Distributed Algorithms on Big Data Frameworks for Alignment and Analysis of Big Data generated by Next-Generation Sequencing

A Thesis

Submitted for the Degree of

Doctor of Philosophy

in

School of Computing and Electrical Engineering

by

SANJAY RATHEE

(E. NO. D13016)

School of Computing and Electrical Engineering

Indian Institute of Technology Mandi

Mandi-175001, India

February 2018

INDIAN INSTITUTE OF TECHNOLOGY MANDI
MANDI- 175 001 (H.P.), INDIA
www.iitmandi.ac.in

Declaration by the Research Scholar

I hereby declare that the entire work embodied in this Thesis is the result of investigations carried out by me in the **School of Computing and Electrical Engineering**, Indian Institute of Technology Mandi, under the supervision of **Dr. Arti Kashyap** and that it has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on finding of other investigators.

Place: Signature:

Date: Name:

Declaration by the Research Advisor

I hereby certify that the entire work in this Thesis has been carried out by **Sanjay Rathee**, under my supervision in the **School of Computing and Electrical Engineering**, Indian Institute of Technology Mandi, and that no part of it has been submitted elsewhere for any Degree or Diploma.

Place: Signature:

Date: Name of the Guide:

Abstract

Name of the student: **Sanjay Rathee** Roll No: **D13016**

Degree for which submitted: **Ph.D** Department: **School of Computing and Electrical Engineering**

Thesis title: **Distributed Algorithms on Big Data Frameworks for Alignment and Analysis of Big Data generated by Next-Generation Sequencing**

Thesis supervisor: **Dr. Arti Kashyap**

Month and year of thesis submission: **February 2018**

During last two decades, a huge amount of data is being produced worldwide by various sources. Genomic data is one of the main sources for this huge data termed as Big Data. Next-Generation Sequencing (NGS) machines are producing up to six billion base pairs per run in very cost-effective manner. Currently, the main challenge is to process this huge genomic data to extract relevant information. During extraction of relevant information from this genomic Big Data, alignment and analysis are two most important tasks. In this thesis, we present very accurate and efficient distributed sequence alignment and analysis algorithms.

To tackle the problem of efficient sequence alignment, two distributed sequence alignment algorithms named as AVLR-Mapper and StreamAligner are proposed and implemented using Big Data framework Apache Spark. AVLR-Mapper is first sequence aligner which has distributed index generation approach. AVLR-Mapper uses most efficient search mechanism based on partitioning to reduce computation during read mapping. It outperforms most of the state-of-the-art sequence alignment algorithms in terms of accuracy and performance. StreamAligner is the first sequence aligner which can directly align stream of

machines as stream and output interesting patterns after alignment and analysis. It has a great scope in future for making sequencing, alignment, and visualization (or analysis) process automated. It showed better execution time (speedup upto 9.97x) due to better load balancing and stream processing engine. AVLR-Mapper and StreamAligner are implemented on Apache Spark and evaluated on IIT Mandi local cluster as well as Amazon EC2 cloud. Source code written in Java is available on GitHub.

To analyze large genomic datasets, three distributed association rule mining algorithms named as Reduced-Apriori (R-Apriori), Adaptive-Apriori (A-Apriori), and Flink-Apriori (F-Apriori) are proposed and implemented using Big Data frameworks Apache Spark and Flink. R-Apriori and A-Apriori are implemented on Big Data framework Apache Spark. R-Apriori uses a reduced approach for the second iteration of Apriori algorithm and minimizes computation to a great extent. R-Apriori outperforms conventional Apriori in terms of accuracy and efficiency. A-Apriori uses an adaptive approach for every iteration where the decision is made to use reduced or conventional Apriori approach before every iteration based on precomputations. A-Apriori always performs better than R-Apriori and conventional Apriori for all datasets. F-Apriori uses Apache Flink to handle iterative computations during Apriori and outperforms all association rule mining algorithms in terms of performance. All these association rule mining algorithms are written in Scala and evaluated on local cluster as well as Amazon EC2 cloud. These algorithms are used for analyzing large genome datasets to get interesting patterns from them. These algorithms can be used in Bioinformatics applications like cancer detection, SNP discovery, motif discovery and clustering. In summary, this thesis presents the architecture, algorithm, and implementation of two distributed sequence alignment and three distributed association rule mining algorithms targeted towards helping bioinformatic scientists to reduce the time and cost for alignment and analysis of large genomic datasets.

