Universelle Modelle und ausgewählte Algorithmen des DNA-Computing

Vortragender: Dipl.-Inform. Thomas Hinze

1. **Motivation** – Idee und Potential des DNA-Computing
2. **Mathematische Grundlagen** des DNA-Computing
3. **Molekularbiologische Grundlagen** des DNA-Computing
4. **Universelle Modelle** des DNA-Computing
5. **TT6** – ein anwendungsorientiertes universelles verteiltes Splicing-System
6. **Sisyphus** – ein labornahes Modell des DNA-Computing
7. **Zusammenfassung**, Veröffentlichungen

Motivation - Idee des DNA-Computing

Entwicklungsstand

- Komponente des „Future Computing" $\rightarrow$ alternative Hardware
- **Operationen:** molekularbiologische Prozesse auf Datenträger DNA
- **Potential:** massive Datenparallelität (bis $10^{20} ops/s$)
- Anfangserfolge und Fortschritte im Labor (NP in Polynomialzeit)
- Entwicklung und Analyse von Modellen des DNA-Computing

Vision

- Etablierung eines DNA-basierten Universalcomputers in Theorie und Labor

Herausforderungen

- Behandlung der Seiteneffekte von DNA-Operationen
- Annäherung formaler Mod. des DNA-Computing an Laborimpl.

Motivation – DNA als Datenträger

Eigenschaften im Hinblick auf DNA-Computing

- hohe Speicherkapazität und -dichte (bis $10^{21} bp/l$, $\approx 1 bit/nm^3$)
- redundante, dezentrale, verlustsichere Datenspeicherung mgl.
- Langlebigkeit $\rightarrow$ persistentes Speichermedium
- richtungsbehaftete lineare DNA $\rightarrow$ leichte Kodierung von Daten
- breites Spektrum molekularbiol. Operationen, in-vitro-Handling (Synthese, Rekombination, Separation, Analyse)
- Wiederverwendbarkeit von DNA in Berechnungsprozessen mgl.
- umweltfreundliche, energieeffiziente Verarbeitung ($2 \cdot 10^{19} ops/J$)
- DNA-Konservierung, -verarbeitung $\rightarrow$ kein mech. Verschleiß
- Operationsspektrum gestattet massive Datenparallelität und Simulation universeller Modelle der Berechenbarkeit

Mathematische Grundlagen des DNA-Computing

Berechenbarkeitstheorie

- mathematische (formale, abstrakte) Beschreibung von Rechenvorgängen auf Basis zugrundeliegender Konzepte
- Klassifizierung resultierender Modelle der Berechenbarkeit nach ihrer Berechnungsstärke, Turing-Berechenbarkeit, Universalität
- konventionelle universelle Modelle der Berechenbarkeit, auf denen universelle Modelle des DNA-Computing aufsetzen:
  - deterministische und nichtdeterministische Turingmaschine
  - Klasse der μ-rekursiven Funktionen
  - Klasse der WHILE-Programme
  - Chomsky-Grammatiken für rekursiv aufzählbare Sprachen
- ihre gegenseitige Simulation (Modelltransformationen)
- Algorithmusbegriff

Mathematische Grundlagen des DNA-Computing

Komplexitätstheorie

- Klassifizierung algorithmisch lösbarer Probleme hinsichtlich des Ressourcenbedarfs
  - Zeit
  - Speicherplatz
  - Anzahl Verarbeitungseinheiten

  bei Bearbeitung auf Modellen der Berechenbarkeit
- Komplexitätsmaße für Algorithmen
- Komplexitätsklassen $\mathcal{P}$ und $\mathcal{NP}$
- Problemtransformationen
- Effizienzoptimierung von Algorithmen

Molekularbiol. Grundlagen des DNA-Computing

Operationen auf DNA (Auswahl)

Gewinnen von DNA-Strängen

- **Synthesis** (Einzelstrangsynth.)

