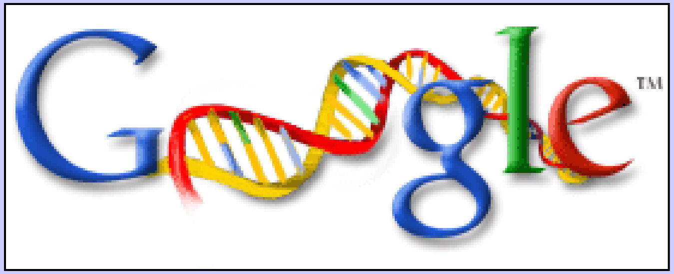
HUMBOLDT-UNIVERSITÄT ZU BERLIN

Introduction to Bioinformatics

Ulf Leser

Bioinformatics



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

VERMISCHTES

Menschliches Erbgut entziffert

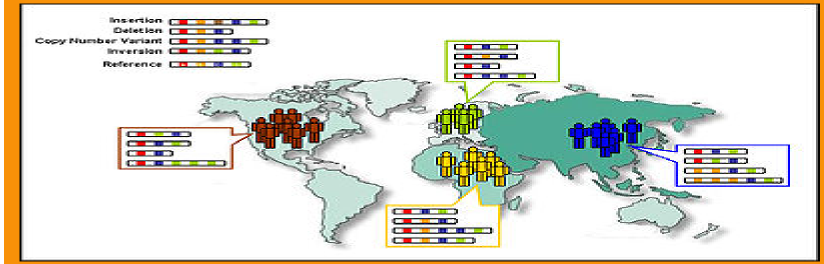
Ein halbes Jahrhundert nach der Entdeckung der DNA-Struktur ist das menschliche Erbgut entschlüsselt. Forscher und Politiker feierten am Montag den Erfolg als „Meilenstein" in der Geschichte der Menschheit.

14.4.2003
Humanes Genom zu 99% sequenziert mit 99.99% Genauigkeit

Published online 16 April 2008 | **452**, 788 (2008) | doi:10.1038/452788b

James Watson's genome sequenced at high speed

2008
Genom of J. Watson finished
4 Months, 1.5 Million USD



2010
1000 Genomes Project

Example: Int. Cancer Genome Cons.

- Large-scale, international endeavor
- Planned for 50 different cancer types
- Cancer types are assigned to co
- Distribute infrastruct
- First federate n to a large int. me project [HAA+08]

50 different cancer types, 500 samples per type, always control + cancer

> 50.000 genomes

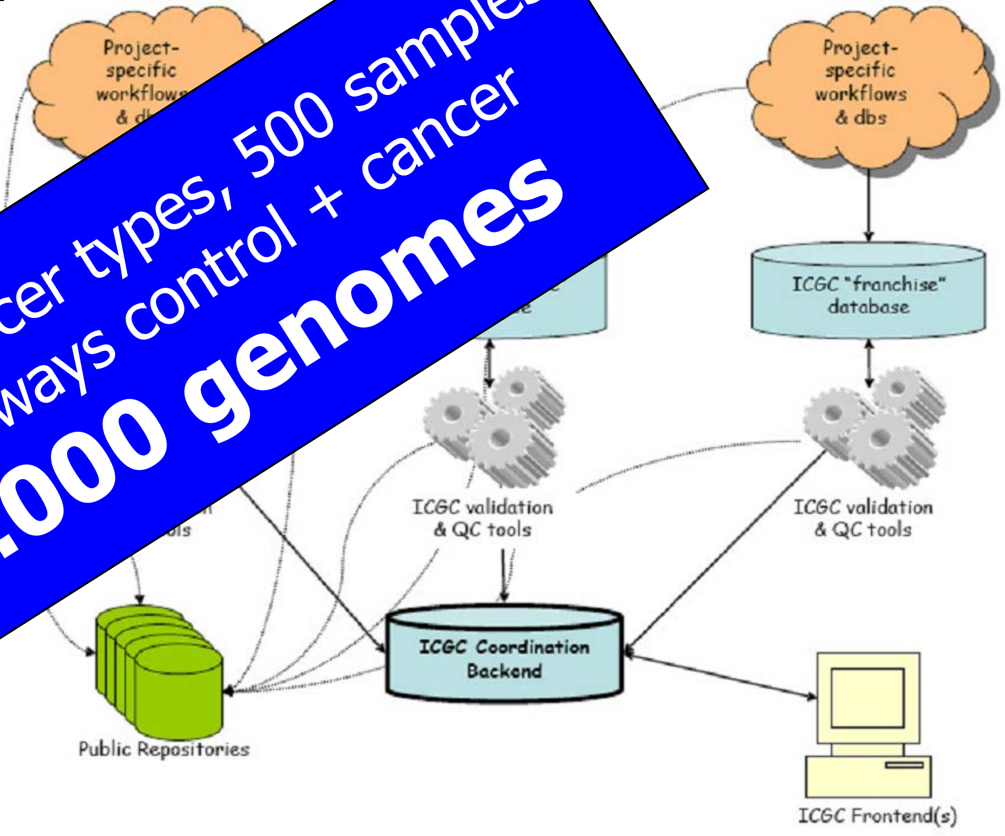
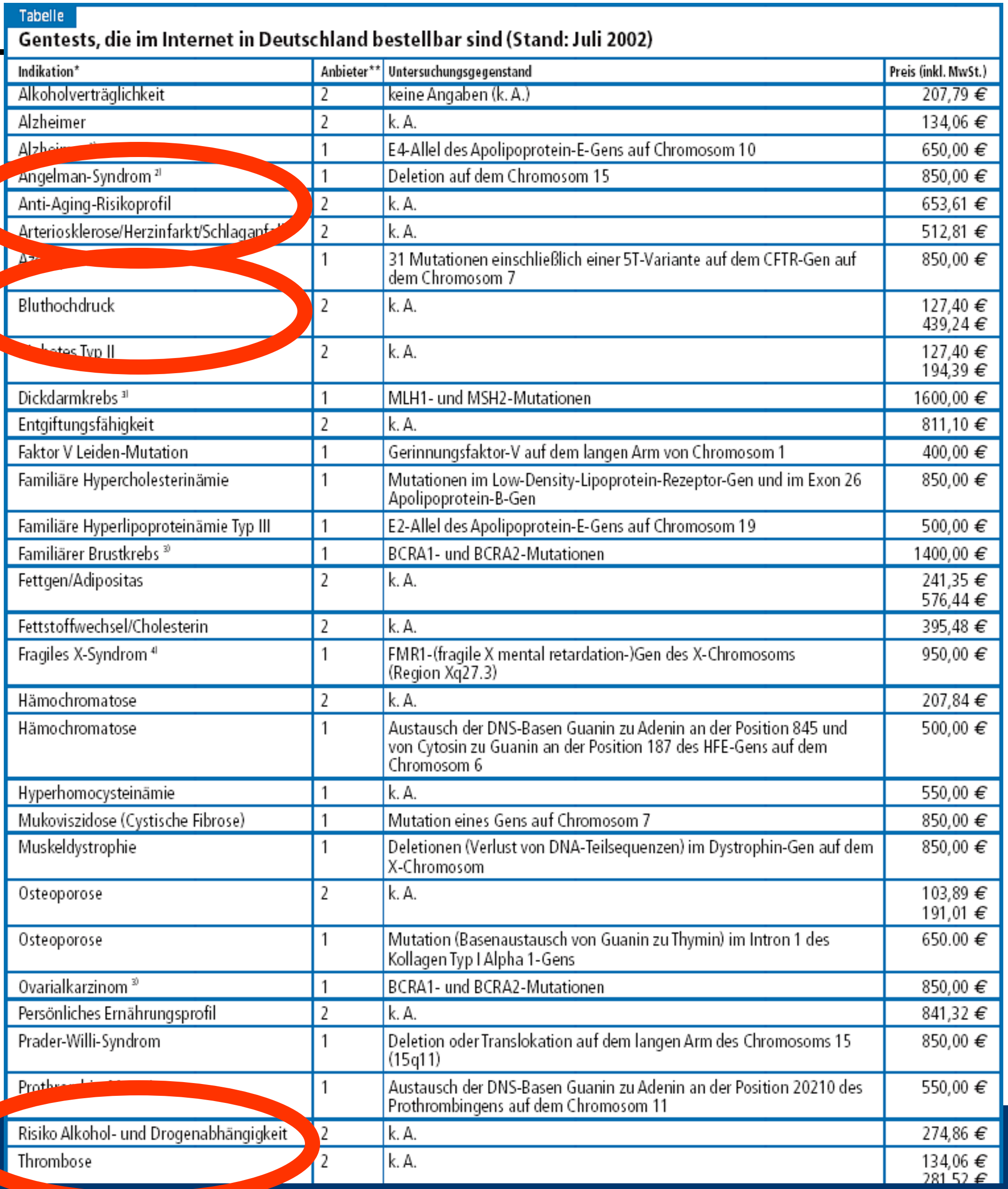


