

Bioinformatik

Einleitung
Überblick



Ulf Leser
Wissensmanagement in der
Bioinformatik



H5N1

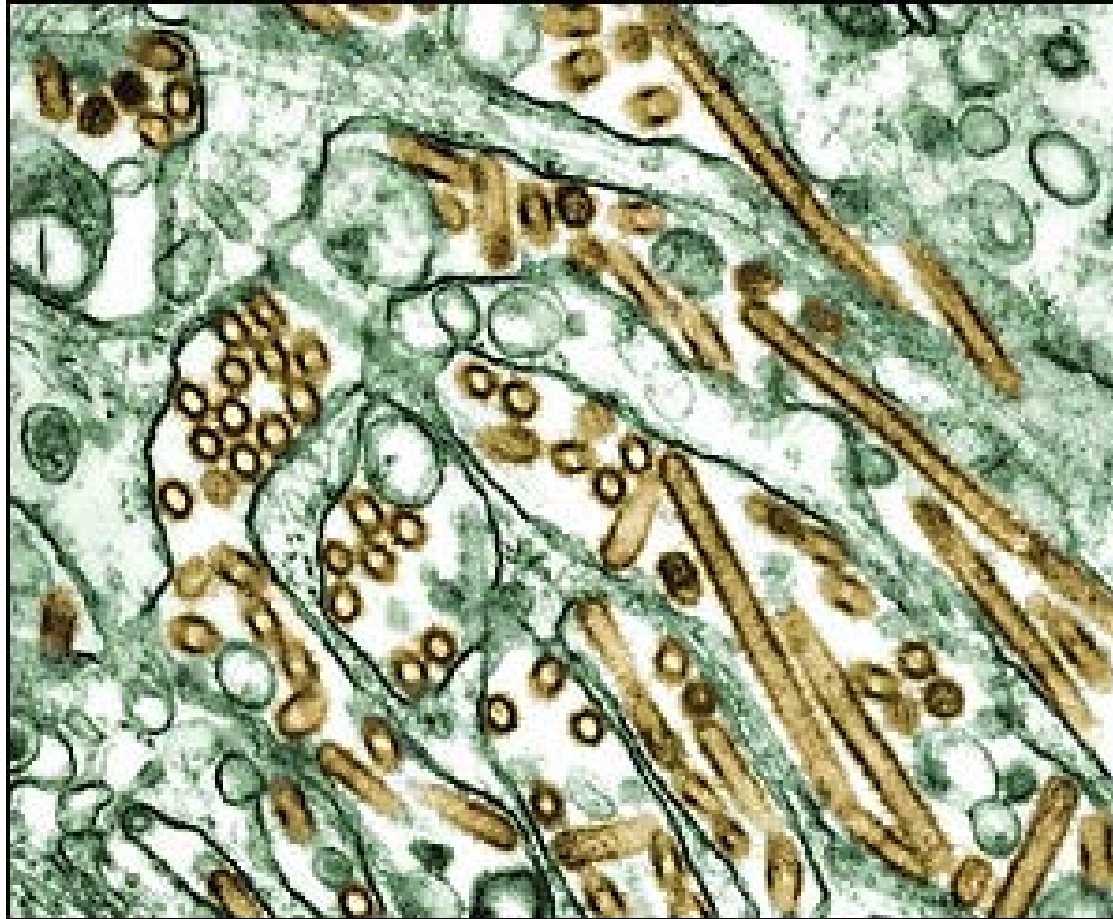
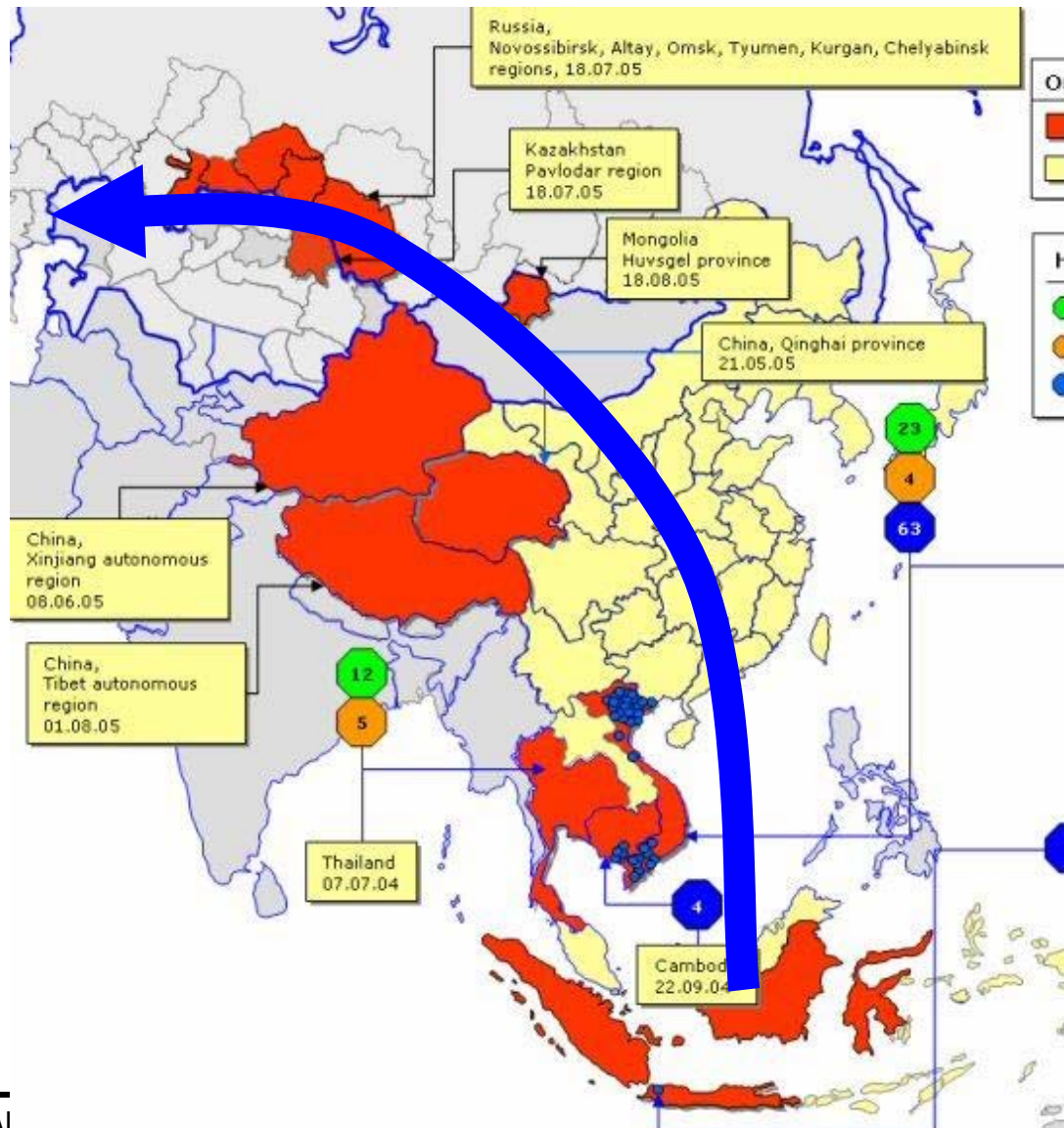
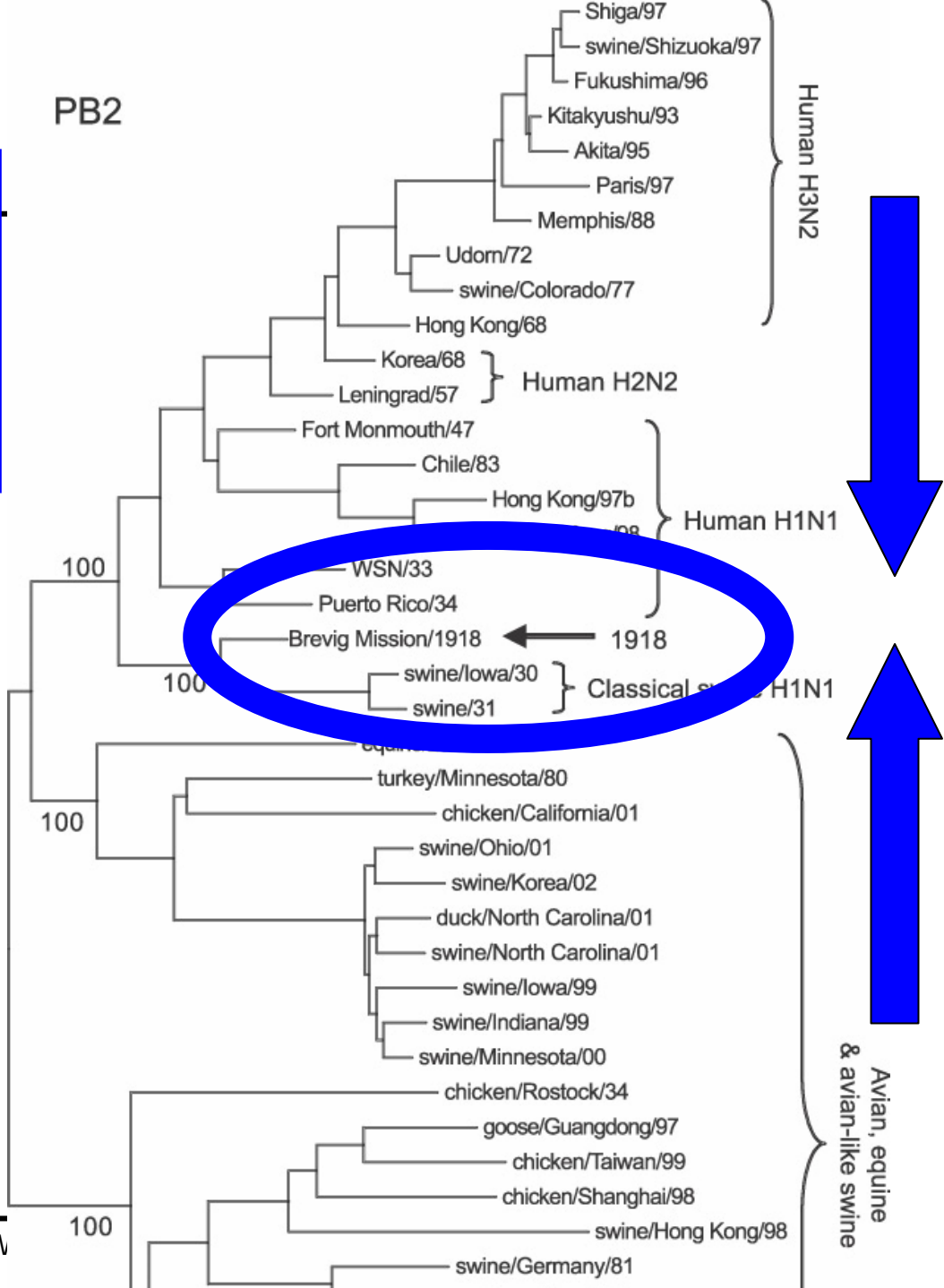


