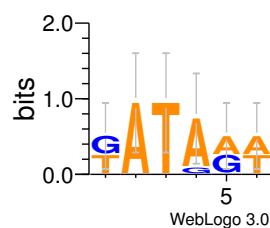


Suggested solutions to exam March 12, 2009, in DD2396 Bioinformatics

Lars Arvestad

Part 1

- (a) True. Multialignment programs are, by necessity, heuristics which means they must take "shortcuts" in computing an answer. The optimal result is sometimes out-of-reach due to the search space.
 - (b) False. Masking is used to "hide" low-complexity regions.
 - (c) True. TM proteins have a fairly constrained structure and have relatively strong statistical signals for important structural features.
 - (d) False. Synteny data can help corroborating homology, but is in itself not sufficient since genes move around, duplicate and get lost. You have to look at sequence similarity first and foremost.
- (a) The principle is that the number of mutations that the phylogeny implies is as small as possible. Some formulate this as finding as short of a tree as possible.
 - (b) Short sequences does not always have enough mutations in them to give a reliable statistic signal. Single mutations will have too much impact on the estimation.
- (a) Both shotgun sequencing and EST sequencing gives data with what we can describe as different starting points. If we expect sequences to overlap rather than cover an orthologous region, then we do not want to penalize for differences in the ends.
 - (b) Local alignment.
- (a) Coverage is the average number of times a single position has been sequenced.
 - (b) A contig is a sequence created by merging several overlapping reads.
 - (c) For paired-ends reads, or mate pairs, we know the approximate distance between pairs of reads.
- Sequence logo, as created by WebLogo: The features I want to see in solutions are:



- Conserved columns are higher than variable columns.
 - The main characters are indicated in the columns.
6. APOD_HUMAN is the most significant hit. The number of significant hits depends on how strict you are. If we use the common, but somewhat arbitrary, rule of thumb of thresholding at 10^{-5} there are seven significant hits. Other answers, if justified, are also accepted.
 7. By translating the compared sequences, we essentially discard so-called silent mutations that do not change the amino acid. We can also find that wildly different codons give related amino acids. In all, we get a more sensitive comparison.

Part 2

8. Multifurcations in bootstrap analysis means that there is not statistical signal in the input sequences to determine the actual branching order, at least not using bootstrapping.
9. Please see the course book or the chapter by Anders Krogh that we have been using. The sequence features I am expecting are coding regions, start and stop codons, introns, acceptor and donor sites.
10. Any alignment of two random sequences will give a score according to some random distribution. With thousands of comparisons, as in the question, we are likely to get approximately the same best score every time. Therefore, the best score is really expected to occur every time, meaning that the E -value of that comparison will be around 1.
11. A common methodology, used by many projects, is to perform an all-against-all Blast comparison. This means that you make a Blast database for all your sequences and then blast the same set of sequence against that same database. All hits above some (significant) score threshold (and sometimes an alignment length threshold) are noted, and other pairs are discarded. Then, proteins that directly or indirectly are linked by the noted hits are grouped as a *family*.

Pro: We get homologous groups using a relatively fast and straightforward method. Significantly similar pairs are guaranteed to be in the same family.

Cons: Multi-domain proteins can cause problems and link together proteins that are only partially homologous. It is also very difficult, if not impossible, to choose a scoring threshold which is neither too specific nor too including.

12. (a) GBlocks and TrimAl remove columns that are either too variable, according to some heuristically chosen criteria, or too gappy (as in: too many indels).
These columns are supposed to be so variable or unreliable that they do not contribute any information to a phylogenetic estimate.
- (b) If you are working with a small dataset (one protein family, for example), then you are supposed to make a manual review of the alignment before you proceed. Obviously misaligned sections can be corrected or removed.
Such work cannot be done on large datasets containing thousands of alignments. An automatic procedure that simply removes parts of an alignment when there are serious doubts on the quality is likely to improve end-results.
- (c) Suppose we have an alignment over a set of sequences which we can split into two subsets, S_1 and S_2 , and that S_1 is generally variable, but S_2 is quite conserved with very few changes. If the S_1 part of the alignment is extremely variable and uncertain in the part where S_2 is most of its variability, then we might unwillingly remove the parts needed to resolve the S_2 clade.
- (d) There are many ways of doing such an evaluation. Here is one suggestion.
I would want to make simulate evolution to generate a large set of multialignments based on known (constructed) phylogenies. We would then run both programs to produce a new, reduced, alignment before a phylogenetic method is applied. We can then simply count how often we recover the correct phylogeny. One would assume that if there is a qualitative difference between the two methods, it would show in that number.