DNA-Computing

eine populärwissenschaftliche Einführung

Dipl.-Inform. T. Hinze, e-mail: th1@tcs.inf.tu-dresden.de

Arbeitsgemeinschaft DNA-Computing an der TU Dresden:
http://wwwtcs.inf.tu-dresden.de/dnacomp

Dresden, 11. Januar 2000

Gliederung

1. Die *Grundidee* des DNA-Computing
   – ein junges, unkonventionelles, massiv datenparalleles Computingkonzept
2. *Weshalb* DNA-Computing? – Grenzen der herkömmlichen Rechentechnik
3. *Einordnung* – DNA-Computing als Komponente des Future-Computing
4. Die DNA als *Datenträger*
5. *Operationen* auf DNA
   – angewandte Rekombinationstechnik und Molekularbiologie
6. Praktisches Rechnen mit DNA – Beispiel *Rucksackproblem*
7. DNA-Computing als *universelles Berechnungsmodell*
8. *Ausblick* und Visionen – Quo vadis DNA-Computing?

Die Grundidee des DNA-Computing

- DNA-Computing heißt „Rechnen im Reagenzglas“.
- Realisierung: Datenspeicherung, Operationen, Algorithmus, Ein-Ausgabe, Hardware
- **Daten:** repräsentiert durch DNA-Stränge
  - DNA: Desoxyribonucleinsäure, Träger der Erbinformation in zelluaren Organismen
  - Eingabedaten werden in DNA-Strängen kodiert (durch Bausteinabfolge und/oder Bausteinanzahl im Strang)
  - die entsprechenden Stränge entweder künstlich hergestellt (synthetisiert) oder aus Organismen gewonnen (isoliert)

Die Grundidee des DNA-Computing

- **Operation:** biochemische Reaktion, die auf alle DNA-Stränge im Reagenzglas gleichzeitig wirkt
  - dadurch massiv datenparallele Abarbeitung!
- **Ausgabe:** Testoperation, die feststellt, ob sich DNA-Stränge im Reagenzglas befinden
  - Bestimmung von Stranglängen (Bausteinanzahl) und Strangsequenzen (Bausteinabfolgen) ebenfalls möglich

Die Grundidee des DNA-Computing

- **DNA-Algorithmus:** Abarbeitungsvorschrift für den Laboranten (oder den Simulationsrechner),
  - welche biochemischen Reaktionen
  - in welchen Reagenzgläsern
  - in welcher Reihenfolge

  auszuführen sind.
- **Hardware:** molekularbiologisches Labor, in dem die zur Ausführung der Reaktionen benötigten Geräte und Chemikalien vorhanden sind
  - SIMD-Modell (Single Instruction Multiple Data, nach Flynn)

Weshalb DNA-Computing – Grenzen der konventionellen Rechentechnik

- Heutige Supercomputer erreichen bis zu 10^{12} ops/s.
- Taktfrequenzen und Integrationsgrad lassen sich nicht grenzenlos steigern.
- elektronische Parallelrechner versprechen Abhilfe, verlagern den exponentiellen Aufwand zur Lösung von kombinatorischen Suchproblemen jedoch von der Zeit zur Prozessoranzahl.

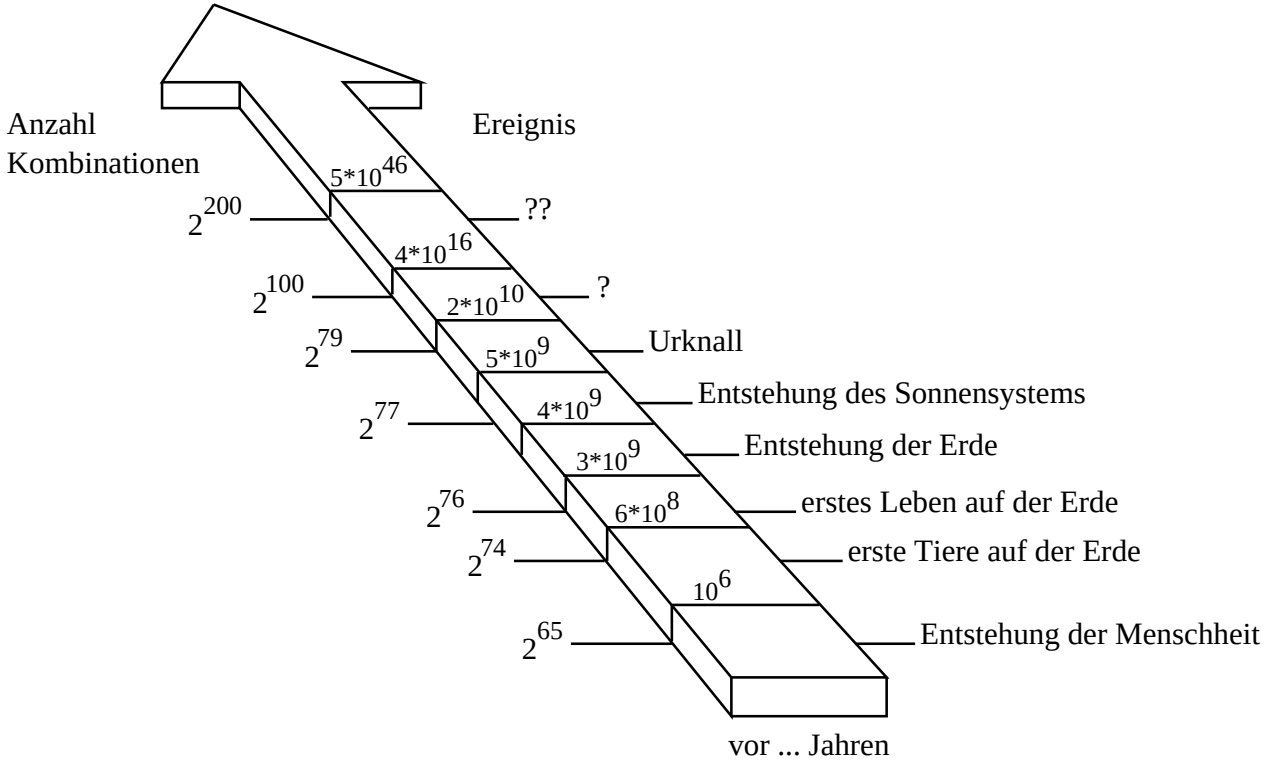
Neue Konzepte für das „zukünftige Computing“ gesucht!

$\longrightarrow$ Future-Computing

Kombinatorische Suchprobleme als rechenintensive Anwendung

Kombinatorische Suchprobleme - die Nadel im Heuhaufen finden

Wann hätte ein sehr schneller PC mit der vollständigen Enumeration beginnen müssen, um bis heute alle 2^n Kombinationen zu generieren und zu prüfen (1µs pro Kombination)?



Merke: Der unterstellte PC bildet alle 13.983.816 Tippmöglichkeiten des Lottos "6 aus 49" in nur 14s!

DNA-Computing als Komponente des Future-Computing

Future-Computing: Lösen bestimmter Probleme auf der Basis
alternativer Hardware

Ausprägungen:

- Quanten-Computing
  Ausnutzung der Fähigkeit eines Quantensystems, sich in mehr als einem Quantenzustand befinden zu können (Superposition)
  $\longrightarrow$ Quantencomputer
- Neural-Computing
  Aufbau, Training und Einsatz Neuronaler Netze
  $\longrightarrow$ Künstliche Intelligenz, Fuzzy-Logik
- Molekular-Computing mit den Spezialformen *DNA-Computing*, RNA-Computing und Protein-Computing

