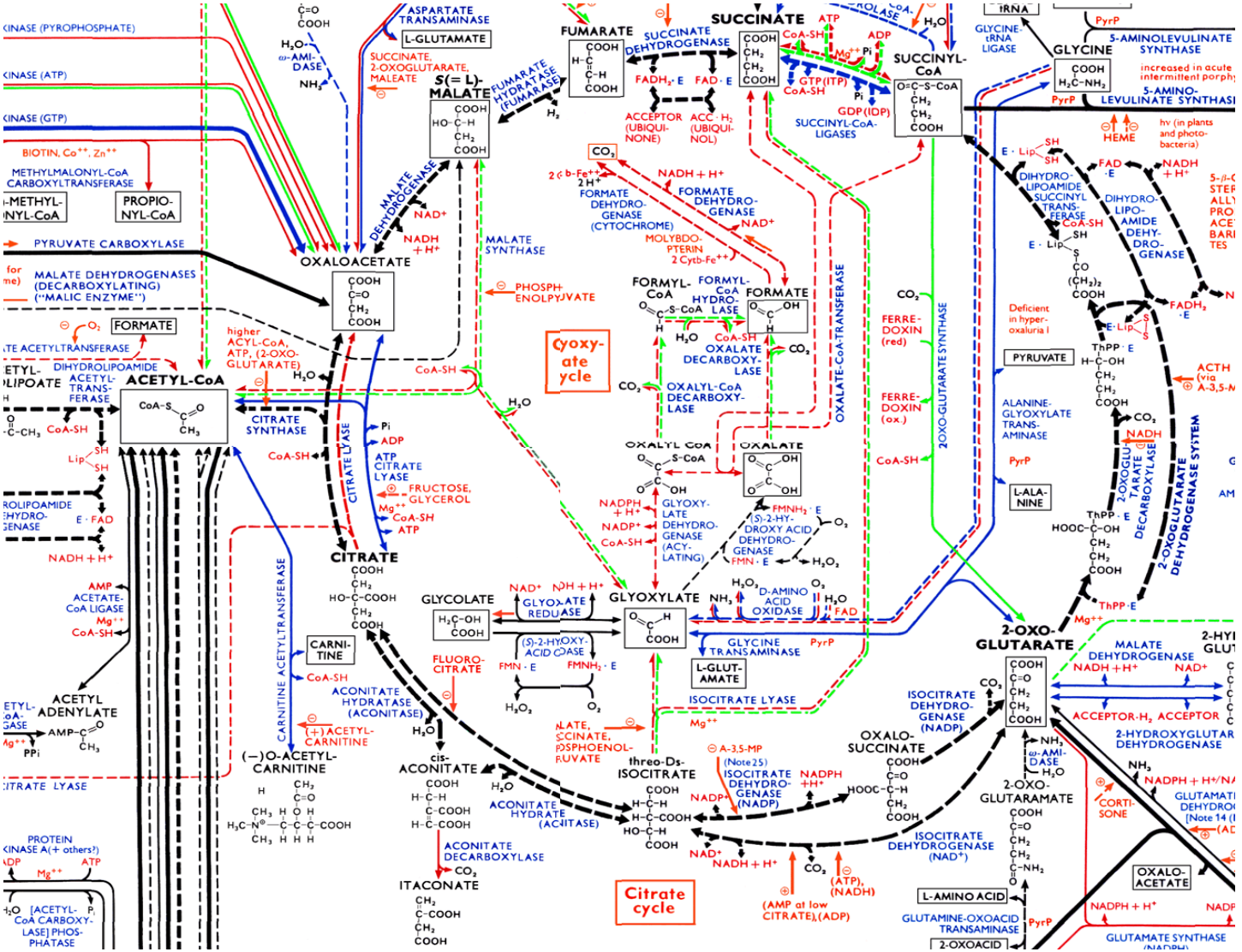




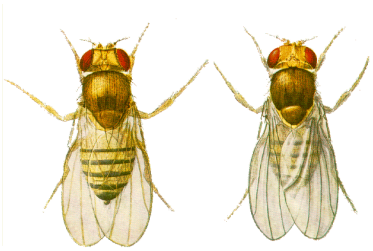
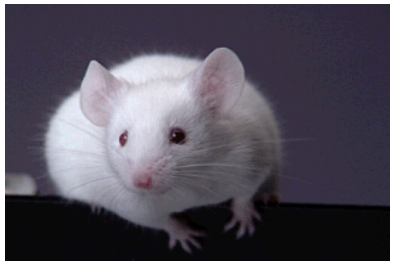
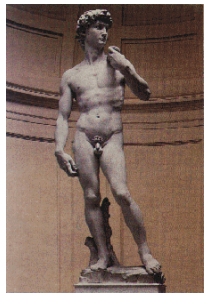




Warum Informatik in der Biologie?



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

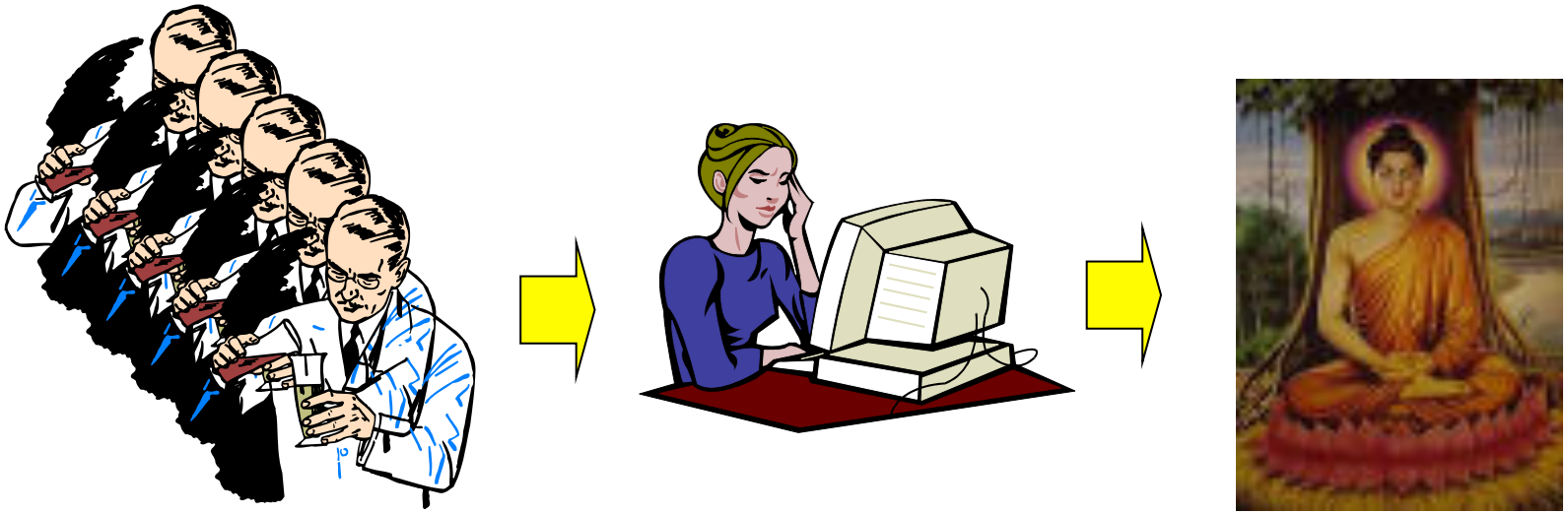
BioInformatik

Genomforschung	Algorithmen*
Molekularbiologie	Datenbanken
Biochemie	Visualisierung
Physiologie	Simulation

> Verständnis biologischer Zusammenhänge
> Kenntnis informatischer Methoden

*eine Menge eindeutiger Anweisungen zur Lösung eines Problems

Die Vision...



www.systemsbiology.org

Literaturauswahl für Tage 1-3

Zvelebil M, Baum JO: *Understanding Bioinformatics.* Garland Science 2008

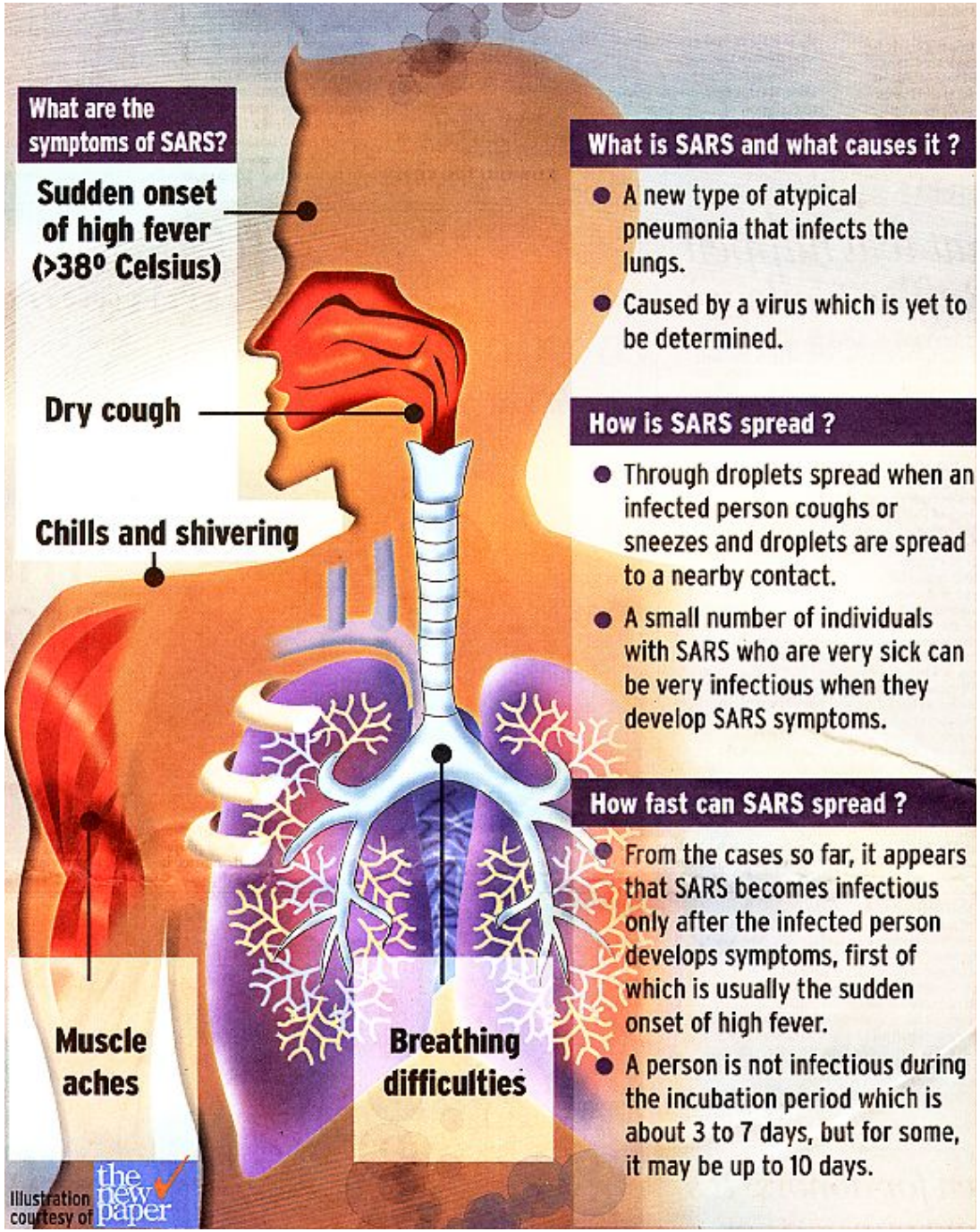
Mount, D.M. *Bioinformatics*. Cold Spring Harbor Press 2004 (für den -zukünftigen- Profi, z. T. kompliziert)

Hansen, A. *Bioinformatik*. *Ein Leitfaden für Natur-wissenschaftler*. Birkhäuser 2004

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

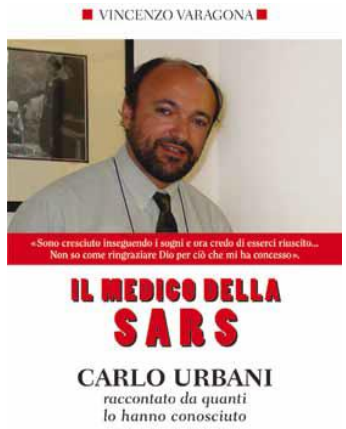

THE TIMES
NEW MYSTERY ILLNESSES
SIGNS OF THINGS TO COME?
CDC puzzled by "SARS" outbreak
The CDC and world health authorities are
puzzled by the new severe acute respirat-
March 2003
Severe acute respiratory syndrome

Symptome

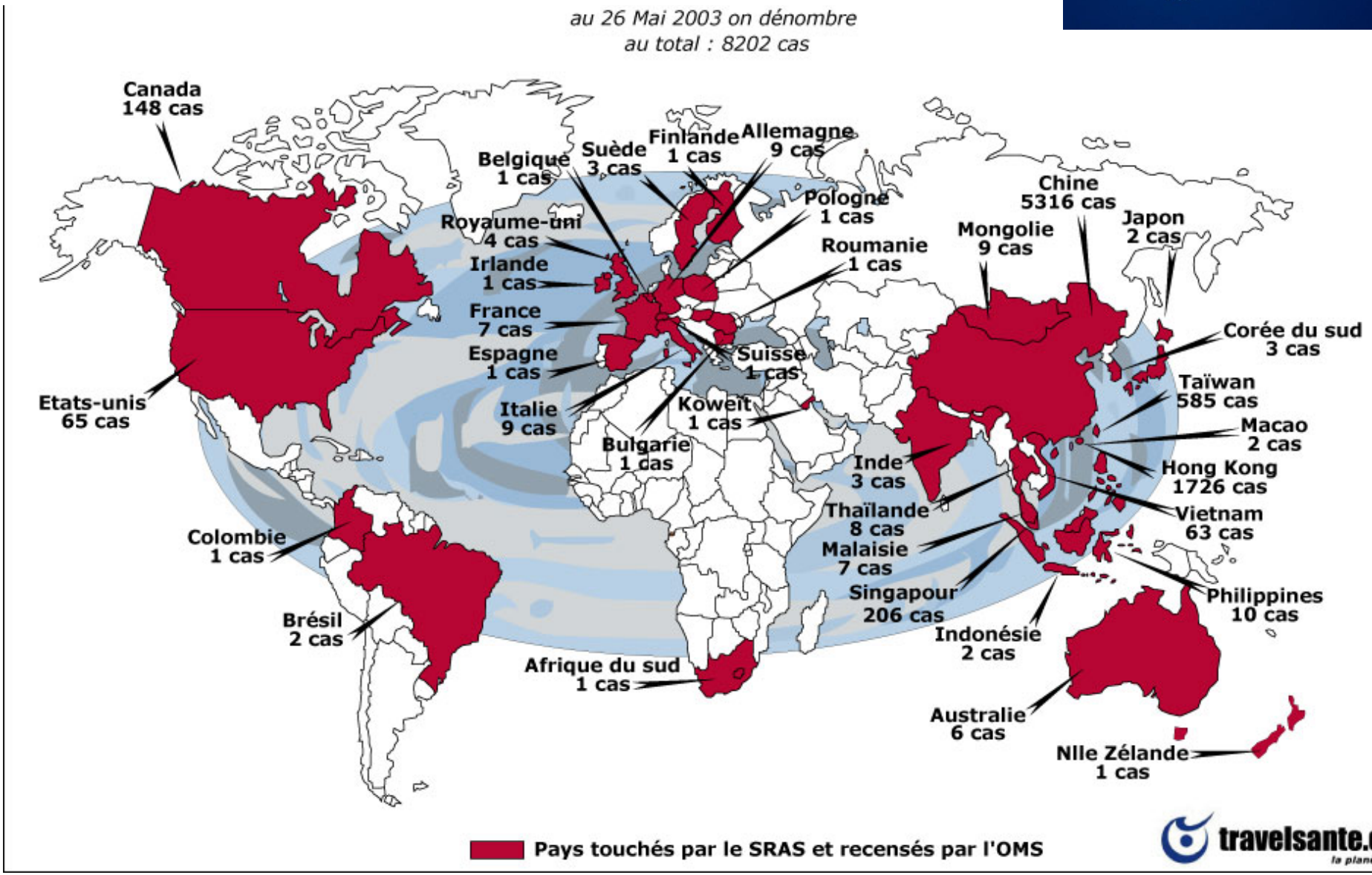

What are the symptoms of SARS?
Sudden onset of high fever (>38° Celsius)
Dry cough
Chills and shivering
Muscle aches
Breathing difficulties
Illustration courtesy of the new paper
What is SARS and what causes it ?
A new type of atypical pneumonia that infects the lungs.
Caused by a virus which is yet to be determined.
How is SARS spread ?
Through droplets spread when an infected person coughs or sneezes and droplets are spread to a nearby contact.
A small number of individuals with SARS who are very sick can be very infectious when they develop SARS symptoms.
How fast can SARS spread ?
From the cases so far, it appears that SARS becomes infectious only after the infected person develops symptoms, first of which is usually the sudden onset of high fever.
A person is not infectious during the incubation period which is about 3 to 7 days, but for some, it may be up to 10 days.

Chronologie der Ereignisse

Nov 2002	first cases of mysterious respiratory illness in Guangdong, China
Jan 31 2003	first major outbreak in Guanzhou hospital: super-spreader infects 130 persons
21 Feb 2003	Doctor A from hospital G visits Hongkong hotel M > dies two days later > other hotel guests infected > travel to Vietnam, Singapore, Canada, US
Early Mar 2003	more newspaper reports
Mar 9	**Carlo Urbani (WHO) called to Hanoi realizes a new disease, prevents local spreading and convinces WHO to get active**
	„I have a hospital full of crying nurses. People are running and screaming and totally scared. We don't know what it is, but it's not flu."
Mar 15	WHO official alert
Mar 15	2 infected pasengers at Frankfurt airport > Sputum PCR-negative for tropical viruses
Mar 18	EM pictures suggest paramyxovirus PCR negative (too much human DNA contamination)
Mar 20	Frankfurt researchers cultivate virus > to Hamburg lab
Mar 22-25	PCRs with degenerate primers give ca 20 sequences, 2 of them are coronavirus sequences
Mar 25	US group reports typical coronavirus image for SARS
Mar 26	C. Drosten (Hamburg) publishes PCR assay for SARS on institute website and distributes positive control material for PCR to 150 labs worldwide
Mar 29	**Carlo Urbani dies of lung failure**
Apr 2003	4300 SARS cases, 250 deaths in 25 countries



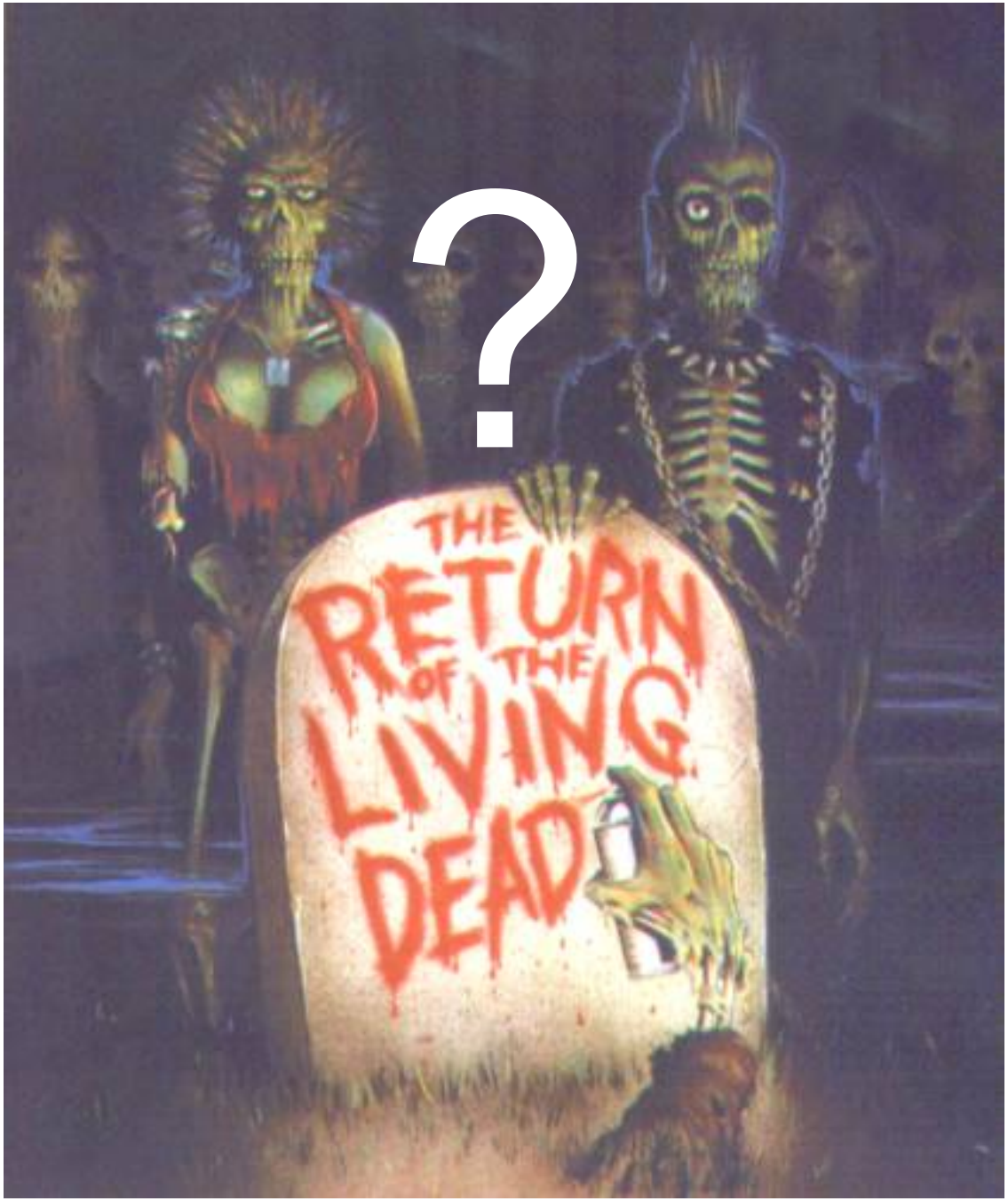
Eine Pandemie



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…


?
THE
RETURN
OF THE
LIVING
DEAD

MERS

Coronavirus

Zehn Jahre nach Sars: Ungewöhnliche Lungenentzündungen alarmieren Forscher

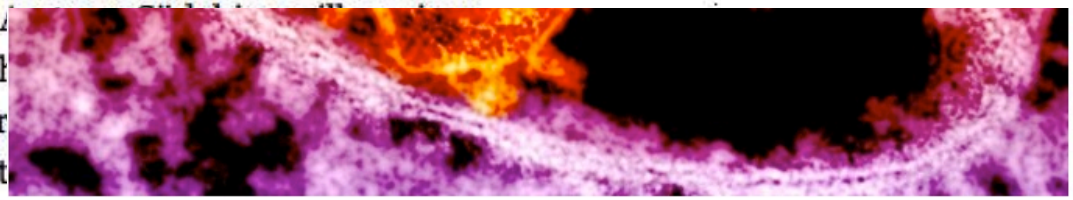
von Jana Schlütter

Angst vor Ansteckung. Als im März 2003 immer mehr Menschen in Hong Kong an Sars erkrankten, trauten sich viele nur noch mit Atemschutzmasken vor die Tür. Foto: AFP - FOTO: AFP

Während die Erforschung von Sars ein Musterbeispiel internationaler Zusammenarbeit war, beklagen Infektionsbiologen beim neuen Virus eine noch stockende Informationspolitik.

