

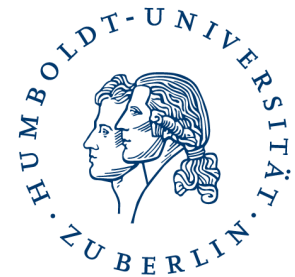
Algorithmische Bioinformatik

Wintersemester 2018 / 2019

Master: 10 SP Modul

Ulf Leser

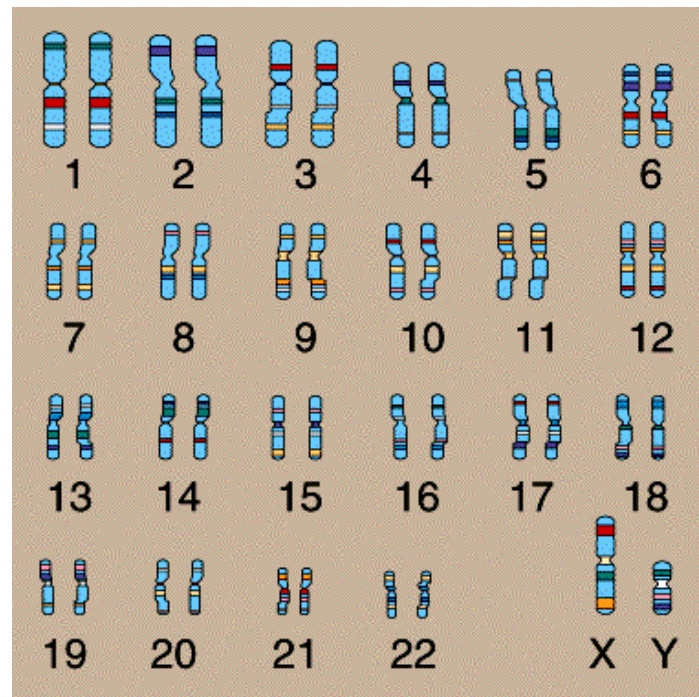
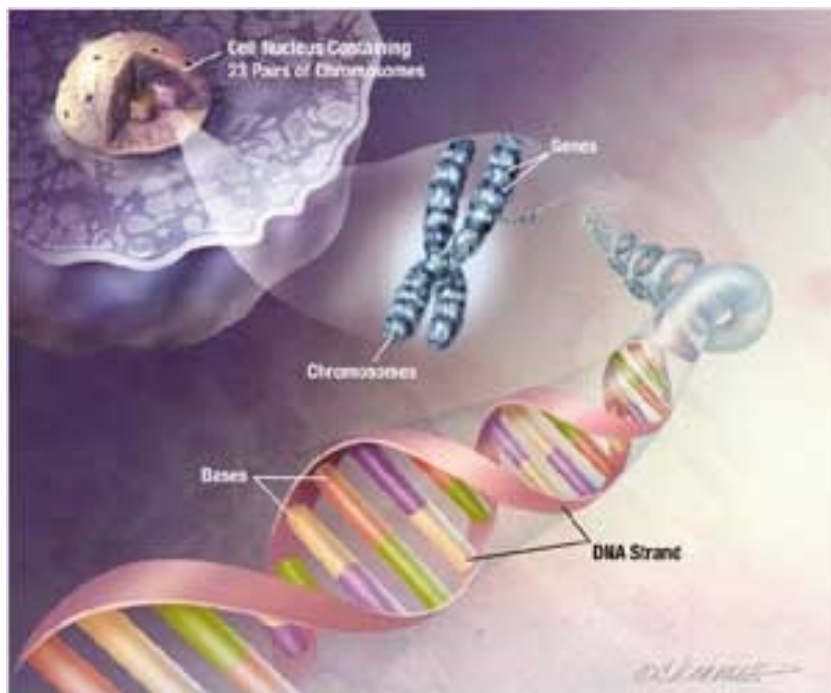
Wissensmanagement in der
Bioinformatik



Ziele für heute

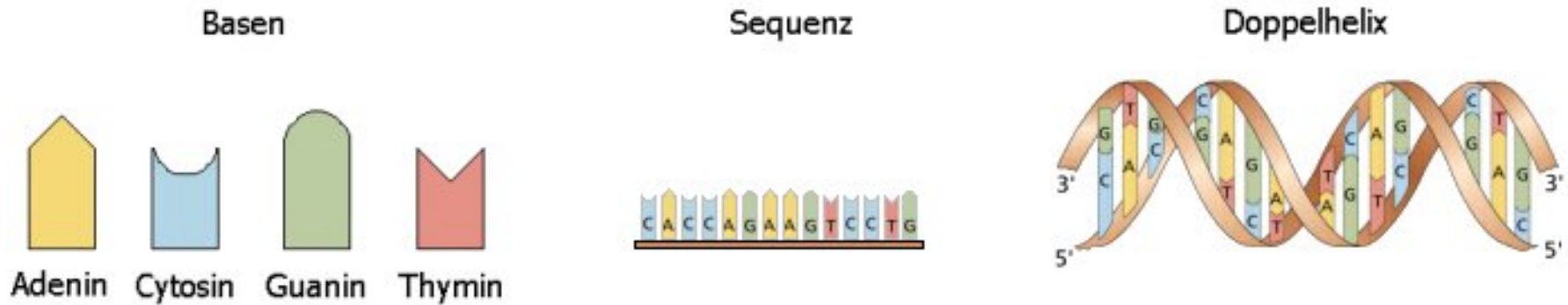
- Lust auf das Thema machen
- Gefühl für Rasananz der Entwicklung geben
- Überblick über die Vorlesung

Genome \approx String



- Human genome: app. 3.000.000.000 letters $\in \{A,C,G,T\}$

DesoxyriboNucleicAcid



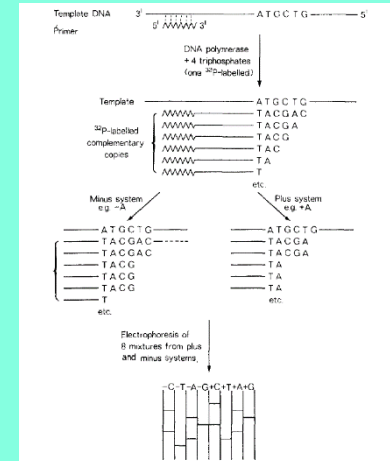
- DNA: Desoxyribonukleinsäure
- Träger der vererbten Information – Genom
- **Alles Leben** verwendet DNA (RNA) aus den selben 4 (5) Molekülen

Fast Development



1953
Double helix structure of DNA,
Watson/Crick

1975
High-throughput
sequencing
Sanger/ Coulson



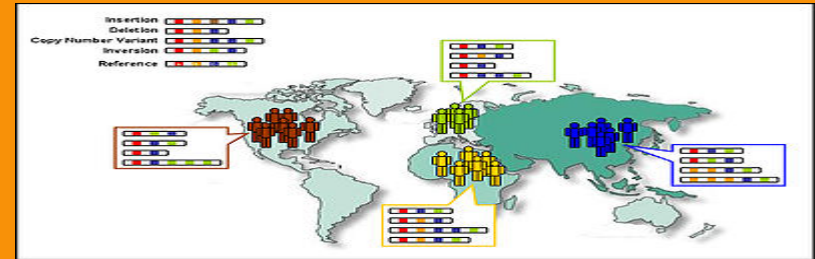
2003
First human genome sequenced
Took ~14 years, ~3 billion USD

Fast Development 2



2008

Genome of J. Watson finished
4 Months, 1.5 Million USD



2010

1000 Genomes Project

1000GP releases more data in first 6 months than EMBL collected in the 25 years before

Large Scale Sequencing Projects



ICGC

50.000 samples: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

Genomics
england



Genomics England ... is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: [the 100,000 Genomes Project](#).



The Veterans Affairs (VA) Office of Research and Development is launching the [Million Veteran Program \(MVP\)](#) The goal of MVP is to better understand how genes affect health and illness in order to improve health care.

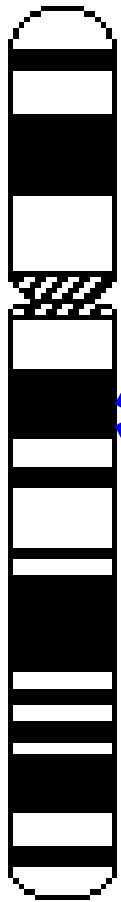
Was ist ein Gen?

