Introduction to Smith-Waterman Algorithm

1.Input

a.DNA sequences

Two DNA sequences in text consist of “A”,”C”,”G”,”T”, for example:

ATCATGAGCTA
ATGGGCCT

b.Similarity scoring matrix

A 4 x 4 matrix, and each element denotes the similarity score of a pair, for example:

S	A	C	T	G
A	1	-1	-1	-1
C	-1	1	-1	-1
T	-1	-1	1	-1
G	-1	-1	-1	1

c.Gap Penalty Function

A function compute the penalty for a gap of width k. Normally we, set it as a linear function, for example:

$P(k)= ak + b$, where $a \geqslant 0$, $a+b \geqslant 0$, $k \geqslant 1$

2.Output

The output consists of two aligned sequences, which means the similar region of two original sequences are matched with a ‘|’ . Also, the gaps (skiped in alignment) are denoted by “_”.

In this software, we also output a brief report about the alignment.

3.Algorithm

a. Let m and n denote the length of two input sequences. Let H be a (m+1) x (n+1) matrix.

b. Set the first row and first column of H to 0, i.e. if i=0 or j=0, $H_{i,j}=0$.

c. Iteratively compute the element of H, by:

```
$H_{i,j}$= max{
    $S(SequenceA_{i-1,}\ SequenceB_{j-1})+H_{i-1,j-1}$,
    max{ $H_{i,j-k}$-penalty(k) |$k \geqslant 1$ },
    max{ $H_{i-k,j}$-penalty(k) |$k \geqslant 1$ },
}, $i \geqslant 1$ and $j \geqslant i$
```

d. Traceback from the maxmium of H.

```
    [ x, y ]=arg $\max_{i,j} H_{i,j}$
```

For $H_{i,j}$, we call $H_{i-1,j}$, $H_{i,j-1}$, $H_{i-1,j-1}$ the neighbor of $H_{i,j}$

Traceback step by step, until meet a 0 in the beighbor.

```
While( none of $H_{x,y}$'s neighbor is 0 )
{
    [ x , y ]= arg $\max_{i,j} H_{i,j}$ , $H_{i,j}$ is $H_{x,y}$'s neighbor.
    Record this step.
}
```

e. Determine the aligned sequences.

Determine the aligned sequences according to the traceback record, where diagonal steps represent matches and vertical or horizental steps represent gaps.

4. Improvement on the Algorithm

It's obvious that parallel computation of H will be far more time effiecient.

Some heuristics can also improve the algorithm, like dynamically abandoning computing brinks and corners of H.