Report Number: WUCSE-2003-76

2003-12-04

An Iterative Beam Search Algorithm for Degenerate Primer Selection

Richard Souvenir

Single Nucleotide Polymorphism (SNP) Genotyping is an important molecular genetics process in the early stages of producing results that will be useful in the medical field. Due to inherent complexities in DNA manipulation and analysis, many different methods have been proposed for a standard assay. One of the proposed techniques for performing SNP Genotyping requires amplifying regions of DNA surrounding a large number of SNP loci. In order to automate a portion of this particular method, it is necessary to select a set of primers for the experiment. Selecting these primers can be formulated as the Multiple Degenerate Primer... **Read complete abstract on page 2.**

An Iterative Beam Search Algorithm for Degenerate Primer Selection

Richard Souvenir

Complete Abstract:

Single Nucleotide Polymorphism (SNP) Genotyping is an important molecular genetics process in the early stages of producing results that will be useful in the medical field. Due to inherent complexities in DNA manipulation and analysis, many different methods have been proposed for a standard assay. One of the proposed techniques for performing SNP Genotyping requires amplifying regions of DNA surrounding a large number of SNP loci. In order to automate a portion of this particular method, it is necessary to select a set of primers for the experiment. Selecting these primers can be formulated as the Multiple Degenerate Primer Design (MDPD) problem. In this thesis, we describe an iterative beam-search algorithm, Multiple, It-erative Primer Selector (MIPS), for MDPD. Theoretical and experimental analyses show that this algorithm performs well compared to the limits of degenerate primer design. Furthermore, MIPS outperforms an existing algorithm which was designed for a related degenerate primer selection problem. Further analysis shows that, due to the composition of the human genome, the results from MIPS may not be realized in practice. Consequently, we address the challenges involved in selecting a suitable set of degenerate primers and possible future improvements to the algorithm.

Department of Computer Science & Engineering

Washington University in St.Louis
School of Engineering & Applied Science

2003-76

An Iterative Beam Search Algorithm for Degenerate Primer Selection - Master's Thesis, December 2003

Authors: Souvenir, Richard

December 4, 2003

Abstract: Single Nucleotide Polymorphism (SNP) Genotyping is an important molecular genetics process in the early stages of producing results that will be useful in the medical field. Due to inherent complexities in DNA manipulation and analysis, many different methods have been proposed for a standard assay. One of the proposed techniques for performing SNP Genotyping requires amplifying regions of DNA surrounding a large number of SNP loci. In order to automate a portion of this particular method, it is necessary to select a set of primers for the experiment. Selecting these primers can be formulated as the Multiple Degenerate Primer Design
(MDPD) problem.

In this thesis, we describe an iterative beam-search algorithm, Multiple, Iterative Primer Selector (MIPS), for MDPD. Theoretical and experimental analyses show that this algorithm performs well compared to the limits of degenerate primer design. Furthermore, MIPS outperforms an existing algorithm which was designed for a related degenerate primer selection problem. Further analysis shows that, due to the composition of the human genome, the results from MIPS may not be realized in practice. Consequently, we address the challenges involved in selecting a suitable
set of degenerate primers and possible future improvements to the algorithm

Short Title: Multiple Degenerate Primer Design Souvenir, M.Sc. 2003

WASHINGTON UNIVERSITY

SEVER INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

AN ITERATIVE BEAM SEARCH ALGORITHM FOR DEGENERATE PRIMER SELECTION

by

Richard M. Souvenir, B.S.

Prepared under the direction of Professor W. Zhang

A thesis presented to the Sever Institute of
Washington University in partial fulfillment
of the requirements for the degree of
Master of Science

December, 2003

Saint Louis, Missouri

WASHINGTON UNIVERSITY
SEVER INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ABSTRACT

AN ITERATIVE BEAM SEARCH ALGORITHM FOR DEGENERATE PRIMER SELECTION

by Richard M. Souvenir

ADVISOR: Professor W. Zhang

December, 2003
Saint Louis, Missouri

Single Nucleotide Polymorphism (SNP) Genotyping is an important molecular genetics process in the early stages of producing results that will be useful in the medical field. Due to inherent complexities in DNA manipulation and analysis, many different methods have been proposed for a standard assay. One of the proposed techniques for performing SNP Genotyping requires amplifying regions of DNA surrounding a large number of SNP loci. In order to automate a portion of this particular method, it is necessary to select a set of primers for the experiment. Selecting these primers can be formulated as the *Multiple Degenerate Primer Design (MDPD)* problem.

In this thesis, we describe an iterative beam-search algorithm, *Multiple, Iterative Primer Selector (MIPS)*, for MDPD. Theoretical and experimental analyses show that this algorithm performs well compared to the limits of degenerate primer design. Furthermore, MIPS outperforms an existing algorithm which was designed for a related degenerate primer selection problem. Further analysis shows that, due to the composition of the human genome, the results from MIPS may not be realized in practice. Consequently, we address the challenges involved in selecting a suitable set of degenerate primers and possible future improvements to the algorithm.

Contents

List of Tables . . . . **v**

List of Figures . . . . **vi**

List of Abbreviations . . . . **viii**

Acknowledgments . . . . **ix**

1 Introduction **1**

2 Related Work **4**

2.1 Primer Selection Problem 4

2.2 Degenerate Primer Design Problem 4

3 Problem Description and Complexity **7**

4 Multiple, Iterative Primer Selector **11**

5 Algorithm Complexity **15**

5.1 Space 15

5.2 Time 16

6 MIPS Experiments **18**

6.1 Beam Size Parameter 18

6.2 Human Dataset 19

6.3 Comparison to HYDEN 20

7 Degenerate Primer Design 23

7.1 Degenerate Primer Efficacy 24

7.1.1 Occurrence probability for one fixed primer 25

7.1.2 Computing the expectation 26

7.1.3 Results 27

7.2 Mispriming 28

7.2.1 Mispriming in i.i.d. Random and Human Genome 29

7.2.2 Reducing Mispriming Events 31

7.2.3 Alternate Strategies to Mispriming 34

8 Conclusions 35

Appendix A MIPS Pseudocode 36

Appendix B Supplemental Data 38

References 47

Vita 51

List of Tables

2.1 IUPAC-IUB symbols for nucleotide nomenclature 6

5.1 Properties of an execution of the MIPS algorithm 15

6.1 Results on a dataset of 95 human SNP regions. 20

7.1 Actual and predicted coverage of 20-mer primers found on sets of 190 random sequences of length 211. Avg Coverage: average coverage of primer found over 20 random trials. Max Predicted: largest coverage m such that $E_{20,\alpha,m} > 1$. 28

7.2 Predicted and actual mispriming rates in simulated PCR experiments with i.i.d. random and human genome. 30

7.3 Mispriming rates in simulated PCR experiments with original and masked input sets 33

7.4 Comparison of MIPS-PT results on original and masked input sets. 33

List of Figures

1.1 Single Nucleotide Polymorphism (SNP) diagram 1

1.2 Example of a typical PCR experiment. The area in bold on the original double-stranded sequence represents the DNA fragment which is the region of interest. Normally this continues for 30 cycles, which would result in a theoretical maximum of 2^{30} fragments for each original molecule. (Diagram from [27]) 3

4.1 Pruning of the search space by MIPS-TT 13

5.1 Timing graphs for various input sizes. 17

6.1 The effect of beam size on the (a) running time of the algorithm and (b) number of primers discovered. 19

6.2 The number of degenerate primers selected by HYDEN and MIPS for 20 randomly-generated datasets in solving PT-MDPD for degeneracy thresholds of 10,000 (a) and 100,000 (b). 21

7.1 Results of RepeatMasker on human SNP input dataset. 32

B.1 MIPS-PT output with $\alpha = 4^6$. 39

B.2 MIPS-PT output with $\alpha = 4^7$. 40

B.3 MIPS-PT output with $\alpha = 4^8$. 41

B.4 MIPS-PT output with $\alpha = 4^9$. 42

B.5 MIPS-PT output on masked input set with $\alpha = 4^6$. 43

B.6 MIPS-PT output on masked input set with $\alpha = 4^7$. 44

B.7 MIPS-PT output on masked input set with $\alpha = 4^8$. 45

B.8 MIPS-PT output on masked input set with $\alpha = 4^9$. 46

List of Abbreviations

- **i.i.d.** independent and identically distributed
- **IUB** International Union of Biochemistry
- **IUPAC** International Union of Pure and Applied Chemistry
- **MDPD** Multiple Degenerate Primer Design
- **MIPS** Multiple, Iterative Primer Selector
- **MP-PCR** Multiplex Polymerase Chain Reaction
- **PCR** Polymerase Chain Reaction
- **PT-MDPD** Primer-Threshold MDPD
- **SNP** Single Nucleotide Polymorphism
- **TT-MDPD** Total-Threshold MDPD

Acknowledgments

I would very much like to thank my adviser, Dr. Weixiong Zhang, for his support of my graduate study financially and more importantly intellectually through our discussions and collaborative efforts. I would also like to acknowledge the other members of my committee, Dr. Jeremy Buhler and Dr. Gary Stormo, who have acted as co-advisors for most of my studies and have offered their time and expertise to assist in the evaluation of my research.

I would like to thank Dante Cannarozzi, Steve Donahue, Kevin Goodier, Christine Julien, Matthew Hampton, Jamie Payton, and Aaron Tenney who have always been willing to lend assistance throughout my progression towards a graduate degree. As friends and colleagues, they have always been willing to share their time by answering questions or providing critiques on practice talks and paper drafts.

I would also like to extend my gratitude to Sharlee Climer, Jianhua Ruan, and Xiaotao Zhang who, as fellow members of the Computational Intelligence Center, have taken the time away from their own research endeavors to share insights which have bolstered the research efforts of us all.

I appreciate very much the efforts of the department faculty who provide an intellectually stimulating environment in which to pursue a graduate education.

I would also like to acknowledge Peggy Fuller, Jean Grothe, Myrna Harbison, and Sharon Matlock, whose efforts ensure that the department functions effectively.

I would especially like to thank my parents, Yves and Elvire L. Souvenir, for all of their love and support.

The research represented in this thesis was supported in part by the NIH under training grant GM08802 and the NSF under ITR/EIA-0113618.

Richard M. Souvenir

Washington University in Saint Louis
December, 2003

Chapter 1

Introduction

Single Nucleotide Polymorphisms (SNPs) are singular base differences among DNA sequences from the same species that are partially responsible for individualization. Figure 1 shows an example of a SNP between two sequences. It is estimated that there are roughly three million SNPs in the human genome [15]. Research investigating associations between SNPs and various diseases, along with studies of differences in how individuals respond to common therapies, promise to revolutionize medical science in the coming years [2]. Another interesting biological facet of SNPs is that recent work suggests there may be only a few hundred thousand ”blocks” of SNPs in the human genome rather than a random dispersion. These “blocks” provide most of the variability seen in human populations [6]. In spite of all this effort, it is still a daunting task to identify the specific genetic variations occurring in specific individuals in order to determine their associations with important phenotypes. Currently, there

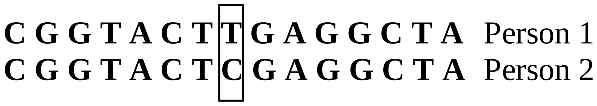
C G G T A C T T G A G G C T A Person 1
C G G T A C T C G A G G C T A Person 2

Figure 1.1: Single Nucleotide Polymorphism (SNP) diagram

are many proposed techniques for the process of determining the SNP composition of a given genome, which is known as SNP Genotyping. In order for these assaying techniques to be effective in large-scale genetic studies of hundreds or thousands of SNPs, they must be scalable, automated, robust, and inexpensive [12].

