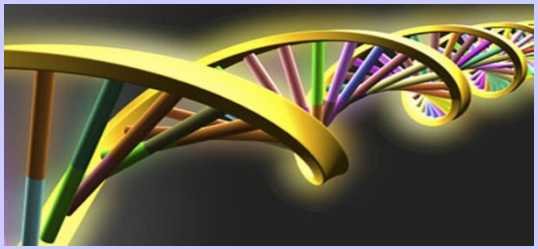
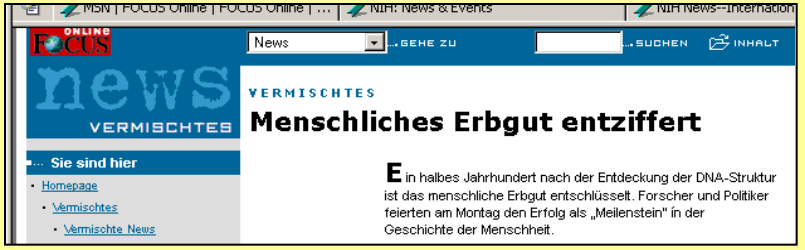
Introduction to Bioinformatics

Ulf Leser

Bioinformatics

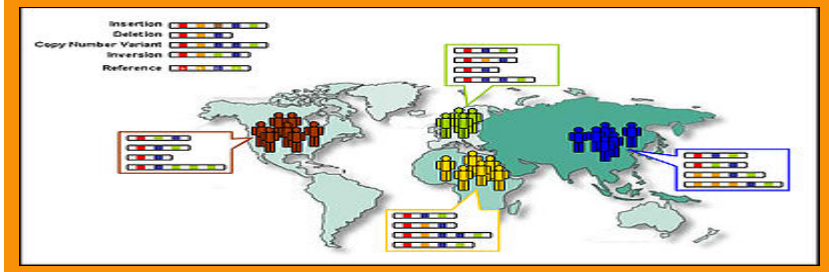


25.4.2003
50. Jubiläum der Entdeckung der Doppelhelix durch Watson/Crick



14.4.2003
Humanes Genom zu 99% sequenziert mit 99.99% Genauigkeit

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 federat
  to a large int.
  project [HAA+08]

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

> 50.000 genomes

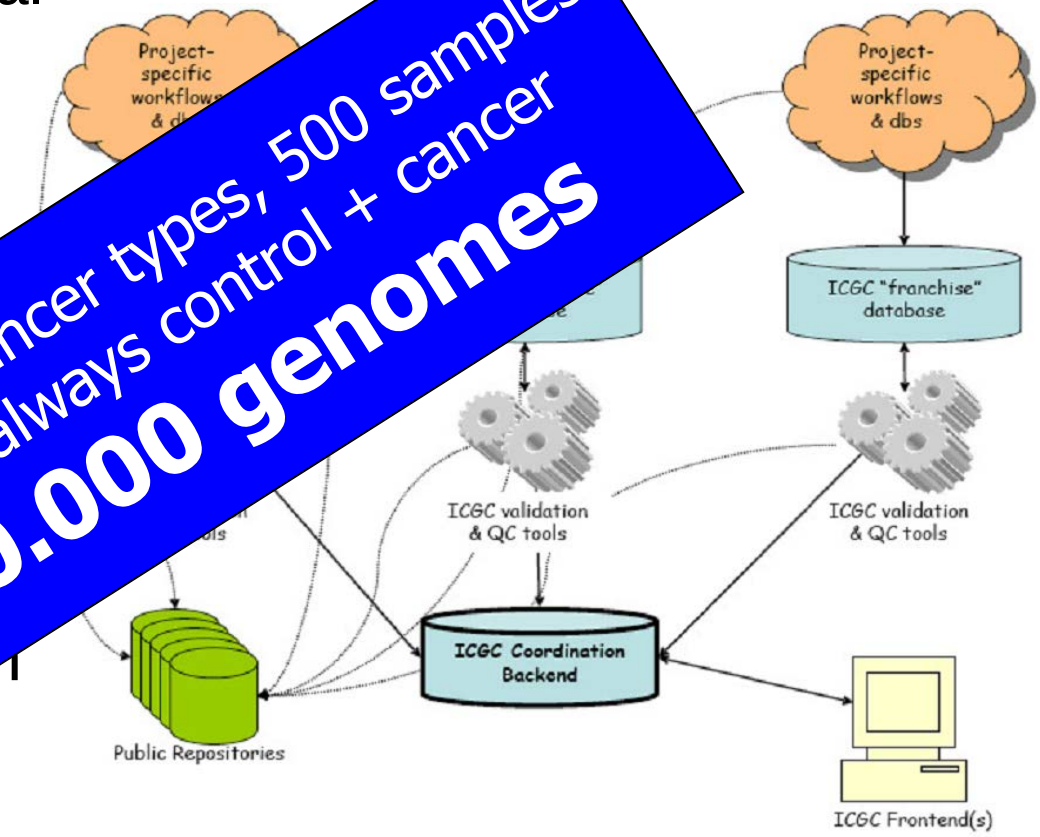
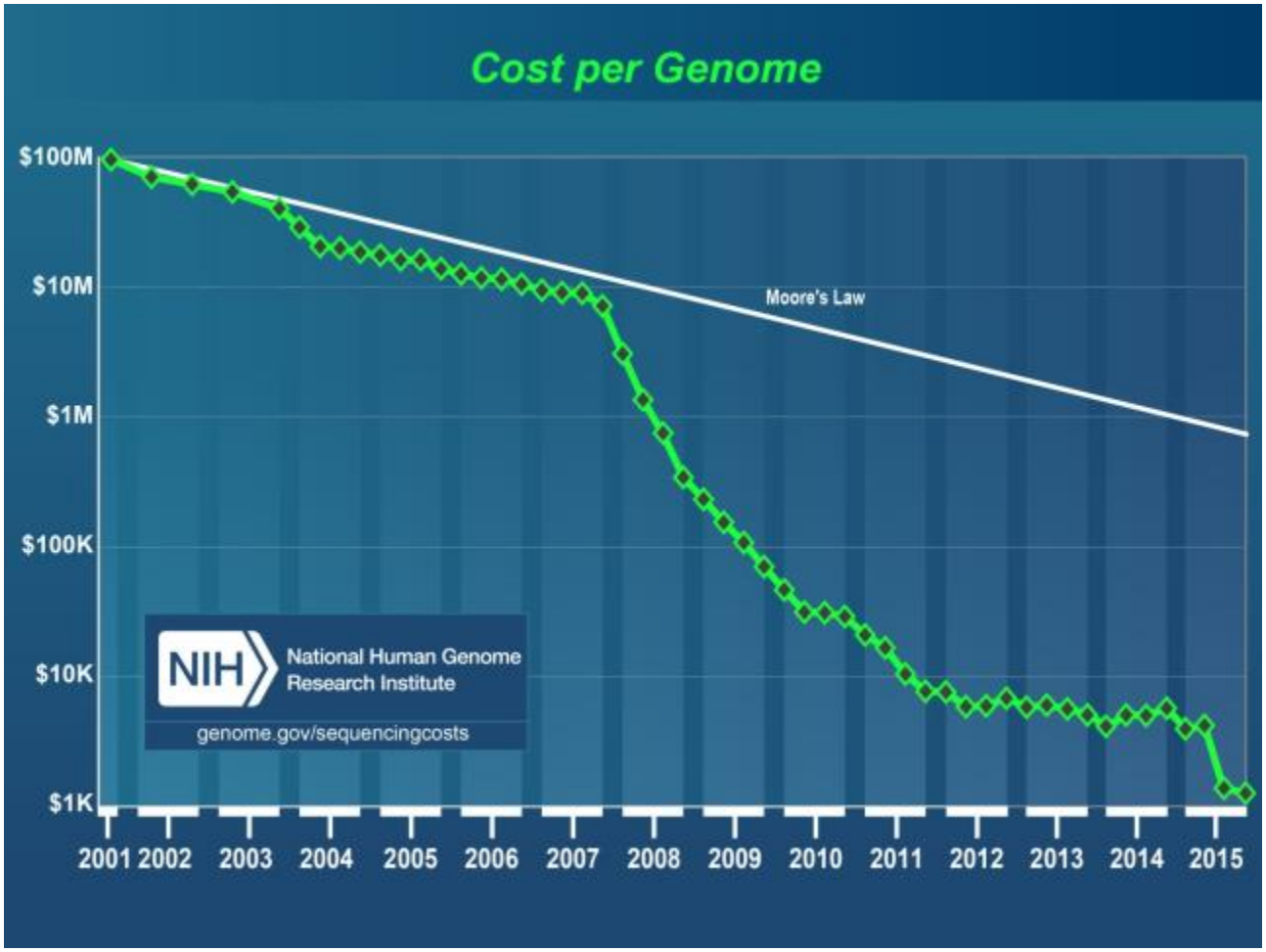


