



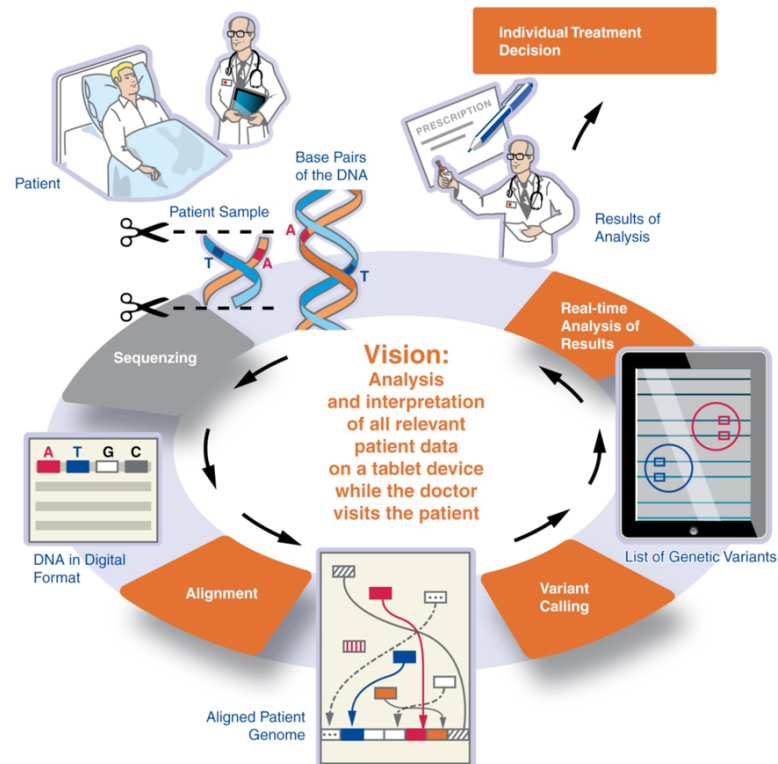
Reconstruct the Genome: Alignment

Dr. Matthieu-P. Schapranow
Code of Life

November 2016

From Raw Genome Data to Analysis

- **DNA Sequencing:** Transformation of analogues DNA into digital format
- **Alignment:** Reconstruction of complete genome with snippets
- **Variant Calling:** Identification of genetic variants
- **Data Annotation:** Linking genetic variants with research findings



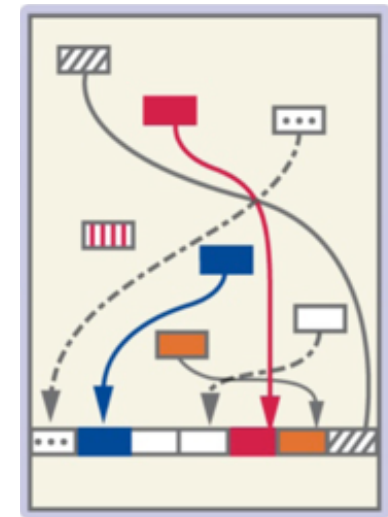
**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Alignment Overview

- Purpose: Mapping of DNA reads to a reference
- Input:
 - DNA reads := Sequence of nucleotides with a length of 100 bp up to some 1 kbp
 - Reference genome := Blueprint for alignment of DNA reads
- Output: Mapped DNA reads

- Bear in mind:
 - Less fraction in DNA reads, i.e. longer reads, allows more precise alignment
 - Reference from same origin improves mapping quality



**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Selected Alignment Algorithms

Needleman-Wunsch Algorithm

- Global alignment strategy
- Alignment score is defined by the value in the most lower right cell of the matrix
- Needleman, S. B. and Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins" in "Molecular Biology", 48(3): 443-53

Needleman-Wunsch Algorithm

- What is the best global alignment for sequences CTG and ACTGC?