One technique involves the use of Multiplex PCR (MP-PCR) to amplify the regions around the SNP [12]. Polymerase Chain Reaction (PCR) [16] is a powerful molecular genetics technique which rapidly amplifies a small segment of DNA using two additional DNA fragments known as primers. Figure 1.2 shows how PCR cyclically creates a large number of DNA fragments. MP-PCR is a variation of PCR where multiple DNA fragments are replicated simultaneously. MP-PCR, like all PCR variations, makes use of oligonucleotide primers to define the boundaries of amplification. For each region of DNA that is to be amplified, two primers, generally referred to as the forward and reverse primers, are needed. In MP-PCR, it is necessary to select a pair of forward and reverse primers for each of the regions to be replicated, and for the large-scale amplification required in SNP Genotyping, there can be hundreds, or perhaps thousands, of those regions. The process of selecting such a large set of primers by current methods, including trial-and-error [12], can be time-consuming and difficult.

In this thesis, we begin with a description of the related work in the area. Next, we describe the Multiple Degenerate Primer Design (MDPD) problem and present an algorithm, the *Multiple, Iterative Primer Selector (MIPS)* [24], to solve this problem. We continue by showing how MIPS performs relative to another solution in the domain on real and simulated data. Next, we discuss the difficulty of solving this problem in general by calculating the theoretical limits of any solution and dealing with the issue of erroneous amplification. Finally, we conclude with comments about avenues for possible improvement.

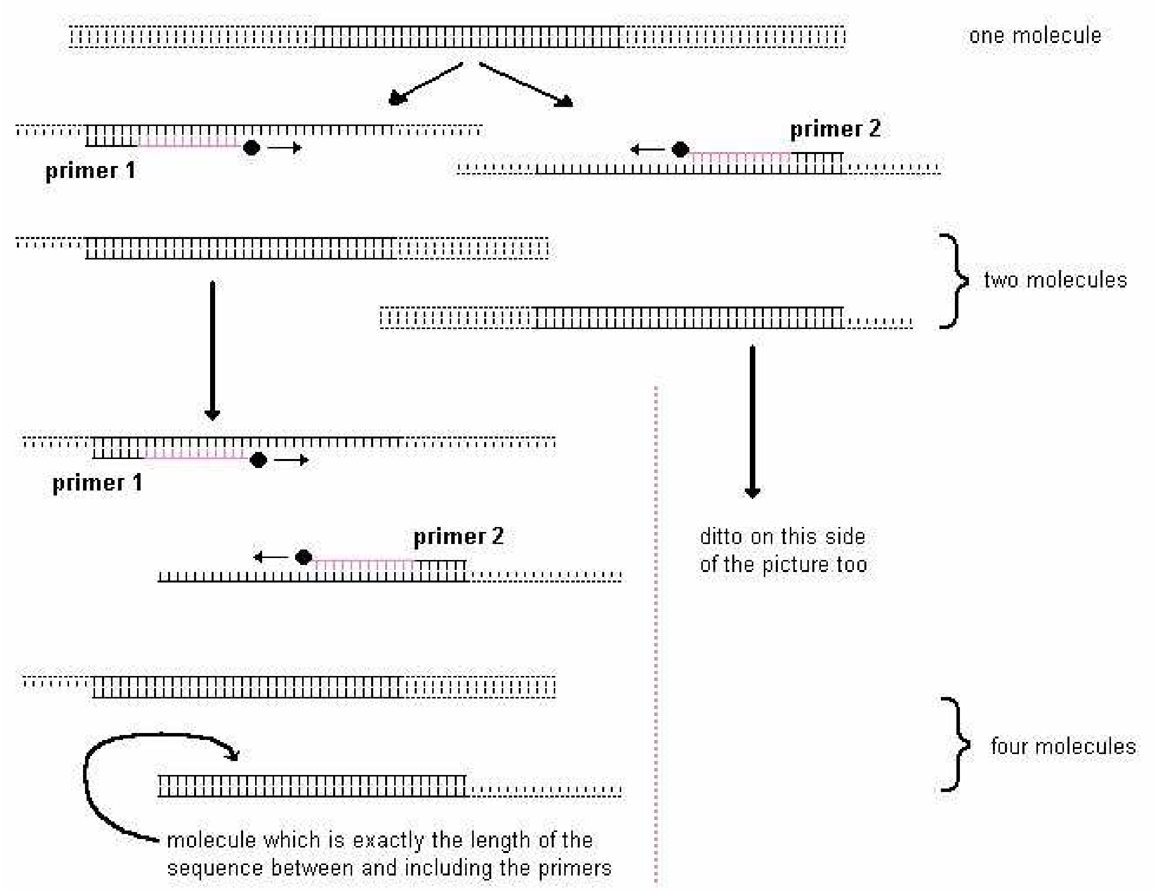

Figure 1.2: Example of a typical PCR experiment. The area in bold on the original double-stranded sequence represents the DNA fragment which is the region of interest. Normally this continues for 30 cycles, which would result in a theoretical maximum of 2^{30} fragments for each original molecule. (Diagram from [27])

Chapter 2

Related Work

There are two problems in primer selection similar to the main problem of this thesis, the Primer Selection Problem and the Degenerate Primer Design Problem.

2.1 Primer Selection Problem

The Primer Selection Problem [20] involves minimizing the number of primers needed to amplify regions of DNA in a set of sequences. It has been shown that this is an NP-hard problem [7] in reductions from other hard problems, including Set Cover and Graph Coloring [4]. There have been a number of proposed heuristics to solve this problem, including a branch-and-bound search algorithm [19]. Also, algorithms have been proposed which incorporate biological data about the primers into the search [17, 5].

2.2 Degenerate Primer Design Problem

Figure 1.2 from the previous chapter shows that in order to perform PCR both forward and reverse primers are needed for the fragment being amplified. Therefore, in a typical MP-PCR experiment where, the number of primers needed is equal to twice the number of sequences in the input set. In general, the algorithms mentioned above reduce the number of primers needed to 25-50% of this value, which can still be rather high for the large-scale amplification needed for SNP Genotyping.

The desire to design fewer primers leads to the use of degenerate primers. Degenerate primers [13] are primers that make use of the degenerate nucleotides [3], which can be found in Table 2.1. The number of primers that a degenerate primer represents is referred to as its *degeneracy*. For example, consider this degenerate primer, *ACMCM*, where *M* is a degenerate nucleotide which represents either of the bases, *A* or *C*. This degenerate primer is actually representative of the set of 4 primers {*ACACA, ACACC, ACCCA, ACCCC*}, and so its degeneracy is 4.

Degenerate primers are as easy to produce as regular primers, and therefore save the molecular biologist time during the primer design phase of the experiment. The use of degenerate primers, however, introduces two new problems. First, the effective concentration of the desired primers is decreased by the presence of undesired primers. Second, the presence of undesired primers can lead to erroneous amplification. Therefore, it is important to use primers of relatively low degeneracy to realize the inherent benefits of degenerate primer design while minimizing the effects of these two problems.

The Degenerate Primer Design (DPD) Problem, is the decision problem of determining whether or not there exists a single degenerate primer below some given degeneracy threshold which can amplify regions of DNA for some number of a set of input sequences. There are two variations of DPD. Maximum Coverage DPD (MC-DPD) is the related maximization problem where the goal is to find the maximum number of sequences that can be amplified by a degenerate primer whose degeneracy falls below some threshold. Minimum Degeneracy DPD (MD-DPD) is the second

Table 2.1: IUPAC-IUB symbols for nucleotide nomenclature

Symbol	Meaning
A	A
C	C
G	G
T	T
U	U
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
X	G or A or T or C
N	G or A or T or C

variation of DPD whose goal is to find the degenerate primer of minimum degeneracy that amplifies all of the input sequences. Both MC-DPD and MD-DPD have been shown to be NP-Hard problems [14].

Chapter 3

Problem Description and Complexity

Some of the notation from [14] is used to describe this problem. To maintain consistency, lower-case symbols (e.g. l, b, i) represent numerical values, counting variables, or individual characters (possibly degenerate) in a sequence. Upper-case symbols (e.g. P, S) denote primers, sequences, or subsequences. Finally, calligraphic symbols (e.g. $\mathcal{S}, \mathcal{C}$) represent sets of sequences or primers.

Let $\Sigma = \{A, C, G, T\}$ be the finite fixed alphabet of DNA. A *degenerate primer* is a string P with several possible characters at each position, i.e., $P = p_1 p_2 \cdots p_l$, where $p_i \subseteq \Sigma, p_i \neq \emptyset$ and l is the length of primer P. The *degeneracy* of P is $d(P) = \prod_{i=1}^{l} |p_i|$. Consider the degenerate primer $P' = \{A\}\{A,C\}\{A,C\}\{A,C\}$. The length of P' is 4 and $d(P') = 8$. For the sake of clarity, we use the IUPAC symbols from Table 2.1 for degenerate nucleotides to represent degenerate primers. Therefore, P' can be represented as $AMMM$ where M is the degenerate nucleotide which represents $\{A,C\}$. Degenerate primers can be constructed by *primer addition*. For any two primers, P^1 and P^2, their sum P^3 equals $(p_1^1 \cup p_1^2)(p_2^1 \cup p_2^2) \cdots (p_l^1 \cup p_l^2)$.

For any sequence S_i in an input set $\mathcal{S}$, we say that a degenerate primer P *covers* S_i if there is a substring F of length l in S_i where for each character f_i in F, $f_i \in p_i$. We now describe the three problems at the heart of this thesis.

Problem 1 (Multiple Degenerate Primer Design(MDPD)). *Given a set of n sequences over an alphabet Σ and integers l and k, is there a set of primers, $\mathcal{P}$, for which each element is of length l that covers all of the input sequences, where $|\mathcal{P}| \leq k$?*

We now describe two optimization problems that are variants of the MDPD problem which add additional constraints to the final solution $\mathcal{P}$.

Problem 2 (Primer-Threshold MDPD (PT-MDPD)). *Given a set of n sequences over an alphabet Σ and integers l and α, find a set of primers, $\mathcal{P}$, for which each element is of length l that covers all of the input sequences, where $\forall P_i \in \mathcal{P}, d(P_i) \leq \alpha$.*

In PT-MDPD, we want a small set of degenerate primers where the degeneracy of each primer in that set is less than some threshold. In the next variation, TT-MDPD, we want a small set of degenerate primers where the sum of the degeneracies of the set is below some threshold.

Problem 3 (Total-Threshold MDPD (TT-MDPD)). *Given a set of n sequences over an alphabet Σ and integers l and α, find a set of primers, $\mathcal{P}$, for which each element is of length l that covers all of the input sequences, where $\sum_{P_i \in \mathcal{P}} d(P_i) \leq \alpha$.*

We now show that optimal, efficient algorithms for these problems do not likely exist since both are NP-complete [7]. In order to show the necessary proofs, we will restate each problem as a decision problem where we wish to determine whether the solution set, $\mathcal{P}$, has size less than a given value, k.

