BLAST : Basic Local Alignment Search Tool

Projet court

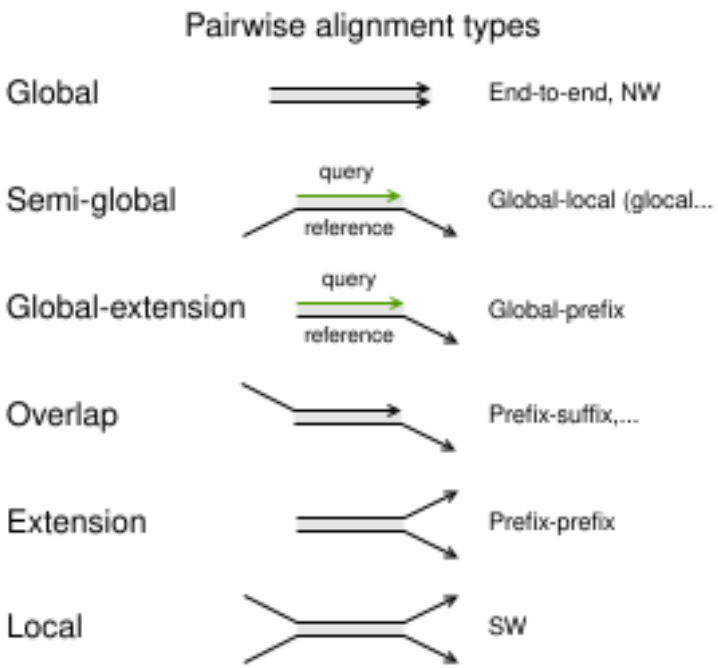
Amory Antao
amoryantao@net-c.com

Université de Paris Cité
M2 Biologie Informatique

September 12, 2025

Introduction and Motivation

- Local alignments → function and evolution




```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Introduction and Motivation

- Local alignments $\rightarrow$ function and evolution
- Explosion of biological data

Introduction and Motivation

- Local alignments $\rightarrow$ function and evolution
- Explosion of biological data
- Limits of exact methods $\rightarrow$ need for heuristics

BLAST: General Principles

Three main steps:

1. Construction of the word list

BLAST: General Principles

Three main steps:

1. Construction of the word list
2. Search for hits

BLAST: General Principles

Three main steps:

1. Construction of the word list
2. Search for hits
3. Extension of hits (X-drop)

BLAST I: Protein Example

Query:

M	A	T	G	L	A

BLAST I: Protein Example

- $w = 3$, word generation.

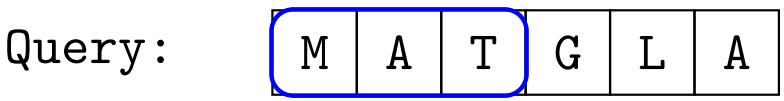
Query:

M	A	T	G	L	A

Words =

BLAST I: Protein Example

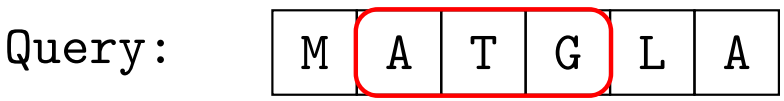
- $w = 3$, word generation.



Words = MAT

BLAST I: Protein Example

- $w = 3$, word generation.



Words = MAT ATG

BLAST I: Protein Example

- $w = 3$, word generation.

Query:

M	A	T	G	L	A

Words = MAT ATG ...

BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T

	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R		5	0	-2	-3	1	0
N			6	1	-3	0	0
D				6	-3	0	2
C					9	-3	-4
Q						5	2
E							5

Extract of BLOSUM62

BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T

	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R		5	0	-2	-3	1	0
N			6	1	-3	0	0
D				6	-3	0	2
C					9	-3	-4
Q						5	2
E							5

Extract of BLOSUM62

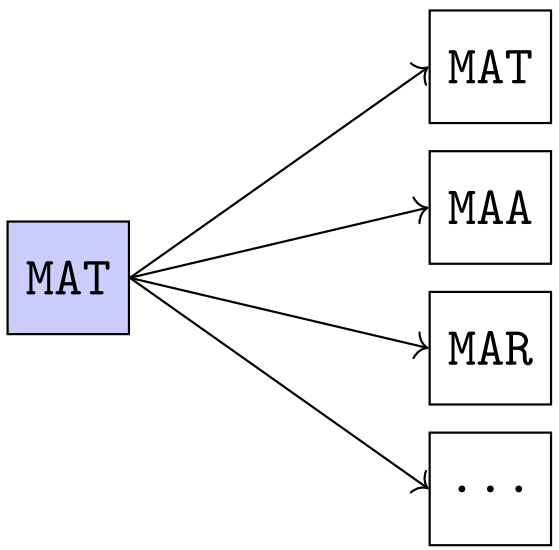
MAT

BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T

	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R		5	0	-2	-3	1	0
