TECHNISCHE UNIVERSITÄT DRESDEN

**Fakultät Informatik** Institut für Technische Informatik, Professur für VLSI-Entwurfssysteme

# BESCHLEUNIGUNG DES SABA-ALIGNMENTS

## Diplomvortrag

**Frank Hoffmann**

Dresden, 27.08.2014

Problemstellung

Parallelisierung

Auswertung/Performance

Zusammenfassung

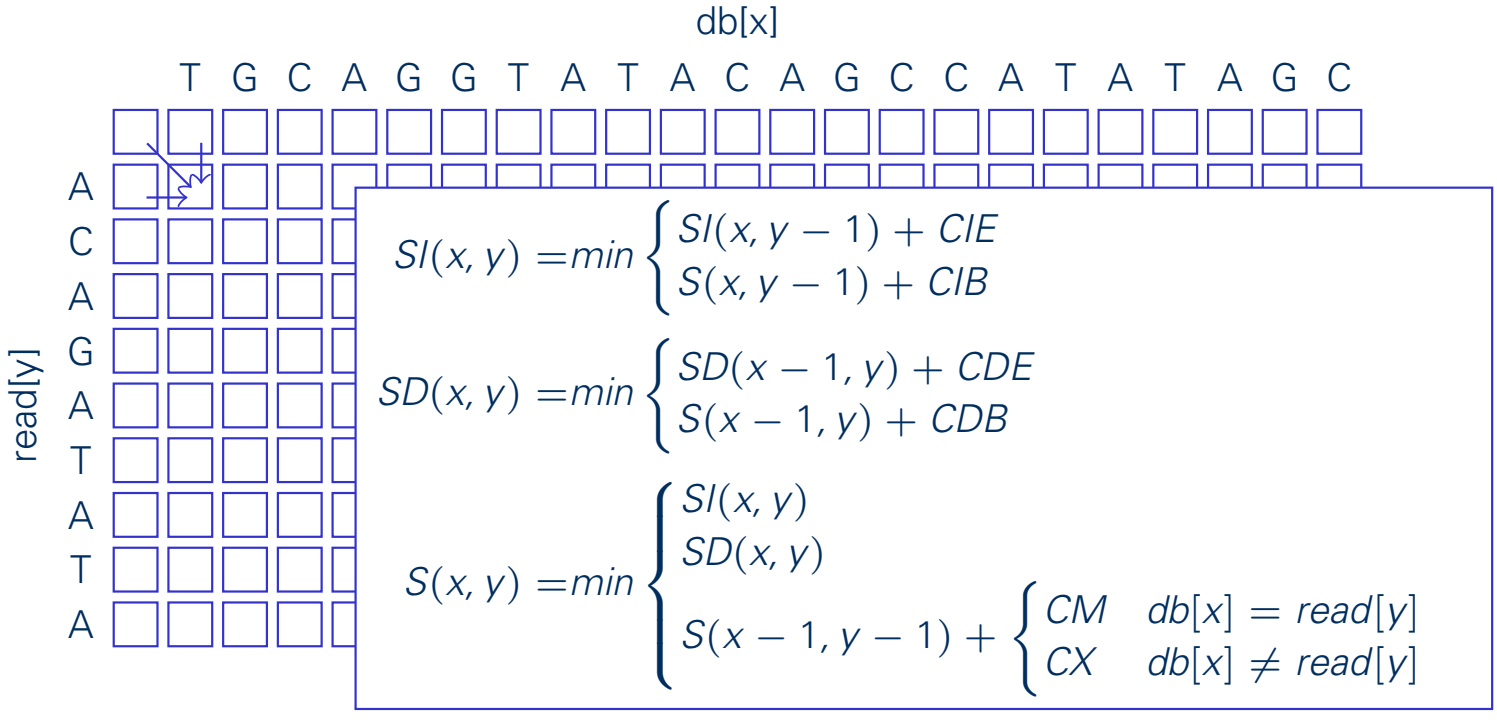
# Problemstellung

Alignment zweier Sequenzen von Symbolen. Als Symbole dienen die Basen Adenin (*A*), Guanin (*G*), Cytosin (*C*) und Thymin (*T*).

| | |
|---|---|
| db | `ACCGTA` |
| read | `A-CGTT` |
| *align*(*db*, *read*) | `MDMMMX` |

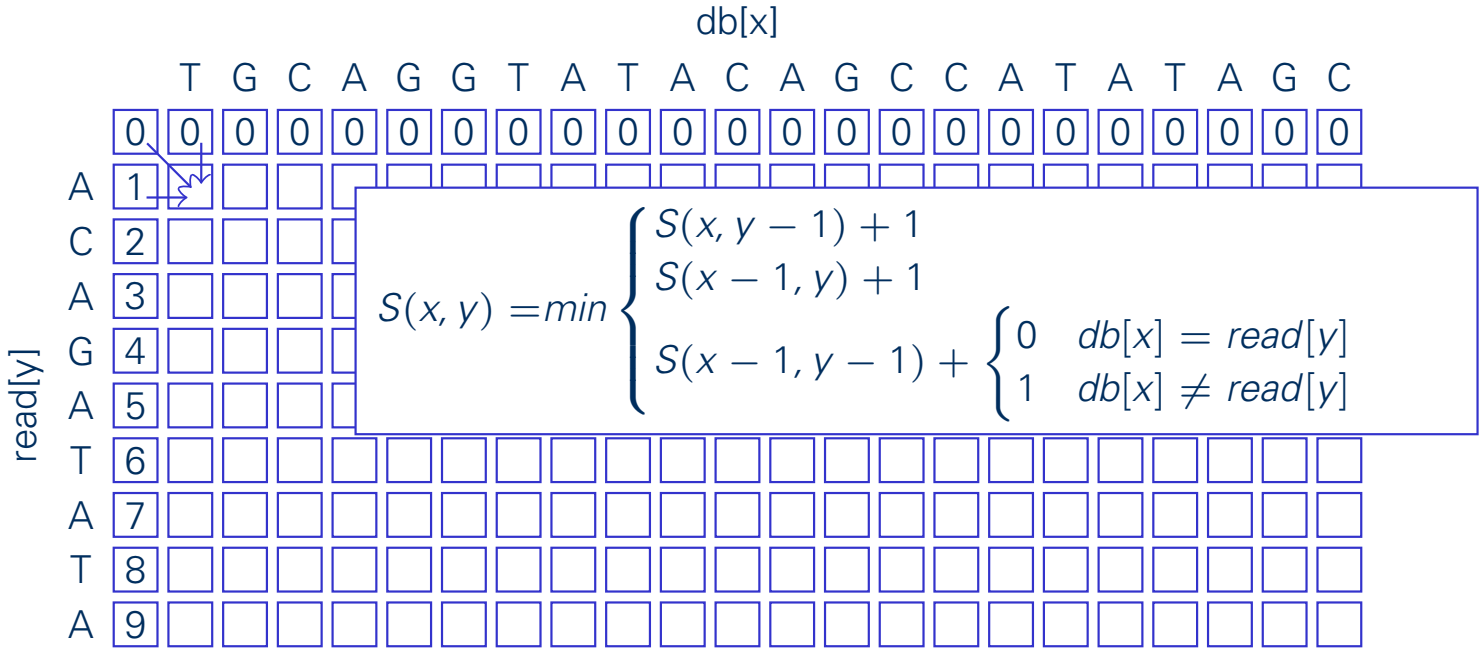
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



