

# Subset Seeds on a Reconfigurable Architecture

## LAW 2007

Gilles Georges\*    Mathieu Giraud•    Julien Jacques\*  
Gregory Kucherov•    Dominique Lavenier\*    Laurent Noé•  
Pierre Peterlongo\*

INRIA, (\* IRISA Rennes Symbiose & • Lille LIFL Séquoia)



21 Septembre 2006



# Overview

Motivations

Subset seeds

Where are we now ?

Specialized architecture

Conclusion

# Overview

Motivations

Subset seeds

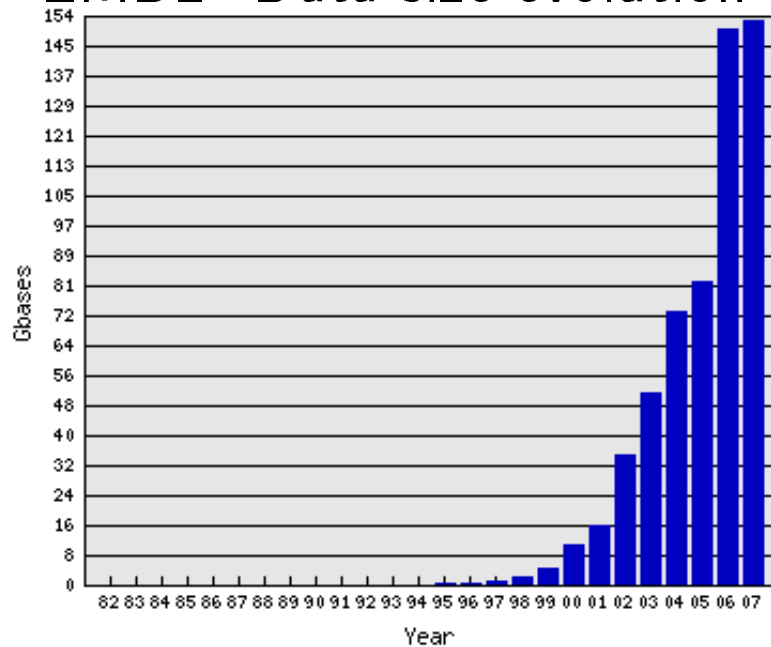
Where are we now?

Specialized architecture

Conclusion

# Biological Data evolution

EMBL<sup>a</sup> Data size evolution :



<sup>a</sup><http://www.ebi.ac.uk/embl/> - Europe's primary nucleotide sequence resource

Data type evolution :

- Take the "*Junk DNA*" into account

Large augmentation of amount of data

# Main goal of the study

Biological similarities detection

Improvements of BLAST<sup>1</sup>-like programs :

- Faster execution
- Larger amount of data
- More sensitive results

---

<sup>1</sup>Altschul, S.; Gish, W.; Miller, W.; Myers, E. & Lipman, D. Basic Local Alignment Search Tool *Journal of Molecular Biology*, **1990**, 215, 403-410

# Overview

Motivations

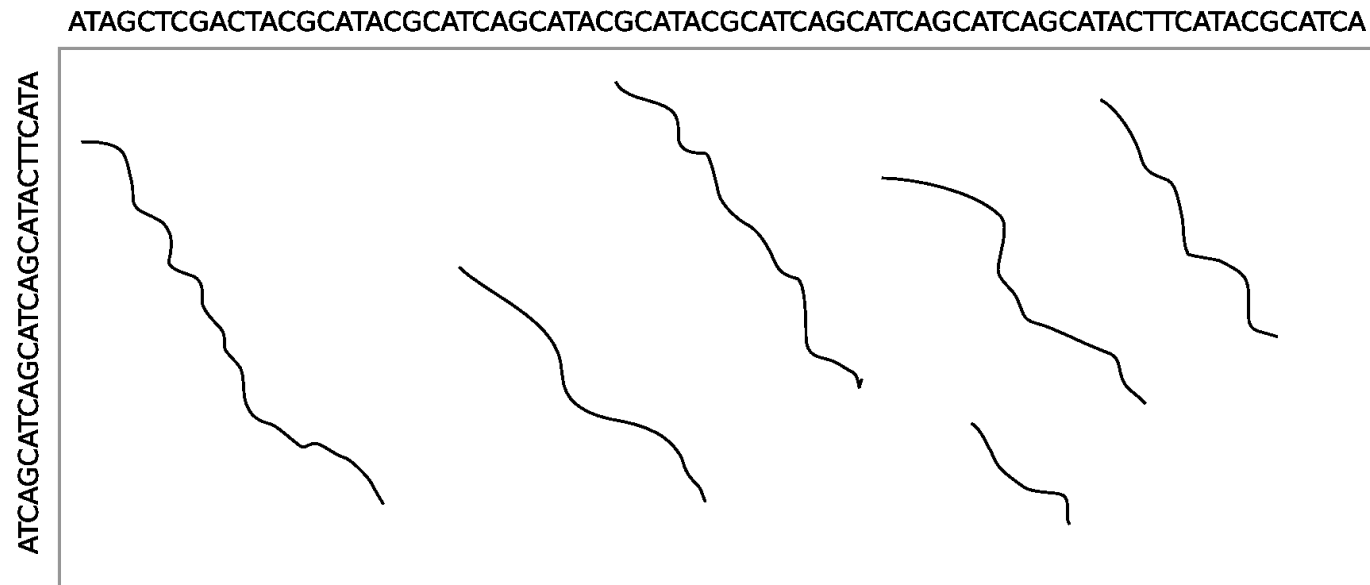
**Subset seeds**

Where are we now ?