	-	A	C	T	G	C
-						
C						
T						
G						

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm Matrix Initialization

- $M(0,0) = 0$

	-	A	C	T	G	C
-	0					
C						
T						
G						

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Matrix Initialization

- $M(0,0) = 0$
- $M(i,0) = M(i-1,0) + f(i), 1 \leq i \leq m$
- $M(0,j) = M(0,j-1) + f(j), 1 \leq j \leq n$
- $f(x) :=$ gap cost function, e.g. $f(x) = -1$

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1					
T	-2					
G	-3					

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Fill Matrix

- Weight function $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $D(i,j)$ defines value of matrix at coordinates (i,j)

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1					
T	-2					
G	-3					

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Fill Matrix

- Weight function $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $D(1,1) := \text{maximum of}$
 - ① $D(0,0) + w(A,C) = 0 + (-1) = -1$
 - ② $D(1,0) + w(A,-) = -1 + (-1) = -2$
 - ③ $D(0,1) + w(-,C) = -1 + (-1) = -2$

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1	$D(1,1)$				
T	-2					
G	-3					

Reconstruct the
Genome: Alignment

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Fill Matrix

■ Weight function $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$

■ $D(1,1) := \text{maximum of}$

① $D(0,0) + w(A,C) = 0 + (-1) = -1$

② $D(1,0) + w(A,-) = -1 + (-1) = -2$

③ $D(0,1) + w(-,C) = -1 + (-1) = -2$

	-	A	C	T	G	C
-	0 ^①	-1 ^②	-2	-3	-4	-5
C	-1 ^③	-1				
T	-2					
G	-3					

Reconstruct the
Genome: Alignment

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Fill Matrix

- Repeat for all $D(i,j)$ until matrix is filled

- Bear in mind: Filling the matrix can be performed in parallel

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1	-1	0	-1	-2	-1
T	-2	-2	-1	1	0	-1
G	-3	-3	-2	0	2	1

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Determine Best Global Alignment

- Trace path back from $D(m,n)$ to origin $D(0,0)$

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1	-1	0	-1	-2	-1
T	-2	-2	-1	1	0	-1
G	-3	-3	-2	0	2	1

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Needleman-Wunsch Algorithm

Determine Best Global Alignment

- Reference: ACTGC
- Alignment: -CTG-
- Score of the alignment is: 1

- Match
- Mismatch
- Gap

- Bear in mind: Backtracing can be performed in parallel

	-	A	C	T	G	C
-	0	-1	-2	-3	-4	-5
C	-1	-1	0	-1	-2	-1
T	-2	-2	-1	1	0	-1
G	-3	-3	-2	0	2	1

Reconstruct the
Genome: Alignment

Schapranow, Nov 2016

Selected Alignment Algorithms

Smith-Waterman Algorithm

- Determine optimal local alignments
- Adaption of Needleman-Wunsch algorithm
 - Initialize all cells within first row and column with zero
 - Alignment score is defined by highest value somewhere in the matrix
 - Backtracing from cell with alignment score to first cell containing zero
- Smith, T. F. and Waterman, M. S. (1981). "Identification of Common Molecular Subsequences" in "Molecular Biology", 147: 195-7.

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm

- What is the best local alignment for sequences CTG and ACTGC?

	-	A	C	T	G	C
-						
C						
T						
G						

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm Matrix Initialization

- $M(0,0) = M(i,0) = M(0,j) = 0 \mid 0 \leq i \leq m, 0 \leq j \leq n$

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0					
T	0					
G	0					

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm

Fill Matrix

- Weight function $w(a,b) := \{ +1 \text{ if } a == b, -1 \text{ else} \}$
- $D(i,j)$ defines value of matrix at coordinates (i,j)

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0					
T	0					
G	0					

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm

Fill Matrix

■ $D(1,1) := \text{maximum of}$

- ① $D(0,0) + w(A,C) = 0 + (-1) = -1$
- ② $D(1,0) + w(A,-) = 0 + (-1) = -1$
- ③ $D(0,1) + w(-,C) = 0 + (-1) = -1$
- ④ 0

