## **Smith-Waterman Search**

The Smith-Waterman Search plugin adds a complete implementation of the Smith-Waterman algorithm to UGENE.

To use the plugin open a nucleotide or protein sequence in the Sequence View and select the Analyze Find pattern [Smith-Waterman] item in the context menu. The Smith-Waterman Search dialog appears:

Enter pattern here			
1			
Search in	Strand	t	Desire
Sequence			Region
<ul> <li>Translation</li> </ul>	• E	Both	Whole sequence 😒
	<b>O</b> (	Direct	
	0.0	Complement	1 - 16571
		o mpionione	
Smith-Waterman algorithm	parameters		
Algorithm version	Scoring matrix	Gap scores	Results filtering strategy
SSE2	dna 😔	Gap open -	10   Report results filter-inter
dvanced	View	Gap extension -	1 🗘 Minimal score 90% 🗘

## Algorithm

This search algorithm has a **huge** calculation time advantage over the usual *Search in sequence* algorithm for <u>large sequences</u>. It happens, because the "S *earch in sequence*" algorithm represents a suffix algorithm - one sequence moves along the other and areas opposite each other are compared. Using this approach, the computation time increases in direct proportion to the length of the sequence. The *Smith-Waterman algorithm* represents absolutely another approach - searching using a certain kind of matrices.

The algorithm works in a following way:

 A matrix is createad where the sequence in which the search is carried out is located vertically, and the searched sequence is located horizontally. Fill the adjacent row and column with 0. Example: <u>The sequence in which the search is carried out</u> - ACGCGAT. <u>The searched sequence</u> - CGCG.

D <sub>(i,j)</sub>		с	G	с	G
	0	0	0	0	0
Α	0				
С	0				
G	0				
С	0				
G	0				
Α	0				

- т 0
- 2. Start filling the matrix from the left top corner moving from left to right and from top to bottom using the following formula:

Where:

D(S1, S2) - current cell,

 $D_{(S1-1, S2)}$  - the cell on the left,

 $D_{(S1, S2-1)}$  - the cell above,

 $D_{(S1-1, S2-1)}$  - the on the top and left,

gap - the Gap open penalty if we meet gap for the first time and the Gap extension penalty if we meet gap the second time or more. score - the value from the Scoring matrix table.

The following example uses nucl Scoring matrix (+5 for match, -4 for mismatch) and -1 for Gap open and Gap extension penalties:



3. Find the greater number and move to the top, left or top-left to the next greater number:

D(i,j)		c	G	с	G	D	6.0		с	G	с	G	D(i,j)		с	G	С	G	D(i,j)		с	G	с	G	D(ij)		С	G	с	G	D(i,j)		С	G	с	G
	0	0-1	0-1	0-1	0-1			0	0.1	0.1	0'1	0.1		0	0.1	0.1	0.1	0-1		0	0-1	0-1	0-1	0.1		0 0	0 1	0 1	0 1	0.1		0	0.4	0-1	0-1	0
A	0.4	04	04	04	04		A	0.1	0.4	0.4	0.4	0.4	A	0.1	0	0	0	0	A	0-1	0	04	04	04	A	0	04	0 4	04	0.4	A	0'1	0	0 4	0	0
С	0	5 *	4	6	4		с	0-1	6	4	5	4	С	0.1	5	4	5	4	С	0-1	5 ິ	4	5	4	С	0	σ	4	5	4	С	0.1	5	4	5	4
G	0	4	10	94	10 6		G	0-1	4	10	9.4	10	G	0.1	4	10	9.4	10	G	0 1	4	0	94	10 5	G	0-1	4	10	9.4	10 5	G	0.1	4	10	9	10
С	0	5	9	15	14		с	0-1	5	9	15	14 <sup>-4</sup>	С	0.1	5	.4 9	Ø	14	С	0-1	5	94	15	14	С	0-1	5	9	15	14	с	0.1	5	94	15	14
G	0	4	10	14	20		G	0 1	4	10 6	14 <sup>-4</sup>	ିତ	G	0.4	4	10	-4 14	20	G	0-1	4	10	14	20	G	0-1	4	10	14	20	G	0.1	4	10	14	20
A	0.4	34	94	13	19 <sup>-4</sup>		A	0.1	3	9.4	13	.4 19	A	0.1	3 4	9	13	19	A	0-1	3	9	13	19 <sup>.4</sup>	A	0-1	34	9 4	13	19 <sup>.4</sup>	A	0.1	3	9	13	19
т	0	2	8	12 4	18		т	0-1	24	8.4	12	18 <sup>-4</sup>	т	0 1	2 4	84	12	18	т	0-1	2	8	12 4	18	т	0-1	24	84	12	18	т	0.1	24	8 4	12	18

Matching green symbols indicate the desired sequence intersection.