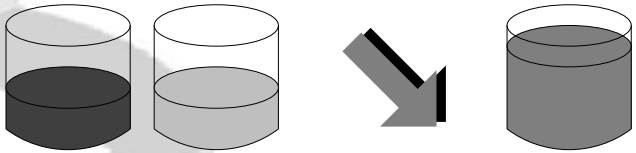
Knüpfen und Aufbrechen von Wasserstoffbrückenbindungen

- **Annealing** (Hybridisierung)

- **Melting** (Denaturierung)

Mischen und Aufteilen von DNA-Lösungen

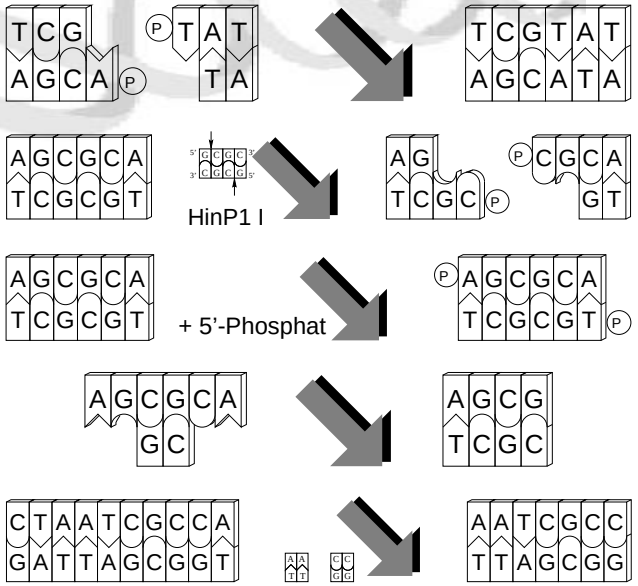
- **Union** (Mischen)



Molekularbiol. Grundlagen des DNA-Computing

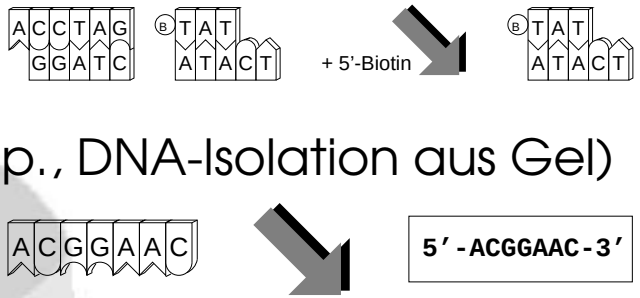
Enzymatische Reaktionen

- **Ligation** (endenkomp. Verkettung)
- **Digestion** (gezielte Spaltung)
- **Labeling** (Strangendenmodifik.)
- **Polymerisation** (Endenglättung)
- **PCR** (Polymerase-Kettenreaktion)



Separieren und Analysieren von DNA-Strängen

- **Affinity Purification** (Biotinsep.)
- **Gel Electrophoresis** (Längenbest., -sep., DNA-Isolation aus Gel)
- **Sequencing** (Nucleotidfolgebest.)



Universelle Modelle des DNA-Computing

Modelle des DNA-Computing

- spezielle Modelle der Berechenbarkeit, bei denen verarbeitete Daten durch abstrahierte Abbildungen von DNA modelliert sind (DNA-basierte Daten), die durch spezifische Sätze darauf einwirkender Operationen modifiziert werden (Generierung und selektive Modifikation DNA-basierter Daten)

Ziele

- Analyse des DNA-Computing aus berechenbarkeitstheoretischer und komplexitätstheoretischer Sicht (**welche Problemklassen** mit **welchem Aufwand wie** lösbar)
- Schaffung eines mgl. labornahen Modells eines **DNA-basierten Universalcomputers**, für effiziente Lösung rechen- und speicherintensiver algorithmischer Probleme vorteilhaft einsetzbar
- spezifische Modelloptimierungen

Universelle Modelle des DNA-Computing

Entwicklungsstand – Vielzahl Modelle, Unterschiede hinsichtlich

- laborpraktischer Implementierbarkeit, Abstraktionsniveau
- Existenz einer formalen Modellbeschreibung
- Berechnungsstärke des Modells
- Organisation von Daten- und Steuerfluß
- Portierbarkeit, Transformierbarkeit und Effizienz von Algorithmen

Auswahlkriterien

- universelle und platzbeschränkt universelle Modelle des DNA-Computing
- formal beschreibbar, beruhen auf ausschließlicher Nutzung
- linearer DNA, modelliert durch geeignete formale Sprachen

Universelle Modelle des DNA-Computing

Untersuchte Modelle und deren Eigenschaften

ausgewählte Modelle des DNA-Computing	universell	platzbeschränkt universell	restriktiv	multimengenbasiert	deterministisch	imperativ	regelbasiert	Multiple-Instruction-fähig	Multiple-Data-fähig	direkte Algorithmen-transformation von
Filtering-Modelle		■			■	■			■	WHILE
Parallel Associative Memory		■		■	■	■			■	TM
DNA-Pascal	■				■	■			■	WHILE
DNA Equality Checking	■				■	■			■	WHILE
Insertion-Deletion-Systeme	■						■	■	■	G
Watson-Crick D0L-Systeme	■		■		■				■	μ
Splicing-Systeme (H-, EH-)	■						■	■	■	G

