Performance improving Extensions of the Smith-Waterman Algorithm

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

Carsten Tanke (B.Eng.)

Kiel University of Applied Sciences

Institute for Communications Technology and Microelectronics

Contents

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

- Similarity of two sequences
- The Smith-Waterman-Algorithm
- Modification of the Smith-Waterman-Algorithm with submatrices
- Results

Similarity of two sequences

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

Sequence 1: CATCGTCCTGCCTAAT
Sequence 2: ATCTCGGTGCTCTTCT

- Hamming distance
- Alignment is the main tool in bioinformatics
- Finding the optimal alignment

Smith – Waterman (1)

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

- Algorithm to find a local alignment
- Developed by Temple F. Smith and Michael S. Waterman in 1981

Procedure:

- Calculation of a matrix with dimensions (n+1) x (m+1)
- Finding the maximum value in the matrix
- Finding the alignment path (“backtracking”)

Smith – Waterman (2)

$$gap = -1$$

$$w = \begin{cases} +2 \; if \; Seq1(a) = Seq2(b) \\ -1 \; if \; Seq1(a) \neq Seq2(b) \end{cases}$$

$$F(i,j) = max \begin{cases} F(i-1, j-1) + w \\ F(i-1, j) + gap \\ F(i, j-1) + gap \\ 0 \end{cases}$$

		A	T	C	T
	0	0	0	0	0
C	0				
A	0				
T	0				
C	0				

		A	T	C	T
	0	0	0	0	0
C	0	0			
A	0	2			
T	0				
C	0				

Smith – Waterman (Example)

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

		A	T	C	T	C	G	G	T	G	C	T	C	T	T	C	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	2	1	0	0	0	2	1	2	1	0	2	1
A	0	2	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1
T	0	1	4	3	3	2	1	0	2	1	0	3	2	3	3	2	3
C	0	0	3	6	5	5	4	3	2	1	3	2	5	4	3	5	4
G	0	0	2	5	5	4	7	6	5	4	3	2	4	4	3	4	4
T	0	0	2	4	7	6	6	6	8	7	6	5	4	6	6	5	6
C	0	0	1	4	6	9	8	7	7	7	9	8	7	6	5	8	7
C	0	0	0	3	5	8	8	7	6	6	9	8	10	9	8	7	7
T	0	0	2	2	5	7	7	7	9	8	8	11	10	12	11	10	9
G	0	0	1	1	4	6	9	9	8	11	10	10	10	11	11	10	9
C	0	0	0	3	3	6	8	8	8	10	13	12	12	11	10	13	12
C	0	0	0	2	2	5	7	7	7	9	12	12	14	13	12	12	12
T	0	0	2	1	4	4	6	6	9	8	11	14	13	16	15	14	14
A	0	2	1	1	3	3	5	5	8	8	10	13	13	15	15	14	13
A	0	2	1	0	2	2	4	4	7	7	9	12	12	14	14	14	13
T	0	1	4	3	2	1	3	3	6	6	8	11	11	14	16	15	16

Smith – Waterman (Example)

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

		A	T	C	T	C	G	G	T	G	C	T	C	T	T	C	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	2	1	0	0	0	2	1	2	1	0	2	1
A	0	2	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1
T	0	1	4	3	3	2	1	0	2	1	0	3	2	3	3	2	3
C	0	0	3	6	5	5	4	3	2	1	3	2	5	4	3	5	4
G	0	0	2	5	5	4	7	6	5	4	3	2	4	4	3	4	4
T	0	0	2	4	7	6	6	6	8	7	6	5	4	6	6	5	6
C	0	0	1	4	6	9	8	7	7	7	9	8	7	6	5	8	7
C	0	0	0	3	5	8	8	7	6	6	9	8	10	9	8	7	7
T	0	0	2	2	5	7	7	7	9	8	8	11	10	12	11	10	9
G	0	0	1	1	4	6	9	9	8	11	10	10	10	11	11	10	9
C	0	0	0	3	3	6	8	8	8	10	13	12	12	11	10	13	12
C	0	0	0	2	2	5	7	7	7	9	12	12	14	13	12	12	12
T	0	0	2	1	4	4	6	6	9	8	11	14	13	**16**	15	14	14
A	0	2	1	1	3	3	5	5	8	8	10	13	13	15	15	14	13
A	0	2	1	0	2	2	4	4	7	7	9	12	12	14	14	14	13
T	0	1	4	3	2	1	3	3	6	6	8	11	11	14	**16**	15	**16**

Smith – Waterman (Example)