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0	$D(1,1)$				
T	0					
G	0					

Reconstruct the
Genome: Alignment

Schapranow, Nov 2016

Smith-Waterman Algorithm

Fill Matrix

■ $D(1,1) := \text{maximum of}$

① $D(0,0) + w(A,C) = 0 + (-1) = -1$

② $D(1,0) + w(A,-) = 0 + (-1) = -1$

③ $D(0,1) + w(-,C) = 0 + (-1) = -1$

④ 0

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0	0				
T	0					
G	0					

Diagram illustrating the Smith-Waterman algorithm fill matrix. The matrix shows the maximum alignment score for each pair of prefixes. The values are 0 for all cells. Red arrows indicate the path from (0,0) to (1,1) via (1,0) and (0,1). A yellow arrow indicates the path from (0,0) to (1,1) via (0,1) and (1,0). The value 0 is highlighted in yellow in the cell (1,1).

Reconstruct the
Genome: Alignment

Schapranow, Nov 2016

Smith-Waterman Algorithm Fill Matrix

- Repeat for all $D(i,j)$ until matrix is filled

- Bear in mind: Filling the matrix can be performed in parallel

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0	0	1	0	0	1
T	0	0	0	2	1	0
G	0	0	0	1	3	2

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm

Determine Local Alignments

- Trace path back from $\max(D(i,j))$ to first $D(i,j) = 0$

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0	0	1	0	0	1
T	0	0	0	2	1	0
G	0	0	0	1	3	2

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Smith-Waterman Algorithm

Determine Local Alignments

- Reference: ACTGC
- Local alignment: CTG
- Score of the alignment is: 3

- ← Match
- ← Mismatch
- ← Gap

- Bear in mind: Backtracing can be performed in parallel for multiple local optima

	-	A	C	T	G	C
-	0	0	0	0	0	0
C	0	0	1	0	0	1
T	0	0	0	2	1	0
G	0	0	0	1	3	2

**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

Selected Alignment Algorithms

Burrows-Wheeler Aligner

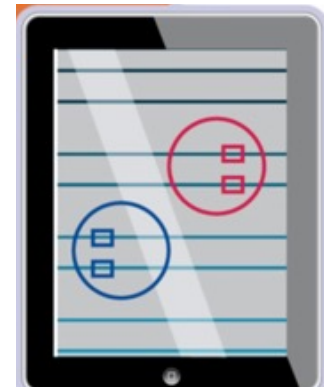
- Alignment of short read against long reference sequence
 - Uses Burrows-Wheeler Transform (BWT) to optimize search
 - Output of BWT supports compression as it contains repeated characters
-
- Li, H. and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform" in "Bioinformatics", 25(14): 1754-60

Selected Alignment Tools

- BWA: Smith-Waterman + BWT to keep memory footprint low
- Bowtie: Similar to Smith-Water/Needleman-Wunsch + BWT
- HANA Aligner (based on IMDB): BWA + FM index/BWT to speed-up match detection
- Isaac (commercialized by Illumina): Smith-Waterman
- Torrent Mapping Alignment Program (TMAP) (commercialized by IonTorrent): Smith-Waterman + FM index/BWT

What is Next?

- How to handle gaps/mismatches in the aligned genome?



**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016

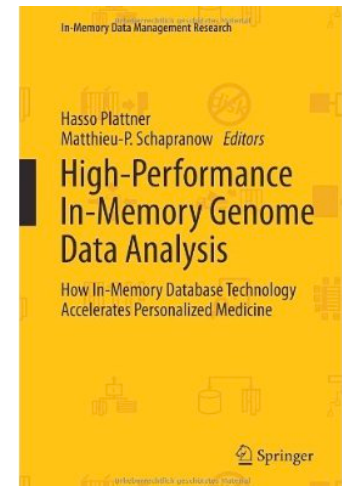
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



**Reconstruct the
Genome: Alignment**

Schapranow, Nov 2016