# ABSTRACT

Title of Dissertation: OPTIMIZING THE ACCURACY OF LIGHTWEIGHT METHODS FOR SHORT READ ALIGNMENT AND QUANTIFICATION

Mohsen Zakeri
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Rob Patro
Department of Physics

The analysis of the high throughput sequencing (HTS) data includes a number of involved computational steps, ranging from the assembly of reference sequences, mapping or alignment of the reads to existing or assembled sequences, estimating the abundance of sequenced molecules, performing differential or comparative analysis between samples, and even inferring dynamics of interest from snapshot data. Many methods have been developed for these different tasks that provide various trade-offs in terms of accuracy and speed, because accuracy and robustness typically come at the expense of sacrificing speed and vice versa. In this work, I focus on the problems of alignment and quantification of RNA-seq data, and review different aspects of the available methods for these problems. I explore finding a reasonable balance between these competing goals, and introduce methods that provide accurate results without sacrificing speed.

Alignment or mapping of sequencing reads to known reference sequences is a challenging computational step in the RNA-seq pipeline mainly because of the large size of sample data and reference sequences, and highly-repetitive sequence. Recent

quantification methods introduced the concept of lightweight alignment in order to accelerate the mapping step, and therefore, the whole quantification pipeline. I collaborated with my colleagues to explore some of the shortcomings of the lightweight alignment methods, and to address those with a new approach called the selective-alignment. Moreover, we introduce an aligner, Puffaligner, which benefits from both the indexing approach introduced by the Pufferfish index and also selective-alignment to producing accurate alignments in a short amount of time compared to other popular aligners.

To improve the speed of RNA-seq quantification given a collection of alignments, some tools group fragments (reads) into equivalence classes which are sets of fragments that are compatible with the same subset reference sequences. Summarizing the fragments into equivalence classes factorizes the likelihood function being optimized and increases the speed of the typical optimization algorithms deployed. I explore how this factorization affects the accuracy of abundance estimates, and propose a new factorization approach which demonstrates higher fidelity to the non-approximate model.

Finally, estimating the posterior distribution of the transcript expressions is a crucial step in finding robust and reliable estimates of transcript abundance in the presence of high levels of multi-mapping. To assess the accuracy of their point estimates, quantification tools generate inferential replicates using techniques such as Bootstrap sampling and Gibbs sampling. The utility of inferential replicates has been portrayed in different downstream RNA-seq applications, i.e., performing differential expression analysis. I explore how sampling from both observed and unobserved data points (reads) improves the accuracy of Bootstrap sampling. I demonstrate the utility of this approach in estimating allelic expression with RNA-seq reads, where the absence of unique mapping reads to reference transcripts is a major obstacle for calculating robust estimates.

# Optimizing the accuracy of lightweight methods for short read alignment and quantification

by

Mohsen Zakeri

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Rob Patro, Chair/Advisor
Professor Mihai Pop
Professor John Dickerson
Professor Erin Molloy
Professor Michael Cummings, Dean's representative

# Preface

If needed.

# Foreword

If needed.

# Dedication

If needed.

# Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Rajarshi Roy for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. He has always made himself available for help and advice and there has never been an occasion when I've knocked on his door and he hasn't given me time. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank my co-advisor, Dr. Parvez Guzdar. Without his extraordinary theoretical ideas and computational expertise, this thesis would have been a distant dream. Thanks are due to Professor Robert Gammon, Professor Edward Ott and Professor Thomas Antonsen for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript.

My colleagues at the nonlinear optics laboratory have enriched my graduate life in many ways and deserve a special mention. David DeShazer helped me start-off by rewriting the basic simulation code in a user-friendly format. Christian Silva provided help by setting up the GRENOUILLE apparatus and performing some of the simulations. My interaction with Rohit Tripathi, Ryan McAllister, Vasily Dronov, Min-Young Kim, Elizabeth Rogers, William Ray, Jordi Garcia Ojalvo, Riccardo Meucci, Atsushi Uchida,

and Fabian Rogister has been very fruitful. I'd also like to thank Wing-Shun Lam and Benjamin Zeff for providing the LaTex style files for writing this thesis.

I would also like to acknowledge help and support from some of the staff members. Donald Martin's technical help is highly appreciated, as is the computer hardware support from Edward Condon, LaTex and software help from Dorothea Brosius and purchasing help from Nancy Boone.

I owe my deepest thanks to my family - my mother and father who have always stood by me and guided me through my career, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them. I would also like to thank Dr. Mohan Advani, Dr. Vasudeo Paralikar and Dr. Vinod Chaugule who are like family members to me.

My housemates at my place of residence have been a crucial factor in my finishing smoothly. I'd like to express my gratitude to Sivasankar Pandeti, Jayakumar Patil, Amit Trehan and Punyaslok Purakayastha for their friendship and support.

I would like to acknowledge financial support from the Office of Naval Research (ONR), Physics, for all the projects discussed herein.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all and thank God!

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| **ARD** | **A**bsolute **R**elative **D**ifference |
| **BWT** | **B**urrows-**W**heeler **T**ransform |
| **DP** | **D**ynamic-**P**rogram |
| **EM** | **E**xpectation **M**aximization |
| **HTS** | **H**igh **T**hroughput **S**equencing |
| **Indel** | **In**sertion (and) **D**eletion |
| **MARD** | **M**ean **A**bsolute **R**elative **D**ifference |
| **MEM** | **M**aximal **E**xact **M**atch |
| **NGS** | **N**ext **G**eneration **S**equencing |
| **SA** | **S**uffix **A**rray |
| **SGS** | **S**econd **G**eneration **S**equencing |
| **SMEM** | **S**uper **M**aximal **E**xact **M**atch |
| **uni-MEM** | **U**nique **M**aximal **E**xact **M**atch |
| **VBEM** | **V**ariational **B**ayesian - **E**xpectation **M**aximization |

# Chapter 1: Introduction

Out of four major biological macromolecules (proteins, carbohydrates, lipids and nucleic acids), nucleic acids carry the most significant information about the identity of each organism. Even within each organism, the different content of nucleic acids in different organs and cells, defines their main functions and characteristics, also known as phenotypes. There are four types of nucleic acids (**A**denine, **C**ytosine, **T**hymine(**U**racil), **G**uanine) which are the main components of the **D**eoxyribo**N**ucleic **A**cids (DNA) and **R**ibo**N**ucleic **A**cid (RNA) in living organisms (Archaea, Bacteria, and Eukarya). Each DNA or RNA molecule is formed by a sequence of the nucleic acids. While the DNA content of different organisms are distinct, the DNA molecules across all cells of each individual are almost identical. Even during the cell division, all the DNA molecules are duplicated and preserved in each new cell's nucleus in the form of chromosomes. On the other hand, there exists different types of RNA molecules in different cells of each organism, leading to their different functions. RNA molecules are created from specific regions of chromosomes, called genes, in a process called transcription. A gene is called expressed in a specific cell if it is transcribed to RNA molecules. Different genes being expressed in different cell types leads to their vastly various functions. The set of all the genes, and the set of all the RNA molecules present in a cell are called the genome and the transcriptome respectively.

