Pairwise Sequence Alignments

Patrick Aboyoun
Gentleman Lab
Fred Hutchinson Cancer Research Center
Seattle, WA

November 30, 2016

Contents
1 Introduction 2
2 Pairwise Sequence Alignment Problems 2
3 Main Pairwise Sequence Alignment Function 3
  3.1 Exercise 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
4 Pairwise Sequence Alignment Classes 5
  4.1 Exercise 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
5 Pairwise Sequence Alignment Helper Functions 6
  5.1 Exercise 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
6 Edit Distances 11
  6.1 Exercise 4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
7 Application: Using Evolutionary Models in Protein Alignments 12
  7.1 Exercise 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
8 Application: Removing Adapters from Sequence Reads 13
  8.1 Exercise 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
9 Application: Quality Assurance in Sequencing Experiments 17
  9.1 Exercise 7 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
10 Computation Profiling 20
  10.1 Exercise 8 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
11 Computing alignment consensus matrices 23
12 Exercise Answers 24
  12.1 Exercise 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
  12.2 Exercise 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
  12.3 Exercise 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
  12.4 Exercise 4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
  12.5 Exercise 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
1 Introduction

In this document we illustrate how to perform pairwise sequence alignments using the Biostrings package through the use of the pairwiseAlignment function. This function aligns a set of pattern strings to a subject string in a global, local, or overlap (ends-free) fashion with or without affine gaps using either a fixed or quality-based substitution scoring scheme. This function’s computation time is proportional to the product of the two string lengths being aligned.

2 Pairwise Sequence Alignment Problems

The (Needleman-Wunsch) global, the (Smith-Waterman) local, and (ends-free) overlap pairwise sequence alignment problems are described as follows. Let string $S_i$ have $n_i$ characters $c_{(i,j)}$ with $j \in \{1, \ldots, n_i\}$. A pairwise sequence alignment is a mapping of strings $S_1$ and $S_2$ to gapped substrings $S'_1$ and $S'_2$ that are defined by

$$S'_1 = g(1,a_1)c_{(1,a_1)} \cdots g(1,b_1)c_{(1,b_1)}g(1,b_1+1)$$
$$S'_2 = g(2,a_2)c_{(2,a_2)} \cdots g(2,b_2)c_{(2,b_2)}g(2,b_2+1)$$

where

$a_i, b_i \in \{1, \ldots, n_i\}$ with $a_i \leq b_i$

$g_{(i,j)} = 0$ or more gaps at the specified position $j$ for aligned string $i$

$\text{length}(S'_1) = \text{length}(S'_2)$

Each of these pairwise sequence alignment problems is solved by maximizing the alignment score. An alignment score is determined by the type of pairwise sequence alignment (global, local, overlap), which sets the $[a_i, b_i]$ ranges for the substrings; the substitution scoring scheme, which sets the distance between aligned characters; and the gap penalties, which is divided into opening and extension components. The optimal pairwise sequence alignment is the pairwise sequence alignment with the largest score for the specified alignment type, substitution scoring scheme, and gap penalties. The pairwise sequence alignment types, substitution scoring schemes, and gap penalties influence alignment scores in the following manner:

Pairwise Sequence Alignment Types: The type of pairwise sequence alignment determines the substring ranges to apply the substitution scoring and gap penalty schemes. For the three primary (global, local, overlap) and two derivative (subject overlap, pattern overlap) pairwise sequence alignment types, the resulting substring ranges are as follows:

Global - $[a_1, b_1] = [1, n_1]$ and $[a_2, b_2] = [1, n_2]$

Local - $[a_1, b_1]$ and $[a_2, b_2]$

Overlap - $\{[a_1, b_1] = [a_1, n_1], [a_2, b_2] = [1, b_2]\}$ or $\{[a_1, b_1] = [1, b_1], [a_2, b_2] = [a_2, n_2]\}$

Subject Overlap - $[a_1, b_1] = [1, n_1]$ and $[a_2, b_2]$

Pattern Overlap - $[a_1, b_1]$ and $[a_2, b_2] = [1, n_2]$
Substitution Scoring Schemes: The substitution scoring scheme sets the values for the aligned character pairings within the substring ranges determined by the type of pairwise sequence alignment. This scoring scheme can be fixed for character pairings or quality-dependent for character pairings. (Characters that align with a gap are penalized according to the “Gap Penalty” framework.)

Fixed substitution scoring - Fixed substitution scoring schemes associate each aligned character pairing with a value. These schemes are very common and include awarding one value for a match and another for a mismatch, Point Accepted Mutation (PAM) matrices, and Block Substitution Matrix (BLOSUM) matrices.

Quality-based substitution scoring - Quality-based substitution scoring schemes derive the value for the aligned character pairing based on the probabilities of character recording errors. Let $\epsilon_i$ be the probability of a character recording error. Assuming independence within and between recordings and a uniform background frequency of the different characters, the combined error probability of a mismatch when the underlying characters do match is $\epsilon_c = \epsilon_1 + \epsilon_2 - \left(\frac{n}{n-1}\right) \epsilon_1 \epsilon_2$, where $n$ is the number of characters in the underlying alphabet (e.g. in DNA and RNA, $n = 4$). Using $\epsilon_c$, the substitution score is given by $b \log_2 \left( \frac{\gamma(x,y) \cdot (1 - \epsilon_c) \cdot n + (1 - \gamma(x,y)) \cdot \epsilon_c \cdot \left(\frac{n}{n-1}\right)}{\gamma(x,y)} \right)$, where $b$ is the bit-scaling for the scoring and $\gamma(x,y)$ represents the same underlying letters (e.g. using IUPAC, $\gamma(A,A) = 1$ and $\gamma(A,N) = 1/4$).

Gap Penalties: Gap penalties are the values associated with the gaps within the substring ranges determined by the type of pairwise sequence alignment. These penalties are divided into gap opening and gap extension components, where the gap opening penalty is the cost for adding a new gap and the gap extension penalty is the incremental cost incurred along the length of the gap. A constant gap penalty occurs when there is a cost associated with opening a gap, but no cost for the length of a gap (i.e. gap extension is zero). A linear gap penalty occurs when there is no cost associated for opening a gap (i.e. gap opening is zero), but there is a cost for the length of the gap. An affine gap penalty occurs when both the gap opening and gap extension have a non-zero associated cost.

3 Main Pairwise Sequence Alignment Function

The pairwiseAlignment function solves the pairwise sequence alignment problems mentioned above. It aligns one or more strings specified in the pattern argument with a single string specified in the subject argument.

> library(Biostrings)
> pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede")

Global PairwiseAlignmentsSingleSubject (1 of 2)
| pattern: | [1] succ--eed |
| subject: | [1] supersede |
| score:   | -33.99738     |

The type of pairwise sequence alignment is set by specifying the type argument to be one of "global", "local", "overlap", "global-local", and "local-global".

> pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+ type = "local")

Local PairwiseAlignmentsSingleSubject (1 of 2)
| pattern: | [1] su       |
| subject: | [1] su       |
| score:   | 5.578203     |
The gap penalties are regulated by the `gapOpening` and `gapExtension` arguments.

```
> pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+                      gapOpening = 0, gapExtension = 1)

Global PairwiseAlignmentsSingleSubject (1 of 2)
pattern: [1] su-cce--ed
subject: [1] sup--ersed
score: 7.945507
```

The substitution scoring scheme is set using three arguments, two of which are quality-based related (`patternQuality`, `subjectQuality`) and one is fixed substitution related (`substitutionMatrix`). When the substitution scores are fixed by character pairing, the `substitutionMatrix` argument takes a matrix with the appropriate alphabets as dimension names. The `nucleotideSubstitutionMatrix` function translates simple match and mismatch scores to the full spectrum of IUPAC nucleotide codes.

```
> submat <-
+    matrix(-1, nrow = 26, ncol = 26, dimnames = list(letters, letters))
> diag(submat) <- 0
> pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+                     substitutionMatrix = submat,
+                     gapOpening = 0, gapExtension = 1)

Global PairwiseAlignmentsSingleSubject (1 of 2)
pattern: [1] su-cce--ed
subject: [1] sup--ersed
score: -5
```

When the substitution scores are quality-based, the `patternQuality` and `subjectQuality` arguments represent the equivalent of \([x - 99]\) numeric quality values for the respective strings, and the optional `fuzzyMatrix` argument represents how closely two characters match on a \([0, 1]\) scale. The `patternQuality` and `subjectQuality` arguments accept quality measures in either a `PhredQuality`, `SolexaQuality`, or `IlluminaQuality` scaling. For `PhredQuality` and `IlluminaQuality` measures \(Q \in [0, 99]\), the probability of an error in the base read is given by \(10^{-Q/10}\) and for `SolexaQuality` measures \(Q \in [-5, 99]\), they are given by \(1 - 1/(1+10^{-Q/10})\). The `qualitySubstitutionMatrices` function maps the `patternQuality` and `subjectQuality` scores to match and mismatch penalties. These three arguments will be demonstrated in later sections.

