**Computational Biology Lecture Questions**

**Lecture 1**

*What are the weaknesses and strengths of the different alignment methods (dot-matrix method, smith-waterman, Needleman-wunsch,BLAST)?*

Dot-matrix method is very easy to construct, but it does not give you an optimal alignment, it only gives you an idea of the alignment.

For SW the pros are that it gives an optimal alignment, but this alignment is local.

For NW the pros are that is gives a global alignment.

For BLAST you make comparisons and do not necessarily make alignments.

*With which alignment methods do you get an optimal alignment?*

Smith-Waterman, Needleman-Wunsch. Optimal in this case means given your scoring scheme.

*Do you obtain the same alignments when using different scoring schemes in the NW and the SW algorithm?*

No, but if you increase each penalty with the same number, the optimal alignment is still the same.

**Lecture 2**

*Does the odds ratio in a GWAS prove that a minor variant at a SNP position causes a genetic disease?*

No, it just gives a hint whether there might be an association between a SNP position and a genetic disease.

*In a GWAS why can you not reject your null hypothesis if your p-value is less than alfa?*

You would divide alfa by the number of hypotheses you want to test. This is called correction for multiple testing (Bonferroni correction).

**Lecture 3**

*What are the differences between the JC69 and the TN93 models?*

JC69 only has one parameter and TN93 has three parameters for the different transitions and transversions and the equilibrium frequencies. The equilibrium frequencies of JC69 are 0.25.

*Which of these two models would you choose if you were to perform a phylogenetic analysis based on sequence distances?*

Your model can evolve under JC69, but statistical variation can make you trick into seeing different parameters that are not actually there. For real life data we will however probably choose TN93.

*In lecture 3 we tried to naively reconstruct a phylogeny based on three sequences. The pairwise Hamming distances were all the same. Would you expect that any of the presented nucleotide sequence models would result in different trees?*

Yes, if you use a more complex model different pairs of sequences will end up with different other sequences.

**Lecture 5**

*What is the minimal number of cherries in a phylogenetic tree of 99 tips? What is the maximum number?*

The minimum number is 1, the maximum number is 49.

*In how many ways can you write the newick string for a rooted tree with species A,B,C? In how many ways can you write the Newick string for a rooted tree with n species?*

For A,B,C we have 4 different representations. 2^(n-1) for n species.

*Consider the least squares method. Why would we use wi,j which are not equal to 1?*

We know that there is more variance over a longer period of time, so for the larger distances. This means they will produce a larger number. But we want to give equal weight to the large and small distances, so we might give a higher weight to a lower distance and a smaller weight to a larger distance.

**Lecture 6**

*Consider the Fitch algorithm. Do you obtain all most parsimonious ancestral sequences when choosing the different nucleotides in the curly brackets?*

No, because in the example above you can have the C in the root already, but then you still have 2 mutations.



*Does the maximum likelihood tree reconstruction method return estimates for the internal sequences?*

We have P(Dk|X) at every node, which is the likelihood of X. This is not the same as P(X) which we would want in this case. Therefore, we cannot use this to return estimates for the internal sequences.

*Does the Fitch algorithm return the parsimony score for any phylogenetic tree and any sequence alignment? Or are there situations when the Fitch algorithm does not return the smallest number of mutations required?*

It always returns the minimum value of mutations.

**Lecture 7**

*Is there a way to test how to best root a maximum likelihood tree?*

If your model is time reversible then the likelihoods of the trees are the same for all roots. If you don’t have a time reversible substitution model then you can try out different roots and calculate the likelihood. Otherwise not.

*Can you use the bootstrapping ideas for assessing confidence in a UPGMA tree?*

Yes, you subsample new sequences from the original sequence and you build your new tree. Then you look how many times the splits between individuals were observed.

*What is required to infer the direction of transmission from phylogeny?*

If you have one cluster within another cluster, then you know that the second likely arose from the first. So in the example below you can be sure that patient 2 got infected before patient 1.



**Lecture 8**

*In a Fisher’s exact test, how would you calculate which values for one of the cells in the contingency table would lead to a rejection of the null hypothesis, given that row and column sums remain the same?*

You take one entry and fill in there a zero,one,two and based on the hypergeometric distribution you can then calculate the probability. Then to get p-values for a certain X you look in the probability distribution above X.

*Is the Brownian motion model a good model for all continuous traits? Could you imagine a situation where this is not the case and which assumption in this model could be violated?*

It can be a good model because it models drift. If your trait evolved under such a process without selection it is a good model. If your process is not neutral and there is selection towards some optimum, then the model is not good. Then one wants to extend to account for selection.

*Do you think it is a good strategy to first determine the species tree and then look at character evolution, or would a co-estimation of characters and the phylogeny make more sense?*

The substitution models assume neutrality. We would assume if we would co-estimate them that the phenotype evolved neutrally. Which might be the case, but if there is selection then it might not be useful to use the characters to infer the phylogeny because you assume neutrality in the inference.

**Lecture 9**

*How does the approximate number of steps required to calculate the phylodynamic likelihood depend on the number of leaves in a phylogenetic tree?*

It is linear proportional to the number of leaves

*What kind of population dynamic process could a decrease in slope in the LTT plot reflect?*

Increase in b, decrease in d.

