School of Computer Science
University of Technology Sydney

Unsupervised learning for high performance compression of genomic data collections

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Tao Tang

June 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Tao Tang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature of Candidate

Production Note: Signature
removed prior to publication

Date 11/03/2021

Acknowledgments

First and foremost, I offer my deepest gratitude to my supervisor, Professor Jinyan Li, whose expertise was invaluable in formulating the research questions and methodology, for his extensive support and guidance during last three years of my PhD studies. His guidance including scientific writing and research ideas made all of this work possible. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I am very thankful to Dr. Yuansheng Liu, who is our team member. Dr. Yuansheng Liu has vast knowledge in the area of bioinformatics, especially the compression of NGS data, he gives many insightful suggestions and comments to improve the quality of our projects. I would also like to thank other team members: Dr. Chaowang Lan, Dr. Hui Peng, Xiaocai Zhang, Xuan Zhang, Dr. Yi Zheng and Dr. Zhixun Zhao, for their kind help during my studies. The fantastic activities with them are enjoyable and relaxing under the heavy pressure of research.

This work would not be possible without the people with whom I collaborated and whose help was essential in many projects I participated in: Professor Gyorgy Hutvagner, Dr. Kun Yu.

Furthermore, I gratefully acknowledge all the organizations provided the scholarships including International Research Scholarships, ARC Discovery Scholarship.

I am deeply grateful to my wonderful parents for their support and love throughout all my life. I also extend my deep gratefulness to my relatives and friends in China.

Tao Tang

UTS, Sydney, Australia

Feb 2021

Contents

Certificate . . . . **i**

Acknowledgment . . . . **iii**

List of Figures . . . . **ix**

List of Tables . . . . **xi**

List of Publications . . . . **xiii**

Abstract . . . . **xv**

Chapter 1 Introduction **1**

1.1 Compression of Genome Sequence Data 3

1.2 Compression of Reads Data 6

1.3 Effect of Error Correction on Compression and De Novo Assembly of NGS Data 8

1.4 Research Contributions 10

1.5 Thesis Organization 12

Chapter 2 Related Work and Literature Review **13**

2.1 Reference-based Genome Compression 13

2.2 Short Reads Compression 15

2.3 De Novo Assembly and Error Correction of Reads Data 18

2.3.1 De Novo Assembly 18

2.3.2 Error Correction 20

2.4 Summary 22

Chapter 3 Sketch Distance-based Clustering of Chromosomes for Large Genome Database Compression **23**

3.1 Introduction 23
3.2 Method 27
3.2.1 Construction of Distance Matrix for a Set of Chromosome Sequences 27
3.2.2 Clustering of Chromosomes from the Distance Matrix . 28
3.2.3 Compression 30
3.2.4 Data 32
3.3 Experiment and Results 32
3.3.1 Test Methodology 32
3.3.2 Gains of Compression Performance 35
3.3.3 Speed Performance 37
3.4 Summary 41

Chapter 4 Unsupervised Learning for Compression of Short Reads Databases 43
4.1 Introduction 43
4.2 Theoretical Analysis 44
4.3 Method 46
4.3.1 Transformation of a Reads Set into a Feature Vector . 47
4.3.2 Clustering of the Feature Vectors 49
4.3.3 Group-by-group Compression 50
4.4 Result and Performance Analysis 51
4.4.1 Compression Performance Gains 52
4.4.2 3D Visualization for the Clusters of Reads Data Sets after Transformation 54
4.4.3 PCA Analysis on the Feature Vectors Converted from the Reads Datasets 56
4.4.4 Time Complexity and Speed Performance 58
4.5 Summary 59

Chapter 5 Assembly-improved Compression of Genomic Short Reads via Error Correction 61

5.1 Introduction 61
5.2 Error Correction and Compression 63
5.3 Error Correction and Genome Assembly: Current Observations 64
5.4 Correction-integrated Method for Multiple Reads Sets Compression: An advancement for Both Data Compression and Genome Assembly 67
5.4.1 Distance Matrix Construction 68
5.4.2 Path Graph Construction 68
5.4.3 Divide Path Graph into Groups 69
5.5 Advancement: Error-corrected Compression of Reads Databases and their Assembly Performance 71
5.5.1 Compression Ratio Gain 71
5.5.2 Effect on De Novo Assembly 74
5.5.3 Complexity Analysis and Speed Performance 76
5.6 Summary 77

Chapter 6 Conclusions and Future Work 78
6.1 Conclusions 78
6.2 Future Work 79

Bibliography 81

List of Figures

1.1 An example of FASTA format data 4

1.2 An example of FASTQ format data 7

3.1 Workflow of our ECC approach for compressing a database of genomic sequences 26

4.1 3D visualisation of the feature space after Multi-Dimensional Scaling from (a) database-17, (b) database-17 by clustering of MRC; (c) database-21, (d) database-21 by clustering of MRC; (e) database-50, (f) database-50 by clustering of MRC; (g) database-100, (h) database-100 by clustering of MRC 55

