03_genomics

June 5, 2024

Bayesian applications in genomics

0.0.1 Reconstructing genomes from sequencing data

You are going to develop and implement a Bayesian approach to reconstruct genomes from data produced from high-throughput sequencing machines.

Specifically, you will be doing **genotype calling** from short-read NGS data.

Load the *R* functions needed with `source("Data/functions.R")`.

Among these functions, we provide one that calculates the likelihood of a certain sequence of bases for diploid individuals. This function is called *calcGenoLikes* and takes 5 paramaters in input: * the sequence itself (collection of bases) * the first allele of the genotype * the second allele of the genotype * the sequencing error rate * a boolean indicating whether the results should be returned in logarithmic scale (TRUE) or not (FALSE)

```
[ ]: source("Data/functions.R")
```

For instance, assuming that your sequence is `AAGAGGA`, your alleles are `A` and `G` (meaning that you want to calculate the likelihood for genotypes `{AA,AG,GG}`, and your sequencing error rate is 0.01, then the likelihood (not in logarithm) for each genotype can be calculated with `calcGenoLikes("AAGAGGA", "A", "G", 0.01, FALSE)`

```
[ ]: calcGenoLikes("AAGAGGA", "A", "G", 0.01, FALSE)
```

Complete all the following tasks using *R* when necessary. The key point of these exercises is to not recalculate quantities that you have already computed. The aim is that you should be able to understand whether the likelihood or the prior is the same (or not) between different scenarios.

A)

Using Bayes' theorem, write the formula for the posterior probability of genotype G being AA given the sequencing data D. Write the explicit denominator assuming that your alleles are A and G and all possible genotypes are only AA, AG, GG.

```
[ ]: # P(G=AA|D) = P(D|G=AA) * P(G=AA) / P(D) #
     # P(D) = sum_g={AA, AT, TT}  P(D|G=g) * P(G=g)
```

B)

Assuming that your data is `AAAG`, your alleles are A and G, and the sequencing error rate is 0.01, calculate genotype posterior probability using a uniform prior, e.g. $P(G = AA) = P(G = AG) = P(G = GG) = ?$

```
[ ]: #...
```

C)

With the same assumptions as in point B, calculate genotype posterior probabilities using prior probabilitties based on Hardy Weinberg Equilibrium with a frequency of G of 0.1. Do you need to calculate a new likelihood or is it the same one as in point B?

```
[ ]: #...
```

D)

With the same priors used in point D but with a sequencing error rate of 0.05, calculate the genotype posterior probabilities. Do you need to calculate a new likelihood or is it the same one as in point C?

E) (bonus question)

With the same assumptions as in point C, calculate genotype posterior probabilities using a prior based on Hardy Weinberg Equilibrium with a frequency of G of 0.1 and an inbreeding coefficient of 0.2. In this case, we need to modify our previous priors. Specifically, if f is the frequency of allele A and I is the inbreeding coefficient, then the prior probabilities for all genotypes are: * $p(AA) = f^2 + I \times f \times (1-f)$ * $p(AT) = 2 \times f \times (1-f) \times (1-I)$ * $p(TT) = (1-f)^2 + I \times f \times (1-f)$

Do you need to calculate a new likelihood or is it the same one as in points B and C?

```
[ ]: #...
```

F)

Assuming that our collection of sequenced bases is `AAAGAGAAAAAAAGGGGAAAGGA`, calculate the genotype posterior probabilities using the same priors as in point C and sequencing error rate of 0.05. What happens if we have more data? What is the **confidence** in our genotype inference?

```
[ ]: #...
```

G)

What happens if we have a lot of data? Assume that your sequenced bases are `bases <- paste(c(rep("A",1e3),rep("G",1e3)), sep="", collapse="")`. Calculate the genotype likelihoods for this data. What is happening here?

It is convenient to use numbers in log-scale and you can do that by selecting TRUE as the last parameter in the *calcGenoLikes*. Remember that if you want to calculate proper probabilities (in log) you have to approximate the sum of logs.

Without calculating posterior probabilities, what is the effect of the prior here in your opinion?

```
[ ]: #...
```