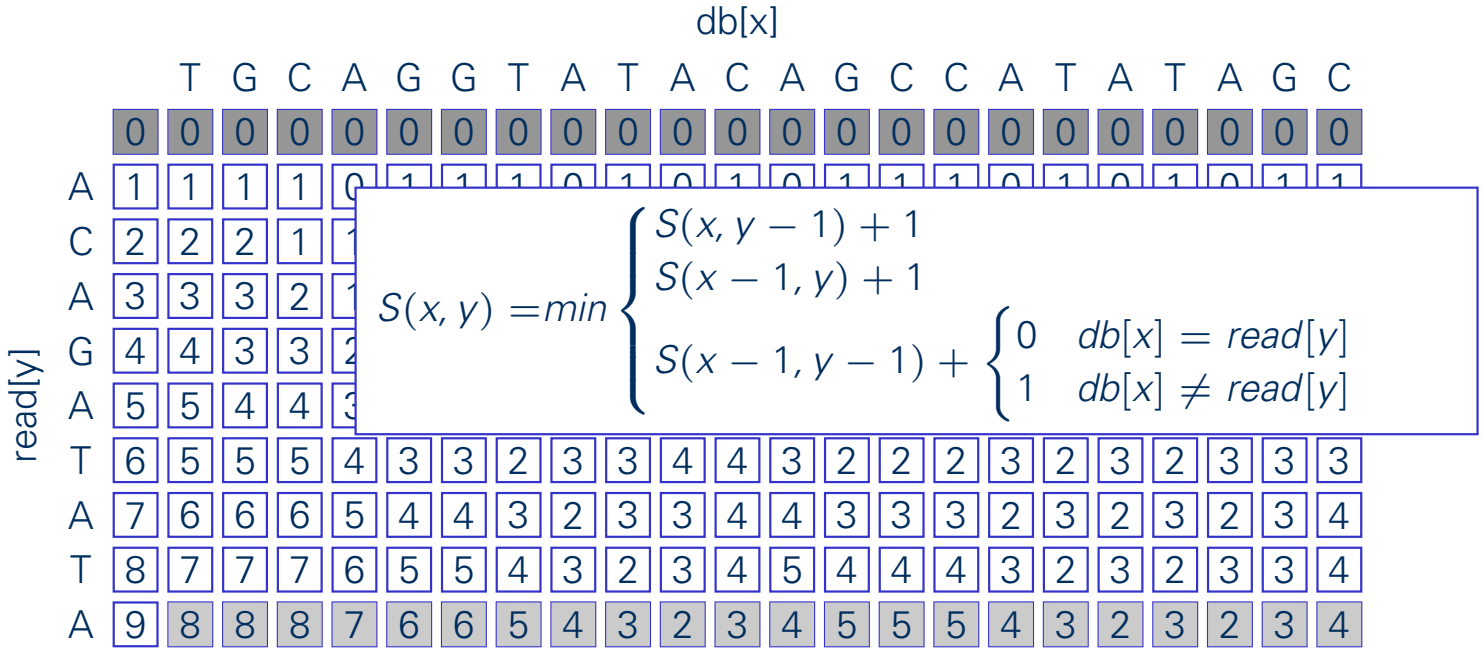
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



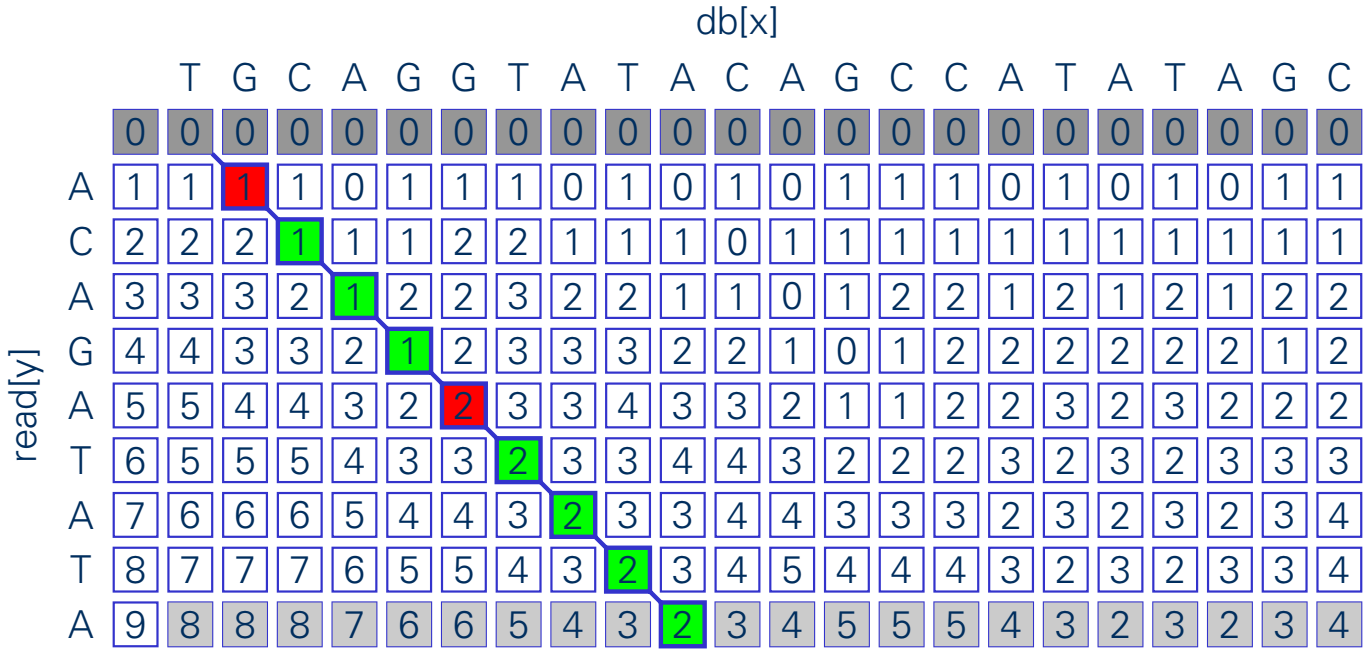
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



