

# Algorithmen auf Sequenzen

07.04.2011

Prof. Dr. Sven Rahmann

## Team

Prof. Dr. Sven Rahmann  
Dipl.-Inf. Dominik Kopczynski (Übungen)

**Vorlesung** Do 8:30-10 in OH14, R104

**Übungen** Do 14-15 in OH14, R202

## Alle Informationen

Webseite zur Vorlesung: <http://www.rahmannlab.de/lehre/v-algo-seq>  
Bitte regelmäßig dort nachlesen!

## Sprechstunde

Mo 16-17 in OH14, R214 und nach Vereinbarung!  
Bitte möglichst per e-mail anmelden, sonst evtl. sehr lange Wartezeiten!  
Sven.Rahmann /at/ tu-dortmund.de

## Prüfungsleistungen

Dies ist Bachelor-Wahlmodul INF-BSc-315 (2V+1Ü, 4 LP).

Für die DPO 2001 kann ein LNW erworben werden (4,5 LP).  
Prüfung mit erweitertem Umfang 3V+1Ü, 6 LP in SpGs 4,6,7.

### Prüfung

Klausur oder mündlich (je nach Teilnehmerzahl).  
vermutlich mündlich, Festlegung nächste Woche in der Übung.

### Übungen

Nicht Voraussetzung zur Prüfungsanmeldung,  
aber wichtig zum Verständnis und Bestehen der Prüfung  
Bitte nehmen Sie aktiv teil!  
Die Übungen sind für Sie!  
Nähere Hinweise von Herrn Kopczynski.

## Literatur (Auswahl)

Gonzalo Navarro, Mathieu Raffinot (2002)

### **Flexible Pattern Matching in Strings**

Cambridge University Press

Richard Durbin et al. (1998)

### **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids**

Cambridge University Press

Dan Gusfield (1997)

### **Algorithms on Strings, Trees and Sequences**

Cambridge University Press

Nello Christianini and Matthew W. Hahn (2007)

### **Introduction to Computational Genomics – a Case Studies Approach**

Cambridge University Press

## Sequenzen sind überall – Anwendungen der Sequenzanalyse

Biosequenzen (DNA, RNA, Proteine)

(aber: Genome sind komplexer als „nur“ eine DNA-Sequenz)



Texte (Literatur, wissenschaftliche Texte)

Die Kunst hinter guter Literatur und hinter guten wissenschaftlichen Arbeiten besteht darin, schwierige, komplex zusammenhängende Sachverhalte in eine logische Abfolge von einzelnen Sätzen zu bringen.

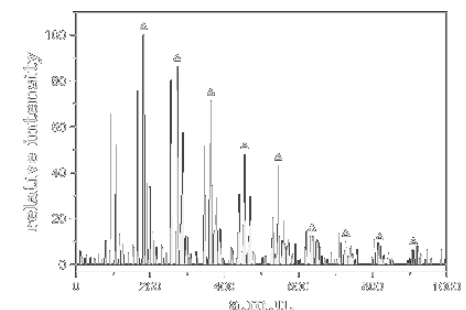
Programme

Dateien, Datenströme

Komplexe Datenstrukturen werden serialisiert.

Zeitreihen, Spektren

Audiosignale, Massenspektren, ...



## Probleme auf Sequenzen (Auswahl)

Suche nach Mustern in Sequenzen (z.B. Textverarbeitung: find-replace)

Approximative Mustersuche (Meier vs. Mayer)

Suche nach einem ähnlichem Wort in einem Wörterbuch

Sequenzvergleich: Quantifikation von Ähnlichkeit / Unterschieden

Anwendung: Revisionskontrolle, Verfolgen von Änderungen (z.B. subversion)

Entdecken von auffälligen Mustern, z.B. wiederholten Teilstrings  
wichtig für Genomanalyse, Kompression

Quantifikation  
von Sequenzkomplexität,  
Datenkompression

### Brief Biography

**Sven Rahmann** became a professor for [Bioinformatics for High-Throughput Technologies](#) at the [Chair of Algorithm Engineering, Computer Science Department, TU Dortmund](#) in October 2007.

Between August and December 2007, he spent four months at [HHMI Janelia Farm Research Campus](#) as a visiting scientist in [Gene Myers' lab](#).

**Sven** was an independent Junior Research Group leader of the [Computational Methods for Emerging Technologies \(COMET\)](#) group, formerly known as the [Algorithms and Statistics for Systems Biology](#) group, at Bielefeld University from March 2004 till September 2007. The group closely collaborated with the [Genome Informatics](#) group at the [Faculty of Technology \(Technische Fakultät\)](#) at [Bielefeld University](#). During the same time, **Sven** was also a member of the [Institute of Bioinformatics \(IFB\)](#) at the [Center for Biotechnology \(CeBITec\)](#), and part of the faculty of the [Graduate School in Bioinformatics and Genome Research](#) at Bielefeld University.

From January 2001 till February 2004, **Sven** wrote his [doctoral thesis on oligonucleotide design for microarrays](#) in the [Computational Molecular Biology group](#) at the [Max Planck Institute for Molecular Genetics](#) in Berlin.

Between 1994 and the end of 2000, **Sven** studied mathematics and computer science with a focus on statistical methods in bioinformatics at the universities of Göttingen, UC Santa Cruz, and Heidelberg. During this time, he worked as a freelance programmer for the [Gothaer insurance company](#) and as a student assistant in the [Theoretical Bioinformatics group](#) of the [German National Cancer Research Center \(DKFZ\)](#), where he wrote his [Diploma \(M.Sc.\) thesis on word statistics in random texts](#).

## Grundlegende Definitionen

Alphabet

Indexmenge

Sequenz = String = Wort = Tupel = -gram

$A^n$ ,  $A^+$ ,  $A^*$

die leere Sequenz

Präfix, Suffix

Teilstring

Teilsequenz

## Beispiele für Sequenztypen

In dieser Vorlesung (und in konkreten Anwendungen fast immer)

- Endliches Alphabet  $A$
- Indexmenge  $I = \{0, 1, \dots, N-1\}$  für ein endliches  $N$

### Beispiele

DNA-Sequenz	$A = \{A, C, G, T\}$
Protein-Sequenz	$A = 20$ Standard-Aminosäuren
C-Programme	$A =$ ASCII-Zeichen (7-bit)
Java-Programme	$A =$ Unicode-Zeichen
Audiosignal	$A = \{0, \dots, 2^{16}-1\}$
Massenspektrum	$A = [0, 1]$ (unendlich) oder Double

### Darstellung im Computer (Java)

String (wenn  $A$  aus Unicode) oder  $A[]$  oder  $\text{ArrayList}\langle A \rangle$  oder  $\text{Map}\langle I, A \rangle$

## Ein Beispiel zur DNA-Sequenzanalyse

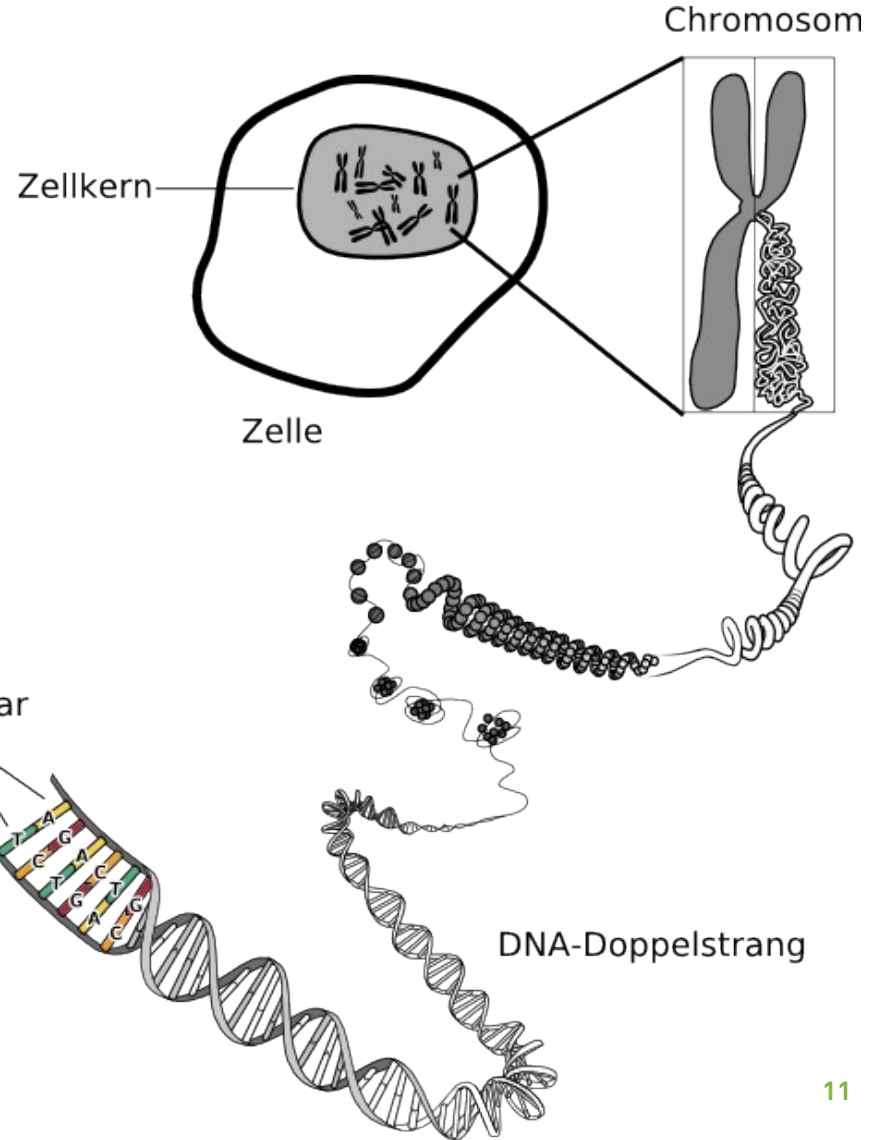
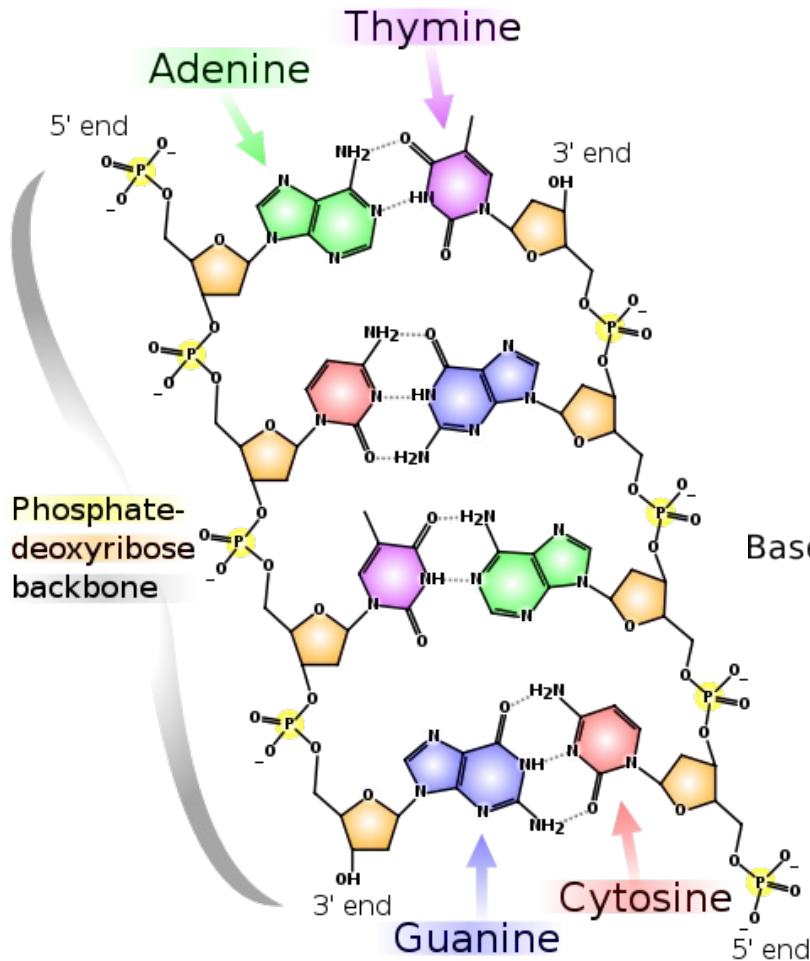
- März 2009:  
Neuer Subtyp des Influenza-A H1N1-Virus („Schweinegrippe“) tritt auf
- Isolierung und Sequenzierung des Erregers in verschiedenen Patienten
- Anlegung einer zentralen Datenbank am NCBI  
<http://www.ncbi.nlm.nih.gov/genomes/FLU/aboutdatabase.html>
- Vergleich mit älteren Varianten des H1N1 möglich
- Was kann man (mit einfachen Mitteln!) lernen?