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

		A	T	C	T	C	G	G	T	G	C	T	C	T	T	C	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	2	1	0	0	0	2	1	2	1	0	2	1
A	0	2	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1
T	0	1	4	3	3	2	1	0	2	1	0	3	2	3	3	2	3
C	0	0	3	6	5	5	4	3	2	1	3	2	5	4	3	5	4
G	0	0	2	5	5	4	7	6	5	4	3	2	4	4	3	4	4
T	0	0	2	4	7	6	6	6	8	7	6	5	4	6	6	5	6
C	0	0	1	4	6	9	8	7	7	7	9	8	7	6	5	8	7
C	0	0	0	3	5	8	8	7	6	6	9	8	10	9	8	7	7
T	0	0	2	2	5	7	7	7	9	8	8	11	10	12	11	10	9
G	0	0	1	1	4	6	9	9	8	11	10	10	10	11	11	10	9
C	0	0	0	3	3	6	8	8	8	10	13	12	12	11	10	13	12
C	0	0	0	2	2	5	7	7	7	9	12	12	14	13	12	12	12
T	0	0	2	1	4	4	6	6	9	8	11	14	13	**16**	15	14	14
A	0	2	1	1	3	3	5	5	8	8	10	13	13	15	15	14	13
A	0	2	1	0	2	2	4	4	7	7	9	12	12	14	14	14	13
T	0	1	4	3	2	1	3	3	6	6	8	11	11	14	**16**	15	**16**

Resulting alignments

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

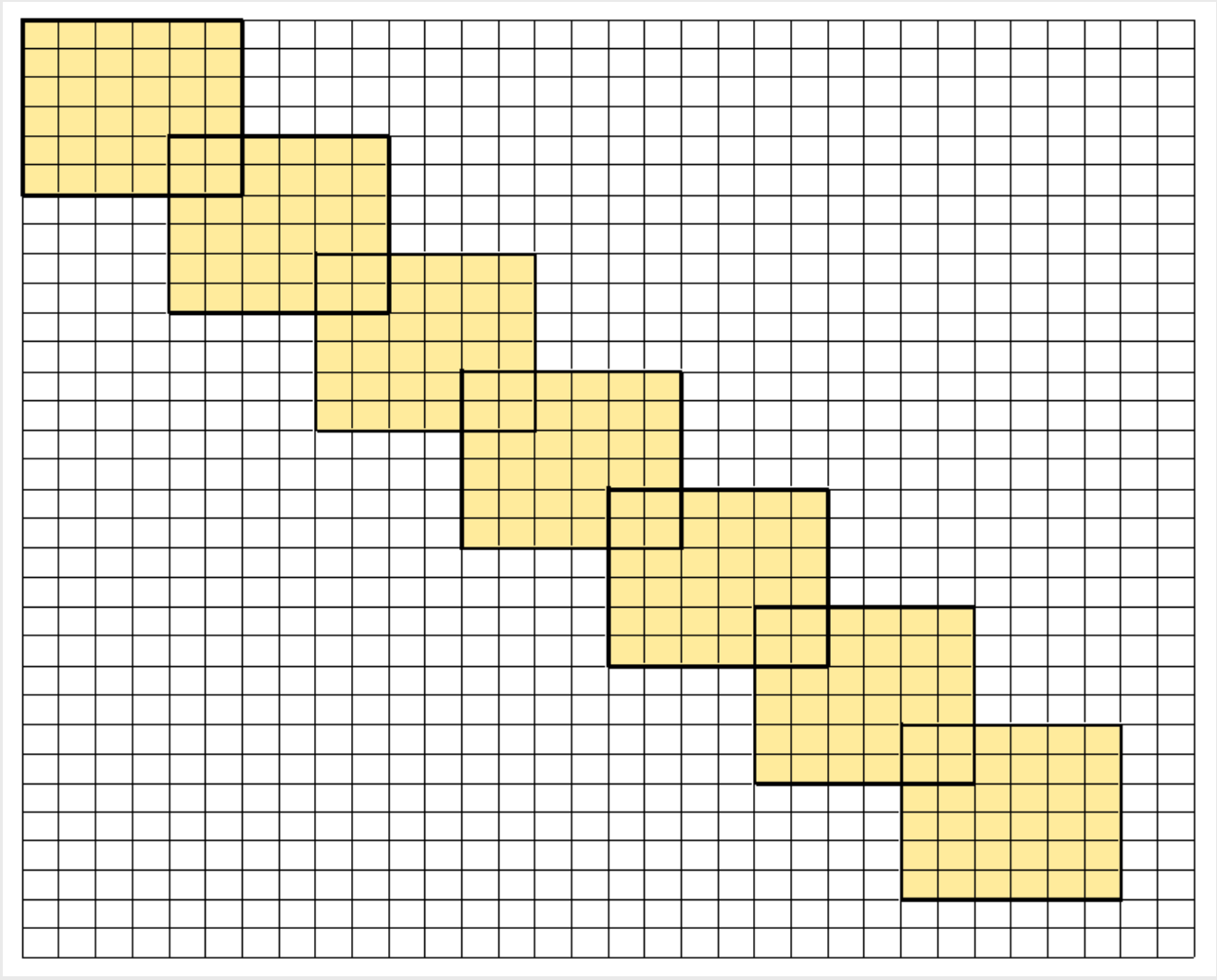
```
C A T C G T C - C T G C - C T
  | | |   | |       | | |   | |
- A T C - T C G G T G C T C T

C A T C G T C - C T G C - C T A A T
  | | |   | |       | | |   |     |
- A T C - T C G G T G C T C T T C T

C A T C G T C - C T G C - C T A A T
  | | |   | |       | | |   |     |
- A T C - T C G G T G C T C - - T T
```