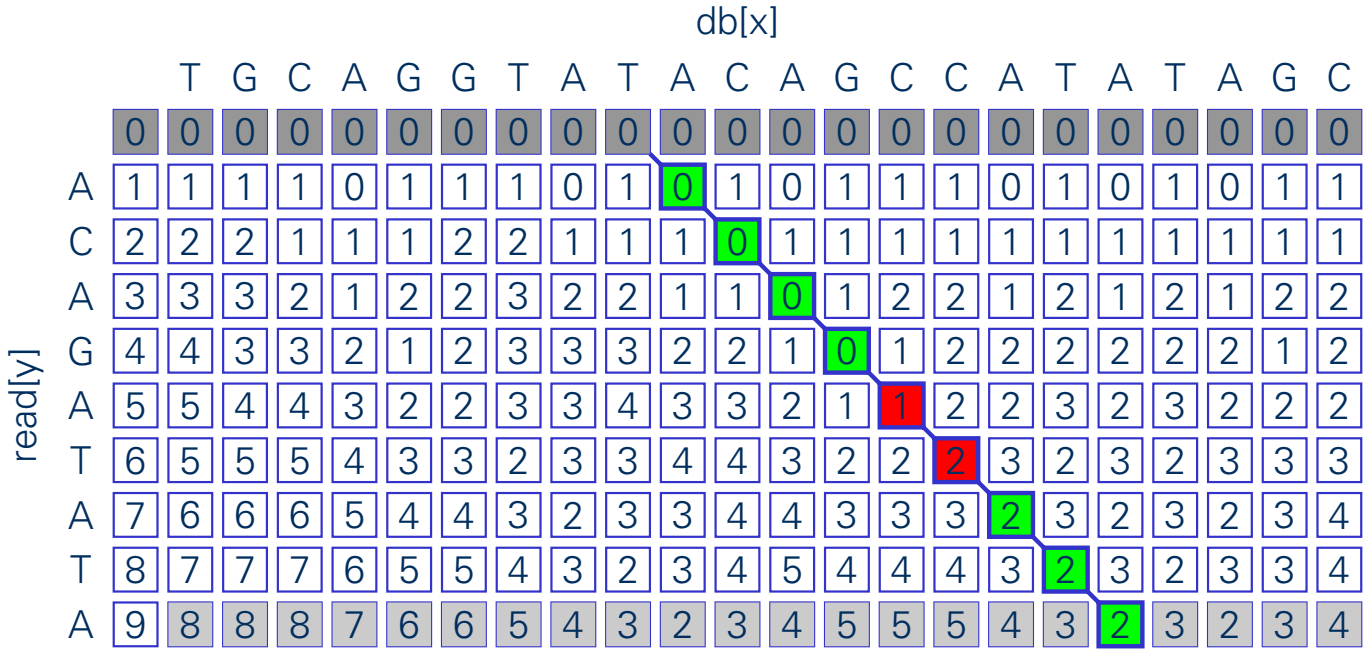
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



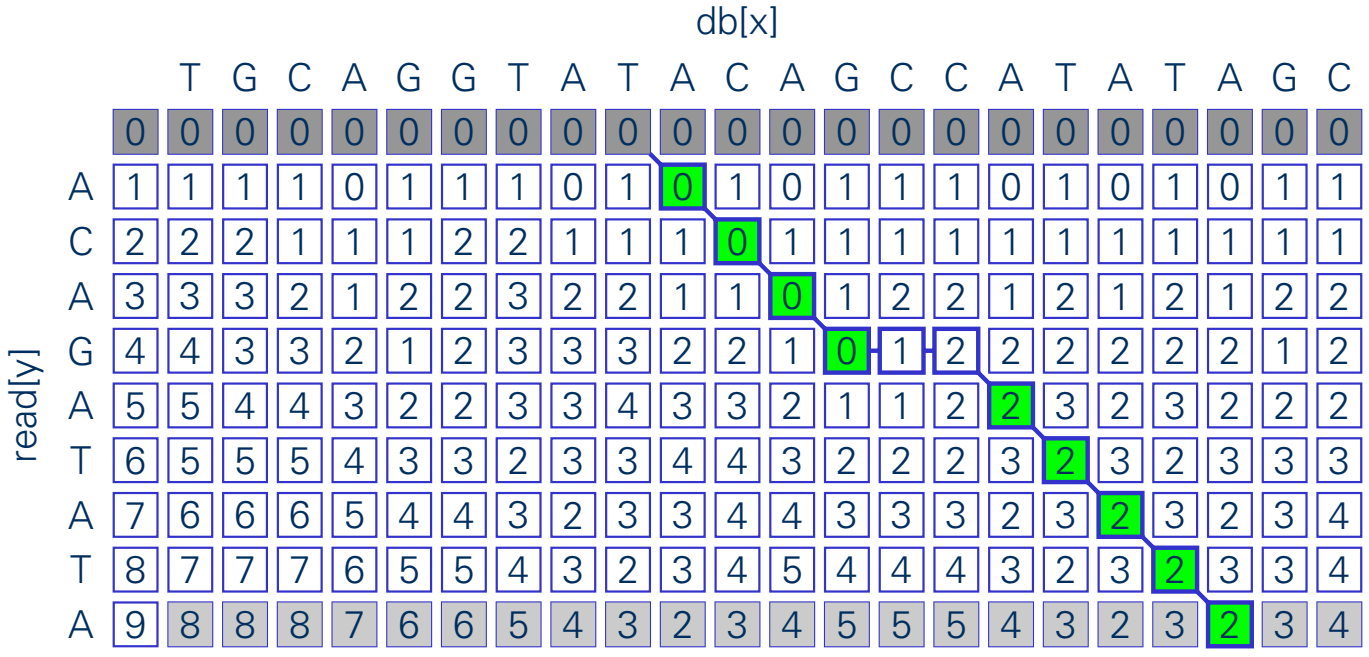
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