The transcription of a gene is started by a specific protein called the RNA polymerase. This protein copies the sequence of nucleic acids from a gene into a new RNA-sequence, called the pre-mRNA. There are two main types of subsequences present in each pre-mRNA, introns (intragenic regions) and exons (expressed regions).The pre-mRNA molecules turn into the mRNA molecules after the intronic regions are spliced out. Alternative splicing of the set of introns and exons generates various mRNA molecules from a single gene. The set of all mRNA molecules generated from a single gene are called the isoforms or transcripts of the gene. fig:altsplice shows how two different isoforms are generated from a single gene through alternative splicing.

Many technologies have been proposed for gathering information about the transcriptomic contents of an organism. RNA sequencing (RNA-seq) is a powerful sequencing technique, and has become very popular since its introduction mortazavi. In the RNA-seq protocols, RNA sequences are first fragmented into smaller pieces, then these fragments are amplified through the PCR process, and finally the set of amplified fragments are sequenced and read sequences are generated. If both ends (5' and 3') of the fragments are sequenced paired end reads will be generated, while single end reads are sequenced only from a single end of each fragment. Transcriptome assembly, detecting novel isoforms, and measuring the expression level of any isoform in a sample, are some of the main important applications of RNA-seq data. Mapping or alignment of RNA-seq reads to the set of known references is one of the early computational steps in many of these applications. Throughout this section, some famous computational approaches proposed for this critical step are introduced.

The number of reads generated from each isoform of a gene depends on the gene's level of expression in the cell. Each gene consists of encoding segments called exons. Each transcript or isoform of a gene consists of a set of particular exons. To illustrate the alternative splicing, consider the example in fig:altsplice which shows two different isoforms of $gene_a$ which is a gene with three exons. Different splicing events can lead to new combinations of exons in each isoform, e.g., $t_1$ and $t_2$ are two isoforms generated from $gene_a$, each containing specific subset of $gene_a$'s exons. The number of reads mapping to specific exons and exon junctions indicate the expression level of the isoforms containing those exons. According to sequence similarities in different genes or exons shared between different isoforms of a gene, a read may map to multiple isoforms in the transcriptome, this ambiguity makes abundance estimation challenging. An example of reads generated from different isoforms of a gene in RNA-seq experiments is displayed in fig:altsplice. In this example, green and blue reads suggest expressions of both $t_1$ and $t_2$, while red reads are only compatible with $t_1$.

Reads spanning two different exons can be evidence for the splicing events, e.g., the green-blue read in fig:altsplice spans the red and green exons which suggest the existence of an isoform containing these two exons next to each other, i.e., $t_2$. In this example, $t_1$ and $t_2$ are the known isoforms (also called transcripts) of the gene $gene_a$. In an RNA-seq experiment reads might be generated which do not map to any known isoforms of genes, e.g., the gray reads mapping to the intronic gray region of $gene_a$ in fig:altsplice or green-gray read which maps to the exon-intron junction. Presence of such reads could be evidence for discovering new exons or isoforms of $gene_a$, or preserved intronic regions in the isoforms (intron retention). Finding evidence for novel isoforms or intron

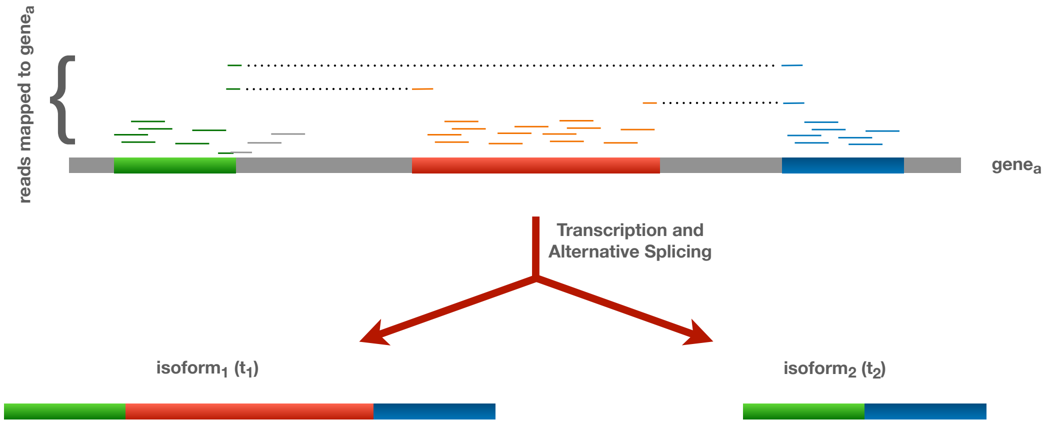

Figure 1.1: An example of splicing events in the gene $gene_a$, resulting in two transcripts $t_1$ and $t_2$. According to sequence similarities the RNA-seq reads might be mapped to single or multiple isoforms.

retention events is an important advantage of RNA-seq assay over alternative sequencing techniques.

In the following sections of this chapter I review the existing tools which are widely used for alignment (computing the compatibility of the reads to the reference sequences) and quantification (estimating the expression ratio of the isoforms) of RNA-seq samples. Both of these steps greatly affect the accuracy of the downstream analysis in RNA-seq pipelines srivastava2019alignment. In chapter 2, I present a fast mapping technique called selective-alignmentwhich is built alongside *quasi-mapping*to achieve a higher sensitivity which results in more accurate abundance estimates. Chapter 2 also introduces a new tool for computing alignment of DNA-seq and RNA-seq reads to known reference sequences. In chapter 3, an improved factorization of the quantification likelihood function is presented which maintains a high fidelity to the underlying data and will result in higher confidence for more fine grained analysis of transcripts.

## 1.1 Mapping RNA-seq reads to a known transcriptome

Alignment is a crucial and expensive computations step in an RNA-seq analysis pipeline. The main goal of this step is to find, for each read, the region on the reference genome (or transcriptome) from which it is originally fragmented. The length of the short RNA-seq reads is usually between 100 and 200 bases, while the latest human genome sequence is about three billion bases. The goal, for each read, is to find a substring in the reference sequences which best matches the read. Often it is not possible to find an exact match for a read because of the variations present in the samples with respect to the reference sequences. The variations are divided into two main types, variations introduced by technical errors or the biological variations existing in each new sample. Therefore, the alignment procedure most of the time results in an inexact match for each read rather than a region which exactly matches the read. In order to find the most compatible reginos in the reference sequences to each read, we should define the compatibility of two sequences. The compatibility is measured by the number of edits required to convert one sequence to the other one. Different types of edits are considered which are substitution, insertion and deletion. The penalties assigned to each type of edit might be different and can be usually configured in most of the aligners. The sum of all the penalties is considered to be the edit distance between two sequences. Therefore, we can define the alignment problem as finding the region on the reference with the minimum edit distance to each read. This can be achieved by classical algorithms such as Smith Waterman smith1981identification in $O(n * m)$ time and $O(n * m)$ space, where n and m are the length of the reference sequence and the read sequence respectively. This solution becomes super expensive when the length of reference sequences and the number read sequences are very large which is the common case in the RNA and DNA sequences. Therefore, a number of additional solutions proposed to accelerate this procedure while maintaining the accuracy of this approach. In the following sections, some of the main common approaches will be discussed briefly.

