

Introduction to Bioinformatics

Ulf Leser

Bioinformatics



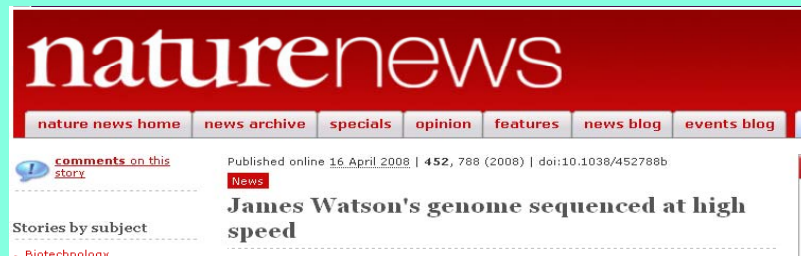
25.4.2003

50. Jubiläum der Entdeckung der Doppelhelix durch Watson/Crick



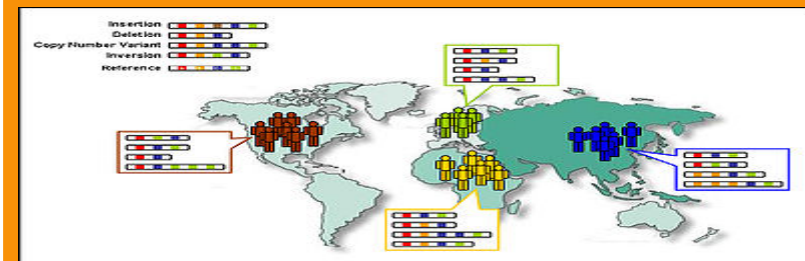
14.4.2003

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 countries
- Distributed infrastructure
- First federated genome project [HAA+08]

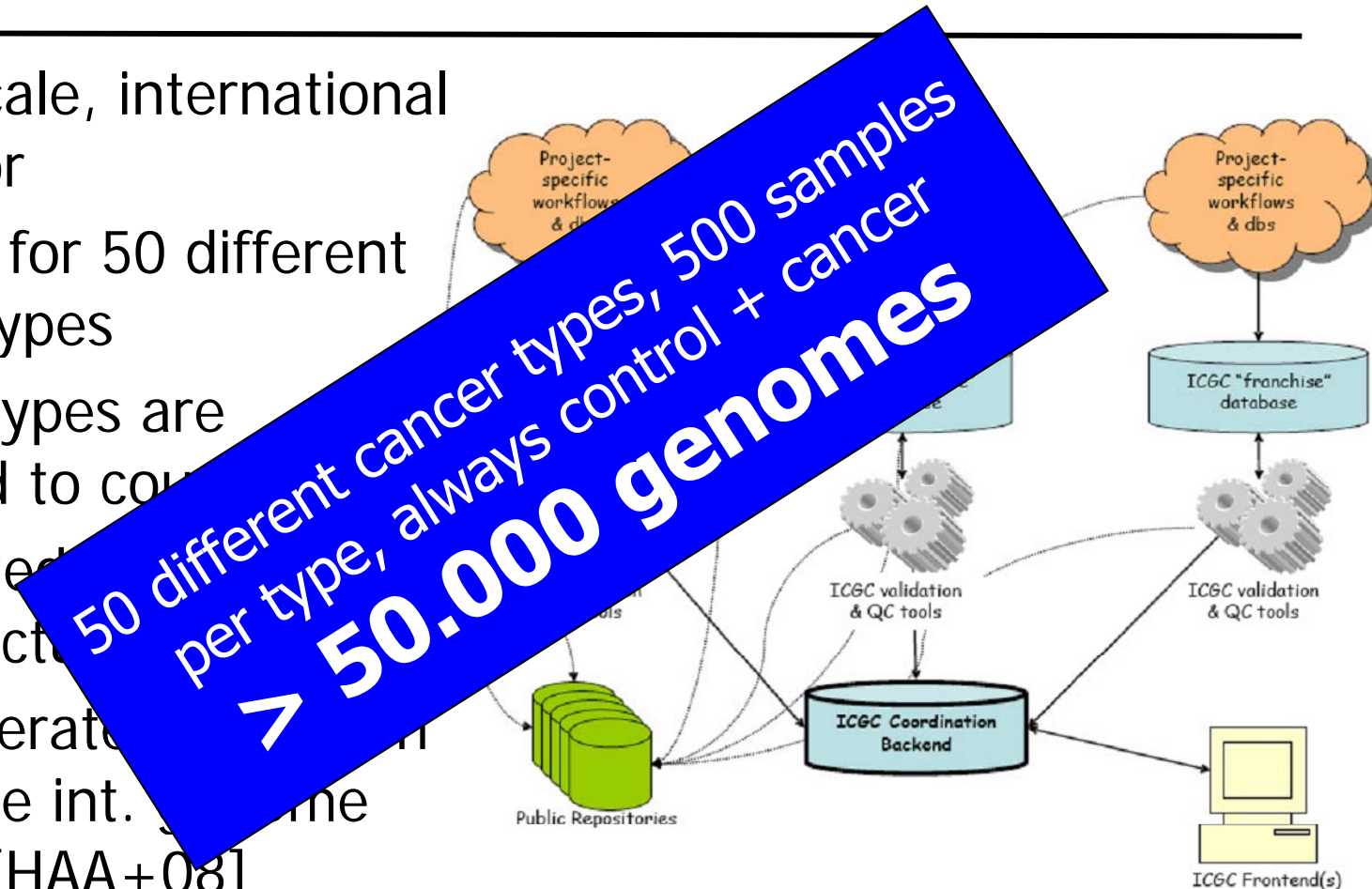


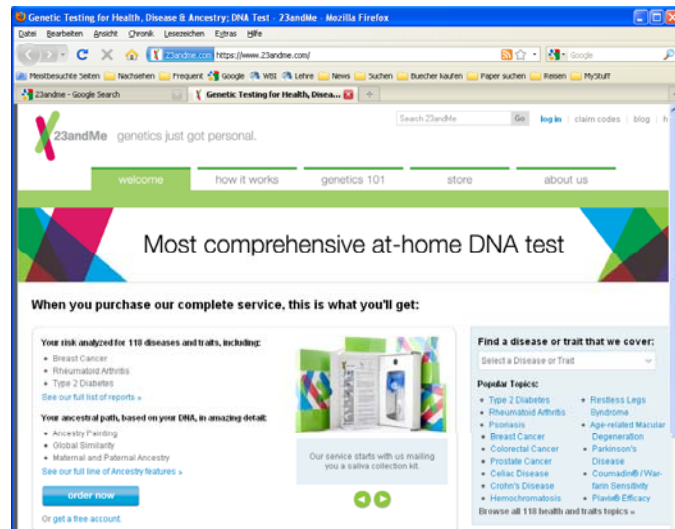
Figure 2: ICGC data coordination as a franchise system

