Package ‘baseq’

March 11, 2026

Title Basic Sequence Processing Tool for Biological Data

Version 0.2.0

Description Primarily created as an easy and understanding way to do basic sequences surrounding the central dogma of molecular biology.

License GPL-3

URL https://github.com/ambuvjyn/baseq

BugReports https://github.com/ambuvjyn/baseq/issues

Encoding UTF-8

RoxygenNote 7.3.3

Imports ggplot2

Suggests testthat (>= 3.0.0), rmarkdown, knitr, Biostrings

VignetteBuilder knitr

Config/testthat/edition 3

LazyData true

NeedsCompilation no

Author Ambu Vijayan [aut, cre] (ORCID: <https://orcid.org/0000-0001-8924-5685>),
J. Sreekumar [aut] (ORCID: <https://orcid.org/0000-0002-4253-6378>, Principal Scientist, ICAR - Central Tuber Crops Research Institute)

Maintainer Ambu Vijayan <ambuvjyn@gmail.com>

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2026-03-11 22:30:18 UTC

Contents

as_baseq_dna 3
as_baseq_rna 3
as_Biostrings 4

calculate_assembly_stats . . . 4
calculate_charge . . . 5
calculate_codon_usage . . . 5
calculate_identity . . . 6
calculate_mw . . . 6
calculate_pi . . . 7
calculate_tm . . . 7
clean_file . . . 8
clean_seq . . . 8
count_bases . . . 9
count_kmers . . . 9
count_pattern . . . 10
dna_to_protein . . . 10
dna_to_rna . . . 11
fastq_to_fasta . . . 11
filter_fastq_quality . . . 12
find_cpg_islands . . . 12
find_longest_orf . . . 13
gc_content . . . 13
get_genetic_code . . . 14
plot_aa_composition . . . 14
plot_dotplot . . . 15
plot_gc_skew . . . 15
plot_hydrophobicity . . . 16
read_seq . . . 16
reverse_translate . . . 17
rev_comp . . . 17
rna_to_dna . . . 18
rna_to_protein . . . 18
sars_fragment . . . 19
search_motif . . . 19
shuffle_sequence . . . 20
simulate_digestion . . . 20
simulate_fasta . . . 21
simulate_fastq . . . 21
simulate_pcr . . . 22
simulate_sequence . . . 23
summarize_fasta . . . 23
translate . . . 24
write_seq . . . 24

Index **25**

as_baseq_dna *S3 DNA Class*

Description

Creates an S3 object of class baseq_dna.

Usage

```
as_baseq_dna(s)
```

Arguments

s	A character string containing the sequence

Value

A baseq_dna object

as_baseq_rna *S3 RNA Class*

Description

Creates an S3 object of class baseq_rna.

Usage

```
as_baseq_rna(s)
```

Arguments

s	A character string containing the sequence

Value

A baseq_rna object

as_Biostrings *Bioconductor Bridge*

Description

Converts baseq sequences to Biostrings format.

Usage

```
as_Biostrings(s)
```

Arguments

s	A character vector or list of sequences

Value

A DNAStringSet object

calculate_assembly_stats *Assembly Stats*

Description

Computes N50, L50, and other assembly statistics.

Usage

```
calculate_assembly_stats(seqs)
```

Arguments

seqs	A character vector or list of sequences (contigs)

Value

A named numeric vector of statistics

Examples

```
contigs <- c("ATGC", "ATGCATGC", "ATGCATGCATGC")
calculate_assembly_stats(contigs)
```

calculate_charge *Protein Net Charge*

Description

Calculates the net electrical charge of a protein at a given pH.

Usage

```
calculate_charge(s, ph = 7.4)
```

Arguments

s	A character string containing the protein sequence
ph	Numeric pH value (default: 7.4)

Value

Numeric net charge

calculate_codon_usage *Codon Usage RSCU*

Description

Calculates Relative Synonymous Codon Usage (RSCU).

Usage

```
calculate_codon_usage(s)
```

Arguments

s	A character string containing the coding DNA sequence

Value

A dataframe with codon statistics

Examples

```
data(sars_fragment)
calculate_codon_usage(sars_fragment)
```

calculate_identity *Sequence Identity*

Description

Compares two sequences of equal length.

