

Algorithmische Bioinformatik

Halbkurs

Wintersemester 2007 / 2008

Ulf Leser

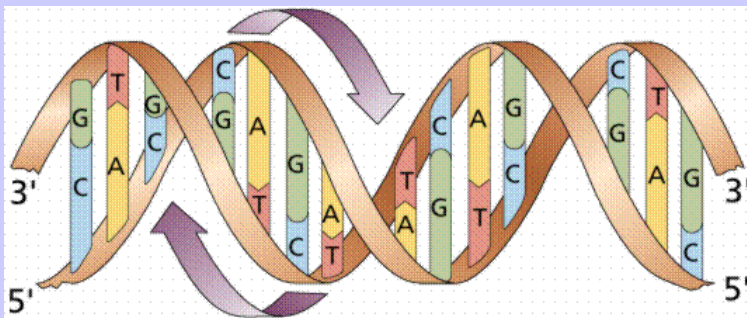
Wissensmanagement in der
Bioinformatik



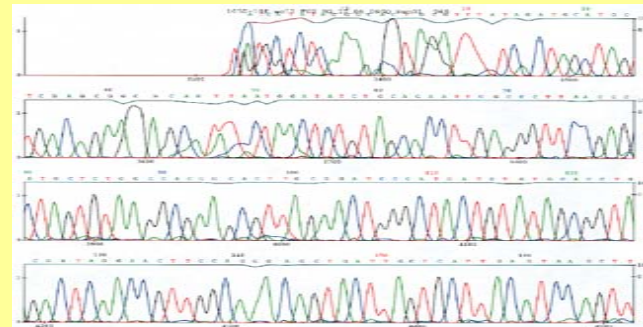
Bioinformatik



25.4.2003
50. Jubiläum der
Entdeckung der
Doppelhelix durch
Watson/Crick



14.4.2003
Humanes Genom zu 99%
sequenziert mit 99.99%
Genauigkeit



Fragen ...

- Was heißt das?
- Warum sind 3.000.000.000 Zeichen für uns so wichtig?
- Wenn das Zeichen sind – gibt es Wörter? Eine Sprache?
- Wie wurden die Daten erhoben?
- Wie kann man daraus Wissen gewinnen?

- Welche Rolle spielt dabei die Informatik?

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 ²⁾	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 €
Azotämie	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	FMRI-(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 €

Halbkurs Algorithmisch Bioinformatik

- Vorlesung 4SWS
- Übung 2SWS
- Sprechstunde: Nach Vereinbarung
Ulf Leser
Raum: IV.103
Tel: (030) 2093 – 3902
eMail: leser (..) informatik . hu-berlin . de

Termine und Prüfung

- Vorlesung
 - Dienstag, 11-13 Uhr
 - Donnerstag, 11-13 Uhr
- Übung
 - Donnerstag, 13-15 Uhr
- Voraussetzung für Prüfung
 - Je nach Studienordnung: Vordiplom
 - Bestehen der Übung (alle Aufgaben)
 - **Verständnis der Algorithmen**

Gäste

- Die nächsten 5 Doppelstunden
 - Prof. Dr. Thomas Börner
 - Professor für Genetik
 - Zellen, Chromosomen, Gene, Transcription und Translation, differentielles Splicen, ...
- Im Verlaufe des Semesters
 - Prof. Krause, FH Bingen: Clustering von Proteinsequenzen
 - Dr. Schuchart, MicroDiscovery GmbH: Industrielle Bionformatik

Literatur

- Primär
 - Dan Gusfield: „Algorithms on Strings, Trees, and Sequences“, Cambridge University Press, 1997 (ca. 60 Euro)
- Weitere
 - Böckenhauer, Bongartz: „Algorithmische Grundlagen der Bioinformatik“, Teubner, 2003 (ca. 40 Euro)
 - David Mount: „Bioinformatics. Sequence and Genome Analysis“, Cold Spring Harbour Press, 2001 (ca. 70 Euro)
 - Setubal & Meidanis: „Introduction to Computation Molecular Biology“, PWS Publishing, 1997 (ca. 90 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 address bar containing the URL: http://zope.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws0708/hk_algbio/. The browser's menu bar includes 'Datei', 'Bearbeiten', 'Ansicht', 'Chronik', 'Lesezeichen', 'Extras', and 'Hilfe'. The address bar also shows a search engine dropdown set to 'Google' and a search input field. Below the address bar, there are several bookmarked sites: 'Nachsehen', 'Frequent', 'Google', 'WBI', 'Lehre', 'News', 'Suchen', 'Buecher kaufen', 'Paper suchen', 'Reisen', and 'MyStuff'. The browser's tab bar shows two tabs: 'lord semantic goble dblp - Google-Suche' and 'Halbkurs Algorithmische Bioinfor...'. The main content area of the browser displays the website 'Halbkurs Algorithmische Bioinformatik'. The website has a blue header with the Humboldt University logo and the text 'Institut für Informatik'. Below the header, there is a navigation menu with links for 'Deutsch', 'English', 'Wissensmanagement in der Bioinformatik', 'Kontakt', 'Mitarbeiter', 'Veranstaltungen', 'Lehre', 'Archiv', and 'WS 07/08'. The main content area features a logo for the 'Center for Genome Based Bioinformatics' and the title 'Halbkurs Algorithmische Bioinformatik'. The text describes the course as 'Halbkurs im Wintersemester 2007/2008' led by 'Professor Ulf Leser'. It details the course content, which includes algorithms for solving fundamental questions in modern molecular biology, such as exact string matching, approximate matching, index structures, sequence databases, edit distance, alignment, multiple alignment, phylogenetic trees, and sequence assembly. The course also covers evolutionary phylogenetic trees. The website lists prerequisites, examinations, location and time, literature, and topics and dates for individual sessions. The footer of the website includes the Humboldt University logo and the text 'Humboldt Universität Informatik'.

Halbkurs Algorithmische Bioinformatik

Halbkurs im Wintersemester 2007/2008
Professor Ulf Leser

Der Halbkurs "Algorithmische Bioinformatik" behandelt Algorithmen zur Lösung grundlegender Fragestellungen moderner Molekularbiologie. Nach einer ausführlichen 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.

Voraussetzungen

Voraussetzung für den Besuch sind grundlegende Kenntnisse in Algorithmen. Kenntnisse in der Molekularbiologie werden nicht vorausgesetzt, sondern vermittelt.

Prüfungen

Prüfungen sind mündlich. Die Vorlesung ist als Halbkurs der praktischen Informatik anrechenbar.

Ort / Zeit:

- Dienstag, 11.00 - 13.00, RUD26, Raum 1'303
- Donnerstag, 11.00 - 13.00, RUD26, Raum 1'313

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).

- Einleitung und Überblick
- [Einführung in die Molekularbiologie](#), Prof. Thomas Börner, Institut für Biologie
Findet statt am 18.10.2007: 11-15 Uhr; 23.10.2007: 11-13 Uhr; 25.10.2007: 11-15 Uhr
- Sequenzierung, cDNA Clustering, Funktionsvorhersage; Strings und Stringvergleiche
- Lineares Stringmatching: Z-Box Algorithmus
- Boyer-Moore Algorithmus: Apostolico-Giancarlo Variante

Fragen ?

Fragen meinerseits

- Diplominformatiker?
- Semester?
- Prüfung?

- Spezielle Erwartungen?