### 1.1.1 The main approaches for computing read alignments

One of the most common approaches for accelerating the alignment problem is “seed and extend”. The seeds are supposed to reduce the search space for the alignment problem into smaller regions that are most likely to include a reasonable alignment for the queried read rather than comparing each read sequence to all the reference sequences. Seeds are often shorter than the reads and represent an exact match from a substring in the read to some region in reference. The seeds are later extended into full alignments for the read by computing the full alignment of the reads to the regions identified by each read.

Finding the seeds requires the pre-processing of the reference sequences which is called the indexing step. Different indexing strategies are used in different aligners which have various space and time requirements. There are two main types of indices, full-text indices and hash-based indices. The full-text indices are often smaller in size, and a sequence of any different size can be queried in them, while the hash based indices take more space and only accept queries with a fixed size. The main benefit of the hash based indices are their speed compared to the full text indices. Full text indices are used in popular RNA-seq aligners such as Bowtie bowtie, Bowtie2 bowtie2, BWA bwamem and STAR Dobin2013Star. It is worth noting that STAR's index employs a hybrid approach of both full text and hash based indices which results in being faster at the expense of larger memory requirements. Other aligners such as deBGA debga, Minimap2 minimap2, are based on hash based indices and often use exact k-mer searches as the first step of the alignment step to find the seeds. K-mers of each read are all the substrings of length k in the read sequence. Querying a k-mer into the index results in finding all the locations on the reference sequences where the k-mer exists.

FM-index and Suffix Array are two closely related full text indices. The suffix Array (SA) of a string S with the length n is defined by the sorted order of all suffixes of string S concatenated with a terminal character which is lexicographically smaller than all other characters in the alphabet. Adding the terminal character ($) ensures that no suffix is the prefix of any other suffix in S. In practice, the suffix array stores only the indices corresponding to each suffix in an array. If pattern p exist in the string S, then, it will be a prefix of some suffix in S. Lexicographically sorting the suffixes in SA, provides this property that all suffixes which include some pattern p as their prefix, will appear in consecutive rows in the SA. Therefore, to query a pattern in S, it suffices to find an interval [a,b) in the SA which are all the suffixes that include p as their prefix. Each pattern can be searched in the SA by a binary search, which takes O(log(—S—)x—p—). The search process can be enhanced by keeping some extra information like the longest common prefix lengths (LCP) for some pair of suffixes, as a result the query time will be O(log—S—+—p—) instead. STAR is one of the most popular aligners which use the Suffix Array to index the reference sequences. [cite] STAR uses some hash tables for accelerating the search process as well which comes at the cost of increasing the index size.

FM-Index is another full text index which consists of some auxiliary data structures alongside the Burrows Wheeler Transform (BWT) of the reference string S. BWT(S) is closely related to the Suffix Array. To enable efficient search for every pattern using BWT, the LF mapping property in the BWT is utilized with the help of storing the occurrence information of every character in the BWT. Using the succinct data structures, this can be stored in O(—S—) space. One other useful characteristic of the BWT is that it tends to put repetitions of each character next two each other, this doesn't mean that all repetitions are put next to each other, but it is common to find longer substrings of A or any other character in the BWT of a sequence compared to the original sequence. This property of the BWT makes it more compressible compared to the original sequence which results in smaller index size. Bowtie, Bowtie2, and BWA are some of the popular aligners which index the reference sequences with a FM-Index.

The other main type of indices used for finding the alignment of a query in a large set of reference sequences are hash based indices. Hash based indices use substrings of a fixed size from the reference sequence to search each new query. The substrings of length k from the reference sequences are called k-mers. K-merss constitute the keys in the hash based indices. Hash based indices store the location where each k-mer occurs in the reference sequences. During the query time, we are able to extract all the k-mers from each read and query those e in the set of keys of the hash based index. The index will then retrieve all the positions each k-mer occurs in the reference which will later play the role of the seeds for the seed and extend procedure. It is important to note that queries of

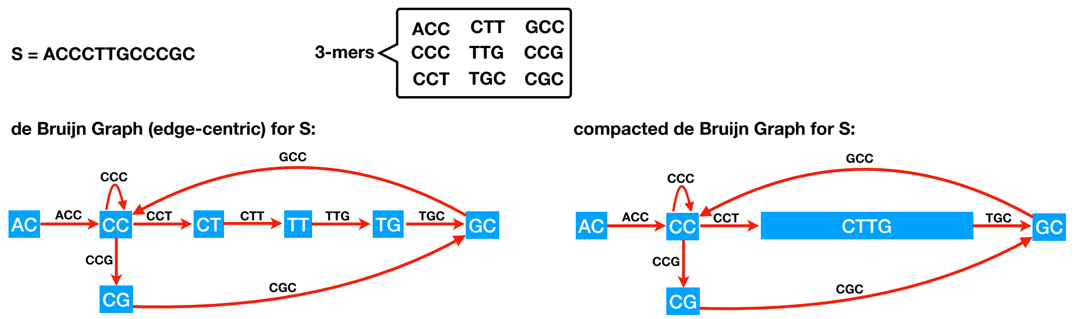

Figure 1.2: The edge-centric de Bruijn graphand the compacted De Bruijn graphfor the sequence S. There are 9 3-mers in this sequence which correspond to the edges in the de Bruijn graph. 2-mers form the nodes of the de Bruijn graph. 3 nodes (CT, TT and TG) are combined together in the compacted De Bruijn graphsince they are on a non-branching path of the de Bruijn graph.

length smaller than k cannot be made into these hash based indices, so, one drawback of these types of indices is in order t o find a seed for the query in the reference, there needs to be at least one substring of length k in the read which match the reference sequences without any edit distance. Therefore, the length of k should be carefully selected based on the error rate of the sequencing technology, so that with high probability at least one match from each read is found on the reference, if the read is actually originating from a position on the reference sequences.