Specialized architecture

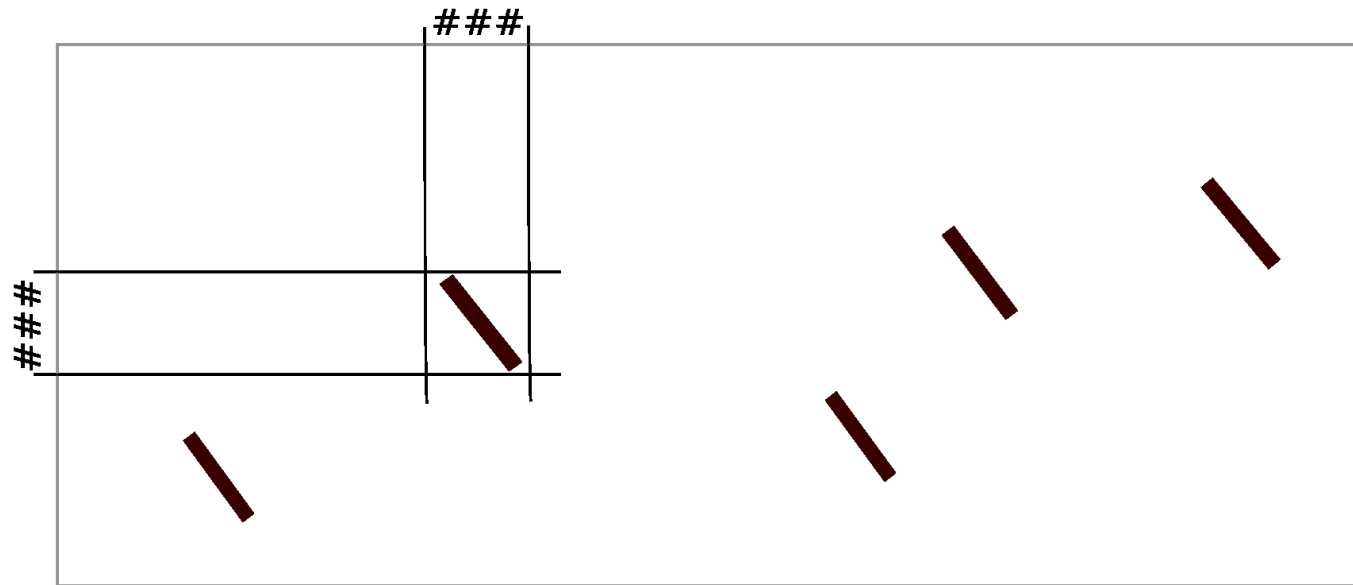
Conclusion

# Seeds - Basic Ideas



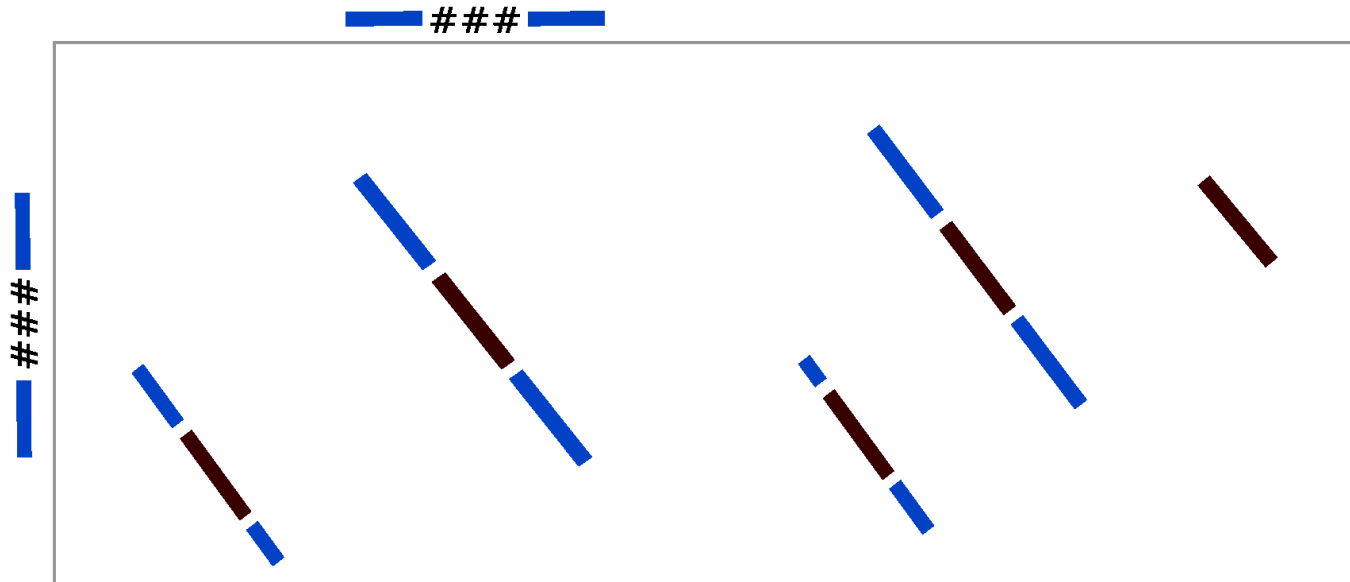
Computing local alignments, using dynamic programming