Foto: Centers for Disease Control

Migration

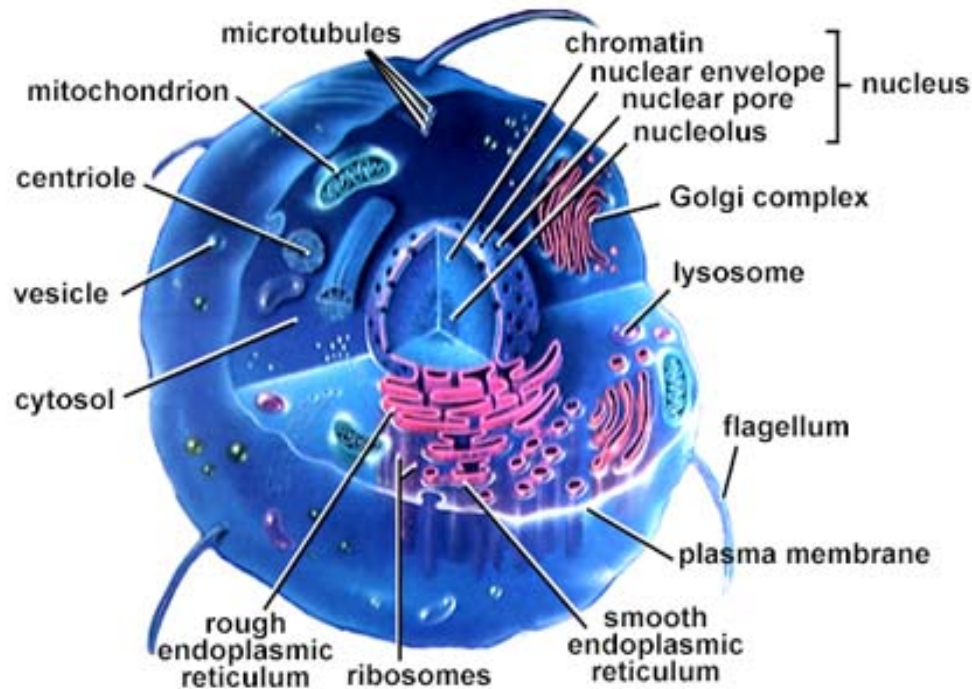


Vogelgrippe
beim Menschen?



