

# Package ‘seqtrie’

September 1, 2023

**Title** Radix Tree and Trie-Based String Distances

**Version** 0.2.5

**Date** 2023-8-31

**Description** A collection of Radix Tree and Trie algorithms for finding similar sequences and calculating sequence distances (Levenshtein and other distance metrics). This work was inspired by a trie implementation in Python: “Fast and Easy Levenshtein distance using a Trie.” Hanov (2011) <<http://stevehanov.ca/blog/index.php?id=114>>.

**License** GPL-3

**Biarch** true

**Encoding** UTF-8

**Depends** R (>= 3.5.0)

**LazyData** true

**SystemRequirements** GNU make

**LinkingTo** Rcpp, RcppParallel, BH

**Imports** Rcpp (>= 0.12.18.3), RcppParallel (>= 5.1.3), R6

**Suggests** knitr, rmarkdown, stringdist, stringfish, qs, dplyr, Biostrings, igraph, ggplot2

**VignetteBuilder** knitr

**RoxygenNote** 7.2.3

**Copyright** This package includes code from the ‘span-lite’ library owned by Martin Moene under Boost Software License 1.0. This package contains data derived from Adaptive Biotechnologies “ImmuneCODE” dataset under Creative Commons Attribution 4.0.

**URL** <https://github.com/traversc/seqtrie>

**BugReports** <https://github.com/traversc/seqtrie/issues>

**NeedsCompilation** yes

**Author** Travers Ching [aut, cre, cph],  
Martin Moene [ctb, cph] (span-lite C++ library),  
Hanov Steve [ctb] (Trie levenshtein implementation in Python)

**Maintainer** Travers Ching <traversc@gmail.com>

**Repository** CRAN

**Date/Publication** 2023-09-01 11:00:02 UTC

## R topics documented:

covid_cdr3	2
dist_matrix	3
dist_pairwise	4
dist_search	6
generate_cost_matrix	7
RadixForest	8
RadixTree	11

**Index** **16**

---

covid_cdr3	<i>Adaptive COVID TCRB CDR3 data</i>
------------	--------------------------------------

---

### Description

Unique TCRB CDR3 sequences from the Nolan et al. 2020. CDR3s were extracted via IgBLAST. The license for this data is Creative Commons Attribution 4.0 International License.

### Usage

```
data(covid_cdr3)
```

### Format

A character vector of length 133,034.

### References

Nolan, Sean, et al. "A large-scale database of T-cell receptor beta (TCRB) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2." (2020). doi: 10.21203/rs.3.rs-51964/v1.

### Examples

```
data(covid_cdr3)
# Average CDR3 length
mean(nchar(covid_cdr3)) # [1] 43.56821
```

---

dist_matrix	<i>Compute distances between all combinations of two sets of sequences</i>
-------------	--

---

## Description

Compute distances between all combinations of query and target sequences

## Usage

```
dist_matrix(  
  query,  
  target,  
  mode,  
  cost_matrix = NULL,  
  gap_cost = NULL,  
  gap_open_cost = NULL,  
  nthreads = 1,  
  show_progress = FALSE  
)
```

## Arguments

query	A character vector of query sequences.
target	A character vector of target sequences.
mode	The distance metric to use. One of hamming (hm), global (gb) or anchored (an).
cost_matrix	A custom cost matrix for use with the "global" or "anchored" distance metrics. See details.
gap_cost	The cost of a gap for use with the "global" or "anchored" distance metrics. See details.
gap_open_cost	The cost of a gap opening. See details.
nthreads	The number of threads to use for parallel computation.
show_progress	Whether to show a progress bar.

## Details

This function calculates all combinations of pairwise distances based on Hamming, Levenshtein or Anchored algorithms. The output is a NxM matrix where N = length(query) and M = length(target). Note: this can take a *really* long time; be careful with input size.

Three types of distance metrics are supported, based on the form of alignment performed. These are: Hamming, Global (Levenshtein) and Anchored.

An anchored alignment is a form of semi-global alignment, where the query sequence is "anchored" (global) to the beginning of both the query and target sequences, but is semi-global in that the end of either the query sequence or target sequence (but not both) can be unaligned. This type of alignment is sometimes called an "extension" alignment in literature.