Chromosom

DNA

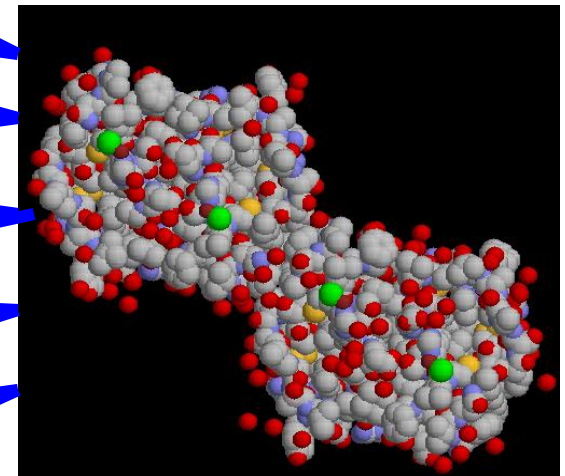
RNA

Protein



A
C
G
T
T
G
A
T
G
A
C
C
A
G
A
G
C
T
T
G
T

A
C
G
U
U
G
A
C
A
G
A
G
C
U
U
C
U

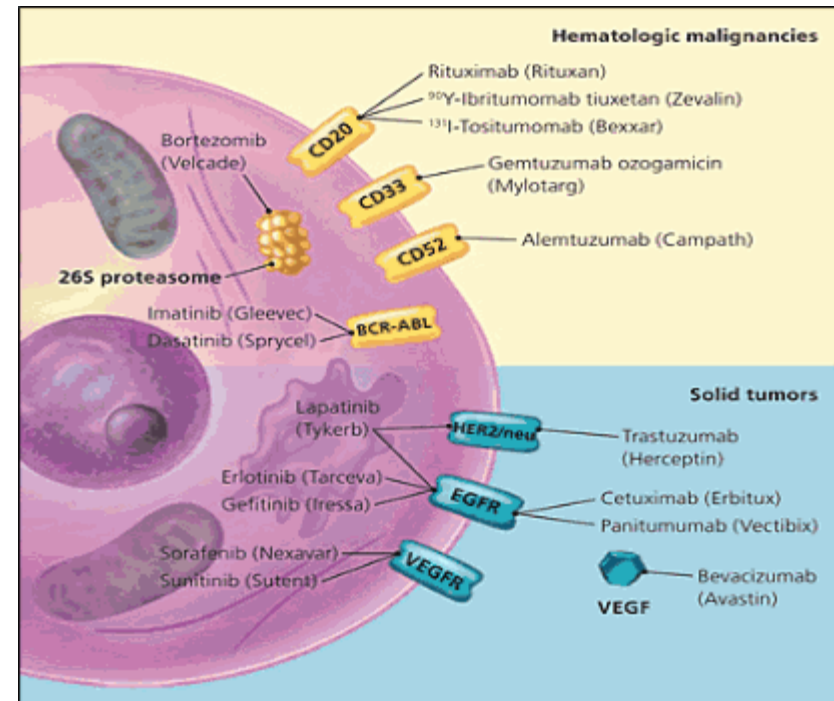


Genomics in a Nutshell

- **~2% are coding** – genes being translated into proteins
 - Whole Genome Sequencing – WGS
 - Whole Exome Sequencing - WES
- **~20.000 genes**, forming maybe 500K different proteins
 - ~3000 genes are conserved since ever (yeast)
 - We share ~95% of our genes with mice, rats, dogs, ...
 - ~25% of our genes have a still unknown function
- It's not only genes: **miRNA, enhancer, binding sites, chromatin structure, epigenomics, ...**

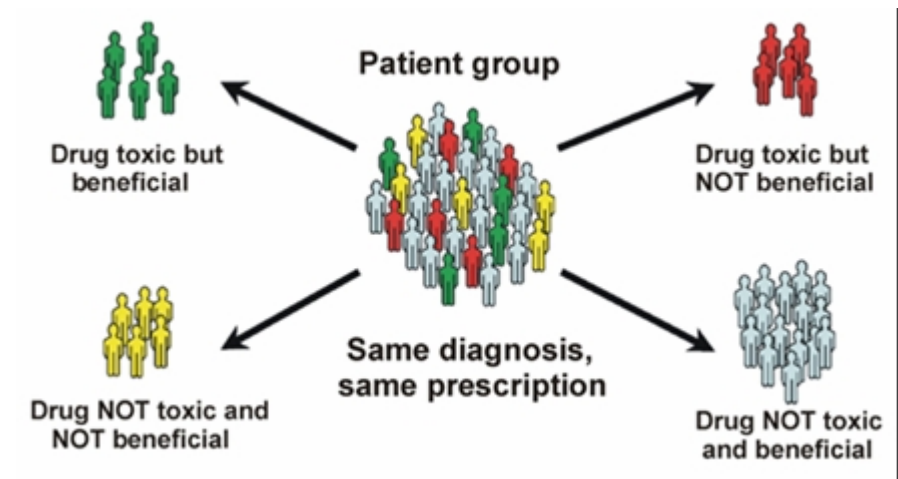
Genomics for Medicine

- **Cancer**, immunology, genetic diseases, infections
- **Cancer**
 - Cells proliferating uncontrolled, leaving their tissue
 - What goes wrong? Cell division, DNA repair, surface adhesion, cellular signaling
 - ~200 core cancer genes
- Targeted therapy: Drug attacking a **mutated gene**
- “Cancer is becoming a chronic diseases”



Precision Medicine, Personalized Medicine

- Tailor treatment to the **individual patient's genome**
- "Genome" – **mutation profile**
 - We know 10s of Millions of human mutations
 - Mutation – deviation from the norm?
 - Mutation – genomic subsequence rarely seen
- Requires **many genomes**
 - What is rare?
 - Often enough to obtain a **statistically robust association**
 - Most effects involve many mutations / genes
 - Combinatorial explosion



Wofür kann man sie benutzen?

- Kommerzielle Gentests im WWW

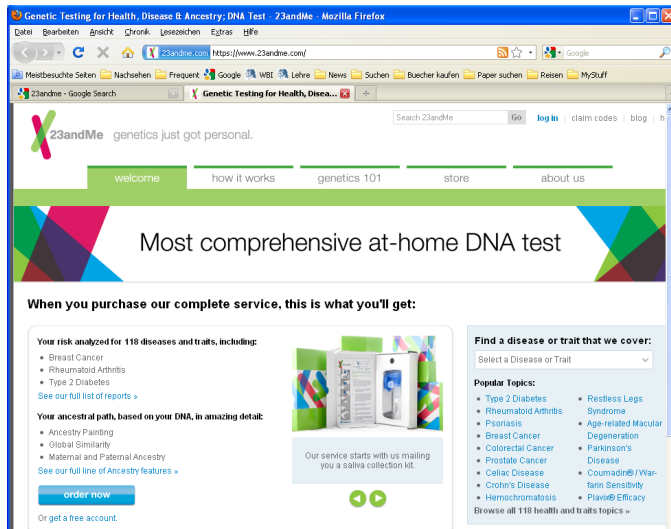
- 2 Firmen
- 32 Tests
- Preise: 100–1400€

Tabelle

Gentests, die im Internet in Deutschland bestellbar sind (Stand: Juli 2002)

Indikation*	Anbieter**	Untersuchungsgegenstand	Preis (inkl. MwSt.)
Alkoholverträglichkeit	2	keine Angaben (k. A.)	207,79 €
Alzheimer	2	k. A.	134,06 €
Alzheimer ³⁾	1	E4-Allel des Apolipoprotein-E-Gens auf Chromosom 10	650,00 €
Angelman-Syndrom ²⁾	1	Deletion auf dem Chromosom 15	850,00 €
Anti-Aging-Risikoprofil	2	k. A.	653,61 €
Arteriosklerose/Herzinfarkt/Schlaganfall	2	k. A.	512,81 €
Azosemie	1	31 Mutationen einschließlich einer 5T-Variante auf dem CFTR-Gen auf dem Chromosom 7	850,00 €
Bluthochdruck	2	k. A.	127,40 € 439,24 €
Diabetes Typ 2	2	k. A.	127,40 € 194,39 €
Dickdarmkrebs ³⁾	1	MLH1- und MSH2-Mutationen	1600,00 €
Entgiftungsfähigkeit	2	k. A.	811,10 €
Faktor V Leiden-Mutation	1	Gerinnungsfaktor-V auf dem langen Arm von Chromosom 1	400,00 €
Familiäre Hypercholesterinämie	1	Mutationen im Low-Density-Lipoprotein-Rezeptor-Gen und im Exon 26 Apolipoprotein-B-Gen	850,00 €
Familiäre Hyperlipoproteinämie Typ III	1	E2-Allel des Apolipoprotein-E-Gens auf Chromosom 19	500,00 €
Familiärer Brustkrebs ³⁾	1	BCRA1- und BCRA2-Mutationen	1400,00 €
Fettgen/Adipositas	2	k. A.	241,35 € 576,44 €
Fettstoffwechsel/Cholesterin	2	k. A.	395,48 €
Fragiles X-Syndrom ⁴⁾	1	FMR1-(fragile X mental retardation-)Gen des X-Chromosoms (Region Xq27.3)	950,00 €
Hämochromatose	2	k. A.	207,84 €
Hämochromatose	1	Austausch der DNS-Basen Guanin zu Adenin an der Position 845 und von Cytosin zu Guanin an der Position 187 des HFE-Gens auf dem Chromosom 6	500,00 €
Hyperhomocysteinämie	1	k. A.	550,00 €
Mukoviszidose (Cystische Fibrose)	1	Mutation eines Gens auf Chromosom 7	850,00 €
Muskeldystrophie	1	Deletionen (Verlust von DNA-Teilsequenzen) im Dystrophin-Gen auf dem X-Chromosom	850,00 €
Osteoporose	2	k. A.	103,89 € 191,01 €
Osteoporose	1	Mutation (Basenaustausch von Guanin zu Thymin) im Intron 1 des Kollagen Typ I Alpha 1-Gens	650,00 €
Ovarialkarzinom ³⁾	1	BCRA1- und BCRA2-Mutationen	850,00 €
Persönliches Ernährungsprofil	2	k. A.	841,32 €
Prader-Willi-Syndrom	1	Deletion oder Translokation auf dem langen Arm des Chromosoms 15 (15q11)	850,00 €
Prothrombin-Mutation	1	Austausch der DNS-Basen Guanin zu Adenin an der Position 20210 des Prothrombingens auf dem Chromosom 11	550,00 €
Risiko Alkohol- und Drogenabhängigkeit	2	k. A.	274,86 €
Thrombose	2	k. A.	134,06 € 281,52 €

State of the "Art"



- 6/2010: „Gentest-Firma vertauscht DNA-Ergebnisse ihrer Kunden“ (Nature Blog)
- 7/2010: US general accounting office compared 15 (4) companies: totally **contradicting results**

Modul Algorithmische Bioinformatik

- Vorlesung 4 SWS
- Übung 2 SWS

- Sprechstunde: Nach Vereinbarung
Ulf Leser
Raum: IV.105
Tel: (030) 2093 – 3902
eMail: leser (..) informatik . hu-berlin . de

Termine und Prüfung

- Vorlesung
 - Montag, 13-15 Uhr
 - Donnerstag, 9-11 Uhr
- Übung
 - Montag, 11-13 Uhr
 - Donnerstag, 11-13 Uhr
- **Erste Übung für alle:** Donnerstag, 18.10.2015, 11 Uhr
- Voraussetzung für Prüfung
 - Bestehen aller Übungsaufgaben
 - **Verständnis der Algorithmen**

Gäste

- Die vier nächsten Doppelstunden
 - Raik Otto
 - Zellen, Chromosomen, Gene, Transcription und Translation, differentielles Splicen, ...
- Im Verlaufe des Semesters
 - ?