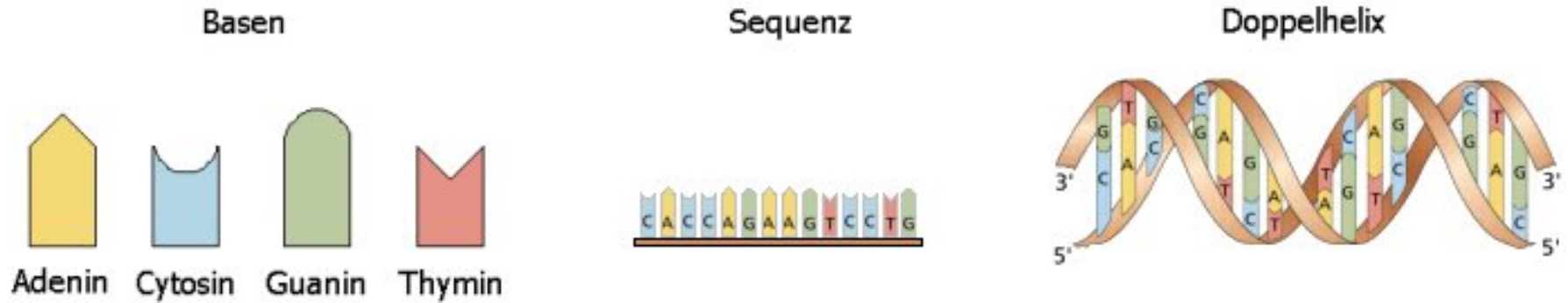
Nature. 2005 Oct 6;437(7060):889-93.

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“

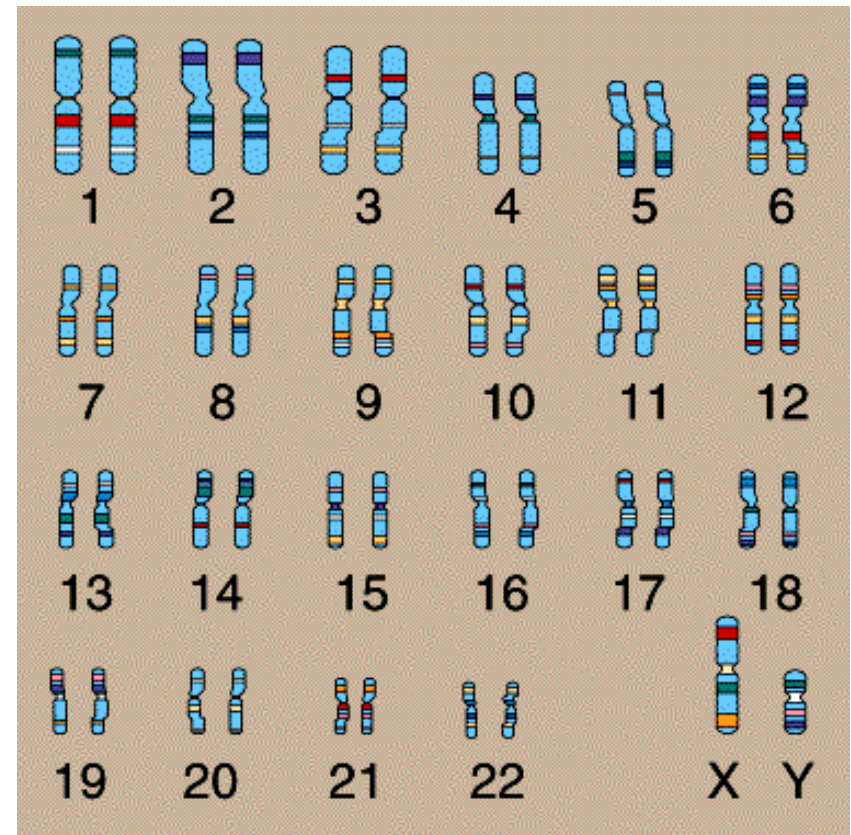
~~• ~ 100.000 Gene~~

~~• ~ 56.000 Gene~~

~~• ~ 30.000 Gene~~

~~• ~ 24.000 Gene~~

• ~ 20.000 Gene



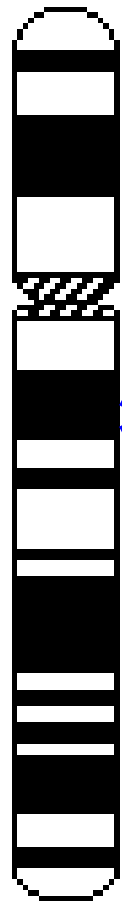
Was ist ein Gen ?

Chromosom

DNA

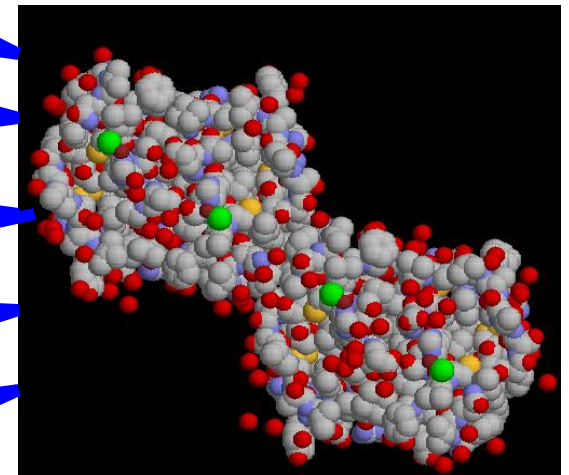
RNA

Protein

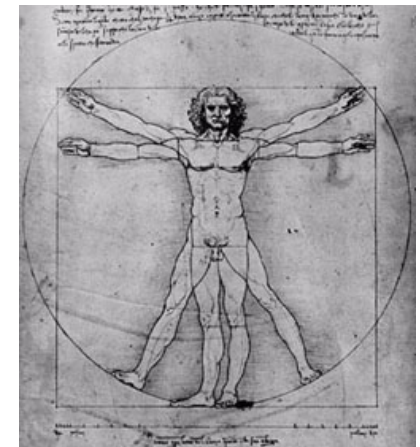
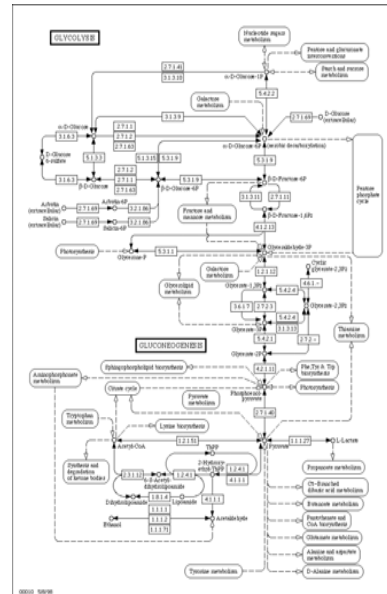
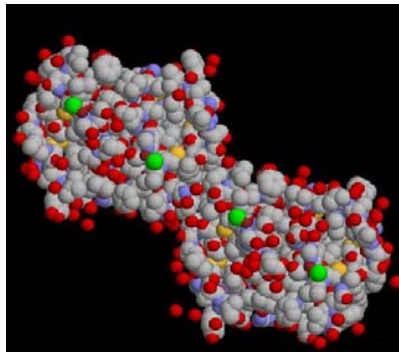
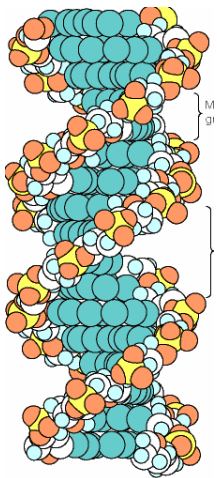
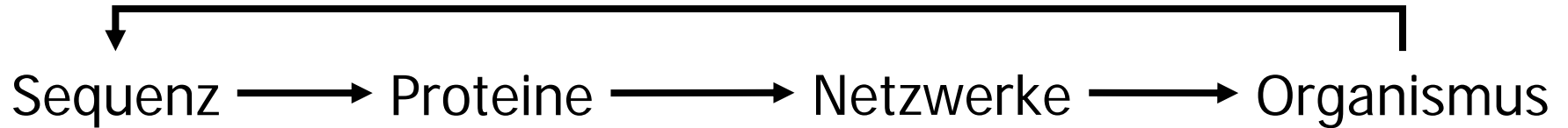