For PT-MDPD, we will use a reduction from the Primer Selection Problem (PSP) [20]. The input to PSP is a set of input sequences $\mathcal{S}'$ and a threshold k', and the goal is to find a set of (non-degenerate) primers $\mathcal{P}'$ which cover all the sequences in $\mathcal{S}'$ and where $|\mathcal{P}'| \leq k'$.

Theorem 1. *PT-MDPD is NP-complete.*

Proof. To show that PT-MDPD $\in$ NP, given an input set $\mathcal{S}$, we use the solution $\mathcal{P}$ as a certificate for $\mathcal{S}$. Checking whether the primers in $\mathcal{P}$ cover all of the sequences in $\mathcal{S}$ can be accomplished in polynomial time in the number of sequences, given the observation that $|\mathcal{P}| \leq |\mathcal{S}|$.

We next prove that PSP $\leq_P$ PT-MDPD, which shows that our problem is NP-Complete. The reduction begins with an instance of PSP $= < \mathcal{S}', k' >$. To construct an instance of PT-MDPD $= < \mathcal{S}, \alpha, k >$, we simply let $\mathcal{S} = \mathcal{S}'$, $\alpha = 1$, and $k = k'$.

At this point, it should be obvious that a valid solution for PSP yields a valid solution for this construction of PT-MDPD, and vice-versa. Therefore, the remainder of the proof is trivial and therefore omitted. □

For TT-MDPD, we will use a reduction from the related primer design problem, Degenerate Primer Design (DPD), which has been shown to be NP-Complete [14]. An instance of DPD is a set $\mathcal{S}'$ of n' strings, and integers l', α', and m'. A solution is a degenerate primer of length l' and degeneracy at most α' that matches at least m input strings. For this reduction, we consider a special case of DPD where $m' = n'$.

Theorem 2. *TT-MDPD is NP-complete.*

Proof. To show that TT-MDPD $\in$ NP, given an input set $\mathcal{S}$, we use the solution $\mathcal{P}$ as a certificate for $\mathcal{S}$. Checking whether the primers in $\mathcal{P}$ cover all of the sequences in $\mathcal{S}$ and the total weight is less than α can be accomplished in polynomial time.

We next prove that DPD $\leq_P$ TT-MDPD, which shows that our problem is NP-Complete. The reduction begins with an instance of DPD $= < \mathcal{S}', n', l', \alpha' >$.

The instance of TT-MDPD $= < \mathcal{S}, \alpha, l, k >$ is constructed as follows. We simply let $\mathcal{S} = \mathcal{S}'$, $\alpha = \alpha'$, $l = l'$ and $k = 1$.

In this special case of DPD where $m' = n'$ and $k = 1$ for TT-MDPD, the goal of each problem is identical: to find a single degenerate primer of length $l = l'$ with degeneracy $\alpha = \alpha'$ which covers all of the sequences. Therefore a valid solution for one problem yields a valid solution for the other. □

Chapter 4

Multiple, Iterative Primer Selector

To overcome the difficulty caused by the NP-hardness of MDPD problems, we propose an iterative beam search heuristic, the Multiple, Iterative Primer Selector (MIPS) to make a tradeoff between optimality and tractability. In order to solve PT-MDPD and TT-MDPD, MIPS can run in either of two modes, MIPS-PT and MIPS-TT, respectively. This chapter focuses on MIPS-TT. However, we will highlight how MIPS-PT operates differently.

MIPS progressively constructs a set of primers that covers all the input sequences. Define a *k-primer* to be a degenerate primer that covers k input sequences. The basic algorithm first generates a set of candidate 2-primers, each having some degeneracy value, then iteratively extends all candidate k-primers into $(k+1)$-primers by generalizing them to cover an additional sequence. Generalization stops when no primer can be extended without exceeding the degeneracy threshold α. At this point, the set of remaining primers cover k_{last} sequences, so we retain the primer of minimum degeneracy, remove the input sequences it covers from consideration, and repeat the algorithm until all sequences are covered.

To guide the search, MIPS uses the degeneracy of a primer as a scoring function. The set of primers that are stored for further extension are known as a *beam.*

Beam search [1] differs from greedy or best-first search in that multiple nodes, degenerate primers in this case, are saved for extension instead of just one. This model of progressively adding to a beam of degenerate primers and updating the scoring function is similar to the Consensus motif-finding model [10].

It is important to note that the degeneracy of a given k-primer increases or remains the same with the addition of new sequence fragments. This observation encourages us to employ a strategy which ignores degenerate primers with high degeneracy, in order to speed up the algorithm. Therefore, the search is restricted only to the primers with the lowest degeneracy. In this algorithm, the number of the candidate primers to restrict the search to at each level can be specified. This constant, b, describes the number of k-primers to save for each level. Increasing b can possibly improve the quality of the solution, but lengthens the running time of the algorithm. In section 6.1, we examine the effect of this parameter, b, on the speed and quality of the solution produced by MIPS.

The constructive search continues until one of two cases occurs. In the first case, all sequences are covered by a single n-primer, where n is the number of sequences in the input set. The algorithm then terminates with that primer as the result. In the second case, no k-primer can be extended to a $(k+1)$-primer without exceeding the degeneracy threshold and there exists at least one sequence uncovered. At this point, k_{last} sequences have been covered. The algorithm chooses the best degenerate (k_{last})-primer, P_0, from the set $\mathcal{P}$ of primers sorted by degeneracy value. The problem then reduces to a smaller instance where the input set is the original set of sequences minus those covered by P_0. In MIPS-PT, the degeneracy threshold for this subproblem is equivalent to the original threshold, α. In MIPS-TT, the degeneracy threshold is reduced by the degeneracy of P_0. *The algorithm then restarts on the reduced problem.*

For MIPS-PT, iteratively applying this procedure will eventually return a set of primers to cover the set of input sequences. However, this is not necessarily the case for MIPS-TT. After P_0 is discovered and its sequences are removed from consideration, the new threshold may be too low to cover the rest of the sequences. In this case, MIPS-TT *backtracks* to the previous level, $k_{last} - 1$, and selects the next best primer P_0' as part of the final solution. Again, MIPS restarts on the sequences that P_0' has not covered and with a degeneracy limit that is the original α minus the degeneracy of P_0'.

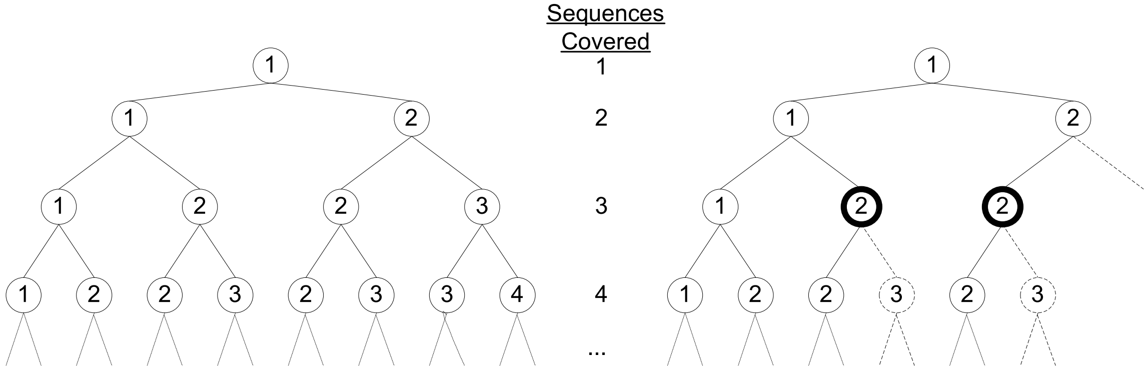

Figure 4.1: Pruning of the search space by MIPS-TT

Figure 4.1 shows, schematically, an execution of MIPS-TT. For these graphs, the depth of a node represents the number of sequences from the input set covered and the number in a node represents the number of degenerate primers that will be used to cover those sequences. Each node can be expanded into two child nodes. The left child represents covering an additional sequence using an existing degenerate primer and the right child represents covering an additional sequence using a new degenerate primer. The left tree in Figure 4.1 shows a full search. The right tree shows the pruning that takes place in MIPS-TT during the backtracking phrase. Consider the two bold nodes. Both of these cover the same number of sequences with the same number of primers. MIPS-TT will therefore only expand the node whose total score is better. While this greedy choice may not be optimal, it avoids the exponential expansion seen on the full tree by not exploring the nodes represented by dotted circles.

The pairwise comparison of two sequence fragments is the dominating operation and a rate-limiting step of the algorithm. A majority of these comparisons are between two fragments that share few, if any, nucleotides. To avoid comparisons between dissimilar fragments, the exhaustive pairwise comparison is replaced with a similarity lookup. All of the primer candidates are added to a FASTA-style [18] lookup table. In general, for DNA, a FASTA table fragment length of 6 is recommended [9]. Using the table, each fragment is compared only to the other fragments that are returned.

A pseudo-code description of the MIPS algorithm is given in Appendix A.

Chapter 5

Algorithm Complexity

We now examine the theoretical bounds of MIPS and compare these values to the computing resources consumed in practice. Consult Table 5.1 for a list of variables used in the chapter and what they represent.

Table 5.1: Properties of an execution of the MIPS algorithm

Variable	Represents
n	number of input sequences
m	average sequence length
b	beam size
l	primer length

5.1 Space

From the input set, each primer is stored individually which requires space $O(nml)$. In the implementation, there are four $n \times n$ matrices that are needed for back-tracking and storing degenerate primers that could eventually become part of the final solution. This adds an additional $O(n^2)$ of storage. Therefore, the total amount of space is $O(n^2 + nml)$.

5.2 Time

The time complexity is analyzed in a bottom-up fashion. The procedure of comparing the fragments in the beam to the remaining sequences is called ONE_PASS which is described in Algorithm 3 of Appendix A. ONE_PASS makes $O(bnm)$ primer additions, since there are $O(nm)$ total fragments and b fragments in the beam. Each primer addition requires comparing every character in each of the two primers. Therefore, this portion requires $O(bnml)$ time.

The process of generating new beams of k-primers, for increasing k, is called MIPS_SEARCH, which is described in Algorithm 2 of Appendix A. MIPS_SEARCH uses ONE_PASS to build new beams, and could, in the worst case, build n beams. Therefore, the overall time complexity is $O(bn^2ml)$. The number of times MIPS_SEARCH is executed depends on the amount of back-tracking. This is directly related to the number of primers in the final solution. In the best case, if the solution only requires one primer, there will be only one call to MIPS_SEARCH. In the worst case, if the solution requires n primers (one primer for each input sequence) there will be $n^2/2$ calls to MIPS_SEARCH. Let p be the number of primers in the final solution. The best approximation to the number of MIPS_SEARCH calls is $O(pn)$. This brings the overall time complexity to $O(bn^3mlp)$.

