A coverage criterion for spaced seeds and its applications to SVM string-kernels and k-mer distances

Laurent Noé, Donald E. K. Martin

LIFL (UMR 8022 Lille 1/CNRS) - Inria Lille, Villeneuve d'Ascq, France
Department of Statistics, North Carolina State University, Raleigh, NC, USA

AlgoB group seminar
December 3, 2014 - Marne-la-Vallée

SeqBio 2014
November 4&5, 2014 - Montpellier
Outline

1. Introduction to spaced seeds . . .
2. Spaced seed *coverage*
   - Definition
   - Associated automaton
   - Possible use (as a seed “quality” measure).
3. Experimental results
   - SVM classifiers
   - Alignment-free distances
4. Conclusion
Spaced Seeds
[Burkhardt and Kärkkäinen, 2001, Ma et al., 2002]...

Definition

A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:
- $1$: accepts only match symbol $|$
- $*$: accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of 1).

Example

$$\pi = 111*1*11$$

ATCAGTGCAGAATGCAGCAAGA
A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:
- $1$: accepts only match symbol $|$,
- $*$: accepts all alignment symbols ($joker$).

$s$: span (length), $w$: weight (number of 1).

Example

$\pi = 111*1*11$

111*1*11
ATCAGTGCAGAATGCGCAAGA
\[ \vdots \]
ATCAGCGCAAAATGCAGAAGA
A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:

- $1$: accepts only match symbol $|$,
- $*$: accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of 1).

Example

$\pi = 111*1*1$
Definition

A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1,*\}$:
- $1$: accepts only match symbol $\|$,
- $*$: accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of $1$).

Example

$\pi = 111*1*11$

```
ATCAGTGCGAATGCGCAAGA
ATCAGCGCAATGCTCAAGA
```

Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Definition

A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:

- $1$: accepts only match symbol $|$,
- $*$: accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of 1).

Example

$\pi = 111*1*11$

\[
\begin{align*}
111*1*11 & \\
ATCAGT GCGAATGCG & GCAAGA \\
ATCAGCGCAATGCT & TCAAGA
\end{align*}
\]
Definition

A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:

- $1$: accepts only match symbol $\|$,
- $*$: accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of $1$).

Example

$\pi = 111*1*11$

```
ATCAGTGCGAATGCGCAAGA
|||:||:||:||:||:||:||
ATCAGCGCAATGCCTCAAGA
```
A spaced seed $\pi$ is defined as a binary word over the alphabet $\{1, *\}$:
- $1$ : accepts only match symbol $|$
- $*$ : accepts all alignment symbols (joker).

$s$: span (length), $w$: weight (number of 1).

Example

$\pi = 111*1*11$

ATCAG\textcolor{green}{T}G\textcolor{green}{C}G\textcolor{green}{A}AT\textcolor{green}{G}C\textcolor{green}{G}CA\textcolor{green}{A}GA$

$\textcolor{green}{111*1*11}$

ATCAGCGC\textcolor{orange}{A}AATGCT\textcolor{orange}{G}CA\textcolor{orange}{A}GA
Example

ATCAGTGCAAATGCTCAAGA

ATCAGTGCAAATGCTCAAGA
Example

ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA

111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111

ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA
ATCAGTGCAAATGCTCAAGA

111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Example

Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Example

Laurent Noé, Donald E. K. Martin
A coverage criterion for spaced seeds and its applications
Example

Laurent Noé, Donald E. K. Martin  

A coverage criterion for spaced seeds and its applications
Example

Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Example

ATCAGTGCAATTGCAGCAAGA
| | | | | : | | | | | | | | |
ATCAGCGCAGTTGCAATGCAAGA

111111
111111
111111
111111
111111
111111
111111
111111
111111
111111
111111

ATCAGTGCAATTGCAGCAAGA
| | | | | : | | | | | | | | |
ATCAGCGCAGTTGCAATGCAAGA

111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11
111*1*11

Laurent Noé, Donald E. K. Martin
A coverage criterion for spaced seeds and its applications
Example

Laurent Noé, Donald E. K. Martin
A coverage criterion for spaced seeds and its applications
Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Laurent Noé, Donald E. K. Martin

A coverage criterion for spaced seeds and its applications
Recent work related to spaced seeds