Usage

```
calculate_identity(s1, s2)
```

Arguments

s1	First sequence
s2	Second sequence

Value

A list with Identity percentage and Hamming Distance

Examples

```
calculate_identity("ATGC", "ATGG")
```

calculate_mw *Protein MW*

Description

Calculates the molecular weight of a protein sequence.

Usage

```
calculate_mw(s)
```

Arguments

s	A character string containing the protein sequence

Value

Numeric molecular weight in Daltons

calculate_pi *Protein pI*

Description

Estimates the isoelectric point of a protein sequence.

Usage

```
calculate_pi(s)
```

Arguments

s	A character string containing the protein sequence

Value

Numeric pI value

calculate_tm *Primer Tm*

Description

Calculates the melting temperature of a primer sequence.

Usage

```
calculate_tm(s, salt = 50)
```

Arguments

s	A character string containing the sequence
salt	Numeric salt concentration in mM (default: 50)

Value

Numeric Tm in Celsius

clean_file *Batch File Cleaner*

Description

Cleans all sequences in a FASTA or FASTQ file.

Usage

```
clean_file(input_file, type = "auto", output_dir = "")
```

Arguments

input_file	Path to input file
type	Sequence type ("DNA", "RNA", or "auto")
output_dir	Optional output directory

Value

Path to the cleaned file

clean_seq *Universal Sequence Cleaner*

Description

Removes non-standard characters from DNA or RNA sequences.

Usage

```
clean_seq(sequence, type = "auto")
```

Arguments

sequence	A character string containing the sequence
type	A string "DNA", "RNA", or "auto"

Value

A character string of the cleaned sequence

count_bases *Count Bases*

Description

Returns a frequency table of the bases in a sequence.

Usage

```
count_bases(s)
```

Arguments

s	A character string containing the sequence

Value

A table object with base counts

Examples

```
data(sars_fragment)
count_bases(sars_fragment)
```

count_kmers *K-mer Counting*

Description

Counts all possible substrings of length k.

Usage

```
count_kmers(s, k = 3)
```

Arguments

s	A character string containing the sequence
k	Integer length of k-mer

Value

A table of k-mer counts

Examples

```
data(sars_fragment)
count_kmers(sars_fragment, k = 3)
```

count_pattern *Count Pattern*

Description

Counts the occurrences of a specific pattern in a sequence.

Usage

```
count_pattern(s, p)
```

Arguments

s	A character string containing the sequence
p	A character string containing the pattern to count

Value

Integer count of occurrences

Examples

```
data(sars_fragment)
count_pattern(sars_fragment, "ATTA")
```

dna_to_protein *Translate DNA to Protein*

Description

Translates a DNA sequence into protein in all 6 reading frames.

Usage

```
dna_to_protein(s, table = 1)
```

Arguments

s	A character string containing the DNA sequence
table	Integer indicating the NCBI genetic code table (default: 1)

Value

A list of translated protein sequences

dna_to_rna *DNA to RNA*

Description

Transcribes a DNA sequence into RNA.

Usage

```
dna_to_rna(s)
```

Arguments

s	A character string containing the DNA sequence

Value

A character string of the RNA sequence

fastq_to_fasta *Convert FASTQ to FASTA*

Description

Converts a FASTQ file to FASTA format.

Usage

```
fastq_to_fasta(fastq_file)
```

Arguments

fastq_file	Path to input FASTQ

Value

Path to output FASTA

filter_fastq_quality *Quality Filter FASTQ*

Description

Filters FASTQ reads based on average quality score.

Usage

```
filter_fastq_quality(
  input_file,
  output_file,
  min_avg_quality = 20,
  phred_offset = 33
)
```

Arguments

input_file	Path to input FASTQ
output_file	Path to output FASTQ
min_avg_quality	Minimum average Phred score (default: 20)
phred_offset	Phred offset (default: 33)

find_cpg_islands *CpG Island Detection*

Description

Identifies candidate CpG islands in a DNA sequence.

Usage

```
find_cpg_islands(s, window = 200)
```

Arguments

s	A character string containing the DNA sequence
window	Sliding window size (default: 200)

Value

A dataframe with start and end positions

find_longest_orf *Find Longest ORF*

Description

Scans a DNA sequence in all 6 reading frames to find the longest open reading frame.

