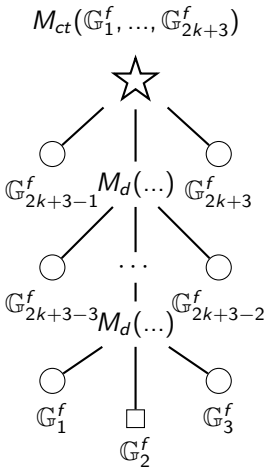
Algorithms in Comparative Genomics

Universität Bielefeld, WS 2021/2022
Dr. Marília D. V. Braga · Leonard Bohnenkämper

https://gi.cebitec.uni-bielefeld.de/teaching/2021winter/cg

Exercise sheet for the holidays, 23.12.2021

Exercise 1 (Christmas Tree Median) (10* pts)



Given $2k+3$ canonical genomes $\mathbb{G}^f_1, \mathbb{G}^f_2, \mathbb{G}^f_3, ...\mathbb{G}^f_{2k+3}$ and an algorithm to compute the median $M_d(\mathbb{A}^f, \mathbb{B}^f, \mathbb{C}^f)$ of three genomes $(\mathbb{A}^f, \mathbb{B}^f, \mathbb{C}^f)$ under a distance model d. The *Christmas Tree Median*[1] is defined as

$$M_{ct}(\mathbb{G}^f_1, \mathbb{G}^f_2, \mathbb{G}^f_3, ...\mathbb{G}^f_{2k+3-2}, \mathbb{G}^f_{2k+3-1}, \mathbb{G}^f_{2k+3}) = M_d(\mathbb{G}^f_{2k+3-1}, M_{ct}(\mathbb{G}^f_1, \mathbb{G}^f_2, \mathbb{G}^f_3, ...\mathbb{G}^f_{2k+3-2}), \mathbb{G}^f_{2k+3}) \quad (1)$$

with recursion base

$$M_{ct}(\mathbb{G}^f_1, \mathbb{G}^f_2, \mathbb{G}^f_3) = M_d(\mathbb{G}^f_1, \mathbb{G}^f_2, \mathbb{G}^f_3) \quad (2)$$

1. Compute the Christmas Tree Median of the following genomes under the breakpoint distance ($d = d_{BP}$): $\mathbb{G}^f_1 = (\ 1\ 2\ 3\ 4\ )$, $\mathbb{G}^f_2 = (\ \bar{2}\ \bar{1}\ \bar{4}\ \bar{3}\ )$, $\mathbb{G}^f_3 = (\ 2\ 3\ 4\ 1\ )$, $\mathbb{G}^f_4 = [\ 1\ \bar{3}\ \bar{2}\ 4\ ]$, $\mathbb{G}^f_5 = [\ \bar{2}\ \bar{1}\ \bar{4}\ \bar{3}\ ]$
2. Disprove (e.g. via counter example): The Christmas Tree Median under the breakpoint distance is always a breakpoint median.
3. Given the order of the genomes may not be permuted, is the Christmas Tree Median under the SCJ distance ($d = d_{SCJ}$) unique? Argue why/why not (Spoiler).
4. Prove or disprove: No metric d on a set with two or more distinct elements exists, under which the Christmas Tree Median is always a true Median[2] (Spoiler 1, 2, 3, 4, 5).

Exercise 2 (Double Distances) (5* pts)

Regard the genomes $\mathbb{S}^f = (\ 1\ 2\ 3\ 4\ 5\ )$ and $\mathbb{D}^f = (\ \bar{1}\ 5\ \bar{5}\ 1\ 2\ \bar{4}\ \bar{3}\ 2\ 3\ 4\ )$.

1. Calculate the breakpoint double distance $d^2_{BP}(\mathbb{S}^f, \mathbb{D}^f) = d_{BP}(\mathbb{S}^f \oplus \mathbb{S}^f, \mathbb{D}^f)$ and give an optimal matching $\mathbb{M}^f_{opt}$ on $\mathbb{S}^f \oplus \mathbb{S}^f$ and $\mathbb{D}^f$.
2. Calculate the SCJ double distance $d^2_{SCJ}(\mathbb{S}^f, \mathbb{D}^f)$ between $\mathbb{S}^f$ and $\mathbb{D}^f$.

[1] which I made up; don't look for this in the literature ;)
[2] The true median of a set $K \subseteq S$ under metric d on space S being the element $M_d \in S$ that minimizes $\sum_{k \in K} d(M_d, k)$.

3. Computing the DCJ double distance is NP-hard. Using the matching from subtask 1 what is the DCJ distance $d_{DCJ}(\mathbb{M}^f_{opt}, \mathbb{D}^f)$ between $\mathbb{M}^f_{opt}$ and $\mathbb{D}^f$?

4. Find another matching $\tilde{\mathbb{M}}^f_{opt}$ on $\mathbb{S}^f \oplus \mathbb{S}^f$, which minimizes $d_{SCJ}(\tilde{\mathbb{M}}^f_{opt}, \mathbb{G}^f)$, but produces a different DCJ distance from the one computed in subtask 3, i.e. $d_{DCJ}(\tilde{\mathbb{M}}^f_{opt}, \mathbb{G}^f) \neq d_{DCJ}(\mathbb{M}^f_{opt}, \mathbb{G}^f)$.