# Seeds - Basic Ideas



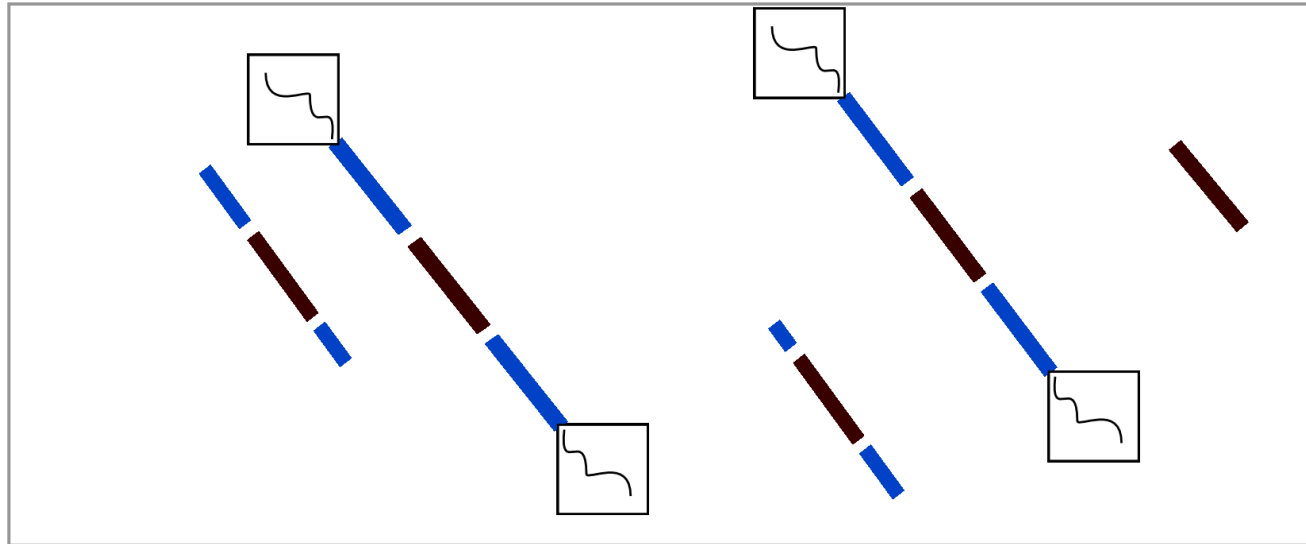
1. Detect matching seeds - **indexation**

# Seeds - Basic Ideas



1. Detect matching seeds - **indexation**
2. Extend to neighbors

# Seeds - Basic Ideas



1. Detect matching seeds - **indexation**
2. Extend to neighbors
3. Perform alignments

# Seeds - Sensitivity and Specificity

## Representation, specificity and sensitivity

- ### : “*classical*” seed

# Seeds - Sensitivity and Specificity

## Representation, specificity and sensitivity

- ### : “*classical*” seed
- # : Low specificity & High sensitivity  
(Slow and precise)

# Seeds - Sensitivity and Specificity

## Representation, specificity and sensitivity

- ### : “classical” seed
- # : Low specificity & High sensitivity  
(Slow and precise)
- ##### : High specificity & Low sensitivity  
(Fast and imprecise)

# Seeds - Sensitivity and Specificity

## Representation, specificity and sensitivity

- ### : “classical” seed
- # : Low specificity & High sensitivity  
(Slow and precise)
- ##### : High specificity & Low sensitivity  
(Fast and imprecise)

## What is the good seed?

Main difficulty : **design** seeds to have best ratio  
specificity v.s. sensitivity.

# Spaced<sup>2</sup> seeds

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#	#	#	#													

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#																

---

<sup>2</sup>Burkhardt, S., & Kärkkäinen, J. Better filtering with gapped q-grams. *Fundamenta Informaticae* 56, 1-2 (2003), 51-70. Preliminary version in *Combinatorial Pattern Matching* **2001**

# Spaced<sup>2</sup> seeds

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#	-	-	#	-	#	-	#	-	#	-	#	-	#	-	#	#
#	#	#	-	-	#	-	#	-	#	-	#	-	#	-	#	-	#	#

---

<sup>2</sup>Burkhardt, S., & Kärkkäinen, J. Better filtering with gapped q-grams. *Fundamenta Informaticae* 56, 1-2 (2003), 51-70. Preliminary version in *Combinatorial Pattern Matching* **2001**

# Spaced<sup>2</sup> seeds

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#	#	#	#													
	#	#	#	#	#													
		#	#	#	#	#												

A	T	C	A	G	T	G	C	A	A	T	G	C	T	C	A	A	G	A
					.			.					.					
A	T	C	A	G	C	G	C	G	A	T	G	C	G	C	A	A	G	A
#	#	#	-	-	#	-	#	#										
	#	#	-	-	#	-	#	#										
		#	#	#	-	-	#	-	#	#								

---

<sup>2</sup>Burkhardt, S., & Kärkkäinen, J. Better filtering with gapped q-grams. *Fundamenta Informaticae* 56, 1-2 (2003), 51-70. Preliminary version in *Combinatorial Pattern Matching* **2001**



# Spaced seeds

```

ATCAGTGCAATGCTCAAGA
||| | | . || . || | | . | | | |
ATCAGCGCGATGCGCAAGA
### - - # - ##
          ### - - # - ##
                ### - - # - ##

```

Spaced seeds have better sensitivity

# Multiple<sup>3</sup> spaced seeds

Instead of a unique Spaced seed  
{###- -##-##} (weight 6)

---

<sup>3</sup>M. Li, B. Ma, D. Kisman & J. Tromp PatternHunter II : Highly sensitive and fast homology search. *J. of Bioinformatics and Comp. Biol.*, 2(3), 417-439, 2004.

# Multiple<sup>3</sup> spaced seeds

Instead of a unique Spaced seed  
 $\{\#\#\#- -\#-\#\#\}$  (weight 6)

Use a set of spaced seeds

{  
 $\#\#\#- -\#-\#\#\#$ , (weight 7)  
 $\#- -\#\#\#-\#-\#\#$ , (weight 7)  
 $\#- -\#\#\#- -\#\#-\#$ , (weight 7)  
 ... }

---

<sup>3</sup>M. Li, B. Ma, D. Kisman & J. Tromp PatternHunter II : Highly sensitive and fast homology search. *J. of Bioinformatics and Comp. Biol.*, 2(3), 417-439, 2004.

