Computational Biology Summary

# Open questions for CB

Lecture 4, 49: How do we get these numbers?

Lecture 6, go over Felsenstein, Fitch and MLE again.

# Sequencing methods (excluded from the exam)

# Sequence alignments and BLAST

**What happens when DNA/RNA is copied?**

* Mutation
* Insertion
* Deletion
* Repeats
* Inversions
* Inverted repeats.

**What is the premise of an alignment?**

It’s based on the fact that there is a common ancestor. This assumption is impossible to know for sure.

**What is a homologous sequence?**

It’s a sequence with shared ancestry (= which has the highest probability/score under a certain model).

**What are the types of alignment that can be made?**

Pairwise alignment types:

* P-P
* DNA-DNA
* RNA-DNA
* DNA/RNA-P -> be careful for gaps and frameshifts within codons

It’s also possible to have a multiple sequence alignment (MSA).

**What is the difference between a local and a global alignment?**

Global alignment aligns one sequence to the other from the start to the end.

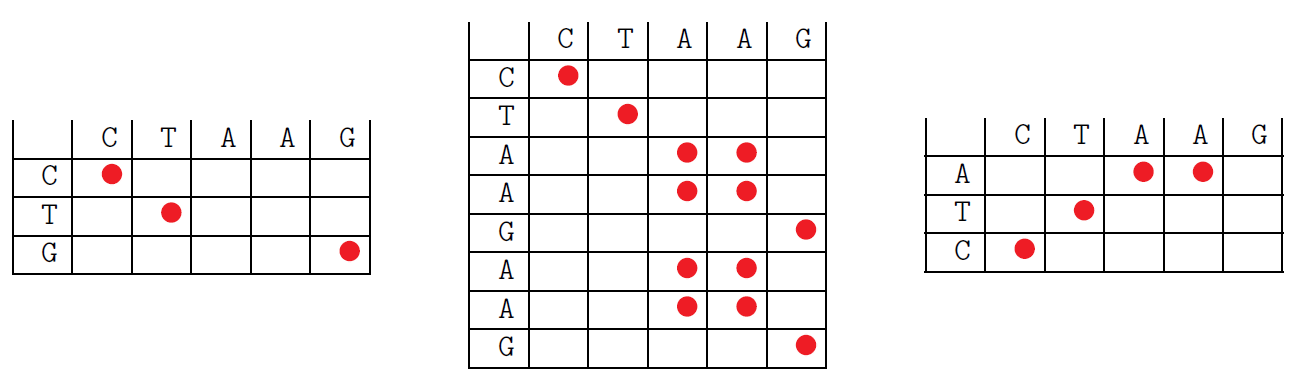
Local alignment finds the longest subsequences with highest similarity.

**What are the different strategies to find alignments?**

* Qualitative methods: dot-matrix method
* Exact method w dynamic programming: N-W (global) & S-W (local).
* Heuristic and fast methods (e.g. BLAST)

**What is the dot matrix method?**

Simplest way to align sequences.



Gap in sequence repeats inversions.

**What are some of the (dis)advantages of the dot matrix method?**

Pro:

* Visually easy method to identify sequence features such as repeats, inversions, and inverted repeats.

Cons:

* Time consuming
* Does not give one optimal alignment.

**What is a big trade-off to make when doing local alignments?**

Do we accept that a position in our alignment has a mutation or do we introduce a gap?

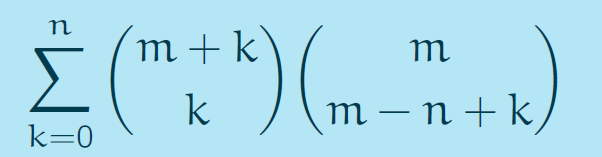
**What strategy can be adopted to tackle this tradeoff?**

Assign costs for different actions in the alignment process.

**How many possible alignments are there between sequence a and b, each of length m and n, respectively?**

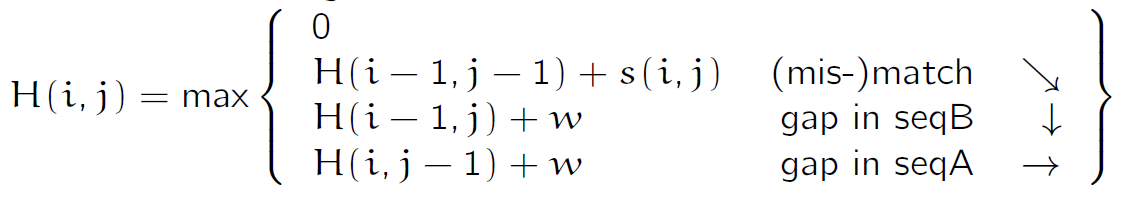
If k is the number of gaps that can be introduced, then there is (m+k k) possibilities to place these gaps between the m letters of a.

The total number of possible alignments is:



**What are the three steps in the S-W algorithm?**

1. Initialization: 0th row and 0th column set to 0, remaining rows and columns correspond to nucleotides.
2. Iteratively calculate the score H(i,j) of the optimal alignment with ai and bj at the end of the field (i,j) and keep track from where the optimal alignment comes according to:



s(i,j)=3 if ai=bi

s(i,j)=-1 if ai≠bi

w=-2 (gap penalty)

1. Start from the highest number, which corresponds to the best local alignment and walk backwards until 0 is reached.

**What are some of the advantages of the S-W algorithm?**

Pros:

* Fast w.r.t. brute force search -> O(mn)
* Finds optimal local alignment or one of the optimal alignments if there is more than one with the same highest score.

