Midterm Exam

Zachary Foster

May 2, 2014

1

Assuming that the 8bp motif can be located at any position in the intergenic spacer and can be located on either strand, there is $1000 \times 2 \times (100 - 8 + 1) = 186000$ positions at which an 8bp motif can occur. The chances of a random 8bp motif being any one permutation of base pairs is $4^{-8} \approx 1.53 \times 10^{-5}$. For a particular permutation of motif locations, the chances of exactly n number of 8bp motifs being identical is...

$$\left(4^{-8}\right)^{n} \times \left(1 - 4^{-8}\right)^{186000-n}$$

and the number of possible permutations of locations that the motifs can appear in is...

$$\frac{186000!}{n!\,(186000-n)!}$$

Therefore, the chances that 5 or more identical 8bp motifs is...

$$1 - \sum_{n=0}^{4} \left[\left(\frac{186000!}{n!\,(186000-n)!} \right) \times \left(4^{-8}\right)^{n} \times \left(1 - 4^{-8}\right)^{186000-n} \right] = 0.1583097$$

Below is the R code I used to do the calculation:

```
1 - sum(choose(186000, 0:4) * ((4^(-8))^(0:4)) * (1 - 4^(-8))^(186000 - 0:4))
```

```
## [1] 0.1583
```

2

Genome assembly is the process of finding overlaps between random fragments of DNA in order to infer the larger sequene they arose from. There are three general types of assebly algorithyms: greedy, overlap-layout-consensus, and de bruijn. Greedy assembliers, such as Phrap, iterativly combine the sequencs with the greatest overlaps until no more combinations are possible given a alignment quality threshold. They were the first invented and were designed for the long reads made by sanger sequencing. Overlap-layout-consensus assembliers, such as Phusion, find the shortest path along a graph constucuted from alignments of subsequences. They are much faster than greedy assebliers and allow for some sequencing error. De Bruijn assembliers, such as Velvet, work by finding the shortest path througha graph contruscted of fixed-length k-mers. They are able to handle extreamly large amounts of data, such as is produced by the Ilummina HiSeq, but do not tolderate erros well. There are also reference-based assembliers that map reads to a closly related genome. Repetitve sequenecs tend to cause breaks in assemblies if the repeat length is longer than the read length used in sequenceing, since reads will overlape equally well in multilple places and the unmber and location of repeats in the genomic sequence cannot always be determined. This is particularly a problem for short reads, such as those produced by Illumina HiSeq. Mate pair sequencing allows for two reads to be sequenced from the ends of a longer molecule of known length. The

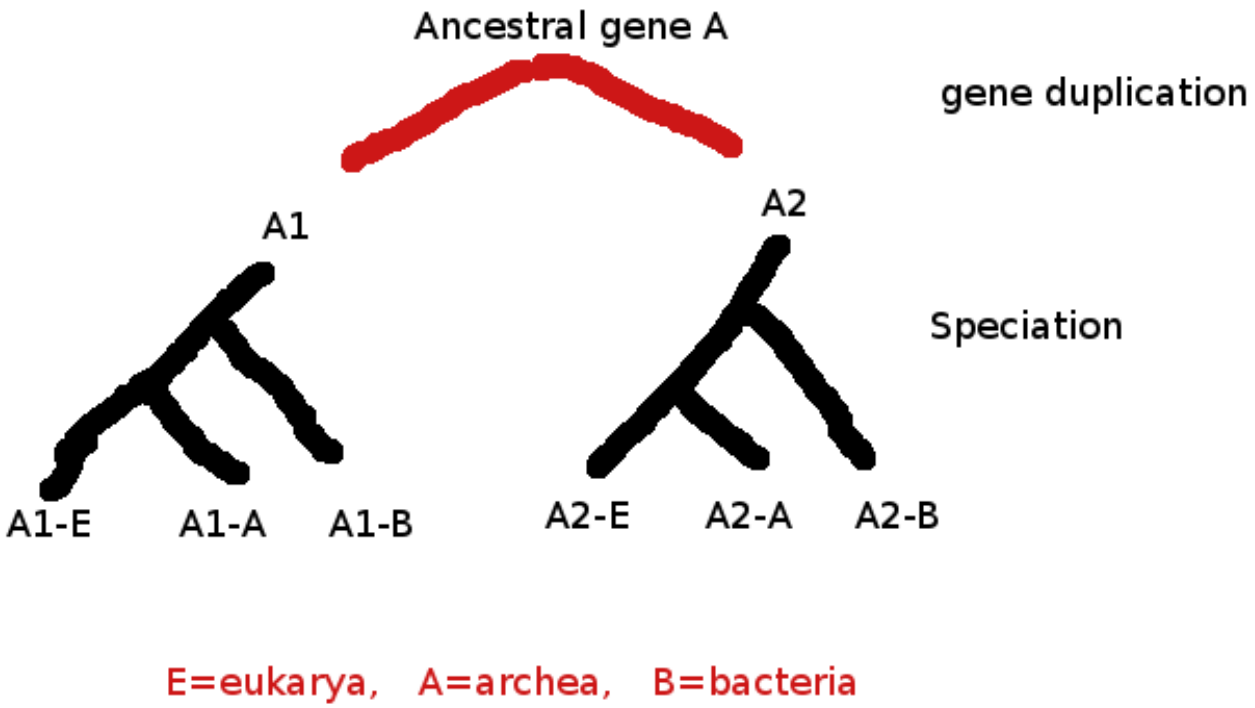
known length and origintation between the two reads can then be used to bridge gaps in the assembliy and acertain the location and number of repeats if a sufficent range of inserty sizes are used. Breaks in assemblies caused by repetitive sequences can also be spanned in retrospect by directed sequences (e.g. primer walking).

3

Ab initio gene prediction tools idenitify genes based off of sufficiently large open reading frams (ORFs) and organism-specific gene start/stop sequence motifs. Gene prediciton can also be done using BLAST to find homologous genes in refernece databases. The advantages of *ab initio* tools is that they can find genes that do not have analogs in refnereces database and do not require a refenrece database to work, however they do not rovide functional information. By contrast BLAST-based gene prediciton provides biologically-relevant functional information based off of putative homogolog annotations, however it requires homologs to be present in public databases.

4

The gene must have duplicated before the divergence on the ancestral organisms into eukarya, archea, and bacteria.



5

BLASTP detects more distanst protein homologs due to the noise introduced by the degenercy of the condon code. Since the thrid base of every codon often does not effect primary amino acid seuquence, the DNA sequence diverges faster than the protein sequence. Profile based approches contruct a model of sequenecs realtivly closely realted to the quesry sequecne in order to idenitfy the characteristics associated with the protein. the profile is then used to detect more distant homologs. Profile-based methods return

more results because they, in effect, use the information for multiple queries, which were derived from the primary query.