Conditional Probability

Com Sci C121

Bayes Theorem: Allows is to condition on the likelihood

You have reads, and using conditional probability you want to find its location. You need the probability that there is an error to where it maps to. SNP-> Single Nucleotide Polymorphism.

Humans have genomes that are around 3.2 Billion bases long

Any two humans are 99.9% similar

Many types of genetic variations exist: insertions, deletions, substitutions, inversions, etc.

We will be focusing on single nucleotide polymorphisms-> these are single base mutations

We have sequencing reads that align to a genome with some mismatches. Are these mismatches errors from the sequencer or are they genetic variants? Further there are two chromosomes, so are there one or two copies of the variant

Common variants are often called the major allele and the less common one is called the minor allele. Frequency of the minor allele is called the dosage.

How do we know if a single mismatch is a error or a SNP?

Give me different models that align with the reads. This would give a scoring algorithms.

Specific problem setup:

We have reads that overlap a position in the genome

For simplicity, we are assuming we know this position varies in the human genome

We are restricting ourselves to biallelic single nucleotide polymorphisms

They can have only two possibilities, genotype A or genotype B. You assume you know the frequency of the alleles.

We need P(probablesequence|data) -> this is further simplified using bayes

The oracle-> how is the data generated? Make the following assumptions

1. Select a chromosome
2. Pick a starting point on the chromosome
3. Is it an error ?
4. What is the observation?

We write down some probabilities->

i = index for the read I

ci = underlying “true” base

P( ci = A | G= AA ) = 1

P( ci = B | G= AA ) = 0

Ei -> Random variable that there is an error at read I

P(Ei = 1) = Ei

P(Ei = 0) = 1-Ei

WE need to start with the likelihood of one observation:

Oi : the observation of read I

P(Oi = A | Ei ci) = P(Oi = A | Ei =0, ci  = A, G = AA) = 1

Probability that the observation is an A given that the error probability is zero, underlying “true base” is A, and that the previous data was AA.

P(Oi = B | Ei ci) = P(Oi = A | Ei =0, ci  = A, G = AA) = 1