An Improved Needleman–Wunsch Algorithm for Pairwise Sequence of Protein–Albumin

Lailil Muflikhah, Dian Eka R.
Faculty of Computer Science, Brawijawa Universiy

Introducción

- Alineación Global
  - Needleman-Wunsch O(MN)
- Alineación Local
  - Smith-Waterman O(MN)
- Algoritmos que usan heurísticas
  - FASTA
  - BLAST (solo para alineación local)

Introducción

$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(A_i, B_j, p) \\ F(i-1,j) \\ F(i,j-1) \end{cases}$$

```
EARDFNQYYSSIKRSGSI
        :::::::
EPKLFIQYYSSIKRTMGH
```

Figure 1. Local alignment

```
EARDFNQYYSSIKRSGSI
:...:..:::::::::....
EPKLFIQYYSSIKRTMGH
```

Figure 2. Global alignment

```
A = GTASC-DG
     :::::. :
B = GTASNND-
```

Figure 3. Sequence alignment

Algoritmo Propuesto (INWA)

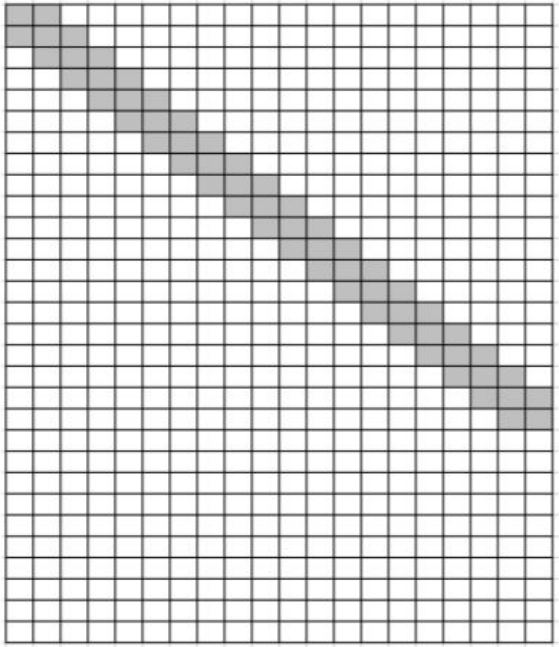

Figure 4. Three main diagonal of 35x20 matrix

Algoritmo Propuesto (INWA)

D	D+1			
D-1	D	D+1		
	D-1	D	D+1	
		D-1	D	D+1
			D-1	D

Figure 5. Marked area for the same length of input sequences

D1	x	x	x	D2	D2+1		
D1-1	D1	x	x	x	D2	D2+1	
	D1-1	D1	x	x	x	D2	D2+1
		D1-1	D1	x	x	x	D2

Figure 6. Marked area for the different length of input sequences

Algoritmo Propuesto

A.1. Marking area to fill in the matrix

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣						
3	G						
4	E						
5	S						
6	K						
7	C						

Figure 7. Marked area from case study 1

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑					
4	E						
5	S						
6	K						
7	C						

Figure 8. Initialization from case study 1

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑	2↖	0←			
4	E		0↑	1↖	-1←		
5	S			-1↑	0↖	1↖	
6	K				-2↑	-1↑	0↖
7	C					-3↑	1↖

Figure 9. All marked area has been filled from case study 1

B.1. Marking area to fill in the matrix

The marked area in this case can be seen at Figure 11.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣						
3	G						
4	P						
5	T						
6	G						
7	T						
8	G						
9	E						
10	S						
11	K						
12	C						

Figure 11. Marked area from case study 2

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑					
4	P	-4↑					
5	T	-6↑					
6	G	-8↑					
7	T	-10↑					
8	G	-12↑					
9	E						
10	S						
11	K						
12	C						

Figure 12. Initialization from case study 2

B.3. Filling the marked area in matrix

All marked cells of matrix are filled as in Figure 13.

