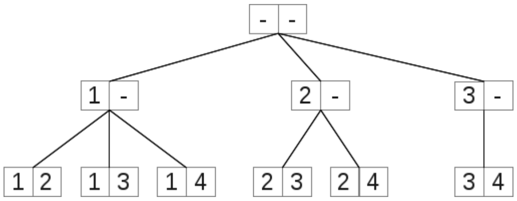
Introduction to Bioinformatics Algorithms
Homework 5
Solution

Saad Mneimneh, Computer Science, Hunter College of CUNY

Problem 1: Partial digest with "Branch and Bound"

(a) Given a set of size n, generate all subsets of size $k \leq n$. Think of the elements as being $\{1, 2, \ldots, n\}$. For example, the following tree illustrates all subsets when $k = 2$ and $n = 4$ (the leaves are at level k).



Observe that for each node in the tree, if the largest value is a and the level is l, then $n - a \geq k - l$ (otherwise the node cannot have any descendants). Using a function similar to the one covered in HW1, you can generate all subsets of a given size without having to explicitly generate the tree.

(b) Implement the brute force algorithm for the partial digest problem. You will need to generate all subsets of $L - \{\max L\}$ of size $n-2$ (check the pseudocode). You will also be able to prune certain subtrees if a partial subset cannot be extended, e.g. some distance is not in L. For this, adapt an implement of a *next* and a *skip* on the tree as explained in class for the branch-and-bound technique. Try the brute force algorithm with and without the pruning on a large example, e.g. generate a set of n random integers X and their pairwise distances L, then try to retrieve X from L.

Problem 2: Sorting by reversals
The algorithm on page 133 repeatedly chooses a reversal ρ that minimizes the number of breakpoints $b(\pi\rho)$ until the number of breakpoints is 0. The analysis following the presentation of this algorithm gives an approximation factor of 4.

(a) Show by example that when all strips are increasing, it is possible to choose a ρ as described above such that $\pi\rho$ has one decreasing strip, which when reversed, all strips are increasing again.

Solution: Here's a permutation with only increasing strips.

01.67.23.89.45

Still, there is a reversal that will minimize the number of breakpoints, in this case, reduces them by 1. For instance the reversal of 7.23. We get:

$$01.6.32.789.45$$

With only one decreasing strip that can be reversed, and when reversed, all strips will be increasing again:

$$0123.6789.45$$

(b) Suggest a modification to the algorithm so that if every strip is increasing, there will be a decreasing strip in each of the next two steps. How does that affect the approximation factor?

Solution: The algorithm that chooses the decreasing strip with the smallest number k and performs a reversal that puts $k-1$ and k together (thus reducing the number of breakpoints), or a increasing strip when no decreasing strips exist and reverses it (even though it many not reduce the number of breakpoints), works. In this case, one could show by simple analysis that once an increasing strip is reversed, the next two moves can work with decreasing strips. So we remove at least 2 breakpoints every 3 moves. Therefore, the approximation factor is 2/(2/3)=3, which is better than 4, since at most 2 breakpoints can be removed with 1 move.

Problem 3: Sequencing
Do problem 8.6 in the book.

Problem 4: Shortest superstring (optional)
The greedy algorithm for shortest covering string described in class essentially does this: repeatedly merge a pair of strings with maximum overlap until one string remains. Define the compression of the algorithm as the number of symbols saved compared to plainly concatenating all strings. Prove that the greedy algorithm achieves at least 1/2 the compression of the optimal superstring.