Figure 2: ICGC data coordination as a franchise system

Things you can do with it

- 2002
  - 2 companies
  - 32 Tests
  - Price: 100–1400€

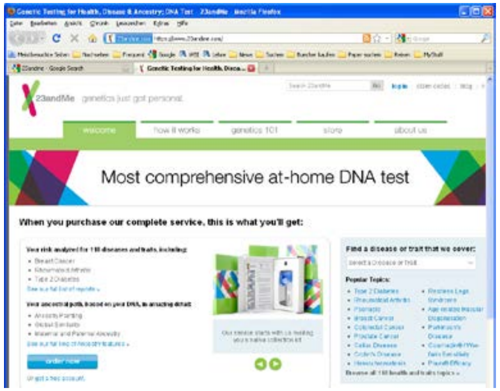


Tabelle

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 €
Alzhei[illegible]	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/Schlaganf[illegible]	2	k. A.	512,81 €
A[illegible]	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 €
[illegible]betes Typ II	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 €
Prothr[illegible]	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”



23andMe genetics just got personal.

welcome | how it works | genetics 101 | store | about us

Most comprehensive at-home DNA test

When you purchase our complete service, this is what you'll get:

Find a disease or trait that we cover:

Popular Topics:

- 6/2010: „Gentest-Firma vertauscht DNA-Ergebnisse ihrer Kunden“ (Nature Blog)
- 7/2010: US general accounting office compared 15 (4) companies: totally contradicting results
- 2013: FDA closes main business line of 23andme
  - “...as 23andMe had not demonstrated that they have "analytically or clinically validated the PGS for its intended uses" and the "FDA is concerned about the public health consequences of inaccurate results from the PGS device"

This Lecture

- Formal stuff
- A very short introduction in Molecular Biology
- What is Bioinformatics?
  - And an example
- Topics of this course

This course

- Is mandatory for students of Biophysics Bachelor
- Is open for Bachelor students in computer science
- Brings 5 SP and will be held as 2+2
- Does assume basic knowledge in computer science
  - Will not teach programming – you need to know it already
- Does not assume knowledge in biology
- Is introductory – many topics, often not much depth
  - Visit “Algorithmische Bioinformatik” afterwards ...
- Ask questions! leser (a) informatik.hu ... berlin...

Exercises

- Taught by Yvonne Mayer / Lichtblau
- Registration through Goya
- There will be 5 assignments
- We build teams
- No grades
- System
  - First week: 2-3 presentations of results of previous assignment and discussion of new assignment
  - Next week: Questions
  - ...
- You need to pass all but one assignment to be admitted to the exam

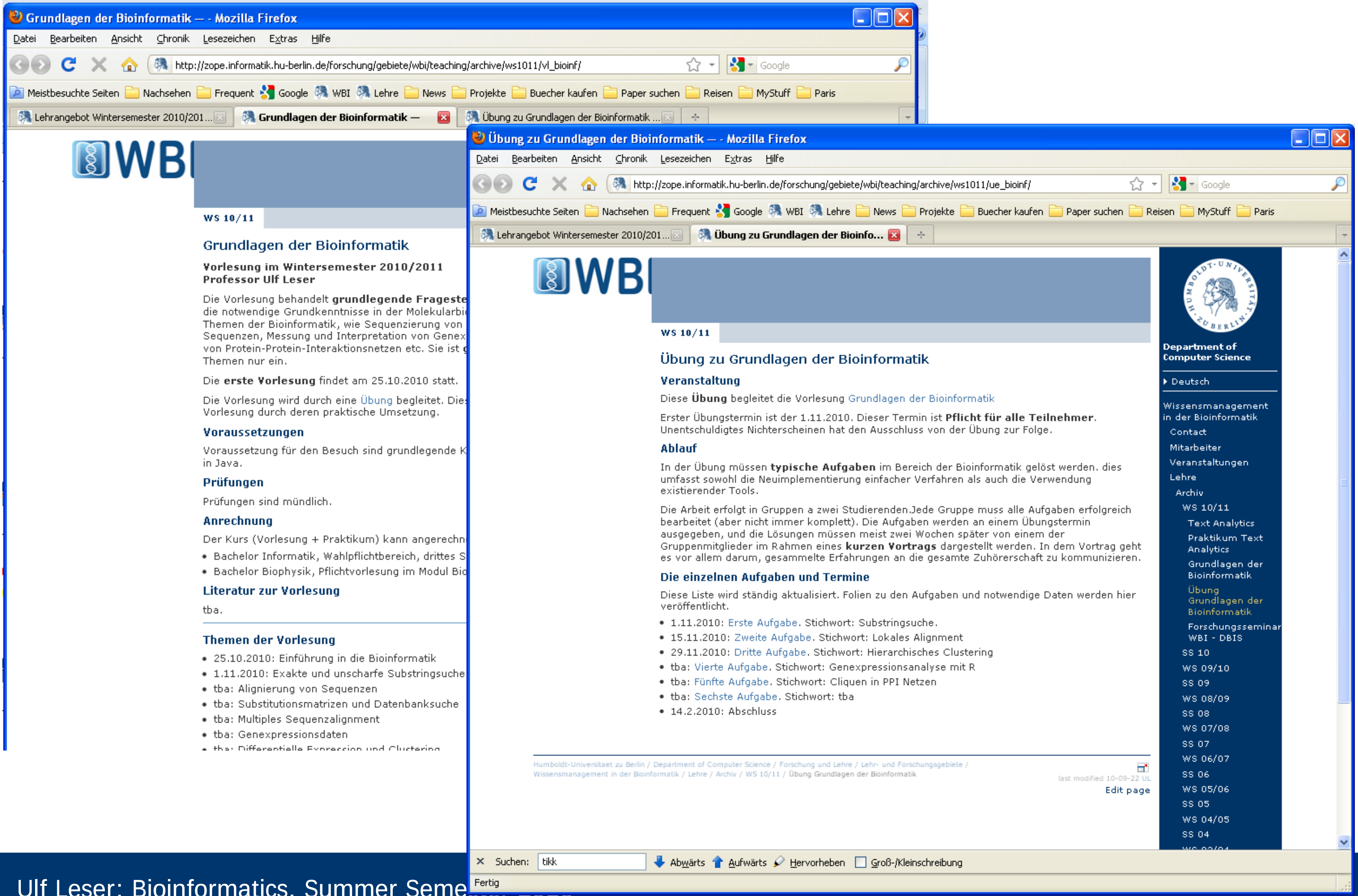
Exams

- Written examination
- Friday, 29.7.2016, 11-14 o’clock, room 3.001

Literature

- For algorithms
  - Gusfield (1997). „Algorithms on Strings, Trees, and Sequences“, Cambridge University Press
  - Böckenhauer, Bongartz (2003). „Algorithmische Grundlagen der Bioinformatik“, Teubner
- For other topics
  - Lesk (2005). „Introduction to Bioinformatics“, Oxford Press
  - Cristianini, Hahn (2007). "Introduction to Computational Genomics - A Case Study Approach", Cambridge University Press
  - Merkl, Waack (2009). "Bioinformatik Interaktiv", Wiley-VCH Verlag.
- For finding motivation and relaxation
  - Gibson, Muse (2001). "A Primer of Genome Science", Sinauer Associates.
  - Krane, Raymer (2003). "Fundamental Concepts of Bioinformatics", Benjamine Cummings.
- These slides

Web Sides



WS 10/11

Grundlagen der Bioinformatik

Vorlesung im Wintersemester 2010/2011
Professor Ulf Leser

Die Vorlesung behandelt **grundlegende Frageste**
die notwendige Grundkenntnisse in der Molekularbi
Themen der Bioinformatik, wie Sequenzierung von
Sequenzen, Messung und Interpretation von Genex
von Protein-Protein-Interaktionsnetzen etc. Sie ist
Themen nur ein.

Die **erste Vorlesung** findet am 25.10.2010 statt.

Die Vorlesung wird durch eine Übung begleitet. Die
Vorlesung durch deren praktische Umsetzung.

Voraussetzungen

Voraussetzung für den Besuch sind grundlegende K
in Java.

Prüfungen

Prüfungen sind mündlich.

Anrechnung

Der Kurs (Vorlesung + Praktikum) kann angerechn

- Bachelor Informatik, Wahlpflichtbereich, drittes S
- Bachelor Biophysik, Pflichtvorlesung im Modul Bio

Literatur zur Vorlesung

tba.

Themen der Vorlesung

- 25.10.2010: Einführung in die Bioinformatik
- 1.11.2010: Exakte und unscharfe Substringsuche
- tba: Alignierung von Sequenzen
- tba: Substitutionsmatrizen und Datenbanksuche
- tba: Multiples Sequenzalignment
- tba: Genexpressionsdaten
- tba: Differentielle Expression und Clustering

WS 10/11

Übung zu Grundlagen der Bioinformatik

Veranstaltung

Diese **Übung** begleitet die Vorlesung Grundlagen der Bioinformatik

Erster Übungstermin ist der 1.11.2010. Dieser Termin ist **Pflicht für alle Teilnehmer**. Unentschuldigtes Nichterscheinen hat den Ausschluss von der Übung zur Folge.

Ablauf

In der Übung müssen **typische Aufgaben** im Bereich der Bioinformatik gelöst werden. dies umfasst sowohl die Neuimplementierung einfacher Verfahren als auch die Verwendung existierender Tools.

Die Arbeit erfolgt in Gruppen a zwei Studierenden.Jede Gruppe muss alle Aufgaben erfolgreich bearbeitet (aber nicht immer komplett). Die Aufgaben werden an einem Übungstermin ausgegeben, und die Lösungen müssen meist zwei Wochen später von einem der Gruppenmitglieder im Rahmen eines **kurzen Vortrags** dargestellt werden. In dem Vortrag geht es vor allem darum, gesammelte Erfahrungen an die gesamte Zuhörerschaft zu kommunizieren.

Die einzelnen Aufgaben und Termine

Diese Liste wird ständig aktualisiert. Folien zu den Aufgaben und notwendige Daten werden hier veröffentlicht.

- 1.11.2010: Erste Aufgabe. Stichwort: Substringsuche.
- 15.11.2010: Zweite Aufgabe. Stichwort: Lokales Alignment
- 29.11.2010: Dritte Aufgabe. Stichwort: Hierarchisches Clustering
- tba: Vierte Aufgabe. Stichwort: Genexpressionsanalyse mit R
- tba: Fünfte Aufgabe. Stichwort: Cliquen in PPI Netzen
- tba: Sechste Aufgabe. Stichwort: tba
- 14.2.2010: Abschluss

Humboldt-Universitaet zu Berlin / Department of Computer Science / Forschung und Lehre / Lehr- und Forschungsgebiete / Wissensmanagement in der Bioinformatik / Lehre / Archiv / WS 10/11 / Übung Grundlagen der Bioinformatik

last modified 10-09-22 UL
Edit page

Department of Computer Science

- Deutsch

Wissensmanagement in der Bioinformatik
- Contact
- Mitarbeiter
- Veranstaltungen
- Lehre
  - Archiv
    - WS 10/11
      - Text Analytics
      - Praktikum Text Analytics
      - Grundlagen der Bioinformatik
      - Übung Grundlagen der Bioinformatik
      - Forschungsseminar WBI - DBIS
    - SS 10
    - WS 09/10
    - SS 09
    - WS 08/09
    - SS 08
    - WS 07/08
    - SS 07
    - WS 06/07
    - SS 06
    - WS 05/06
    - SS 05
    - WS 04/05
    - SS 04

Suchen: tikk

My Questions

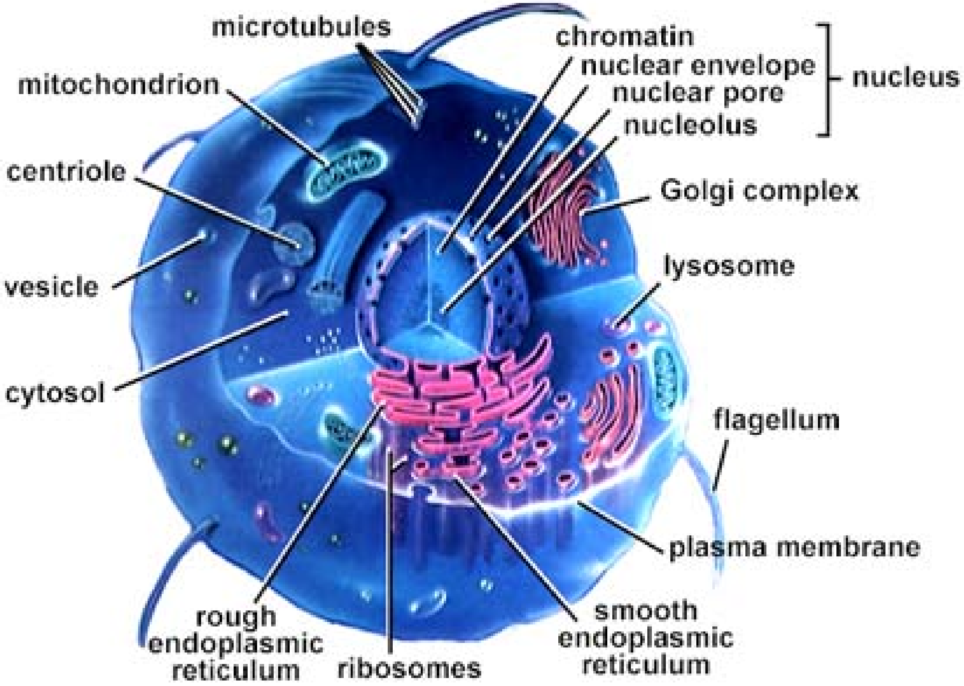
- Diplominformatiker?
- Bachelor Informatik?
- Kombibachelor?
- Biophysik?
- Other?

- Semester?
- Prüfung?
- Spezielle Erwartungen?

This Lecture

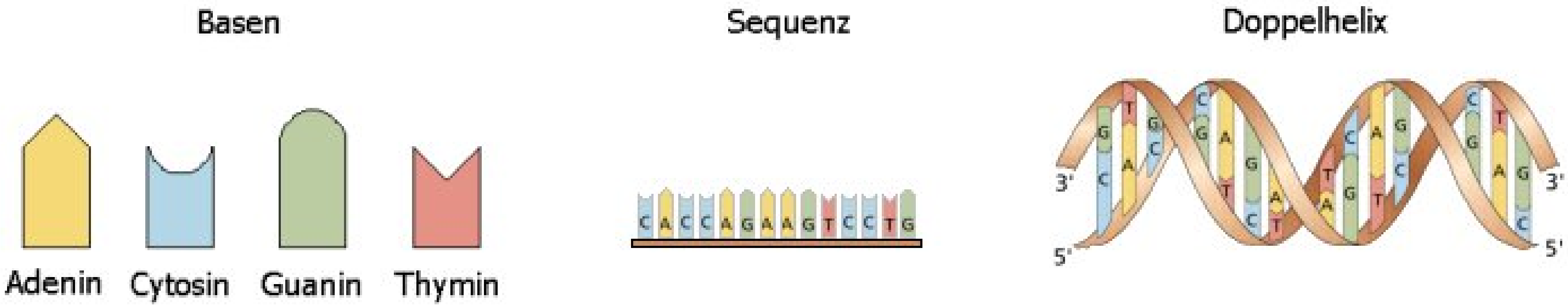
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
- Topics of this course

Cells and Bodies



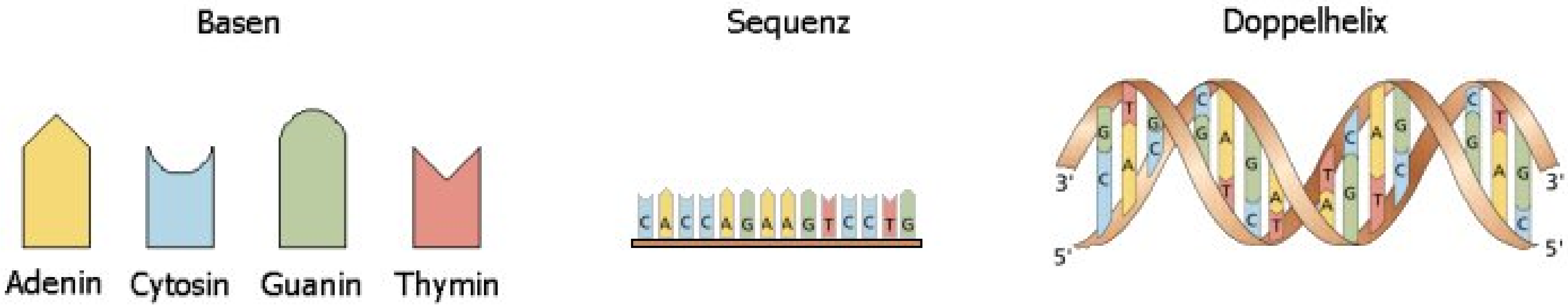
- App. 75 trillion cells in a human body
- App. 250 different types: nerve, muscle, skin, blood, ...

DesoxyriboNucleicAcid



- DNA: Desoxyribonukleinsäure
  - Four different molecules
  - The DNA of all chromosomes in a cell forms its genome
  - All cells in a (human) body carry the same genome
  - All living beings are based on DNA for proliferation
  - There are always always always exceptions

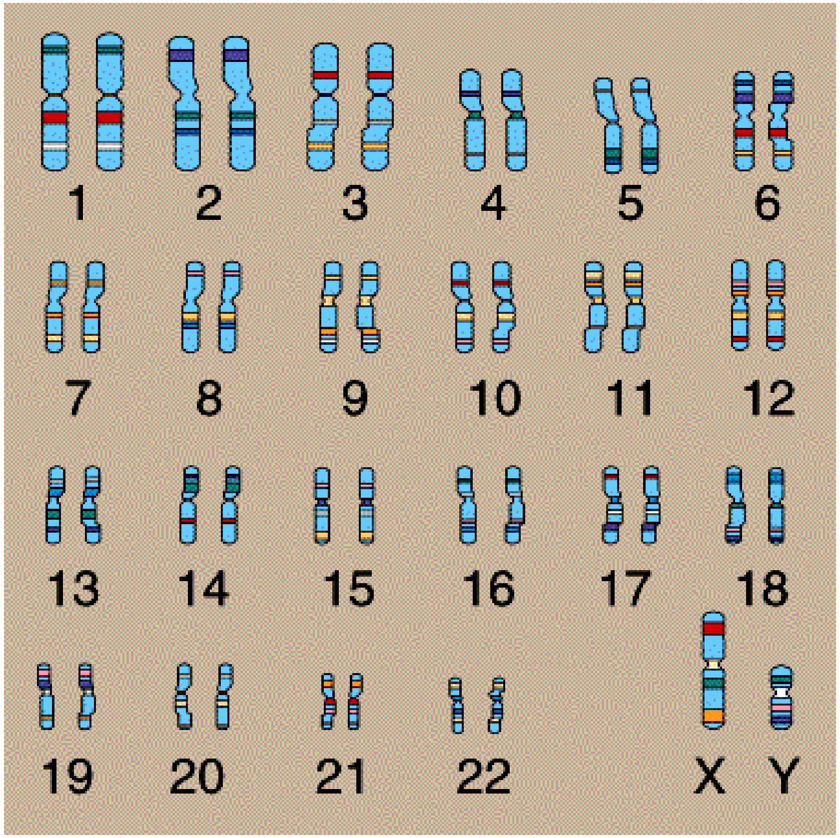
DesoxyriboNucleicAcid



- DNA: Desoxyribonukleinsäure
  - Four different molecules (one replaced in RNA)
  - The DNA of all chromosomes in a cell together with the mitochondria-DNA forms its genome
  - Almost all cells in a (human) body carry almost the same genome
  - All living beings are based on DNA or RNA for proliferation

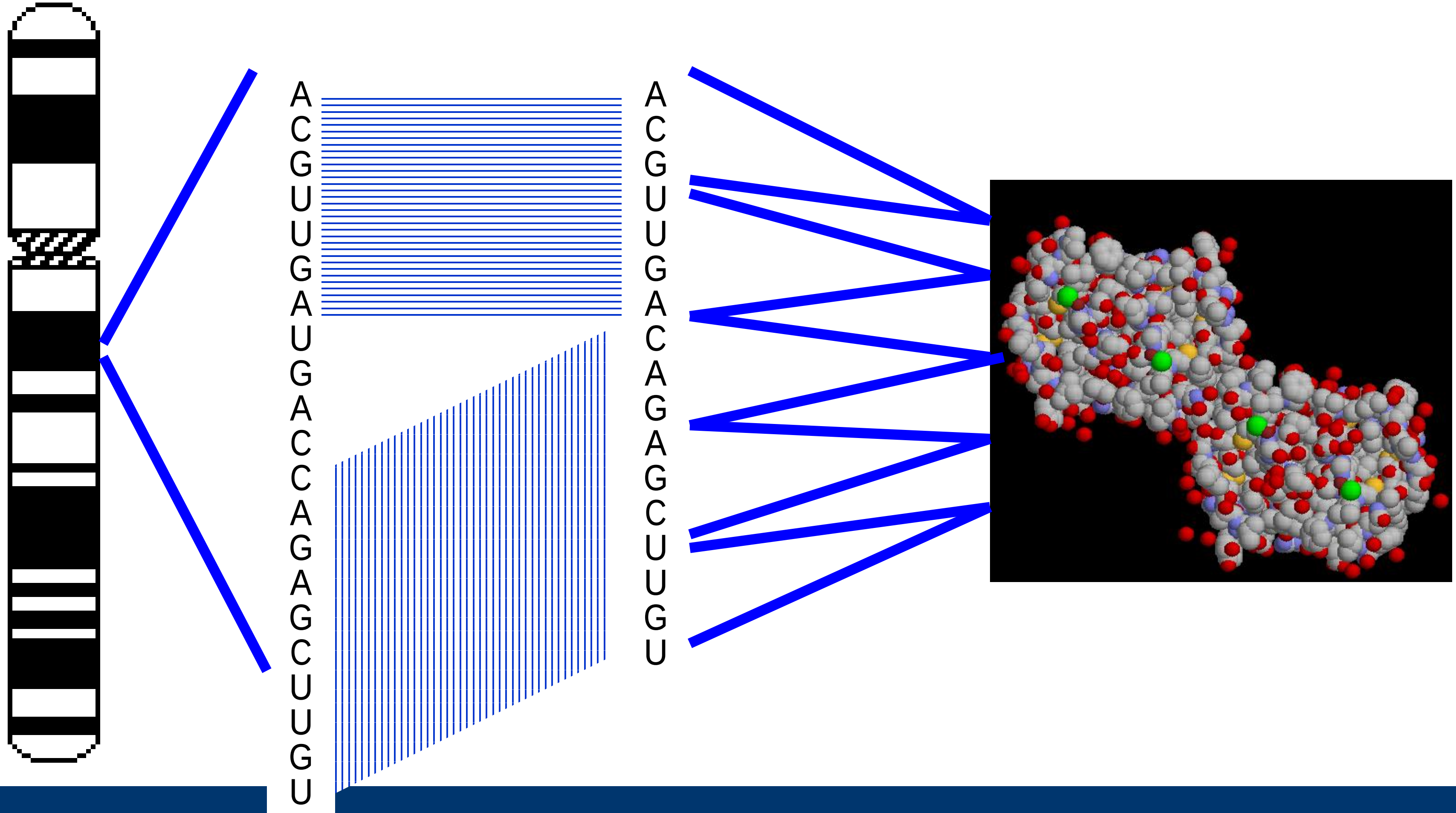
The Human Genome

- 23 chromosomes
  - Most in pairs
- ~3.000.000.000 letters
- ~50% are repetitions of 4 identical subsequences
  - ~~~100.000 genes~~
  - ~~~56.000 genes~~
  - ~~~30.000 genes~~
  - ~~~24.000 genes~~
- ~20.000 genes

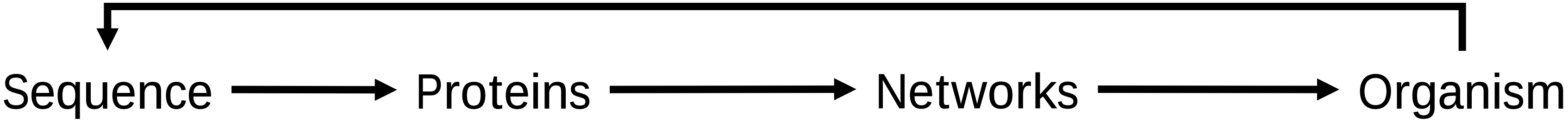


(Protein-Coding) Genes

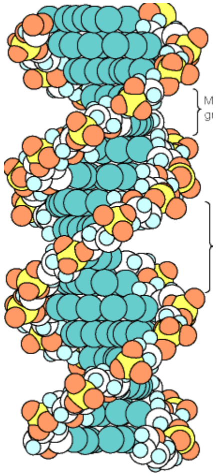
Chromosome RNA mRNA Proteine

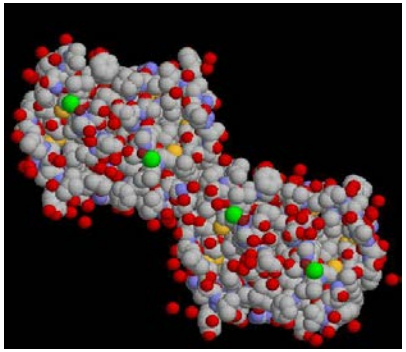


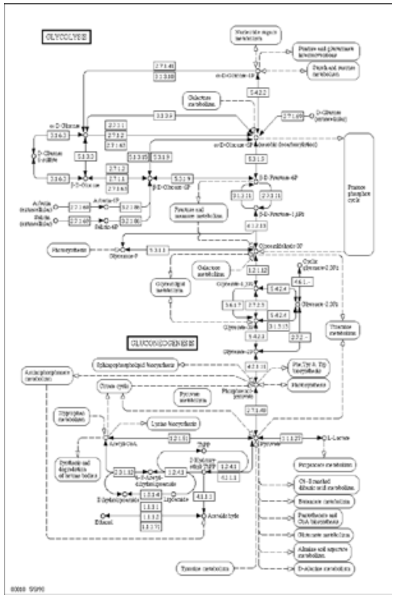
Proliferation

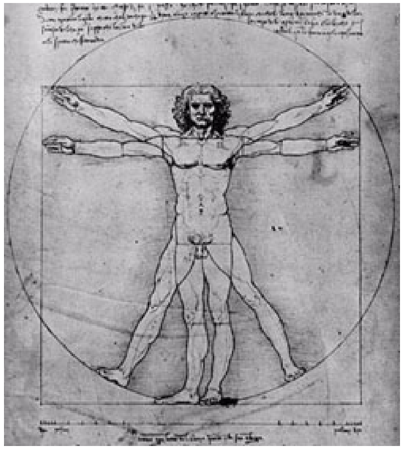


Sequence → Proteins → Networks → Organism

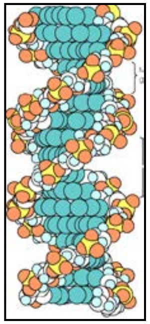








Computer Science in Molecular Biology / Medicine



Genomics

Sequencing

Gene prediction

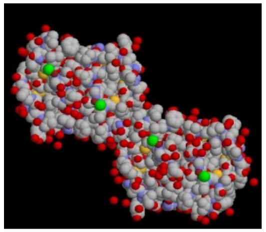
Evolutionary relationships

Motifs - TFBS

Transcriptomics

RNA folding

...



Proteomics

Structure prediction

... comparison

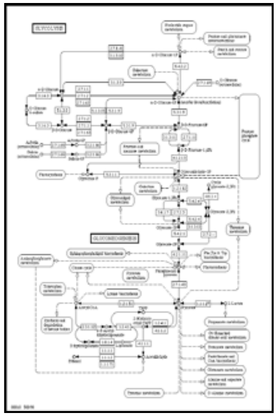
Motives, active sites

Docking

Protein-Protein Interaction

Proteomics

...



Systems Biology

Pathway analysis

Gene regulation

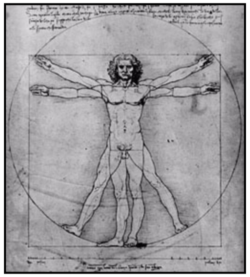
Signaling

Metabolism

Quantitative models

Integrative analysis

...



Medicine

Phenotype – genotype

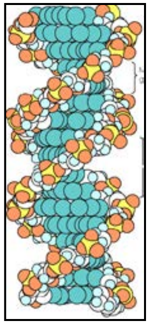
Mutations and risk

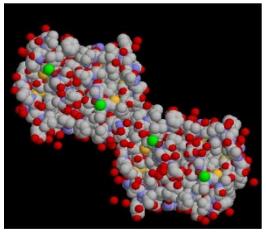
Population genetics

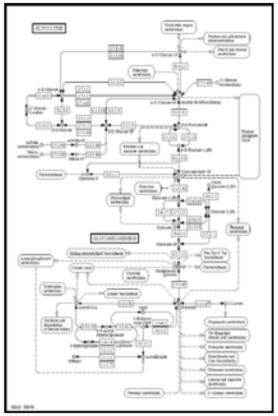
Adverse effects

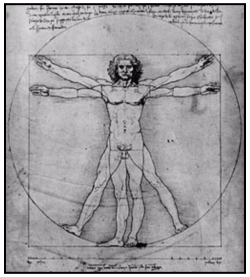
...

This Lecture









Genomics

Sequencing
Gene prediction
Evolutionary relationships
Motifs - TFBS
Transcriptomics
RNA folding
...

Proteomics

Structure prediction
... comparison
Motives, active sites
Docking
Protein-Protein Interaction
Proteomics
...

Systems Biology

Pathway analysis
Gene regulation
Signaling
Metabolism
Quantitative models
Integrative analysis
...

Medicine

Phenotype – genotype
Mutations and risk
Population genetics
Adverse effects
...

This Lecture

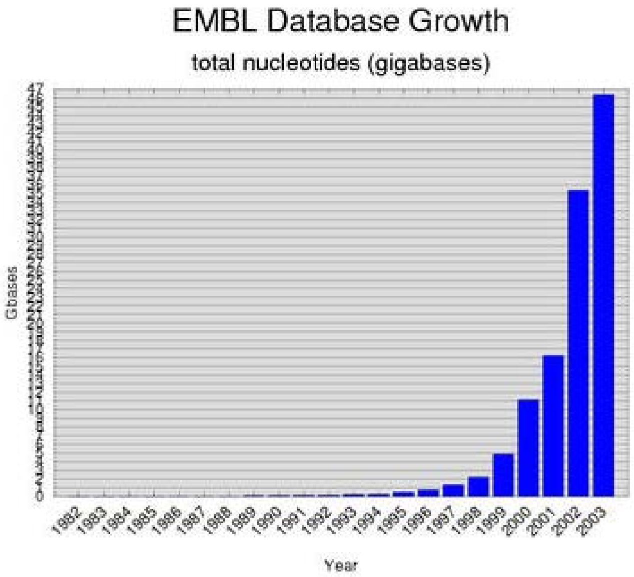
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
  - And an example
- Topics of this course

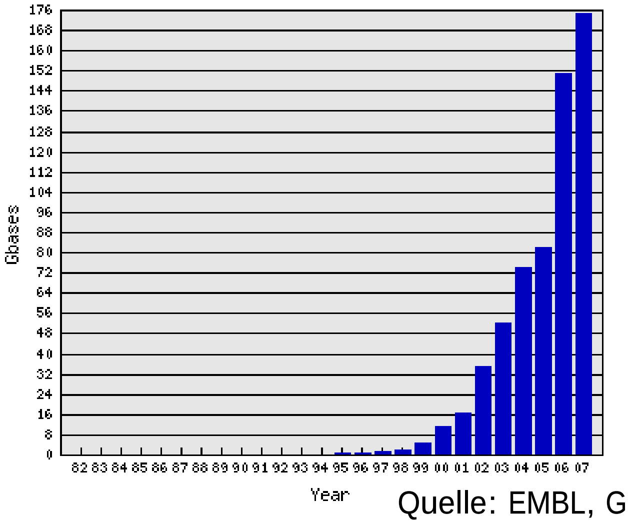
Bioinformatics / Computational Biology

- Computer Science methods for
  - Solving biologically relevant problems
  - Analyzing and managing experimental data sets
- Empirical: Data from high throughput experiments
- Focused on algorithms and statistics
- Problems are typically complex, data full of errors – importance of heuristics and approximate methods
- Strongly reductionist – Strings, graphs, sequences
- Interdisciplinary: Biology, Computer Science, Physics, Mathematics, Genetics, …

History

- First protein sequences: 1951
- Sanger sequencing: 1972
- Exponential growth of available data since end of 70th
  - Bioinformatics is largely data-driven – new methods yield new data requiring new algorithms





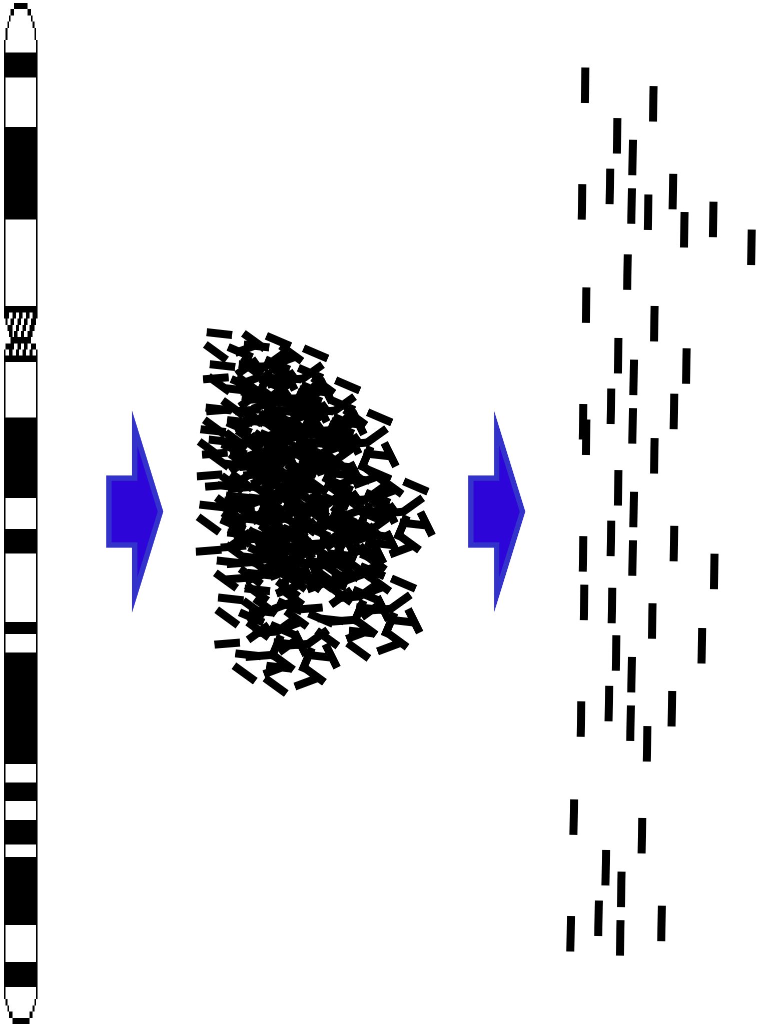
Quelle: EMBL, Genome Monitoring Tables

History 2

- First papers on sequence alignment
  - Needleman-Wunsch 1970, Gibbs 1970, Smith-Waterman 1981, Altschul et al. 1990
- Large impact of the Human Genome Projekt (~1990)
- Only 14 mentions of „Bioinformatics“ before 1995
- „Journal of Computational Biology“ since 1994
- First professorships in Germany: end of 90th
- First university programs: ~2000
- First German book: 2001
- Commercial hype: 1999 – 2004

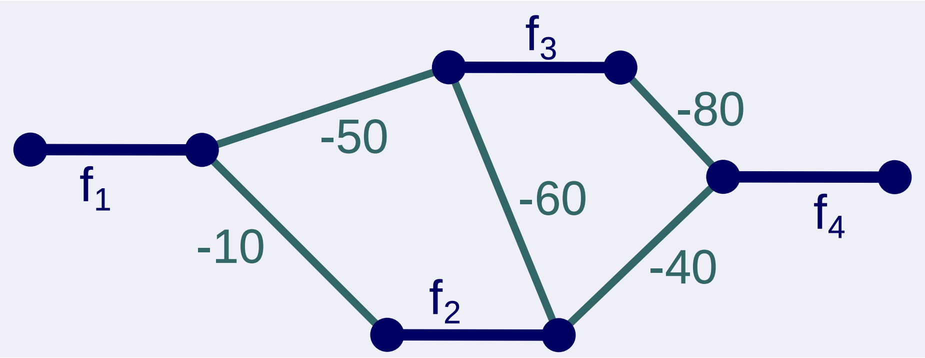
A Concrete Example: Sequencing a Genome

- Chromosomes (yet) cannot be sequenced entirely
  - Instead: Only small fragments can be sequenced
- But: Chromosomes cannot be cut at position X, Y, ...
  - Instead: Chromosomes only can be cut at certain subsequences
- But: We don’t know where in a chromosome those subsequences are
  - Sequence assembly problem



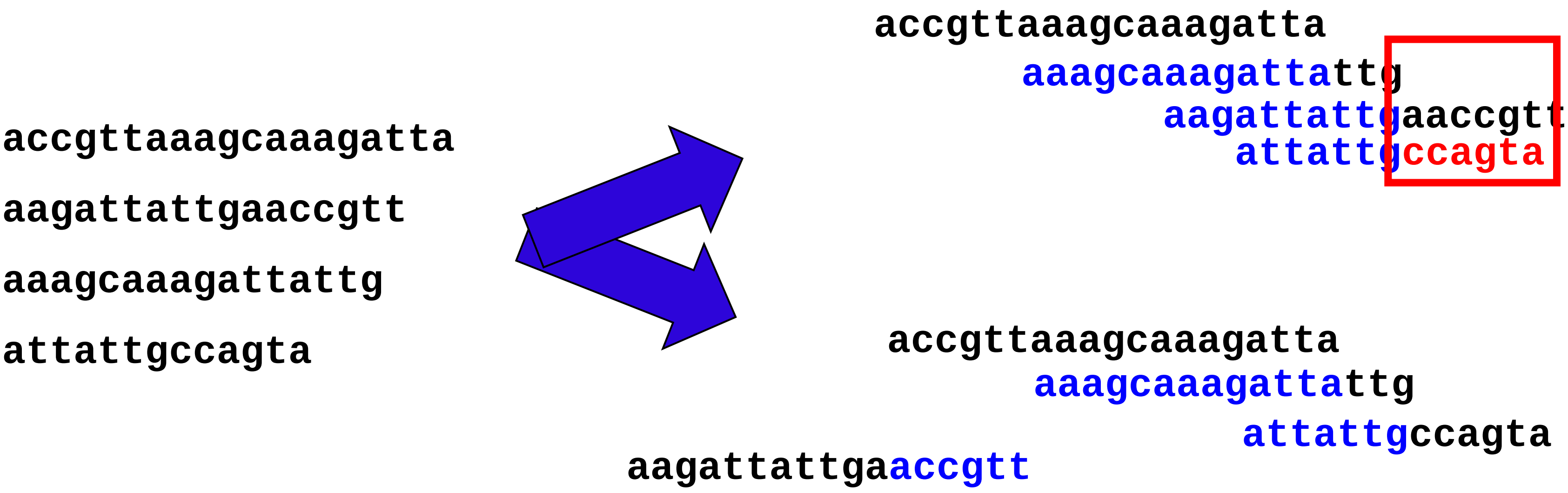
Problem

- Given a large set of (sub)sequences from randomly chosen positions from a given chromosome of unknown sequence
- Assembly problem: Determine the sequence of the original chromosome
  - Everything may overlap with everything to varying degrees
  - Let‘s forget about orientation and sequencing errors



Greedy?