Figure 2: ICGC data coordination as a franchise system

Possible Through Cost Reduction



What does this mean ?

http://www.genome.gov

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…	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…	2	k. A.	512,81 €
A…	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 €
…hates 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…	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 €

This Lecture

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

This course

- Bachelor computer science, Wahlpflichtbereich
- 5 SP, lecture / exercises are 2+2
- Does assume basic knowledge in computer science
  - Programming, algorithms, complexity
- 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 Raik Otto
- There will be 5 assignments
- We build teams
- 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

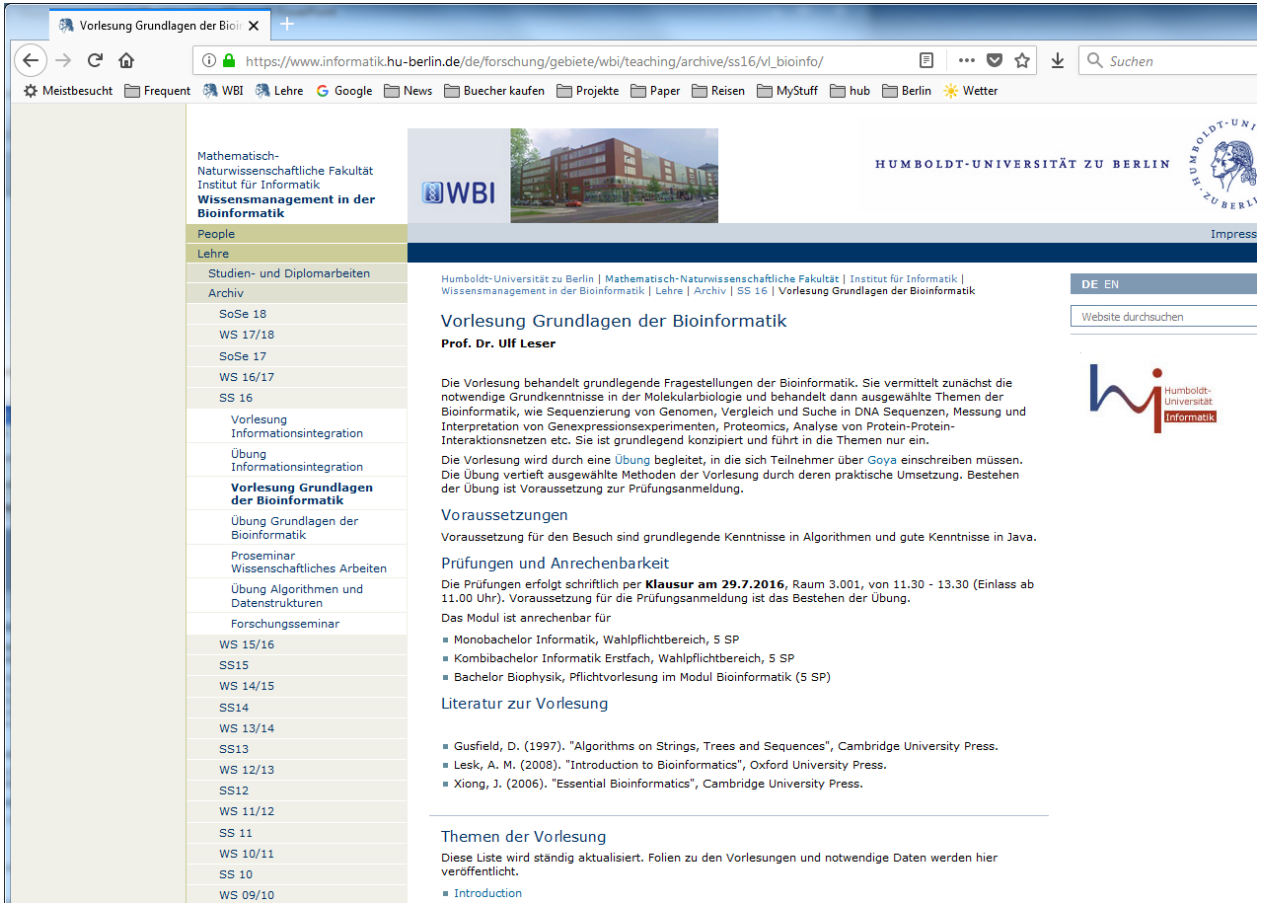
Exams

- Written examination
- Date to be announced

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 Side


Vorlesung Grundlagen der Bioi
https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ss16/vl_bioinfo/
Suchen
Meistbesucht Frequent WBI Lehre Google News Buecher kaufen Projekte Paper Reisen MyStuff hub Berlin Wetter
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik
Wissensmanagement in der Bioinformatik
WBI
HUMBOLDT-UNIVERSITÄT ZU BERLIN
People
Lehre
Studien- und Diplomarbeiten
Archiv
SoSe 18
WS 17/18
SoSe 17
WS 16/17
SS 16
Vorlesung Informationsintegration
Übung Informationsintegration
Vorlesung Grundlagen der Bioinformatik
Übung Grundlagen der Bioinformatik
Proseminar Wissenschaftliches Arbeiten
Übung Algorithmen und Datenstrukturen
Forschungsseminar
WS 15/16
SS15
WS 14/15
SS14
WS 13/14
SS13
WS 12/13
SS12
WS 11/12
SS 11
WS 10/11
SS 10
WS 09/10
Impress
Humboldt-Universität zu Berlin | Mathematisch-Naturwissenschaftliche Fakultät | Institut für Informatik | Wissensmanagement in der Bioinformatik | Lehre | Archiv | SS 16 | Vorlesung Grundlagen der Bioinformatik
DE EN
Website durchsuchen
Humboldt-Universität Informatik
Vorlesung Grundlagen der Bioinformatik
Prof. Dr. Ulf Leser
Die Vorlesung behandelt grundlegende Fragestellungen der Bioinformatik. Sie vermittelt zunächst die notwendige Grundkenntnisse in der Molekularbiologie und behandelt dann ausgewählte Themen der Bioinformatik, wie Sequenzierung von Genomen, Vergleich und Suche in DNA Sequenzen, Messung und Interpretation von Genexpressionsexperimenten, Proteomics, Analyse von Protein-Protein-Interaktionsnetzen etc. Sie ist grundlegend konzipiert und führt in die Themen nur ein.
Die Vorlesung wird durch eine Übung begleitet, in die sich Teilnehmer über Goya einschreiben müssen. Die Übung vertieft ausgewählte Methoden der Vorlesung durch deren praktische Umsetzung. Bestehen der Übung ist Voraussetzung zur Prüfungsanmeldung.
Voraussetzungen
Voraussetzung für den Besuch sind grundlegende Kenntnisse in Algorithmen und gute Kenntnisse in Java.
Prüfungen und Anrechenbarkeit
Die Prüfungen erfolgt schriftlich per Klausur am 29.7.2016, Raum 3.001, von 11.30 - 13.30 (Einlass ab 11.00 Uhr). Voraussetzung für die Prüfungsanmeldung ist das Bestehen der Übung.
Das Modul ist anrechenbar für
Monobachelor Informatik, Wahlpflichtbereich, 5 SP
Kombibachelor Informatik Erstfach, Wahlpflichtbereich, 5 SP
Bachelor Biophysik, Pflichtvorlesung im Modul Bioinformatik (5 SP)
Literatur zur Vorlesung
Gusfield, D. (1997). "Algorithms on Strings, Trees and Sequences", Cambridge University Press.
Lesk, A. M. (2008). "Introduction to Bioinformatics", Oxford University Press.
Xiong, J. (2006). "Essential Bioinformatics", Cambridge University Press.
Themen der Vorlesung
Diese Liste wird ständig aktualisiert. Folien zu den Vorlesungen und notwendige Daten werden hier veröffentlicht.
Introduction

My Questions

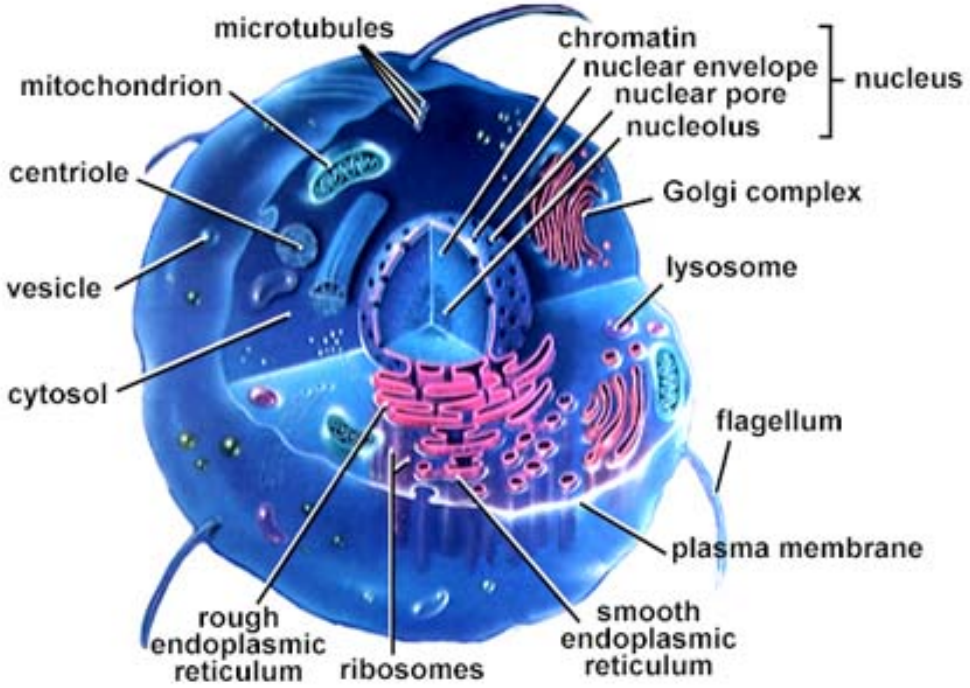
- Diplominformatiker?
- Bachelor Informatik?
- Kombibachelor?
- Biophysik?
- Other?

- Semester?
- Prüfung?
- Spezielle Erwartungen?

