Biochemisch inspirierte Berechnungen

Einführung

Peter Leupold

Vorlesung Wintersemester 2009/2010

Termine

Vorlesung:	Mittwoch	8-10 Uhr	WAalt 1332
	Donnerstag	16-17 Uhr	WAalt 1332
Übung:	Donnerstag	17-18 Uhr	WAalt 1332

Übungsaufgaben: Aus- und Abgabe jeweils Donnerstag

Prüfungszulassung: aktive Teilnahme an der Tafel, mindestens 50% der Aufgaben gelöst.

Natural Computing

Mit keinem anderen Wissenschaftsfeld hat die Informatik in den letzten Jahren intensiveren Austausch betrieben als mit den Biowissenschaften, und dies in beiden Richtungen.

Die Biowissenschaften benötigen immer effizientere Algorithmen zur Bewältigung immer größerer Datenmengen zum Beispiel bei der Gensequenzierung oder der Simulation komplexer Systeme und Vorgänge.

Andererseits hat das genauere Verständnis von Vorgängen in der Natur die Informatik in vielfältiger Weise zur Entwicklung neuer Methoden inspiriert.

Natural Computing

Der Begriff Natural Computing umfasst inzwischen ein weites Feld verschiedener Forschungsbereiche. Eine mögliche Einteilung

Durch die Natur inspiriertes Rechnen Auf herkömmlichen Rechnern werden Algorithmen entwickelt, die sich an Prinzipien wie Evolution, natürliche Auslese und dergleichen orientieren.

Simulation der Natur Dient vor allem dem Erkenntnisgewinn für Biologen und Chemiker, z.B. Bevölkerungsentwicklung, Proteinfaltung.

Berechnen mit natürlichen Materialien Hauptsächlich biochemische Substanzen werden verwendet um neue Berechenbarkeitsmodelle zu entwickeln - theoretisch oder auch praktisch.

Natural Computing

Einige Beispiele, die wir hier nicht behandeln werden sind:

Durch die Natur inspiriertes Rechnen
- Evolutionary Computing
- Biologisch inspirierte Netzwerkmodelle

Simulation der Natur
- Schwarmintelligenz
- Künstliches Leben
- Künstliche Immunsysteme

Berechnen mit natürlichen Materialien
- Zelluläre Automaten
- Neuronale Netze
- Quantenberechnungen

Unverbindlicher und grober Überblick über die Vorlesung

- Einführung
  - Motivation
  - Adlemans Experiment und DNA Computing
- Biochemisch inspirierte Berechenbarkeitsmodelle
  - Splicing Systeme
  - String-basierte Modelle
  - Membransysteme
  - Beobachtersysteme
- Biochemisch inspirierte algorithmische Probleme
  - Grundlagen — Kombinatorik von Wörtern
  - Textsuche und Textalgorithmen
  - Duplication-Wurzeln

Voraussetzungen

Berechenbarkeitstheorie
- Turingmaschine, berechnungstechnische Vollständigkeit, Churchsche These
- Chomsky-Hierarchie, generative Grammatiken
- Determinismus / Nichtdeterminismus

Komplexitätstheorie
- Laufzeiten, exponentielle, polynomielle, lineare
- $\mathcal{O}$-Notation und ähnliche
- Die Klassen NP und P, Vollständigkeit

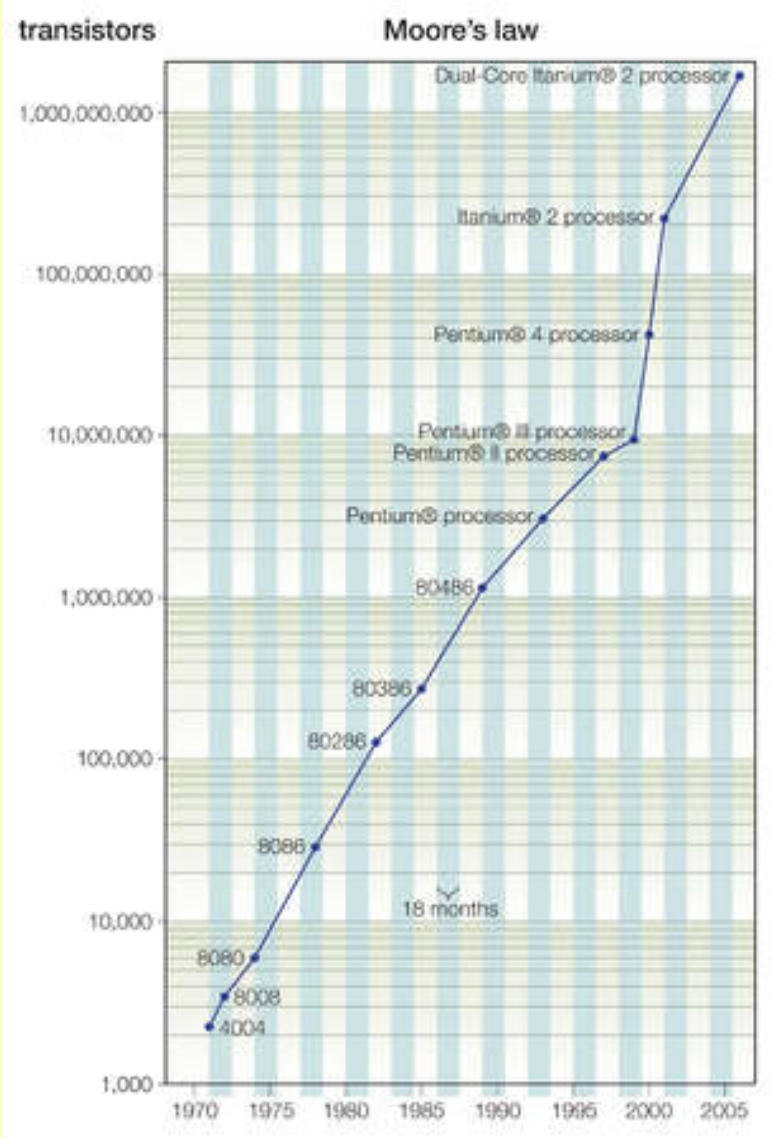
Sonstiges
- Grundwissen Biochemie, Gymnasium Oberstufe

Das Mooresche Gesetz

Das Mooresche Gesetz

Die Komplexität integrierter Schaltkreise mit minimalen Komponentenkosten verdoppelt sich etwa alle zwei Jahre.

Was sind die Grenzen des Wachstums für die Prozessorleistung?



Tunnelstrom und Rauschen

Bei einer anliegenden äusseren Spannung fliesst normalerweise elektrischer Strom durch einen elektrischen Leiter oder auch Halbleiter. Durch einen Nichtleiter findet ein solcher Stromfluss in der Regel nicht statt.

Wenn jedoch ein solcher Nichtleiter als Barriere zwischen zwei Leitern sehr schmal ist, dann kommt es doch zu einem Stromfluss, dem sogenannten *Tunnelstrom*.

Bei sehr eng benachbarten Bauelementen kann dieser Tunnelstrom zu Fehlern, sogenanntem *Rauschen* führen.

Dies zeigt, dass das Mooresche Gesetz an Grenzen stossen wird, und zwar in einem Grössenbereich von mindestens Tausenden oder Millionen von Atomen in der Leitungsbreite und -länge.

Die Versprechen des DNA Computing

Eine Motivation für das DNA Computing rührt von der Idee, Rechenelemente weit unterhalb der minimalen Grösse von elektrischen Bauteilen zu konstruieren. Im Idealfall sollten einzelne Moleküle oder gar Atome Funktionen erfüllen oder Informationen tragen.

Eine wesentliche Funktion der DNA in Organismen ist das Speichern der Erbinformation, und zwar mit einem bis zwei Bit pro Molekül.

Weiterhin existieren zahlreiche Enzyme, Mechanismen und Organellen, die diese Information lesen, verarbeiten und auch verändern.

Somit ist die DNA ein natürlicher Kandidat als Baustein für Rechner auf Molekularebene.

