Algorithmische Bioinformatik

Abschluss

Ulf Leser

Wissensmanagement in der Bioinformatik

- Themen des Semesters
- Ihr Feedback
- Prüfung
- Werbung
- DNA Computing

Zusammenfassung 1

- Exakte Suche
  - Z-Box: Geschachtelten Z-Boxen: O(n+m)
  - Boyer-Moore: Falschrum suchen, Bad-Character und Good-Suffix Rule; Worst Case O(m*n) oder O(m+n), Average Case sublinear
  - Knuth-Morris-Pratt: Naiver Algorithmus ++, Shift Regeln
- Mehrere Strings
  - Keyword-Trees / Aho-Corasick: KMP weitergedacht, einfache Konstruktion, lineare Suche erfordert komplexe Hilfskonstrukte: Failure- und Output Links
  - Suffix-Bäume: Einfache Suche, komplexe Konstruktion mit Ukkonen: Extensionsregeln, Suffix-Links und Skip-Count Trick
  - Suffixarrays: Platz sparend und schnell, Konstruktion in O(m*log(m))

Zusammenfassung 2

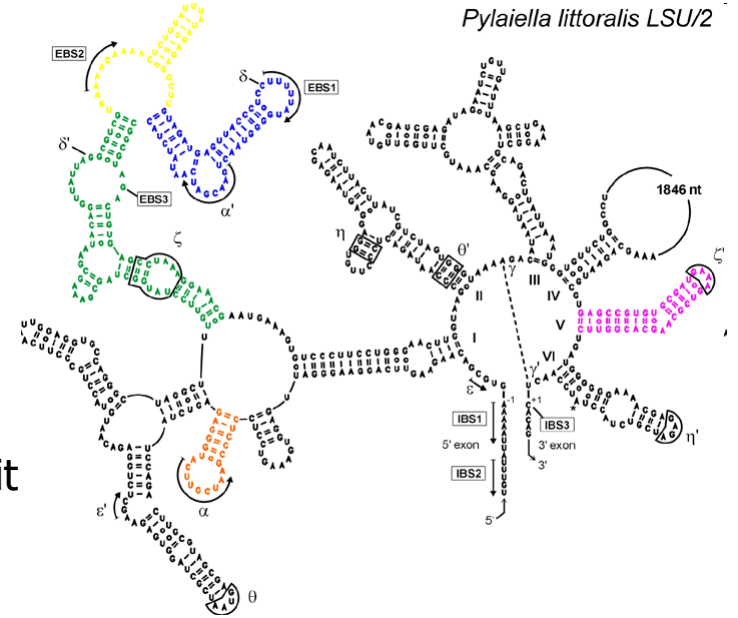
- Approximative Suche
  - Dotplots, Alignment, Ähnlichkeit, Editabstand, Editgraphen
  - Dynamische Programmierung
  - Globales, lokales, gapped, End-Free Alignment
  - Alignment in linearem Platz, k-Banded Alignment
- Das Reich der Heuristiken
  - Substitutionsmatrizen PAM und Blossum
  - Datenbanksuche: BLAST und BLAST2, BLAT
- Statistische Sequenzanalyse: HMMs
  - Markov Ketten und CpG Inseln
  - Viele Zustände mit gleicher Ausgabe: Hidden Markov Modelle. Dekodierung (Viterbi), Evaluation (Forward), Lernen (Baum-Welch)

Zusammenfassung 3

- Multiple Sequenz Alignments
  - Sum-Of-Pairs – komplex und an der Biologie vorbei
  - CLUSTAL W – Guide Trees und progressives MSA
  - Suche mit einem MSA: Profilalignment, RegExp, Profil-HMM
- Phylogenie
  - Ultrametriken – wenn die Welt sehr einfach wäre
  - Additive Bäume – wenn die Welt einfach wäre
  - Perfect Phylogeny – wenn die Welt schön wäre
  - Maximum Parsimony – groß und klein