Beispiel 1: H5N1

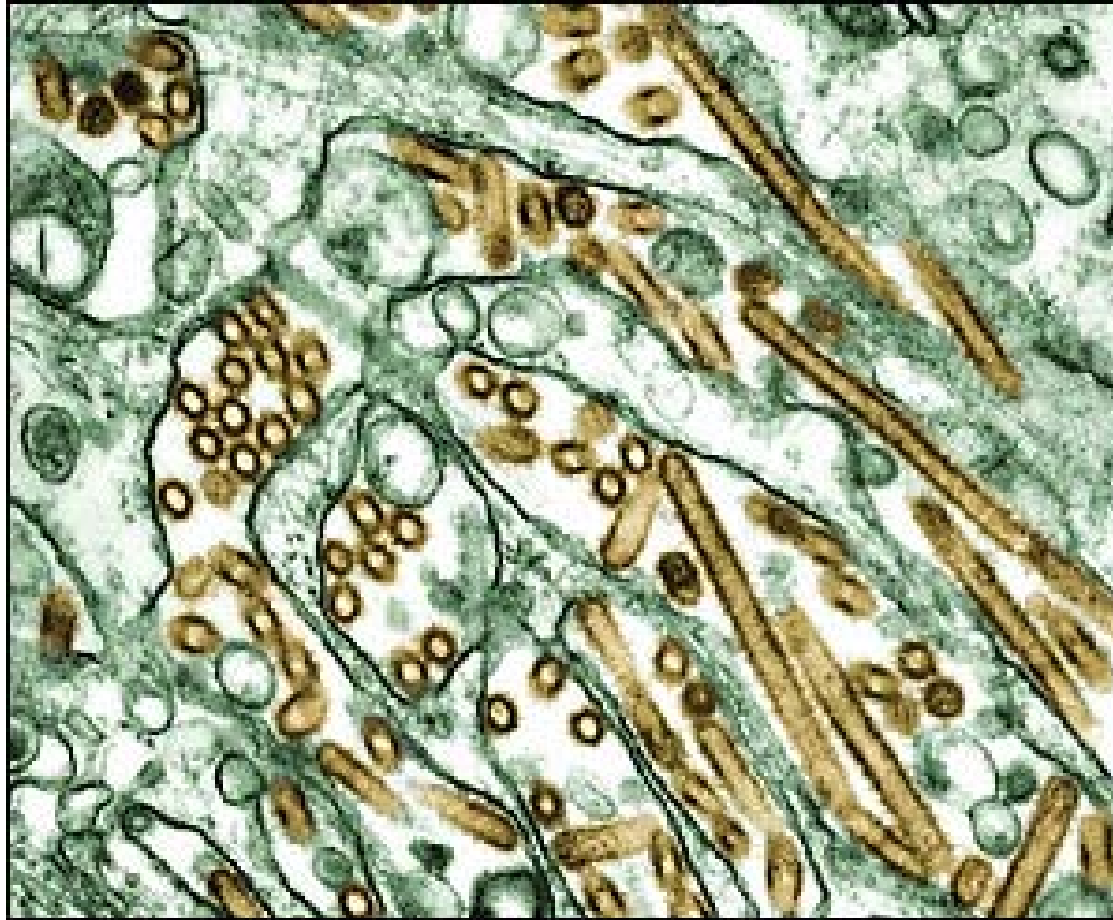
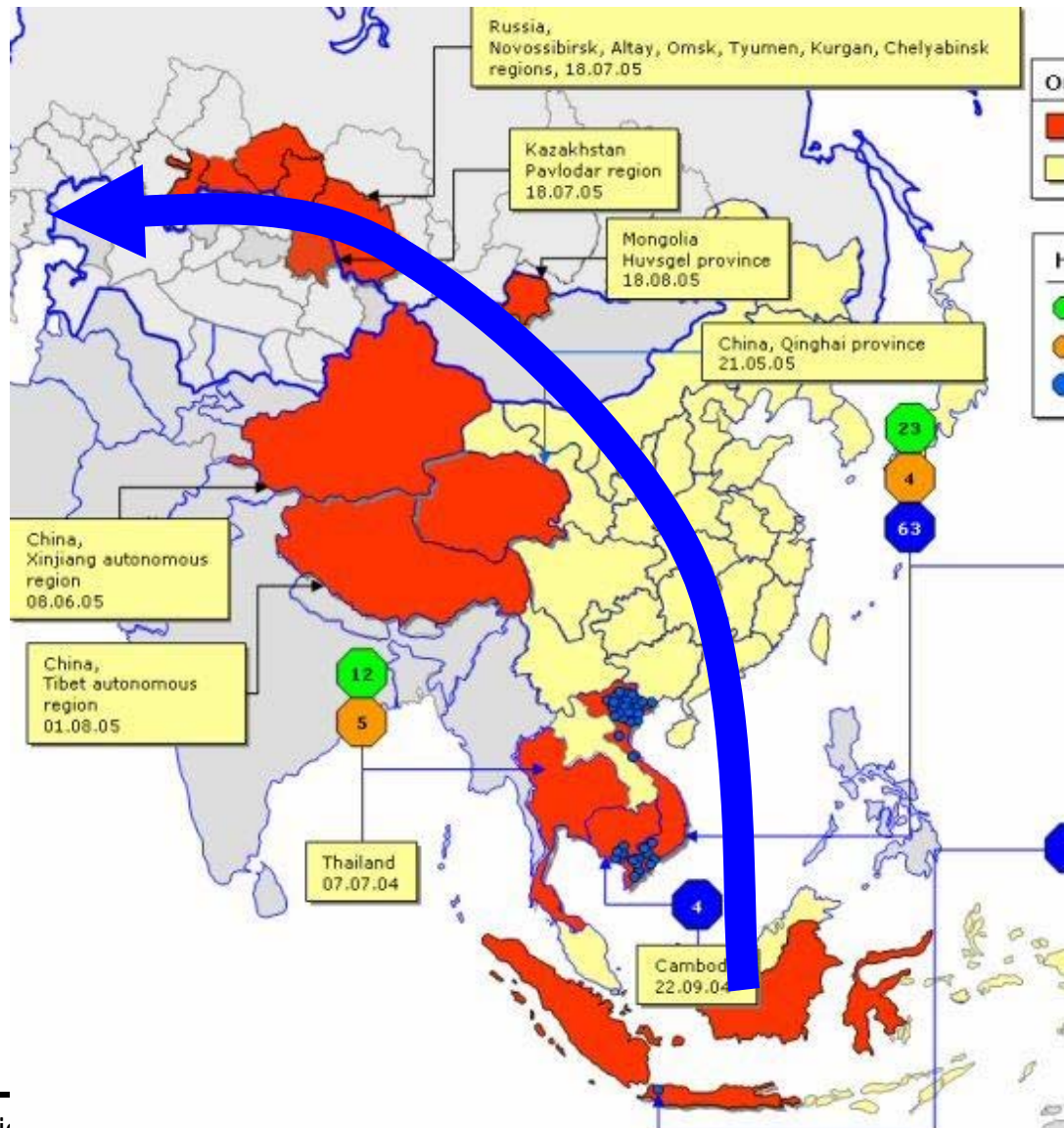
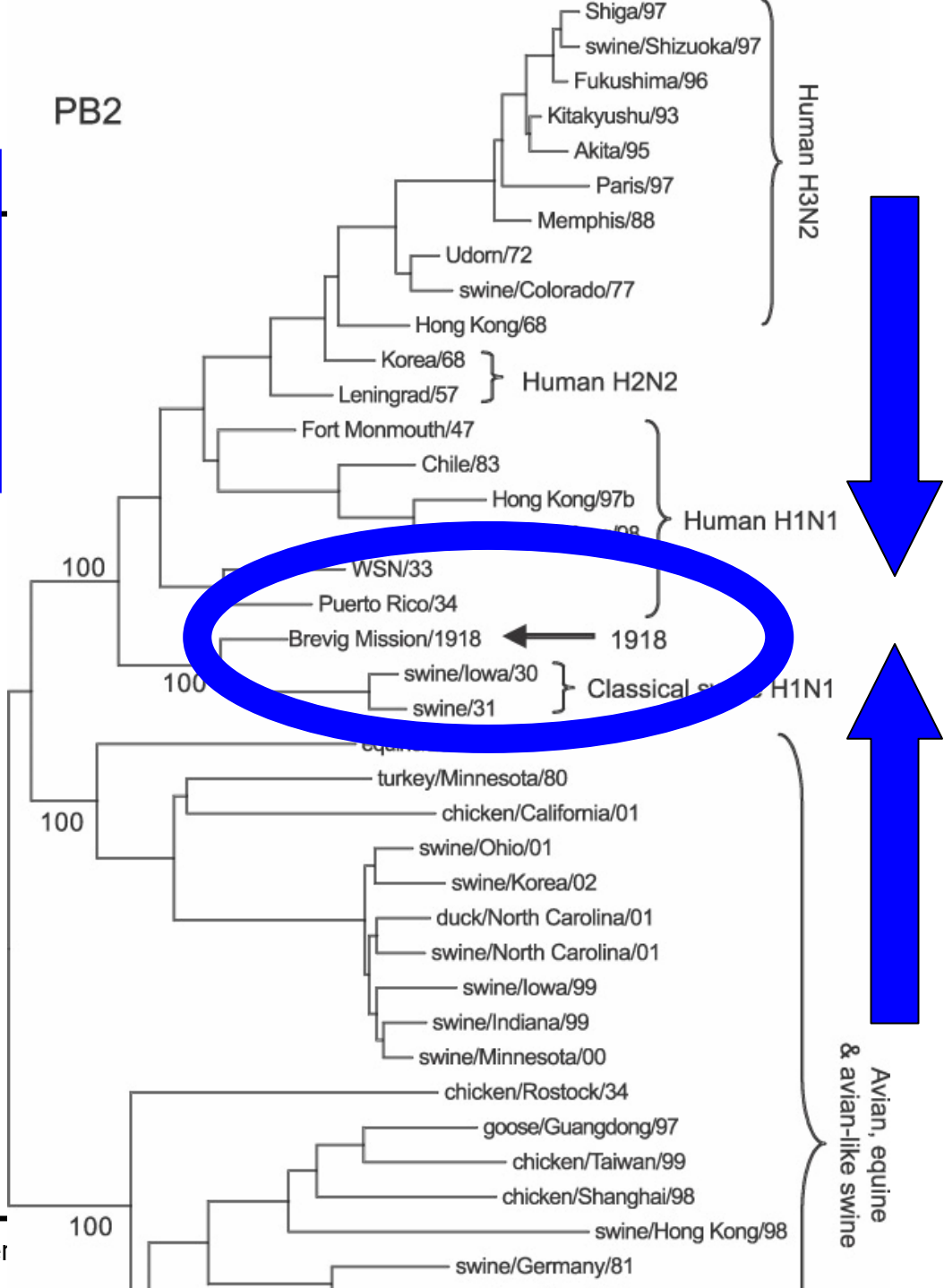


Foto: Centers for Disease Control

Migration

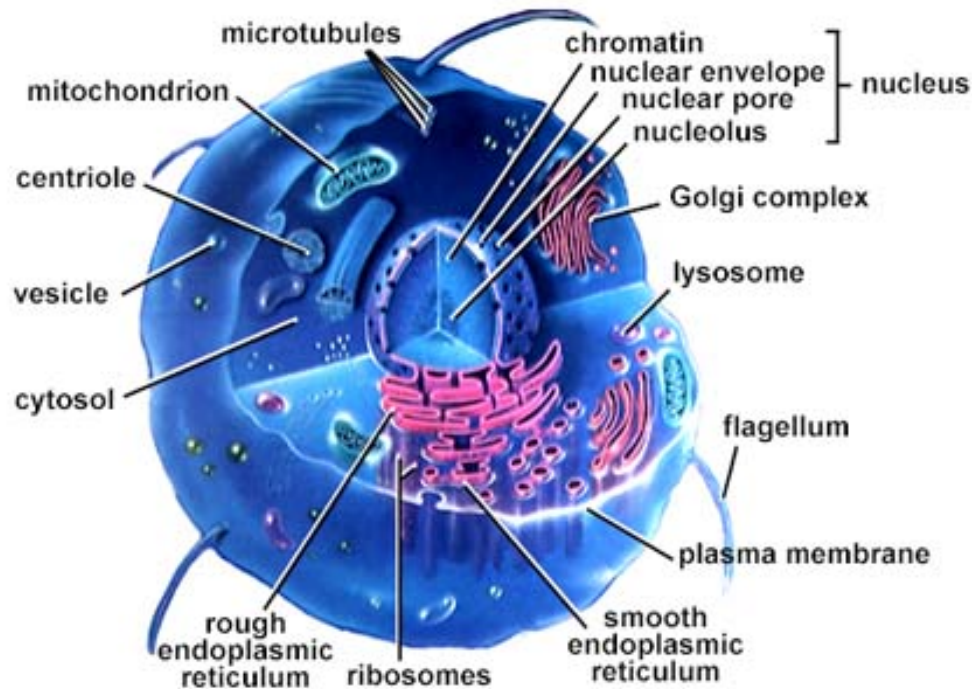


Vogelgrippe
beim Menschen?