Empfehlen · Twittern · +1

Der
Hoch
Metr
fühlt
trotz
nächsten Tag wird er mit Atemwegsproblemen auf d
Krankenhauses behandelt. Er warnt seine Ärzte, das
Ansteckendes hat – er hatte zuvor in der Provinz Gu
Lungenentzündungen behandelt. Unterdessen reist
aus Toronto, nach Kanada zurück. Sie infiziert etlich
ebenfalls kurz darauf an einer Krankheit, die wir he



Mers zählt, wie auch Sars, zu den Coronaviren. (Foto: Reuters)

Samstag, 03. Mai 2014

Virus kam aus Saudi-Arabien

USA melden ersten Mers-Fall

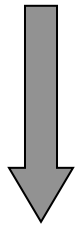
Die WHO warnt vor dem tödlichen Mers-Virus - es sei eine "Gefahr für die ganze Welt". Bislang konzentriert sich die Ausbreitung vor allem auf die arabische Halbinsel. Nun wird das Virus auch in den USA nachgewiesen.

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und Sequenzierung

- Computer: Erkennen der Virusgene (*de novo* Genvorhersage)

Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)



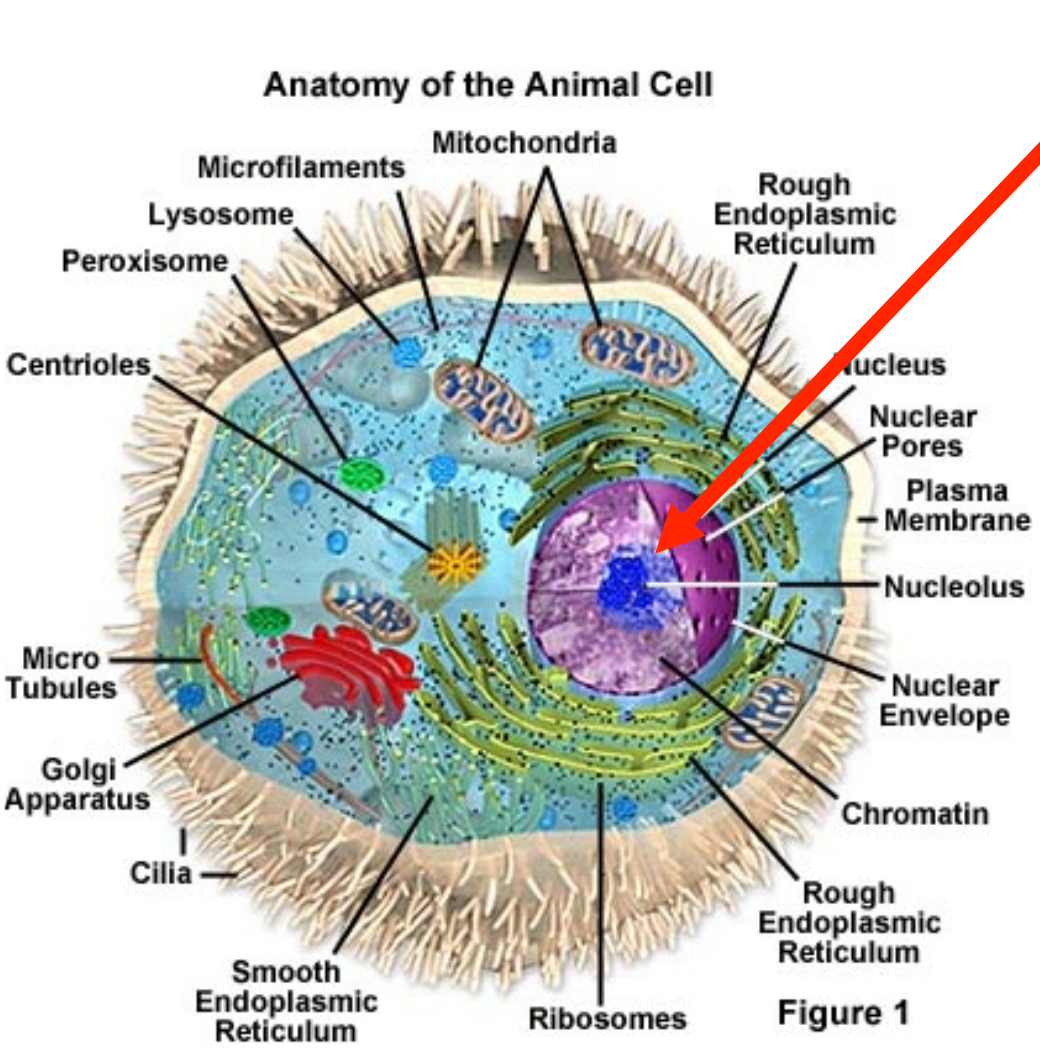
Verwandtschaft? Ausbreitung? Herkunft?
(Phylogenetische Rekonstruktion)

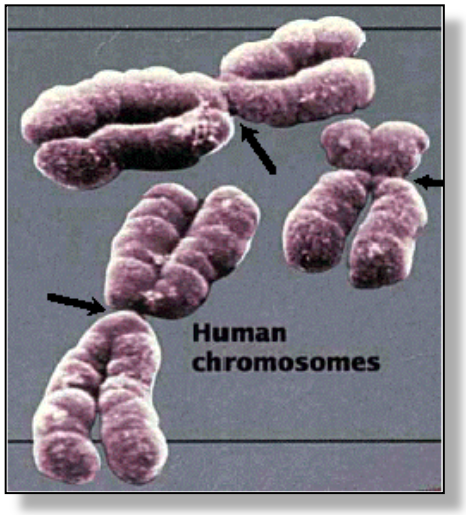
Struktur der Proteine? (Struktur-Vorhersage,
-Modellierung)

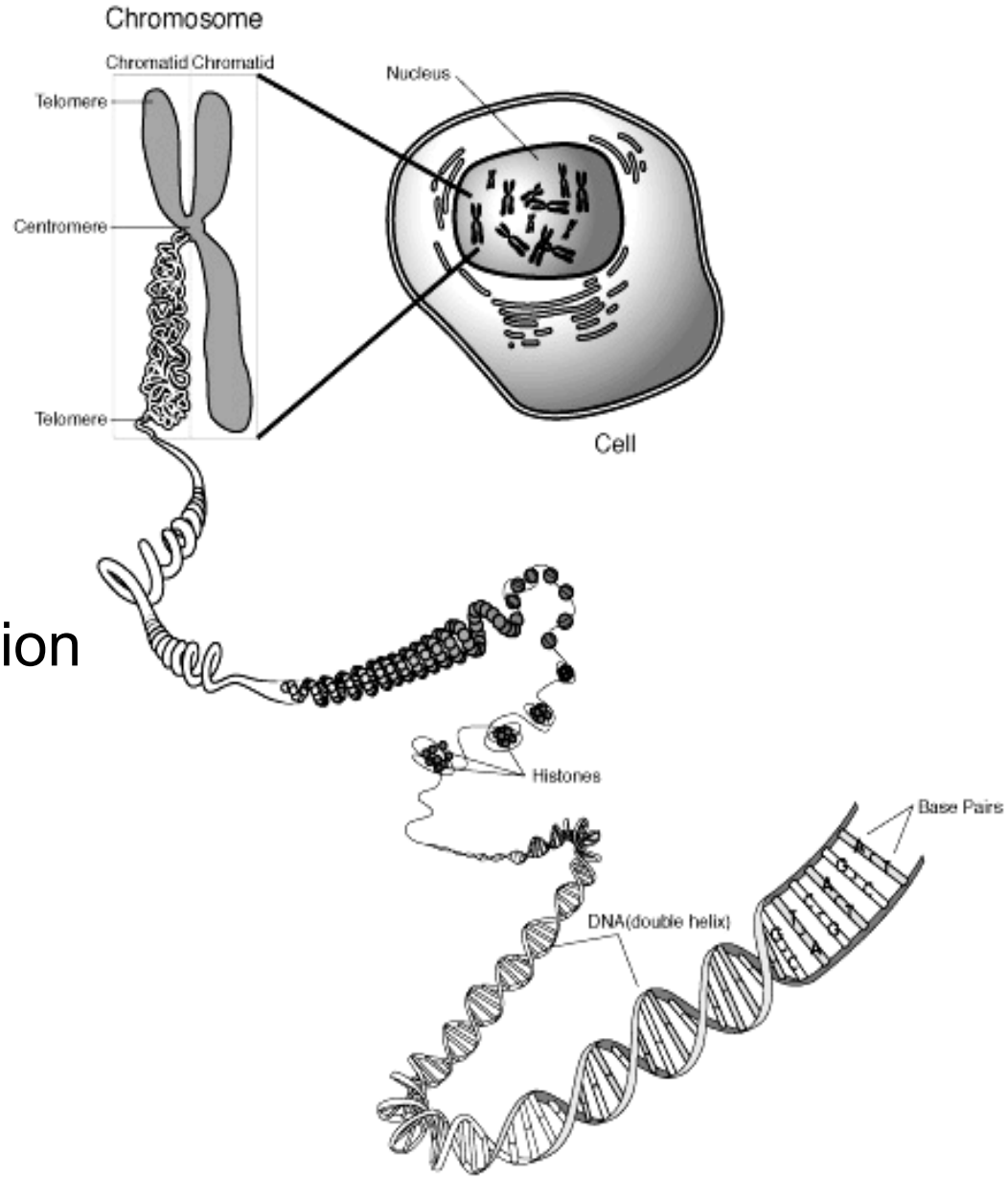
Wirkstoff-Design

- Labor: Wirkstoff-Test

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







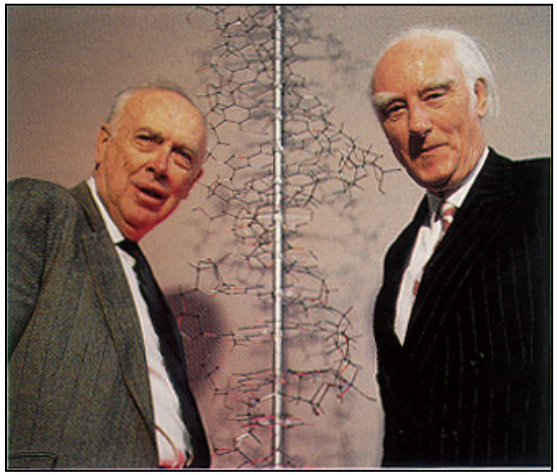
Das **Genom** ist die Gesamtheit der Erbinformation einer Zelle.

Gene sind funktionelle Abschnitte auf der **DNA**.

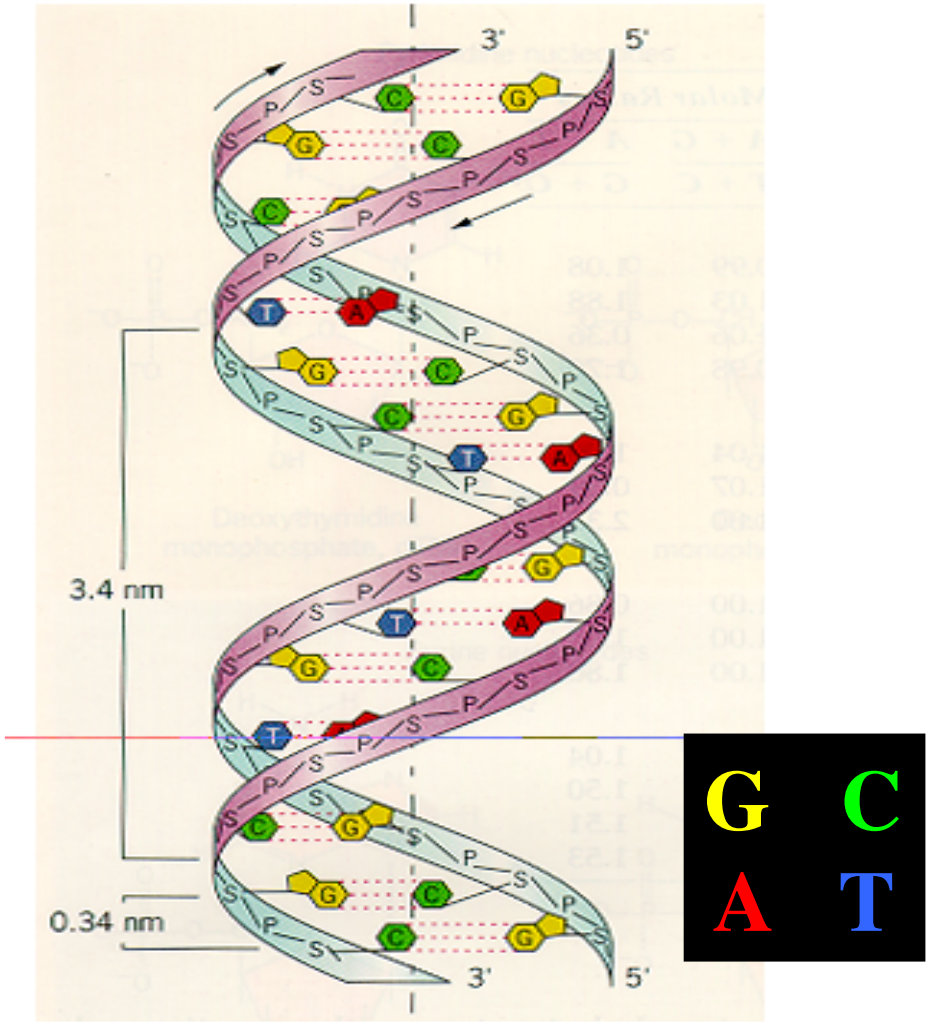
Die DNA ist auf **Chromosomen** aufgeteilt.

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“)

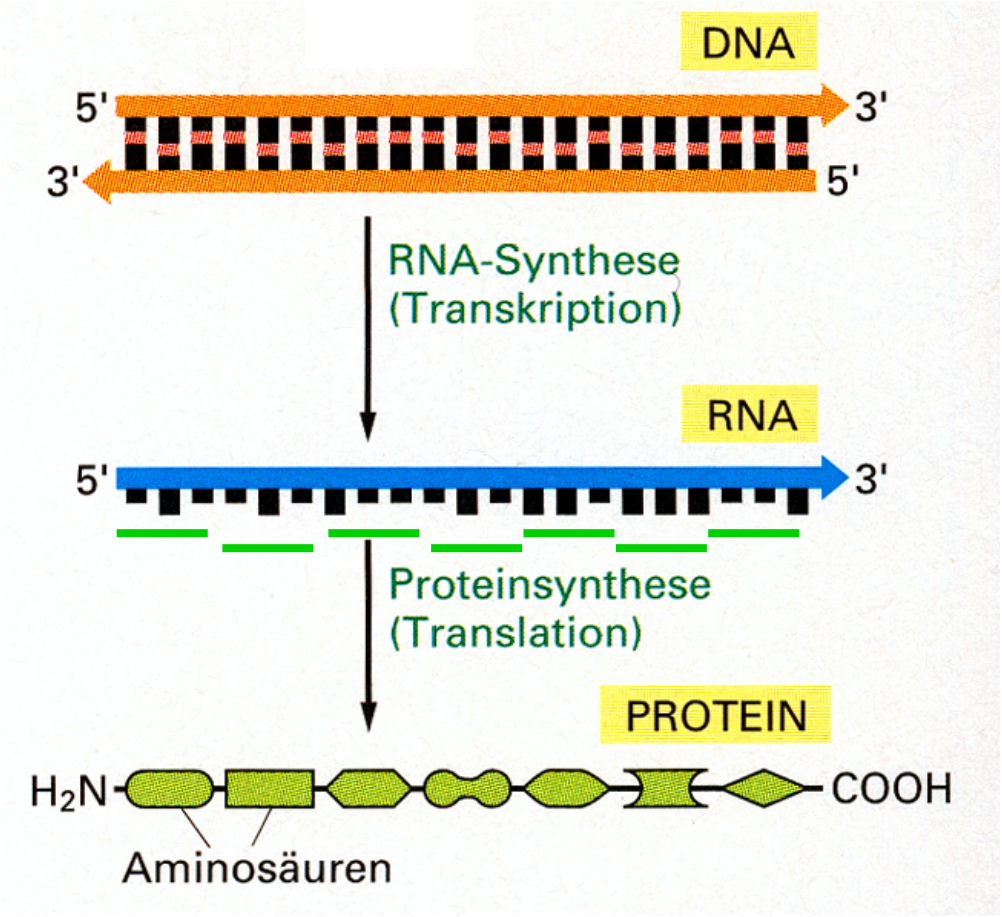
Beispiel:

5‘-GAGGGCTACTGCA-3‘

oder

5‘-TGCAGTAGCCCTC-3‘

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



Informationsspeicher

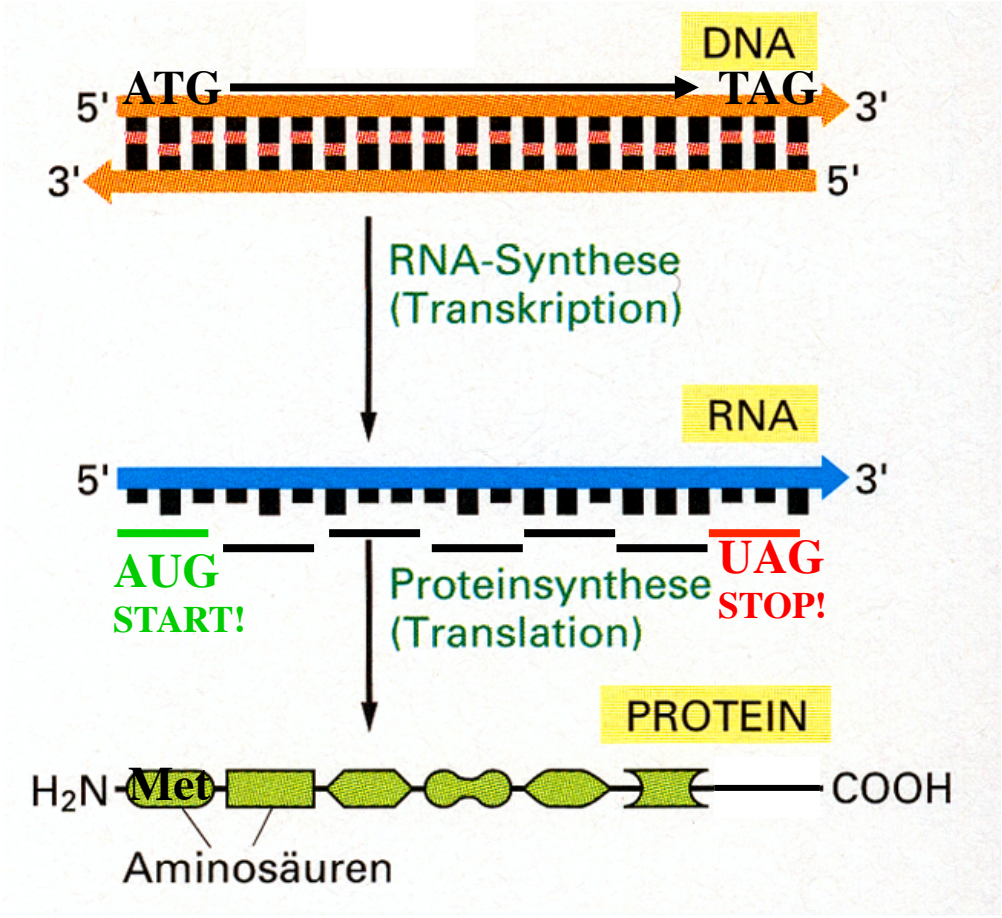
Informationsabschrift

Produkt

Q:

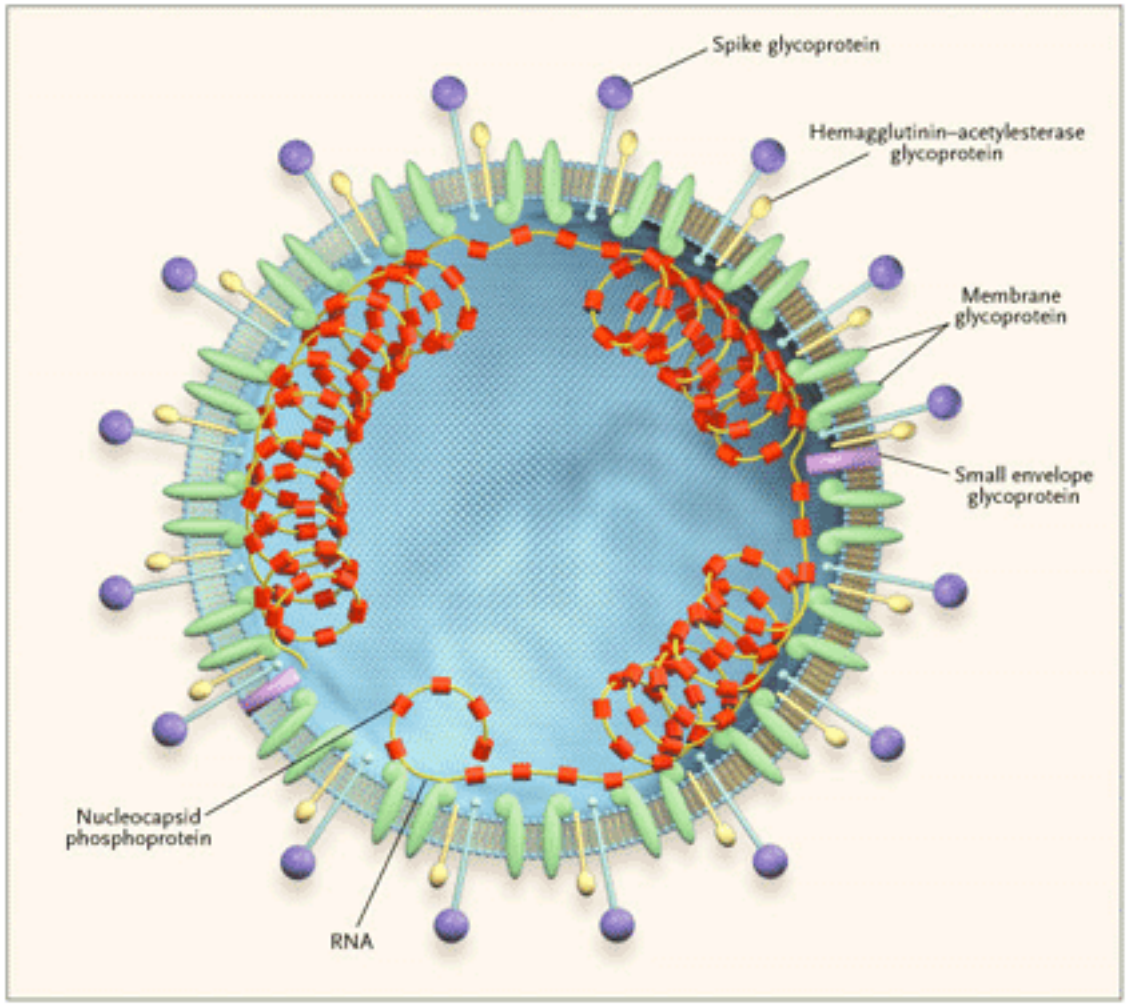
Wie erkenne ich, dass ein DNA-Abschnitt ein Protein-kodierendes Gen enthält?

Wie erkenne ich ein proteinkodierendes Gen?