Another approach for indexing all the $k$-mers of a set reference sequence is to build the de Bruijn graph. Each $k$-mer forms an edge in the de Bruijn graph between the prefix and suffix k-1 mers which it consists of. In the fig:dbg, a de Bruijn graphis shown which is built from the sequence S = “ACCCTTGCCCGC” and the k equal to 3. One main property in the de Bruijn graph is that every k-mer appears exactly once in the graph. This property helps to reduce the redundancy of repeats which usually exist in the DNA and RNA sequences. The compacted de Bruijn graph is built after compacting the non branching paths from the original de Bruijn graph, as shown in the fig:dbg. The length of the nodes in the compacted de Bruijn graph might be larger than k-1, but still each k-mer appears at most once in the graph (either as an edge or as a substring in one node). If the de Bruijn graph is built from multiple sequences (e.g., the set of human transcripts, or a collection of microbial genomes), one is interested to know in which reference sequences each k-mer appears. A k-mer might appear in multiple reference sequences due to shared exons in transcripts, or sequence similarities in different strains of some species.

A colored de Bruijn graphstores this data for each $k$-mer (edge) of the graph as the color information. For example, in the case of the transcriptome, each color represents the set of transcripts in which the $k$-mer corresponding to that edge exists. $K$-mers appearing in the same set of transcripts (e.g., due to shared exons) will have the same colors. A contig in the compacted De Bruijn graphis a non branching path in the graph, and if the edges are colored, not all the k-mers in a contig might have the same color. If all the $k$-mers that are part of a non branching path also have the same color set assigned, those $k$-mers can be combined together and form a unitig in the colored compacted de Bruijn graph.

Pufferfish pufferfish is a space and time efficient index built on top of the compacted colored de Bruijn graph. So, it can be used to perform efficient k-mer queries from each read sequence to large transcriptomic or genomic references. In the second chapter of this manuscript, we introduce Puffaligner which uses the Pufferfish index to query k-mers from the short RNA-seq reads to compute read alignments through the seed and extend strategy.

### 1.1.2 Alignment Free Approaches for Mappings reads

Exploiting the methods of abundance estimation for RNA seq reads revealed the fact that although alignment is very useful for finding the candidate transcripts to which reads map, the full alignment information (The exact details of all the gaps and mismatches) are unnecessary for performing quantification. In fact the position, orientation and length of the matching fragment on each transcript is adequate for achieving accurate estimates. Therefore, lightweight methods were developed to avoid performing full alignments upstream of the quantification. These methods typically build an index over the reference sequence (similar to the FM index in alignment tools). Then, in the quantification step, the reads are mapped to the reference on the fly, rapidly, using the built-in index. Therefore, the memory peak footprint of such methods is bounded by the reference size and complexity and scales well as the number of reads increases. Note that for performing quantification over multiple samples to the same reference, the index needs to be built only once.

The first lightweight algorithm for mapping reads to reference transcripts was introduced in *Sailfish*Patro2014Sailfish. *Sailfish*is an alignment-free quantification tool which builds an index over all subsequences of size k from the reference, called $k$-mers . The *Sailfish*index consists of a perfect hash function mapping each $k$-mer in the reference to a unique integer, an array indicating the counts of each $k$-mer , an index mapping each $k$-mer to the set of transcripts in which it appears, and another index mapping each transcript to the multiset of $k$-mers it contains. In the quantification step, *Sailfish*explores all the $k$-mers the read contains and keeps the count for the ones appearing in the reference. So instead of mapping the whole reads, *Sailfish*only maps the $k$-mers , and uses their count for each transcript as evidence for the relative expression of the transcripts. Although *Sailfish*'s approach is 30 times faster than fastest quantification tools which perform alignment upstream of quantification, it suffers from increased ambiguity of a large rate of multi mapping $k$-mers , which sometimes reduces the accuracy of abundance estimation. The smaller k size causes a higher multi mapping rate, while the larger size of the k results in less robustness to sequencing errors because each $k$-mer is mapped with no error using the perfect hash function.

The idea of mapping $k$-mers instead of the whole reads to reference transcriptome introduces a huge improvement in the speed of quantification tools. However, it is sub-optimal to only consider occurrence of subsequences of size k as evidence of expression while the observed data is of larger length. Hence, the developers of *Sailfish*introduced a new idea for mapping the whole reads using the perfect hash function of $k$-mers by benefiting from the suffix array data structure. The new mapping approach is called *Quasi-mapping*Srivastava2016rapmap and is utilized in the newer version of *Sailfish*software and also in the new quantification tool called SalmonPatro2017Salmon.

The suffix array of a sequence is a sorted array of all of the sequence's suffixes. Therefore, all suffixes starting with the same prefix are located in adjacent positions of the suffix array. Note that only the position of the occurrence of suffixes in $T$ are stored in the suffix array. The number of elements of the array is equal to the length of the sequence and there is a one-to-one mapping from each row of the suffix array to a character in the BWT of the sequence. We can also introduce suffix array intervals similar to intervals of the Burrows Wheeler transform. In the *quasi-mapping*index the suffix array $SA(T)$ is built from reference transcriptome $T$. Therefore, each row in the $SA$ starts at a unique transcript of the transcriptome. There is a hash function $I(k_i) = [b, c)$ from each $k$-mer $k_i$ in $T$ to a suffix array interval from row $b$ until row $c$; all rows that contain $k_i$ as a prefix. For mapping each read, the $k$-mers , $k_i$ of the read (existing in the hash table) are hashed to find SA intervals. It is often possible to extend a match between the query and a subset of rows of the SA interval. The *quasi-mapping*algorithm attempts to find the longest subsequence of the query starting with the $k_i$ as a prefix in the interval (also called maximum mappable prefix of $k_i$ ($MMP_i$) Dobin2013Star) with a binary search, as the SA is sorted lexicographically. *Quasi-mapping*retrieves a set of transcripts for each $k$-mer , the transcripts appearing in all sets are reported as mapping candidates for the read. The reverse complement of the read is also mapped and the sequence (either forward or reverse complement) with the higher number of matching $k$-mers determines the mapping orientation. For paired end reads, the other end of the read is also quasi-mapped to the reference. Then, transcripts appearing as candidates for both ends of the reads are reported as the mapping possibilities for the paired end reads.

It has been demonstrated that *quasi-mapping*finds very high quality mappings which result in highly-accurate abundance estimations. However, there are different aspects of the algorithm which can be modified in order to retrieve mappings with higher specificity and sensitivity. In fact, this idea is presented in chapter 2 as selective-alignment. *Quasi-mapping*extends the $k$-mer matches by the MMP length in order to find legitimate matches for queries. However, if an extension is not possible and no other $k$-mer match exists in the read, *quasi-mapping*may report all transcripts of the interval as the mapping candidates for the read. These low quality matches introduce a number of spurious mappings. A filtering process shall be introduced in order to filter the spurious hits in this case. Other than suffering from spurious mappings *quasi-mapping*could also miss true mappings of the read in rare cases where errors are positioned adversarially on the read. An obvious case of losing the true mapping is if a read contains no subsequence of size $k$ from the true transcript. In another case, the true mapping of the read might be lost from the SA interval by performing MMP extension, if a longer exact match of the read to the interval masks the match to the row with true location. The hits in the reverse complement of the read are only considered if there are less number of hits in forward strand compared to reverse complement. Therefore, spurious mappings in the forward strand might mask the true hit in the reverse complement. For some reads, multiple positions might be found on the same transcript where the read maps. *Quasi-mapping*greedily considers the left most one as the true mapping while that might not be the best possible matching of the read to that transcript. To address these challenges in *quasi-mapping*, selective-alignmentis introduced as a new lightweight approach to achieve both higher sensitivity and specificity than *quasi-mapping*.