## Multiple<sup>3</sup> spaced seeds

Instead of a unique Spaced seed  
 $\{\#\#\#- -\#-\#\#\}$  (weight 6)

Use a set of spaced seeds

{  
 $\#\#\#- -\#-\#\#\#$ , (weight 7)  
 $\#- -\#\#\#-\#-\#\#$ , (weight 7)  
 $\#- -\#\#\#- -\#\#-\#$ , (weight 7)  
 ... }

### Advantages and drawback

- :-) Better sensitivity
- :-( Higher memory usage, (and slower)

---

<sup>3</sup>M. Li, B. Ma, D. Kisman & J. Tromp PatternHunter II : Highly sensitive and fast homology search. *J. of Bioinformatics and Comp. Biol.*, 2(3), 417-439, 2004.

# Subset seeds (protein example)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Kucherov, G.; Noé, L. & Roytberg, M. A Unifying Framework for Seed Sensitivity and its Application to Subset Seeds *WABI*, 2005, 3692 of LNCS, 251-263

# Subset seeds (protein example)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

All amino acids are not equivalents

- Create groups of characters

# Subset seeds (protein example)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

All amino acids are not equivalent

- Create groups of characters

```
# | A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
- | CFYWMLIVGPATSNHQEDRK
@1 | C, STPAG, NDEQ, HRK, MILV, FYW
@2 | CFYWMLIV, GPATSNHQEDRK
# | A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
```

# So what?

Design good set(s) of subset (spaced) seeds

- Speed
- Memory
- Specificity
- Sensitivity

# Overview

Motivations

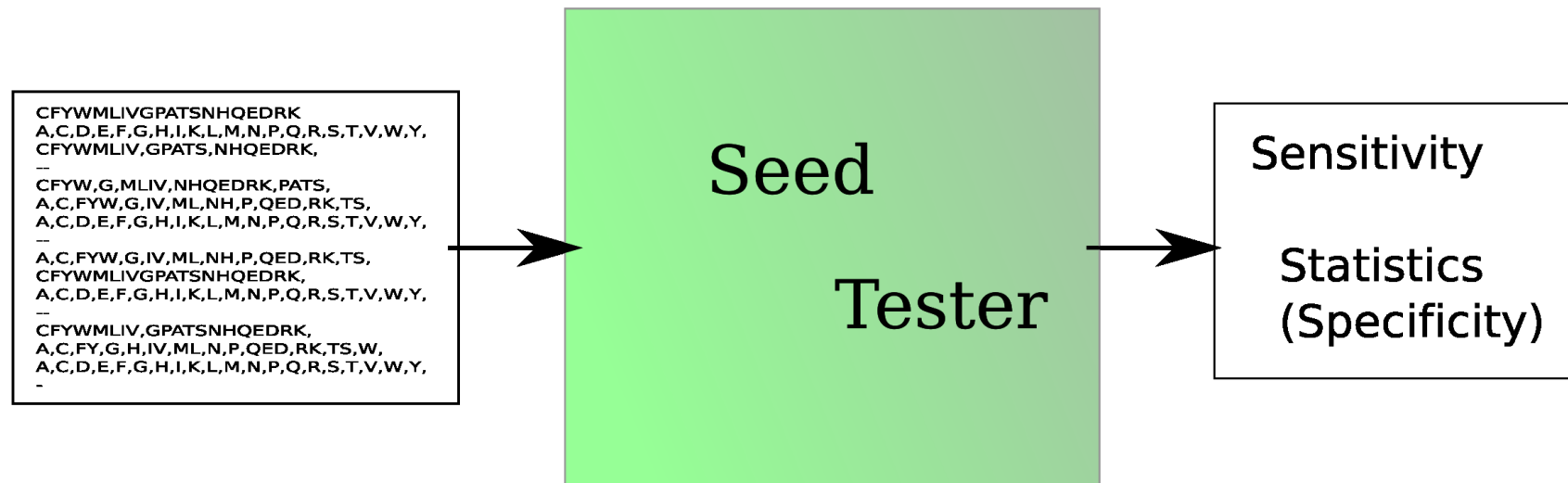
Subset seeds

Where are we now ?