A
C
G
T
T
G
A
T
G
A
C
C
A
G
A
G
C
T
T
G
T

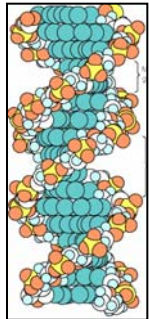
A
C
G
T
T
G
A
C
A
G
A
G
C
T
T
G
T



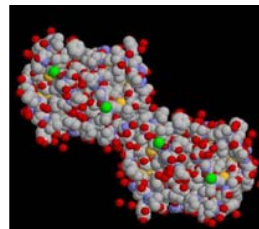
Der Weg des Lebens



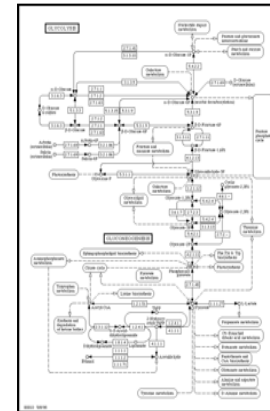
Welche Rolle spielt dabei die Informatik ?



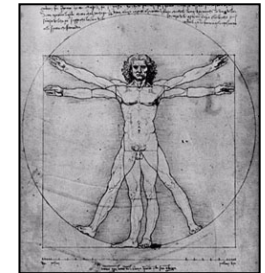
Sequenzierung
Erkennung von
Genen
Verwandtschaft
zw. Spezies
Regulation &
Expression
RNA Gene –
„Dark matter“



Dreidimensionale
Faltung
Strukturvergleich
und -ähnlichkeit
Interaktion
Sekundärstruktur
Proteinidenti-
fikation

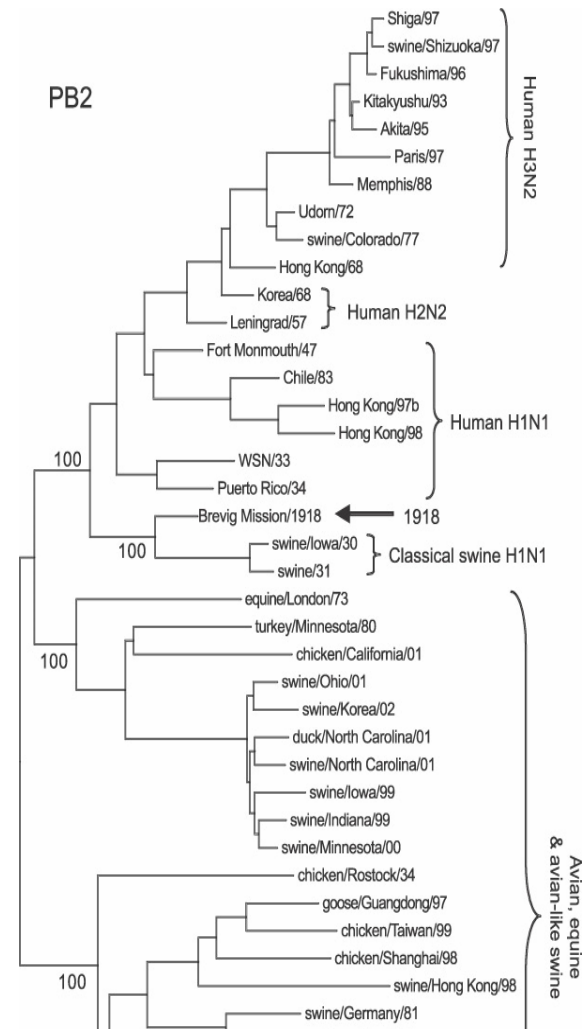
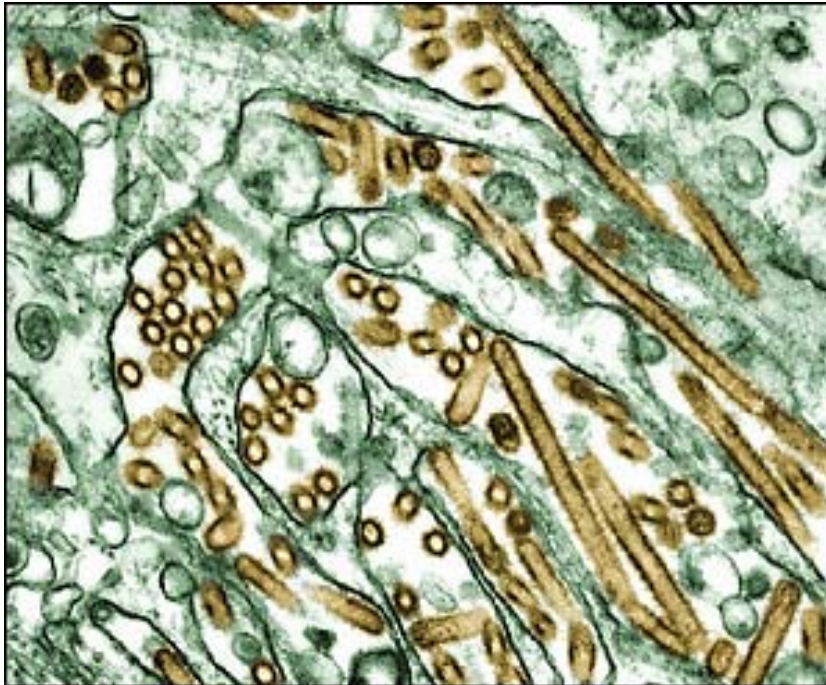


Netzwerkanalyse
Geschwindigkeit
von
Reaktionsketten
Stoffumsatz
Kompartamente
Muster und
Redundanz