4.2 3D Visualisation of the feature space after Principal Component Analysis from (a) database-17 (b) database-21; (c) database-50 (d) database-100 57

5.1 Common characteristics of reference-free compression, error correction and de novo assembly 62

5.2 Example of de Bruijn based assembly of corrected and uncorrected data. (a) A set of 17 reads with potential errors. (b) The de-Bruijn graph based on reads set, red arrows indicates chimeric connection, yellow arrows indicates "tip" (nodes that disconnected on one end). (c) The assembly result based on original set, include redundant (the first three) and short (the last) contigs. (d) The assembly result based on original set after removing erroneous structures. (e) The MSA and corresponding correction, only the alignment of corrected reads are shown. (f) The de-Bruijn graph based on corrected reads set. Erroneous connection and nodes are removed. (g) The assembly result based on corrected set with continuous contigs. 65

List of Tables

3.1 Compression ratio for the H. sapiens dataset-60 (171GB) . . . 33

3.2 Compression ratio gain of ECC against different cases on H. sapiens dataset-60 (171GB) 34

3.3 Compression ratio on H. sapiens dataset-1152 (3128GB) . . . 36

3.4 Compression ratio gain of ECC against different cases on H. sapiens dataset-1152 (3128GB) 37

3.5 Compression ratio on the Oryza sativa dataset-2818(1,012GB) 38

3.6 Compression ratio gain of ECC against different cases on Oryza sativa dataset-2818(1,012GB) 39

3.7 Reference selection time of ECC (in hours) 39

3.8 Compression time of each algorithm on the three datasets . . . 40

4.1 Two sequences that share the same minimizer 46

4.2 First 10 elements of the feature vector transformed from a reads set 46

4.3 Example of the minimizers and corresponding feature vector of a set of 10 reads with k=2 48

4.4 Example of the first 10 features of feature vectors from two clusters in the database of 21 reads datasets 51

4.5 Four collections of reads data sets 52

4.6 Compressed file size (in byte) comparison between the straightforward *one-by-one* approach and our MRC approach (with $k = 7$ for the setting of k-minimizer) 53

4.7 Minimizers of high importance in our PCA analysis 56
4.8 Transformation and clustering time of MRC (in second), red indicates the shortest compression time of each dataset. 59

5.1 Four collections of reads data sets 72
5.2 Compressed file size (in byte) comparison between the straightforward *one-by-one* approach and error correction (via Karect) + PMRC approach 73
5.3 Number of mismatched nucleotides of minicom 74
5.4 NGA50 of the assembled by SPAdes in different modes and with correction of Karect in advance 75
5.5 Computation time of PMRC and compression time of PgRC1.2 and minicom on each corrected database (in second). 76

List of Publications

Below is the list of journal papers associated with my PhD research:

Journal Papers Published

- **Tang, T.**, Liu, Y., Zhang, B., Su, B., & Li, J. (2019). Sketch distance-based clustering of chromosomes for large genome database compression. *BMC genomics*, 20(10), 1-9.

- **Tang, T.**, & Li, J. (2021). Transformation of FASTA files into feature vectors for unsupervised compression of short reads databases. *Journal of Bioinformatics and Computational Biology*, 2050048-2050048.

- **Tang, T.**, Hutvagner. G., Wang, W. & Li, J. Assembly-improved compression of genomic short reads via error correction: survey and advancement. Under review.

Abstract

The advanced next-generation sequencing (NGS) technologies have launched a new era of all fields of genetics. However, the vast quantity of data generated by NGS technologies also proposed great challenges to data storage, transmission and analysis. In this thesis, we focus on the compression of multiple data collections of short reads and assembled genome, we also explore the relationship between compression, error correction and de novo assembly of short reads data. First, we introduce an efficient clustering-based reference selection algorithm for the compression of genome databases. This method clusters the genomes into subsets of highly similar genomes using MinHash sketch distance, then applies a two-level compression based on the clustering result. The compression ratio gain of our approach can reach up to 20-30% in most cases for the datasets from NCBI, the 1000 Human Genomes Project and the 3000 Rice Genomes Project.

Furthermore, we propose a new clustering-based method for the compression of short reads datasets. Our approach transforms each file into a feature vector for clustering, then compresses the files in the same group together to increase the total number of detected overlappings during compression. The experiments show that our method achieves 20%-30% improvements in compression ratio than the previous one-by-one compression.

Finally, we review the relationship between reference-free compression, MSA based error correction and de novo assembly of short reads data. We demonstrate that high quality error correction can significantly reduce the number of mismatched nucleotides during reference-free compression and hence improve the final compression ratio. The experiment results verify our estimation and show that the same error correction also has a positive effect on de novo assembly in most cases. In addition, we also propose a path graph based method for compression of short reads datasets.