Bioinformatics Algorithms

(Fundamental Algorithms, module 2)

Zsuzsanna Lipták

Masters in Medical Bioinformatics
academic year 2017/18, spring term

Pairwise Alignment 2

Semiglobal alignment

match: 1,
mismatch: -1,
gap: -1

```
CAGCGTACACT        CAGCGTACACT
---CCTA----        C--C-T--A--
-----------        -----------
 score −5           score −3
```

Semiglobal alignment

match: 1,
mismatch: -1,
gap: -1

```
CAGCGTACACT        CAGCGTACACT
---CCTA----        C--C-T--A--
-----------        -----------
 score −5           score −3
```

- The left alignment seems better, but it has a lower score.
- We would like the extremal gaps (before and after the second string) not to count at all.
- Note that this is not covered by local alignment (why?).

Semiglobal alignment

match: 1,
mismatch: -1,
gap: -1

If we do not count the extremal gaps, then we get:

```
CAGCGTACACT        CAGCGTACACT
---CCTA----        C--C-T--A--
-----------        -----------
  score 2           score −1
```

. . . as desired, the score now reflects that the left alignment is better than the right one.

Semiglobal alignment: algorithm

gaps matched here should be free	action
beginning of *s*	0s in first column
end of *s*	maximize over last column
beginning of *t*	0s in first row
end of *t*	maximize over last row

Semiglobal alignment: algorithm

gaps matched here should be free	action
beginning of *s*	0s in first column
end of *s*	maximize over last column
beginning of *t*	0s in first row
end of *t*	maximize over last row

Analysis

time and space $O(nm)$

Semiglobal alignment: example

The global similarity of the two strings $s = \texttt{ACGC}$ and $t = \texttt{GCTC}$ is 0, with (unique) optimal alignment $\binom{\texttt{ACGC}}{\texttt{GCTC}}$. Let us compute an optimal semiglobal alignment of s and t, where we set all four types of external gaps as free, and match: $+1$, mism., gap $=$ -1.

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

optimal semiglobal alignment:

```
ACGC--
--GCTC
```

score $= 2$

Semiglobal alignment

N.B.

- Semiglobal alignment is also called *end-space-free alignment*.
- It is not *one* algorithm, but (strictly speaking) 15 different ones, depending on where we want to have charge-free gaps (e.g. beginning and end of first sequence; beginning of first, end of second; etc.)

Applications include:

- find a prefix of s with maximum similarity to t - which variant do we need?
- overlap finding (e.g. for sequence assembly): find prefix s' of s and suffix t' of t s.t. $sim(s', t')$ maximal, or vice versa (prefix of t with suffix of s) - which variant do we need?
- substring match: find a substring s' of s with $sim(s', t)$ maximal - which variant do we need?