In contrast a global alignment must align the entire query and target sequences. When mismatch and indel costs are equal to 1, this is also known as the Levenshtein distance.

By default, if mode == "global" or "anchored", all mismatches and indels are given a cost of 1. However, you can define your own distance metric by setting the cost\_matrix and gap parameters. The cost\_matrix is a strictly positive square integer matrix and should include all characters in query and target as column- and rownames. To set the cost of a gap (insertion or deletion) you can include a row and column named "gap" in the cost\_matrix *OR* set the gap\_cost parameter (a single positive integer). Similarly, the affine gap alignment can be set by including a row and column named "gap\_open" in the cost\_matrix *OR* setting the gap\_open\_cost parameter (a single positive integer). If affine alignment is used, the cost of a gap is defined as: TOTAL\_GAP\_COST = gap\_open\_cost + (gap\_cost \* gap\_length).

If mode == "hamming" all alignment parameters are ignored; mismatch is given a distance of 1 and gaps are not allowed.

### Value

The output is a distance matrix between all query (rows) and target (columns) sequences. For anchored searches, the output also includes attributes "query\_size" and "target\_size" which are matrices containing the lengths of the query and target sequences that are aligned.

### Examples

```
dist_matrix(c("ACGT", "AAAA"), c("ACG", "ACGT"), mode = "global")
```

---

dist_pairwise	<i>Pairwise distance between two sets of sequences</i>
---------------	--

---

### Description

Compute the pairwise distance between two sets of sequences

### Usage

```
dist_pairwise(
  query,
  target,
  mode,
  cost_matrix = NULL,
  gap_cost = NULL,
  gap_open_cost = NULL,
  nthreads = 1,
  show_progress = FALSE
)
```

## Arguments

query	A character vector of query sequences.
target	A character vector of target sequences.. Must be the same length as query.
mode	The distance metric to use. One of hamming (hm), global (gb) or anchored (an).
cost_matrix	A custom cost matrix for use with the "global" or "anchored" distance metrics. See details.
gap_cost	The cost of a gap for use with the "global" or "anchored" distance metrics. See details.
gap_open_cost	The cost of a gap opening. See details.
nthreads	The number of threads to use for parallel computation.
show_progress	Whether to show a progress bar.

## Details

This function calculates pairwise distances based on Hamming, Levenshtein or Anchored algorithms. *query* and *target* must be the same length.

Three types of distance metrics are supported, based on the form of alignment performed. These are: Hamming, Global (Levenshtein) and Anchored.

An anchored alignment is a form of semi-global alignment, where the query sequence is "anchored" (global) to the beginning of both the query and target sequences, but is semi-global in that the end of either the query sequence or target sequence (but not both) can be unaligned. This type of alignment is sometimes called an "extension" alignment in literature.

In contrast a global alignment must align the entire query and target sequences. When mismatch and indel costs are equal to 1, this is also known as the Levenshtein distance.

By default, if mode == "global" or "anchored", all mismatches and indels are given a cost of 1. However, you can define your own distance metric by setting the *cost\_matrix* and *gap* parameters. The *cost\_matrix* is a strictly positive square integer matrix and should include all characters in query and target as column- and rownames. To set the cost of a gap (insertion or deletion) you can include a row and column named "gap" in the *cost\_matrix* *OR* set the *gap\_cost* parameter (a single positive integer). Similarly, the affine gap alignment can be set by including a row and column named "gap\_open" in the *cost\_matrix* *OR* setting the *gap\_open\_cost* parameter (a single positive integer). If affine alignment is used, the cost of a gap is defined as:  $TOTAL\_GAP\_COST = gap\_open\_cost + (gap\_cost * gap\_length)$ .

If mode == "hamming" all alignment parameters are ignored; mismatch is given a distance of 1 and gaps are not allowed.

## Value

The output of this function is a vector of distances. If mode == "anchored" then the output also includes attributes "query\_size" and "target\_size" which are vectors containing the lengths of the query and target sequences that are aligned.

