

BACSEQUER DOKUMENTACE

ÚVOD

BacSequer je nástroj určený pro simulaci čtení bakteriální RNA-Seq. Tento dokument obsahuje průvodce použitím skriptu, včetně detailního popisu funkcí a příkladů použití.

VSTUPNÍ A VÝSTUPNÍ FORMÁTY

Simulátor využívá jako vstup formáty FASTA a GTF/GFF3.

Výstupními formáty pak může být FASTA nebo FASTQ.

POPIS FUNKCÍ

READ_FASTA

Funkce *read_fasta* slouží k načtení genetických sekvencí z FASTA souboru. Funkce vrátí slovník, kde klíče představují identifikátory sekvencí a hodnoty jsou samotné sekvence.

PARAMETRY

file_path (str): Cesta k FASTA souboru.

VÝSTUP

seqs (dict): Slovník obsahující identifikátory a sekvence.

EXTRACT_SEQUENCES

Funkce *extract_sequences* extrahuje sekvence ze souboru FASTA na základě anotací v souboru GTF/GFF3. Jako pomocné funkce využívá *read_fasta_ll*, *parse_gtf_gff* a *extract_id_from_attributes*. Tato funkce je základem pro další zpracování v rámci simulací.

PARAMETRY

fasta_file (str): Cesta k souboru ve formátu FASTA.

gtf_gff_file (str): Cesta k souboru GTF/GFF3 s anotacemi.

VÝSTUP

extracted_seqs (dict): Slovník, kde klíče jsou ID anotací a hodnoty jsou extrahované sekvence.

CALCULATE_GC_CONTENT

Funkce spočítá procentuální obsah GC pro načtenou sekвени.

PARAMETRY

seq (str): Vstupní sekvence.

VÝSTUP

float: %GC obsah načtené sekvence.

ADJUST_GC_CONTENT

Upravuje obsah GC v čtení na požadovaný obsah GC při zachování jeho délky.

PARAMETRY

read (str): Sekvence, kterou chceme upravit.

gc_content (float): Požadovaný obsah GC (v rozsahu 0 až 1).

VÝSTUP

read (str): Upravená sekvence.

EXTRACT_OPERONS

Extrahuje operony a jejich umístění ze souboru GTF/GFF3.

PARAMETRY

file_path (str): Cesta k souboru.

VÝSTUP

operons (dict): Slovník s názvy operonů jako klíči a jejich umístěními jako hodnotami.

EXTRACT_RRNA

Funkce zpracovává soubor GTF/GFF3 za účelem extrakce anotací rRNA.

PARAMETRY

file_path (str): Cesta k souboru.

VÝSTUP

rrna_dict (dict): Slovník, kde klíče jsou seq_id s indexem a hodnoty jsou slovníky s klíči 'start' a 'end'.

EXTRACT_CDS

Funkce slouží k extrakci názvů a lokací kódujících oblastí GTF/GTFF3 souboru. Vybranou oblast lze navíc na začátku i na konci rozšířit o zadaný počet nukleotidů.

PARAMETRY

file_path (str): Cesta k souboru.

start_extension (int, optional): Počet nukleotidů přidaných na začátek CDS.

end_extension (int, optional): Počet nukleotidů přidaných na konec CDS.

VÝSTUP

cds_dict (dict): Slovník, kde klíče jsou identifikátory genů nebo transkriptů a hodnoty jsou seznamy tuple (start, end) kódujících oblastí.

CALCULATE_RRNA_PERCENTAGE

Spočítá procentuální zastoupení rRNA v sadě sekvencí.

PARAMETRY

fasta_path (str): Cesta k souboru FASTA.

gff_path (str): Cesta k souboru GTF/GFF3.

read_fasta_func (funkce): Funkce pro čtení souboru FASTA (dostupná v tomto programu - *read_fasta*).

PARSE_FOR_STRAND

Zpracovává soubor GTF nebo GFF3 a vrací slovník, kde jsou klíče názvy genů a hodnoty informace o vláknech ('+' nebo '-').

PARAMETRY

file_path (str): Cesta k souboru.

VÝSTUP

strand_info (dict): Slovník s názvy genů jako klíči a informacemi o vláknech jako hodnotami.

CREATE_LONG_MRNA

Vytvoří dlouhou mRNA sekvenci na základě lokací operonů.

PARAMETRY

seqs (dict): Slovník sekvencí, kde klíče jsou identifikátory genů a hodnoty jsou DNA sekvence.

operon_locations (dict): Slovník s lokacemi operonů, kde klíče jsou identifikátory operonů a hodnoty jsou seznamy identifikátorů genů v operonu.

VÝSTUP

long_mrna_sequences (dict): Slovník s dlouhými mRNA sekvencemi pro každý operon.

REVERSE_COMPLEMENT

Vrací reverzní komplementární sekvenci pro zadanou DNA sekvenci.

PARAMETRY

seq (str): DNA sekvence, která má být převedena na její reverzní komplement.

