Thomas Holland

Problem

The problem my project aims to solve is regarding degenerate primer design for a subset of sequences. When creating degenerate primers there is the issue of making copies of DNA from organisms that the researcher does not want to copy due to primers that are too degenerate. By too degenerate I mean that they attach to DNA strands that we aren't interested in due to the primer attaching to DNA that we don't need nor want to be copied. This problem occurs often as its common to have mixed samples when taking them from an environment other than in a lab.

The first step of this problem is first actually marking DNA that you want to make copies of with degenerate primers. This problem has been thoroughly explored with solutions such as CODEHOP[1], ICODEHOP[2], Hyden[3], DePiCt[4], Primer BLAST[5], and Primer3[6] which are used to find a degenerate primer that matches sets of DNA. There is no research currently available that has approached the problem of stopping matches with confounding DNA sets.

The problem that I am aiming to solve is locating primers that amplify a given set of sequences. I will then compare these primers with the other DNA sets we don't want to attach to and if it attaches to any of them, I will log which places it attaches. After checking all of the DNA in the set that I don't want to make copies of my program will check to see which change in the primer would allow it to match with as many DNA strands as we want copies of while not making copies of the excluded set. It will then make another primer that will match with those not covered using the original. This will create two primers in place of one, so we will lose degeneracy but gain more primers to stop unintended attachments.

We will use a primer of minimum length 18 as the optimal range for primers is between 18 and 25. The length of the segment between the two primers will be equal to or larger than 140 nucleic acids long.

My approach

I am going to first attempt to use CODEHOP, ICODEHOP, Hyden, DePiCt, Primer BLAST, and Primer3 as they all make optimal primers when given a set of DNA. Assuming that one of them will provide an optimal enough solution for my needs I will be working with a set of degenerate primers and a set of DNA that I don't want the primer to attach to. I will have the Primers checked against each strand to see if they attach anywhere using the fixed pattern variable texts methodology by shifting it over by the amount that we can be sure it will not match by using the Knuth-Morris-Pratt algorithm. After finding the set of DNA that matches the primers from the set, we don't want the primers to match to I will take that subset of the points that match with the primers and mark which changes will have the least most in terms of removing degeneracy. This means whichever nucleic acid causes the most attachments with the set we don't want to attach to will have its degeneracy removed and we will move to having two primers to work with instead. It will then check the set left and continue this until we are left with a set of degenerate primers that attach to none of the DNA in the set that we don't want copied or at least the minimum number.

If CODEHOP, ICODEHOP, Hyden, DePiCt, Primer BLAST, and Primer3 don't provide a viable solution then this creating of the optimal primer with the set of DNA we don't want to attach to will be done while finding an optimum degenerate primer

Progress Report

When attempting to use the software's available from the list of CODEHOP, ICODEHOP, Hyden, DePiCt, Primer BLAST, and Primer3 some provided usable solutions. Unfortunately based on what I have read in the papers available for them I have determined that their solutions are not optimal for our purposes. The ones that I could use all provided different solutions to the problem because of the different processes they used. Some of the variance may also have been due to the variety of settings available. I attempted to make the settings as standardized as possible, but many provided some settings that the others did not, so they were unable to be edited. For my program it currently provides no settings. It reads in the fasta file and divides the different segments. Currently there aren't any markers to identify which sequences are in the set of sequences that we want copies of and which are not. The program sets the first 10 as the set that we want copies of from the 9745 sequences in the file. It currently checks for primers in the first 10 by brute force comparing them against all others in the set and marks where there are matches across the sequences. It does not shift the sequence currently, so it does not account for the likely instance that they are not perfectly lined up. It takes what is selected as the best degenerate primer and attempts to attach it to all of the sequences that we do not want it to make copies of. All of the sequences it attaches to are stored in an array for use later.

Future work

For the rest of the project the goal is to complete basic functionality of the program. This will be done by using the sequences that it attached to that it should not of and finding the areas that can have their degeneracy removed in order to not attach to as many of these unwanted sequences as possible. This will create more primers but will optimize for the problem of not wanting copies of these other sequences.

Possible improvements

This is a list of improvements that should be made but based on the timeline may not all be reached:

- Optimization of original primer creation: The original primer creation is currently brute forced which is not optimal as it leads to an extremely long-time complexity. It also does not select the best degenerate primer for the set as it currently takes the set and finds the areas in the arrays that are most similar. This does not consider the fact that they may be out of order sequences and need to be shifted over to the left or right.
- Check for whether or not a primer is possible: creating a primer that covers the subset but not the set may be impossible as may creating a set of primers that attach to only the subset. This requires further research but as it is based on an np-hard problem (finding a degenerate primer for a set of sequences) this is also likely an np-hard problem.
- Replacing the arrays with a better data structure. Will discuss this in a meeting with you after the break. I have been looking into whether there might be a more optimal data structure such as a modified tree to use for comparison between the sequences. One option may be a modified trie but based on my research I think that a modified set of automata's may provide the best results.

  If there is an automata model made for each sequence, then the program can hand off each comparison of two sequences to a thread that returns the area with the best match. This would be more optimal because the construction of the automata would be slow but once they are made, they would help to increase the speed of comparisons. These modified automata would have to add a degenerate link to the primer whenever there is a mismatch. I am not entirely sure if this is possible and so I hope to meet with you to discuss it after the thanksgiving break.

Bibliography

[1] T. M. Rose, J. G. Henikoff, and S. Henikoff, "CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design," *Nucleic Acids Res.*, 2003.

[2] R. Boyce, P. Chilana, and T. M. Rose, "iCODEHOP: A new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences," *Nucleic Acids Res.*, 2009.

[3] C. Linhart and R. Shamir, "The degenerate primer design problem," *Bioinformatics,* 2002.

[4] X. Wei, D. N. Kuhn, and G. Narasimhan, "Degenerate primer design via clustering," in *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003,* 2003.

[5] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden, "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.," *BMC Bioinformatics,* 2012.

[6] A. Untergasser *et al.*, "Primer3-new capabilities and interfaces," *Nucleic Acids Res.*, 2012.