MSc Biologie Modul 7A:

„Genomforschung und Bioinformatik“

T. Hankeln (AG Molekulargenetik & Genomanalyse)
mit A. Bicker, D. André, L. Hellmann, C. Osterhof, A. Prothmann,
Michel Seiwert, Benjamin Rieger

& Holger Herlyn (Anthropologie) sowie Julian König (IMB)

14+X Tage, ganztägig, 12.11.18 - 30.11.18
Seminar inklusive
Seminarraum Genetik, J. J. Becherweg 32, EG
Beginn Mo 12.11.18, 9.00 Uhr

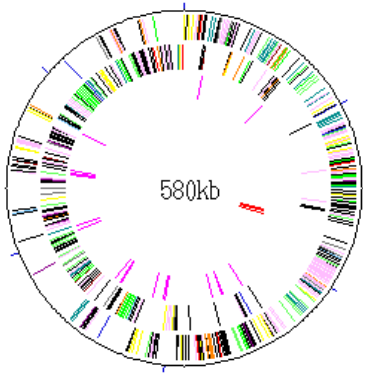
- Anwendung von Literatur- und Sequenzdatenbanken
- Sequenzierprojekte (RNA, DNA)
- Genvorhersage und komparative Genomik
- Phylogenetische Rekonstruktion
- Detektion adaptiver Evolution
- Next-Gen Sequencing (RNA-Seq, Exon-Seq etc)

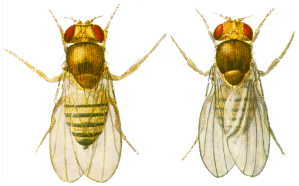
WS2018/2019

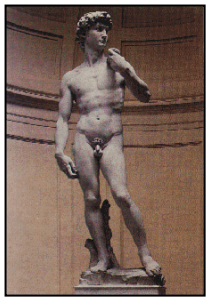
„Genomforschung und Sequenzanalyse

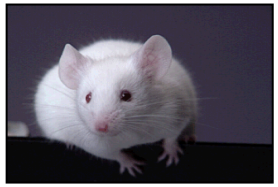
- Einführung in Methoden der Bioinformatik- “

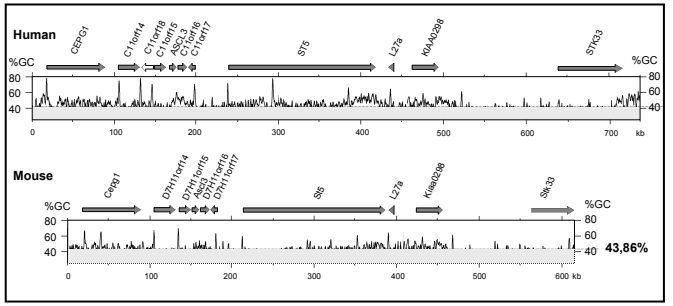
Thomas Hankeln

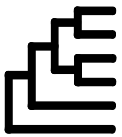


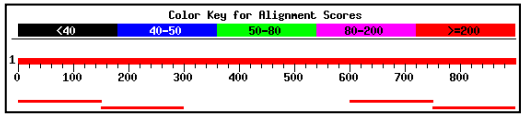




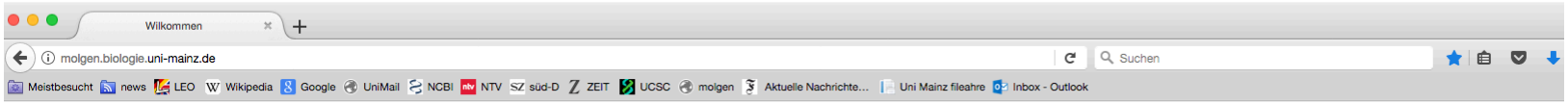








pdfs http://molgen.biologie.uni-mainz.de



Institut für Organismische und Molekulare Evolutionsbiologie
Fachbereich Biologie
Johannes Gutenberg Universität Mainz

AG Molekulargenetik und Genomanalyse

Die AG ist im Jahre 2017 aus dem ursprünglich 1994 gegründeten Institut für Molekulargenetik, gentechnologische Sicherheitsforschung und Beratung (IMSB) hervorgegangen.
Ihre gegenwärtigen **Forschungsaktivitäten** umfassen:

- Funktionsanalyse und molekulare Evolution der Globin-Genfamilie
- Genomanalyse und Genregulationsmechanismen bei hypoxietoleranten, krebs- und alterungsresistenten Nagetieren
- Genomsequenzierung und molekulare Phylogenomik weitgehend unerforschter Tierstämme
- Entwicklung von Methoden für die Artendiagnostik in Nahrungsmitteln durch Next-Generation-Sequencing
- Aufklärung der Bedeutung von repetitiver „Junk DNA" in der Evolution von Genomen und Chromosomen bei Insekten

Informationen und Material zu Lehrveranstaltungen

• **MSc-Modul 7A/B „Genomforschung und Sequenzanalyse" (AG Hankeln)**
Nächster Termin: Nov./Dez. 2017 (Anmeldung per JOGUstine)

PDFs zur Modul 7A-Vorlesung "**Genomforschung und Sequenzanalyse" (T. Hankeln)**
(1)(2)(3)(4) (5)(6)(7)(8)(9)(10)(11) (12) Seq-Testfile zu VL1
Ergänzende PDFs zum MSc-Modul 7A (F1-Praktikum „Genomforschung und Sequenzanalyse")
(1)(2)(3)(4)(5)(6)
PDFs zur ergänzenden Vorlesung **„Molekulare Evolution von Genen und Genomen" (T. Hankeln)**
(1) (2) (3) (4) (5) (6) (7) (8)

• **MSc-Modul 8A/B „Genexpressionsanalyse in der Entwicklungsgenetik" (AG Hankeln / Dr. C. Berger)**
Nächster Termin: Nov./Dez. 2017 (Anmeldung per JOGUstine)

• **BSc-Modul 13/14 „Analyse von Eukaryotengenen" (AG Hankeln).**
Nächster Termin: 13. – 24. Februar 2017, ganztags (Anmeldung per JOGUstine)

• **BSc-Modul 13/14 „Molekulargenetik der Eukaryoten" (Hankeln, Kraemer, Rapp, Bicker).**
Nächster Termin: 08. Mai – 02. Juni 2017, Durchführung halbtags (Anmeldung per JOGUstine)

• **BSc-Modul 8 Grundvorlesung & Grundpraktikum „Allgemeine und Molekulare Genetik"**
PDFs zur Grundvorlesung: DNA Chromatin Replikation Genorganisation/Trankription1 Transkription2 Transkription3 Gentechnologie

bzw. im READER-Verzeichnis

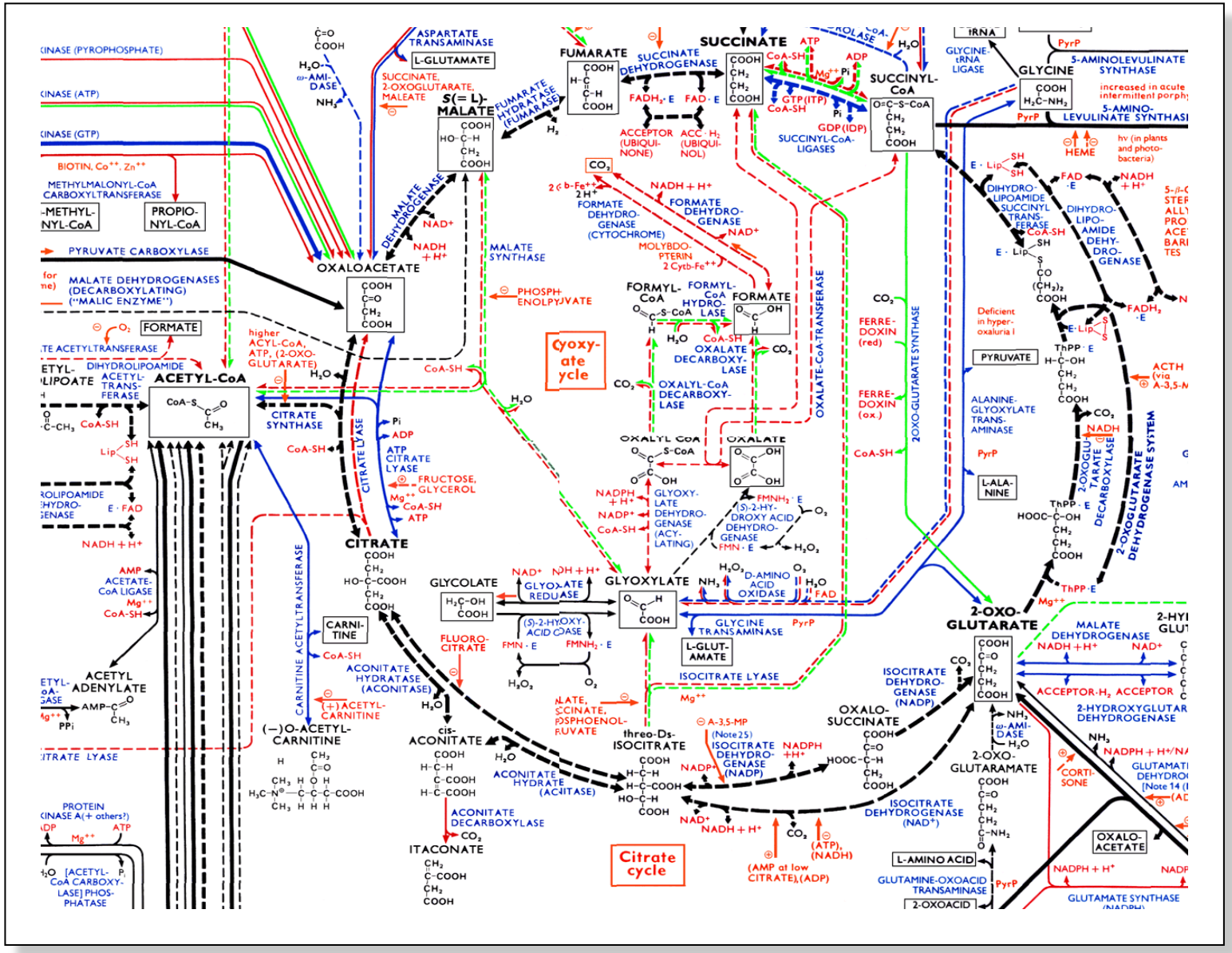
Termine und voraussichtliche Themen:

22.10. Mo	Gene, Genome, Sequenzierung: molekularbiolog. Grundlagen
25.10. Do	Strategien der Gen-Suche, Datenbanken und Sequenzformate
26.10. Fr	Sequenzvergleiche („alignment“)
29.10. Mo	Datenbank-Suchen
30.10. Di	Multiples Alignment
31.10. Mi	Phylogenetische Rekonstruktion 1
06.11. Di	Phylogenetische Rekonstruktion 2
07.11. Mi	Methoden der Genomsequenzierung
08.11. Do	Genvorhersage und -Annotation

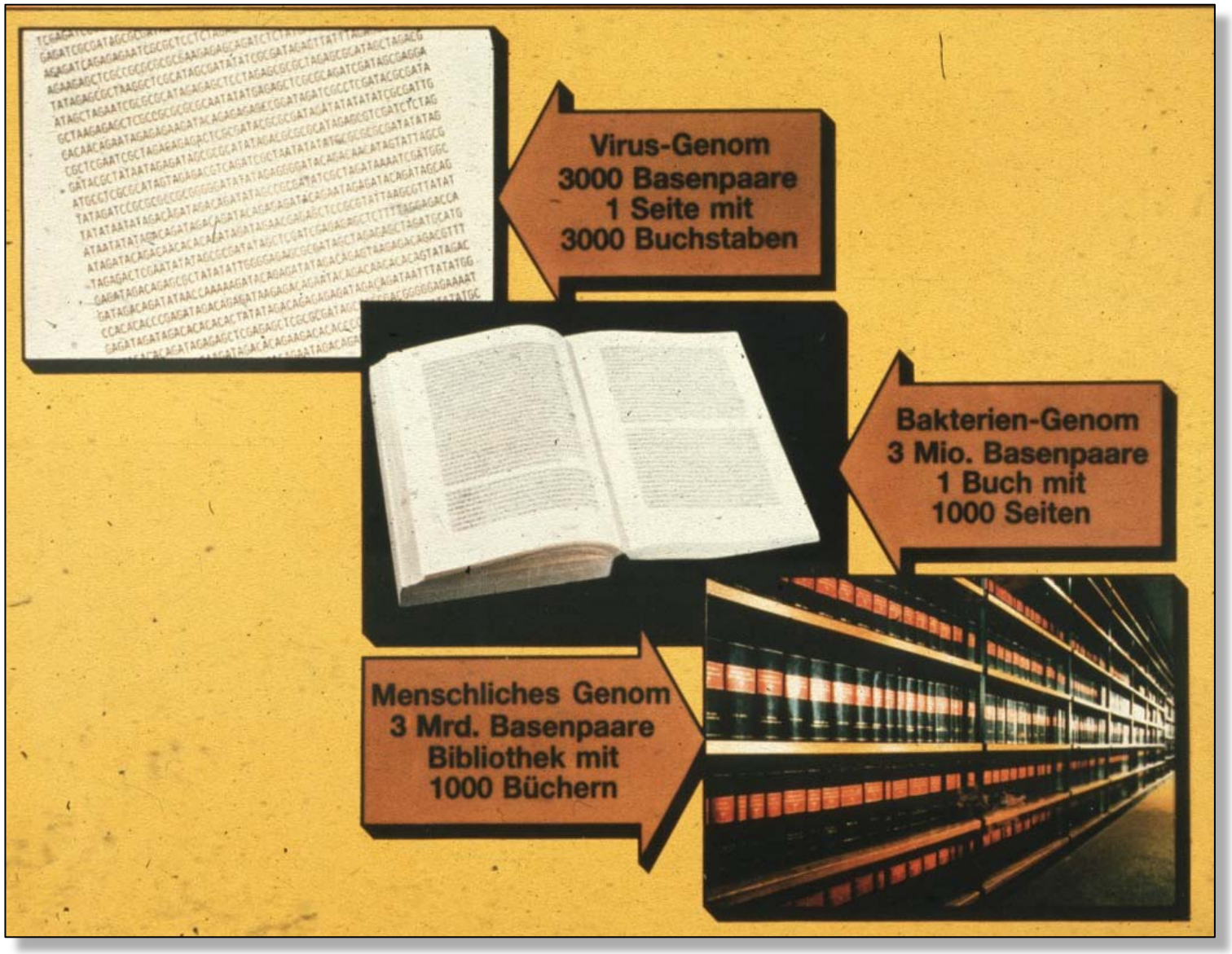
Weitere VL-Teile im Rahmen des Kurses:

Verarbeitung großer Sequenzdatensätze (NGS), Funktionelle Genomforschung (SNPs, Mikroarrays, RNA-Seq, Chip-Seq etc.)