Cons:

* S-W works only for local alignment
* Only pairwise alignments are possible
* Too slow for scanning big libraries

**What are the changes to the S-W algorithms to transform it to the N-W algorithm for global alignment?**

1. Initialization: H(0,j)=j x w and H(i,0)=i x w
2. Has to start from the bottom right field (m,n) and follow until the top left field (0,0).

**What is the most widely used heuristic approaches to find alignments?**

BLAST: Basic Local Alignment Search Tool.

**What are the steps in the BLAST algorithm?**

1. Split the query sequence into subsequences of length k.
2. Search these k-letter words in the database sequences. Similar words are allowed but scored, e.g. match +5 and mismatch -3.
3. Keep only the sequenes with the highest scores
4. Expand the k-letter word to the right and left, note the scores
5. Stop if the score drops below a certain threshold.
6. Keep only the pairwise alignments that are above the threshold.
7. Report sequences.

**What are the (dis)advantages of BLAST?**

pros:

* Up to 50-100x faster than direct alignment
* Allows searches for exact matches but also for similarity up to a pre-defined degree

cons:

* Does not guarantee the optimal pairwise alignments of the query and database sequences
* Limitation: can only find a match when the gene/sequence is available in the database

**When do we want to perform MSA?**

* Reconstruct the evolutionary history of individuals in a phylogeny
* Assess the sequence conservation of proteins

**What are the approaches that allow us to perform MSA?**

Heuristic, computationally demanding.

1. Ad hoc approach: pairwise alignment against a reference sequence.s
2. Arrays: perform SW on multiple dimensions (!: extremely slow, k sequences of length m requires mk steps)
3. Other dynamic programming based approaches: ClustalW, Muscle, Malign.

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

SW/NW

**Do you obtain the same alignments when using different scoring schemes in the S-W and N-W algorithms?**

No, but it’s possible.

# Nucleotide substitution models

**What is the RQ that GWAS addresses?**

Can we identify genetic risk factors for common diseases?

**What is the most widely used strategy?**

* Two large groups of individuals: one healthy control group and one group with a certain disease.
* All individuals are genotyped for the majority of known SNP locations.
* For each SNP, count the number of healthy individuals without this specific mutation and with the mutation as well as the number of diseased individuals without the mutation and with the mutation.
* Calculate the odds ratio of diseased and healthy people w.r.t. the abundance of each SNP.

**What does this OR tell us?**

This can only inform us about a potential association between an SNP and a genetic disease.

**How can we test for the significance of the OR?**

We perform a contingency table test (Fisher’s exact test or Pearson’s chi^2) to calculate the p-value.

**What is the definition of the p-value?**

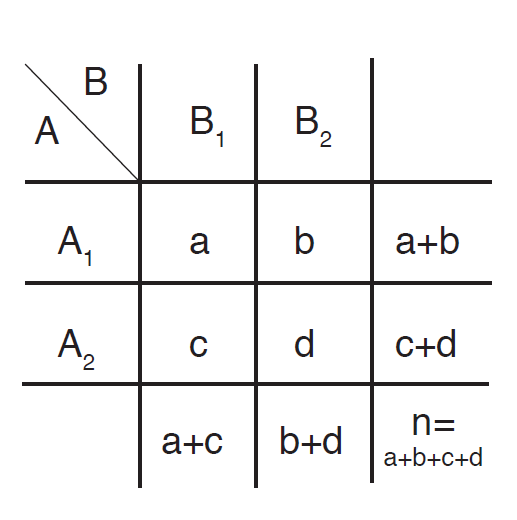
Given a random variable X and an observation x. Let us assume a null hypothesis H0, which is a statement on the distribution of X. The p-value is the probability of observing x or a more extreme realization under the assumption the null hypothesis was true. P(X=x or more extreme|H0)

**What is the definition of the significance level?**

The significance level alpha is a value that one defines before performing a statistical test. If the p-value lies below this significance level, the observed result of the r.e. cannot support the null hypothesis.

**How is Fisher’s exact test performed?**

* A contingency table is constructed



* Our hypothesis is that we want to test whether class A is linked to class B. H0: The random variable expressing A1 and B1 is distributed according to a hypergeometric distribution.

**What are the pitfalls for Fisher’s exact test?**

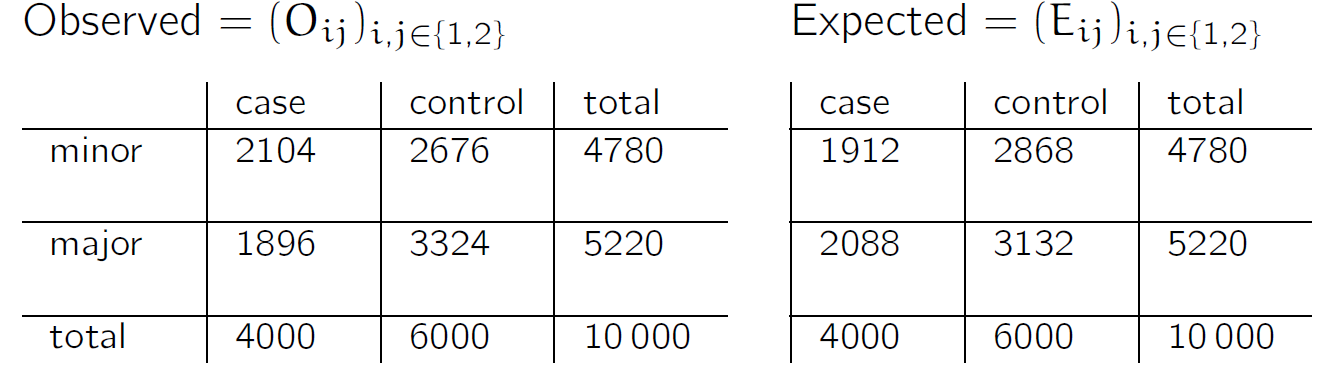
It only works for small numbers, n choose k cannot be computed for larger numbers.