Was wir nicht gemacht haben

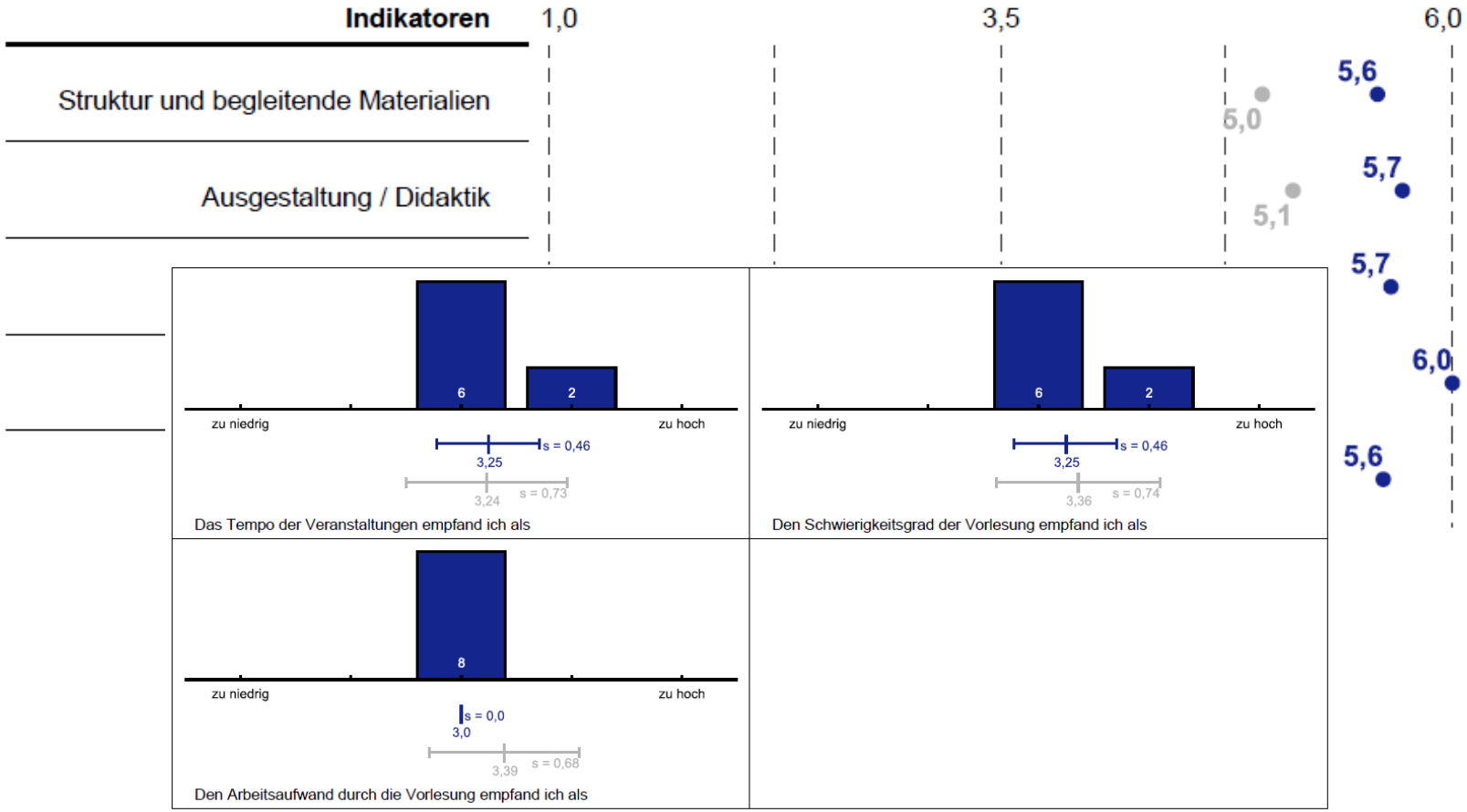
- String Matching: Shift-And, Karb-Rabin, …
- Genome Mapping (QT-Bäume …)
- Rearrangements: Sort by reversal
- Lokales MSA, Branch&Bound, T-Coffee
- Indexierung von Sequenzen; metrische Suchbäume
- Maximum Likelihood Phylogenie, Konsensus-Bäume
- Phylogenetische Netzwerke
- RNA-Struktur
- Read Mapping
- Probabilistische Methoden in der Phylogenie
- Proteine: 2D, 3D, Docking, Strukturähnlichkeit
- Genome Graphs
- Genome Assembly
- …
- Machine Learning in Bioinformatics



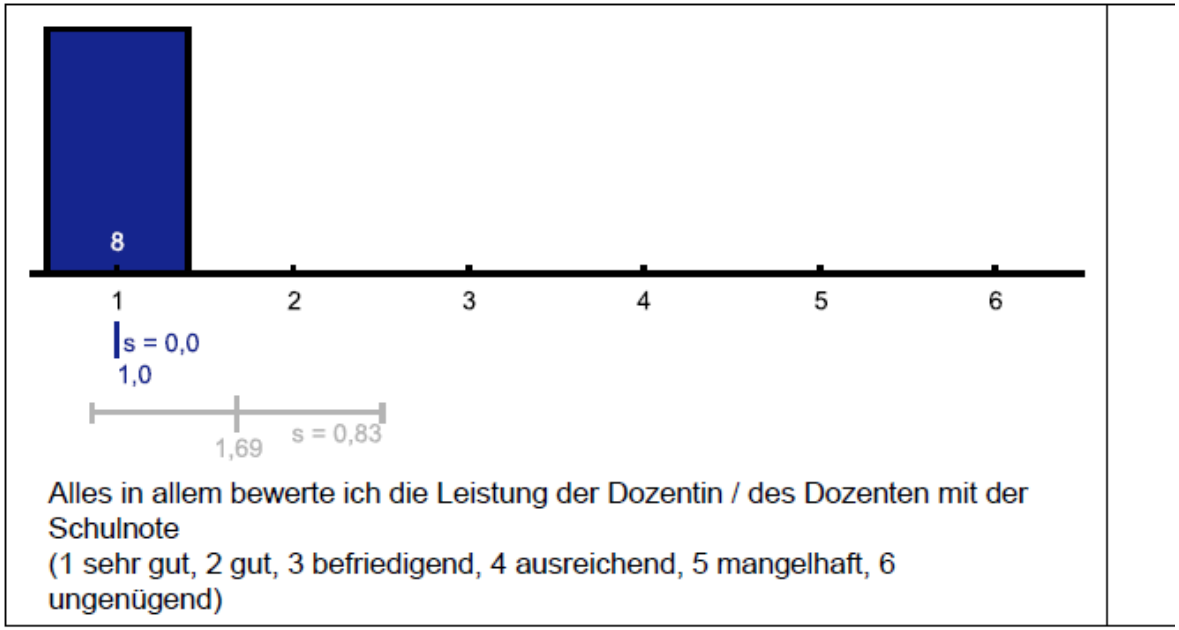
Quelle: http://www.cgm.cnrs-gif.fr/

- Themen des Semesters
- Ihr Feedback
- Prüfung
- Werbung
- DNA Computing

Bewertung (n=8)