N			6	1	-3	0	0
D				6	-3	0	2
C					9	-3	-4
Q						5	2
E							5

Extract of BLOSUM62

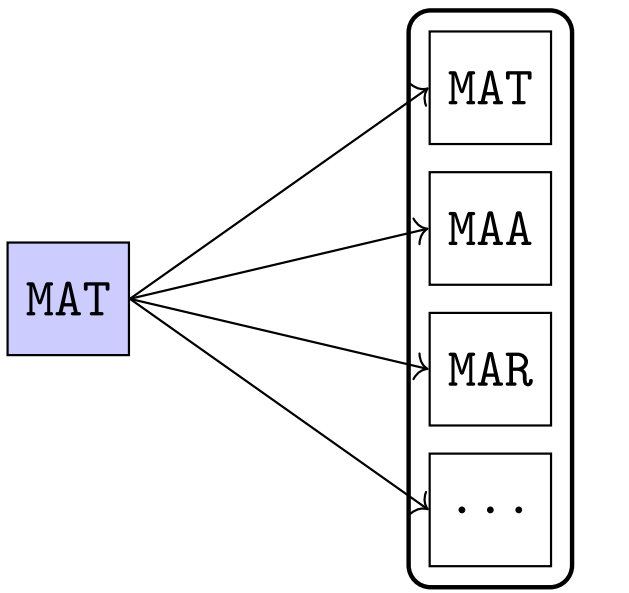


BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T

	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R		5	0	-2	-3	1	0
N			6	1	-3	0	0
D				6	-3	0	2
C					9	-3	-4
Q						5	2
E							5

Extract of BLOSUM62

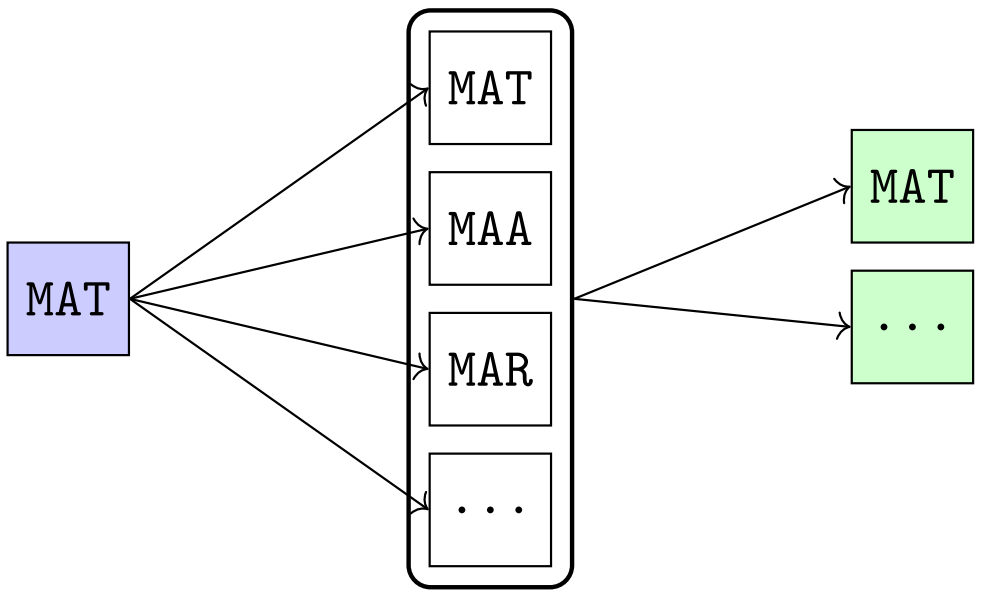


BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T

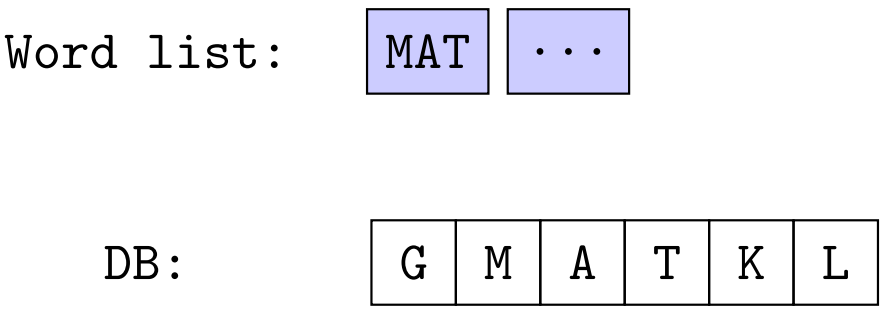
	A	R	N	D	C	Q	E
A	4	-1	-2	-2	0	-1	-1
R		5	0	-2	-3	1	0
N			6	1	-3	0	0
D				6	-3	0	2
C					9	-3	-4
Q						5	2
E							5