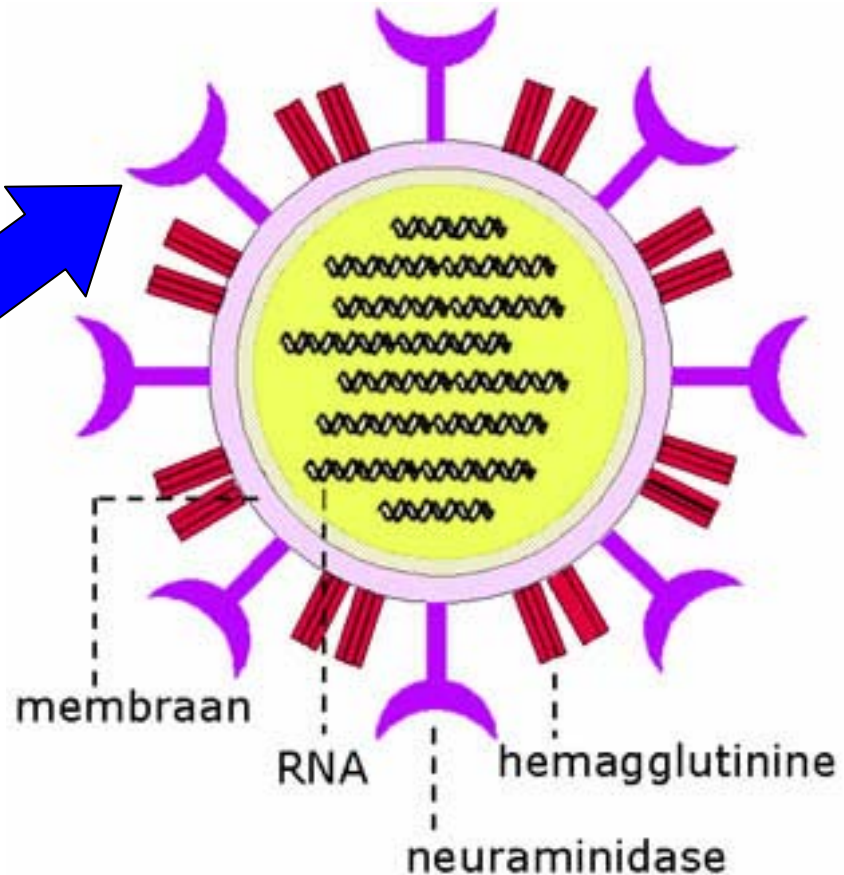
Korrelation
Phänotyp/Genotyp
Arzneimittel-
empfänglichkeit
Studienstatistik
und -verwaltung

Wo kommt der Stammbaum her?



Grundlage für Stammbaumberechnung

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

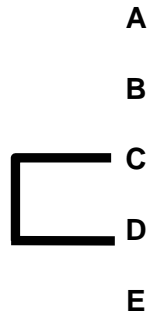


Abstandsbasierte Phylogenie

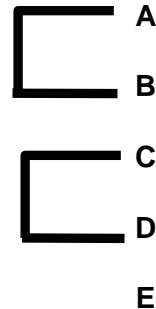
- Gegeben k Sequenzen
- Berechne die Ähnlichkeitsmatrix M
 - $O(k^2)$ paarweisen Alignment scores
- Hierarchisches Clustering
 - Wähle Zelle (i,j) mit kleinstem Abstand aus Matrix M
 - Das ist das erste Paar
 - Erzeuge M' : Lösche die Sequenzen i und j aus M und füge neue Spalte/Zeile (ij) ein
 - Für 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

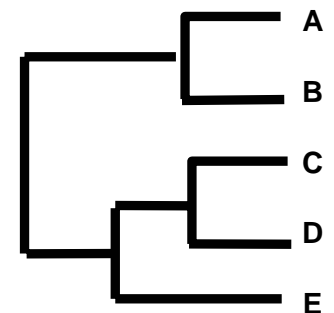
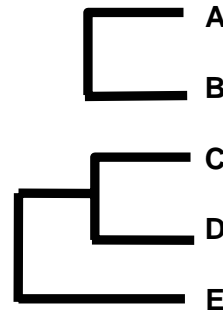
	A	B	C	D	E
A		17	59	59	77
B			37	61	53
C				13	41
D					21



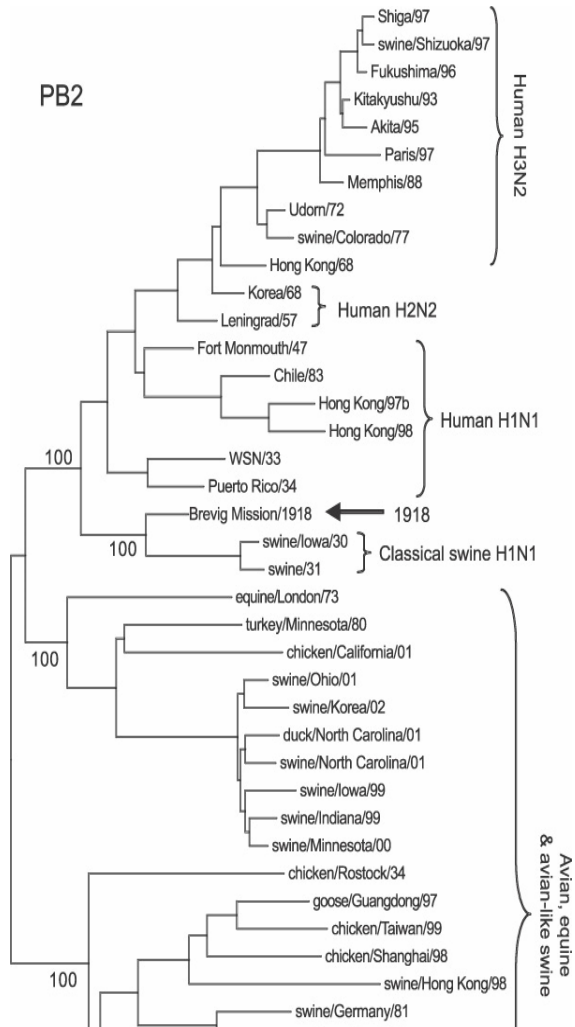
	A	B	E	CD
A		17	77	59
B			53	49
E				31



	E	CD	AB
E		31	65
CD			54



Was bedeutet also 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
- Also wahrscheinlich auch ähnliche Wirkung
- Zu benachbarten Knoten sind nur noch wenige Mutationen notwendig

Beispiel 2: Sequenzbestimmung

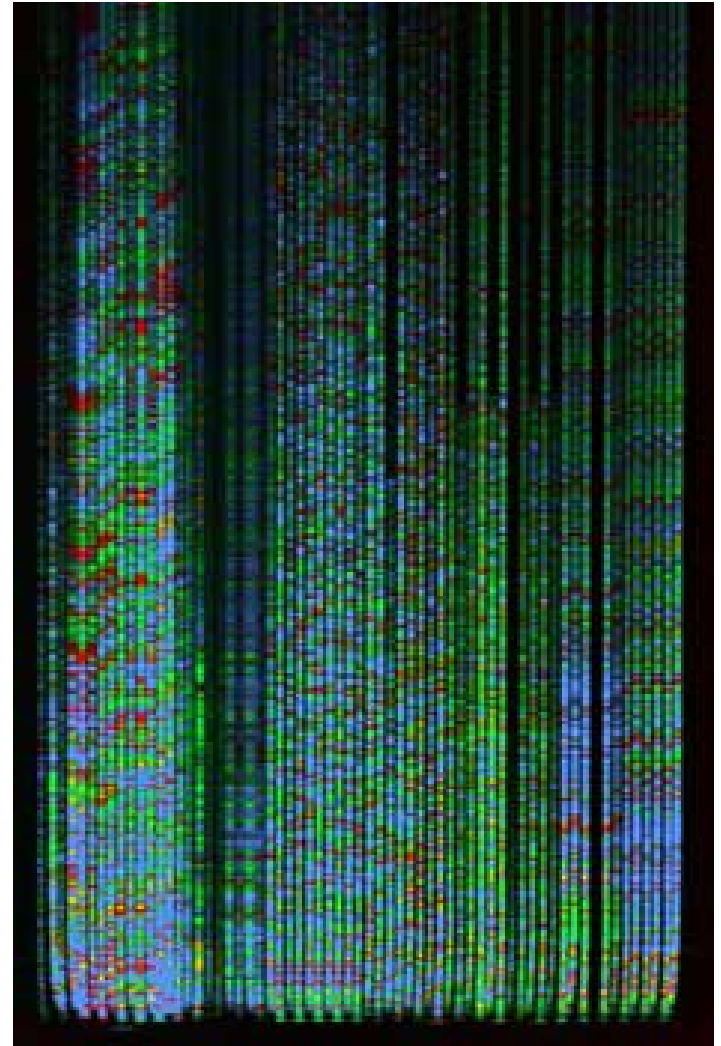
- Genom wird nicht entschlüsselt, sondern „gemessen“
- Direkt messen kann man nur sehr kurze Stücke
- Also wird ein Chromosom erst in viele Einzelteile zerlegt
- Diese werden einzeln gemessen
- Problem: Daraus die Sequenz des Chromosom berechnen