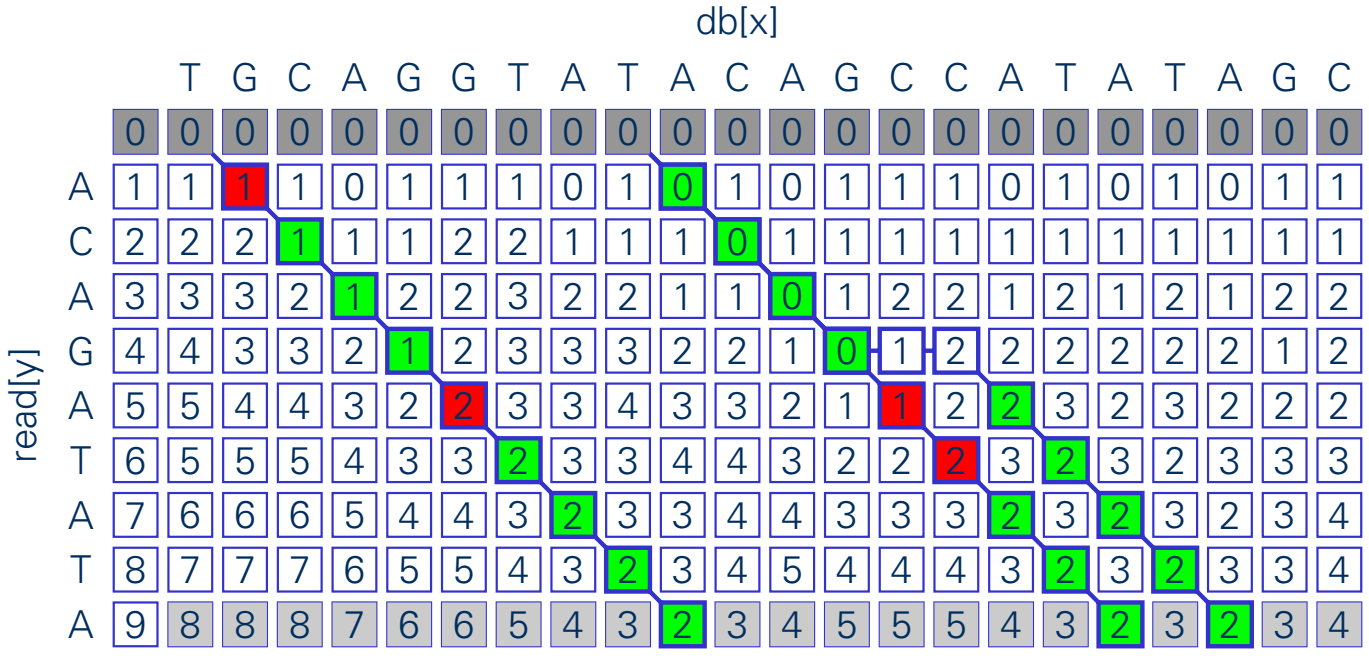
# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



# Problemstellung

Berechnet das optimale Alignment zweier Sequenzen (db,read).
Ein Alignment besteht aus Editieroperationen, welche mit Kosten behaftet sind. Für die Berechnung wird eine Matrix benutzt, in welcher die benötigten Kosten aufaddiert werden.



# Parallelisierung

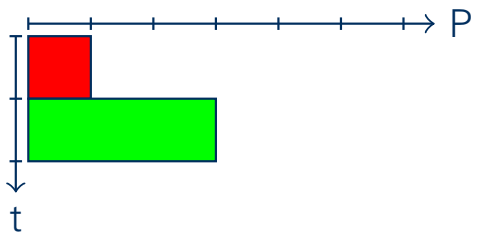
Das Einlesen der Reads benötigt 3,5 % der Zeit und das Ausgeben der Alignments 2,32 %.

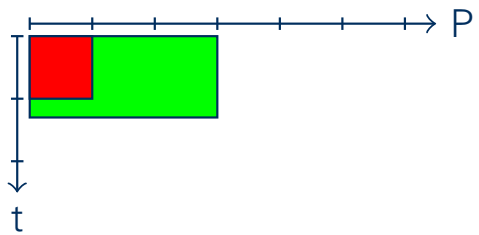
$$S_{p1}^{max}(P) = \frac{1}{f + \frac{(1-f)}{P}}$$

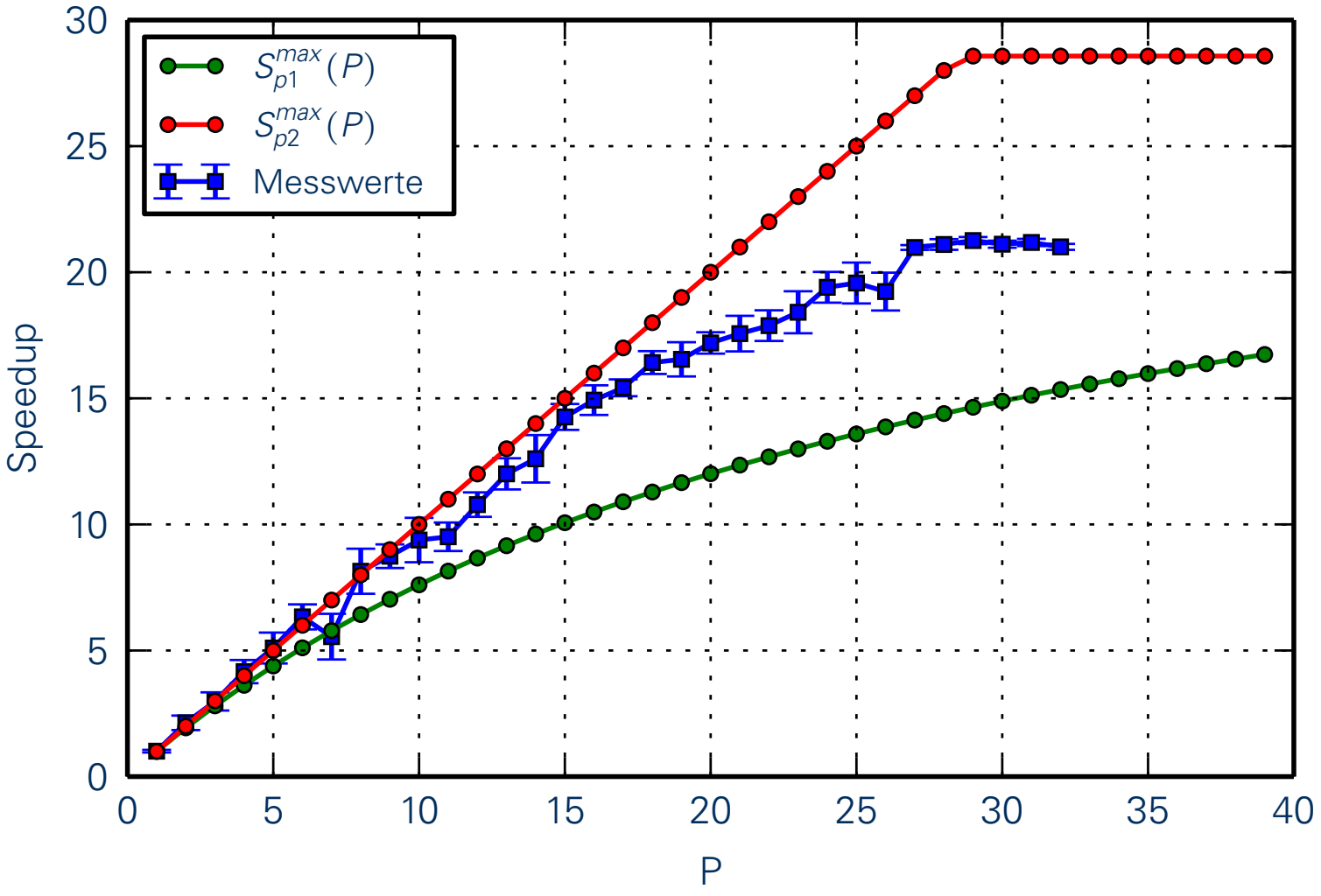
$$f = 0{,}035$$

(Amdahl's Law)

