Algorithmische Bioinformatik

Wintersemester 2015 / 2016

Master: 10 SP Modul

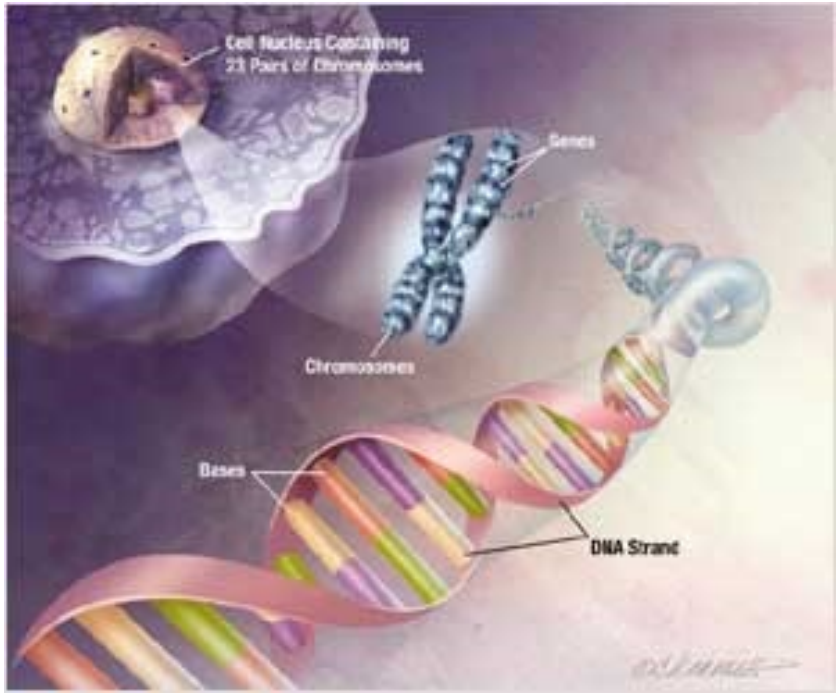
Ulf Leser

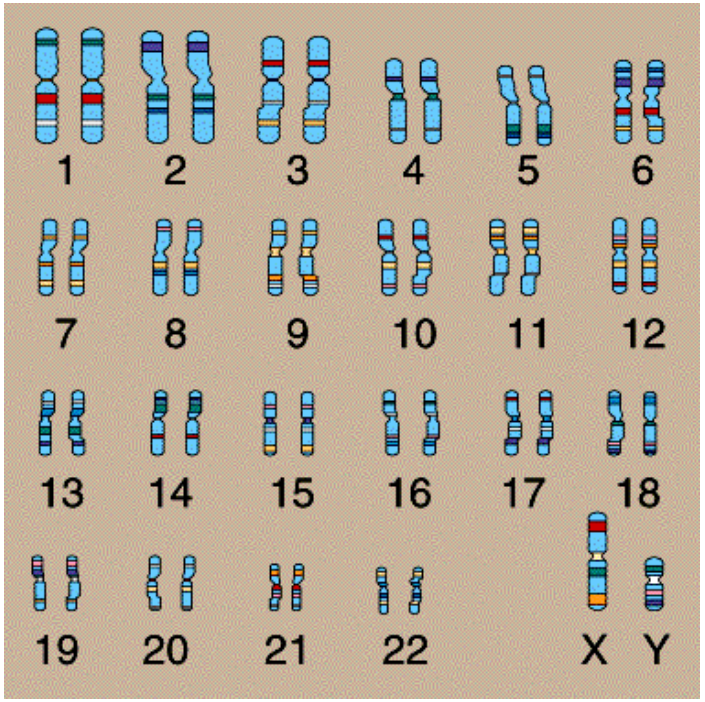
Wissensmanagement in der Bioinformatik

Ziele für heute

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

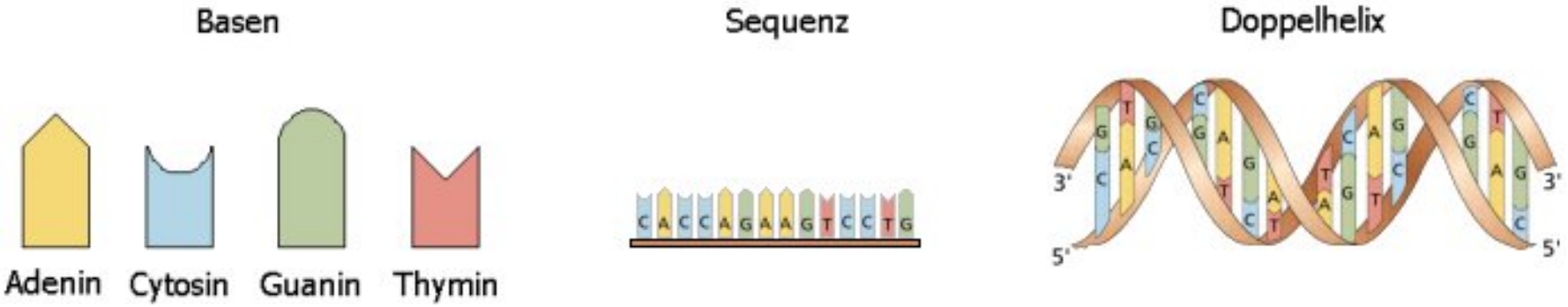
Genome ≈ String





- Human genome: app. 3.000.000.000 letters ∈ {A,C,G,T}

DesoxyriboNucleicAcid

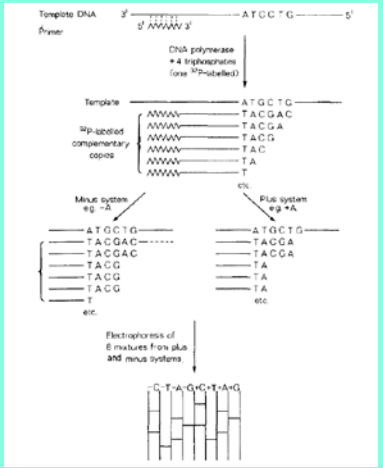


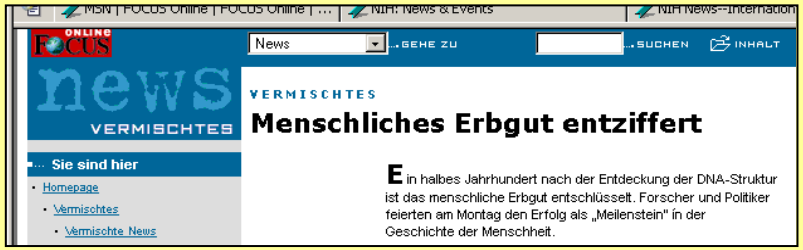
- DNA: Desoxyribonukleinsäure
- Träger der vererbten Information – Genom
- Alles Leben verwendet DNA (RNA) aus den selben 4 (5) Molekülen