Warum Informatik in der Biologie?



Warum Informatik in der Biologie?



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)

Bioinformatik

- die etwas engere Sichtweise-

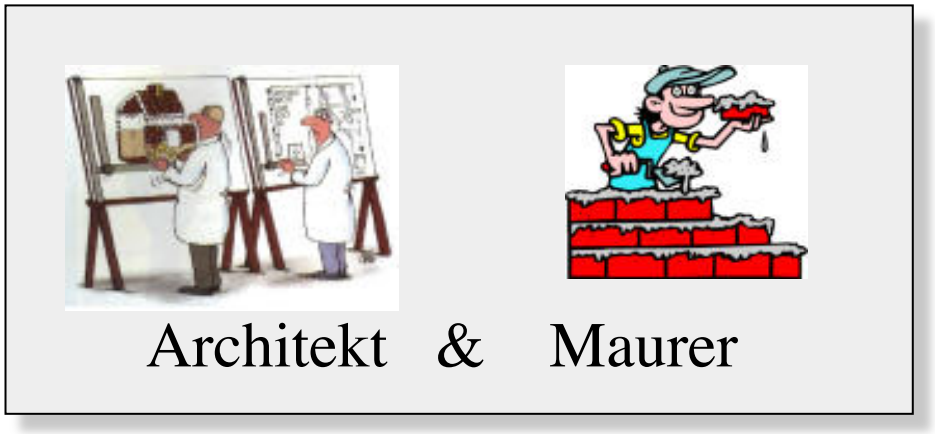
S. O‘ Brien:
(Neapel 2002)

Deposition
Curation
Accessing
Manipulation
Interpretation

} of linear genetic information

also: Entwickeln und Benutzen von Sequenz-Datenbanken, Such-Werkzeugen und Tools zur Datenauswertung

Muss ich programmieren können?



Nützlich sind:

> Web sites basteln
> PERL als Programmiersprache
> UNIX/Linux als Betriebssysteme
> SQL als Datenbankformat

Literatur

Zvelebil M, Baum JO, Understanding bioinformatics. Garland Science 2008 (gute Mischung...)

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

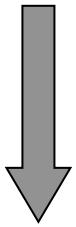
- Symptome: ähnlich Lungenentzündung
- 114 Tage-Epidemie (2002/2003)
- 8098 Erkrankungen, 774 Tote
- 29 Länder betroffen
- eine paralysierte asiatische Volkswirtschaft…

Das Szenario ...ein neues tödliches Virus!

- Labor: Isolierung der Erbsubstanz, Sequenzierung

- Computer: Ähnlichkeit zu bekannten Genen? (Datenbanksuchen)

  Verwandschaft? (Phylogenetische Rekonstruktion)



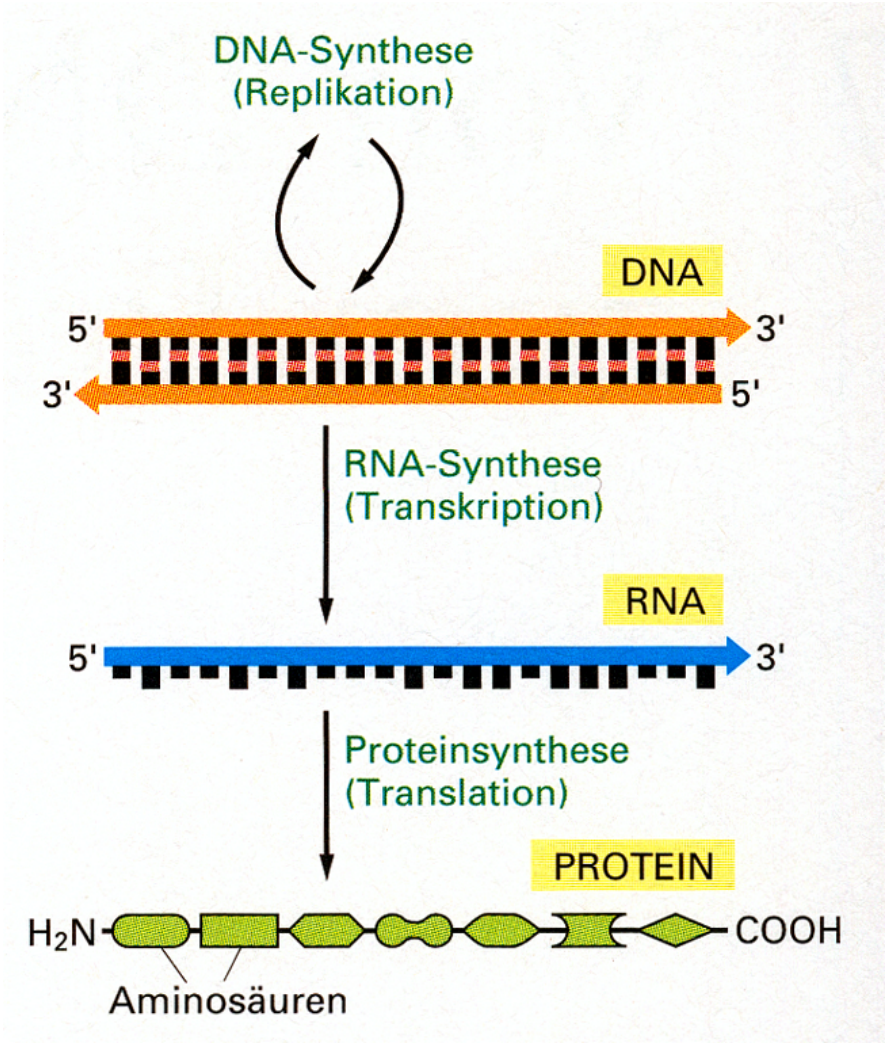
  De-Kodierung der Virusproteine (Genvorhersage)

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

  Wirkstoff-Design

- Labor: Wirkstoff-Test

DNA als Speicher der genetischen Information



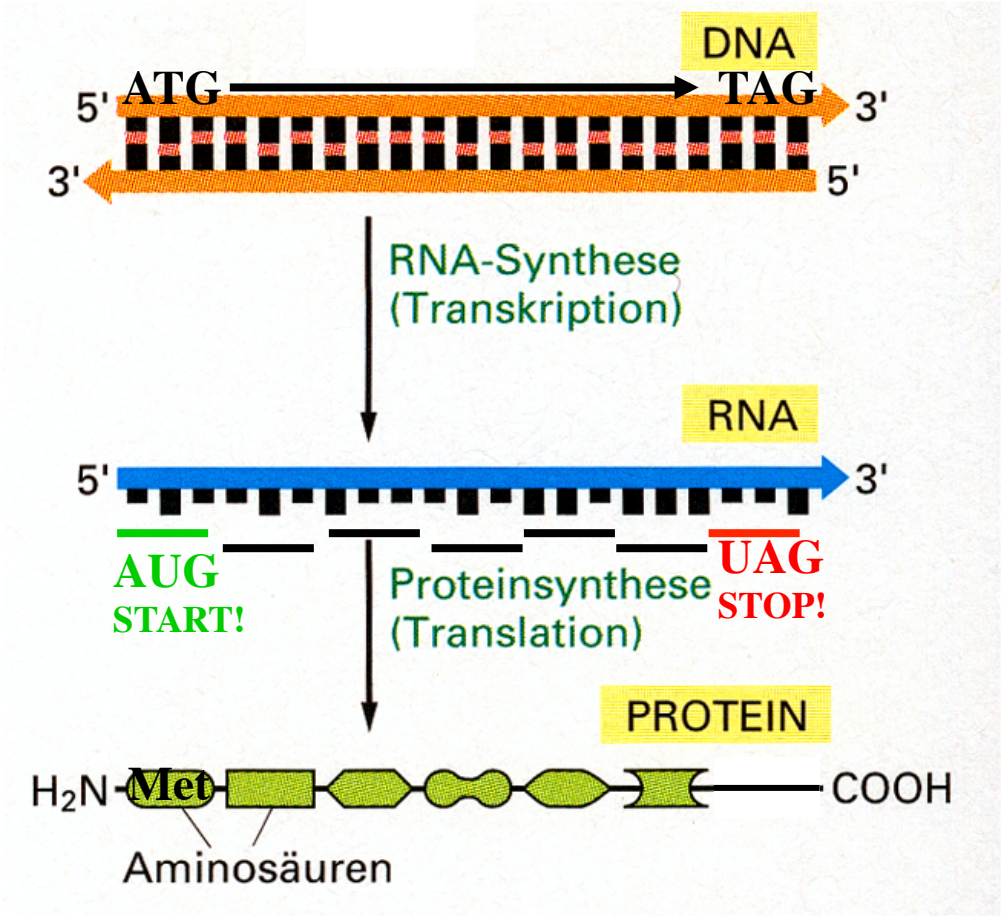
Informationsspeicher

Informationsabschrift

Q:

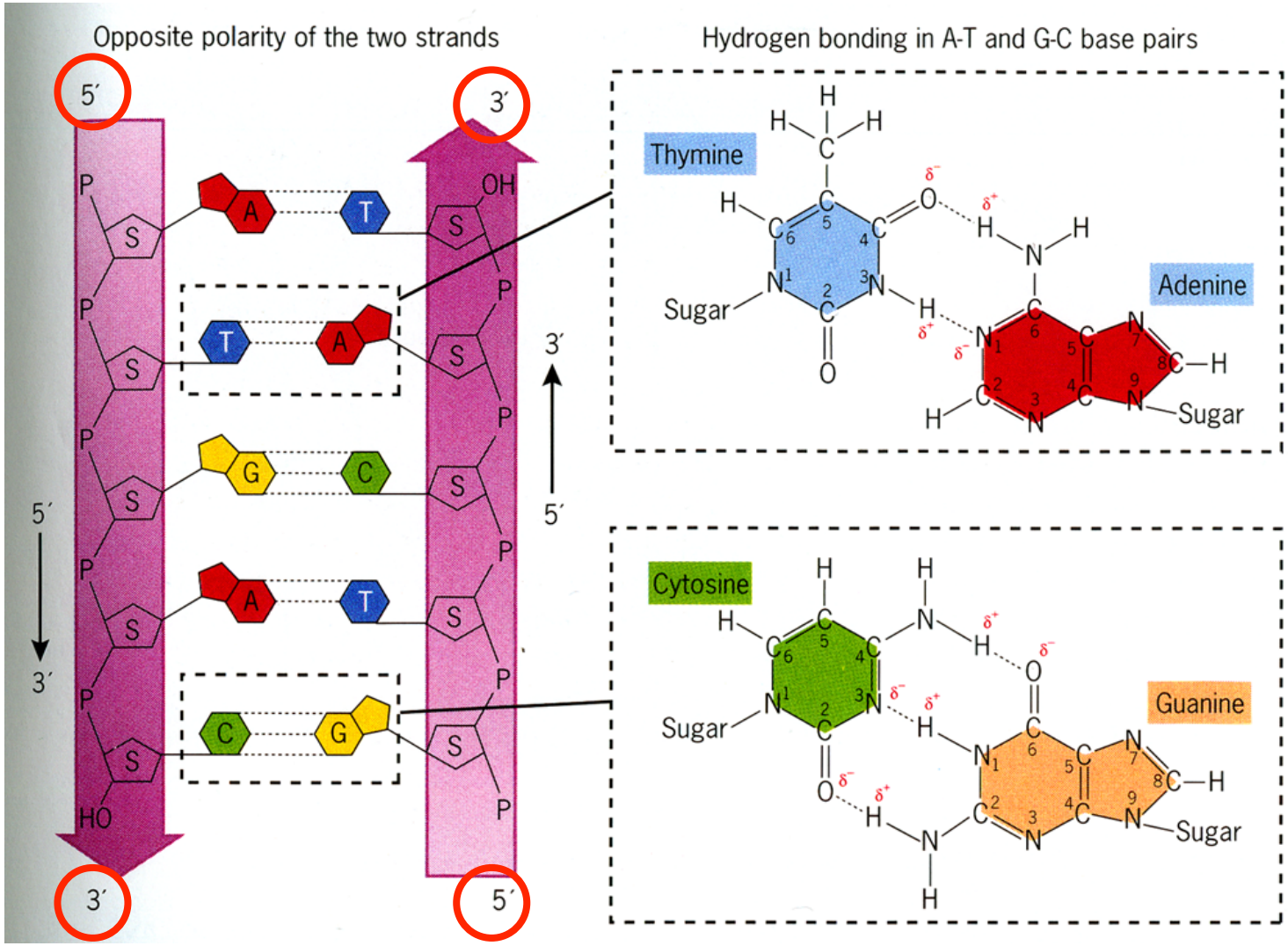
Wie erkenne ich (oder der Computer),
dass ein DNA-Abschnitt
ein Protein-kodierendes Gen enthält?

Wie erkenne ich ein proteinkodierendes Gen?