Specialized architecture

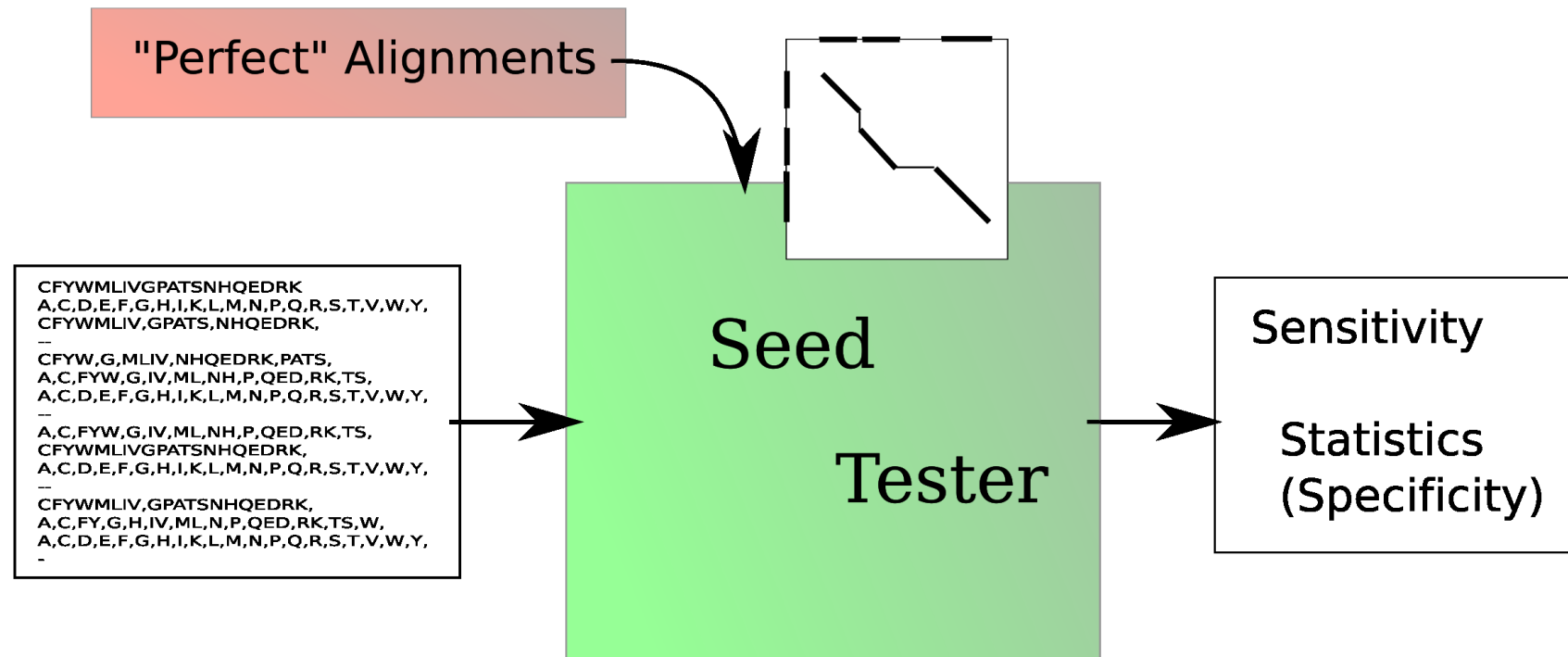
Conclusion

# Seed tester



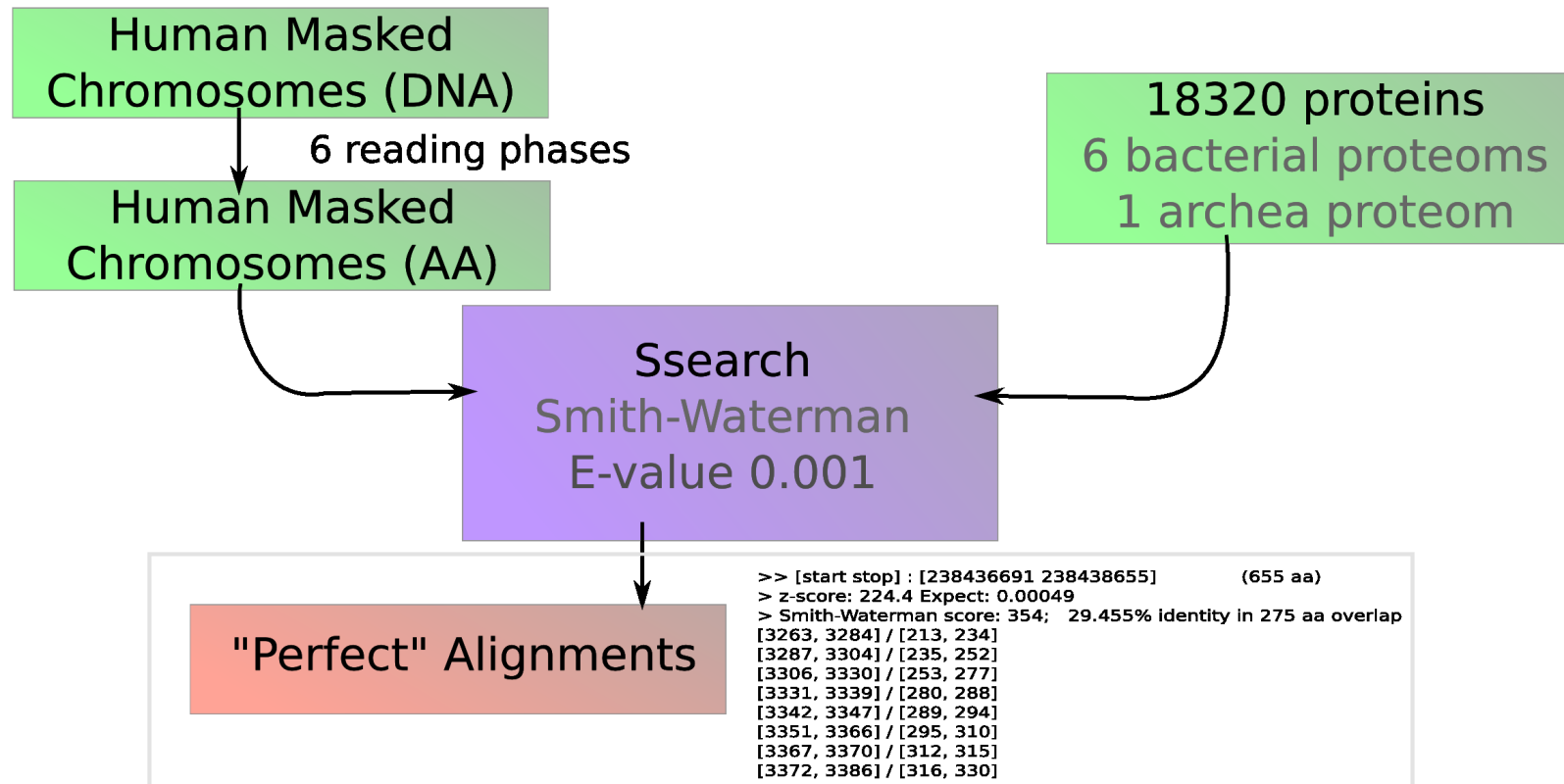
- Test quickly sensitivity (and possibly specificity)
- A few minutes
- Specificity is longer,  $\Rightarrow$  statistical computation is preferred

## Seed tester



- Necessity to pre-compute *perfect* alignments

# Seed tester



## Today

- Chromosomes 1, 2 and 19 treated
- 3273 Alignments found

Lipman, D.J. & Pearson, W.R. Rapid and Sensitive Protein Similarity Searches *Science*, **1985**, 227, 1435-1441