The graphs in Figure 5.1 show how the running time of MIPS changes when various parameters of the input set are manipulated. These graphs correlate with the theoretical predictions of time dependencies. All of these experiments were run on a computer running Red Hat Linux 7.3 with an AMD 1.6GHz CPU and 2GB RAM.

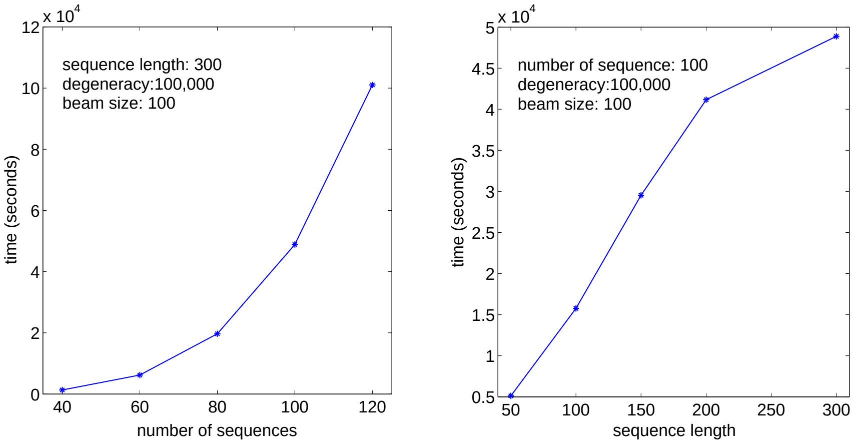

Figure 5.1: Timing graphs for various input sizes.

Chapter 6

MIPS Experiments

MIPS has been applied to both human DNA sequences and randomly generated datasets. The primary dataset is a database of sequences containing regions of human DNA surrounding 95 known SNPs. The sequences varied in length from a few hundred nucleotides to well over one thousand. The location of a SNP on a sequence was marked in order to provide a reference for the forward and reverse primers. To ensure effective PCR product analysis, each primer could not be located within 10 bases of the SNP and the entire PCR product length could not exceed 400 bases.

In this chapter, we perform three experiments. First, we show how the beam size affects the speed and quality of the solution produced by MIPS. Second, we show some results of MIPS on the human dataset of 95 sequences. Finally, we compare MIPS to an algorithm designed to solve the DPD problem considered in [14].

6.1 Beam Size Parameter

Figure 6.1a shows that increasing the beam size linearly increases the running time of the algorithm. Figure 6.1b shows the effect of beam size on the solution quality, or number of primers. These figures show the trade-off between the quality of the

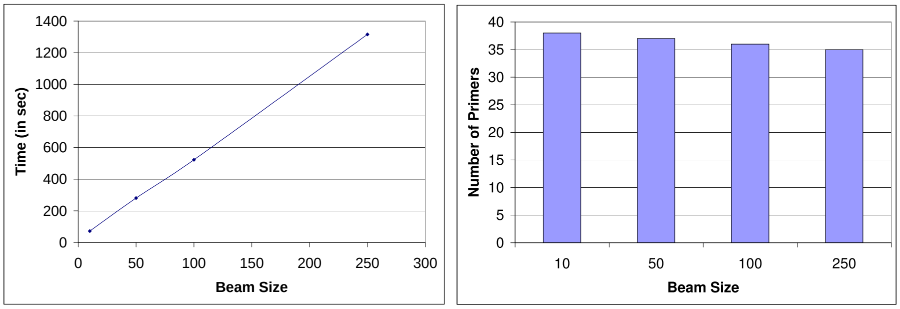

Figure 6.1: The effect of beam size on the (a) running time of the algorithm and (b) number of primers discovered.

solution and the running time of the algorithm. For this particular dataset, there was a decrease of two degenerate primers in the final solution when the beam size was increased from 10 to 100. Moreover, only a slightly better solution was discovered when the beam size was increased to 250. For the average desktop computer, beam sizes larger than a few hundred result in impractical running times. For the input set we used, which contained 95 human DNA sequences, using a beam size of 100 produced solutions that did not significantly improve as the beam size increased. Empirically, a beam size close, in value, to the number of sequences in the input set seems to produce a solution that is balanced in running time and quality.

6.2 Human Dataset

In an unpublished laboratory experiment, a set of degenerate primers of length 20 was manually constructed where each primer was a mixture of 8 specific bases and 12 fully degenerate nucleotides (*e.g.* $AGTCGGTANNNNNNNNNNNN$.) For this experiment, the total degeneracy would be $\approx 4^{12}$. MIPS was originally designed to automate this procedure and, possibly, reduce the total degeneracy and/or number of primers used. In practice the desired accuracy in the experiment determines the actual parameter values used for MIPS. Table 6.2 shows the results. For 95 sequences, 190 primers

Table 6.1: Results on a dataset of 95 human SNP regions.

PT-MDPD		*TT-MDPD*	
Degeneracy	*# Primers*	*Degeneracy*	*# Primers*
$4^6 \approx 4K$	53	$4^9 \approx 262K$	44
$4^7 \approx 16K$	44	$4^{10} \approx 1M$	37
$4^8 \approx 64K$	36	$4^{11} \approx 4M$	30
$4^9 \approx 262K$	29	$4^{12} \approx 16M$	23

would be needed in the general case. MIPS-PT decreased the total number of primers to 15% of this unoptimized value for a degeneracy limit of $4^9 = 262,144$. Table 6.2 includes the similar results for PT-MDPD and TT-MDPD.

6.3 Comparison to HYDEN

The HYDEN algorithm [14] is a heuristic designed for finding approximate solutions to the DPD problems. Recall that DPD is a set of problems where the general goal is to find a *single* degenerate primer that either covers the most sequences while having a degeneracy value less than a specified threshold or covers all of the sequences with minimum degeneracy. The DPD problem is the most closely related one to our MDPD problem.

HYDEN can solve the PT-MDPD problem indirectly by iteratively solving the MC-DPD problem on smaller and smaller sets. After selecting a pair of degenerate primers under a given bound that covers a certain subset of the sequences in an input set, the algorithm runs again on the remaining sequences. For the reasons described below, iteratively solving MC-DPD is not the most effective way to solve the PT-MDPD problem. However, this was the most reasonable comparison that was possible given the implementation available to us at the time of testing. The graphs in Figure 6.2 shows the number of primers that each algorithm found from

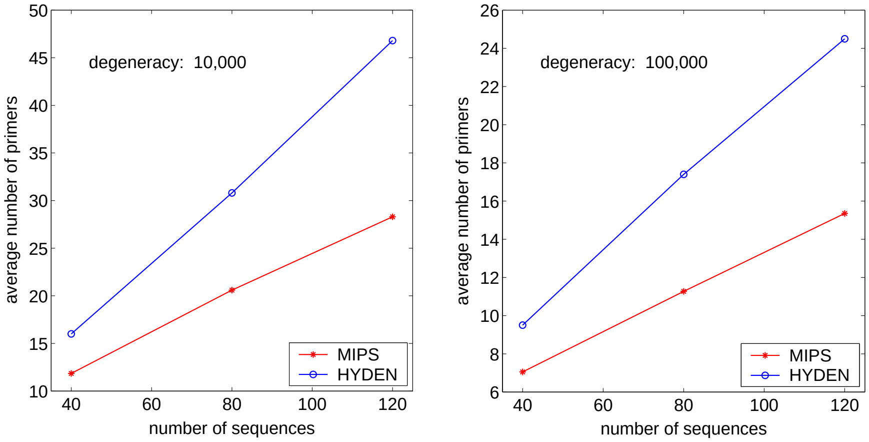

Figure 6.2: The number of degenerate primers selected by HYDEN and MIPS for 20 randomly-generated datasets in solving PT-MDPD for degeneracy thresholds of 10,000 (a) and 100,000 (b).

a randomly generated set of sequences of varying lengths with varying degeneracy thresholds. They are uniformly-distributed i.i.d. sequences of equal length. Each program searched for degenerate primers without allowing any mismatches at any positions.

In general, HYDEN produced more primers than MIPS in attempting to solve PT-MDPD. For a primer degeneracy value of 100,000 and over 100 sequences, the difference was as large as 60% more primers. These results can be partially explained by the differing design requirements of the DPD and MDPD problems. Even when applied iteratively, the goal of the DPD problems is to have a result which could be divided into distinct PCR experiments. The goal of the MDPD problems is to have a set of primers for one large-scale PCR experiment. Specifically, to solve the DPD problem, the HYDEN algorithm must ensure that for any given degenerate forward primer that is discovered, exactly one degenerate reverse primer is used to cover the sequences covered by the forward primer. Therefore, a given degenerate forward primer is restricted to which sequences it is reported to cover based on the presence of a suitable degenerate reverse primer, and vice-versa. Moreover, the HYDEN algorithm has an additional restriction in which any given degenerate primer is limited to covering either a set of forward or reverse primers, but not both.

Chapter 7

Degenerate Primer Design

The results of the previous chapter suggest that MIPS and other algorithms in the domain can be useful in selecting a set of degenerate primers. However, as previously mentioned, the use of degenerate primers generally introduces problems into the biological assay. In this section, we will discuss these problems in depth, show how they are amplified when the background base composition is non-random (such as in the human genome), re-examine the quality of the solutions MIPS produces, and finally suggest improvements to MIPS.

Representing an unnecessarily large set of primers is a problem introduced by the use of degenerate primers. For this discussion, we define *target primers* to be primers that are intended to be used in the PCR assay, and *auxiliary primers* to be primers represented by degenerate primers, which may or may not bind to fragments in the input set, but are not intended to be used in the PCR assay. The two main problems associated with degenerate primer usage are a decrease in the effective concentration of the target primers and an increase in the possibility of amplifying an unexpected region, or mispriming. In order to explore these problems, in the next two sections we consider the following questions:

- How effective do we expect a given degenerate primer to be? In other words, for the set of primers that a given degenerate primer represents, what is the ratio of target primers to auxiliary primers?

- Given the presence of these auxiliary primers, how often do we expect to see an unexpected PCR product?

7.1 Degenerate Primer Efficacy

Multiplex primer design demands that many input sequences share sites complementary to some common (possibly degenerate) primer. In the general case, the sequences to be co-amplified are not related, so their complementarity to a common primer is largely a matter of chance. We therefore explore the chance-imposed limits of multiplexing, that is, how many unrelated DNA sequences are likely to be covered by a single PCR primer of a given degeneracy?

Let $\mathcal{S}$ be a collection of n DNA sequences of common length m. Call a primer P an *(l, α, k)-primer* for $\mathcal{S}$ if it has length l and degeneracy at most α and covers at least k sequences of $\mathcal{S}$. A natural way to quantify the limits of multiplexing is to compute the probability that an (l, α, k)-primer exists for $\mathcal{S}$. However, this probability is difficult to compute, even assuming that $\mathcal{S}$ consists of i.i.d. random DNA with equal base frequencies. We instead compute the *expected* number of (l, α, k)-primers for $\mathcal{S}$. If this expectation is much less than one, Markov's inequality implies that $\mathcal{S}$ is unlikely to contain any such primer.