*Assume a birth-death process where each individual at present is sampled with probability p. How is the derivation of p(0|t,p), the probability of sampling no individual at present, different compared to the derivation of p(0|t)?*

P(0|t) is the probability that an individual at time t in the past gave rise to 0 descendants in the present. We fixed p(0|0)=0 and P(1|0)=1. We have to replace p(0|0)=1-p and p(1|0)=p. So the boundary conditions change.

**Lecture 10**

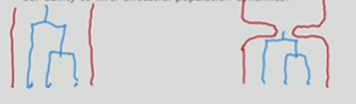
*Under WF model, how many generations do we have to go back before we find the common ancestor of a pair of genes sampled from a haploid population size of N?*

We have the probability of P(m) for coalescence m generations ago.

*Suppose you had a tree inferred using present-day samples from a population that experienced a severe bottleneck in its recent past. How and why would this bottleneck likely affect our ability to infer ancestral population dynamics?*

We assume our individuals are at a demographic equilibrium and this bottleneck would have shifted this equilibrium, which means our ability would be wrong to infer population dynamics.

The situation might look like the tree on the right side.



So inferring anything beyond the bottleneck point is almost impossible.

*Imagine spreading a WF population across islands in an archipelago, so that movement between the islands is restricted but within each island the population is well-mixed. Qualitatively, how would you expect this population structure to influence estimates of the effective population size?*

Depending on the timescale you’re looking at, your samples that you take within your island will be similar to other islands. If you take it on a short timescale, you cannot infer much about the total effective population size, whereas if you have it on a long timescale, then you can assume the population is well mixed again.

**Lecture 11**

*Does a Bayesian phylogenetic analysis of the kind described here allow one to directly infer ancestral sequences? Why/Why not?*

No. The way we set up that likelihood was that we integrated over all the ancestral states. We used this to set up the likelihood of the present day sequences given a tree and a substitution model. Implicit is the idea that we’ve averaged over all possible histories, so we can’t directly recover them.

*How might we test to see whether a bayesian MCMC analysis has explored the full state space supported by the posterior?*

In practice we run the algorithm several times, rather than just once. If we get different quantitatively distinct results, we know that our algorithm has not converged and we have not explored the entire state space, so the states need to be run longer.

*Suppose you have conducted a bayesian phylodynamic analysis and recovered a 95% HPD interval for the birth rate parameter. If you take this result and use it to construct a new prior for this parameter and use this prior to analyze the same data, would the resulting second posterior be valid?*

We know with 95% probability that the birth rate is between a certain interval. It would not be valid, because you are using the same data twice. You would end up with a false sense of confidence. You would end up with a tightened posterior, even though you’ve only analyzed one set of data.

**Lecture 12**

*Under a structured birth-death model, how do the sub-population sizes vary (if at all) through time?*

They vary according to the birth/death rates and the migration rates. If you have a birth rate that is greater than the death rate, they will grow exponentially. Otherwise they will decay exponentially.

*Suppose you perform a structured coalescent analysis on sequences collected from a relatively unstructured population. Would you expect the posterior migration rate to be very low or very high? Why?*

If you get types from both populations, the migration rate would be extremely large. This means there is free movement between those two types. You can see this by only sampling from one island, but pretending it is from different islands. Since they are not structured, you would infer a very high migration rate (they are all on the same island). If, however, You would only sample from a single deme, then it could go either way. Both would yield a mixed population.

*How might the evolution of languages violate the assumption of a substitution+birth/death phylodynamic model?*

The borrowing of words between languages is very important and is not easily modelled.

**Lecture 13**

**Homework 1**

*Why are gaps generally more penalized than mismatches?*

A gap would mean that there was a deletion or insertion at some point, which has a much lower probability than a snp.

*Why does it make sense to disallow gaps at the start and end of a local alignment, considering the -2 score for a gap?*

It’s a local alignment, which means that the start and end of one of the two sequences is completely aligned in the ‘middle’ of another sequence. On the start and end the alignment stops, it is not a gap anymore.

*What is a potential problem that could arise when trying to locally align a very short sequence to a much longer one?*

You could potentially have many different possible places to align it in.

*There are (look at slide) possible alignments for two sequences of length m and n, m>=n. For sequences of lengths m=n=100 this amounts to 2.05x10ˆ75 possible alignments. Give a rough estimate for the number of steps needed to align two sequences of this length using the Needleman-wunsch algorithm. For sequences of length m=n-100 roughly how many times less than exchaustive search is this?*

You need to fill out a matrix. This means you have two for loops of each 100 iterations. So 10^2\*10^2=10^4 iterations. This is 10^71 less than brute force.

*How could you extend NW to k sequences?*

You could extend NW into k dimensions.

**Homework 2**

*Imagine that the ancestral sequence of a tree of 10 species is composed of 50% T and 50% C. How do you expect the distribution of nucleotides to change of the overall rate of change is extremely low compared to the time scale of the tree? What if it is relatively high?*

If the rate is low, the distribution will not change much. If the rate is high it will converge to the equilibrium distribution.