Literatur

- Primär
 - Dan Gusfield: „Algorithms on Strings, Trees, and Sequences“, Cambridge University Press, 1997 (ca. 60 Euro)
- Weitere
 - Ohlebusch: "Bioinformatics Algorithms", Verlag Enno Ohlebusch.
 - David Mount: „Bioinformatics. Sequence and Genome Analysis“, Cold Spring Harbour Press, 2001 (ca. 70 Euro)
 - Gibson & Muse: „A primer of genome sciences“, Sinauer Associates, 2001 (ca. 50 Euro)
- Sowie **Originalliteratur**

Webseite

The screenshot shows a Mozilla Firefox browser window with the following details:

- Title Bar:** Halbkurs Algorithmische Bioinformatik -- Mozilla Firefox
- Menu Bar:** Datei, Bearbeiten, Ansicht, Chronik, Lesezeichen, Extras, Hilfe
- Address Bar:** zope.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1112/hk_algbio/
- Search Bar:** Google
- Navigation Bar:** Meistbesuchte Seiten, Nachsehen, Frequent, WBI, Lehre, Google, News, Projekte, Buecher kaufen, Paper suchen, Reisen, MyStuff, Lymphomexplorer
- Page Content:**
 - Logo:** WBI (with a DNA helix icon)
 - Section:** WS 11/12
 - Page Title:** Halbkurs Algorithmische Bioinformatik
 - Author:** Professor Ulf Leser
 - Description:** Der Halbkurs "Algorithmische Bioinformatik" behandelt Algorithmen zur Lösung grundlegender Fragestellungen moderner Molekularbiologie. Nach einer Einführung in die Grundlagen der Molekularbiologie (Gene und Genome, Expression, Proteine, Regulation und Transkription) werden die folgenden algorithmischen Probleme behandelt: Exaktes Stringmatching, Stringmatching mit mehreren Pattern, approximatives Matching, Indexstrukturen für Sequenzdatenbanken, Editabstand und Alignment, Multiples Alignment, Phylogenetische Bäume. Die Algorithmen werden jeweils anhand der zugrundeliegenden biologischen Fragestellung erklärt, wie z.B. Patternsuche in DNA- und Proteinsequenzen, Assembly von Teilsequenzen, Homologiesuche in Sequenzdatenbanken, und Berechnung evolutionärer Stammbäume. Die Vorlesung wird durch eine Übung begleitet.
 - Erste Vorlesung:** ist am Mittwoch, den 19.10.2011.
 - Voraussetzungen:** Voraussetzung für den Besuch sind gute Kenntnisse in Algorithmen (z.B. Modul Algorithmen & Datenstrukturen). Kenntnisse in der Molekularbiologie werden nicht vorausgesetzt, sondern vermittelt.
 - Prüfungen:** Prüfungen sind mündlich. Die Vorlesung ist anrechenbar für:
 - Diplomstudiengang Informatik, Halbkurs praktischen Informatik, 8SP
 - Master Informatik, 8SP
 - Masterstudiengang Biophysik, Vertiefungsrichtung Bioinformatik, 8SPVoraussetzung für die Zulassung zur Prüfung ist das Bestehen der Übung.
 - Literatur zur Vorlesung:** Dan Gusfield: "Algorithms on Strings, Trees, and Sequences", Cambridge University Press. Die Vorlesung folgt in grossen Teilen diesem Buch. Zusätzliche Literatur wird in den jeweiligen Stunden angegeben.
 - Themen und Termine im Einzelnen:** (Folien sind hier jeweils vor der Vorlesung als PDF verfügbar. Änderungen möglich).
 - 19.10.11: Einleitung und Überblick

- Right Sidebar (Institut für Informatik):**
- English
- Wissensmanagement in der Bioinformatik
- Kontakt
- Mitarbeiter
- Veranstaltungen
- Lehre
- Archiv
- WS 11/12
- Modul Data Warehousing und Data Mining
- Übung Data Warehousing und Data Mining
- Halbkurs Algorithmische Bioinformatik
- Übung Algorithmische Bioinformatik
- Seminar Data and Text Mining
- Forschungsseminar WBI - DBIS
- SS 11
- WS 10/11
- SS 10
- WS 09/10
- SS 09
- WS 08/09
- SS 08
- WS 07/08
- SS 07
- Search Bar (Bottom):** Suchen: publication

Ihre Bewertung 13/14

	Freundlich	Fragen	Sprache	Präsentation	Beispiele	Konzeption	Überblick	Viel neues	Kritische Auseinan	Nützlich	Lernziele	Materialien	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	Abweichung vom C	Abweichung pro F1	Alter	Geschlecht	Gefehlt	Teilnehmerzahl	Warum kommen?	Studiengang	Fachsemester	Korrektur 6er	Korrektur 3er	
1																													
4	5	5	5	5	4	5	4	5	5	4	5	4	3	3	4	4	2	2	20	1,18	23	M	1	3	3,4	M	7	0	0
5	6	6	6	5	5	6	6	6	6	6	6	5	3	4	4	1	1	5	0,31	20	M	0	3	2,3	BA	3	1	0	
6	6	6	6	6	6	6	6	6	6	6	6	6	3	4	4	1	1	2	0,12	27	M	3	2	2,3	DI	13	0	0	
7	6	5	6	5	5	5	6	5	5	6	5	6	4	4	4	1	2	11	0,65	23	M	2	2	2,3	M	9	0	0	
8	Durchschnitt	5,83	5,67	5,83	5,50	5,33	5,67	5,67	5,67	4,80	5,40	5,60	5,33	3,17	3,67	3,67	1,17	1,33		25,7		1,8	2,7			6,2			
9	Wunschzahl	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	3,00	3,00	3,00	1,00	1,00					3,0						
10	Abweichung	0,17	0,33	0,17	0,50	0,67	0,33	0,33	0,33	1,20	0,60	0,40	0,67	-0,17	-0,67	-0,67	-0,17	-0,33					0,33						
11		0,17	0,33	0,17	0,50	0,67	0,33	0,33	0,33	1,20	0,60	0,40	0,67	0,17	0,67	0,67	0,17	0,33	7,70										
12																													
13	Besonders gut																												
14	Praxisbezug																												
15	2 Sehr gute Folien																												
16	Selbsttests																												
17	gastvorträge																												
18	2 Atmosphäre																												
19	Uni-Palaver																												
20	3 Dozent																												

Ihre Bewertung 15/16

Konzeption	Viel neues	Lernziele	Materialien	Rhetorik	Beispiele	Klare Struktur	Verbindungen	Klar verständlich	Übung hilft	Verständliche Antworten	Kritische Auseinandersetzung	Gute Atmosphäre	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	Abweichung vom Optm	Abweichung pro Frage	Geleitet	Warum kommen?	Studiengang	Fachsemester
5	5	4	5	6	5	5	5	6	6	6	5	5	3	4	4	2	2	14	0,82	1	2,4	M	3
6	6	6	6	6	6	6	6	6	6	6	6	6	3	4	3	1	1	1	0,06	0	3,4	M	3
6	6	6	6	6	6	6	6	6	6	6	6	6	3	3	3	1	1	2	0,11	0	3,4	M	1
5	5	5	4	6	5	5	6	6	5	5	4	3	3	3	4	1	2	14	0,78	1	4	M	13
6	6	6	5	6	5	5	5	6	5	6	5	6	3	3	3	1	1	6	0,33	1	4	M	5
6	6	6	6	5	5	6	6	6	6	6	5	6	3	3	3	1	1	4	0,22	0	2,4	M	1
5	6	6	6	6	6	5	6	6	6	6	5	6	3	3	3	1	1	3	0,17	0	2,4	M	1
6	6	5	5	6	6	6	6	6	5	5	6	6	4	4	4	1	1	7	0,39	0	3,4	M	3
5	5	5	6	6	5	5	6	6	5	6	6	6	2	3	3	1	1	7	0,39	1	3,4	DI	15
6	5	6	6,0	6	6	6	6	6	6	6	6	6	3	3	3	1	1	1	0,06	1	2,3,4	M	3
6	6	4	4	6	2	4	6	6	5	5	6	6	3	4	5	2	2	15	0,83	1	3,4	M	4
6	5	5	5	6	6	6	6	6	6	6	6	5	3	2	3	1	1	5	0,28	0	4	BA	12
6	6	5	6	6	6	5	5	6	6	5	6	6	4	3	4	1	1	6	0,33	0	2,4	M	1
5	5	5	4	5	5	5	5	5	4	5	6	5	4	4	4	2	2	19	1,06	1	2,4	M	3
5,64	5,57	5,50	5,21	5,86	5,29	5,36	5,71	5,93	5,43	5,38	5,64	5,71	3,14	3,29	3,50	1,21	1,29						
6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	3,00	3,00	3,00	1,00	1,00						
0,36	0,43	0,50	0,79	0,14	0,71	0,64	0,29	0,07	0,57	0,62	0,36	0,29	-0,14	-0,29	-0,50	-0,21	-0,29						
0,36	0,43	0,50	0,79	0,14	0,71	0,64	0,29	0,07	0,57	0,62	0,36	0,29	0,14	0,29	0,50	0,21	0,29	7,19					
Beispiele	Aufbau	Hochschulpolitik	WDH am Anfang	Viele konkrete Algorithmen	Biologie am Anfang	Erklärungsstil																	
Häufige Ausfälle; GOYA funktioniert nicht	Biologie am Anfang zu gedrängt																						
Biologie																							

Was wir ändern wollen

- Compressive Genomics
- Ev. Burrows-Wheeler Transform, Read Mapping
- Folien zu Editabstand kürzen, aus 3 macht 2
- Folien Neighbor Joining ausführlicher
- Übung: Aufgabe 5 wird ersetzt

Fragen

- Diplominformatiker?
- Biophysiker?
- Bachelor?
- Semester?