The final argument, `scoreOnly`, to the `pairwiseAlignment` function accepts a logical value to specify whether or not to return just the pairwise sequence alignment score. If `scoreOnly` is `FALSE`, the pairwise alignment with the maximum alignment score is returned. If more than one pairwise alignment has the maximum alignment score exists, the first alignment along the subject is returned. If there are multiple pairwise alignments with the maximum alignment score at the chosen subject location, then at each location along the alignment mismatches are given preference to insertions/deletions. For example, `pattern`: [1] ATTA; `subject`: [1] AT-A is chosen above `pattern`: [1] ATTA; `subject`: [1] A-TA if they both have the maximum alignment score.

```
> submat <-
+    matrix(-1, nrow = 26, ncol = 26, dimnames = list(letters, letters))
> diag(submat) <- 0
> pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+                     substitutionMatrix = submat,
+                     gapOpening = 0, gapExtension = 1, scoreOnly = TRUE)

[1] -5 -5
```
3.1 Exercise 1

1. Using `pairwiseAlignment`, fit the global, local, and overlap pairwise sequence alignment of the strings "syzygy" and "zyzzyx" using the default settings.

2. Do any of the alignments change if the `gapExtension` argument is set to `-Inf`?

[Answers provided in section 12.1]

4 Pairwise Sequence Alignment Classes

Following the design principles of Bioconductor and R, the pairwise sequence alignment functionality in the Biostrings package keeps the end-user close to their data through the use of five specialty classes: `PairwiseAlignments`, `PairwiseAlignmentsSingleSubject`, `PairwiseAlignmentsSingleSubjectSummary`, `AlignedXStringSet`, and `QualityAlignedXStringSet`. The `PairwiseAlignmentsSingleSubject` class inherits from the `PairwiseAlignments` class and they both hold the results of a fit from the `pairwiseAlignment` function, with the former class being used to represent all patterns aligning to a single subject and the latter being used to represent elementwise alignments between a set of patterns and a set of subjects.

```r
> psa1 <- pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede")
> class(psa1)
[1] "PairwiseAlignmentsSingleSubject"
attr(,"package")
[1] "Biostrings"

and the `pairwiseAlignmentSummary` function holds the results of a summarized pairwise sequence alignment.

> summary(psa1)

Global Single Subject Pairwise Alignments

Number of Alignments: 2

Scores:

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-34.00</td>
<td>-31.78</td>
<td>-29.56</td>
<td>-29.56</td>
<td>-27.34</td>
<td>-25.12</td>
</tr>
</tbody>
</table>

Number of matches:

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.00</td>
<td>3.25</td>
<td>3.50</td>
<td>3.50</td>
<td>3.75</td>
<td>4.00</td>
</tr>
</tbody>
</table>

Top 7 Mismatch Counts:

<table>
<thead>
<tr>
<th>SubjectPosition</th>
<th>Subject</th>
<th>Pattern</th>
<th>Count</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>p</td>
<td>c</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>e</td>
<td>c</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>e</td>
<td>r</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>r</td>
<td>e</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>s</td>
<td>c</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>d</td>
<td>e</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>e</td>
<td>d</td>
<td>1</td>
</tr>
</tbody>
</table>

> class(summary(psa1))
The `AlignedXStringSet` and `QualityAlignedXStringSet` classes hold the “gapped” \( S'_i \) substrings with the former class holding the results when the pairwise sequence alignment is performed with a fixed substitution scoring scheme and the latter class a quality-based scoring scheme.

```r
> class(pattern(psa1))
[1] "QualityAlignedXStringSet"
attr(,"package")
[1] "Biostrings"
```

```r
> submat <- +  matrix(-1, nrow = 26, ncol = 26, dimnames = list(letters, letters))
> diag(submat) <- 0
> psa2 <- +  pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+                              substitutionMatrix = submat,
+                              gapOpening = 0, gapExtension = 1)
> class(pattern(psa2))
[1] "AlignedXStringSet"
```

4.1 Exercise 2

1. What is the primary benefit of formal summary classes like `PairwiseAlignmentsSingleSubjectSummary` and `summary.lm` to end-users?

   [Answer provided in section 12.2.]

5 Pairwise Sequence Alignment Helper Functions

Tables 1 and 3 show functions that interact with objects of class `PairwiseAlignments`, `PairwiseAlignmentsSingleSubject`, and `AlignedXStringSet`. These functions should be used in preference to direct slot extraction from the alignment objects.

The `score`, `nedit`, `nmatch`, `nmismatch`, and `nchar` functions return numeric vectors containing information on the pairwise sequence alignment score, number of matches, number of mismatches, and number of aligned characters respectively.

```r
> submat <- +  matrix(-1, nrow = 26, ncol = 26, dimnames = list(letters, letters))
> diag(submat) <- 0
> psa2 <- +  pairwiseAlignment(pattern = c("succeed", "precede"), subject = "supersede",
+                              substitutionMatrix = submat,
+                              gapOpening = 0, gapExtension = 1)
> score(psa2)
[1] -5 -5```
<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[</td>
<td>Extracts the specified elements of the alignment object</td>
</tr>
<tr>
<td>alphabet</td>
<td>Extracts the allowable characters in the original strings</td>
</tr>
<tr>
<td>compareStrings</td>
<td>Creates character string mashups of the alignments</td>
</tr>
<tr>
<td>deletion</td>
<td>Extracts the locations of the gaps inserted into the pattern for the alignments</td>
</tr>
<tr>
<td>length</td>
<td>Extracts the number of patterns aligned</td>
</tr>
<tr>
<td>mismatchTable</td>
<td>Creates a table for the mismatching positions</td>
</tr>
<tr>
<td>nchar</td>
<td>Computes the length of “gapped” substrings</td>
</tr>
<tr>
<td>nedit</td>
<td>Computes the Levenshtein edit distance of the alignments</td>
</tr>
<tr>
<td>indel</td>
<td>Extracts the locations of the insertion &amp; deletion gaps in the alignments</td>
</tr>
<tr>
<td>insertion</td>
<td>Extracts the locations of the gaps inserted into the subject for the alignments</td>
</tr>
<tr>
<td>nindel</td>
<td>Computes the number of insertions &amp; deletions in the alignments</td>
</tr>
<tr>
<td>nmatch</td>
<td>Computes the number of matching characters in the alignments</td>
</tr>
<tr>
<td>nmismatch</td>
<td>Computes the number of mismatching characters in the alignments</td>
</tr>
<tr>
<td>pattern, subject</td>
<td>Extracts the aligned pattern/subject</td>
</tr>
<tr>
<td>pid</td>
<td>Computes the percent sequence identity</td>
</tr>
<tr>
<td>rep</td>
<td>Replicates the elements of the alignment object</td>
</tr>
<tr>
<td>score</td>
<td>Extracts the pairwise sequence alignment scores</td>
</tr>
<tr>
<td>type</td>
<td>Extracts the type of pairwise sequence alignment</td>
</tr>
</tbody>
</table>