VÝSTUP

str: Reverzní komplementární sekvence k zadané DNA sekvenci.

SIMULATE_READS

Simuluje čtení na základě poskytnutých sekvencí.

Simulace čtení je provedena na základě slovníku *seqs*.

Pro úpravu procentuálního GC obsahu je potřeba zadat parametr *gc_content* v rozsahu od 0 do 1.

Pro simulaci pouze z kódujících oblastí je žádoucí zadat parametr *cds_locations*.

Parametr *operon_locations* umožňuje vytvořit polycistronní molekuly RNA a obohatit tak vstupní knihovnu.

Parametry *rrna_locations* a *rrna_percentage* slouží k doplnění vstupní knihovny o požadované procento rRNA a simulaci kontaminace.

Pomocí *strand_ori* a *strand_info* parametrů lze provést strand-specific simulaci.

PARAMETRY

seqs (dict): Slovník sekvencí, kde klíče jsou názvy sekvencí a hodnoty jsou sekvence. Slouží jako vstupní knihovna pro simulaci.

read_length (int): Délka čtení.

num_reads (int): Počet čtení, která mají být simulována pro každou sekvenci.

gc_content (float nebo None): Požadovaný obsah GC.

cds_locations (dict nebo None): Slovník s lokacemi kódujících oblastí.

operon_locations (dict nebo None): Slovník s lokacemi operonů.

rrna_locations (dict nebo None): Slovník s lokacemi rRNA.

rrna_percentage (float nebo None): Požadované procentuální zastoupení rRNA ve vstupní knihovně (v rozsahu 0 až 100).

strand_ori (str nebo None): Požadovaná strand orientation. (+ nebo -)

strand_info (dict nebo None): Informace o orientaci vláken pro každou sekvenci.

VÝSTUP

reads (list): Seznam čtení jako dvojic (název sekvence, čtení).

PHRED_TO_ASCII

Umožňuje převod Phred skóre do hodnot ASCII.

PARAMETRY

phred_score (int): Phred quality score (běžně v rozmezí 0 až 40).

VÝSTUP

str: Odpovídající hodnota ASCII.

WRITE_OUTPUT

Zapíše sekvence do souboru ve formátu FASTA nebo FASTQ. Jako pomocná funkce je použita *generate_quality_scores*.

Pro FASTQ formát je možné nastavit maximální a minimální kvalitu čtení, přičemž pokud tyto hodnoty nejsou zadány, jsou odhadnuty na základě platformy Illumina.

PARAMETRY

reads (list of tuples): Seznam sekvencí a jejich názvů. Každý prvek seznamu je tuple (seq_name, read), kde 'seq_name' je název sekvence a 'read' je samotná sekvence.

output_format (str): Formát výstupního souboru. Může být 'fasta' nebo 'fastq'.

output_file (str): Název výstupního souboru.

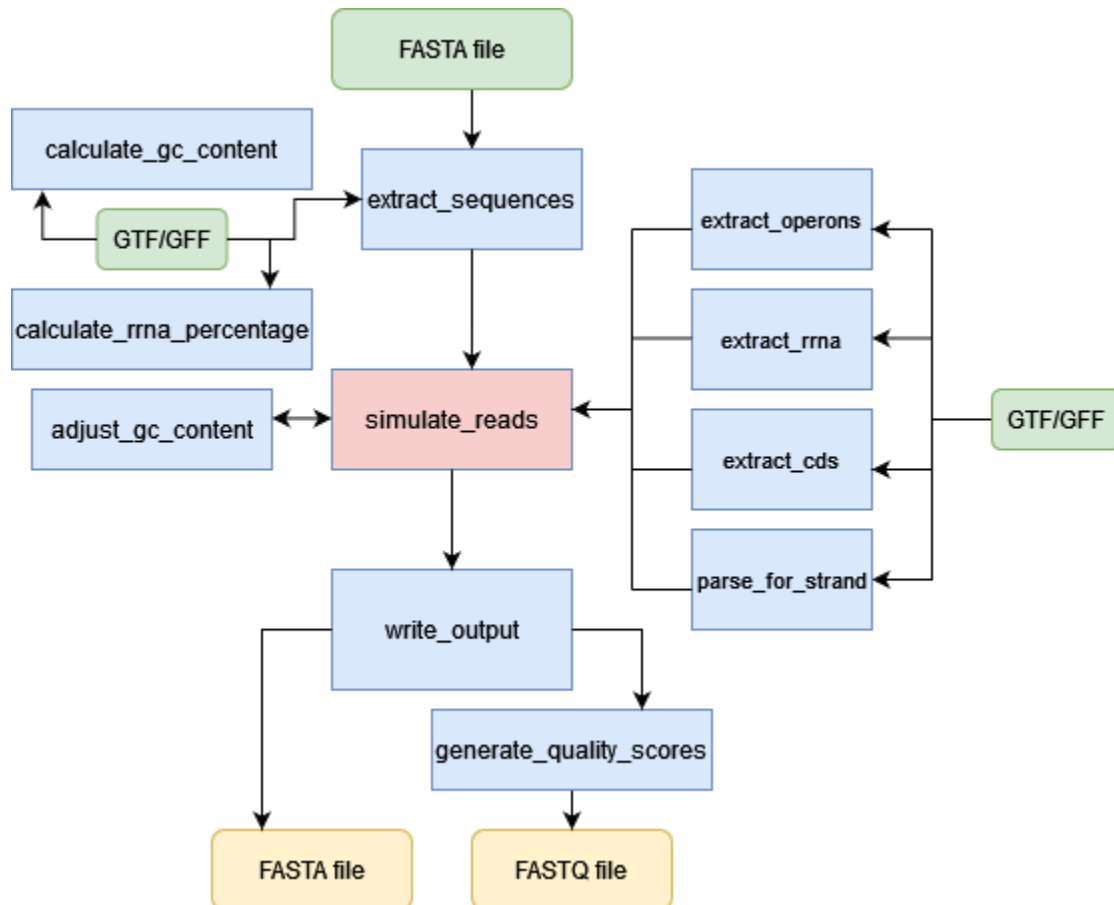
max_quality (int, optional): Maximální ASCII hodnota pro kvalitní skóre na začátku sekvence (pouze pro FASTQ).