Smith-Waterman with submatrices

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences



Smith-Waterman with submatrices

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

Parameters:

- Dimension of submatrix: > 1/3 of original matrix
- Overlap: > 1/4 of submatrix

Optimal results:

- Reduction of computational costs
- No loss in alignment quality

(1) SW with submatrices along main diagonal

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

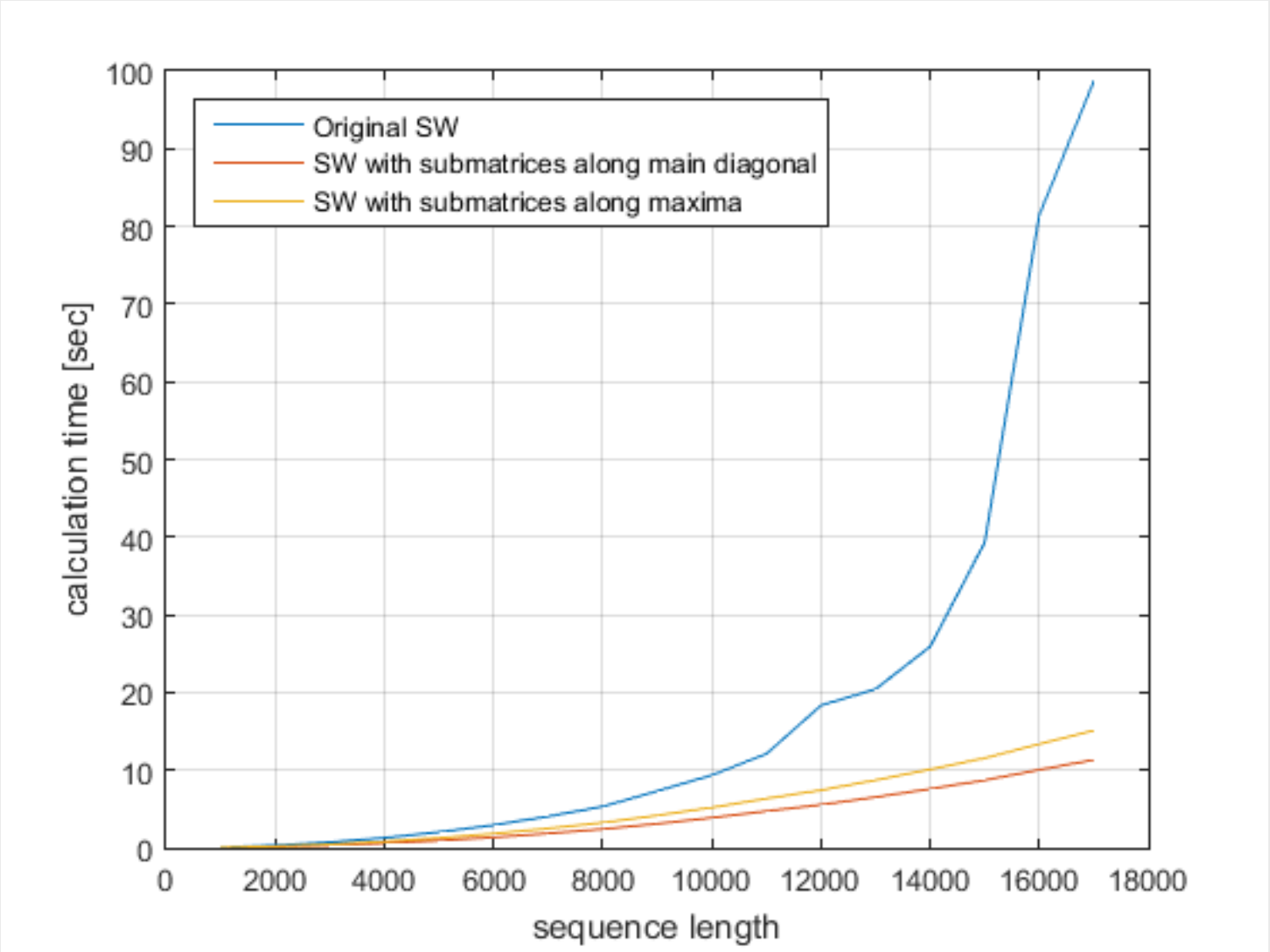
		A	T	C	T	C	G	G	T	G	C	T	C	T	T	C	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	2	1	0	0	0	2	1	2	1	0	2	1
A	0	2	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1
T	0	1	4	3	3	2	1	0	2	1	0	3	2	3	3	2	3
C	0	0	3	6	5	5	4	3	2	1	3	2	5	4	3	5	4
G	0	0	2	5	5	4	7	6	5	4	3	2	4	4	3	4	4
T	0	0	2	4	7	6	6	6	8	7	6	5	4	6	6	5	6
C	0	0	1	4	6	9	8	7	7	7	9	8	7	6	5	8	7
C	0	0	0	3	5	8	8	7	6	6	9	8	10	9	8	7	7
T	0	0	2	2	5	7	7	7	9	8	8	11	10	12	11	10	9
G	0	0	1	1	4	6	9	9	8	11	10	10	10	11	11	10	9
C	0	0	0	3	3	6	8	8	8	10	13	12	12	11	10	13	12
C	0	0	0	2	2	5	7	7	7	9	12	12	14	13	12	12	12
T	0	0	2	1	4	4	6	6	9	8	11	14	13	16	15	14	14
A	0	2	1	1	3	3	5	5	8	8	10	13	13	15	15	14	13
A	0	2	1	0	2	2	4	4	7	7	9	12	12	14	14	14	13
T	0	1	4	3	2	1	3	3	6	6	8	11	11	14	16	15	16

