1 Representing Trees

(a)

I can specify an ordering by choosing a parity for $n-1$ node merging events. Thus, there are 2^{n-1} isomorphic trees.

(b)

A simple method to do this is implemented in my routine "canonicalize" located in the file tree.py. First I associate a numerical index with each leaf. For each internal node, I ask whether the minimal leaf index of the subtree is less than the minimal leaf index of the right subtree. If the right min-index is lower than the left min-index, I swap the left and right subtrees of the internal node.

This algorithm proceeds recursively from the top of the tree to the bottom calling the recursion exactly once at each node. It is efficient and used in the construction of NJ and UPGMA trees in the next problem.

(c)

I would simply run the same algorithm with a root arbitrarily placed at the lowest index.

2 Tree Building

a-c

See figs 2a, 2b_upgma, 2b_nj.png and the python programs tree.py/upgma.py. For fig 2a sequence costs for each tree position are computed via a simple dynamic program and the total tree costs per position are represented by a tuple at the (artificial) root node.

upgma.py computes both NJ and upgma trees based on whether the method 'run_nj' or 'run_upgma' is selected. If 'run' is selected, it runs with randomized sample data. NJ computes distance correct trees by considering average distance between a node and its neighbors.

3 Positive Seletion

(a)

i

Long haplotypes indicate that a derived locus has risen to prominence in a population on a shorter time scale than recombinative disruption. For a large population this is unlikely to happen unless selection is at work.

ii

Because in a long haplotype many individual polymorphisms may have been derived and hitchihiked to prominence by chance alone.

iii

cM is preferable here because rates of recombination can vary across the chromosome - recombination frequency measure a genetic distance with recombination taken into account. (see figs/3aiii.png)

iv

C, Europe appears to be under selection as it scores high in EHH in comparison with each population.

v

79 snps appear above 2.0.

(b)

(i)

Back of the envelope: the ancestor is about halfway between chimp and human thus .66% of the traits that we label “ancestral” will have in fact been derived by the chimp. For these we will usually observe the true ancestral trait in humans which we will misidentify as derived.

ii

See figs/3bii_all.png and figs/3bii_europe.png.

iii

See figs/3biii_europe.png and figs/3biii_all.png.

(c)

i

$$\begin{aligned}
H_S &= \overline{p \times (1-p)} \\
&= \frac{(p+d)-(p+d)^2+(p-d)-(p-d)^2}{2} \\
\implies F_{ST} &= \frac{H_T - H_S}{H_T} \\
&= \frac{p(1-p)-(p-p^2-d^2)}{p(1-p)} \\
&= \frac{d^2}{p(1-p)}
\end{aligned}$$

ii

I can estimate p for a pairwise comparison by averaging derived frequency over the two subpopulations. d is simply $\frac{f_1-f_2}{2}$ with the appropriate sign. By the formula above, I can compute F_{ST}.

iii

See figs/3cii.png. 7 SNPs pass $F_{ST} > .6$ and one of these is excluded by the previous tests. Thus, 6 SNPs pass the above thresholds and have $F_{ST} > .6$.

(d)

i

Of 2687 SNPs, 76 fall in exons, 1121 fall in introns, and 105 total are conserved.

ii

Only one of the SNPs passing all thresholds is associated with an exons (the remaining five are intron associated). This SNP, rs16891982 is also strongly conserved according to the annotation given. I presume that it is functional. The single SNP indicated by F_{ST} to be significant but not passing previous thresholds also sits in an exonic region but it is not strongly conserved.

(e)

The weight of the evidence: Long haplotype, derived allele frequency, population differentiation, and annotation appear to suggest that rs16891982, living on the gene MATP otherwise known as Q9UMX9 is under selection.

Present in Europeans but not Africans or Asians, associated with a gene encoding a transporter protein mediating melanin synthesis, it seems that rs16891982 may be responsible for some part of European palor. For such a small bit of biology, a recent switch from a C to a G nucleotide appears to have caused substantial mischief in centuries past and present.

4 Variant Discovery

(a)

This implies that if we try to identify SNPs from a naive search of the entire genome, we will find many false positives unless we have some way to discriminate sequencing errors from actual variation.

(b)

The easiest way to compute empirical error rate is to examine all sequence positions where coverage is above some c_{min} and count the frequency with which reads at a given position differ from the consensus read of a position.

Choosing $c_{min} = 10$, $(n = 238,000)$, I see that 92% of positions are completely conserved and at only 260 positions do more than two reads differ. Thus for nearly 999 out of 1000 positions, a consensus can be established over more than 80% of reads and I can compute a mean error rate over all reads with coverage $> c_{min}$: $\epsilon_0 = .005$.