Table 1: Functions for *PairwiseAlignments* and *PairwiseAlignmentsSingleSubject* objects.

```r
> nedit(psa2)
[1] 4 5
> nmatch(psa2)
[1] 4 4
> nmismatch(psa2)
[1] 3 3
> nchar(psa2)
[1] 8 9
> aligned(psa2)

A BStringSet instance of length 2

width seq
[1] 9 succe-ed-
[2] 9 pr-ec-ede

> as.character(psa2)
[1] "succe-ed-" "pr-ec-ede"

> as.matrix(psa2)

[1,] "s" "u" "c" "e" "-" "e" "d" "-"
[2,] "p" "r" "-" "e" "c" "e" "d" "e"
```

7
Function | Description
--- | ---
aligned | Creates an `XStringSet` containing either “filled-with-gaps” or degapped aligned strings
as.character | Creates a character vector version of `aligned`
as.matrix | Creates an “exploded” character matrix version of `aligned`
consensusMatrix | Computes a consensus matrix for the alignments
consensusString | Creates the string based on a 50% + 1 vote from the consensus matrix
coverage | Computes the alignment coverage along the subject
mismatchSummary | Summarizes the information of the `mismatchTable`
summary | Summarizes a pairwise sequence alignment
toString | Creates a concatenated string version of `aligned`
Views | Creates an `XStringViews` representing the aligned region along the subject

Table 2: Additional functions for `PairwiseAlignmentsSingleSubject` objects.

```r
> consensusMatrix(psa2)
-  0  0  1  0  0  2  0  0  1
c  0  0  1  1  1  0  0  0  0
d  0  0  0  0  0  0  2  0  0
e  0  0  0  1  1  0  2  0  1
p  1  0  0  0  0  0  0  0  0
r  0  1  0  0  0  0  0  0  0
s  1  0  0  0  0  0  0  0  0
u  0  1  0  0  0  0  0  0  0
```

The `summary`, `mismatchTable`, and `mismatchSummary` functions return various summaries of the pairwise sequence alignments.

```r
> summary(psa2)
Global Single Subject Pairwise Alignments
Number of Alignments: 2
Scores:
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  -5    -5    -5    -5    -5    -5
Number of matches:
  Min. 1st Qu. Median Mean 3rd Qu. Max.
   4     4     4     4     4     4
Top 6 Mismatch Counts:
  SubjectPosition Subject Pattern Count Probability
  1          1 s p  1  0.5
  2          2 u r  1  0.5
  3          3 p c  1  0.5
  4          4 e c  1  0.5
  5          5 r c  1  0.5
  6          5 r e  1  0.5
```

```r
> mismatchTable(psa2)

> mismatchTable(psa2)
```

8
The `pattern` and `subject` functions extract the aligned pattern and subject objects for further analysis. Most of the actions that can be performed on `PairwiseAlignments` objects can also be performed on `AlignedXStringSet` and `QualityAlignedXStringSet` objects as well as operations including `start`, `end`, and `width` that extracts the start, end, and width of the alignment ranges.

```r
> class(pattern(psa2))
[1] "AlignedXStringSet"
```

A BStringSet instance of length 2
width seq
<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ aligned, unaligned</td>
<td>Extracts the specified elements of the alignment object</td>
</tr>
<tr>
<td>alphabet</td>
<td>Extracts the allowable characters in the original strings</td>
</tr>
<tr>
<td>as.character, toString</td>
<td>Converts the alignments to character strings</td>
</tr>
<tr>
<td>coverage</td>
<td>Computes the alignment coverage</td>
</tr>
<tr>
<td>end</td>
<td>Extracts the ending index of the aligned range</td>
</tr>
<tr>
<td>indel</td>
<td>Extracts the insertion/deletion locations</td>
</tr>
<tr>
<td>length</td>
<td>Extracts the number of patterns aligned</td>
</tr>
<tr>
<td>mismatch</td>
<td>Extracts the position of the mismatches</td>
</tr>
<tr>
<td>mismatchSummary</td>
<td>Summarizes the information of the mismatchTable</td>
</tr>
<tr>
<td>mismatchTable</td>
<td>Creates a table for the matching positions</td>
</tr>
<tr>
<td>nchar</td>
<td>Computes the length of “gapped” substrings</td>
</tr>
<tr>
<td>nindel</td>
<td>Computes the number of insertions/deletions in the alignments</td>
</tr>
<tr>
<td>nmismatch</td>
<td>Computes the number of mismatching characters in the alignments</td>
</tr>
<tr>
<td>rep</td>
<td>Replicates the elements of the alignment object</td>
</tr>
<tr>
<td>start</td>
<td>Extracts the starting index of the aligned range</td>
</tr>
<tr>
<td>toString</td>
<td>Creates a concatenated string containing the alignments</td>
</tr>
<tr>
<td>width</td>
<td>Extracts the width of the aligned range</td>
</tr>
</tbody>
</table>

| Table 3: Functions for \textit{AlignedXString} and \textit{QualityAlignedXString} objects. |

\[1\] 8 succe-ed  
\[2\] 9 pr-ec-ede

\[
> \text{nindel(pattern(psa2))}
\]

\[
\begin{array}{ll}
\text{Length} & \text{WidthSum} \\
[1,] & 1 & 1 \\
[2,] & 2 & 2 \\
\end{array}
\]

\[
> \text{start(subject(psa2))}
\]

\[
[1] 1 1
\]

\[
> \text{end(subject(psa2))}
\]

\[
[1] 8 9
\]

5.1 Exercise 3

For the overlap pairwise sequence alignment of the strings ”syzygy” and ”zyzzyx” with the \texttt{pairwiseAlignment} default settings, perform the following operations:

1. Use \texttt{nmatch} and \texttt{nmismatch} to extract the number of matches and mismatches respectively.
2. Use the \texttt{compareStrings} function to get the symbolic representation of the alignment.
3. Use the \texttt{as.character} function to get the character string versions of the alignments.
4. Use the \texttt{pattern} function to extract the aligned pattern and apply the \texttt{mismatch} function to it to find the locations of the mismatches.
5. Use the \texttt{subject} function to extract the aligned subject and apply the \texttt{aligned} function to it to get the aligned strings.

[Answers provided in section 12.3]
6 Edit Distances

One of the earliest uses of pairwise sequence alignment is in the area of text analysis. In 1965 Vladimir
Levenshtein considered a metric, now called the Levenshtein edit distance, that measures the similarity
between two strings. This distance metric is equivalent to the negative of the score of a pairwise sequence
alignment with a match cost of 0, a mismatch cost of -1, a gap opening penalty of 0, and a gap extension
penalty of 1.

The stringDist uses the internals of the pairwiseAlignment function to calculate the Levenshtein edit
distance matrix for a set of strings.

There is also an implementation of approximate string matching using Levenshtein edit distance in the
agrep (approximate grep) function of the base R package. As the following example shows, it is possible to
replicate the agrep function using the pairwiseAlignment function. Since the agrep function is vectorized
in x rather than pattern, these arguments are flipped in the call to pairwiseAlignment.

```r
> agrepBioC <-
  + function(pattern, x, ignore.case = FALSE, value = FALSE, max.distance = 0.1)
  + {
  +   if (!is.character(pattern)) pattern <- as.character(pattern)
  +   if (!is.character(x)) x <- as.character(x)
  +   if (max.distance < 1)
  +     max.distance <- ceiling(max.distance / nchar(pattern))
  +   characters <- unique(unlist(strsplit(c(pattern, x), "", fixed = TRUE)))
  +   if (ignore.case)
  +     substitutionMatrix <-
  +       outer(tolower(characters), tolower(characters), function(x,y) -as.numeric(x!=y))
  +   else
  +     substitutionMatrix <-
  +       outer(characters, characters, function(x,y) -as.numeric(x!=y))
  +   dimnames(substitutionMatrix) <- list(characters, characters)
  +   distance <-
  +     - pairwiseAlignment(pattern = x, subject = pattern,
  +       substitutionMatrix = substitutionMatrix,
  +       type = "local-global",
  +       gapOpening = 0, gapExtension = 1,
  +       scoreOnly = TRUE)
  +   whichClose <- which(distance <= max.distance)
  +   if (value)
  +     whichClose <- x[whichClose]
  +   whichClose
  + }
> cbind(base = agrep("laysy", c("1 lazy", "1", "1 LAZY"), max = 2, value = TRUE),
  +   bioc = agrepBioC("laysy", c("1 lazy", "1", "1 LAZY"), max = 2, value = TRUE))

base   bioc
[1,] "1 lazy" "1 lazy"

> cbind(base = agrep("laysy", c("1 lazy", "1", "1 LAZY"), max = 2, ignore.case = TRUE),
  +   bioc = agrepBioC("laysy", c("1 lazy", "1", "1 LAZY"), max = 2, ignore.case = TRUE))

base   bioc
[1,] 1 1
[2,] 3 3
```
6.1 Exercise 4

1. Use the `pairwiseAlignment` function to find the Levenshtein edit distance between "syzygy" and "zyzzyx".

2. Use the `stringDist` function to find the Levenshtein edit distance for the vector c("zyzzyx", "syzygy", "succeed", "precede", "supersede").