## Das zentrale Dogma der Molekularbiologie

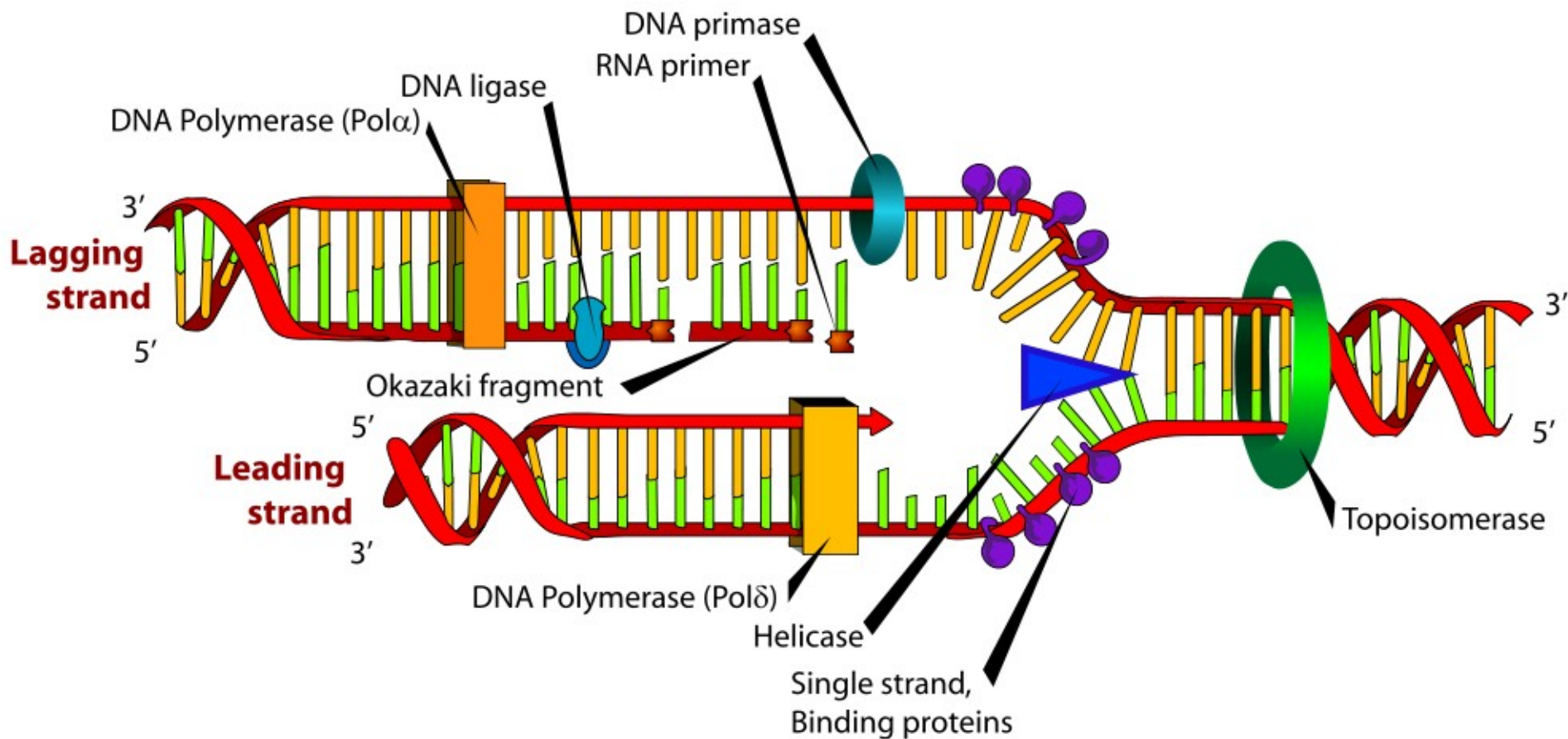
- DNA enthält Erbinformation (bei Viren oft RNA)
- Diese wird „abgelesen“ (Transkription) in mRNA
- Anhand der mRNA-Sequenz wird ein Proteinmolekül erstellt (Translation)
- Proteine führen Funktionen in der Zelle aus.
- Der Informationsfluss ist also DNA -> RNA -> Protein.
- (Diese Sichtweise ist nicht immer korrekt.)

Die folgenden Abbildungen sind selbst erstellt oder von Mariana Ruiz Villarreal, die diese in Wikipedia freigegeben hat.