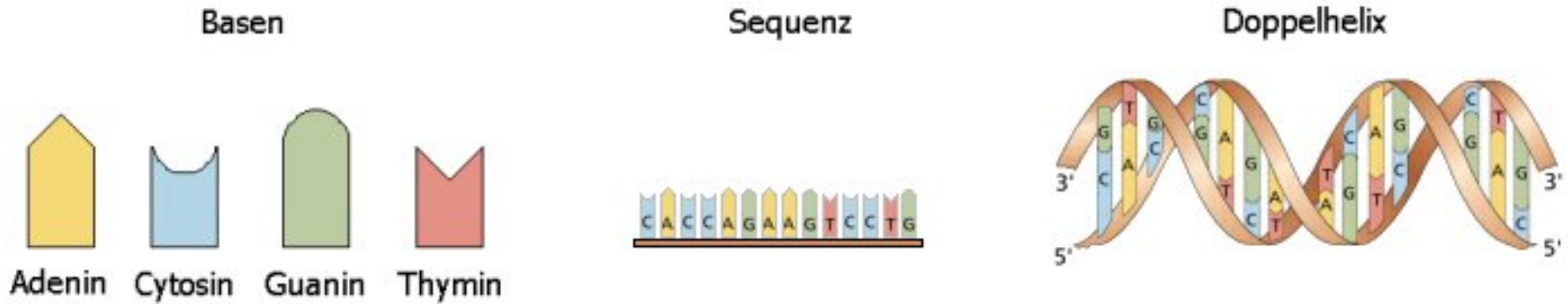
Nature. 2005 Oct 6;437(7060):889-93.

Zellen



- Ca. 75 Billionen Zellen im menschlichen Körper
- Ca. 250 verschiedene Typen: Nerven, Haut, Muskeln, ...

DesoxyriboNucleicAcid



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

Das menschliche Genom

• ... AGGCTGATGGATTAGAGACC ...

• 23 Chromosomenpaare

• ~ 3.000.000.000
Buchstaben

• ~ 50% bestehen
aus 4 „Parasiten“

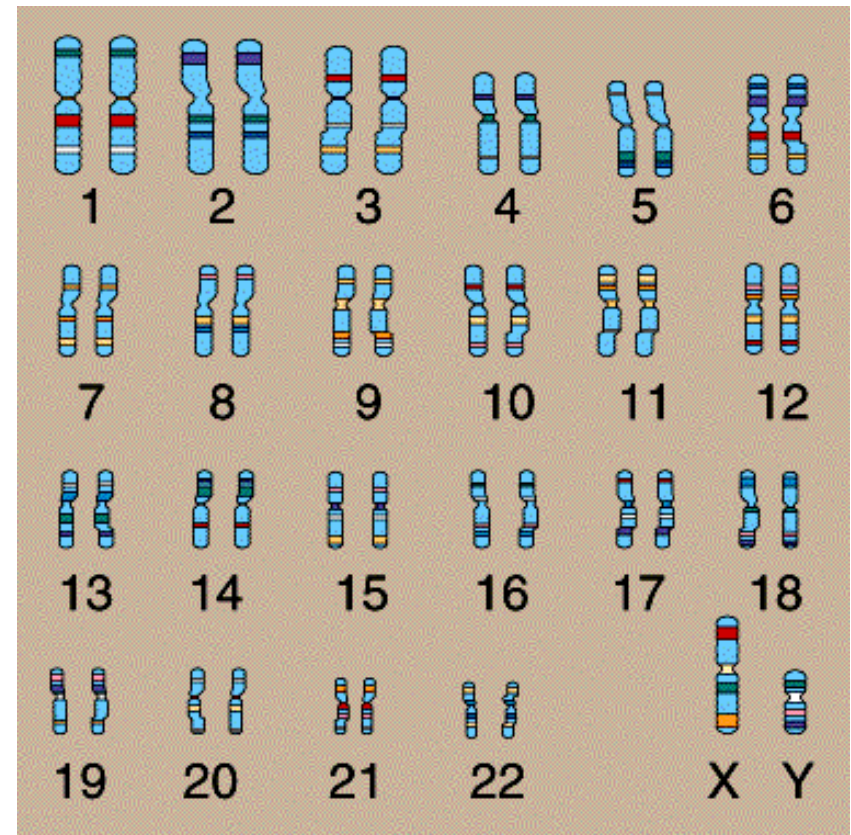
~~• ~ 100.000 Gene~~

~~• ~ 56.000 Gene~~

~~• ~ 30.000 Gene~~

~~• ~ 24.000 Gene~~

• ~ 20.000 Gene



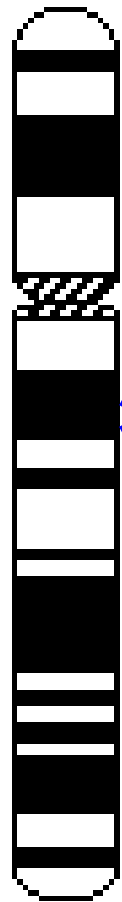
Was ist ein Gen?

Chromosom

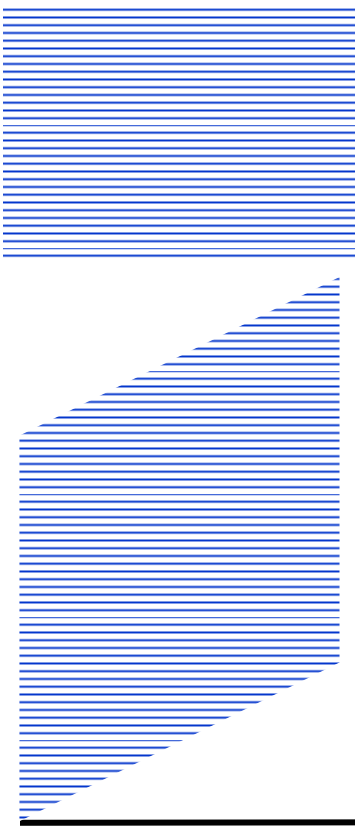
DNA

RNA

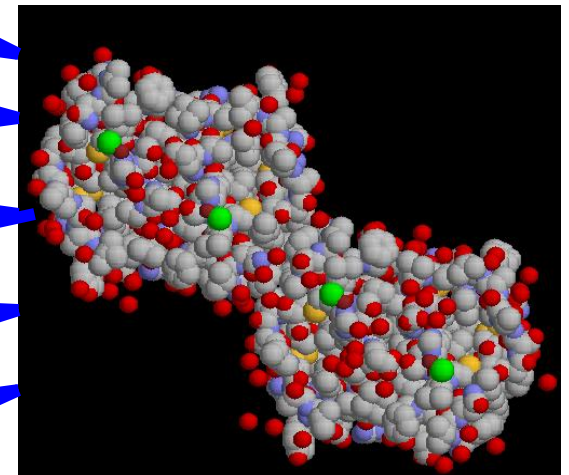
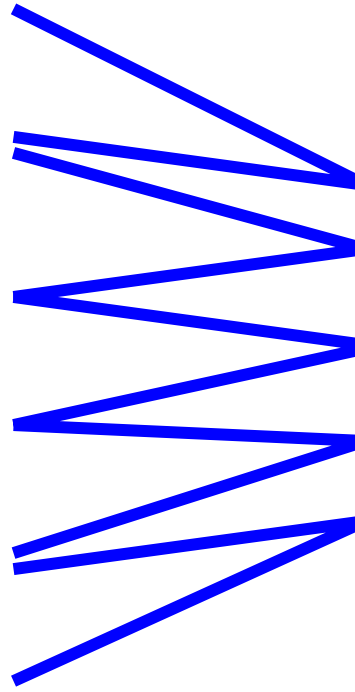
Protein