A similar approach to *quasi-mapping*is employed in *kallisto*Bray2016Kallisto called pseudoalignment. *kallisto*index consists of a colored deBruijn graph from reference where nodes are $k$-mers and each node receives the colors of transcripts in which it appears. The contigs in the graph are formed from the linear stretches of the nodes ($k$-mers ) with identical sets of colors. Kallistoalso maintains a hash table mapping each $k$-mer to the contig it is contained in and the $k$-mer 's position in the contig. Using this index, the reads' $k$-mers are mapped to contigs. Since all the $k$-mers appearing in the same contig receive the same set of colors, and therefore the same transcripts, the rest of the $k$-mers in the contig can be skipped for mapping, similar to the idea of NIP skipping in *quasi-*

*mapping*.

## 1.2 Abundance estimation of the transcriptome

In this section, we formalize the problem of abundance estimation with RNA-seq reads according to the model laid out by Li2010RSEM. There are $M$ transcript types in transcriptome $T$, $t_1, t_2, ..., t_M$. In a given sample there are $c_i$ copies of transcript type $t_i$, which are not observed directly.

### 1.2.1 The generative model of a sequencing experiment

The generative model of RNA-seq experiments states that the number of fragments sequenced from $i^{th}$ transcript type is proportional to the total number of sequenceable nucleotides belonging to transcripts of type $t_i$. If the length of the $i^{th}$ transcript is given by $l_i$, assuming all the reads have the size $l_r$, we can define effective length, $\tilde{l_i} = l_i - l_r + 1$ which is all possible start positions on transcript $t_i$ for sequencing a read of size $l_r$. The portion of sequenceable nucleotides of transcript type $t_i$ is $\eta_i = \frac{c_i l_i}{\sum_j c_j l_j}$ and $\alpha_i \propto \eta_i$, where $\alpha_i$ is the number of fragments drawn from transcripts of type $t_i$.

If $\mathcal{F}$ with $|\mathcal{F}| = N$, is the set of sequenced fragments, assuming independence for drawing each fragment, the likelihood of the underlying transcript abundances, $\theta$, can be written as:

$$\mathcal{L}\left(\theta; \mathcal{F}\right) = \prod_{f_i \in \mathcal{F}} \sum_{j=1}^{M} \Pr\left(t_i \mid \theta\right) \Pr\left(f_i \mid t_j\right). \tag{1.1}$$

The conditional probability of drawing a particular fragment $f_i$, given transcript

$t_j$, $\Pr(f_i|t_j)$, is particularly critical for reaching accurate estimates and is derived from mapping information. This term encodes, given parameters of the model and experiment, how likely it is to observe a specific fragment $f_i$ arise from transcript $t_j$. Many terms can be included in such a conditional probability, some common terms include:

$$\Pr(d_i \mid f_i, t_j) = \frac{\Pr_D(d_i)}{\sum_{k=1}^{\tilde{l}_j} \Pr_D(k)}, \tag{1.2}$$

the probability of observing a mapping of implied length $d_i$ for $f_i$ given that it derives from $t_j$, where $\Pr_D(k)$ is the probability of observing a fragment of length $k$ under the empirical fragment length distribution $D$;

$$\Pr(p_i \mid d_i, f_i, t_j) = \frac{1}{l_j - d_i + 1}, \tag{1.3}$$

the probability of a observing a mapping starting at position $p_i$ for fragment $f_i$ given that it has implied length $d_i$ and is derived from $t_j$;

$\Pr(o_i \mid f_i, t_j) = \{ \ cases 0.5 if unstranded \{ 1 \ .0 if compatible orientation \epsilon if incompatible orientation$

, (1.4)

the probability of observing a mapping with a specific orientation $o_j$ (i.e., forward or antisense) with respect to the underlying transcript for $f_j$, given $t_i$, $\epsilon$ (a user-defined constant), and knowledge of the underlying protocol, and

$$\Pr\left(a_i \mid f_i, o_i, d_i, p_i, t_j\right), \tag{1.5}$$

the probability of observing the particular alignment (e.g., `CIGAR` string) $a_i$ for $f_i$ given it is sampled from transcript $t_j$, has orientation $o_i$, implied length $d_i$ and starts at position $p_i$—such a probability is calculated from a model of alignments, like those presented in Li2010RSEM,Roberts2013Express,Patro2017Salmon.

In fact, one can conceive of many such general models of "fragment-transcript agreement" Patro2017Salmon. However, here we consider that $\Pr\left(f_j \mid t_i\right)$ is simply the product of the conditional probabilities defined in eqn:pr$_{l}en1, eqn : pr_{s}tart1, eqn : orient1, eqn : align1, appropriatelynormalized.$

### 1.2.2 Expectation-Maximization for optimizing the model parameters

Exact inference from the likelihood function is intractable for the large scale of RNA seq data. Local optimization methods, like expectation maximization (EM), are often applied to fit the best parameters in the model. The parameters of the model indicate the rate of expression for each transcript in the underlying samples. The EM approach is employed by both alignment based tools such as *RSEM*Li2010RSEM, *mmseq* Turro2011Haplotype, *IsoEM* Nicolae2011Estimation and also non-alignment based tools like *Sailfish*Patro2014Sailfish, SalmonPatro2017Salmon and *kallisto*Bray2016Kallisto.

[H] transcriptome $T$, fragment set $\mathcal{F}$,conditional properties $\Pr\left(f_i | t_j\right)$ for fragment transcript pairs $\theta$, relative abundance of transcripts uniform initialization not converged $t_i \in T$ $\alpha_i = 0$, $\theta_i = \frac{1}{|T|}$ E-step:

```
    $f_i \in \mathcal{F}$ $sum = \sum_{t_k \in T} \theta_k \times \Pr\left(f_i|t_k\right)$
$t_j \in T$ $\alpha_j + = \frac{\theta_j \times \Pr\left(f_i|t_j\right)}{sum}$ M-step:
$sum = \sum_{t_k \in T} \frac{\alpha_k}{\tilde{l_k}}$
$t_i \in T$ $\theta_i = \frac{\alpha_i/\tilde{l_i}}{sum}$ Overview of the EM algorithm for optimizing the generative model
```

The overview of the EM algorithm for optimizing eqn:likelihood$_f m1 isdisplayedinalgorithm$**??**$.I$
$step, theexpectednumberof f ragmentssequencedfromeachtranscripttypeinthesampleiscalculated$
$step.Thisiterativeprocessisrepeateduntiltheconvergenceon \theta$ values is reached.