[Answers provided in section 12.4.]

7 Application: Using Evolutionary Models in Protein Alignments

When proteins are believed to descend from a common ancestor, evolutionary models can be used as a guide in pairwise sequence alignments. The two most common families of evolutionary models of proteins used in pairwise sequence alignments are Point Accepted Mutation (PAM) matrices, which are based on explicit evolutionary models, and Block Substitution Matrix (BLOSUM) matrices, which are based on data-derived evolution models. The Biostrings package contains 5 PAM and 5 BLOSUM matrices (PAM30, PAM40, PAM70, PAM120, PAM250, BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80, and BLOSUM100) that can be used in the `substitutionMatrix` argument to the `pairwiseAlignment` function.

Here is an example pairwise sequence alignment of amino acids from Durbin, Eddy et al being fit by the `pairwiseAlignment` function using the BLOSUM50 matrix:

```r
> data(BLOSUM50)
> BLOSUM50[1:4,1:4]
 A R N D
A 5 -2 -1 -2
R -2 7 -1 -2
N -1 -1 7 2
D -2 -2 2 8
```

```r
> nwdemo <- pairwiseAlignment(AAString("PAWHEAE"), AAString("HEAGAWGHEE"), substitutionMatrix = BLOSUM50, gapOpening = 0, gapExtension = 8)
```

7.1 Exercise 5

1. Repeat the alignment exercise above using BLOSUM62, a gap opening penalty of 12, and a gap extension penalty of 4.

2. Explore to find out what caused the alignment to change.

[Answers provided in section 12.5.]
8 Application: Removing Adapters from Sequence Reads

Finding and removing uninteresting experiment process-related fragments like adapters is a common problem in genetic sequencing, and pairwise sequence alignment is well-suited to address this issue. When adapters are used to anchor or extend a sequence during the experiment process, they either intentionally or unintentionally become sequenced during the read process. The following code simulates what sequences with adapter fragments at either end could look like during an experiment.

```r
> simulateReads <-
+ function(N, adapter, experiment, substitutionRate = 0.01, gapRate = 0.001) {
+   chars <- strsplit(as.character(adapter), "")[[1]]
+   sapply(seq_len(N), function(i, experiment, substitutionRate, gapRate) {
+     width <- experiment["width"][[i]]
+     side <- experiment["side"][[i]]
+     randomLetters <-
+       function(n) sample(DNA_ALPHABET[1:4], n, replace = TRUE)
+     randomLettersWithEmpty <-
+       function(n)
+         sample(c("", DNA_ALPHABET[1:4]), n, replace = TRUE,
+           prob = c(1 - gapRate, rep(gapRate/4, 4)))
+     nChars <- length(chars)
+     value <-
+       paste(ifelse(rbinom(nChars,1,substitutionRate), randomLetters(nChars), chars),
+         randomLettersWithEmpty(nChars),
+         sep = "", collapse = "")
+     if (side)
+       value <-
+         paste(c(randomLetters(36 - width), substring(value, 1, width)),
+           sep = "", collapse = "")
+     else
+       value <-
+         paste(c(substring(value, 37 - width, 36), randomLetters(36 - width)),
+           sep = "", collapse = "")
+     value
+   }, experiment = experiment, substitutionRate = substitutionRate, gapRate = gapRate)
+ }
> adapter <- DNAString("GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA")
> set.seed(123)
> N <- 1000
> experiment <-
+   list(side = rbinom(N, 1, 0.5), width = sample(0:36, N, replace = TRUE))
> table(experiment["side"], experiment[["width"]])

           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
    0    13 10  8  7 11 18  9 15 15 18 16 10 11  9 13 13 18 18 14  9 19 13
    1    15 21 21 12 17 14  8 11 12 10 14 16  7 14 19 14 14 16 14 16 16 16

           22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
    0    19 19 12  6 15 18 12 15 16 17 19  6 13 18 15
    1    13 11  9 11 15 10 10 16 15 15 11  7 12  7 14
> adapterStrings <-
+   simulateReads(N, adapter, experiment, substitutionRate = 0.01, gapRate = 0.001)
```

---

13
These simulated strings above have 0 to 36 characters from the adapters attached to either end. We can use completely random strings as a baseline for any pairwise sequence alignment methodology we develop to remove the adapter characters.

Since edit distances are easy to explain, it serves as a good place to start for developing a adapter removal methodology. Unfortunately given that it is based on a global alignment, it only is useful for filtering out sequences that are derived primarily from the adapter.

One improvement to removing adapters is to look at consecutive matches anywhere within the sequence. This is more versatile than the edit distance method, but it requires a relatively large number of consecutive matches and is susceptible to issues related to error related substitutions and insertions/deletions.
> adapterAligns2 <-
+   pairwiseAlignment(adapterStrings, adapter, substitutionMatrix = submat2,
+   type = "local", gapOpening = 0, gapExtension = Inf)
> table(score(adapterAligns2) > quantile(randomScores2, 0.999), experiment["width"])

> # Determine if the correct end was chosen
> table(start(pattern(adapterAligns2)) > 37 - end(pattern(adapterAligns2)),
+   experiment["side"])

# Limiting consecutive matches to the ends provides better results, but it doesn’t resolve the issues related
# to substitutions and insertions/deletions errors.

> ## Method 3: Use consecutive matches on the ends with an FDR of 1e-03
> submat3 <- nucleotideSubstitutionMatrix(match = 1, mismatch = -Inf, baseOnly = TRUE)
> randomScores3 <-
+   pairwiseAlignment(randomStrings, adapter, substitutionMatrix = submat3,
+   type = "overlap", gapOpening = 0, gapExtension = Inf,
+   scoreOnly = TRUE)
> quantile(randomScores3, seq(0.99, 1, by = 0.001))

> adapterAligns3 <-
+   pairwiseAlignment(adapterStrings, adapter, substitutionMatrix = submat3,
+   type = "overlap", gapOpening = 0, gapExtension = Inf)
> table(score(adapterAligns3) > quantile(randomScores3, 0.999), experiment["width"])

> # Determine if the correct end was chosen
> table(end(pattern(adapterAligns3)) == 36, experiment["side"])

15
Allowing for substitutions and insertions/deletions errors in the pairwise sequence alignments provides much better results for finding adapter fragments.

```r
# Method 4: Allow mismatches and indels on the ends with an FDR of 1e-03
randomScores4 <- pairwiseAlignment(randomStrings, adapter, type = "overlap", scoreOnly = TRUE)
quantile(randomScores4, seq(0.99, 1, by = 0.001))

99% 99.1% 99.2% 99.3% 99.4% 99.5% 99.6%
7.927024 7.927024 7.927024 7.927024 7.927024 7.927024 7.927024
99.7% 99.8% 99.9% 100%

adapterAligns4 <- pairwiseAlignment(adapterStrings, adapter, type = "overlap")
table(score(adapterAligns4) > quantile(randomScores4, 0.999), experiment["width"])

FALSE 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
28 31 29 19 28 32 17 1 2 5 1 0 1 0 1 0 0 0 0
TRUE 29 32 30 21 17 30 28 22 31 31 32 30 13 25 25 29

# Determine if the correct end was chosen
table(end(pattern(adapterAligns4)) == 36, experiment["side"])

FALSE 0 1
484 17
TRUE 23 476

Using the results that allow for substitutions and insertions/deletions errors, the cleaned sequence fragments can be generated as follows:

```r
# Method 4 continued: Remove adapter fragments
fragmentFound <- score(adapterAligns4) > quantile(randomScores4, 0.999)
fragmentFoundAt1 <- fragmentFound & (start(pattern(adapterAligns4)) == 1)
fragmentFoundAt36 <- fragmentFound & (end(pattern(adapterAligns4)) == 36)
cleanedStrings <- as.character(adapterStrings)
cleanedStrings[fragmentFoundAt1] <- as.character(narrow(adapterStrings[fragmentFoundAt1], end = 36,
+ width = 36 - end(pattern(adapterAligns4[fragmentFoundAt1]) - 1)))
cleanedStrings[fragmentFoundAt36] <- as.character(narrow(adapterStrings[fragmentFoundAt36], start = 1,
+ width = start(pattern(adapterAligns4[fragmentFoundAt36])) - 1))
cleanedStrings <- DNAStringSet(cleanedStrings)
cleanedStrings
```
A DNAStringSet instance of length 1000
width seq
[1]  26 TTGCACGATAGTTGCAATATGCTACAA
[2] 15 ATTTTCTCCTCTCAG
[3]  36 TGAACCGGGTCACCATCATCATCTAGTAGGACCTT
[4]   5 GCGTT
[5]   5 GCTTA
... ... ...
[996]  3 TAA
[997]  25 TTGTTTGGGAAGGCTTCAACCAAC
[998]  9 CAGCTAATA
[999] 22 CGGATGGGACGCACATGACT
[1000] 20 TCGCTATCTGAAGTCTGTTT

8.1 Exercise 6

1. Rerun the simulation time using the simulateReads function with a substitutionRate of 0.005 and gapRate of 0.0005. How do the different pairwise sequence alignment methods compare?

2. (Advanced) Modify the simulateReads function to accept different equal length adapters on either side (left & right) of the reads. How would the methods for trimming the reads change?

[Answers provided in section 12.6]

9 Application: Quality Assurance in Sequencing Experiments

Due to its flexibility, the pairwiseAlignment function is able to diagnose sequence matching-related issues that arise when matchPDict and its related functions don’t find a match. This section contains an example involving a short read Solexa sequencing experiment of bacteriophage $\phi$ X174 DNA produced by New England BioLabs (NEB). This experiment contains slightly less than 5000 unique short reads in srPhiX174, with quality measures in quPhiX174, and frequency for those short reads in wtPhiX174.

In order to demonstrate how to find sequence differences in the target, these short reads will be compared against the bacteriophage $\phi$ X174 genome NC_001422 from the GenBank database.

```r
> data(phiX174Phage)
> genBankPhage <- phiX174Phage[[1]]
> nchar(genBankPhage)

[1] 5386
```

```r
> data(srPhiX174)
> srPhiX174

A DNAStringSet instance of length 1113
width seq
[1]  35 GTTATTATACCGTCAAGGACTGTGACTATTGAC
[2]  35 GGTGGTTATACCGTCAAGGACTGTGACTATTGAC
[3]  35 TACCGTCAAGGACTGTGACTATTGACGTCCTTC
[4]  35 GTACGCCGGCAATAATATTGTGTTGTTTCATG
[5]  35 GTTTTCACTGTTTGGTCTAATTTTACCCGTACTAA
... ... ...
[1109]  35 ATAATGTCTTATATGTGTTTTTACTGTTTCTATC
```

17
For these short reads, the `pairwiseAlignment` function finds that the small number of perfect matches is due to two locations on the bacteriophage φX174 genome.

Unlike the `countPDict` function, the `pairwiseAlignment` function works off of the original strings, rather than `PDict` processed strings, and to be computationally efficient it is recommended that the unique sequences are supplied to the `pairwiseAlignment` function, and the frequencies of those sequences are supplied to the `weight` argument of functions like `summary`, `mismatchSummary`, and `coverage`. For the purposes of this exercise, a substring of the GenBank bacteriophage φX174 genome is supplied to the `subject` argument of the `pairwiseAlignment` function to reduce the computation time.
Scores:

\[
\begin{array}{cccccc}
\text{Min.} & \text{1st Qu.} & \text{Median} & \text{Mean} & \text{3rd Qu.} & \text{Max.} \\
-45.08 & 35.81 & 50.07 & 41.24 & 59.50 & 67.35 \\
\end{array}
\]

Number of matches:

\[
\begin{array}{cccccc}
\text{Min.} & \text{1st Qu.} & \text{Median} & \text{Mean} & \text{3rd Qu.} & \text{Max.} \\
21.00 & 31.00 & 33.00 & 31.46 & 34.00 & 35.00 \\
\end{array}
\]

Top 10 Mismatch Counts:

<table>
<thead>
<tr>
<th>SubjectPosition</th>
<th>Subject</th>
<th>Pattern</th>
<th>Count</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
<td>C</td>
<td>T</td>
<td>22965</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>C</td>
<td>T</td>
<td>22849</td>
</tr>
<tr>
<td>3</td>
<td>76</td>
<td>G</td>
<td>T</td>
<td>1985</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>A</td>
<td>T</td>
<td>1296</td>
</tr>
<tr>
<td>5</td>
<td>79</td>
<td>C</td>
<td>T</td>
<td>1289</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
<td>A</td>
<td>C</td>
<td>1153</td>
</tr>
<tr>
<td>7</td>
<td>72</td>
<td>G</td>
<td>A</td>
<td>1130</td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>G</td>
<td>A</td>
<td>1130</td>
</tr>
<tr>
<td>9</td>
<td>67</td>
<td>T</td>
<td>G</td>
<td>1130</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>A</td>
<td>G</td>
<td>1103</td>
</tr>
</tbody>
</table>

```r
> revisedPhage <- + replaceLetterAt(genBankPhage, c(2793, 2811), "TT")
> table(countPDict(srPDict, revisedPhage))

0  1
6768 47034
```

The following plot shows the coverage of the aligned short reads along the substring of the bacteriophage \( \phi \) X174 genome. Applying the `slice` function to the coverage shows the entire substring is covered by aligned short reads.

```r
> genBankCoverage <- coverage(genBankAlign, weight = wtPhiX174)
> plot((2793-34):(2811+34), as.integer(genBankCoverage), xlab = "Position", ylab = "Coverage", +    type = "l")
> nchar(genBankSubstring)

[1] 87
> slice(genBankCoverage, lower = 1)
```

Views on a 87-length Rle subject

views:

```r
  start end width
[1]  1  87  87 [ 8899 9698 10484 11228 11951 12995 13547 ...]
```
9.1 Exercise 7

1. Rerun the global-local alignment of the short reads against the entire genome. (This may take a few minutes.)

2. Plot the coverage of these alignments and use the `slice` function to find the ranges of alignment. Are there any alignments outside of the substring region that was used above?

3. Use the `reverseComplement` function on the bacteriophage ϕ X174 genome. Do any short reads have a higher alignment score on this new sequence than on the original sequence?

   [Answers provided in section 12.7]

10 Computation Profiling

The `pairwiseAlignment` function uses a dynamic programming algorithm based on the Needleman-Wunsch and Smith-Waterman algorithms for global and local pairwise sequence alignments respectively. The algorithm consumes memory and computation time proportional to the product of the length of the two strings being aligned.
```r
N <- as.integer(seq(500, 5000, by = 500))
timings <- rep(0, length(N))
names(timings) <- as.character(N)
for (i in seq_len(length(N))) {
  string1 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
  string2 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
  timings[i] <- system.time(pairwiseAlignment(string1, string2, type = "global"))["user.self"]
}
timings

500 1000 1500 2000 2500 3000 3500 4000 4500 5000
0.136 0.140 0.164 0.196 0.252 0.316 0.384 0.580 0.544 0.616

coefficients:

```

```

> coef(summary(lm(timings ~ poly(N, 2))))

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 0.33280000 | 0.01440273 | 23.106731 | 7.211145e-08 |
| poly(N, 2)1 | 0.54255494 | 0.04554543 | 11.912391 | 6.678712e-06 |
| poly(N, 2)2 | 0.09783164 | 0.04554543 | 2.148001 | 6.881642e-02 |

> plot(N, timings, xlab = "String Size, Both Strings", ylab = "Timing (sec.")", type = "l",
+     main = "Global Pairwise Sequence Alignment Timings")
```
When a problem only requires the pairwise sequence alignment score, setting the `scoreOnly` argument to `TRUE` will more than halve the computation time.