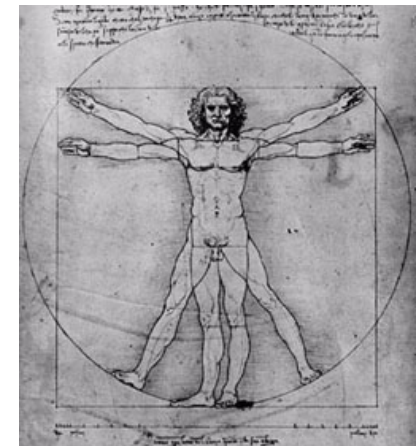
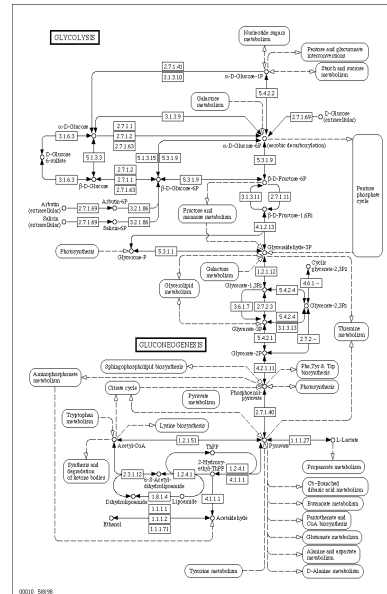
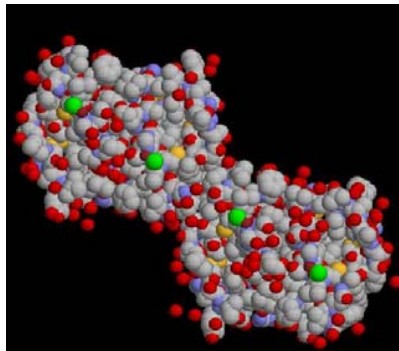
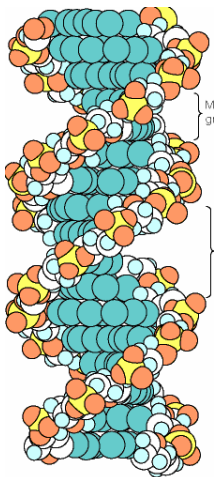
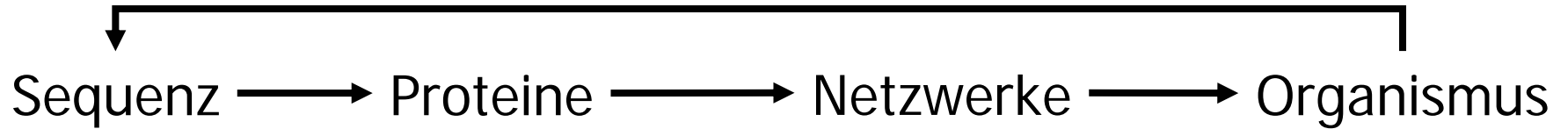
A
C
G
T
T
G
A
T
G
A
C
C
A
G
A
G
C
T
T
G
T



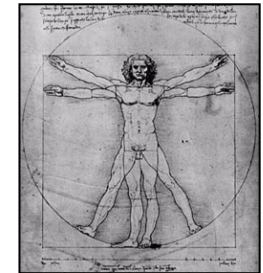
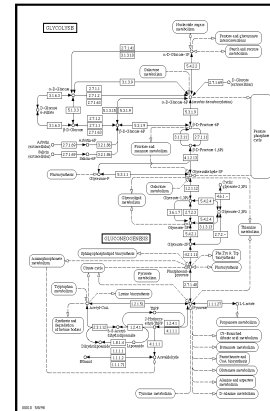
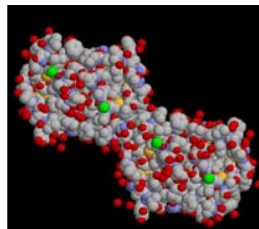
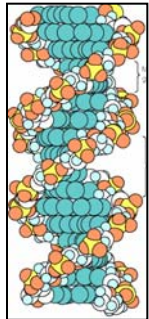
A
C
G
T
T
G
A
C
A
G
A
G
C
T
T
G
T



Der Weg des Lebens



Welche Rolle spielt dabei die Informatik ?



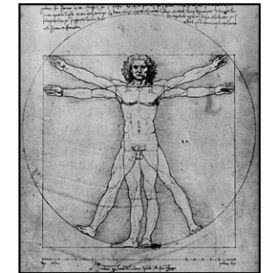
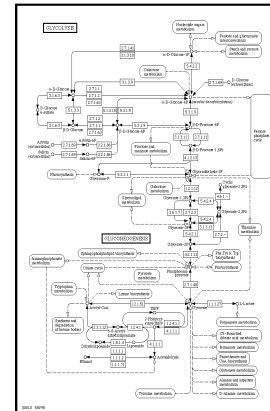
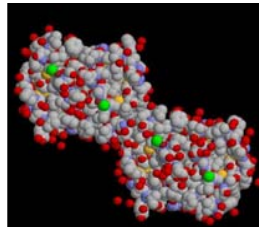
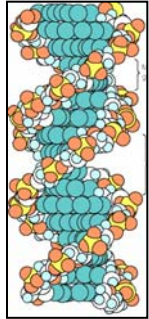
Sequenzierung
Erkennung von Genen
Verwandtschaft zw. Spezies
Regulation & Expression
RNA Gene – „Dark matter“

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

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

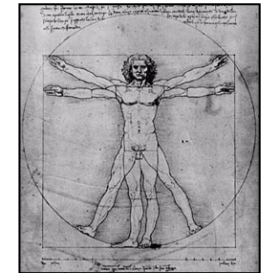
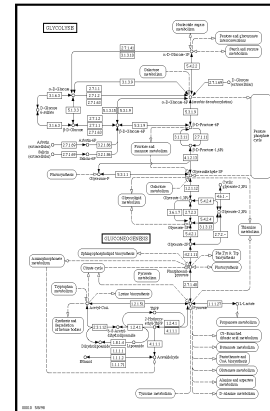
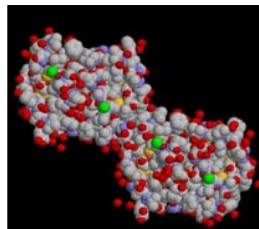
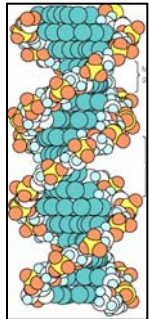
Muster und
Redundanz

Korrelation
Phänotyp/Genotyp

Arzneimittel-
empfindlichkeit

Studienstatistik
und -verwaltung

Seminar: Data Mining in der Bioinformatik



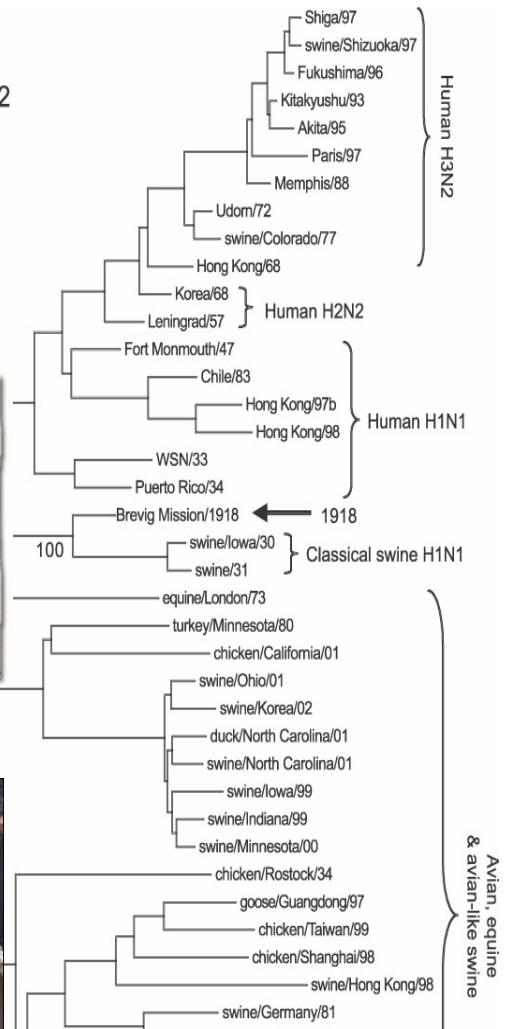
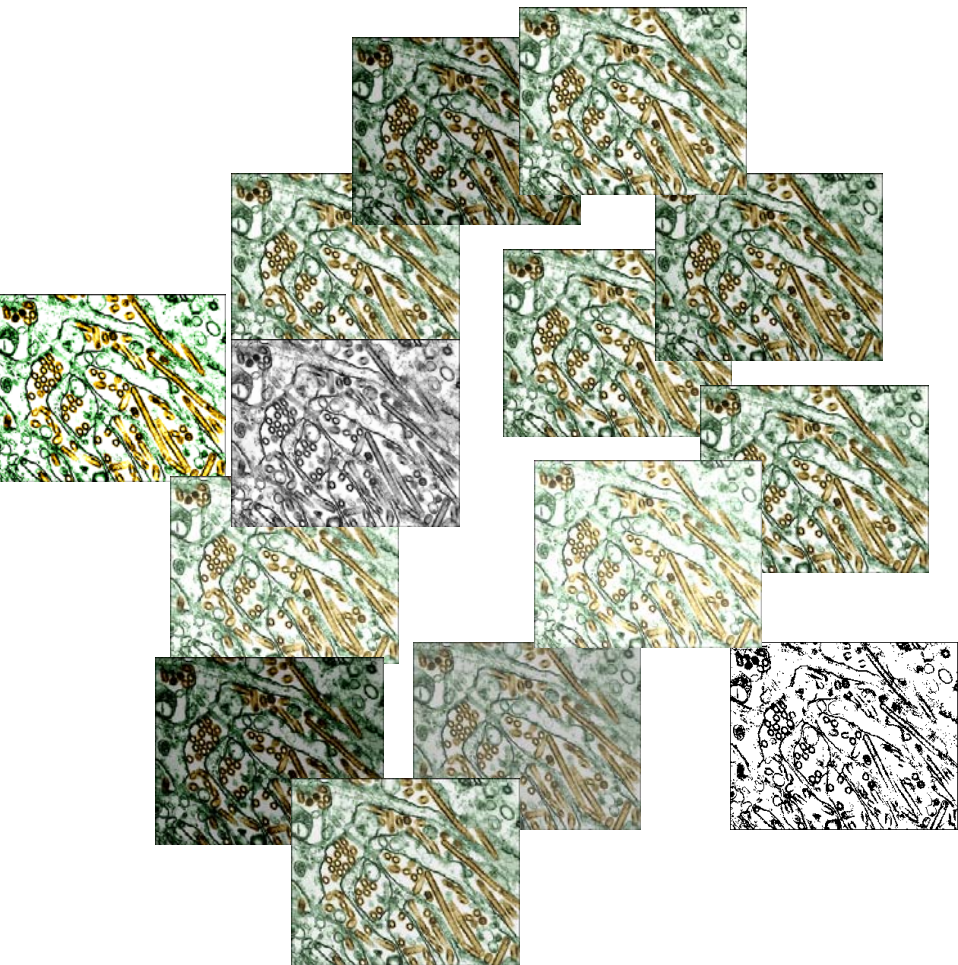
Sequenzierung
Erkennung von Genen
Verwandtschaft zw. Spezies
Regulation & Expression
RNA Gene – „Dark matter“