**How does Pearson’s chi^2 test compare to Fisher’s exact test?**

It was invented for bigger numbers. It’s used to calculate the deviance between observed and expected numbers based on a hypergeometric distribution.

**How is the test performed?**

We compared observed vs. expected results:

****

Then we compute the following test statistic:

Which should be chi^2-distributed. The p-value is then the probability of a chi^2-distributed random variable being higher or equal to x.

**What are the criticisms that GWAS receive today?**

* Missing QC steps, e.g. test for biases for tested group structure
* Multiple testing correction required
* Correlation only between single SNPs but not between genes tested.
* Methogological problems have been addressed and used to identify over 4,000 SNPs.

**Why do we want to quantify variation between sequences?**

In order to construct the phylogeny between sequences.

**What is the distance between two sequences?**

The number of differing nucleotides substitutions per site.

**What are the requirements for the mathematical model estimating the distance?**

* Stochastic process modelling the substitution through time.
* Substitution rates.

**How can nucleotide substitution be modelled?**

They can be modelled as a Markov chain:

* It’s a stochastic process
* Lives on a state space and jumps to the different states
* Is memoryless: the probability of jumping to a state only depends on the current state.

**When is a Markov chain called time homogenous?**

If the transition probabilities on the state space do not change over time.

**What is the difference between a rate and a probability?**

Rates: measures events per time unit

* Rates are fixed, deterministic quantities.
* Describes averages

Probabilities: measure of chance that a random event occurs

* Stochastic
* Describes an exact event

**How do rates relate to probabilities?**

**Assume we have the transition matrix Q, how can we infer probabilities from it?**

**What does**  **mean?**

It means that every possible series of states is visited in time t.

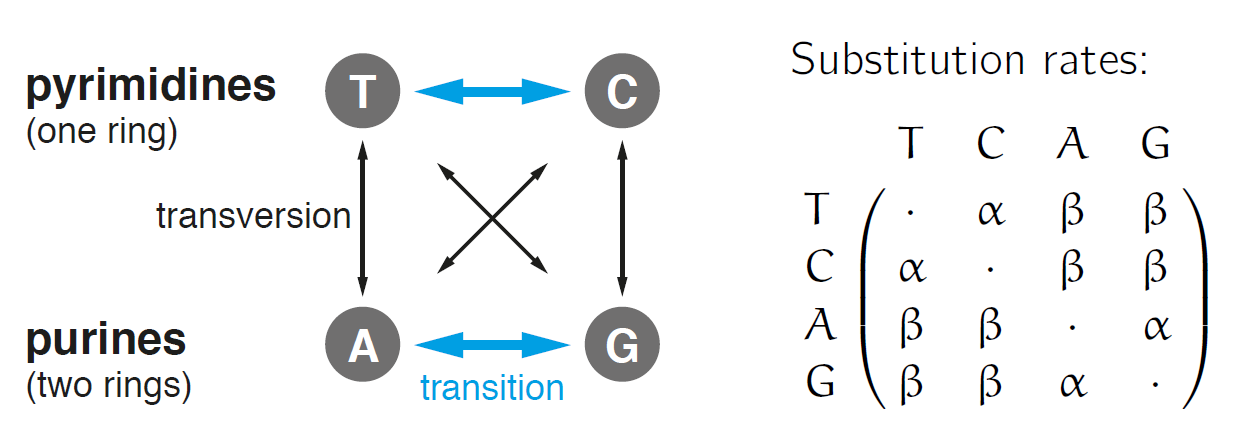
**What are Markov chains a great model for nucleotide substitutions?**

1. They are memoryless: a nucleotide substitution happends independently from the substitution history at this site.
2. Substitution rate matrix defines the transition probabilities
3. The transition probabilities take into account every possible substitution path.

**What is the JC69 model?**

All substitution rates are the same, lambda. Probability of staying the same is 1-3lambda

**What is the K80 model?**

* Transitions: pyrimidines-pyrimidine & purine-purine: alpha
* Tranversions: pyrimidines-purine & purine-pyrimidine : beta

**What is the TN93/HKY model?**

Adds to the K80 model, multiply each rate of K80 by nucleotide equilibrium frequency. Additionally, T-C & A-G happen at different rates. HKY considers alpha1=alpha2.

**What is the GTR model?**

It’s the generalized time-reversible model.

Pros:

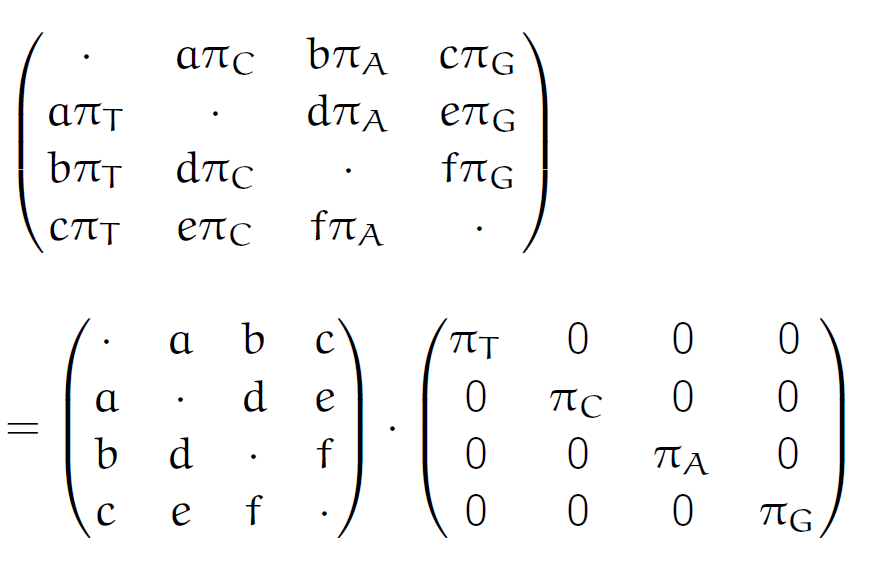
* Flexible
* Time-reversible

Cons:

* Not completely general

**What does time-reversibility enable?**

Makes calculations and simulations easier.



**What is the most general substitution model?**

It’s the unrestricted mode, where each substitution has a different rate, not handy to use & not time-reversible.

**In a GWAS, why can you not reject your null hypothesis if the p-value is <alpha?**

Because of multiple testing, this is called a Bonferroni correction.

**Why is it not advisable to reconstruct a phylogeny based on the Hamming distance?**

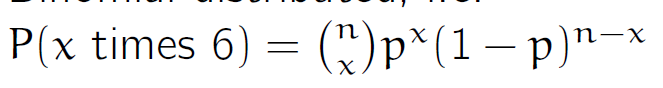
Does not include substitution rates or stochasticity.

# Nucleotide, amino acid and codon substitution models

**What is the maximum likelihood estimator?**

It is an estimator of a model parameter that maximises the probability to obtain the observed results.

**How is it obtained?**

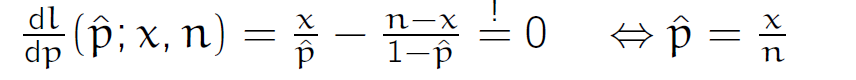


**How can the log likelihood be derived from this example?**



**How can the maximum be found?**

Set first derivative to 0:



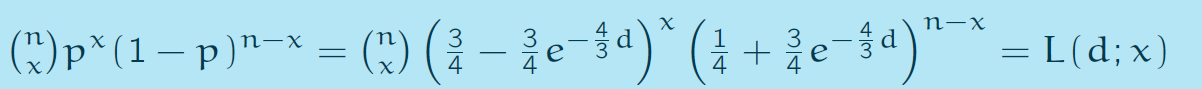
**How can the CI of the MLE for the parameters of the distribution of a random variable X be found?**

Under certain conditions, 2(l(theta hat)-l(theta)) is chi^2\_k distributed, where k is the length of theta.

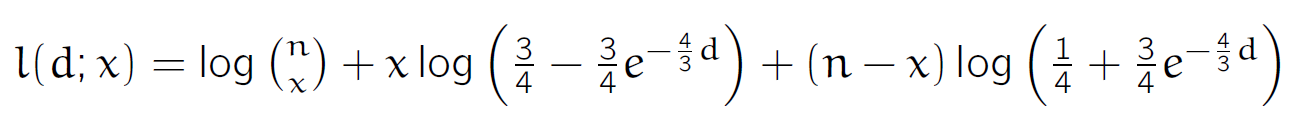
1. Determine the value of l(theta hat; x)
2. Substract 0.5chi^2\_(k,5%)
3. Determine theta value for which

l(theta;x)=l(theta hat;x)-0,5chi^2\_{k,5%}

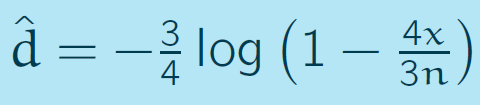
**What is the MLE for the sequence distance under the JC69 model?**



Log likelihood:



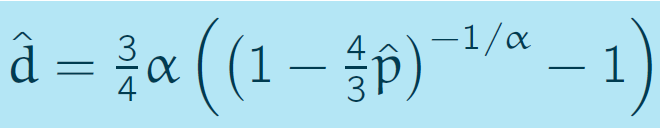
MLE of the JC69 distance:



**Why could the JC69 model be supplemented with a gamma distribution?**

To encode for the variability in substitution rates.

**What is the distance function under the JC69+gamma model?**



This avoids ignoring site variation, which leads to an underestimation of the sequence distance.

**What differs when trying to model amino acid substitution?**

Q needs to be 20x20 -> increases computational complexity.

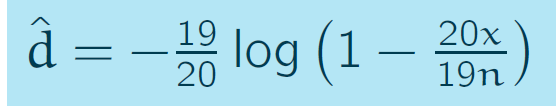
**What are the two approaches that can be adopted in AA substitution modelling?**

Empirical

Mechanistic

Desired: time-reversibility.

**How does the JC69 distance look like when doing AA substitution?**



The distance can be estimated with an empirical or mechanistic Q matrix.