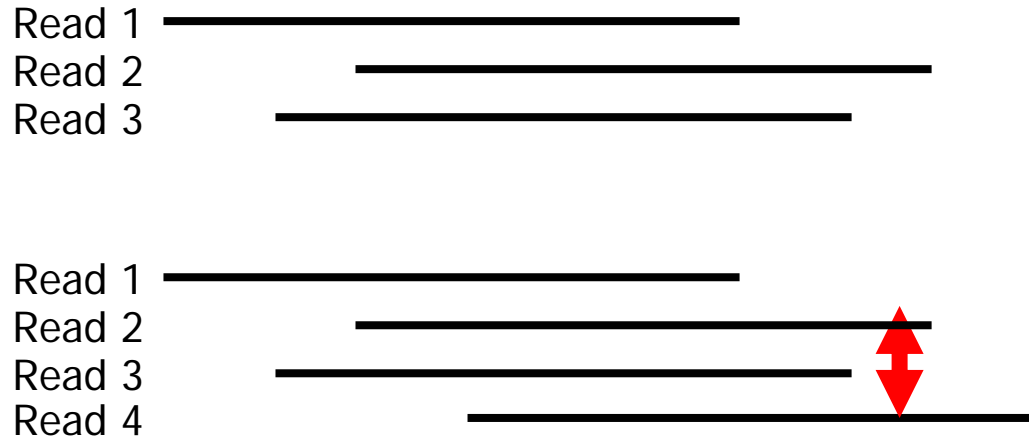
Quelle: <http://www.geneticsplace.com>

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



- Die DNA Stücke überlappen sich sehr stark
 - Entstanden durch „zufällige“ Schnitte durch Chromosom
 - Das ist gut – Redundanz, höhere Gewissheit, Überlappung
 - Das ist schlecht - potentielle Konflikte
- **Typische geforderte Überlappung im HGP: 7-10 Reads**
- 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 notwendig
 - Finde minimalen String T so, dass T alle Einzelsequenzen enthält

Abstrakte Formulierung

- SUPERSTRING

- Geg.: Menge S von Strings

- Ges.: String T 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:

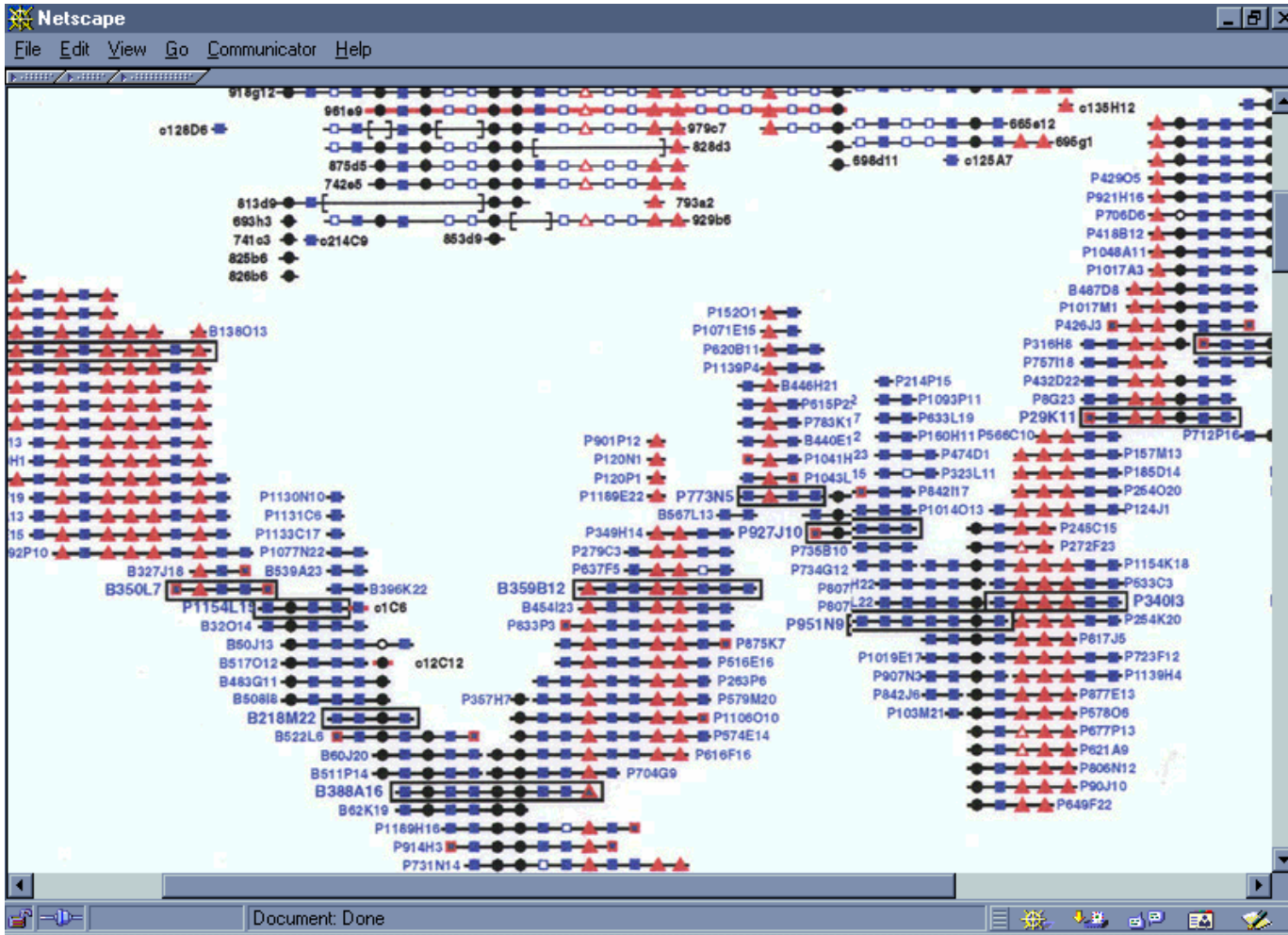
- Fehler in Sequenzen (s „ungefähr Substring“ von T)