Fast Development

1953
Double helix structure of DNA, Watson/Crick

1975
High-throughput sequencing
Sanger/ Coulson

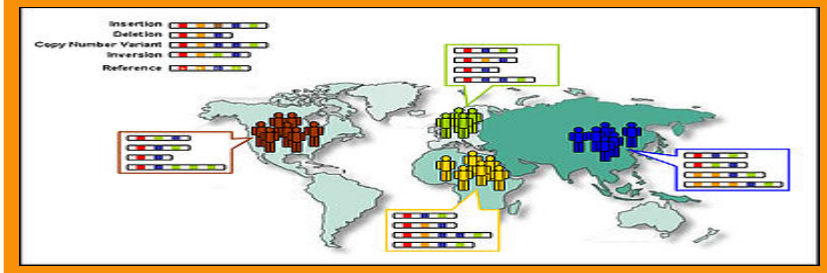




2003
First human genome sequenced
Took ~14 years, ~3 billion USD

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

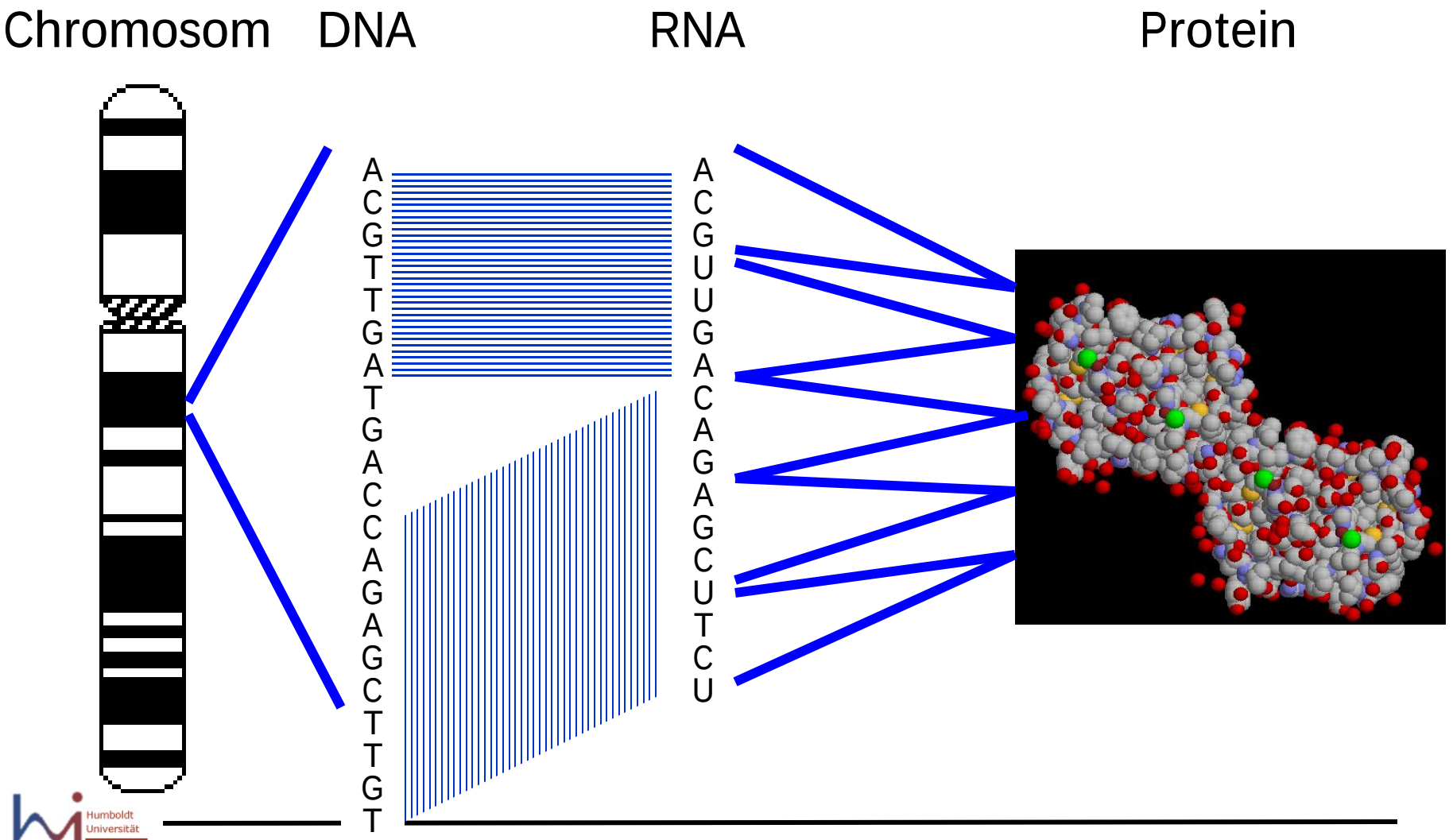
Large Scale Sequencing Projects

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 … 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?

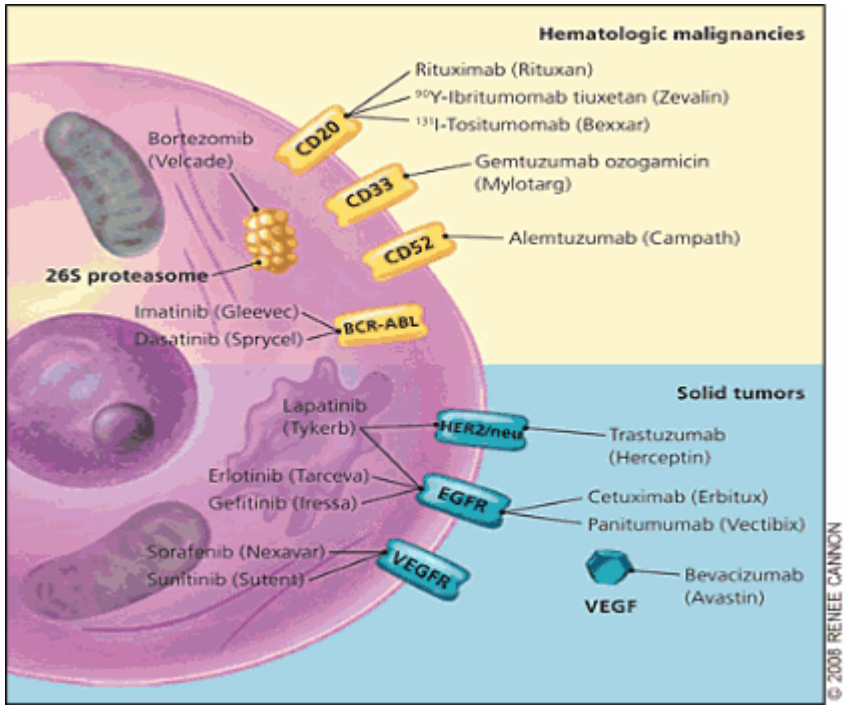


Genomics in a Nutshell

- ~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

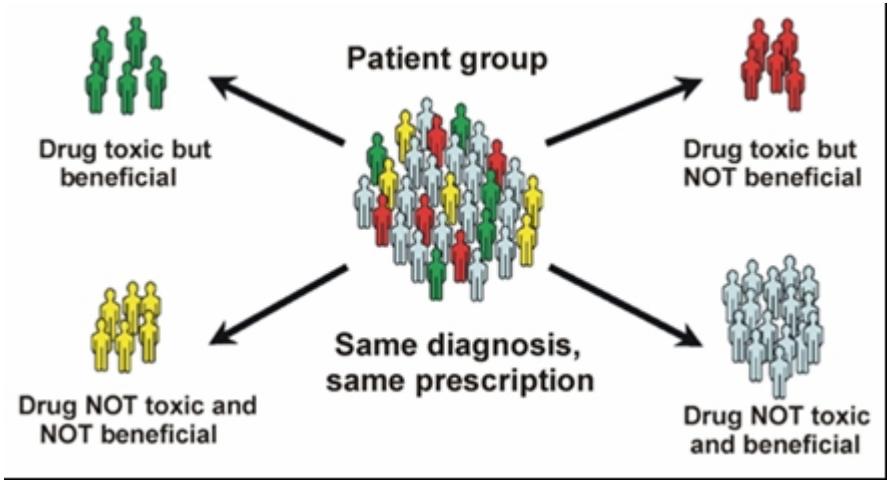
- Cancer, immunology, genetic diseases, infections
- Cancer
  - Cells proliferating uncontrolled, leaving their tissue
  - What goes wrong? Cell division, DNA repair, surface adhesion, cellular signaling
  - ~200 core cancer genes



- 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

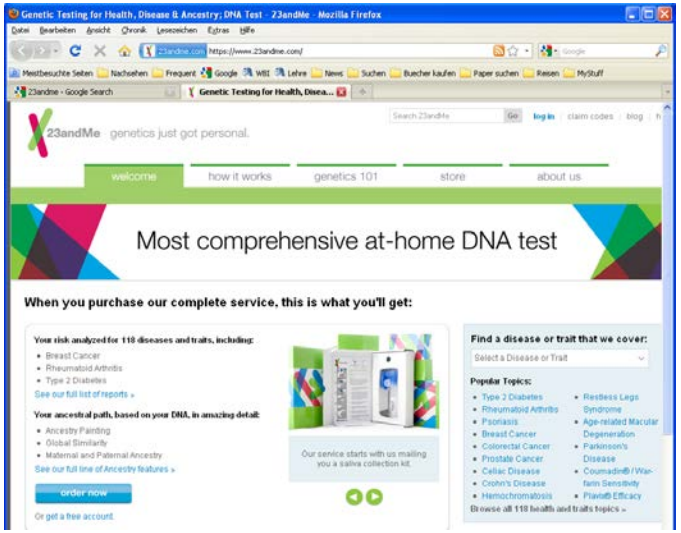
- 2 Firmen
- 32 Tests
- Preise: 100–1400€

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 [1]	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/Schl...	2	k. A.	512,81 €
Azoo...	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	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
  - Dienstag, 13-15 Uhr
  - Donnerstag, 13-15 Uhr
- Übung
  - Donnerstag, 15-17 Uhr
- Erste Übung für alle: Donnerstag, 22.10.2015, 15 Uhr
- Voraussetzung für Prüfung
  - Bestehen aller Übungsaufgaben
  - Verständnis der Algorithmen