$$S_{p2}^{max}(P) = min(P, \frac{1}{max(0{,}035,\ 0{,}023\,2)})$$



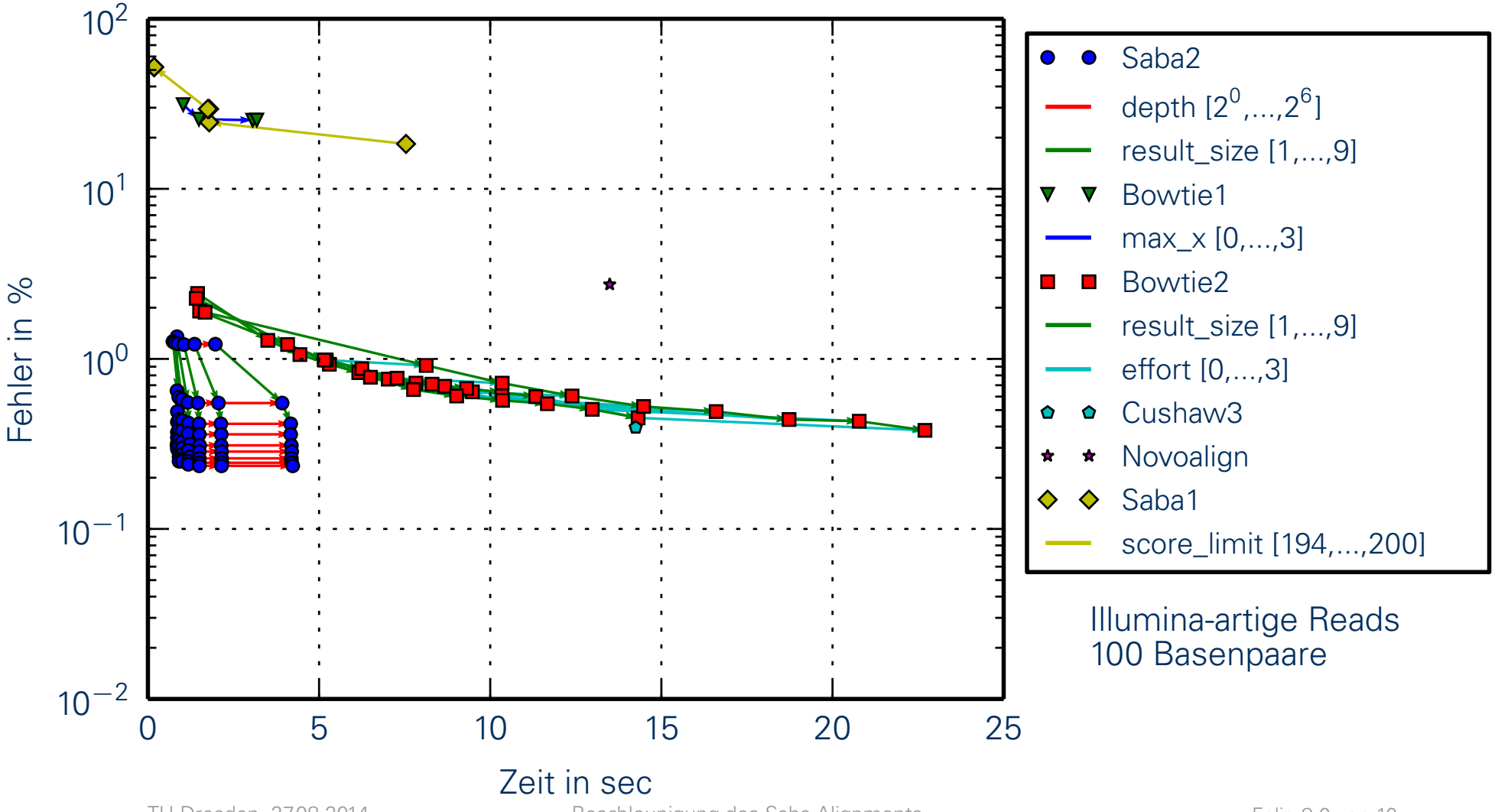

$S_{p1}^{max}(P)$
$S_{p2}^{max}(P)$
Messwerte
Speedup
P
0
5
10
15
20
25
30
35
40