ORF
= offener Leserahmen

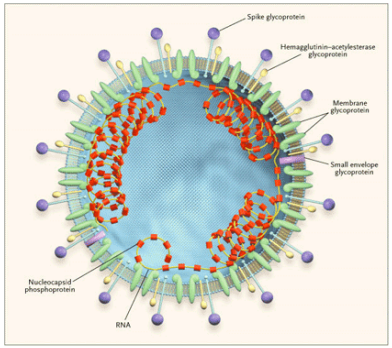
Viren haben eigene Erbinformation (manchmal aus RNA)

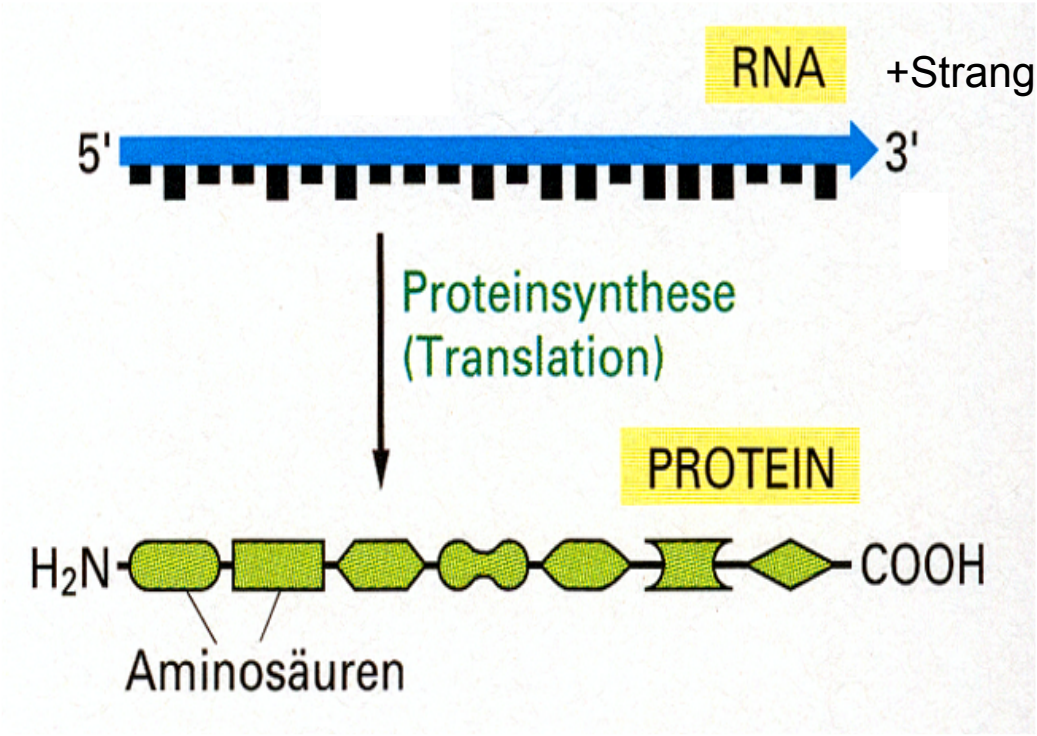


SARS Coronavirus

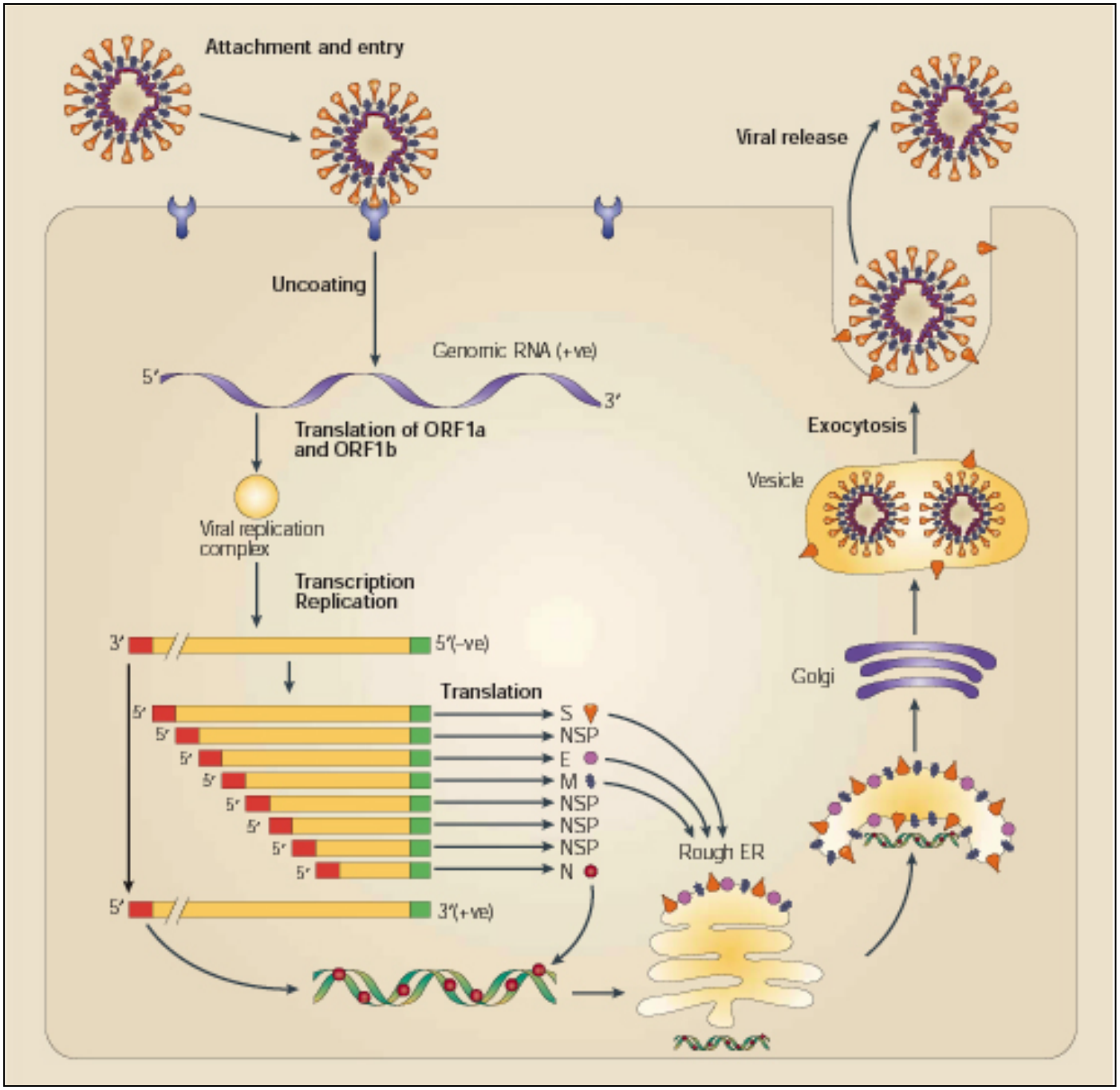
- RNA-Genom, umhüllt von Nukleo-Capsidprotein
- Lipidhülle mit 4 Proteinen

Viren haben eigene Erbinformation (manchmal aus RNA)

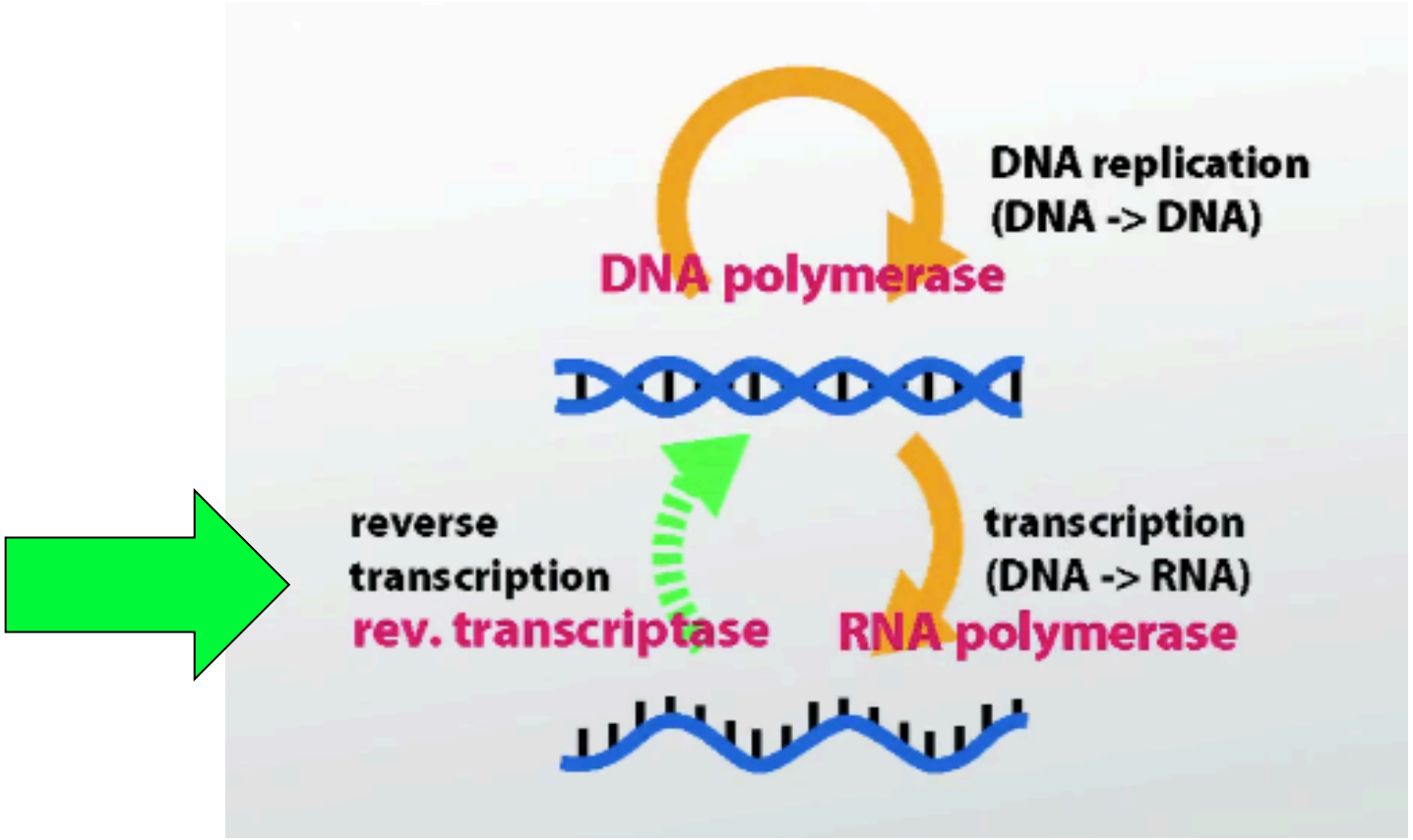




SARS-Lebens-Zyklus



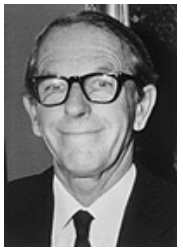
Für die Sequenzaufklärung wird RNA in DNA umgeschrieben!



http://study.com/academy/lesson/reverse-transcriptase-definition-function-structure.html

Methoden der DNA-Sequenzierung

1977



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

synonym: > Kettenabbruch-Sequenzierung
> Didesoxy-Sequenzierung

2000: Human Genome Project

ws
Print"

The New York Times

National Edition

Arizona and New Mexico: M
cloudy in New Mexico, thunderst
in the mountains. Partly sunny
where. Highs 80 mountains, ove
deserts. Weather map is on Page

. No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 Printed in Arizona ONE DOLL

tic Code of Human Life Is Cracked by Scientist

become part that Congress was entitled to the last
word because Miranda's presump-
tion that a confession was not volun-

The Book of Life

The 3 billion base pairs ... of the intertwining double helix of DNA ... that make up the set of chromosomes in our cells, have been sequenced.

BASE PAIRS
Rungs between the strands of the double helix

BASES
A adenine
C cytosine
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCCE

2 Rivals' Announcem
Marks New Medica
Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 —
achievement that represents a
nacle of human self-knowledge
rival groups of scientists said t
that they had deciphered the he
tary script, the set of instruc
that defines the human organis

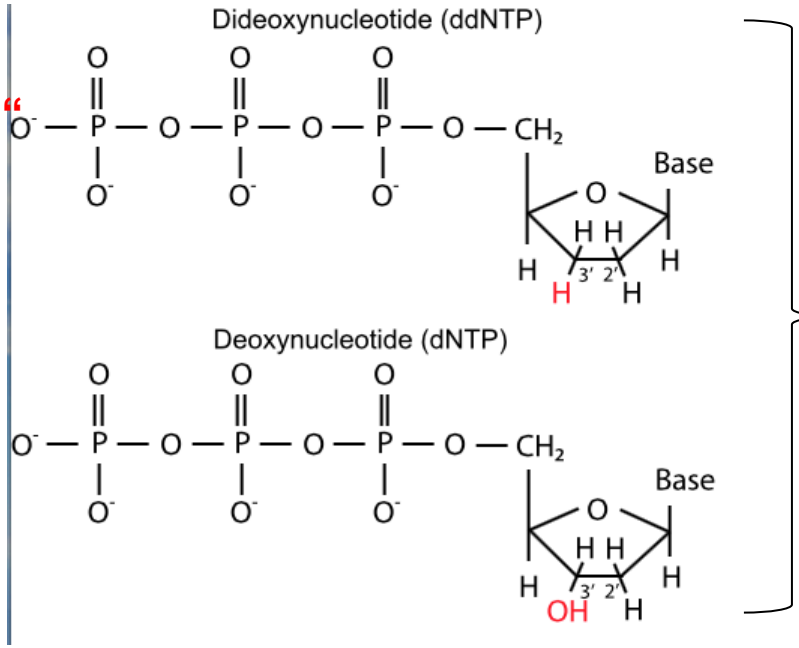
Das Sanger-Verfahren

- Replikation in vitro! Zutaten?

  Primer, Polymerase, dNTPs

- ...der nobelpreiswürdige Trick:

„Terminatoren“



...die Mischung macht‘s!!

Das Sanger-Verfahren

Sequenz bekannt Sequenz unbekannt

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

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

5‘-CTAGGACTGTAC **T**^{Stop}

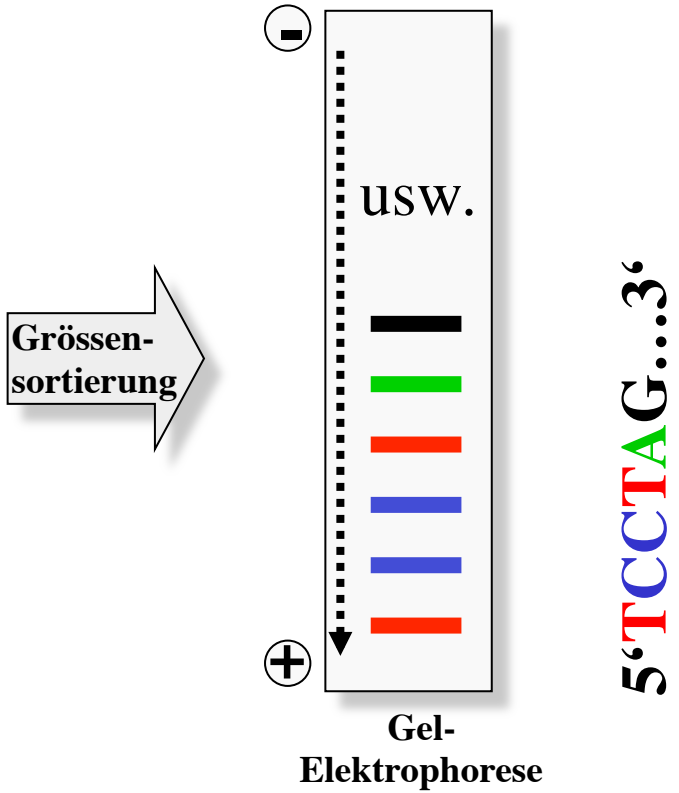
5‘-CTAGGACTGTAC **TC**^{Stop}

5‘-CTAGGACTGTAC **TCC**^{Stop}

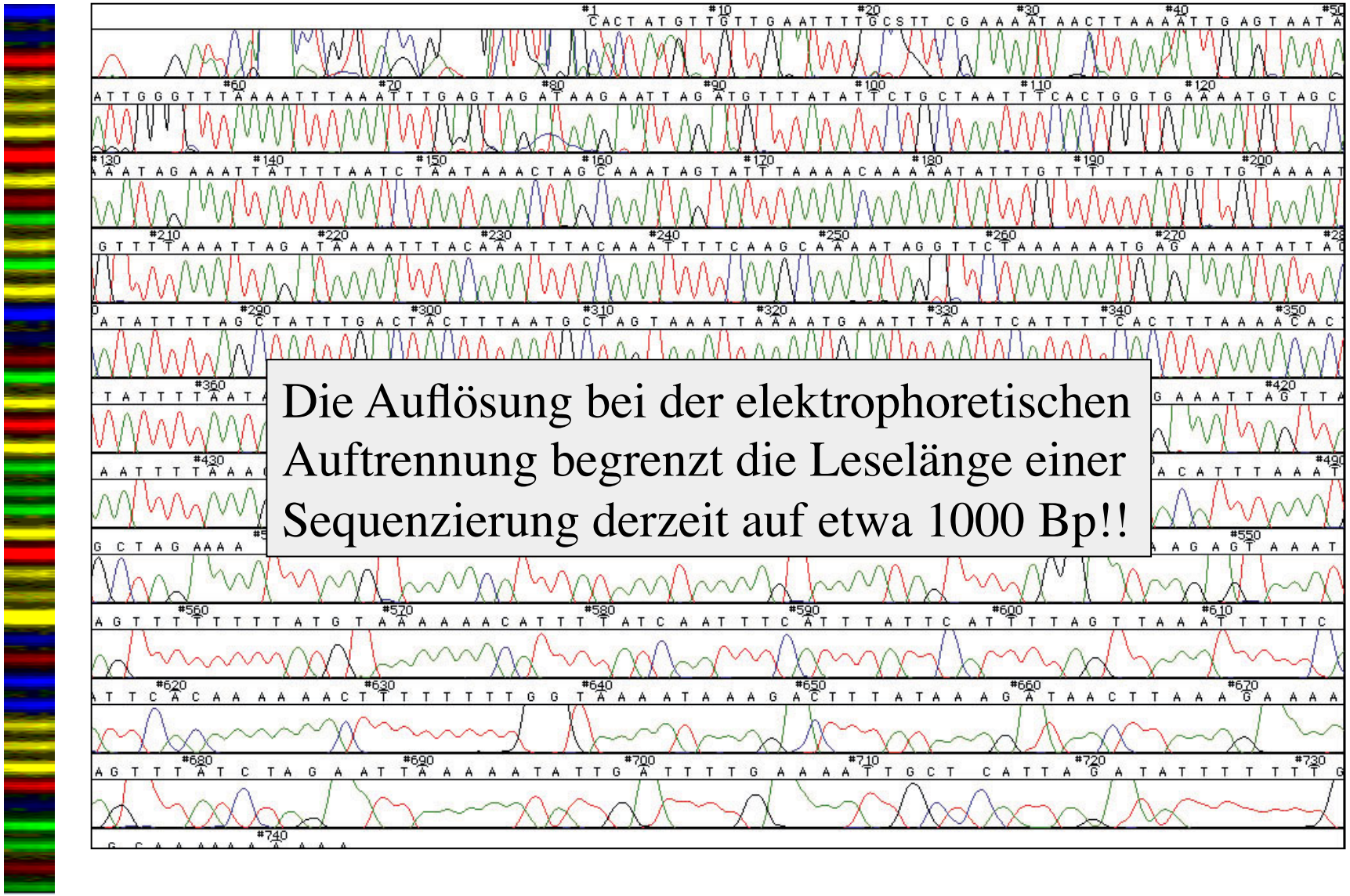
5‘-CTAGGACTGTAC **TCCT**^{Stop}

5‘-CTAGGACTGTAC **TCCTA**^{Stop}

5‘-CTAGGACTGTAC **TCCTAG**^{Stop}



Sequenzdaten-Chromatogramm



DNA ist ein kompliziertes Molekül!

„WATSON“

„CRICK“

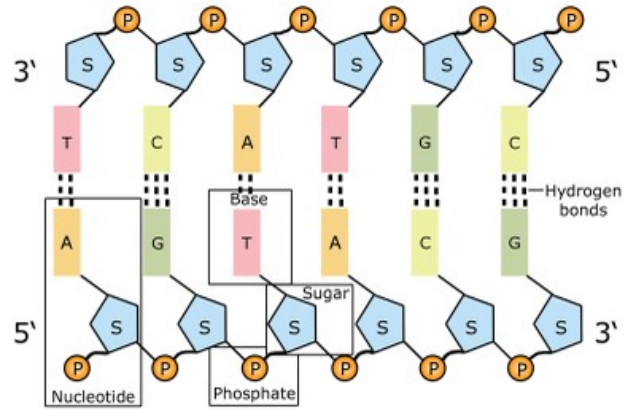


Image adapted from: National Human Genome Research Institute.

Basen sind komplementär!

Rückgrat ist chemisch gesehen anti-parallel! (5‘>3‘ bzw 3‘>5‘)

Q : Welchen Vorteil hat es für uns Genomforscher, dass die DNA aus **zwei** Strängen besteht?

„Doppelsträngige“ Sequenzierung!!

„WATSON“

„CRICK“

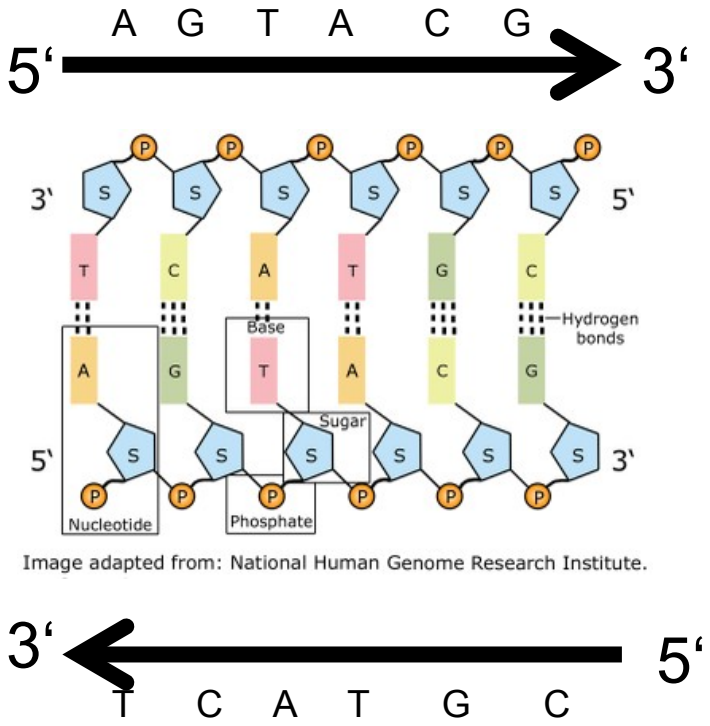


Image adapted from: National Human Genome Research Institute.

Passen die beiden Sequenzen fehlerlos zueinander?

Sequenzvergleich durch Alignment: die Schlüssel-Technik der Bioinformatik!