Things you can do with it

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

Indikation*	Anbieter**	Untersuchungsgegenstand	Preis (inkl. MwSt.)
Alkoholverträglichkeit	2	keine Angaben (k. A.)	207,79 €
Alzheimer	2	k. A.	134,06 €
Alkoholismus	1	E4-Allel des Apolipoprotein-E-Gens auf Chromosom 10	650,00 €
Angelman-Syndrom ²¹	1	Deletion auf dem Chromosom 15	850,00 €
Anti-Aging-Risikoprofil	2	k. A.	653,61 €
Arteriosklerose/Herzinfarkt/Schlaganfall	2	k. A.	512,81 €
Aspermie	1	31 Mutationen einschließlich einer ST-Variante auf dem CFTR-Gen auf dem Chromosom 7	850,00 €
Bluthochdruck	2	k. A.	127,40 € 439,24 €
Diabetes Typ II	2	k. A.	127,40 € 194,39 €
Dickdarmkrebs ²¹	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 ²⁰	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 ⁴¹	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 ²⁰	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-Mutationen	1	Austausch der DNS-Basen Guanin zu Adenin an der Position 20210 des Prothrombingens auf dem Chromosom 11	550,00 €
Risiko Alkohol- und Drogenabhängigkeit	2	k. A.	274,86 €
Thrombose	2	k. A.	134,06 € 281,52 €

State of the "Art"



- 6/2010: „Gentest-Firma vertauscht DNA-Ergebnisse ihrer Kunden“ (Nature Blog)
- 7/2010: US general accounting office compared 15 (4) companies: totally **contradicting results**

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 not assume much knowledge in [computer science](#)
- Does not assume knowledge in [biology](#)
- Will not teach programming – you need to know it already
- Is introductory – many topics, often not much depth
 - Visit “Algorithmische Bioinformatik” afterwards ...
- Ask questions! [Leser \(a\) informatik.hu ... berlin...](#)

Exercises

- Taught by Samira Jaeger
- Registration through Goya
- There will be 6 assignments
- We build teams of 2 students
- 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

- Examination will be oral
- 20-30 minutes per student
- Days will be set in June

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

The image displays two overlapping screenshots of a Mozilla Firefox browser window. The top window shows the main page of 'Grundlagen der Bioinformatik' for the Wintersemester 2010/2011, featuring a blue header with the WBI logo and a sidebar with navigation links. The bottom window shows a specific exercise page titled 'Übung zu Grundlagen der Bioinformatik', which includes a list of tasks and a sidebar with a detailed navigation menu. Both windows show the browser's address bar, menu bar, and toolbar.

Grundlagen der Bioinformatik — Mozilla Firefox

http://zope.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1011/vl_bioinf/

Meistbesuchte Seiten | Nachsehen | Frequent | Google | WBI | Lehre | News | Projekte | Buecher kaufen | Paper suchen | Reisen | MyStuff | Paris

Lehrangebot Wintersemester 2010/2011... | Grundlagen der Bioinformatik — | Übung zu Grundlagen der Bioinformatik...

WBI

WS 10/11

Grundlagen der Bioinformatik

Vorlesung im Wintersemester 2010/2011
Professor Ulf Leser

Die Vorlesung behandelt **grundlegende Fragestellungen** die notwendige Grundkenntnisse in der Molekularbiologie, die Themen der Bioinformatik, wie Sequenzierung von Genen, Sequenzen, Messung und Interpretation von Genexpressionsdaten, von Protein-Protein-Interaktionsnetzen etc. Sie ist für alle Themen nur ein.

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

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

Voraussetzungen

Voraussetzung für den Besuch sind grundlegende Kenntnisse in Java.

Prüfungen

Prüfungen sind mündlich.

Anrechnung

Der Kurs (Vorlesung + Praktikum) kann angerechnet werden auf:

- Bachelor Informatik, Wahlpflichtbereich, drittes Semester
- Bachelor Biophysik, Pflichtvorlesung im Modul Bioinformatik

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

Übung zu Grundlagen der Bioinformatik — Mozilla Firefox

http://zope.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1011/ue_bioinf/

Meistbesuchte Seiten | Nachsehen | Frequent | Google | WBI | Lehre | News | Projekte | Buecher kaufen | Paper suchen | Reisen | MyStuff | Paris

Lehrangebot Wintersemester 2010/2011... | Übung zu Grundlagen der Bioinformatik...

WBI

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 Neimplementierung einfacher Verfahren als auch die Verwendung existierender Tools.

Die Arbeit erfolgt in Gruppen zu zwei Studierenden. Jede Gruppe muss alle Aufgaben erfolgreich bearbeiten (aber nicht immer komplett). Die Aufgaben werden an einem Übungstermin ausgeben, 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-Universität 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

Edit page

Suchen: Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung Fertig

Ihr Feedback

1	Alter	Geschlecht	Geführt	Teilnehmerzahl	Warum kommen?	Fachsemester	Freundlich	Fragen	Sprache	Präsentation	Beispiele	Gliederung	Überblick	Neue Erkenntnisse	Kritische Auseinandersetzung	Nützlich	Lernziele	Materialien	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	
2	22	W	0	3	1	7	5,00	5,00	6,00	5,00	3,00	4,00	5,00	6,00	4,00	5,00		5,00	3,00	3,00	3,00	2,00	2,00	
3	22	W	0	3	1	5	5,00		5,00	5,00	5,00	6,00	6,00	5,00	5,00	4,00	5,00	5,00	3,00	3,00	5,00	2,00	2,00	
4	22	W	0	4	1	5	5,00	5,00	6,00	6,00	6,00	6,00	6,00	6,00	5,00			5,00	3,00	3,00	4,00	1,00	2,00	
5	22	W	1	3	1,2	5	5,00	5,00	5,00	4,00	5,00	5,00	4,00	5,00	4,00	4,00	5,00	5,00	4,00	3,00	3,00	2,00	2,00	
6	22		0	3	1,2	5	6,00	6,00	6,00	6,00	5,00	6,00	6,00	4,00	5,00	5,00	6,00		3,00	3,00	3,00	1,00	1,00	
7	35	M	0	3	1,3	15	6,00	6,00	6,00	4,00	2,00	4,00	5,00	6,00	5,00	5,00		5,00	5,00	4,00	4,00	1,00	3,00	
8	22	M	0	4	1	5	6,00	6,00	6,00	6,00	6,00	5,00	6,00	6,00	5,00	6,00	6,00		3,00	4,00	3,00	1,00	2,00	
9	22	W	0	3	1,2	5	6,00	5,00	6,00	6,00	6,00	6,00	6,00	5,00	4,00	4,00	5,00	6,00	3,00	3,00	3,00	1,00	1,00	
10	25	M	2	3	1,2,3	5	6,00	6,00	6,00	5,00	4,00	6,00	5,00	5,00	6,00	4,00	5,00	5,00	3,00	3,00	4,00	1,00	2,00	
11	23	M	2	2	1,2	5	6,00	6,00	6,00	5,00	4,00	6,00	6,00	6,00	4,00	4,00		5,00	3,00	3,00	4,00	2,00	2,00	
12	20	M	0	3	1,2	5	6,00	6,00	5,00	5,00	5,00	6,00	6,00	6,00	6,00	5,00	6,00	5,00	3,00	3,00	4,00	1,00	1,00	
13	21	W	1	3	1,2	5	5,00	5,00	5,00	6,00	5,00	4,00	4,00	5,00	5,00	5,00	5,00	4,00	3,00	3,00	3,00	2,00	2,00	
14	23,2		0,5	3,1	1,0	6,0	5,58	5,55	5,67	5,25	4,67	5,33	5,42	5,42	4,83	4,64	5,38	5,00	3,25	3,17	3,58	1,42	1,83	
15				3,0			6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	3,00	3,00	3,00	1,00	1,00	
16				###			0,42	0,45	0,33	0,75	1,33	0,67	0,58	0,58	1,17	1,36	0,63	1,00	-0,25	-0,17	-0,58	-0,42	-0,83	
17							0,4	0,5	0,3	0,8	1,3	0,7	0,6	0,6	1,2	1,4	0,6	1,0	0,3	0,2	0,6	0,4	0,8	9,6