# First results

## Selection of a Subset seeds :

- Find **98%** of alignments while BLAST finds **96%** of alignment.
- The 2% difference is biologically relevant

# Overview

Motivations

Subset seeds

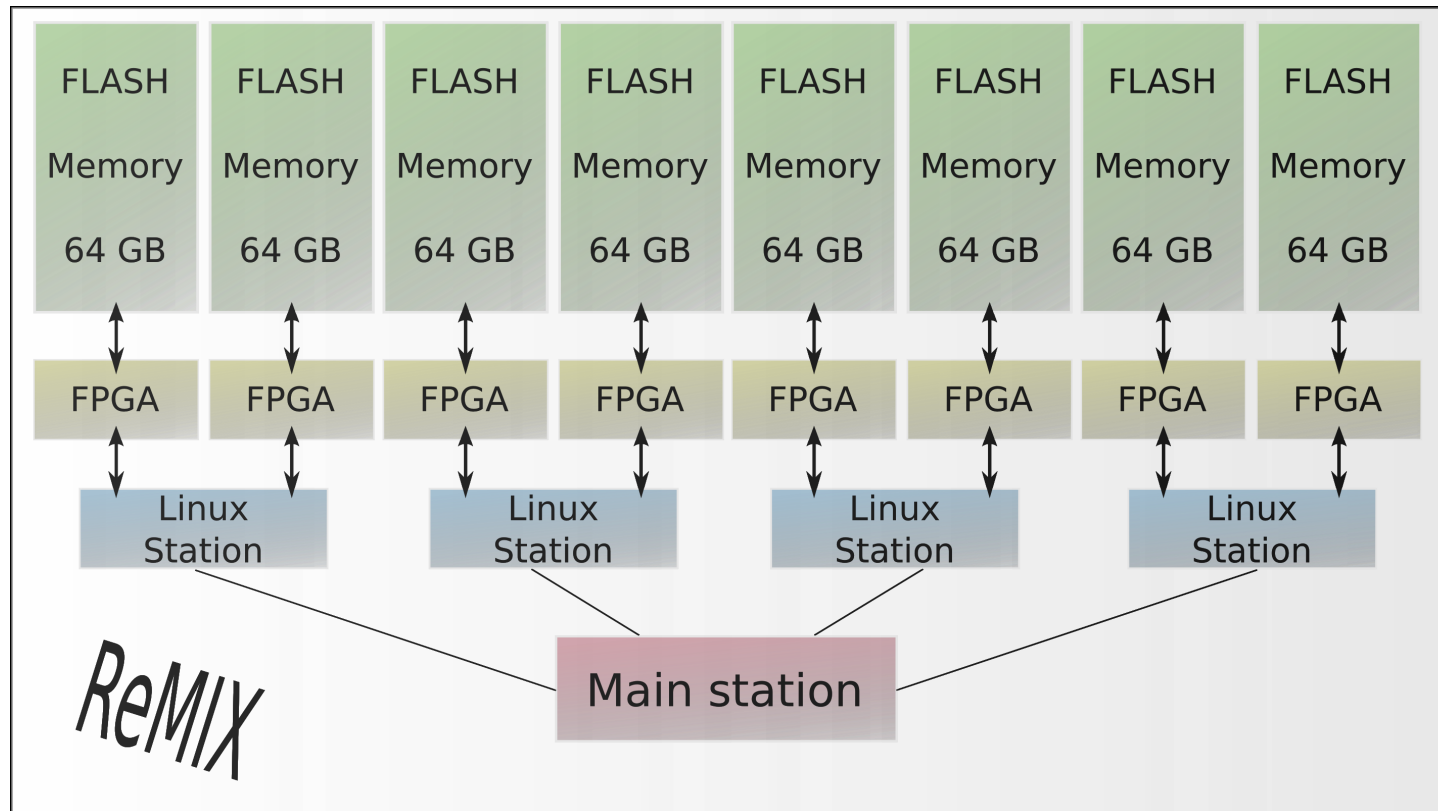
Where are we now ?

