Center for Computational Sciences and Simulation CCSS

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Genominformatik: DNA-Sequenzanalysen mit GPUs

Johannes Köster, Sven Rahmann
Lehrstuhl für Genominformatik, Medizinische Fakultät

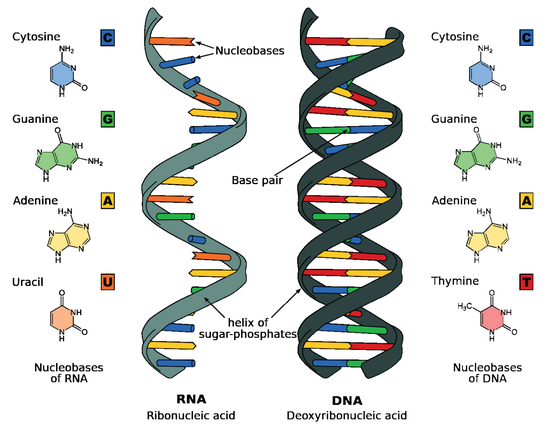
DNA- und RNA-Sequenzierung

DNA: Träger der genetischen Information einer Zelle.
Kettenmolekül aus 4 Basen: A,C,G,T, Doppelstrang
Menschliches Genom: ca 3.2 Gbp (Giga Basenpaare).

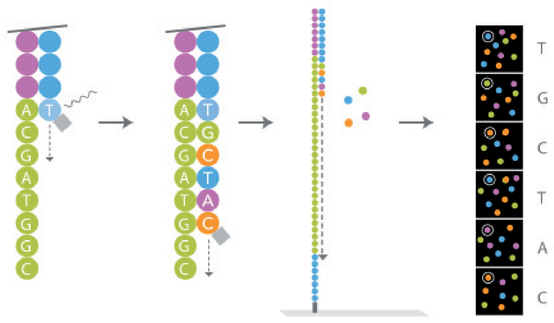
RNA: Kopien von Abschnitten (Genen) der DNA.
Kettenmolekül aus 4 Basen: A, C, G, U, Einzelstrang

Sequenzierung: Bestimmung der DNA-/RNA-Sequenzen einer Probe (Tumorgewebe, Blut, aber auch: Bakterien).
Liefert kurze Abschnitte (“Reads”, 50 – 400 bp)

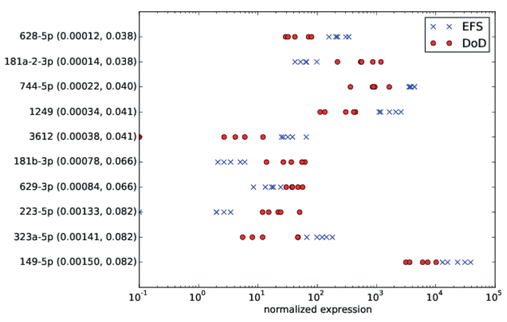
Anwendungen (Medizin, Biologie), Beispiele:
- Vergleich Normalgenom, Tumorgenom
- Analyse der Genaktivität (Genexpression) in verschiedenen Gewebetypen
- Identifikation bisher unbekannter Gene
- Quantifizierung von Biodiversität in Umweltproben



Unterschiede RNA / DNA



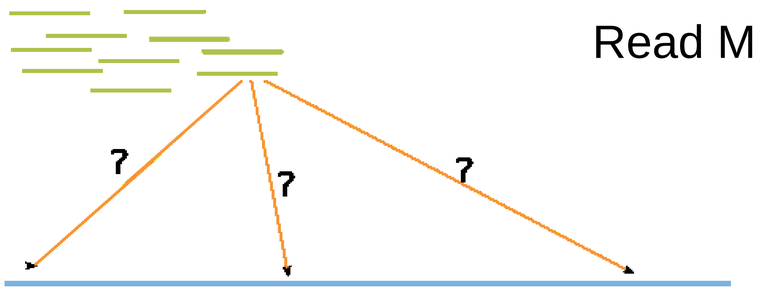
DNA-Sequenzierung



Differentiell exprimierte microRNAs im Tumor Neuroblastom

Read Mapping Problem

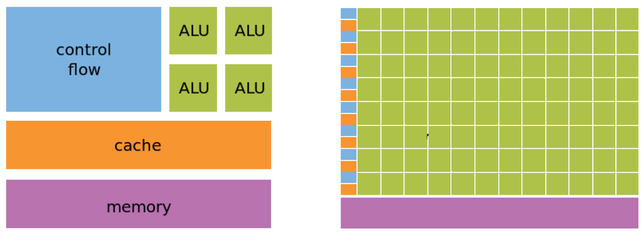
Gegeben: Sequenz des Genoms (3 Gbp),
Sequenzen von hunderten Millionen Reads (10 – 100 Gbp).
Gesucht: Lokalisierung jedes Reads im Genom.
Anforderungen: schnell, fehlertolerant.
Naiv: 3G * 100G = 3 * 10^{21} Zeichenvergleiche



Read Mapping

Unsere Lösungsansätze

1. Indexbasierte Verfahren:
   Vorberechnung der Orte von *k*-meren im Genom
2. Parallelisierung der Vergleiche mit GPUs



Aufbau CPU vs. Aufbau GPU

Herausforderungen

- wenig(er) Speicher auf der Grafikkarte
- Kontrollflussanweisungen (if-else) stören Parallelität

Vorgehen

- Hierarchischer binärer *k*-mer Index:
  Für verschiedene Regionen, kommt das k-mer vor?
- Alignments berechnen
  mit bit-parallemem Algorithmus von Myers