- Zwei Orientierungen von s möglich

- Gaps

- Heuristische Verfahren

Resultat



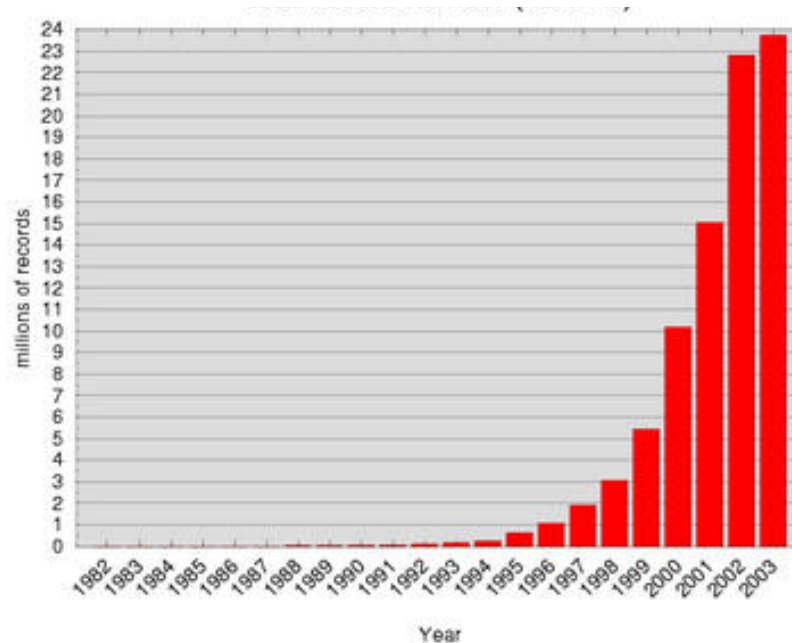
-
- Was ist nun Bioinformatik?

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 typischerweise sehr schwer – **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 stetigem Wachstum



Quelle: EMBL, Genome Monitoring Tables, Stand 12.2.2003

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 ...