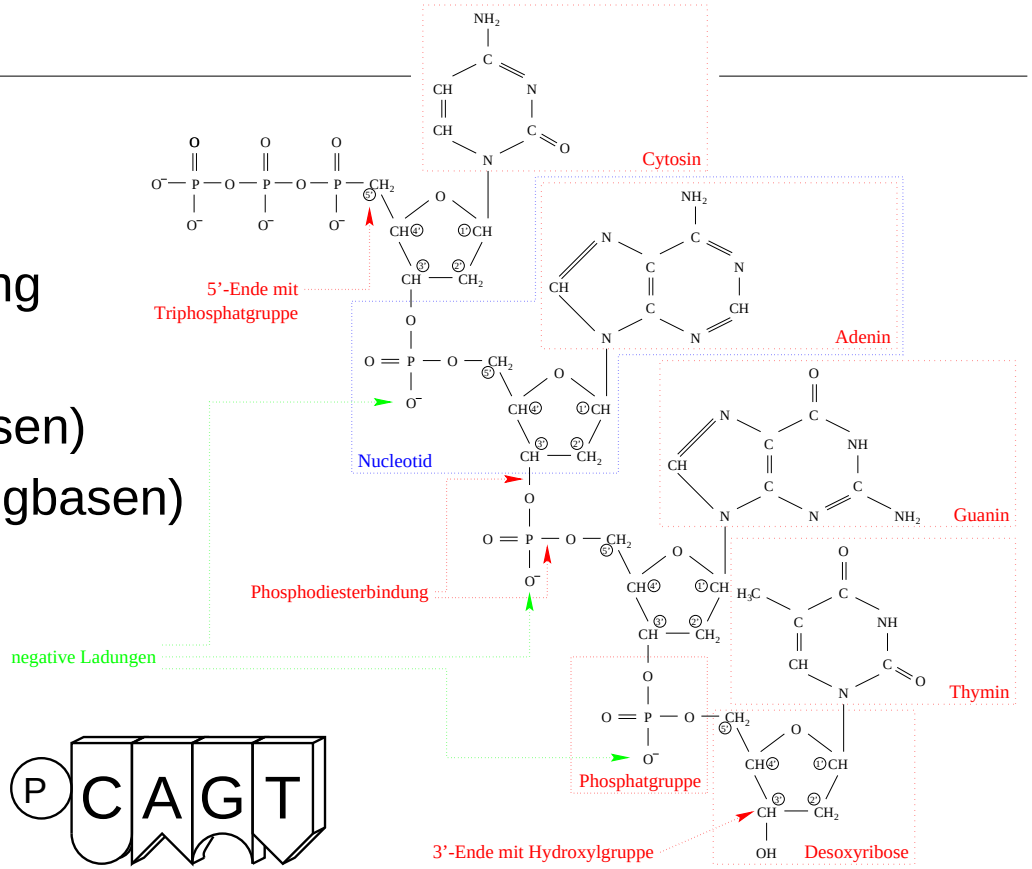
Die DNA als Datenträger

- Nucleinsäuren: im chemischen Sinne Polynucleotide
- aufgebaut aus heterozyklischen Basen, Kohlenhydrat und Phosphorsäure
- nach Art des Kohlenhydrats unterschieden:
  - Desoxyribonucleinsäuren (Desoxyribose als Kohlenhydrat)
  - Ribonucleinsäuren (Ribose als Kohlenhydrat)
- als Basen dienen:
  - bei DNA: Adenin, Cytosin, Guanin und Thymin
  - bei RNA: Adenin, Cytosin, Guanin und Uracil
- Begriff „Nuclein“: Komponente aus dem Zell*kern*
- Moleküle des Lebens

Chemische Struktur eines DNA-Einzelstranges

Begriffe:

- Desoxyribose
- Phosphatgruppe
- Phosphodiesterbindung
- negative Ladung
- Purine (Doppelringbasen)
- Pyrimidine (Einfachringbasen)
- Nucleotid (A, C, G, T)
- 3’-Ende
- 5’-Ende
- Hydroxylgruppe
- Leserichtung



DNA-Einzelstrang

Definitionen:

DNA-Einzelstrang: Ein *DNA-Einzelstrang* ist eine Sequenz der Nucleotide A, C, G und T. Die Enden der Nucleotidsequenz werden mit 5’ und 3’ angegeben. Jedes der beiden DNA-Einzelstrangenden ist durch eine geeignete chemische Gruppe markiert.

sense: Die Leserichtung 5’-3’ wird als *sense* bezeichnet.

antisense: Die Leserichtung 3’-5’ wird als *antisense* bezeichnet.

Es ist üblich, Nucleotidsequenzen in 5’-3’-Richtung zu notieren.

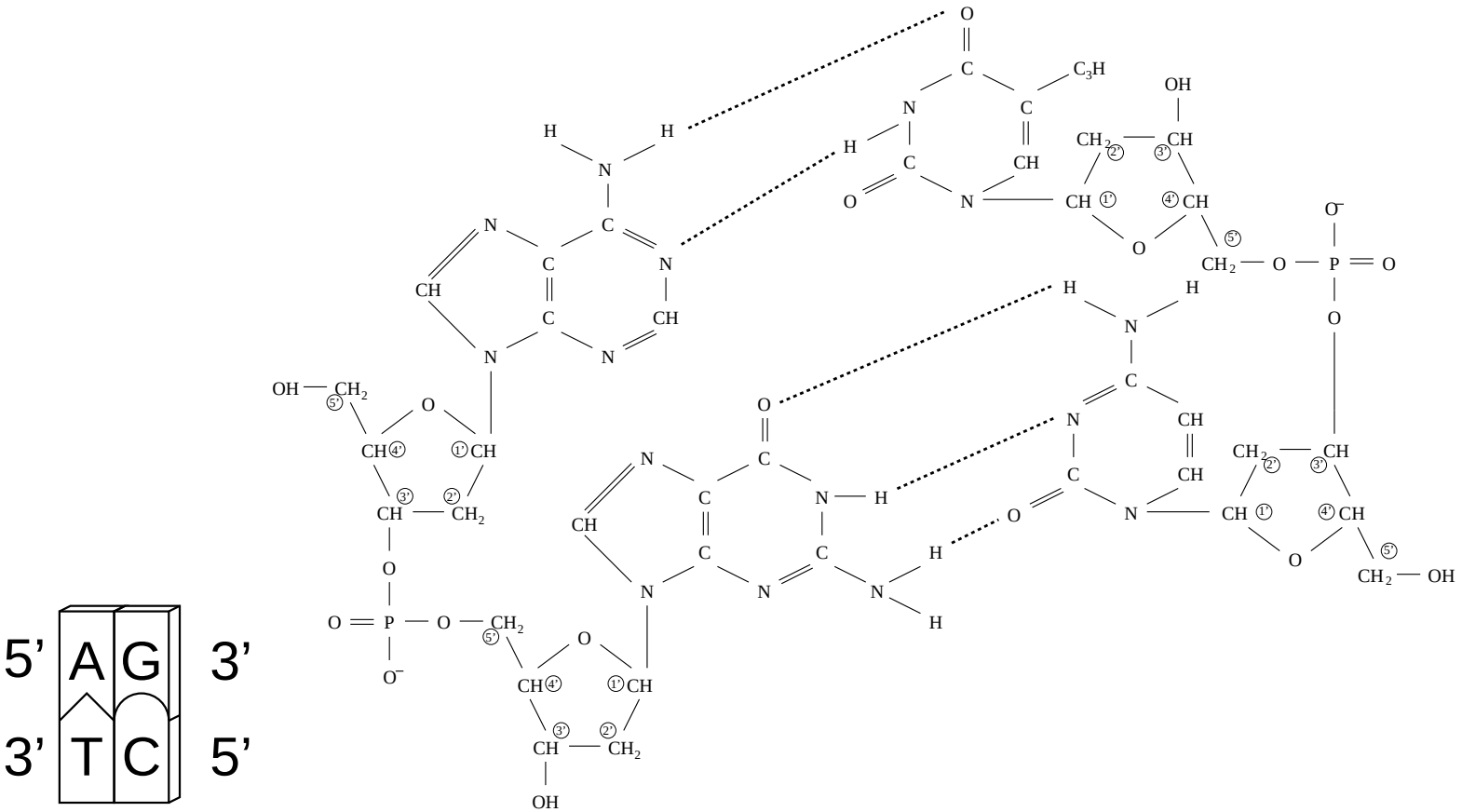
Basenpaarung durch Wasserstoffbrückenbindungen

Definition:

Wasserstoffbrückenbindung: ist eine chemische Bindung zwischen einem Proton (H^+) einer Hydroxylgruppe (-OH) oder HN-Gruppe und dem einsamen Elektronenpaar eines O-Atoms oder eines N-Atoms. Sie bildet sich aus, wenn die Gruppen sich auf eine Entfernung von $0,28nm$ nähern.

- Die Bindungsenergie beträgt nur $1/10$ der Hauptvalenzbindung.
- Basenpaarung kann zwischen benachbarter einzelsträngiger DNA auftreten.
- Dabei bilden sich zwischen je zwei gegenüberliegenden Basen Wasserstoffbrückenbindungen wie folgt aus:
- Adenin und Thymin können sich paaren (2 Wasserstoffbrücken).
- Cytosin und Guanin können sich paaren (3 Wasserstoffbrücken).

Chemische Struktur eines DNA-Doppelstranges



Komplementarität und Antiparallelismus

Definitionen:

Infolge der Möglichkeiten zur Basenpaarung:

Komplementarität von Nucleotiden: Die Nucleotide A und T sind zueinander *komplementär*, ebenso die Nucleotide C und G.

Nur komplementäre Nucleotide können sich über Wasserstoffbrückenbindungen zusammenlagern!

antiparallel: Zwei DNA-Einzelstränge, die entgegengesetzt ausgerichtet sind (ein Strang in 5’-3’-Richtung, der andere Strang in 3’-5’-Richtung), werden als *antiparallel* bezeichnet.