If a transcript $t_j$ is not present in the set of transcripts to which fragment $f_i$ is mapped, the value of $\Pr\left(f_i|t_j\right)$ is equal to zero. The EM updates can benefit from the sparsity of $\Pr\left(f_i|t_j\right)$ matrix by only performing updates in the E-step for the set of transcripts that $f_i$ maps to instead of the whole set of transcripts. Hence, if $\Omega(f_i)$ is the set of compatible transcripts with read $f_i$, in algorithm **??**, the line 8 shall be modified to : $sum = \sum_{t_k \in \Omega(f_i)} \theta_k \times \Pr\left(f_i|t_k\right)$ and the loop iteration in line 9 to : $t_j \in \Omega(f_i)$.

### 1.2.3 Factorizations of the likelihood function

Each iteration of the EM algorithm updates the $\alpha$ values for each fragment independently. Although each update cost has collapsed considerably by benefiting from the sparsity of mapping matrix, the number of EM updates still scales with the number of fragments (and alignments). Sequence similarities in reads shall be utilized for factorizing the likelihood function, which results in bounding the number of updates as the number of fragments grows.

If a set of fragments, $F'$, exactly map to the same set of transcripts, $T'$, with the same conditional probabilities (meaning that for each pair $f_i, f_j \in F'$, $\Omega(f_i) = \Omega(f_j) = T'$ and the equation $\Pr(f_i|t_k) = \Pr(f_j|t_k)$ holds for all $t_k \in T'$), then all fragments in $F'$ are exactly equivalent, and they result in the same update rule in the EM iterations. Hence, they can be grouped together to apply the update once for all such fragments. This factorization, which is introduced by *IsoEM*isoem, maintains full fidelity to information regarding fragment mappings to transcripts because all fragments in a group are identical.

A more popular approach for factorizing the likelihood is employed by *mmseq*Turro2011Haplotype and also later in alignment-free tools like *Sailfish*Patro2014Sailfish, SalmonPatro2017Salmon and *kallisto*Bray2016Kallisto. *mmseq*introduced a notion of fragment equivalence classes, which treats as equivalent any fragments that map to the same set of transcripts. Unlike *IsoEM*, the equivalence notion does not depend on the values of conditional probabilities. According to this definition, every set of fragments like $\mathcal{F}^q$ such that for all $f_i, f_j \in \mathcal{F}^q$, $\Omega(f_i) = \Omega(f_j) = \Omega(\mathcal{F}^q)$, form an equivalence class with the label $\Omega(\mathcal{F}^q)$. Define $N^q = |\mathcal{F}^q|$ to be the number of fragments in class $\mathcal{F}^q$. Then, the likelihood function based on these equivalence classes, can be approximated as:

$$\mathcal{L}\left(\theta; \mathcal{F}\right) \approx \prod_{[q] \in \mathcal{C}} \left( \sum_{\langle i, t_i \rangle \in \Omega([q])} \Pr\left(t_i \mid \theta\right) \cdot \Pr\left(f \mid [q], t_i\right) \right)^{N^q}, \tag{1.6}$$

where $\mathcal{C}$ is the set of all equivalence classes, and $\Pr(f \mid [q], t_i)$ is the probability of generating a fragment $f$ given that it comes from equivalence class $[q]$ and transcript $t_i$. The key to the efficiency of likelihood evaluation (or optimization) under this factorization, is that the probability $\Pr(f \mid [q], t_i)$ is assumed to be identical for each of the $N^q$ fragments in each equivalence class $[q]$—hence, we do not subscript $f$ in eqn:likelihood$_{f}act.Thisallowsonetoreplac$

in full model (eqn:likelihood$_{f}m1)withaproductoverallequivalenceclassesineqn : likelihood_{f}act.The$

$\Pr(f_j \mid t_i)$ $thatisarbitrarilydifferentfrom$ $\Pr(f \mid [q], t_i)$ $.Moreover, themostcommonapproxima$

$levelinformation(e.g., itissettoonedividedbytheeffectivelengthoft_i).$

After applying any factorization which groups a set of fragments together in equivalence classes $\mathcal{F}^q$, the fragments in the EM iteration can be substituted with equivalence classes (groups) and each update would increase the $\alpha$ values based on the number of fragments in each equivalence class. The modified version of the E step in algorithm **??** is displayed in algorithm **??**.

[H] E-step:

```
$\mathcal{F}^q \in \mathcal{C}$ $sum = \sum_{t_k \in \Omega(\mathcal{F}^q)} \theta_k \times \Pr(f|\mathcal{F}^q, t_k)$
$t_j \in \Omega(\mathcal{F}^q)$ $\alpha_j + = \frac{\theta_k \times \Pr(f|\mathcal{F}^q, t_k)}{sum}$   Modified E step after employing factorization
```

### 1.2.4 Online EM for optimizing the likelihood function

The conditional probability values are stored in memory during the EM iterations in order to avoid expensive I/O operations and re-computation in each EM round. If no factorization is used, for each existing mapping pair $f_i$ and $t_j$, a value is stored. This makes the memory requirement scale with the number of mappings. *eXpress*Roberts2013Express attempts to bound the memory requirement by benefiting from an online-EM algorithm rather than a batch-EM. An online-EM consists of a single iteration over all fragments in the sample, updating $\alpha$ values once for each fragment. Fragments are not stored in memory after being observed, which makes *eXpress*'s memory requirement independent of the number of fragments in the sample. The large number of fragments in RNA-seq samples lets *eXpress*often achieve high quality abundance estimates with a single run over the data. *eXpress*requires the output of an alignment tool for mapping reads to transcripts to run the online phase. Again, here, the mapping information shall limit the number of updates performed for each fragment.

*eXpress*applies a modified version of online updates which prevents the algorithm from performing updates for each transcript in each iteration. The online update rules for each fragments are:

$$\alpha^{i+1} = \alpha^i + m_i \tilde{\tau}^i, \tag{1.7}$$

where:

$$\tilde{\tau}_t^i = \Pr\left(T = t | F = f_i\right), \tag{1.8}$$

and

$$m_{i+1} = m_i \times \frac{\gamma_{i+1}}{1 - \gamma_i} \times \frac{1}{\gamma_i}, \tag{1.9}$$

$\alpha^i$ is the optimized value after observing the first i fragments. Bayes' rule can be applied to obtain the probability in eqn:online2 from the conditional probabilities. The value $m_i$ is called forgetting mass and depends on the forgetting factor $\gamma_i$. The $\gamma$ values are set as $\gamma_i = \frac{1}{i^c}$ where $0.5 < c < 1.0$. After observing all $N$, fragments the relative counts of fragments from each transcript type can be obtained from the vector $\alpha^N$.