Extract of BLOSUM62



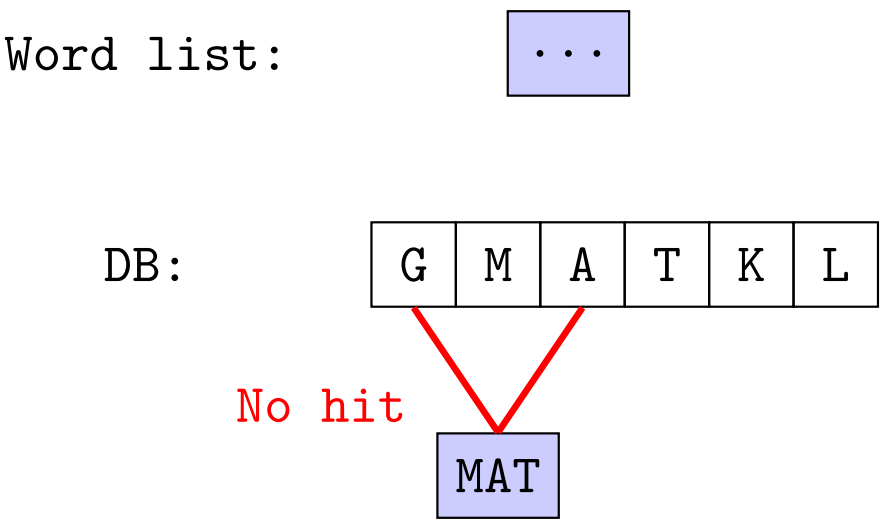
BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database



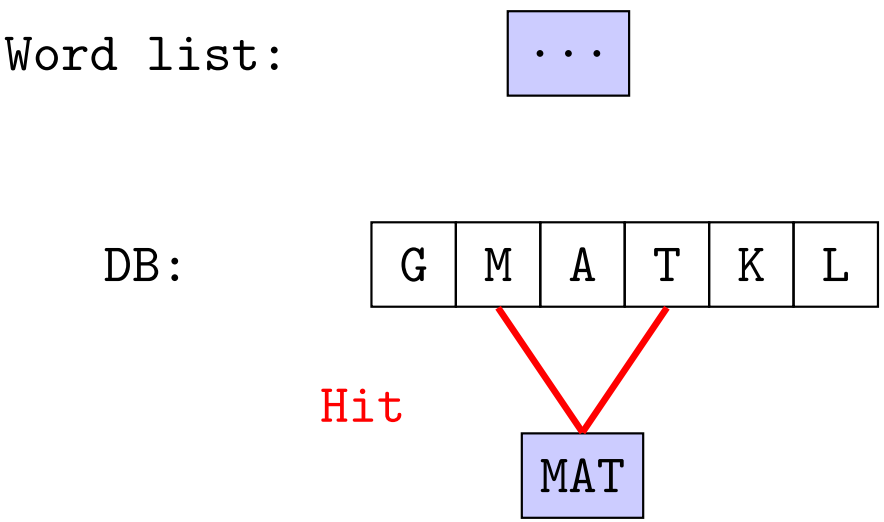
BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database



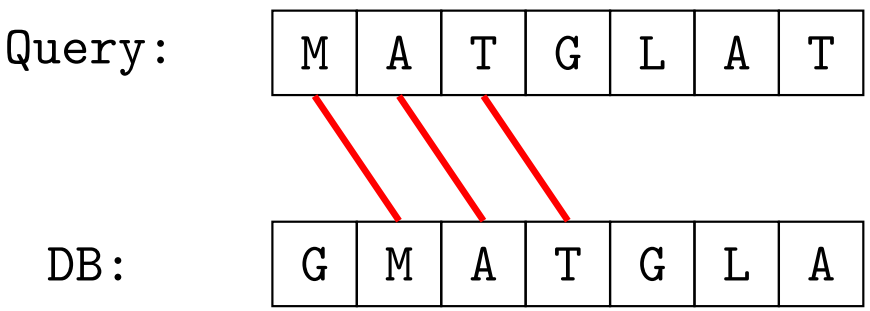
BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database



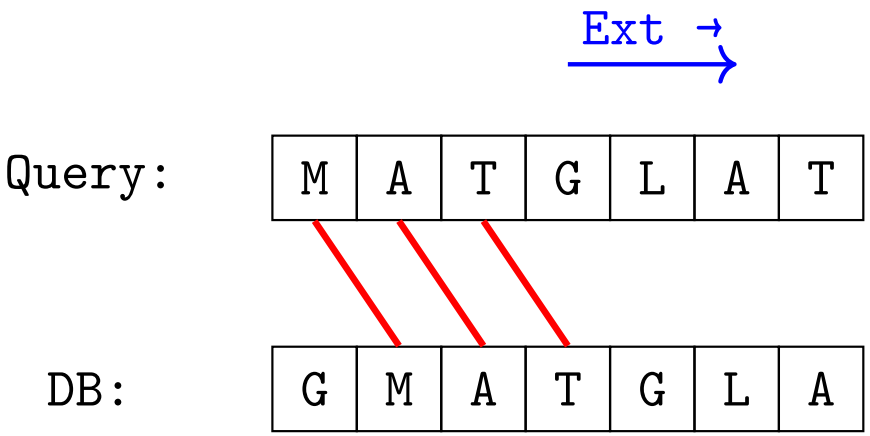
BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database
- Ungapped extension with X-drop



BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database
- Ungapped extension with X-drop

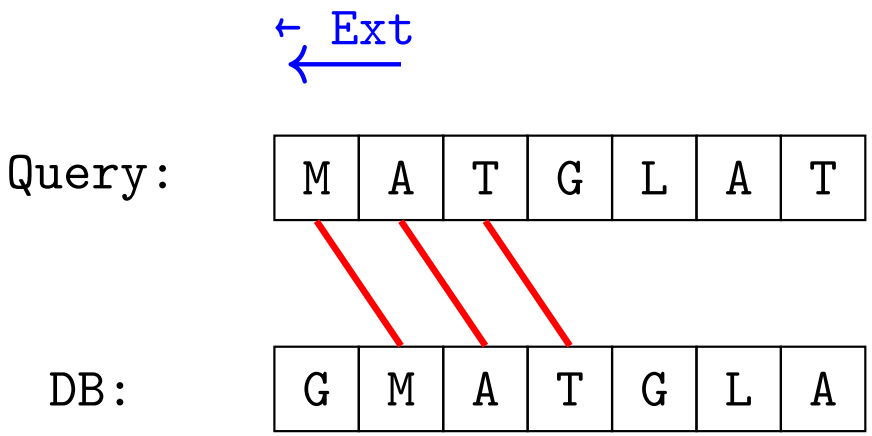