```r
> scoreOnlyTimings <- rep(0, length(N))
> names(scoreOnlyTimings) <- as.character(N)
> for (i in seq_len(length(N))) {
+   string1 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
+   string2 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
+   scoreOnlyTimings[i] <- system.time(pairwiseAlignment(string1, string2, type = "global", scoreOnly = TRUE))[["user.self"]]
+ }
> scoreOnlyTimings

500 1000 1500 2000 2500 3000 3500 4000 4500 5000
0.112 0.120 0.144 0.272 0.216 0.224 0.292 0.316 0.388 0.440

> round((timings - scoreOnlyTimings) / timings, 2)

500 1000 1500 2000 2500 3000 3500 4000 4500 5000
0.18 0.14 0.12 -0.39 0.14 0.29 0.24 0.46 0.29 0.29
10.1 Exercise 8

1. Rerun the first set of profiling code, but this time fix the number of characters in string1 to 35 and have the number of characters in string2 range from 5000, 50000, by increments of 5000. What is the computational order of this simulation exercise?

2. Rerun the second set of profiling code using the simulations from the previous exercise with scoreOnly argument set to TRUE. Is it still twice as fast?

[Answers provided in section 12.8]

11 Computing alignment consensus matrices

The consensusMatrix function is provided for computing a consensus matrix for a set of equal-length strings assumed to be aligned. To illustrate, the following application assumes the ORF data to be aligned for the first 10 positions (patently false):

```r
> file <- system.file("extdata", "someORF.fa", package="Biostrings")
> orf <- readDNAStringSet(file)
> orf

A DNAStringSet instance of length 7

<table>
<thead>
<tr>
<th></th>
<th>seq</th>
<th>names</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACTTAAATATCTTTTT...TCGACCTTTCTTGATAT</td>
<td>YAL001C TFC3 SGDI...</td>
</tr>
<tr>
<td>2</td>
<td>TTCAAGGCGATGAATTC...AATTTTTCTATCTCTT</td>
<td>YAL002W VPS8 SGDI...</td>
</tr>
<tr>
<td>3</td>
<td>CTTCATGTCAGCTGCCTCAGCT...ACTCATGTAGCTGCCTCAT</td>
<td>YAL003W EFB1 SGDI...</td>
</tr>
<tr>
<td>4</td>
<td>AGAGAGAGATTTACTCTTTCA...CTTTTTTTTCTATTCTTT</td>
<td>YAL005C SSA1 SGDI...</td>
</tr>
<tr>
<td>5</td>
<td>TTTTCTCCTTTACTTTACTTTATTCTTCTT</td>
<td>YAL007C ERP2 SGDI...</td>
</tr>
<tr>
<td>6</td>
<td>CTTAGCGGCTGCTGCGGCTGAGC...TTTGGCAGAATGTACTTTT</td>
<td>YAL008W FUN14 SGD...</td>
</tr>
<tr>
<td>7</td>
<td>AGAGAGAAAAAATCAG</td>
<td>YAL009W SP07 SGDI...</td>
</tr>
</tbody>
</table>

> orf10 <- DNAStringSet(orf, end=10)
> consensusMatrix(orf10, as.prob=TRUE, baseOnly=TRUE)

<table>
<thead>
<tr>
<th>,1</th>
<th>,2</th>
<th>,3</th>
<th>,4</th>
<th>,5</th>
<th>,6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.2857143</td>
<td>0.2857143</td>
<td>0.2857143</td>
<td>0.0000000</td>
<td>0.5714286</td>
</tr>
<tr>
<td>C</td>
<td>0.4285714</td>
<td>0.1428571</td>
<td>0.2857143</td>
<td>0.2857143</td>
<td>0.2857143</td>
</tr>
<tr>
<td>G</td>
<td>0.1428571</td>
<td>0.1428571</td>
<td>0.1428571</td>
<td>0.1428571</td>
<td>0.2857143</td>
</tr>
<tr>
<td>T</td>
<td>0.1428571</td>
<td>0.4285714</td>
<td>0.2857143</td>
<td>0.4285714</td>
<td>0.0000000</td>
</tr>
<tr>
<td>other</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>,7</th>
<th>,8</th>
<th>,9</th>
<th>,10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.4285714</td>
<td>0.4285714</td>
<td>0.2857143</td>
</tr>
<tr>
<td>C</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.2857143</td>
</tr>
<tr>
<td>G</td>
<td>0.4285714</td>
<td>0.4285714</td>
<td>0.1428571</td>
</tr>
<tr>
<td>T</td>
<td>0.1428571</td>
<td>0.1428571</td>
<td>0.2857143</td>
</tr>
<tr>
<td>other</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

The information content as defined by Hertz and Stormo 1995 is computed as follows:

```r
> informationContent <- function(Lmers) {
+   zlog <- function(x) ifelse(x==0,0,log(x))
+   co <- consensusMatrix(Lmers, as.prob=TRUE)
+   lets <- rownames(co)
+   for (i in c("A", "C", "G", "T", "other")) {
+     for (j in c("A", "C", "G", "T", "other")) {
+       EC[(i, j)] <- 2 * zlog(co[(i, j)] / co[lets[i], lets[j]])
+     }
+   }
+   return(EC)
+ }
```
+ fr <- alphabetFrequency(Lmers, collapse=TRUE)[lets]
+ fr <- fr / sum(fr)
+ sum(co*zlog(co/fr), na.rm=TRUE)
+
} 
> informationContent(orf10)

[1] 2.167186

12 Exercise Answers

12.1 Exercise 1

1. Using `pairwiseAlignment`, fit the global, local, and overlap pairwise sequence alignment of the strings "syzygy" and "zyzzyx" using the default settings.

```r
> pairwiseAlignment("zyzzyx", "syzygy")
Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] zyzzyx
subject: [1] syzygy
score: -19.3607
```

```r
> pairwiseAlignment("zyzzyx", "syzygy", type = "local")
Local PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [2] yz
subject: [2] yz
score: 4.607359
```

```r
> pairwiseAlignment("zyzzyx", "syzygy", type = "overlap")
Overlap PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] ""
subject: [1] ""
score: 0
```

2. Do any of the alignments change if the `gapExtension` argument is set to `-Inf`? Yes, the overlap pairwise sequence alignment changes.

```r
> pairwiseAlignment("zyzzyx", "syzygy", type = "overlap", gapExtension = Inf)
Overlap PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] ""
subject: [1] ""
score: 0
```

12.2 Exercise 2

1. What is the primary benefit of formal summary classes like `PairwiseAlignmentsSingleSubjectSummary` and `summary.lm` to end-users? These classes allow the end-user to extract the summary output for further operations.

```r
> ex2 <- summary(pairwiseAlignment("zyzzyx", "syzygy"))
> nmatch(ex2) / nmismatch(ex2)
[1] 0.5
```
12.3 Exercise 3

For the overlap pairwise sequence alignment of the strings "syzygy" and "zyzzyx" with the `pairwiseAlignment` default settings, perform the following operations:

```r
> ex3 <- pairwiseAlignment("zyzzyx", "syzygy", type = "overlap")
```

1. Use `nmatch` and `nmismatch` to extract the number of matches and mismatches respectively.
   ```r
   > nmatch(ex3)
   [1] 0
   > nmismatch(ex3)
   [1] 0
   ```

2. Use the `compareStrings` function to get the symbolic representation of the alignment.
   ```r
   > compareStrings(ex3)
   [1] ""
   ```

3. Use the `as.character` function to get the character string versions of the alignments.
   ```r
   > as.character(ex3)
   [1] "-------"
   ```

4. Use the `pattern` function to extract the aligned pattern and apply the `mismatch` function to it to find the locations of the mismatches.
   ```r
   > mismatch(pattern(ex3))
   IntegerList of length 1
   [[1]] integer(0)
   ```

5. Use the `subject` function to extract the aligned subject and apply the `aligned` function to it to get the aligned strings.
   ```r
   > aligned(subject(ex3))
   A BStringSet instance of length 1
   width seq
   [1] 0
   ```

12.4 Exercise 4

1. Use the `pairwiseAlignment` function to find the Levenshtein edit distance between "syzygy" and "zyzzyx".
   ```r
   > submat <- matrix(-1, nrow = 26, ncol = 26, dimnames = list(letters, letters))
   > diag(submat) <- 0
   > ex3 <- pairwiseAlignment("zyzzyx", "syzygy", substitutionMatrix = submat,
   +               gapOpening = 0, gapExtension = 1, scoreOnly = TRUE)
   ```
   ```r
   [1] 4
   ```
2. Use the `stringDist` function to find the Levenshtein edit distance for the vector `c("zyzzyx", "syzygy", "succeed", "precede", "supersede")`.