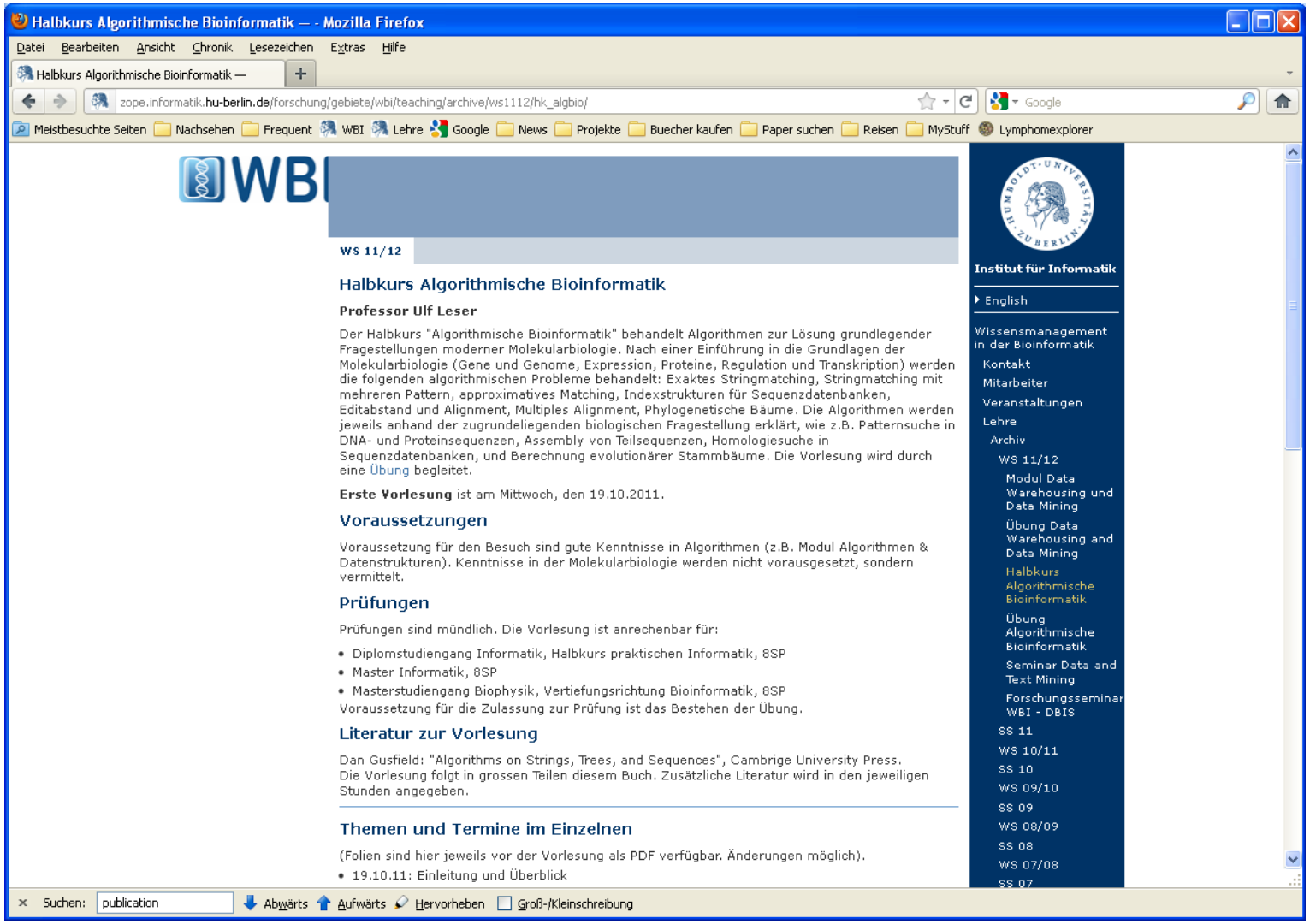
Gäste

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

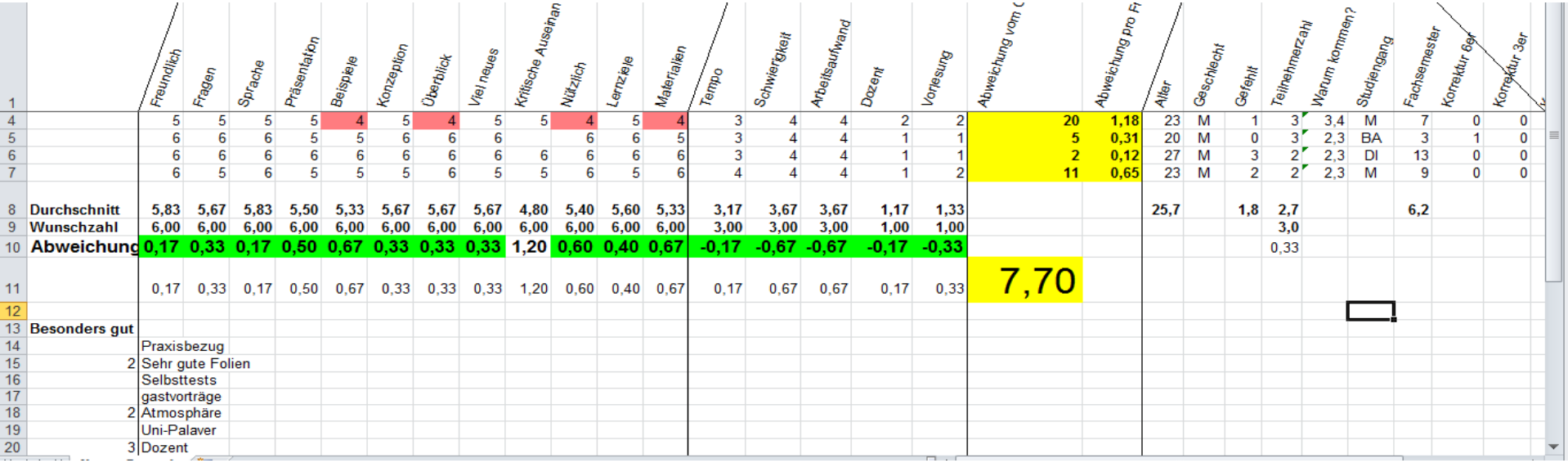
Literatur

- 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


Halbkurs Algorithmische Bioinformatik — - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Halbkurs Algorithmische Bioinformatik —
zope.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1112/hk_algbio/
Google
Meistbesuchte Seiten Nachsehen Frequent WBI Lehre Google News Projekte Buecher kaufen Paper suchen Reisen MyStuff Lymphomexplorer
WBI
WS 11/12
Halbkurs Algorithmische Bioinformatik
Professor Ulf Leser
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, 8SP
Voraussetzung für die Zulassung zur Prüfung ist das Bestehen der Übung.
Literatur zur Vorlesung
Dan Gusfield: "Algorithms on Strings, Trees, and Sequences", Cambrige 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
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 and 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
Suchen: publication
Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung

Ihre Bewertung



		Freundlich	Fragen	Sprache	Präsentation	Beispiele	Konzeption	Überblick	Viel neues	Kritische Auseinan	Nützlich	Lernziele	Materialien	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	Abweichung vom D	Abweichung pro Fr	Alter	Geschlecht	Gefehlt	Teilnehmerzahl	Warum kommen?	Studiengang	Fachsemester	Korrektur 6er	Korrektur 3er
4		5	5	5	5	4	5	4	5	5	4	5	4	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	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																											