### 1.2.5 Dual phase optimization

The inference algorithm in SalmonPatro2017Salmon consists of two phases. First, Salmonruns an online EM optimization to obtain high quality primary estimates of abundances. In this phase, Salmonis able to achieve a good estimation of fragment length distribution by examining many fragments as they are streamed in the online EM. Therefore Salmoncan derive good estimates of conditional probabilities using the fragment length distribution and other information provided with mappings. The equivalence classes over sets of fragments are also created in the online phase. Salmonintroduces the notion of rich equivalence classes by assigning a single scalar to each transcript in an equivalence class, by averaging the conditional probabilities of all fragments in the class to the transcript. This value is equal to $\frac{1}{|\Omega(F^q)|}$ in non-rich equivalence classes.

Salmonuses the estimates obtained in the online phase as a starting point for performing a batch EM algorithm in the second phase. This two-phase optimization allows Salmonto rich very high quality estimates compared to other existing quantification tools. The online phase of the Salmonenables deriving a new factorization of the likelihood function to be optimized in the batch EM phase, which does not discard any necessary information for accurate abundance estimation. The details of this factorization is discussed in chapter 3.

### 1.2.6 Metrics for evaluating quantification accuracy

The formula for calculating the metrics used for evaluating the abundance estimation results in the manuscript are as follows. The metrics are Mean Absolute Relative Difference

(MARD), Mean Absolute Error (MAE), and Mean Squared Log Error (MSLE).

$$MARD(y,\hat{y}) = \frac{1}{n_{refs}} \sum_{i=0}^{n_{refs}-1} \frac{|y_i - \hat{y}_i|}{y_i + \hat{y}_i} . MAE(y,\hat{y}) = \frac{1}{n_{refs}} \sum_{i=0}^{n_{refs}-1} |y_i - \hat{y}_i| . MSE(y,\hat{y}) = \frac{1}{n_{refs}} \sum_{i=0}^{n_{refs}-1} \quad (1.10)$$

All of these metrics compute the difference of the estimated abundances with the truth. In addition to these metrics, we also evaluate the correlation between the estimations and truth by computing the Spearman correlation. Spearman correlation is computed using the pandas library reback2020pandas in Python.

### 1.2.7 Inference of the Posterior Distribution of RNA-seq quantification

Estimating the inference uncertainty of the RNA-seq quantification is one of the crucial steps for many downstream analysis, e. g., finding the differentially expressed genes or transcripts, i. e., DE analysis. In fact, methods like Swish [**?**] directly use the inferential replicates created by RNA-seq quantification tools for finding the DE genes with a higher precision compared to other approaches.

There are two main approaches for sampling the posterior distribution for estimating the uncertainty of quantification estimates; Gibbs samplingand Bootstrap sampling. The Gibbs samplingis a MCMC procedure which walks through the space that the EM explores for finding the maximum likelihood estimationsof the $\mathcal{T}$expressions. At the end of each iteration of the Gibbs sampling, the $t_e$xpression vector could be identified as a new inferential replicate for estimating the posterior distribution. To decrease between replicate correlations, the sampling could take place after every fixed number of iterations which is called the thinning factor for the sampling. Running the Gibbs samplingprocedure for long enough could reach to the convergance of the posterior estimate, this will be only reached only after the Gibbs sampler has explored all the posterior samples. The number of iterations which is required for Gibbs samplingto converge usually depends on the properties of the sample, and as the number of $\mathcal{T}$in the sample increases the convergance usually takes longer.

Bootstrap samplingis the other main approach for estimating the posterior distribution of the abundance estimations in RNA-seq. The bootstrap procedure [**?**] is a widely-used and computationally straightforward procedure for calculating measures of accuracy of an estimator. It works by resampling (with replacement) from the observed data, and treating these as population samples. The procedure has been used in many contexts for non-parametric estimation. In RNA-seq, Computing the abundance estimation of all Bootstrapsample leads to a estimating a posterior distribution for the abundances.

RNA-seq quantification tools ( [**?**,**?**] have implemented the regular bootstrap sampling by resampling the equivalent classcounts. equivalent classesare a summerized representations of the reads, therefore, sampling the equivalent classcounts instead of each individual read improves the efficiency of the Bootstrap sampling. positional Bootstrap samplingis also another way of generating Bootstrapsamples by sampling the positions where the reads map to on each transcript [**?**]. Furthermore, the RNA-seq quantification tool, Salmon [**?**] also includes a Gibbs samplingprocedure for estimating the posterior distribution. Bitseq [**?**] applies a MCMC Gibbssampler to generate samples from the posterior probabilty distribution.

# Bibliography

[1] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 1.

[2] N. Bloembergen, *Nonlinear Optics* (Benjamin, Reading, MA, 1977).

[3] Y.R. Shen, *Principles of Nonlinear Optics* (Wiley, New York, 1984).

[4] P.N. Butcher and D.N. Cotter, *The Elements of Nonlinear Optics* (Cambridge University Press, Cambridge, UK, 1990).

[5] R.W. Boyd, *Nonlinear Optics* (Academic Press, San Diego, CA, 1992).

[6] A.C. Newell and J.V. Moloney *Nonlinear Optics (Advanced Topics in the Interdisciplinary Mathematical Sciences)* (Westview Press, Boulder, CO, April 1992).

[7] D. Marcuse,*Light Transmission Optics* (Van Nostrand Reinhold, New York, 1982), Chaps. 8 and 12.

[8] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 2.

[9] P. Diament, *Wave Transmission and Fiber Optics* (Macmillan, New York, 1990).

[10] V.E. Zakharov and A.Shabat, Sov. Phys. JETP **34**, 62 (1972)

[11] R.H. Hardin and F.D. Tappert, SIAM Rev. Chronicle **15**, 423 (1973).

[12] R.A.Fisher and W.K. Bischel, Appl. Phys. Lett. **23**, 661 (1973); J. Appl. Phys **46**, 4921 (1975).

[13] J.W. Cooley and J.W. Tukey, Math. Comput. **19**, 297 (1965).

[14] R. Trebino, D.J. Kane, ”Using phase retrieval to measure the intensity and phase of ultrashort pulses: frequency resolved optical gating,” J. Opt. Soc. Am. B **10**, 1101 (1993).

[15] D.J. Kane, R. Trebino, ”Characterization of Arbitrary Femtosecond Pulses Using Frequency-Resolved Optical Gating,” IEEE J. Quant. Elect. **29**, 571 (1993).

[16] D.J. Kane, R. Trebino, ”Single-shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating,” Opt. Lett. **10**, 1101 (1993).

[17] P. O’Shea, M. Kimmel, X. Gu, R. Trebino, ”Highly simplified device for ultrashort-pulse measurement,” Opt. Lett. **26**, 932 (2001).

[18] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 3.

[19] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 4.

[20] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 10.

[21] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 7.

[22] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 6.

[23] R.H. Stolen, E.P. Ippen, and A.R. Tynes, Appl. Phys. Lett. **20**, 62 (1972).

[24] E.P. Ippen and R.H. Stolen, Appl. Phys. Lett. **21**, 539 (1972).

[25] R.G. Smith, Appl. Opt. **11**, 2489 (1972).

[26] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 9.