```r
> stringDist(c("zyzzyx", "syzygy", "succeed", "precede", "supersede"))
```

1 2 3 4 5
2 4
3 7 6
4 7 7 5
5 9 8 5 5

12.5 Exercise 5

1. Repeat the alignment exercise above using `BLOSUM62`, a gap opening penalty of 12, and a gap extension penalty of 4.

```r
> data(BLOSUM62)
> pairwiseAlignment(AAString("PAWHEAE"), AAString("HEAGAWGHEE"), substitutionMatrix = BLOSUM62, +
                   gapOpening = 12, gapExtension = 4)
```

Global PairwiseAlignmentsSingleSubject (1 of 1)
pattern: [1] P---AWHEAE
subject: [1] HEAGAWGHEE
score: -9

2. Explore to find out what caused the alignment to change. The shift in gap penalties favored infrequent long gaps to frequent short ones.

12.6 Exercise 6

1. Rerun the simulation time using the `simulateReads` function with a `substitutionRate` of 0.005 and a `gapRate` of 0.0005. How do the different pairwise sequence alignment methods compare? The different methods are much more comparable when the error rates are lower.

```r
> adapter <- DNAString("GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA")
> set.seed(123)
> N <- 1000
> experiment <- list(side = rbinom(N, 1, 0.5), width = sample(0:36, N, replace = TRUE))
> table(experiment[["side"]], experiment[["width"]])
```

```
  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
0  13 10  8 11 18  9 15 15 18 16 10 11  9 13 13 18 18 14  9 19 13
1 15 21 12 17 14  8 11 12 10 14 16  7 14 19 14 14 16 14 16 16 16
2 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
0 19 19 12  6 15 18 12 15 16 17 19  6 13 18 15
1 13 11  9 11 15 10 16 15 15 11  7 12  7 14
> ex6Strings <-
+  simulateReads(N, adapter, experiment, substitutionRate = 0.005, gapRate = 0.0005)
> ex6Strings <- DNAStringSet(ex6Strings)
> ex6Strings
```
A DNAStringSet instance of length 1000

```r
> # Method 1: Use edit distance with an FDR of 1e-03
> submat1 <- nucleotideSubstitutionMatrix(match = 0, mismatch = -1, baseOnly = TRUE)
> quantile(randomScores1, seq(0.99, 1, by = 0.001))
```

<table>
<thead>
<tr>
<th>99%</th>
<th>99.1%</th>
<th>99.2%</th>
<th>99.3%</th>
<th>99.4%</th>
<th>99.5%</th>
<th>99.6%</th>
<th>99.7%</th>
<th>99.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-16.996</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-16.000</td>
<td>-15.998</td>
</tr>
<tr>
<td>99.9%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
> ex6Aligns1 <- pairwiseAlignment(ex6Strings, adapter, substitutionMatrix = submat1, gapOpening = 0, gapExtension = 1)
> table(score(ex6Aligns1) > quantile(randomScores1, 0.999), experiment["width"])
```

|     | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|-----|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| FALSE | 28 | 31 | 29 | 19 | 28 | 32 | 17 | 26 | 27 | 28 | 30 | 26 | 18 | 23 | 32 | 27 | 32 | 28 | 25 | 35 |
| TRUE  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |

```
> # Method 2: Use consecutive matches anywhere in string with an FDR of 1e-03
> submat2 <- nucleotideSubstitutionMatrix(match = 1, mismatch = -Inf, baseOnly = TRUE)
> quantile(randomScores2, seq(0.99, 1, by = 0.001))
```

<table>
<thead>
<tr>
<th>99%</th>
<th>99.1%</th>
<th>99.2%</th>
<th>99.3%</th>
<th>99.4%</th>
<th>99.5%</th>
<th>99.6%</th>
<th>99.7%</th>
<th>99.8%</th>
<th>99.9%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

```
> ex6Aligns2 <- pairwiseAlignment(ex6Strings, adapter, substitutionMatrix = submat2, type = "local", gapOpening = 0, gapExtension = Inf)
> table(score(ex6Aligns2) > quantile(randomScores2, 0.999), experiment["width"])
```

|     | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|-----|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| FALSE | 28 | 31 | 29 | 19 | 28 | 32 | 17 | 26 | 27 | 28 | 4  | 1  | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 26 | 25 | 15 |
| TRUE  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 26 | 25 | 15 | 23 | 31 | 26 | 32 | 34 | 27 | 35 | 36 |
```
> # Determine if the correct end was chosen
> table(start(pattern(ex6Aligns2)) > 37 - end(pattern(ex6Aligns2)), +
>     experiment[["side"]])
>
>  0  1
> FALSE 473 48
> TRUE  34 445

> ## Method 3: Use consecutive matches on the ends with an FDR of 1e-03
> submat3 <- nucleotideSubstitutionMatrix(match = 1, mismatch = -Inf, baseOnly = TRUE)
> ex6Aligns3 <-
>   pairwiseAlignment(ex6Strings, adapter, substitutionMatrix = submat3,
>                     type = "overlap", gapOpening = 0, gapExtension = Inf)
> table(score(ex6Aligns3) > quantile(randomScores3, 0.999), experiment[["width"]])

>  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
> FALSE 28 31 29 19 28 32 17 0 1 2 1 3 1 1 2 2 3 2 2 3
> TRUE 0 0 0 0 0 0 0 26 26 26 25 16 17 18 19 20

>  21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
> FALSE 5 2 3 2 1 3 5 3 3 2 2 4 3
> TRUE 24 30 27 19 16 27 23 19 28 28 30 27 11 21 21 26

> # Determine if the correct end was chosen
> table(end(pattern(ex6Aligns3)) == 36, experiment[["side"]])
>
>  0  1
> FALSE 473 42
> TRUE  34 451

> ## Method 4: Allow mismatches and indels on the ends with an FDR of 1e-03
> quantile(randomScores4, seq(0.99, 1, by = 0.001))

>  99%  99.1%  99.2%  99.3%  99.4%  99.5%  99.6%
>  7.927024  7.927024  7.927024  7.927024  7.927024  7.927024  7.927024
>  99.7%  99.8%  99.9%  100%

> ex6Aligns4 <- pairwiseAlignment(ex6Strings, adapter, type = "overlap")
> table(score(ex6Aligns4) > quantile(randomScores4, 0.999), experiment[["width"]])

>  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
> FALSE 28 31 29 19 28 32 17 0 1 2 1 0 1 0 1 0 0 0 0 0 0 0
> TRUE 0 0 0 0 0 0 0 26 26 26 29 26 17 23 31 27 32 34 28 25 35

>  21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
> FALSE 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> TRUE 29 32 30 21 17 30 28 22 31 32 30 13 25 25 29

> # Determine if the correct end was chosen
> table(end(pattern(ex6Aligns4)) == 36, experiment[["side"]])

28
2. (Advanced) Modify the simulateReads function to accept different equal length adapters on either side (left & right) of the reads. How would the methods for trimming the reads change?