- Spezielle Erwartungen?

Inhalte der Vorlesung

- Einführung
- Ein konkretes Beispiel
- Themen der Vorlesung

Beispiel 1: H5N1

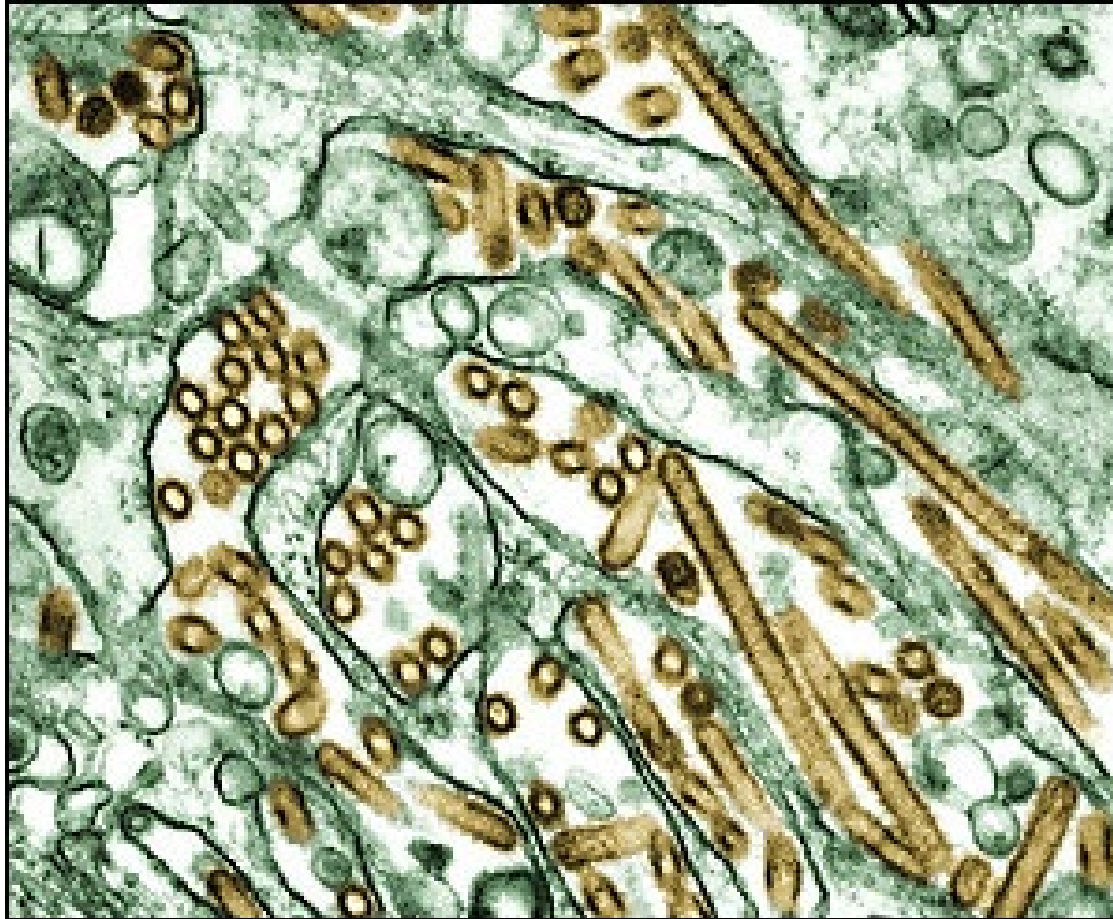
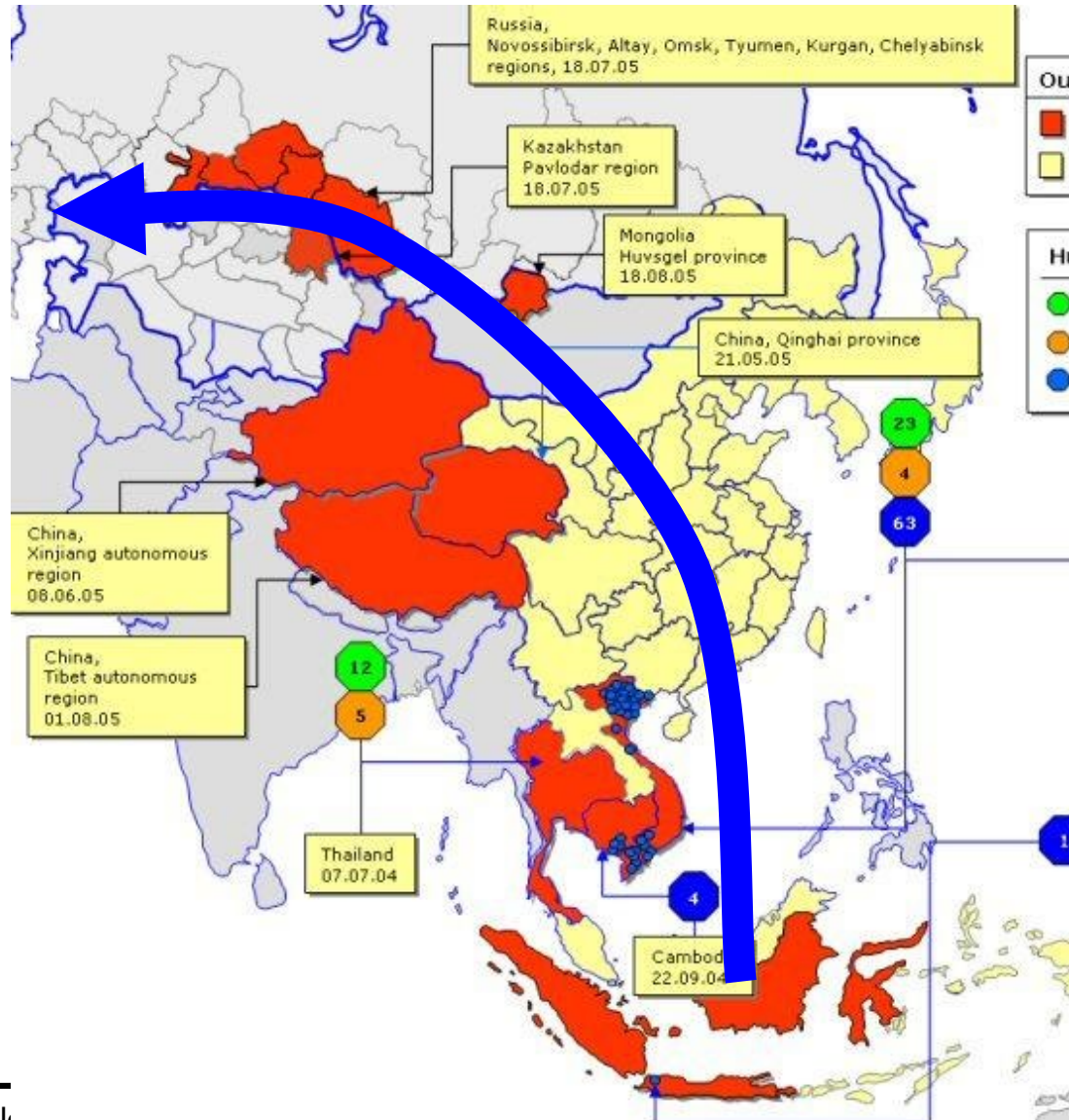
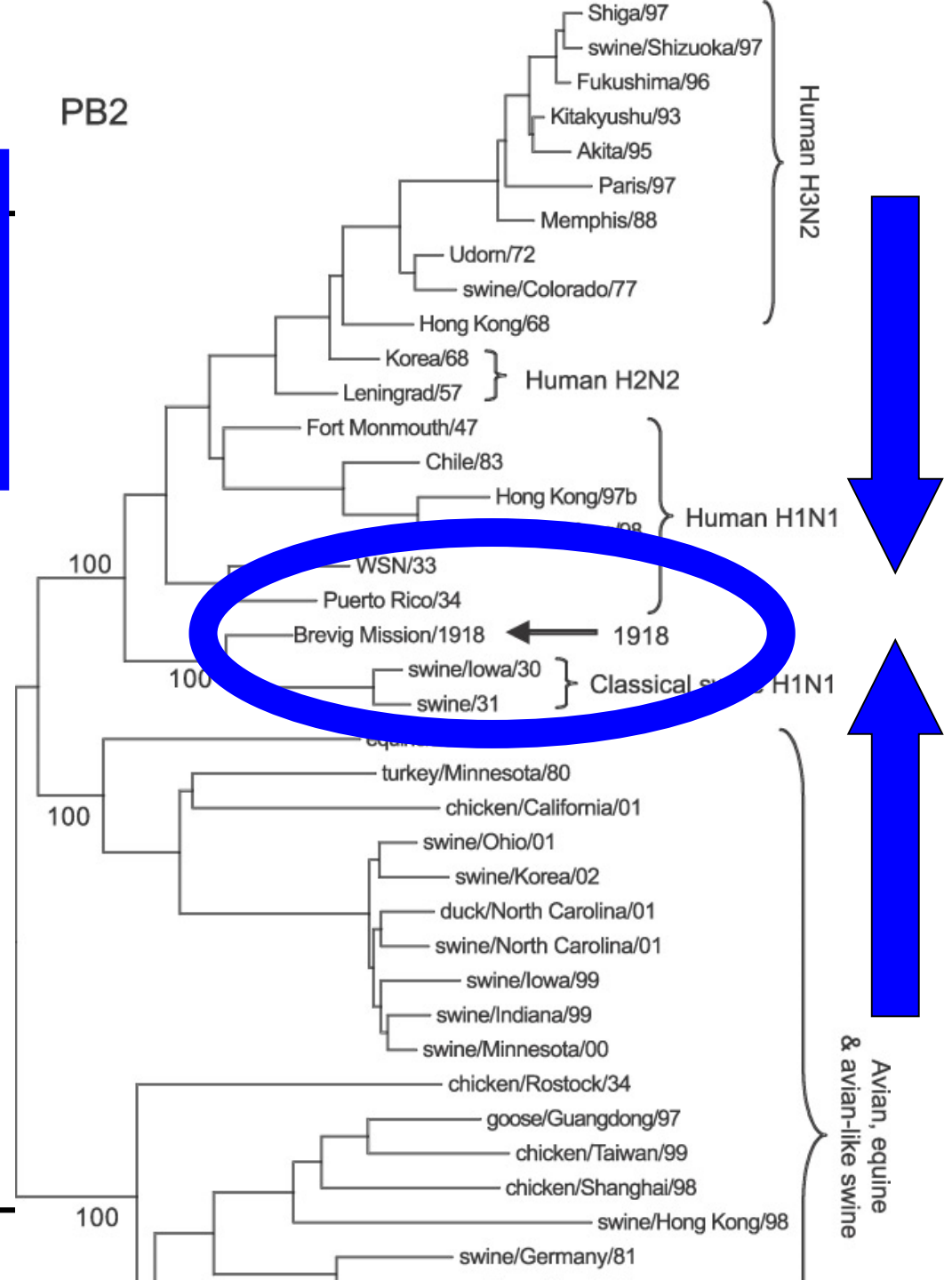