Dreidimensionale Faltung
Strukturvergleich und -ähnlichkeit
Interaktion
Sekundärstruktur
Proteinidentifikation

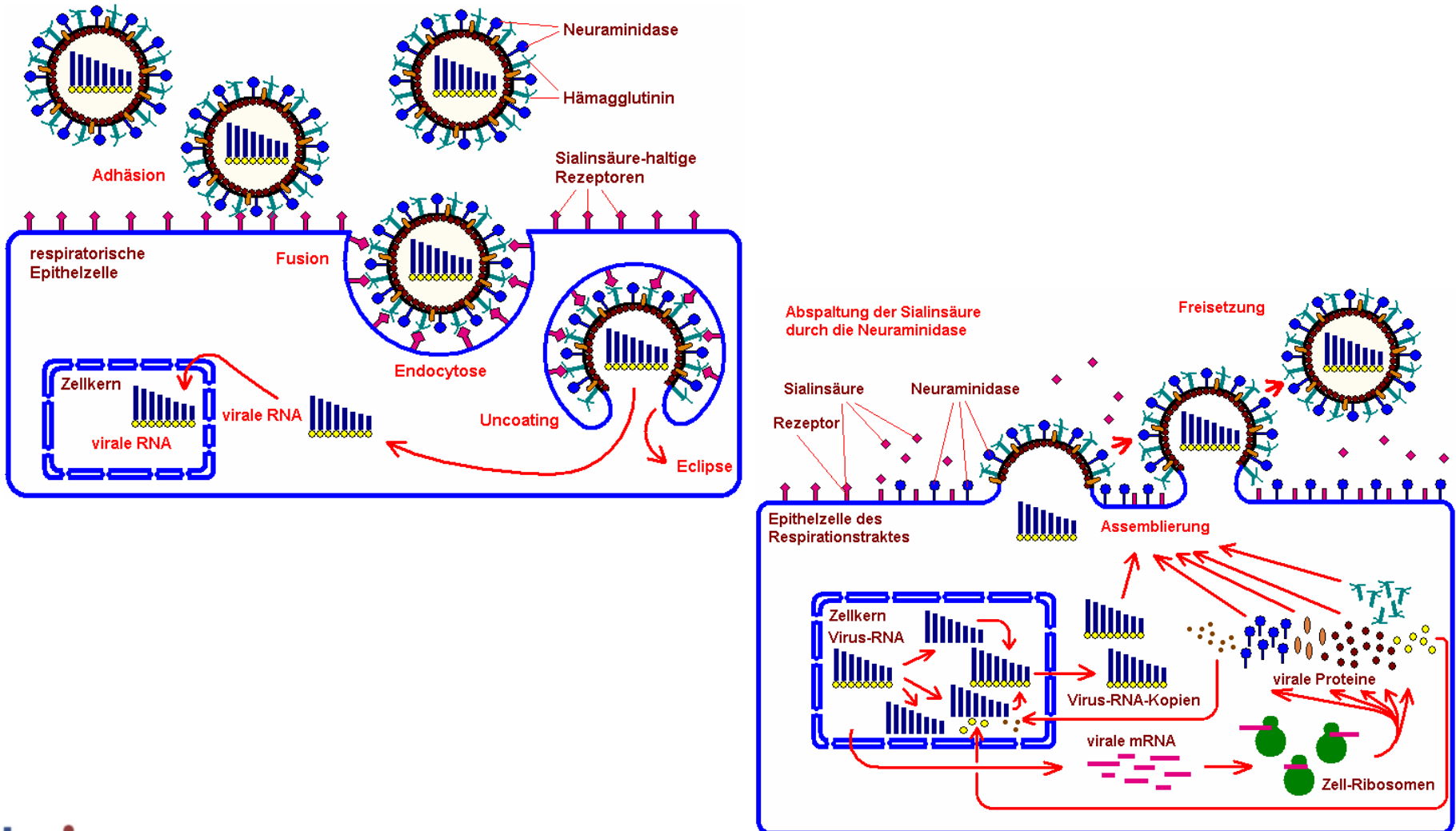
Netzwerkanalyse
Geschwindigkeit von Reaktionsketten
Stoffumsatz
Kompartamente
Muster und Redundanz

Korrelation Phänotyp/Genotyp
Arzneimittel-empfindlichkeit
Studienstatistik und -verwaltung

Wo kommt der Stammbaum her?

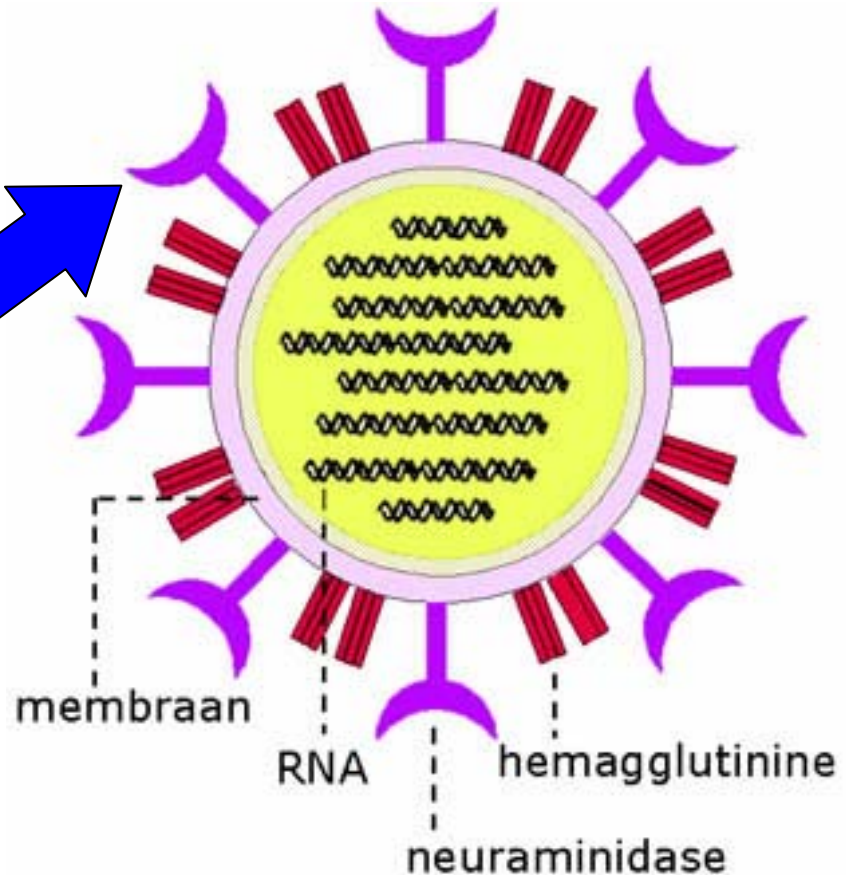


Viren



Grundlage für Stammbaumberechnung

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

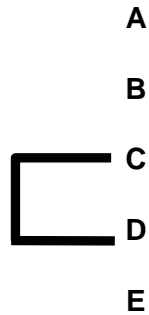


Abstandsbasierte Phylogenie

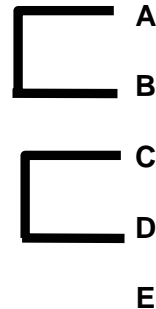
- Gegeben k Sequenzen
- Berechne die Ähnlichkeitsmatrix M
 - $O(k^2)$ paarweisen Alignment scores
- Hierarchisches Clustering
 - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
 - Das ist das erste Paar
 - Erzeuge M' : Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (ij) ein
 - Für alle $k \neq ij$: $M'[ij,k] = (M[i,k] + M[j,k]) / 2$
 - Mittlerer Abstand zu i und j
 - Iteriere, bis Matrix nur noch 2×2 groß ist
 - Baue parallel bei jeder Entnahme eines Paares einen Baum auf

Konstruktion des Guide Trees

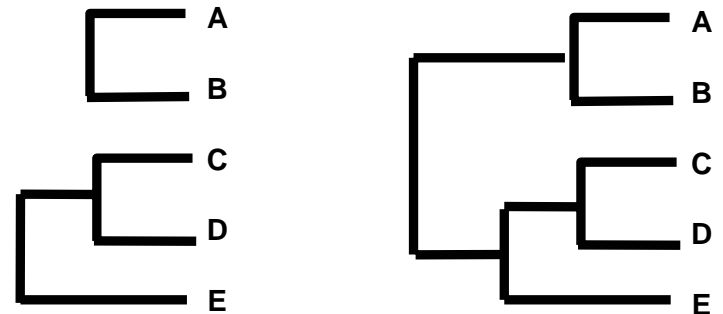
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