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

Nt-Substitution

As-Austausch

Gap bzw. InDel

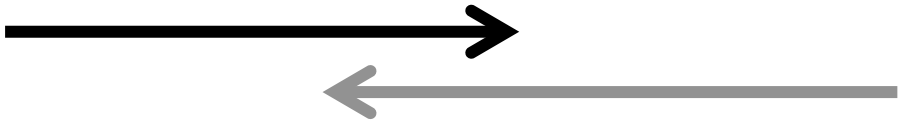
```
Query: 5   EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQY--NCRQFSSPEDCLSSPEFL 62
           + ELIR SW ++ ++ + HG +LF+RLF L+P+LL LF Y  NC     S +DCLSSPEFL
Sbjct: 8   DKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYTTNC---GSTQDCLSSPEFL 64
```

ähnliche As

identische As

Alignments können auf Nukleotid- oder Aminosäure-Ebene erfolgen

Sequenzvergleich durch Alignment



5‘-TTACTAC-3‘ und 5‘-TGCGGTA-3‘

5‘-TTACTAC-3‘
| | |
3‘-ATGGCGT-5‘

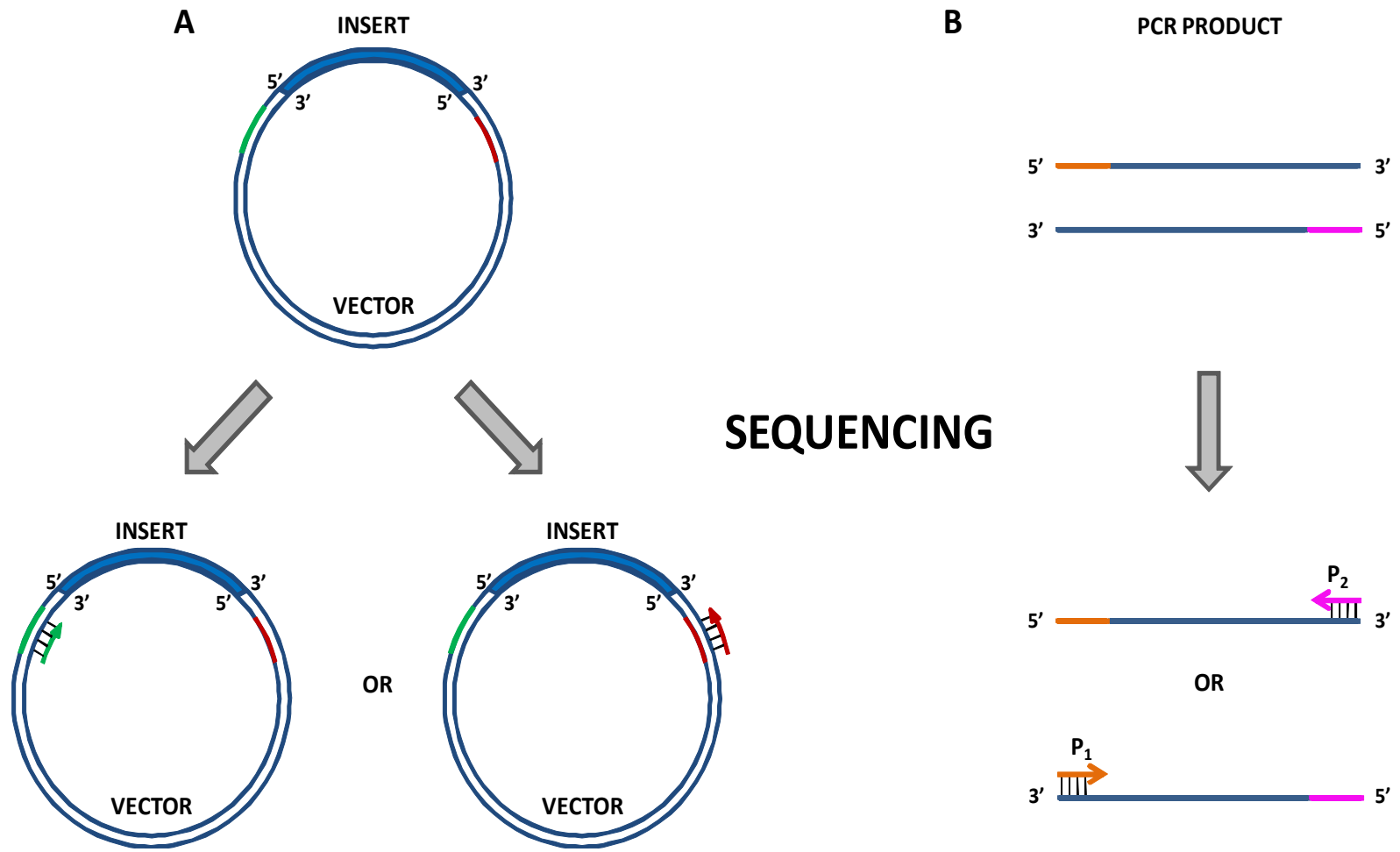
Sequenzvergleich durch Alignment

5‘-TTACTAC-3‘
und
5‘-TGCGGTA-3‘

5‘-TACCGCA-3‘ „Reverse Complement“

5‘-TTACTAC-3‘
5‘-TACCGCA-3‘

Welche Moleküle können wir so sequenzieren?



Practical Exercises 1

- Editing of Sanger chromatograms
- Spotting and explaining uncertainties...
- Alignment of forward and reverse reads
- Producing a consensus sequence

Examples:

a)Reads from a plasmid clone

b)Reads from a PCR amplificate obtained from SARS virus

c)Optional 1: „historic“ first PCR reads from SARS discovery

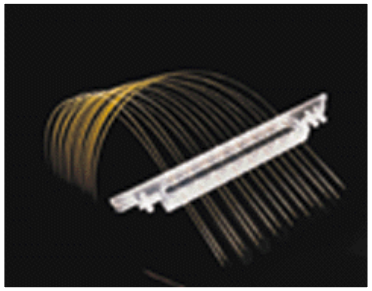
d)Optional 2: reads from a PCR amplificate produced to diagnose a human disease condition

Seit 20 Jahren...

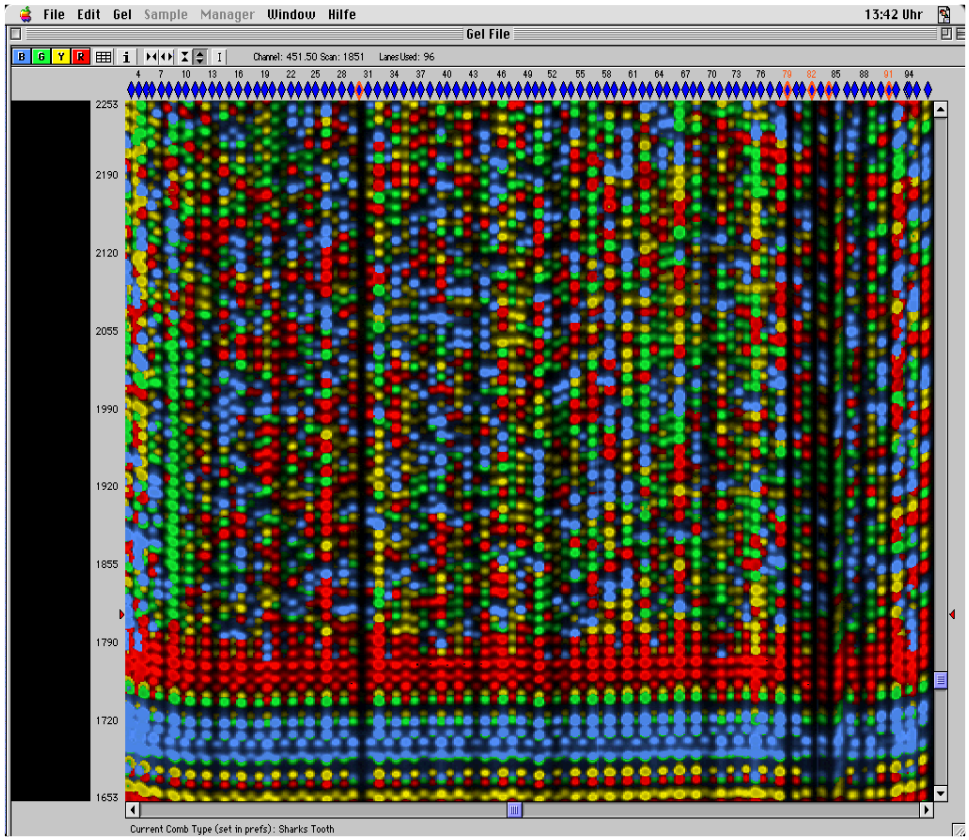
Sanger-DNA-Sequenzierung

Applied Biosystems 3730*xl* DNA Analyzer.

ABI 3730 Sequencer



Kapillaren

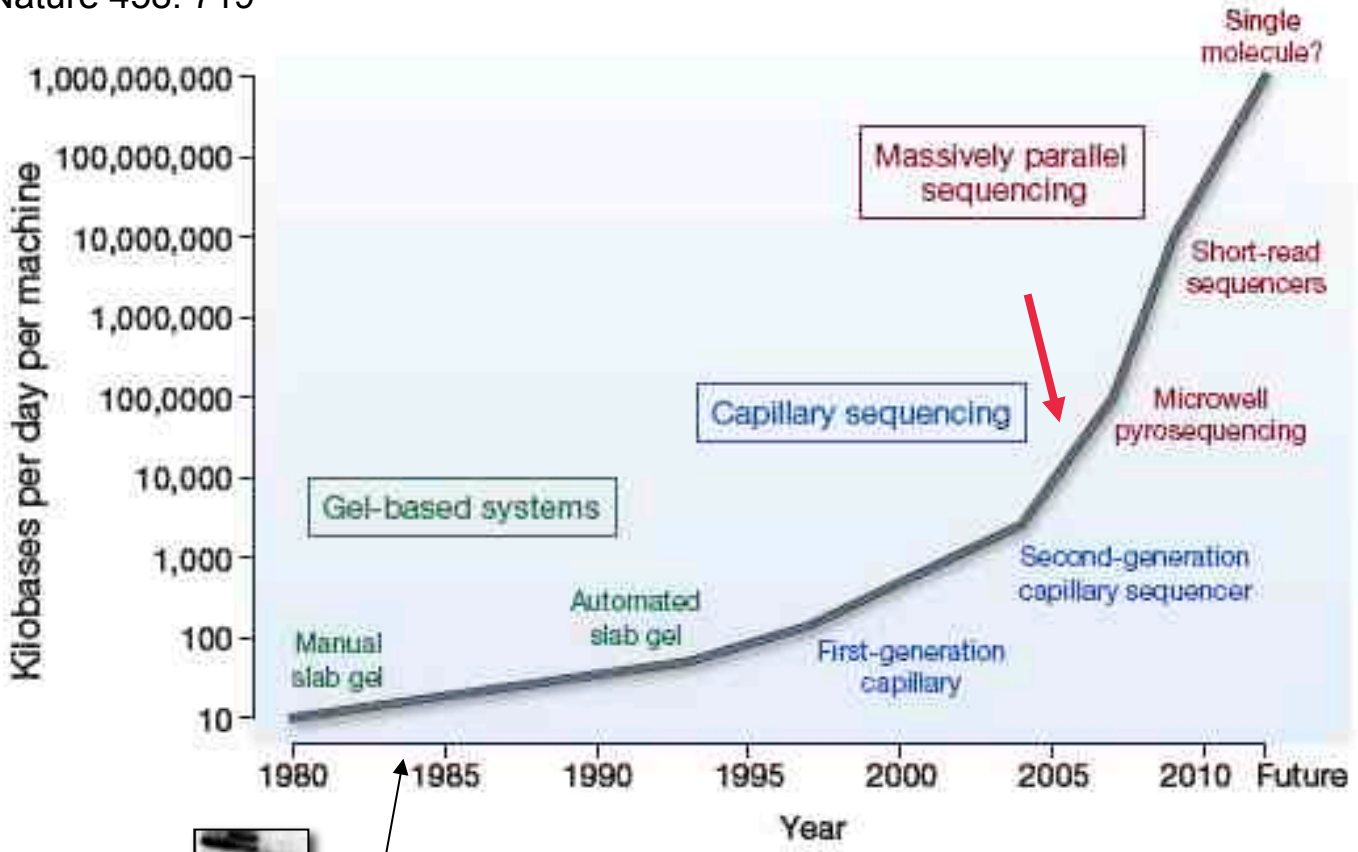


96 Spuren x 1000 Basen = ca. 100 000 Basen in ca. 2 Std

Next-Generation-Sequencing:

A million-fold improvement!

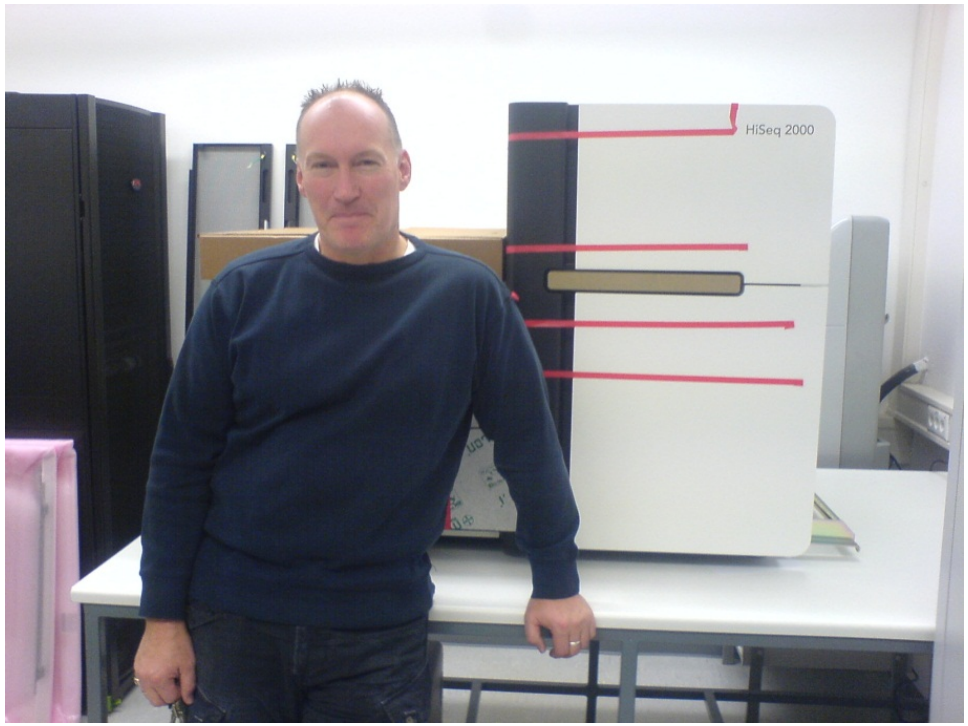
Nature 458: 719



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

Next-Generation Sequencing (NGS)

Illumina HiSeq 2500



2 Milliarden Reads
à jeweils 100 Bp

= 200 000 000 000 Bp

= ca. 60 x Humangenome
in 9 Tagen

(bzw. 2 Humangenome
mit je 30facher Abdeckung)

Celebrity genomics

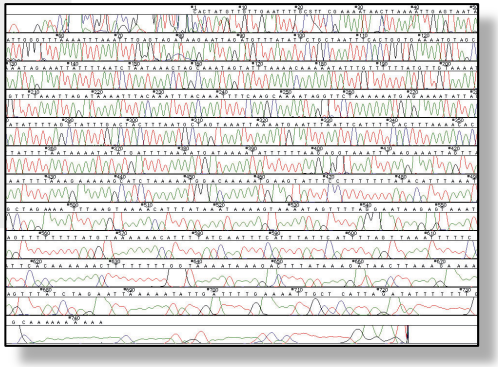
Science confirms the Neanderthal in Ozzy Osbourne

Among the findings that will presented Friday: Osbourne's genes reveal a probability of alcohol dependency "six times higher than the average person."

"All this is big news for blokes everywhere, I think: If the Neanderthals could get laid, there's hope for us all."

Sequenzierungsstrategien für große Genome sind erforderlich!

Aus technischen Gründen ist die Leselänge einer Sequenzierung begrenzt!!
Sanger: ca. 1000 Bp
Illumina: 100-300 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 ‚shotgun'-Strategie

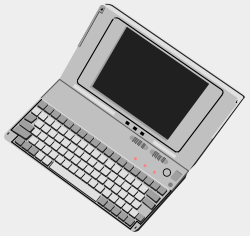
Ausgangs-DNA-Moleküle

‚zufälliges' Zerbrechen der DNA

DNA-Fragmente (überlappend!)

Sequenzreaktionen

Assemblierung der Teilsequenzen



```
ATGGCGAATGCCTTGACGCCGATGCAATTCAAG
   GGCGAATGCCTTGACGCC        ATTCAAGTGCATGTA..
```

Überlappung Überlappung

Konsensus-Sequenz


www.dnalc.org

Das Genom des SARS-Virus

- 1 Monat nach Virus-Identifikation 2 Genome sequenziert!
- Länge : 29 740 Bp (RNA), 13 Gene
- nach 3 Monaten > 20 Virus-Isolate sequenziert

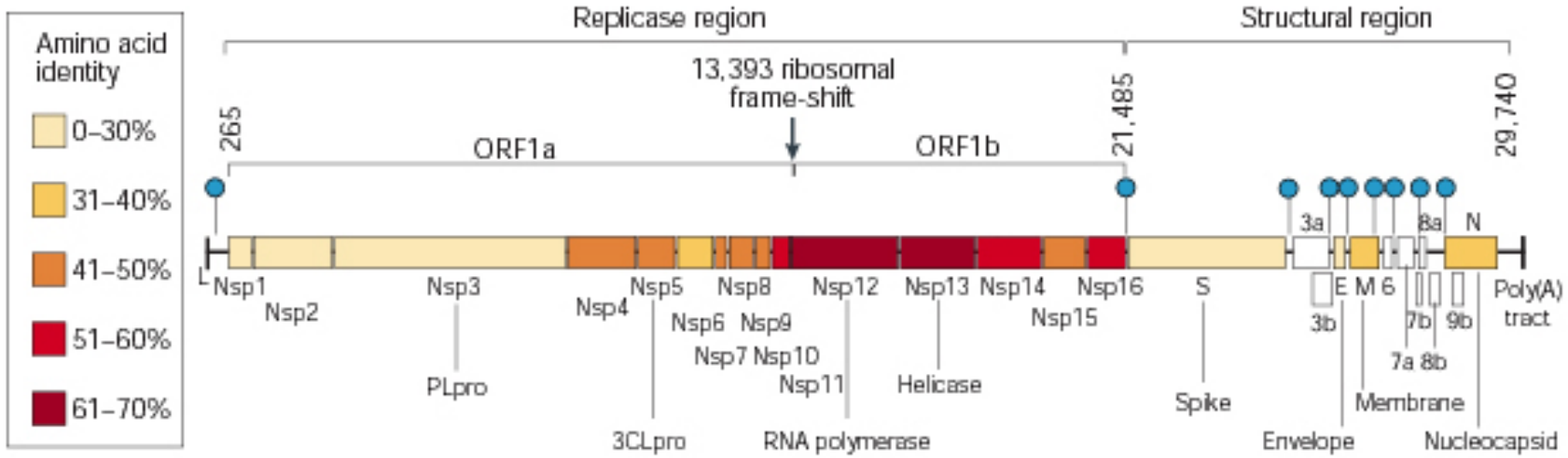


Figure 2 | **Genome structure of SARS coronavirus.** Replicase and structural regions are shown together with the predicted cleavage products in ORF1a and ORF1b. The position of the leader sequence (L), the 3′ poly(A) tract and the ribosomal frameshift site between ORF1a and ORF1b are also indicated. 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′).

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

news feature

First past the post

From the moment the mysterious illness known as SARS was declared a global threat to health, virologists were racing to develop a diagnostic test. Alison Abbott visits the tiny German lab that got there first.

A. ABBOTT

World-beater: Christian Drosten burned the midnight oil to be the first to produce a rapid test for SARS.

Christian Drosten is exhausted, but last week he was putting on a brave face for the television crews trying to squeeze into his poky lab. Drosten's fatigue, and his sudden celebrity, both stem from the fact that his team developed the first diagnostic test for severe acute respiratory syndrome (SARS).

Remarkably, Drosten and his colleagues pulled off this feat just 11 days after the World Health Organization (WHO) issued its alert about the disease. And since the test was unveiled on 26 March it has been distributed to more than 150 labs around the world. All in all, it's a considerable achievement for such a small team, given the high-powered virology labs that were engaged in the same quest.

For Drosten, who develops diagnostic tests for viruses and bacteria at the Bernhard Nocht Institute for Tropical Medicine in Hamburg, Germany, the past few weeks have been a whirlwind. The story began in early March, when Drosten and his team — research scientist Stefan Günther and a handful of students — were reading daily Internet postings about a mysterious respiratory illness in Vietnam. "Then, on 15 March, the WHO issued its global SARS alert, and two infected people, a doctor and his wife, landed at Frankfurt airport," Drosten recalls.

Drosten's speciality is polymerase chain reaction (PCR) diagnostics, in which a 'primer' corresponding to a distinctive sequence from a known virus or bacterium is used to amplify the pathogen's genetic material, if it is present in a sample. Initial tests in Frankfurt and elsewhere drew a blank in the search for a viral culprit in sputum taken from the doctor. So when a second sample, taken on 17 March, was sent to the Hamburg institute to see if a tropical virus might be involved, it fell into Drosten's hands.

Having ruled out the obvious tropical viruses, Drosten began to think about rare viruses that might cause symptoms similar to SARS. He first fixated on the family of paramyxoviruses, whose members include the rare Hendra and Nipah viruses, which can jump from animals to people. "Frankfurt colleagues had looked at the patient's sputum sample under the electron microscope, and the shape, like a squashed sphere, was reminiscent of a typical paramyxovirus," says Drosten.

Think again

Having worked through the night of 18 March on paramyxovirus tests, Drosten got nothing but negative results. He couldn't quite believe it — particularly when Canadian researchers working on samples from patients in Toronto suggested that the causal agent was a metapneumovirus, a type of paramyxovirus. So Drosten sent his sample to Europe's leading paramyxovirus expert, Albert Osterhaus at Erasmus University in Rotterdam, the Netherlands, who confirmed the negative diagnosis.

The remaining option was to try to fish out the elusive virus with a less-specific series of PCR reactions that can amplify the genetic material of a wide range of viruses. This required a pure culture of the virus — the patient's sample would be too full of human genetic material to yield a clean result. Here, Drosten's connections to Frankfurt were crucial. He went to medical school in the city, and many of his former colleagues still work there. And on 20 March, while he was in Frankfurt preparing a presentation with one such colleague, Drosten learned that a culture set up at the university had just begun to yield the virus.

Drosten sped back to Hamburg with a sample on 22 March. His new series of PCR tests took him through a sleepless weekend, during which the machine he was using to sequence the genetic material being amplified failed. But by 25 March, the recalcitrant device had delivered 20 or so sequences.

Two of these matched up with sequences from the coronavirus family. But just as the coronavirus sequence was coming off the machine in Hamburg, researchers at the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, announced that they had identified the SARS agent as a coronavirus. They also had an electron micrograph, which showed the typical crown-like appearance of a coronavirus that had been masked in the Frankfurt image.

Having been pipped by the CDC in implicating a coronavirus, Drosten wasted no time in creating primers that would allow other labs to test for its presence. The next day, he described them on the institute's website and offered to provide them for free. He also made available synthetic sequences that mimic the behaviour of the virus in the PCR test. These 'positive controls' are important, as negative PCR results can often mean simply that the procedure has failed. Although scientists in Hong Kong and at the CDC have since developed similar tests, they have not distributed them as widely, nor provided positive controls.

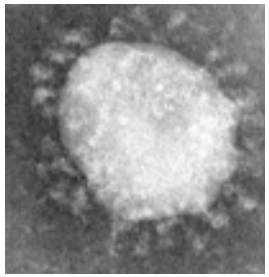
After such a hectic few weeks, Drosten is looking forward to life returning to normal. His colleagues, meanwhile, are basking in the reflected glory. "But we think he needs a long rest and a good feeding up," says one. ■

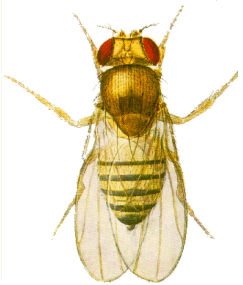
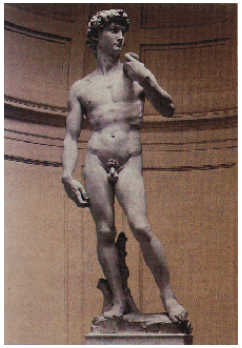
Alison Abbott is *Nature*'s senior European correspondent.

• www.bni-hamburg.de

Diagnostik durch PCR!

Genomgrößen im Vergleich



• SARS-Virus	29 kb	13 Gene*
• E. coli	463 kb	4 288 Gene
• Hefe	12 157 kb	6 692 Gene
• Fadenwurm	103 022 kb	20 362 Gene
• Fliege	142 573 kb	13 918 Gene
• Homo sapiens	3 555 000 kb	20 310 Gene
• Reis	374 424 kb	35 679 Gene !
• Ackerschmalwand	135 670 kb	27 655 Gene
• Huhn	1 285 000 kb	18 346 Gene

Stand Juli 2017

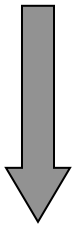
* Nur proteinkodierende Gene!