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

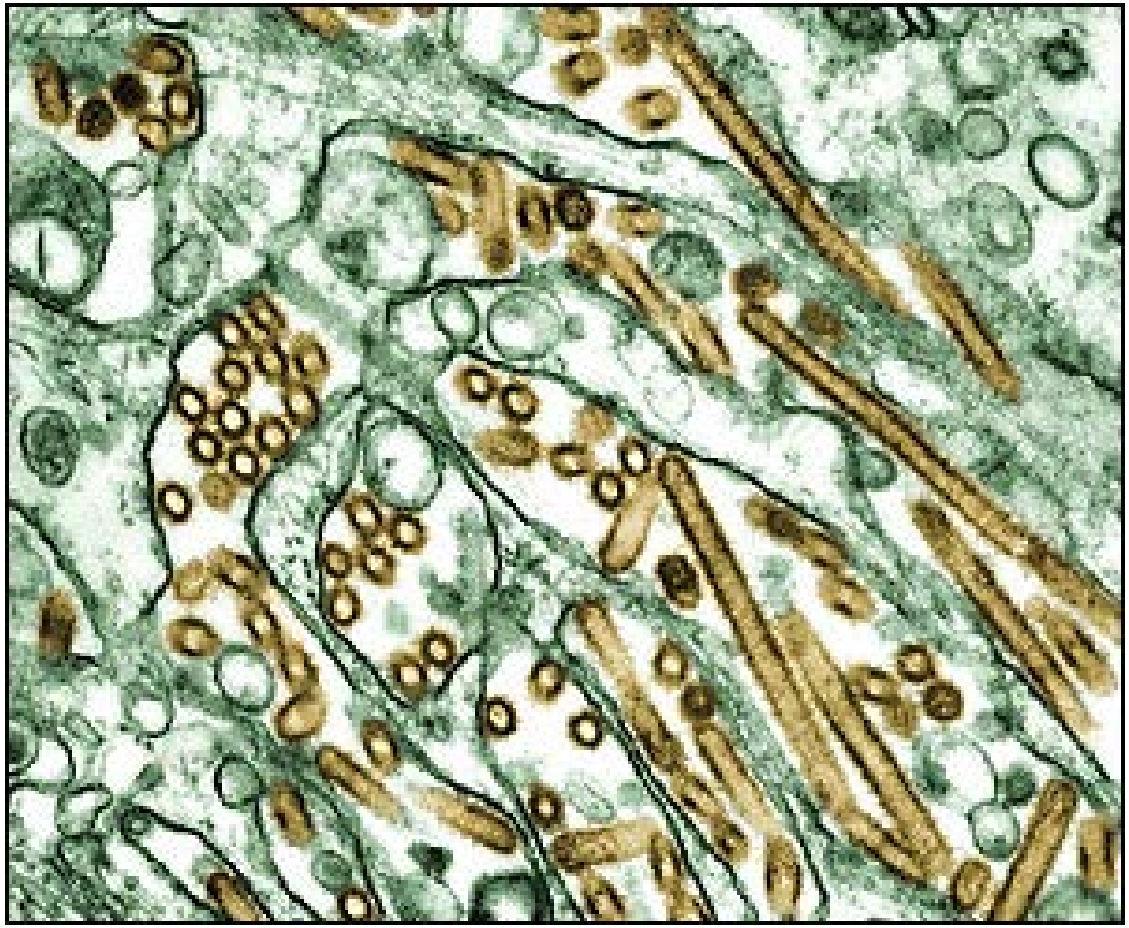
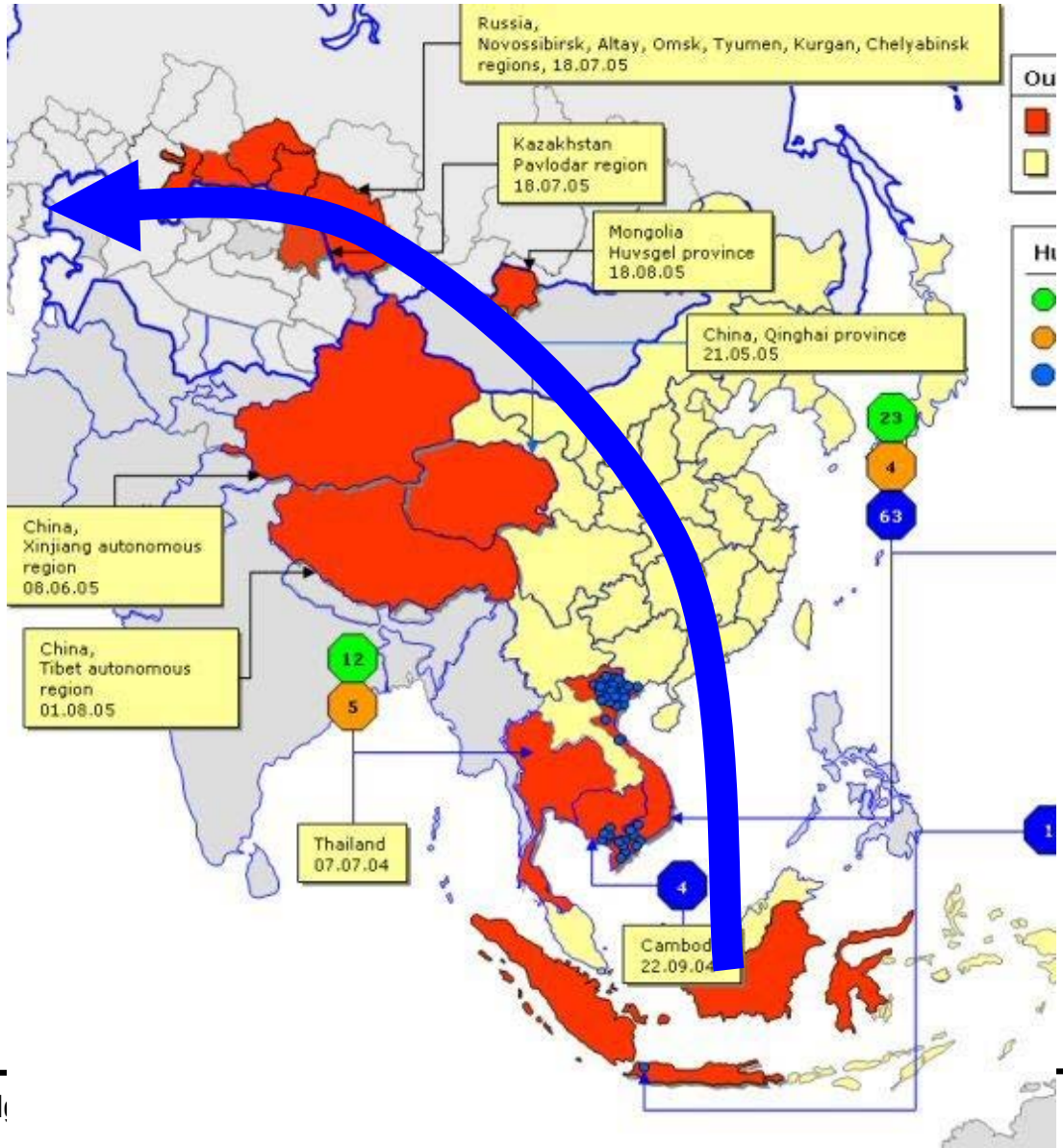
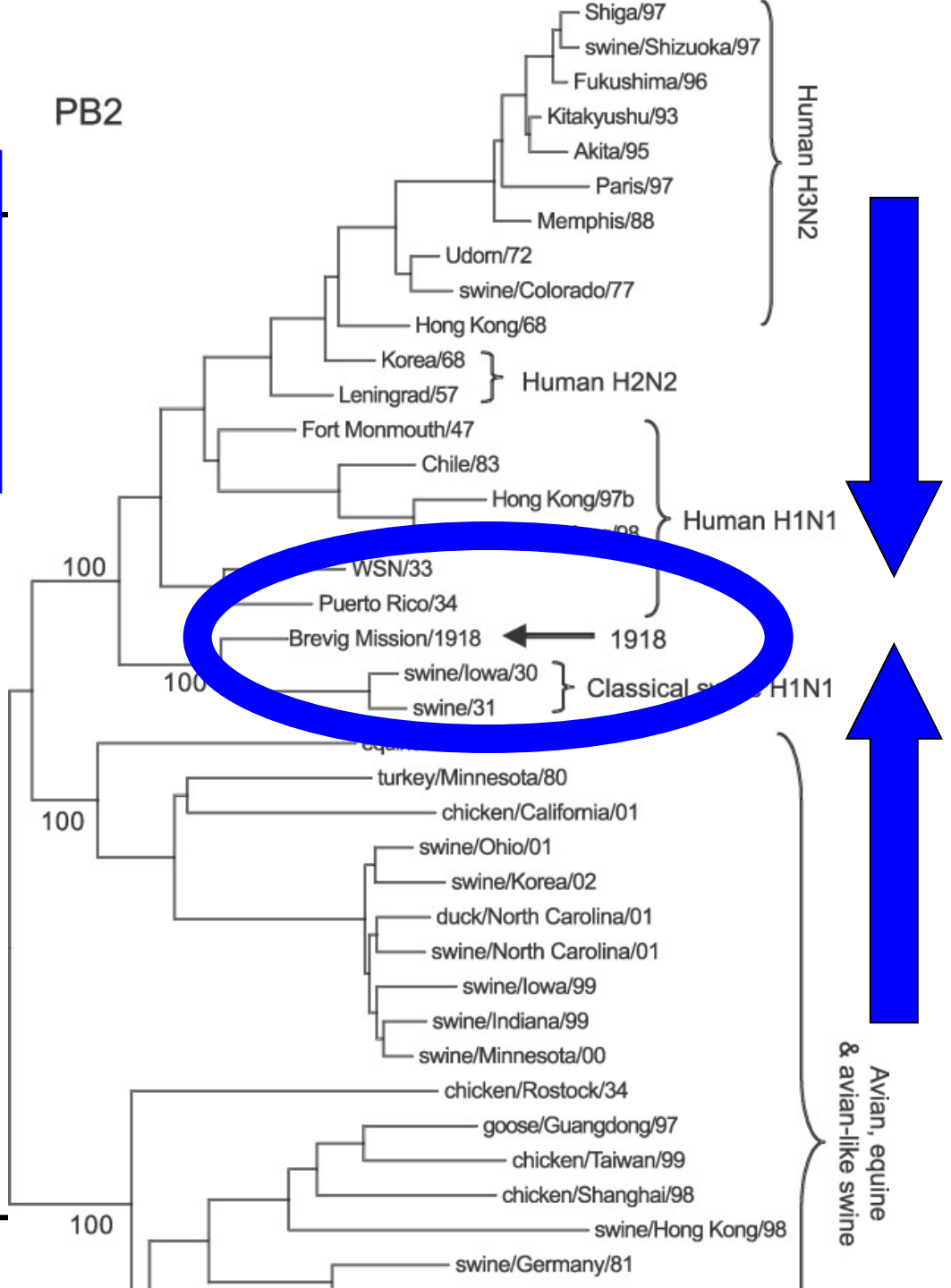


