When are two sequences homologous?
When they evolved from the same ancestral sequence. They are orthologs, if this sequence was in the most recent common ancestor (MRCA) of the organisms that harbor the sequences you are comparing; they are paralogs, when the ancestral sequence predates the MRCA.

Do all sequences that are homologous show significant sequence similarity.
No. Many homologous sequences have been saturated with substitutions. In many of these instances PSI blast or HMMER searches succeed in identifying homologs.

If two sequences (that do not contain regions of low complexity) show significant similarity in their primary sequence, are they homologous?
Yes, provided the significance is accurately assessed. (E-value, multiple tests, PRSS)

Are orthologous sequences homologs?
Yes

Can a sequence in one organism have many orthologous sequences in another organism?
Surprisingly the answer is yes. In case of the in-paralogs in globin gene evolution, all the animal hemo and myglobins (while paralogs of one another) are orthologous to the leg-hemoglobin molecule in plants.

Two sequences (each 300 amino acids long) in a pairwise sequence alignment have 70% identical residues (distributed rather evenly along the sequences), another 15% of the residues show conservative substitution. Which is correct?
A) the sequences are 70% homologous
B) the sequences are 85% homologous
C) the sequences are homologous
D) There is a good chance the sequences are not homologous.

What is the difference between a global and a local alignment.
Local: find significant matches within two sequences (e.g., A and B). This ignore the non-matching parts of the two sequences. Also, one sequence within sequence A can match to several sequences contained in B (e.g., the proteolipid example).
Global: Align two sequences from beginning to end. There are different strategies to handle the non-matching parts. The human brain seems to prefer to align these, even if there isn't a good match, the alternative to align the non-matching stretches to gaps in the other sequence. Looks ugly, but creates fewer artifacts in phylogenetic reconstruction. (clustal and muscle tend towards the former; prank tends towards the latter)

In a blast search, how is the chance expressed that a match between a query and a database sequence is a false positive?
This is the E-value, or rather if the E-values is small it gives the chance that the match is a false positive. (E-values can be larger than 1, the probability to be a false positive can never be larger than 1).

In blast searches, is the E-value of a match propositional to the database that was searched to obtain the match?
Yes.

What does the abbreviation "ssh" mean, and what can you do with ssh?
Secure shell – connect to a computer and interact with it using the command line (or using an X-windows interface)

You have several sequence files in your directory that have very, very long filenames (GCF_000196515.1_ASM19651v1_cds_from_genomic.fna GCF_000020025.1_ASM2002v1_cds_from_genomic.fna and GCF_000009705.1_ASM970v1_cds_from_genomic.fna). You want to copy the contents of these files into a single file. How can you do this without typing the very long names at the command line (and without using a Graphics User Interface)?
Use wildcards (e.g., cat *.fna > all.fna) or automatic line completion

What is automatic line completion, and how can you invoke it?
Unix completes the command line in the only way possible, if there are alternative, it can list the alternatives. You invoke this by pressing the tab-key once or repeatedly.

At the command line you start typing a command, e.g. `"cd ~/D"`. You then hit the <tab> key on your keyboard. What will happen?
If you have a folder called Desktop in your home directory, the line will be completed to
`cd ~/Desktop/`

You do a pan genome analysis of 28 genomes.
One gene family has members in all 28 genomes, the genes in this family would be considered to be part of the _core or strict core genome of the group_
One gene family has members in 10 genomes, genes of this family would be considered to be part of the _pan-genome_

One gene family has members in 27 genomes, these genes would be considered to be part of the _pan-genome_ or of __the relaxed core__

What are possible reasons for the recombination events within a genome to frequently occur between sites equidistant to the origin of replication.

A) Recombination might occur at the same time as replication
B) Other recombination events might misplace terminus of replication relative to the origin of replication, or they might place AIMES in the wrong position relative to the origin and terminus of replication.

In comparing two genomes, what are differences between a mummer and a gene plot (Which is based on nucleotide sequences, and which on encoded proteins? Which provides information on the encoding DNA strand? Which is better in detecting paralogs?
Mummer:: based on nucleotide sequences; provides information on the encoding DNA strand

gene plot: based on encoded proteins; better in detecting paralogs.

What is strand bias?
The two DNA strands (either called leading and lagging, or + and -, or Watson and Crick strand) have different composition.

G pairs with C and A pairs with T, how can there be a bias in Gs and Cs?
The double stranded DNA has no bias, but the individual strand can.

Why is plotting cumulative bias less affected by noise that plotting the G over C bias in a window?
Through plotting the cumulative bias, one integrates over the bias. Thus the noise is turned into variation of the slope.

Why is a PSI blast search more effective in finding distant homologs than a normal blast search?
Using many homologous sequences, this approach focuses on key conserved residues, and also allows for site specific substitution patterns (e.g., in one site Ser and Thr frequently replace one another, another site might have frequent substitution between Ser, Thr, and Cys).

Why do PSI blast searches sometimes have a problem with estimating the probability of false positives in a useful way?
The PSSM can be corrupted by a non-homologous sequence

Why is % identity between a query and a match in the database not a good criterion to evaluate the significance of a match?
Short matches can have a high percent identity without being significant.

You write a script that for each entry in array (e.g., a nucleotide in a genome) performs an activity (e.g., increasing the counter for this base by 1). You start the loop with
foreach (@array) { }
What is the standard name for each array entry inside the loop?
$_

You want to perform a set of steps in a program only if two numerical variables ($a and $b) have the same value. Which is correct:
if ($a = $b) {some commands}
if ($a == $b) {some commands}
if ($a eq $b) {some commands} (this works too, at least most of the time, but it considers the variable as a string, not a number

Which character at the beginning of the line denotes a comment line in perl?
(except for the shebang, which also starts with #, and tell the operating system where to look for the interpreter or shell)

Which character at the beginning of the line denotes the name/annotation of a sequence in a fasta formatted sequence file?

>

What is a potential problem with progressive alignment programs (e.g., clustalw)?
The alignment biases future phylogenetic reconstruction in favor of the alignment order.

If an intron does not have a length that is a multiple of 3, what problem does this cause for alternative splicing?
If the next exon is skipped, the following exon would be in a different reading frame.

Briefly describe the following approaches to phylogenetic reconstruction from aligned sequences:
Distance matrix analysis using distances that were corrected for multiple substitutions
Calculate the distance between all pairs of sequences in an alignment (usually all sequences are aligned in an MSA (=multiple sequence alignment), usually one corrects the distances for multiple substitutions (% divergence is not additive). The correction can be quite complex, as in maximum likelihood distances. Then calculate a tree in which the distances *match* the distances in the pair wise distance matrix. (match, could mean to have a weighted minimal error, or the local minimum evolution as in neighbor joining, or any other measure to describe the fit. A problem of the non-algorithmic approaches is that usually one cannot be certain that one has indeed found the tree that best fulfills the chosen criterion – this is also true for the next two …).

Maximum parsimony analysis
Given a multiple sequence alignment find the tree that can explain the MSA with the least number of substitution event. Given a tree and an MSA, one can calculate the minimum number of necessary substitutions, however, there is no formula to calculate the most parsimonious tree, i.e. one needs to search tree-space to find the most parsimonious tree, and usually this search is heuristic and not guaranteed to find the most parsimonious tree.

Maximum likelihood analysis
Instead of just counting steps, this approach gives different weight (probabilities) to different events. These probabilities can either be provided by the user as a model of evolution, or these probabilities can be estimated from the data (in case of the latter, the parameter is chosen that maximizes the probability of the MSA). Tree space (and possibly parameter space) is searched to find the tree under which the MSA is most probable.