Acknowledgements

On this moment of submission of my thesis, I would like to thank all those good persons, who have guided me, supported me, inspired me and from whom I have learnt to live the life. I am grateful to my supervisor Dr. Arti Kashyap for letting me in, great guidance, support, encouragement, and the opportunity, she gave me to get exposure during my Ph.D. time. I thank her sincerely for everything, she has done for me. She has guided me with her invaluable suggestions and encouraged me a lot in the academic life. I am extremely grateful for her confidence in me and the freedom she gave for me to work. It was a great time for me to work with her. I am also grateful to Dr. Manohar Kaul, Department of Computer Science, Indian Institute of Technology Hyderabad, India for his support, excellent suggestions and the opportunity, he gave me to work with him. I express my special thanks to him for his suggestions, corrections and discussions on my projects. I am thankful to Dr. Ashutosh Sharma, Scientist at Agilent Inc., USA for the support, fruitful discussions on research problems and providing me large bioinformatics datasets for evaluation. I would like to thank my Doctoral Committee members Dr. Tulika Srivastava, Dr. Arnav Bhavsar, Prof. Banshi Dhar Chaudhary, and Dr. Bharat Singh Rajpurohit for their support and evaluation of my research work. I am thankful to Dr. Pankaj for helping me during early days of my research work. I am thankful to Dr. Renu for helping me throughout my research work. I thank my fellow lab members Rohit, Imran, Rajneesh, Yogesh, Ruchika and UHL lab members Pooja, Pawan and Rakesh for their helping nature and discussion on research topics. I am thankful to Robin, Shiwani, Monika and all my friends who have made my stay memorable at IIT Mandi. I am thankful to the Ministry of Human Resource Development (MHRD) and IIT Mandi for the financial support and fellowship during my Ph. D. tenure. I am thankful to IIT Mandi for providing the excellent computing facilities. Finally, I would like to thank my parents, sisters, and brother for allowing me to get a higher education, their beliefs in me, affection, love, support, good wishes and encouragement they gave me to fulfill the dreams of my life. All above, I am thankful to the God who has given me the strength to fight against all challenges which I faced all through my life and courage to carry out this work.

Sanjay Rathee

School of Computing and Electrical Engineering

Indian Institute of Technology Mandi, Himachal Pradesh, India

List of Publications

The following publications and manuscripts are included in the thesis:

1. **S. Rathee**, M. Kaul, and A. Kashyap, "R-Apriori: An efficient apriori based algorithm on spark," In *Proceedings of the 8th Workshop on Ph.D. Workshop in Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 27-34. DOI=http://dx.doi.org/10.1145/2809890.2809893.

2. **S. Rathee** and A. Kashyap, "Exploiting Apache Flink's Iteration Capabilities for Distributed Apriori: Community Detection Problem as an example," In *Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE Explore, pp. 739-745, (2016).

3. **S. Rathee** and A. Kashyap, "An Adaptive Mapreduce based Apriori Algorithm for Frequent Itemset Mining," In *Business Analytics and Intelligence*, Springer Ebook, M. Mathirajan and U.D. Kumar (Eds), December (2016).

4. **S. Rathee** and A. Kashyap, "SDC-Miner: An association rule mining algorithm for crowd mining of uncertain data using Apache Spark," In proceedings of *5th International Conference on Business Analytics and Intelligence*, IIM Bangalore, December (2017).

5. **S. Rathee** and A. Kashyap, "StreamAligner: A Stream based read-mapping tool on Apache Spark," In *Journal of Big Data*, Springer, 5(1), p.8, Available: `https://github.com/sanjaysinghrathi/StreamAligner`

6. **S. Rathee** and A. Kashyap, "Adaptive-Miner: an efficient distributed association rule mining algorithm on Spark," In *Journal of Big Data*, Springer, 5(1), p.6, Available: `https://github.com/sanjaysinghrathi/Adaptive-Miner`

7. **S. Rathee** and A. Kashyap, "Spark-Indexer: A distributed Suffix array and BWT index generation tool on Apache Spark," Communicated in *Big Data Research*, Elsevier, Available: `https://github.com/sanjaysinghrathi/Spark-Indexer`

8. **S. Rathee**, M. Kaul, A. Sharma, and A. Kashyap, "AVLR-Mapper: An accurate and variable length read-mapping tool on Apache Spark," Communicated in *IEEE/ACM Transactions in Bioinformatics and Computational Biology*, Available: `https://github.com/sanjaysinghrathi/AVLR-Mapper`

Oral/Poster Presentations

1. **In 8th Ph.D. Workshop in Information and Knowledge Management (PIKM'15)**, ACM, Melbourne, Australia, November 2015

   Title – *R-Apriori: An Efficient Apriori based Algorithm on Spark*, Oral Presentation

2. **Research Fair 2016**, Indian Institute of Technology Mandi, India, February 2016

   Title – *Reduce approach for Apriori algorithm on Apache Spark*, Oral Presentation

3. **In International Conference on Advances in Computing, Communications and Informatics (ICACCI)**, Jaipur, India, September 2016

   Title – *Exploiting Apache Flink's Iteration Capabilities for Distributed Apriori: Community Detection Problem as an example*, Oral Presentation

4. **In International Conference on Business Analytics and Intelligence (ICBAI)**, IISc Bangalore, India, December 2016

   Title – *An Adaptive Mapreduce based Apriori Algorithm for Frequent Itemset Mining*, Oral Presentation

5. **In 11th Inter-Research-Institute Student Seminar in Computer Science (IRISS 2017)**, ACM Kolkata, India, January 2017

   Title – *Reduced-Apriori algorithm for association rule mining*, Poster Presentation

6. **Research Fair 2017**, Indian Institute of Technology Mandi, India, March 2017

   Title – *An adaptive approach for association rule mining on Apache Spark*, Poster Presentation

7. **In Intl. Conference on Business Analytics and Intelligence (ICBAI)**, IIM Bangalore, India, December 2017

   Title – *SDC-Miner: An association rule mining algorithm for crowd mining of uncertain data using Apache Spark*, Oral Presentation

Contents