Exercise 3 (Santa's Unsigned Inversion Distance) (8* pts)

A snowstorm has caused chaos in Santa's workshop! The n presents which are usually nicely ordered from $1, ..., n$ are now in a random permutation $r_1, .., r_n$. Fortunately the elves can use magic to invert any segment $r_i, r_{i+1}, ...r_{i+k-1}, r_{i+k}$ to $r_{i+k}, r_{i+k-1}, ..., r_{i+1}, r_i$ in constant time.

1. Sort the following pile of presents twice using
   (a) signed inversions (as discussed in the lecture)
   (b) unsigned inversions (as the elves use)

   1 4 2 3 5.

2. Give a short description of how and why a signed inversion sorting scenario can be mapped to an unsigned one and why it is therefore possible to sort the presents in $\mathcal{O}(n)$ time.

3. After a while you notice that the elves seem to be doing a lot of unnecessary inversions. You start to suspect that sorting unsigned permutations that are not already sorted with unsigned inversions can always be done in fewer steps than with signed inversions. Why might that be?

4. The algorithm you described in subtask 2 sorts a list of length n in $\mathcal{O}(n)$ time. Why doesn't this conflict with the the theoretical bound of $\Omega(nlog(n))$ you are familiar with for sorting? (Spoiler)

Exercise 4 (Secret Santa is a Genome) (12* pts)

A common holiday tradition in the West is the game[3] of Secret Santa[4]. The activity works as follows: Everyone participating writes their name on a piece of paper. After the papers have been mixed, everyone draws a name in secret. One then acquires a gift for the person whose name one drew. The gifts are then handed out on another day, typically also in secret, such that no one knows who bought whose gift. A common annoyance is drawing ones own name from the pile. This is just a sub-category of a broader phenomenon. In a normal game of Secret Santa it is possible that there are subsets of the participants, between which no gift is given. For example, two people might draw each other's names, therefore being separated from the rest of the group. Let us refer to the maximal subsets between which no gift is given as *cohorts* and to cohorts of size 1 as *unicycles*.

1. One could try to remedy the problem of too small cohorts without redrawing the names by letting people swap the name they drew with someone else. This, of course, sacrifices some anonymity. Given the number of cohorts c, what is the minimum number of swaps needed in order to unify everyone into one cohort?

2. Can you find the connection to genomes, that is can you find an equivalent formulation of the problems described above in the language of comparative genomics? What is the equivalent of the cohort? What is the equivalent of the swapping operation?[5] Bonus: Can you meaningfully translate the deletion operation from comparative genomics to the Secret Santa situation?

3. In a game of Secret Santa you happen to know who gets whom a gift:

Person	Gives gift to
Alice	Eve
Bob	Gerald
Cathy	Alice
Dennis	Bob
Eve	Hugo
Fernando	Cathy
Gerald	Dennis
Hugo	Fernando

[3] It's not technically a game, but we will refer to it as such :)

[4] In Germany typically known as *Wichteln*.

[5] There are no exact equivalents as Secret Santa is not a signed permutation while genomes are. However, there are comparative genomic constructs you know that are able to mimic the two concepts.

Who needs to swap with whom, such that each person has a gift for the next person in alphabetical order, i.e. Alice → Bob, Bob → Cathy, ..., Hugo → Alice. Because the game should still stay as secret as possible, try to find a series of swaps of minimal length and show that there can be no shorter series of swaps[6].

4. Given n names $a_1, ..., a_n$ and an initial distribution of gift-giving (as in Sub-task 3). We call a series of swaps a *sorting* if it transforms the distribution to an alphabetical one, i.e. $a_1 \to a_2, ..., a_n \to a_1$. A sorting is *optimal* if it contains equally many or fewer swaps than any other sorting.

   (a) Prove or disprove: If $n = 3$ there is always an optimal sorting, such that no intermediate distribution has a unicycle, given the initial distribution was unicycle-free.

   (b) Prove or disprove: If $n > 3$ there is always an optimal sorting, such that no intermediate distribution has a unicycle, given the initial distribution was unicycle-free (Spoiler 1, 2, 3, 4,5).

5. A procedure to avoid unicycles in the distribution from the start was presented by Hannah Fry in a youtube-video by the channel `Numberphile`[7] in 2016. It works as follows: Cards labeled as $1|1, 2|2, ..., n|n$ are shuffled into a random permutation $k_1|k_1, ..., k_n|k_n$. Each card $k_j|k_j$ is split into its two components k_j, k_j and the upper half is shifted by 1 (modulo n), i.e. one obtains $k_1|k_n$, $k_2|k_1$, $k_3|k_2$, ..., $k_n|k_{n-1}$. Each participant draws such a recombined card. The upper half is public information and signifies which person is associated with which number. The lower half is private information and signifies to which number one has to give a gift.

   (a) What is the minimum size of a cohort in a game of Secret Santa generated by this procedure?

   (b) Can you design a procedure that is able to create cohorts of arbitrary, random sizes, but not unicycles? Your procedure should **not** rely on a game-master, a person or computer, that knows more than the average player.

☆☆☆ **Have fun, stay safe and enjoy your holidays!** ☆☆☆

[6] It is recommended that you do Sub-task 2 first

[7] `https://youtu.be/5kC5k5QBqcc?t=483`