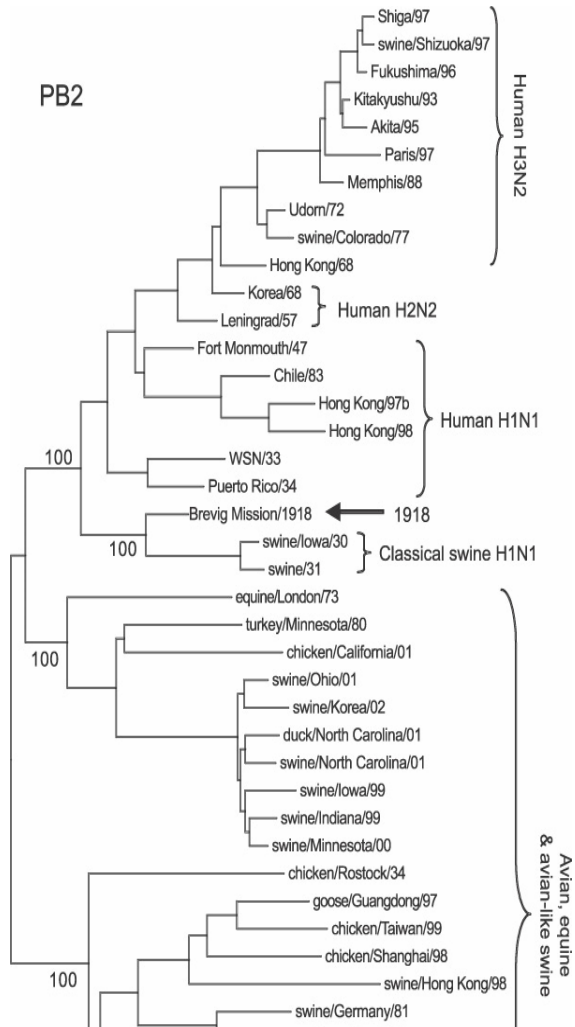
	A	B	E	CD
A		17	77	59
B			53	49
E				31



	E	CD	AB
E		31	65
CD			54



Was bedeutet der Stammbaum?

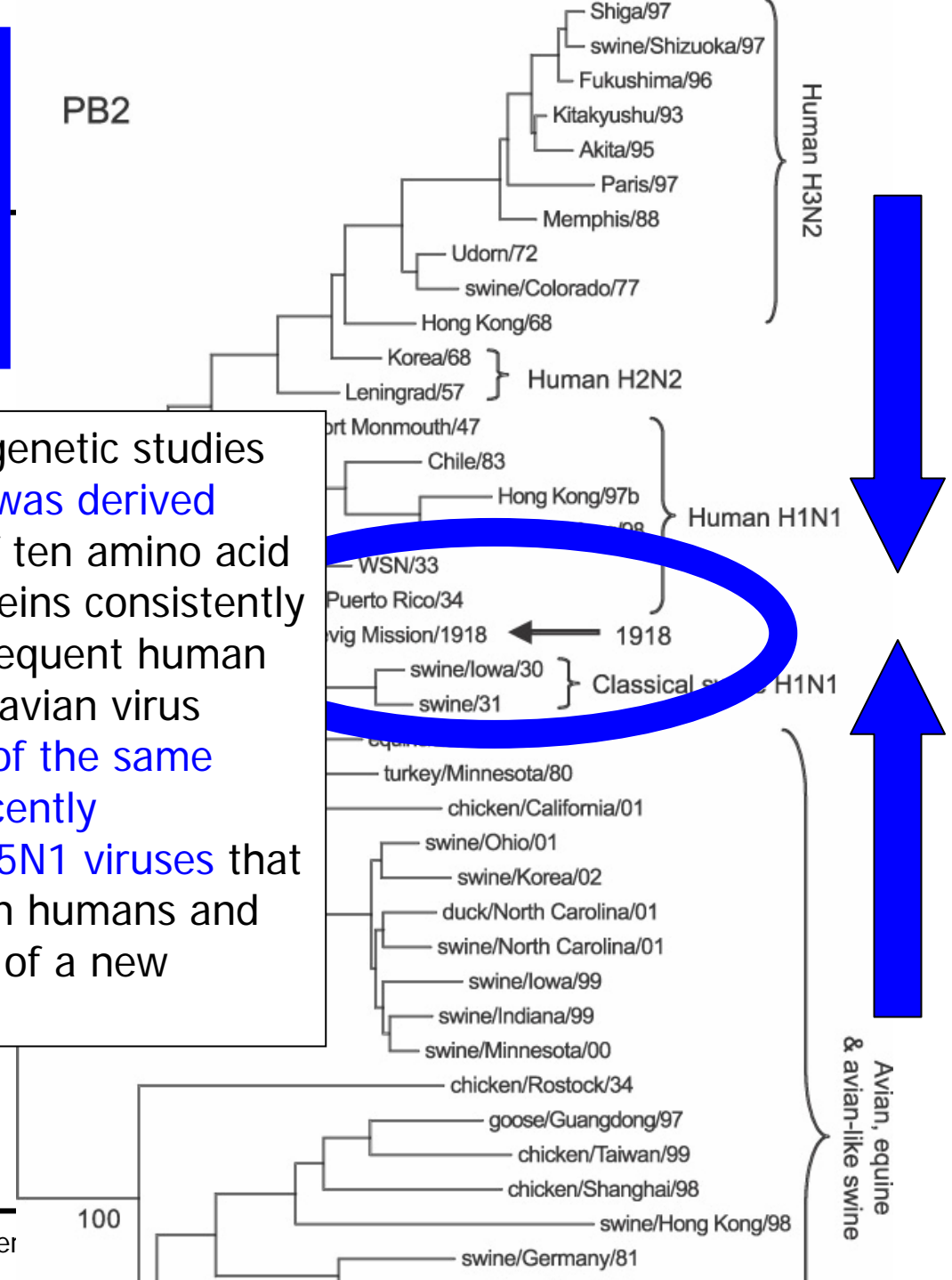


- Grundidee ist Evolution
 - Ur-Virus und Speziation
 - Richtungslose Mutationen
 - Selektion durch Umwelt (Host!)
- Benachbarte Stämme haben ähnliche Sequenzen
- Ähnliche Sequenzen bedingen ähnliche Funktion der Proteine und wahrscheinlich auch ähnliche Wirkung
- Um benachbarte Knoten zu „erreichen“,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

Beispiel 2: Sequenzbestimmung

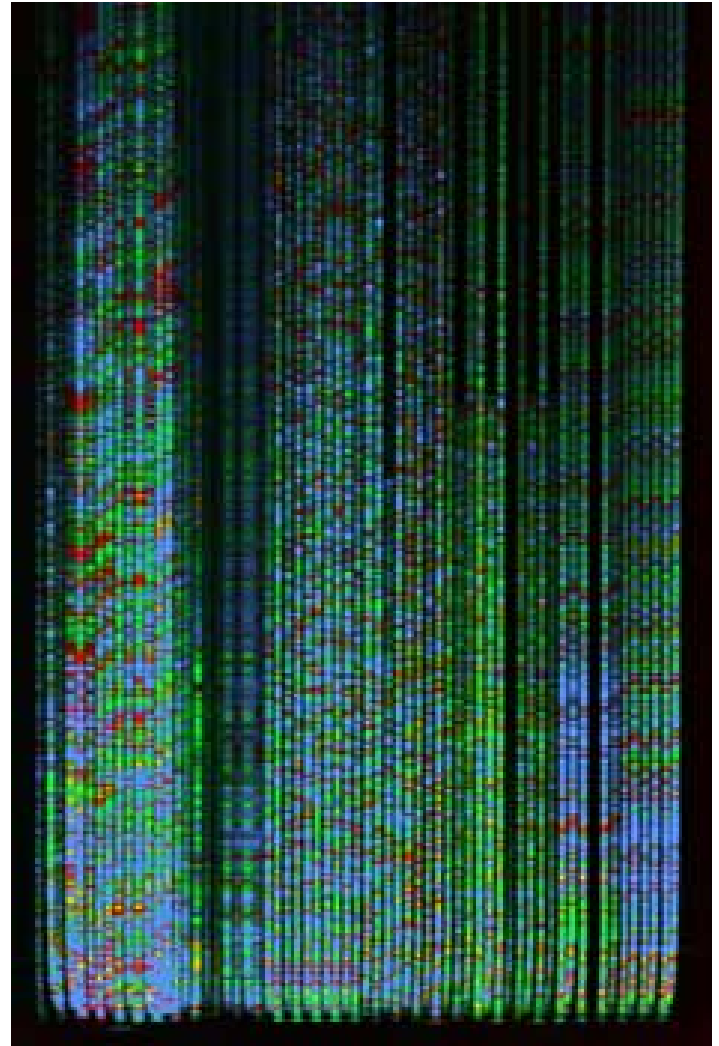
- Ein Genom wird nicht entschlüsselt, sondern „gemessen“
- Direkt messen kann man nur sehr kurze Stücke
- Also wird ein Chromosom in viele Einzelteile zerlegt
- Diese werden einzeln gemessen
- **Bioinformatik-Problem:** Aus den Sequenzen der Bruchstücke die Sequenz des ursprünglichen Chromosoms berechnen



Quelle:
<http://www.geneticsplace.com>

Experimentelles Ergebnis

- Parallele Messung von 36 kurzen DNA Stücken
- Jede Farbe steht für eine Base
 - Adenin, Cytosin, Guanin, Thymin
- Erste Aufgabe
 - Bilderkennung
 - Signalerkennung
- Ergebnis: $\{A,C,G,T\}^*$
- **Fehlerrate:** Ca. 1% pro Read
 - Also 30.000.000 Fehler im Genom?