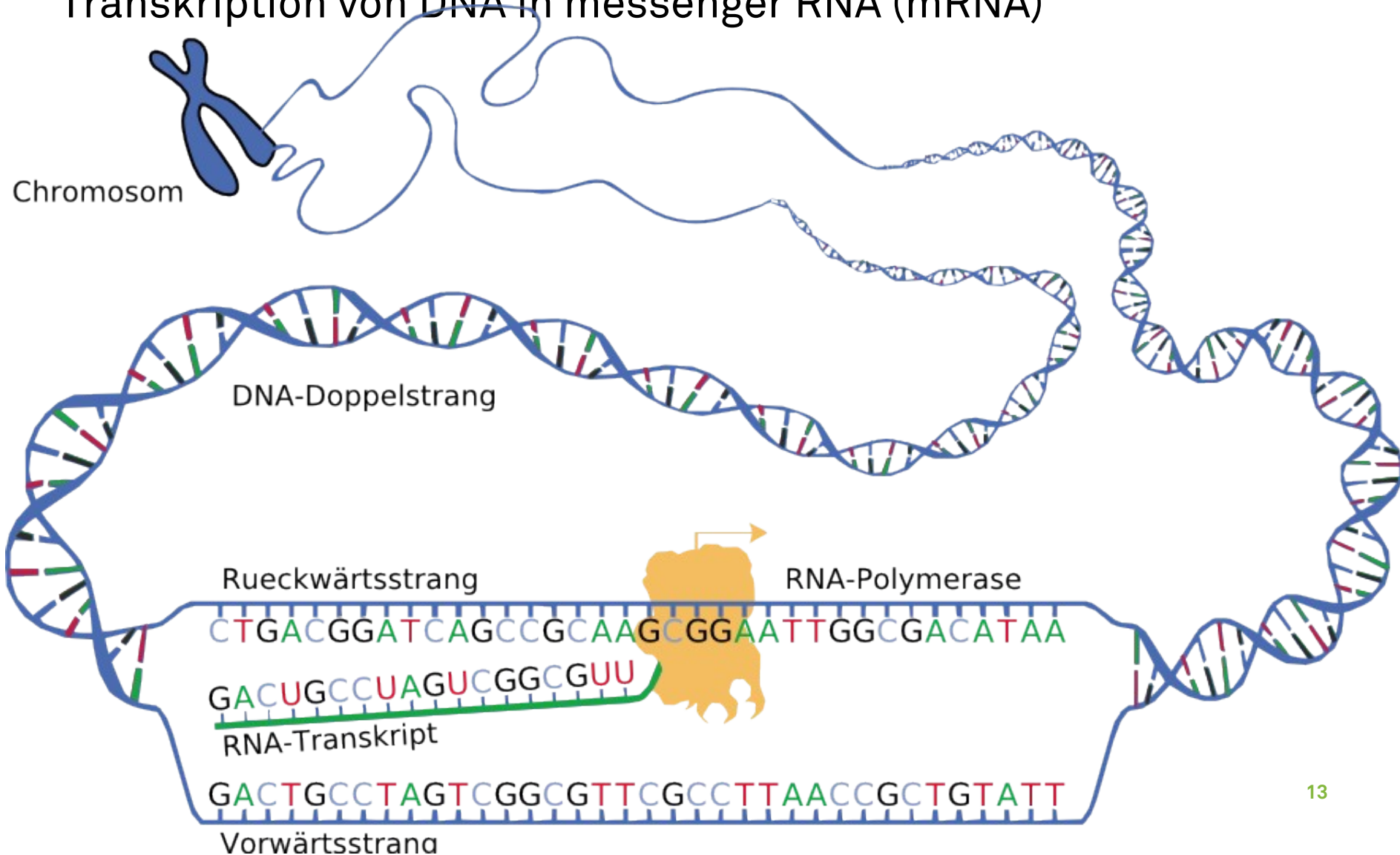
# DNA-Struktur: Biopolymer aus 4 Basen



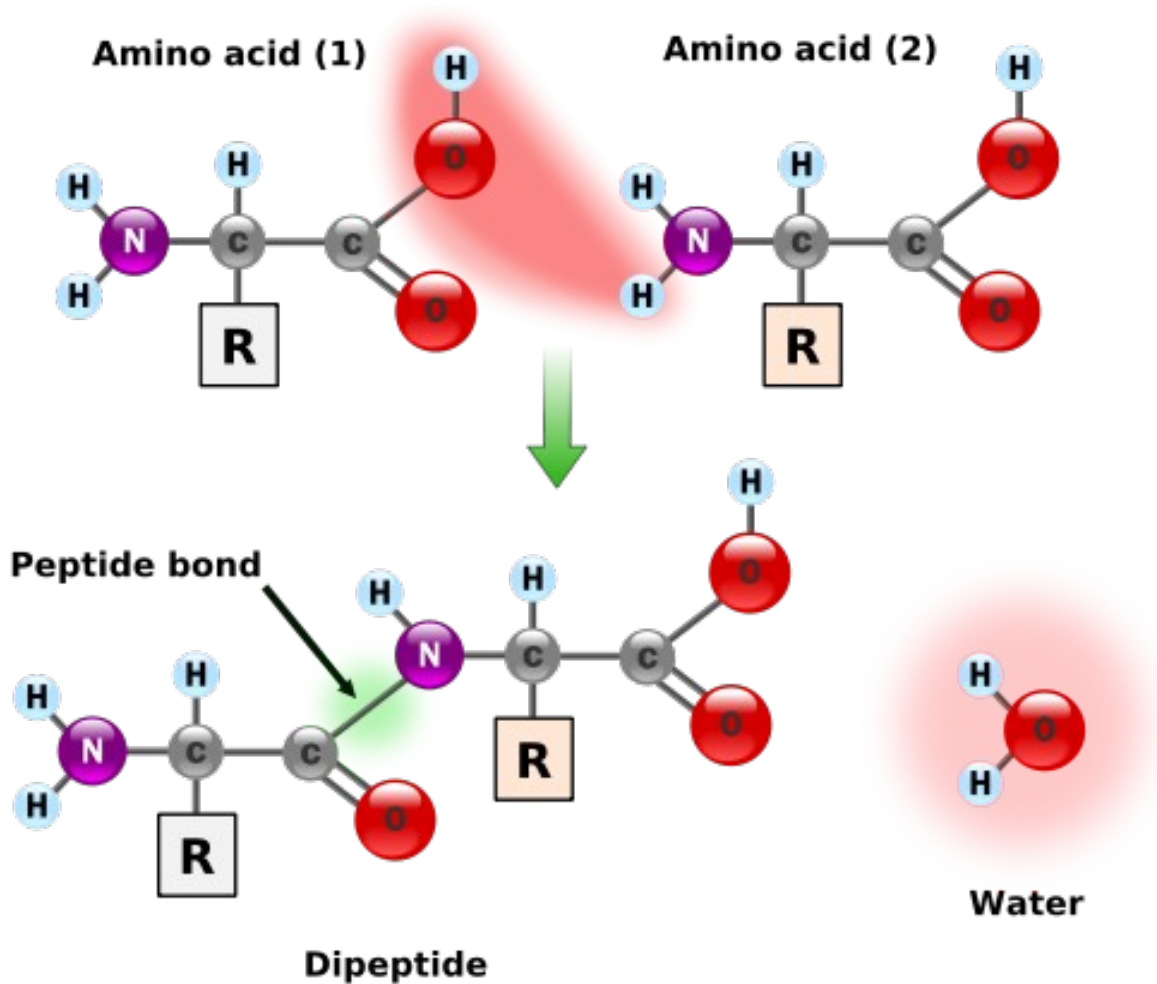
## DNA-Replikation



## Transkription von DNA in messenger RNA (mRNA)

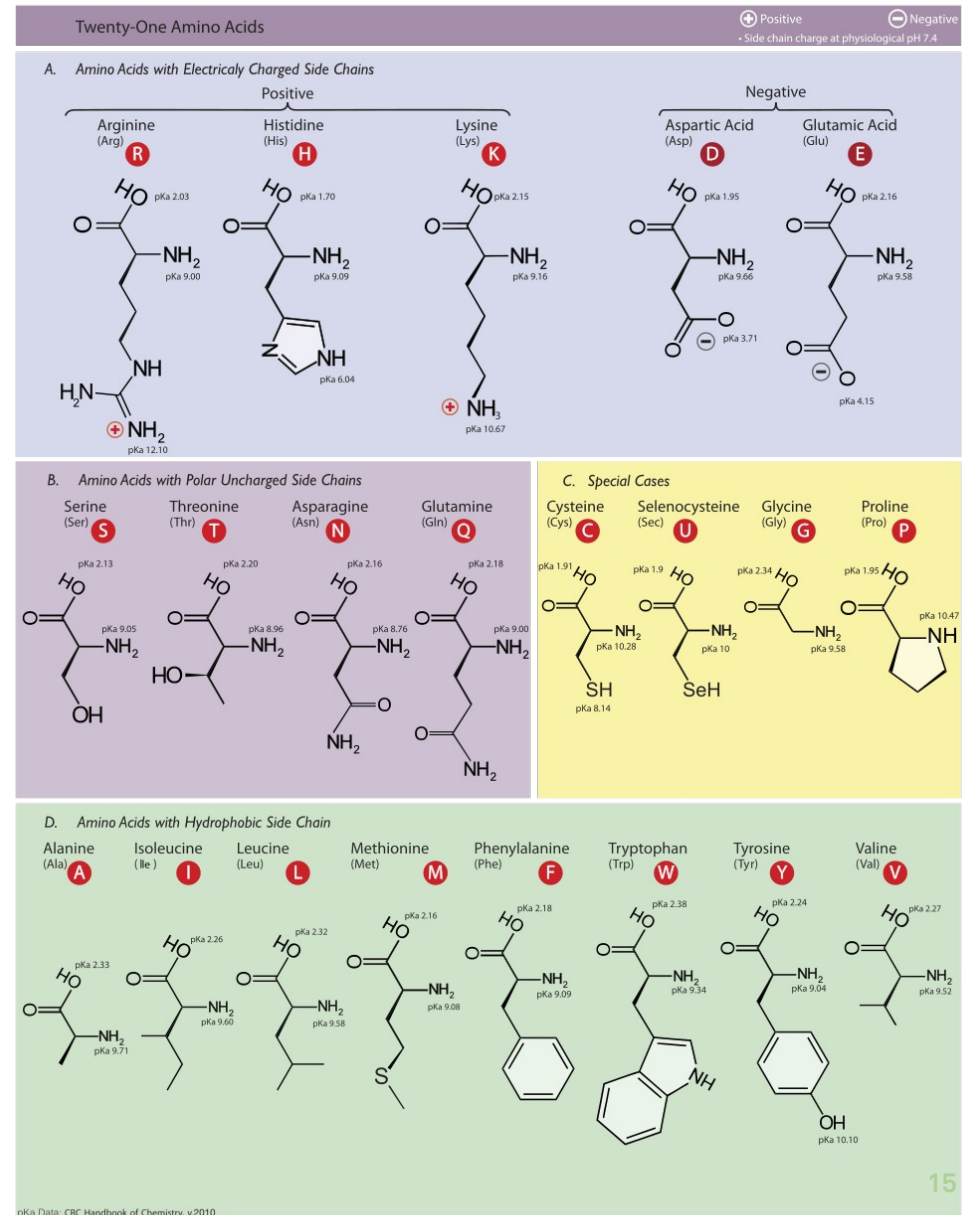


# Proteine sind Ketten von Aminosäuren

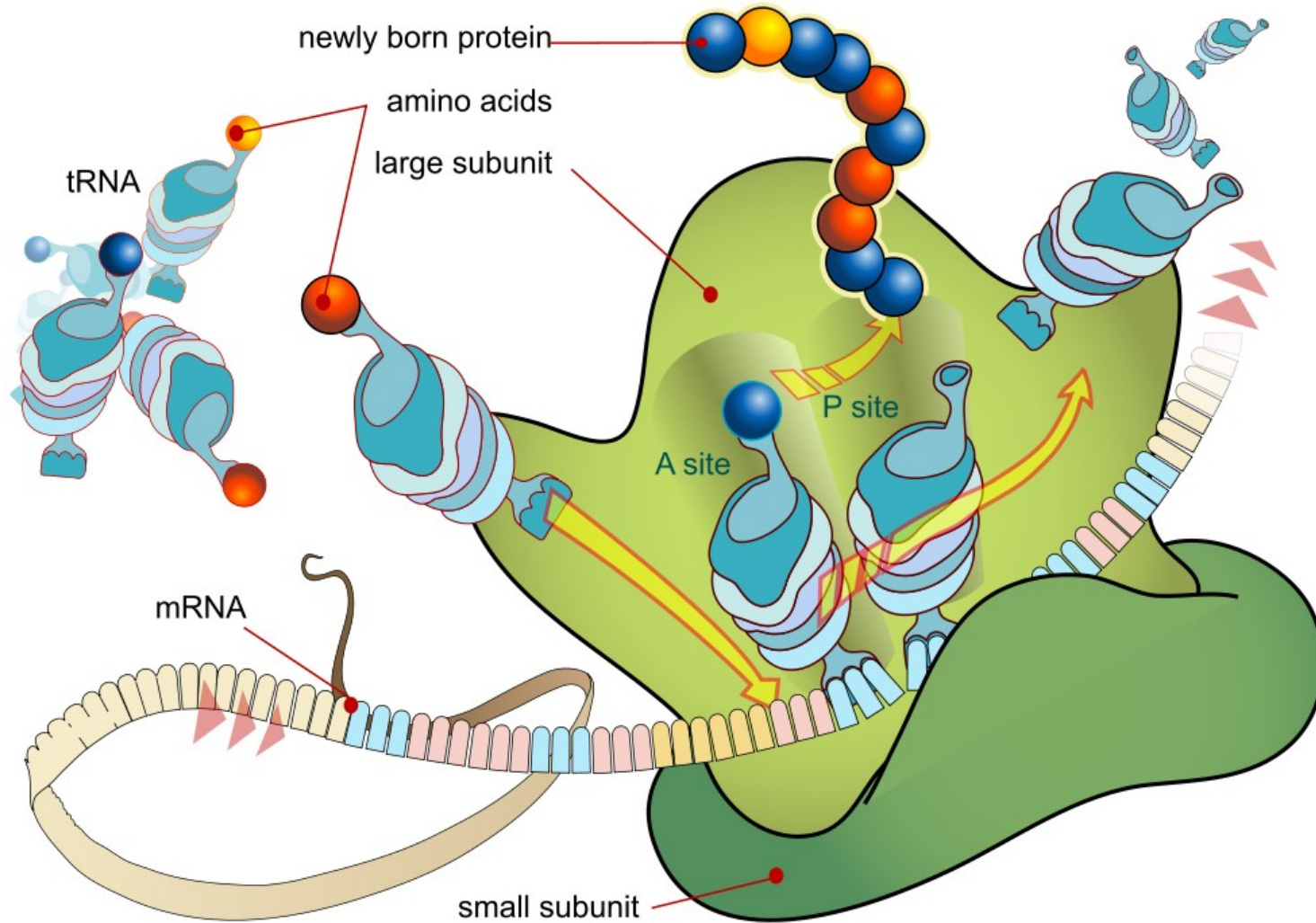


# 21 Aminosäuren

- Illustration:  
Dan Cojocari, Toronto



## Translation von mRNA in Protein am Ribosom

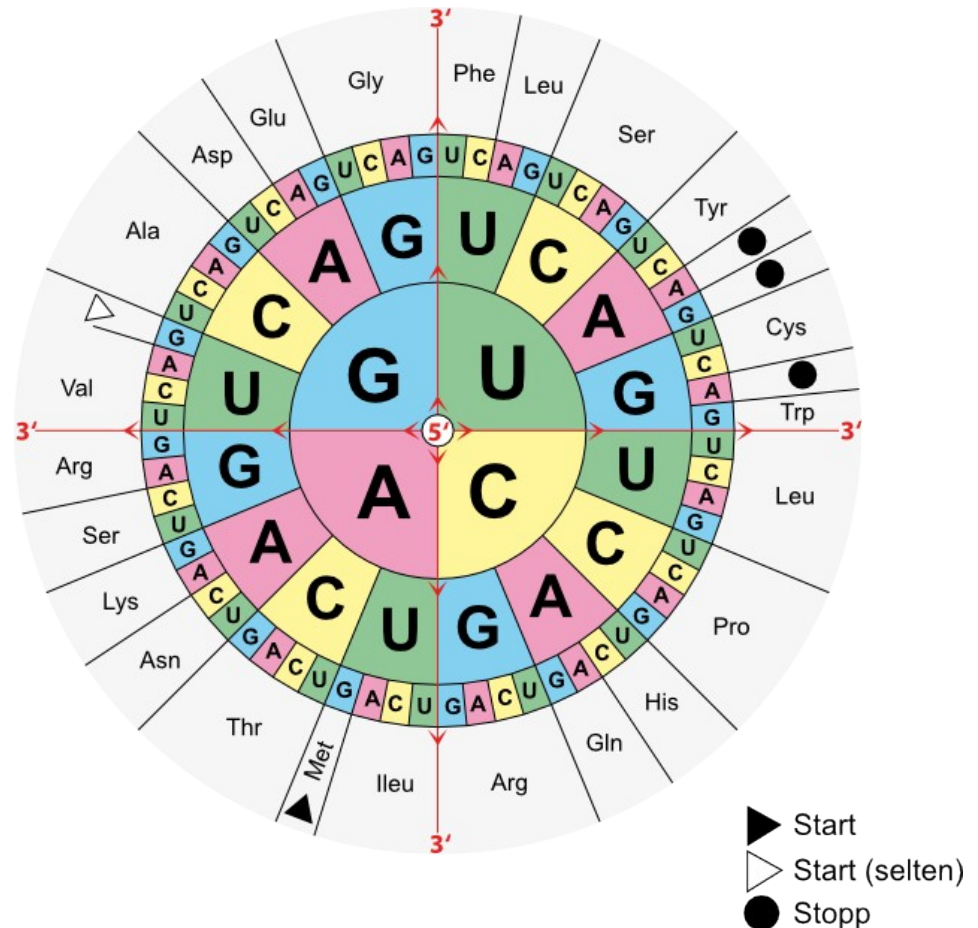


## Genetischer Code:

3 nt => 1 aa

### Formaler Übersetzungsprozess

- Lies jeweils 3 nt
- Suche dazu die passende Aminosäure
- Hänge diese aneinander, bis die DNA-Sequenz erschöpft ist.
- Beachte auch: Start, Stopp



Base1 = TTTTTTTTTTTTTTTT CCCCCCCCCCCCCCCC AAAAAAAAAAAAAAAAAA GGGGGGGGGGGGGGGG  
 Base2 = TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG  
 Base3 = TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG  
 AA = FFLLSSSSYY\*\*CC\*W LLLLPPPPHHQRRRR I IIMTTTTNKKSSRR VVVVAAAADDEEGGGG

## Übungen

- Wie lautet die Aminosäuresequenz zu
  - ATGCTTGGG ?
  - GAGATTAC ?
- Wie viele Möglichkeiten gibt es, diese Aminosäuresequenz zu erzeugen?
  - ILSW