ORF
= offener Leserahmen

Der DNA-Doppelstrang



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‘

„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

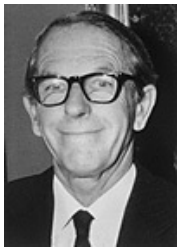
Phi-X 174 sequenced, *Nature* 1977

Methoden der DNA-Sequenzierung

1977 (*old school*)

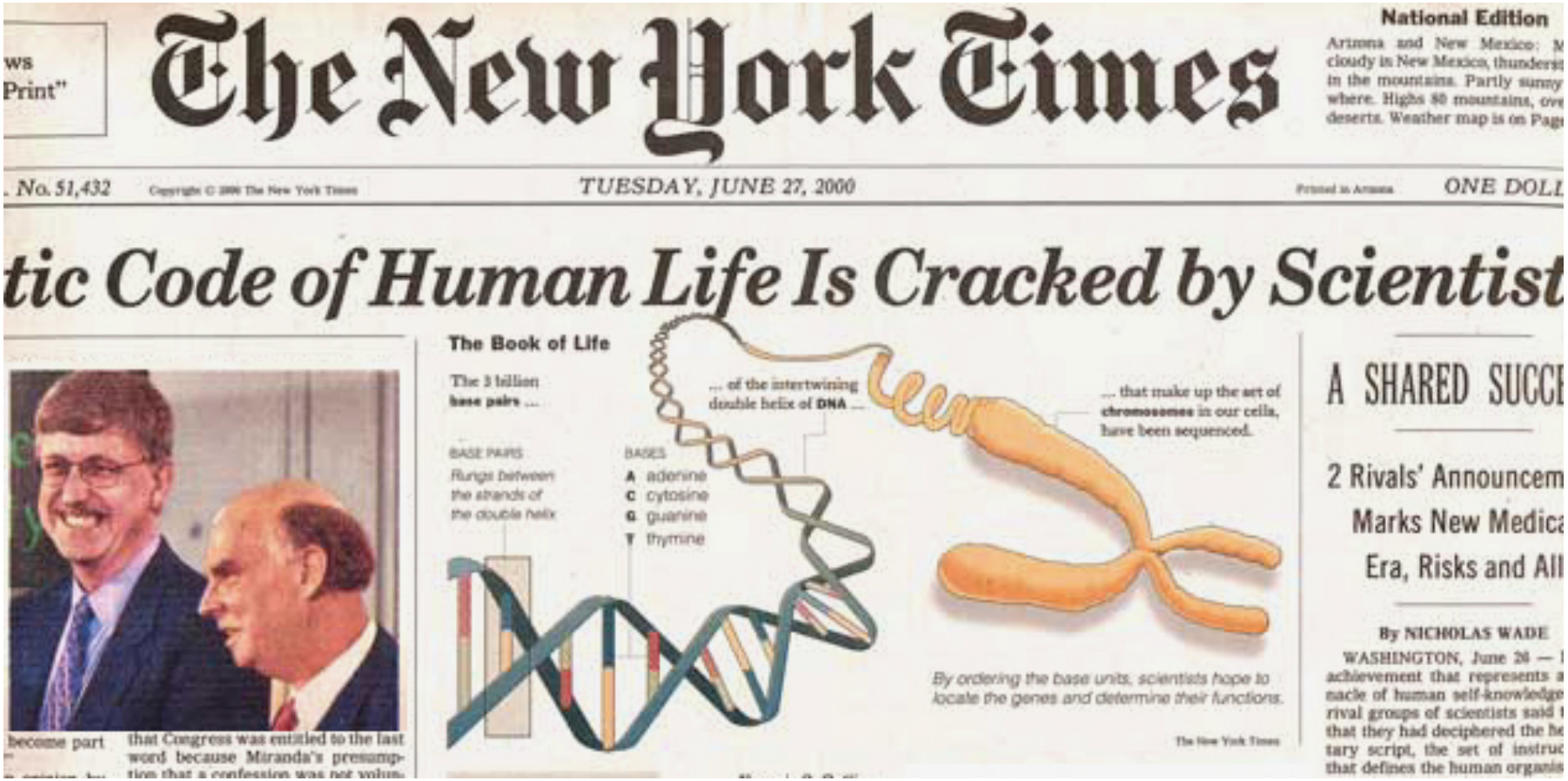
- chemische Sequenzierung (**Maxam & Gilbert**)
- enzymatische Sequenzierung (**Sanger**)

synonym: > Kettenabbruch-Sequenzierung
> Didesoxy-Sequenzierung



1918-2013

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-

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

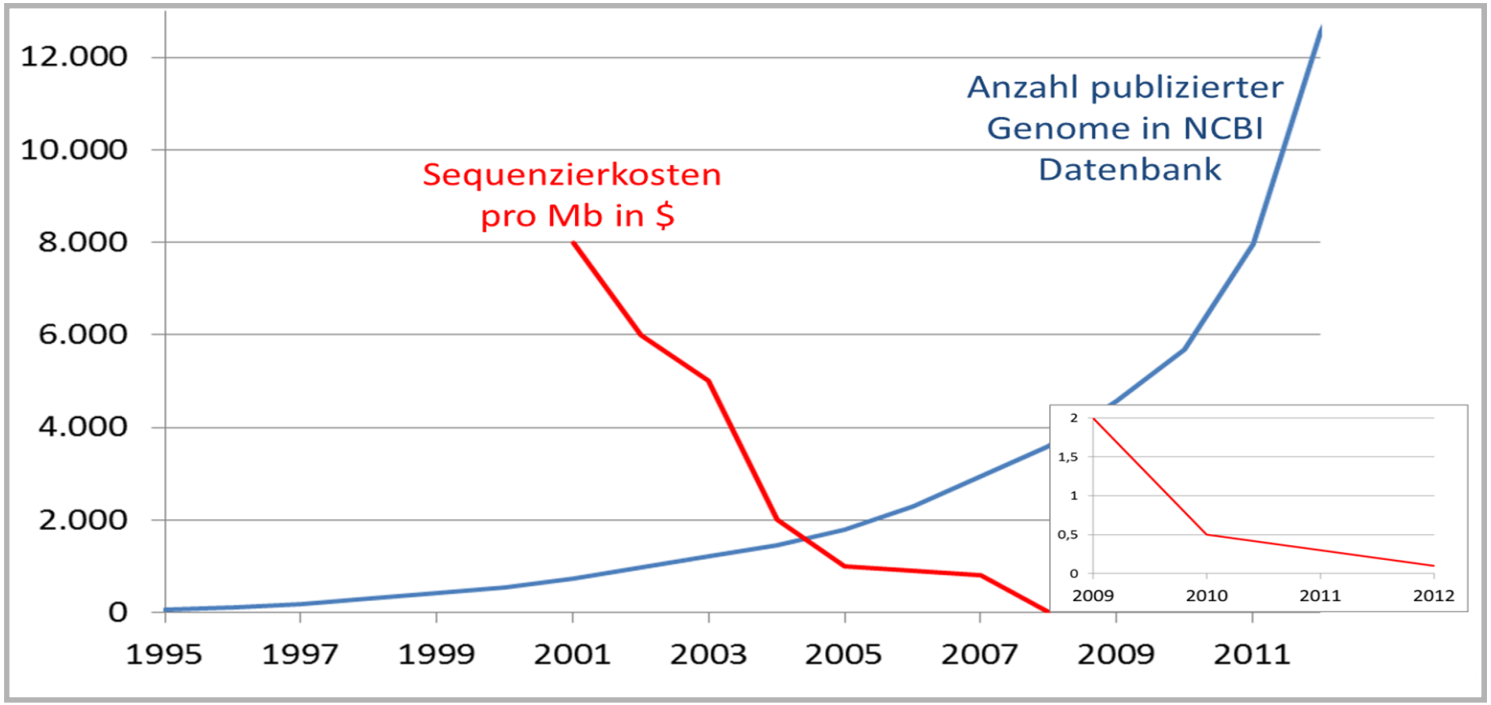
A SHARED SUCCE

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

By NICHOLAS WADE

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

Next-Generation Sequencing



- Pyrosequencing (454), ion-based sequencing (Ion Torrent)
- **seq-by-synthesis with reversible terminators (Illumina)**
- single molecule sequencing (PacBio, Nanopore)

...werden später erklärt!

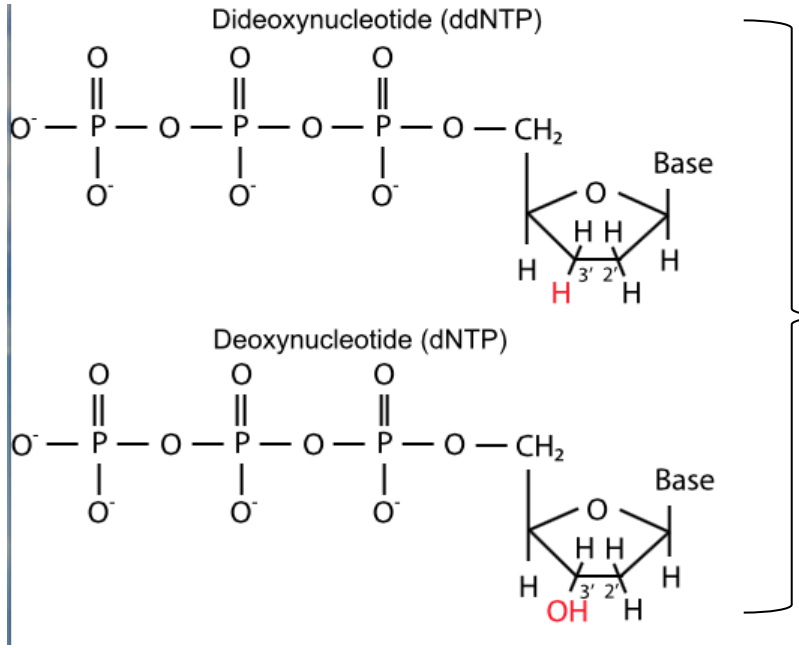
Das Sanger-Verfahren

- Replikation *in vitro*! Zutaten?

  Matrize (einzelsträngig),
  Primer, DNA-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}

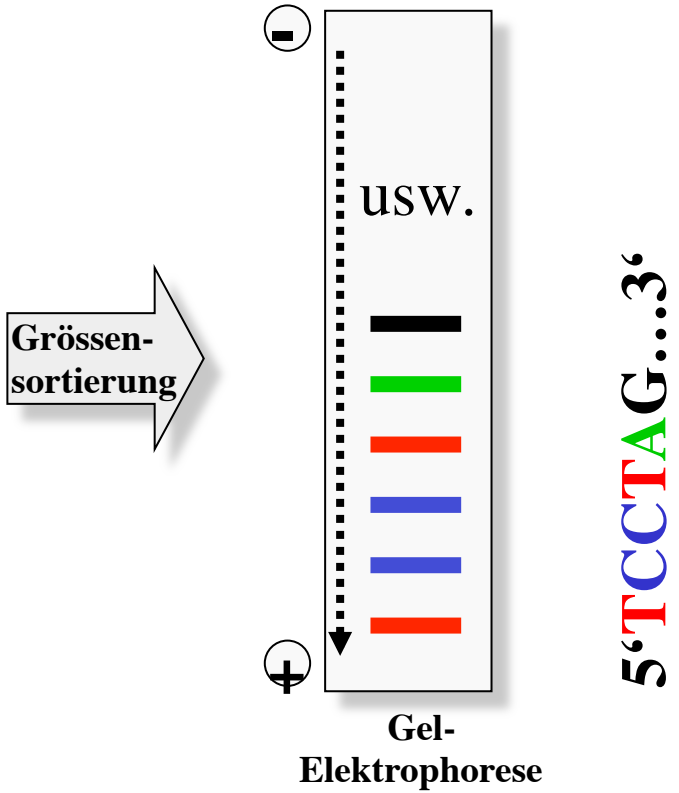
5‘-CTAGGACTGTAC **TC**^{Stop}

5‘-CTAGGACTGTAC **TCC**^{Stop}

5‘-CTAGGACTGTAC **TCCT**^{Stop}

5‘-CTAGGACTGTAC **TCCTA**^{Stop}

5‘-CTAGGACTGTAC **TCCTAG**^{Stop}



Das Sanger-Verfahren

Q: in welcher Richtung wird eine DNA mit dem Sanger-Verfahren entschlüsselt?

Wer bestimmt diese Richtung?

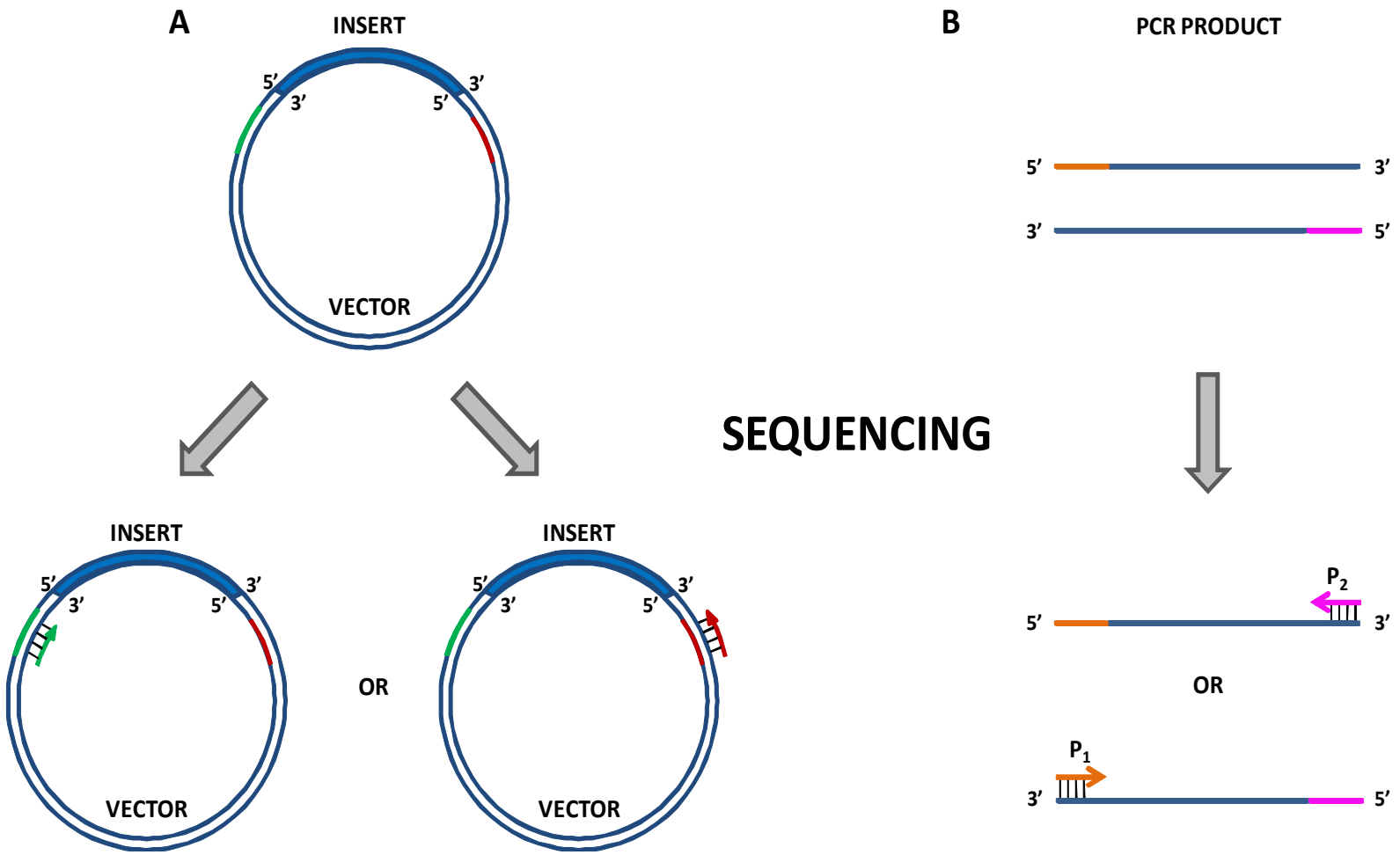
Eine Sanger-Sequenzierungsreaktion wird immer in 5‘ > 3‘ -Richtung (Polymerase!) gelesen!

(egal, welcher der beiden Stränge gerade sequenziert wird)

Sanger-Chromatogramm

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

Welche Matrizen-Moleküle können wir so sequenzieren?



„Doppelsträngige“ Sequenzierung!!