**Specialized architecture**

Conclusion

# ReMIX, Overall presentation

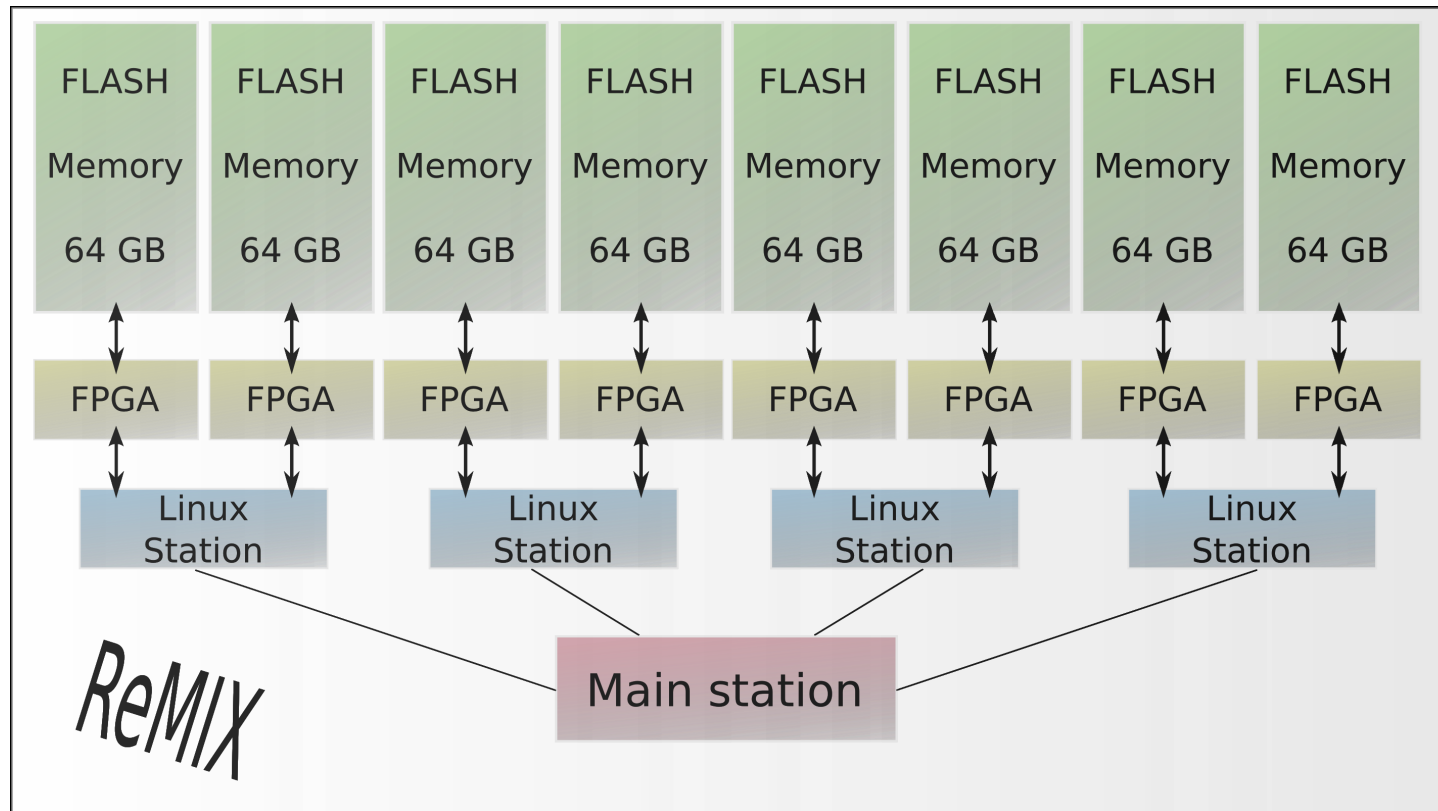
Lavenier, D. ; Liu, X. & Georges, G. Seed-Based Genomic Sequence Comparison Using a FPGA/FLASH Accelerator *IEEE FPT* 2006.



- 512 GB FLASH memory (indexation, step 1)
- FPGA : Compute approximatively  $8 \times 160$  ungaped alignments simultaneously in 50 clock cycles (step 2)
- A clock cycle  $\Rightarrow 25.10^{-9}$  seconds

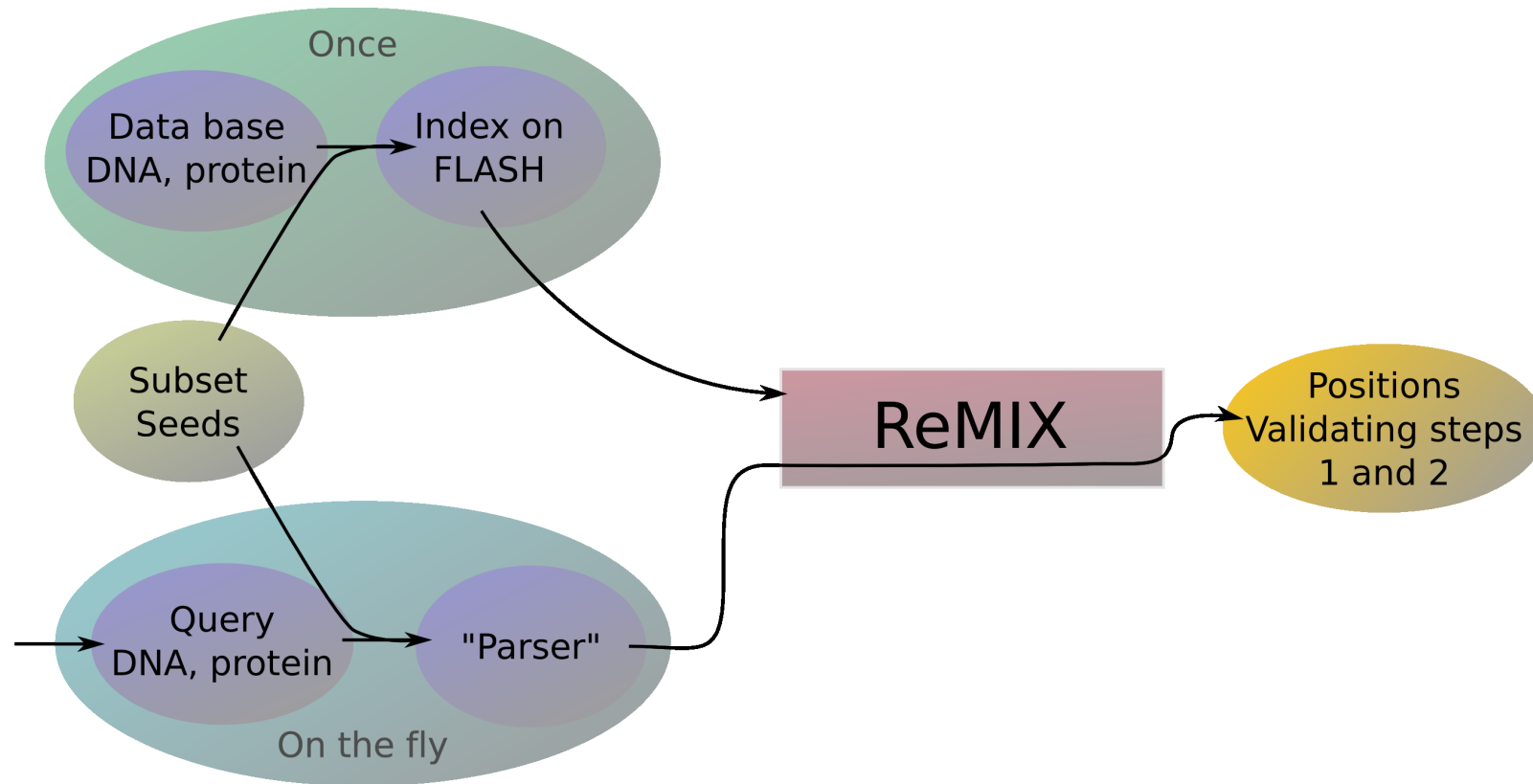
# ReMIX, Overall presentation

Lavenier, D. ; Liu, X. & Georges, G. Seed-Based Genomic Sequence Comparison Using a FPGA/FLASH Accelerator *IEEE FPT* 2006.