WHILE: WHILE-Programm
TM: Turingmaschine
G: Chomsky-Grammatik vom Typ 0
μ: μ-rekursive Funktion, beschrieben im μ-Rekursionsschema

Rucksackproblem

kombinatorisches Suchproblem, NP-vollständig

Problemdefinition

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

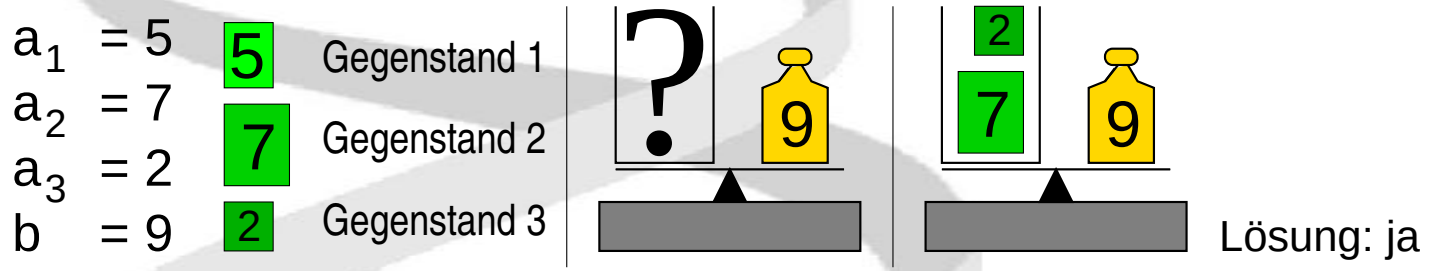
gesucht: Gibt es eine Teilmenge $I \subseteq \{1, \ldots, n\}$ mit $\sum_{i \in I} a_i = b$?

Veranschaulichung

a_1 bis a_n: Gewichte der Gegenstände 1 bis n.

Gibt es eine Packmöglichkeit mit einer Auswahl aus diesen Gegenständen, so daß das Referenzgewicht b entsteht?

Beispiel



Splicing-System (EH-System)

Definition: Ein **EH-System** ist ein Quadrupel $\gamma = (V, \Sigma, A, R)$ mit dem Alphabet V, $\Sigma \subseteq V$, $A \subseteq V^*$, $R \subseteq V^*\#V^*\$V^*\#V^*$ und $\#, \$ \notin V$.

- V bezeichnet das Alphabet von γ,
- Σ das Alphabet der Terminalsymbole,
- die formale Sprache A die Menge der Axiome,
- die formale Sprache R die Menge der Splicing-Regeln.
- Die Elemente von Σ werden als Terminalzeichen und die Elemente von $V \setminus \Sigma$ als Nichtterminalzeichen bezeichnet.

Ausgehend von den Axiomen erfolgt eine fortlaufende Anwendung von Splicing-Regeln, wobei Zeichenketten erzeugt werden.

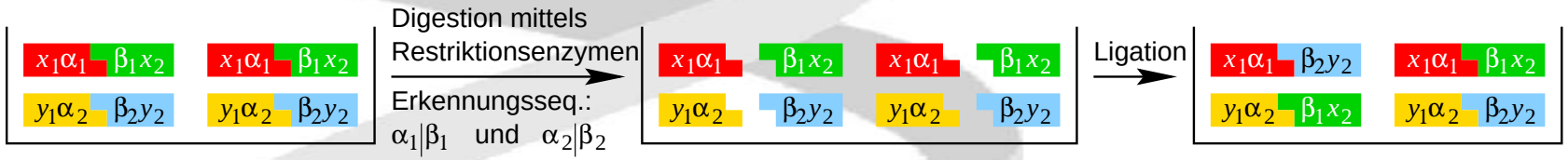
EH-Systeme γ generieren somit formale Sprachen $L(\gamma)$.