Ein Vergleich

Der Mensch

Das menschliche Genom hat einen Informationsgehalt von ca. 750 Megabyte.

Windows XP

Windows XP umfasst etwa 1,5 Gigabyte.

Ist Windows wesentlich komplizierter als ein Mensch,
oder ist es wesentlich schlechter organisiert?

Die Versprechen des DNA Computing

"The excitement DNA computing incited was mainly caused by its capability of **massively parallel** searches. This in turn showed its potential to yield tremendous advantages from the point of view of **speed, energy consumption and density of stored information**. For example, in Adleman's model the number of operations per second was up to $1.2x10^{18}$. This is approximately 1,200,000 times faster than the fastest supercomputer. While existing super computers execute 10^9 operations per Joule the energy efficiency of a DNA computer could be $2x10^{19}$ operations per Joule. That means that a DNA computer could be about 10^{10} times more energy efficient. Finally, storing information in molecules of DNA could allow for an information density of approximately 1 bit per cubic Nanometer, while existing storage media store information at a density of approximately 1 bit per 10^{12} nm^3. A single DNA memory could hold more words than all the computer memories ever made."

(Process of Bio-Computing and Emergent Technology, 1997, Lila Kari.)

Die Versprechen des DNA Computing

Im Idealfall verspricht das Rechnen mit Molekülen also mehrere Vorteile:

- Minimale Grösse – bis zu ein Bit pro Molekül oder gar per Atom.
- Massive Parallelität – jedes Molekül kann prinzipiell jederzeit mit jedem anderen interagieren.
- Sehr geringer Energieverbrauch, geringe Abwärme.

Wir betrachten diese Möglichkeiten am Beispiel des Experiments von Leonard Adleman, das als Geburtsstunde des DNA Computing gesehen werden kann.

Das Experiment

In einem 1994 in *Science* erschienenen Artikel beschrieb Leonard M. Adleman die Lösung einer Instanz des NP-vollständigen Hamiltonpfadproblems mithilfe von DNA-Molekülen.

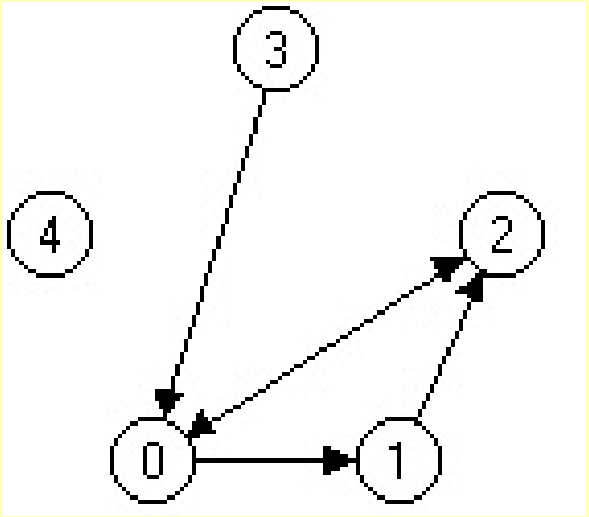
Damit wollte er die Machbarkeit von Berechnungen auf molekularer Ebene demonstrieren.

Um das Experiment verstehen zu können, betrachten wir zunächst das Hamiltonpfadproblem und relevantes Wissen über DNA.

Das Gerichtete Hamiltonpfadproblem

Sei $G = (V, E)$ ein Graph mit n Knoten und m Kanten.

Ein *Kreis* ist eine Folge von Kanten, die keinen Knoten mehrfach durchläuft, und bei der Start- und Endknoten identisch sind.



Ein Kreis heisst *hamiltonisch*, wenn er alle Knoten des Graphen durchläuft. Ist er nicht geschlossen, so liegt ein *Hamiltonpfad* vor.

Das Gerichtete Hamiltonpfadproblem

Das Gerichtete Hamiltonpfadproblem wird hier in folgender Form behandelt:

Eingabe Ein gerichteter Graph G, zwei Knoten v_{in} und v_{out}.

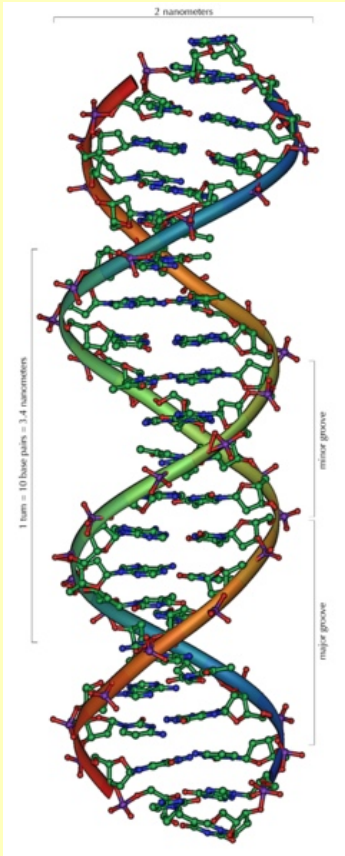
Fragestellung Hat G einen Hamiltonpfad, d.h. einen Pfad der kreuzungsfrei alle Knoten durchläuft, von v_{in} nach v_{out}?

Ausgabe JA oder NEIN.

DNS/DNA

Bemerkung vorab: Da die gesamte Literatur in Englisch verfasst ist, und wir auch direkt englische Begriffe wie *DNA Computing* verwenden, werden wir durchgehend das englische Kürzel DNA anstelle des deutschen DNS verwenden.

Im Normalzustand ist die DNA in Form einer Doppelhelix organisiert. Chemisch gesehen handelt es sich um eine Nukleinsäure, ein langes Kettenmolekül (Polymer) aus Einzelstücken, sogenannten Nukleotiden. Jedes Nukleotid besteht aus einem Phosphat-Rest, einem Zucker und einer von vier organischen Basen mit den Kürzeln A, T, G und C.



Quelle: Wikipedia

DNA

Die wesentlichen Merkmale von DNA in unserem Zusammenhang sind:

Doppelstrang Im Normalfall liegt DNA in Organismen in Form eines Doppelstranges vor. Dabei sind die Bindungen innerhalb eines Stranges normale chemische Bindungen. Zwischen den beiden Strängen bestehen lediglich Wasserstoffbrücken und andere Wechselwirkungen, so dass diese Bindung schwächer ist.

Komplementarität Die beiden Stränge lagern sich nicht beliebig aneinander. Nur komplementäre Stränge bilden die notwendigen Wasserstoffbrücken. Die jeweils komplemantären Basen sind Adenin (A) und Thymin (T) sowie Guanin (G) und Cytosin (C).

Richtung Die DNA-Stränge haben eine Richtung vom sogenannten $5'$-zum $3'$-Ende.

DNA – Richtung

Die Namen $5'$- und $3'$-Ende rühren von der Nummerierung der Kohlenstoffmoleküle in der Desoxyribose her.

Bei der Schreibung als Zeichenkette liegt in der Regel das $5'$-Ende links, das $3'$-Ende rechts.

DNA – Komplementarität

Zur mathematischen Formulierung der Komplementarität verwenden wir eine Abbildung, die wir mit einem Überstrich notieren:

$$\overline{A} = T, \overline{T} = A, \overline{G} = C, \overline{C} = G$$

Für längere Zeichenketten ist weiter zu beachten, dass sich die Reihenfolge der Zeichen spiegelbildlich umkehrt.