Foto: Centers for Disease Control

Migration

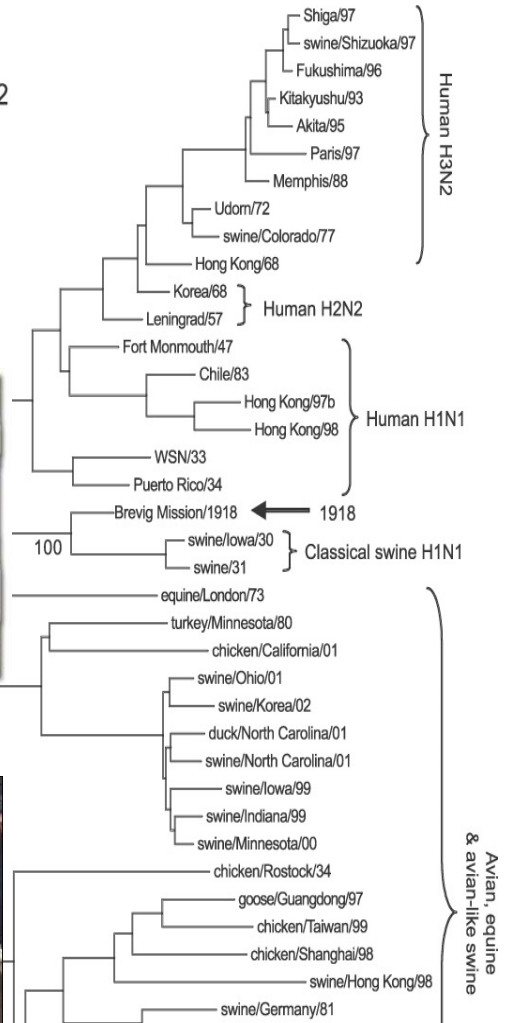
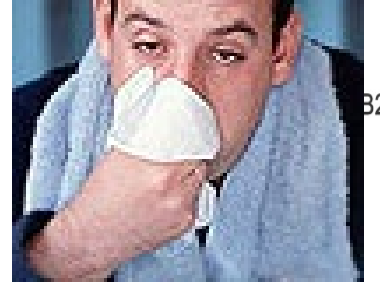
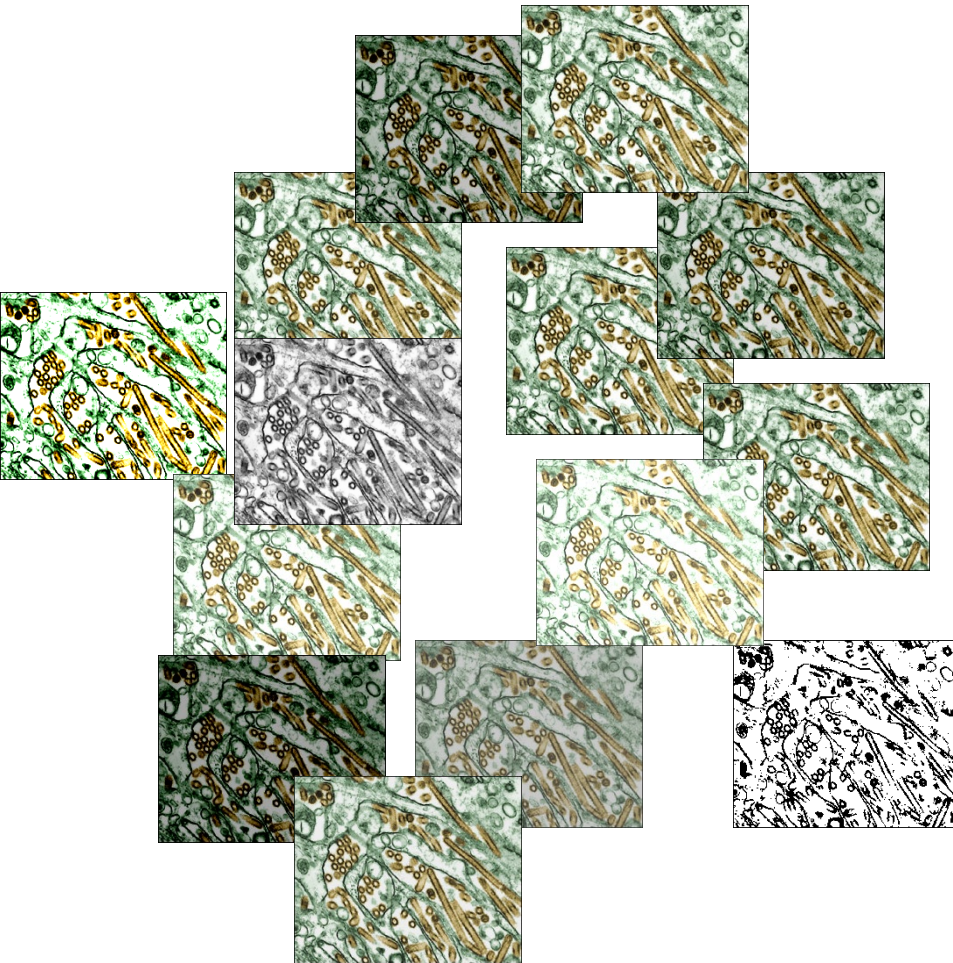


Vogelgrippe beim Menschen?

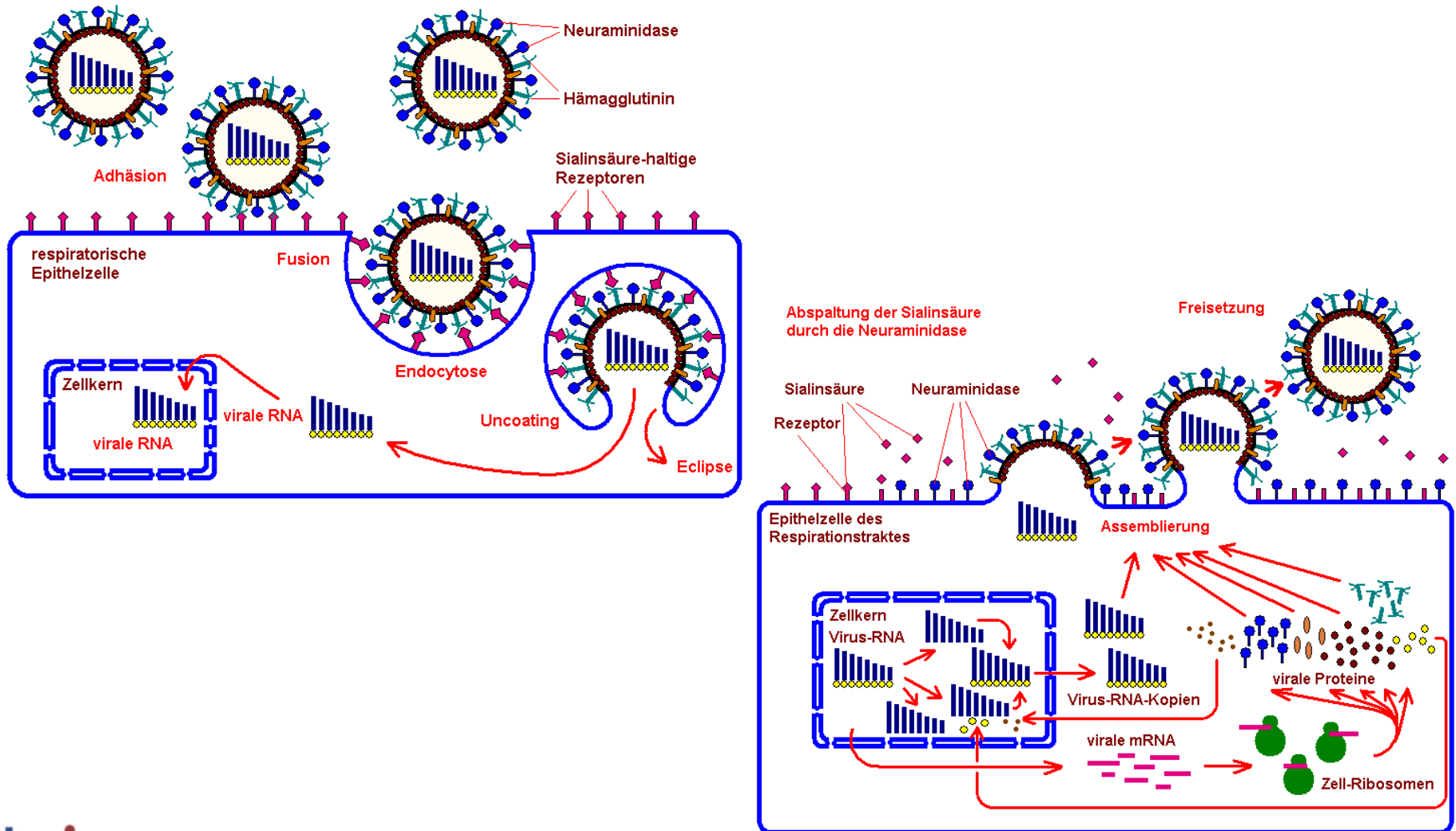


Nature. 2005 Oct 6;437(7060):889-93.