1. Alignment-free distances
   [Leimeister et al., 2014, Horwege et al., 2014, Boden et al., 2013]

2. SVM classification
   [Onodera and Shibuya, 2013, Ghandi et al., 2014]

3. Read clustering
   [Bao et al., 2011, Chong et al., 2012, Hauser et al., 2013]

4. Metagenonomic classification, . . .
“New Uses for Old Things”

little boy

frying pan

1

1

1

1

1

http://arch5541.wordpress.com/2012/11/16/and-then-there-was-teflon/
"New Uses for Old Things"

little boy

frying pan

1

ATCAGTGCAGATTGGGAAGAG

111*1*11

http://arch5541.wordpress.com/2012/11/16/and-then-there-was-teflon/
“New Uses for Old Things”

Little boy

Frying pan

ATCAGTGCGAATGCGCAAGA
ATCAGCGCAAATGCTCAAGA

[Image of a missile with Teflon]
“New Uses for Old Things”

little boy

frying pan

ATCAGTGCAGATGCACCAAGA

ATCAGCGCAGATGCTCAAGA

111*1*11

111*1*11

111*1*11

111*1*11

111*1*11

111*1*11

111*1*11

1

http://arch5541.wordpress.com/2012/11/16/and-then-there-was-teflon/
"New Uses for Old Things"

http://arch5541.wordpress.com/2012/11/16/and-then-there-was-teflon/
Coverage measure for a seed

**Definition**

Number of match symbols **covered** by *at least one* 1 symbol from any seed hit [Benson and Mak, 2008, Martin, 2013]
Coverage measure for a seed

**Definition**

Number of match symbols **covered** by *at least one* 1 symbol from any seed hit [Benson and Mak, 2008, Martin, 2013]

**Example**

ATCAGTGCGAATGCACCAAGA


ATCAGCAGCAGAATGCCTCAAGA

111*1*11

111*1*11

111*1*11

111*1*11
Coverage measure for a seed

**Definition**

Number of match symbols **covered** by *at least one* 1 symbol from any seed hit [Benson and Mak, 2008, Martin, 2013]

**Example**

```
ATCAGTGCGAATGC\textcolor{green}{G}\textcolor{orange}{C}AAGA
\textcolor{green}{|} | | | | | :| | | | | | | | \textcolor{orange}{|} | | | | | | | | |
ATCAG\textcolor{green}{C}G\textcolor{orange}{C}A\textcolor{green}{A}AATGCT\textcolor{orange}{C}AAGA
111*1*11
111*1*11
111*1*11
111*1*11
```

Coverage is of 15
Coverage measure for a seed

Definition

Number of match symbols **covered** by *at least one* 1 symbol from any seed hit [Benson and Mak, 2008, Martin, 2013]

Example

```
ATCAGTGCAGATGCGCAAGA
111*1*11
ATCAGCGCAATGCTCAAGA
111*1*11
```

Coverage is of 15
Coverage measure for a seed

Definition

Number of match symbols covered by at least one 1 symbol from any seed hit [Benson and Mak, 2008, Martin, 2013]

Example

ATCA\textcolor{red}{G}\textcolor{red}{T}\textcolor{red}{G}\textcolor{red}{C}\textcolor{red}{G}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{T}\textcolor{red}{G}\textcolor{red}{C}\textcolor{red}{G}\textcolor{red}{C}\textcolor{red}{A}\textcolor{red}{A}\textcolor{red}{G}\textcolor{red}{A}

\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
\textcolor{red}{G}&\textcolor{red}{G}&\textcolor{red}{C}&\textcolor{red}{A}&\textcolor{red}{A}&\textcolor{red}{T}&\textcolor{red}{G}&\textcolor{red}{C}&\textcolor{red}{G} \\
\hline
\end{tabular}

Coverage is of 15
Coverage measure for a seed

alignment : \( x = 101111001011111 \)
Coverage measure for a seed

alignment: \( x = 101111001011111 \)

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>seed: ( \pi = 11*1 )</td>
</tr>
<tr>
<td>( x = 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 )</td>
</tr>
</tbody>
</table>
Coverage measure for a seed

alignment : $x = 101111001011111$

Example

seed : $\pi = 11*1$

$\pi_{occ1}$

\[
\begin{array}{ccccccccc}
& & 1 & 1 & * & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \\
1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1
\end{array}
\]
Coverage measure for a seed

alignment : \( x = 101111001011111 \)