## Parameters

First of all you need to specify the pattern to search for. The rest parameters are optional:

Search in - select either to search in the sequence or in its amino acid translation.

Strand - select the strand to search in: direct, reverse-complementary or both strands.

Region — specifies the region of the sequence that will be used to search for the pattern. By default, if a subsequence has been selected when the dialog has been opened, then the selected subsequence is searched for the pattern. Otherwise, the whole sequence is used. You can also input a custom range.

Algorithm version — version of the algorithm implementation. All versions produce the same result - only the inner technology is used in a different ways:

- *Classic 2* classical implementation of the algorithm. Works everywhere.
- SSE2 the algorithm, which works much faster, but requires you to have SSE2 processor.

Scoring matrix - can be chosen from a bunch of matrices supplied with UGENE. To view a matrix selected click the View button.

Gap open — penalty for opening a gap.

Gap extension — penalty for extending a gap.

*Report results* — simple heuristic which allows to filter intersected hits. If it is set to *none*, the algorithm may report large set of almost identical results in the same region.

Minimal score — another simple heuristic which measures sequences similarity. It is more convenient than using some abstract scores. If set to 100%, the algorithm will search for exact substring match.

## Input and output

The results of the search are saved as annotations or as multiple alignment. To set the saving parameters go to the Input and output tab of the dialog.

- If you want to save the results as annotations input *the annotations saving* parameters (*Annotation name*, *Group name*, *Annotation type*, *Description* and a file to save the annotation to). Also you can add qualifier with corresponding pattern subsequences to result annotations. Check the corresponding checkbox for it.
- If you want to save the results as multiple alignment select the following parameters:

			Smith-Waterman Search	
		Smith-Wate	terman parameters Input and output	
Save results as	Multiple a	ignment		0
Aligner options				
Alignment files d	irectory path	/Users/dsukhomli	linov/	5
Set advanced o	ptions			
Template for a	lignment files	names	[PN]_[SN]_[C]	
Template for r	eference subs	equences names	[SN]_[S]_[E]	
Template for p	attern subsec	quences names	[PN]_[S]_[E]	
Pattern seque	nce name		P1	
[SN	] Reference s	equence name pre	refix [PN] Pattern sequence name prefix	
	[S] Subseque	nce start position	[E] Subsequence end position	
	[hm	s] Time	[L] Subsequence length	
	[C]	Counter	[MDY] Date	
				A 12
Help			Cancel	Alig

Here you can select a file to save the alignment to (Alignment files directory path parameter).

Using the Set advanced options checkbox you can select the saving options.

You can set the different templates for files names: create your own or create by using the following: [E] — adds a subsequence end position, [hms] — adds a time, [MDY] — adds a date, [S] — adds a subsequence start position, [L] — adds a subsequence length, [SN] — adds a reference sequence name prefix, [PN] — adds a pattern sequence name prefix, [C] — adds a counter.

You can create templates for alignment files names, reference subsequence names, pattern subsequence names and for pattern sequence name:

8	File Ad	tions	Settir	ngs	Tools	Win	dow	Help	)											- 8	
	3 🗁	H	•	9	. 🔍		Ţ	Ŷ	10	1	9	9	٠	\$3							
	Project						×	<b>د</b>													
roject	Name filter									Con	sensi										
<u>1</u> :	Objects				R	efere	nce	sub	sequ	ienc	е				AC	GΤ	A C	<b>G</b> -			
	4 🛚 🖁	hum	an_T1.fa	9						_					1 2	3 4	5 6	7 8			
		8 [	;] huma	n_T1	(UCSC	: April	2002	Chu	ıman	T1	5727	74 57	281	[1	A C	GT	A C	GT	)	8]	
	🛯 🖺	P1_h	uman_1	1_1.	aln			P1	17	,				[1	A C	GΤ	A C	G -	5	7]	
		(i	nj P1_h	iuma	n_11																
								Pa	Pattern subsequence												
																			Þ		
								Find	d: 🤙		Alig	nme	nt		<b>\</b>	Ln 1	2 (	Col 1/8	Pos 1	1/8	