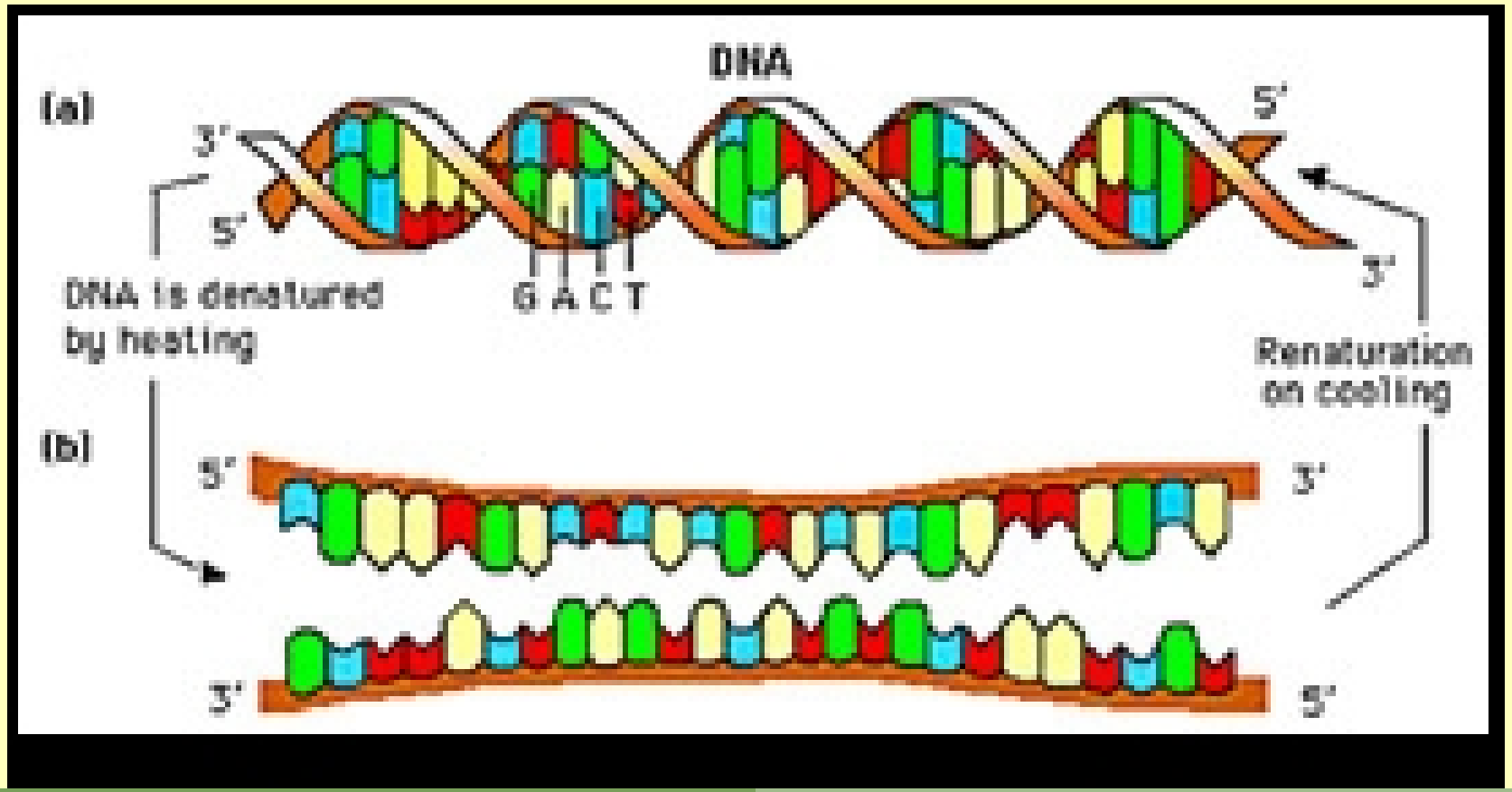
$$\overline{\overline{X}} = X$$

Die Abbildung ist eine sogenannte *Involution*.

$$\overline{ATGGC} = GCCAT$$

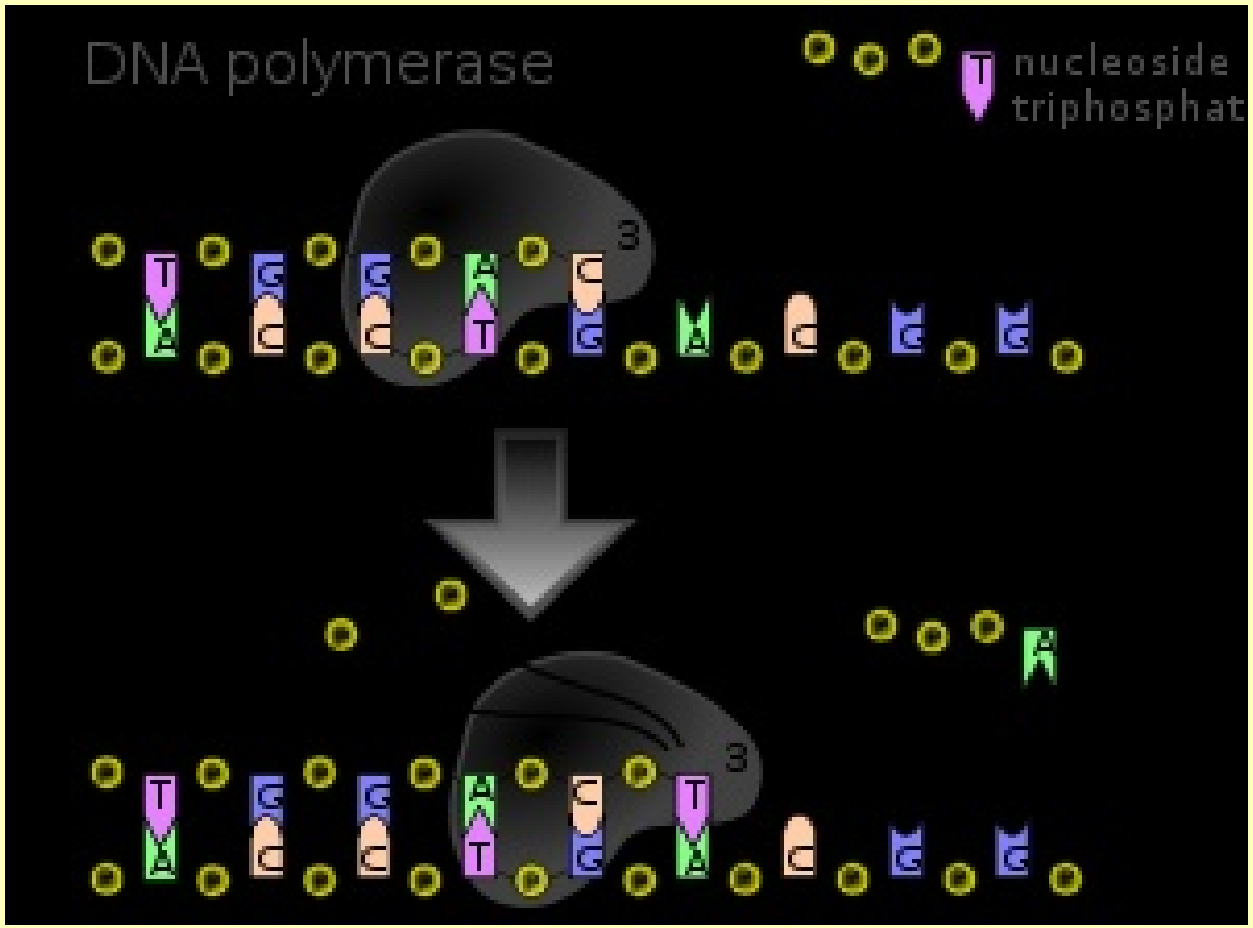
Denaturierung und Renaturierung

Bei Erhöhung der Temperatur werden schliesslich die Bindungskräfte zwischen den beiden Einzelsträngen überwunden, und die Doppelhelix spaltet sich in die beiden Einzelstränge auf. Erneute Abkühlung kann diese sich wieder aneinanderlagern lassen. Diese Vorgänge heissen *Denaturierung* und *Renaturierung*.

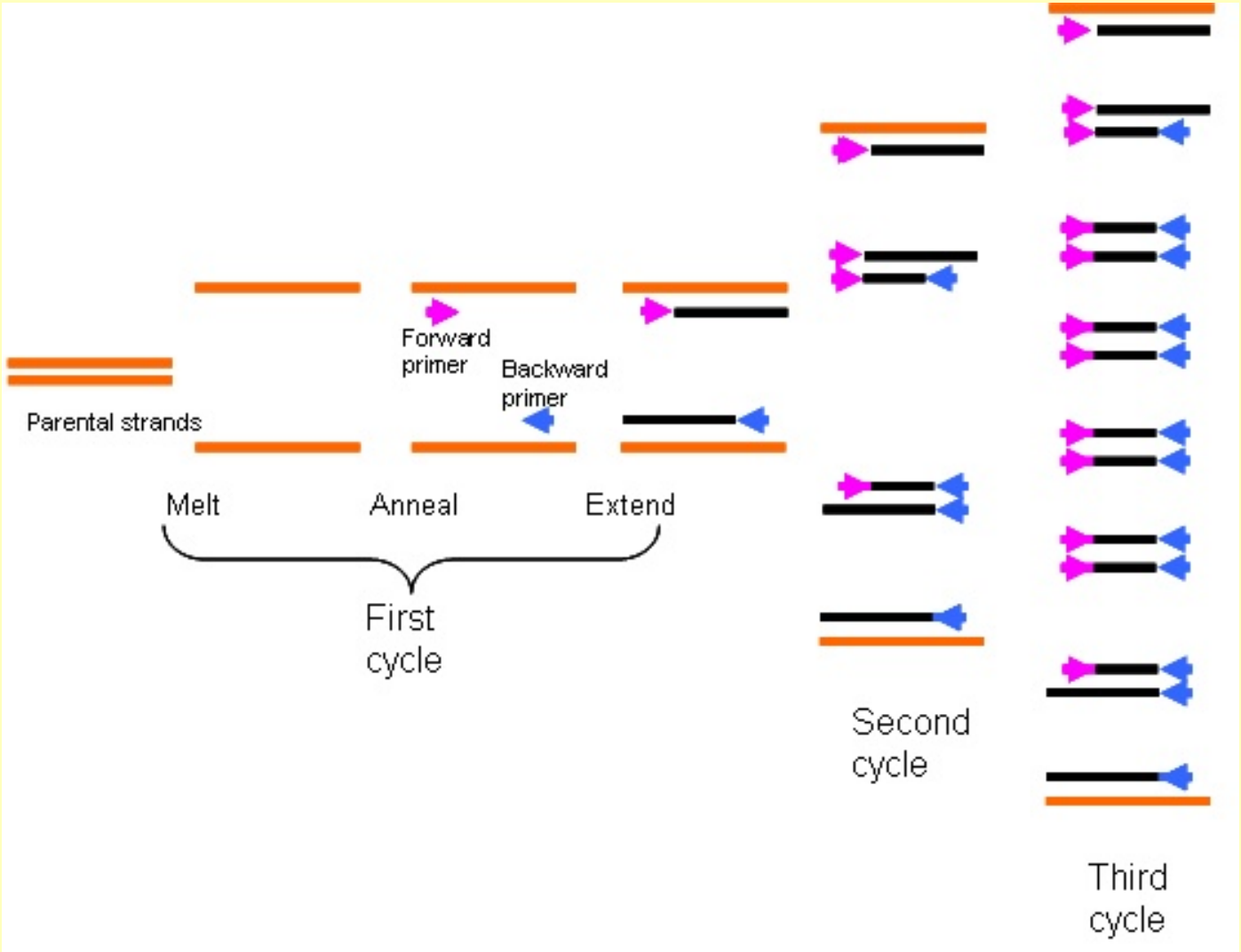


Die Polymerase Kettenreaktion (PCR)

Die Polymerase Kettenreaktion (englisch Polymerase Chain Reaction, PCR) ist eine Methode, um die Erbsubstanz DNA in vitro, d.h. im Reagenzglass, zu vervielfältigen. Dazu wird das Enzym DNA-Polymerase verwendet.



Die Polymerase Kettenreaktion (PCR)



Die Polymerase Kettenreaktion (PCR)

Wichtig:

- Exponentielles Wachstum. Im Idealfall verdoppelt sich die Anzahl der gewünschten Stränge in jedem Schritt.
- Primer sind die kurzen Startsequenzen, die die Polymerase ihre Arbeit starten lassen. Sie müssen stets in hinreichender Zahl vorhanden sein, da die Polymerase an Einzelsträngen nicht mit ihrer Arbeit beginnt.

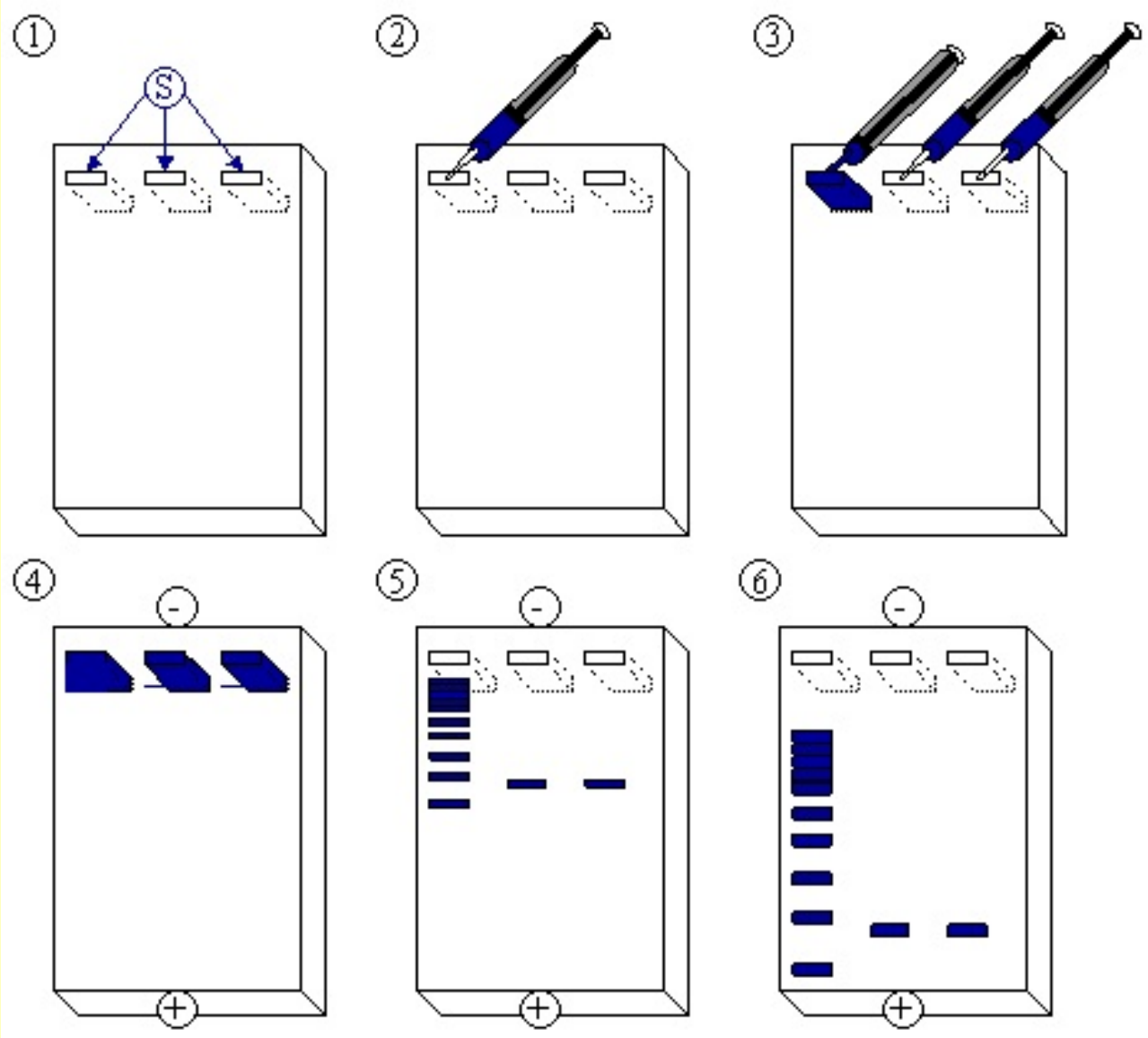
Agarose-Gelelektrophorese

Agarose-Gelelektrophorese ermöglicht es, Nukleinsäure-Stränge nach ihrer Grösse zu trennen und um ihre Grösse durch Vergleich mit Strängen bekannter Grösse zu bestimmen.

Lange Fäden aus Agarosepolymeren werden zu einem Gel vernetzt. Je höher die Agarose konzentriert ist, desto kleiner sind die Poren, die sich in dem Gel befinden.

Die Gelelektrophorese funktioniert wie ein Sieb für Moleküle; ein elektrisches Feld wird verwendet, um die negativ geladenen Nukleinsäure-Moleküle durch die Gelmatrix zu ziehen, wobei die kleineren Moleküle sich schneller durch das Gel bewegen können und somit eine Auftrennung der Stränge nach ihrer Grösse ermöglicht wird.

Agarose-Gelelektrophorese



Das Lesen von DNA

Die sogenannte *DNA-Sequenzierung* bestimmt die Basenfolge eines DNA-Stranges. Die gebräuchlichste Methode ist die *Didesoxymethode nach Sanger*, die inzwischen weitgehend automatisiert abläuft.

In jeder einzelnen Sequenzierreaktion können nur relativ kurze DNA-Abschnitte von weniger als 1000 Basenpaaren gelesen werden. Längere Sequenzen müssen zunächst zerlegt und im Anschluß wieder rekonstruiert werden.

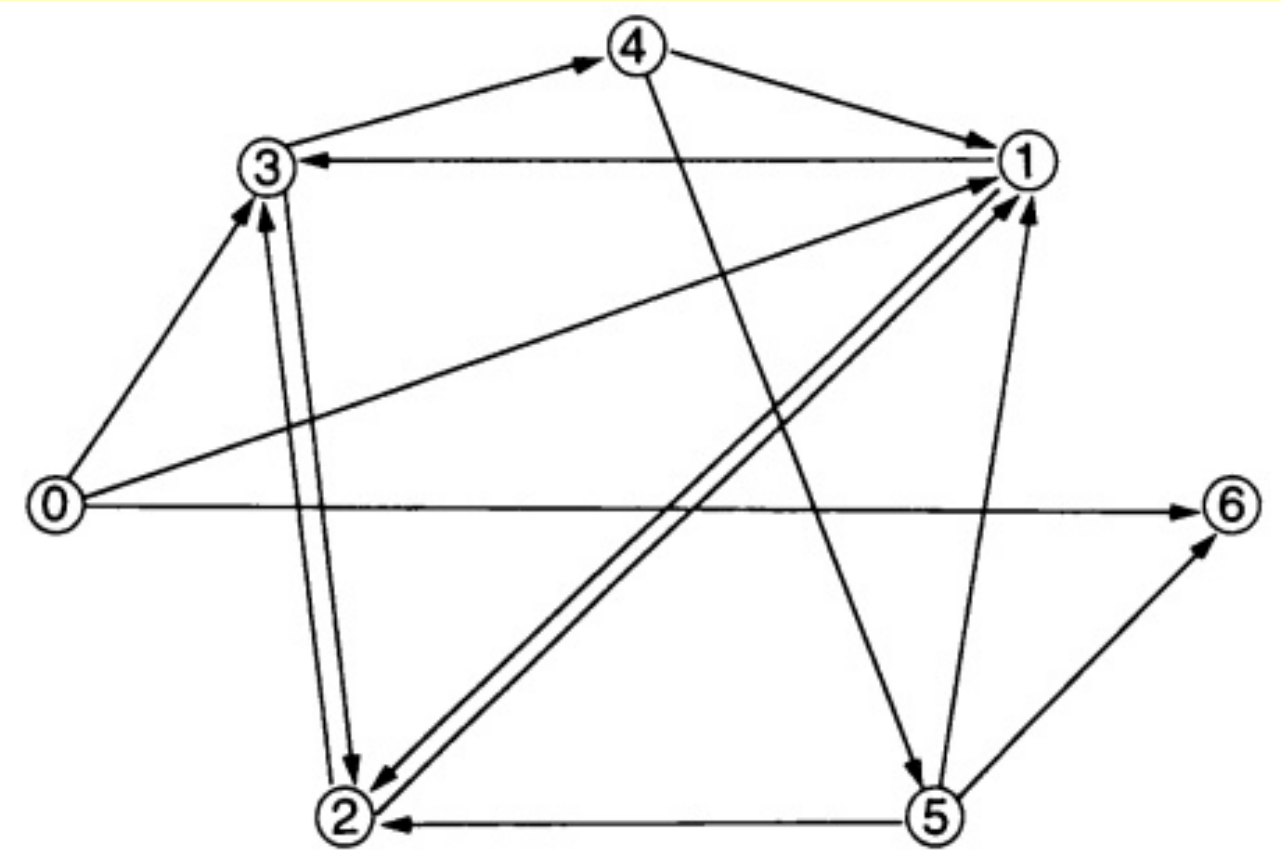
Die Sequenzierung efolgt mithilfe einer Variante der PCR, die Varianten der Basen als mit Fluoreszenz-Farbstoffen markierte Didesoxynukleosidtriphosphate (ddNTP) verwendet. Bei deren Einbau lässt der Strang sich nicht mehr weiter fortsetzen (kein $3'$-Ende). So entstehen Stränge unterschiedlicher Länge.

Hamiltonpfadproblem – der Algorithmus

1. Erzeuge zufällige Wege durch den Graphen.
2. Filtere die Pfade von v_{in} nach v_{out} heraus.
3. Filtere die Pfade heraus, die genau so viele Knoten durchlaufen wie der Graph hat.
4. Filtere die Pfade heraus, die alle Knoten des Graphen durchlaufen.
5. Wenn ein Pfad übrig ist, dann gib JA aus, anderenfalls gib NEIN aus.

Es ist zu beachten, dass in Schritt eins ein existierender Hamiltonpfad evtl. nicht erzeugt wird, und somit eine Ausgabe NEIN im letzten Schritt falsch sein kann.

Der Beispielgraph



Der Hamiltonpfad hier ist $0 \longrightarrow 1 \longrightarrow 2 \longrightarrow 3 \longrightarrow 4 \longrightarrow 5 \longrightarrow 6$.

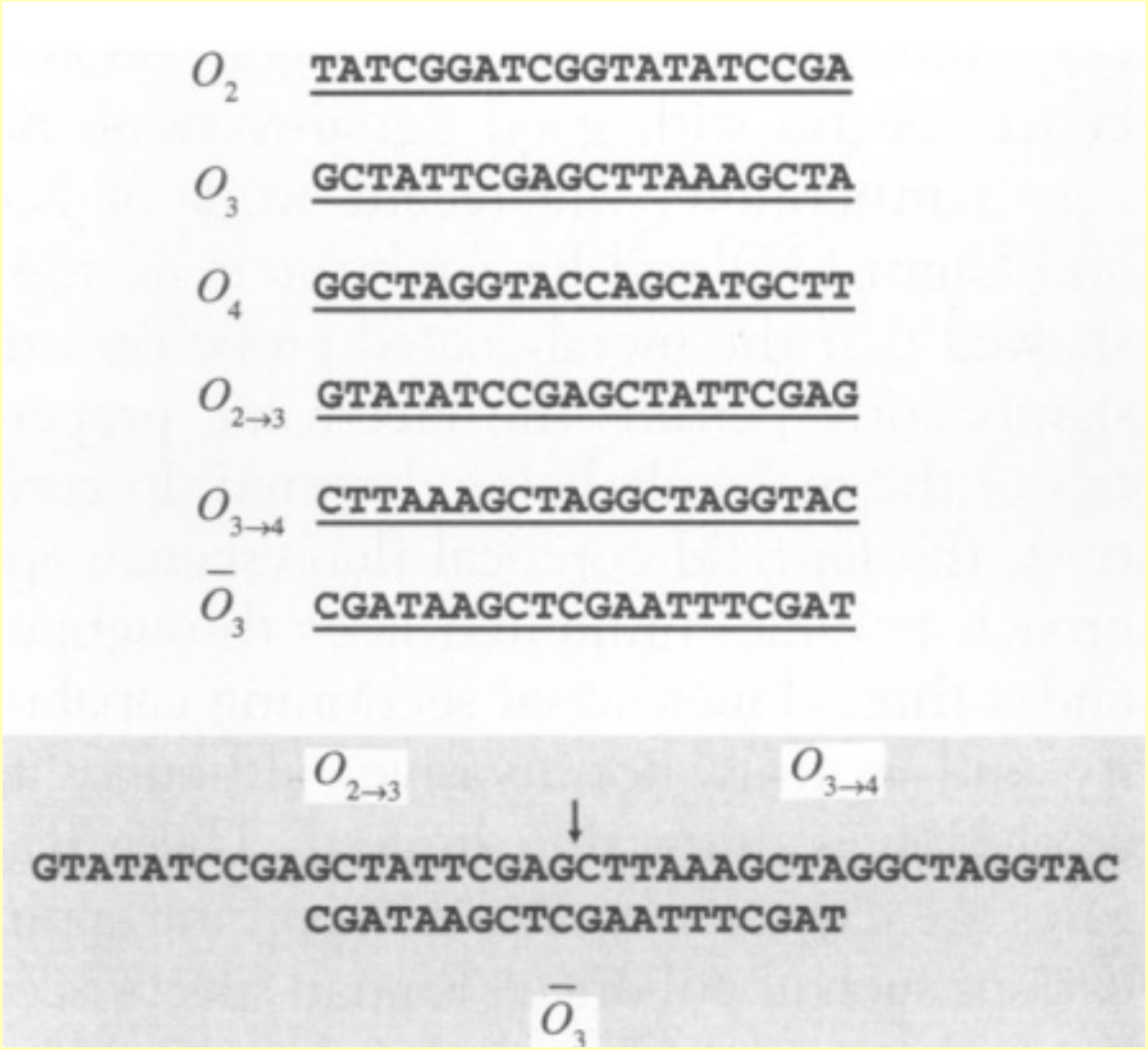
Die Codierung

- Zunächst werden den sechs Knoten zufällige (verschiedene) Sequenzen O_i von je zwanzig Basen zugeordnet.
- Jeder Kante wird dann folgende Sequenz von ebenfalls 20 Basen zugeordnet:

  Eine Kante $O_{i\rightarrow j}$ von Knoten v_i nach v_j beginnt mit den letzten zehn Basen des Knoten v_i und endet mit den ersten zehn Basen des Knoten v_j.Ausnahmen sind Kanten, die in v_{in} beginnen und mit allen

  20 Basen dieses Knoten anfangen, sowie Kanten, die in v_{out} enden und mit allen zwanzig Basen dieses Knoten enden.
- Im Experiment wurden dann Sequenzen $\overline{O_i}$ und $O_{i\rightarrow j}$ gemischt.

Die Codierung



Die Codierung

Durch das Mischen der Sequenzen $\overline{O_i}$ und $O_{i\rightarrow j}$ wird der erste Schritt des Algorithmus, die Erzeugung zufälliger Pfade implementiert. Da pro Kante etwa $3 \cdot 10^{13}$ Kopien der jeweiligen Sequenz verwendet werden, wird mit an Sicherheit grenzender Wahrscheinlichkeit jeder mögliche (kreisfreie) Pfad erzeugt.

Der zweite Schritt des Algorithmus, das Herausfiltern von Pfaden von v_{in} nach v_{out}, erfolgt mittels PCR mit den Primern O_0 und $\overline{O_6}$.

Somit werden nur Pfade von v_0 nach v_6 vermehrt.

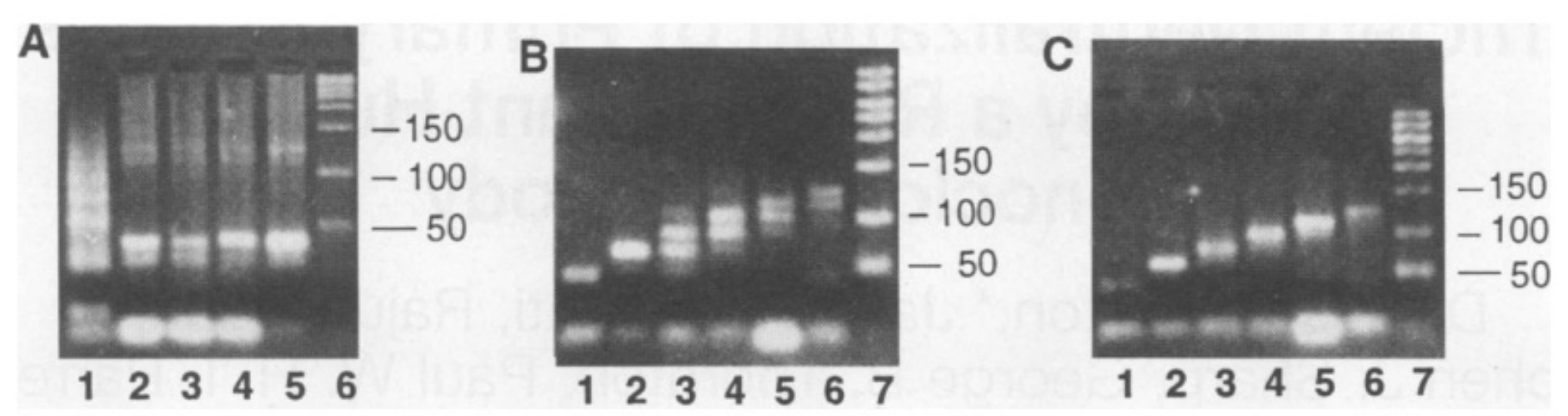
Implementierung

Die Pfade der Länge 6 wurden dann mittels Agarose-Gelelktrophorese heraussortiert (Stränge mit 140 Basenpaaren), was Schritt 3 des Algorithmus implementiert.

Das herausfiltern der Pfade die alle Knoten durchlaufen erfolgte mithilfe der kurzen Komlementstränge $\overline{O_i}$ mit i zwischen 0 und 6.

Schritt 5 erfolgte erneut mittels Agarose-Gelelktrophorese.

Das Ergebnis



A: Produkte des Schritts 1.

B: Produkte des Schritts 3.

C: Endprodukt mit dem Hamiltonpfad.

Fazit – Rechenzeit

Die Durchführung des gesamten Experiments nahm 1994 sieben Tage Laborarbeit in Anspruch.

Seither sind die Labortechniken in der Biochemie allerdings um ein Vielfaches beschleunigt worden.

Wichtiger jedoch ist der Fakt, dass die Rechenzeit mit grösseren Probleminstanzen lediglich linear ansteigt – verglichen mit wahrscheinlich exponentieller Laufzeit auf herkömmlichen Rechnern.

Dafür steigt die benötigte Menge an Molekülen wohl exponentiell an.

Fazit – Fehlerquote

Mögliche Fehlerquellen sind:

Codierung Die Codewörter müssen nicht nur verschieden, sondern verschieden genug sein. Was ist genug?

Haarnadeln DNA-Stränge können mit sich selbst einen teilweisen Dopplestrang bilden und so die Doppelstrangbildung verhindern.

Unvollständigkeit Weder die Erzeugung aller Pfade noch alle Filterschritte garantieren die vollständige Betrachtung aller und nur der gewünschten Moleküle. Aber grosse Molekülzahl führt zu minimaler Fehlerwahrscheinlichkeit.

Vergleich zu Berechnugen herkömmlicher Computer

Keine klare Trennung von Speicher und Prozessor (wie in der gewohnten von Neumann-Architektur).

Dies ermöglicht die massive Paralleliät, bringt in diesem Fall jedoch eine Spezialisierung auf nur eine Instanz eines einzigen Problems mit sich – oder die Entwicklung der Codierung muss mit einbezogen werden.

Perspektiven

Besteht trotz der offenkundigen Probleme Hoffnung auf nützllche DNA-Berechnungen? Auch die ersten elektronischen Rechner hatten vergleichbare Defizite.

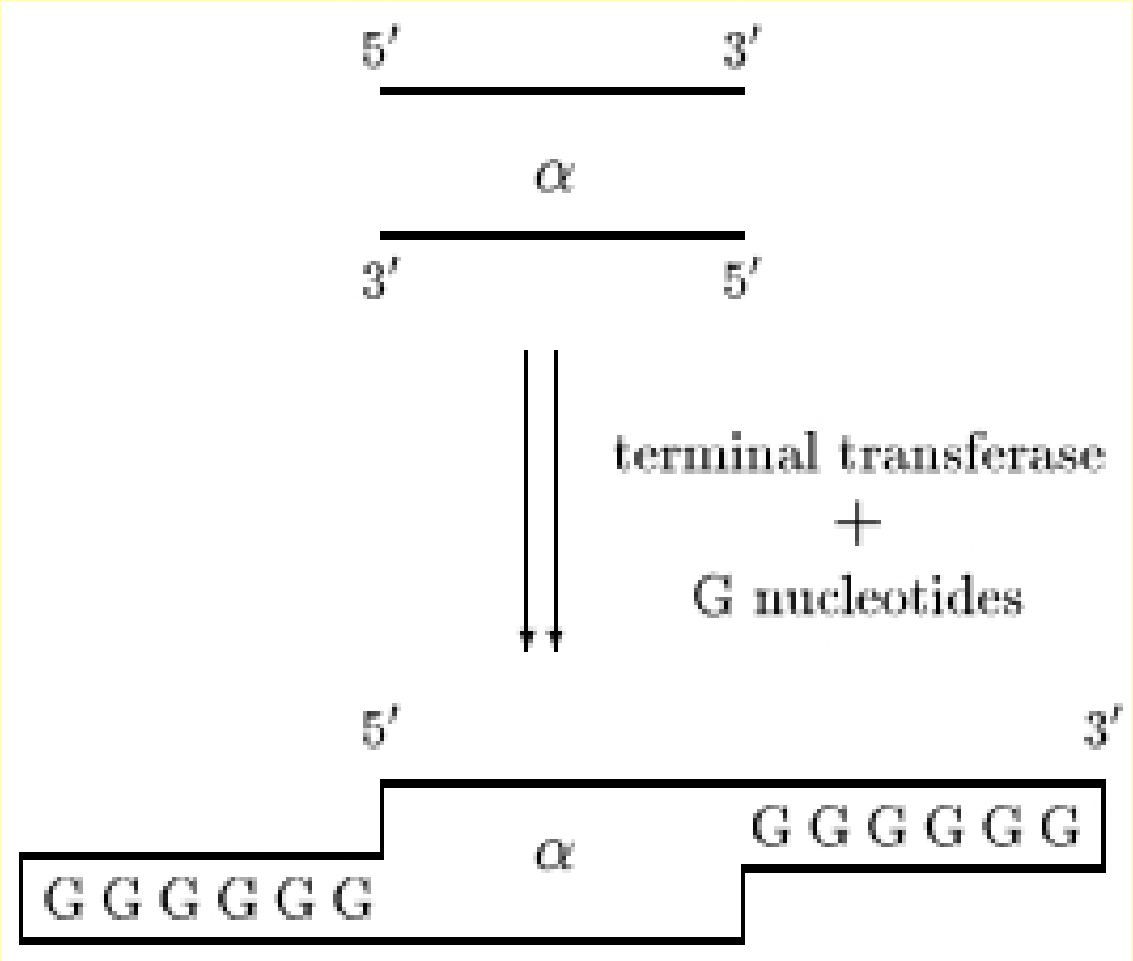
Mühsame Eingabe: Vgl. z.B. gestanzte Lochkarten.

Spezialisierung: Auch die ersten mechanischen und elektrischen Rechenwerke waren keineswegs universelle Turingmaschinen.

Taktung: Röhrenrechner hatten auch keine Taktzeiten im Millisekundenbereich. Das Lesen und Schreiben von DNA hat sich in den letzten Jahrzehnten enorm beschleunigt und automatisiert.

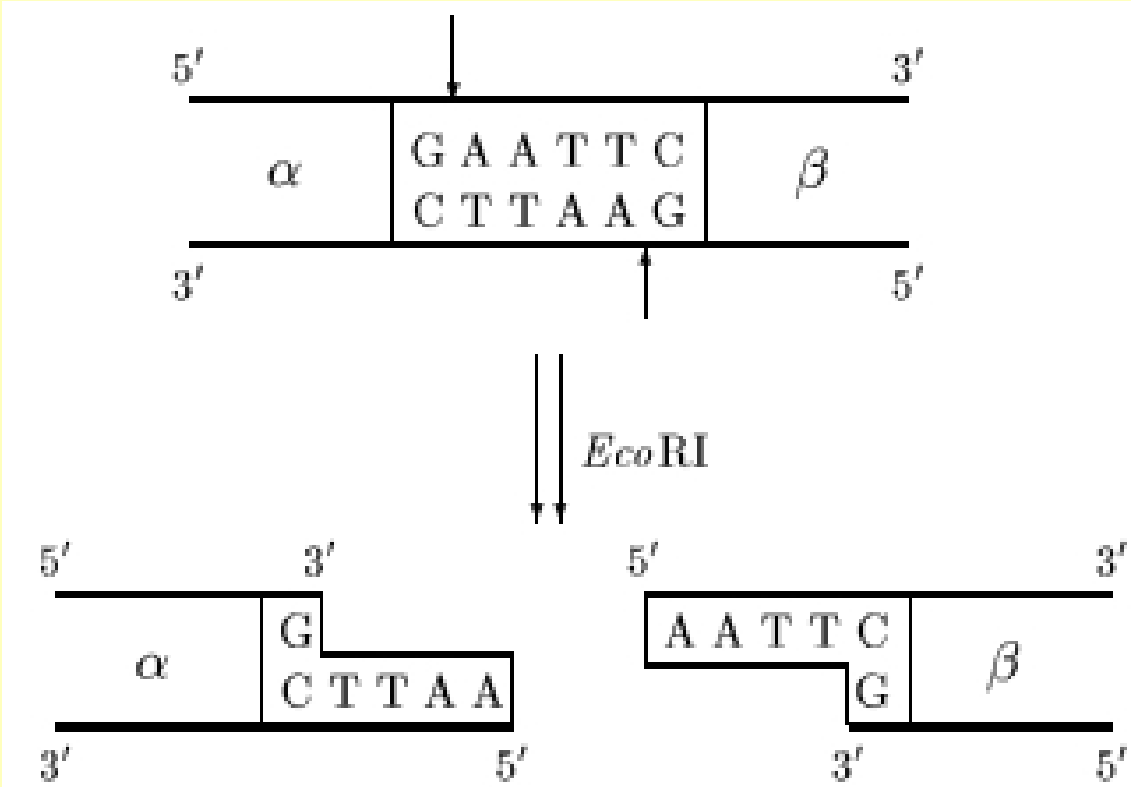
Weitere Enzyme – Terminale Transferase

Die *Terminale Transferase* ist eine besondere Polymerase, die auch an einen vollständigen Doppelstrang noch Nukleotide anhängen kann.



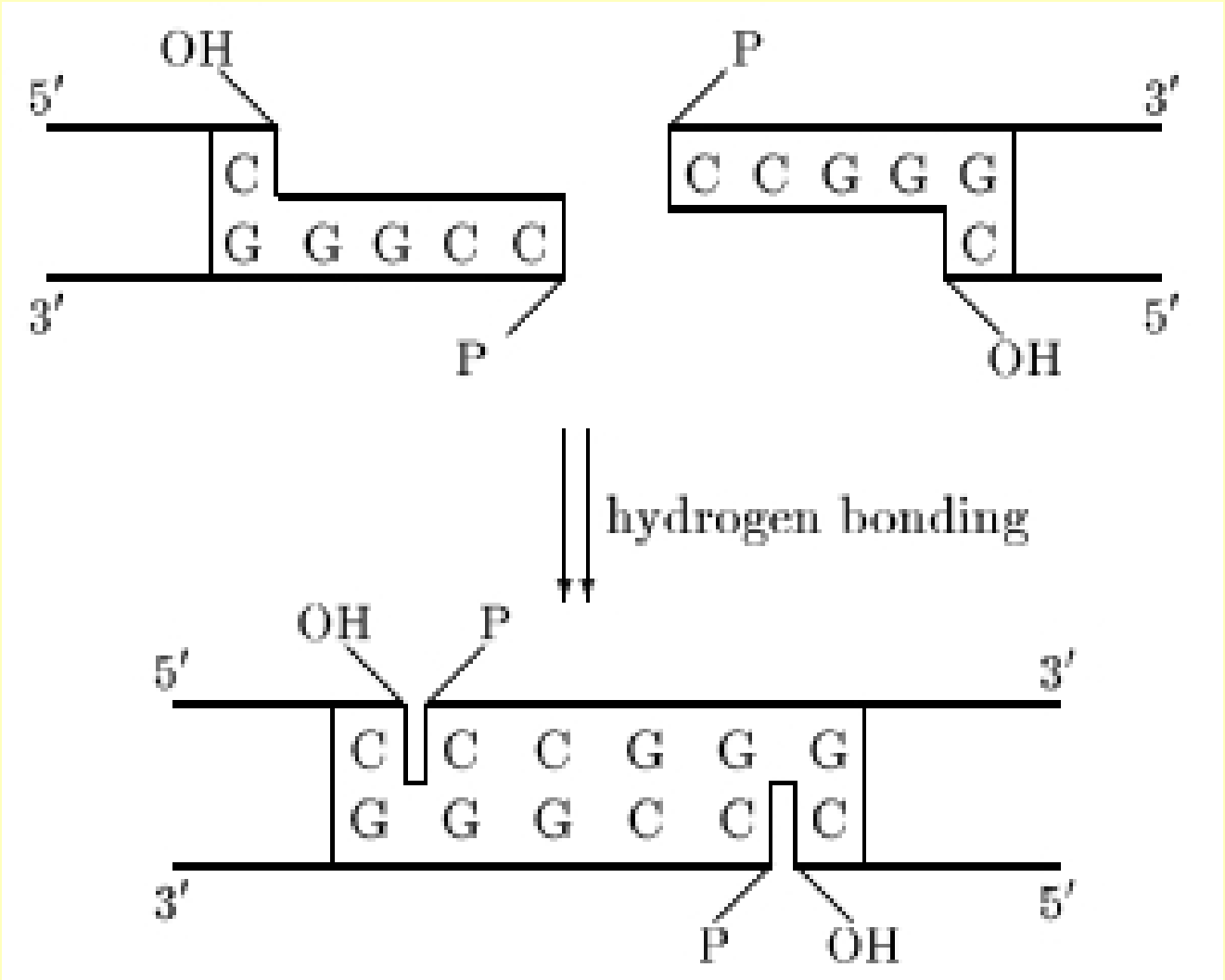
Weitere Enzyme – Nukleasen

- Exonukleasen entfernen einzelne Nukleotide an den Enden eines DNA-Stranges.
- Endonukleasen zerschneiden DNA-Stränge.
- Restriktionsenzyme sind Endonukleasen die nur Doppelstränge an Stellen mit bestimmten Sequenzen zerschneiden.



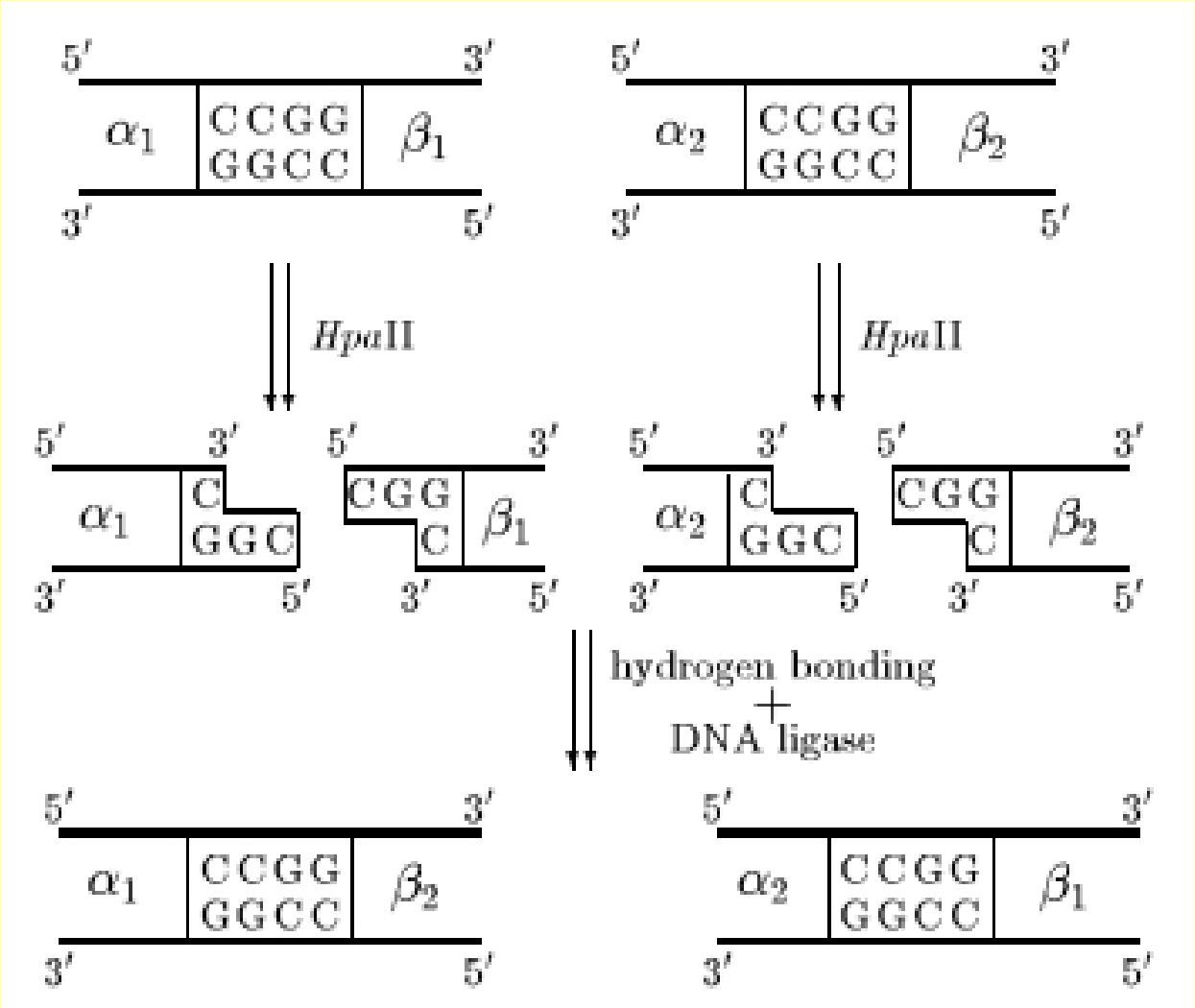
Weitere Enzyme – Ligasen

Ligasen verbinden zwei Stränge, die sich über sogenannte *sticky ends* aneinandergelagert haben.



Weitere Enzyme

Ein Beispiel für die mögliche Zusammenarbeit von Enzymen:



Weitere Enzyme

Die vorgestellten Enzymklassen stellen, zusammen mit der Komplementarität, die wesentlichen von der Natur zur Verfügung gestellten Werkzeuge zur Arbeit mit DNA dar.

In unserem Zusammenhang bilden sie die elementaren Operationen für Berechnungen, vergleichbar den Maschinenbefehlen auf herkömmlichen Rechnern.

Modelle des DNA Computing beruhen in aller Regel auf einer Auswahl dieser Operationen.