**Base1 = TTTTTTTTTTTTTTTTTT CCCCCCCCCCCCCCCC AAAAAAAAAAAAAAAAAA GGGGGGGGGGGGGGGG**  
**Base2 = TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG TTTTCCC AAAAGGGG**  
**Base3 = TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG**  
**AA = FLLSSSSYY\*\*CC\*W LLLLPPPPHHQRRRR I IIMTTTTNKKSSRR VVVVAAAADDEEGGGG**

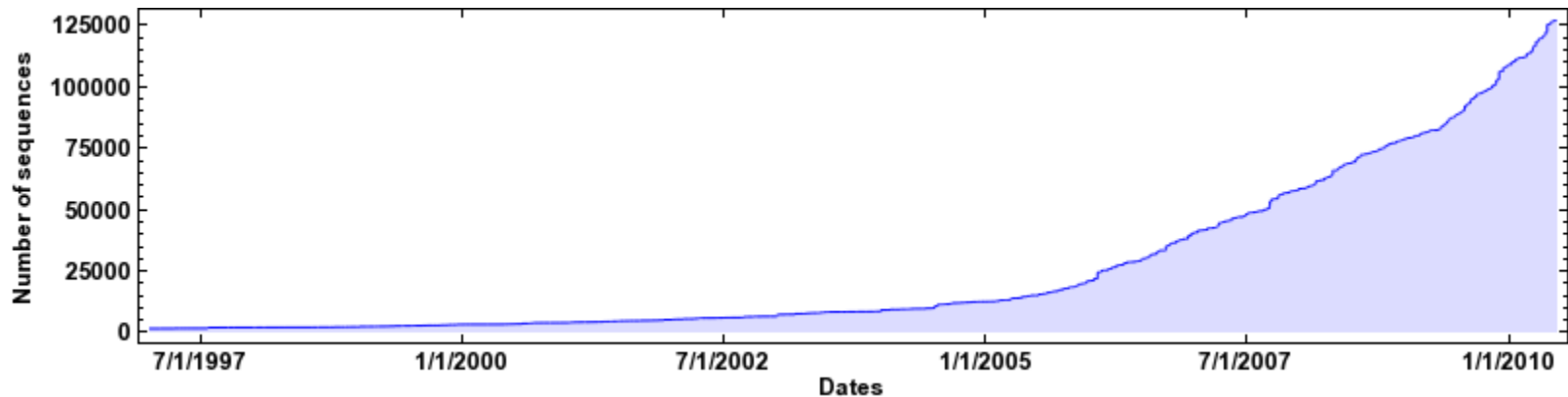
## Beispiel: Übersetzung DNA -> Protein (Nichtstrukturprotein NS1 des H1N1-Grippevirus, Patient aus Puerto Rico im Jahr 1934)

```
ATGGATCCAAACACTGTGTCAAGCTTTCAGGTAGATTGCTTTCTTTGGCATGTCCGCAAACGAGTTGCAG  
ACCAAGAAGTAGGTGATGCCCCATTCTTGATCGGCTTCGCCGAGATCAGAAATCCCTAAGAGGAAGGGG  
CAGCACTCTTGGTCTGGACATCGAGACAGCCACACGTGCTGGAAAGCAGATAGTGGAGCGGATTCTGAAA  
GAAGAATCCGATGAGGCACTTAAAATGACCATGGCCTCTGTACCTGCGTCGCGTTACCTAACCGACATGA  
CTCTTGAGGAAATGTCAAGGGAATGGTCCATGCTCATAACCAAGCAGAAAGTGGCAGGCCCTCTTTGTAT  
CAGAATGGACCAGGCGATCATGGATAAAAACATCATACTGAAAGCGAACTTCAGTGTGATTTTTGACCGG  
CTGGAGACTCTAATATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGAAATTTACCATTC  
CTTCTCTTCCAGGACATACTGCTGAGGATGTCAAAAATGCAGTTGGAGTCCTCATCGGAGGACTTGAATG  
GAATGATAACACAGTTTCGAGTCTCTGAAACTCTACAGAGATTTCGCTTGGAGAAGCAGTAATGAGAATGGG  
AGACCTCCACTCACTCCAAAACAGAAACGAGAAATGGCGGGAACAATTAGGTCAGAAGTTTGA
```

```
MDPNTVSSFQVDCFLWHVRKRVADQELGDAPFLDRLRRDQKSLRGRGSTLGLDIETATRAGKQIVERILKE  
ESDEALKMTMASVPASRYLTDMTLEEMSREWSMLIPKQKVAGPLCIRMDQAIMDKNIILKANFSVIFDRLE  
TLILLRAFTEEGAIVGEISPLPSLPGHTAEDVKNVGVVLIGGLEWNDNTVRVSETLQRFWRSSNENGRPP  
LTPKQKREMAGTIRSEV.
```