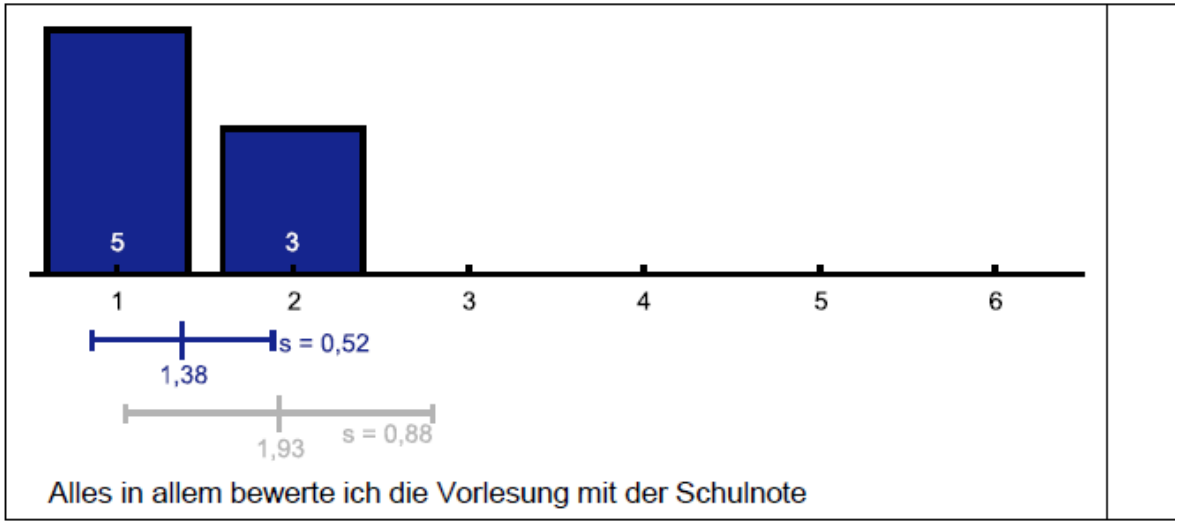


Gesamt



Alles in allem bewerte ich die Leistung der Dozentin / des Dozenten mit der Schulnote
(1 sehr gut, 2 gut, 3 befriedigend, 4 ausreichend, 5 mangelhaft, 6 ungenügend)

Gesamteindruck -Vorlesung



Alles in allem bewerte ich die Vorlesung mit der Schulnote

Positiv

Das hat mir gefallen: (DozentInnenfrage)

- Alles außer das was ich bei der anderen frage genannt habe.
- Der stetige Bezug zu realen Aufgabenstellungen, den biologischen Hintergründen und aktueller Forschung hat sehr zu meinem Interesse an d
  beigetragen.
- Die Beispiele bei Algorithme
  Die Erzählweise
  Die Unipolitik und neuen Erkenntnisse der Bioinformatik
- Die biologische Einführung.
  Die Struktur und Themenauswahl der Vorlesung.
  Lockere Atmosphäre.
- schrittweise Beispiele; Fragen kompetent beantwortet

Negativ

Welche konstruktiven Anregungen und Verbesserungsvorschläge haben Sie? (DozentInnenfrage)

- Die Foliensätze regelmäßiger hochladen - vielleicht auch schon vorher, dann kann man sich in den Folien Sachen markieren und Notizen m
  Den letzen Foliensatz auch Prüfungsrelevant machen, den fand ich sehr interessant. Ich weiß nur leider nicht, wie er heißt, da er noch nicht
- Die für Suffixarrays vorgestellten Algorithmen scheinen nicht mehr dem wissenschaftlichen Stand der Dinge zu entsprechen. Eine Erneuer
  wäre schön.
- Keine 9 uhr VL mehr. Die Folien sollten gleich nach der VL zur Verfügung stehen.
  Die Beispiele könnten etwas ausführlicher über mehr Folien gehen.
- recycelte Folien mit kleineren Fehlern; mehr Beispiele; deutlichere Unterschiede zum Bachelor Modul; konsequent bleiben in der Sprache

Was ich ändern will

- Read-Mapping mit BTW als Thema dazu
- O(n) Konstruktion von Suffixarrays
- Ultrametrische Bäume in $O(n^2)$
- Beweis Approximationsfaktor Center-Star

- Themen des Semesters
- Ihr Feedback
- Prüfung
- Werbung
- DNA Computing

Prüfung

- Ablauf
  - 30 Minuten
  - Sind Sie fit?
- Tiefe statt Breite
  - Verstehen, nicht auswendig lernen
- Erst denken, dann antworten
- Manches muss man schnell abrufen können
  - 1.0 braucht auch große Stoffabdeckung
- Nervosität
  - Einordnen: Es geht um ein 10/90 ihrer Gesamtnote
- Vergessen Sie die Biologie nicht

Werbung

- Studienprojekte und Masterarbeiten
  - Text Mining (in biomedizinischen Texten)
  - Verteilte Analyse sehr großer Datenmengen (BigData)
  - Statistische Analyse biomedizinischer Daten (Bioinformatik)
  - Diverse Datenbankthemen, insb. Indexierung
  - Ihr Lieblingsthema (wenn es halbwegs zum Profil passt)
- Oft interdisziplinär
  - Charite, MPI's, FUB, Linguisten
- Gerne in Kooperation mit Firmen
- Immer: Intensive Betreuung

Laufende Studien-, Bachelor-, Magister-, Master- und Diplomarbeiten

Diplom- / Magister/ Masterarbeiten

Anne-Kathrin Kruschke: Process Management in eHealth: Grundlagen, Anwendung, Regulierung
Diplomarbeit Informatik
Oktober 2018 - April 2019
Betreuung: Ulf Leser, Herbert Zech (Uni Basel)

