bio334_day2_part4

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 Part4

Tajima's D calculation

[ ]:
```
import IPython.display
IPython.display.Audio("voice/day2_part4.mp3")
```

In this part, we will finally complete Tajima's D implementation.

The remaining part is the denominator of the fraction, the standard deviation of the small *d*.

But actually, we cannot know the value. So, we will estimate the value by using the formula in the original paper.

The same equation is written in the Wikipedia.

The equation looks complicated, but it is not so difficult. It has only addition, subtraction, multiplication, and division.

Just one thing note that the large S in the Wikipedia equation is the total number of segregating sites, it is not the small s in the Tajima's D formula below.

k hat is not the small *pi*, but it is usually expressed by large *PI*, which is not divided by the length of the sequence.

If you have a question, please do not hesitate to ask in Slack.

3 Tajima's D

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

- 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

4 Estimated SD of d

Wikipedia:TajimasD

Note - S (large S) in the formula above is the number of segregating sites, **NOT** the rate of segregating sites - i.e. The small d above is not same to the small d in the the following equation:

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

$\pi = \frac{\hat{k}}{length(sequence)} = \frac{\Pi}{length(sequence)}$

$s = \frac{S}{length(sequence)}$

$S = $ the number of segregating sites

5 What is Tajima's D?

- calculated pi vs. estimated pi (from the number of segregating sites)
- constant population and no natural selection $\Rightarrow d = 0$
- higher heterozygous sites$\Rightarrow d > 0$
- lower heterozygous sites $\Rightarrow d < 0$

Q: What is high/low heterozygous site? Singleton?

- Explain the typical cases in white board

[ ]:
```
IPython.display.Audio("voice/tajimasd_meaning1.mp3")
```

Now you can start the exercise, but just please let me quickly check again the meaning of the Tajima's D.

Tajima's D is actually the comparison between the calculated nucleotide diversity and the estimated nucleotide diversity from the number of segregating sites.

In theory, under the constant population and no natural selection, I mean, under the neutral condition, the Tajima's D should be plus/minus zero, which means the calculated nucleotide diversity and the estimation from the number of SNP sites should be the same.

The positive D means the calculated actual nucleotide diversity is higher than the estimation, that is, the population has more high heterozygous sites than neutral, and negative D suggests the opposite, more low heterozygous sites than the neutral.

I will show you this effect by simulation below.

Another good point to use computer programming is simulation. We can test a mathematical model numerically by setting any kind of condition virtually.

If you are not interested in a computer simulation or numerical experiment, please skip below and go on to the exercise.

6 Why programming?

- **Reusability**: Reproducibility, you can calculate with just one command type
- **Batch process**: Automation, computer can work while you are sleeping
- **Understanding**: It helps you to learn algorithms
- **Computer simulation**

7 Wright-Fisher Model

Assuming * Haploid population * Constant population size * No natural selection (Random selection) * Point nucletide mutation

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

Here, I will introduce the most popular model in population genetics, called the Wright-Fisher model. In this model, the following conditions are assumed.

1. Haploid population
2. Constant population size
3. No natural selection
4. Only point mutation happens, no recombination, no structural variant happens, and so on.

It means that this model can simulate the effect of genetic drift.

If you want to know more detail, please refer to Wikipedia, search by Google, or just ask me.

8 Marbles in a Jar (Analogy)

- 20 marbles in a jar
- colors correspond to two different alleles
- In each new generation the organisms reproduce at random

Wikipedia, genetic drift

9 Simulation

1. Population of nucleotide sequences
2. Produce a new generation
    1. select one sequence at random
    2. point mutation
    3. add it in a new population
3. Repeat (2)

WF Model

Parameters

- Population size (number of sequences)
- Length of sequence
- Point mutation rate

- Generation (iteration time)

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

In this simulation, the calculation steps are as follows:

1. Initializing the population
2. repeating the generations

in one generation, three things happen

1. selecting one sequence at random
2. inserting point mutations
3. adding it in a new population

This is continued until it becomes the fixed population size in the next generation.

To simplify in this simulation, I have not inserted the cross-over step.