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und Sequenzierung

- Computer: Erkennen der Virusgene (*de novo* Genvorhersage)

  Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)



Verwandtschaft? Ausbreitung? Herkunft?
(Phylogenetische Rekonstruktion)

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

Wirkstoff-Design

- Labor: Wirkstoff-Test

Genvorhersage
& Genomannotation

Die (vereinfachte) Aufgabe...

- gegeben sind uncharakterisierte DNA-Sequenzen eines Genoms

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

- Mache daraus ein Modell für die Struktur der Gene!

Warum „vereinfacht“?

- gar nicht alle Gene werden in Proteine übersetzt!
  (Das Humangenom hat ca 20 000 RNA-Gene!!!)

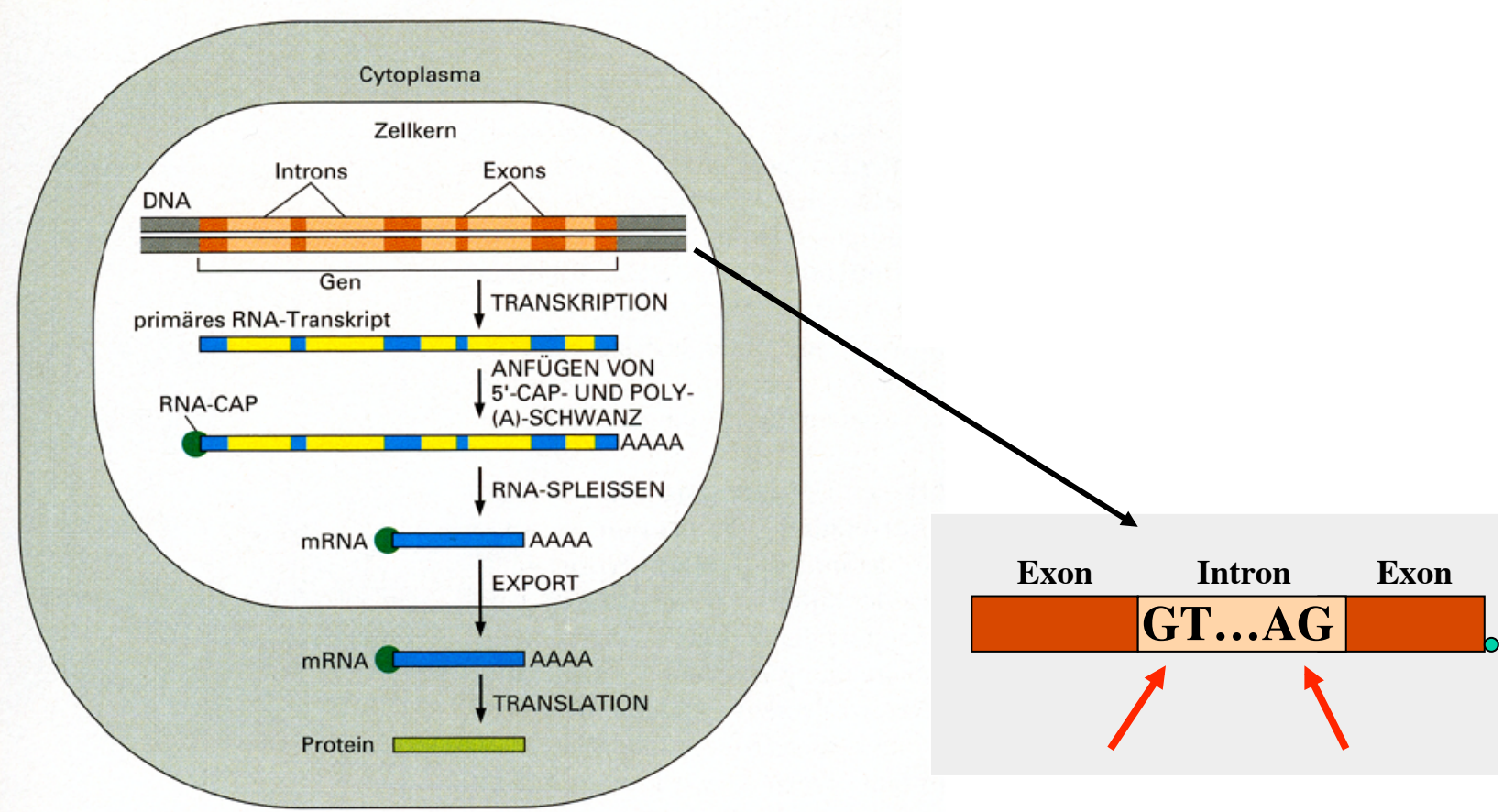
- auch nicht alle Genregionen proteinkodierender Gene werden in Proteine übersetzt (5‘ und 3‘-untranslatierte Exons)

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

Wo steckt denn nun das Gen?

```
  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 tcaggaaat 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
```

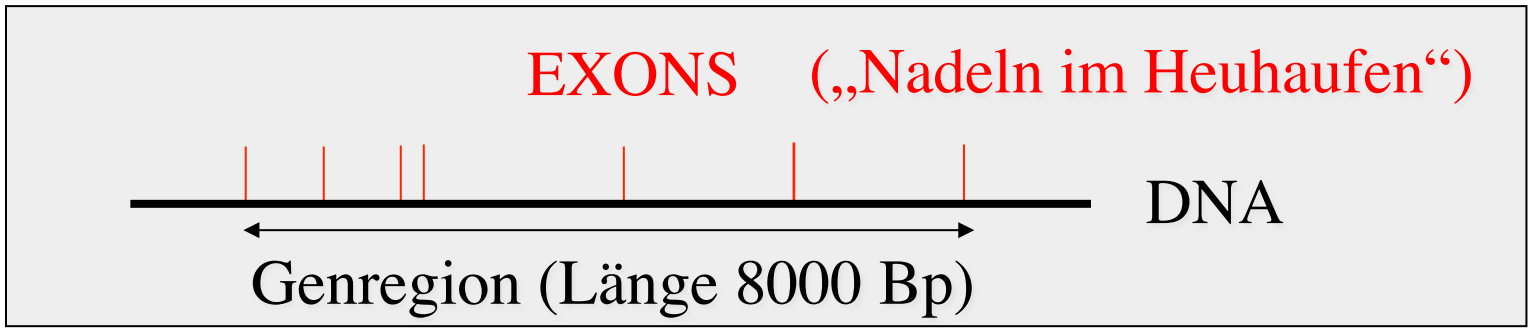
Bei Eukaryoten-Genomen ist Generkennung besonders schwierig



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

Mosaikgene erschweren die Gen-Identifizierung in Eukaryoten

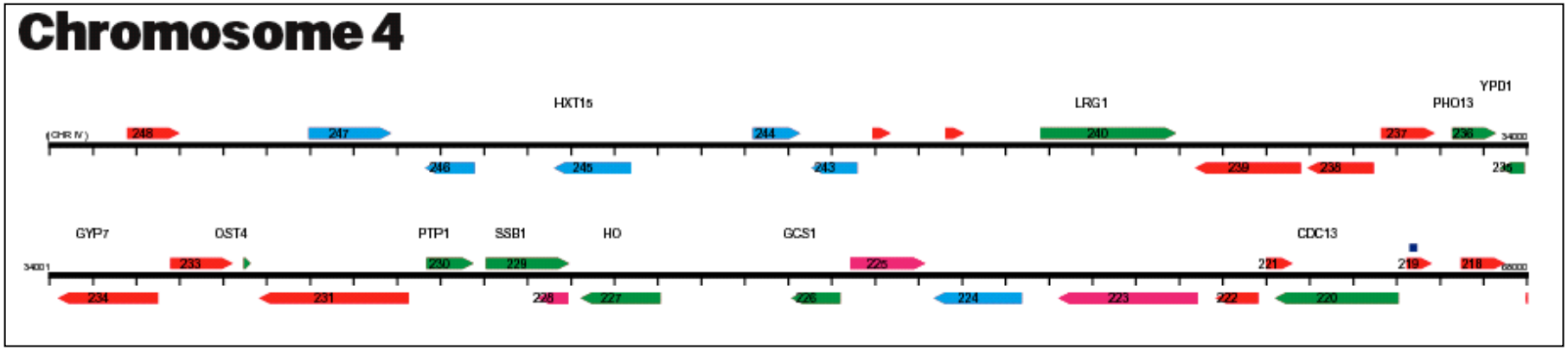
Ein typisches menschliches Gen:



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

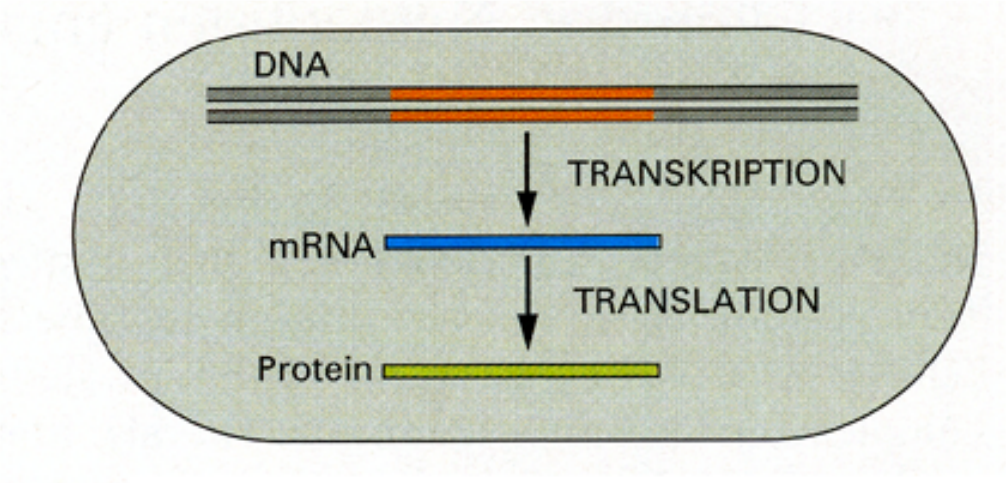
Beide Stränge der DNA können Gene kodieren!

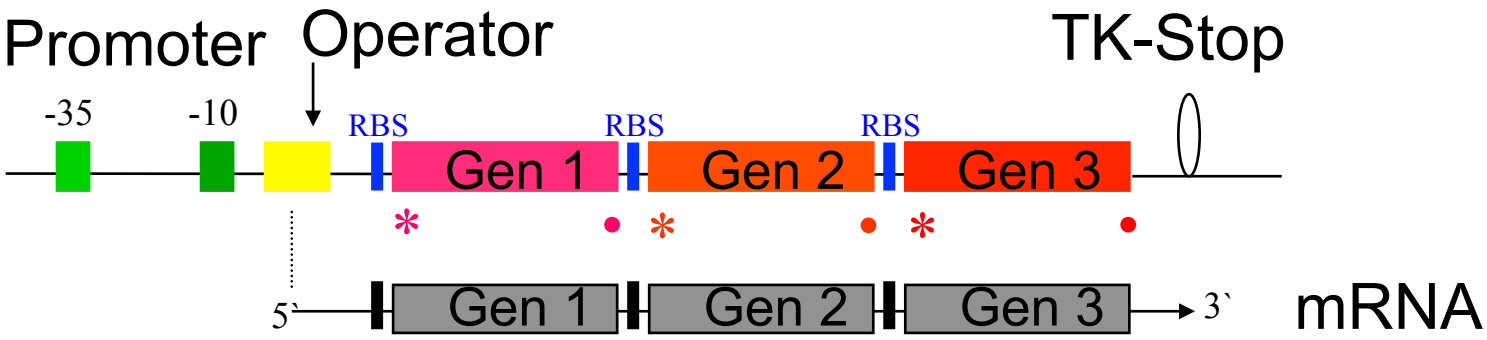
Ausschnitt aus dem Genom der Hefe:



- Pfeile zeigen Transkriptionsrichtungen an
- manche Gene überlappen, d.h. beide Stränge der DNA kodieren dort

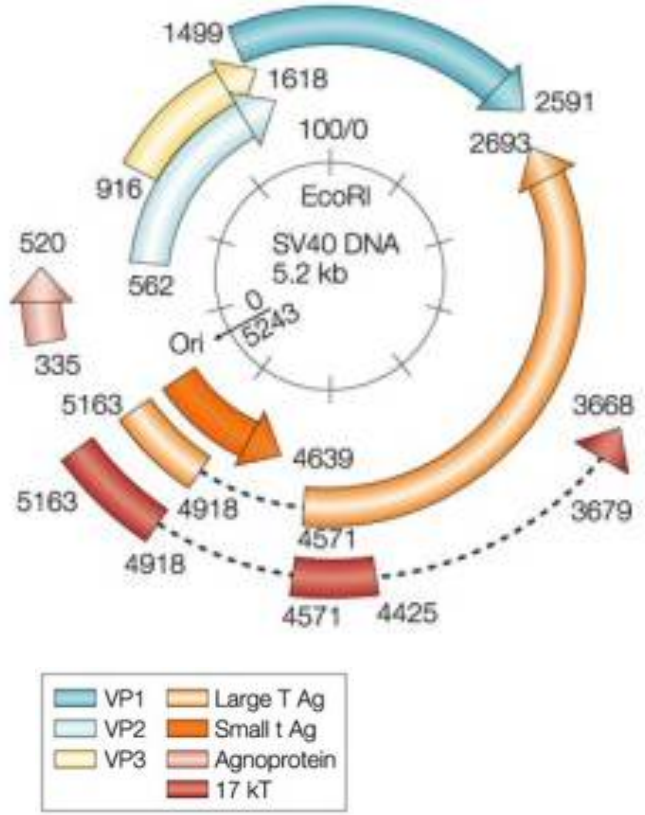
Bakteriengene sind einfacher gebaut und oft in Operons arrangiert





RBS ribosome binding site (Shine-Dalgarno-Box) * Translations-Startkodon (ATG) • Translations-Stopkodon

Virengenome sind kompakt und besonders ökonomisch genutzt



Gazdar et al. 2002

- Gene liegen sehr dicht gepackt
- Gene können partiell überlappen
- beide Stränge der DNA kodieren

Die zwei Strategien der Gensuche

- ich wende Allgemeinwissen an, wie Gene „gebaut“ sind

> *de novo* Genvorhersage

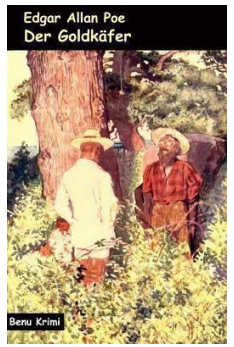
- ich suche: gibt es woanders schon ähnliche Gene?

> Datenbank-Suche

Alles geht! Oder:
Edgar Allen Poe und der DNA-Geheimcode

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

```
53‡‡†305))6*;4826)4‡.)4‡):806*;48†8¶60))8
5;1‡(;:‡*8†83(88)5*†;46(88*96*?;8)*‡(;485
);5*†2:*‡(;4956*2(5*--4)8¶8*;4069285);)6†8
)4‡‡;1(‡9;4808 1;8:8‡1;48†85;4)485†528806
*81(‡9;48;(88;4(‡?34;48)4‡;161;:188;‡?;
```

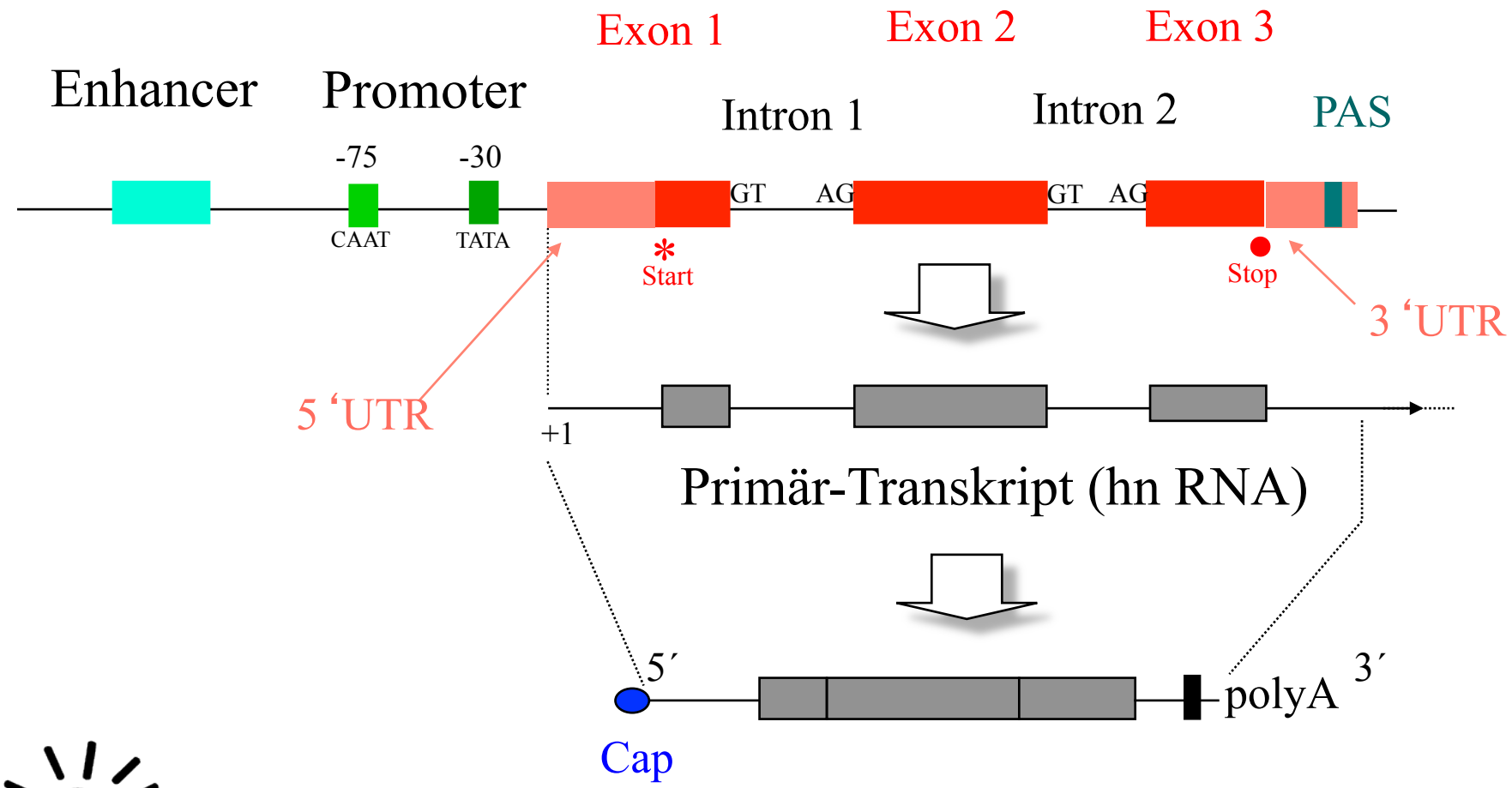


- häufigstes engl. Wort? ;48 the

```
53‡‡†305))6*THE26)H‡.)H‡):E06*THE†E¶60))E
5T1‡(T:‡*E†E3(EE)5*†T46(EE*96*?TE)*‡(THE5
)T5*†2:*‡(TH956*2(5*--H)E¶E*TH0692E5)T)6†E
)H‡‡T1(‡9THE0E1TE:E‡1THE†E5TH)HE5†52EE06
*E1(‡9THET(EETH(‡?3HTHE)H‡T161T:1EET‡?T
```

Remember Grundvorlesung Modul 8...

Struktur von RNA Pol II-Genen



Wir kennen also „Signale“ in Genen

- 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

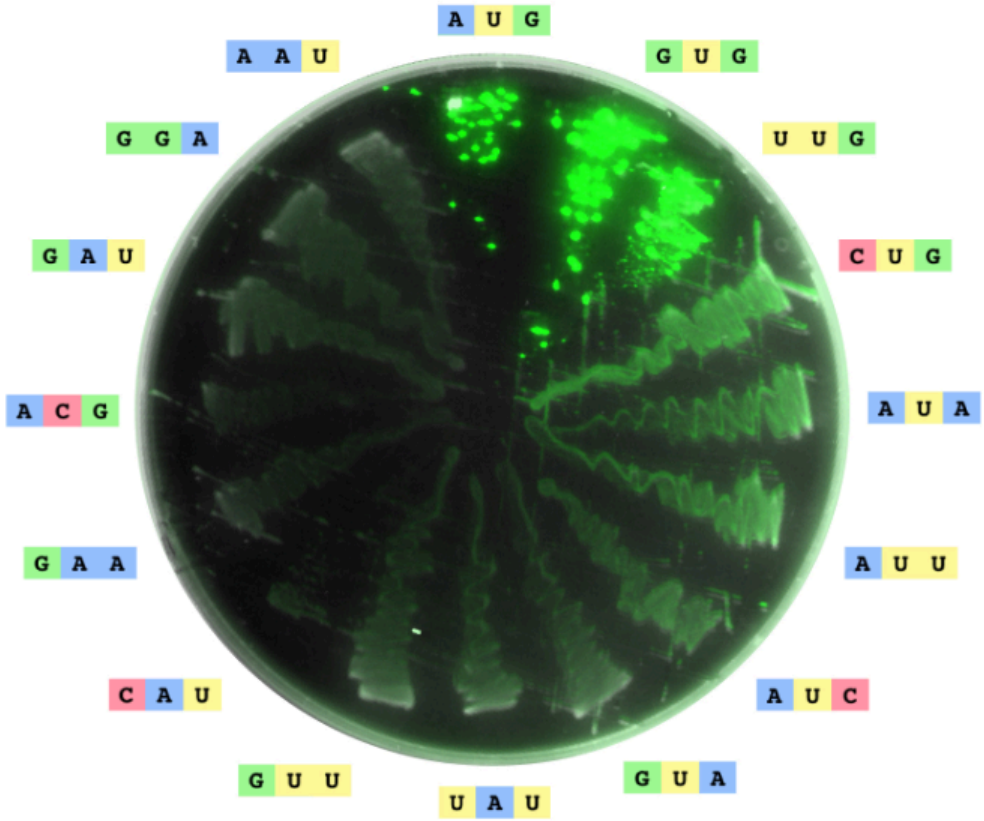
Published online 21 February 2017

Nucleic Acids Research, 2017, Vol. 45, No. 7 ***3615–3626***
doi: 10.1093/nar/gkx070

NAR Breakthrough Article

Measurements of translation initiation from all 64 codons in *E. coli*

Ariel Hecht[1,2,3,†], Jeff Glasgow[1,2,3,†], Paul R. Jaschke[3,4,†], Lukmaan A. Bawazer[1,2,3,†], Matthew S. Munson[1,2,3], Jennifer R. Cochran[1,3], Drew Endy[1,3,*] and Marc Salit[1,2,3,*]