Wo kommt der Stammbaum her?

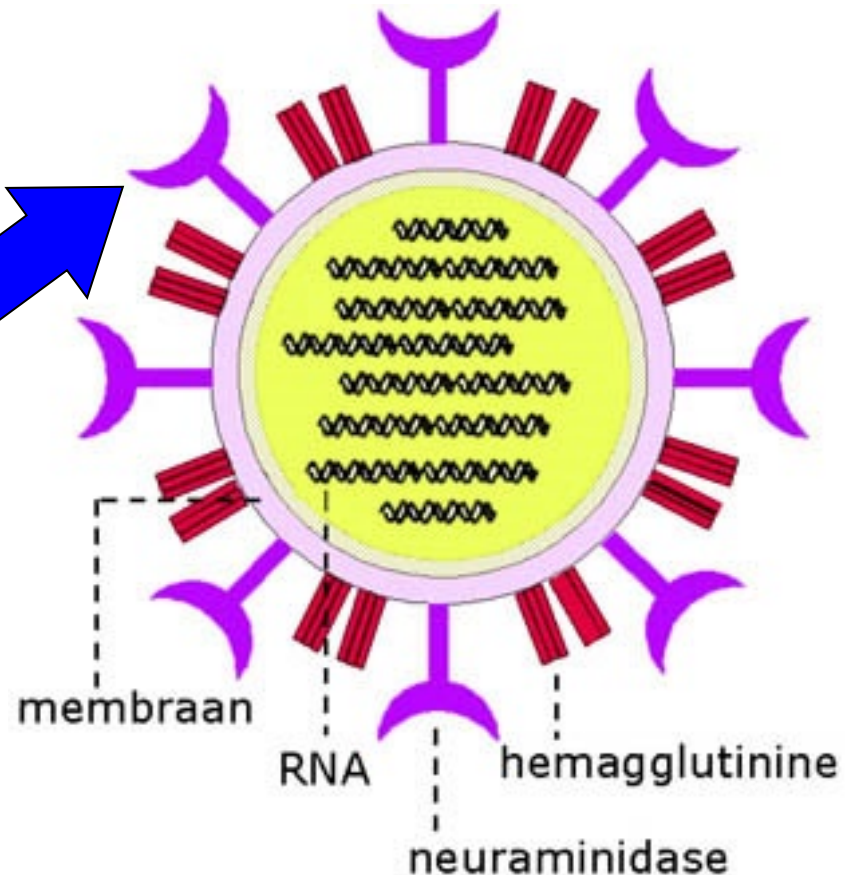


Viren



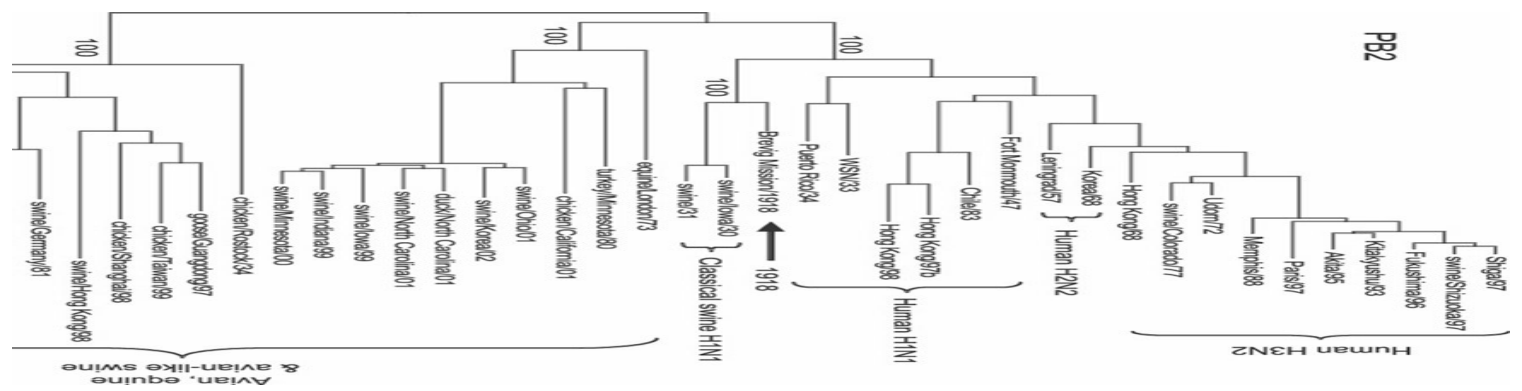
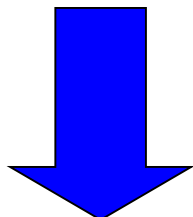
Grundlage für Stammbaumberechnung

The influenza A viral heterotrimeric polymerase complex (... , PB2) ... having a **role in host specificity.**



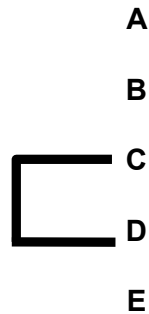
Informatik?

M E R I K E L R D L M S O S R T R E I L T K T T V D H M A I-3028-ATGGAAGAATAAAAAGAACTAAGGGATCTAATGTCCGAGTCTCGCAGTCCGCGAGATCTCACAAAAACCAGTGGACCATATGGCCATA-117 I K K
 Y T S G R Q E K N P A L R M K W M M A M K Y P I T A D-60 ATCAAGAAGTACACATCAGGAAGACAGGAGAAAGAAACCCCGCACTTAGGATGAAATGGATGATGGCAATGAAATATCCAAATTACAGCAGAC-207 K R I M E
 M I P E R N E Q G O T L W S K T N D A G S D R V M-90 AAGAGGATAATGGAAATGATTCCTGAGAGAAATGAGCAAGAAATGAGCAAGGACAACTCTATGGAGTAAAACGAAACGATGCCGGATCAGACCGAGTATG-297 V S P L A V T
 W W N R N G P T T S A V H Y P K I Y K T Y F E-120 GTATCACCTCTGGCTGTGACATGGTGAATAGGAAATGAGCAAGGACAACTCTATGGAGTAAAACGAAACGATGCCGGATCAGACCGAGTATG-297 V S P L A V T
 F G P V H F R N Q V K I R R V D I N P G-150 AAAGTCAAAGCTTAAAACATGGAACCTTGGCCCTGCTCATTTCCAGAAACCAAGTCAAATACCTCGAAGAGTGGACATAACATCGGAA-567 S O L T I T K E K K E E L Q
 V I M E V V F P N E V G A R I L T S E-180 CATGCAGATCTCAGTCCAAAGAGCAGAGGATGTAATCATCTGCTGAGAGTGTGTTTCCCAACCAAGTGGAGCCAGGATACATACATCGGAA-567 S O L T I T K E K K E E L Q
 D C K I S P L M V A Y M L E R E-210 TCGCAACTGACAATAACCAAGAAAAGAAAGAACTCAAAGATGTAATAATTCCTCTTAATGGTGGCAGATGTTGGAGAGAGAA-657 L V R K T R F L P V A G G T S S V
 Y I E V L H L T O G T C W-240 CTGGTCCGAAAAACAAGATTCTCCAGTGGCTGGTGGGACAAGCAGTGTGTATATTGAAGTGTTCATTTGACTCAAGGAACATGCTGG-747 E Q M Y T P G G E V R N D D V D O S
 L I I A A R N I V R R A-270 GAACAGATGTACTCCAGGAGGGGAAGTGAAGGAATGATGATGTGATCAAAGCTAATTTATGCTGCTAGGAACATAGTGAAGAGCGG-837 T V S A D P L A S L L E M C H S T O I G
 G I R M V D I L R Q-300 ACAGTGTACAGCAGTCCACTAGCATCTCTGTGGAAATGTGCCACAGCACAGATTTGGTGAATAAGGATGGTAGACATCCTTAGGGCAG-927 N P T E E Q A V D I C K A A M G L R I S S S
 R V H E G Y-360 AAGAGGACAAGTGGCTCATCAGTCAAGAGGGAGGAAGAGTGCCTACAGGCAATCTTCAACCAATTGAAGATAAGGATGGTAGACATCCTTAGGGCAG-927 N P T E E Q A V D I C K A A M G L R I S S S
 S G R D-390 GAAGAGTTCACAAATGTTGGGAGAGGACACAGCTATACTCAGAAAAGGACAGCAGGAGATTGATTCAGCTGATAGTGAAGTGGAGAGAC-1197 E Q S I A E A I I V A M V F S Q E D C M I K A V R G D
 L N F-420 GAACAGTTCGATGCCGAGGCAATAATTGGCCATGGTATTTTCAAGAGGATTGATGATAAAGGCAGTATAGGGGTGATCTGAATTTTC-1287 V N R A N Q R L N P M H Q L L R H F Q K D A K V L F O N
 W G-450 GTTAATAGGGCGAATCAGCGATTGAATCCTATGCAATCACTTTGAGGCAATTTCAAAGGATGCGAAAGTGTCTTTTCAAATGGGG-1377 I E P I D N V M G M I G I L P D M T P S T E M S M R G V R
 V-480 ATTGAACCCATCGACAATGTGATGGGAATGATTGGGATCTGCCGACATGACTCCAAGTACTGAGATGTCAATGAGAGGAGTGAAGATC-1467 S K M G V D E Y S S T E R V V S I D R F L R V R D Q R G N-
 510 AGCAAAATGGGAGTAGATGAGTACTCCAGCAGAGAGGGTGGTGGTGAAGTGTGACTTCAAGGCAATCTTCAACCAATTGAAGATAAGGATGGTAGACATCCTTAGGGCAG-927 N P T E E Q A V D I C K A A M G L R I S S S
 540 GTACTACTGTCTCCTGAGGAGGTCAGCGAAAACACAGGGAACAGAGAAATGACGATAACTTATTCATCTCAATGATGGGAGGTTAAT-1647 G P E S V L V N T Y Q W I I R N W E T V K I O W S Q N P T M-
 570 GGCCCTGAATCAGTGTGGTCAACACTATCAGTGGATCATCAGAACTGGGAAACTGTTAAAATCTCAGTGGTCAAGAACTCCTCAATG-1737 L Y N K M E F E P F O S L V P K A A R G Q Y S G F V R T L F-600
 CTATACAATAAAATGGAAATTTGAGCCATTTAGTCTTTAGTTCCTAAGGCCGCTAGAGGTCATACAGTGGGTTTGTGAGAACTCTGTTTC-1827 Q O M R D V L G T F D T V Q I I K L L P F A A A P K Q S R-630
 CAGCAATGAGGAGTGTCTGGGACATTTGACCCGTTGAGATAATAAAACTTCTCCCTTTGAGCCCTCCACCAAGCAAGTGA-1917 M O F S S L T V N V R G S G M R I L V R G N S P V F N Y N K-660
 ATGCAGTCTCTCTGACTGTGAATGTGAGAGGATCAGGAATGAGAACTACTTGAAGGGCAATTTCTCCGATTTCAACTCAACAAAG-2007 A T K R L T V L G K D A G A L T E D P D E G T A G V E S A V-690
 GCACTAAGAGACTCAGCTCTCGGAAAGGATGCAAGTGTCTTAACTGAAGACCCAGATGAAGGACAGCTGGAGTGGAGTCTGTGTT-2097 L R G F L I L G K E D R R Y G P A L S I N E L S N L A K G E-720
 CTGAGGATTCCTACTCTGGCAAAGAGCAGGAGATGAGCCAGCATTAAGCATCAATGAAGTGAAGCAATCTCGAAAGGGGAG-2187 K A N V L I G O G D V V L V M K R K R D S I L L T D S Q T A-750
 AAGGCTAATGTGCTAATTTGGCAAGGAGACGTGGTGTGTAATGAAACGGAAACGGGACTCTAGCATCTACTGACAGCCAGACAGCG-2277 T K R I R M A I N Stop ACCAAAAAATTCGGATGCCATCAATTAG-2307