(2) SW with submatrices along last maximum

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

		A	T	C	T	C	G	G	T	G	C	T	C	T	T	C	T
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	2	1	2	1	0	0	0	2	1	2	1	0	2	1
A	0	2	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1
T	0	1	4	3	3	2	1	0	2	1	0	3	2	3	3	2	3
C	0	0	3	6	5	5	4	3	2	1	3	2	5	4	3	5	4
G	0	0	2	5	5	4	7	6	5	4	3	2	4	4	3	4	4
T	0	0	2	4	7	6	6	6	8	7	6	5	4	6	6	5	6
C	0	0	1	4	6	9	8	7	7	7	9	8	7	6	5	8	7
C	0	0	0	3	5	8	8	7	6	6	9	8	10	9	8	7	7
T	0	0	2	2	5	7	7	7	9	8	8	11	10	12	11	10	9
G	0	0	1	1	4	6	9	9	8	11	10	10	10	11	11	10	9
C	0	0	0	3	3	6	8	8	8	10	13	12	12	11	10	13	12
C	0	0	0	2	2	5	7	7	7	9	12	12	14	13	12	12	12
T	0	0	2	1	4	4	6	6	9	8	11	14	13	16	15	14	14
A	0	2	1	1	3	3	5	5	8	8	10	13	13	15	15	14	13
A	0	2	1	0	2	2	4	4	7	7	9	12	12	14	14	14	13
T	0	1	4	3	2	1	3	3	6	6	8	11	11	14	16	15	16

Reduction of calculation times in Matlab



Results

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

- **Calculation costs / time reduced ✓**
- **Alignment quality for submatrices along main diagonal nearly identical (0.04 %) ✓**
- **Alignment quality for submatrices along maxima is inferior to 0.37 % ✗**

→ **Smith-Waterman with submatrices along main diagonal provides best results in quality and calculation time**

Thanks for your attention !

Time for questions...

References

FACHHOCHSCHULE KIEL
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN
University of Applied Sciences

- **Smith, T.F. and Waterman, M.S.** Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981, 147.
- **Lesk, Arthur M.** Introduction to bioinformatics. Oxford: University Press, 2005.