bbcggbcbaggbbaacabaaabgbaac
c**ab**a**ab**g**baa**

bbcggbcbaggbbaacabaaabgbaac
c**a**b**a**a**ab**g**ba**a

bbcggbcbaggbbaacabaaabgbaac
c**a**b**aa**b**g****baa**

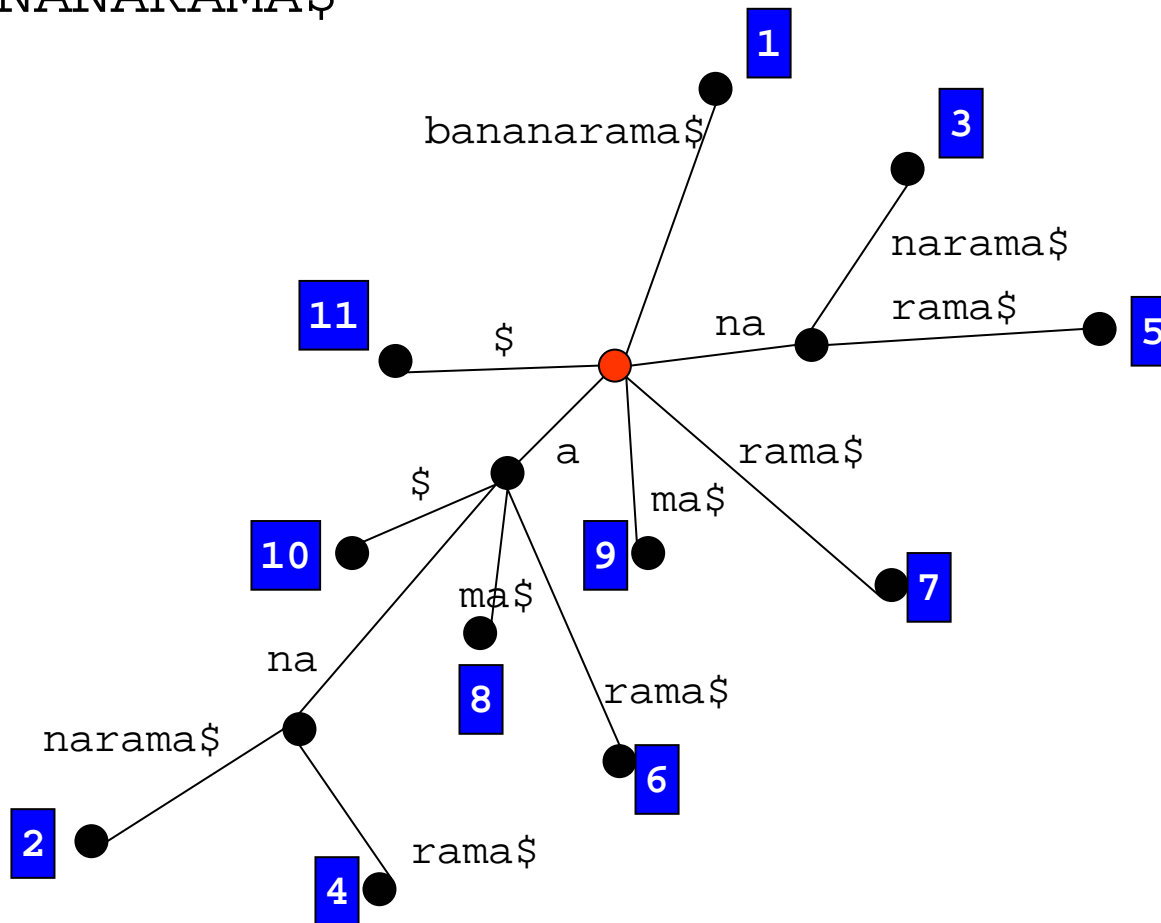
2. Suffixbäume und -arrays

- 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

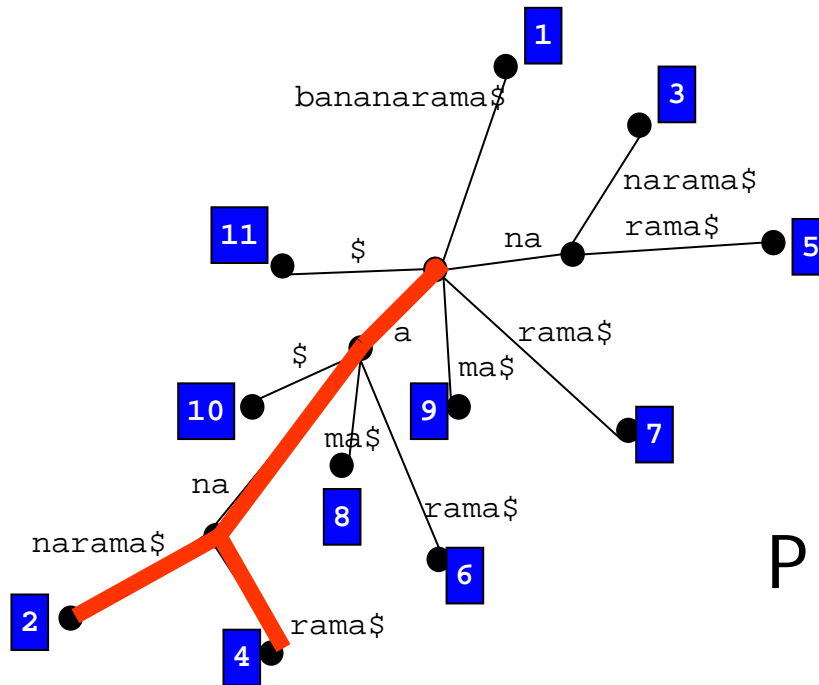
Beispiel

12345678901

- S= BANANARAMA\$



Suche im Suffixbaum



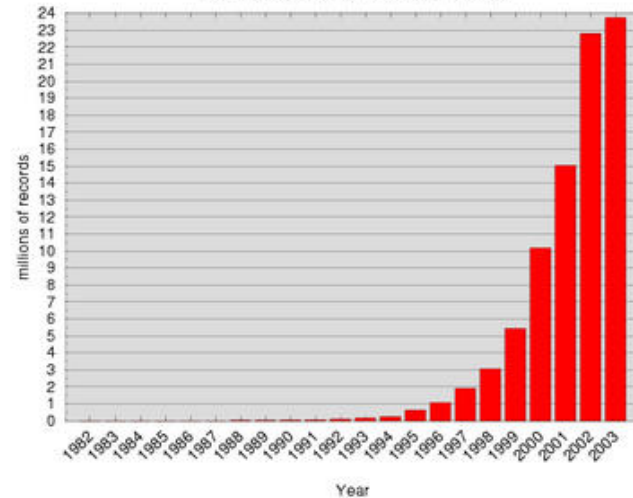
P = „ana“

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**



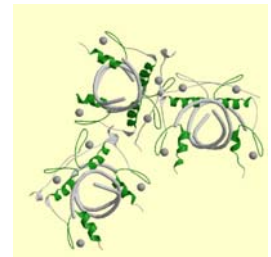
- Also: Heuristiken, z.B. BLAST
 - Suche nach „Seeds“ mit exakten Matches
 - Verlängerung und Zusammenfügung der Seeds zu Matches
 - Findet nicht alle Hits, aber die meisten „interessanten“
 - „BLAST“ ist fast **Synonym für Bioinformatik** geworden

4. Multiples Alignment

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

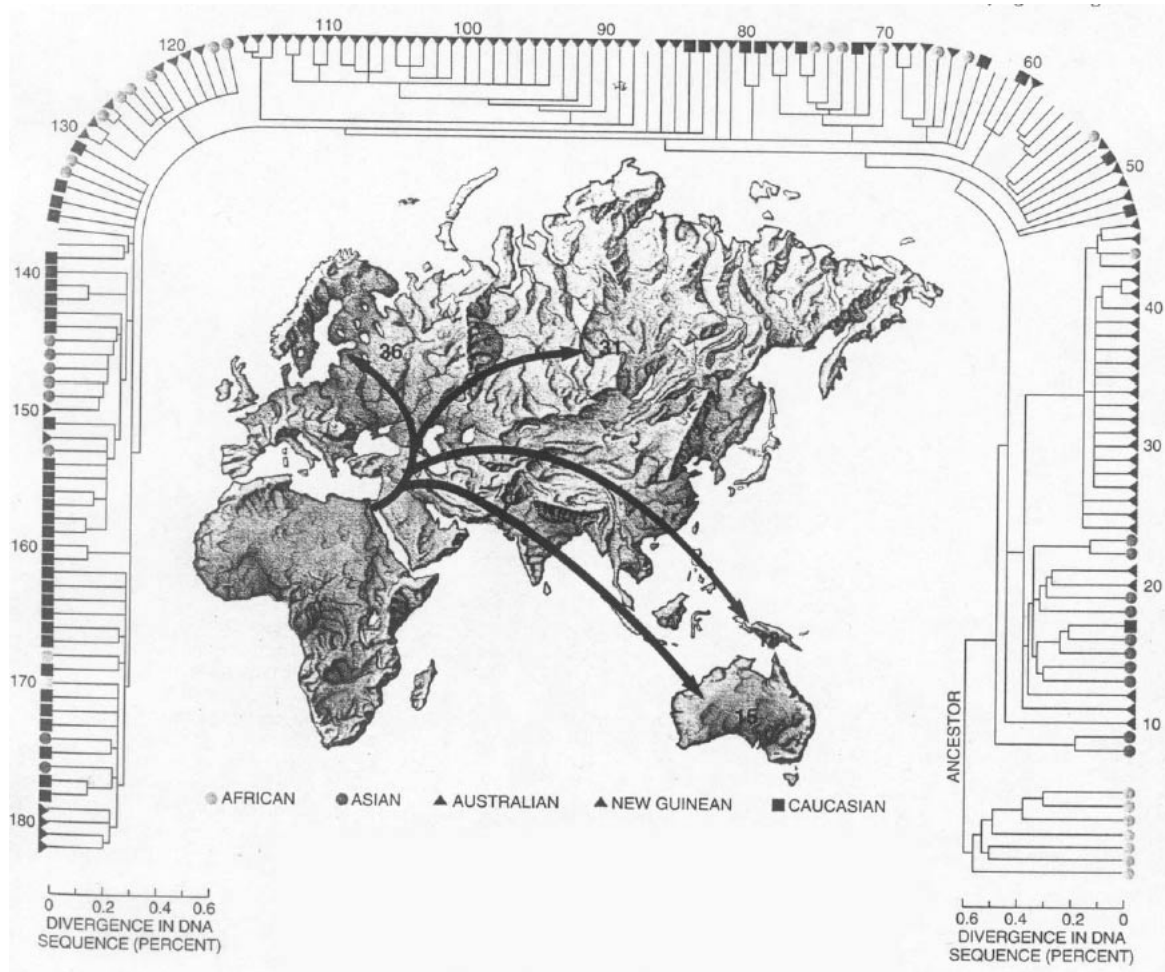
```
YVCK...LCN...FAFKTKGNLTKHMRSK..AH
YRCP...R.ENC...RTYTTKFNLSHILT..FH
FRCGY.KCGG...RLYTTAHLKVHERA...H
YRCE...KCG...KMYKTERCLKVHNLV...H
FSCS...QCD...ESFVORSELELHRQL...H
FPCE...QCD...EKFKTEKQLERHVKT...H
FQCN...QCG...ASFQKGNLLRHIKL...H
FKCH...LCY...RCFGQQTNLDRHLK...H
FRCK...RCR...TRFRQOSELKHKMKT...H
FECN...VCG...SAFRLQLYLSEHQKT...H
MSCKV...CD...RVFYRLDNLRSHLKQ...H
FSCQ...HCH...RAFADRSNLRHLQT...H
FRCG...YCG...RAFVVKDYLNKHLTT...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

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.