*Imagine you are trying to simulate evolution along a tree using only the substitution rate matrix Q, without ever computing the transition probability matrix P. The overall rate of change from a nucleotide I can be read as -qii in the matrix Q. How would you randomly draw the time when the next substitution event happens?*

You can draw a time from an exponential distribution with rate -qii.

*Following the previous question, assume you have drawn the time of the next subsititution event for a particular nucleotide. How could you sample the nucleotide it is substituted by?*

You can sample from a probability distribution with [%nuc1,%nuc2,%nuc3].

**Homework 3**

*What happens if you try to compute the K80 distance between two very dis-*

*similar sequences? Take for instance the sequences AACTCA and TTAGTG.*

2V or 2S may be larger than one, which gives a negative value in the logarithm which is impossible to evaluate.

*Assume you are using the UPGMA algorithm, and your initial distance matrix*

*has multiple equal minimal entries. Does your implementation of the algorithm*

*influence the output tree?*

Yes, since it matters which one you choose first to merge.

*Say we know that several sampled sequences have evolved under a JC69 model.*

*We calculate a JC69 distance matrix for these sequences, apply the UPGMA*

*algorithm to construct a tree, and calculate distances from the tree. However,*

*our tree distances do not match the JC69 distances. Name two features of*

*the sampling scheme or the true evolutionary process that might cause such a*

*discrepancy.*

A) UPGMA assumes a strict molecular clock  
B) UPGMA assumes all sequences were sampled at the same point in time and it will force an ultrametric tree.

*Can any of the potential problems you raised in your answer to question 3 be*

*taken into account using an alternative algorithm mentioned in the lectures?*

*If yes, which problem(s) and with which algorithm? If no, please give a short*

*justifcation.*

You can use the neighbor joining algorithm which doesn’t assume this strict molecular clock.

***Homework 4***

*Figure1, taken from lecture 6, shows likelihood calculations performed on a tree*

*with pendant branch lengths of 0.2 and internal branch lengths of 0.1. The K80*

*model was used. The cells correspond, from left to right, to nucleotides T, C,*

*A and G. Explain why there are two di  
erent probabilities at node 6 for A and G.*

There is no G in any of the children so it’s less likely that we had a G at the root of this subtree.

*Imagine you have computed the likelihood of a tree with Felsenstein's pruning*

*algorithm and have written down the results of intermediary calculations on*

*internal nodes. You now alter the tree with a nearest-neighbour interchange*

*move. If you want to calculate the likelihood of the new tree using the same*

*algorithm, can you reuse: none, some, or all of the calculations you performed*

*on the previous tree? Give a concise explanation.*

You can reuse some of the calculations. You can reuse the ones for the nodes whose children haven’t changed.

*List at least three key differences between the UPGMA tree reconstruction and the maximum-likelihood tree search.*

1. A strict molecular clock. 2. UPGMA is much less demanding because you don’t have to search through the tree space. 3. UPGMA is a phenetic approach (only takes pairwise distances into account), maximum likelihood is a probabilistic approach.

*How would the runtime of Felsenstein's algorithm change if you had 5 nucleotides instead of 4?*

Since we have two nested for loops for which you compute the amount of nucleotides, this would become O(5^2) instead of O(4^2)

*In lecture 4 we learnt about the condition of \time reversibility". Could you*

*place the root anywhere in your tree and still obtain the same likelihood if the*

*substitution model was not time reversible?*

No, in a time reversible model the direction of descendance matters, which is not the case in a non time reversible model.

***Homework 5***

*Why can't Fisher's exact test be directly used to check for correlation on the*

*values of discrete traits at the tips of a phylogenetic tree?*

We want to see whether they are correlated independently of evolution. Fisher’s test looks at correlation in general, and does not take into account where this correlation comes from.

*In the independent contrast method, the contrasts Zk are mutually indepen-*

*dent and have identical variance. Explain why this is done.*

They are independent because this is a requirement for performing linear regression.

*Explain which steps of the algorithm focus on defining a set of independent*

*variables, and which steps ensure that these have identical variance.*

The taking the difference of the traits the independence is defined. By dividing over the square root of the sum of the branch lengths you make sure they have identical variance.

*You are studying the height of canines and the length of claws among apes.*

*What could make you think that they evolved independently, or on the con-*

*trary, that they evolved in a correlated way ? In which case would the two*

*traits show a correlation ? In which case would they show correlated normal-*

*ized contrasts ?*

You would think they have evolved in an uncorrelated way of the traits were used for something different, whereas if they evolved in a correlated way they would have been used for the same thing. You would observe a correlation in both cases, but you would only observe a correlation in normalized constrasts in the second case.

*A strategy for comparing discrete characters while accounting for relatedness*

*between individuals was presented in lecture 8. Imagine you want to perform*

*a discrete character comparison using this strategy on two discrete traits that*

*you are able to accurately observe on individuals at the tips of a known tree.*

*What problem might you encounter when trying to ll in the contingency table*

*of character changes?*

Differences in branch lengths are not taken into account. Changes are more likely to happen on longer branches than on shorter ones.