[27] G.P. Agrawal, *Nonlinear Fiber Optics* (Academic Press, San Diego, CA, 2001), Chap. 8.

[28] D.L. Hart, Arthur F. Judy, Rajarshi Roy and James W. Beletic, Phys. Rev. E **57**, 4757 (1998); D.L. Hart, Arthur F. Judy, T.A.B. Kennedy, Rajarshi Roy and K. Stoev, Phys. Rev. A **50**, 1807 (1994).

[29] K. Ito, *Lectures on Stochastic Processes* (Tata Institute of Fundamental Research, Bombay, 1960).

[30] R.L. Stratanovich, *Topics in the Theory of Random Noise*, Vols I. and II. (Gordon & Breach, New York, 1963).

[31] H. Risken, *The Fokker-Planck Equation* (Springer-Verlag, Berlin, 1989).

[32] M.J. Werner and P.D. Drummond, J. Comput. Phys. **132**, 312 (1997).

[33] P.D. Drummond and I.K. Mortimer, J. Comput. Phys. **93**, 144 (1991).

[34] S.J. Carter, Phys. Rev. A. **51**, 3274 (1995).

[35] J.R. Thompson and Rajarshi Roy, Phys. Rev. A **43**, 4987 (1991).

[36] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in Fortran: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1992).

[37] C. Headley, G.P. Agrawal, IEEE J. Quantum Electron. **QE-31**, 2058 (1995), C. Headley, G.P. Agrawal J. Opt. Soc. Am. B. **13**, 2170 (1995).

[38] S.H. Perlmutter, M.D. Levenson, R.M. Shelby and M.B. Weisman, Phys. Rev. Lett. **61** 1388, 1988.

[39] F. Kh. Abdullaev, J.H. Hensen, S. Bischoff and M.P. Sorensen, J. Opt. Soc. Am. B. **15**, 2424 (1998); F. Kh. Abdullaev, J.G. Caputo, and Nikos Flytzanis, Phys. Rev E. **50**, 1552 (1994).

[40] William H. Glenn, IEEE J. Quantum Electron. **QE-25**, 1218 (1989).

[41] R. Trebino. *Frequency-Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses* (Kluwer Academic 2002).

[42] G.P. Agrawal *Nonlinear Fiber Optics* (Academic, San Diego, 2001).

[43] J.M. Dudley, X. Gu, L. Xu, M. Kimmel, E. Zeek, P. O'Shea, R. Trebino, S. Coen, R.S. Windeler, "Cross-correlation frequency resolved optical gating analysis

of broadband continuum generation in photonic crystal fiber: simulations and experiments," Opt. Express **10**, 1215 (2002).

[44] Q.D. Liu, J.T. Chen, Q.Z. Wang, P.P. Ho, and R.R. Alfano, "Single pulse degenerate-cross-phase modulation in a single-mode optical fiber," Opt. Lett. **20**, 542 (1995).

[45] T. Sylvestre, H. Maillotte, E. Lantz, and D. Gindre "Combined spectral effects of pulse walk-off and degenerate cross-phase modulation in birefringent fibers", Journal of Nonlinear Optical Physics and Materials 6, 313-320 (1997).

[46] Q.D. Liu, L. Shi, P.P. Ho, R.R. Alfano, R.J. Essiambre, and G.P. Agrawal, "Degenerate cross-phase modulation of femtosecond laser pulses in a birefringent single-mode fiber," IEEE Photon. Tech. Lett. **9**, 1107 (1997).

[47] F.G. Omenetto, B.P. Luce, D. Yarotski and A.J. Taylor, "Observation of chirped soliton dynamics at l= 1.55 mm in a single-mode optical fiber with frequency-resolved optical gating," Opt. Lett. **24**, 1392 (1999).

[48] F.G. Omenetto, Y. Chung, D. Yarotski, T. Shaefer, I. Gabitov and A.J. Taylor, "Phase analysis of nonlinear femtosecond pulse propagation and self-frequency shift in optical fibers," Opt. Commun. **208**, 191 (2002).

[49] F.G. Omenetto, J.W. Nicholson, B.P. Luce, D. Yarotski, A.J. Taylor, "Shaping, propagation and characterization of ultrafast pulses in optical fibers," Appl. Phys. B **70**[Suppl.], S143 (2000).

[50] N. Nishizawa and T. Goto, "Experimental analysis of ultrashort pulse propagation in optical fibers around zero-dispersion region using cross-correlation frequency resolved optical gating," Opt. Express **8**, 328 (2001).

[51] N. Nishizawa and T. Goto, "Trapped pulse generation by femtosecond soliton pulse in birefringent optical fibers," Opt. Express **10**, 256 (2002).

[52] N. Nishizawa and T. Goto, "Characteristics of pulse trapping by use of ultrashort soliton pulses in optical fibers across the zero-dispersion wavelength," Opt. Express **10**, 1151 (2002).

[53] N. Nishizawa and T. Goto, "Ultrafast all optical switching by use of pulse trapping across zero-dispersion wavelength," Opt. Express **11**, 359 (2003).

[54] , K. Ogawa, M.D. Pelusi, "Characterization of ultrashort optical pulses in a dispersion-managed fiber link using two-photon absorption frequency-resolved optical gating," Opt. Commun. **198**, 83-87 (2001).

[55] R.A. Altes, ”Detection, estimation, and classification with spectrograms,” J. Acoust. Soc. Am. **67**(4), 1232 (1980).

[56] A. Christian Silva, ”GRENOUILLE - Practical Issues,” unpublished.

[57] J. Garduno-Mejia, A.H. Greenaway, and D.T. Reid, ”Designer femtosecond pulses using adaptive optics,” Opt. Express **11** 2030 (2003).

[58] P. O’Shea, M. Kimmel, X. Gu, R. Trebino, ”Increased-bandwidth in ultrashort-pulse measurement using an angle-dithered nonlinear-optical crystal,” Opt. Express **7**, 342 (2000).

[59] P. O’Shea, M. Kimmel, R. Trebino, ”Increased phase-matching bandwidth in simple ultrashort-laser-pulse measurements,” J. Opt. B **4**, 44 (2002).

[60] S. Akturk, M. Kimmel, P. O’Shea, R. Trebino, ”Measuring pulse-front tilt in ultrashort pulses using GRENOUILLE”, Opt. Express **11**, 491 (2003).

[61] K. J. Blow, D. Wood, ”Theoretical Description of Transient Stimulated Raman Scattering in Optical Fibers,” IEEE J. Quant. Elect. **25**, 2665 (1989).

[62] R.H. Stolen, J.P. Gordon, W.J. Tomlinson, J. Opt. Soc. Am. B **6**, 1159 (1989).

[63] P.V. Mamyshev and S.V. Chernikov, Sov. Lightwave Commun. **2**, 97 (1992).

[64] C. Headley III, *Ultrafast Stimulated Raman Scattering in Optical Fibers, Ph.D. Thesis, University of Rochester, NY (1995).*