	1	2	3	4	5	6	7
1		␣	G	T	A	S	C
2	␣	0	-2←				
3	G	-2↑	2↖	0←			
4	P	-4↑	0↑	1↖	-1←		
5	T	-6↑	-2↑	2↖	0←	-2←	
6	G	-8↑	-4↑	0↑	1↖	-1←	-3←
7	T	-10↑	-6↑	-2↑	-1↑	0↖	-2←
8	G	-12↑	-8↑	-4↑	-3↑	-2↑	-1↖
9	E		-10↑	-6↑	-5↑	-4↑	-3↑
10	S			-8↑	-7↑	-3↖	-5↑
11	K				-9↑	-5↑	-4↖
12	C					-7↑	-3↖

Figure 13. All marked area has been filled from case study 2

	1	2	3	4	5	6	7
1		_	G	T	A	S	C
2	_	0	-2				
3	G	-2	2↖	0			
4	P	-4	0↑	1	-1		
5	T	-6	-2	2↖	0	-2	
6	G	-8	-4	0	1↖	-1	-3
7	T	-10	-6	-2	-1↑	0	-2
8	G	-12	-8	-4	-3↑	-2	-1
9	E		-10	-6	-5↑	-4	-3
10	S			-8	-7	-3↖	-5
11	K				-9	-5↑	-4
12	C					-7	-3↖

Figure 14. Backtracking path from case study 2

isoform CRA_q :

```
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEE
NFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADES
AENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERN
ECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLY
EIARRHPYFYAPELLFFAACCQSSMNFGMKGRLRLPNRDSS
VPVSKNLEKELSKHGQ
```

isoform CRA_p :

```
MSQLKICELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEV
SRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKT
PVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFT
FHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMD
DFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

The both sequences are aligned globally using the proposed method with default parameter value (match=1; mismatch=-1; gap=-1) and the result is as below:

```
. Alignment score      :-9
  Length of sequence 1:221
  Length of sequence 2:200
  Sequence identity    :29/274 (0.11%)
  Positives            :49/274 (0.18%)
  Gaps                 :127/274 (0.46%)
  HSSP                 :-9.74 (not similar)
  Filled Matrix        :4822
  Execution Time : 49 ms
```

Table 1
The number of filled matrix areas

Pair sequence no.	Original NW	Proposed method (INWA)	The score of result
1	99645	52339	Same
2	89790	42484	Same
3	118479	71173	Same
4	94827	47521	Same
5	271998	224692	Same
6	114756	67450	Same
7	131400	84094	Sane
8	36792	9070	Same
9	76212	28906	Same
10	139941	92635	Same
…	…	…	…
…	…	…	…
1250	41925	37893	same

Resultados

- La evaluación se implementa mediante diez ensayos contra 1250 alineaciones secuenciales por pares de proteína humana-albúmina.
- Además, el tiempo computacional utilizando Needleman-Wunsch original es de 5.382 segundos y el otro con INWA de 3.988 segundos.
- Significa que hay una reducción de tiempo del 25.9% del método original.
- Los resultados de alineación por INWA son los mismos que Needleman-Wunsch.
- Esto significa que el tiempo computacional del método propuesto se reduce pero realiza el resultado de alineación óptimo.
- Además, la complejidad de tiempo y espacio del método propuesto (INWA) es O (N). Son menos que el método original NW de O (MN).

Conclusiones

- Se ha aplicado una mejora del algoritmo de Needleman-Wunsch (INWA) para alinear la secuencia de pares para la proteína-albúmina humana.
- La idea principal del método propuesto es omitir los datos no utilizados por el área restante en blanco a fin de obtener el menor tiempo de cálculo y reducir la complejidad del espacio.
- INWA solo llena parcialmente la celda de la matriz.
- Este algoritmo se puede aplicar a ambos tipos de alineación de secuencia, ya sea que las secuencias de entrada tengan la misma longitud o no.
- Además, la complejidad de espacio y tiempo de INWA es O (N).
- Es mejor que el algoritmo original de Needleman-Wunch que es O (MN).
- Además, el tiempo de ejecución de INWA es un 25.9% más rápido que el algoritmo original de Needleman-Wunsch.