**Example**

seed : \( \pi = 11*1 \)

\[
\begin{array}{ccccccc}
\pi_{occ_1} & 1 & 1 & * & 1 \\
\pi_{occ_2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

\[
\begin{array}{cccccccccccccc}
x = & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
& \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]
Coverage measure for a seed

alignment: \( x = 101111001011111 \)

Example

<table>
<thead>
<tr>
<th>( \pi_{occ_1} )</th>
<th>( \pi_{occ_2} )</th>
<th>( \pi_{occ_3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 * 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>: : : :</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ x = \begin{array}{ccccccccccc}
1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array} \]

seed: \( \pi = 11*1 \)
Coverage measure for a seed / a set of seeds

alignment : \( x = 101111001011111 \)

Example

seed : \( \pi = 11*1 \)

\[
\begin{array}{cccc}
\pi_{occ_1} & 1 & 1 & * & 1 \\
\pi_{occ_2} & \cdot & \cdot & \cdot & \cdot \\
\pi_{occ_3} & \cdot & \cdot & \cdot & 1 & 1 & * & 1 \\
x = 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
\]

set of seeds : \( \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \)
Coverage measure for a seed / a set of seeds

alignment : \( x = 101111001011111 \)

Example

seed : \( \pi = 11*1 \)

\[
\begin{array}{cccccccccccc}
\pi_{occ_1} & 1 & 1 & * & 1 \\
\pi_{occ_2} & : & : & : & : \\
\pi_{occ_3} & : & : & : & : \\
x = & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

set of seeds : \( \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \)

\[
\begin{array}{cccccccccccc}
x = & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]
Coverage measure for a seed / a set of seeds

alignment : \( x = 101111001011111 \)

Example

seed : \( \pi = 11*1 \)

\[
\begin{array}{cccccc}
\pi_{oc1} & 1 & 1 & * & 1 \\
\pi_{oc2} & * & * & * & * \\
\pi_{oc3} & 1 & 1 & * & 1 \\
x & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

set of seeds : \( \{ \pi_1, \pi_2 \} = \{ 11*1, 1*1*1 \} \)

\[
\begin{array}{cccccc}
\pi_2 \quad oc_{c1} & 1 & * & 1 & * & 1 \\
\pi_1 \quad oc_{c2} & 1 & 1 & * & 1 \\
\pi_2 \quad oc_{c3} & 1 & 1 & * & 1 \\
\pi_1 \quad oc_{c4} & 1 & 1 & * & 1 \\
\pi_2 \quad oc_{c5} & 1 & 1 & * & 1 \\
\pi_1 \quad oc_{c6} & 1 & 1 & * & 1 \\
x & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1
\end{array}
\]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

$$\{\pi_1, \pi_2\} = \{11*1, 1*1*1\}$$
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{1\#1, 1\#1\#1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage measure for a seed / a set of seeds

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]

\[ t_i(q) = \bigoplus w_i(a^\circ) \otimes t(q^\circ) \]

\( \circ \in "going \ to"(q) \)
That’s how coverage can be measured, estimated, computed on several models...
That’s how coverage can be measured, estimated, computed on several models…

But, …is coverage useful?
Experimental results

1. SVM classifiers
2. Alignment-free distances
Are spaced seeds better with *string kernels* classifiers?
SVM classifiers

- Are spaced seeds better with *string kernels* classifiers?
  Yes: see [Onodera and Shibuya, 2013, Ghandi et al., 2014]
SVM classifiers

- Are spaced seeds better with *string kernels* classifiers? Yes: see [Onodera and Shibuya, 2013, Ghandi et al., 2014]
- Which spaced seed patterns are better? Does coverage help here?
RFAM 11.0 database (50% training, 50% testing)
SVM classifiers

1. RFAM 11.0 database (50% training, 50% testing)
2. Single/double seeds of weight $w = 3 \ldots 4$, span up to $w + 4$
SVM classifiers

1. RFAM 11.0 database (50% training, 50% testing)
2. Single/double seeds of weight \( w = 3 \ldots 4 \), span up to \( w + 4 \)
3. For each seed,
   - Learn (classical string kernel with linear classifier).
   - Measure the SVM zero/one error.
SVM classifiers