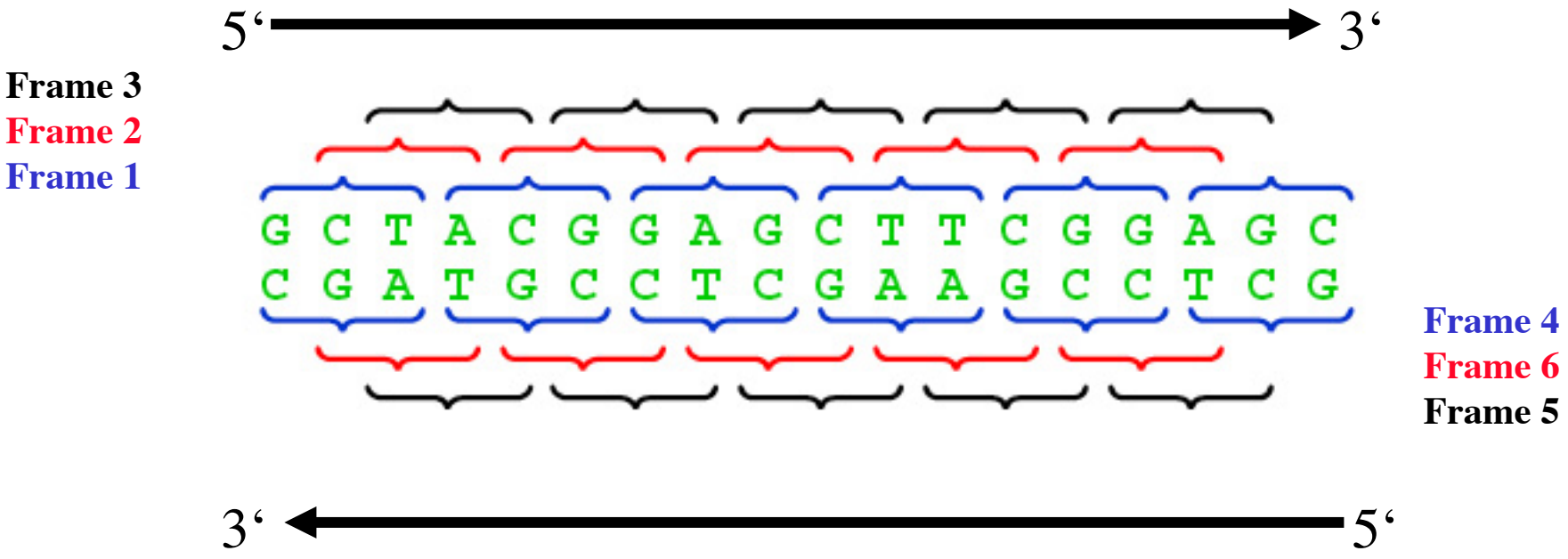


ABSTRACT

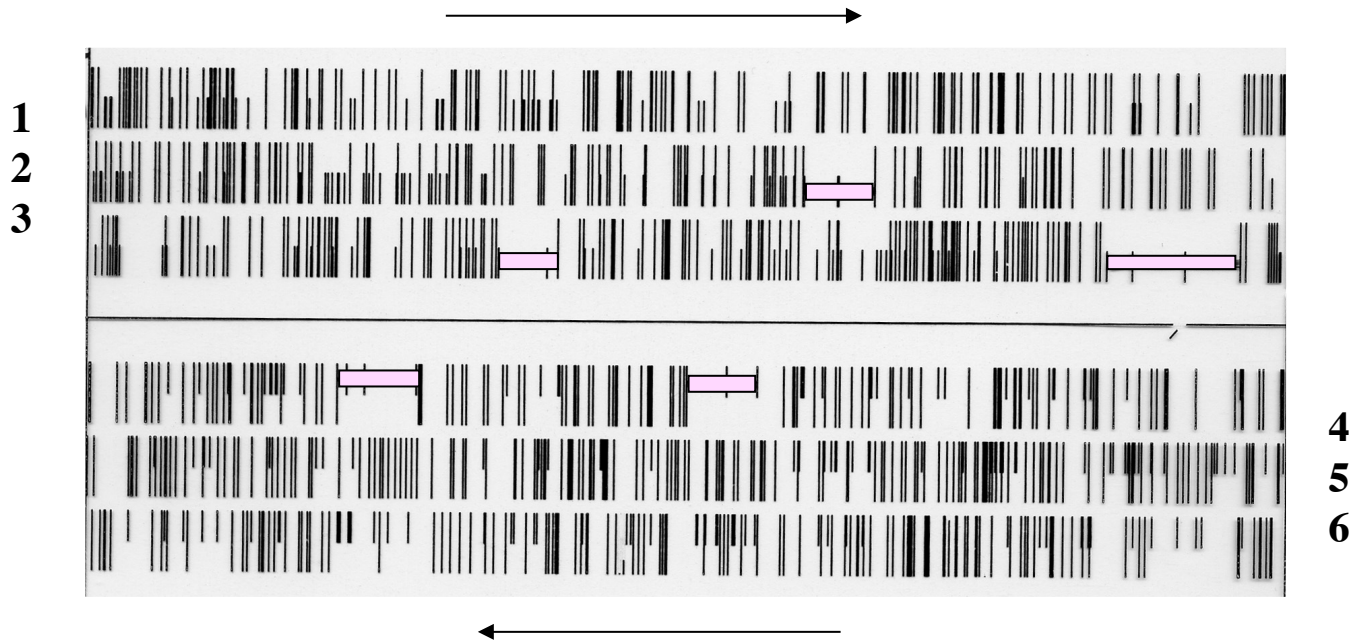
Our understanding of translation underpins our capacity to engineer living systems. The canonical start codon (AUG) and a few near-cognates (GUG, UUG) are considered as the 'start codons' for translation initiation in *Escherichia coli*. Translation is typically not thought to initiate from the 61 remaining codons. Here, we quantified translation initiation of green fluorescent protein and nanoluciferase in *E. coli* from all 64 triplet codons and across a range of DNA copy number. We detected initiation of protein synthesis above measurement background for 47 codons. Translation from non-canonical start codons ranged from 0.007 to 3% relative to translation from AUG. Translation from 17 non-AUG codons exceeded the highest reported rates of non-cognate codon recognition. Translation initiation from non-canonical start codons may contribute to the synthesis of peptides in both natural and synthetic biological systems.

Proteinkodierende Gene haben auch einen „*besonderen Inhalt*“

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



Computer-Suche nach ORFs



DNAstar Programmpaket

Start

Stop

Potenzielle Gene

Reine ORF-Suche ist nicht ausreichend, um exakte Modelle komplizierter Gene vorherzusagen!

Moderne integrierte Genvorhersage-Programme suchen in „anonymen“ Sequenzen nach Hinweisen auf Gene.

Dabei werden statistische Kenntnisse zur Architektur von Genen benutzt...

...Hidden Markov Models (HMM)

Markov WER??

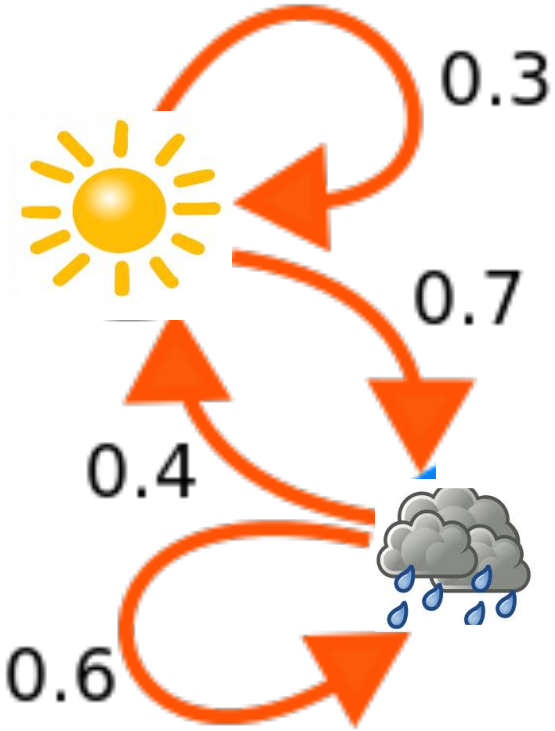
- Andrei Andreyevich Markov (1856-1922)
- Markov-Kette:

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

Markov-Kette



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%.

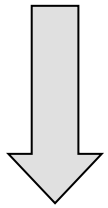
  ⇒ zähle die e‘s im Text, um zu berechnen mit welcher Wahrscheinlichkeit es sich um einen deutschen Text handelt

Hidden Markov Models

Q: 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“)

Aufgabe: suche ein Gen-Signal

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

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

Suche weitere, dazu passende Gen-Signale mithilfe einer einfachen *Textsuche* :

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

Diese Textsuche kann aber nicht unterscheiden zwischen...

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

`ACAC--ATC`

b) einer höchst unwahrscheinlichen Sequenz

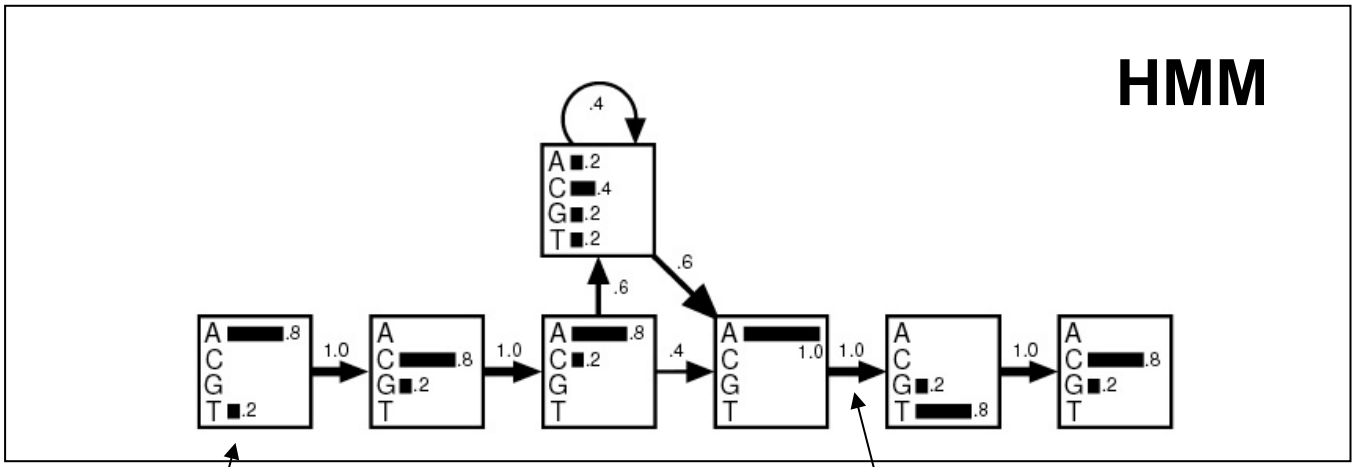
`TGCT--AGG`

Funktionsprinzip eines HMM bei der Suche nach dem Gen-Signal

Besser ist:

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

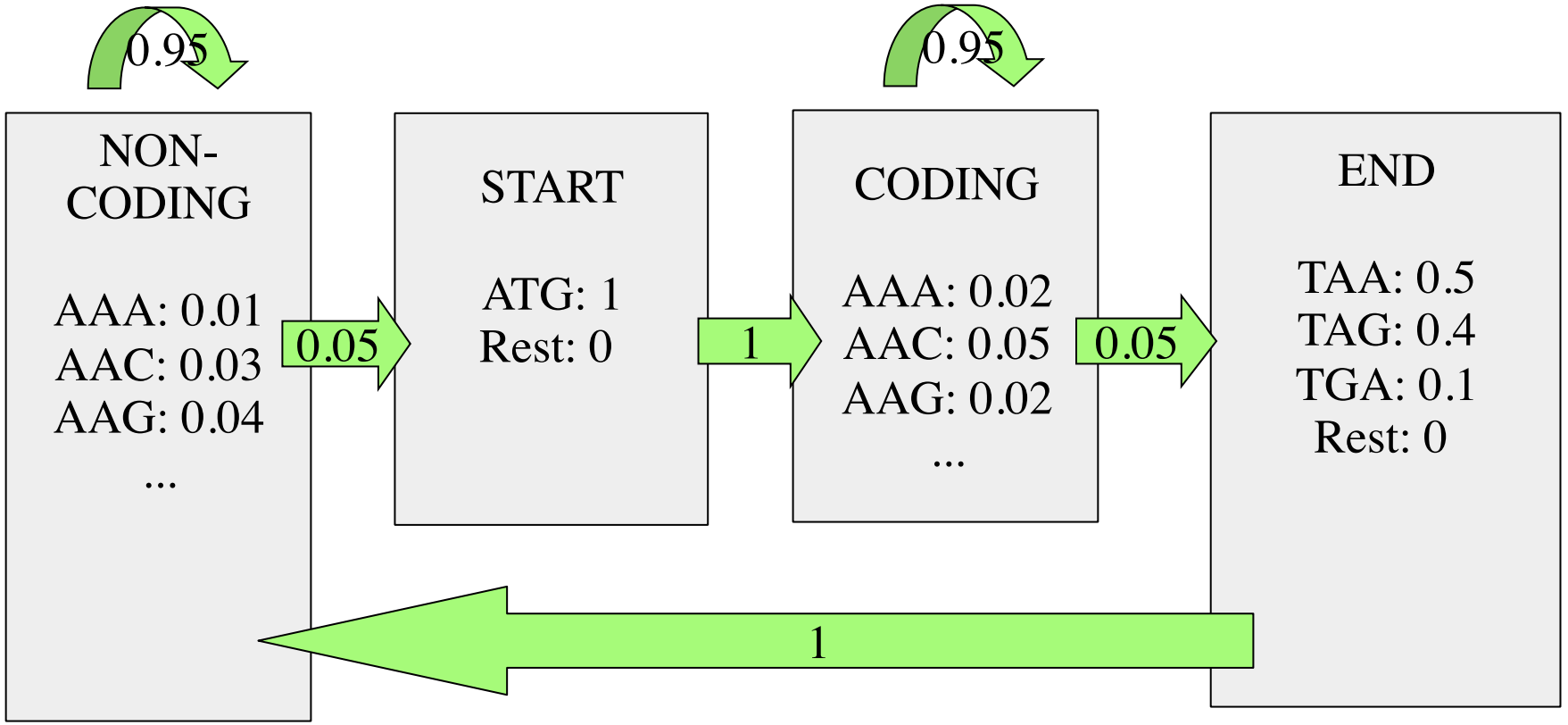
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



Emissions-
wahrscheinlichkeit

Übergangs-
wahrscheinlichkeit

Beispiel: HMM-Sensor für ein einfaches prokaryotisches Gen



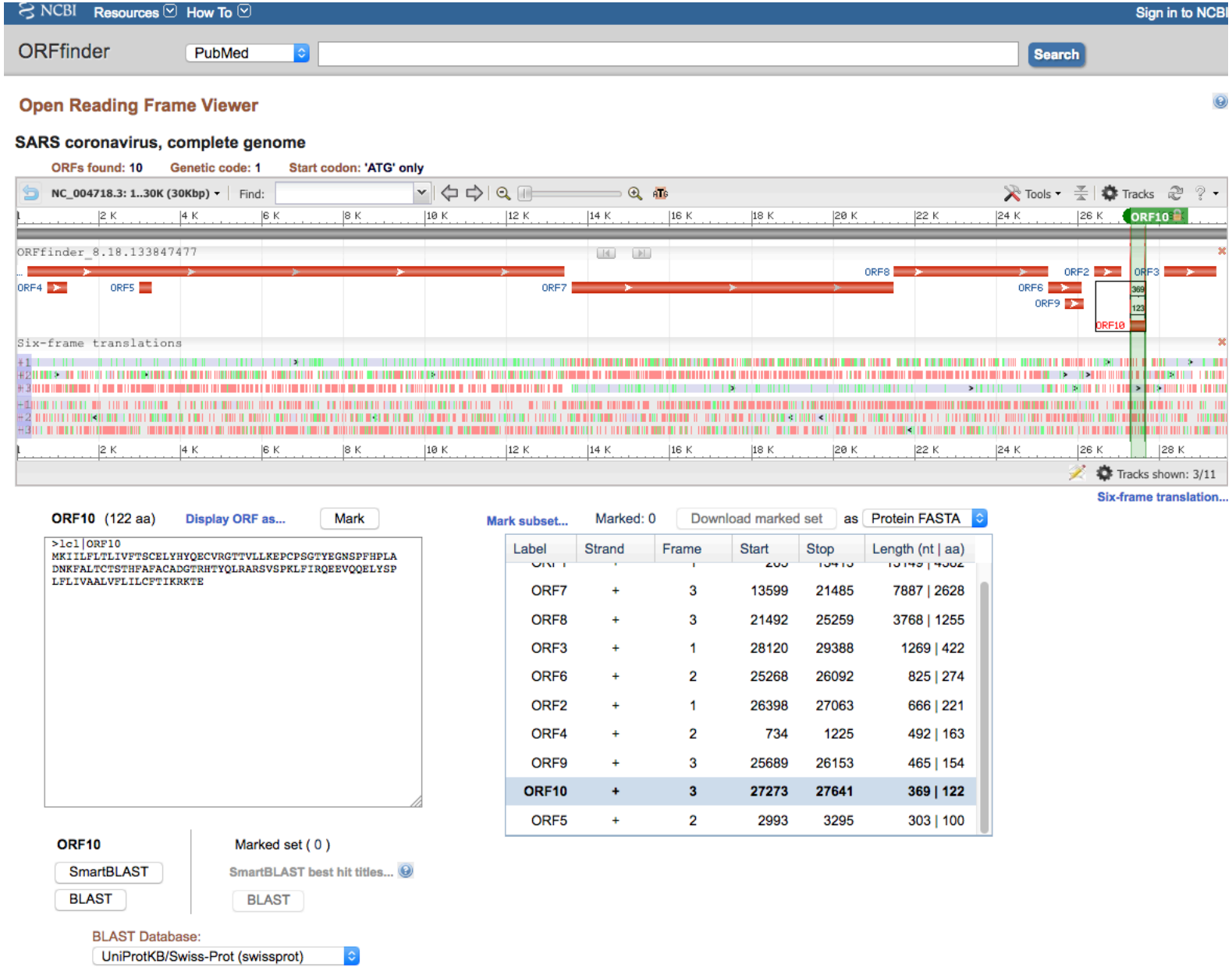
Practical Exercises 2

- Detect possible ORFs in the SARS virus genome:
  how many gene candidates do you find?

- Search for ORFs in the „anonymous“ cloned sequence from excercise 1

- Apply a HMM-based gene prediction tool:
  Do you find gene signals in the sequence from excercise 1?
  Which type of gene is it?

Der NCBI-ORFfinder



SARS-Genom und seine Gene

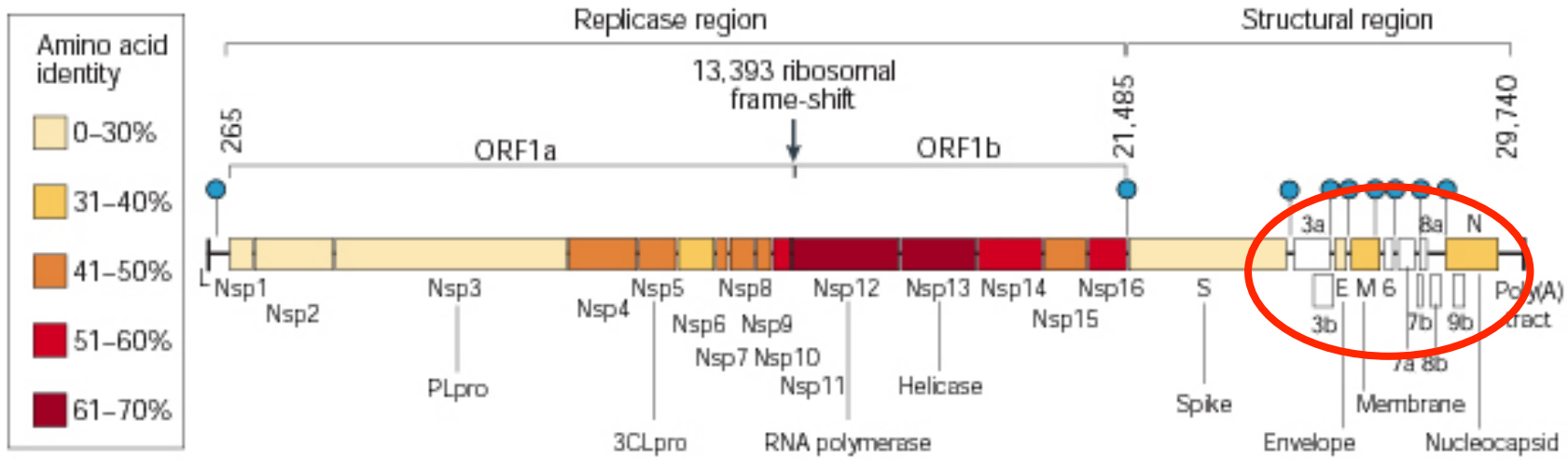


Figure 2 | **Genome structure of SARS coronavirus.** Replicase and structural regions are shown together with the predicted cleavage products in ORF1a and ORF1b. The position of the leader sequence (L), the 3′ poly(A) tract and the ribosomal frameshift site between ORF1a and ORF1b are also indicated. 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 überlappende proteinkodierende Genbereiche!

SARS-ORFs: was kodieren sie?

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.

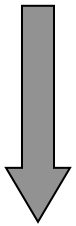
ORFs mit z.T. unbekannter Identität und Funktion

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz und „Sequenzierung“

- Computer: Erkennen der Virusgene (*de novo* Genvorhersage)

  Ähnlichkeit zu **bekannten** Genen? (Datenbanksuchen)



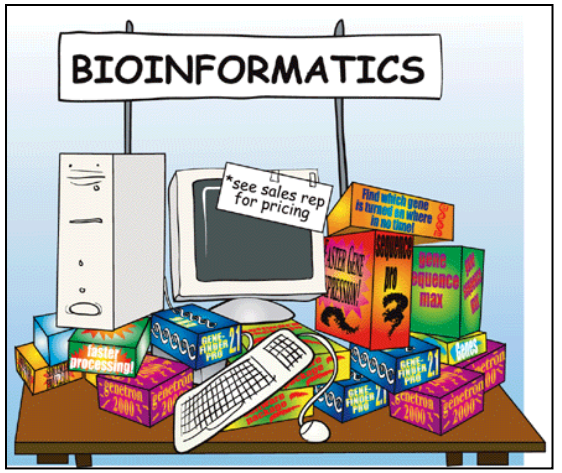
Verwandtschaft? Ausbreitung? Herkunft?
(Phylogenetische Rekonstruktion)

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

Wirkstoff-Design

- Labor: Wirkstoff-Test

Datenbanken in der Molekularbiologie



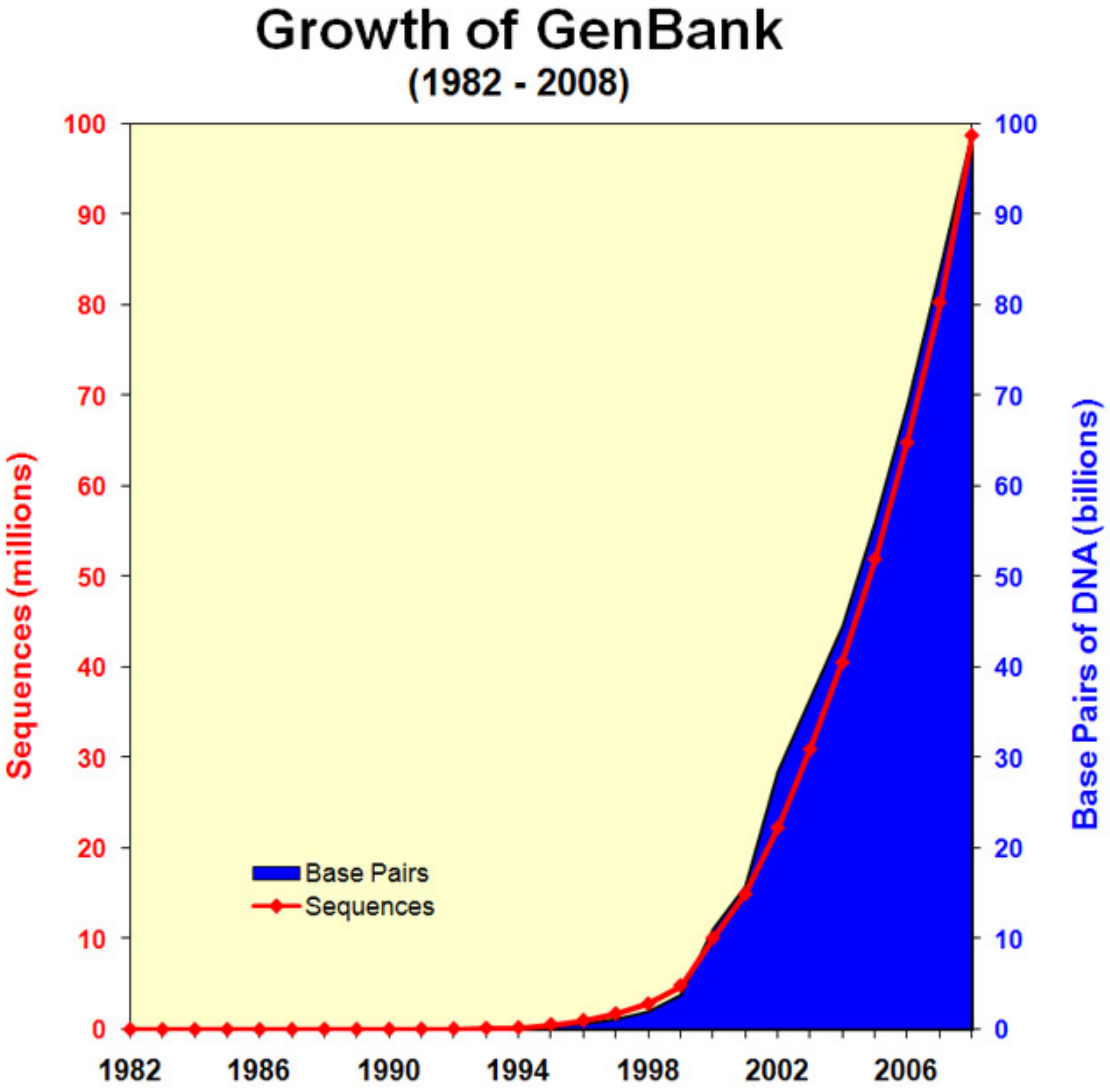
- **Literaturdatenbanken (z.B. PubMed)**
- **Sequenzdatenbanken**