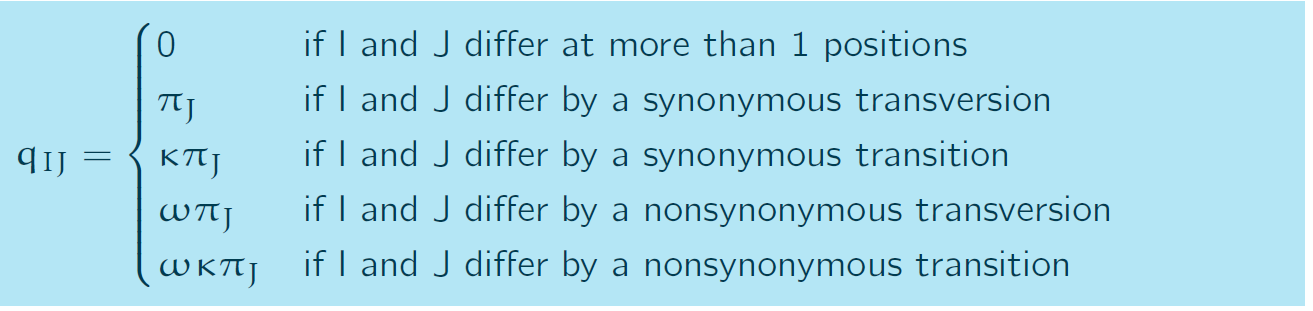
**What changes when doing codon substitution models?**

There are multiple codons leading to the same AA, which are empirically more likely to occur. Codon changes leading to the same AA are called synonymous substitutions. Non-synonymous substitutions are when the resulting AA changes.

Transversions and transitions can also be modelled, making some of the codon changes more likely. Also, equilibrium frequencies can be incorporated.

**What are the dimensions of the substitution matrix in the case of codon transitions?**

61x61. No stop codon included. (43-3)



**How can selection be uncovered from these substitutions?**

Comparing the amount of nonsynonymous and synonymous substitutions.

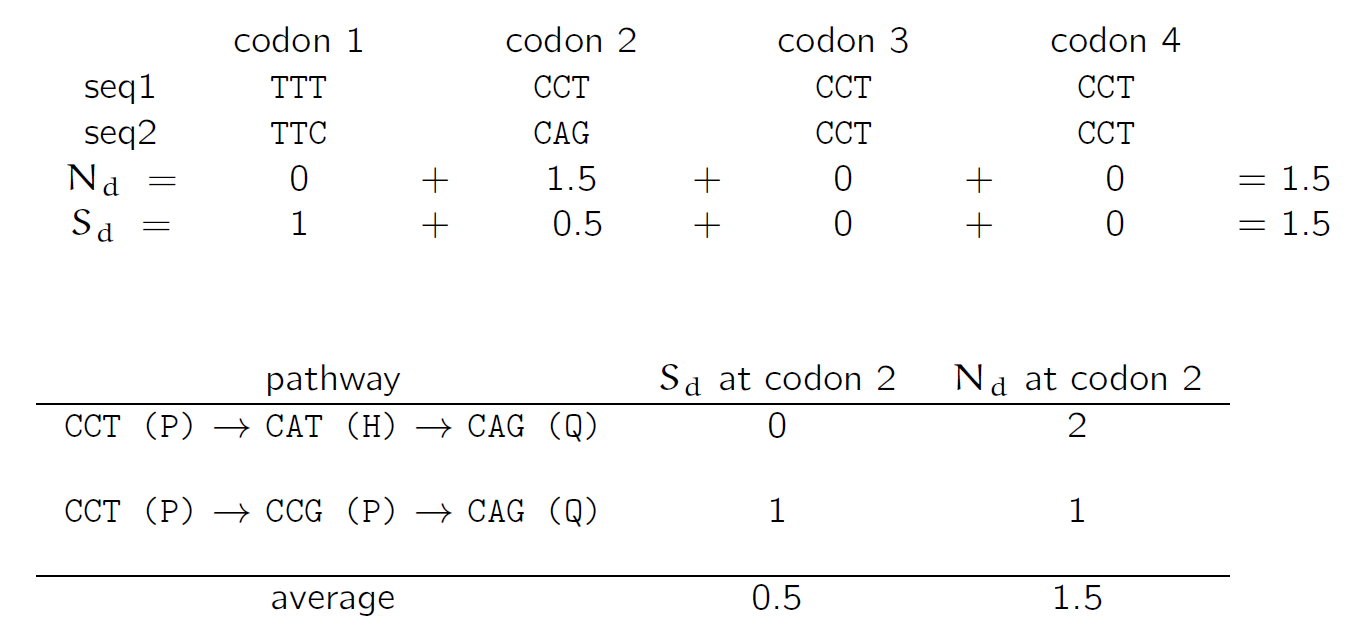
dN: distance at nonsynonymous codon positions