This Lecture

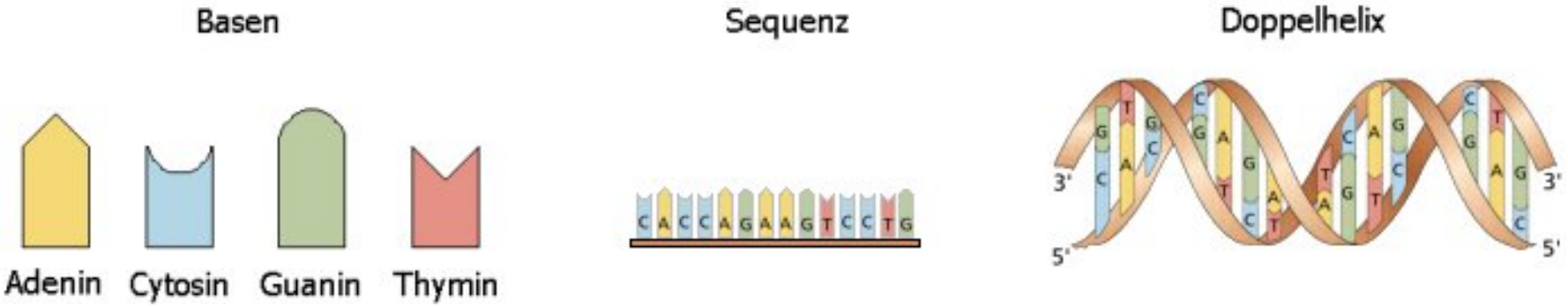
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
- Topics of this course

Cells and Bodies



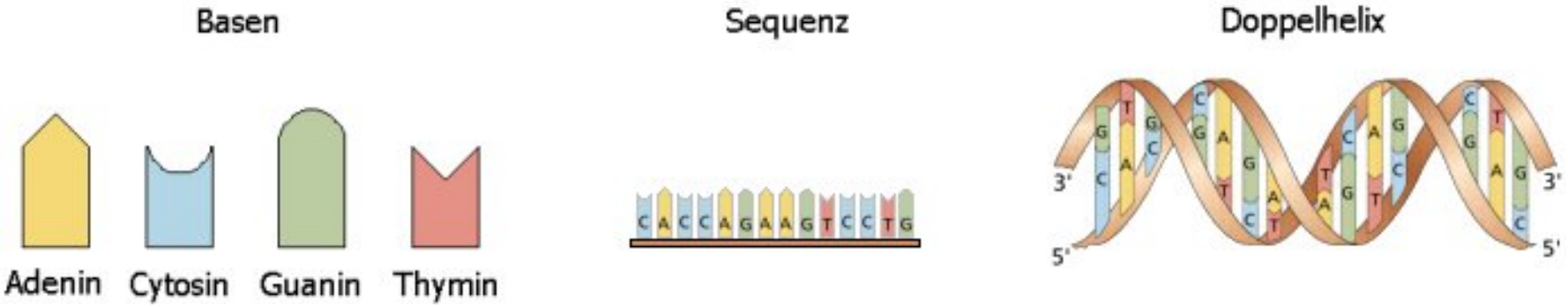
- App. 75 trillion cells in a human body
- App. 250 different types: nerve, muscle, skin, blood, ...

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

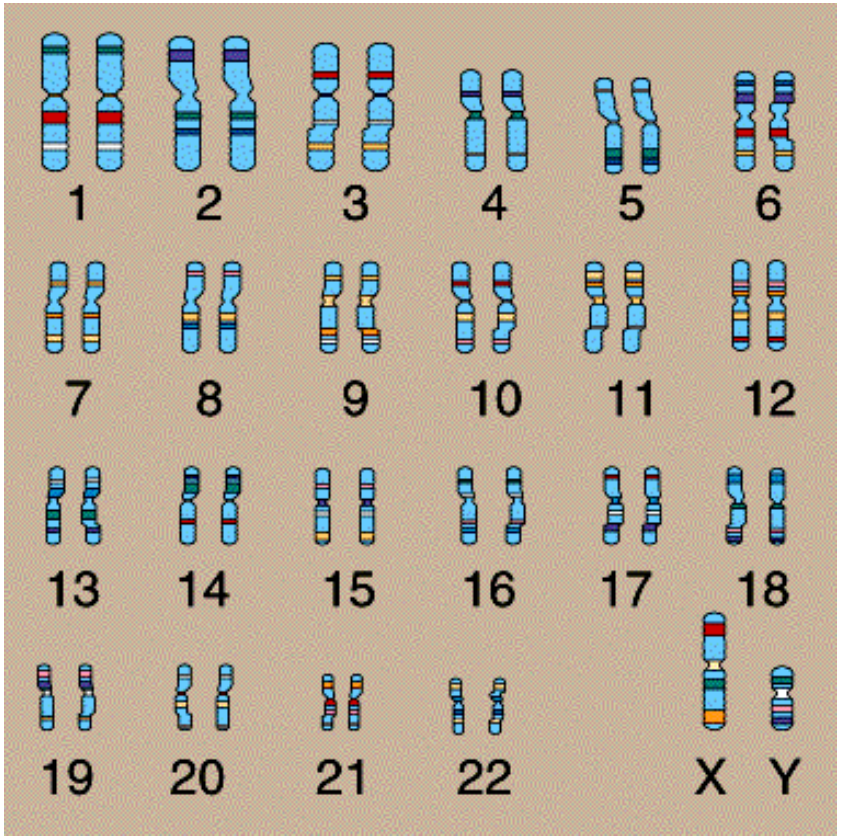
DesoxyriboNucleicAcid



- 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

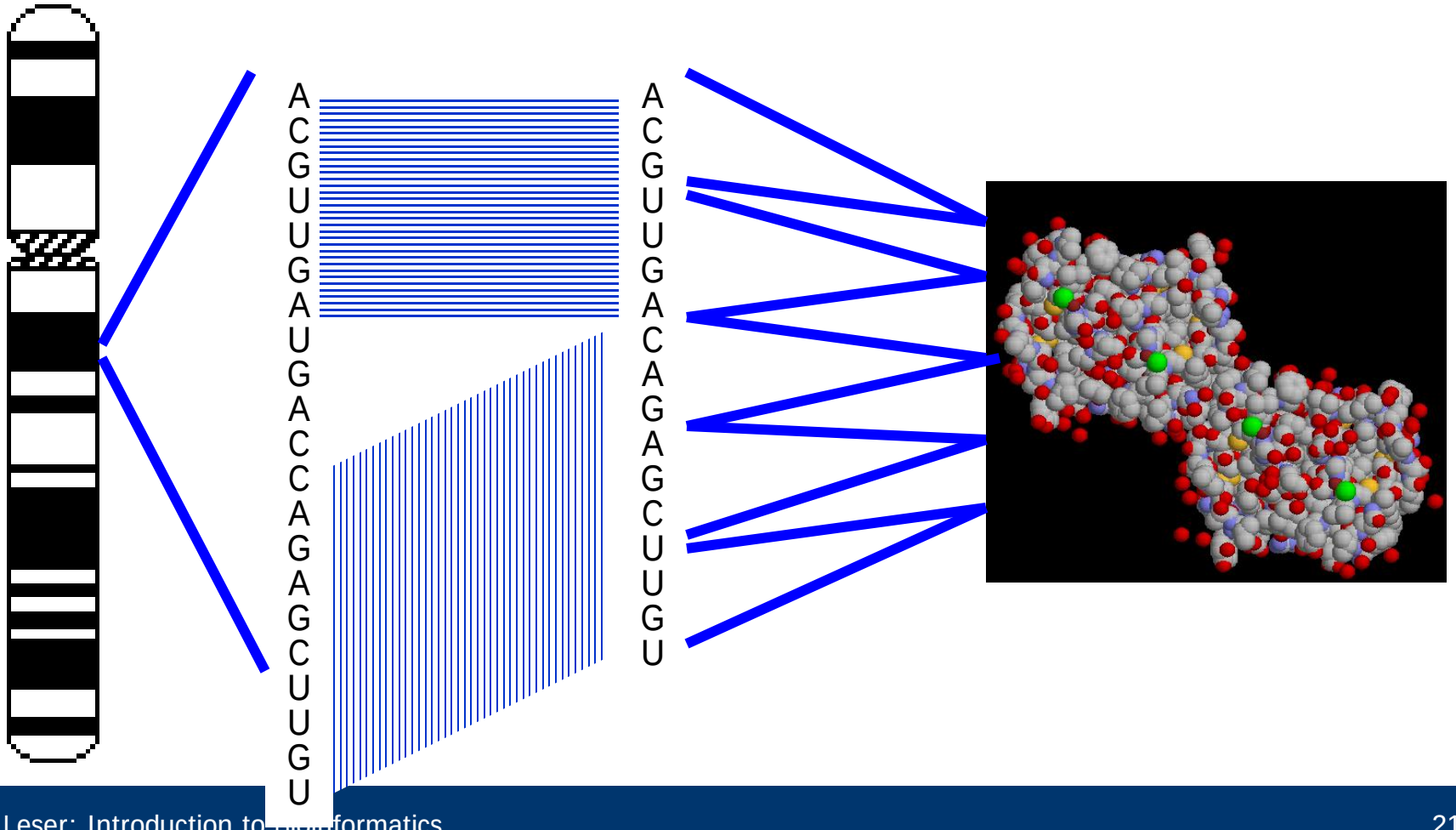
The Human Genome

- 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



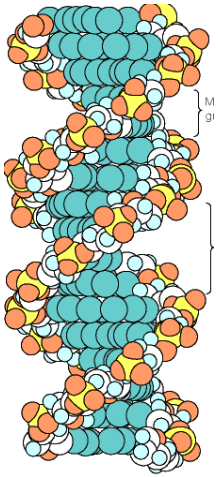
(Protein-Coding) Genes

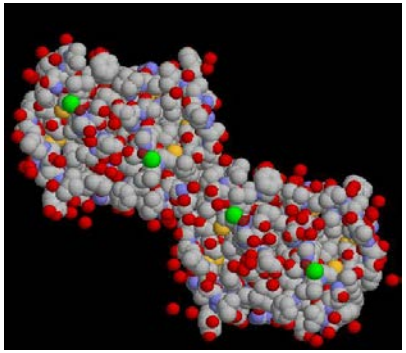
Chromosome RNA mRNA Proteine

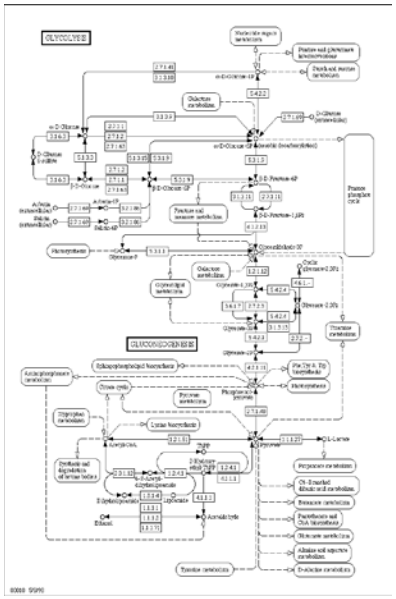


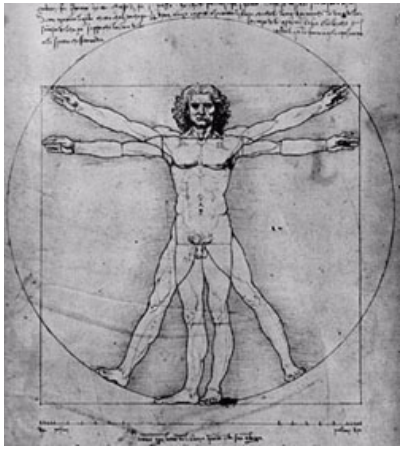
Proliferation

Sequence → Proteins → Networks → Organism

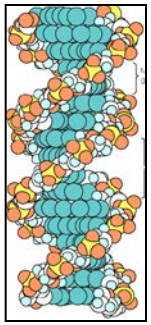


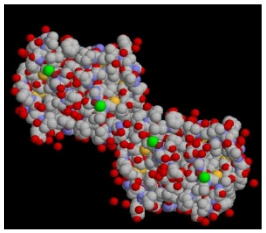


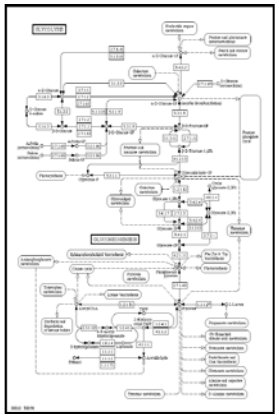


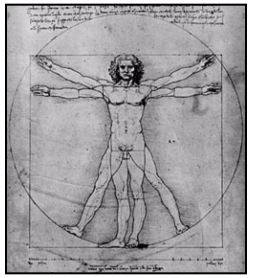


This Lecture









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

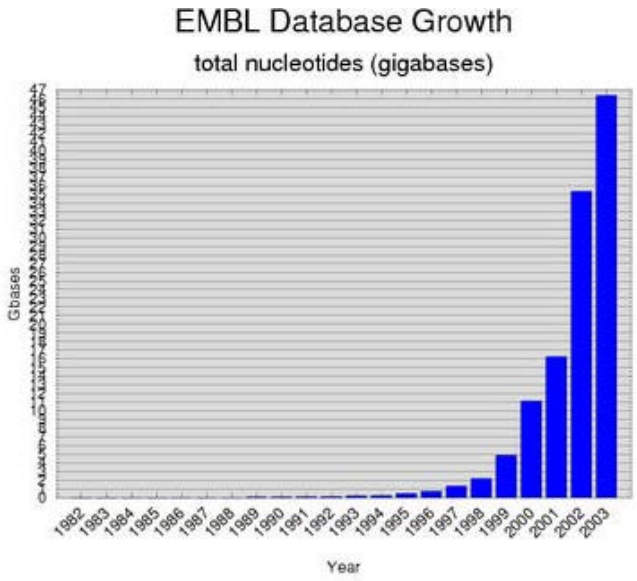
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
  - And an example
- Topics of this course

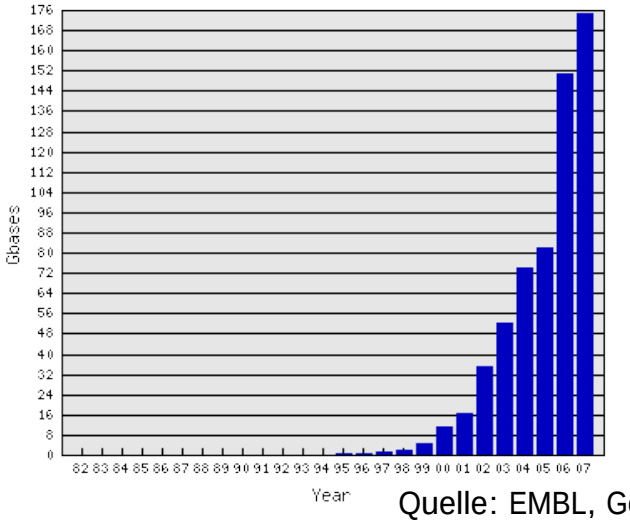
Bioinformatics / Computational Biology

- 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





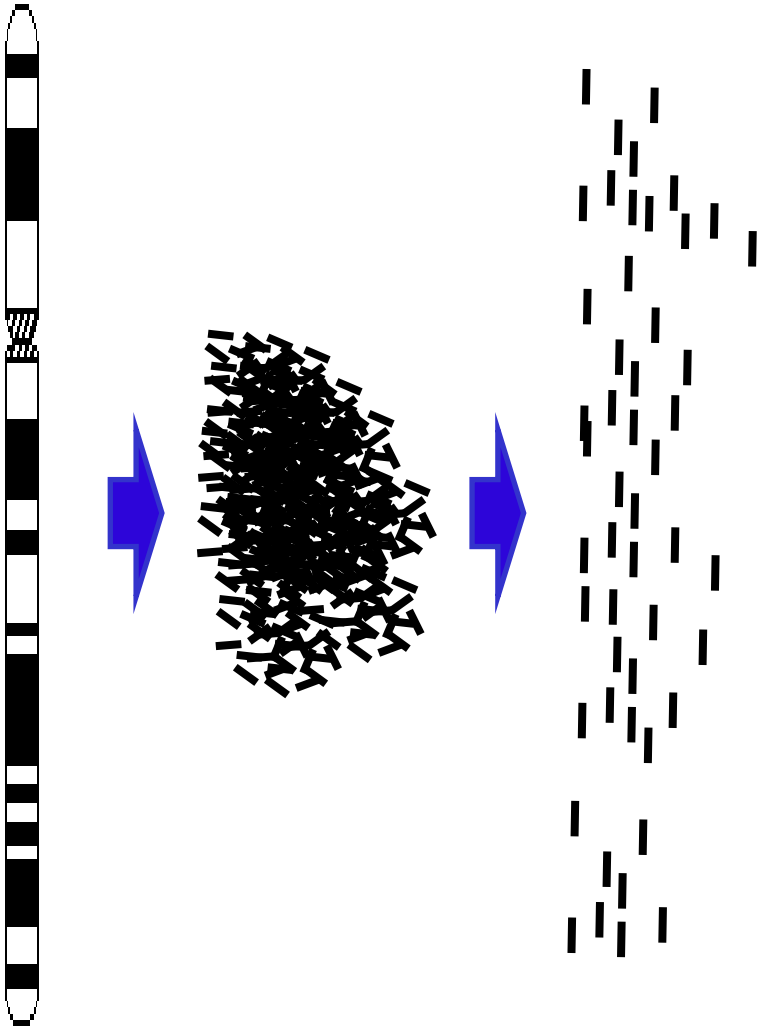
Quelle: EMBL, Genome Monitoring Tables

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 90’s
- First university programs: ~2000
- First German book: 2001
- Commercial hype: 1999 – 2004

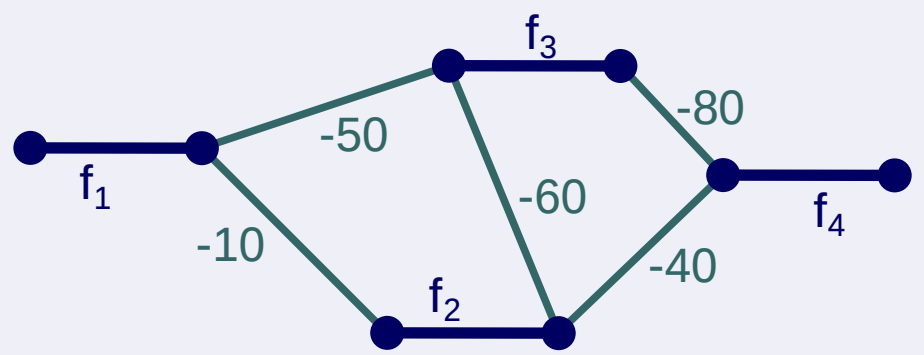
A Concrete Example: Sequencing a Genome

- Chromosomes (still) 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



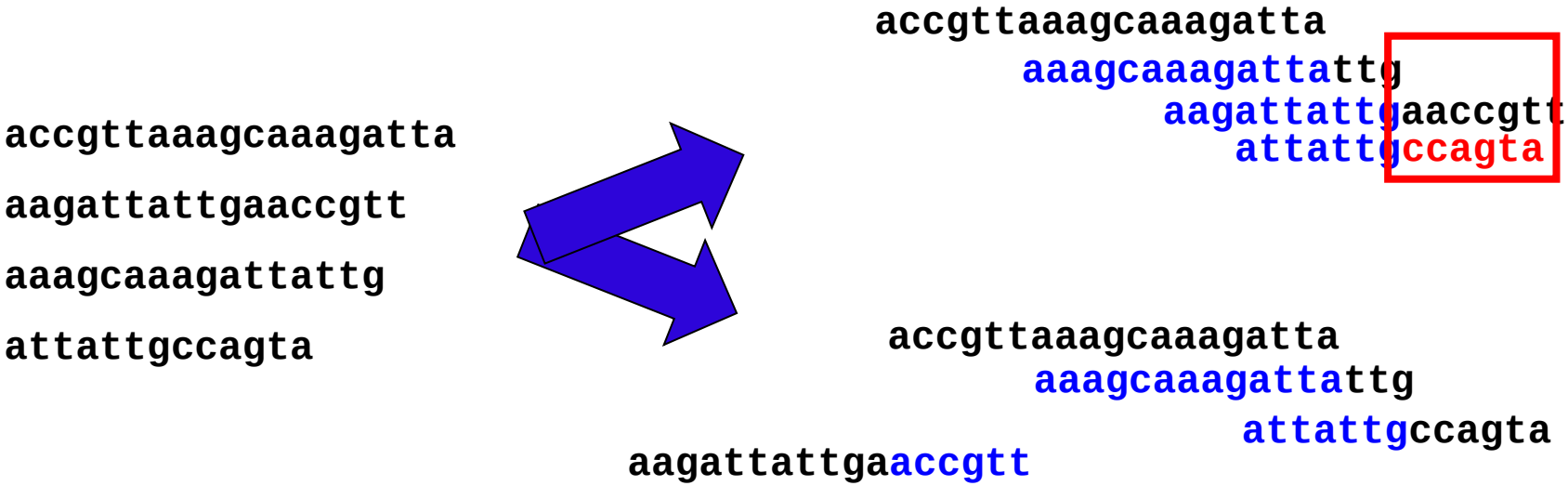
Problem

- 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?

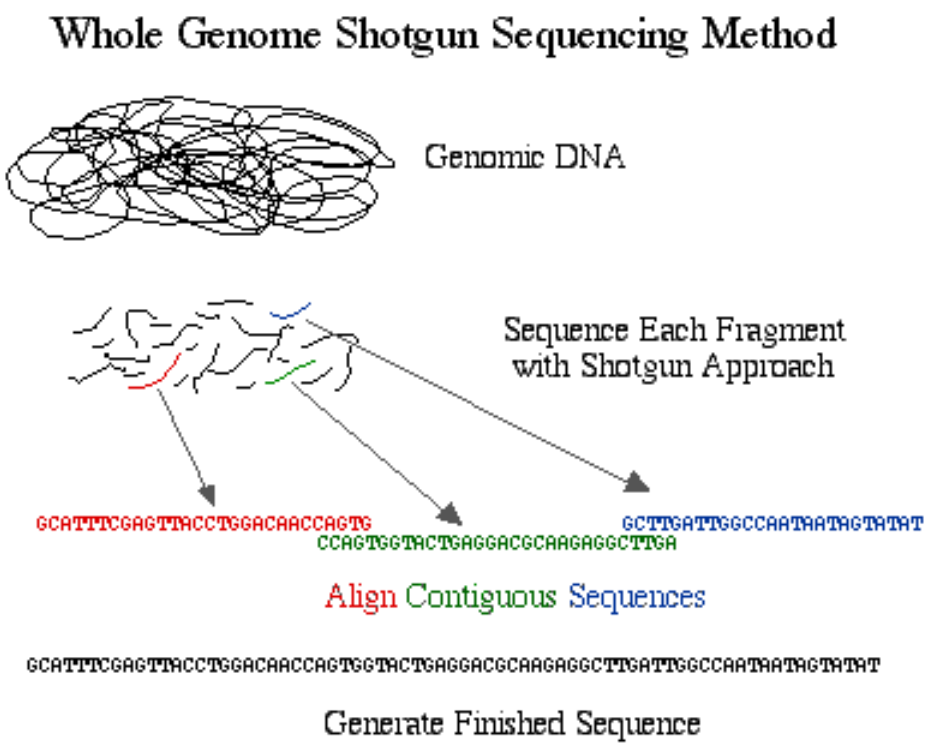
- Take one sequence and compute overlap with all others
- Keep the one with largest overlap and align
- Repeat such extensions until no more sequences are left
  - Note: This would work perfectly if all symbols of the chromosome were distinct



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

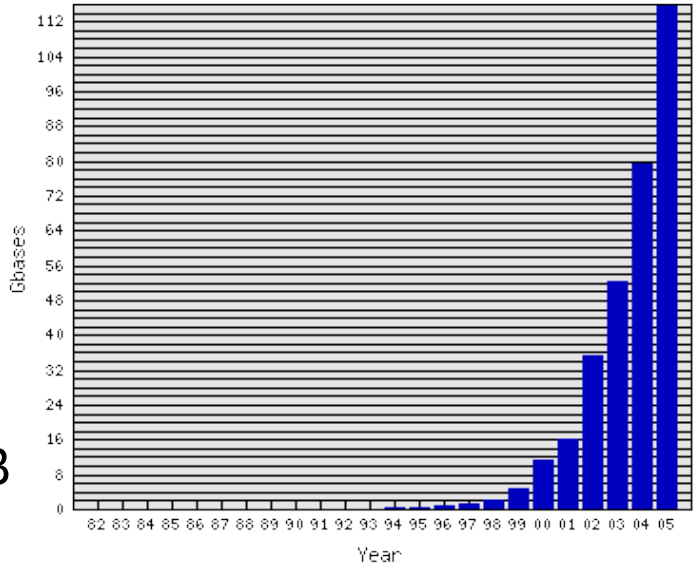
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
  - And an example
- Topics of this course

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

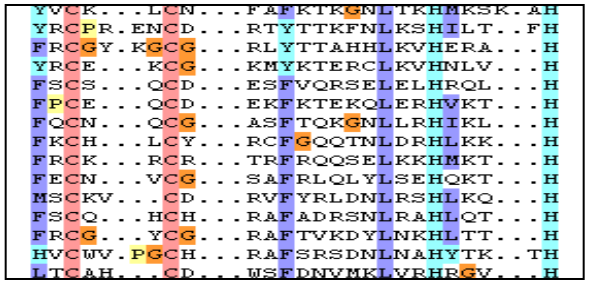
- Comparing two sequences is costly
- Given s, assume we want to find the most similar s' in a database of all known sequences
  - Naïve: Compare s with all strings in DB
  - Will take years and years

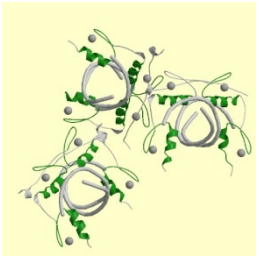


- BLAST: Basic local alignment search tool
  - Ranks all strings in DB according to similarity to s
  - Similarity: High is s, s' contain substrings that are highly similar
  - Heuristic: Might miss certain similar sequences
  - Extremely popular: You can "blast a sequence"

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





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?