## Examples

```
dist_pairwise(c("ACGT", "AAAA"), c("ACG", "ACGT"), mode = "global")
```

---

 dist\_search

*Distance search for similar sequences*


---

### Description

Find similar sequences within a distance threshold

### Usage

```
dist_search(
  query,
  target,
  max_distance = NULL,
  max_fraction = NULL,
  mode = "levenshtein",
  cost_matrix = NULL,
  gap_cost = NULL,
  gap_open_cost = NULL,
  tree_class = "RadixTree",
  nthreads = 1,
  show_progress = FALSE
)
```

### Arguments

query	A character vector of query sequences.
target	A character vector of target sequences.
max_distance	how far to search in units of absolute distance. Can be a single value or a vector. Mutually exclusive with max_fraction.
max_fraction	how far to search in units of relative distance to each query sequence length. Can be a single value or a vector. Mutually exclusive with max_distance.
mode	The distance metric to use. One of hamming (hm), global (gb) or anchored (an).
cost_matrix	A custom cost matrix for use with the "global" or "anchored" distance metrics. See details.
gap_cost	The cost of a gap for use with the "global" or "anchored" distance metrics. See details.
gap_open_cost	The cost of a gap opening. See details.
tree_class	Which R6 class to use. Either RadixTree or RadixForest (default: RadixTree)
nthreads	The number of threads to use for parallel computation.
show_progress	Whether to show a progress bar.

## Details

This function finds all sequences in *target* that are within a distance threshold of any sequence in *query*. This function uses either a RadixTree or RadixForest to store *target* sequences. See the R6 class documentation for additional details.

Three types of distance metrics are supported, based on the form of alignment performed. These are: Hamming, Global (Levenshtein) and Anchored.

An anchored alignment is a form of semi-global alignment, where the query sequence is "anchored" (global) to the beginning of both the query and target sequences, but is semi-global in that the end of either the query sequence or target sequence (but not both) can be unaligned. This type of alignment is sometimes called an "extension" alignment in literature.

In contrast a global alignment must align the entire query and target sequences. When mismatch and indel costs are equal to 1, this is also known as the Levenshtein distance.

By default, if mode == "global" or "anchored", all mismatches and indels are given a cost of 1. However, you can define your own distance metric by setting the *cost\_matrix* and *gap* parameters. The *cost\_matrix* is a strictly positive square integer matrix and should include all characters in query and target as column- and rownames. To set the cost of a gap (insertion or deletion) you can include a row and column named "gap" in the *cost\_matrix* OR set the *gap\_cost* parameter (a single positive integer). Similarly, the affine gap alignment can be set by including a row and column named "gap\_open" in the *cost\_matrix* OR setting the *gap\_open\_cost* parameter (a single positive integer). If affine alignment is used, the cost of a gap is defined as:  $TOTAL\_GAP\_COST = gap\_open\_cost + (gap\_cost * gap\_length)$ .

If mode == "hamming" all alignment parameters are ignored; mismatch is given a distance of 1 and gaps are not allowed.

## Value

The output is a data.frame of all matches with columns "query" and "target". For anchored searches, the output also includes attributes "query\_size" and "target\_size" which are vectors containing the portion of the query and target sequences that are aligned.

## Examples

```
dist_search(c("ACGT", "AAAA"), c("ACG", "ACGT"), max_distance = 1, mode = "levenshtein")
```

---

generate\_cost\_matrix *Generate a simple cost matrix*

---

## Description

Generate a cost matrix for use with the search method

**Usage**

```
generate_cost_matrix(
    charset,
    match = 0L,
    mismatch = 1L,
    gap = NULL,
    gap_open = NULL
)
```

**Arguments**

charset	A string representing all possible characters in both query and target sequences (e.g. "ACGT")
match	The cost of a match
mismatch	The cost of a mismatch
gap	The cost of a gap or NULL if this parameter will be set later.
gap_open	The cost of a gap opening or NULL. If this parameter is set, gap must also be set.