Foto: Centers for Disease Control

Migration

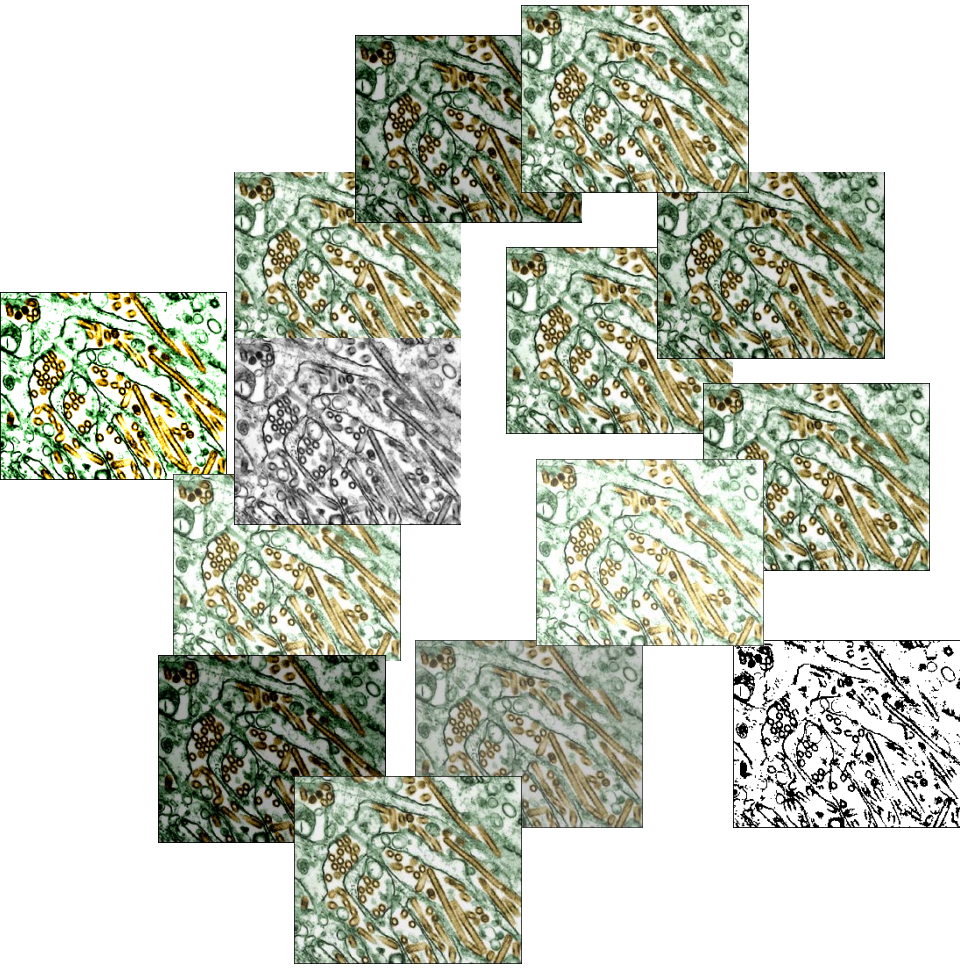

Russia,
Novossibirsk, Altay, Omsk, Tyumen, Kurgan, Chelyabinsk regions, 18.07.05
Kazakhstan
Pavlodar region
18.07.05
Mongolia
Huvsgel province
18.08.05
China, Qinghai province
21.05.05
China,
Xinjiang autonomous region
08.06.05
China,
Tibet autonomous region
01.08.05
Thailand
07.07.04
23
4
63
12
5
4

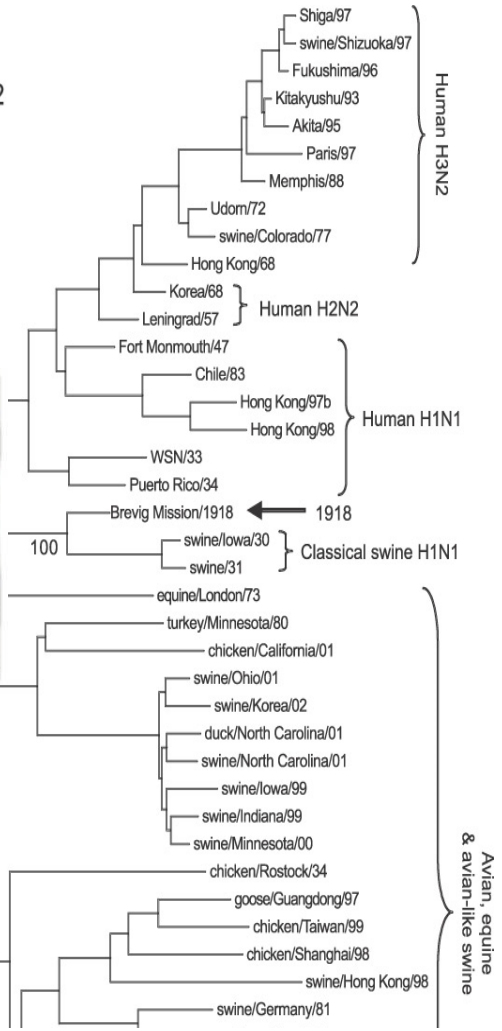
Vogelgrippe beim Menschen?



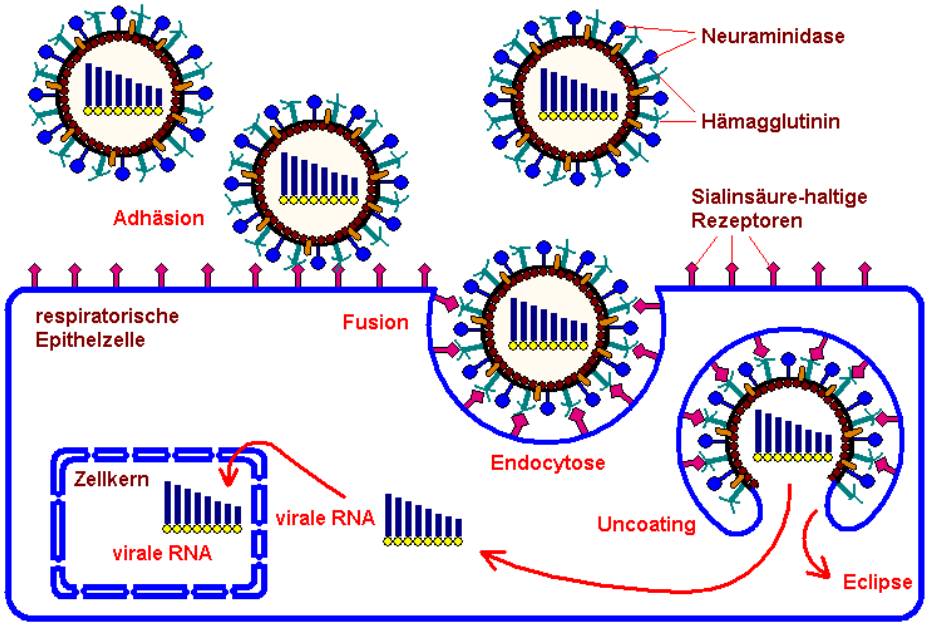
Nature. 2005 Oct 6;437(7060):889-93.

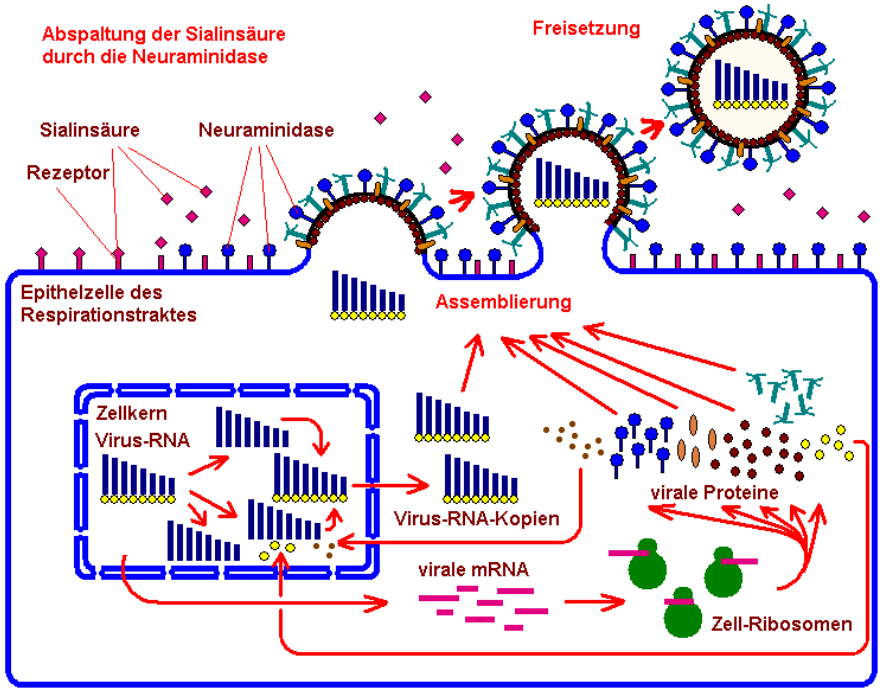
Wo kommt der Stammbaum her?