  - primäre DB: DNA- u. Proteinsequenzen

  - abgeleitete DB: interpretierte Sequenzdaten

    (z.B. Proteindomänen oder Stoffwechselwege)

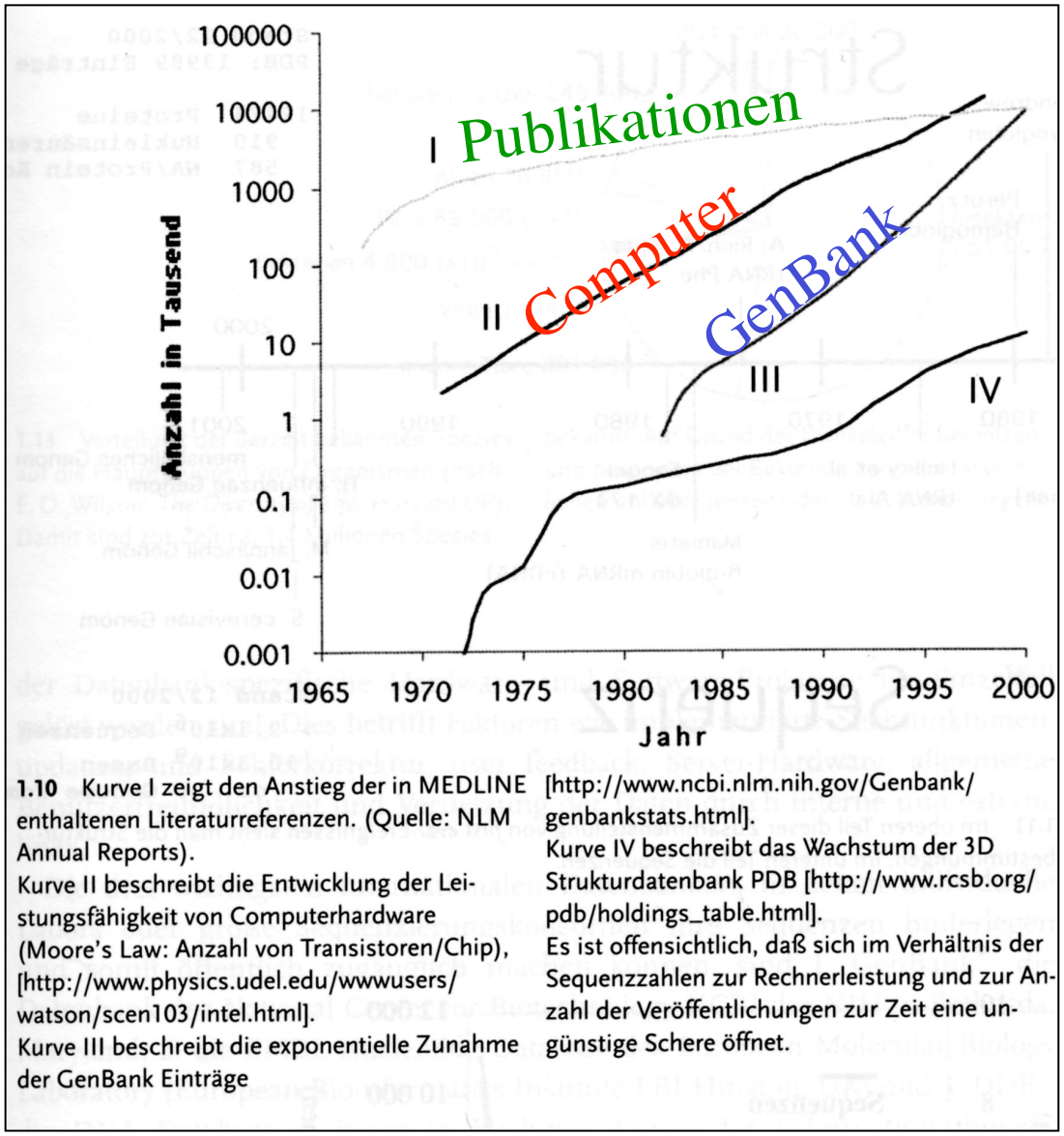
Datenbank-Wachstum



September 2017:

? Einträge
? Bp

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.

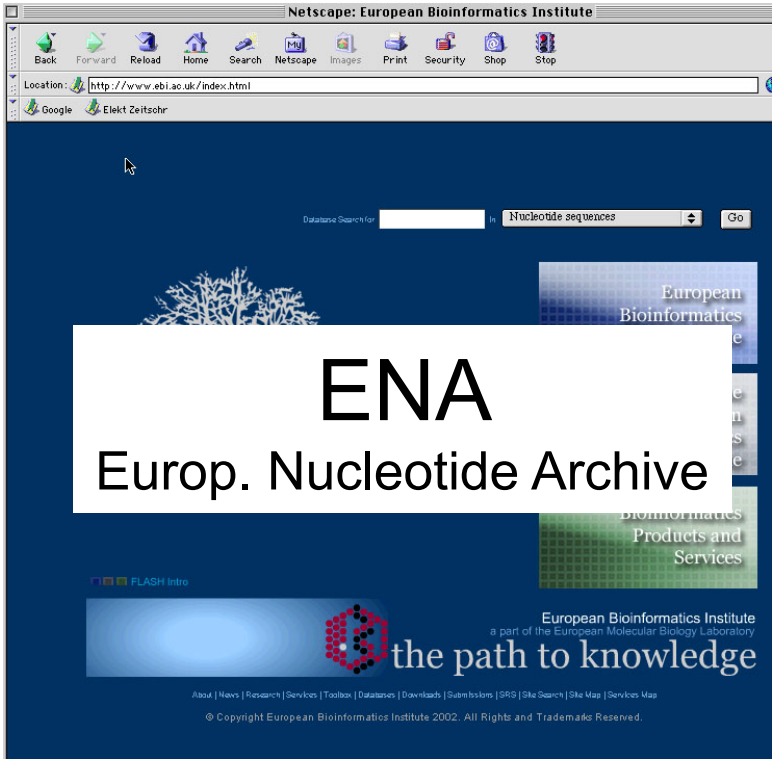
Datenbanken in der Molekularbiologie

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



Ein GenBank-Eintrag

accession no.

Version

Zitat

CDS = coding sequence

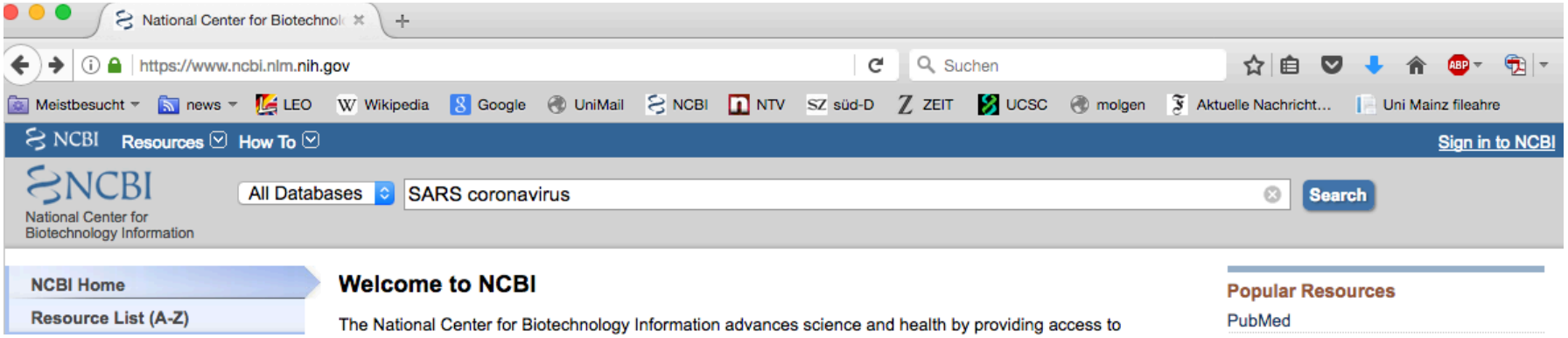
übersetzte Protein-sequenz

Nukleotidsequenz

```
1: AJ315164. Mus musculus Cygb..

LOCUS       MMU315164               9488 bp    DNA     linear   ROD 09-JUL-2002
DEFINITION  Mus musculus Cygb gene for cytoglobin.
ACCESSION   AJ315164
VERSION     AJ315164.1  GI:21727817
KEYWORDS    CYGB gene; cytoglobin.
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE   1
  AUTHORS   Ebner,B., Burmester,T. and Hankeln,T.
  TITLE     Comparative sequence analysis of the mouse cytoglobin gene
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 9488)
  AUTHORS   Hankeln,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-JUL-2001) Hankeln T., Inst. Molekulargenet., Univ.
            Mainz, J.J. Becherweg 32, Mainz, D-55099, GERMANY
FEATURES             Location/Qualifiers
     source          1..9488
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
     gene            1736..8693
                     /gene="Cygb"
     mRNA            join(<1736..1878,5637..5868,6206..6369,8660..>8693)
                     /gene="Cygb"
     CDS             join(1736..1878,5637..5868,6206..6369,8660..8693)
                     /gene="Cygb"
                     /codon_start=1
                     /product="cytoglobin"
                     /protein_id="CAC86190.1"
                     /db_xref="GI:21727818"
                     /translation="MEKVPGDMEIERRERSEELSEAERKAVQATWARLYANCEDVGVA
                     ILVRFFVNFPSAKQYFSQFRHMEDPLEMERSPQLRKHACRVMGALNTVVENLHDPDKV
                     SSVLALVGKAHALKHKVEPMYFKILSGVILEVIAEEFANDFPVETQKAWAKLRGLIYS
                     HVTAAYKEVGWVQQVPNTTTPPATLPSSGP"
     exon            <1736..1878
                     /gene="Cygb"
                     /number=1
     intron          1879..5636
                     /gene="Cygb"
                     /number=1
     exon            5637..5868
                     /gene="Cygb"
                     /number=2
     intron          5869..6205
                     /gene="Cygb"
                     /number=2
     exon            6206..6369
                     /gene="Cygb"
                     /number=3
     intron          6370..8659
                     /gene="Cygb"
                     /number=3
     exon            8660..>8693
                     /gene="Cygb"
                     /number=4
BASE COUNT     2078 a    2830 c    2633 g    1947 t
ORIGIN
        1 ttttgttatt agtgtgtgtg tgtgtgtgtg tgtgtgtgtg tgtgtgtgtg tgtgtgtgtg
       61 tgtgagaaag gacagcttgt aagagtcaat ctcgggctgg tgagatggca cagtgggtaa
      121 gagcacccga ctgctcttct gaaggtccgg agttcaaatc ccagcaacca catggtggct
      181 cacaaccatc cgtaacaaga tctgacgccc tcttctggag tgtctgaaga cagctacagt
      241 gtacttacat ataaataaat aaatctttaa aaaaaaaaaa aaaaagagtc aatctcttcc
      301 ttccaccccg tgggtcctag ggccggaact cttcagatca tcaagttttg tgaggcaaat
      361 acctttaaag attttaaatc aacagggacc caagctgaag aggagagcca ctcacttcct
      421 gggcagcagg gctcctcctc atcttgagct gggagccttg aggagaacca gagacagtgc
      481 acttatctgt gtcaggacca ggcaggccct ctgtggtctc agggtccctg ctgtctacag
      541 ccaggtctga gcaactgctg gcagggtgag ggttctgggt ctctcaagac tgcactcctc
      601 ccctctgtcc cagtgtcacc tctccctgag cagtctaaga aggagatgaa ggatctgcct
      661 tctggttctc aaactgaact ctcgatggtg acaaatgtct tcactgtccg gtgccttact
      721 caggacttcc ggctccccag agcctctcca tcatacctga ctgactgcct ctctgtggaa
      781 cttcactcca ccacggctag ggctaggacg gtaattcagg acagtgtctg gtccttatct
      841 acctatcatt aactaccttc tcaggactcc ctgctcgcag gcgtagggag ggaagtgggg
      901 cacgggtctc tggctcc                                    aagagc cagctggacc
      961 agcagatggg aagagag                                   actcga ggtccccacg
     1021 tctgacaacg ggtgcct                                   ggggct acaggggaag
     1081 cccggatttg tctgcaa                                   gggtgt gtttgacttt
     1141 gagtcagatc cccacctcgt gaagggtgta cacacacaca cacacacaca cacacacaca
     1201 cacacacgtc cagacaccac acgtcgagag ctgagacccg caccctcagc tccgcaccag
     1261 cccggggcgc acgccagtac acccggcgct gtccatcacc gcgtgacccc ctgatcctcc
     1321 cagccccttc tgcacattgg ccacacctac ctctcccagc cgggaccgcg gtggccttgc
     1381 taacgggtgg gtggtcaggc aggcagacgc cagccgtgac acccccatcc cgcctaacct
     1441 tcaccttgcc aaaattgact cccagaaaca ggactggatt ttttggagcg gatttttttt
     1501 aaaaaacatt ttttccccag cagaccacat ctccgccccc agctcgcagt cccccgcccc
     1561 ccgcacatat accctgcaca cccgcgcgca cacacacccg gcgcgcacag acacacgctc
     1621 cctccctccg gcgctctact cctcgcccgc ccgccctcct gcccgcctcg cacgcaccgg
     1681 cctcggctcc ccgcgcccgc cgccgagcga agctgcgctg ggctcggagc tgctcatgga
     1741 gaaagtgccg ggcgacatgg agatagagcg tagggagagg agcgaggagc tgtccgaggc
     1801 ggagaggaag gcggttcagg ctacgtgggc ccggctgtat gccaactgcg aggacgtggg
```

Ein *bekanntes* Genom, Gen oder Protein auffinden...



Alternativ...

human myoglobin
Mb
NM_001164048
hankeln t AND myoglobin

Gene identifizieren durch Datenbanksuchen

- Nimm Deine neue, uncharakterisierte Sequenz („query“)
- Vergleiche sie mit allen (!) Sequenzeinträgen der Datenbank

Ein passender ‚Match‘ mit einem bekannten Gen (auf Nukleotidebene) oder Protein (auf Aminosäure-ebene) ist der direkte Beweis, dass in deiner Suchsequenz ein Gen liegt !!!

Sequenzvergleich durch Alignment: die Schlüssel-Technik der Bioinformatik!

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

Nt-Substitution

As-Austausch

Gap bzw. InDel

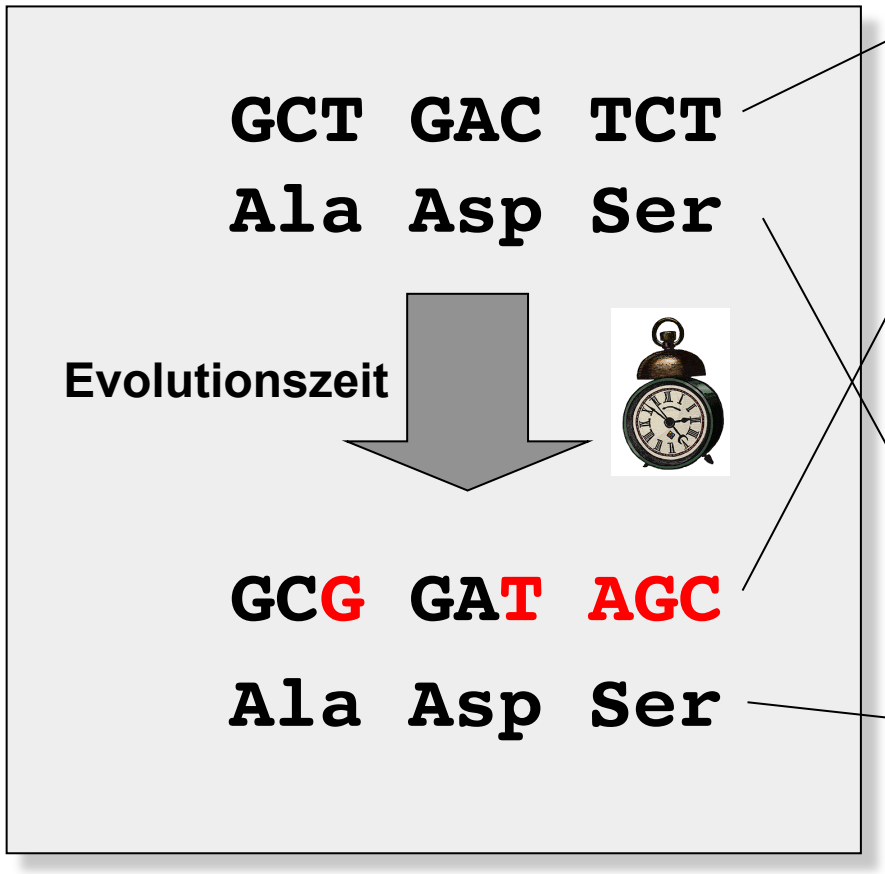
```
Query: 5   EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQY--NCRQFSSPEDCLSSPEFL 62
           + ELIR SW ++ ++ + HG +LF+RLF L+P+LL LF Y  NC     S +DCLSSPEFL
Sbjct: 8   DKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYTTNC---GSTQDCLSSPEFL 64
```

ähnliche As

identische As

Alignments können auf Nukleotid- oder Aminosäure-Ebene erfolgen

Suche ich auf DNA- oder auf Protein-Ebene?



DNA mutiert schnell:
„stille“ Mutationen sind „selektiv neutral“ und häufen sich an

Aminosäuren bleiben lange Zeit gleich („konserviert“):
Selektion auf Funktion!

Suche ich auf DNA- oder auf Protein-Ebene?

GCT GAC TCT

Ala Asp Ser

Evolutionszeit

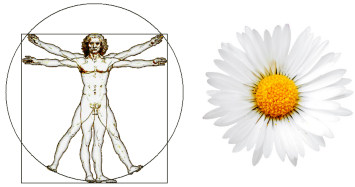
GCG GAT AGC

Ala Asp Ser

Konsequenz:

- Suche auf DNA-Ebene funktioniert gut zwischen nahe verwandten Taxa oder Genen

- Suche auf Aminosäureebene kann auch noch Ähnlichkeiten von entfernt verwandten Sequenzen detektieren



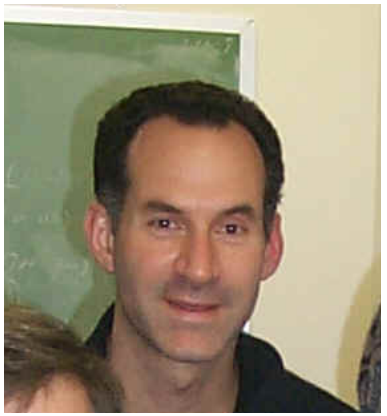
Suche in Sequenzdatenbanken

Populärstes (und schnellstes) Werkzeug:

BLAST
(Altschul et al. 1991, 1997)

