bio334_day2_part3

May 13, 2025

1 Bio334 Practical Bioinformatics

The 2nd module, 14-16, May, 2025

1.1 Masaomi Hatakeyama

- GitHab https://github.com/masaomi/bio334_2025
- TAs: Narjes Yousefi, Kenji Yip Tong

2 Day2 Part3

Count the number of segregating site

[ ]:
```
import IPython.display
IPython.display.Audio("voice/day2_part3.mp3")
```

In this part, we will focus on the counting of the segregating sites, the SNP sites.

As you can see the definition of Tajima's D formula below, we need to count the number of SNP sites.

3 Tajima's D

One of summary statistics

$Tajima's D = \frac{d}{\sqrt{V(d)}}, d = \pi - s/\sum_{i=1}^{n-1} \frac{1}{i}$

- π: nucleotide diversity
- s: rate of segregating site = (number of segregating sites) / (length of sequence)
- n: number of sequences
- Reference: Tajima, F. Genetics 123, 585–95, 1989
- http://en.wikipedia.org/wiki/Tajima's_D
- What is segregating site?

4 Algorithm Design

Tajima's D

1. load sequences (already done)
2. calc nucleotide diversity (already done)
3. count the number of segregating sites
4. calc rate of the segregating sites

[ ]:
```
IPython.display.Audio("voice/tajimasd_calculation_step1.mp3")
```

Let's think about the process of calculating Tajima's D.

You have already implemented the nucleotide diversity, pi.

The remaining thing is the calculation of the number of segregating sites.

The segregating site means a different nucleotide site of some individuals from the others in the population.

How can we detect it?

5 Count segregating sites

What is the segregating site?

Break it into 2 parts 1. detecting alleles in each site 2. if the number of alleles >= 2, then it is counted as a segregating site

[ ]:
```
IPython.display.Audio("voice/tajimasd_calculation_step2.mp3")
```

Let's think about how to count the number of the segregating sites.

How about the following idea?

1. detecting alleles in each site
2. if there are the number of alleles, in other words, if you find more than two different nucleotides, then the position is counted as a segregating site.

In the example below, it uses *set* conversion from a list object.

However, this is just one solution. You might be able to find another solution.

6 Example1 (remember Set)

```
lst = [1, 2, 3, 2, 3, 4, 5, 1, 5]

print("list =", lst)
print("set =", set(lst)) #=> ??
print(len(set(lst)))     #=> ??
```

- Set is useful to discard duplicated elements

Q: How can we use it for Tajima'sD ?

[ ]:
```
lst = [1, 2, 3, 2, 3, 4, 5, 1, 5]

print("list =", lst)
print("set =", set(lst))
print(len(set(lst)))
```

[ ]:
```
# Example2

seq1 = "ATGC"
seq2 = "ATAT"
sequences = [seq1, seq2]
for i in range(0, len(seq1)):
    alleles = []
    for seq in sequences:
        alleles.append(seq[i])
    if len(set(alleles)) > 1:
        print("segregating")
    else:
        print("not segregating")
    print(alleles)
```

7 Idea (Inspiration, *abduction*)

Point

- **Set** object is used for detecting segregating site
- *probably, the two things (1) set has uniq elements, (2) need to detect different types of allele, hit my head, and Aha! came*

8 Idea

Logical thinking, *deduction*

1. Keep all nucleotides at same position in a List object
2. Convert it into a Set object
3. Count the number of Set elements
4. A segregating site must have more than two elements of the Set

[ ]:
```
IPython.display.Audio("voice/tajimasd_calculation_step3.mp3")
```

Let me just explain the algorithm that I constructed here in natural language.

1. keep all the nucleotides at the same position in a *List* object
2. convert the *List* object into a *Set* object.
3. count the number of the *Set* elements.
4. if the number of the *Set* elements is more than two, it is counted as a segregating site.

Do you think it makes sense?

Now you are ready to calculate the numerator of the fraction in Tajima's D formula.

Let's do the exercise, loading the FASTA file and calculating the number of segregating sites.

9 Tajima's D

One of summary statistics

$Tajima'sD = \frac{d}{\sqrt{V(d)}}, d = \pi - s/\sum_{i=1}^{n-1} \frac{1}{i}$

- π: nucleotide diversity
- s: rate of segregating site = (number of segregating sites) / (length of sequence)
- n: number of sequences
- Reference: Tajima, F. Genetics 123, 585–95, 1989
- http://en.wikipedia.org/wiki/Tajima's_D