10 Allele Frequency

without mutation, Population n=20

Wikipedia, genetic drift

Without mutation, Population n=200

Wikipedia, genetic drift

Without mutation, Popuation n=2000

Wikipedia, genetic drift

[ ]:
```
IPython.display.Audio("voice/wright_fisher_result1.mp3")
```

As you can see the results above, the dynamics is critically depending on the population size.

If the population size becomes bigger and bigger, the allele frequency gets closer to a constant, that is, the infinite population size and the constant allele frequency, which are actually a part of the assumptions in Hardy-Weinberg law.

11 No Mutation

Number of segregating sites, it becomes zero

12 No Mutation

Nucleotide Diversity, it becomes zero

13 No Mutation

Tajima's D, it goes to minus

[ ]:
```
IPython.display.Audio("voice/wright_fisher_result2.mp3")
```

If there is no mutation, at the end of the generations, all the alleles will be gone, and no segregating site is left, which means that the nucleotide diversity becomes zero.

I know this will never happen in the real world, but it is good to understand the model by simulation.

14 With Mutation

Number of segregating sites, it becomes constant but flactuating

15 With Mutation

Nucleotide Diversity, both becomes same

16 With Mutation

Tajima's D, it keeps around zero, but fluctuating

[ ]:
```
IPython.display.Audio("voice/wright_fisher_result3.mp3")
```

If the mutation is inserted in the model, the nucleotide diversity will be constant with fluctuation. and Tajima's D becomes zero, but fluctuated. The range of the fluctuation depends on the mutation rate and the population size.

This is the effect of genetic drift under neutral. The critical parameters in this model are mutation rate and population size.

So, what can we learn from the model?

I would say, the nucleotide diversity is fluctuated depending on the population size or population dynamics and mutation rate, and it will take some generations to become a constant, which is called transient time (or transient state).

In the actual situation, the population size is not constant, so, I would say, the sampling strategy and/or sample size might become very critical for the calculation of the nucleotide diversity.

17 Tajima's D > 0

$\pi > \theta_w = \frac{s}{\sum_{i=1}^{n-1} \frac{1}{i}}$

- **Excess of intermediate-frequency variants:**
  - **Population bottleneck or structure:**
    - Recent bottlenecks can reduce the number of rare variants (e.g., singletons), while older variants at moderate frequencies may persist.
    - Subdivided populations may maintain different allele frequencies, leading to an excess of intermediate-frequency polymorphisms.
  - **Balancing selection (e.g., heterozygote advantage):** Maintains multiple alleles at moderate frequencies, increasing the average pairwise differences ( ).

18 Tajima's D < 0

$\pi < \theta_w = \frac{s}{\sum_{i=1}^{n-1} \frac{1}{i}}$

- **Excess of rare variants (e.g., singletons):**
  - **Rapid population expansion:**
    - New mutations accumulate, introducing many rare alleles (singletons), lowering the average pairwise difference ( ).
  - **Positive or purifying selection:**
    - **Directional selection / Selective sweep:** A beneficial mutation rises quickly in frequency, reducing variation in nearby regions.
    - **Purifying selection:** Deleterious mutations are removed, reducing intermediate-frequency variants.

[ ]:
```
IPython.display.Audio("voice/tajimasd_meaning2.mp3")
```

What kinds of factors can cause deviations in nucleotide diversity and Tajima's D from the neutral expectation?

Two major factors are changes in population size and natural selection.

A **population bottleneck** reduces rare alleles, leaving only common ones. As a result, **(nucleotide diversity)** may decrease slightly, but **(based on segregating sites)** decreases even more, leading to **a positive Tajima's D**.

In contrast, **a rapid population expansion** increases the number of rare variants (singletons), which **increases more than** , making **Tajima's D negative**.

Similarly, **directional selection** (positive selection) can reduce diversity at selected loci, leading to **negative Tajima's D**.

On the other hand, **balancing selection** maintains multiple alleles at intermediate frequencies, increasing heterozygosity and resulting in **positive Tajima's D**.

19 Example1

The remaining part of d calculation

```
x = 0; n = 10
for i in range(1, n):
    x += 1.0/i

a = [1.0/i for i in range(1, n)]
y = sum(a)

print(x); print(sum(a))
```