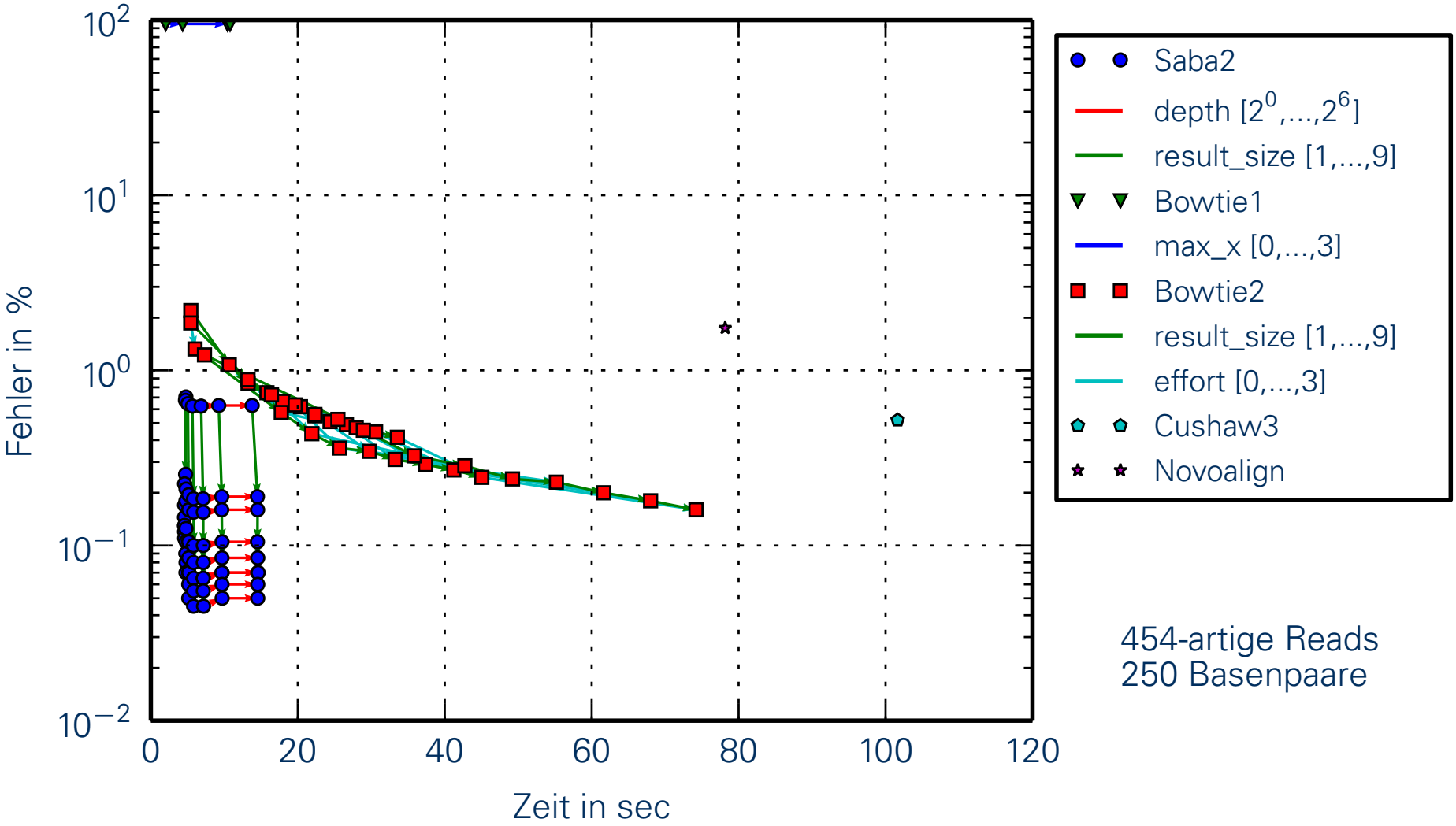
# Auswertung/Performance

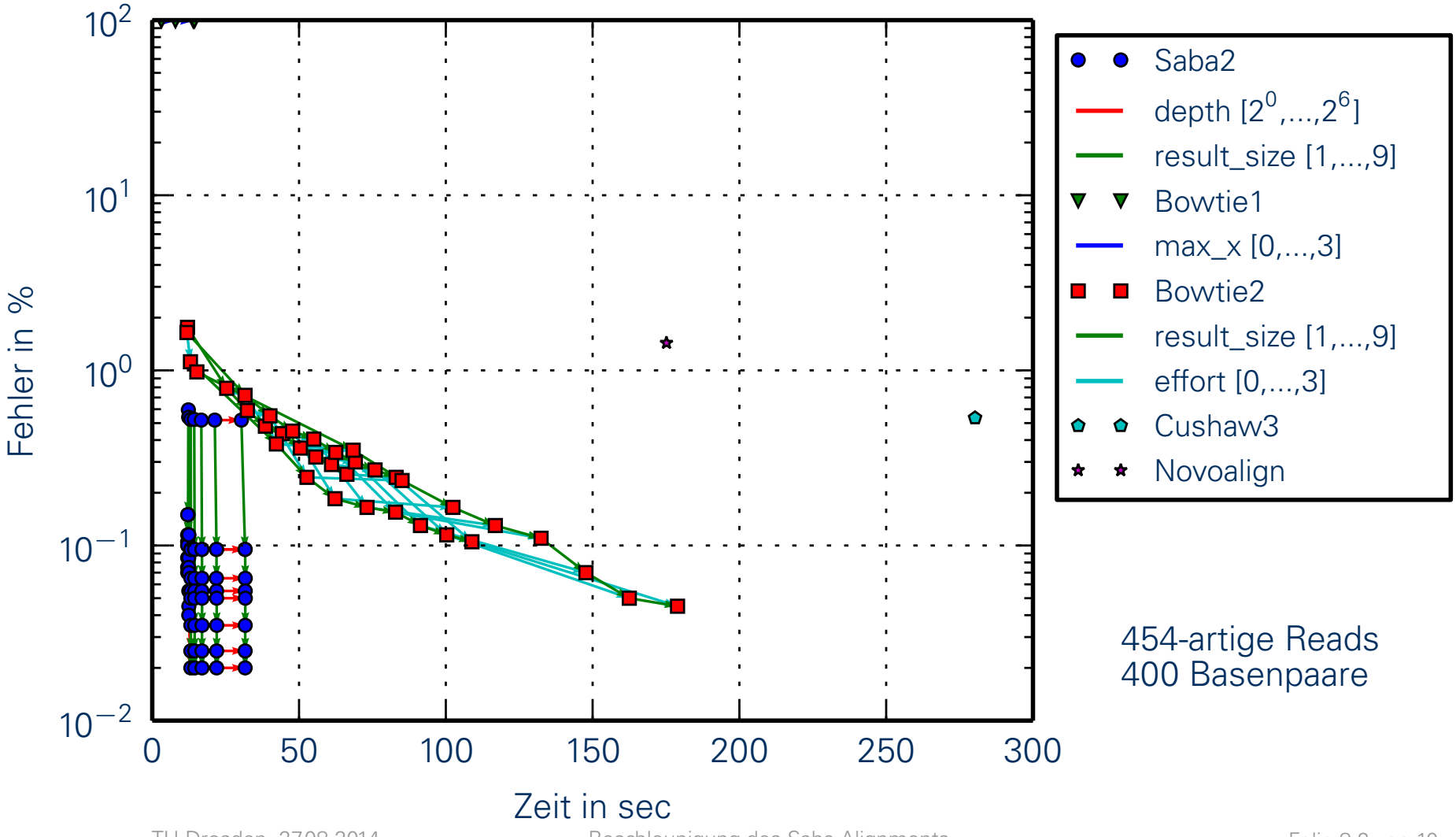
Analyse von Reads, welche mit verschiedenen Seqenzierungsarten simuliert wurden.