```r
> simulateReads <-
+ function(N, left, right = left, experiment, substitutionRate = 0.01, gapRate = 0.001) {
+ leftChars <- strsplit(as.character(left), "")[[1]]
+ rightChars <- strsplit(as.character(right), "")[[1]]
+ if (length(leftChars) != length(rightChars))
+ stop("left and right adapters must have the same number of characters")
+ nChars <- length(leftChars)
+ sapply(seq_len(N), function(i) {
+ width <- experiment["width"][[i]]
+ side <- experiment["side"][[i]]
+ randomLetters <-
+ function(n) sample(DNA_ALPHABET[1:4], n, replace = TRUE)
+ randomLettersWithEmpty <-
+ function(n)
+ sample(c("", DNA_ALPHABET[1:4]), n, replace = TRUE,
+ prob = c(1 - gapRate, rep(gapRate/4, 4)))
+ if (side) {
+ value <-
+ paste(ifelse(rbinom(nChars,1,substitutionRate), randomLetters(nChars), rightChars),
+ randomLettersWithEmpty(nChars),
+ sep = "", collapse = "")
+ value <-
+ paste(c(randomLetters(36 - width), substring(value, 1, width)),
+ sep = "", collapse = "")
+ } else {
+ value <-
+ paste(ifelse(rbinom(nChars,1,substitutionRate), randomLetters(nChars), leftChars),
+ randomLettersWithEmpty(nChars),
+ sep = "", collapse = "")
+ value <-
+ paste(c(substring(value, 37 - width, 36), randomLetters(36 - width)),
+ sep = "", collapse = "")
+ }
+ value
+ })
+ }
> leftAdapter <- adapter
> rightAdapter <- reverseComplement(adapter)
> ex6LeftRightStrings <- simulateReads(N, leftAdapter, rightAdapter, experiment)
> ex6LeftAligns4 <-
+ pairwiseAlignment(ex6LeftRightStrings, leftAdapter, type = "overlap")
> ex6RightAligns4 <-
+ pairwiseAlignment(ex6LeftRightStrings, rightAdapter, type = "overlap")
> scoreCutoff <- quantile(randomScores4, 0.999)
> leftAligned <-
```
12.7 Exercise 7

1. Rerun the global-local alignment of the short reads against the entire genome. (This may take a few minutes.)

   > genBankFullAlign <-
   + pairwiseAlignment(srPhiX174, genBankPhage,
   +     patternQuality = SolexaQuality(quPhiX174),
   +     subjectQuality = SolexaQuality(99L),
   +     type = "global-local")
   > summary(genBankFullAlign, weight = wtPhiX174)

Global-Local Single Subject Pairwise Alignments
Number of Alignments: 53802

Scores:
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
   -45.08 56.72 59.89 60.59 69.56 69.85

Number of matches:
   Min. 1st Qu. Median  Mean 3rd Qu.   Max.
   24.00 33.00 34.00 34.01 35.00 35.00

Top 10 Mismatch Counts:
   SubjectPosition Subject Pattern Count Probability
   1   2811   C   T   2296 5 0.999912919
   2   2793   C   T   22845 0.999693681
   3   2834   G   T   1985 0.106800818
   4   2835   G   T    605 0.033570081
   5   2829   G   T    489 0.023314580
   6   2782   G   T    325 0.013882363
   7   2839   A   T    287 0.018648473
2. Plot the coverage of these alignments and use the slice function to find the ranges of alignment. Are there any alignments outside of the substring region that was used above? Yes, there are some alignments outside of the specified substring region.

```r
> genBankFullCoverage <- coverage(genBankFullAlign, weight = wtPhiX174)
> plot(as.integer(genBankFullCoverage), xlab = "Position", ylab = "Coverage", type = "l")
> slice(genBankFullCoverage, lower = 1)
```

Views on a 5386-length Rle subject

views:

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1195</td>
<td>1230</td>
<td>36</td>
</tr>
<tr>
<td>2514</td>
<td>2548</td>
<td>35</td>
</tr>
<tr>
<td>2745</td>
<td>2859</td>
<td>115</td>
</tr>
<tr>
<td>3209</td>
<td>3247</td>
<td>39</td>
</tr>
<tr>
<td>3964</td>
<td>3998</td>
<td>35</td>
</tr>
</tbody>
</table>

3. Use the reverseComplement function on the bacteriophage \( \phi \) X174 genome. Do any short reads have a higher alignment score on this new sequence than on the original sequence? Yes, there are some strings with a higher score on the new sequence.

```r
> genBankFullAlignRevComp <- pairwiseAlignment(srPhiX174, reverseComplement(genBankPhage),
+ patternQuality = SolexaQuality(quPhiX174),
+ subjectQuality = SolexaQuality(99L),
+ type = "global-local")
> table(score(genBankFullAlignRevComp) > score(genBankFullAlign))

FALSE  TRUE
1112  1
```

12.8 Exercise 8

1. Rerun the first set of profiling code, but this time fix the number of characters in string1 to 35 and have the number of characters in string2 range from 5000, 50000, by increments of 5000. What is the computational order of this simulation exercise? As expected, the growth in time is now linear.

```r
> N <- as.integer(seq(5000, 50000, by = 5000))
> newTimings <- rep(0, length(N))
> names(newTimings) <- as.character(N)
> for (i in seq_len(length(N))) {
+   string1 <- DNAString(paste(sample(DNA_ALPHABET[1:4], 35, replace = TRUE), collapse = ""))
+   string2 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
+   newTimings[i] <- system.time(pairwiseAlignment(string1, string2, type = "global"))[["user.self"]]
+ }
> newTimings
```

<table>
<thead>
<tr>
<th>5000</th>
<th>10000</th>
<th>15000</th>
<th>20000</th>
<th>25000</th>
<th>30000</th>
<th>35000</th>
<th>40000</th>
<th>45000</th>
<th>50000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.120</td>
<td>0.116</td>
<td>0.124</td>
<td>0.128</td>
<td>0.132</td>
<td>0.136</td>
<td>0.132</td>
<td>0.132</td>
<td>0.144</td>
<td>0.148</td>
</tr>
</tbody>
</table>
> coef(summary(lm(newTimings ~ poly(N, 2))))

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 0.1324000000 | 0.001202019 | 110.1480547 | 1.340164e-12 |
| poly(N, 2)1 | 0.0306067927 | 0.003801116 | 8.0520538 | 8.744956e-05 |
| poly(N, 2)2 | -0.0006963106 | 0.003801116 | -0.1831858 | 8.598452e-01 |

> plot(N, newTimings, xlab = "Larger String Size", ylab = "Timing (sec.")

2. Rerun the second set of profiling code using the simulations from the previous exercise with `scoreOnly` argument set to `TRUE`. Is it still twice as fast? Yes, it is still over twice as fast.

```r
> newScoreOnlyTimings <- rep(0, length(N))
> names(newScoreOnlyTimings) <- as.character(N)
> for (i in seq_len(length(N))) {
+     string1 <- DNAString(paste(sample(DNA_ALPHABET[1:4], 35, replace = TRUE), collapse = ""))
+     string2 <- DNAString(paste(sample(DNA_ALPHABET[1:4], N[i], replace = TRUE), collapse = ""))
+     newScoreOnlyTimings[i] <- system.time(pairwiseAlignment(string1, string2, type = "global", score
```
> newScoreOnlyTimings

5000 10000 15000 20000 25000 30000 35000 40000 45000 50000
0.120 0.120 0.116 0.112 0.124 0.144 0.160 0.152 0.128 0.136

> round((newTimings - newScoreOnlyTimings) / newTimings, 2)

5000 10000 15000 20000 25000 30000 35000 40000 45000 50000
0.00 -0.03 0.06 0.13 0.06 -0.06 -0.21 -0.06 0.14 0.06

13 Session Information

All of the output in this vignette was produced under the following conditions:

> sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C

attached base packages:
[1] stats4 parallel stats graphics grDevices utils
[7] datasets methods base

other attached packages:
[1] Biostrings_2.42.1 XVector_0.14.0 IRanges_2.8.1
[4] S4Vectors_0.12.1 BiocGenerics_0.20.0

loaded via a namespace (and not attached):
[1] zlibbioc_1.20.0 tools_3.3.2

References