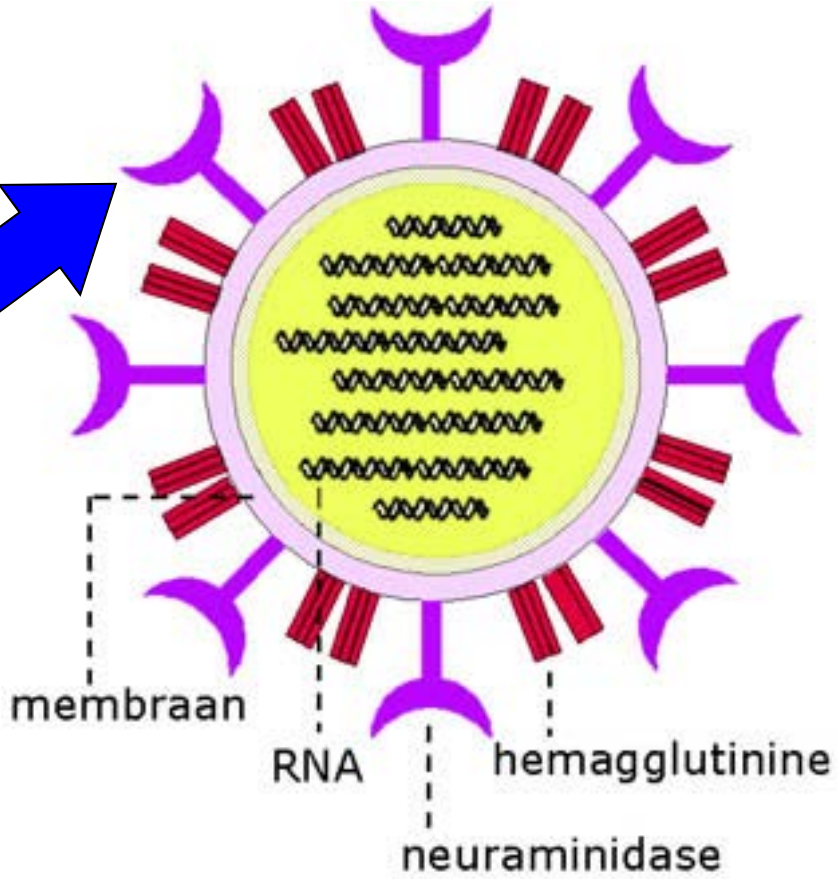
Viren


Neuraminidase
Hämagglutinin
Adhäsion
Sialinsäure-haltige Rezeptoren
respiratorische Epithelzelle
Fusion
Endocytose
Zellkern
virale RNA
virale RNA
Uncoating
Eclipse


Abspaltung der Sialinsäure durch die Neuraminidase
Freisetzung
Sialinsäure
Neuraminidase
Rezeptor
Epithelzelle des Respirationstraktes
Assemblierung
Zellkern
Virus-RNA
Virus-RNA-Kopien
virale Proteine
virale mRNA
Zell-Ribosomen

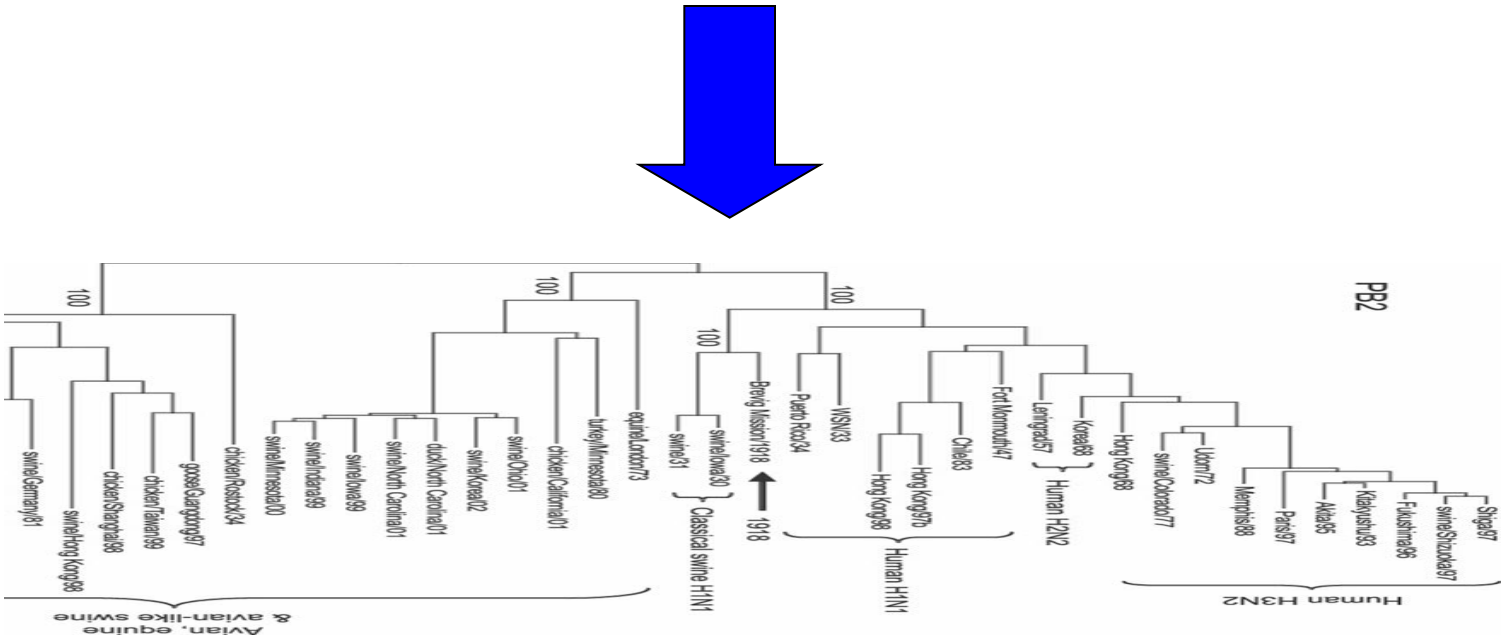
Grundlage für Stammbaumberechnung

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



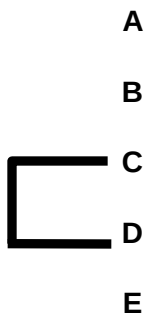
Informatik?

- M E R I K E L R D L M S Q S R T R E I L T K T T V D H M A I-3028-ATGGAAAGAATAAAAGAACTAAGGGATCTAATGTCGCAGTCTCGCACTCGCGAGATACTCACAAAAACCACCGTGGACCATATGGCCATA-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 ATCAAGAAGTACACATCAGGAAGACAGGAGAAGAACCCCGCACTTAGGATGAAATGGATGATGGCAATGAAATATCCAATTACAGCAGAC-207 K R I M E
  M I P E R N E Q G Q T L W S K T N D A G S D R V M-90 AAGAGGATAATGGAAATGATTCCTGAGAGAAATGAGCAAGGACAAACTCTATGGAGTAAAACGAACGATGCCGGATCAGACCGAGTGATG-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 GTATCACCTCTGGCTGTGACATGGTGGAATAGGAATGGACCAACGACAAGTGCAGTTCACTATCCAAAAATCTACAAAACTTATTTTGAA-387 K V E R L K H G T
  F G P V H F R N Q V K I R R R V D I N P G-150 AAAGTCGAAAGGTTAAAACATGGAACCTTTGGCCCTGTCCATTTCAGAAACCAAGTCAAAATACGTCGAAGAGTTGACATAAATCCTGGT-477 H A D L S A K E A Q D
  V I M E V V F P N E V G A R I L T S E-180 CATGCAGATCTCAGTGCCAAAGAGGCACAGGATGTAATCATGGAAGTTGTTTTCCCAAACGAAGTGGGAGCCAGGATACTAACATCGGAA-567 S Q 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 TCGCAACTGACAATAACCAAAGAAAAGAAAGAAGAACTCCAAGATTGTAAAATTTCTCCTTTAATGGTGGCATACATGTTGGAGAGAGAA-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 Q G T C W-240 CTGGTCCGAAAAACAAGATTCCTCCCAGTGGCTGGTGGGACAAGCAGTGTGTATATTGAAGTGTTGCATTTGACTCAAGGAACATGCTGG-747 E Q M Y T P G G E V R N D D V D Q S
  L I I A A R N I V R R A-270 GAACAGATGTACACTCCAGGAGGGGAAGTGAGGAATGATGATGTTGATCAAAGCTTAATTATTGCTGCTAGGAACATAGTGAGAAGAGCG-837 T V S A D P L A S L L E M C H S T Q I G
  G I R M V D I L R Q-300 ACAGTGTCAGCAGATCCACTAGCATCTCTGTTGGAAATGTGCCACAGCACACAGATTGGTGGAATAAGGATGGTAGACATCCTTAGGCAG-927 N P T E E Q A V D I C K A A M G L R I S S S
  F S F G G F T F-330 AACCCGACAGAAGAGCAAGCCGTGGATATATGCAAGGCAGCAATGGGCCTGAGAATTAGCTCATCCTTTAGCTTTGGCGGATTCACATTT-1017 K R T S G S S V K R E E E V L T G N L Q T L K I
  R V H E G Y-360 AAGAGGACAAGTGGCTCATCAGTCAAGAGGGAGGAAGAAGTGCTTACAGGCAATCTTCAAACATTGAAGATAAGAGTGCATGAGGGATAT-1107 E E F T M V G R R A T A I L R K A T R R L I Q L I V
  S G R D-390 GAAGAGTTCACAATGGTTGGGAGAAGAGCAACAGCTATACTCAGAAAAGCGACCAGGAGATTGATTCAGCTGATAGTGAGTGGGAGAGAC-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 GAACAGTCGATTGCCGAGGCAATAATTGTGGCCATGGTATTTTCACAAGAGGATTGTATGATAAAGGCAGTTAGGGGTGATCTGAATTTC-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 Q N
  W G-450 GTTAATAGGGCGAATCAGCGATTGAATCCTATGCATCAACTTTTGAGGCATTTTCAAAAGGATGCGAAAGTGCTTTTTCAAAATTGGGGA-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 ATTGAACCCATCGACAATGTGATGGGAATGATTGGGATACTGCCCGACATGACTCCAAGTACTGAGATGTCAATGAGAGGAGTGAGAGTC-1467 S K M G V D E Y S S T E R V V V S I D R F L R V R D Q R G N-
  510 AGCAAAATGGGAGTAGATGAGTACTCCAGCACAGAGAGGGTGGTGGTGAGCATTGACCGCTTTTTAAGAGTCCGGGACCAACGAGGAAAC-1557 V L L S P E E V S E T Q G T E K L T I T Y S S S M M W E V N-
  540 GTACTACTGTCTCCTGAGGAGGTCAGCGAAACACAGGGAACAGAGAAATTGACGATAACTTATTCATCGTCAATGATGTGGGAGGTTAAT-1647 G P E S V L V N T Y Q W I I R N W E T V K I Q W S Q N P T M-
  570 GGCCCTGAATCAGTGTTGGTCAACACCTATCAGTGGATCATCAGAAACTGGGAAACTGTTAAAATTCAGTGGTCACAGAATCCTACAATG-1737 L Y N K M E F E P F Q S L V P K A A R G Q Y S G F V R T L F-600
  CTATACAATAAAATGGAATTTGAGCCATTTCAGTCTTTAGTTCCTAAGGCCGCTAGAGGTCAATACAGTGGGTTTGTGAGAACTCTGTTC-1827 Q Q M R D V L G T F D T V Q I I K L L P F A A A P P K Q S R-630
  CAGCAAATGAGGGATGTGCTTGGGACATTTGACACCGTTCAGATAATAAAACTTCTTCCCTTTGCAGCCGCTCCACCAAAGCAAAGTAGA-1917 M Q 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
  ATGCAGTTCTCCTCTCTGACTGTGAATGTGAGAGGATCAGGAATGAGAATACTTGTAAGGGGCAATTCTCCCGTATTCAACTACAACAAG-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
  GCCACTAAGAGACTCACAGTTCTCGGAAAGGATGCAGGTGCTTTAACTGAAGACCCAGATGAAGGCACAGCTGGAGTGGAGTCTGCTGTT-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
  CTGAGAGGATTCCTCATTCTGGGCAAAGAAGACAGGAGATATGGGCCAGCATTAAGCATCAATGAACTGAGCAATCTTGCGAAAGGGGAG-2187 K A N V L I G Q G D V V L V M K R K R D S S I L T D S Q T A-750
  AAGGCTAATGTGCTAATTGGGCAAGGAGACGTGGTGTTGGTAATGAAACGGAAACGGGACTCTAGCATACTTACTGACAGCCAGACAGCG-2277 T K R I R M A I N Stop ACCAAAAGAATTCGGATGGCCATCAATTAG-2307