min_quality (int, optional): Minimální ASCII hodnota pro kvalitní skóre na konci sekvence (pouze pro FASTQ).

VÝSTUP

Funkce vytvoří soubor s daným názvem a zapíše do něj sekvence.

VÝVOJOVÝ DIAGRAM SIMULÁTORU



SPUŠTĚNÍ SIMULÁTORU

Před spuštěním skriptu se ujistěte, že máte nainstalovanou potřebnou knihovnu:

pip: pip install biopython.

PŘÍKLADY VOLÁNÍ FUNKCÍ

```
fasta = "cesta_k_vasemu_fasta_souboru"
```

```
gff = "cesta_k_vasemu_gff_souboru"
```

EXTRAKCE SEKVENCÍ Z FASTA SOUBORU

ZÁKLADNÍ

```
seqs = read_fasta(fasta)
```

S POMOCÍ GTF/GFF3 SOUBORU

```
seqs = extract_sequences(fasta, gff)
```

ZÍSKÁNÍ LOKACÍ OPERONŮ

```
operon_locs = extract_operons(gff)
```

ZÍSKÁNÍ LOKACÍ KÓDUJÍCÍCH OBLASTÍ

ZÁKLADNÍ

```
cds_locs = extract_cds(gff)
```

S ROZŠÍŘENÍM

```
cds_locs = extract_cds(gff, 15, 18)
```

ZÍSKÁNÍ INFORMACÍ O STRAND ORIENTACI

```
strand = parse_for_strand(gff)
```

ZÍSKÁNÍ LOKACÍ RRNA OBLASTÍ

```
rrna_locs = extract_rrna(gff)
```


VÝPOČET PROCENTUÁLNÍHO ZASTOUPENÍ RRNA

```
rRNA_percentage = calculate_rrna_percentage(fasta, gff, read_fasta)

print(f"rRNA Percentage: {rRNA_percentage:.2f}%")
```

SIMULACE

ZÁKLADNÍ

```
reads = simulate_reads(seqs, 75, 100)
```

S VYUŽITÍM DALŠÍCH ROZŠÍŘENÍ

Zadání požadovaného GC obsahu, simulace z kódujících oblastí s kontaminací rRNA a zohledněním operonových oblastí, strand specifita pro dopředná vlákna.

```
reads = simulate_reads(seqs, 75, 100, gc_content=0.35, cds_locations=cds_locs,
operon_locations=operon_locs, rrna_locations=rrna_locs, rrna_percentage=30,
strand_ori="+", strand_info=strand)
```

VYTVOŘENÍ VÝSTUPNÍHO SOUBORU

FASTA FORMÁT

```
write_output(reads, output_format='fasta', output_file='output.fasta',
max_quality=None, min_quality=None)
```

FASTQ FORMÁT

```
write_output(reads, output_format='fastq', output_file='output.fastq',
max_quality=32, min_quality=28)
```

KOMPLETNÍ VOLÁNÍ

```
fasta = "cesta_k_vasemu_fasta_souboru"
gff = "cesta_k_vasemu_gff_souboru"
seqs = extract_sequences(fasta, gff)
operon_locs = extract_operons(gff)
cds_locs = extract_cds(gff, 15)
strand = parse_for_strand(gff)
rrna_locs = extract_rrna(gff)
reads = simulate_reads(seqs, 75, 100, gc_content=0.35, cds_locations=cds_locs,
operon_locations=operon_locs, rrna_locations=rrna_locs, rrna_percentage=30,
strand_ori="+", strand_info=strand)
write_output(reads, output_format='fastq', output_file='output.fastq',
max_quality=None, min_quality=None)
```