1. RFAM 11.0 database (50% training, 50% testing)
2. Single/double seeds of weight $w = 3 \ldots 4$, span up to $w + 4$
3. For each seed,
   - Learn (classical string kernel with linear classifier).
   - Measure the SVM zero/one error.
4. Compute the correlation coefficient between the SVM zero/one error and:
   - the single hit criterion (at least one seed hit)
   - the multi hit criterion (at least $n$ seed hits)
   - the coverage criterion (at least $n$ seed coverage)
3 SVM zero/one error

weight 3 single seed

Good

Bad

svm zero/one error
4 SVM zero/one error vs multi hit sensitivity for $n = 5$
SVM zero/one error vs coverage sensitivity for $n = 5$
SVM zero/one error vs single hit

single hit sensitivity vs SVM zero/one error

weight 3 single seed

111 1*111 1*1 11**1 1***1 1****1
SVM zero/one error vs sensitivity (3 criteria)
SVM zero/one error vs sensitivity (3 criteria)
Are spaced seeds better in estimating the “true” alignment distance?
Alignment-free distances

- Are spaced seeds better in estimating the “true” alignment distance? Yes: see [Leimeister et al., 2014, Horwege et al., 2014, Boden et al., 2013] …
Are spaced seeds better in estimating the “true” alignment distance? Yes: see [Leimeister et al., 2014, Horwege et al., 2014, Boden et al., 2013] . . .

Which spaced seed patterns are better? Does coverage help here?
Alignment-free distances

1. Alignment length (e.g. $l = 16, 32, 64$)
Alignment-free distances

1. Alignment length (e.g. $l = 16, 32, 64$)
2. Single/double seeds of weight $w = 2 \ldots 9$, span up to $w + 4$
Alignment-free distances

1. Alignment length (e.g. \( l = 16, 32, 64 \))
2. Single/double seeds of weight \( w = 2 \ldots 9 \), span up to \( w + 4 \)
3. For each seed,
Alignment-free distances

1. Alignment length (e.g. $l = 16, 32, 64$)
2. Single/double seeds of weight $w = 2 \ldots 9$, span up to $w + 4$
3. For each seed,
   - Generate **any possible alignment** of length $l$ and measure its percentage of identity.
Alignment-free distances

1. Alignment length (e.g. \( l = 16, 32, 64 \))

2. Single/double seeds of weight \( w = 2 \ldots 9 \), span up to \( w + 4 \)

3. For each seed,
   - Generate **any possible alignment** of length \( l \) and measure its percentage of identity.
   - Compute the *correlation coefficient* between the true percentage of identity of any alignment and
     - the *multi-hit* value of the seed (next plot: x-axis)
     - the *coverage* value of the seed (next plot: y-axis)

on this alignment.
True distance Correlation with MultiHit (x) vs Coverage (y) distance

- double seed (id ≥ 0.0%)
- single seed (id ≥ 0.0%)

The graph shows the correlation between true distance and multi-hit distance divided by true distance, with points scattered across the plot. There are two categories indicated by different colors: red for double seeds and black for single seeds. The diagonal line represents x = y, indicating perfect correlation.
Alignment-free distances

- Fixed alignment length and Variable Minimal % id: http://www.youtube.com/watch?v=YfQcF_GJ1jM
- Variable alignment length and Variable Minimal % id: http://www.youtube.com/watch?v=LDenQv6H1EM
Coverage /vs/ Percentage : correlation computation

\( \{ \pi_1, \pi_2 \} = \{ 11*1, 1*1*1 \} \)
Coverage /vs/ Percentage: correlation computation

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage /vs/ Percentage: correlation computation

\[ \{ \pi_1, \pi_2 \} = \{11*1, 1*1*1\} \]
Coverage /vs/ Percentage : correlation computation

\[ \{\pi_1, \pi_2\} = \{11*1, 1*1*1\} \]
Coverage \(\text{vs}\) Percentage: correlation computation

\[
\{\pi_1, \pi_2\} = \{11*1, 1*1*1\}
\]
Coverage /vs/ Percentage: correlation computation

$$\{\pi_1, \pi_2\} = \{11*1, 1*1*1\}$$
A coverage criterion for spaced seeds

its applications to SVM string-kernels and k-mer distances
Perspectives

- Automaton size / building ways / generating function

- Guessing most likely matches/mismatches distribution?

- Phylogenetic studies?


kClust: fast and sensitive clustering of large protein sequence databases.
*BMC Bioinformatics*, 14(248).

Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches.

Fast alignment-free sequence comparison using spaced-word frequencies.
PatternHunter: Faster and more sensitive homology search.

Coverage of spaced seeds as a measure of clumping.

Faster exact probabilities for statistics of overlapping pattern occurrences.

The gapped spectrum kernel for support vector machines.