We do not count the total number of (l, α, k)-primers for $\mathcal{S}$ but only the number of *maximal* primers. A primer P of degeneracy at most α is said to be *maximal* if increasing P's degeneracy at any position would cause its total degeneracy to exceed α. The expected number of maximal (l, α, k)-primers for $\mathcal{S}$ is in general less than the total number of (l, α, k)-primers, but a primer of this type exists for $\mathcal{S}$ if and only if a maximal primer exists. Hence, the former expectation is more useful than the latter for bounding the probability that at least one (l, α, k)-primer exists.

7.1.1 Occurrence probability for one fixed primer

Let P be a primer of length l, such that the jth position of P permits $|p_j|$ different bases. Let $\mathcal{S}$ be a collection of n i.i.d. random DNA sequences of common length m with equal base frequencies, and let T be a single l-mer at a fixed position in some sequence $S_i \in \mathcal{S}$. Say that P *matches* T if P would hybridize to T. We have that

$$\begin{aligned}\Pr[P \text{ matches } T] &= \prod_{j=1}^{l} \frac{|p_j|}{4} \\ &= \frac{d(P)}{4^l}.\end{aligned}$$

The probability that P covers S_i, i.e., that it matches *at least one* l-mer of S_i, depends in a complicated way on P's overlap structure, but if S_i is not too short and $d(P)/4^l \ll 1$ (both of which are typically true), then using Poisson approximation [26],

$$\Pr[P \text{ occurs in } S_i] \approx 1 - e^{-\frac{d(P)}{4^l}(m-l+1)}.$$

Let q be the probability that P matches somewhere in a single sequence of length m, and let $c(P)$ be P's coverage of $\mathcal{S}$, i.e., the number of sequences of $\mathcal{S}$ in which P matches at some position. Because the sequences of $\mathcal{S}$ are independent, the probability that P matches in at least k sequences given by the binomial tail probability

$$\Pr[c(P) \geq k] = 1 - \Pr[B(n, q) < k],$$

where $B(n, q)$ is the sum of n independent Bernoulli random variables, each with probability q of success.

7.1.2 Computing the expectation

Let $\Pi(l, \alpha)$ be the set of all maximal primers of length l and degeneracy at most α. To count the expected number $E_{l,\alpha,k}$ of (l, α, k)-primers for $\mathcal{S}$, observe that

$$E_{l,\alpha,k} = \sum_{P \in \Pi(l,\alpha)} \Pr[c(P) \geq k].$$

Enumerating all $P \in \Pi(l, \alpha)$ to compute this expectation would be computationally expensive, but this enumeration is not needed for i.i.d. sequences with equal base frequencies. Given these assumptions about $\mathcal{S}$'s sequences, the probability that P matches a given l-mer does not change if we rearrange its positions (e.g. "AMC" versus "MCA") or change the precise nucleotides matched (e.g. "RTG" versus "MCA"). Let W be a multiset of l values drawn from $\{1, 2, 3, 4\}$ that lists the degeneracies n_j (in any order) of a primer from $\Pi(l, \alpha)$. Then every primer described by the same W has the same probability of covering at least k sequences in $\mathcal{S}$. Hence, the desired expectation is given by

$$E_{l,\alpha,k} = \sum_{W} \#(W) \Pr[c(P) \geq k \mid P \text{ described by } W].$$

where the sum ranges over all feasible W for $\Pi(l, \alpha)$ and $\#(W)$ denotes the number of degenerate primers described by W. The probability is computed as described above, so we need only describe how to compute $\#(W)$.

Let W be a multiset with n_1 1's, n_2 2's, n_3 3's, and n_4 4's. If we fix *which* positions in P permit 1, 2, 3, and 4 nucleotides respectively, then there are $4^{n_1} \times$

$6^{n_2} \times 4^{n_3}$ ways of assigning nucleotide sets to these positions. Hence,

$$\#(W) = \begin{pmatrix} & l & \\ n_1 & n_2 & n_3 \end{pmatrix} 4^{n_1+n_3} 6^{n_2}.$$

Enumerating all feasible W for $\Pi(l, \alpha)$ is straightforward, so the expectation can be computed.

7.1.3 Results

The theoretical estimates of the previous section can be used to evaluate whether a particular primer-design algorithm performs well on the MC-DPD problem, that is, whether it finds degenerate primers with coverage close to the maximum predicted for a given set of input sequences. We evaluated the MIPS algorithm's performance on MC-DPD by comparing the primers it found in random DNA with those expected to exist in theory. For these experiments, we generated test sets of i.i.d. random DNA sequences with equal base frequencies with $n = 190$, and $m = 211$, so that the number and average length of the test sequences roughly matched those of the human DNA test sequences.

We used MIPS to find a single primer with maximum coverage in each test set, subject to varying degeneracy bounds α. Table 7.1 compares the average coverage of primers found by MIPS in 20 trials to the largest coverage k such that $E_{l,\alpha,k}$ for test sets of the specified size is > 1. Primers with coverage exceeding this value of k are not expected to occur in the test sets, while primers with slightly smaller coverage may or may not occur frequently.

MIPS proved adept at finding primers close to the maximum predicted coverage for relatively small degeneracies ($\alpha \leq 10000$). We therefore have considerable confidence in its ability to find high-coverage primers if they are present. The gap

Table 7.1: Actual and predicted coverage of 20-mer primers found on sets of 190 random sequences of length 211. Avg Coverage: average coverage of primer found over 20 random trials. Max Predicted: largest coverage m such that $E_{20,\alpha,m} > 1$.

degeneracy α	Avg Coverage	Max Predicted
1000	6.30	7
10000	10.55	12
100000	19.30	26

between the best primers found by MIPS and those predicted to occur in theory grows with the degeneracy bound, but we cannot say with certainty whether this fact represents a limitation of the algorithm or of the theoretical estimates, since primers with expectation greater than one may with significant probability still fail to occur. Moreover, the high degeneracies where MIPS might perform poorly are of less practical interest, since single primers with such high degeneracies are experimentally more difficult to work with.

Overall, MIPS appears to be operating close to the theoretical limit for MC-DPD problems of small degeneracy. Although our analysis does not directly address the MDPD problems, any large gap between the most efficient design and the designs produced by MIPS is unlikely to arise from failure to find single high-coverage primers when they exist.

7.2 Mispriming

Due to the presence of auxiliary primers, it is possible that a pair of primers binds to an undesired location and results in an erroneous amplification. *Mispriming* is the occurrence of this event where the unwanted PCR product is indistinguishable, by size, from the targeted products.

Suppose we design a set of degenerate primers with length l, such that the *total degeneracy* of the set is α. We wish to estimate the expected number of mispriming events when our primer set is applied to a genome of length g. The background model greatly influences the calculations, therefore, we will consider two models separately, an i.i.d. random genome with equal base frequencies and the human genome.

A pair of l-mers cause a mispriming event if and only if they bind to the genome within δ bases of each other, in the appropriate orientations to permit amplification of the sequence between them. Let i index the positions of the genome on its forward strand. Let the 0-1 random variable x_i indicate the event that an l-mer from our primer set is complementary to the forward strand at position i, and let $\overline{x}_i$ be the event that an l-mer is complementary to the reverse-complement strand at i. We say that a mispriming event occurs at i if $\overline{x}_i \cap \bigcup_{j=i}^{i+\delta-1} x_j = 1$. Denote this event by the 0-1 indicator M_i. The total number of mispriming events M in a genome of size g is simply $\sum_{i=1}^{g} M_i$.

7.2.1 Mispriming in i.i.d. Random and Human Genome

In this section we consider the background model where the genome consists of i.i.d. random sequence with equal base frequencies. For the expectation of a matching event to occur at a position i we have that $E[x_i] = E[\overline{x}_i] = \frac{\alpha}{4^l}$.

Note that the two matching events are independent in an i.i.d. random DNA sequence when the two primers do not overlap. To simplify our calculations, we ignore the effect of overlapping primer boundaries. Using Poisson approximation to estimate the probability of the matching event on the forward strand, we have that

$$\begin{aligned} E[M_i] &= E[\overline{x}_i \cap \bigcup_{j=i}^{i+\delta-1} x_j] \\ &= E[\overline{x}_i] E\left[\bigcup_{j=i}^{i+\delta-1} x_j\right] \\ &\approx E[\overline{x}_i]\left(1 - e^{-\sum_{j=i}^{i+\delta-1} E[x_j]}\right). \end{aligned}$$

Finally, setting $\rho = \alpha/4^l$, we derive the expected mispriming rate as

$$\begin{aligned} E[M] &= \sum_{i=1}^{g} E[M_i] \\ &\approx g\rho\left(1 - e^{-\delta\rho}\right). \end{aligned}$$

To test the accuracy of these calculations, we constructed a human-size genome ($g \approx 3 \times 10^9$) of i.i.d. random sequence of equal base frequencies. We obtained results from MIPS-PT on the human dataset used in Chapter 6. Finally, we simulated a PCR experiment using both the test genome and the human genome (10 Apr 2003) [25, 11], assuming that the primers in the solution would all bind to complementary fragments, thus ignoring inexact binding. In accordance with the calculations, we considered a mispriming event an instance of a matching event occurring in one strand and another matching event occurring on the opposite strand within $\delta = 500$ bp. Table 7.2 shows the total degeneracy of the solution, the number of predicted mispriming events, and finally the number of mispriming events seen in the simulation of the test and human genomes.

Table 7.2: Predicted and actual mispriming rates in simulated PCR experiments with i.i.d. random and human genome.

Total Degeneracy	Predicted	Random Genome	Human Genome
84720	0.009	0	82254
321456	0.133	1	112162
1262260	2.063	6	64938
4824870	30.12	81	201209

The model predicts the mispriming rate well for the test genome, however fails to predict the same for the human genome. In the next section we discuss implications of these results, the complexity involved in properly calculating the expected human mispriming rate and possible heuristics that can be employed to select effective degenerate primers which do not misprime with such high frequency.

Early data from the results of the Human Genome Project [25] strongly suggest that the sequence and frequency of the bases of the human genome is not random. The evidence lies in the presence of interspersed repeats and regions of low-complexity sequence [21, 22]. The presence of repetitive elements in the human genome can affect the mispriming rate of the MIPS solver by violating the implicit assumption that a degenerate primer's mispriming rate is solely determined by the degeneracy of the primer.

7.2.2 Reducing Mispriming Events

Consider an input sequence that contains fragments which are overrepresented in the genome. If MIPS selects any of these loci as the primer binding site in the final solution, the likelihood of a mispriming event increases when screened against the sequence of the human genome. The solution to this problem, therefore, is not to allow MIPS to select these fragments. For this, we processed the input sequences with RepeatMasker [23], which masks sequence fragments which are overrepresented in certain genomes, in our case 'Primates'. Using the human SNP input set, Figure 7.1 shows the results of applying RepeatMasker and the percentage of bases that were masked by the algorithm.