The main errors to worry about are mismatches and insertion/deletions. In either case, I would worry that such errors could cause errors in reference sequence alignment itself. With newer sequencing technologies using shorter reads, we will tend to have higher error rates (lower quality scores), more difficulty correcting mismatches, and inevitably less certainty in sequence alignment - especially in repeating regions.

Some genomic regions appear to frequently mismatch the reference genome. figs/4b_selected_region shows such a region where nearly 30 nucletides in a row are identified as possible SNPs. Since there is high disagreement among sequencing reads, it is likely that some factor is causing poor sequencing accuracy in this region. Checking this region and its neighborhood for motifs using sequence finding code from the first problem set, I see that a yeast motif 'GATGAG' or ESR2 - a repressor of protein synthesis genes - sits in the middle of the region with high sequencing errors. Presumably GATGAG induces an epigenetic alteration that causes sequencing errors. See figs/4b_motif_errors

Per base, the raw error frequencies vary from .45% got C and G nucleotides to .66/.7% got T and A nucleotides respectively. Mismatch errors are more common than errors resulting in undetermined bases, "N". Some bases are more likely to be mismatched to one another in a non-reflexive fashion (see figs/4b_base_errors.png)

I can also compute a mean error rate per sequence position. The easest way to do this is to check reference reads along the sequence and compare to every read hitting that position - hashing errors according to position relative to read start. Normalizing by the total umbers of reads at each position gives the plot in figs/4b_positional_errors.png.

(c)

First I map the references sequences across the genome and then I do the same thing for sequence reads. To keep track of discrepancies between reads, I map reads to an $n \times 5$ array that functions as a hash table incremented whenever a read indicates a base (or "N") at a particular position. From this hash table, I find the most common read at a site but in order to distinguish SNPs with high certainty - say, 95% - I demand that the probability of the correct labeling exceeds the probabillity of all other labels by a factor of 20. Using the single valued empirical error rate $\epsilon_0 = .005$ estimated above, I can approximate relative probability of correct labeling by $\epsilon_0^{|\text{consensus}|-|\text{other}|}$. In order to minimize misidentifications of SNPs, I can apply a bonferri correction and demand that correct labeling be more probable by a factor of $n \times 20$.

With bonferri correction applied, 299 SNPs are significant to 95% confidence. The top 50 are listed below:

ref base	snp base	hits	misses	prob	context
A	T	20	1.0	100.668	GAAAATACGA
A	C	20	1.0	100.668	AAATACGATT
T	A	20	1.0	100.668	AAAATACGAT
G	A	16	0.0	84.773	CGGTGAAACT
A	T	16	0.0	84.773	TAGAATTAAT
T	A	16	0.0	84.773	GTAGAATTAA
G	C	15	0.0	79.475	ACCGGCGCAG
A	C	16	1.0	79.475	CAGCCCACTA
G	C	15	0.0	79.475	AGGAGCCAAA
C	G	16	1.0	79.475	AATACGATTT
T	A	15	0.0	79.475	AACAAATATA
A	C	15	0.0	79.475	ATGCGCAATA
T	A	14	0.0	74.176	ATTCGAACCG
G	A	14	0.0	74.176	AGTAGAAAAT
T	A	14	0.0	74.176	TTCGAACCGG
G	A	15	1.0	74.176	ATACGATTTG
G	A	13	0.0	68.878	ATTTGACCAA
G	T	13	0.0	68.878	CTCGATGTCA
G	C	13	0.0	68.878	AAAAGCGCAG
G	C	13	0.0	68.878	AAGCGCAGGG
A	G	13	0.0	68.878	GAAAAGCGCA
C	G	13	0.0	68.878	AAAGCGCAGG
A	T	13	0.0	68.878	GGAGTTCAGG
A	G	13	0.0	68.878	GCGCAGGGCA
C	A	13	0.0	68.878	AGCGCAGGGC
T	G	12	0.0	63.58	CATTTGAAAT
A	T	14	2.0	63.58	TACGATTTGT
C	T	12	0.0	63.58	AAATTTTTTT
C	T	13	1.0	63.58	CTTCGTGTTT
T	G	12	0.0	63.58	TGTTTGGTCT
G	A	12	0.0	63.58	AAGCGAACAA
T	G	13	1.0	63.58	GATTTGTTTG
C	T	12	0.0	63.58	AACAGTCTGG
A	C	12	0.0	63.58	ATTCCCAAAT
T	C	11	0.0	58.281	TGGAACCCAG
G	A	11	0.0	58.281	CTCTAAATCA
G	A	11	0.0	58.281	ACAAGATCTG
G	T	13	2.0	58.281	ATTTGTTTGT
T	A	10	0.0	52.983	CTTCAACCGT
G	C	10	0.0	52.983	CAGGGCATTC
A	G	10	0.0	52.983	GAATTGGACA
G	A	10	0.0	52.983	TACCGAAGGA
A	T	10	0.0	52.983	GTTGTTCATT
A	G	10	0.0	52.983	GACTGGAAAT
C	G	10	0.0	52.983	CAACCGCCTC
G	C	10	0.0	52.983	GACAACCGCC
T	A	10	0.0	52.983	CTCAAAACAA
A	T	12	2.0	52.983	AGGAGTTCAG
A	G	9	0.0	47.685	GAAAGGACTA
C	T	9	0.0	47.685	TTTTTTCAAA