„WATSON“

„CRICK“

5‘ A G T A C G 3‘

„Forward read“

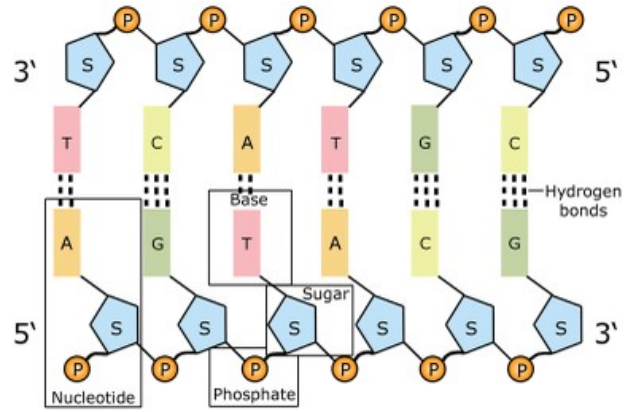


Image adapted from: National Human Genome Research Institute.

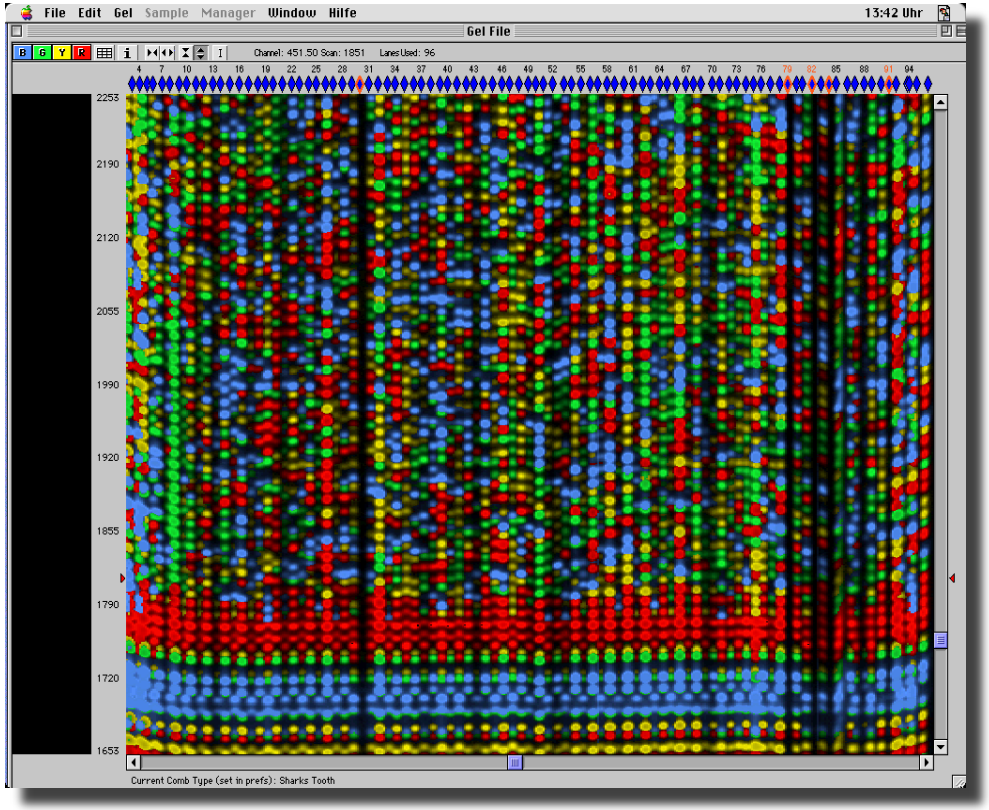
3‘ T C A T G C 5‘

„Reverse read“

Wir überprüfen also:
Passen die beiden Sequenzen fehlerlos zueinander?

DNA-Sequenzierung

vor etwa 20 Jahren



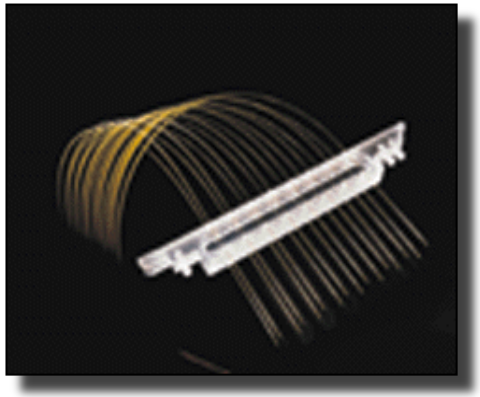
- Gerät: ABI 377
- konventionelles Gel (0,4 mm dick)
- Problem: „Tracking“ der Spuren bei der Auswertung durch Computer

96 Spuren x 600 Basen = ca. 60 000 Basen in ca. 12 Std

Durchsatz limitiert durch zu hohe Hitze bei hohen Feldstärken (>50V/cm) in 0,4 mm Gelen

Hochdurchsatz-DNA-Sequenzierung vor 10 Jahren

Applied Biosystems 3730*xl* DNA Analyzer.



Kapillar-Elektrophorese

- kein tracking-Problem!!!
- mehr Oberfläche/Vol.
  > besserer Hitzeabtransport
- höhere Feldstärken möglich
  > ein Run ca. 2 Std.
- Run bei ca. 70°C minimiert Rückfaltungen der Sequenzierprodukte („Kompressionen“)
- „lineares“ Polyacrylamid als Matrix ist erneuerbar in Kapillaren

„Base calling“

1. Idealisierte Peak-Vorhersage:
   - ausgehend von gleichmäßig angeordneten Peak-Regionen
   - werden beidseitig idealisierte Peak-Positionen vorhergesagt
2. beobachtete Peaks werden identifiziert
3. Anpassen von beobachteten an die vorhergesagten Peaks
   - > Weglassen oder Splitten von Peaks
   - > Liste von „matched“ Peaks ergibt Sequenz
4. „unpassende“ Peaks werden überprüft und u.U. eingepasst

PHRED-Base caller:
Ewing et al. (1998) Genome Res. 8, 175-185

Der Computer liefert parallel einen File mit der „reinen“ DNA-Sequenz (ASCII-Format; „name.seq“), sowie den Chromatogramm-File („name.abd“ oder „name.scf“ /für standard chromatogram format)

Qualitätsbewertung in Chromatogrammem

Ewing and Green (1998) Genome Res. 8, 186-194

$$\text{Phred-Wert } q = -10 \times \log_{10}(p)$$

p = Irrtums-Wahrscheinlichkeit

Phred-Wert 20 > error rate 0,01

Phred-Wert 30 > error rate 0,001

gute Qualität: mind. Phred 20

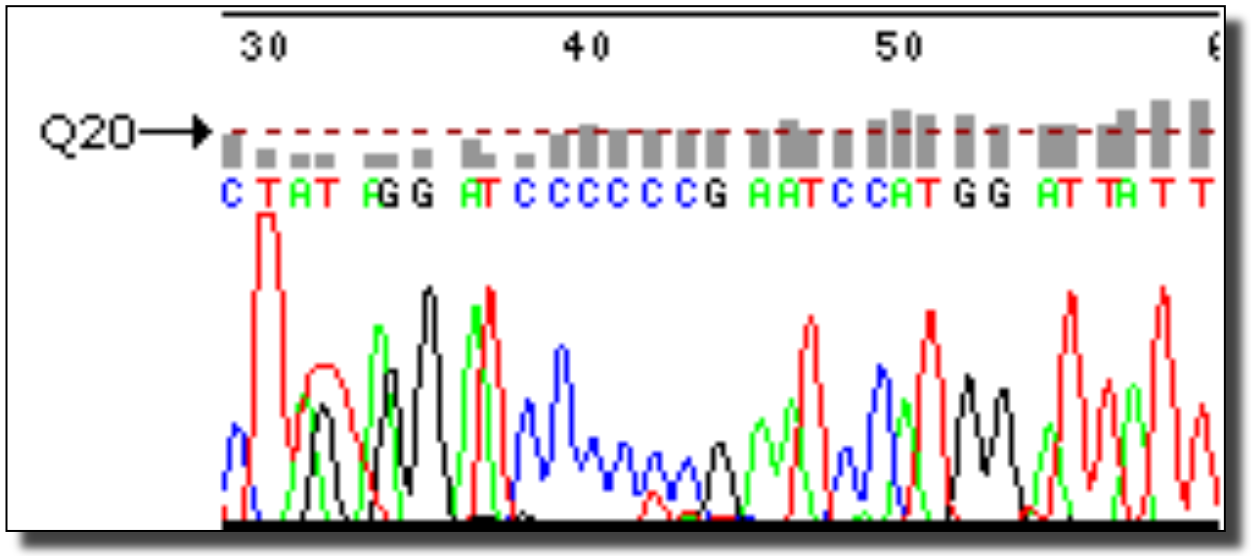


Table 2.1. *Base–nucleic acid codes*

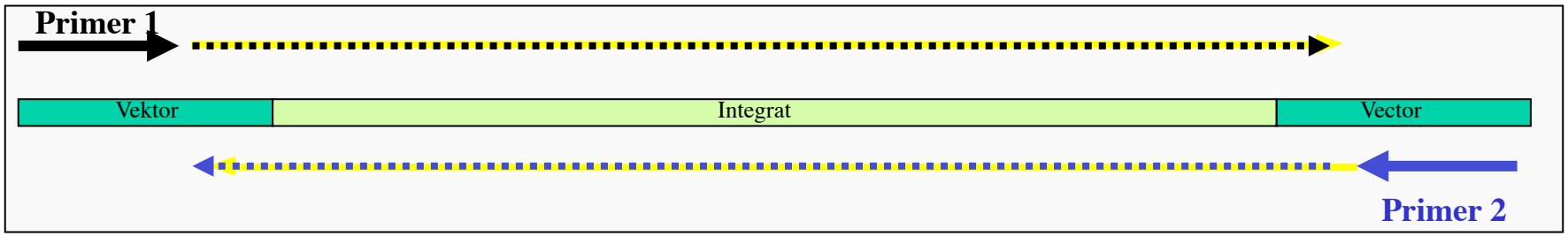
Symbol	Meaning	Explanation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	A or G	puRine
Y	C or T	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	C or G	Strong interactions 3 h bonds
W	A or T	Weak interactions 2 h bonds
H	A, C or T not G	H follows G in alphabet
B	C, G or T not A	B follows A in alphabet
V	A, C or G not T (not U)	V follows U in alphabet
D	A, G or T not C	D follows C in alphabet
N	A,C,G or T	Any base

Adapted from NC-IUB (1984).

Der IUB-*Ambiguity-Code* für DNA

Sequenzierung „kurzer“ DNA-Fragmente

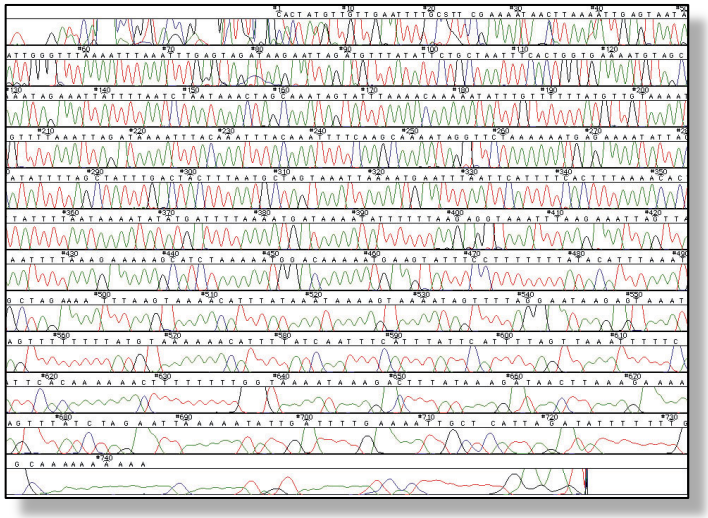
Bei Integrat-Längen bis etwa 1000 bp ist es möglich, mit zwei Sequenzierungsreaktionen die vollständige Basenabfolge auf beiden Strängen zu ermitteln.



„doppelsträngige Sequenzierung“ = Gold-Standard

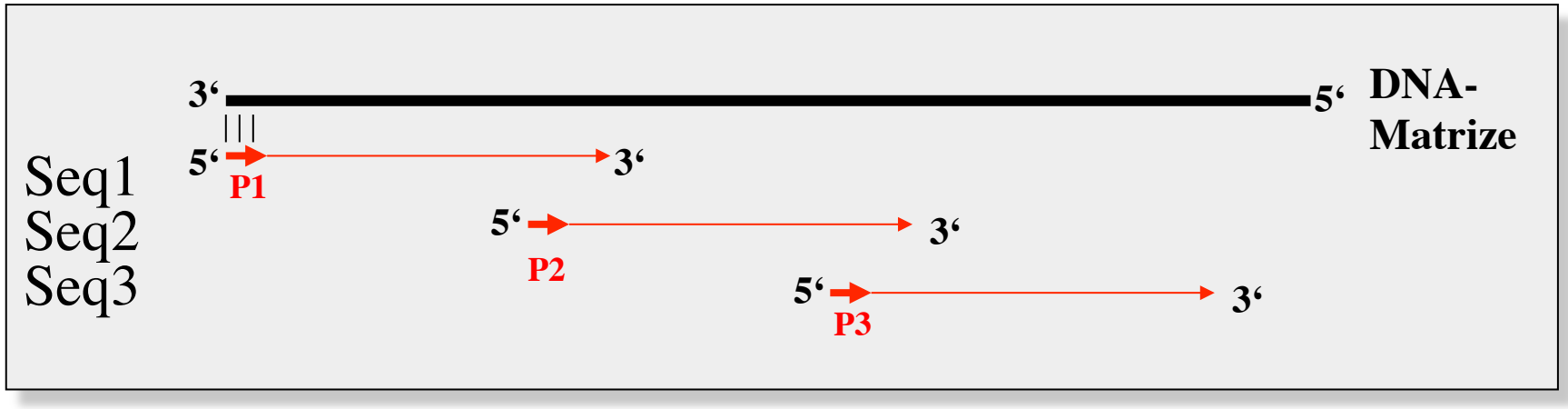
Sequenzierungsstrategien sind erforderlich!