Assembly

Read 1 **AGGTCTCTTGGAAGCTAGATA**

Read 2 _____

Read 3 _____

Read 1 _____

Read 2 _____

Read 3 _____

Read 4 _____



- Die DNA Stücke überlappen sich
 - Entstanden durch „zufällige“ Schnitte durch Chromosom
 - Das ist gut – Redundanz, höhere Gewissheit, Überlappung
 - Das ist schlecht - potentielle Konflikte
- Geforderte Überlappung im HGP: 7-10 Reads pro Base
- Problem: Berechne die originale Anordnung der Teilstücke

Ideen?

attggtatggctcca

agcttgtgtaagcacca

ttgaagtggatatattggtatggctccaagtca

ttaagcggctggctcca

cggtcgatga

- Was ist die originale Sequenz bei
 - Sequenz 1: cggtcgatgatgatgatgat
 - Sequenz 2: gatgatgatgatcggcattatac
- Möglichkeiten
 - cggtcgatgatgatgatgatgatgatgatcggcattatac
 - cggtcgatgatgatgatgatgatgatcggcattatac
 - cggtcgatgatgatgatgatgatgggtgtgagcagtgacgatgatgatcggcattatac
- Präzisierung der Fragestellung notwendig
 - Finde **minimalen String T** so, dass T alle Einzelsequenzen enthält

Abstrakte Formulierung

- SUPERSTRING

- Geg.: Menge S von Strings (Reads)
- Ges.: String T (Originalsequenz) so, dass
 - (a) $\forall s \in S: s \in T$ (s Substring von T)
 - (b) $\forall T'$, für die (a) gilt, gilt: $|T| \leq |T'|$ (T ist minimal)

- NP-vollständiges Problem

- Verschärfungen:

- Fehler in Sequenzen (s „ungefähr Substring“ von T)
- Zwei Orientierungen von s möglich
- Gaps

- Heuristische Verfahren

-
- Was ist nun Bioinformatik?

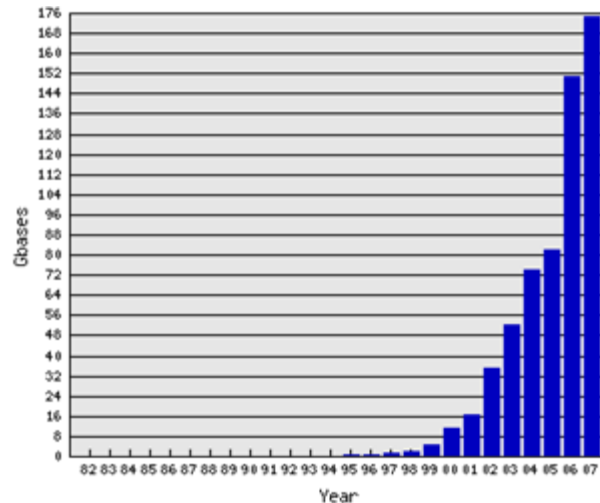
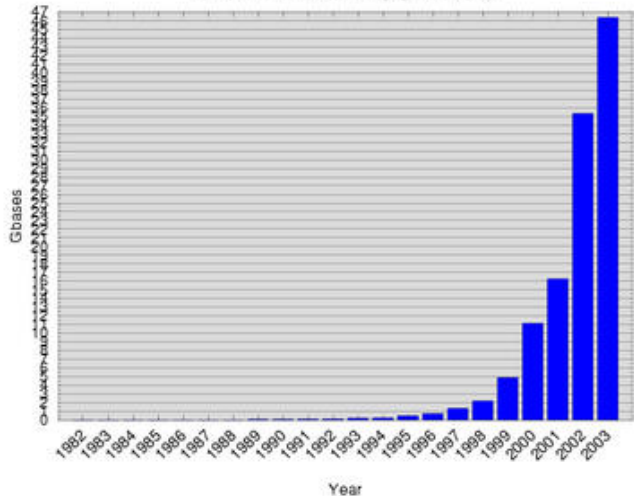
Bioinformatik

- Bioinformatics, Computational Biology
 - Verfahren der Informatik zur Lösung biologisch relevanter Probleme
 - Verfahren der Informatik zum Speichern und Suchen in biologischen Daten
 - **Algorithmisch** geprägt: kombinatorische oder probabilistische Algorithmen
 - Probleme sind typischerweise sehr schwer und viele Fehler in den Daten vorhanden – **Heuristiken** und Näherungsverfahren
 - Bioinformatik stellt Algorithmen zur Verfügung – die **Interpretation der Ergebnisse** ist ein anderes Thema
 - Bioinformatik ist **reduktionistisch** – Strings, Graphen, ...
 - Interdisziplinär – Informatik, Molekularbiologie, Mathematik, Physik, Chemie, ...

Geschichte der Bioinformatik

- Erste Proteinsequenzen seit 1951 bekannt
- Doppelhelix der DNA: 1953 (Watson, Crick)
- Entwicklung DNA Sequenzierung: Sanger, 1972
- DNA **Sequenzdatenbanken** seit Ende der 70er mit exponentiellem Wachstum

EMBL Database Growth
total nucleotides (gigabases)



Quelle: EMBL, Genome Monitoring Tables

Geschichte der Bioinformatik

- Klassische Papiere behandeln [Sequenzvergleich](#)
 - Needleman-Wunsch 1970, Gibbs 1970, Smith-Waterman 1981, BLAST `90
- Boom begann mit [Human Genome Projekt](#) (~1990)
- Erste „Bioinformatik“ Veröffentlichungen Ende der 80er Jahre
 - 14 Erwähnungen von „Bioinformatics“ vor 1995 in Medline
 - „Computational biology“ erstmals 1988
- „Journal of Computational Biology“ seit 1994
 - Vorher Veröffentlichungen in der Mathematik, Informatik, Molekularbiologie, Biochemie, Biophysik
- Erste [Professuren für Bioinformatik](#) in Deutschland Ende der 90er Jahre
- Erstes deutsches Bioinformatik Buch 2001 (!)
- Boom an Universitäten seit ca. 2000
- „Boom“ der Biotechnologieunternehmen 2000 – 2002
- [Wir beobachten das Entstehen einer neuen Wissenschaft](#)

-
- Themen der Vorlesung

Beispiel

- Ziel: Statt

T ctgagatcgcgta
P gagatc
gagatc
gagatc
gagatc
gagatc
gatatc
gatatc
gatatc

- ... lieber ...

bbcggbcbagggbbaacabaabgbaac
c**ab**a**ab**g**baa**

bbcggbcbagggbbaacabaabgbaac
c**a**b**a**a**ab**g**ba**a

bbcggbcbagggbbaacabaabgbaac
c**a**b**a**a**ab**g**baa**

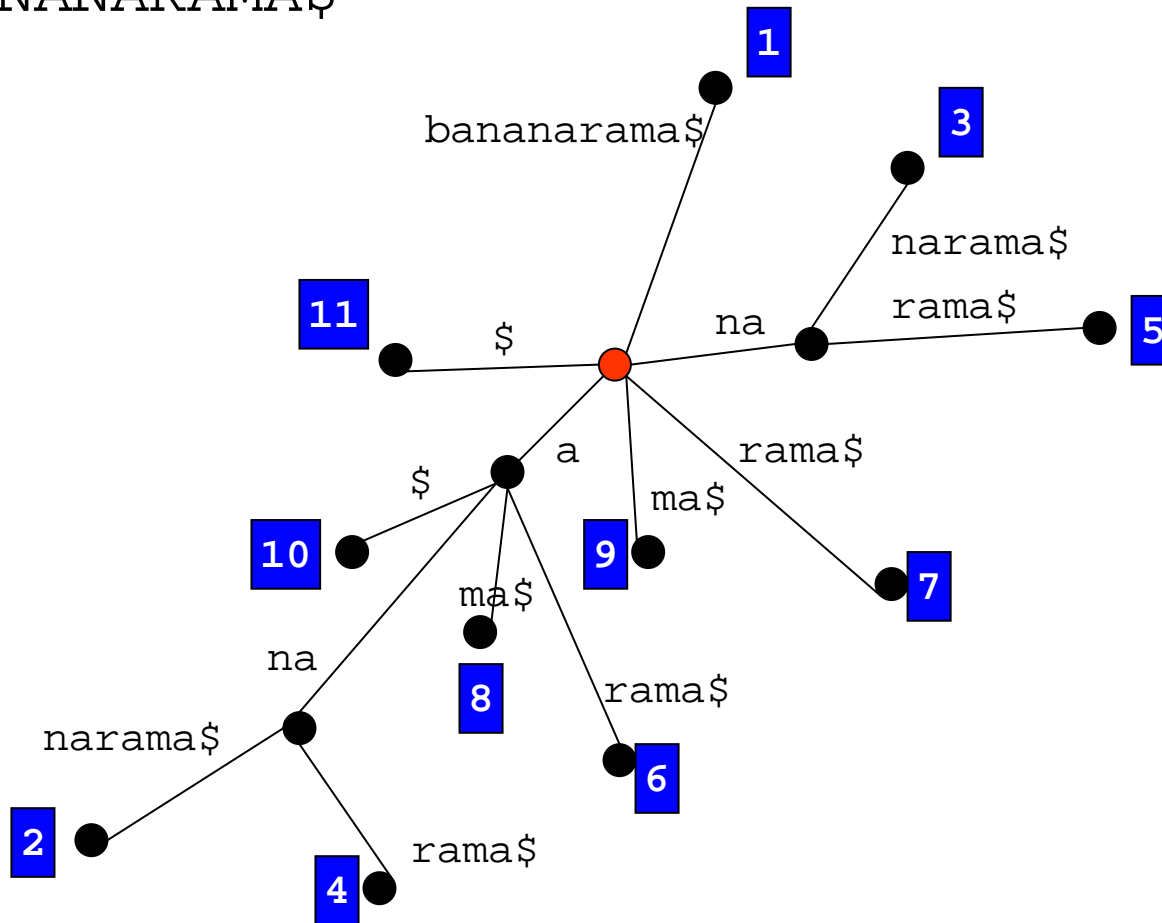
2. Suffixbäume und -arrays

- 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
- Verschiedene Erweiterungen
 - [Suffixarrays](#)
 - Suffixbäume auf Sekundärspeichern

