Greedy approach for list alignment

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import pyranges
```

In [ ]:
```
path_maternal="data/shared_contigs_maternal_coordinates.tsv"
path_paternal="data/shared_contigs_paternal_coordinates.tsv"


path_grandmother_RM = "data/assembly.v1.0.PAN010.diploid.RM.bed"
path_grandfather_RM = "data/assembly.v1.0.PAN011.diploid.RM.bed"
path_mother_RM = "data/assembly.v1.0.PAN027.diploid.RM.bed"
path_proband_RM = "data/assembly.v1.0.PAN028.diploid.RM.bed"
```

In [ ]:
```
def split_info(info): # PAN010.chr1.haplotype1:142532907-142550224
    if pd.isna(info):
        return pd.Series([None, None, None])
    else:
        info = info.split(":")
        chr = info[0]
        positions = info[1].split("-")
        start = int(positions[0])
        end = int(positions[1])
        return pd.Series([chr, start, end])
def preprocess_rm(rm_df):
    """
    Get rid of simple repeats and satellites
    """
    rm_filt = rm_df[rm_df["ThickEnd"] != "unknown"]

    return rm_filt

def extract_repeat_annotations(info, rm_bed):
    if pd.isna(info):
        return None
    else:
        chr, start, end = split_info(info)
        rm_filtered = rm_bed[chr, start:end]
        rm_df = rm_filtered.df.sort_values(by="Start")
        rm_preprocessed = preprocess_rm(rm_df)
        rm_list = rm_preprocessed[["Chromosome", "Start", "End", "Name", "Score", "Strand", "ThickStart", "ThickEnd", "Ite

        return rm_list

def match_repeats(match1, match2):
    if match1[5] != match2[5]: # Strand
        return False
    if match1[7] != match2[7]: # Repeat Type (2nd level)
        return False
    return True

def backtrack_previous(repeat, previous_repeats, index):
    """
    Walk backward from index, collecting matching repeats into list
    """
    if index < 0:
        return []

    current = previous_repeats[index]
    if match_repeats(repeat, current):
        return backtrack_previous(repeat, previous_repeats, index - 1) + [current]
    else:
        return []

def match_previous(repeat, previous_repeats, neighbor_num=3):
    if not previous_repeats:
        return []

    start = max(0, len(previous_repeats) - neighbor_num)

    for i in reversed(range(start, len(previous_repeats))):
        if match_repeats(repeat, previous_repeats[i]):
            return backtrack_previous(repeat, previous_repeats, i)

    return []


def find_novel_insertions(grandparent, mother, proband, neighbor_num=3):
    candidates = []
    matched_repeats = []
```

```
    candidate_chain_counter = 0

    proband = proband or mother  # fallback if proband is None

    proband_idx = 0
    grandparent_idx = 0

    while proband_idx < len(proband):
        repeat = proband[proband_idx]
        matched = False
        # Check at most neighbor_num elements in grandparent from the current pointer
        for offset in range(neighbor_num):
            g_idx = grandparent_idx + offset
            if g_idx >= len(grandparent):
                break

            if match_repeats(repeat, grandparent[g_idx]):
                # Advance the grandparent pointer to skip matched elements
                grandparent_idx = g_idx + 1
                matched_repeats.append(repeat)
                matched = True
                candidate_chain_counter = 0
                break

        if not matched:
            candidates.append(repeat)
            matching_previous = match_previous(repeat, matched_repeats)
            candidates.extend(matching_previous)
            candidate_chain_counter += 1

        if candidate_chain_counter >= neighbor_num:
            print("This is a long chain of candidates indicating that the matching algorithm is not working properly.")
            # If we have a long candidate chain (likely due to annotation misclassification)
            # move the grandparent pointer forward to avoid breaking the matching system
            grandparent_idx += neighbor_num
            candidate_chain_counter = 0
        # Advance proband pointer
        proband_idx += 1

    return candidates
```

In [ ]:
```
df_maternal = pd.read_csv(path_maternal, sep="\t", header=None, names=["Contig", "Grandparent", "Mother", "Proband"])

# Load repeat annotations as PyRanges objects
rm_grandfather = pyranges.readers.read_bed(path_grandfather_RM)
rm_grandmother = pyranges.readers.read_bed(path_grandmother_RM)
rm_mother = pyranges.readers.read_bed(path_mother_RM)
rm_proband = pyranges.readers.read_bed(path_proband_RM)
```

In [5]:
```
for row in df_maternal.itertuples(index=False):
    contig, grandparent_info, mother_info, proband_info = row
    grandparent_chr, grandparent_start, grandparent_end = split_info(grandparent_info)
    mother_chr, mother_start, mother_end = split_info(mother_info)
    proband_chr, proband_start, proband_end = split_info(proband_info)

    grandparent_repeats = extract_repeat_annotations(grandparent_info, rm_grandmother)
    mother_repeats = extract_repeat_annotations(mother_info, rm_mother)
    proband_repeats = extract_repeat_annotations(proband_info, rm_proband) if not pd.isna(proband_info) else None

    insertion_candidates = find_novel_insertions(grandparent_repeats, mother_repeats, proband_repeats)
```

This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.
This is a long chain of candidates indicating that the matching algorithm is not working properly.