A side effect of using the masked input set was that two of the sequences of the input set were rendered unusable. The masking process effectively reduces the size of the input sequences and therefore the possible binding sites. Two of the input sequences did not contain 20 consecutive unmasked bases, or any possible binding sites, so they were omitted. Table 7.3 shows the reduction in mispriming events when the input sequences are masked by comparing the mispriming rates of the results of MIPS-PT on the unmasked input set versus the masked input set.

```
RepeatMasker summary:
==================================================
file name: RMemail6411.seq
sequences:          190
total length:     34874 bp  (34874 bp excl N-runs)
GC level:         40.59 %
bases masked:      2756 bp (  7.90 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               14         1217 bp     3.49 %
      ALUs            8          601 bp     1.72 %
      MIRs            6          616 bp     1.77 %

LINEs:                4          493 bp     1.41 %
      LINE1           3          392 bp     1.12 %
      LINE2           1          101 bp     0.29 %
      L3/CR1          0            0 bp     0.00 %

LTR elements:         5          551 bp     1.58 %
      MaLRs           2          289 bp     0.83 %
      ERVL            1           97 bp     0.28 %
      ERV_classI      2          165 bp     0.47 %
      ERV_classII     0            0 bp     0.00 %

DNA elements:         2          109 bp     0.31 %
      MER1_type       0            0 bp     0.00 %
      MER2_type       1           78 bp     0.22 %

Unclassified:         0            0 bp     0.00 %

Total interspersed repeats:     2370 bp     6.80 %


Small RNA:            2          218 bp     0.63 %

Satellites:           0            0 bp     0.00 %
Simple repeats:       3           95 bp     0.27 %
Low complexity:       3           73 bp     0.21 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The sequence(s) were assumed to be of primate origin.
RepeatMasker version 07/07/2001 , default mode
run with cross_match version 0.990329
RepBase Update 6.3, vs 05152001
```

Figure 7.1: Results of RepeatMasker on human SNP input dataset.

Table 7.3: Mispriming rates in simulated PCR experiments with original and masked input sets

Degeneracy Threshold	Original	Masked
$4^6 \approx 4K$	82254	164
$4^7 \approx 16K$	112162	1104
$4^8 \approx 64K$	64938	2043
$4^9 \approx 262K$	201209	17337

Empirically, these results seem to indicate that simply removing overrepresented fragments from the input set render the results of MIPS far more useful in practice by reducing the number of predicted PCR artifacts.

Another interesting result of masking the input sequences for this particular dataset is the resulting solution from MIPS-PT. Intuitively, it is expected that reducing the size of the input set would likely increase both the size and total degeneracy of the final solution when compared to the original data set since the likelihood of finding similar fragments is decreased. Table 7.4 shows the number of primers selected and total degeneracy of the final solutions for both the original and masked input set. The full output of MIPS-PT for both of these input sets can be found in Appendix B. For each degeneracy threshold tested, MIPS-PT selected fewer primers for the masked data set and on two occassions the total degeneracy of those primers was also less than that of the original set.

Table 7.4: Comparison of MIPS-PT results on original and masked input sets.

	Original		*Masked*	
Threshold	*# Primers*	*Degeneracy*	*# Primers*	*Degeneracy*
$4^6 \approx 4K$	53	84720	49	128144
$4^7 \approx 16K$	44	321456	42	319872
$4^8 \approx 64K$	36	$1.262 * 10^6$	34	$1.299 * 10^6$
$4^9 \approx 262K$	29	$4.824 * 10^6$	28	$4.277 * 10^6$

7.2.3 Alternate Strategies to Mispriming

Using RepeatMasker on the input set dramatically reduces the number of expected mispriming events by eliminating input sequence fragments which are overrepresented in the genome. However, it is still possible that one or more of the degenerate primers selected represents an overrepresented fragment which does not occur at all in the input set. Consider this simple example where the sequence $ACACACAC$ is a repetitive element in the human genome. If the final solution of MIPS includes the primer $MMMMMMMM$ where M is the degenerate nucleotide which represents $\{A,C\}$, then this solution could lend itself to a large number of mispriming events even though the particular repetitive sequence is not necessarily a part of the input set.

The problem is that certain degenerate primers represent overrepresented sequence fragments and it is not desirable to select these primers in any final solution. Therefore, we want a method to determine whether or not a given degenerate primer is likely to cause a large number of mispriming events *before* it is selected as part of a final solution. A simple workaround would be to maintain a list of each degenerate primer and its frequency in the human genome. A scoring function could then be generated to calculate the likelihood of a degenerate primer being involved in a mispriming event. However, there are over 10^{24} degenerate primers of length 20 and maintaining such a data structure is currently infeasible.

Another workaround would be to dynamically calculate such a likelihood for each primer as they are encountered in the beam search. It is feasible to estimate the probability of a degenerate sequence appearing in a complex background model such as the human genome using a high-order Markov model and a dynamic programming algorithm, similar to the Viterbi algorithm [8].

Chapter 8

Conclusions

SNP Genotyping is poised to become an important procedure in the future of human genomics. Based on sound theoretical principles, the application ideas in various domains are on the verge of implementation. One of the final barriers to realizing this promise rests in a practical, cost-efficient technique for large-scale DNA analysis. The work of this thesis focuses on a problem that arises in high-throughput multiplex PCR experiments, which is a major part of one of the proposed SNP Genotyping techniques.

We developed an iterative beam-search heuristic, MIPS, for this problem which can be used to select a set of degenerate primers for a given set of sequences. This algorithm compares favorably to an existing algorithm for similar problems. Using both theoretical calculations and experimental analysis, we have shown that MIPS provides results which are close to the theoretical limits of degenerate primer design. We also discussed the practical limitations of the algorithm and the modifications that can be employed to improve upon the solutions. MIPS is neither time nor memory intensive and could conceivably be used as a desktop tool. The overall effectiveness of this algorithm will ultimately be determined by the application of the resulting primers in biological experiments.

Appendix A

MIPS Pseudocode

Algorithm 1 MIPS($\mathcal{S}, \alpha$)

```
Initialize Global variables (2-D matrices): BEST - candidate fragments; COVERED - sequences covered; ALLOWABLE - remaining degeneracy, $ALLOWABLE(0,0) = \alpha$.
for $p = 1$ to the number of degenerate primers that will be used do
    Let $c$ = the maximum number of sequences that the ($p$-1) primers covered
    while $c > 0$ do
        $MIPS\_SEARCH(\mathcal{S} - COVERED(p-1, c), ALLOWABLE(p-1, c), p, c)$
        if this search covers S, print solution and exit
        else c=c-1
    end while
end for
```

Algorithm 2 MIPS_SEARCH($\mathcal{S}, \alpha, p, c$)

```
Input: Sequence set $\mathcal{S}$, degeneracy bound $\alpha$, primer number $p$, sequences covered $c$.
Output: total number of sequences covered
Initialize priority queue $Q$ of size $b$;
Perform pair-wise comparisons.
for all sequence $S_i \in \mathcal{S}$ do
    for all substring $S_i[j, l]$ do
        Let $\mathcal{C} = \{ x | \langle f, x \rangle \in T$ and $f$ is a $k$-length substring of $S_i[j, l] \}$
        for all fragment $C_k \in \mathcal{C}$ do
            $D = S_i[j, L] + C_k$
            Insert $D$ into queue $Q$
        end for
    end for
end for
Let $c' = c$
while queue $Q$ is not empty do
    Let $P =$ the best element of $Q$
    if degeneracy($P$) $<$ degeneracy($BEST(p, c)$) then
        $BEST(p, c') = P$
        $ALLOWABLE(p, c') = \alpha -$ degeneracy($P$)
        $COVERED(p, c') = COVERED(p - 1, c) \cup$ covers($P, \mathcal{S}$)
        $Q = ONE\_PASS(Q, \mathcal{S}, \alpha)$
    end if
    $c' = c' + 1$
end while
return $(c' + 1)$
```

Algorithm 3 ONE_PASS($Q, \mathcal{S}, \alpha$)

```
Input: Priority queue $Q$, set of sequences $\mathcal{S}$, degeneracy bound $\alpha$.
Output: Priority queue $Q'$
for all primer $P \in Q$ do
    for all sequence $S_i \in \mathcal{S}$ do
        if $S_i \notin$ covers($P$) then
            for all substring $S_i[j, l]$ do
                $D = S_i[j, l] + P$
                Insert $D$ into queue $Q'$
            end for
        end if
    end for
end for
return $Q'$
```

Appendix B

Supplemental Data

The following figures are the full output of MIPS-PT on a dataset of regions of human DNA surrounding 95 known SNP locations.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 4100
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 190

Total # Primers: 53
Primers                Degeneracy    # Covered
GARATMWCWWYWRMAGAAAT   512           2
GAAYATAGTARGSYYTCWKT   128           2
YRTSCATTTATMTTRGASTG   64            2
TCYKSMTSTGAAAYYTRSMK   2048          3
ACYTKKRARTYCCTTHSYST   1536          3
ARRWGGKGCWRGRTSYTGRY   2048          3
CAYWAGSCARGABYWRRKGT   1536          3
GHWGSARYHTVTRTCACCCC   864           3
TCASMTGKMCAWCAMASTSY   512           3
MASCWYMVATYSTGTGKCTG   768           3
CWSTHTCTRMWTCTGYCMTM   768           3
TABAMACHTTYMWCAWCAKT   576           3
KATTAKTWVTAAYMAATDAW   576           3
AWBKATGCTSWDTTTTGTSY   576           3
CCTYKMACWTWTMWWAASAG   512           3
TRRCTRARAYAAGWYKCAKG   512           3
CCWMYTCTRSTGRSYKTGCA   512           3
CASARAKSAGGWGGCMWMGW   512           3
AKMSACAGAKDKBTTTGCYG   576           3
AGMCAGAGGTVRGAKMHKRG   576           3
AWRWTWGWAWBRMAAKRTTT   3072          4
TTCTTTYKMATWGVRATSYC   384           3
WGKHYKTTYCTSWBTHTAAA   3456          4
WWAWCMTAWBCMCMCASRSA   3072          4
TKWCTGYRDKYYTBTSCTTG   2304          4
WTWAWAHTAWGCAWTKARTA   384           3
GDAKKGGGWGAYWTYCCTTM   384           3
WGDDRARGAAAKTGAGRVWR   3456          4
WKHAAWARKTYWTMAADATT   2304          4
SCCWKTCTCWTTCAVRCCAR   192           3
WWMCCTBYWWCMTCTCTKMT   1536          4
WHTCTCCACDYCMDMCTSYY   3456          4
AKGRRNAAAGGRAAGWVGVW   2304          4
YTASRRTTTHCWHTYTKCAW   2304          4
YTWSAAWTWWTTACARHMAS   1536          4
KTTKSWGKTYTTHMMMACTR   3072          4
TAAMAWWRRTSAYTGMMDTT   1536          4
AGAVRAGCARARRGRBSWWA   2304          4
WSAAKAWRCYKADGVTTWAA   2304          4
ATRKKRGRMCTKTGGTRRRW   2048          4
CTGVYTKGARRRAMSWCAMT   1536          4
ATWWBTWCTKTKGSYMTTTR   1536          4
MMARAACAVAMACASRYVSA   2304          4
MWKBMARAGVAWWTCATWAA   2304          4
CTTYYYWCCHCCCCTBYYWK   2304          4
WGTGYTSSSWTWASWGSYGT   2048          4
WTATTBYCAMMAMKYTTTSW   1536          4
TAGGCADYVAANAAABAWWT   8634          4
TTKARKDAACTTWHTYWAWG   1152          4
AWMRARARGRARAAMAMRKW   4096          5
WMBATTKTKHDTWTTTAWMT   3456          5
TWKTTTDTTDKTHTDTBTKA   3888          6
AGACYSTGTYWSHWAAAAAD   576           5
-----------------------------------------
Total Degeneracy:      84720
```

Figure B.1: MIPS-PT output with $\alpha = 4^6$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 16400
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 190

Total # Primers: 44
Primers                 Degeneracy  # Covered
TGWADWAABTMHYDARKMAA    10368       3
WMMKCYCADCTRDSTKCYTS    9216        3
AARYCWKSAABATWKTADKS    4608        3
CMYSRWRTCCWGSYTCCCWG    1024        3
BMTBTSAARGSAACYRYWCA    2304        3
WSMHAKMCCTWBACTGTHCA    1728        3
ASMWCYTBHTSTKAAATTWG    1152        3
KKRDDGTGDGTRARMRKRAA    13824       4
ARTASSWDRHGDRRGWTCMC    13824       4
TYMAWRACTGWKDYMWKWTK    12288       4
TDWTAGAAANRMAADDWYTW    6912        4
CTCTYWBHWGYYTGKDTCYW    6912        4
TKTNRSDKATGAGAGDRVWG    6912        4
ASTYTCWWSAYCAKYMMMWY    8192        4
MWGCYTCTKBCMWHYWCABA    6912        4
CCWMYTYWRSTGRVYKTGCW    6144        4
AMWYKWAKGAAHDTSTTTMY    4608        4
TRAAWYYYSTYTMTGWBWTW    6144        4
TTCTTYYKMWYWGVRATBYC    4608        4
MAYTGMTTWTGHRWWWTKWA    3072        4
YHTMATCWKMTKTYWYWTTT    3072        4
ASACAKARGKVASRDCYWRG    4608        4
SYCWKYCTCHTYCMVRCCAR    4608        4
AKTAABTWWTATYTSYWYWW    3072        4
KBTWAAYAGWTTADGWHWWT    3456        4
SCCAKWGWCWGADWTYYTTB    2304        4
YWMCMCTBYWWCHTCTCYYM    9216        5
KGDAKKGRSWGAYWTYCYTT    3072        4
YTASRRTYTHCWHTYTBYAW    13824       5
AGAVRAGCARARRGRBSWWA    2304        4
CTGVYTKGARRRAMSWCAMT    1536        4
WAAMAWWDRTSAYTGMMDYT    9216        5
WRWGYTSSSWTWABWGSYGT    12288       5
GRDGAAABKKARRBTKTAWD    10368       5
SCTKNYYWYCWCMYCTVYCA    12288       5
ATWWNTWYTKTKGBHMTTTR    9216        5
TYMTTYARAHWSAWRRYAWA    6144        5
MMMVAACAVAYRMASRSVCA    13824       5
TTKADKDAWMTTWHTYWAWG    6912        5
KAKGMAATBARDRMHDAAVT    15552       6
WMBATTKTKHDTWTYYAWMT    13824       6
AWMVARARSRARAAMAMRKW    12288       6
TWKTTTDTTDKTHTDTBTKA    3888        6
AKWYYYTKTYTCHWAAAHWD    13824       7
-----------------------------------------
Total Degeneracy: 321456
```