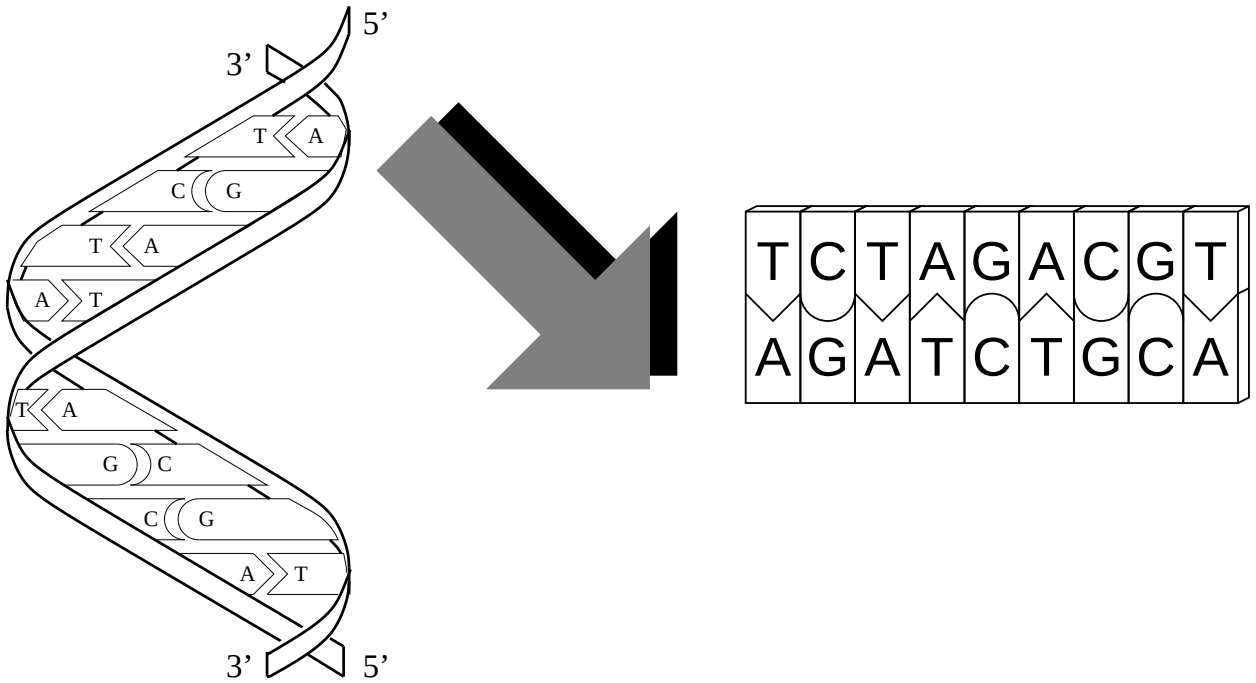
Vom DNA-Einzelstrang zum DNA-Doppelstrang

Basenpaarung bei kompletten Nucleotidsequenzen:

- *jede* Base des einen Stranges paart mit der gegenüberliegenden Base des anderen Stranges
- zweisträngiges Band komplementärer Nucleotide
- gegenüberliegende Stränge sind antiparallel
- gegenüberliegende Stränge zu einer Schraube gewunden
- eine Windung enthält 10 Basenpaare
- Struktur bezeichnet man als DNA-Doppelhelix
- $\longrightarrow$ Kern des DNA-Modells von Watson/Crick

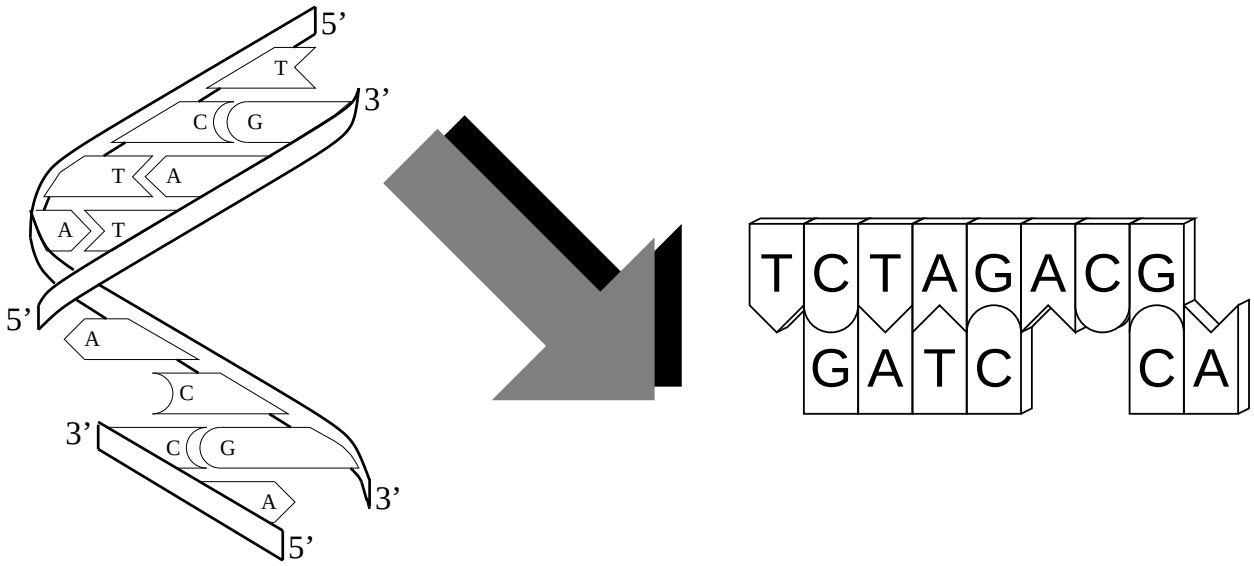
Struktur der DNA-Doppelhelix

- DNA in zelluaren Organismen zumeist als Doppelhelix
- komplementäre Nucleotide durch paßgeformte Bausteine dargestellt



Struktur der DNA-Doppelhelix

- durch die Möglichkeiten der in-vitro-Rekombinationstechnik muß der Begriff DNA-Doppelstrang weiter gefaßt werden
- Beispiel mit Einzelstrangüberhängen und -abschnitten:



DNA-Doppelstrang

Definition:

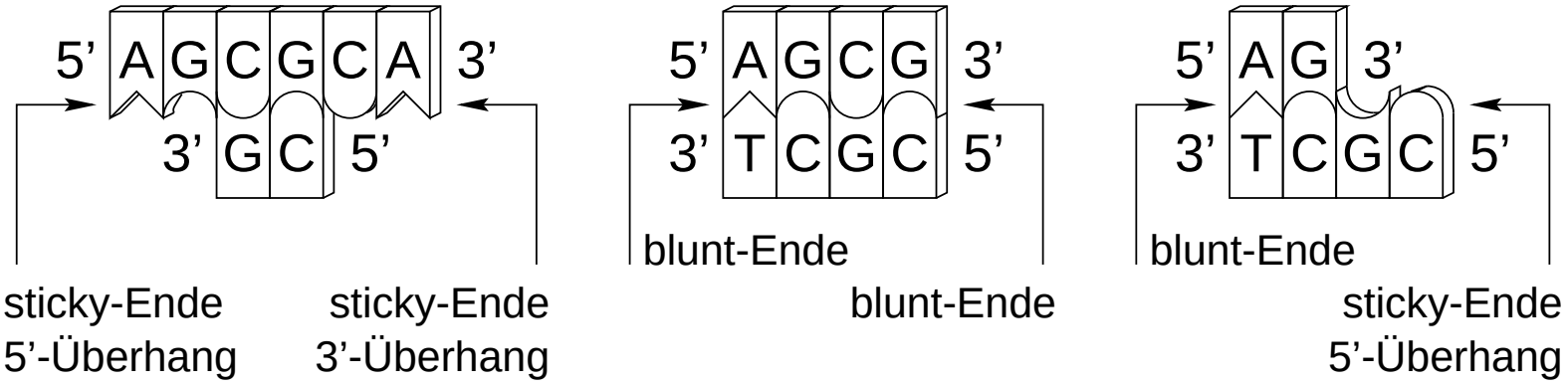
DNA-Doppelstrang: Ein *DNA-Doppelstrang* besteht aus mindestens zwei, jeweils in einer Nucleotidfolge komplementären sowie antiparallelen DNA-Einzelsträngen, die sich unter Bildung von Sequenzen komplementärer Nucleotidpaare verbunden haben. Die Nucleotidfolge reicht dabei bis zu einem Ende eines beteiligten DNA-Einzelstranges. Ein DNA-Doppelstrang läßt sich durch eine Sequenz komplementärer Nucleotidpaare beschreiben, die von Nucleotidsequenzen (DNA-Einzelstrangabschnitten), auch an den Strangenden, unterbrochen sein kann. Jeder integrierte DNA-Einzelstrang erzeugt ein 3'-5'-Endenpaar des DNA-Doppelstranges. Jedes der DNA-Doppelstrangenden ist durch eine geeignete chemische Gruppe markiert.