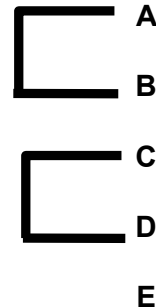


Konstruktion des Guide Trees

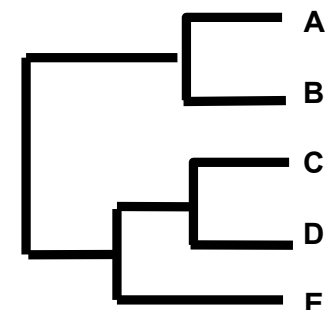
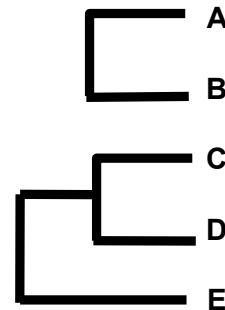
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



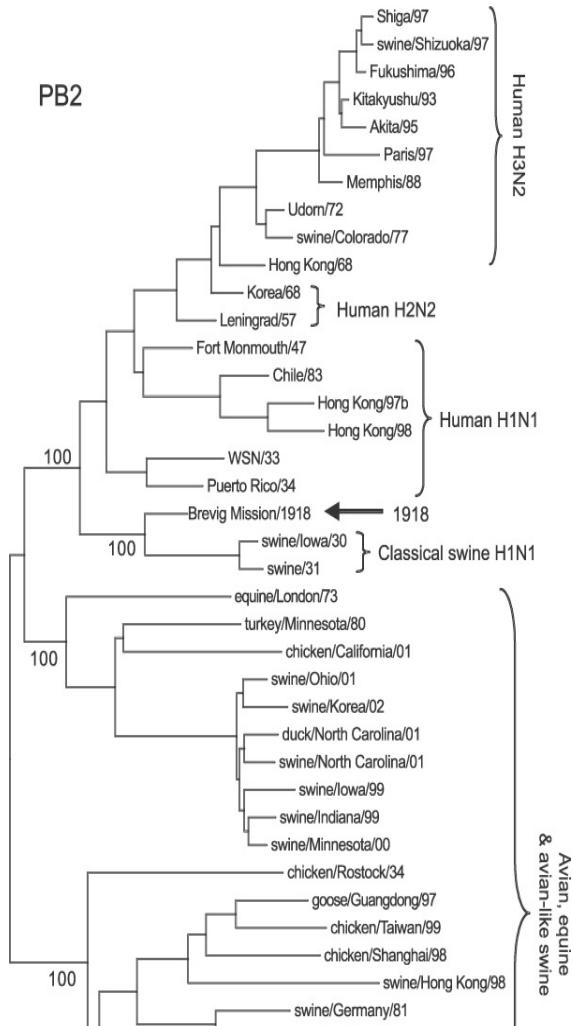
	A	B	E	CD
A		17	77	59
B			53	49
E				31



	E	CD	AB
E		31	65
CD			54



Was bedeutet der Stammbaum?

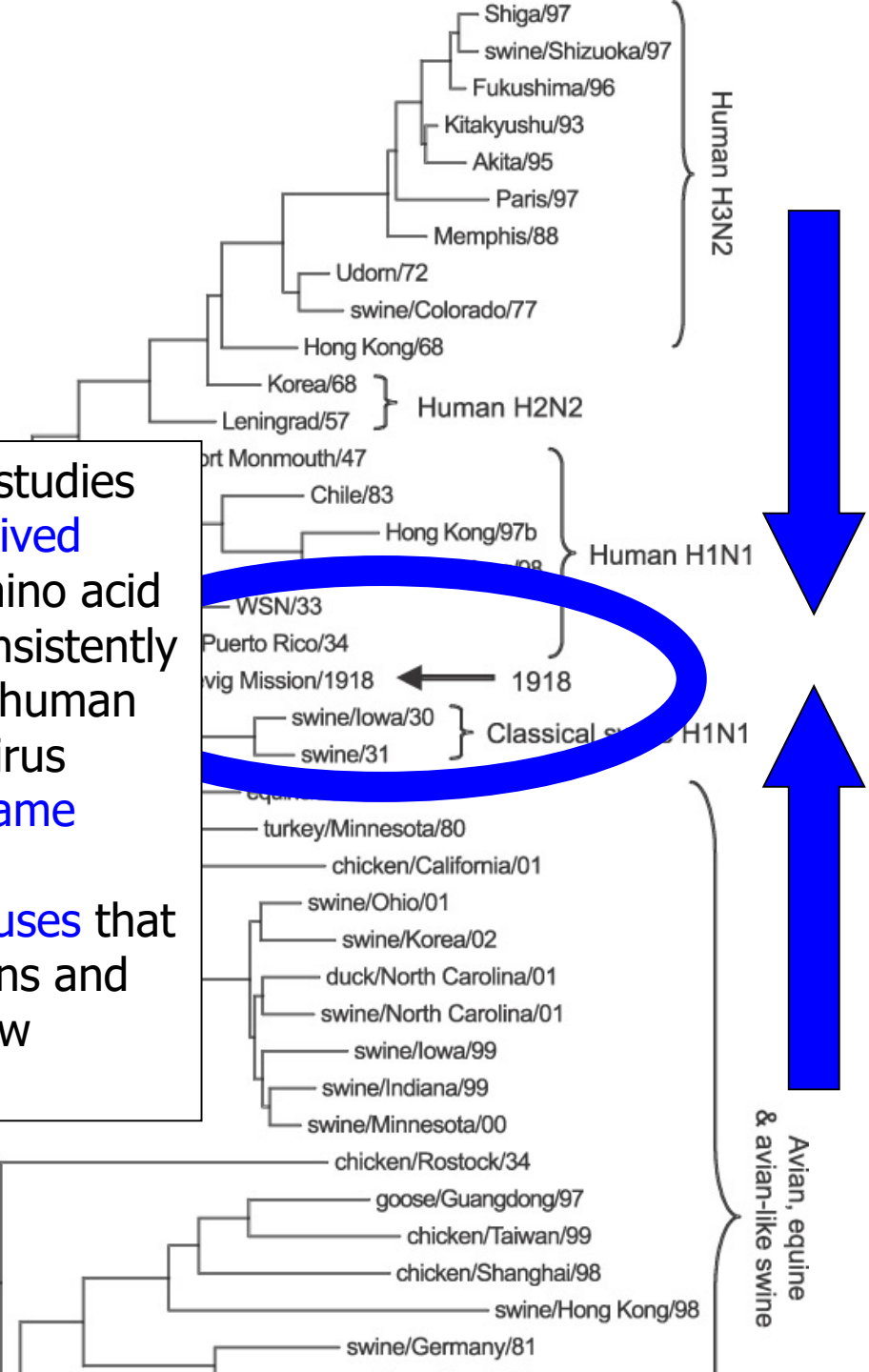


- Grundidee ist Evolution
 - Ur-Virus und Speziation
 - Richtungslose Mutationen
 - Selektion durch Umwelt (Host!)
- Benachbarte Stämme haben ähnliche Sequenzen
- Ähnliche Sequenzen bedingen ähnliche Funktion der Proteine und wahrscheinlich auch ähnliche Wirkung
- Um benachbarte Knoten zu „erreichen“, sind **nur noch wenige Mutationen** notwendig

Vogelgrippe beim Menschen?