Oleksiy Ostapenko: Learning to remember: Dynamic generative memory for continual learning (Expose)
Masterarbeit Informatik
November 2018 - April 2019
Betreuung: Patrick Jähnichen, Ulf Leser

Michel Manthey: Extraktion von Prozessmodellen aus Clinical Guidelines (Expose)
Diplomarbeit Informatik
Juli 2018 - Februar 2019
Betreuung: Ulf Leser, Johannes Starlinger (Charite)

Arne Binder: Training recursive compositional models with hierarchical linguistic information for semantic tasks in NLP (Expose)
Diplomarbeit Informatik
Juni 2018 - Januar 2019
Betreuung: Ulf Leser, tba

Studien-/ Bachelorarbeiten

Lukas Abegg: Neural Information Retrieval with IRGAN (Expose)
Studienprojekt Informatik
November 2018 - Februar 2019
Betreuung: Jurica Seva, Ulf Leser

Rafael Moczalla: Generator of Synthetic Time Series for Motif Discovery (Expose)
Studienprojekt Informatik
November 2018 - Februar 2019
Betreuung: Patrick Schäfer, Ulf Leser

Sebastian Biegel: Automatische Erkennung von epileptischen Anfällen im neonatalen EEG (Expose)
Bachelorarbeit Informatik
September 2018 - Januar 2019
Betreuung: Ulf Leser, Lars Matthäus (eemagine)

Igor Savin: Prosodie in sozialen Medien
Bachelorarbeit Informatik
Juni 2018 - November 2018
Betreuung: Ulf Leser, Susanne Fuchs (ZAS)

Ausblick

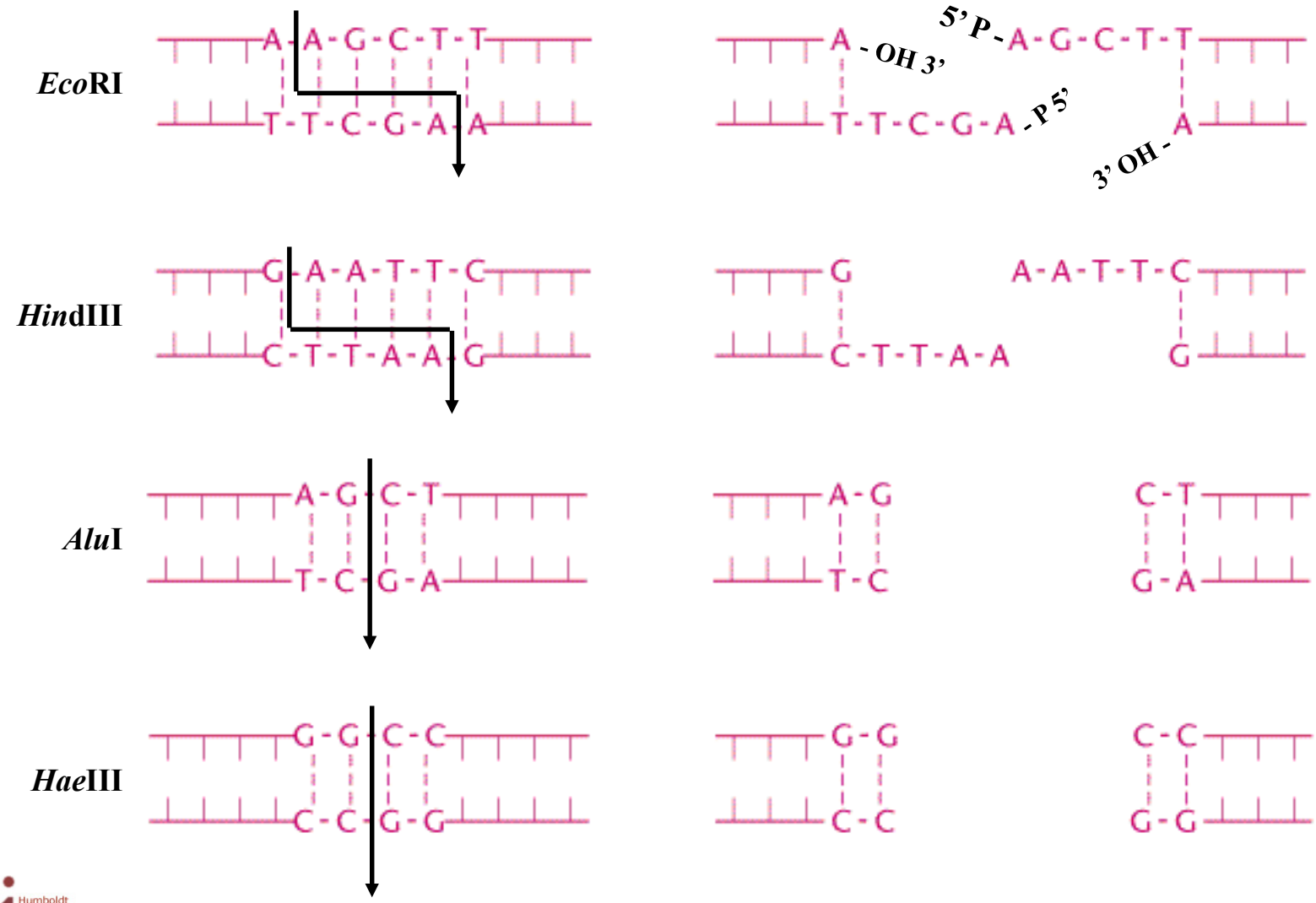
- Sommer 2019
  - BA: Einführung in die Bioinformatik
  - BA: Algorithmen und Datenstrukturen
  - BA: Proseminar Wissenschaftliches Arbeitens

- Themen des Semesters
- Ihr Feedback
- Prüfung
- Werbung
- DNA Computing

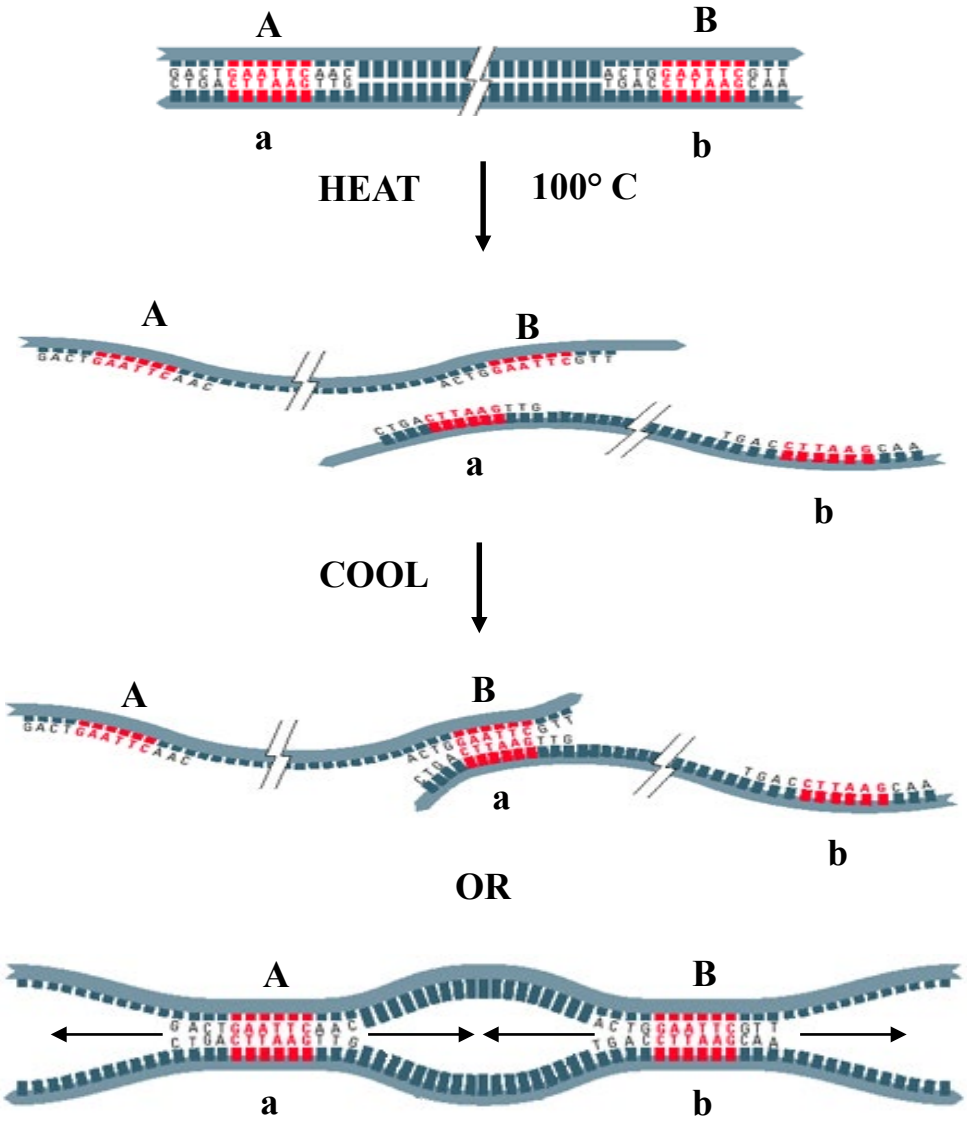
DNA Computing

- *„The use of biological molecules, primarily DNA, DNA analogs, and RNA, for computational purposes"*
- Warum kann das interessant sein?
  - Herkömmliche Chips stoßen an physikalische Grenzen
    - Strukturen kleiner als Lichtwellenfrequenz
    - Wege auf Chips so lange, dass Lichtgeschwindigkeit ein Problem wird
    - Chips entwickeln zu viel Hitze
  - DNA Moleküle: billig, relativ einfach zu handhaben, viele
  - Massive Parallelität möglich
    - 5 Gramm DNA enthalten ca. 10^{21} Basen
  - Primitive Operationen (= biotechnische Verfahren) vorhanden
    - Schneiden, Kopieren, Verlängern, Verkleben

Schneiden: Restriktionsenzyme

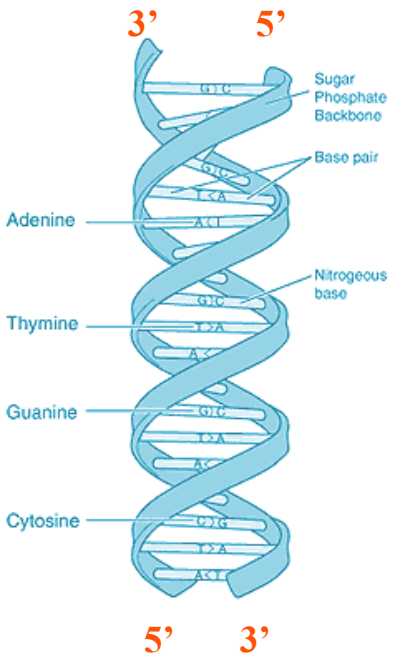


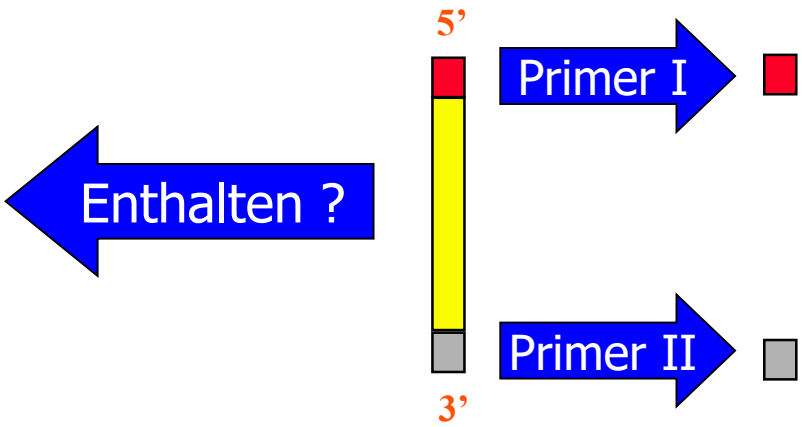
Verkleben: Hybridisierung



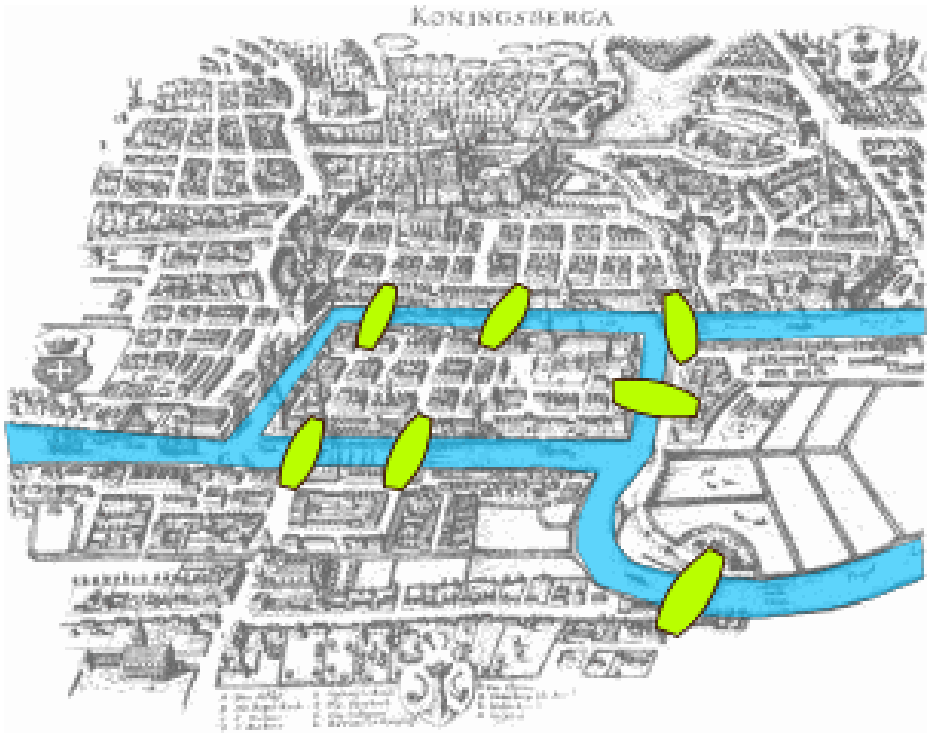
Vervielfältigen: PCR

- Gegeben:
  - Doppelsträngige DNA D
  - Probe S mit bekannter Sequenz
- Frage:
  - $S \in D$?



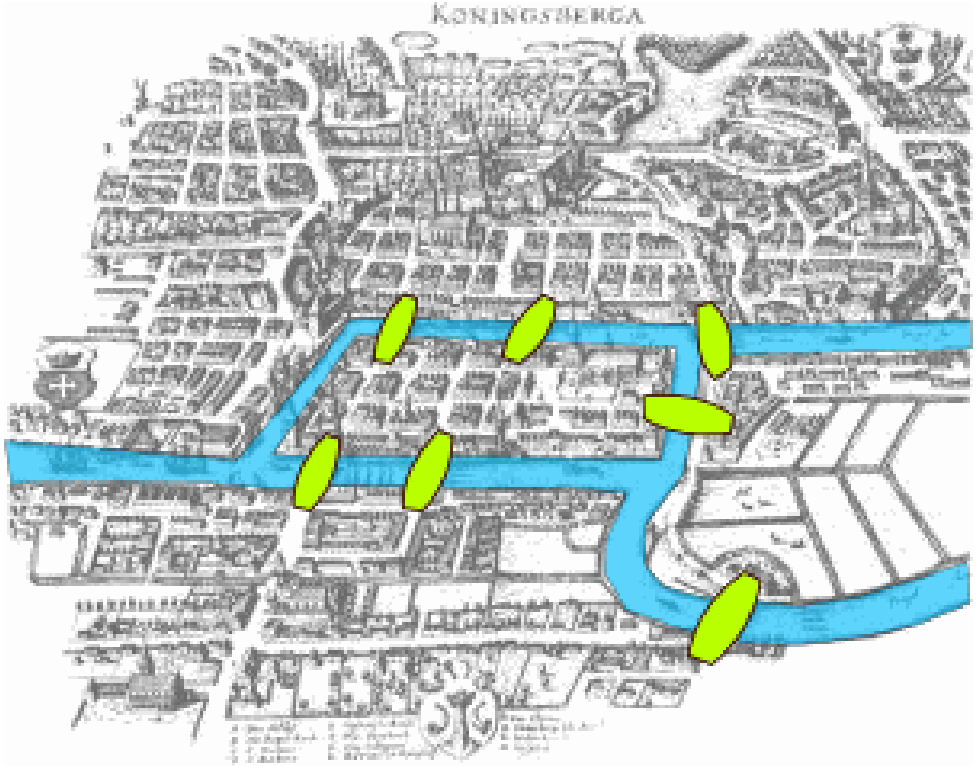


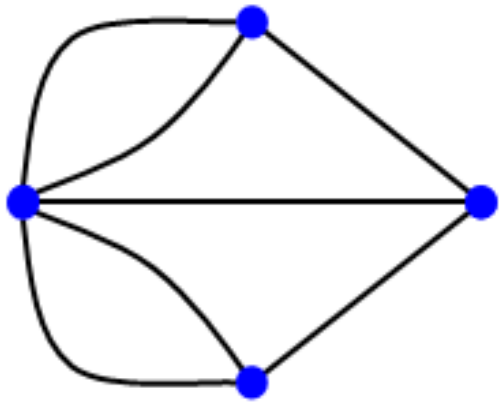
Eulers Problem – Königsberger Brücken



- Gibt es einen Weg, der jede Brücke genau einmal überquert?
- Und wieder da ankommt, wo man losgegangen ist?

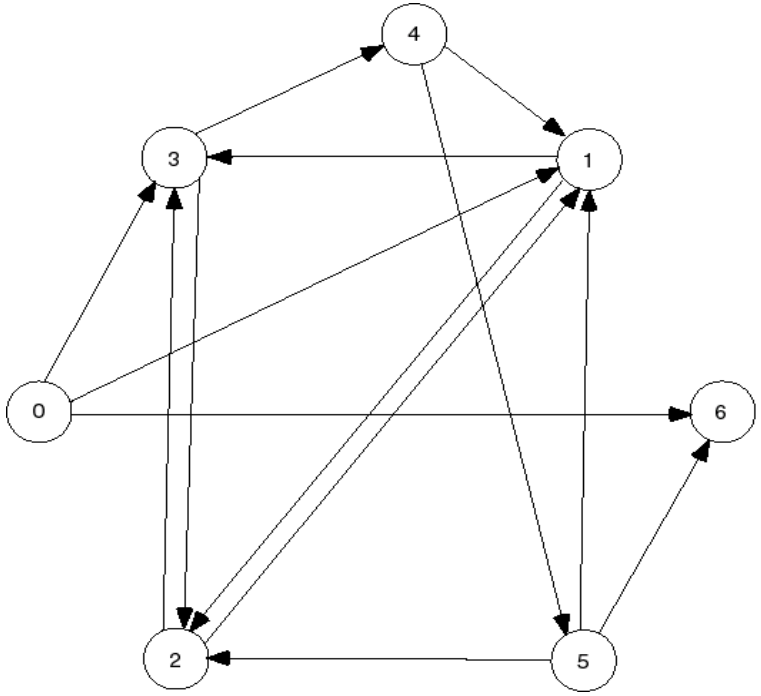
Modellierung





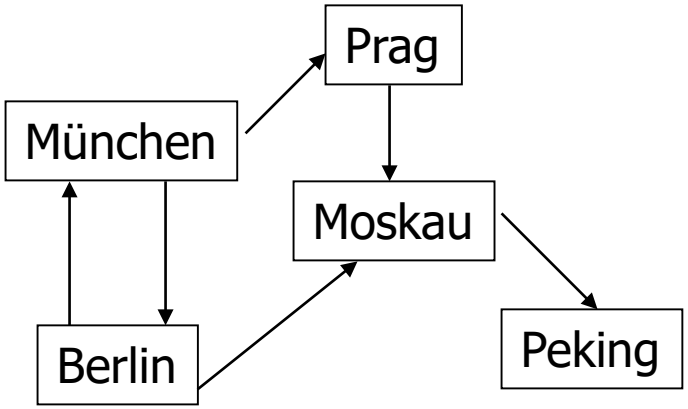
Hamilton Zyklen

- Gegeben ein Graph G mit Knoten V und gerichteten Kanten E
- Gibt es einen Pfad durch G, der an Knoten V_1 anfängt, an V_n aufhört, und jeden Knoten genau einmal berührt?
- NP-vollständig



DNA Computing Methode

- Leonard M. Adleman, *Molecular Computation of Solutions to Combinatorial Problems*, Science 226, 1994
- Abbildung des Graphen in DNA Moleküle
  - Knoten V: eindeutige DNA Sequenz der Länge 10
  - Kante $E(V_i,V_j)$: DNA Sequenz gebildet aus: $h(V_i[6...10])+h(V_j[1...5])$
    - h ist das Komplement einer Sequenz (A-T, C-G)