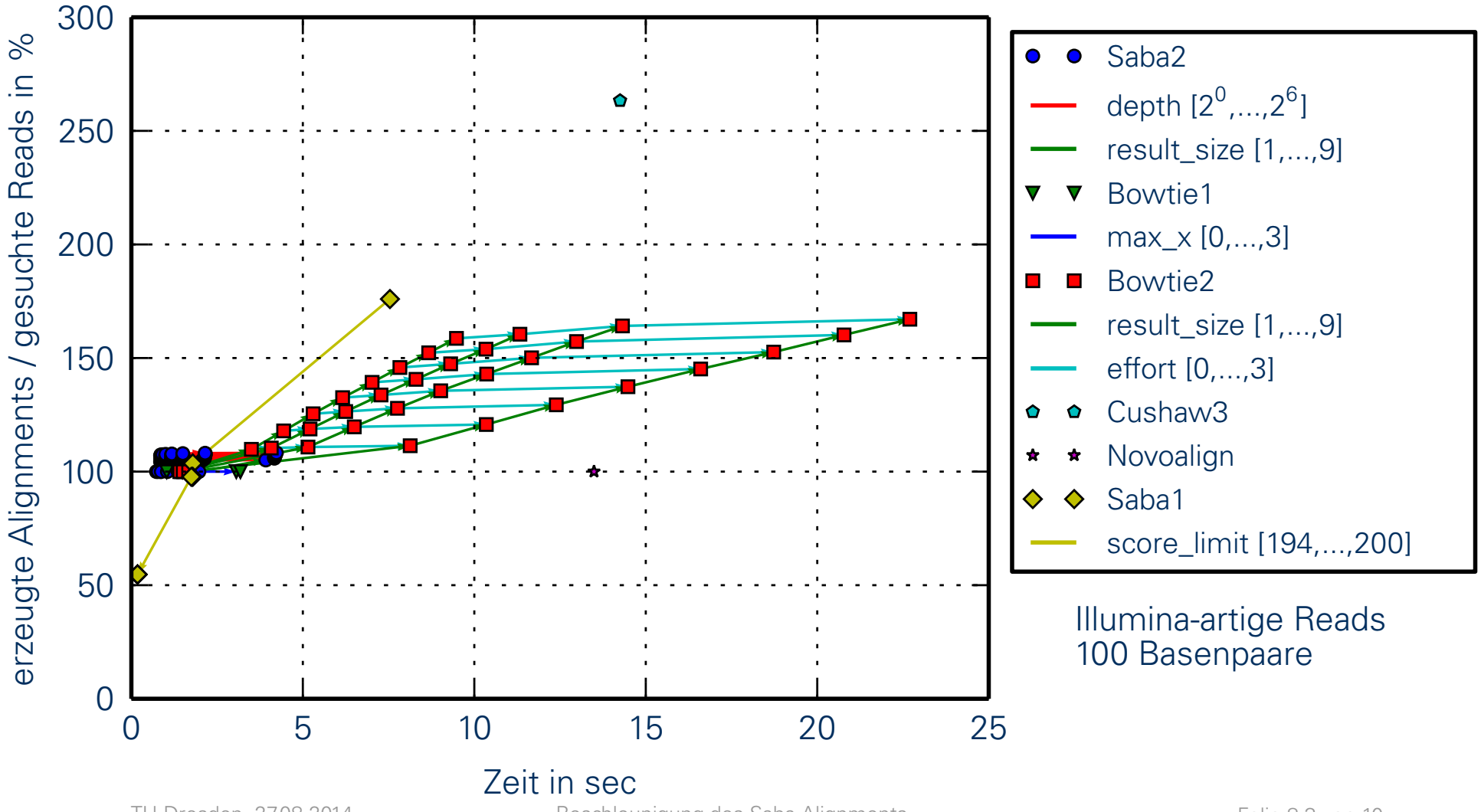
| Algorithmus | Jahr | Basenpaare | Entwickler |
|---|---|---|---|
| Bowtie1 | 2013 | 35-100 | Johns Hopkins University |
| Bowtie2 | 2014 | 50-1000 | Johns Hopkins University |
| Cushaw3 | 2014 | ~100 | Uni Mainz |
| Novoalign | 2014 | short and long | novocraft.com |
| Saba1 | 2013 | | Frank Hoffmann |

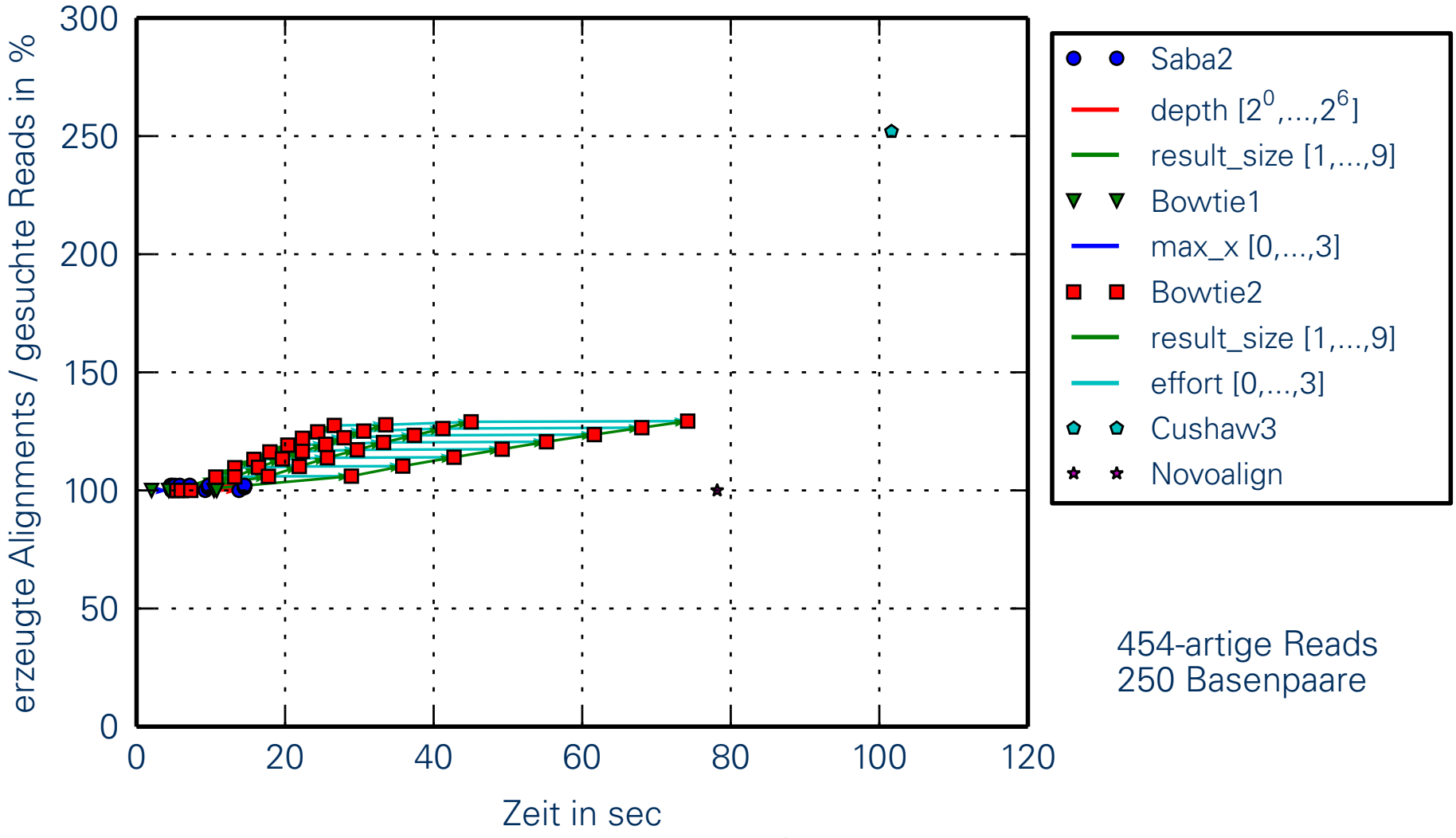
- Illumina-artig mit 100 Basenpaaren
- 454-artig mit 250 Basenpaaren
- 454-artig mit 400 Basenpaaren