Splicing-System (EH-System)

Splicing-Operation (Splicing-Regel)

- basiert unmittelbar auf DNA-Operationen
- **Definition** (T. Head, 1987): Sei V ein Alphabet sowie $\$$ und $\#$ zwei Symbole $\notin V$. Eine Splicing-Regel über V ist ein Wort $r = \alpha_1\#\beta_1\$\alpha_2\#\beta_2$ mit $\alpha_i, \beta_i \in V^*$, $i \in \{1,2\}$. Für jedes $r \in R$ und die Wörter $x, y, w, z \in V^*$ wird definiert:

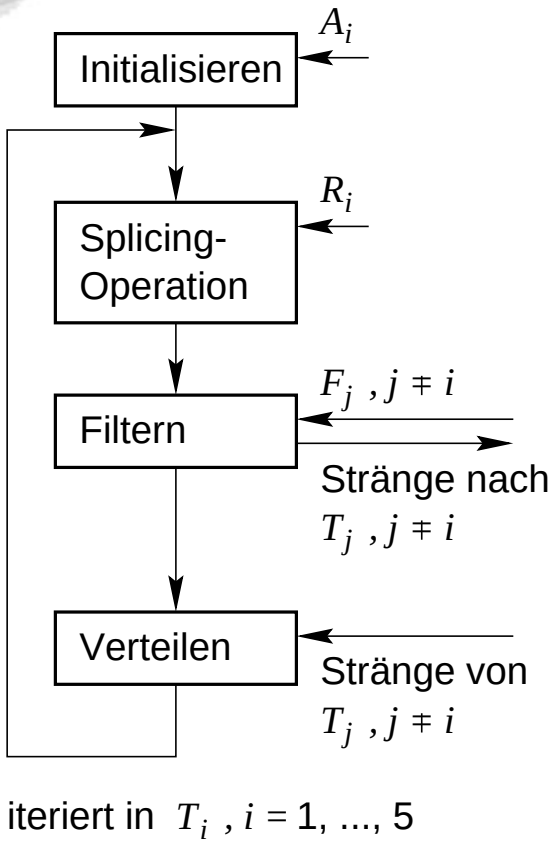
$$(x,y) \to_r (z,w) \text{ gdw. } \quad x = x_1\alpha_1\beta_1x_2, \quad y = y_1\alpha_2\beta_2y_2,$$
$$z = x_1\alpha_1\beta_2y_2, \quad w = y_1\alpha_2\beta_1x_2.$$



Universelles verteiltes TT6-EH-System

Systemeigenschaften

- uneingeschränkte Universalität
- Endlichkeit aller Systemkomponenten
- statischer Systemaufbau
- Ressourcenminimierung
- ausschließliche Nutzung linearer DNA
- kein systeminhärenter Nichtdeterminismus
- Bereitstellung des Berechnungsergebnisses in separatem Reagenzglas T_6
- Beschreibung des Gesamtsystems basiert auf implementierbaren DNA-Operationen
- Anzahlminimierung zu verteilender Stränge
- optimierter PCR-basierter Filtermechanismus



Universelles verteiltes TT6-EH-System

Definition (Auszug): Ein TT6-EH-System (Test Tube 6 Ext. Head System) ist ein 8-Tupel $\Gamma = (V, \Sigma, T_1, T_2, T_3, T_4, T_5, T_6)$ mit folg. Komponenten:

