

Bioinformatics Algorithms

(Fundamental Algorithms, module 2)

Zsuzsanna Lipták

Masters in Medical Bioinformatics
academic year 2018/19, II semester

Pairwise Alignment 1

Alignments

Alignment

- a way of visualizing similarities and differences between two strings
- we want to find a **good** way of doing this

Ex: five different alignments of $s = \text{ACCT}$ and $t = \text{CAT}$

-ACCT	ACCT	ACCT	ACC-T	---ACCT
CA--T	-CAT	CAT-	--CAT	CAT----

2 / 34

Alignments

Alignment

- a way of visualizing similarities and differences between two strings
- we want to find a **good** way of doing this

Ex: five different alignments of $s = \text{ACCT}$ and $t = \text{CAT}$

-ACCT	ACCT	ACCT	ACC-T	---ACCT
CA--T	-CAT	CAT-	--CAT	CAT----

Formal definition

An **alignment** $\mathcal{A}$ of $s, t \in \Sigma^*$ is a matrix with two rows, entries from $\Sigma \cup \{-\}$, s.t.
gap

- deleting all gaps from the first row yields s
- deleting all gaps from the second row yields t
- no column consists of two gaps

2 / 34

Scoring alignments

scoring function

- score of a column: **match** (same char), **mismatch** (diff. chars), **gap**
in general: scoring function $f : \Sigma \cup \{-\} \times \Sigma \cup \{-\} \rightarrow \mathbb{R}$
- score of $\mathcal{A}$** = sum of column scores

Ex.

	match	mismatch	gap
	2	-1	-1

-ACCT	ACCT	ACCT	ACC-T	---ACCT
CA--T	-CAT	CAT-	--CAT	CAT----

3 / 34

Scoring alignments

scoring function

- score of a column: **match** (same char), **mismatch** (diff. chars), **gap**
in general: scoring function $f : \Sigma \cup \{-\} \times \Sigma \cup \{-\} \rightarrow \mathbb{R}$
- score of $\mathcal{A}$** = sum of column scores

Ex.

	match	mismatch	gap
	2	-1	-1

-ACCT	ACCT	ACCT	ACC-T	---ACCT
CA--T	-CAT	CAT-	--CAT	CAT----
1	2	-4	1	-7

3 / 34

Scoring alignments

So acc. to our scoring function, alignment 2 is the best (of the five)!

-ACCT	ACCT	ACCT	ACC-T	---ACCT
CA--T	-CAT	CAT-	--CAT	CAT----
1	2	-4	1	-7

But is it **best possible**?

4 / 34

Optimal alignments

Def.

An **optimal alignment** of s and t is an alignment $\mathcal{A}$ with maximum score, i.e. an alignment $\mathcal{A}$ s.t.

$$\text{score}(\mathcal{A}) = \max\{\text{score}(\mathcal{A}') : \mathcal{A}' \text{ is an alignment of } s \text{ and } t\}$$

Def.

Given $s, t \in \Sigma^*$ and scoring function f , the **similarity** of s and t , is

$$\begin{aligned} \text{sim}(s, t) &= \text{score of an optimal alignment} \\ &= \max\{\text{score}(\mathcal{A}) : \mathcal{A} \text{ is an alignment of } s \text{ and } t\} \end{aligned}$$

5 / 34

Our computational problem: Global alignment

Problem variant 1

Input: Two strings s, t over alphabet Σ , scoring function f .

Output: $\text{sim}(s, t)$.

Problem variant 2

Input: Two strings s, t over alphabet Σ , scoring function f .

Output: An optimal alignment of s and t .

N.B.: In variant 1, we want only a number, we are not interested in an optimal alignment itself.

6 / 34

Our computational problem: Global alignment

For now, let's concentrate on Variant 1 (i.e. only $\text{sim}(s, t)$ is sought).

Global alignment

Input: Two strings s, t over alphabet Σ , scoring function f .

Output: $\text{sim}(s, t)$.

We will see two algorithms for this problem.

7 / 34

Exhaustive search

Algorithm 1: Exhaustive search

1. consider every possible alignment of s and t
2. for each of these, compute its score
3. output the maximum of the scores computed

8 / 34

Number of alignments

List all alignments of $s = AC$ and $t = GA$.

Algorithm Exhaustive search for global alignment

Input: strings s, t , with $|s| = n, |t| = m$; scoring function f

Output: value $\text{sim}(s, t)$

1. `int max = (n + m)g;` // g is the cost of a gap
2. **for** each alignment $\mathcal{A}$ of s and t (in some order)
3. **do if** $\text{score}(\mathcal{A}) > \text{max}$
4. **then** $\text{max} \leftarrow \text{score}(\mathcal{A})$;
5. **return** max ;

Note:

1. The variable max is needed for storing the highest score so far seen.
2. The initial value of max is the score of *some* alignment of s, t (which one?)

9 / 34

10 / 34