- Take one sequence and compute overlap with all others
- Keep the one with largest overlap and align
- Repeat such extensions until no more sequences are left
  - Note: This would work perfectly if all symbols of the chromosome were distinct



```
accgttaaagcaaagatta
aagattattgaaccgtt
aaagcaaagattattg
attattgccagta
```

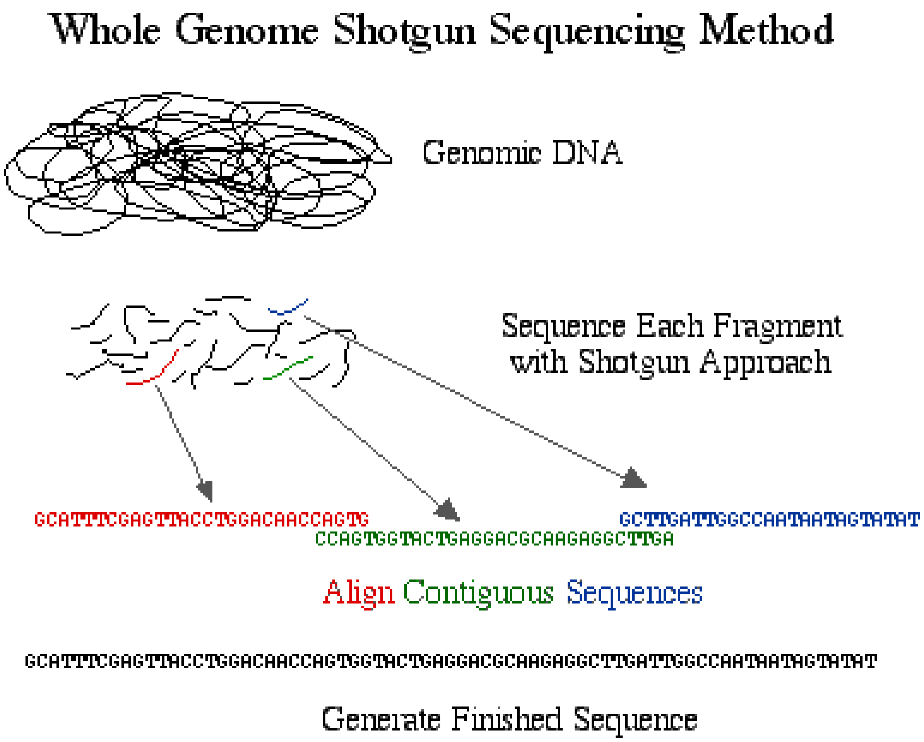
```
accgttaaagcaaagatta
      aaagcaaagattattg
            aagattattgaaccgtt
               attattgccagta
```

```
           accgttaaagcaaagatta
                 aaagcaaagattattg
                          attattgccagta
aagattattgaaccgtt
```

Abstract Formulation

- SUPERSTRING
  - Given a set S of strings
  - Find string t such that
    - (a) $\forall s \in S: s \in t$ (all s are substrings of t)
    - (b) $\forall t'$ for which (a) holds: : $|t| \leq |t'|$ (t ist minimal)
- Problem is NP-complete
  - Very likely, there is no algorithm that solves the problem in less than $k_1 * k_2 2^n$ operations, where k_1, k_2 are constants and $n=|S|$
- Bioinformatics: Find clever heuristics
  - Solve the problem “good enough”
  - Finish in reasonable time

Dimension



- Whole genome shotgun
  - Fragment an entire chromosome in pieces of 1KB-100KB
- Sequence start and end of all fragments
  - Homo sap.: 28 million reads
  - Drosophila: 3.2 million reads
- Eukaryotes are very difficult to assemble because of repeats
  - A random sequence is easy

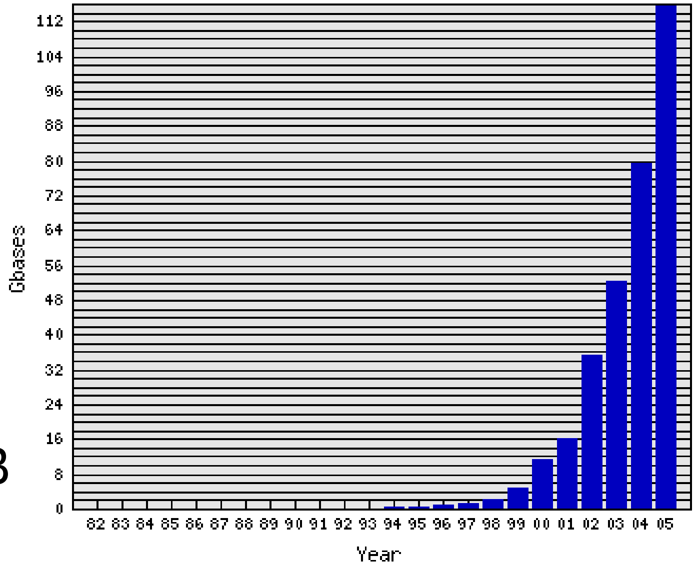
This Lecture

- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
  - And an example
- Topics of this course

Searching Sequences (Strings)

- A chromosome is a string
- Substrings may represent biologically important areas
  - Genes on a chromosome
  - Transcription factor binding sites
  - Similar gene in a different species
  - ...
- Exact or approximate string search

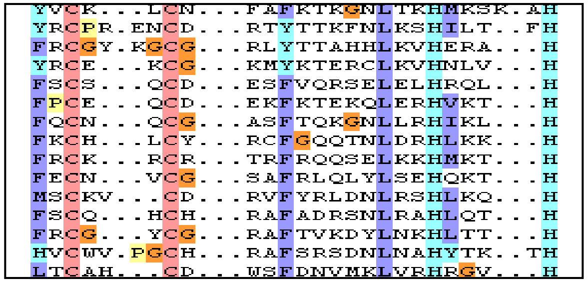
Searching a Database of Strings



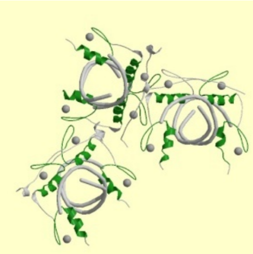
- Comparing two sequences is costly
- Given s, assume we want to find the most similar s’ in a database of all known sequences
  - Naïve: Compare s with all strings in DB
  - Will take years and years
- BLAST: Basic local alignment search tool
  - Ranks all strings in DB according to similarity to s
  - Similarity: High is s, s’ contain substrings that are highly similar
  - Heuristic: Might miss certain similar sequences
  - Extremely popular: You can “blast a sequence”

Multiple Sequence Alignment

- Given a set S of sequences: Find an arrangement of all strings in S in columns such that there are (a) few columns and (b) columns are maximally homogeneous
  - Additional spaces allowed