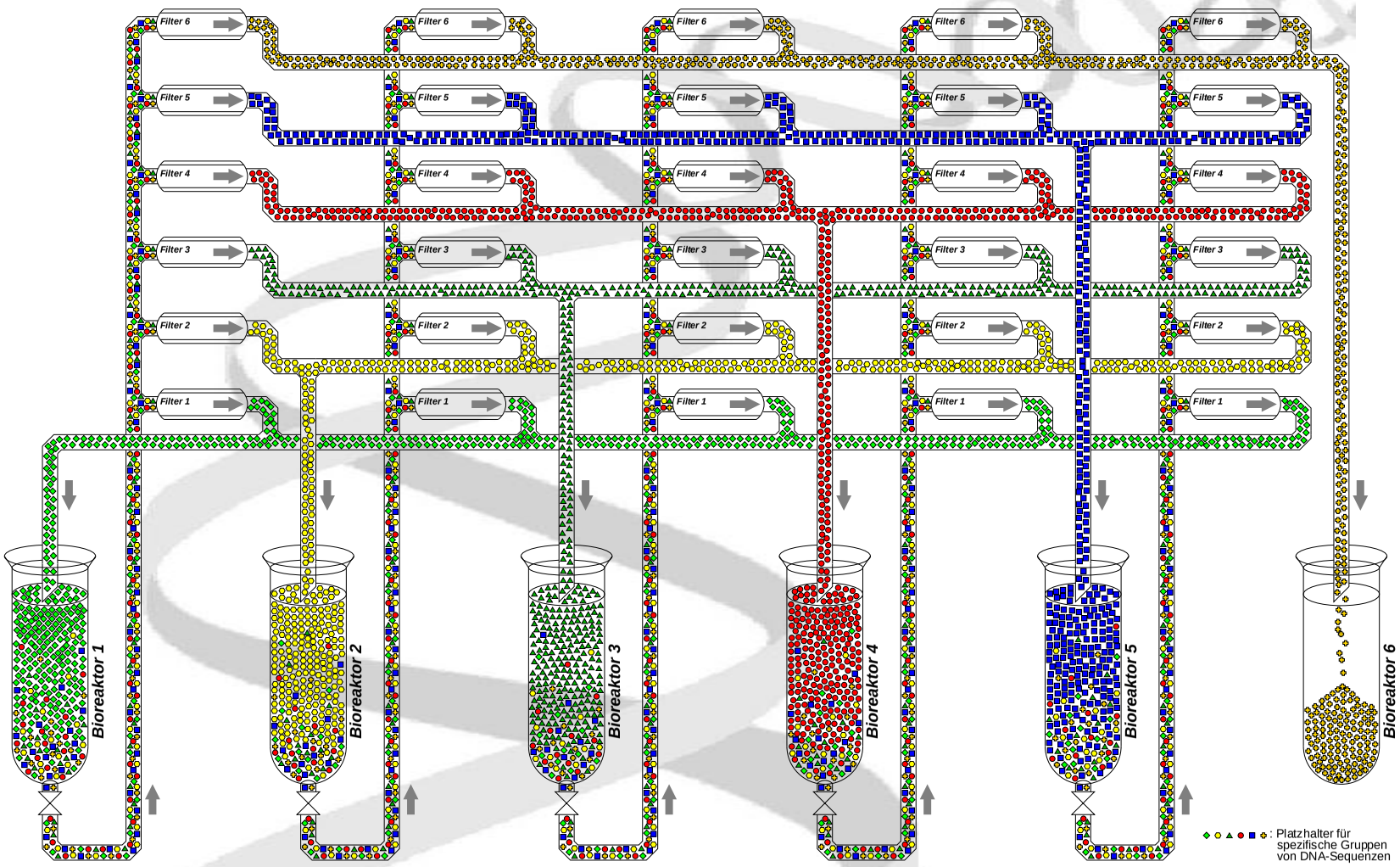
- V bezeichnet das Alphabet von Γ,
- $\Sigma \subseteq V$ das Alphabet der Terminalsymbole,
- jede Komponente T_i mit $i = 1, \ldots, 6$ ein Reagenzglas von Γ.

Jedes Reagenzglas T_i ist def. durch ein Tripel $T_i = (A_i, R_i, F_i)$, wobei

- $A_i \subset V^*$ die endliche Menge der Axiome,
- $R_i \subset V^* \otimes \{\#\} \otimes V^* \otimes \{\$\} \otimes V^* \otimes \{\#\} \otimes V^*$ die endliche Menge der Splicing-Regeln,
- $F_i \subset V^* \times V^*$ die endliche Menge der Filtermuster (Wortpaare) für T_i bezeichnet.

Universalitätsnachweis durch Transformation von Chomsky-Grammatiken des Typs 0 (o.B.d.A. Kuroda-Normalform) in TT6-EH-Systeme.

Universelles verteiltes TT6-EH-System, Prinzipskizze



Sisyphus – ein labornahes DNA-Computing-Modell

Modelleigenschaften

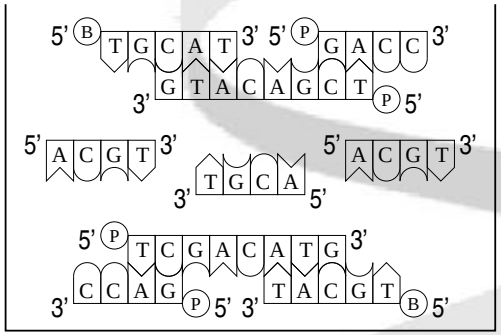
- uneingeschränkte Universalität
- Restriktivität
- Multimengenbasiertheit
- Nutzung des imperativen Paradigmas
- Multiple-Data-Fähigkeit
- niedriges Abstraktionsniveau der Modelloperationen
- Modellierung DNA-basierter Daten auf Ebene von Nucleotiden und Strangendenmarkierungen
- Seiteneffektberücksichtigung (über geeignete Parametrisierung)
- nichtdeterministischer Ansatz