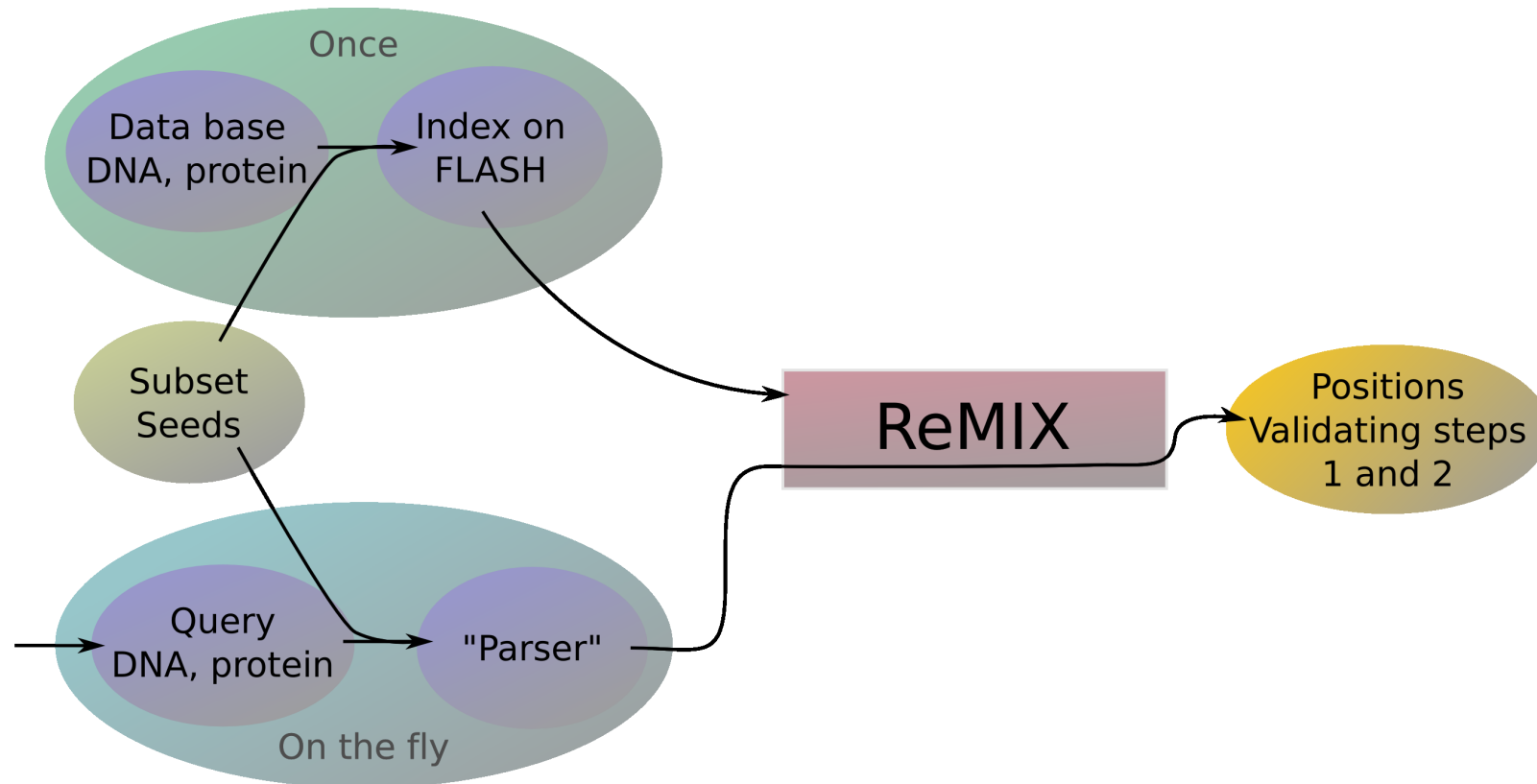
- 512 GB FLASH memory (indexation, step 1)
- FPGA : Compute approximatively  $8 \times 160$  ungaped alignments simultaneously in 50 clock cycles (step 2)
- **1024 millions** of ungaped alignments (neighbor) per second

# ReMIX, Biological application



- Index computed once
- Query, parsed on the fly

# ReMIX, Biological application



- Index computed once
- Query, parsed on the fly
- **done** seed ### (prototype, speed up 75)
- **todo** all others...

# Overview

Motivations

Subset seeds

Where are we now ?

Specialized architecture

**Conclusion**

# Conclusion

## Goal

- BLAST-like programs :
  - Take larger amount of data
  - Increase speed
  - Increase sensitivity

# Conclusion

## Goal

- BLAST-like programs :
  - Take larger amount of data
  - Increase speed
  - Increase sensitivity

## Done

- Subset seeds
- Framework for subset seeds testing
- (In progress) Implementation on ReMIX

# Conclusion

## Goal

- BLAST-like programs :
  - Take larger amount of data
  - Increase speed
  - Increase sensitivity

## Done

- Subset seeds
- Framework for subset seeds testing
- (In progress) Implementation on ReMIX

## To be done

- Investigation on subset seeds (98 %, 99 %, [99.9 % ?])
- Feed the seed tester with new alignements
- Implementation, tests, distribution