$$score \overset{?}{<} score\_max - x\_drop$$

stop

BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database
- Ungapped extension with X-drop

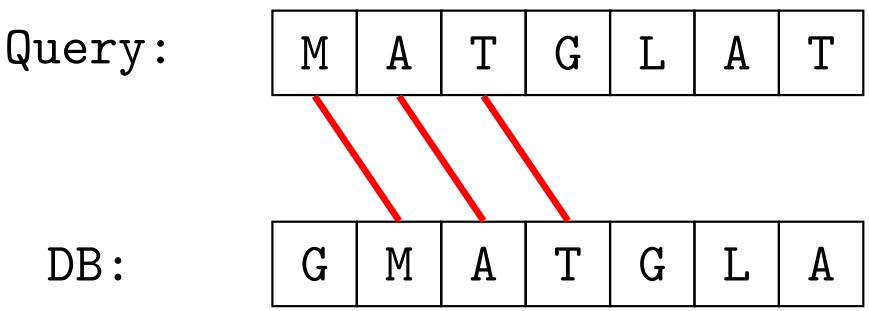


$$score \overset{?}{<} score\_max - x\_drop$$

stop

BLAST I: Protein Example

- $w = 3$, word generation.
- Neighbors via substitution matrix and threshold T
- Search for hits in the database
- Ungapped extension with X-drop



Final result: score_max right + score_max left

BLAST II: Protein Example

- Word generation

BLAST II: Protein Example

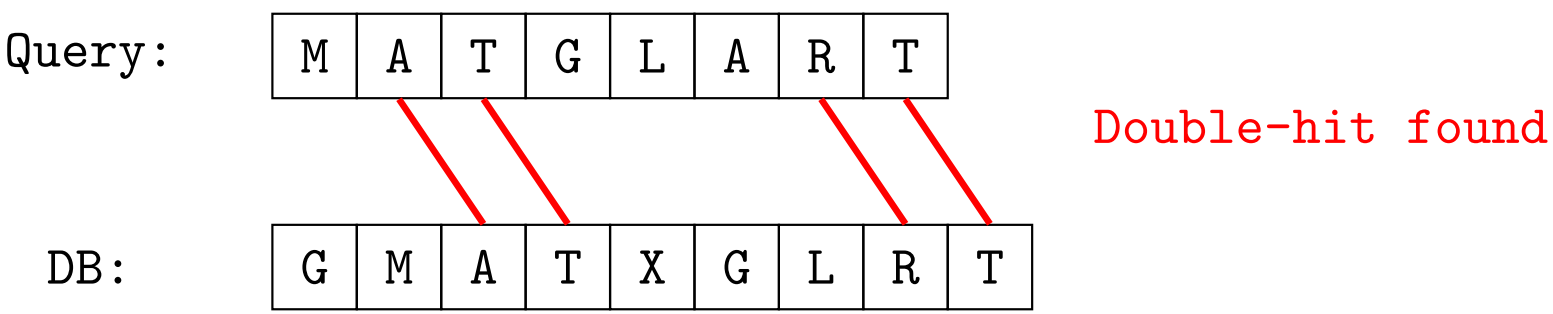
- Word generation