DNA-Doppelstrangenden

Definition:

blunt-Ende: Ein DNA-Doppelstrangende ist *blunt* (glatt), wenn es keinen Einzelstrangüberhang besitzt.

sticky-Ende: Ein DNA-Doppelstrangende ist *sticky* (klebrig), wenn es einen Einzelstrangüberhang besitzt. Dabei können sowohl das 3’-Ende als auch das 5’-Ende überhängen.



Eigenschaften von DNA-Strängen

im Hinblick auf ihre Nutzung als Datenträger im DNA-Computing

Speicherdichte: bis zu 10^{21} Basenpaaren pro Liter, das entspricht etwa 1 bit pro nm^3. DNA durch gewundene Struktur sehr kompakt

Langlebigkeit: DNA unter geeigneten Bedingungen beliebig lange konservierbar, deshalb auch Eignung als persistentes Speichermedium

redundante, dezentrale, verlustsichere Informationsspeicherung: DNA leicht millionenfach duplizierbar und auf mehrere Reagenzgläser verteilbar

In-vitro-Handling: DNA auch außerhalb von Zellen generierbar, verarbeitbar und konservierbar

energieeffiziente Verarbeitung: etwa $2 \cdot 10^{19}$ Operationen pro Joule

Eigenschaften von DNA-Strängen

im Hinblick auf ihre Nutzung als Datenträger im DNA-Computing

Operationsspektrum: Vielzahl von Rekombinationstechniken, großes Repertoire an molekularbiologischen Operationen

richtungsbehaftet: erleichtert Kodierung und Dekodierung von Daten in DNA-Sequenzen

elektrisch negativ geladen: Anwendung elektrophoretischer Analysemethoden

Instrumentarium zur Kodierung und Visualisierung von DNA-Daten: Synthese und Sequenzierung möglich

umweltfreundlich: keine aufwendig zu entsorgenden Nebenprodukte wie z.B. bei Elektronikschrott

Operationen auf DNA

- angewandte Rekombinationstechnik, Molekularbiologie und Biochemie
- Operationen auf DNA:
  - Generieren von DNA-Einzelsträngen $\longrightarrow$ *Synthesis*
  - Knüpfen und Aufbrechen von Wasserstoffbrückenbindungen $\longrightarrow$ *Annealing, Melting*
  - Mischen von DNA-Lösungen $\longrightarrow$ *Union*
  - Enzymatische Reaktionen $\longrightarrow$ *Ligation, Digestion, Labeling, Polymerisation, PCR*
  - Separieren nach Strangendenmarkierung $\longrightarrow$ *Affinity Purification*
  - Längenseparation und -bestimmung $\longrightarrow$ *Agarose Gel Electrophoresis*
  - Sequenzbestimmung $\longrightarrow$ *Sequencing*

Molekularbiologische Operationen für das DNA-Computing

Allgemeine Grundsätze:

in vitro: Reaktionen zumeist im Reagenzglas (Reaktionsgefäß, Tube) ausgeführt und nicht in der lebenden Zelle

Dilution (Verdünnung): DNA in wäßriger Lösung bestimmter DNA-Konzentration (Molarität, $\frac{mol}{l}$) zur Reaktion bereitgestellt

Redundanz: jeder DNA-Strang in sehr vielen identischen Kopien vorhanden

vollständiger Ablauf: jede Reaktion wirkt idealerweise auf alle DNA-Stränge im Reagenzglas

Reinheit des Ansatzes: idealerweise keine Verunreinigungen durch Fremdstoffe und Reste vorangegangener Reaktionen

Abarbeitungsprotokoll: beschreibt die praktische Ausführung der Reaktion reproduzierbar unter Erfassung aller kontrollierbaren Einflußgrößen

Einflußgrößen molekularbiologischer Operationen (Auswahl)

Reaktionen sind parameterbehaftet!

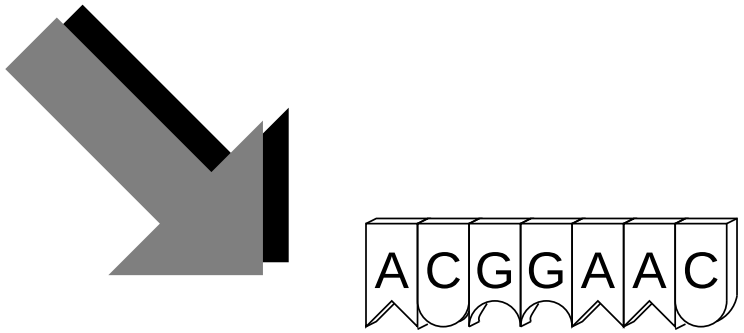
- Ausgangsstoffe und ihre Konzentrationen
- Reihenfolge und Zeitpunkte ihrer Zusammenführung
- Temperatur-Zeit-Verlauf des Reaktionsansatzes
- Inkubationszeiten des Reaktionsansatzes
- pH-Wert, Luftfeuchtigkeit, elektrische Größen
- äußere Kräfte und ihr Verlauf (Vortexen, Zentrifugieren, Schütteln, Durchmischen mit Pipette, Ruhelage)
- …

Es gibt auch sehr viele unkontrollierbare Einflußgrößen, die die Reproduzierbarkeit beeinträchtigen (Seiteneffektanfälligkeit)!

Synthesis

- Synonyme: Synthese

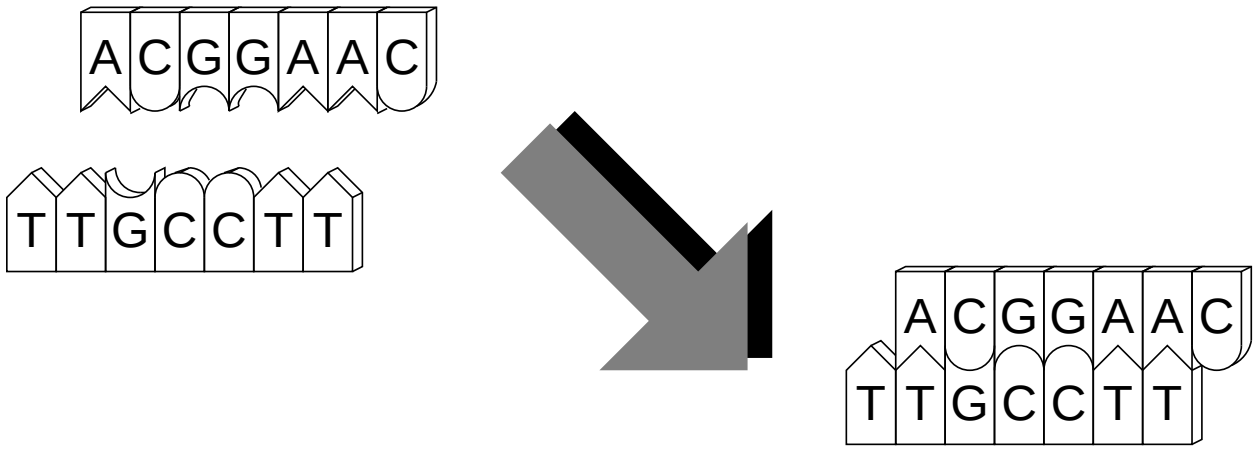
5′-ACGGAAC-3′



- Generieren von DNA-Einzelsträngen (Oligonucleotiden)