**Value**

A cost matrix

**Examples**

```
generate_cost_matrix("ACGT", match = 0, mismatch = 1)
```

---

RadixForest

*RadixForest*


---

**Description**

Radix Forest class implementation

**Details**

The RadixForest class is a specialization of the RadixTree implementation. Instead of putting sequences into a single tree, the RadixForest class puts sequences into separate trees based on sequence length. This allows for faster searching of similar sequences based on Hamming or Levenshtein distance metrics. Unlike the RadixTree class, the RadixForest class does not support anchored searches or a custom cost matrix. See *RadixTree* for additional details.

**Public fields**

forest\_pointer Map of sequence length to RadixTree

char\_counter\_pointer Character count data for the purpose of validating input

**Methods****Public methods:**

- `RadixForest$new()`
- `RadixForest$show()`
- `RadixForest$to_string()`
- `RadixForest$graph()`
- `RadixForest$to_vector()`
- `RadixForest$size()`
- `RadixForest$insert()`
- `RadixForest$erase()`
- `RadixForest$find()`
- `RadixForest$prefix_search()`
- `RadixForest$search()`
- `RadixForest$validate()`

**Method** `new()`: Create a new RadixForest object

*Usage:*

```
RadixForest$new(sequences = NULL)
```

*Arguments:*

`sequences` A character vector of sequences to insert into the forest

**Method** `show()`: Print the forest to screen

*Usage:*

```
RadixForest$show()
```

**Method** `to_string()`: Print the forest to a string

*Usage:*

```
RadixForest$to_string()
```

**Method** `graph()`: Plot of the forest using igraph

*Usage:*

```
RadixForest$graph(depth = -1, root_label = "root", plot = TRUE)
```

*Arguments:*

`depth` The tree depth to plot for each tree in the forest.

`root_label` The label of the root node(s) in the plot.

`plot` Whether to create a plot or return the data used to generate the plot.

*Returns:* A data frame of parent-child relationships used to generate the igraph plot OR a ggplot2 object

**Method** `to_vector()`: Output all sequences held by the forest as a character vector

*Usage:*

```
RadixForest$to_vector()
```

*Returns:* A character vector of all sequences contained in the forest.

**Method** `size()`: Output the size of the forest (i.e. how many sequences are contained)

*Usage:*

```
RadixForest$size()
```

*Returns:* The size of the forest

**Method** `insert()`: Insert new sequences into the forest

*Usage:*

```
RadixForest$insert(sequences)
```

*Arguments:*

`sequences` A character vector of sequences to insert into the forest

*Returns:* A logical vector indicating whether the sequence was inserted (TRUE) or already existing in the forest (FALSE)

**Method** `erase()`: Erase sequences from the forest

*Usage:*

```
RadixForest$erase(sequences)
```

*Arguments:*

`sequences` A character vector of sequences to erase from the forest

*Returns:* A logical vector indicating whether the sequence was erased (TRUE) or not found in the forest (FALSE)

**Method** `find()`: Find sequences in the forest

*Usage:*

```
RadixForest$find(query)
```

*Arguments:*

`query` A character vector of sequences to find in the forest

*Returns:* A logical vector indicating whether the sequence was found (TRUE) or not found in the forest (FALSE)

**Method** `prefix_search()`: Search for sequences in the forest that start with a specified prefix. E.g.: a query of "CAR" will find "CART", "CARBON", "CARROT", etc. but not "CATS".

*Usage:*

```
RadixForest$prefix_search(query)
```

*Arguments:*

`query` A character vector of sequences to search for in the forest

*Returns:* A data frame of all matches with columns "query" and "target".

**Method** `search()`: Search for sequences in the forest that are with a specified distance metric to a specified query.

*Usage:*

```
RadixForest$search(
  query,
  max_distance = NULL,
  max_fraction = NULL,
  mode = "levenshtein",
  nthreads = 1,
  show_progress = FALSE
)
```

*Arguments:*

*query* A character vector of query sequences.

*max\_distance* how far to search in units of absolute distance. Can be a single value or a vector.  
Mutually exclusive with *max\_fraction*.

*max\_fraction* how far to search in units of relative distance to each query sequence length.  
Can be a single value or a vector. Mutually exclusive with *max\_distance*.

*mode* The distance metric to use. One of hamming (hm), global (gb) or anchored (an).