Usage

```
find_longest_orf(s)
```

Arguments

s	A character string containing the DNA sequence

Value

A character string of the longest translated protein sequence

gc_content *GC Content*

Description

Calculates the percentage of G and C bases in a DNA sequence.

Usage

```
gc_content(s)
```

Arguments

s	A character string containing the sequence

Value

Numeric percentage of GC content

Examples

```
data(sars_fragment)
gc_content(sars_fragment)
```

get_genetic_code *Get Genetic Code*

Description

Returns a mapping of codons to amino acids.

Usage

```
get_genetic_code(table = 1)
```

Arguments

table	Integer NCBI genetic code table index

Value

A named character vector

plot_aa_composition *Plot AA Composition*

Description

Visualizes the amino acid composition categorized by biochemical properties.

Usage

```
plot_aa_composition(s)
```

Arguments

s	A character string containing the protein sequence

Value

A ggplot object

Examples

```
prot <- "MKFLVLALAL"
plot_aa_composition(prot)
```

plot_dotplot *Plot Dot Plot*

Description

Generates a dot plot comparison of two sequences.

Usage

```
plot_dotplot(s1, s2, window = 1)
```

Arguments

s1	First sequence
s2	Second sequence
window	Integer word size for matching (default: 1)

Value

A ggplot object

Examples

```
s1 <- "ATGCATGCATGC"
s2 <- "ATGCGTGCATGC"
plot_dotplot(s1, s2, window = 3)
```

plot_gc_skew *Plot GC Skew*

Description

Generates a sliding window plot of GC skew (G-C)/(G+C).

Usage

```
plot_gc_skew(s, window = 100)
```

Arguments

s	A character string containing the DNA sequence
window	Integer window size (default: 100)

Value

A ggplot object

Examples

```
data(sars_fragment)
plot_gc_skew(sars_fragment, window = 100)
```

---

plot_hydrophobicity *Plot Hydrophobicity*

Description

Generates a sliding window plot of protein hydrophobicity using the Kyte-Doolittle scale.

Usage

```
plot_hydrophobicity(s, window = 9)
```

Arguments

s	A character string containing the protein sequence
window	Integer window size (default: 9)

Value

A ggplot object

Examples

```
prot <- "MKFLVLALAL"
plot_hydrophobicity(prot, window = 3)
```

---

read_seq *Universal Sequence Reader*

Description

Reads a FASTA or FASTQ file and returns it as a dataframe or list.

Usage

```
read_seq(file, format = "df")
```

Arguments

file	Path to the input sequence file
format	A string indicating "df" (dataframe) or "list" (default: "df")

Value

A dataframe or list of the sequence data.

reverse_translate *Reverse Translation*

Description

Converts a protein sequence back into DNA using common codons.

Usage

```
reverse_translate(s)
```

Arguments

s	A character string containing the protein sequence

Value

A character string of the resulting DNA sequence

rev_comp *Universal Reverse Complement*

Description

Generates the reverse complement of a DNA or RNA sequence.

Usage

```
rev_comp(sequence)
```

Arguments

sequence	A character string containing the sequence

Value

A character string of the reverse complement

rna_to_dna *RNA to DNA*

Description

Reverse transcribes an RNA sequence into DNA.

Usage

```
rna_to_dna(s)
```

Arguments

s	A character string containing the RNA sequence

Value

A character string of the DNA sequence

rna_to_protein *Translate RNA to Protein*

Description

Translates an RNA sequence into protein in all 6 reading frames.

Usage

```
rna_to_protein(s, table = 1)
```

Arguments

s	A character string containing the RNA sequence
table	Integer indicating the NCBI genetic code table (default: 1)

Value

A list of translated protein sequences

sars_fragment *SARS-CoV-2 Genome Fragment*

Description

A small fragment of the SARS-CoV-2 genome used for examples and testing.

Usage

```
sars_fragment
```

Format

A character string.

Source

NCBI GenBank

search_motif *Motif Searching*

Description

Finds all occurrences of a motif in a sequence.

Usage

```
search_motif(s, p)
```

Arguments

s	A character string containing the sequence
p	A character string containing the motif (regex)

Value

A dataframe with the Start, End, and Match string

shuffle_sequence *Shuffle Sequence*

Description

Randomly permutes the characters of a sequence.

Usage

```
shuffle_sequence(s)
```

Arguments

s	A character string containing the sequence

Value

A character string of the shuffled sequence

simulate_digestion *Virtual Digestion*

Description

Simulates restriction enzyme digestion.

Usage

```
simulate_digestion(s, p)
```

Arguments

s	A character string containing the DNA sequence
p	A character string containing the restriction site (regex)

Value

A numeric vector of fragment lengths

simulate_fasta *Simulate FASTA File*

Description

Generates a dummy FASTA dataset.

Usage

```
simulate_fasta(n_seq = 5, seq_len = 100, gc = NULL, type = "DNA", file = NULL)
```

Arguments

n_seq	Number of sequences
seq_len	Length of each sequence
gc	Target GC content
type	"DNA" or "RNA"
file	Optional file path to save

Value

A dataframe of simulated sequences

simulate_fastq *Simulate FASTQ File*

Description

Generates a dummy FASTQ dataset.

Usage

```
simulate_fastq(
  n_reads = 5,
  read_len = 100,
  gc = NULL,
  type = "DNA",
  file = NULL
)
```

Arguments

`n_reads`	Number of reads
`read_len`	Length of each read
`gc`	Target GC content
`type`	"DNA" or "RNA"
`file`	Optional file path to save

Value

A dataframe of simulated reads

`simulate_pcr` *PCR Simulator*

Description

Simulates a PCR reaction and predicts amplicon sizes.

Usage

```
simulate_pcr(template, fwd, rev_p)
```

Arguments

`template`	A character string containing the DNA template
`fwd`	A character string of the forward primer
`rev_p`	A character string of the reverse primer

Value

A numeric vector of amplicon sizes

simulate_sequence *Simulate Sequence*

Description

Generates a random DNA or RNA sequence.

Usage

```
simulate_sequence(len, gc = NULL, type = "DNA")
```

Arguments

len	Integer length of the sequence
gc	Numeric target GC content (0 to 1)
type	"DNA" or "RNA"

Value

A character string of the simulated sequence

summarize_fasta *FASTA Summary*

Description

Generates a comprehensive summary of a multi-FASTA file.

Usage

```
summarize_fasta(file)
```

Arguments

file	Path to the FASTA file

Value

A summary dataframe

Examples

```
# summarize_fasta("path/to/my.fasta")
```

translate — *Generic Translate*

Description

Generic function to translate DNA or RNA to protein.

Usage

```
translate(x, ...)
```

Arguments

x	A baseq_dna or baseq_rna object
...	Additional arguments

Value

A list of translated sequences

write_seq — *Universal Sequence Writer*

Description

Writes a sequence object (dataframe or list) to a FASTA or FASTQ file.

Usage

```
write_seq(x, file)
```

Arguments

x	A sequence object (dataframe or list)
file	Path to the output sequence file

Value

Invisible TRUE

Index

∗ **datasets**
sars_fragment, 19

as_baseq_dna, 3
as_baseq_rna, 3
as_Biostrings, 4

calculate_assembly_stats, 4
calculate_charge, 5
calculate_codon_usage, 5
calculate_identity, 6
calculate_mw, 6
calculate_pi, 7
calculate_tm, 7
clean_file, 8
clean_seq, 8
count_bases, 9
count_kmers, 9
count_pattern, 10

dna_to_protein, 10
dna_to_rna, 11

fastq_to_fasta, 11
filter_fastq_quality, 12
find_cpg_islands, 12
find_longest_orf, 13

gc_content, 13
get_genetic_code, 14

plot_aa_composition, 14
plot_dotplot, 15
plot_gc_skew, 15
plot_hydrophobicity, 16

read_seq, 16
rev_comp, 17
reverse_translate, 17
rna_to_dna, 18
rna_to_protein, 18

sars_fragment, 19
search_motif, 19
shuffle_sequence, 20
simulate_digestion, 20
simulate_fasta, 21
simulate_fastq, 21
simulate_pcr, 22
simulate_sequence, 23
summarize_fasta, 23

translate, 24

write_seq, 24