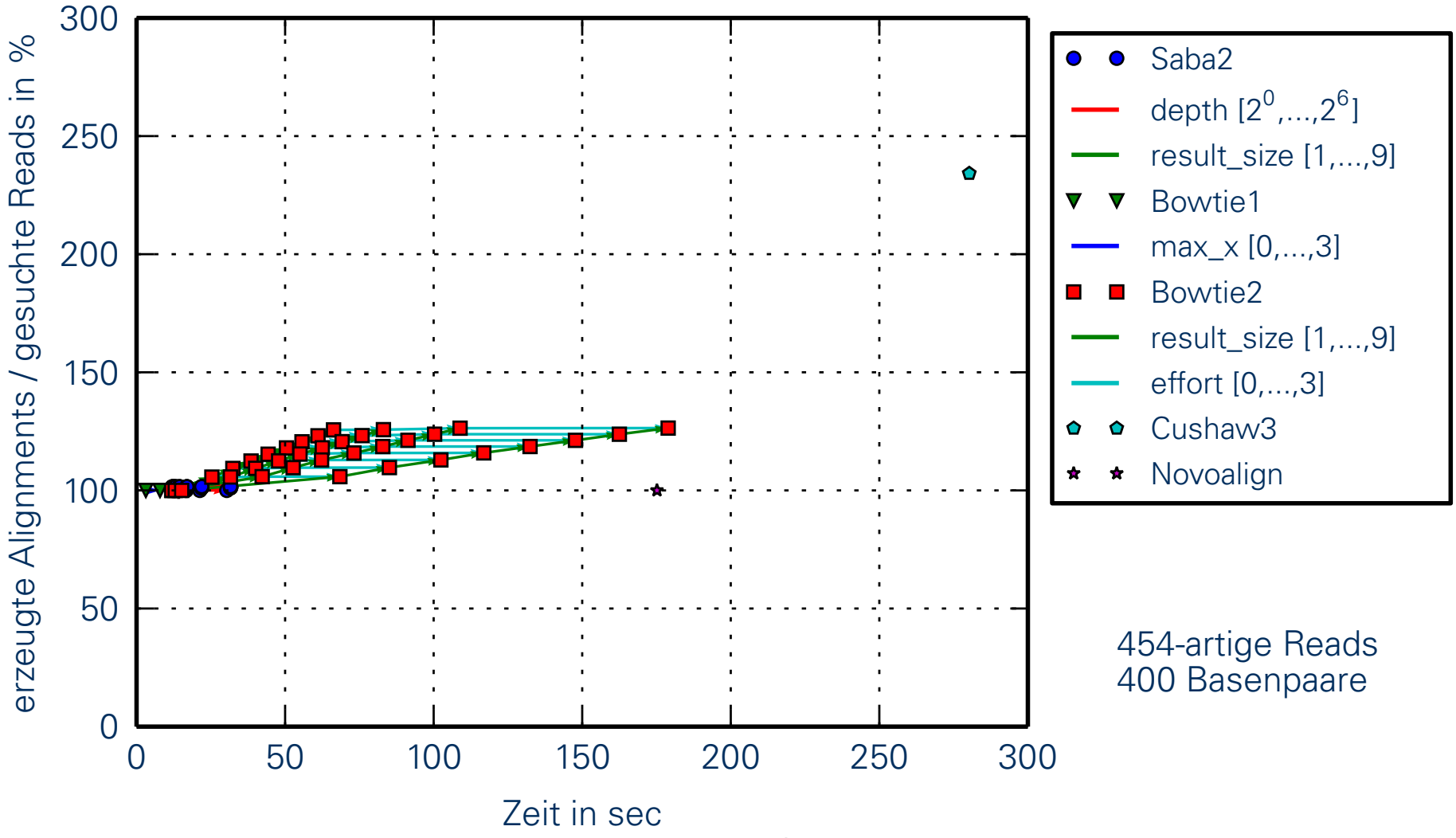

Fehler in %
$10^2$
$10^1$
$10^0$
$10^{-1}$
$10^{-2}$
0
5
10
15
20
25
Zeit in sec
Saba2
depth [$2^0$,...,$2^6$]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
Saba1
score_limit [194,...,200]
Illumina-artige Reads
100 Basenpaare


Fehler in %
$10^2$
$10^1$
$10^0$
$10^{-1}$
$10^{-2}$
0
20
40
60
80
100
120
Zeit in sec
Saba2
depth [$2^0$,...,$2^6$]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
454-artige Reads
250 Basenpaare


Fehler in %
$10^2$
$10^1$
$10^0$
$10^{-1}$
$10^{-2}$
0
50
100
150
200
250
300
Zeit in sec
Saba2
depth [$2^0$,...,$2^6$]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
454-artige Reads
400 Basenpaare


erzeugte Alignments / gesuchte Reads in %
300
250
200
150
100
50
0
0
5
10
15
20
25
Zeit in sec
Saba2
depth [2^0,...,2^6]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
Saba1
score_limit [194,...,200]
Illumina-artige Reads
100 Basenpaare


erzeugte Alignments / gesuchte Reads in %
300
250
200
150
100
50
0
0
20
40
60
80
100
120
Zeit in sec
Saba2
depth [2^0,...,2^6]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
454-artige Reads
250 Basenpaare


erzeugte Alignments / gesuchte Reads in %
300
250
200
150
100
50
0
0
50
100
150
200
250
300
Zeit in sec
Saba2
depth [2^0,...,2^6]
result_size [1,...,9]
Bowtie1
max_x [0,...,3]
Bowtie2
result_size [1,...,9]
effort [0,...,3]
Cushaw3
Novoalign
454-artige Reads
400 Basenpaare

# Zusammenfassung

- Optimierung des Alignments
- Parallelisierung
- Vergleich mit anderen Verfahren

»Wissen schafft Brücken.«