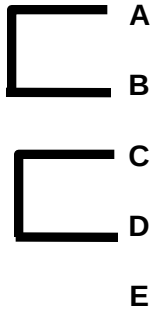


Konstruktion des Guide Trees

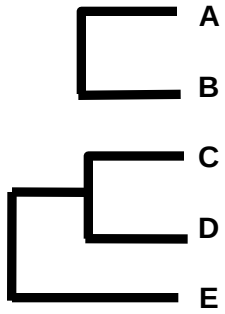
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21

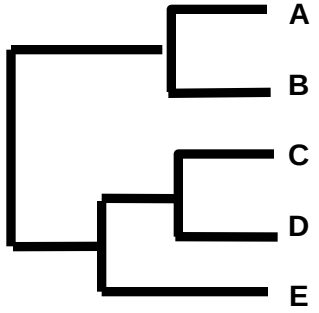


	A	B	E	CD
A		17	77	59
B			53	49
E				31

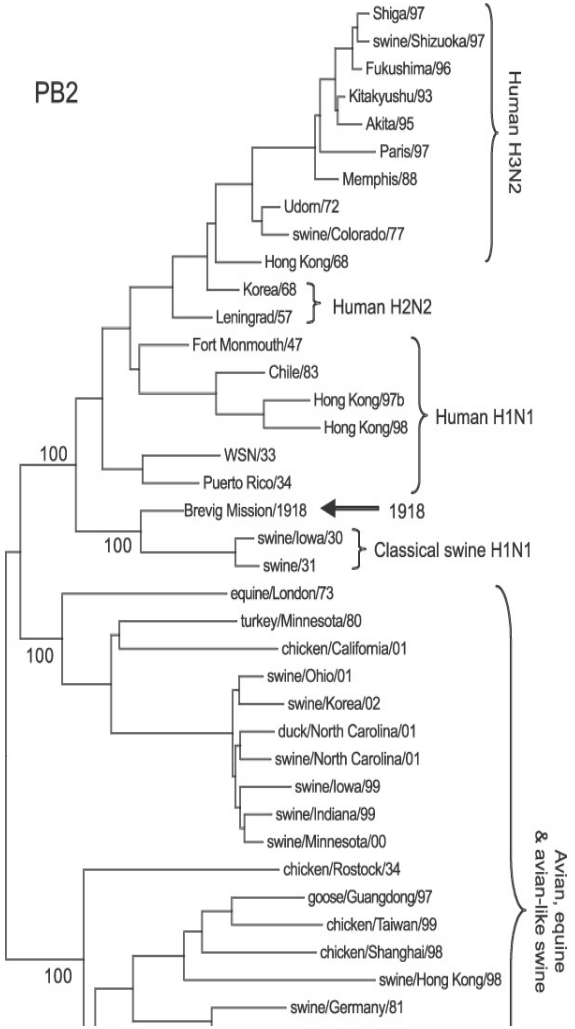


	E	CD	AB
E		31	65
CD			54



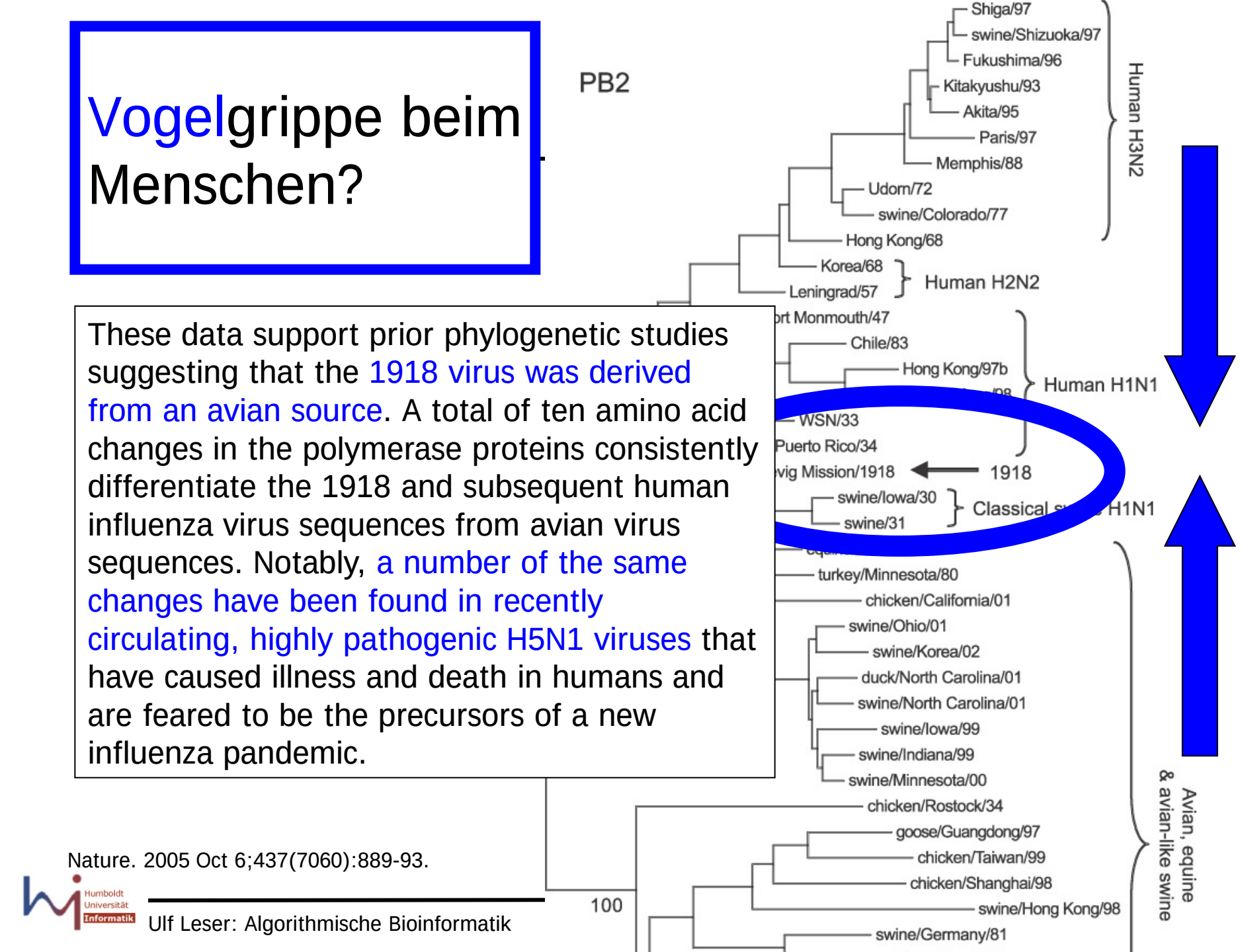


Was bedeutet der Stammbaum?