- Schlecht: Mehr Beispiele, nicht übermäßig nützlich
- Gut: Vieles (Freundlich, Fragen, Sprache, Tempo, Schwierigkeit, Dozent)

Was wir geändert haben

- Datenbanken für Übungen werden in Übungen kurz vorgestellt
- Erste Übung: Schnellkurs in praktischer Softwareentwicklung mit Java
 - Kein Java-Kurs!
- Geringe Verkürzung des Vorlesungsstoffs

Wissensmanagement in der Bioinformatik

- Who am I?
- Our topics in research
 - Management of biomedical data and knowledge
 - Scientific database systems
 - Text Mining
 - Biomedical data analysis
- Our topics in teaching
 - Algorithmische Bioinformatik
 - Text Analytics
 - Data Warehousing und Data Mining
 - Informationsintegration

Questions?

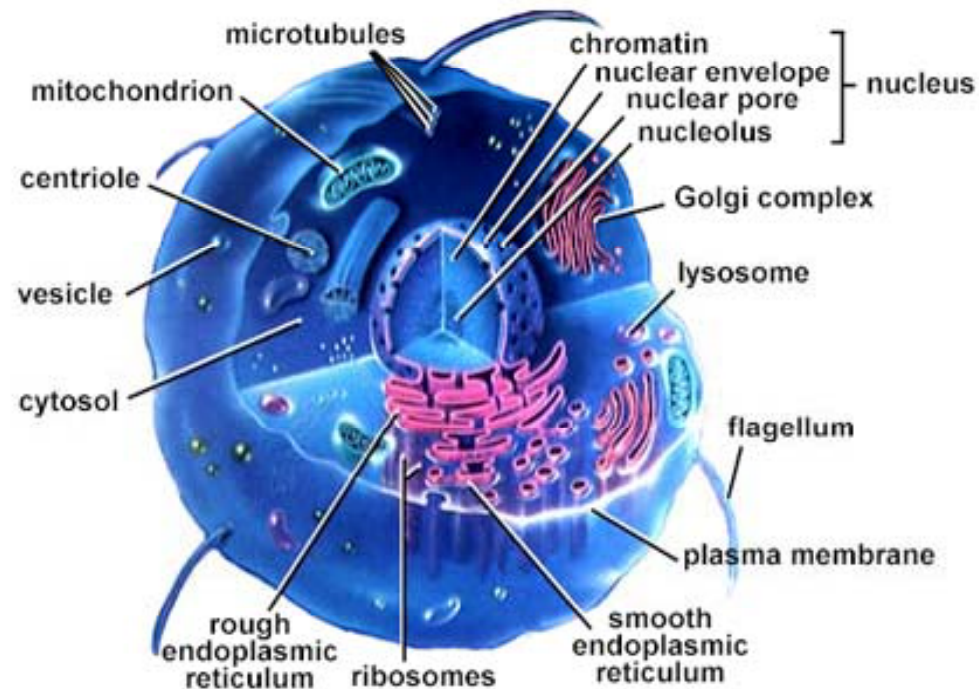
My Questions

- Diplominformatiker?
 - Bachelor Informatik?
 - Kombibachelor?
 - Biophysik?
 - Other?
-
- Semester?
 - Prüfung?
 - Spezielle Erwartungen?

This Lecture

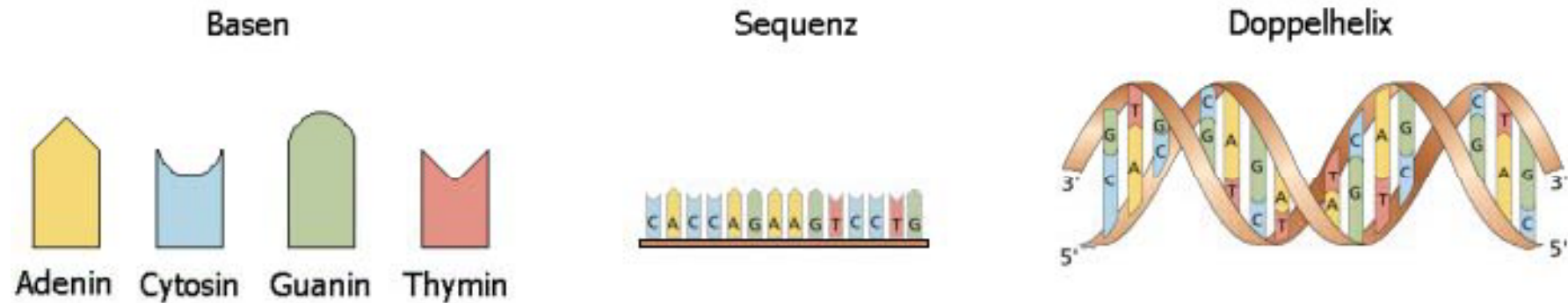
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
- Topics of this course

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*

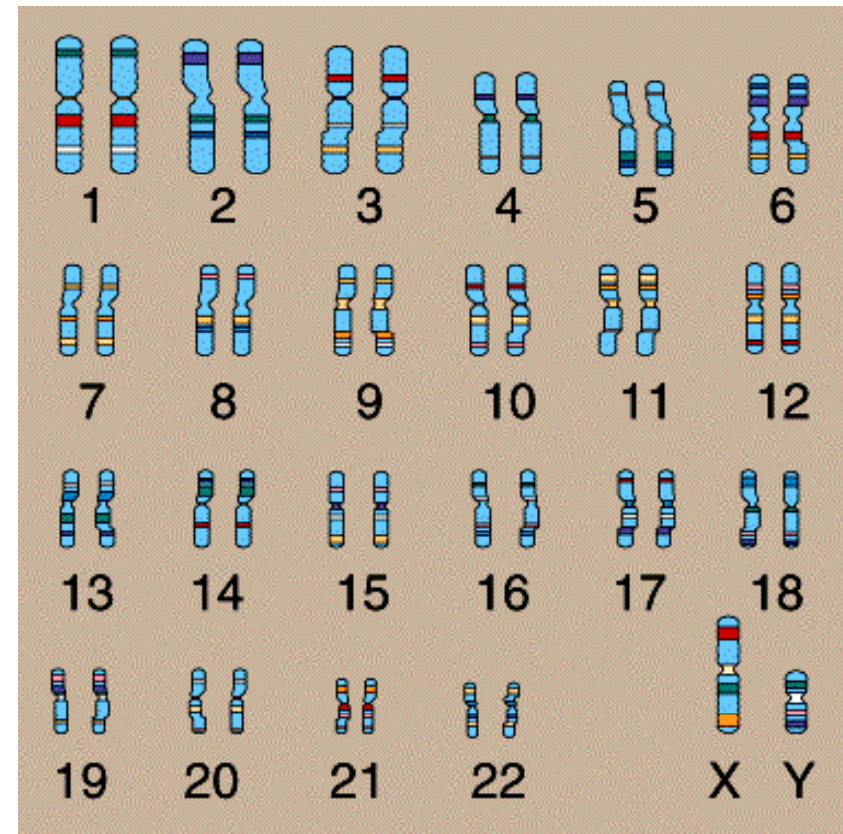
DesoxyriboNucleicAcid



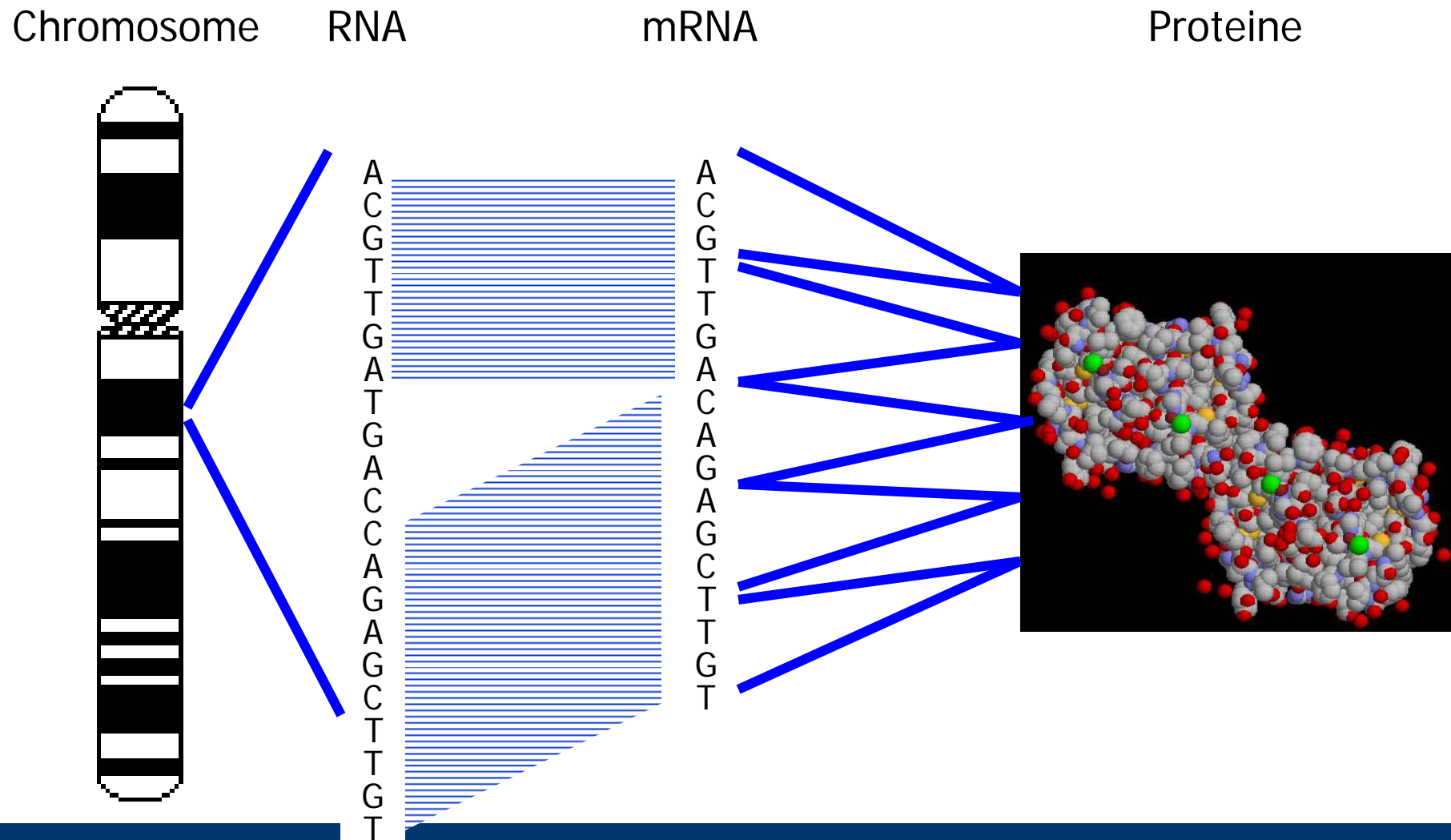
- DNA: Desoxyribonukleinsäure
 - Four different molecules (one swapped in RNA)
 - The DNA of all chromosomes in a cell together with the mitochondria-DNA forms its genome
 - Almost all cells in a (human) body carry almost the same genome
 - All living beings are based on DNA (or RNA) for proliferation

The Human Genome

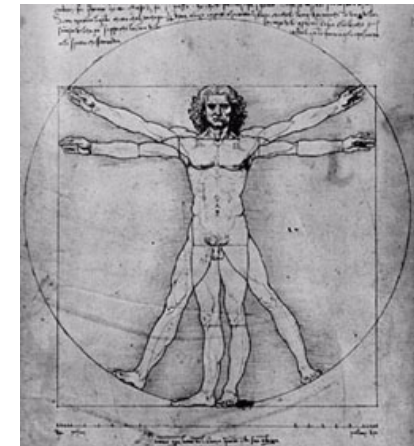
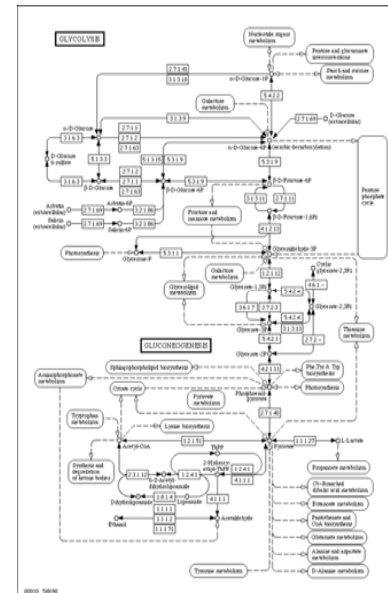
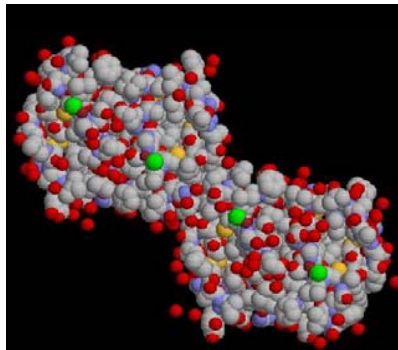
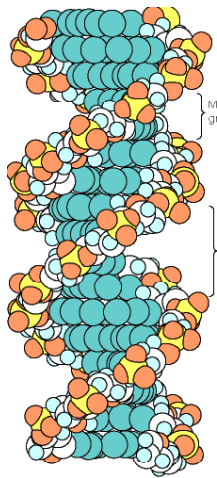
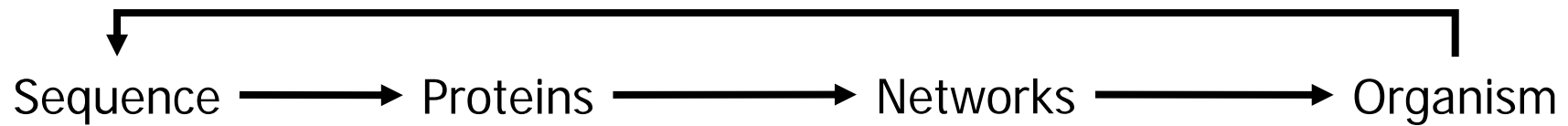
- 23 chromosomes
 - Most in pairs
- ~3.000.000.000 letters
- ~50% are repetitions of 4 identical subsequences
 - ~~~100.000 genes~~
 - ~~~56.000 genes~~
 - ~~~30.000 genes~~
 - ~~~24.000 genes~~
- ~20.000 genes



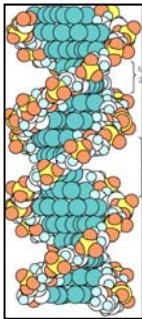
(Protein-Coding) Genes



Proliferation

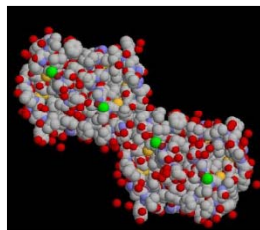


Computer Science in Molecular Biology / Medicine



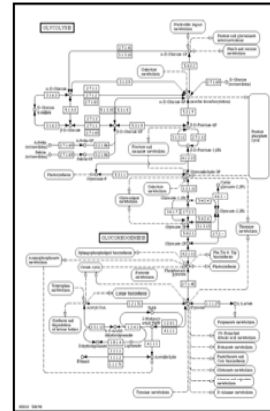
Genomics

Sequencing
Gene prediction
Evolutionary relationships
Motifs - TFBS
Transcriptomics
RNA folding



Proteomics

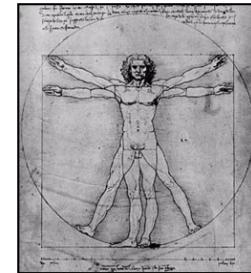
Structure prediction
... comparison
Motives, active sites
Docking
Protein-Protein Interaction
Proteomics



Systems Biology

Pathway analysis
Gene regulation
Signaling
Metabolism
Quantitative models
Integrative analysis

...

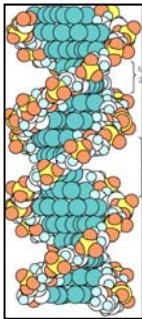


Medicine

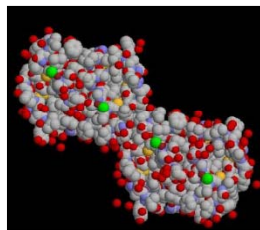
Phenotype – genotype
Mutations and risk
Population genetics
Adverse effects

...

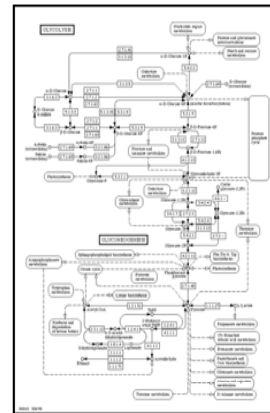
This Lecture



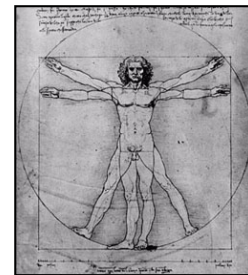
Genomics
Sequencing
Gene prediction
Evolutionary relationships
Motifs - TFBS
Transcriptomics
RNA folding



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

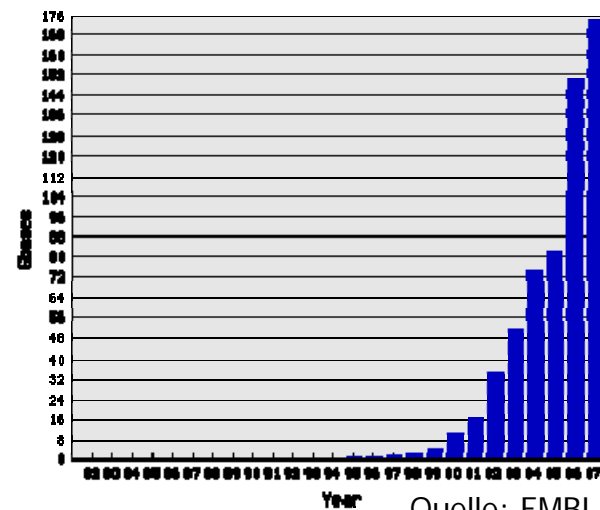
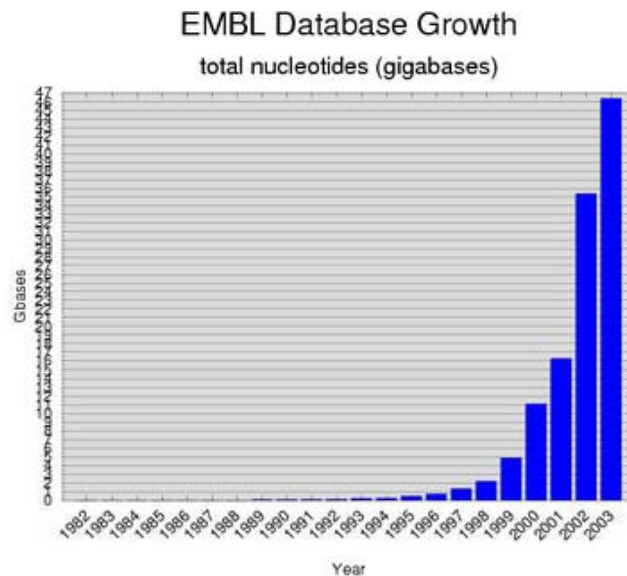
- Formal stuff on the course
- A very short introduction in Molecular Biology
- What is Bioinformatics?
 - And an example
- Topics of this course

Bioinformatics / Computational Biology

- Computer Science methods for
 - Solving biologically relevant problems
 - Analyzing and managing experimental data sets
- **Empirical**: Data from high throughput experiments
- Mostly focused on developing algorithms
- 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

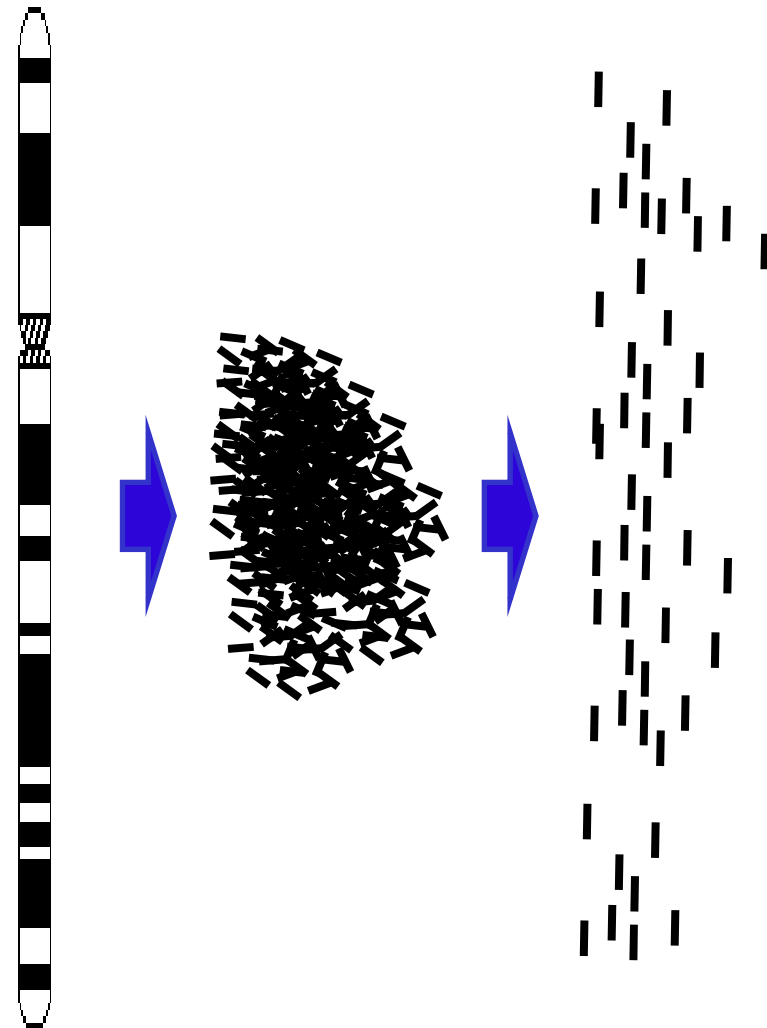
- First papers on sequence alignment
 - Needleman-Wunsch 1970, Gibbs 1970, Smith-Waterman 1981, Altschul et al. 1990
- Large impact of the [Human Genome Projekt](#) (~1990)
- Only 14 mentions of „Bioinformatics“ before 1995
- „Journal of Computational Biology“ since 1994

- First [professorships](#) in Germany: end of 90th
- First university programs: ~2000
- First German book: 2001

- Commercial hype: 1999 – 2004

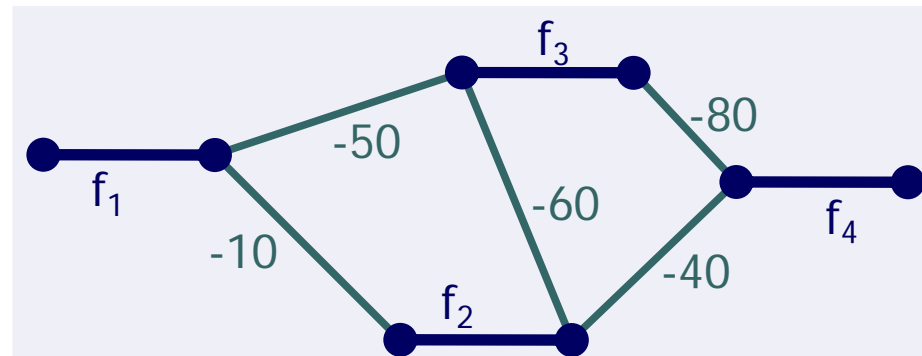
A Concrete Example: Sequencing a Genome

- 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



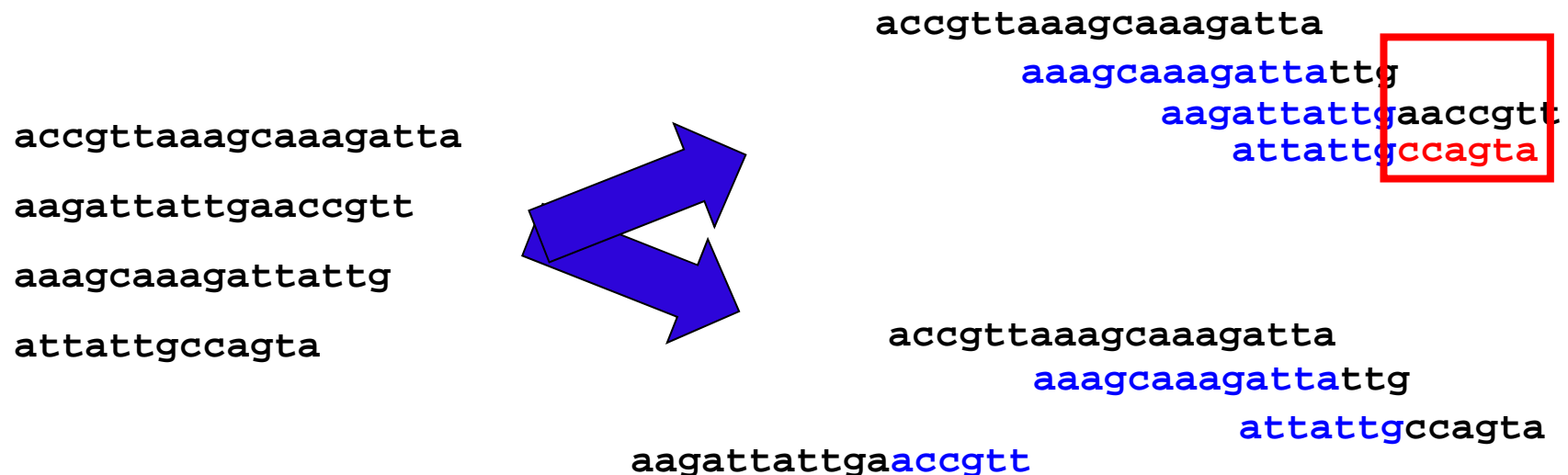
Problem

- 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

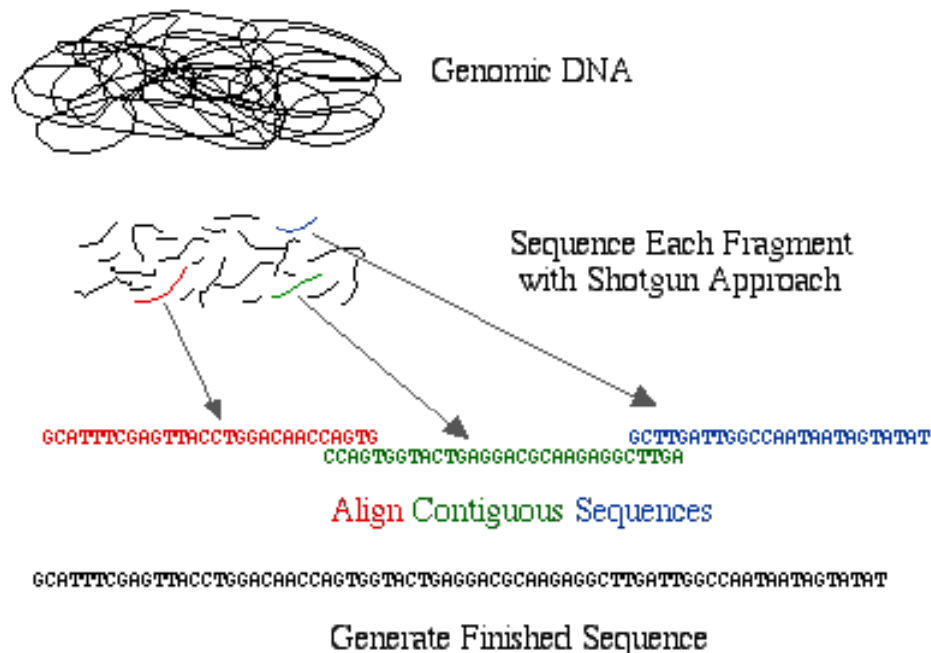


Abstract Formulation

- **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 is minimal)
- Problem is **NP-complete**
 - Very likely, there is no algorithm that solves the problem in less than $k_1 \cdot 2^n$ operations, where k_1, k_2 are constants and $n = |S|$
 - We need exponentially many operations
 - Exact solution not computable for $n > 20$...
- Bioinformatics: Find clever **heuristics**
 - Solve the problem “good enough”
 - Finish in reasonable time

Dimension

Whole Genome Shotgun Sequencing Method



- 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

Plan

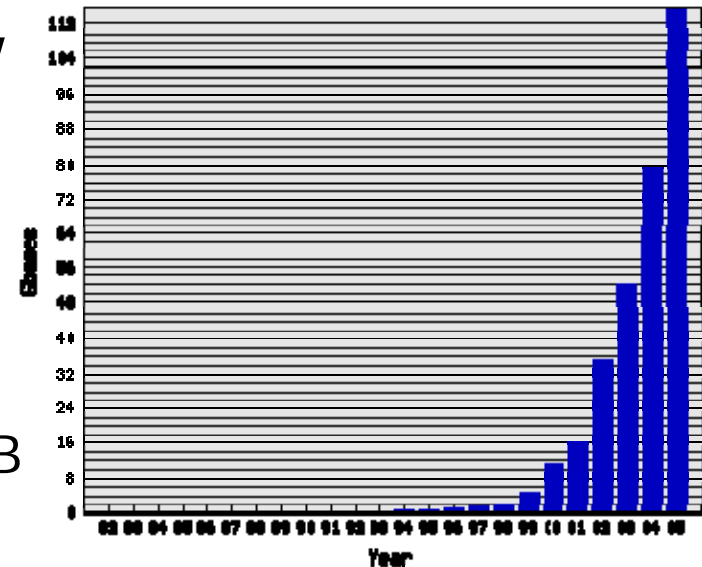
- 22.04.2011: Bank Holiday
- 29.04.2011: Björn: Substringssuche
- 06/13.05.2011 Ulf: Editabstand,
- 20/27.05.2011 Karin/Philippe: Genexpression
- 03/10.06.2011 Ulf: PAM, BLAST, MSA
- 17.06.2011 Ulf: Proteinsekundärstruktur
- 24.06.2011 Samira: Protein-Protein-Interaktionnetze
- 01.07.2011 Ulf: Proteomics
- 08.07.2011 Michael: Metabolische Netzwerke
- 15.07.2011 Ulf: Abschluss, Reserve

Searching Sequences (Strings)

- A chromosome is a string
- Substrings may represent **biologically important areas**
 - Genes on a chromosome
 - Transcription factor binding sites
 - Overlapping sequences in assembly
 - Same gene in a different species
 - Similar gene in a different species
 - ...
- Exact or **approximate string search**
 - Naive and Boyer-Moore algorithm
 - Approximate gap-free matching

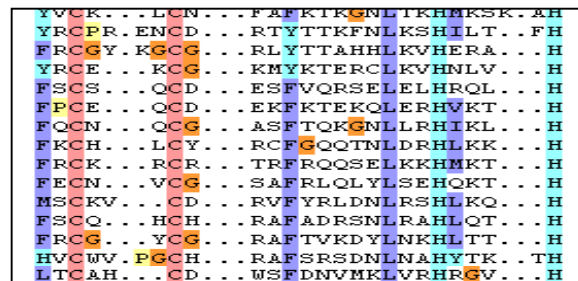
Searching a Database of Strings

- Comparing two sequences is costly
 - $s, t: O(|s| * |t|)$
- 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 if s, s' contain substrings that are highly similar
 - Heuristic: Might **miss certain similar sequences**
 - Extremely popular: You can “blast a sequence”

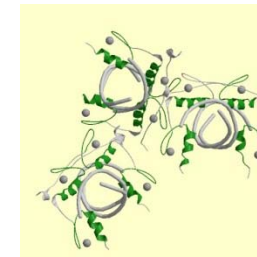


Multiple Sequence Alignment

- 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



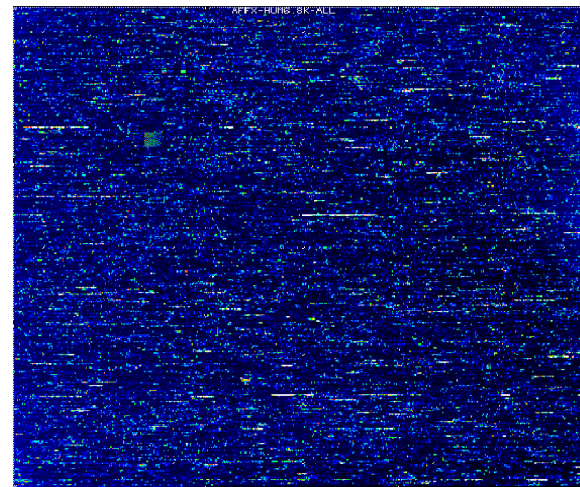
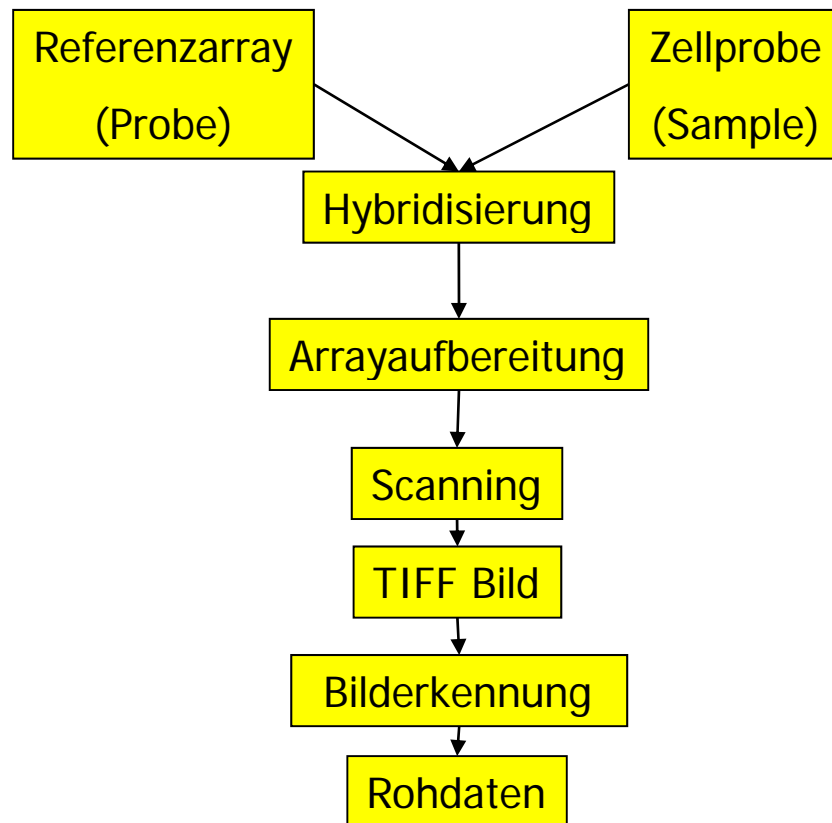
```
YVCK...LCN...FAFKIK...NLTKHMKSK..AH
YRCPR..ENCD...RTYTTKFNLKSHILT..FH
FRGGY..KCG...RLYTTAHLKVHERA...H
YRCE...KCG...KMYKTERCLKVHNLV...H
FSCS...QCD...ESFVQSELELHRQL...H
FPCE...QCD...EKFKTEKQLERHVKT...H
FQCN...QCG...ASFTQKGNLLRHIKL...H
FKCH...LCY...RCFGQQTNLDRHLKK...H
FRCK...RCR...TRFRQSELEKHHMKT...H
FECN...VCG...SAFRLQLYLSEHQKT...H
MSCKV...CD...RVFYRLDNLRSHLKQ...H
FSCQ...HCH...RAFADRSNLRHLQ...H
FRCG...YCG...RAFTVKDYLNKHLTT...H
HVCWV..PGCH...RAFSRSDNLNAHYTK..TH
LTCAH...CD...WSFDNVMKLVRRHGV...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?

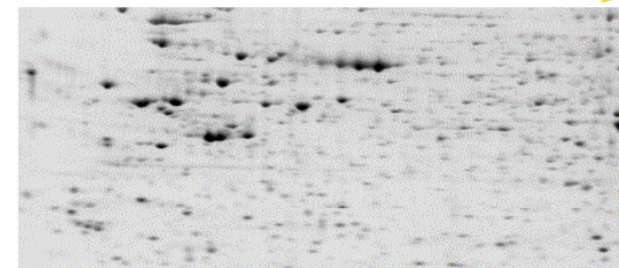
Microarrays / Transcriptomics



Proteomics

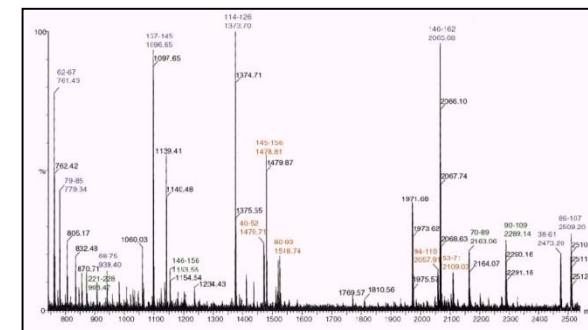
- The real workhorses in a cell are proteins
 - Differential splicing, post-translational modifications, degradation rates, various levels of regulation, ...
- But: Much more difficult to study (compared to mRNA)

- Separation of proteins
 - 2D page, GC / LC



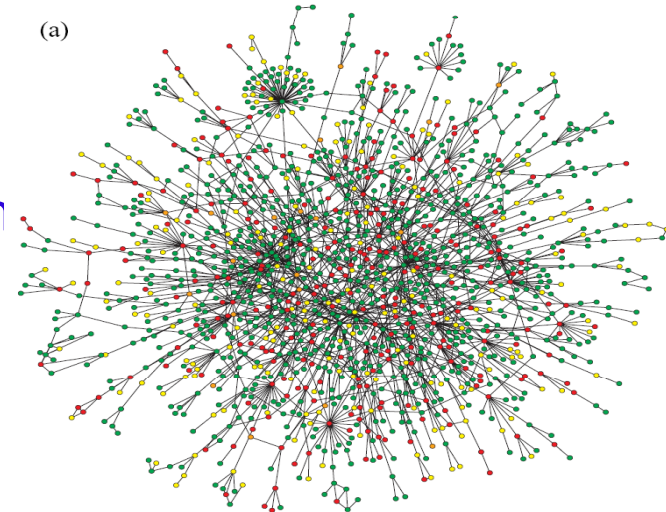
Drug Discovery Today

- Identification of proteins
 - Mass-spectrometry



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



Systems Biology

- Biological networks are more than edges and nodes
- Example: Metabolic networks
 - Graphs capturing **biochemical reactions**
 - Can be **described quantitatively**:

$$\text{N}_2 + 3\text{H}_2 \rightarrow 2\text{NH}_3$$
 - Can be analyzed (Balanced flux? Elements never produced / consumed? ...)
- Dynamic modeling: Kinetics