- Double-hit: two close words on the same diagonal

Query:	M	A	T	G	L	A	R	T

DB:	G	M	A	T	X	G	L	R	T

BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal



BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)

BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)
- Gapped extension

BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)
- Gapped extension
  - Take gaps into account

BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)
- Gapped extension
  - Take gaps into account
  - Gotoh algorithm (1982)

BLAST II: Protein Example

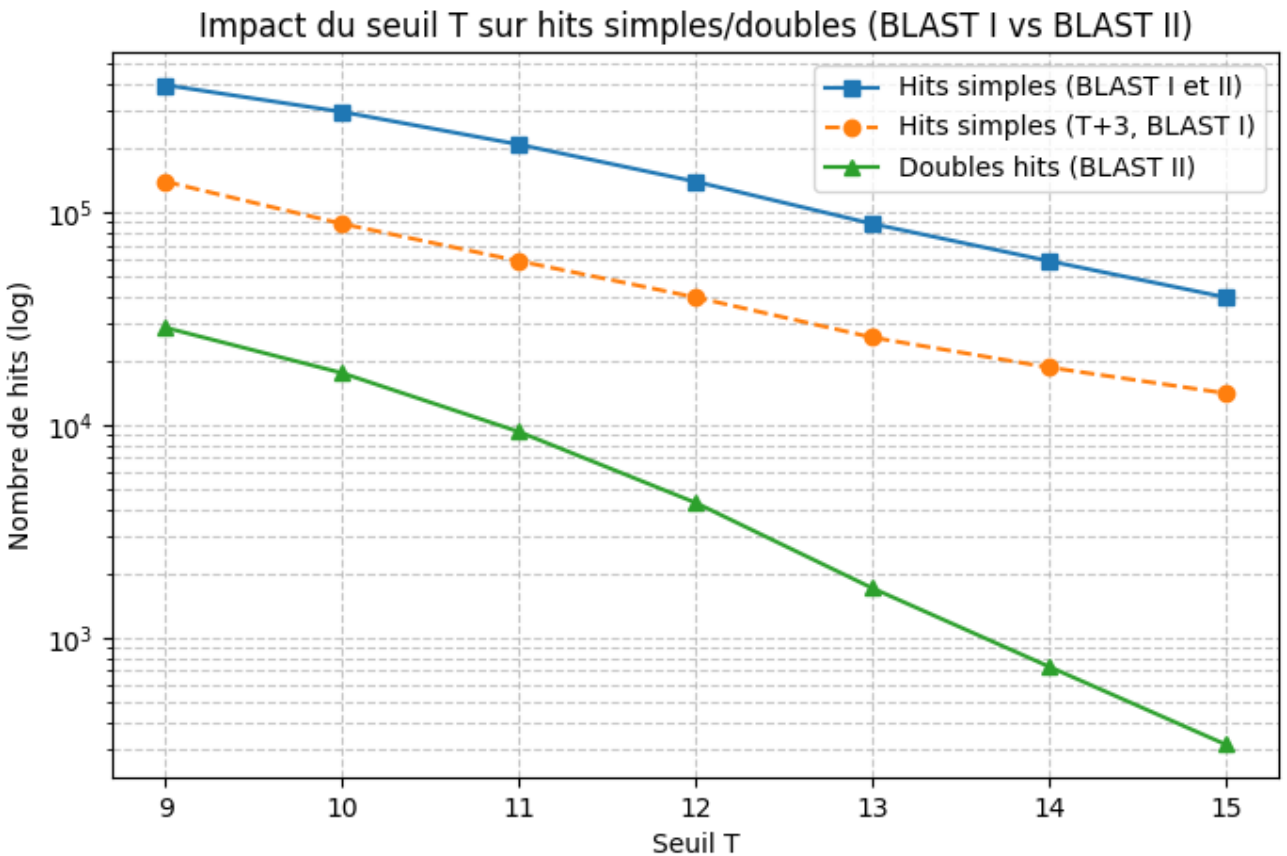
- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)
- Gapped extension
  - Take gaps into account
  - Gotoh algorithm (1982)
  - Gap opening and extension penalties

BLAST II: Protein Example

- Word generation
- Double-hit: two close words on the same diagonal
- Ungapped extension on the second hit
  ($score\_max \geq$ threshold $\rightarrow$ gapped extension)
- Gapped extension
  - Take gaps into account
  - Gotoh algorithm (1982)
  - Gap opening and extension penalties
  - Higher x_drop

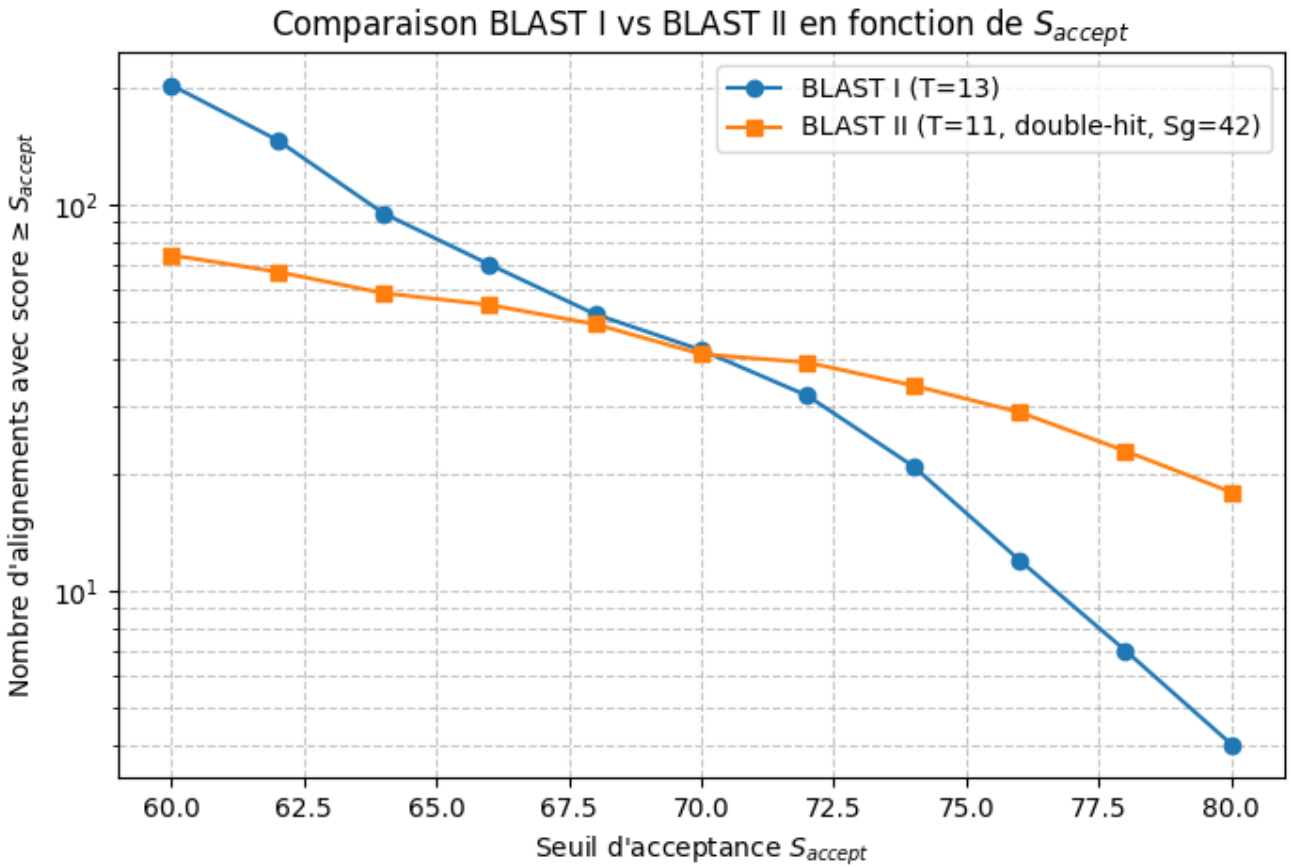
Experimental Results

- Number of hits (single/double, BLAST I vs II).