Sisyphus bildet Basis für Simulationssystem molekularbiol. Prozesse

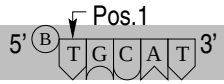
Sisyphus – ein labornahes DNA-Computing-Modell

Abbildung DNA-basierter Daten

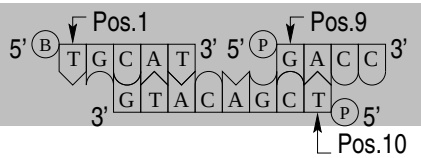
- **Nucleotidsequenz mit beidseitigen Strangendenmarkierungen:**
  Tripel aus $\{H, P, B\} \times \{A, C, G, T\}^+ \times \{H, P\}$
- **DNA-Strang:** definiert zusammengesetzt aus Nucleotidsequenzen mit beidseitigen Strangendenmarkierungen, linear
- **Reagenzglas:** definierte endliche Multimenge von DNA-Strängen



$\langle$ { {((H, ACGT, H),1)},



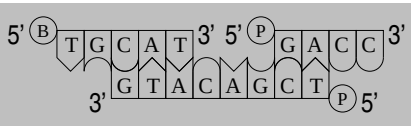
{((B, TGCAT, H),1),
((P, GACC, H),9),
((P, TCGACATG, H),-10)}



},

{ ({((H, ACGT, H),1)}, 3), **3 Exemplare** 5' ACGT 3'

({((B, TGCAT, H),1),
((P, GACC, H),9),
((P, TCGACATG, H),-10)}, 2) **2 Exemplare**



}

$\rangle$

Sisyphus – ein labornahes DNA-Computing-Modell

Parametrisierung der Modelloperationen

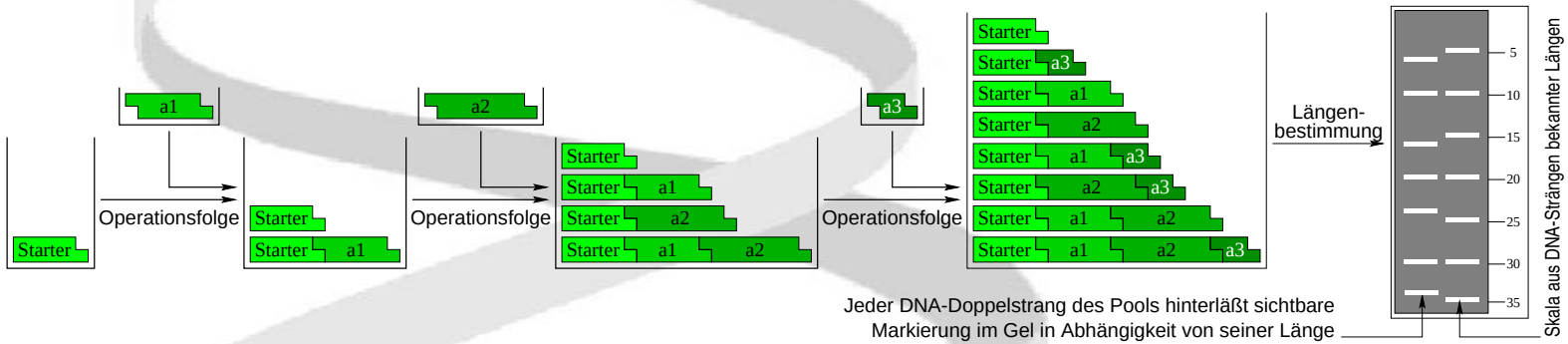
Modelloperationen		Seiteneffekt-parameter (in %): Punktmutationsrate	Deletionsrate	max. relative Deletionlänge	Basenfehlpaarungsrate	Rate unverarb. DNA-Stränge	Strangverlustrate
Synthesis	synthesis: $(\{H\}\times(\{A,C,G,T\}^+)\times\{H\}) \times (\mathbb{N}\setminus\{0\}) \times ([0,100]^2) \times \mathbb{N} \rightarrow$ Reagenzglas	■	■	■			
Annealing	annealing: Reagenzglas $\times (\langle 0,100]) \times ([0,100]^2) \times \mathbb{N} \rightarrow$ Reagenzglas				■	■	
Melting	melting: Reagenzglas $\times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas					■	
Union	union: Reagenzglas $\times$ Reagenzglas $\times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas						■
Split	split: Reagenzglas $\times (\mathbb{N}\setminus\{0\}) \times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas^m						■
Ligation	ligation: Reagenzglas $\times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas					■	
Digestion	digestion: Reagenzglas $\times$ S $\times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas					■	
Labeling	labP5set:, labP5remove:, labB5set: Reagenzglas $\times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas					■	
Polymerisation	polymerisation: Reagenzglas $\times ([0,100]^2) \times \mathbb{N} \rightarrow$ Reagenzglas	■				■	
Affinity Purification	sepB5on:, sepB5off: Reagenzglas $\times ([0,100]^2) \times \mathbb{N} \rightarrow$ Reagenzglas					■	■
Gel Electrophoresis	electrophoresis: Reagenzglas $\times (\mathbb{N}\setminus\{0\}) \times ([0,100]) \times \mathbb{N} \rightarrow P(\mathbb{N})$						■
Separation by Length	seplength: Reagenzglas $\times (\mathbb{N}\setminus\{0\}) \times ([0,100]) \times \mathbb{N} \rightarrow$ Reagenzglas						■