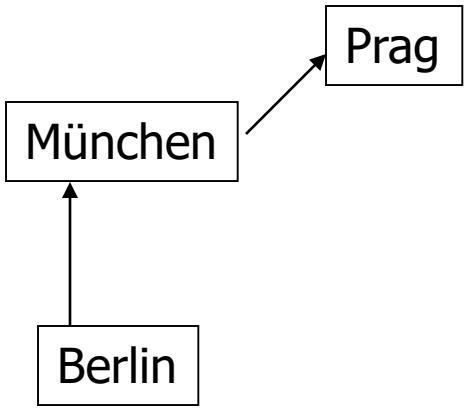


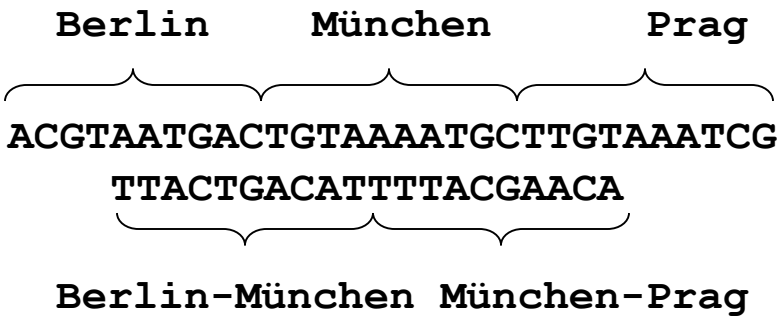

```
Berlin   = ACGTA ATGAC
Moskau   = GTGAC TTGAC
München  = TGTAA AATGC
Peking   = GAATA CGTAA
Prag     = TTGTA AATCG

Berlin-München = TACTG ACATT
Moskau-Peking  = AACTG CTTAT
...
```

Nächster Schritt

- Erzeuge all das in großer Menge und mische es
  - DNA lässt sich gezielt und billig synthetisieren
- Was passiert?
  - Alle Wege werden per Hybridisierung als doppelsträngige DNA entstehen





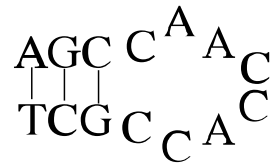
Algorithmus

- Erzeuge alle Pfade wie angegeben
- Filter 1: Selektiere Pfade, die in V_1 beginnen/in V_n aufhören
  - Vervielfältige durch PCR mit Primern aus V_1 und V_n
  - Schneide Banden aus dem Gel
- Filter 2: Selektiere Pfade der Länge n
  - Stränge durch Gelelektrophorese der Länge nach auftrennen
  - Längenbestimmung durch Vergleichs-DNA
- Filter 3: Selektiere Pfade, die jede Stadt enthalten
  - Baue „Städte" mit magnetischem Anhängsel
  - Trenne Doppelstränge; Sonde bindet; Filtern im magnetischen Feld
  - Iterativ für jede Stadt
- Es gibt einen Hamilton Pfad von V_1 nach V_n gdw. mindestens eine Sequenz übrig bleibt

Probleme

- Hybridisierung nicht perfekt
  - Führt zu falsch Positiven und falsch Negativen
- Selbsthybridisierung
- Viele Einzelschritte – fehleranfällig
  - Gerade das Herausfiltern der „richtigen" Pfade
- Größere Probleme verlangen sehr große Mengen DNA
  - Mehr Knoten und mehr Kanten – viel mehr Pfade
  - Wahrscheinlichkeit für Erzeugung jedes Pfades muss sehr hoch sein, damit PCR sie findet

Mismatched Hybridization



Hairpin Hybridization

Neues Problem

- DNA Wort Design (Oligodesign)
  - Berechne n DNA Sequenzen so, dass
  - ... nur die gewollten Hybridisierungen stattfinden
  - ... möglichst wenig Kreuzhybridisierung stattfindet
  - ... keine komplementäre Selbstähnlichkeit vorliegt
  - ... Wahrscheinlichkeit für Hairpins minimal ist
  - ... die Schmelzpunkte ungefähr gleich sind
- Vielfältige Anwendungen
  - Sequenzierung per Hybridisierung
  - Nachweis von Genexpression durch Chips
  - DNA Computing
  - ...

Aktuell?

- Raja Appuswamy, Kevin Le Brigand, Pascal Barbry, Marc Antonini, Olivier Madderson, Paul Freemont, James McDonald, Thomas Heinis: OligoArchive: Using DNA in the DBMS storage hierarchy. CIDR 2019

Synthetic DNA is one such storage media that has received some attention recently due to its high density and durability. In this paper, we investigate the problem of integrating DNA in the database storage hierarchy. More specifically, we ask the following two questions: (i) how can database knowledge help optimize DNA encoding and decoding? and (ii) how can biochemical mechanisms used for DNA manipulation be used to perform in-vitro, near-data SQL query processing?

In answering these questions, we present *OligoArchive*, an architecture for using DNA-based storage system as the archival tier of a relational database. We demonstrate that OligoArchive can be realized in practice by building archiving and recovery tools (pg_oligo_dump and pg_oligo_restore) for PostgreSQL that perform schema-aware encoding and decoding of relational data on DNA, and using these tools to archive a 12KB TPC-H database to DNA, perform in-vitro computation, and restore it back again.

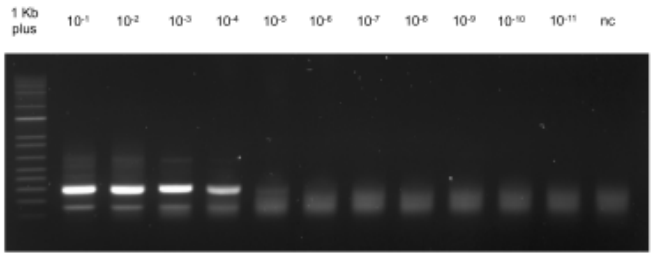


Figure 6: Oligo join sensitivity: specific oligo amplification diluted at different ratios in a background of random oligos. Amplification bands can be seen up to 10^{-5} dilution, meaning that two specific oligos are annealed/joined in a background of 10^5 random oligos.