## Grippeviren - Ein Beispiel zur DNA-Sequenzanalyse

- März 2009:  
Neuer Subtyp des Influenza-A H1N1-Virus („Schweinegrippe“) tritt auf
- Isolierung und Sequenzierung des Erregers in verschiedenen Patienten
- Zentrale Datenbank am NCBI (NIH, Bethesda, Maryland, USA)  
<http://www.ncbi.nlm.nih.gov/genomes/FLU/aboutdatabase.html>

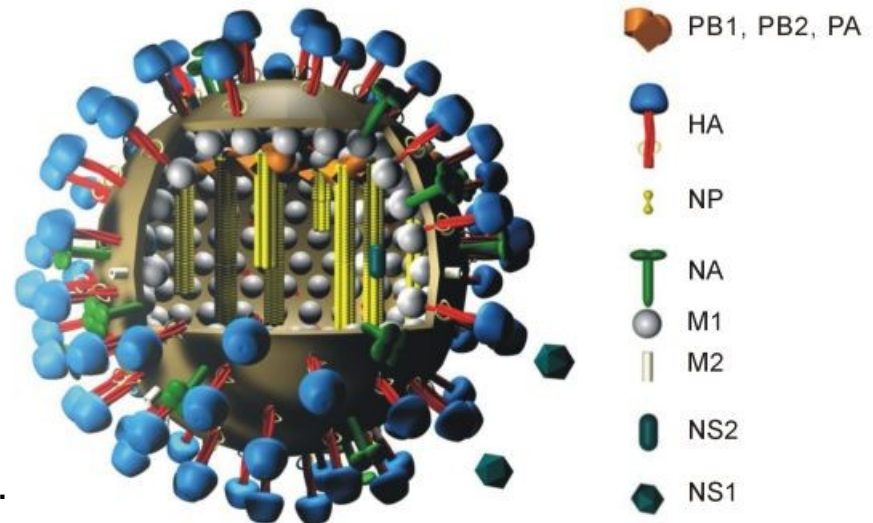


- Dort auch Vergleich mit älteren Varianten des H1N1 möglich.

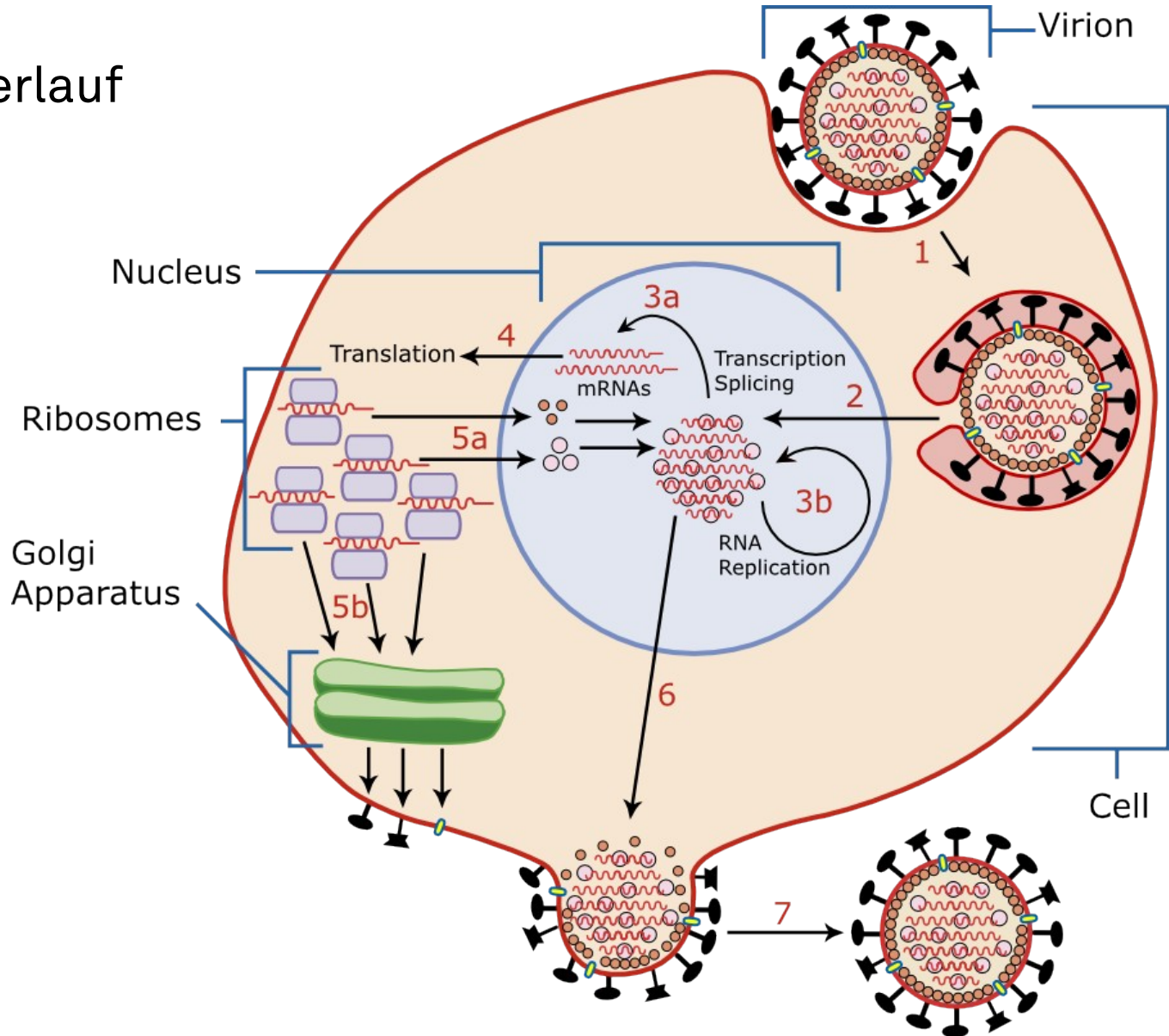
## Virus: Influenza-A H1N1

- Hämagglutinin (HA oder H),
- Neuraminidase (NA oder N),
- Nukleoprotein (NP),
- Matrixproteine (M1) und (M2),
- Polymerase Proteine (PB1, PB2, PA),
- Nichtstrukturproteine (NS1) und (NS2).

Bildquelle: Wikimedia Commons



# Infektionsverlauf



## Vergleich des NS1-Gens zweier verschiedener H1N1-Viren

- Idee: Wähle zwei möglichst unterschiedliche Viren aus Patienten aus Puerto Rico, 1934 und Taiwan, 2002  
Datenbank: <http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi>
- Sequenzen im FASTA-Format, Länge 693 nt = 231 aa

```
>gb|J02150:27-719| /Human/NS1/H1N1/8/Puerto Rico/1934/// nonstructural protein ns1
ATGGATCCAAACACTGTGTCAAGCTTTTCAGGTAGATTGCTTTCTTTGGCATGTCCGCAAACGAGTTGCAG
ACCAAGAAGTAGGTGATGCCCCATTCCTTGATCGGCTTCGCCGAGATCAGAAATCCCTAAGAGGAAGGGG
CAGCACTCTGGTCTGGACATCGAGACAGCCACACGTGCTGGAAAGCAGATAGTGGAGCGGATTCTGAAA
GAAGAATCCGATGAGGCACTTAAAATGACCATGGCCTCTGTACCTGCGTCGCGTTACCTAACCGACATGA
CTCTTGAGGAAATGTCAAGGGAATGGTCCATGCTCATAACCAAGCAGAAAGTGGCAGGCCCTCTTTGTAT
CAGAATGGACCAGGCGATCATGGATAAAAACATCATACTGAAAGCGAAGTTTTCAGTGTGATTTTTGACCGG
CTGGAGACTCTAATATTGCTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGAAATTTACCATTGC
CTTCTCTTCCAGGACATACTGCTGAGGATGTCAAAAATGCAGTTGGAGTCTCATCGGAGGACTTGAATG
GAATGATAACACAGTTTCGAGTCTCTGAAACTCTACAGAGATTCGCTTGGAGAAGCAGTAATGAGAATGGG
AGACCTCCACTCACTCCAAAACAGAAACGAGAAATGGCGGGAACAATTAGGTCAGAAGTTTGA
>gb|DQ249269:27-719| /Human/NS1/H1N1/8/Taiwan/2002/// NS1 protein
ATGGATTCCCACACTGTGTCAAGCTTTTCAGGTAAACTGCTTCCCTTTGGCATGTCCGCAAACAAGTTGCAA
ACCAAGGTCTAGGCGATGCCCCCTTTCTTGATCGGCTTCGCCGAGATCAAAGTCTCTAAAGGGAAAAGG
CAGCACTCTCGGTCTGAACATCAAACAGCCACTTGTGTTGGAAAGCAAATAGTAAAGAGGGTTCTGAAA
AAAAAATCCGATGAGGCATTTAAAATGACAATGGCCTCCGCACTTGTTCGCGGTACCTAACTGACATGA
CTATTGAAAAAATGTCAAGGGACTGGTTCATGCTCATGCCCAAGCAGAAAGTGGCTGGCCCTCTTTGTGT
CAAAATGGACCAGGCGATAATGGATAAGAACATCATACTGAAAGCGAATTTTCAGTGTGATCTTTGATCGG
TTGGAGAATCTGACATTACTAAGGGCTTTCACCGAAGAGGGAGCAATTGTTGGCGAAATTTACCATTGC
CTTCTCTTCCAGGACATACTAATGAGGATGTCAAAAATGCAATTGGGGTCTCATCGGGGACTTGAATG
GAATGATAACACAGTTTCGAGTCTCTGAAACTCTACAGAGATTCGCTTGGAGAAGCAGTAATGAGACTGGG
GGACCTCCATTCACTCCAACACAGAAACGGAAAATGGCGGGAACAATTAGGTCAGAAGTTTGA
```

## Fensterweise Untersuchung der Mutationsrate

- Mutationsrate := Anzahl der Unterschiede / Sequenzlänge
- Untersuchung nicht auf Gesamtsequenz, sondern auf „Fenstern“
- Zeigt, welche Bereiche des Gens sich stärker verändert haben, sowohl auf DNA- als auch auf Proteinebene.
- Wähle Fensterlänge  $99 = 49 + 1 + 49$  bp (Basenpaare), also  $33 = 16 + 1 + 16$  aa (Aminosäuren)
- Für jeden Fenstermittelpunkt:  
Wie viele nt / aa sind im Fenster unterschiedlich?