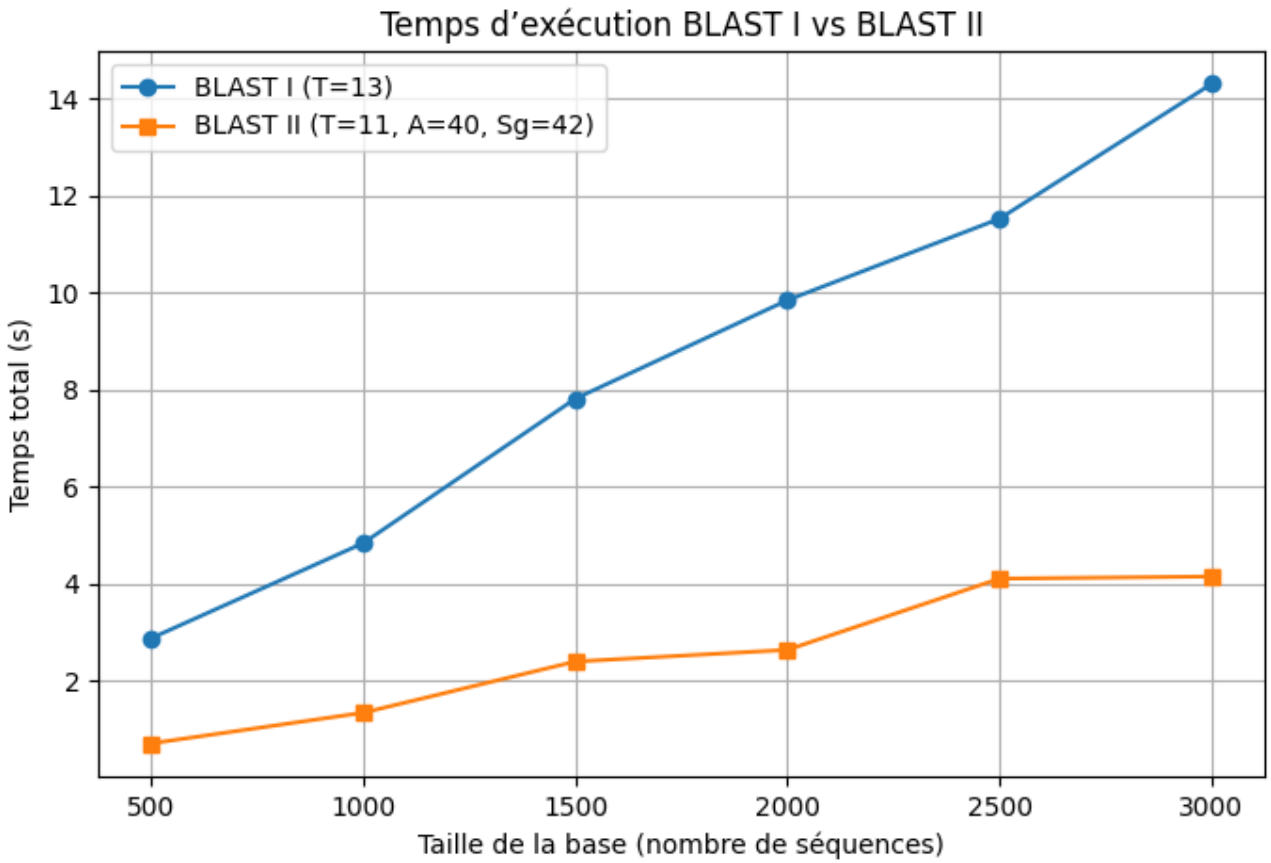
Experimental Results

- Number of hits (single/double, BLAST I vs II).
- Accepted alignments (BLAST I vs II).



Experimental Results

- Number of hits (single/double, BLAST I vs II).
- Accepted alignments (BLAST I vs II).
- Runtime (BLAST II $\approx 3\times$ faster).



Discussion and Perspectives

Discussion

- Results consistent with the literature.

Perspectives

Discussion and Perspectives

Discussion

- Results consistent with the literature.
- BLAST I: simpler but noisier, less robust.

Perspectives

Discussion and Perspectives

Discussion

- Results consistent with the literature.
- BLAST I: simpler but noisier, less robust.
- BLAST II: faster and more sensitive.

Perspectives

Discussion and Perspectives

Discussion

- Results consistent with the literature.
- BLAST I: simpler but noisier, less robust.
- BLAST II: faster and more sensitive.

Perspectives

- Evaluate on real datasets.

Discussion and Perspectives

Discussion

- Results consistent with the literature.
- BLAST I: simpler but noisier, less robust.
- BLAST II: faster and more sensitive.

Perspectives

- Evaluate on real datasets.
- Optimizations in the code.