- 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?



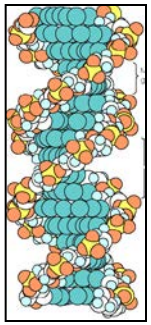
These data support prior phylogenetic studies suggesting that the 1918 virus was derived from an avian source. A total of ten amino acid changes in the polymerase proteins consistently differentiate the 1918 and subsequent human influenza virus sequences from avian virus sequences. Notably, a number 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.

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

Inhalte der Vorlesung

- Einführung
- Ein konkretes Beispiel
- Themen der Vorlesung

Diese Vorlesung



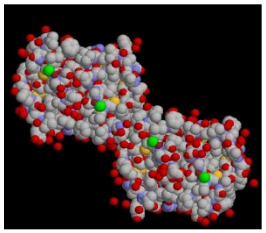
Sequenzierung

Erkennung von Genen

Verwandtschaft zw. Spezies

Regulation & Expression

RNA Gene – „Dark matter“



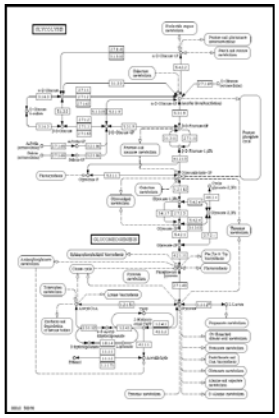
Dreidimensionale Faltung

Strukturvergleich und -ähnlichkeit

Interaktion

Sekundärstruktur

Proteinidenti-fikation



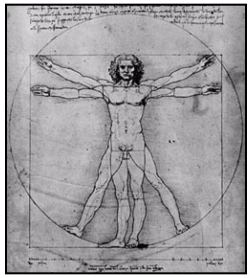
Netzwerkanalyse

Geschwindigkeit von Reaktionsketten

Stoffumsatz

Kompartamente

Muster und Redundanz



Korrelation Phänotyp/Genotyp

Arzneimittel-empfänglichkeit

Studienstatistik und -verwaltung

Stringalgorithmen

- Gegeben ein Template T und ein Pattern P. Finde alle Vorkommen von P in T in möglichst kurzer Zeit
  - Exaktes Matching

- Z-Box — Vorverarbeitung von P
- Boyer-Moore — Schnellster in Praxis
- Knuth-Morris-Prath — Elegante Analyse; Erweiterbar
- Varianten
  - Suche nach mehreren P: Aho-Corasick, Keyword Trees
  - Suche mit Wildcards
  - Suche mit regulären Ausdrücken (= endlichen Automaten)

Indexstrukturen

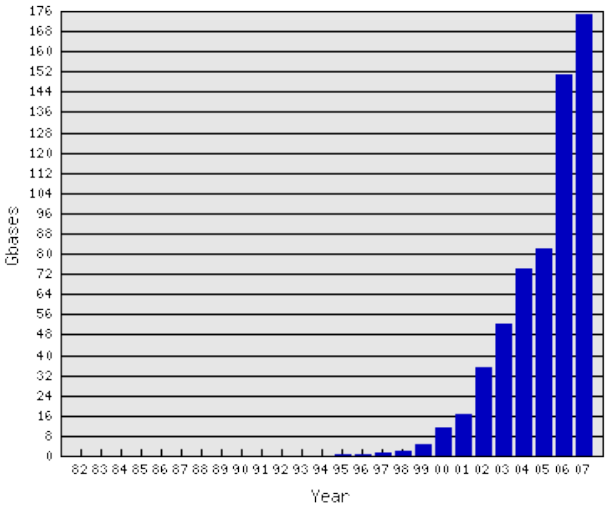
- Gegeben ein festes T und dauernd wechselnde P. Finde eine Datenstruktur für T, die die Suche nach allen P in möglichst kurzer Zeit gestattet
- Grundlegende Datenstruktur: Suffixbäume
- Suche und Konstruktion
  - Ukkonen‘s linearer Algorithmus
- Verschiedene Anwendungen
  - Längster Substring
  - Längster 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

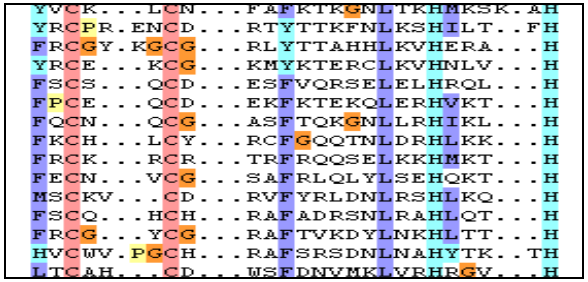
- Quadratische Laufzeit ist zu teuer
  - Genomanalyse benötigt Suche auf allen bekannten Sequenzen
  - Celera Sequenzierung: All-against-all Vergleich von 28.000.000 Teilsequenzen



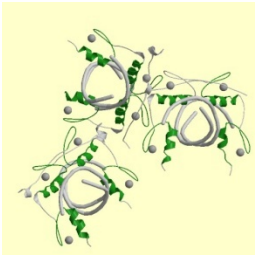
- Also: Heuristiken, z.B. BLAST
  - Suche nach „Seeds“ mit exakten Matches
  - Verlängerung und Zusammenfügung der Seeds zu Matches
  - Findet nicht alle Hits, aber die meisten „interessanten“
  - „BLAST“ ist fast Synonym für Bioinformatik geworden

Multiples Alignment

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



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

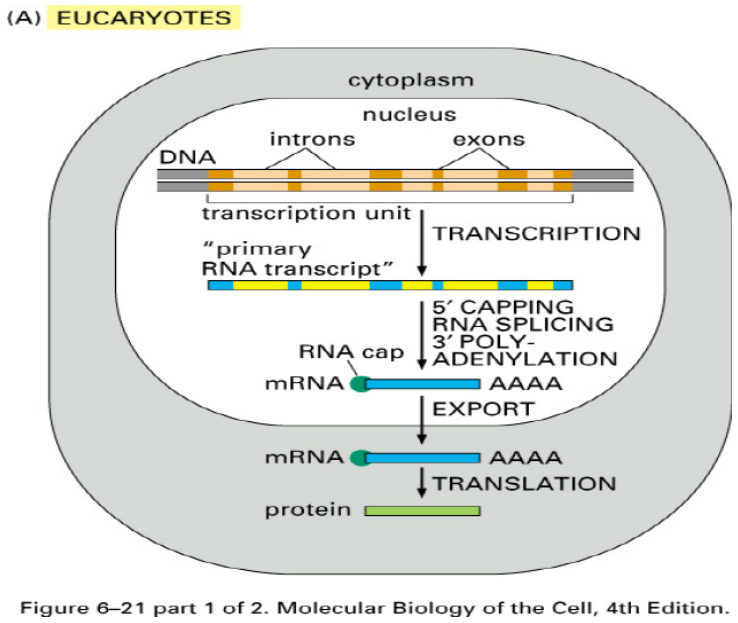
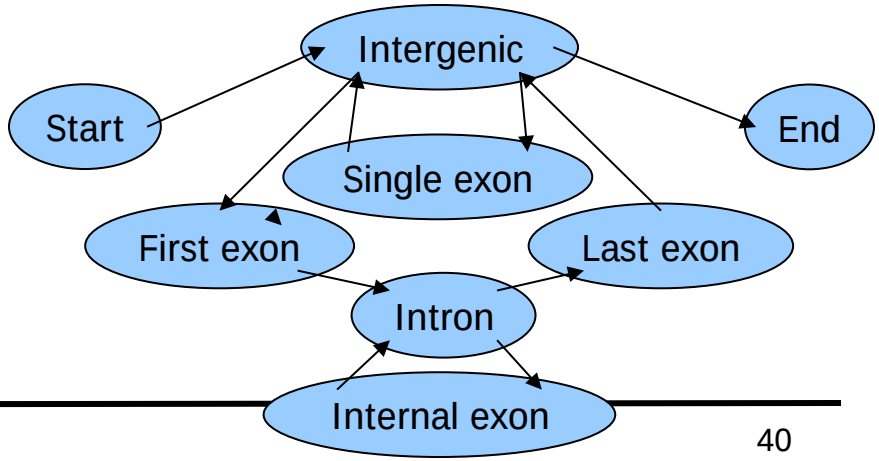
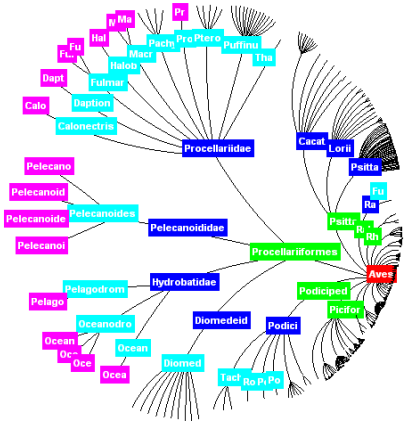


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.



Phylogenetische Bäume

- Grundannahme
  - Spezies entstehen durch Evolution
  - Also gibt es gemeinsame Vorfahren; Spezies stehen in Vater-Kind Beziehungen
- Phylogenie = „Baum der Evolution“
  - Auch: Berechnung des Evolutionsbaums
  - Beantwortung des Taxonomie-problems auf molekularer Basis



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?