*nthreads* The number of threads to use for parallel computation.

*show\_progress* Whether to show a progress bar.

*Returns:* The output is a data.frame of all matches with columns "query" and "target".

**Method** `validate()`: Validate the forest

*Usage:*

```
RadixForest$validate()
```

*Returns:* A logical indicating whether the forest is valid (TRUE) or not (FALSE). This is mostly an internal function for debugging purposes and should always return TRUE.

**Examples**

```
forest <- RadixForest$new()
forest$insert(c("ACGT", "AAAA"))
forest$erase("AAAA")
forest$search("ACG", max_distance = 1, mode = "levenshtein")
# query target distance
# 1   ACG   ACGT       1

forest$search("ACG", max_distance = 1, mode = "hamming")
# query target distance
# <0 rows> (or 0-length row.names)
```

---

 RadixTree

*RadixTree*


---

**Description**

Radix Tree (trie) class implementation

## Details

The RadixTree class is a trie implementation. The primary usage is to be able to search of similar sequences based on a dynamic programming framework. This can be done using the *search* method which searches for similar sequences based on the Global, Anchored or Hamming distance metrics.

Three types of distance metrics are supported, based on the form of alignment performed. These are: Hamming, Global (Levenshtein) and Anchored.

An anchored alignment is a form of semi-global alignment, where the query sequence is "anchored" (global) to the beginning of both the query and target sequences, but is semi-global in that the end of either the query sequence or target sequence (but not both) can be unaligned. This type of alignment is sometimes called an "extension" alignment in literature.

In contrast a global alignment must align the entire query and target sequences. When mismatch and indel costs are equal to 1, this is also known as the Levenshtein distance.

By default, if mode == "global" or "anchored", all mismatches and indels are given a cost of 1. However, you can define your own distance metric by setting the *cost\_matrix* and *gap* parameters. The *cost\_matrix* is a strictly positive square integer matrix and should include all characters in query and target as column- and rownames. To set the cost of a gap (insertion or deletion) you can include a row and column named "gap" in the *cost\_matrix* *OR* set the *gap\_cost* parameter (a single positive integer). Similarly, the affine gap alignment can be set by including a row and column named "gap\_open" in the *cost\_matrix* *OR* setting the *gap\_open\_cost* parameter (a single positive integer). If affine alignment is used, the cost of a gap is defined as:  $TOTAL\_GAP\_COST = gap\_open\_cost + (gap\_cost * gap\_length)$ .

If mode == "hamming" all alignment parameters are ignored; mismatch is given a distance of 1 and gaps are not allowed.

## Public fields

*root\_pointer* Root of the RadixTree

*char\_counter\_pointer* Character count data for the purpose of validating input

## Methods

### Public methods:

- [RadixTree\\$new\(\)](#)
- [RadixTree\\$show\(\)](#)
- [RadixTree\\$to\\_string\(\)](#)
- [RadixTree\\$graph\(\)](#)
- [RadixTree\\$to\\_vector\(\)](#)
- [RadixTree\\$size\(\)](#)
- [RadixTree\\$insert\(\)](#)
- [RadixTree\\$erase\(\)](#)
- [RadixTree\\$find\(\)](#)
- [RadixTree\\$prefix\\_search\(\)](#)
- [RadixTree\\$search\(\)](#)
- [RadixTree\\$validate\(\)](#)

**Method new():** Create a new RadixTree object

*Usage:*

```
RadixTree$new(sequences = NULL)
```

*Arguments:*

sequences A character vector of sequences to insert into the tree

**Method show():** Print the tree to screen

*Usage:*

```
RadixTree$show()
```

**Method to\_string():** Print the tree to a string

*Usage:*

```
RadixTree$to_string()
```

*Returns:* A string representation of the tree

**Method graph():** Plot of the tree using igraph (needs to be installed separately)

*Usage:*

```
RadixTree$graph(depth = -1, root_label = "root", plot = TRUE)
```

*Arguments:*

depth The tree depth to plot. If -1 (default), plot the entire tree.

root\_label The label of the root node in the plot.

plot Whether to create a plot or return the data used to generate the plot.