- Universalitätsnachweis durch Transformation von TT6-EH-Systemen in Operationsfolgen

Algorithmus zur Lösung des Rucksackproblems

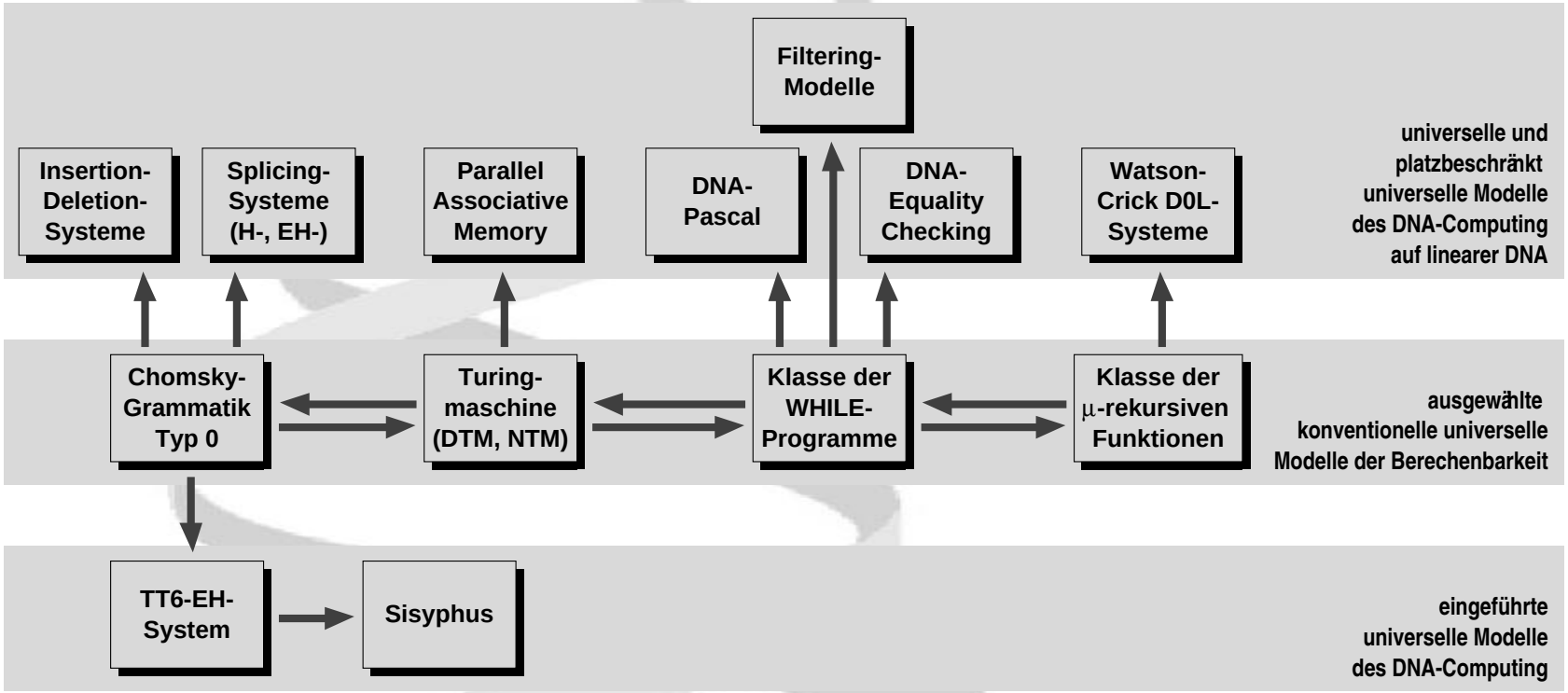
Brute-Force-Ansatz

- Generierung aller Packmöglichkeiten durch gesteuerte Verkettung der DNA-Doppelstr. mittels Folge von DNA-Operationen
- anschließende Längenseparation und Test, ob DNA-Doppelstr. der Länge *Starterlänge* $+ c \cdot b$ vorliegen $\rightarrow$ Lösung ja/nein
- laborpraktisch implementiert für Problemgröße $n = 3$
- Verlagerung des exp. Ressourcenbedarfs: Zeit $\rightarrow$ Speicherplatz
- praktische Skalierbarkeit durch Seiteneffekte der DNA-Operationen und verarbeitbare DNA-Mengen begrenzt ($n \approx 70$)



Zusammenfassung

Beschriebene Transformationen von Modellen der Berechenbarkeit



Zusammenfassung

Ergebnisse

- **durchgängig formale Beschreibung** universeller und platzbeschränkt universeller Modelle des DNA-Computing auf linearer DNA
- **Systematisierung** und **Klassifikation** von Modellen des DNA-Computing
- **einheitliche Konstruktion** und Vergleich von Algorithmen zur Lösung des NP-vollständigen **Rucksackproblems** für diese Modelle
- **Analyse molekularbiologischer Prozesse** und entsprechender Labortechniken, darauf abgestimmte Modellspezifikation
- labornahe Modellierung eines **DNA-basierten Universalcomputers**

Nutzen

- Arbeit liefert einen Beitrag, um die Lücke zwischen laborpraktischem und theoretischem DNA-Computing zu schließen
- Festigung des DNA-Computing als Computingkonzept
- Konkretisierung des Entwurfs DNA-basierter Universalcomputer
- Beitrag zum Berechenbarkeitsbegriff, Adaption auf DNA-Computing

Ausgewählte Veröffentlichungen