Figure B.2: MIPS-PT output with $\alpha = 4^7$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 66000
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 190

Total # Primers: 36
Primers                 Degeneracy  # Covered
TGCTATGCCCAGGTGGCCAG    1           1
YHYAGTWTMAAWBKRYWRMA    18432       3
TDABRMMRYTTTMWTKATSA    4608        3
YMYMCMTTBYBSHRHYAACT    41472       4
WHTVWCCWHYYKBCTSTSAG    20736       4
SMRABCWNHTBWACAKRWWT    55296       5
DATGRHTRTCYTBWTBHABT    11664       4
YTVWGDKGARKAAMDTSAVA    10368       4
CTCTYHBHWKYYTGBDTCYD    46656       5
CASAVAKVAGGHRGSHHHRW    46656       5
HGTSDVSWKGRARGVSCYSC    41472       5
TTBCYDTWCYMYWWHWMABC    41472       5
RTKTGAWKWRNRTGDRWWTR    24576       5
YTSDBAGCHARRSSWWSKWG    55296       5
MDGGARRCCTBYKSMYWYMW    36864       5
SADRSTRASTKYTYCCHDRW    27648       5
MTTCAHSMHTWRRWTKDRSA    27648       5
TSTCTKYDKKYMYBTVCTTK    13824       5
KHAMWWWRTAHKAARMWKTT    18432       5
WDCWKHTYTMTYTMAAWDYH    41472       6
RWWTYHAHWRATATWWHKTB    41472       6
WWMCCTBYWWHMTMTCWKNT    36864       6
TKMTRKTYTBVAWAWMTDKS    27648       5
CTGVYTKKARRRAMSWYMHT    18432       5
WRADDWKCMRAAABKSBAAV    62208       6
TMWCANTGDTKMYKNDADTT    27648       6
WWHWNTHYTKTDKGWMTTTR    55296       6
MHADAHMAVWCAMAVRCNSA    62208       6
WDYYMCMYTCCHVCYWSHCC    41472       6
MYTKYAVWKDAWWWYAWWAA    36864       6
TYBAWKKADCHTRCWTHWWK    41472       6
TATTNBCAMAAHBTWYTVHH    23328       6
WNVANTKHRAKWWTKTADAT    55296       7
AWMRRRARGRARWAMAHDKW    36864       7
TWKTYYDTTDKKHTNTNTKA    55296       8
AKWYYYTKYYTYHWAAAHWD    55296       9
-----------------------------------------
Total Degeneracy: 1.26226e+06
```

Figure B.3: MIPS-PT output with $\alpha = 4^8$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 262200
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 190

Total # Primers: 29
Primers                 Degeneracy  # Covered
KKRWAWAWMTDYWSAARDMA    36864             3
STHTTGKGKVWKKYWBYMYY    110592            4
YWYYWDCWKHRAWMHYTKSA    221184            5
MMARDAHBWGAWKYVRGTKD    124416            5
GASDDRSHAVRKGMTBHYAG    93312             5
CWGKASRHAGNSYDDGVMTS    82944             5
ARHWDWTWKYWCKSMTTYBT    55296             5
GMVDCWRRDAKGWVRMGGVH    186624            6
RYHWSMWTKKHTATKWBMDT    165888            6
KBHDTTTCYDYWCHYWKNKG    248832            6
MTGWWTSTSHMHWASAHDNR    165888            6
HHVVMWSHTHCCYYTCTDHT    209952            6
WADKARMATBKYHBWTBSAY    124416            6
ASAMADWRRKVANVDCYWRG    165888            6
AKTAAYWWHDHTTTNNWSNW    221184            7
GTGDGCYACHGHNYVDDKYY    93312             6
WHHAWRHYWWGMAWWKANTW    221184            8
YVKGDNBCTMMSYYTCHTSH    248832            7
BNNTKVTKKTCTBVAWRDMT    248832            7
WRSKKKBARRKAWGMBWBTS    221184            7
WHAWMBHHTKTKKSYHTWTG    124416            7
DSNAKDGRSDGANWTYYYTK    221184            8
WTWMWTYWRAWWVRWRRHWA    147456            8
HMMNAMCABAYDCHSRCVCA    62208             6
WDYYMSMYTCCHVSYWSHCC    165888            7
WRKHHWTBAWRNWADWAATD    248832            9
AWMVRRARSRARWAMAHDKW    110592            8
HHBWTTNDKWDTWTTTWWRW    248832            10
WNWBYYTBTYTYHWAAAHWD    248832            11
---------------------------------------------
Total Degeneracy: 4.82487e+06
```

Figure B.4: MIPS-PT output with $\alpha = 4^9$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 6600
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 188


Total # Primers: 49
Primers                Degeneracy     # Covered
GWGGKGCTRGGTGCTGRCAG   16             2
TATGAATWTMBTKATKHAKT   288            3
TATGAATWTMBTKATKHAKT   288            3
TATGAATWTMBTKATKHAKT   288            3
CWSTHTCTRMWTCTGYCMTM   768            3
AWAYRGAGHRWWWRAAAAAA   768            3
ATAAMWTGRRAGSMAMRTVA   768            3
CCWYDAACWTTTMWKAAYAK   768            3
CMRCACRAAKMAGGWGRCMW   512            3
YWARAGGAWKAGCTRTGBTS   384            3
MRGTYAWSTTBATAAKMTCT   384            3
SWKTTTCYKYTSMCWKKGGM   4096           4
AMTTCTSSCMRDKRWMTGRY   6144           4
MYHTTTTWYWKHAARAWMTG   4608           4
WDWSTGWKTGAVCWKGRASG   4608           4
RWATGHMATATTKTWRATBH   1728           4
ATWTRSYTWTBCAKTTSHAM   2304           4
GGAWMATRABVAYATBRAWS   3456           4
AMWYKWAKGAAHDTSTTTMY   4608           4
ASACAKARGKVASRDCYWRG   4608           4
WGTWYBTTTMWKAHTDTAWA   3456           4
RTBCYSWDTBTAWAAATRYA   3456           4
RGMWTYTTSMCKWWGSMAGM   4096           4
CMYAGWCTSWYYSARRSCAR   4096           4
TRRATTYTBTBDCTGWTRWM   3456           4
TGWAWTYWTWDATATHWWKT   2304           4
YWWWAWAMWAWKVATTTART   3072           4
AGHMAWWRTTMWYAMAAAYY   3072           4
TWYAAWTARTKACWDAVWCD   3456           4
ARMTTTTYTHTHTSAHTWTB   2592           4
YTASRRTTTHCWHTYTKCAW   2304           4
WACCCTBYWWCHTCTCYYMK   2304           4
KTTKSWGKTYTTHMMMACTR   3072           4
RTAAMATWWKCCCHSASRBA   2304           4
TAAMAWWRRTSAYTGMMDTT   1536           4
CHCHCARGYCASYTWYSWTT   2304           4
AGAVRAGCARARRGRBSWWA   2304           4
ATRKKRGRMCTKTGGTRRRW   2048           4
ATWWBTWCTKTKGSYMTTTR   1536           4
CMADAMMARWCAMARRCNCA   3072           4
MTKKMARWGRAWDTCATWAM   3072           4
TTWHWTTAAAWWMRWGRWWW   6144           5
MTCYYCMYTCYHDCCWCYCC   2304           4
TYTBYWWAMTGTAATADRMM   2304           4
TTTYYCWMKYYWWMCCTTTW   2048           4
TTKARKDAACTTWHTYWAWG   1152           4
AKSHAATKAVDAAWDRAAWG   5184           5
WMBATTKTKHRTATYYAWMT   4608           5
AWMRARARGRARAAMAMRKW   4096           5
-----------------------------------------
Total Degeneracy:      128144
```

Figure B.5: MIPS-PT output on masked input set with $\alpha = 4^6$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 16400
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 188

Total # Primers: 42
Primers                 Degeneracy  # Covered
ADAYRGARHDHWWVAAAAAA    7776        4
RAAANAMWAAATGRSSRHVS    9216        4
TATDWCWKWYTWTWWATVMH    13824       5
ARGMYATSAAANBWMYMTKT    6144        4
KMRADGGRHAAARGRAWGAV    3456        4
YTKVCCTVTGTGNSDCCBBT    7776        4
YTWVTRAMYWYCTTTMTVTH    6912        4
RCARAASSAWBWGYKDTGDT    6912        4
CTRWWTSTSYCYWAGANKHA    6144        4
WRCMCTSHTWCYTCTSYYMK    6144        4
WDWSTGWKTGAVCWKGRASG    4608        4
ASACAKARGKVASRDCYWRG    4608        4
GKDMYCAKKRTADAWMTGCW    4608        4
YABAWHKWDTTYTTMAAAAW    3456        4
GGAWMATGWBSACAKSVMWS    4608        4
WWYTDWTWTWKMRTTWTTTA    3072        4
WRHATKYAYRAATATMWTKW    3072        4
RAYYTCTBBCCAKTMTMYRR    4608        4
CMYAGWCTSWYYSARRSCAR    4096        4
AKTAAYTWWKWTKKSYWCWA    4096        4
WTWYTTHAWAWVAWKDTKWA    13824       5
YTWCADWTWATTWHWAAMAH    3456        4
WTTWVTMWCTYVATATYAYK    2304        4
AGAVRAGCARARRGRBSWWA    2304        4
MWWWAWAMYRWGMATYWMRT    16384       5
KGDAKKGRSWGAYWTYCYTT    3072        4
TMTKTGKKTMKDYRCHTBTS    13824       5
WTWTHYTDTWTYTAWAAHWW    6912        5
CHCHMARGHCABYTWYVWTT    15552       5
TTKYDKKDCWTHAAAAYTDK    10368       5
TGGKGHCYVHRKYTYWSCYC    13824       5
WTWWVTWYTKTKGKWMTKTR    12288       5
RDAACATWTKYSMMMAVRBA    13824       5
TYTHTCTYTYTDDAYYKWYT    6912        5
TTWHWTTAAAWWMRWGRWWW    6144        5
TYTBBWWAMTGTAATWDDMH    15552       5
HTTDMRDARRAWWTMATWAA    6912        5
CTKNYYWYCWCMYCTVYCAS    12288       5
TTKADKDAWMTTWHTYWAWG    6912        5
AKSHAATKAVDAAWDRAAWG    5184        5
WMBATTKTKHRTATYYAWMT    4608        5
AWMRARARGRARWAMAHRKW    12288       6
-----------------------------------------
Total Degeneracy: 319872
```

