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
<table>
<thead>
<tr>
<th>Motivations</th>
<th>Subset seeds</th>
<th>Where are we now?</th>
<th>Specialized architecture</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overview</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Motivations**

Subset seeds

Where are we now?

Specialized architecture

Conclusion
Motivations  Subset seeds  Where are we now?  Specialized architecture  Conclusion

Biological Data evolution

EMBL\textsuperscript{a} Data size evolution:

Data type evolution:

- Take the “Junk DNA” into account

\begin{itemize}
  \item \texttt{http://www.ebi.ac.uk/embl/} - Europe’s primary nucleotide sequence resource
\end{itemize}

Large augmentation of amount of data
Main goal of the study

Biological similarities detection

Improvements of BLAST$^1$-like programs:

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

\[\text{Altschul, S.; Gish, W.; Miller, W.; Myers, E. & Lipman, D. Basic Local Alignment Search Tool Journal of Molecular Biology, 1990, 215, 403-410}\]
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
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\textsuperscript{2} seeds

<table>
<thead>
<tr>
<th>ATCAGTGCAATGCTCAAGA</th>
<th>ATCAGTGCAATGCTCAAGA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>ATCAGCGCGATGCGCAAGA</td>
<td>ATCAGCGCGATGCGCAAGA</td>
</tr>
<tr>
<td># # # # # # # # # #</td>
<td># # # - - # - # #</td>
</tr>
</tbody>
</table>

Spaced$^2$ seeds

ATCAGTGC A ATGCTCA AGA

ATCAGCGCGATGC GCA AGA

ATCAGT GCA ATGCTCA AGA

ATCAGCGCGATGC GCA AGA

---

Spaced$^2$ seeds

<table>
<thead>
<tr>
<th>ATCA GTGCA ATGCTCA AGA</th>
<th>ATCA GTGCA ATGCTCA AGA</th>
</tr>
</thead>
<tbody>
<tr>
<td># # # # # # # # #</td>
<td># # - - # - # #</td>
</tr>
<tr>
<td># # # # # # # #</td>
<td># # - - # - # #</td>
</tr>
<tr>
<td></td>
<td># # - - # - # #</td>
</tr>
</tbody>
</table>

Spaced\textsuperscript{2} seeds

ATCAGTGC AATGCTC AAGA

ATCAGCGCG ATGC GCC AAGA

---

**Spaced seeds**

ATCAGTGCAATGCTCAAGA

- - - - - -

ATCAGCGCGATGCGCAAGA

- - - - - -

- - - - - -

- - - - - -

**Spaced seeds have better sensitivity**
Multiple$^3$ spaced seeds

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

---

Multiple^3 spaced seeds

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

Use a set of spaced seeds
{###- -#-###, (weight 7)
#- -###-#-##, (weight 7)
#- - ###- -##-#, (weight 7)
...}

Multiple\textsuperscript{3} spaced seeds

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

Use a set of spaced seeds
\{
  ###- -#-###, (weight 7)
  #- -####-#-##, (weight 7)
  # - ###- -##-#, (weight 7)
  \ldots
\}\n
Advantages and drawback

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

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

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

**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 |
| 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

**All amino acids are not equivalents**

- Create groups of characters

---

- CFYWMLIVGPATSNHQEDRK

@1 C, STPAG, NDEQ, HRK, MILV, FYW

@2 CFYWMLIV, GPATSNHQEDRK


12/21
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, ⇒ statistical computation is preferred
• Necessity to pre-compute *perfect* alignments
Today

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

### First results

**Selection of a Subset seeds:**

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

<table>
<thead>
<tr>
<th>Motivations</th>
<th>Subset seeds</th>
<th>Where are we now?</th>
<th>Specialized architecture</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

Motivations

Subset seeds

Where are we now?

**Specialized architecture**

Conclusion
Motivations
Subset seeds
Where are we now?
Specialized architecture
Conclusion

ReMIX, Overall presentation

- 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


- 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...
<table>
<thead>
<tr>
<th>Motivations</th>
<th>Subset seeds</th>
<th>Where are we now?</th>
<th>Specialized architecture</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subset seeds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Where are we now?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialized architecture</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conclusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

21/21
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