dS: distance at synonymous codon positions

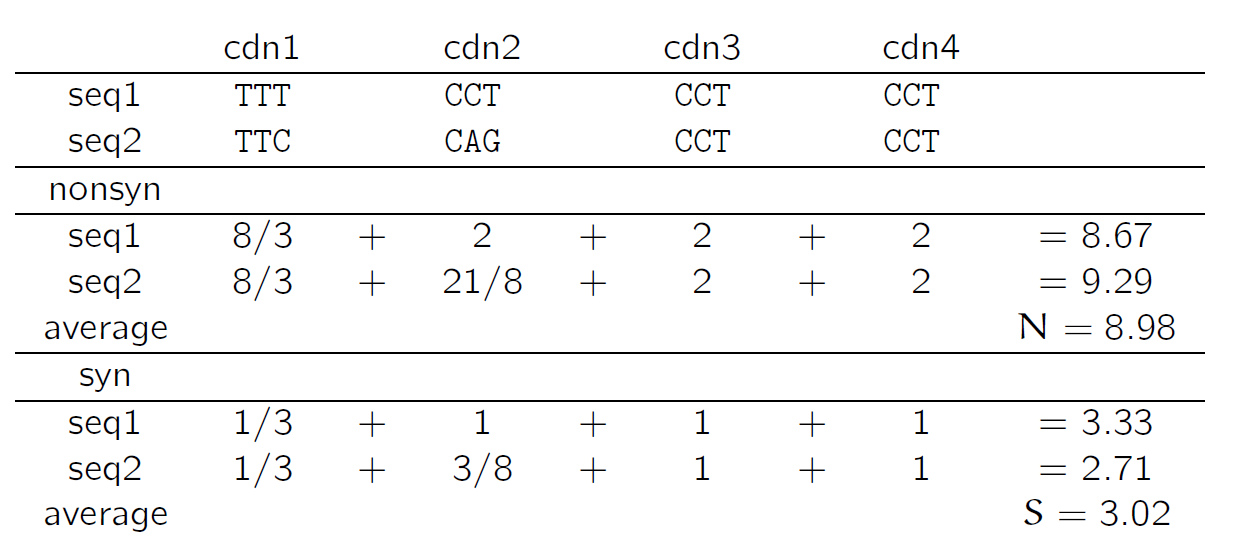
* dN/ dS

Then use dN/ dS counting method:

1. Count the (non-) synonymous differences
   1. Taking into account all possible pathways!



1. Count the (non-) synonymous sites



1. Account for evolutionary history

**How can we interpret the** dN/ dS **ratio?**

dN/ dS<1: nonsynonymous mutation occur less frequently than synonymous mutations (purifying selection)

dN/ dS>1: nonsynonymous mutation occur more frequently than synonymous mutations (positive selection)

**What is the difference between the JC69 and the TN93 models?**

TN93 Considers transitions, transversions & equilibrium frequency, compared to JC69 which does not.

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

TN93 probably, because more realistic.

**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, will cluster sequences differently.

# Phylogenetics: UPGMA and least squares

**What is a tree?**

It’s a graph of nodes and branches without cycles

**What is an unrooted phylogenetic tree?**

It’s a tree with two types of nodes:

* Tip/leaf: node with 1 branch attached
* Internal node: Node with 3 branches attached.

**What is a rooted phylogenetic tree?**

It’s an unrooted tree in which one branch is devided by a new node (root)

**How can unrooted trees be rooted?**

If one adds an outgroup. It means that the branch ending in B is subdivided by the root node. B is chosen by the user as a very distantly related organism to the remaining organisms in the tree.

**What is one condition concerning branches which is required in trees?**

Branch length needs to be positive.

**What is a pendant branch?**

It’s a branch attached to a tip

**What is a cherry?**

It’s a pair of tips which are only separated by one internal node.

**What is a caterpillar tree?**

It’s a rooted tree which only one cherry.

**What is a monophyletic group/clade?**

All ascendants of an internal node in a rooted tree.

**What is an ultrametric tree?**

It’s a rooted tree where the sum of branch lengths from any tip to the root is the same.

**What is a polytomy?**

The definition of a phylogenetic tree is extended such that internal nodes may have more than three branches attached. Such a node is a polytomy. It can be represented as a classic phylogenetic tree with branch lengths of 0.

**How can a rooted Newick tree be constructed?**

Recursively:

* Choose two tips C and D that form a cherry
* Replace the two tips by the new tip C: tx, D:tD where tx is the length of the branch ancestral to the node X.
* The length of the branch ancestral to the new tip is the branch length ancestral to the cherry.

**How can an unrooted Newick tree be constructed?**

* Choose an internal node arbitrarily
* Proceed as in the rooted tree towards the chose internal node
* Connect the three last tips X, Y, Z to (X:tx, Y:ty, Z:tz)

**What properties does a Newick tree have?**

Each tree does not have a unique Newick representation.

**What needs to be respected when representing trees?**

The distance along the evolutionary time axis needs to be respected.

**How are phylogenies reconstructed?**

Similar individuals are clustered together according to sequencing data. Similarity may be defined in different ways:

* Phenetic:
  + Based on overall similarity
  + Pairwise distance based
  + Methods: UPGMA, least squares
* Cladistic:
  + Based on shared characteristics
  + Character-based
  + Methods: parsimony
* Mechanistic:
  + Based on evolutionary model
  + Character-based
  + Methods: ML/Bayesian inference.

**What is the basic idea behind phenetic approaches?**

1. Define how to measure distance between sequences
2. Calculate distances between all pairs of sequences
3. Find a tree where the distances, i.e. branch lengths between the pairs of tips most closely follow the sequence distance matrix.

**What are the two strategies to reconstruct the three in phenetic approaches?**

* Algorithmic approach (UPGMA): sequences separated by the smallest distance are clustered iteratively in a tree.
* Optimality approach: minimise the difference of the pairwise sequence distances to the pairwise distances between the tips in the tree, calculated by summing the length of branches between two tips.

Note: only distances between pairs of sequences are used but not any higher order correlations between sequences.

**What is the I/O of UPGMA?**

Input: distance matrix.

Output: the ultrametric phylogenetic tree

**What are the assumptions behind the UPGMA algorithm?**

* All sequences must come from the same time point
* The algorithm assumes evolution according to a strict molecular clock: the rate of DNA/RNA/protein sequence evolution is constant over time

**Which algorithm relaxes these assumptions?**

Neighbour-joining algorithm, where the branch length corresponds to the number of mutations. The output tree will be unrooted.

**What are the computational steps of the UPGMA algorithm?**

Initialize the size of each node si as ni=1.

