大規模データ解析入門1

- シークエンスのフィルタリング -

データの構造

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

1行 1データ

データの構造

配列名	配列本体
sequence_1	GGCANGTATNGCTGG
sequence_2	GCCGAGTTTAATTGC
sequence_3	GCCGCCCTGGCCGTA
sequence_4	NCNAACTTCGCCCCC
sequence_5	ATGGATATTCAGGAA
sequence_6	CCGCCAAATTGTAAT
sequence_7	AAGNAAGGCTGAACN
sequence_8	TTTTTTTTTTTTTTT
sequence_9	GTAGCTTCAGGTGTA
sequence_10	AAAGCATTGCGTATG

データの構造

タブ区切りテキスト

タブ文字

```
sequence_1 ¥t GGCANGTATNGCTGG
sequence_2 ¥t GCCGAGTTTAATTGC
sequence_3 ¥t GCCGCCCTGGCCGTA
sequence_4 ¥t NCNAACTTCGCCCCC
sequence_5 ¥t ATGGATATTCAGGAA
sequence_6 ¥t CCGCCAAATTGTAAT
sequence_7 ¥t AAGNAAGGCTGAACN
sequence_8 ¥t TTTTTTTTTTTTTTT
sequence_9 ¥t GTAGCTTCAGGTGTA
sequence_10 ¥t AAAGCATTGCGTATG
```

データの構造

タブ区切りテキスト

タブ文字　　改行文字

sequence_1	¥t	GGCANGTATNGCTGG	¥n
sequence_2	¥t	GCCGAGTTTAATTGC	¥n
sequence_3	¥t	GCCGCCCTGGCCGTA	¥n
sequence_4	¥t	NCNAACTTCGCCCCC	¥n
sequence_5	¥t	ATGGATATTCAGGAA	¥n
sequence_6	¥t	CCGCCAAATTGTAAT	¥n
sequence_7	¥t	AAGNAAGGCTGAACN	¥n
sequence_8	¥t	TTTTTTTTTTTTTTT	¥n
sequence_9	¥t	GTAGCTTCAGGTGTA	¥n
sequence_10	¥t	AAAGCATTGCGTATG	¥n

タブ区切りテキスト処理の基本

ファイルを開く

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

タブ区切りテキスト処理の基本

上から順番に一行ずつ読む

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

タブ区切りテキスト処理の基本

上から順番に一行ずつ読む

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

タブ区切りテキスト処理の基本

上から順番に一行ずつ読む

タブ区切りテキスト処理の基本

上から順番に一行ずつ読む

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

タブ区切りテキスト処理の基本

一行を分ける

改行文字を
取り除く

```
sequence_1 ¥t GGCANGTATNGCTGG ¥n
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

タブ区切りテキスト処理の基本

一行を分ける

改行文字を
取り除く

sequence_1 ¥t GGCANGTATTGCTGG ¥n

sequence_2 GCCGAGTTTAATTGC
sequence_3
sequence_4
sequence_5
sequence_6
sequence_7 AAGNAAGGCTGAACN
sequence_8 TTTTTTTTTTTTTTT
sequence_9 GTAGCTTCAGGTGTA
sequence_10 AAAGCATTGCGTATG

今回の演習では、

- 配列の長さを調べる
- Nの数を数える
- N含有率を計算する

タブ区切りテキスト処理の基本

N含有率10%以上

sequence_1
sequence_4

BAD sequences

N含有率10%以上

sequence_2
sequence_3
sequence_5
sequence_6

GOOD sequences

?

sequence_7 AAGNAAGGCTGAACN

タブ区切りテキスト処理の基本

```
sequence_1    GGCANGTATNGCTGG
sequence_2    GCCGAGTTTAATTGC
sequence_3    GCCGCCCTGGCCGTA
sequence_4    NCNAACTTCGCCCCC
sequence_5    ATGGATATTCAGGAA
sequence_6    CCGCCAAATTGTAAT
sequence_7    AAGNAAGGCTGAACN
sequence_8    TTTTTTTTTTTTTTT
sequence_9    GTAGCTTCAGGTGTA
sequence_10   AAAGCATTGCGTATG
```

ファイルを閉じる