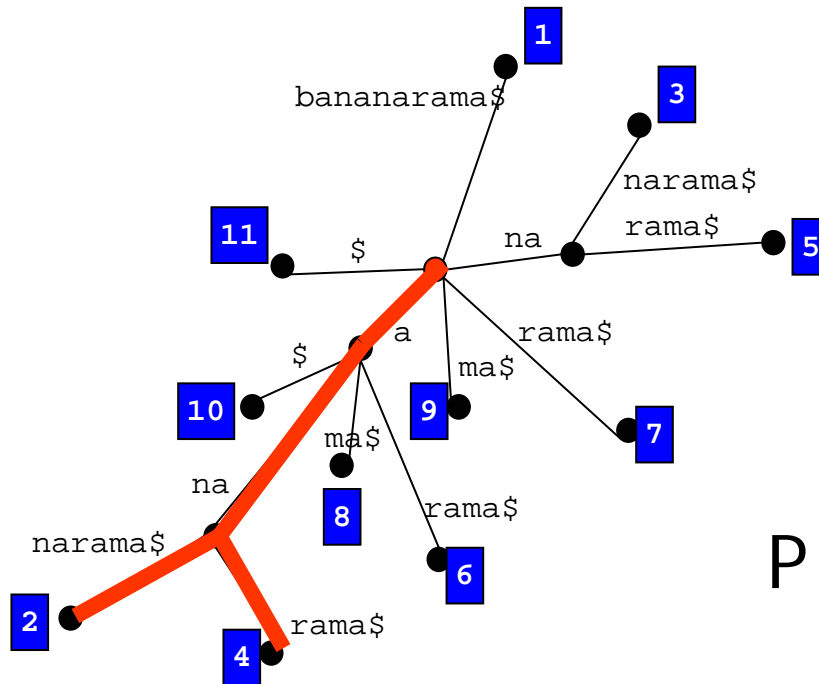
Beispiel

12345678901

- S= BANANARAMA\$



Suche im Suffixbaum



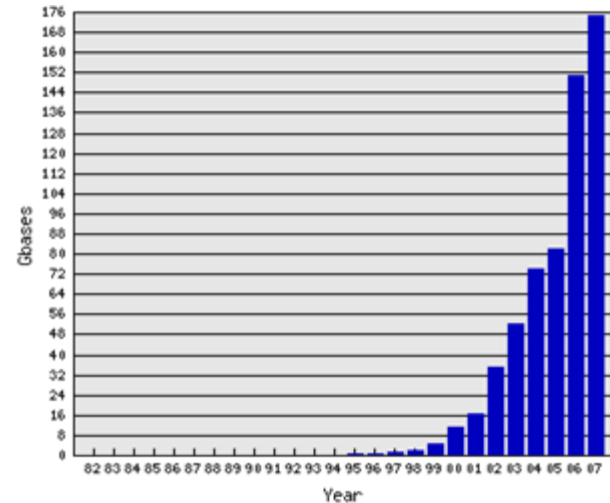
P = „ana“

3. Approximatives Stringmatching

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

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

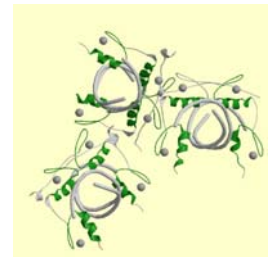


4. Multiples Alignment

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

```
YVCK...LCN...FAFKTKGNLTKHMRSK..AH
YRCP...R.ENC...RTYTTKFNLSHILT...FH
FRCEY.KCG...RLYTTAHLKVHERA...H
YRCE...KCG...KMYKTERCLKVHNLV...H
FSCS...QCD...ESFVORSELELHRQL...H
FPCE...QCD...EKFKTEKQLERHVKT...H
FQCN...QCG...ASFQKGNLLRHIKL...H
FKCH...LCY...RCFQQTNLDRHLK...H
FRCK...RCR...TRFRQOSELKHKMKT...H
FECN...VCG...SAFRLQLYLSEHQKT...H
MSCKV...CD...RVFYRLDNLRSHLKQ...H
FSCQ...HCH...RAFADRSNLRHLQT...H
FRCG...YCG...RAFIVKDYLNKHLTT...H
HVCWV.PGCH...RAFSRSDNLSNAHYTK...TH
LTCAH...CD...WSFDNVMKLVRRHG...H
```

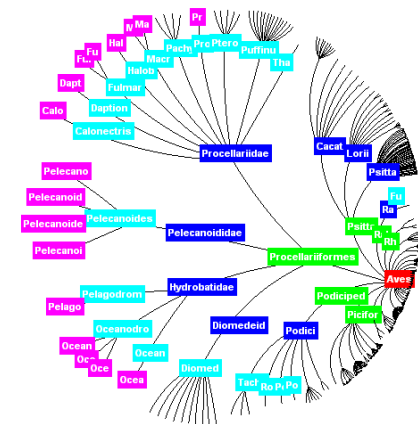
Quelle: Pfam, Zinc finger domain



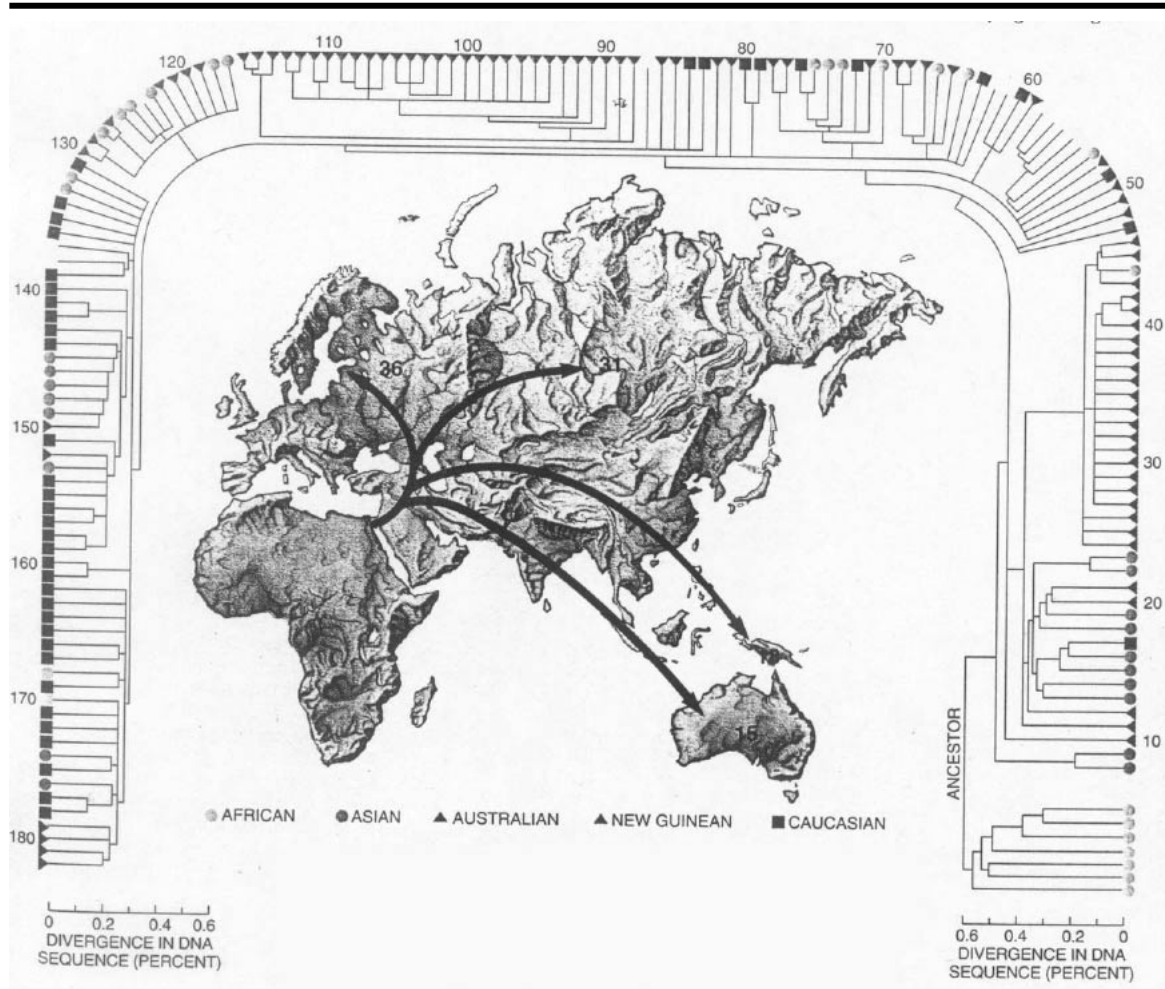
- Hauptziel von MSAs: Finde das „Gemeinsame“ der Sequenzen
 - Funktionen werden oft von sehr kurzen Sequenzstücken bestimmt
 - Welcher Teil eines Proteins bestimmt die Funktion?
 - Wie kann man Proteine in Familien anordnen?
- „Gute MSAs“ sind nicht klar definiert
 - Konkretes Maß zur Güte hängt von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

5. 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 Taxonomieproblems auf molekularer Basis



Beispiel



- Abstammung zwischen Menschen/Gruppen/Völkern
- Basierend auf mitochondrialer DNA
 - Nur mütterlich vererbt
 - Keine Rekombination
- Untersucht wurden 182 Sequenzen
- Evidenz für gemeinsamen Ursprung aller Menschen in Ostafrika: **Eva**

Quelle: A. Wilson and R. Cann. The recent african genesis of humans. *Scientific American*, April, 1992.