These data support prior phylogenetic studies suggesting that the 1918 virus was derived from an avian source. A total of ten amino acid changes in the polymerase proteins consistently differentiate the 1918 and subsequent human influenza virus sequences from avian virus sequences. Notably, a number of the same changes have been found in recently circulating, highly pathogenic H5N1 viruses that have caused illness and death in humans and are feared to be the precursors of a new influenza pandemic.

PB2

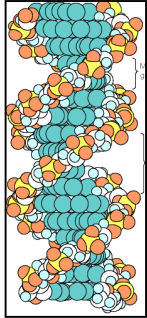


Nature. 2005 Oct 6;437(7060):889-93.

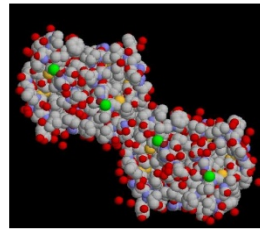
Inhalte der Vorlesung

- Einführung
- Ein konkretes Beispiel
- Themen der Vorlesung

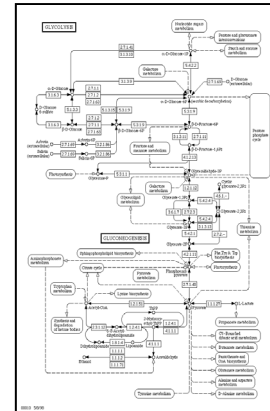
Diese Vorlesung



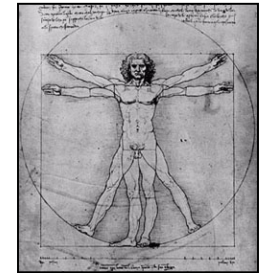
Sequenzierung
Erkennung von Genen
Verwandtschaft zw. Spezies
Regulation & Expression
RNA Gene – „Dark matter“



Dreidimensionale Faltung
Strukturvergleich und -ähnlichkeit
Interaktion
Sekundärstruktur
Proteinidentifikation



Netzwerkanalyse
Geschwindigkeit von Reaktionsketten
Stoffumsatz
Kompartamente
Muster und Redundanz



Korrelation Phänotyp/Genotyp
Arzneimittel-empfindlichkeit
Studienstatistik und -verwaltung

Indexstrukturen

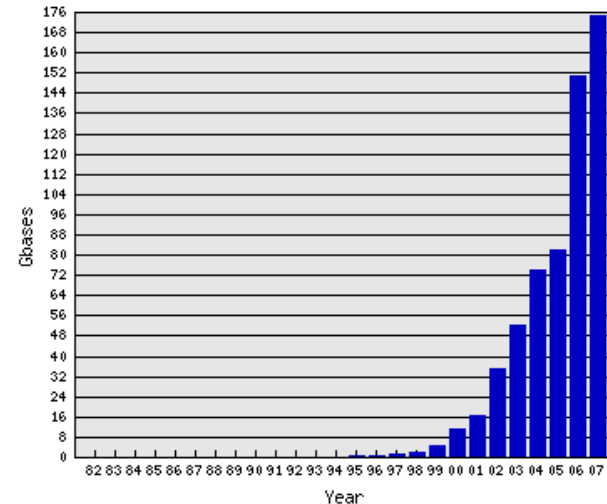
- Gegeben ein festes T und dauernd wechselnde P . Finde eine Datenstruktur für T , die die Suche nach allen P in möglichst kurzer Zeit gestattet
- Grundlegende Datenstruktur: **Suffixbäume**
- Suche und Konstruktion
 - Ukkonen's linearer Algorithmus
- Verschiedene Anwendungen
 - Längster Substring
 - Längster Repeat
- **Suffixarrays**

Approximatives Stringmatching

- Gegeben ein Template T und ein Pattern P. Finde alle Vorkommen von Substrings „ähnlich“ zu P in T in möglichst kurzer Zeit
- Was heißt überhaupt ähnlich?
 - Ähnlichkeitsmaße, Edit-Abstand, Alignierung
- Naiver Algorithmus benötigt exponentielle Laufzeit
 - Verbesserung durch dynamische Programmierung
 - Erreicht quadratische Laufzeit
- Viele Varianten: Globale, lokal, end-free, ...

Heuristiken

- Quadratische Laufzeit ist zu teuer
 - Genomanalyse benötigt Suche auf allen bekannten Sequenzen
 - Celera Sequenzierung: All-against-all Vergleich von 28.000.000 Teilsequenzen



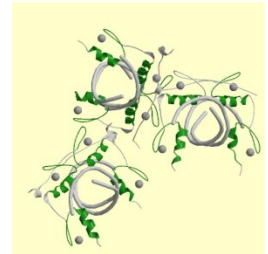
- Also: Heuristiken, z.B. BLAST
 - Suche nach „Seeds“ mit exakten Matches
 - Verlängerung und Zusammenfügung der Seeds zu Matches
 - Findet nicht alle Hits, aber die meisten „interessanten“
 - „BLAST“ ist fast **Synonym für Bioinformatik** geworden

Multiples Alignment

- Gegeben eine Menge von Strings. Ein Multiple Sequence Alignment (MSA) ist eine Anordnung der Strings mit Spaces untereinander

```
YVCK...LCN...FAFKTKGNLTKHMKS...AH
YRCP...ENCD...RTYTTKFNKSHILT...FH
FRCGY...KCGG...RLYTTAHLKVHERA...H
YRCE...KCG...KMYKTERCLKVHNLV...H
FSCS...QCD...ESFVORSELELHRQL...H
FPCE...QCD...EKFKTEKQLERHVKT...H
FQCN...QCG...ASFTQGNLLRHIKL...H
FKCH...LCY...RCFGQQTNLDRHLK...H
FRCK...RCR...TRFRQOSELKHKMT...H
FECN...VCG...SAFRLQLYLSEHQKT...H
MSCKV...CD...RVFYRLDNLRSHLKQ...H
FSCQ...HCH...RAFADRSNLR AHLQT...H
FRCG...YCG...RAFTVKDYLNKHLTT...H
HVCWV...PGCH...RAFSRSDNLN AHYTK...TH
LTC AH...CD...WSFDNVMKLVRRHG...H
```

Quelle: Pfam, Zinc finger domain



- Hauptziel von MSAs: Finde das „Gemeinsame“ der Sequenzen
 - Funktionen werden oft von sehr kurzen Sequenzstücken bestimmt
 - Welcher Teil eines Proteins bestimmt die Funktion?
 - Wie kann man Proteine in Familien anordnen?
- „Gute MSAs“ sind nicht klar definiert
 - Konkretes Maß zur Güte hängt von der Anwendung ab

Genomanalyse - Genvorhersage

- Welche Elemente eines Genoms sind interessant?
- Gene und deren Struktur
 - Promoterregionen, Start Site, Exons, Introns, ...
 - Verschiedene Regionen haben verschiedene Eigenschaften
 - Modellierung als Features
- Aufgabe: Finde die wahrscheinlichste Modulanordnung, gegeben ein Modell

(A) EUCARYOTES

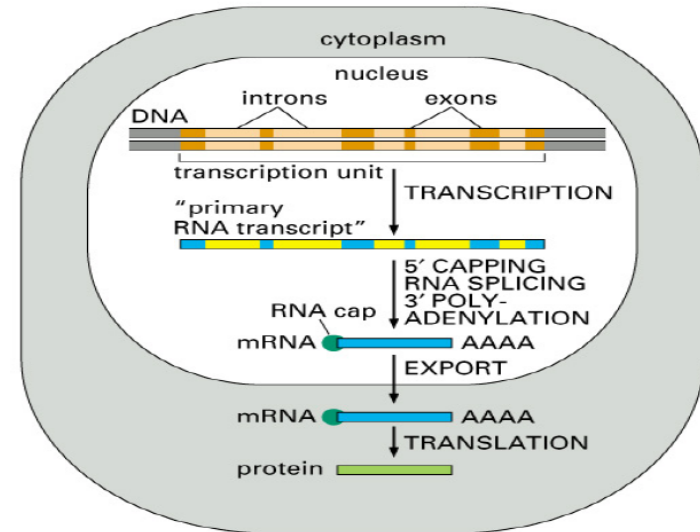
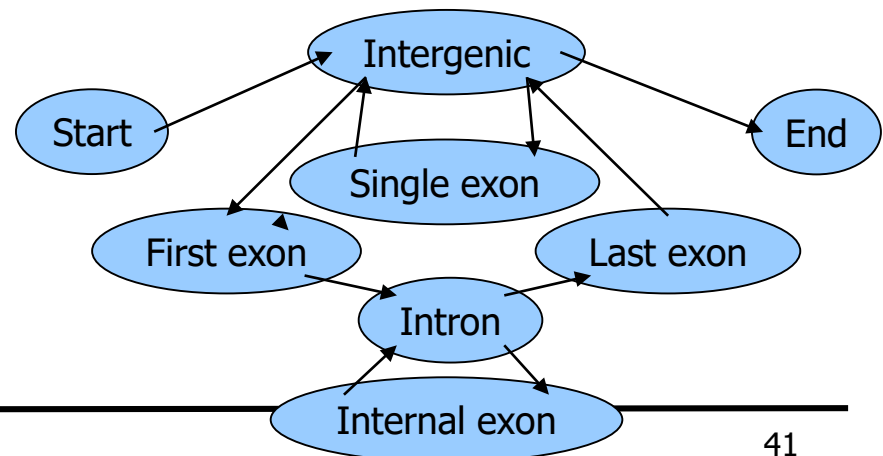


Figure 6-21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.



Selbsttest

- Wie viele Gene gibt es ungefähr in einem menschlichen Genom? Wie viele Basen? Wie viele Proteine?
- Was ist ein phylogenetischer Baum?
- Wie heißen die vier Basen eines Genoms?
- Wie viel Speicherplatz braucht man für 1000 Genome? Wie kann man sparen?