Figure B.6: MIPS-PT output on masked input set with $\alpha = 4^7$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 66000
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 188

Total # Primers: 34
Primers                 Degeneracy  # Covered
GVBRRAWRDHTRHSCTKTST    62208       5
AGBABVTRRMWVTDCYCTWB    46656       5
YNSHRCAMVAAKHAGRWRRC    55296       5
YWTKBAYTKDYRAMTKYMTK    36864       5
CHRKGDMWAADBRCAKRWST    41472       5
KBBTYTWAAAARMTBWKBYR    41472       5
CASHSWSRBTSAGCHTBMVA    31104       5
TBTTHMWTGMAYKTKDMWYW    27648       5
AAWVRMTGWKSYVKKDTGGK    27648       5
AAADHYDDNDCMTTNMAAAT    31104       5
RKDMBCAKKRKADAWMTGCW    27648       5
HAHTRDTKYWGHRKTHTKTM    62208       6
AMAARYAVMYHTMTKHHTKT    20736       5
WWTTWRMWARHATKHYTNMA    36864       6
YTWBMATWWWTTABARDMAB    20736       5
RVARDWKWWGDARAAARBVA    62208       6
YWWWAWWMWWWKNATYTRRT    65536       6
DAAMATHWBCMCMCWKGKVK    20736       5
CCYASTSTSWYBMARVSMMG    18432       5
GMCWKDGHCWGRDDTBTTTK    15552       5
RGARRDGBADAARGRNVWDA    62208       6
TMNNWTTTHMWHWSTWCAYA    36864       6
TTKBDKDNYTHHAAAAYTRK    62208       6
WTDKWDDKAHTDWRGWAAAW    62208       6
WKSKGWCTGHRDTYYTSHBC    41472       6
WTWWVWWYTKTKGKWHTKTR    36864       6
SNAKKGRSDGAHWTYYYTTV    55296       6
CTKNYYWYCWCMYCTVYCAS    12288       5
THTBTSTYTYTDNAYYTWYY    27648       6
TYTBHWWWMTGKAAWADRMH    41472       6
ADVHAATKAVNAADDRAAWG    23328       6
WHBATTKTKHRKATYYMWMT    27648       6
TTKADKDAWHTYWHTYWAWG    20736       6
AWMVARARSRARWAMAHRKW    36864       7
-----------------------------------------
Total Degeneracy: 1.29923e+06
```

Figure B.7: MIPS-PT output on masked input set with $\alpha = 4^8$.

```
MIPS-PT Output
Primer Size: 20
Primer Degeneracy Threshold: 262200
Beam Size: 100
Pair Fragment Size: 6

Number of Sequences: 188

Total # Primers: 28
Primers                 Degeneracy  # Covered
GKDMYCAKKRTADAWMTGCW    4608              4
WTDYCYTKHHYTTWDVWHWA    186624            6
VMADKAMHHATKRHAGWDNT    186624            6
AWARHHWWRDMRTTNHMART    165888            6
AWYTBHBHTWKMVYWWTYTA    124416            6
TWDRHANRMAWTATTNYMRM    73728             6
VNDRGVAAARDGMWCWBWKS    248832            7
TDWWTKNRHRAANWKAWTDW    221184            7
BVADMMYBYTHCCCAMMYSH    186624            6
BHHHWVTTBAKAWKMTSTSK    186624            6
AKHAWHTTKYMDWTBNMCWR    165888            6
NWWSWWWTWWDKAMWAAMAH    147456            7
ADBTWTWTNCAKYWADHDVW    186624            7
CYTBYHYWSSWYCCCWVHCM    82944             6
YWWDAWWMWHWKNATBTRRT    221184            7
ATWTYHTDWHWHYWRNAAHA    124416            7
DAAMAYHWBMMCMCWKGKVK    82944             6
WABVDTTTMMDHTYTWMWWW    124416            7
CMCAGWBKNWHKVARRCMHR    165888            7
ATDKKRGRMMYNKGGKRRVW    147456            7
WTWDVWDYTKTKGKHHTKTR    124416            7
DSNAKDGRSDGANWTYYYTK    221184            8
MYTKHANWKDAWDWYAWWAA    110592            7
HBSKGWCTGHRDTYYTSHBC    93312             7
TTKHNDDAWMTTWHYYHAWR    248832            8
RRSNRAKGRDDAARKRRAWS    147456            8
WHBATTKTKHRDAWHYHWMT    186624            8
AWMVWRARSRARDAMAHRKW    110592            8
-----------------------------------------
Total Degeneracy: 4.27738e+06
```

Figure B.8: MIPS-PT output on masked input set with $\alpha = 4^9$.

References

[1] R. Bisiani. Search, beam. In S. C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, pages 1467–1468. Wiley-Interscience, New York, NY, 2nd edition, 1992.

[2] F.S. Collins and V.A. McKusick. Implications of the human genome project for medical science. *JAMA*, 285:2447–2448, 2001.

[3] Cornish-Bowden. IUPAC-IUB symbols for nucleotide nomenclature. *Nucleic Acids Research*, 13:3021–3030, 1985.

[4] K. Doi and H. Imai. A greedy algorithm for minimizing the number of primers in multiple PCR experiments. *Genome Informatics*, pages 73–82, 1999.

[5] K. Doi and H. Imai. Complexity properties of the primer selection problem for PCR experiments. In *Proceedings of the 5th Japan-Korea Joint Workshop on Algorithms and Computation*, pages 152–159, 2000.

[6] S.B. Gabriel, S.F Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[7] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York, NY, 1979.

[8] Jr. G.D. Forney. The Viterbi algortihm. In *Proc. IEEE*, volume 61, pages 268–278, 1973.

[9] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, chapter 15, page 377. Press Syndicate of the University of Cambridge, 1997.

[10] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

[11] UCSC genome browser. Web site.

[12] P. Kwok. Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Human Genetics*, 2:235–58, 2001.

[13] S. Kwok, S.Y. Chang, J.J. Sninsky, and A. Wang. A guide to the design and use of mismatched and degenerate primers. *PCR Methods and Appl.*, 3:S39–S47, 1994.

[14] C. Linhart and R. Shamir. The degenerate primer design problem. *Bioinformatics*, 18, Suppl. 1:S172–S180, 2002.

[15] G. Marth, R. Yeh, M. Minton, R. Donaldson, Q. Li, S. Duan, R. Davenport, R. Miller, and P. Kwok. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics*, 27, 2001.

[16] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harb Symp Quant Biol*, volume 51(pt 1), pages 263–73, 1986.

[17] P. Nicodeme and J. Steyaert. Selecting optimal oligonucleotide primers for multiplex PCR. In *Proceedings of Fifth Conference on Intelligent Systems for Molecular Biology ISMB97*, pages 210–213, 1997.

[18] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence analysis. In *PNAS*, volume 85, pages 2444–2448, 1988.

[19] W.R. Pearson, G. Robins, D.E. Wrege, and T. Zhang. A new approach to primer selection problem in polymerase chain reaction experiments. In *Third International Conference on Intelligent Systems for Molecular Biology*, pages 285–291. AAAI Press, 1995.

[20] W.R. Pearson, G. Robins, D.E. Wrege, and T. Zhang. On the primer selection problem in polymerase chain reaction experiments. *Discrete Applied Mathematics*, 71, 1996.

[21] A.F.A Smit. Origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Devel.*, 6(6):743–749, 1996.

[22] A.F.A Smit. *Structure and Evolution of Mammalian Interspersed Repeats*. PhD thesis, USC, 1996.

[23] A.F.A. Smit and P. Green. RepeatMasker. http://ftp.genome.washington.edu/RM/RepeatMasker.html.

[24] R. Souvenir, J. Buhler, G. Stormo, and W. Zhang. Selecting degenerate multiplex PCR primers. In *Proceedings of the 3rd Workshop on Algorithms in Bionformatics*, 2003.

[25] The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[26] M.S. Waterman. *Introduction to Computational Biology*. Chapman & Hall, 1995.

[27] J. Wolfe. Web site, 2003. http://www.ucl.ac.uk/ ucbhjow/b241/techniques.html.

Vita

Richard M. Souvenir

Date of Birth September 3, 1979

Place of Birth Chicago, Illinois

Degrees B.S., Applied Science (Computer Science and Biology), August 2001, from Washington University.

Publications

Richard Souvenir, Jeremy Buhler, Gary Stormo, and Weixiong Zhang. “Selecting Degenerate Multiplex PCR Primers” in *Proceedings of the Third Workshop on Algorithms in Bioinformatics (WABI-03)*, Budapest, Hungary, September 2003.

Jeremy Buhler, Richard Souvenir, Weixiong Zhang, and Robi Mitra. “Design of High-Throughput Assay for Alternative Splicing Using Polymerase Colonies” in *Proceedings of the Pacific Symposium on Biocomputing (PSB-04)*, Hawaii, January 2004.

December, 2003