While the distance matrix is not empty, iterate:

1. Choose nodes si and sj such that d(si,sj) is the smallest entry in the distance matrix, in case of several minima, choose one at random.
2. Coalesce si and sj to node sij with size ni,j=ni+nj. The branch length between si , si,j and between sj , si,j is chosen such that all tips descending from si,j have the same distance d(si,sj)/2 so si,j
3. If the distance matrix includes more than 2 nodes, include si,j into the distance matrix with

Where is a node in the distance matrix

1. Delete nodes and from the distance matrix.

**What is the formula for the optimality objective in the least squares method?**

It’s the squared difference between the sequence distance matrix and the tree distance matrix:

Where D is the between sequence distance matrix, d the tree distance matrix for the proposed tree and w weights. It can be a matrix of 1 or inverse proportional to Di,j.

**What is the I/O of the tree?**

Input: distance matrix

Output: tree with smallest S

**What is the algorithm to be used?**

Repeat until all tree topologies were proposed:

1. Propose an unrooted tree topology (without branch lengths)
2. Minimise S by optimizing the branch lengths.

**What is the computational complexity of UPGMA?**

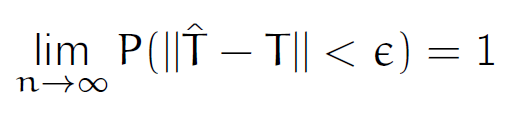
O(n3) for n sequences.

**What is the computational complexity of the least squares method?**

(2n-5)!! + (2n-3)!! -> NP-hard, because each tree architecture needs to be tested.

**What is statistical consistency?**

A phylogenetic reconstruction method is statistically consistent if the true tree is returned for an infinite amount of data, i.e.



**Are UPGMA and least squares consistent?**

Yes.

For UPGMA, the distance matrix tends towards the tree distances. If the sequence distance matrix equals the tree distance matrix, then the true tree is reconstructed.

For least squares, the squared difference between the calculated distance matrix and the tree distance tend towards 0, thus the tree is a least squares tree. Each tree has a unique distance matrix, thus the true tree is the least squares tree.

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

Min: 1.

Max: (n-1)/2

**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?**

2, 2^(n-1).

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

The longer the evolution time, the longer the distance and the higher the variance in distances. A better weight for distances that are short could be a nice alternative. Inverse of distance could be used as weights.

# Phylogenetics: Cladistic tree inferences

**What do cladistic approaches consider?**

They consider the similarity as measured by shared characters in the alignment sequences.

**What method is used to use cladistic approaches?**

The parsimony method: find the tree that needs the minimal number of mutations to get the alignment.

**What are the elements of the parsimony method?**

* Parsimony score of a tree: the minimal number of mutations required to explain the sequences at the tips of the tree.
* Most parsimonious tree: the tree with the minimal parsimony score.

**What is a relationship between unrooted trees and rooted trees in the context of the parsimony method?**

Rooted trees obtained from the same unrooted tree have the same parsimony score.

**What is the I/O of the parsimony tree reconstruction method?**

Input: sequence alignment of n sequences with sequence length m.

Output: Unrooted tree with the lowest parsimony score.

**What is the algorithm for parsimony tree reconstruction?**

* Consider each unrooted tree (2n-5)!!
* Calculate the parsimony score for the unrooted tree, which requires 4n-1m internal sequence alignments. This can be improved by using the Fitch algorithm.

**What is the Fitch algorithm?**

It’s an algorithm to quickly determine the parsimony sore of a tree.

**What is the I/O of the Fitch algorithm?**

Input: Unrooted phylogenetic tree and an alignment of n sequences of length m, corresponding to the n tips of the tree.

Output: Parsimony score k of the tree, i.e. the minimal number of mutations required to explain the sequences at the tips.

**What are the steps in the Fitch algorithm?**

1. Root the tree at an arbitrary edge
2. K=0
3. While the root has no sequence assigned, iterate:
4. Choose a node in the tree where all descending nodes have sequences assigned
5. Assign a sequence to the chosen node. For i=1,…,m:
6. Let Cl and Cr be the sets of nucleotides assigned to the two direct descendants of the chose node for site i.
7. If Cl inter Cr is not empty, assign Cl inter Cr to nucleotide I of the chosen node
8. If C inter Cr is empty, assign Cl union Cr to nucleotide i of the chosen node and set k = k+1

**What is the computational complexity of the Fitch algorithm?**

(n-1)m

**How is the parsimony tree found?**

By calculating the parsimony score on each unrooted subtree = NP-complete.

**Is the parsimony method statistically consistent?**

No, imaging there are two nucleotides (0 and 1) with the same rate of transition 0->1 1->0. The probability of change in the long branch is p and the probability on the short branch is q. Parsimony infers the wrong tree if p,q are in the Felsenstein zone, i.e. p2>q(1-q).

It’s also a phenomenon referred to as “long branch attraction”.

**What is the I/O of the maximum likelihood inference?**

Input: sequence alignment

Output: tree which maximises the probability of the sequences given the tree & the sequence evolution parameters:

* Requires an evolutionary model
* Parameters of the evolutionary model are co-estimated with the tree.

**What is the objective of MLE in the context of phylogenic inference?**

What is the maximum likelihood phylogeny given the sequence data?

**What is the parameter in the context of MLE phylogenetic inference?**

The trees.

**What is L(T, Q; D) := P(D|T,Q)?**

It’s the likelihood function of the parameters T, Q for the given sequence data.

**State the inference problem.**

Determine T, Q which best explains the alignments observed, i.e:

We determine the best tree by evaluating the likelihood for many different proposed trees.