Annealing

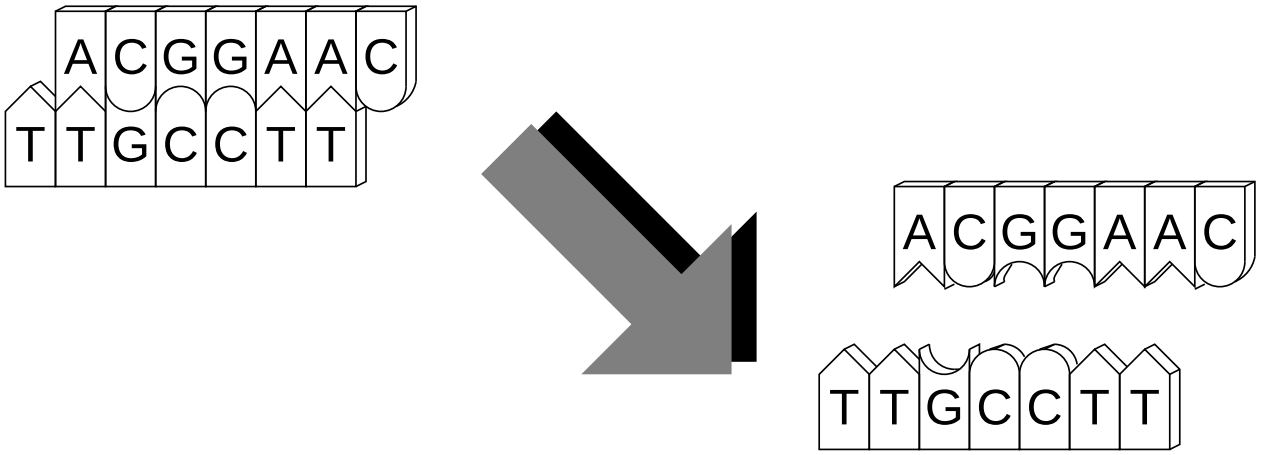
- Synonyme: Hybridisieren, Erstarren



- Zusammenlagern von DNA-Einzelsträngen zu DNA-Doppelsträngen

Melting

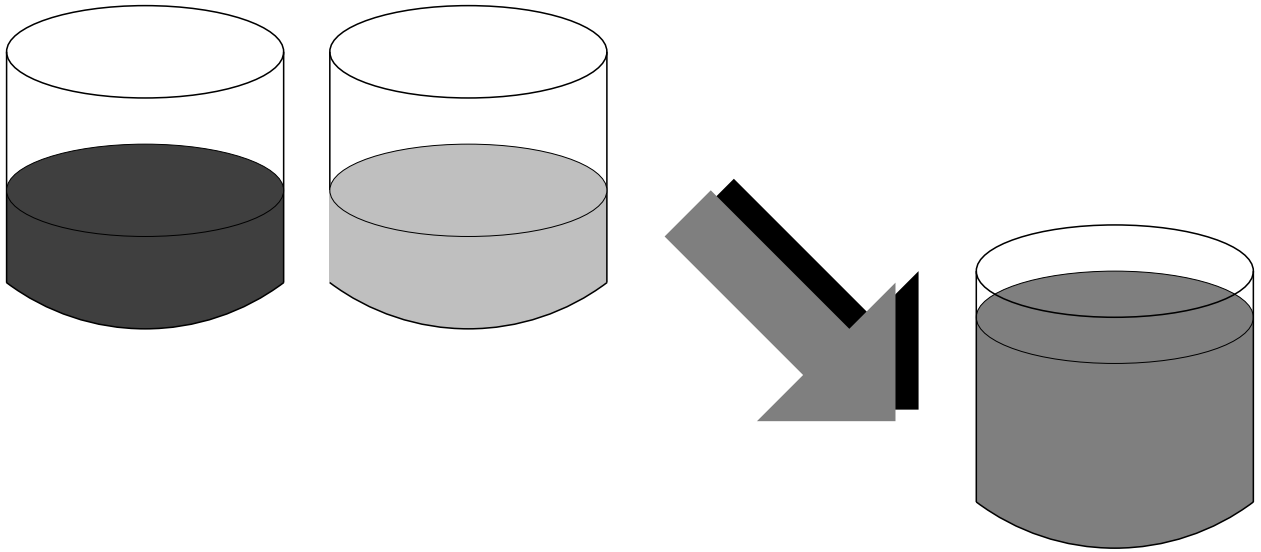
- Synonyme: Denaturieren, Schmelzen



- Aufspalten von DNA-Doppelsträngen in DNA-Einzelstränge

Union

- Synonyme: Merge, Mix, Pour, Mischen, Vereinigen



- Vereinigen von Reagenzglasinhalten

Enzymatische Reaktionen allgemein

- Enzyme gelten als „Katalysatoren des Lebens“ (Biokatalysatoren).
- bestimmte Reaktionen – insbesondere in lebenden Zellen – durch Enzyme milliardenfach beschleunigt; Grundlage für Lebensvorgänge
- Begriff „Enzym“ vom griechischen Wort für Sauerteig abgeleitet
- Vielzahl von Enzymen bekannt, nach ihrer Wirkung klassifiziert (EC)
- Jedes Enzym katalysiert eine bestimmte biochemische Reaktion, geht jedoch selbst unverändert aus der Reaktion hervor.
- chemische Struktur und Wirkungsweise vieler Enzyme aufgeklärt

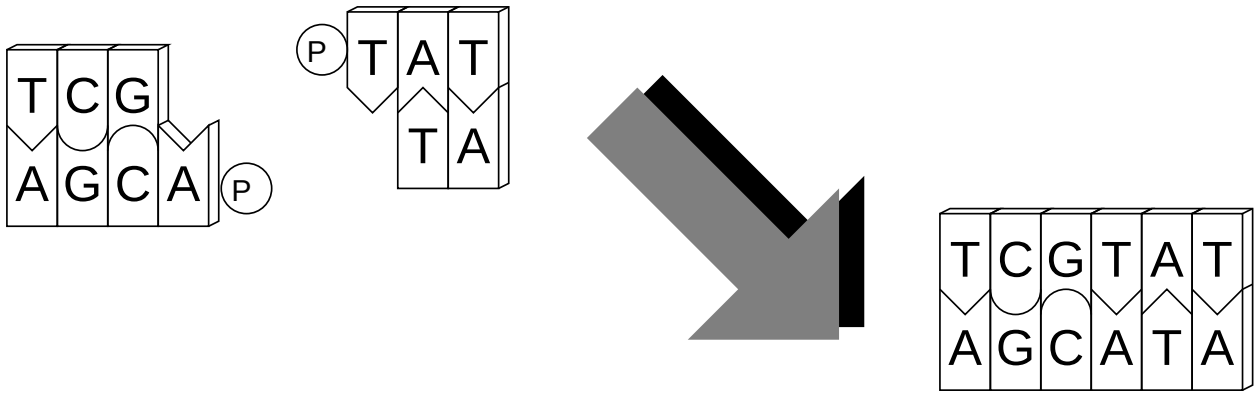
Wie wirken Enzyme?

Prinzip von „Schlüssel und Schloß“

- Ein Teil der Enzymoberfläche ist komplementär zum Substratmolekül (z.B. DNA!) geformt.
- Enzym kann an das Substratmolekül andocken und es eng umschließen.
- Herausbildung eines Enzym-Substrat-Komplexes
- eigentliche Reaktion: eine chemische Bindung des Substratmoleküls aufbrechen und schrittweise umordnen
- Enzym-Substrat-Komplex $\longrightarrow$ Enzym-Produkt-Komplex
- Enzym löst sich wieder vom Substrat ab (Dissoziation) und
- steht für die Katalyse einer gleichartigen Reaktion an einem anderen Substratmolekül zur Verfügung

Ligation

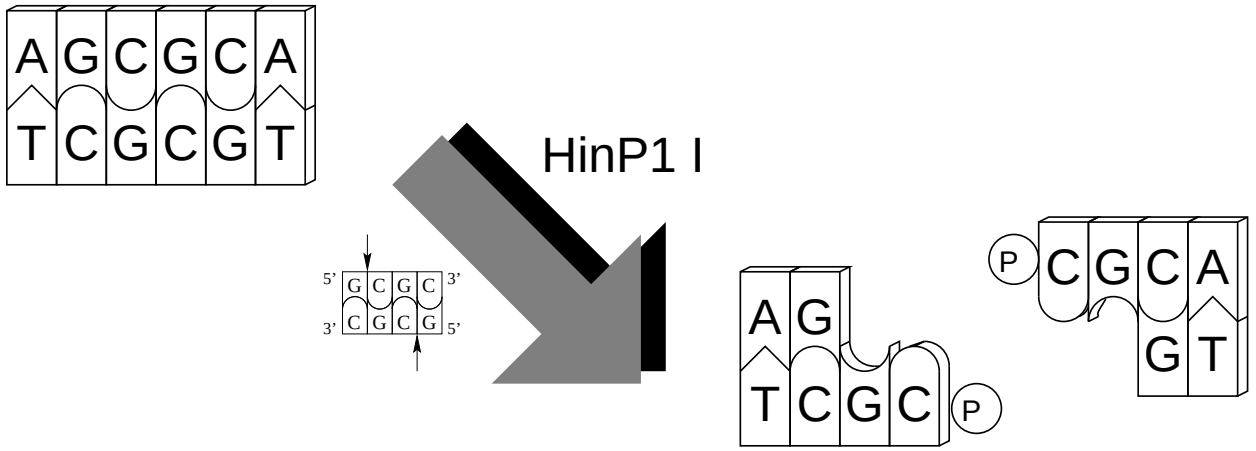
- Synonyme: Concatenate, Verketten



- Verketten von endenkompatiblen DNA-Doppelsträngen

Cut

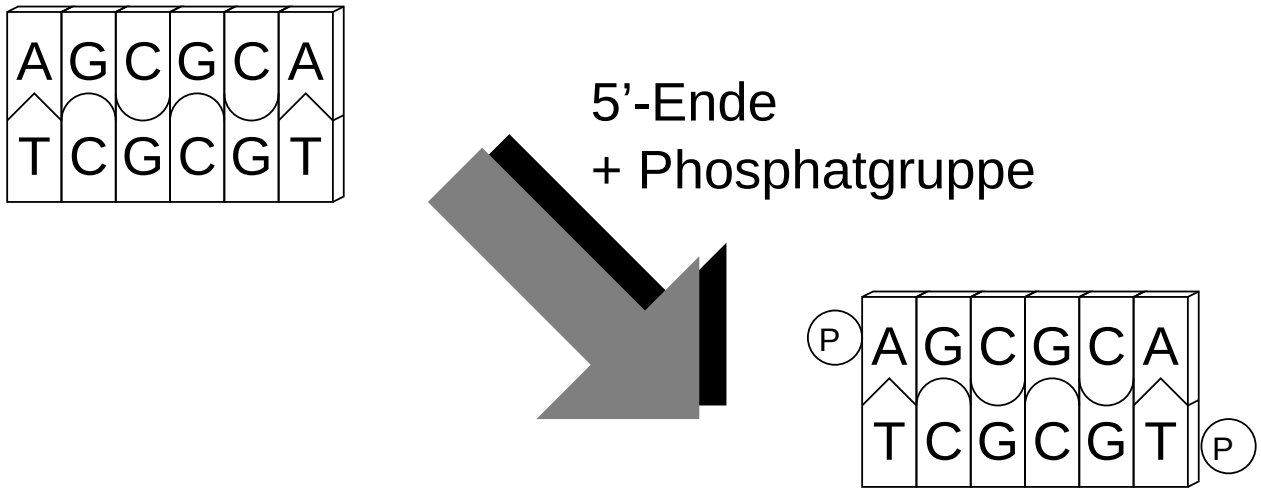
- Synonyme: Digestion, Cleavage, Verdau, Schnitt



- Zerschneiden von DNA-Doppelsträngen

Labeling

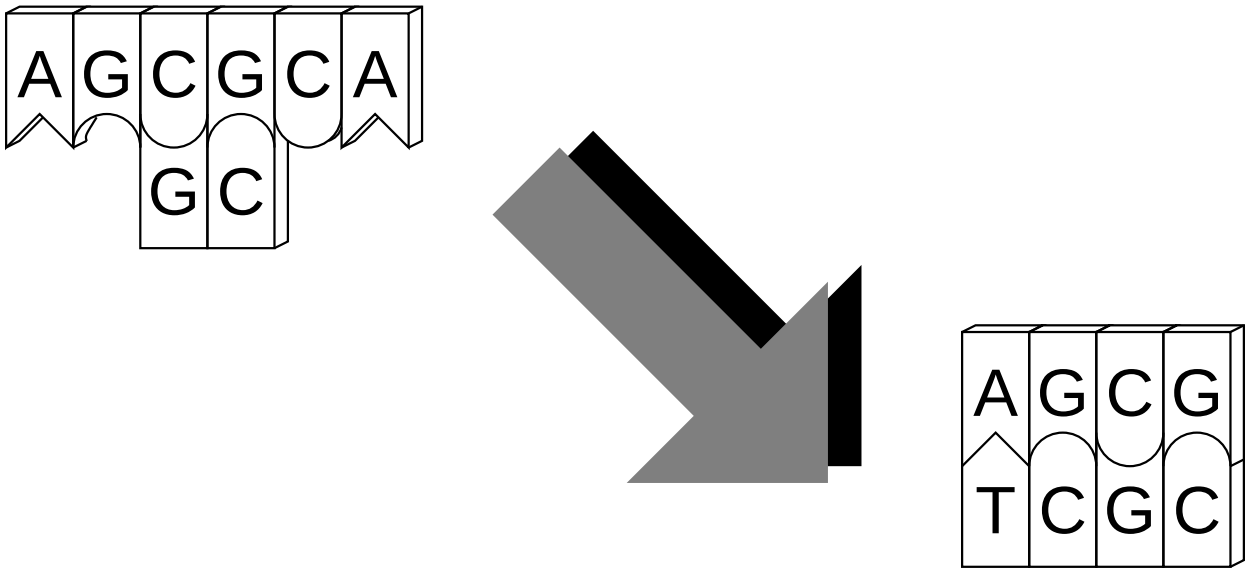
- Synonyme: Markieren, Labeln



- Markieren von Strangenden mit geeigneten Gruppen

Polymerisation

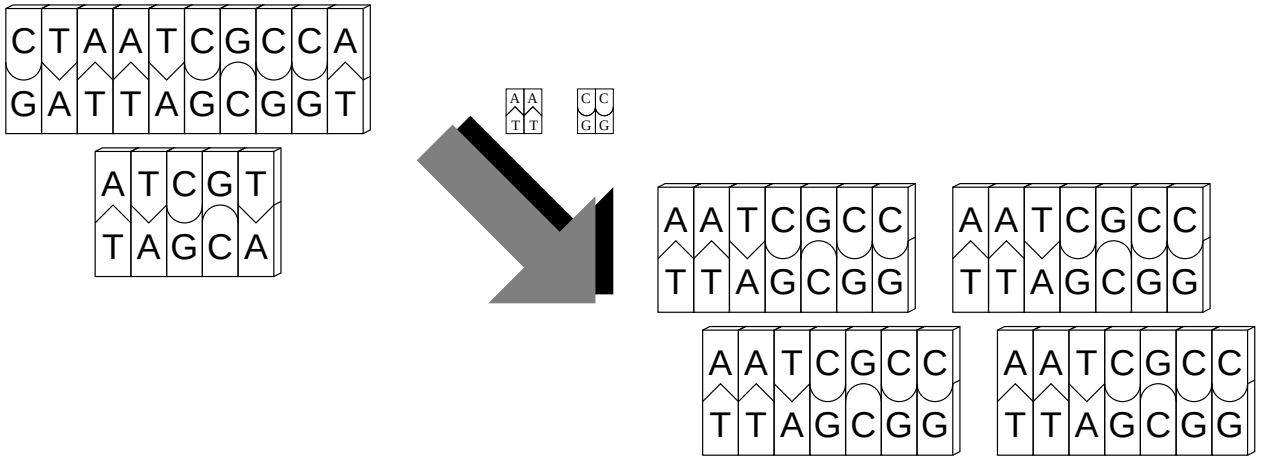
- Synonyme: Blunting



- Auffüllen/Abbauen von Einzelstrangüberhängen

Polymerase Chain Reaction

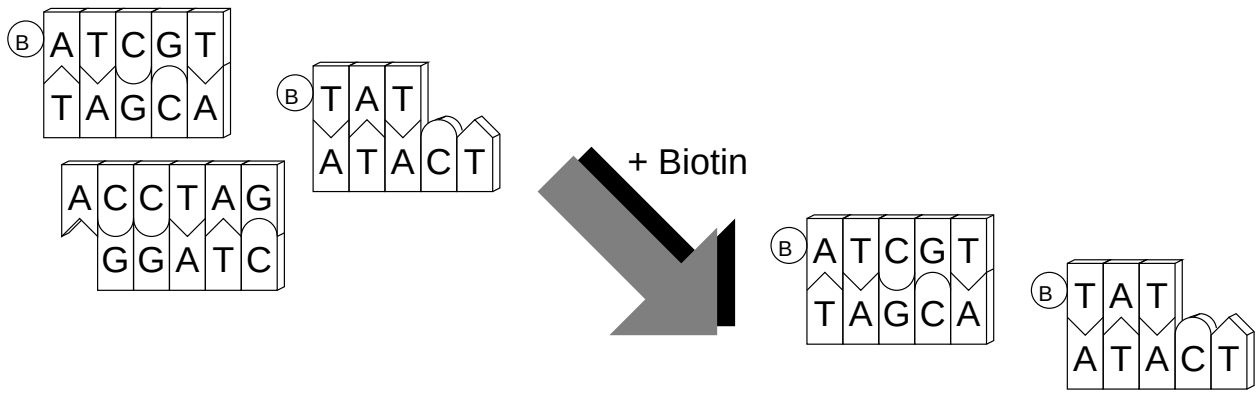
- Synonyme: Polymerase-Kettenreaktion, PCR, Amplify, Duplizieren



- Duplizieren von DNA-Doppelsträngen oder -abschnitten

Affinity Purification

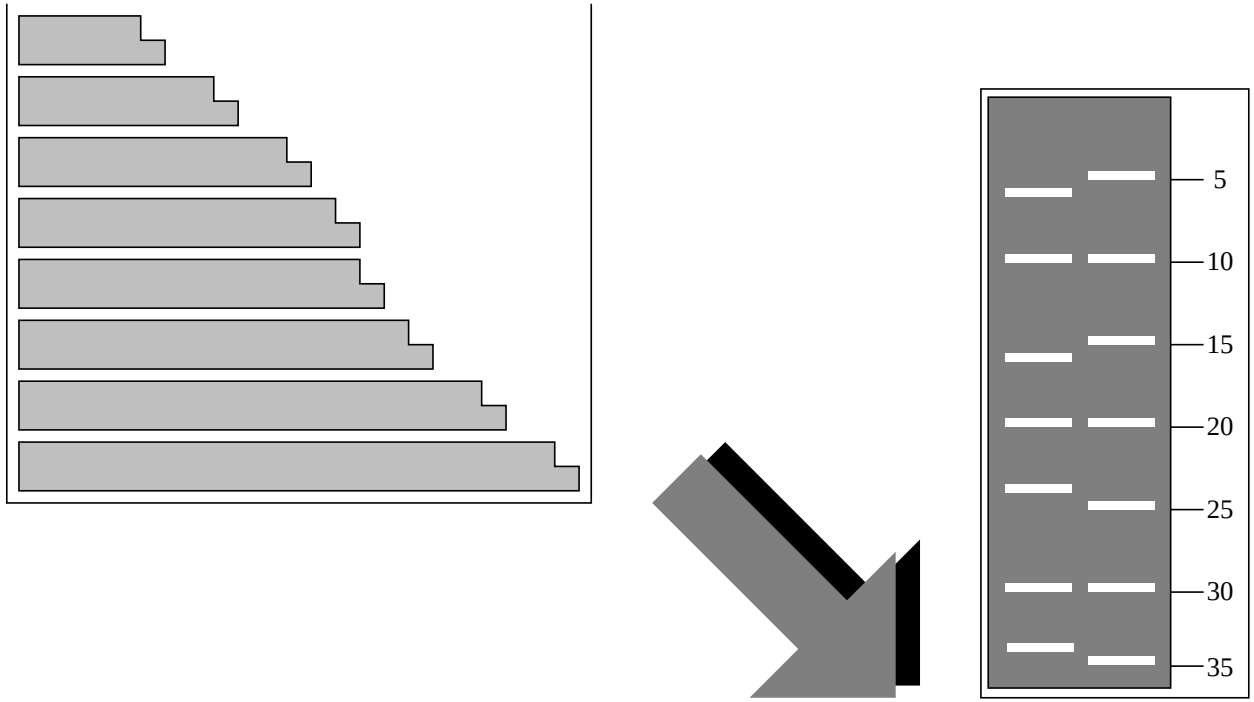
- Synonyme: Auswaschen, Reinigen



- Separieren nach Strangendenmarkierung bzw. Subsequenz

Agarose Gel Electrophoresis

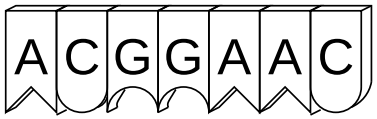
- Synonyme: Detect, Agarosegel-Elektrophorese

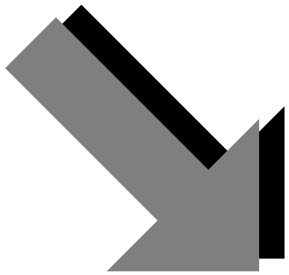


- Stranglängenbestimmung, Längenseparation

Sequencing

- Synonyme: Sequenzieren





5′-ACGGAAC-3′

- Sequenzbestimmung von DNA-Einzelsträngen

Praktisches Rechnen mit DNA am Beispiel

Rucksackproblem (knapsack problem)

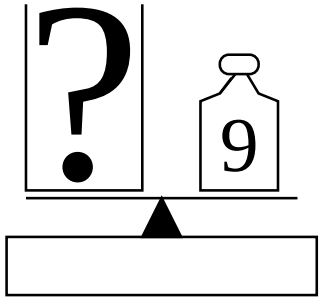
gegeben: n natürliche Zahlen $a_1, \ldots, a_n$ sowie eine natürliche Zahl b

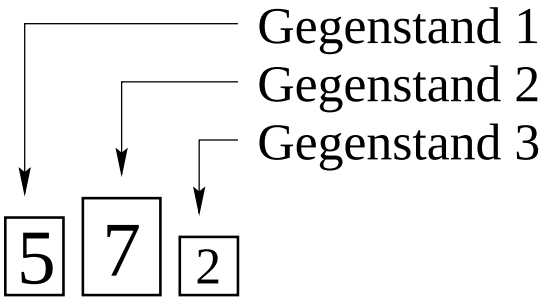
gesucht: Gibt es eine Teilmenge $I \subseteq \{1, 2, \ldots, n\}$ mit $\sum_{i \in I} a_i = b$?

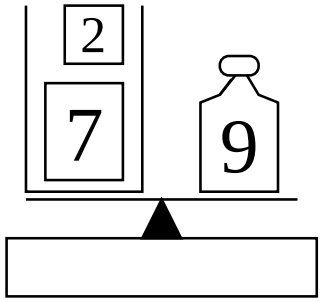
das heißt: Gibt es eine Packmöglichkeit mit einer Auswahl aus diesen Gegenständen, so daß genau das Rucksackgewicht b entsteht?

Beispiel eines Rucksackproblems

$a_1 = 5$
$a_2 = 7$
$a_3 = 2$
$b = 9$







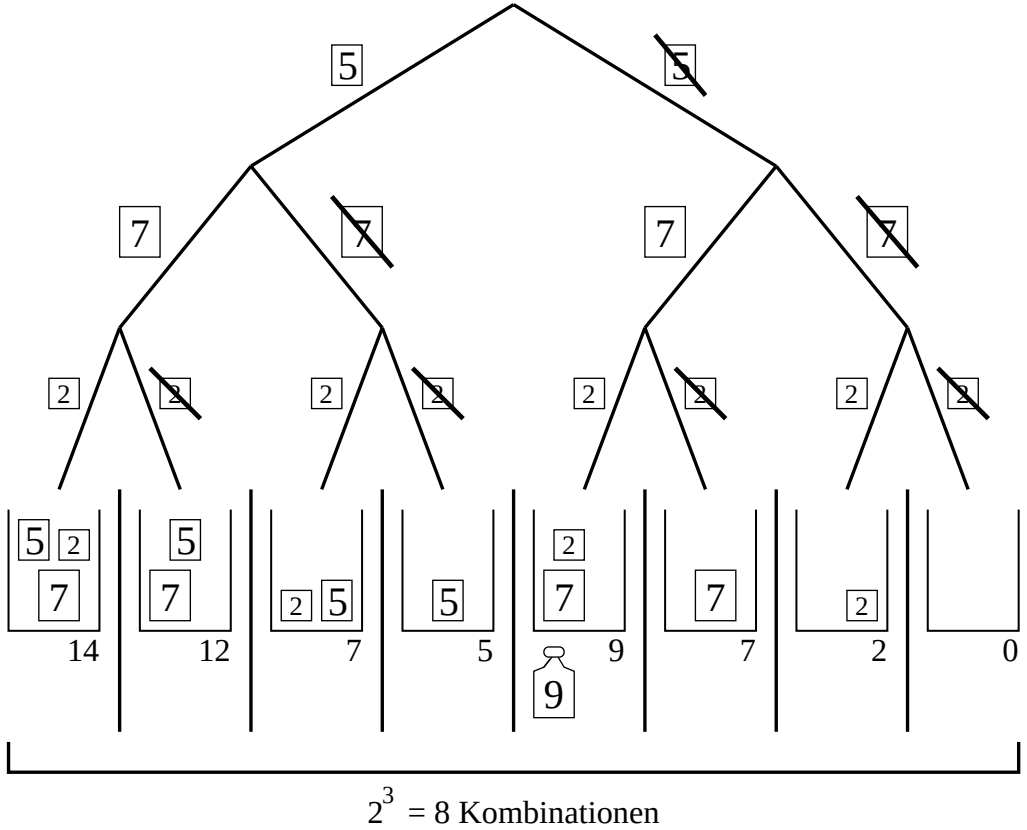
Lösung: JA

Packmöglichkeit:

{Gegenstand2; Gegenstand3}

Rucksackproblem – Lösung durch vollständige Enumeration

Erzeugen und prüfen aller potentiellen Lösungsmöglichkeiten (Kombinationen):



2^3 = 8 Kombinationen

Rucksackproblem – Einordnung und Eigenschaften

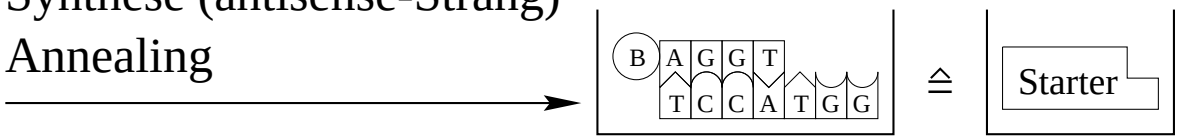
- kombinatorisches Suchproblem
- gehört zur Klasse der NP-vollständigen Probleme
- Erhöhung der Problemgröße um 1 Objekt verdoppelt den Rechenaufwand ($2^{n+1} = 2 \cdot 2^n$)
- bei exakter Lösung auf sequentiellem Rechner exponentieller Zeitbedarf
- auf konventioneller Rechentechnik etwa bis zu einer Problemgröße $n \approx 40$ beherrschbar
- Bei größeren Problemen Einsatz effizienter Näherungsverfahren, die aber nicht in jedem Fall eine exakte Lösung garantieren

Kodierung der Gegenstandsgewichte in DNA-Doppelstränge

Starter:

- DNA-Doppelstrang frei wählbarer Sequenz und Länge,
- ein Strangende
  - ohne Einzelstrangüberhang (blunt) und
  - geblockt (5'-biotinyliert)
- anderes Strangende mit Einzelstrangüberhang (sticky)

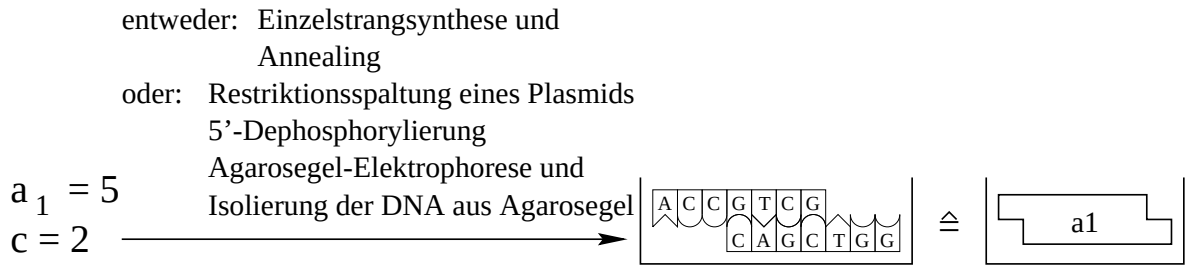
Synthese (sense-Strang) und
Labeling mit 5'-Biotin
Synthese (antisense-Strang)
Annealing



Kodierung der Gegenstandsgewichte in DNA-Doppelstränge

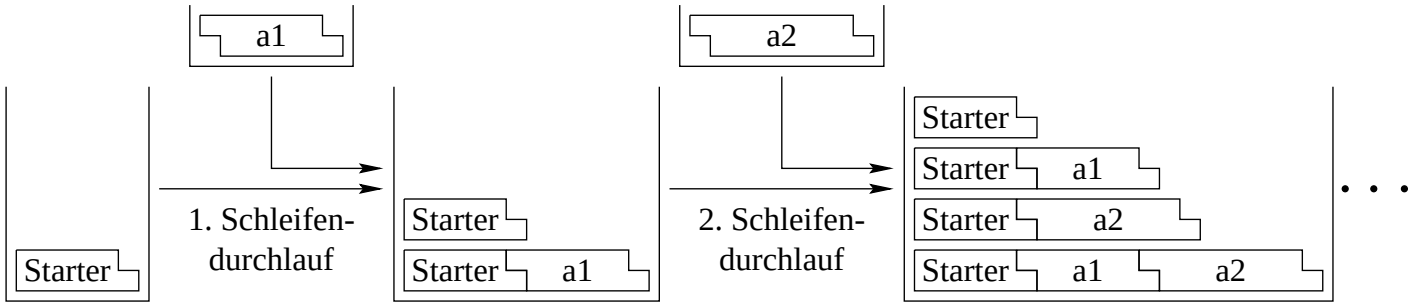
Für jeden Gegenstand i mit $i = 1, \ldots, n$:

- DNA-Doppelstrang frei wählbarer Sequenz
- Länge $l_i = c \cdot a_i$
- c: frei wählbarer konstanter Faktor; $c \geq 1, c \in \mathbf{N}$
- beide Strangenden mit Einzelstrangüberhang (sticky)
- sticky-Enden aller Gegenstands-DNA-Doppelstränge und des Starters müssen kompatibel sein!



DNA-Algorithmus zur Lösung des Rucksackproblems

- Aufbau aller 2^n Kombinationen durch gezieltes Verketten der DNA-Doppelstränge in n Schleifendurchläufen, danach Längentest
- Jeder Schleifendurchlauf nimmt einen bisher ungenutzten Gegenstand hinzu und verdoppelt die Anzahl der Kombinationen.
- In jedem Schleifendurchlauf dieselbe Abfolge biochemischer Reaktionen

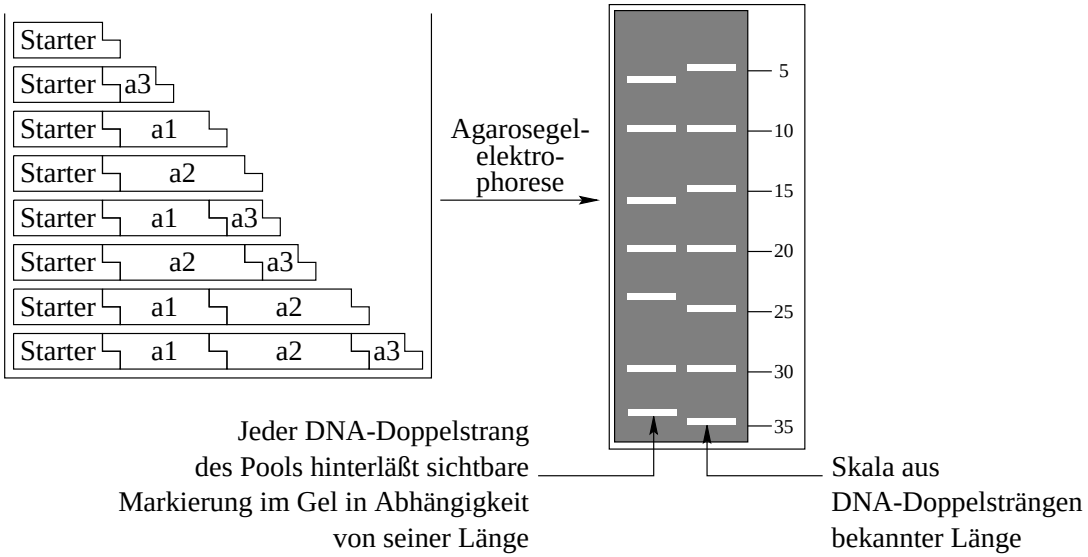


Der i-te Schleifendurchlauf

- Aufteilen des DNA-Pools in zwei Reaktionsgefäße
- 5'-Phosphorylieren der einen Hälfte
  - schafft Voraussetzung für Verkettung mittels Ligation
- Vereinigen beider Poolhälften und des DNA-Doppelstranges für Gegenstand i in ein gemeinsames Reaktionsgefäß
- Ligation
  - verkettet
  - 5'-phosphorylierte und
  - endenkompatible
  - DNA-Doppelstränge

Der Längentest

- Agarosegel-Elektrophorese des finalen DNA-Pools mit 2^n unterschiedlichen DNA-Doppelsträngen, dabei prüfen, ob DNA-Doppelstränge der Länge $Starterlänge + c \cdot b$ existieren
  - Methode zur Längenseparation und Visualisierung von DNA-Doppelsträngen



Bewertung des Algorithmus

- Anzahl Arbeitsschritte linear (und damit polynomiell) in der Anzahl Gegenstände
- prinzipiell skalierbar auf beliebige Problemgrößen (Anzahl Gegenstände), dann jedoch gehäuft Fehler durch Seiteneffekte der biochemischen Reaktionen
- Ausführung mit 3 Gegenständen im Labor beanspruchte 4 Tage, für jeden weiteren Gegenstand ein Tag mehr veranschlagt

DNA-Computing als universelles Berechnungsmodell

- mit dem genannten Satz an Operationen alle NP-Probleme in polynomiellem Zeitaufwand lösbar
- Verschiebung des exponentiellen Aufwandes von der Zeit zum Speicherplatz
- universelle Berechnungsmodelle (Turingmaschine, λ-Kalkül, Klasse der μ-rekursiven Funktionen, Klasse der rekursiv aufzählbaren Sprachen, ...) mittels des DNA-Operationssatzes in der Theorie vollständig simulierbar
- d.h., DNA-Computing besitzt (in der Theorie) die Berechnungskraft des PC, bei laborpraktischer Ausführung längerer Operationsfolgen jedoch Aufsummierung der Seiteneffekte bis zur Unbrauchbarkeit des Ergebnisses
- universelle Berechnungsmodelle des DNA-Computing auf hohem Abstraktionsniveau, z.B. DNA-PASCAL, Splicing-Systeme, Sticker-Modell

Ausblick und Visionen – Quo vadis DNA-Computing?

- bisher eine Reihe kombinatorischer Suchprobleme kleiner Größe im Labor gelöst, man sucht aber die „killer application“
- DNA-Operationen wurden analysiert, sie laufen im Labor jedoch nicht fehlerfrei ab, es treten gehäuft Seiteneffekte auf, die eliminiert werden müssen
- Aufstellung von DNA-Computing-Modellen auf laborpraktisch implementierbarem Abstraktionsniveau unter Erfassung möglichst vieler Seiteneffekte (z.B. DNA-HASKELL)
- Vision: Schaffung des laborpraktisch implementierten handhabbaren DNA-Universalcomputers

vielseitige interdisziplinäre Betätigungsfelder in Informatik, Molekularbiologie und Biochemie