Primer Selection for Polymerase Chain Reactions

Department of
Computer Science
School of Engineering
University of Virginia
Charlottesville, Virginia 22904 (434) 982-2207

William Pearson
Gabriel Robins
Dallas Wrege
Tongtong Zhang

Appeared in the International Conference on Intelligent Systems for Molecular Biology, 1995, pp. 285-291
and Discrete Applied Mathematics, Vol. 71, 1996, pp. 231-246

www.cs.virginia.edu/robins

Contributions

Idea: use PCR to discover new gene
Problem formulation
Analysis of problem complexity
Exact algorithm & effective heuristics
Weighted variants for inexact primers

Overview

Goal:

Discover previously unknown genes

Strategy:

Design PCR primers for large set of known gene family members

➡ *Unknown genes will (hopefully) be amplified*

Problem Formulation

Optimal Primer Selection Problem:

Input: set of DNA sequences
Output: optimal set of primers

Theorem: NP-complete

Proof: reduction from set cover

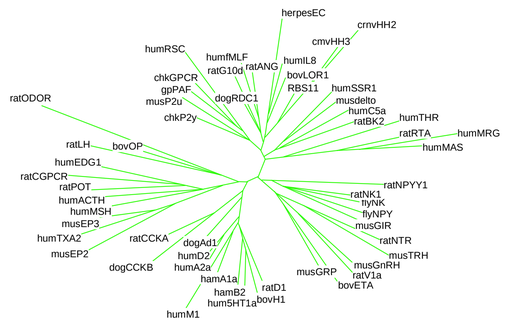
Non-approximability Result:

Can not do better than *(log # sequences)·OPT within polynomial time*

Basic Observations

Gene family:

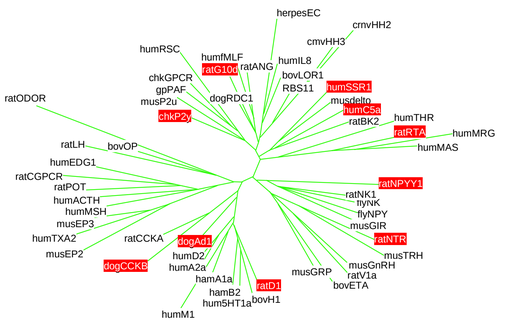
Genes derived from common ancestors



Common regions among family members

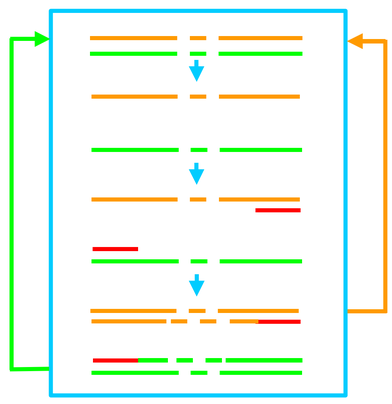
➡ *Common primers*

Primer group:



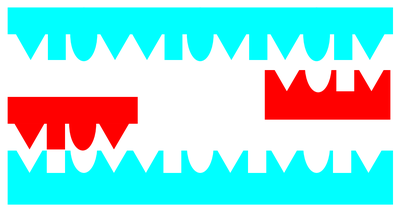
Polymerase Chain Reaction:

An effective method to amplify DNA sequence



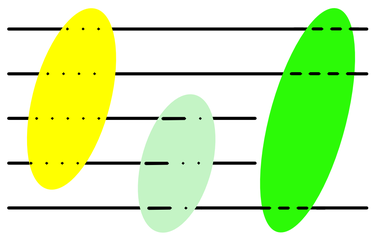
Primer:

DNA sequences of length 15-20
Reverse complementary to target regions



Approaches

Find minimal primer cover:



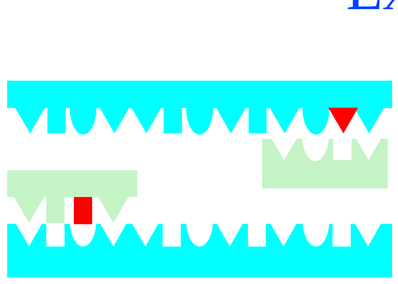
Exact algorithms:

Exhaustive brute-force
Branch-and-bound

Provably-good heuristics:

Solution quality: *log(# sequences) · OPT*
(Best possible within polynomial time)

Extension: Inexact Primers



Goal: Optimize mismatches & #primers

Observation:

Mutually competing objectives

(e.g. melting temperature, primer composition, duplicated subsequences)

Three-phrase approach:

Primer formation
Cover construction
Weight minimization

Sample Output

```
flyNPY     TTCACTGGCTGGCCATGTCGCACTGCTGCTACAATCCGATCATCTACT
dogRDC1    CGCAGTGCCTGTACAGTGTGCACTGCTGCGTCAACCCCGTGCTCTATA
ratNTR     CCAACGCTCTCTTCTACAGTAGCTCCGCCATCAATCCCATCCTCTACA
humfMLF    CAAGTGCCCTGGCCTTCTTCAACAGCTGCCTCAACCCCATGCTCTATG
bovH       CCATCTGGCTGGGCTACATCAACTCCACGCTGAACCCCCTCATCTACC
gpPAF      CCCTCTGCCTCCTTAGCACCAACTGTGTCTTAGACCCTGTCATCTACT
musP2u     CCCGGCCGCTGGCCAGCGCCAACAGTTGTCTTGACCCTGTCATCTACT
ratNPYY    GCCACCTCACGGCCATGATCTCCACCTGCGTCAACCCCATCTTTTATG
```

Solution: 2 primer groups cover 8 sequences

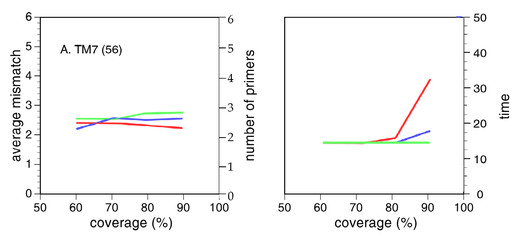
The first primer is: AACCCGATCATCTAC

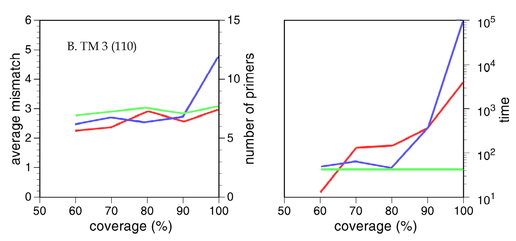
start position		#mismatches
flyNPY (48)	GCTAC..T.....	1
ratNTR (48)	CCATC..T..C...C.....	3
bovH (48)	CGCTG.....CC.....	2
gpPAF (48)	TCTTAG....TG.....	3
musP2u (48)	GTCTTG.....G.AC.....	4

The second primer is: CACAGCTGCGTCAAC

start position		#mismatches
dogRDC1 (24)	GTGGT...T.....	1
humfMLF (24)	TCTTCA.........C....	2
ratNPYY (24)	TGATCTC..C.....	3

Primer Selection: Experimental Results





Resulting Primer Groups