The method that I have described and implemented in q4 uses the estimated sequence wide-error rate and assumes that the effects of read position and nucleotide on SNP confidence are relatively small. The method that I used to compute likelihood of the correct labeling is also only correct for small error rates. Each of these assumptions appears to be the case here but each could be relaxed in a more complicated model. A prior distribution reflecting SNP likelihood across the population and the genome could also be used to further refine results. Presumably the current model will significantly underestimate SNP occurence due to the bonferri correction.

(d)

I would need labeled sequence reads from diseased and healthy individuals. If labeled dataset was small, I would want another dataset with which to establish a prior distribution over SNPs in the population. Presumably I would wish to have a sufficiently large dataset to compute SNP linkage as well. Linked SNPs - even if indivually rare - will not be independently correlated with phenotypes.

In order to assess functional impact of a given SNP$_a$, I would seek a collection examples where SNP$_a$ was delinked from its neighbors and compute relative power of SNP$_a$ to predict disease phenotype. With cell specific expression data I could directly assess correlation of SNP$_a$ gene expression and disease phenotype.

5 Coalescence

ii

I found average total coalesence times and std deviations of $505 \pm 498, 680 \pm 513, 745 \pm 522$ for $k = 2, 3, 4$ respectively. Since coalescence times have a STD of 500 and mean of $500 - 750$, I suppose that running simulations for 3500 generations will be sufficient to ensure coalescence of greater than 99.9% of children and therefore coalescence in 1000 trials.

Too many generations will waste memory and in too few generations, we will be likely to see children that do not share a common ancestor.

iii

My results agree well with predictions of coalescent theory.

1. For $k = 2$, standard deviation is nearly n.
2. For $k = 2$, mean coalescence time is nearly n.
3. For $k > 2$, coalescence theory is approximated by the exponential

$$e^{-\frac{k(k-1)}{2}\left(\frac{1}{N}\right)} \tag{1}$$

giving the mean time for first coalescence for $k = 3$ of 166 and for $k = 4$ of 83. Predicted coalescence times for $k = 3, 4$ are therefore 666 and 749 respectively - a close match to the data I found. Computing expected standard deviation for total coalescence from the sum of expected variances for individual coalescences gives expected stdevs of $500, 526, 534$ which are close to the observed STDs

(b)

i

For $M = F = 250$ the model indeed works the same.

ii

Implementing a model for sexual reproduction where individual males produce children with probability $p_m \propto F$ and individual females produce children with probability $p_f \propto M$, I find total coalescence times and STDs of $314 \pm 327, 430 \pm 345, 474 \pm 351$ for $k = 2, 3, 4$ respectively. Evidently, coalescence times are shortened by the unequal distribution of sexes.

iii

In order to extend the coalescent approximation, consider that the $1/N$ collision frequency for haploid coalescence with $k = 2$ comes from the odds that child org$_2$ will "choose" the same parent as org$_1$ from a single pool of parents with size N.

For M male and F female parents, we must compute collision probabilities in each sexual pool of parents and add them together:

$$\begin{aligned}
p_{\text{collision in M}} &= (p(\text{org}_1 \to M)) \times (p(\text{org}_2 \to M)) \times (p(\text{orgs collide in } M)) \\
&= \frac{1}{2}\frac{1}{2}\frac{1}{M} \\
p_{\text{collision in F}} &= (p(\text{org}_1 \to F)) \times (p(\text{org}_2 \to F)) \times (p(\text{orgs collide in } F)) \\
&= \frac{1}{2}\frac{1}{2}\frac{1}{F} \\
\implies p_{\text{collision}} &= \frac{1}{4}\left(\frac{M+N}{MN}\right)
\end{aligned}$$

Noting tha t$M + F = N$, we see collision likelihood in an unequal sexual population is the same as collision likelihood in a haploid asexual population with size

$$N_{\text{eff}} = \frac{4MF}{N^2} \tag{2}$$

We can test our conclusions by multiplying stds, times by $N/N_{\text{eff}} = 16/25$ to give $490 \pm 510, 671 \pm 539, 740 \pm 548$. Similar to our previous results and suggesting accuracy of our coalescent theory with effective population size. Note also that in the equal sexes case, this extended theory matches exatly the coalescent frequency for diploids - $N_{\text{eff}} = 2N_{\text{M,F}}$.