Read Mapping and Variant Calling

- Identify (single nucleotide) variants in the output of next generation sequencing techniques
  - Taking uncertainty into account
- Characterize identified variants

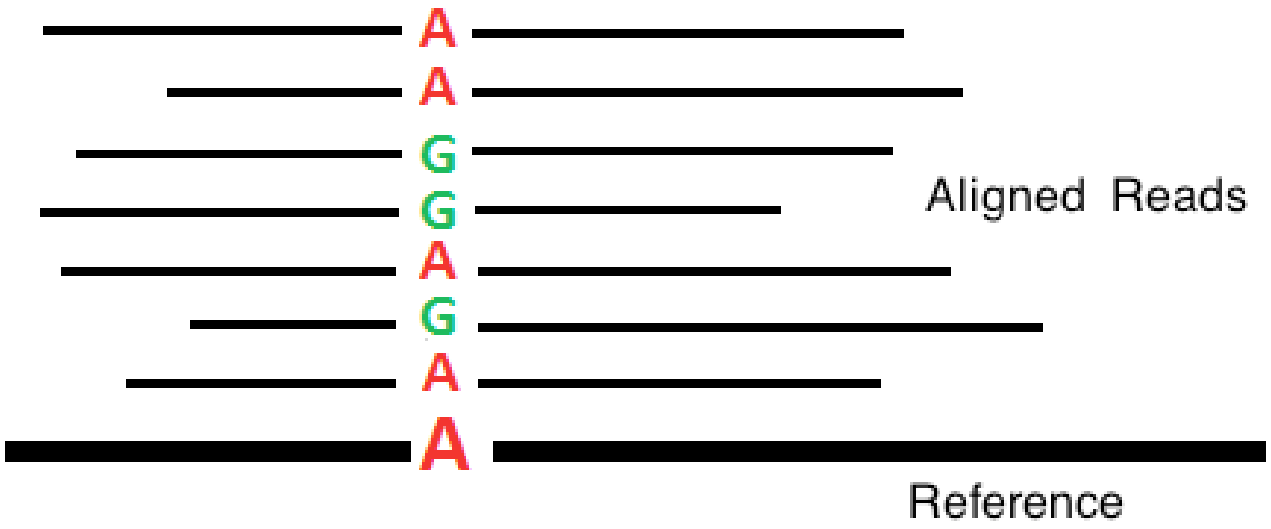
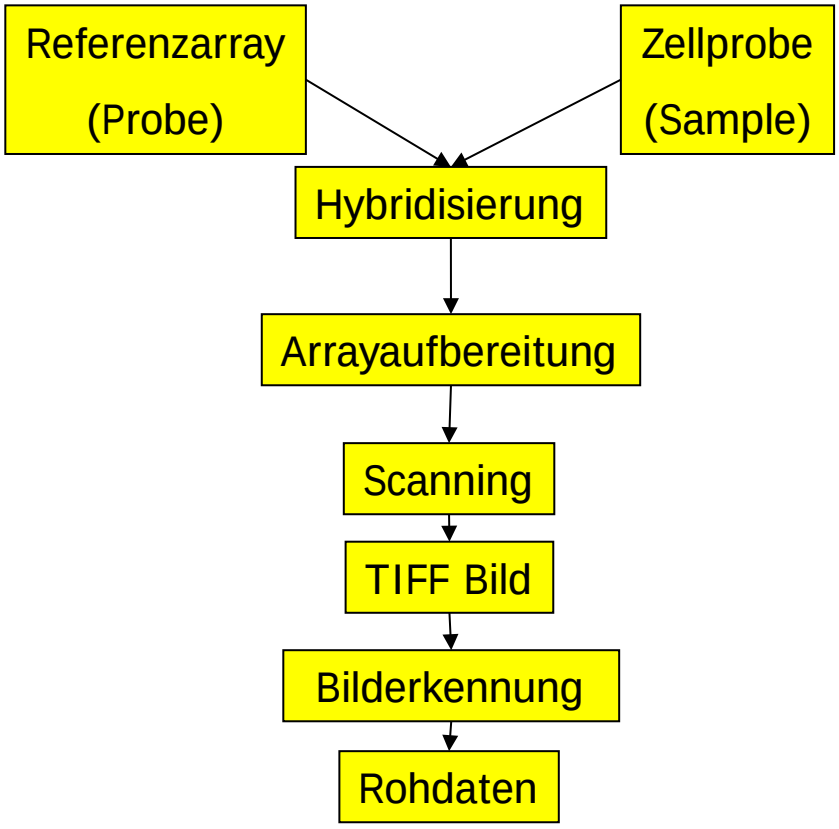
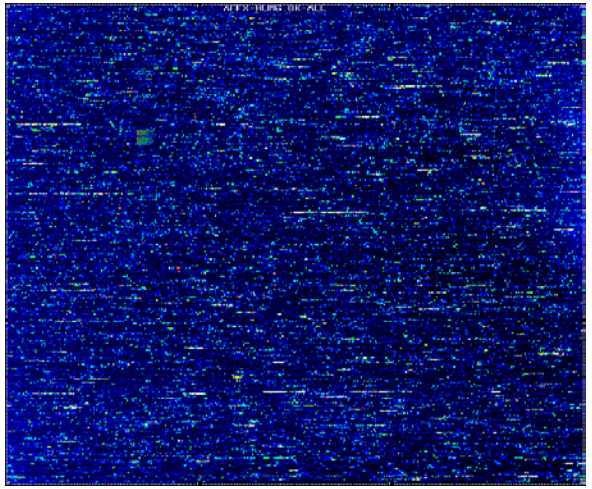


Image Source: Wikipedia

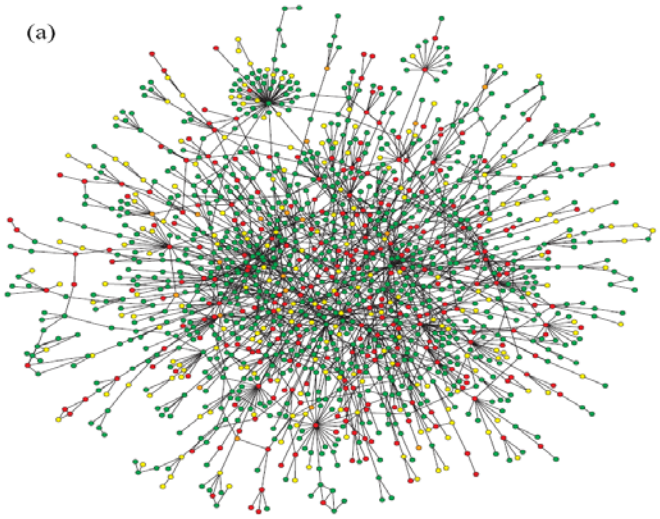
Microarrays / Transcriptomics





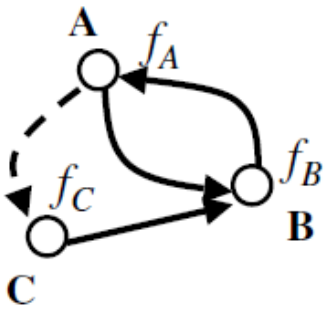
Protein-Protein-Interactions

- Proteins do not work in isolation but interact with each other
  - Metabolism, complex formation, signal transduction, transport, ...
- PPI networks
  - Neighbors tend to have similar functior
  - 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