Thomas Holland

Problem

The problem my project aims to solve is regarding degenerate primer design for large data sets. 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 aren't interested in. 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 primary problem that I am aiming to solve is taking a degenerate primer that matches a set that we want copies of and making it so that it wont match with a set that we don't want copies of. 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 attachs. 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 we want copies of while not making copies of the exclusion 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.

The secondary problem that I aim to solve is the problem that CODEHOP, ICODEHOP, Hyden, DePiCt, Primer BLAST, and Primer3 solve. I will need to resolve the problem of finding degenerate primers if their solutions don't provide an adequate solution that I can use. They may not provide an adequate solution because they attempt to optimize before considering the issue of other primers that we want to make sure we don't match. This optimization may remove options that are more optimal for our use case which may make it less efficient when compared with a solution we make. We will still test them to see if they may provide solutions that are optimal enough for use but if we find that they do not then we will switch to create our own.

In the case that we create our own program to solve the secondary problem we will use a primer of length 20 as the optimal range for primers is between 18 and 25. The length of the segment between the two primers will be between 140 and 220 nucleic acids long to determine which species the DNA is from.

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

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.