T. Hinze, U. Hatnik, M. Sturm. An Object Oriented Simulation of Real Occurring Molecular Biological Processes for DNA Computing and Its Experimental Verification. In N. Jonoska, N.C. Seeman, eds., DNA Computing. Proceedings *Seventh International Meeting on DNA Based Computers (DNA7),* Tampa, FL, USA, 2001. *LNCS Series,* ISBN 3-540-43775-4, Vol. 2340, pp. 1–13, Springer Verlag, 2002

M. Sturm, T. Hinze. Distributed Splicing of RE with 6 Test Tubes. *Romanian Journal of Information Science and Technology,* ISSN 1453-8245, Editura Academiei Romane **4(1–2)**:211–234, 2001

U. Hatnik, T. Hinze, M. Sturm. A Probabilistic Approach to Description of Molecular Biological Processes on DNA and Their Object Oriented Simulation. In V.V. Kluev, N.E. Mastorakis, eds., Proceedings *WSES International Conference on Simulation (SIM2001),* ISBN 960-8052-40-8, Malta, 2001

E.P. Stoschek, M. Sturm, T. Hinze. DNA-Computing – ein funktionales Modell im laborpraktischen Experiment. *Informatik Forschung und Entwicklung,* ISSN 0178-3564, Springer Verlag **16(1)**:35–52, 2001

T. Hinze, M. Sturm. Towards an in-vitro Implementation of a Universal Distributed Splicing Model for DNA Computation. In R. Freund, ed., Proceedings *Theorietag 2000,* ISBN 3-85028-325-9, TU Wien, Austria, pp. 185–189, 2000

T. Hinze, M. Sturm. A universal functional approach to DNA computing and its experimental practicability. In A. Condon, G. Rozenberg, eds., Preliminary Proceedings *Sixth International Meeting on DNA Based Computers (DNA6),* University of Leiden, The Netherlands, p. 257, 2000 und *Technical Report TUD-FI00-05,* ISSN 1430-211X, TU Dresden, 2000

E.P. Stoschek, M. Sturm, T. Hinze, et. al. Molekularbiologisches Verfahren zur Lösung von NP-Problemen. *Deutsches Patent DE 198 53 726 A1,* IPC C12N 15/10, Deutsches Patentamt München, 2000