

Level 1

- Input: Reference genome and 1 copy genome
- Output: Copy genome in reference genome format

```
#!/usr/bin/perl -w
use strict;

my $refgen= <chromo22.txt>; →reference genome
my $toalign; →genome to be aligned
my $aligned = ""; →aligned genome

#open the refgen file
open(FILE, "<chromo22");

#get rid of any extraneous spaces/line breaks
chomp $refgen;
chomp $toalign;

#compare every base pair of $refgen to $toalign using a for loop with if
statements;

#in $aligned- if bp same, record a 0, else, record the bp from $toalign
to $aligned → I don't however think this is a good system. Maybe the
differing base pairs should be stored as numerical values instead? I know
we were discussing this in class, so other ideas?

my $count =0
for ( ;; )
{
    if ((refgen(mycount)).= (toalign(mycount)))
    {
        aligned. "0";
    }
    else
    {
        aligned. toalign(mycount);
    }
}

#print out $aligned;
print "Aligned Sequence: \n$aligned";
```

Level 2

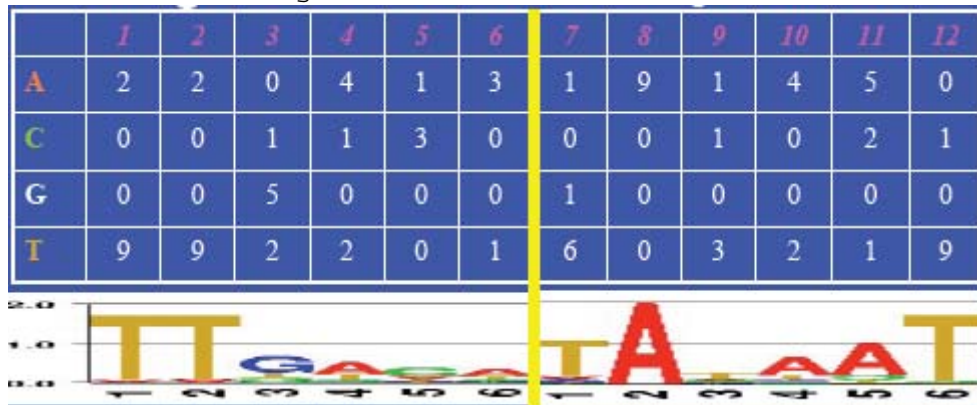
- Input: Reference genome and 100 copy genomes
- Output: List of statistically significant mutation spots

Define statistically significant- let's use the p-value of .0005, as is traditionally used in statistics;

1. Use this value to determine the statistically significant number of differences in base pairs per site. →\$statsig

Assume: Reference genome = consensus sequence (so it has no mutations); otherwise, by comparing the bps between the reference and input genomes, we can write another program to determine the relative presence of various base pairs, and thus generate our own consensus sequence.

1. Make a tally for the differences between the reference and copy genomes labeled by position. (ie. add 1 to \$counter_position as you loop through each copy genome; compare each of these values to \$statsig; output locations with more differences than \$statsig)
 - a. But, this is too cumbersome, as it requires millions of variables labeled \$counter_position)
 - b. If we did however, we could generate a weight matrix, similar to the following:



- c. Such a matrix provides a clear view as to the consensus sequences, hotspots for mutations and the types of mutation.