Leselänge (Sanger): ca. 1000 Bp
Leselänge (Illumina): 50-300 Bp



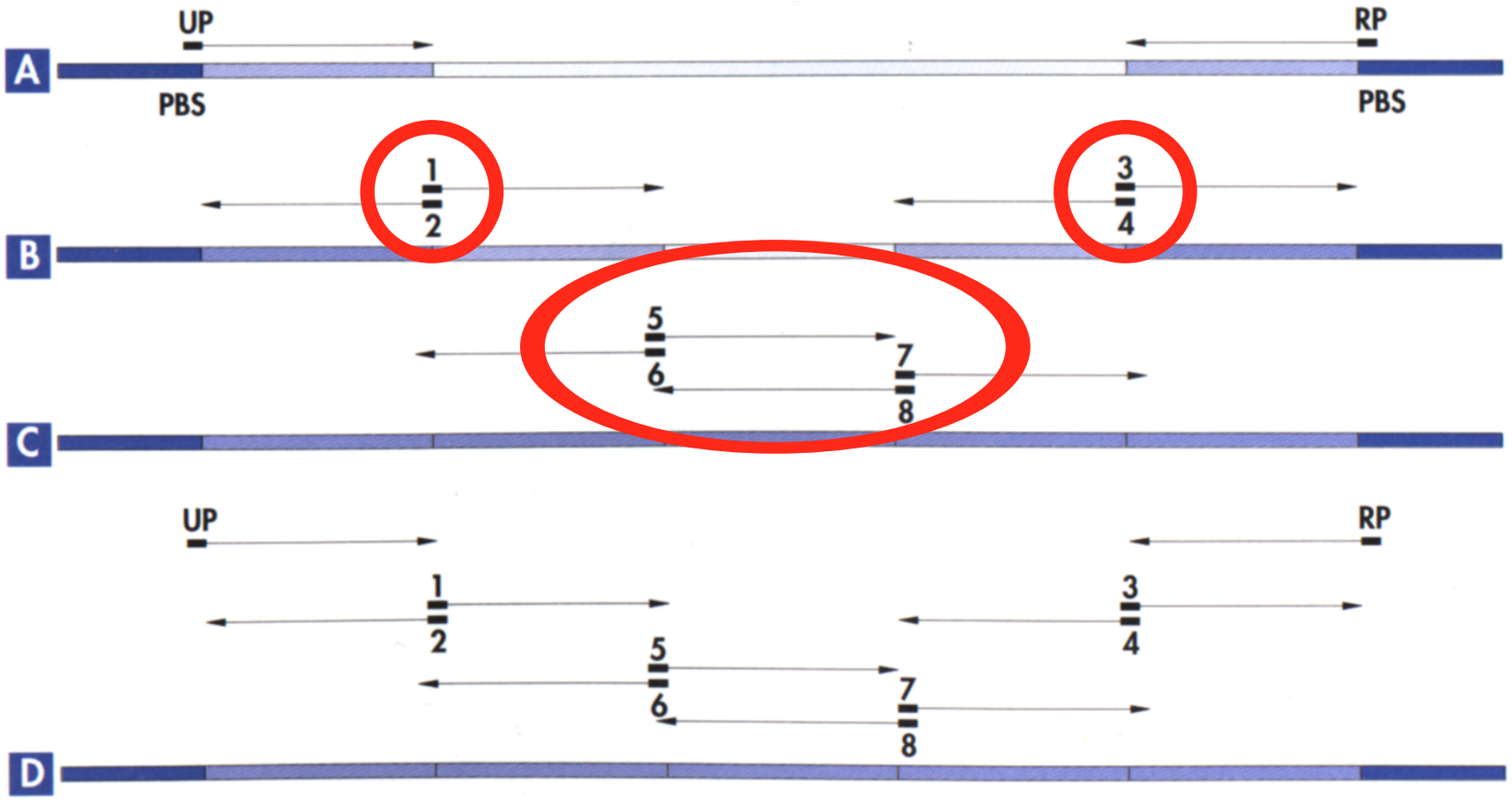
Längere DNA-Moleküle (z. B. ganze Genome) müssen schrittweise (in kleinen Stücken) sequenziert werden. Diese DNA-Sequenzstücke müssen dann zum Genom zusammen-Gesetz werden („Assemblierung“).

Die ‚Primer Walking'-Strategie



- in Kombination mit **Sanger**-Verfahren
- sequentieller Ablauf > langsam
- geordnete Strategie > übersichtlich
- vergleichsweise teuer (Primer kosten Geld)

Die ‚Primer Walking'-strategy

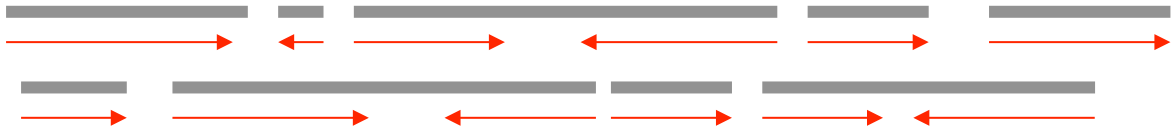


...gleichzeitig VOR und ZURÜCK > ds-Sequenzierung

Die ‚shotgun'-Strategie

Ausgangs-DNA-Moleküle

‚zufälliges' Zerbrechen der DNA



DNA-Fragmente (überlappend!)

Sequenzreaktionen

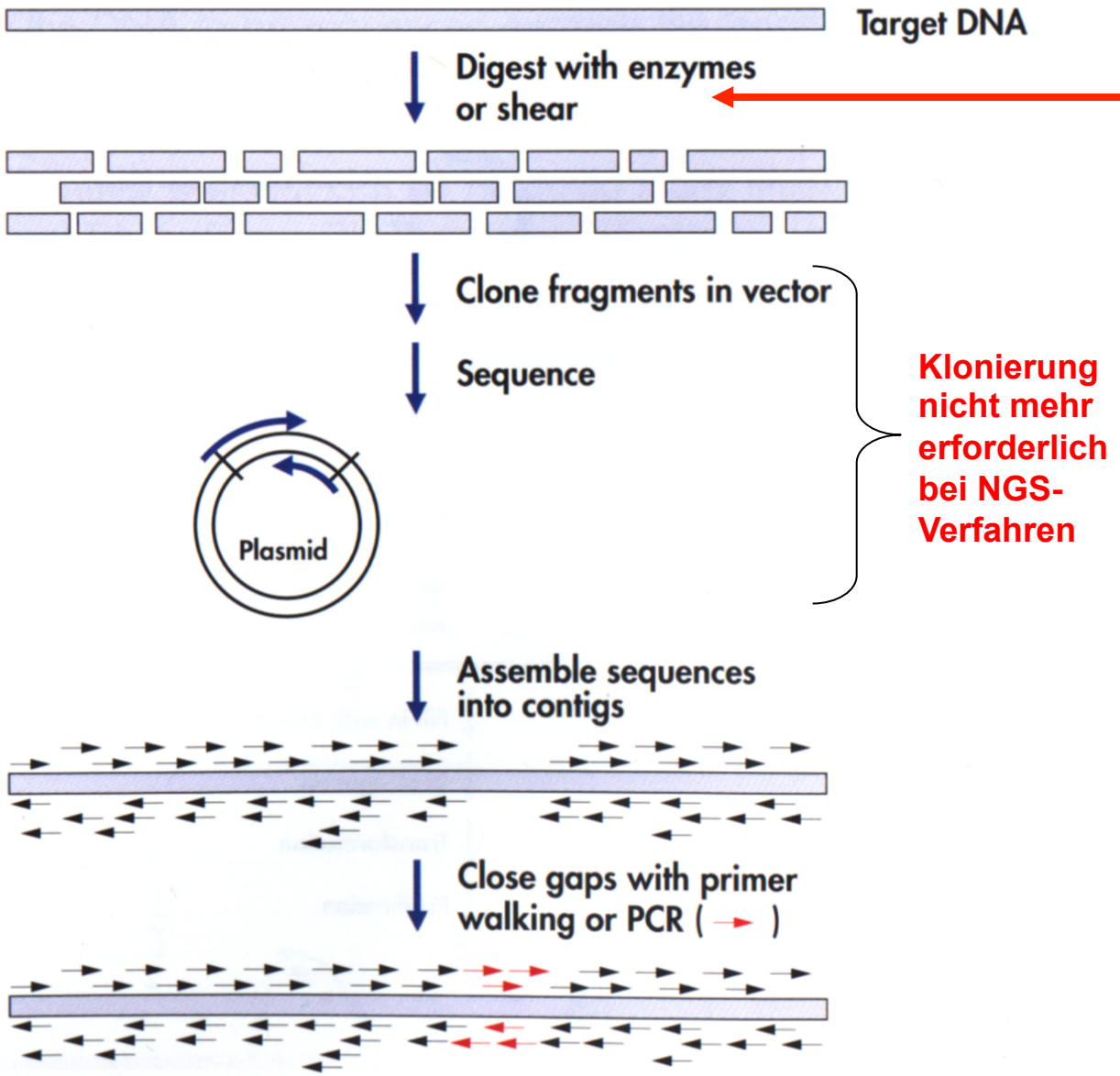
Assemblierung der Teilsequenzen

```
ATGGCGAATGCCTTGACGCCGATGCAATTCAAG
   GGCGAATGCCTTGACGCC          ATTCAAGTGCATGTA..
```

Überlapp Überlapp

Konsensus-Sequenz

Die ‚shotgun'-Strategie



Klonierung nicht mehr erforderlich bei NGS-Verfahren

Einfach mit „Nebulizer"

Sequenzvergleich durch Alignment: die Schlüssel-Technik der Bioinformatik!

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

Nt-Substitution

As-Austausch

Gap bzw. InDel

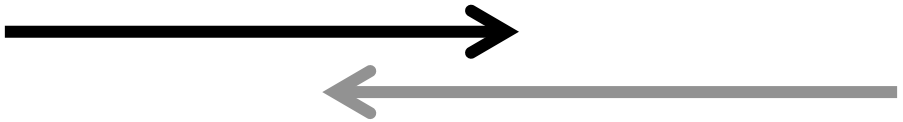
```
Query: 5   EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQY--NCRQFSSPEDCLSSPEFL 62
           + ELIR SW ++ ++ + HG +LF+RLF L+P+LL LF Y  NC     S +DCLSSPEFL
Sbjct: 8   DKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYTTNC---GSTQDCLSSPEFL 64
```

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‘

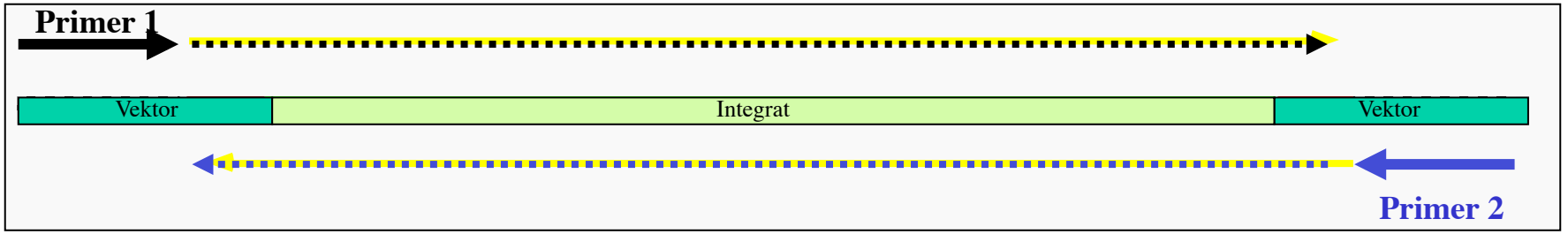
5‘-TACCGCA-3‘ „Reverse Complement“

5‘-TTACTAC-3‘
| | |
5‘-TACCGCA-3‘

Alignment zweier Sequenzen:

„Mensch vs. Computer“

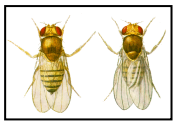
Bsp: **Sequenzierung eines Plasmid-Integrats von beiden Seiten**

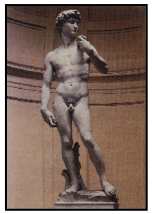


Ausgabe-Files:	Read 1	5‘-GCATTGGCACAT-3‘
	Read 2	5‘-ATGTGCCAATGC-3‘
Mensch:	Read 1	5‘-GCATTGGCACAT-3‘
	Read 2	3‘-CGTAACCGTGTA-5‘
Assembly-Programm	Read 1	5‘-GCATTGGCACAT-3‘
	Read 2RC	5‘-GCATTGGCACAT-3‘
		* ** * ** * * ** * *

rc = reverse complement

Genomgrößen bei Eukaryoten




• Enzephalitozoon	3	MBp	2 000 Gene
• Saccharomyces cerevisiae	12	MBp	6 200 Gene
• Caenorhabditis elegans	97	MBp	19 000 Gene
• Drosophila melanogaster	137	MBp	14 000 Gene
• Gallus gallus	1 000	MBp	23 000 Gene
• Homo sapiens	>3 000	MBp	<25 000 Gene
• Arabidopsis thaliana	125	MBp	25 000 Gene
• Oryza sativa	400	MBp	>50 000 Gene
• Paris japonica	149 000	MBp	?

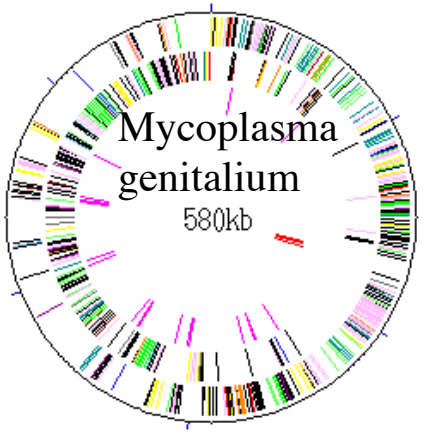
Genomgröße = C-Wert

in bp/Kb/Mb/Gb oder in pg

$1pg = 0.965 \times 10^9$ bp $= 6.1 \times 10^{11}$ Da $= 34$ cm

Genomgrößen in Bakterien

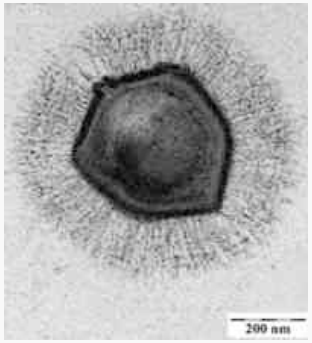
• Eubakterien:	Mycoplasma genitalium	580 kb
	E. coli (K12)	4639 kb
	E. coli (O157:H7)	5529 kb
	Bacillus megaterium	30000 kb
• Archaebakterien:	Thermoplasma acidophilum	1564 kb
	Halobacterium salinarium	4000 kb



Prokaryoten-Genome besitzen nur wenig repetitive DNA und bestehen aus ‚dicht-gepackten' Genen

Tot oder lebendig?

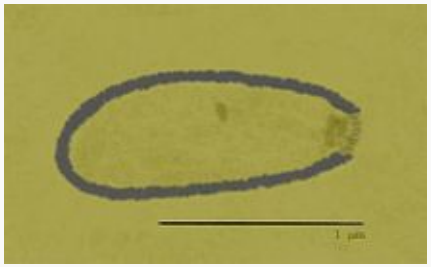
- Mimi*-Virus (befällt Amöben)



1.2 Mb DNA
1260 Gene
< 10% „junk DNA“

400 nm groß!

- Pithovirus sibericum (Bohrkern aus sibirischem Permafrost)



600 Kb DNA
470 Gene

1,5 um groß !!