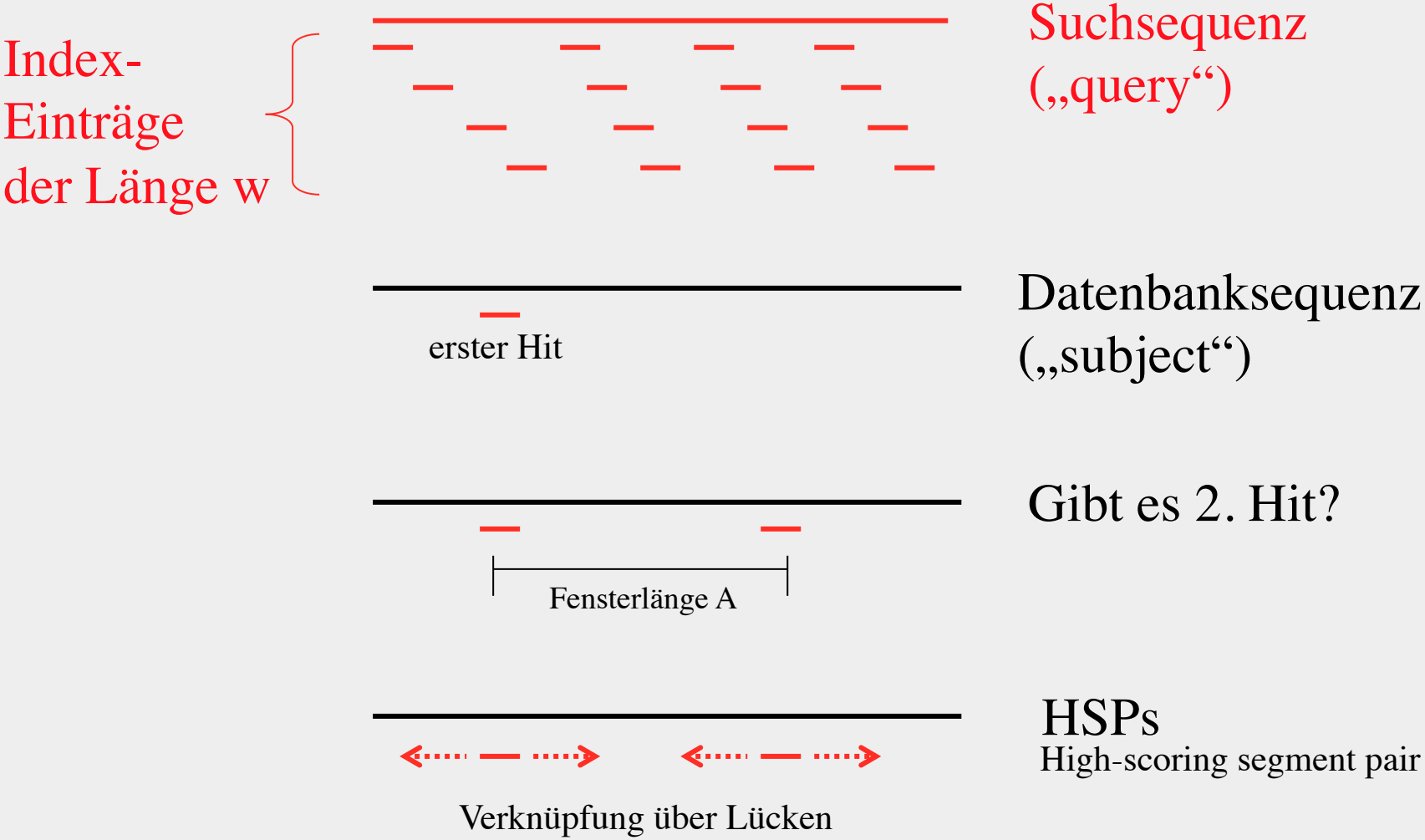
„Basic Local Alignment Search Tool“

Stephen Altschul

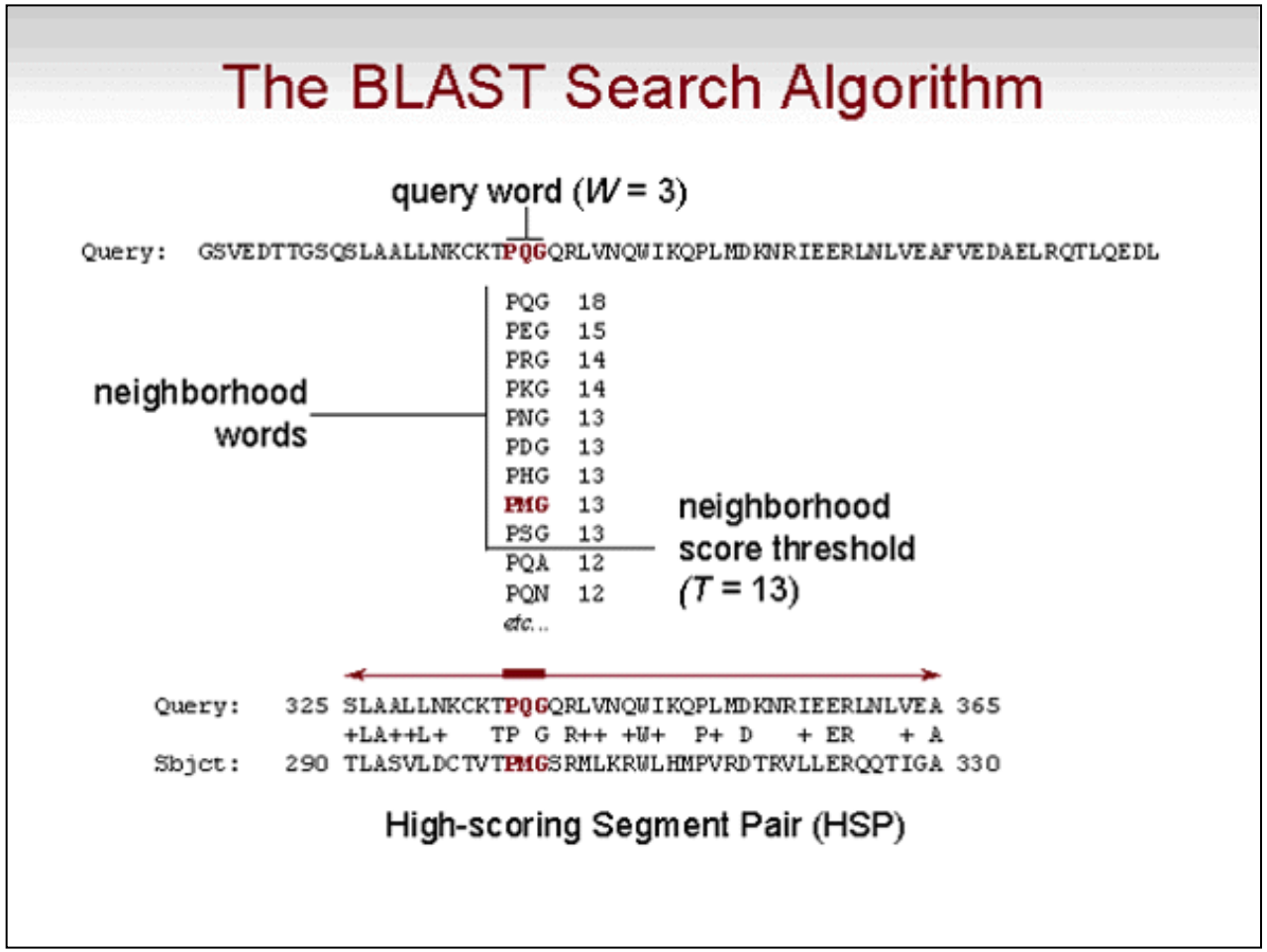


David Lipman

BLAST: teile großes Problem in viele kleine Probleme



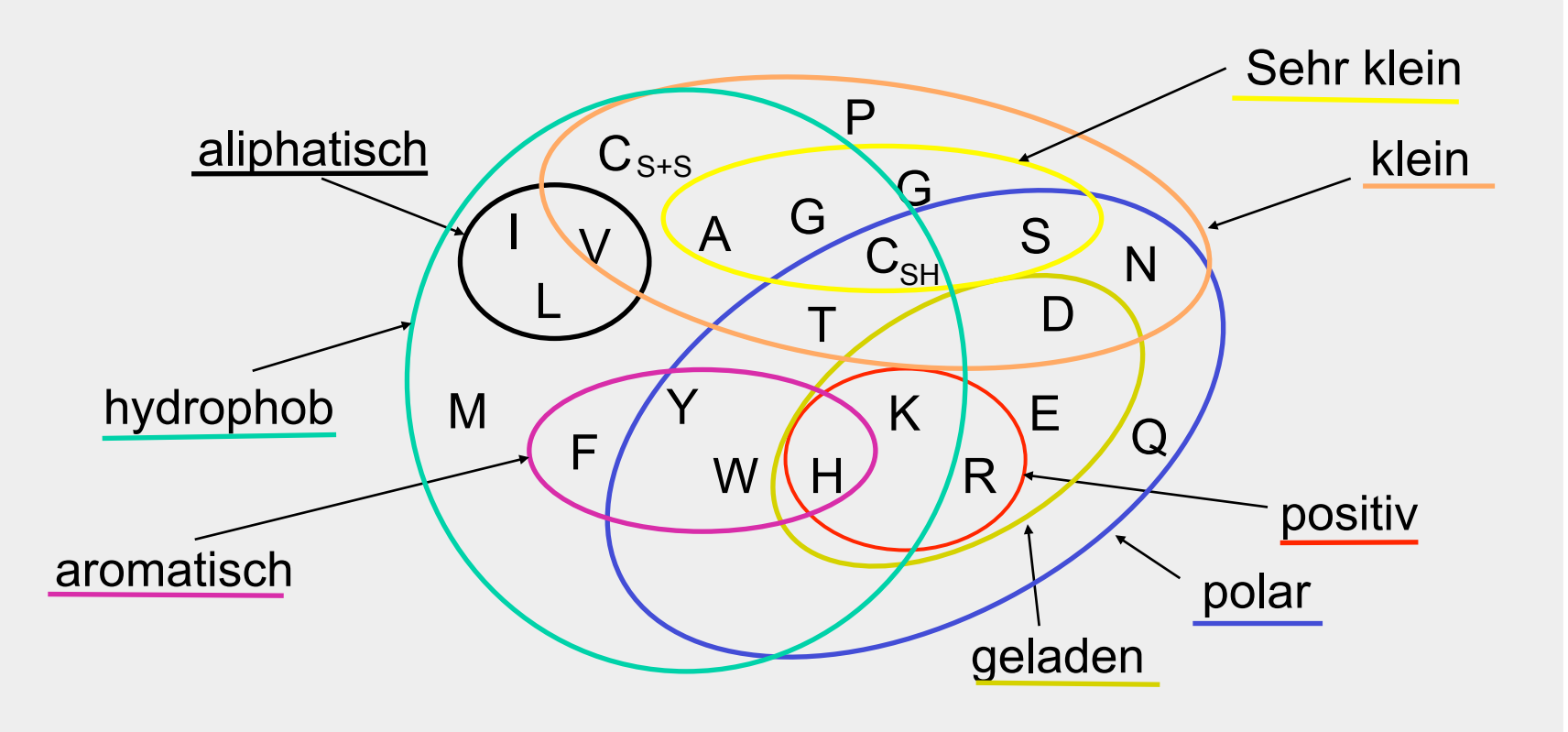
BLAST erkennt bei Proteinen auch biochemische Ähnlichkeiten



(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

Exkurs: Protein-Similarität



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

BLAST

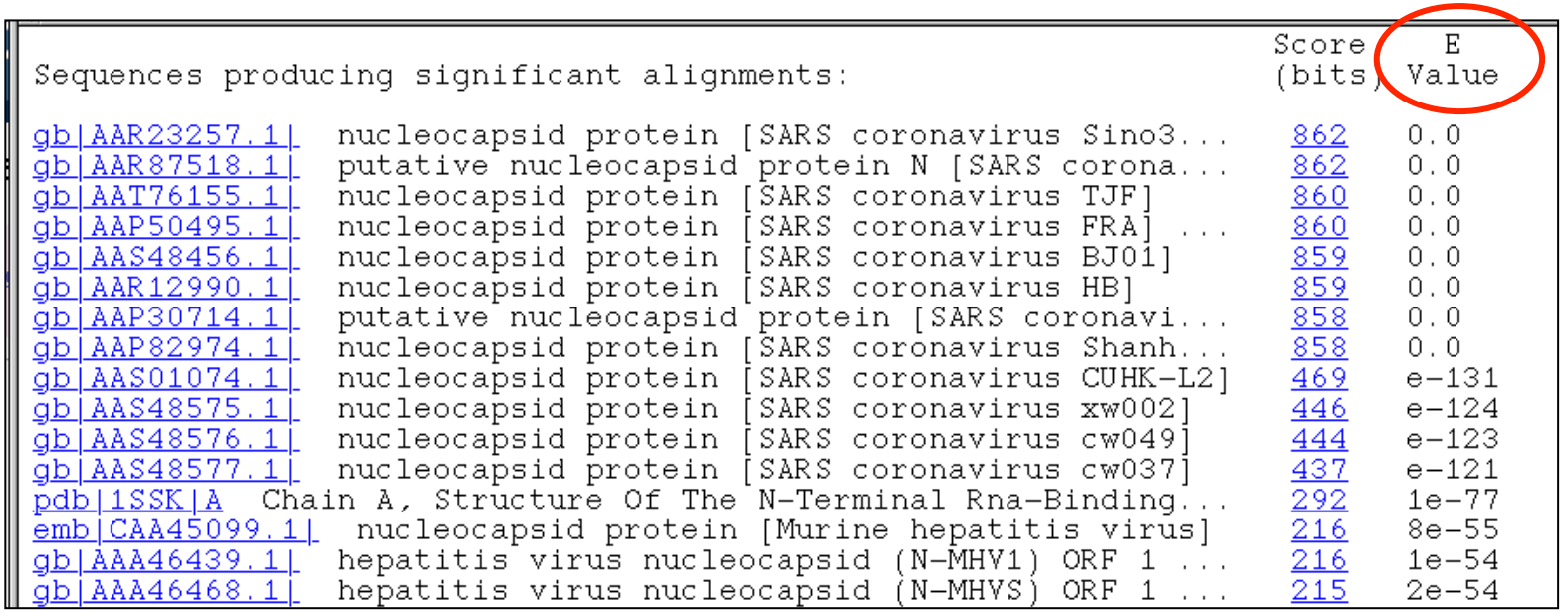
1. Suchsequenz wird in ‚words' der Länge w „zerbrochen"
2. mit Index dieser ‚words' wird Datenbank durchsucht
3. ein „word hit" liegt vor, wenn das ‚word' exakt oder in ähnlicher Form (threshhold-Score >T) erkannt wird

   > 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 bewertet die Signifikanz eines Treffers!



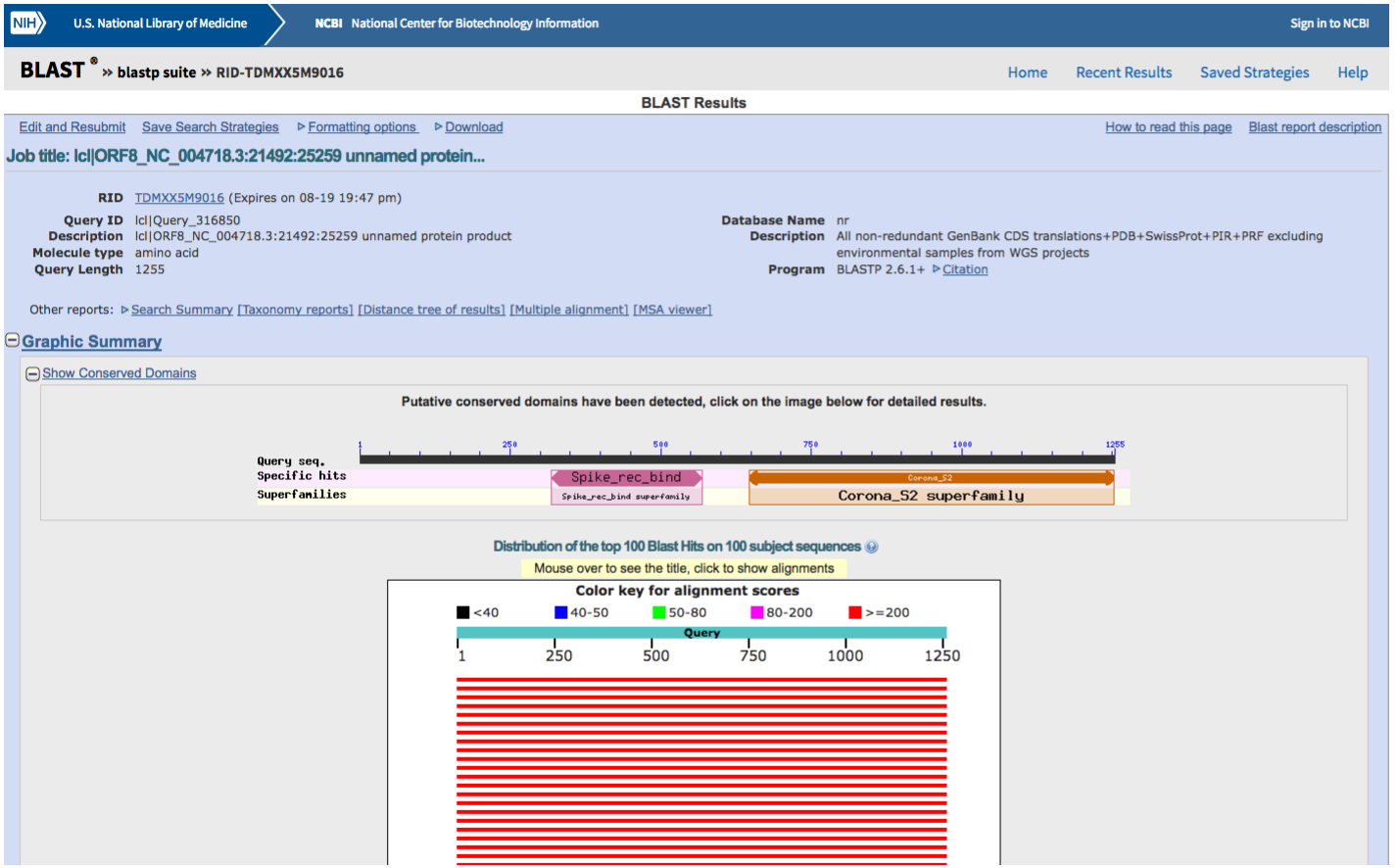
Sequences producing significant alignments:		Score (bits)	E Value
gb|AAR23257.1|	nucleocapsid protein [SARS coronavirus Sino3...	862	0.0
gb|AAR87518.1|	putative nucleocapsid protein N [SARS corona...	862	0.0
gb|AAT76155.1|	nucleocapsid protein [SARS coronavirus TJF]	860	0.0
gb|AAP50495.1|	nucleocapsid 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 Anzahl von Treffern an, die man aus Zufall beim Durchsuchen einer Datenbank der verwendeten Größe erhalten kann.

Practical Exercises 3

- Does SARS ORF1255aa represent a true gene, and which protein does it code for?
- What about the hypothetical Drosophila ORFs and the HMM-predicted Drosophila gene/protein?
- Where is the Drosophila gene located in the fly genome?
- (optional): Don’t mess up the database with rubbish!
- (optional) What‘s in a mystery sequence?

BLASTP-Suche für ORF1255aa des SARS-CoV



BLASTP für ORF1255aa:

die Ergebnisliste

spike glycoprotein [SARS coronavirus PC4-241]	2568	2568	100%	0.0	99%	AAV49723.1
spike glycoprotein [SARS coronavirus civet014]	2566	2566	100%	0.0	99%	AAU04661.1
spike glycoprotein [SARS coronavirus PC4-115]	2566	2566	100%	0.0	99%	AAV49719.1
spike glycoprotein [SARS coronavirus PC4-199]	2566	2566	100%	0.0	99%	AAV49722.1
spike glycoprotein [SARS coronavirus GD03T0013]	2565	2565	100%	0.0	99%	AAS10463.1
spike glycoprotein [SARS coronavirus C013]	2565	2565	100%	0.0	99%	AAV97995.1
spike glycoprotein [SARS coronavirus A021]	2565	2565	100%	0.0	99%	AAV97986.1
spike glycoprotein [SARS coronavirus C029]	2564	2564	100%	0.0	99%	AAV98002.1
spike glycoprotein [SARS coronavirus civet020]	2564	2564	100%	0.0	99%	AAU04664.1
spike glycoprotein [SARS coronavirus A013]	2563	2563	100%	0.0	98%	AAV97985.1
spike glycoprotein [SARS coronavirus PC4-205]	2563	2563	100%	0.0	98%	AAU93319.1
spike glycoprotein [SARS coronavirus B024]	2563	2563	100%	0.0	99%	AAV97990.1
spike glycoprotein [SARS coronavirus B033]	2562	2562	100%	0.0	98%	AAV97992.1
spike glycoprotein [SARS coronavirus civet019]	2561	2561	100%	0.0	98%	AAU04662.1
spike glycoprotein [SARS coronavirus B012]	2560	2560	100%	0.0	98%	AAV97989.1
spike glycoprotein [SARS coronavirus C025]	2560	2560	100%	0.0	98%	AAV98000.1
spike glycoprotein [SARS coronavirus A001]	2560	2560	100%	0.0	98%	AAV97984.1
spike glycoprotein [SARS coronavirus civet010]	2560	2560	100%	0.0	98%	AAU04649.1
spike glycoprotein [SARS coronavirus C028]	2559	2559	100%	0.0	98%	AAV98001.1
spike glycoprotein [SARS coronavirus C018]	2558	2558	100%	0.0	98%	AAV97998.1
spike glycoprotein [Civet SARS CoV 007/2004]	2557	2557	100%	0.0	98%	AAU04646.1
spike glycoprotein [SARS coronavirus A022]	2556	2556	100%	0.0	98%	AAV91631.1
spike protein [SARS-like coronavirus WIV16]	2523	2523	100%	0.0	97%	ALK02457.1
Chain A, Sars-cov Spike Glycoprotein	2472	2472	95%	0.0	99%	5WRG_A
Chain A, Prefusion Structure Of Sars-cov Spike Glycop	2438	2438	94%	0.0	99%	5X58_A
spike protein [Bat SARS-like coronavirus WIV1]	2394	2394	99%	0.0	92%	AGZ48828.1
spike protein [Bat SARS-like coronavirus Rs3367]	2393	2393	99%	0.0	92%	AGZ48818.1
spike protein [Bat SARS-like coronavirus RsSHC014]	2334	2334	99%	0.0	90%	AGZ48806.1
spike protein [Rhinolophus affinis coronavirus]	2327	2327	99%	0.0	90%	AHX37569.1
spike protein [Rhinolophus affinis coronavirus]	2318	2318	99%	0.0	90%	AHX37558.1
spike glycoprotein [recombinant coronavirus]	2223	2223	100%	0.0	84%	ACJ60703.1
spike glycoprotein [Bat coronavirus]	2086	2086	100%	0.0	80%	ARI44799.1
spike glycoprotein [BtRs-BetaCoV/YN2013]	2080	2080	100%	0.0	80%	AIA62330.1
spike protein [Bat coronavirus Cp/Yunnan2011]	2069	2069	100%	0.0	79%	AGC74176.1

...es gibt heute natürlich viele SARS-Datenbankeinträge

Im Jahre 2003 gab es die natürlich noch nicht!

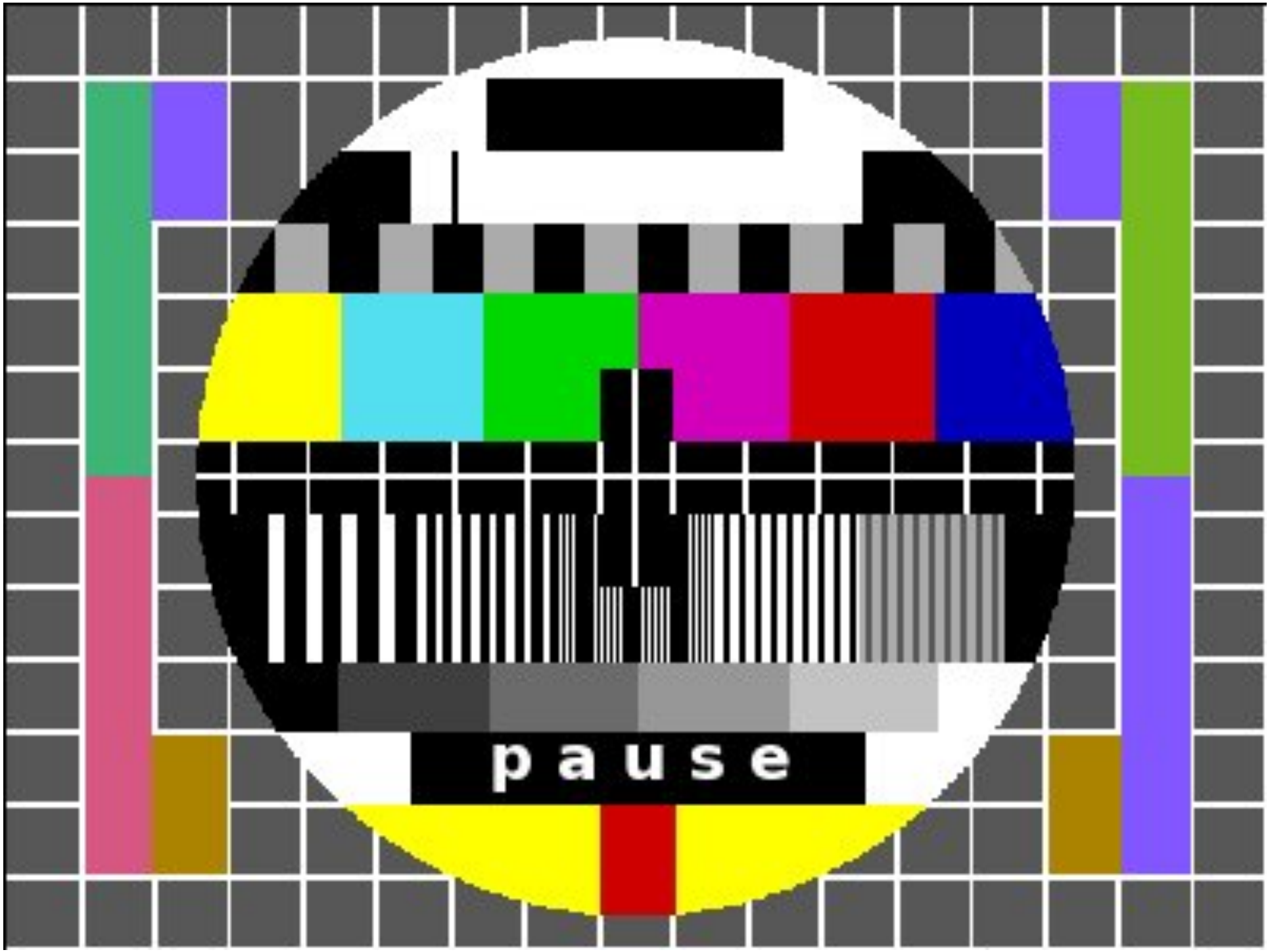
Diese Treffer würden erstmals zeigen: unsere Sequenz ist ein **Nukleocapsid-Protein** aus **Coronaviren!**

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 durch DB-Suche


pause