ATGGATCCA**AACACTGTGTCAAGCTTTCAGGTAGATTGCTTTCCTTTGGCATGTCCGCAAACGAGTTGCAG**  
**ACCAAGA**ACTAGGTGATGCCCCATT**CCTTGATCGGCTT**CGCCGAGATCAGAAATCCCTAAGAGGAAGGGG  
 CAGCACTCTTGGTCTGGACATCGAGACAGCCACACGTGCTGGAAAGCAGATAGTGGAGCGGATTCTGAAA...

vs.

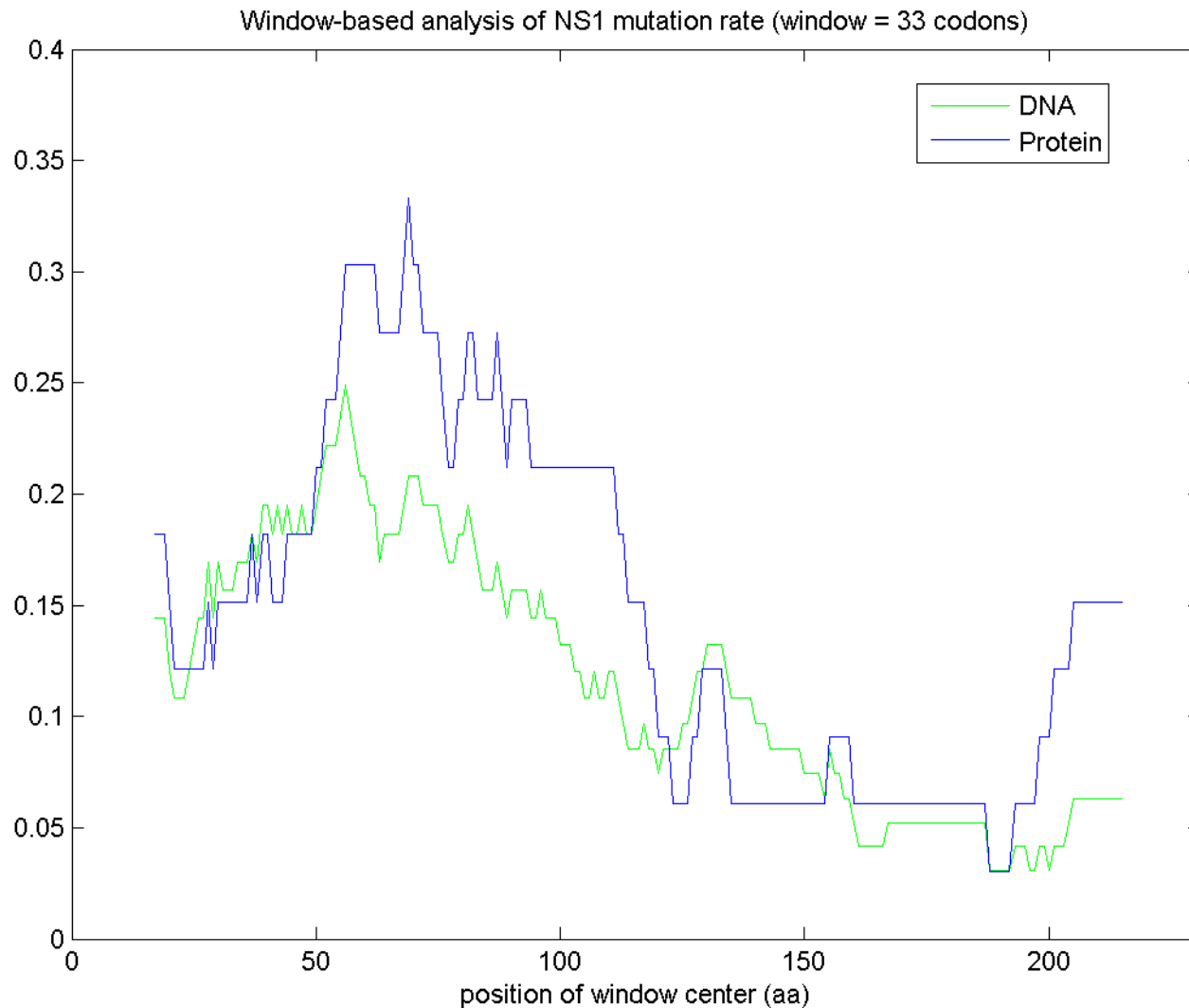
ATGGATTCC**CACACTGTGTCAAGCTTTCAGGTAAACTGCTTCCTTTGGCATGTCCGCAAACAAGTTGCAA**  
**ACCAAGGTCTAGGCGATGCCCCCTTTCCTTGATCGGCTT**CGCCGAGATCAAAAGTCTCTAAAGGGAAAAGG  
 CAGCACTCTCGGTCTGAACATCAAAACAGCCACTTGTGTTGGAAAGCAAATAGTAAAGAGGGTTCTGAAA...

MDP**NTVSSFQVDCFLWHVRKR**VAD**QELGDAPFLDRL**RRDQKSLRGRGSTLGLDIETATRAGKQIVERILK...

vs.

MDS**HTVSSFQVNCFLWHVRKQ**VAN**QGLGDAPFLDRL**RRDQKSLKKGSTLGLNIKTATCVGKQIVKRVLK...

# Mutationsrate auf DNA- und Proteinebene



## Arten von Mutationen (Mutation: Änderung der DNA-Sequenz)

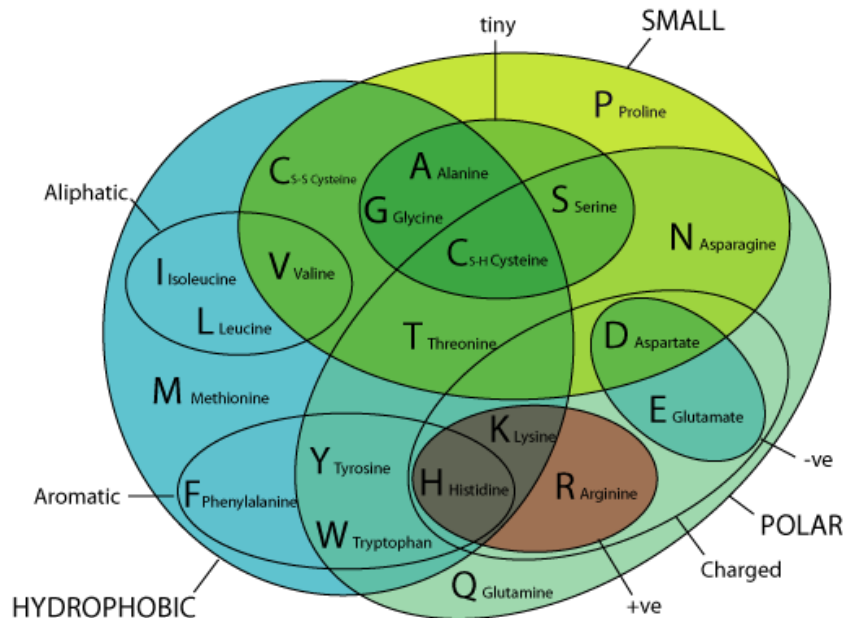
- stumm = synonym
  - Die zugehörige Aminosäure ändert sich nicht.
  - Beispiel: UUG -> UUC (beide codieren Phenlyalanin, F)
  - scheinbar keine Auswirkungen !
- nichtsynonym
  - Die zugehörige Aminosäure ändert sich.
  - Beispiel: UUU -> UUA (Phenylalanin F -> Leuzin L)
  - kann verschiedene Auswirkungen haben.
- Änderung des Leserahmens (idR sehr schädlich)
  - Einfügungen oder Löschungen , deren Länge nicht 0,3,6,... ist, verändern den Leserahmen aller folgenden Tripel und somit alle folgenden Aminosäuren.

```

Base1 = TTTTTTTTTTTTTTTTTT  CCCCCCCCCCCCCCCC  AAAAAAAAAAAAAAAAAA  GGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGG  TTTTCCCCAAAAGGGG  TTTTCCCCAAAAGGGG  TTTTCCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAG  TCAGTCAGTCAGTCAG  TCAGTCAGTCAGTCAG  TCAGTCAGTCAGTCAG
AAs   = FLLSSSSYY**CC*W  LLLLPPPPHHQRRRR  I IIMTTTTNNKKSSRR  VVVVAAAADDEEGGGG
  
```

## Arten nichtsynonymer Mutationen

- neutral: Aminosäure ändert sich, behält aber ähnliche Eigenschaften
- schädlich: Aminosäure wird durch eine mit anderen Eigenschaften ersetzt
- Stopp: Aminosäure wird durch ein Stopp-Signal (\*) ersetzt



```

ILVCAGMFYWHKREODNSTP
XXXXXXXXXXXXX ······X· Hydrophobic
·······XXXXXXXXXX·X Polar
·XXXX········XXXXX Small
····················X Proline
····XX··············X· Tiny
XXX·················· Aliphatic
·······XXXX·········· Aromatic
·········XXX········· Positive (+)
···············X·X···· Negative (-)
···········XXXX·X···· Charged (+/-)

```

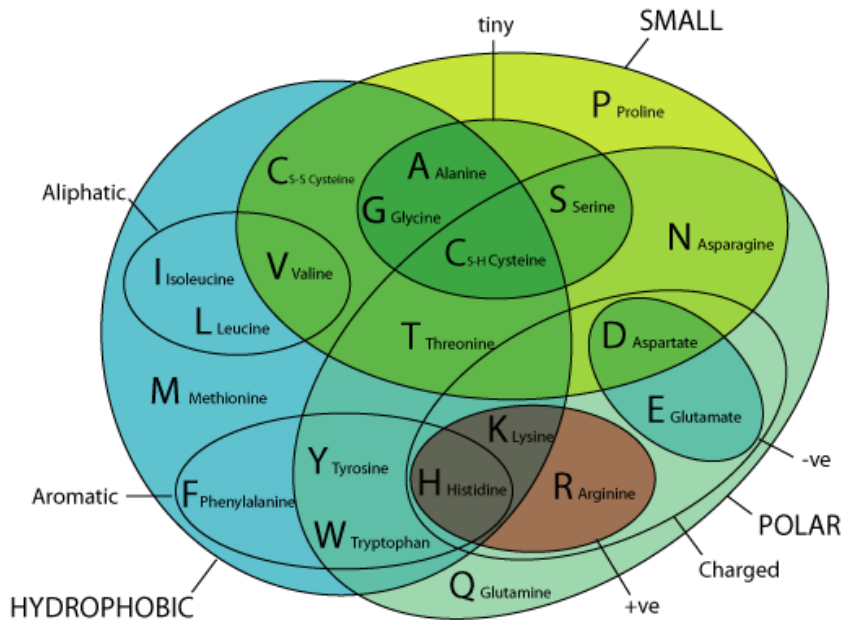
```

Base1 = TTTTTTTTTTTTTTTTTT CCCCCCCCCCCCCCCC AAAAAAAAAAAAAAAAAA GGGGGGGGGGGGGGGG
Base2 = TTTTCCCAAAGGGG TTTTCCCAAAGGGG TTTTCCCAAAGGGG TTTTCCCAAAGGGG
Base3 = TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG
AAs = FLLSSSSYY**CC*W LLLLPPPHHQRRR I IIMTTTTNKKSSR VVVVAAAADDEEGGG

```

## Übungen: Welche Art Mutationen?

- TCG ATT AGG TGG CCC GAG  
AGC CTT CGG TGA CGC GAT



ILVCAGMFYWHKREODNSTP  
 XXXXXXXXXXXX ······X · Hydrophobic  
 ······XXXXXXXXXX ·X Polar  
 ·XXXX ······XXXXX Small  
 ······X Proline  
 ···XX ······X · Tiny  
 XXX ······ Aliphatic  
 ······XXXX ······ Aromatic  
 ······XXX ······ Positive (+)  
 ······X ·X ······ Negative (-)  
 ······XXXX ·X ······ Charged (+/-)

Base1 = TTTTTTTTTTTTTTTT CCCCCCCCCCCCCC AAAAAAAAAAAAAAAAAA GGGGGGGGGGGGGGGG  
 Base2 = TTTTCCCAAAGGGG TTTTCCCAAAGGGG TTTTCCCAAAGGGG TTTTCCCAAAGGGG  
 Base3 = TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG TCAGTCAGTCAGTCAG  
 AAs = FFLSSSSYY\*\*CC\*W LLLLPPPPHHQORRRR I IIMTTTTNNKKSSRR VVVVAAAADDEEGGGG

## Synonyme und nichtsynonyme Stellen

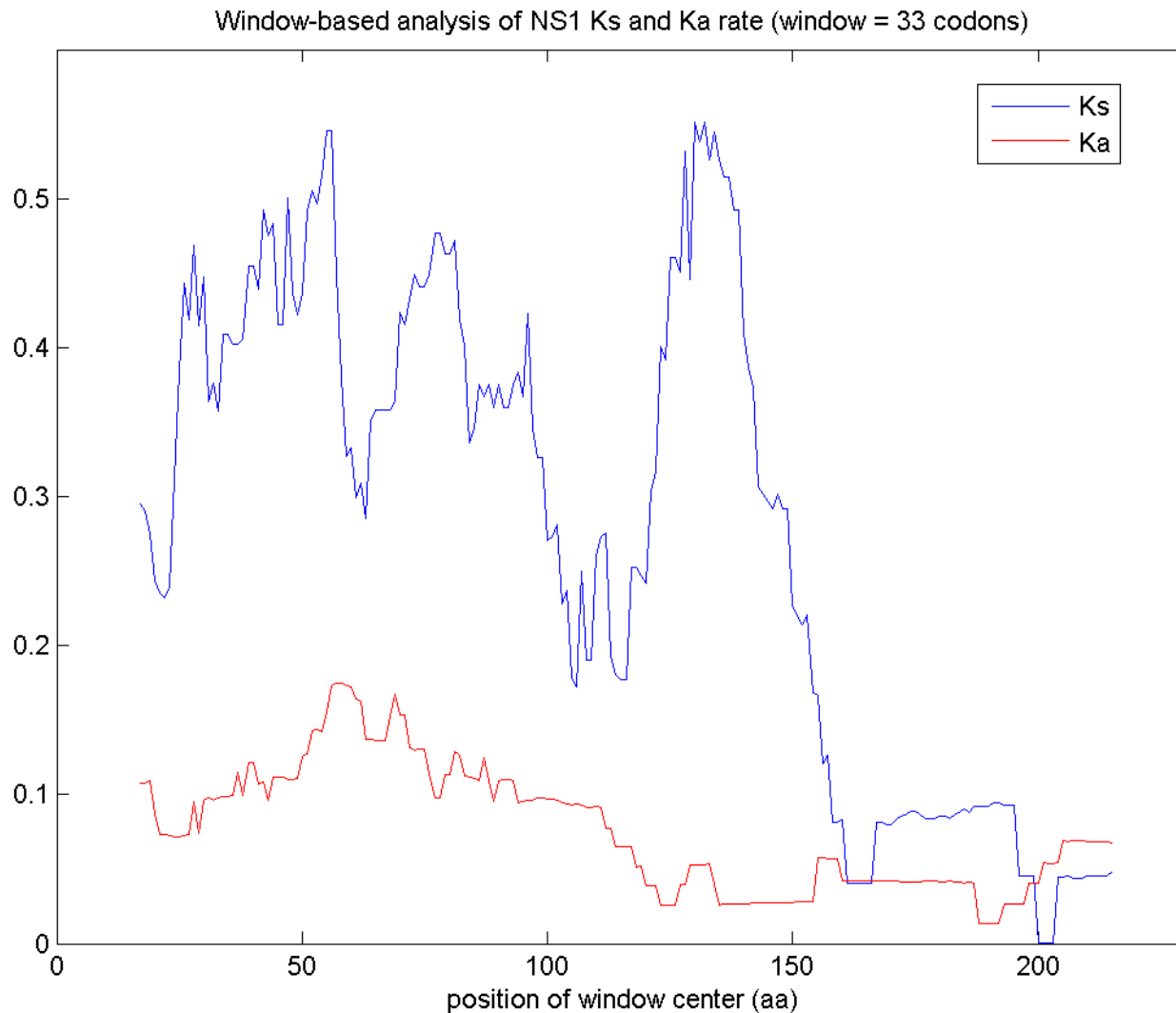
- Betrachte ein Codon (mit seiner Aminosäure).  
 Welche Tendenz besteht zu synonymer/nichtsynonymer Mutation ?
- Betrachte die 9 Codons,  
 die durch eine Substitution an einer der drei Stellen entstehen.  
 Berechne die Anzahl synonyme Codons dieser 9.  
 Dividiere durch 3, um die „Anzahl synonyme Stellen“ zu erhalten.
- Beispiel: TTA (Leucin)
 

ATA	I	CTA	L
GTA	V	TAA	.
TCA	S	TGA	.
TTC	F	TTG	L
TTT	F		
- 2 von 9 stimmen mit Leucin überein,  
 d.h. 2/3 synonyme Stellen, 7/3 nichtsynonyme Stellen

## Synonyme und nichtsynonyme Mutationsraten

- Stellen in einem Fenster:  
Summiere separat synonyme und nichtsynonyme Stellen über alle Codons.
- Für das selbe Fenster in zwei Sequenzen:  
Berechne Durchschnitt über beide Sequenzen.  
→ Wahrscheinlichkeit, dass hier eine (nicht)synonyme Mutation geschieht.
- Mutationen im Fenster:  
Zähle separat synonyme und nichtsynonyme Mutationen im Fenster.
- Berechne separate Mutationsraten:
  - synonym:  $K_S = \text{synonyme Mutationen} / \text{synonyme Stellen}$
  - nichtsynonym:  $K_A = \text{nichtsynonyme Mutationen} / \text{nichtsynonyme Stellen}$

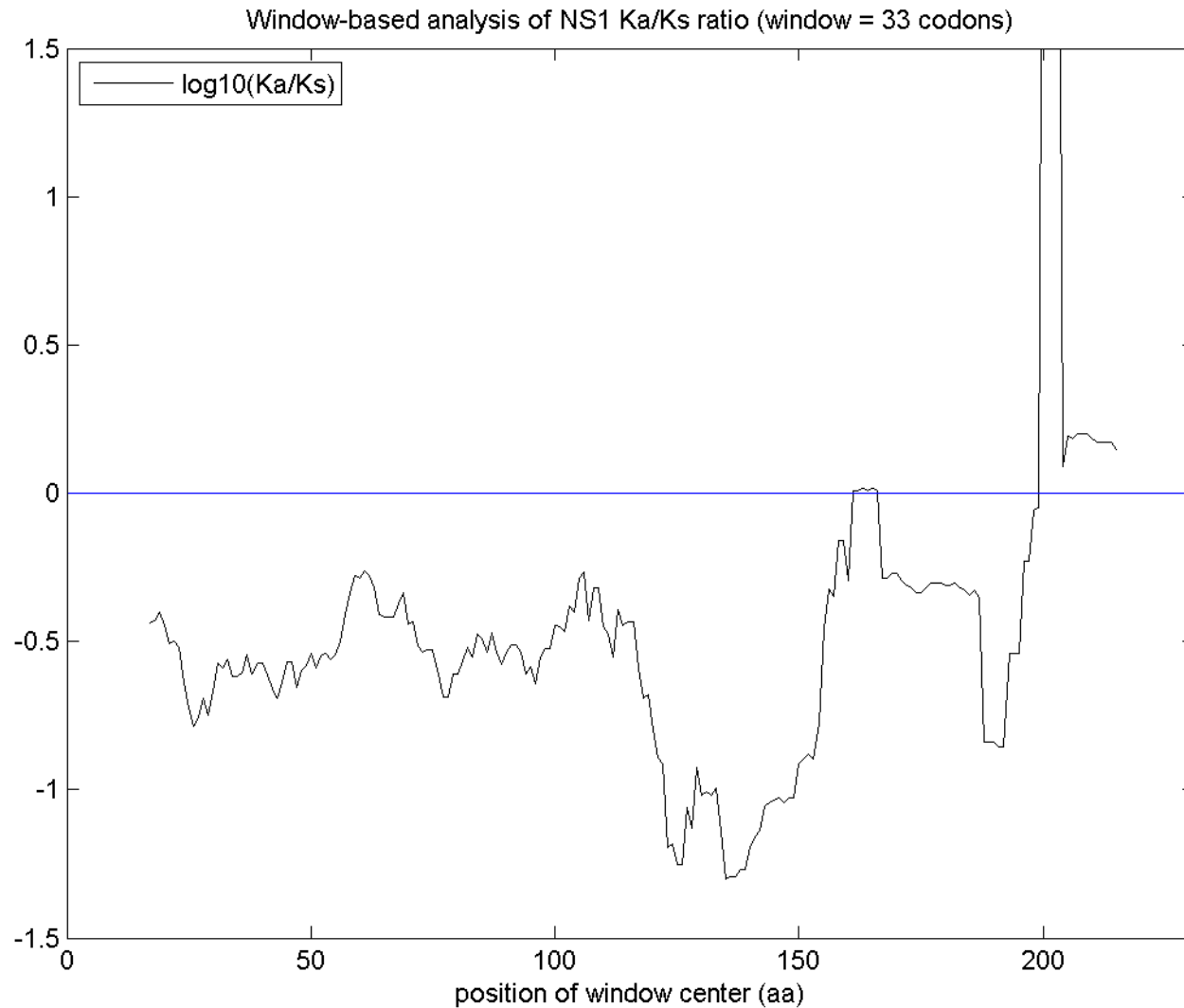
## Ks- und Ka-Rate im NS1-Protein



## KA/KS - Analyse

- Betrachte die Verhältnisse  $K_A$  (nichtsynchron) /  $K_S$  (synonym).
- Verhältnis  $> 1$ : mehr nichtsynonyme Mutationen pro nichtsynonymer Stelle als synonyme Mutationen pro synonyme Stelle
- Verhältnis  $< 1$ : entsprechend
  
- Was bedeutet das?
- Verhältnis  $< 1$  legt den Schluss nahe, dass das Protein unter negativer Selektion steht. Veränderungen der Aminosäure-Sequenz wirken sich vermutlich direkt negativ auf die Fitness des Proteins aus.
  
- Statt Verhältnis  $>= < 1$  betrachte:  $\log(\text{Verhältnis}) >= < 0$

## KA/KS-Analyse des NS1-Gens



## Erkenntnisse

- NS1-Gen des H1N1-Virus steht vorwiegend unter negativer Selektion.
- Funktion ist gestört schon bei wenigen nichtsynonymen Änderungen.
- NS1-Protein nicht übermäßig variabel.
- Gute Nachricht: Virus kann sich Medikamenten, die NS1 angreifen, nur schwer durch Mutationen entziehen.



## Fragen?

- Das zentrale Dogma der Molekularbiologie:  
DNA → RNA → Protein;  
Transkription und Translation
- Translation als formaler Übersetzungsprozess:  
Wörterbücher (dictionaries), Funktionen, Iteratoren, Generatorfunktionen
- Grippeviren und ihr Infektionsverlauf,  
z.B. Influenza A H1N1
- Einfache fensterweise Mutationsanalyse eines H1N1-Proteins
- Detaillierte KA/KS-Analyse:
  - synonyme und nichtsynonyme Mutationen
  - synonyme und nichtsynonyme Stellen
  - synonyme und nichtsynonyme Mutationsrate
- Interpretation der Ergebnisse



## Diskussion: Wirkungen stummer Mutationen

- Stumme Mutationen ändern die Proteinsequenz nicht.
- Also gar kein Effekt?

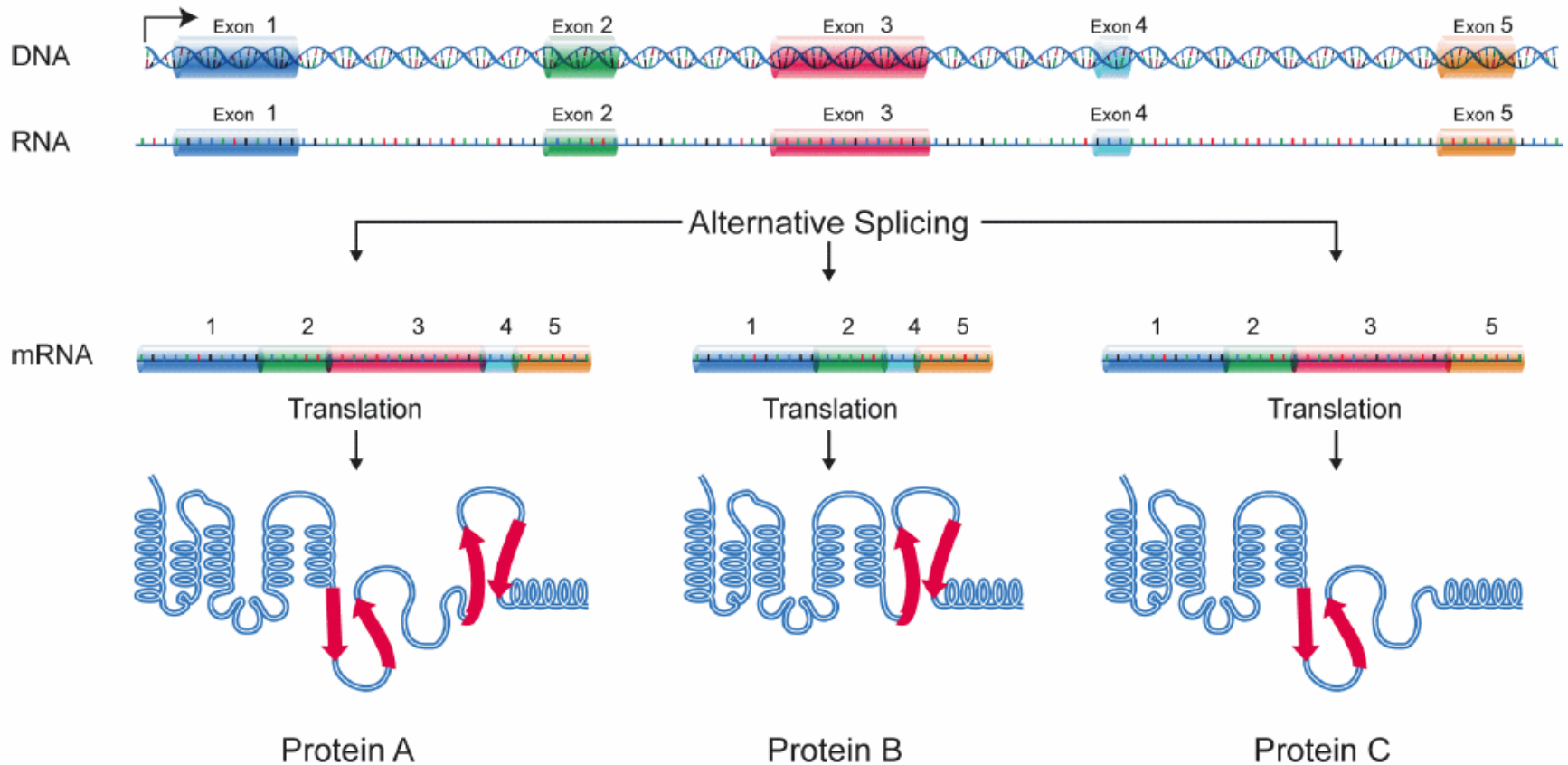
## Diskussion: Wirkungen stummer Mutationen

- Stumme Mutationen ändern die Proteinsequenz nicht.
- Also gar kein Effekt? Nein!
  
- Effekte bei der Translation (Proteinsynthese)
  - Andere tRNAs sind nötig, um das Protein zu synthetisieren.
  - Codons und tRNAs für die selbe Aminosäure sind nicht gleich häufig.
  - Häufiges → Seltenes Codon: Proteinsynthese wird verlangsamt
  - Protein kann sich u.U. anders falten → defekt.

## Diskussion: Wirkungen stummer Mutationen

- Stumme Mutationen ändern die Proteinsequenz nicht.
- Also gar kein Effekt? Nein!
- Effekte bei der Translation (Proteinsynthese)
  - Andere tRNAs sind nötig, um das Protein zu synthetisieren.
  - Codons und tRNAs für die selbe Aminosäure sind nicht gleich häufig.
  - Häufiges → Seltenes Codon: Proteinsynthese wird verlangsamt
  - Protein kann sich u.U. anders falten → defekt.
- Weitere Effekte
  - DNA codiert nicht nur Proteine, sondern enthält auch andere Information.  
Beispiel: Spleiß-Signale
  - Solche Funktionen können durch Mutation beeinträchtigt sein.

## (Alternatives) Spleißen von mRNA



## Marfan-Syndrom

- Protein Fibrilin-1 (Bindegewebe)
- Gen FBN1 auf Chr15
- Dominant vererbt
- dbSNP:rs112989722  
Stumme Mutation
  - 6354C-T
  - ILE2118ILE
  - EX51DEL
- Vermutung: Stumme Mutation beeinflusst das Spleißen des Gens. Exon 51 geht verloren: Protein nicht voll funktionsfähig.