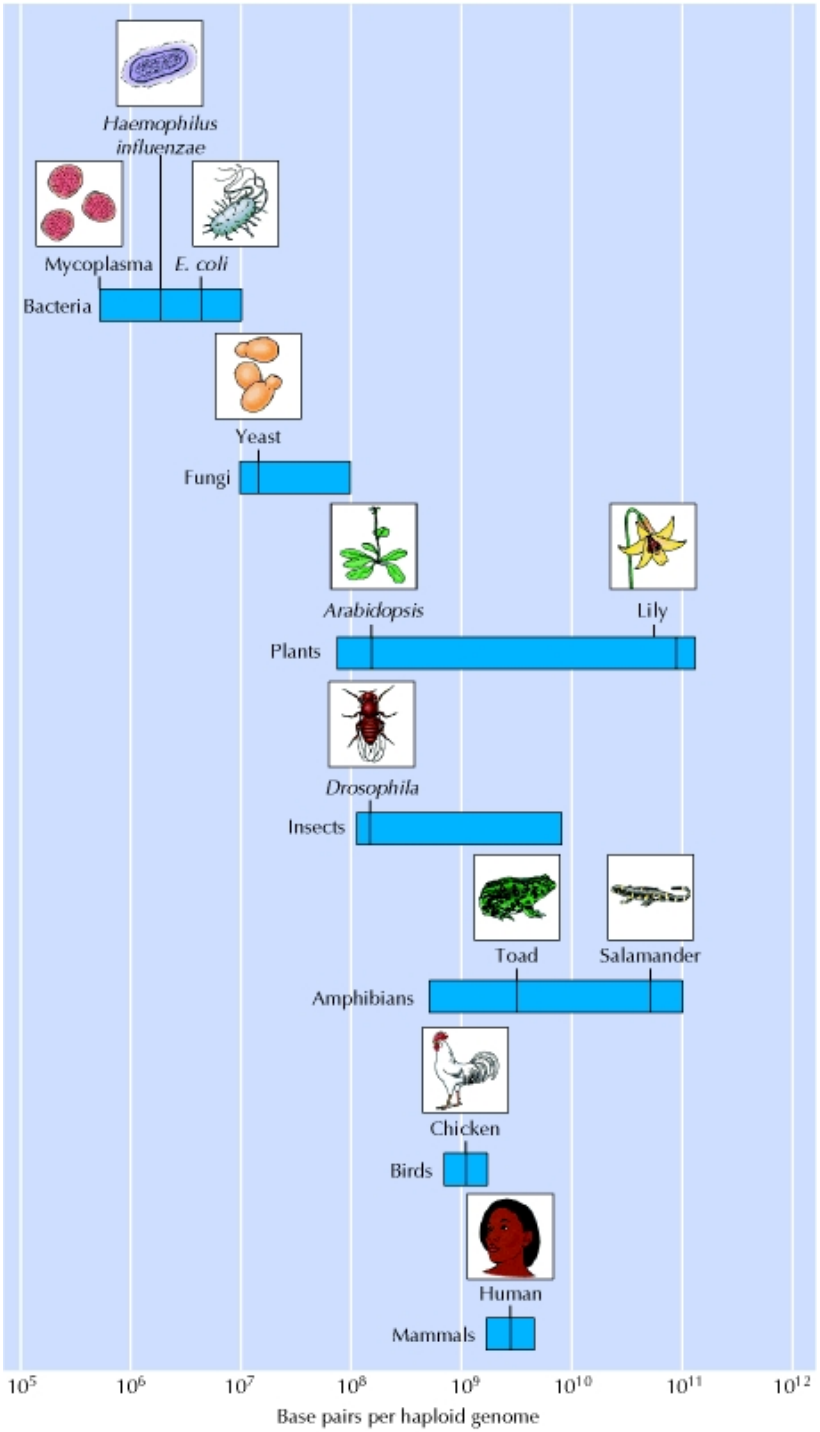
* Microbe-mimicking

Databases of genome sizes

http://www.cbs.dtu.dk/databases/DOGS/
http://www.genomesize.com/
http://data.kew.org/cvalues/CvalServlet?querytype=1
http://www.genomicron.evolverzone.com/2007/04/
genome-size-databases/

http://www.genomesize.com/prokaryotes/
http://www.jcvi.org/cms/research/past-projects/cmr/overview/

https://gold.jgi.doe.gov/

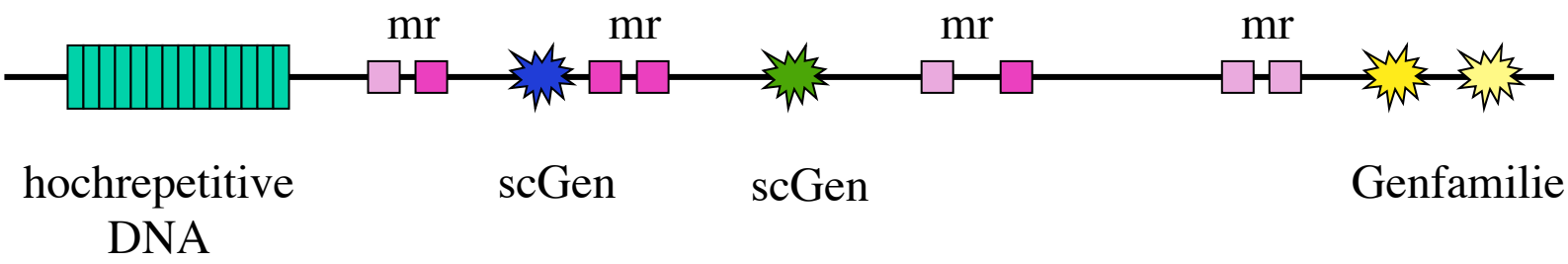


Das C-Wert-Paradoxon*

* Enorme Unterschiede
in der Genomgröße
trotz weitgehend ähnlicher
Komplexität der Organismen

Komplexe Genome höherer Eukaryoten enthalten repetitive und ‚single copy' DNA-Komponenten

- hoch-repetitive DNA ca. 10% des Genoms
- mittel-repetitive DNA ca. 40% des Genoms
- ‚single copy' DNA ca. 50% des Genoms



Genomkomponenten

- **single copy DNA**

  Gene und Intergenregionen

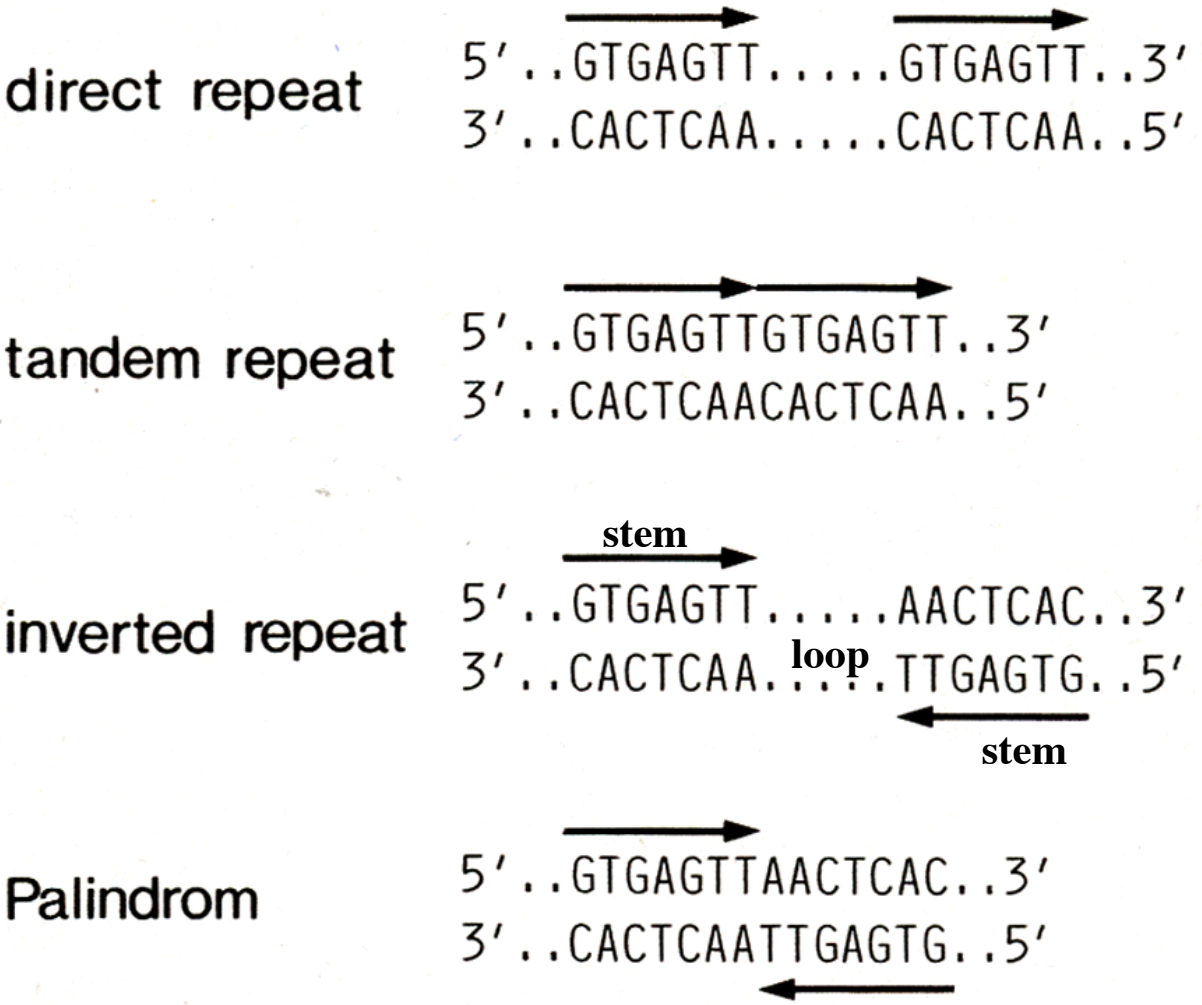
- **mittel-repetitive DNA** (10-1000; meist interspergiert)

  Genfamilien (z. B. Globin/Histon/rDNA-Gene)
  Transposons (= mobile DNA-Abschnitte)

- **hoch-repetitive DNA** (10^3-10^6; oft tandem-repetitiv)

  Satelliten-DNA der Centromer-Regionen

Sequenzwiederholungen in der DNA



Genomkomponenten

Humangenom
3200 MB

Genomweite Repeats, Transposons;
(z.B Alu-SINES und LINES)
1400 MB

Simple Tandem-Repeats
(syn. Microsatelliten)
90 MB

Satelliten-DNA
(Tandem-Repeats in Centromeren)
160 MB

ca. 50% repetitive DNA !

RNA-kodierende Genregionen
? MB

70 % des Genoms?

Proteinkodierende Genregionen
1200 MB

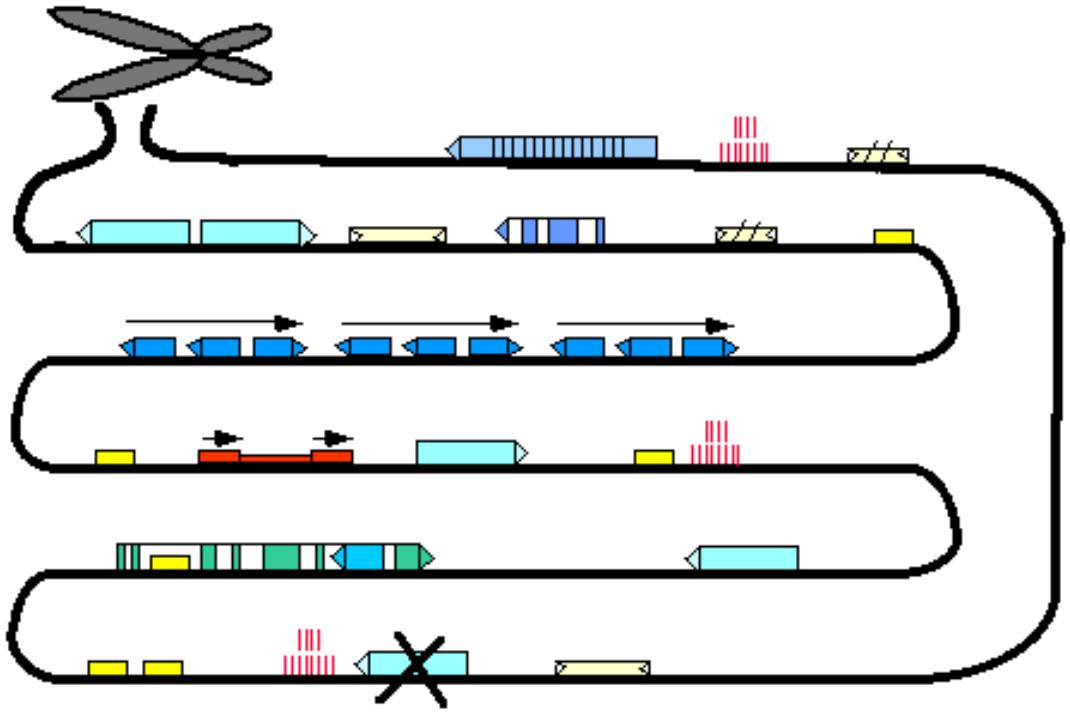
Proteinkodierende Exons (48 MB)

Nur 1,5 % !

Introns, UTR, Promoter
1152 MB

MB = Mega-Basenpaare = 1 Mio Bp

Komponenten des Eukaryotengenoms



intronlose Genfamilie mit inaktivem Pseudogen (X)

tandem-repetitive Genfamilie

single copy-Gen mit Intron-/Exonstruktur

verschachteltes Arrangement zweier single copy-Gene

single copy-Gen mit internen Tandem-Repetitionen

kurze, interspergierte Repetitionen der SINE-Klasse

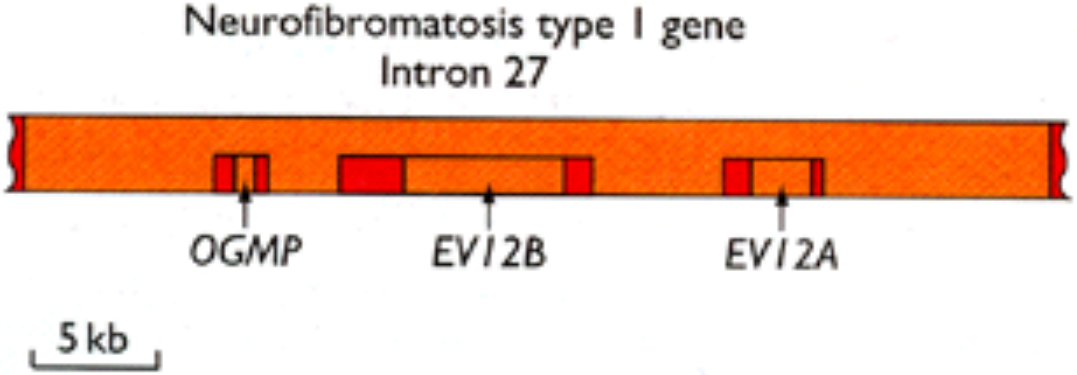
komplette und "verstümmelte" (//) DNA-Transposons

Retro-Transposon mit "long terminal repeats"

längenvariable Mikro-/Mini-Satelliten-Cluster

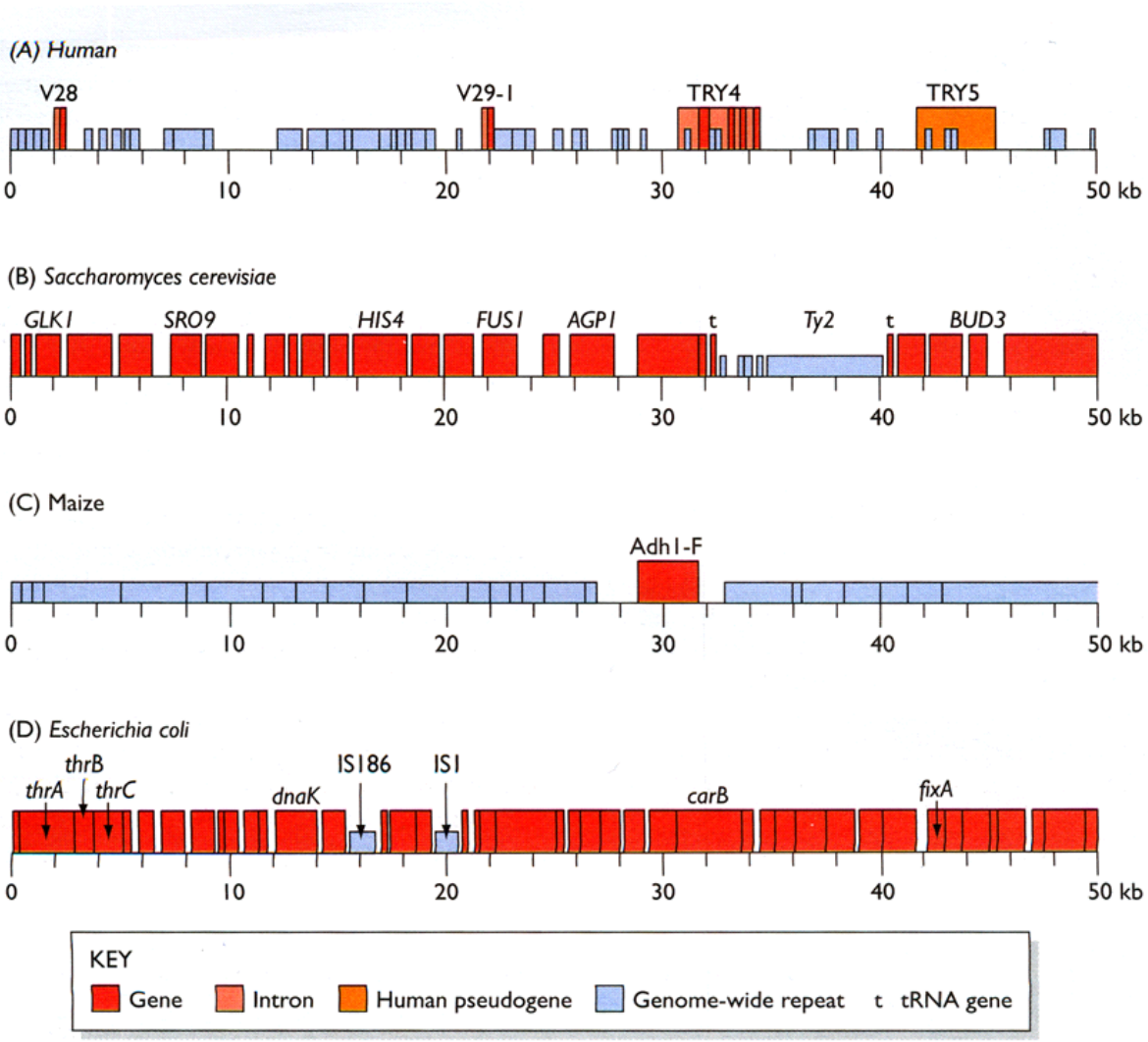
„Nested Genes“

- **Genes-within-genes** are relatively common features of nuclear genomes, one gene being contained within an intron of a second gene. An example in the human genome is the neurofibromatosis type 1 gene, which has three short genes (called *OGMP*, *EVI2A* and *EVI2B*) within one of its introns. Each of these internal genes is also split into exons and introns.

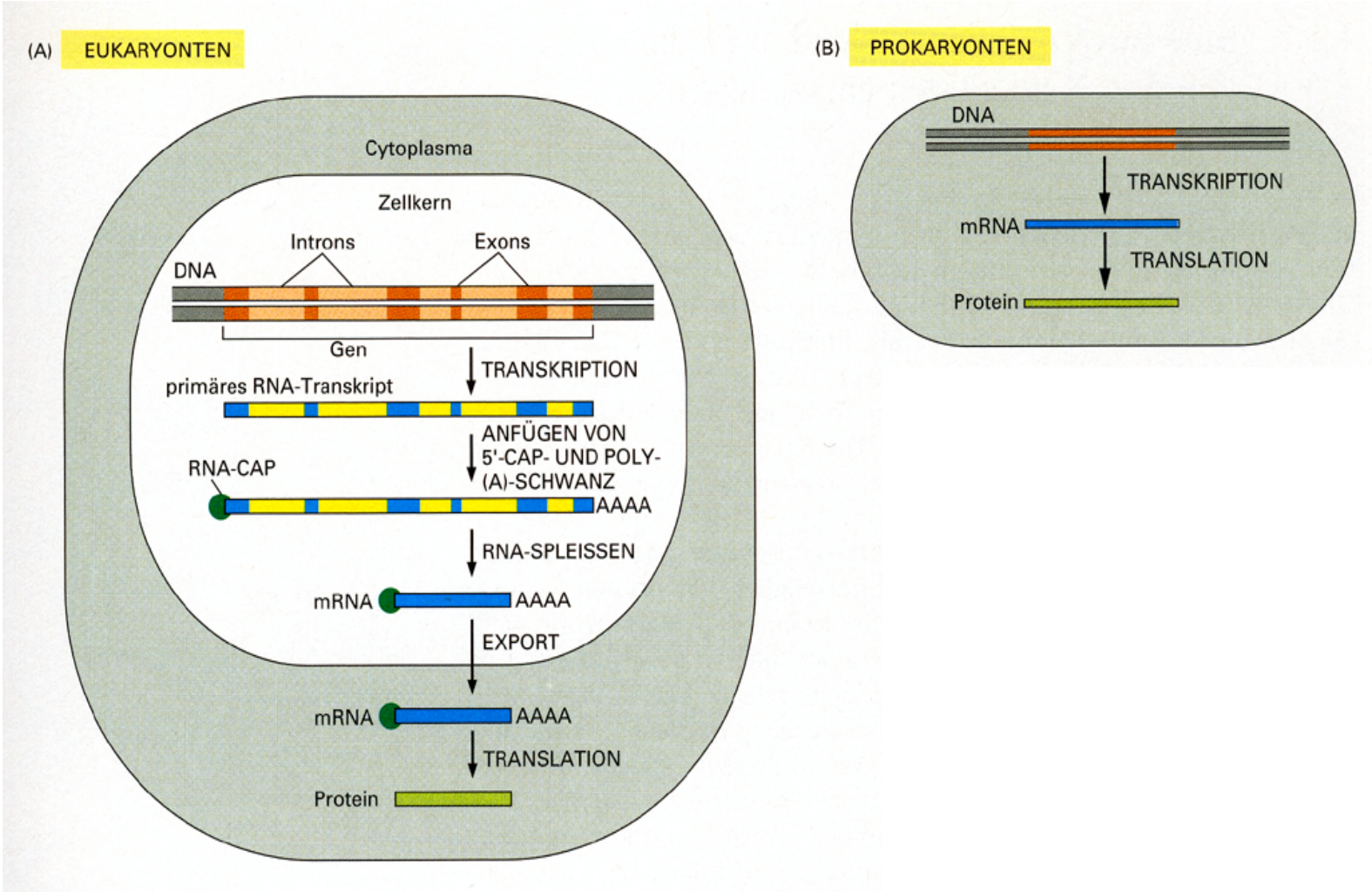


Recently, it has been discovered that many snoRNAs, which are involved in chemical modification of rRNAs (Section 9.4.1), are specified by genes within introns.

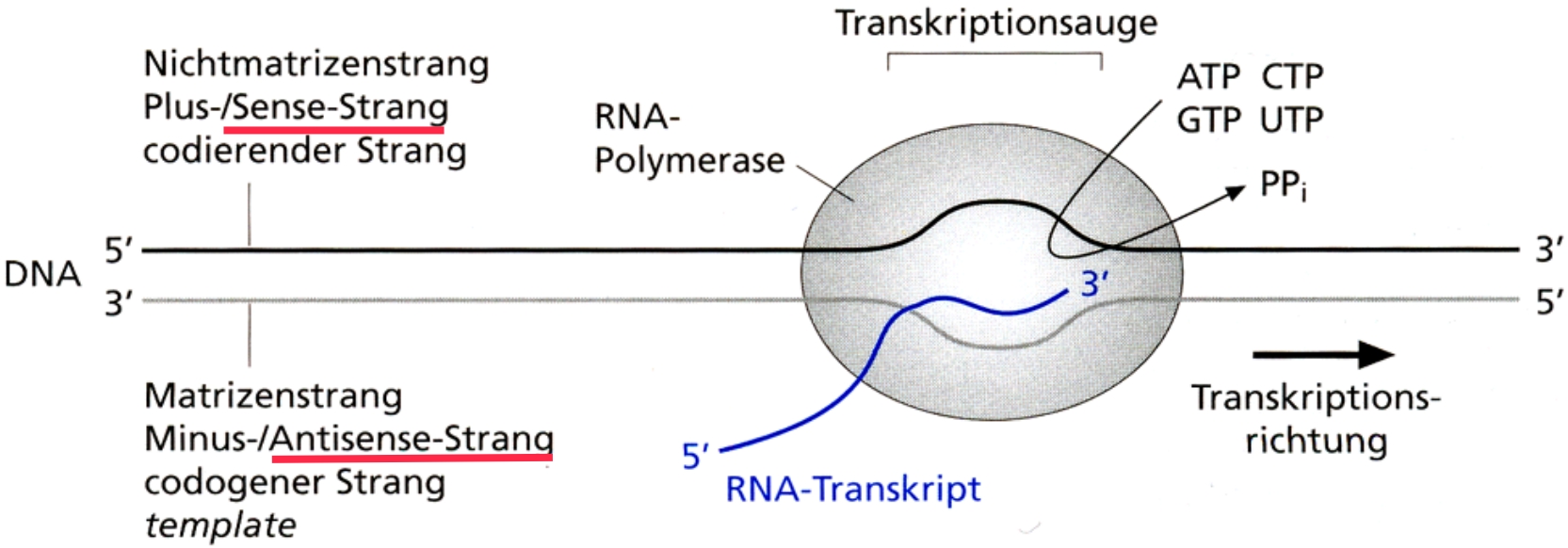
Die Genomstruktur ist taxonspezifisch



Umsetzung der genetischen Information



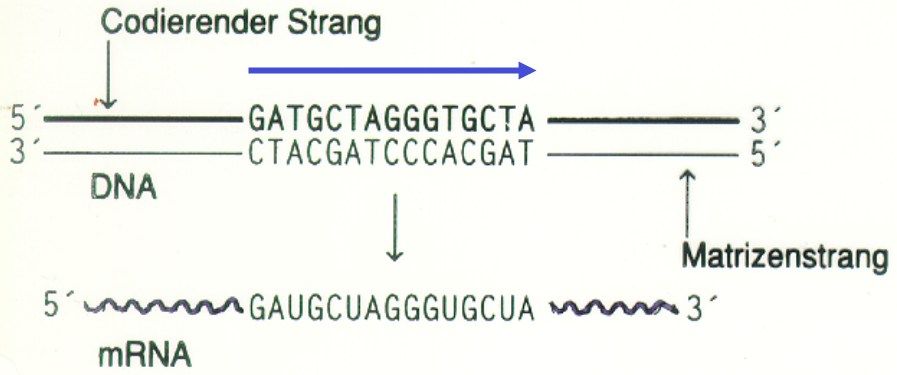
Transkription & Gen-Anordnung



Transkription

Codierende Region: [coding region] Bezeichnung für diejenigen Bereiche auf der DNA, die ein Genprodukt liefern, d. h. für die Synthese einer RNA oder eines Proteins codieren. Siehe Gen.

Codierender Strang: [coding strand, sense strand] Syn. Sinnstrang. Bezeichnung für denjenigen Strang in einem Nucleinsäure-Molekül, der die gleiche Sequenz aufweist, wie die von der entsprechenden DNA-Region abgelesene mRNA. Derjenige Strang, der als Matrize für die Synthese der mRNA dient und daher eine der mRNA komplementäre Sequenz aufweist, wird auch als **Matrizenstrang** [template strand] oder **anticodierender Strang** bzw. **Nicht-Sinnstrang** [anticoding strand, antisense strand] bezeichnet.

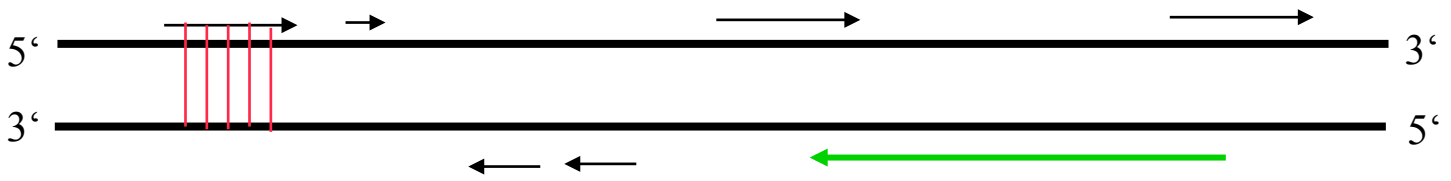


Transkription

Q: wie verhalten sich die Begriffe sense- und nonsense-Strang sowie Watson- und Crick-Strang zueinander?

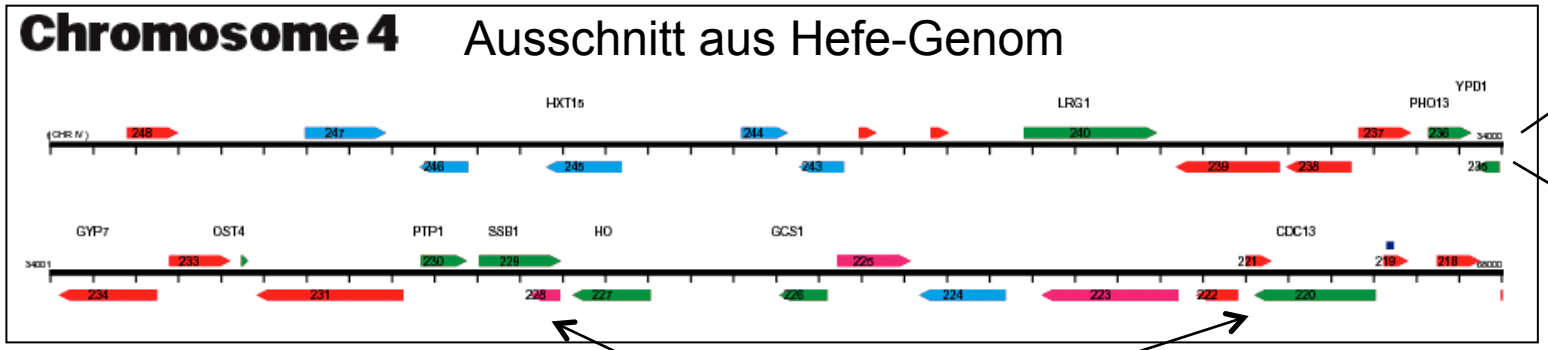
Transkription und Genanordnung

- Beide Stränge der DNA bei Pro-und Eukaryoten können transkribiert werden.
- NEUE Daten! Auch bei Eukaryoten überlappen viele Transkriptionseinheiten (ENCODE-Projekt).



Gene im Eukaryoten-Genom

- Beide Stränge der DNA bei Pro-und Eukaryoten können transkribiert werden.



- Gene können überlappen!

- Transkribierter Genomanteil vermutlich > 70%!
  Protein-kodierender Anteil nur ca. 1,5 %!

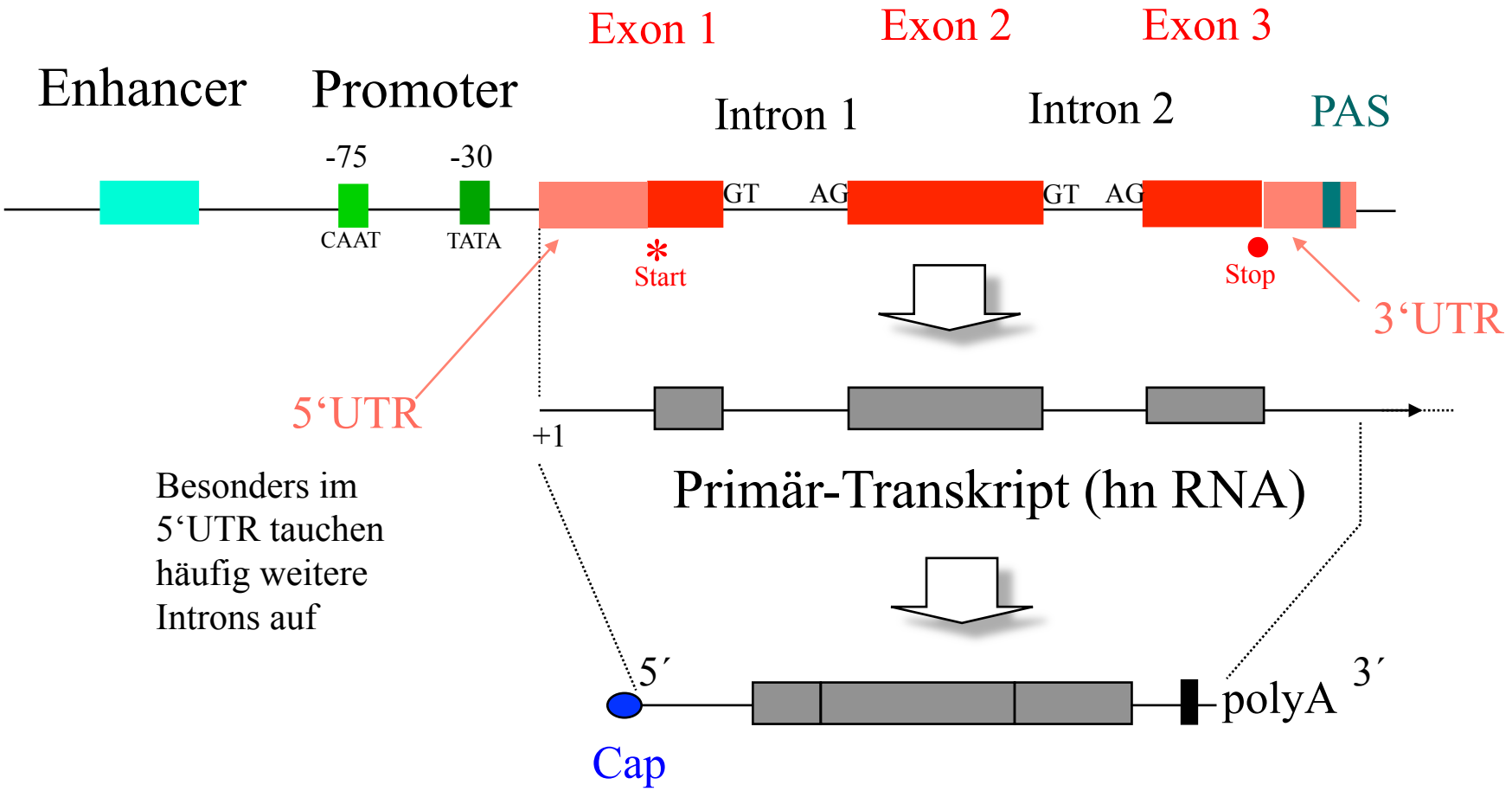
14.6.2007

Ein Schwerpunkt-Thema der Bioinformatik:

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 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
```

Typische Struktur proteinkodierender Gene in Eukaryoten



>> Viele Hinweise auf die Genstruktur!

Definitionen

- Exons sind Teil der reifen mRNA (E. sind nicht immer protein-kodierend!)
- Introns werden aus Primärtranskript (hnRNA) herausgespleißt.
- Introns beginnen meist mit GT und enden mit AG („GT-AG-Regel“)
- hnRNA = (längen)heterogene nukleäre RNA
- 5‘/3‘ UTR = 5‘ bzw. 3‘ liegende nicht-translatierte Regionen der reifen mRNA (können auch eigene Exons sein!)
- Enhancer bestimmen v.a. Spezifität u. Stärke des Transkriptions-vorgangs
- „+1“ = Transkriptionsstart (= erste transkribierte Nt-Position)
- PAS = Polyadenylierungssignal. Ca. 20 Nt abwärts des PAS wird die mRNA geschnitten und polyadenyliert.

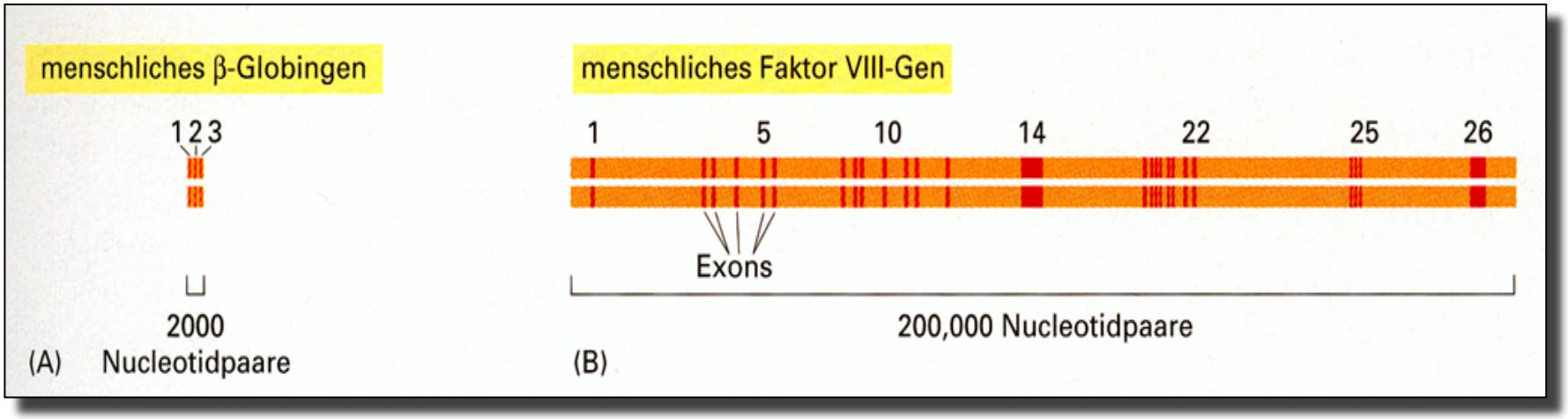
Das durchschnittliche proteinkodierende menschliche Gen

(„…existiert nicht“)

• Größe ‚interner‘ Exons	145 Bp
• Exonanzahl	8.8
• Intronlänge	3365 Bp
• 3‘ UTR	770 Bp
• 5‘ UTR	300 Bp
• CDS	1340 Bp / 447 As
• Genomausdehnung	27 kb

Die Mosaikstruktur erschwert das Erkennen von Genen in Eukaryoten-DNA

Intron-Exon-Struktur



- Duchenne-Muskeldystrophie-Gen:

  78 Exons
  verteilt über 2 400 kb
  99% des Genbereichs besteht aus Introns
  Transkriptionsdauer ca. 20 Std.

- Introns haben 95%-Anteil an menschlichen Genen

Introns in proteinkodierenden Kern-Genen haben Konsensus-Spleißstellen

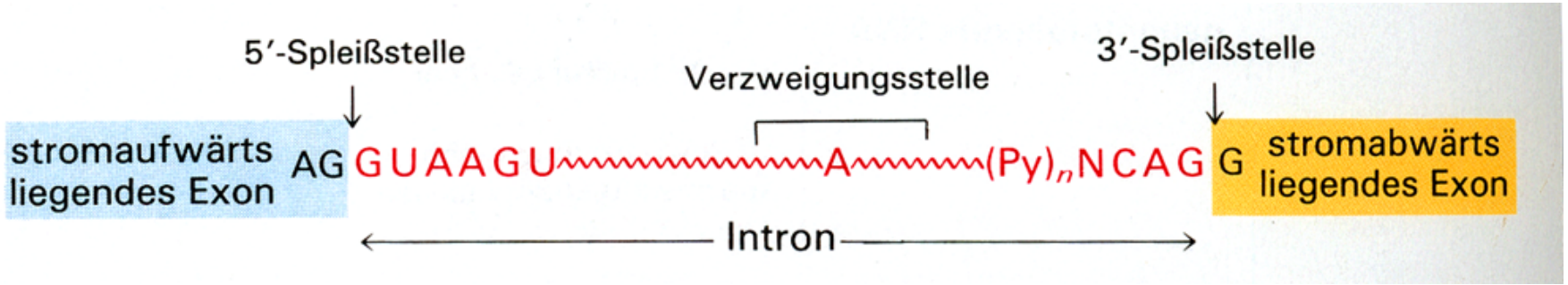
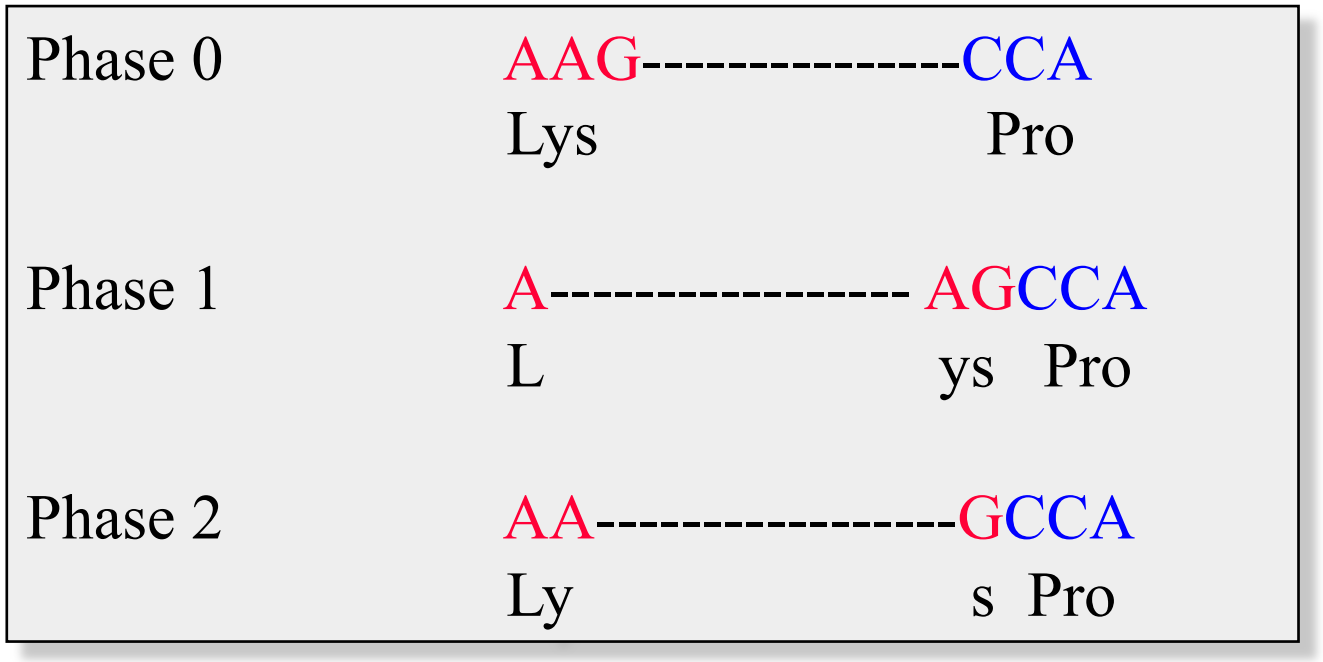


Tabelle 29.3: Basensequenzen von Spleißpunkten bei Transkripten mit Introns

Genregion	Exon	Intron	Exon
Ovalbumin, Intron 2	U A A G G U G A G C	~~~	U U A C A G G U U G
Ovalbumin, Intron 3	U C A G G U A C A G	~~~	A U U C A G U C U G
β-Globin, Intron 1	G C A G G U U G G U	~~~	C C U U A G G C U G
β-Globin, Intron 2	C A G G G U G A G U	~~~	C C A C A G U C U C
Immunglobulin λ_1, Intron 1	U C A G G U C A G C	~~~	U U G C A G G G G C
SV40-Virus, frühes T-Antigen	U A A G G U A A A U	~~~	U U U U A G A U U C

„Intronphasen“ in proteinkodierenden Genen



Introns können also die kodierenden Bereiche an jeder Stelle unterbrechen!