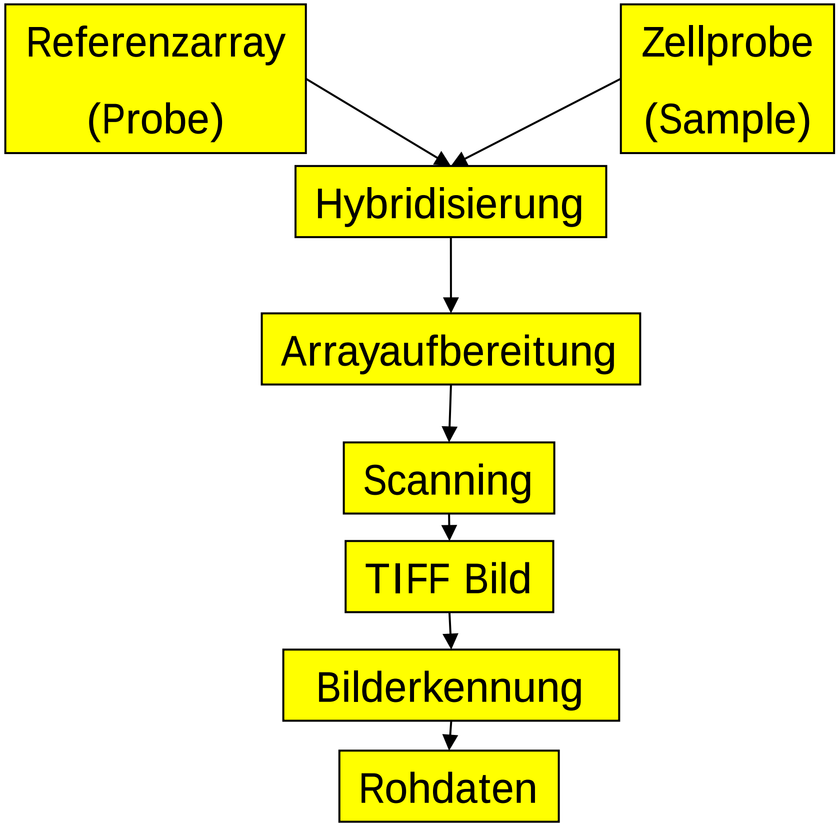
```
YVCK...LCN...FAFKTKGNLTKHMKSK.AH
YRCPR.ENCD...RTYTTKFNLKSHILT..FH
FRCGY.KGCG...RLYTTAHHLKVHERA...H
YRCE...KCG...KMYKTERCLKVHNLV...H
FSCS...QCD...ESFVQRSELELHRQL...H
FPCE...QCD...EKFKTEKQLERHVKT...H
FQCN...QCG...ASFTQKGNLLRHIKL...H
FKCH...LCY...RCFGQQTNLDRHLKK...H
FRCK...RCR...TRFRQQSELKKHMKT...H
FECN...VCG...SAFRLQLYLSEHQKT...H
MSCKV...CD...RVFYRLDNLRSHLKQ...H
FSCQ...HCH...RAFADRSNLRAHLQT...H
FRCG...YCG...RAFTVKDYLNKHLTT...H
HVCWV.PGCH...RAFSRSDNLNAHYTK..TH
LTCAH...CD...WSFDNVMKLVRHRGV...H
```



Source: Pfam, Zinc finger domain

- Goal: Find commonality between a set of functionally related sequences
  - Proteins are composed of different functional domains
  - Which domain performs a certain function?

Microarrays / Transcriptomics



Referenzarray (Probe)

Zellprobe (Sample)

Hybridisierung

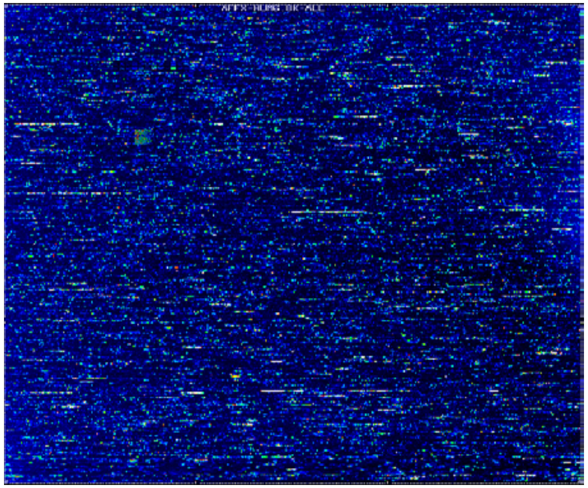
Arrayaufbereitung

Scanning

TIFF Bild

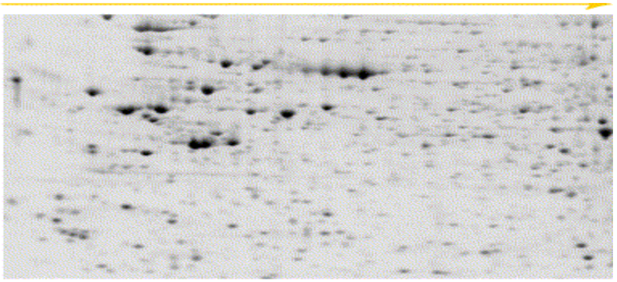
Bilderkennung

Rohdaten

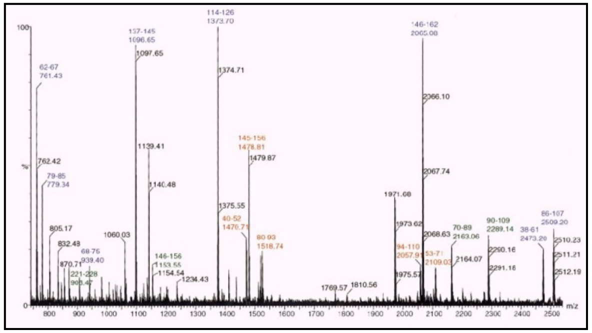


Proteomics

- The real workhorses in a cell are proteins
  - Differential splicing, post-translational modifications, degradation rates, various levels of regulation, ...
- But: Much more difficult to study (compared to mRNA)
- Separation of proteins
  - 2D page, GC / LC
- Identification of proteins
  - Mass-spectrometry

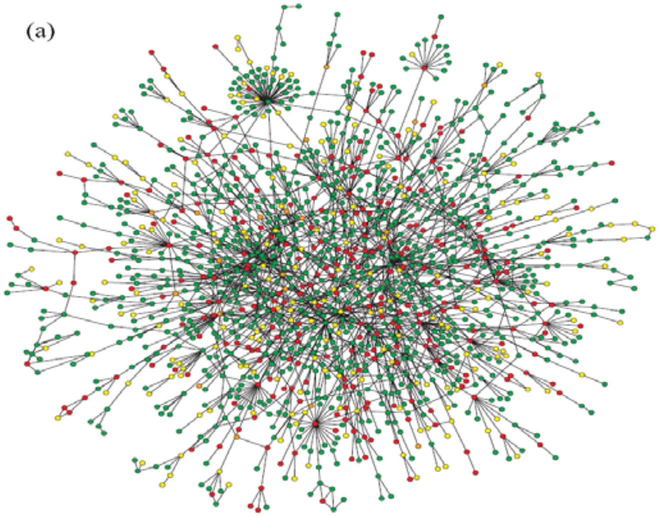


Drug Discovery Today



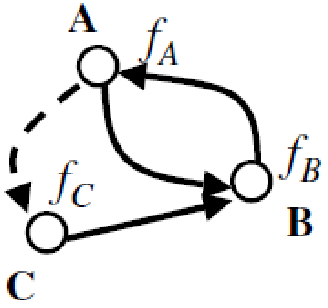
Protein-Protein-Interactions

- Proteins do not work in isolation but interact with each other
  - Metabolism, complex formation, signal transduction, transport, …
- PPI networks
  - Neighbors tend to have similar function
  - Interactions tend to be evolutionary conserved
  - Dense subgraphs (cliques) tend to perform distinct functions
  - Are not random at all



Network Reconstruction

- Molecules perform functions by means of interactions
- Regulation: Networks of genes regulating each other
- Reconstruction: Which gene regulates which other genes in which ways?
- One approach: Boolean networks



$$f_A(B) = B$$

$$f_B(A,C) = A \text{ and } C$$

$$f_C(A) = \text{not } A$$

Boolean Network