*Returns:* A data frame of parent-child relationships used to generate the igraph plot OR a ggplot2 object

**Method to\_vector():** Output all sequences held by the tree as a character vector

*Usage:*

```
RadixTree$to_vector()
```

*Returns:* A character vector of all sequences contained in the tree. Return order is not guaranteed.

**Method size():** Output the size of the tree (i.e. how many sequences are contained)

*Usage:*

```
RadixTree$size()
```

*Returns:* The size of the tree

**Method insert():** Insert new sequences into the tree

*Usage:*

```
RadixTree$insert(sequences)
```

*Arguments:*

sequences A character vector of sequences to insert into the tree

*Returns:* A logical vector indicating whether the sequence was inserted (TRUE) or already existing in the tree (FALSE)

**Method** `erase()`: Erase sequences from the tree

*Usage:*

```
RadixTree$erase(sequences)
```

*Arguments:*

`sequences` A character vector of sequences to erase from the tree

*Returns:* A logical vector indicating whether the sequence was erased (TRUE) or not found in the tree (FALSE)

**Method** `find()`: Find sequences in the tree

*Usage:*

```
RadixTree$find(query)
```

*Arguments:*

`query` A character vector of sequences to find in the tree

*Returns:* A logical vector indicating whether the sequence was found (TRUE) or not found in the tree (FALSE)

**Method** `prefix_search()`: Search for sequences in the tree that start with a specified prefix. E.g.: a query of "CAR" will find "CART", "CARBON", "CARROT", etc. but not "CATS".

*Usage:*

```
RadixTree$prefix_search(query)
```

*Arguments:*

`query` A character vector of sequences to search for in the tree

*Returns:* A data frame of all matches with columns "query" and "target".

**Method** `search()`: Search for sequences in the tree that are with a specified distance metric to a specified query.

*Usage:*

```
RadixTree$search(
  query,
  max_distance = NULL,
  max_fraction = NULL,
  mode = "levenshtein",
  cost_matrix = NULL,
  gap_cost = NULL,
  gap_open_cost = NULL,
  nthreads = 1,
  show_progress = FALSE
)
```

*Arguments:*

`query` A character vector of query sequences.

`max_distance` how far to search in units of absolute distance. Can be a single value or a vector. Mutually exclusive with `max_fraction`.

`max_fraction` how far to search in units of relative distance to each query sequence length. Can be a single value or a vector. Mutually exclusive with `max_distance`.

`mode` The distance metric to use. One of hamming (hm), global (gb) or anchored (an).  
`cost_matrix` A custom cost matrix for use with the "global" or "anchored" distance metrics. See details.  
`gap_cost` The cost of a gap for use with the "global" or "anchored" distance metrics. See details.  
`gap_open_cost` The cost of a gap opening. See details.  
`nthreads` The number of threads to use for parallel computation.  
`show_progress` Whether to show a progress bar.

*Returns:* The output is a data.frame of all matches with columns "query" and "target". For anchored searches, the output also includes attributes "query\_size" and "target\_size" which are vectors containing the portion of the query and target sequences that are aligned.

**Method** `validate()`: Validate the tree

*Usage:*

```
RadixTree$validate()
```

*Returns:* A logical indicating whether the tree is valid (TRUE) or not (FALSE). This is mostly an internal function for debugging purposes and should always return TRUE.

### See Also

[https://en.wikipedia.org/wiki/Radix\\_tree](https://en.wikipedia.org/wiki/Radix_tree)

### Examples

```
tree <- RadixTree$new()
tree$insert(c("ACGT", "AAAA"))
tree$erase("AAAA")
tree$search("ACG", max_distance = 1, mode = "levenshtein")
# query target distance
# 1   ACG   ACGT         1

tree$search("ACG", max_distance = 1, mode = "hamming")
# query target distance
# <0 rows> (or 0-length row.names)
```

# Index

## \* datasets

covid\_cdr3, [2](#)

covid\_cdr3, [2](#)

dist\_matrix, [3](#)

dist\_pairwise, [4](#)

dist\_search, [6](#)

generate\_cost\_matrix, [7](#)

RadixForest, [8](#)

RadixTree, [11](#)