**What is the likelihood calculation for a single tree?**

For an alignment of length m:

Where sk,j is the site j of sequence sk. Then, is defined as:

can be evaluated by calculating for each branch l (with starting sequence sl1 and ending sequence sl2 , and branch length tl) the transition probability from the ancestral nucleotide to the descendant nucleotide. The root nucleotides are then weighted by their equilibrium probabilities pi, so that:

**What is the runtime of the MLE?**

O(m4nn)

**What allows us to speed up calculations?**

Felsenstein’s pruning algorithm. Make the summation more efficient, i.e. polynomial in n through dynamic programming. At tip k, if X is the observed nucleotide, otherwise 0 . Cherries are pruned recursively towards the root. Let k be a node with the descendants l,m.

**What is the probability of the sequences at site j?**

**What is the runtime of Felsentein’s pruning algorithm?**

O(mn), only bottleneck is finding the optimal tree structure, which is NP complete.

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

No, counter-example possible.

**Does the ML tree reconstruction method return estimates for the internal sequences? Give a reason for your answer.**

No, we only have P(Dk|X), not P(X).

**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 possible to always find the smallest number of mutations required.

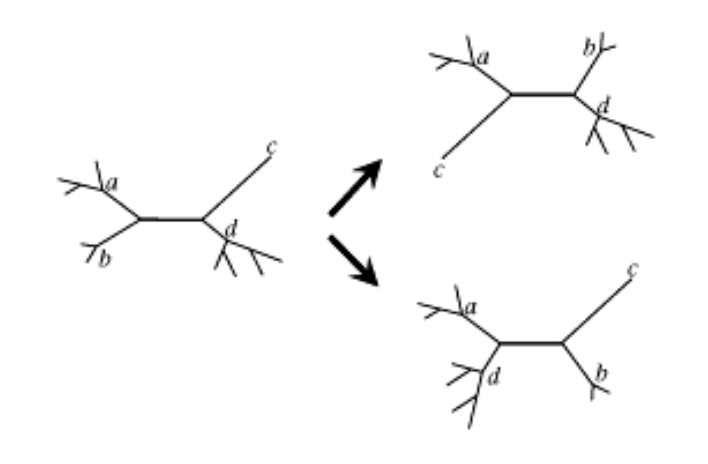
# Maximum likelihood method & testing

**How can we search the tree space for the maximum likelihood tree?**

Use methods like NNI, SPR, and TBR for branching structure. Propose different branch lengths. Use hill-climbing strategies to find the optimum.

**What is the nearest-neighbour interchange (NNI) algorithm?**

Each internal branch in the tree connects four subtrees or nearest neighbours. Interchanging a subtree on one side of the branch with another on the other side constitutes an NNI. Two rearrangements are possible for each internal branch.

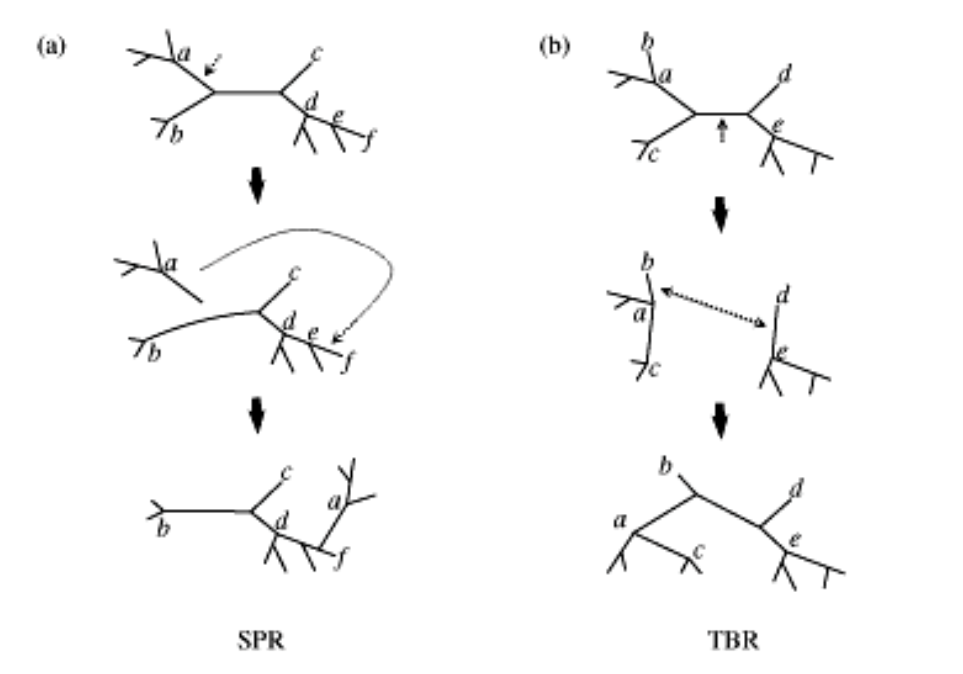


**What is subtree pruning and regrafting (SPR)?**

A subtree is pruned, and then reattached to a different location on the tree.

**What is tree bisection and reconnection (TBR)?**

The tree is broken into two subtrees by cutting an internal branch. Two branches, one from each subtree, are then chosen and rejoined to form a new tree.



**What question is answered by model testing?**

Which evolutionary models are appropriate for our data?

**What are the ways by which we can answer this question?**

Via likelihood ratio tests & AIC.

**How can we assess our confidence in the phylogeny and substitution rates?**

Via

* Likelihood ratios
* Bootstrapping (only method not requiring ML)

**How can we perform likelihood ratios for models?**

* Assume the ata evolved under a model H0
* Assume model H1 within which H0 is nested
* For the given data, let the ML parameter estimate under the model H0 be θ0hat and under H1 θ1hat.
* Assume large amount of data, then:

Where df is the difference in the number of parameter in the general and in the nested model. If that difference is either infinity or 0, then choose 0.5.

* Note that: