Problem Analysis Assignment 02

Aisha Hassan shah 45920842

Macquarie University Sydney Australia

1 Introduction

1.1 BACKGROUND: [1]

This report demonstrates utilization of context aware, graph data mining technique in the Study of "Prenatal cell-free DNA (cfDNA) screening, also known as noninvasive prenatal screening in pregnant women in order to identify specific chromosomal abnormalities in the fetus such as trisomy 21 (Down Syndrome), trisomy 18 and trisomy 13, this particular screening also reveal the gender and blood type (Rh) of the fetus. In this specific testing (cfDNA), the main task is to extract DNA from the mother and fetus from maternal blood sample and examine the chromosome pattern (inside the DNA) of fetus. With normal distribution of chromosomes which is 46 chromosomes in each cell (23 from father side and 23 from mother side) and if there is any abnormality in chromosomes count and pattern than it is a flag shows increased chance for mentioned chromosome disorders and requires medical attention for further investigation. In this context, Graph data mining participates in detailed analysis on chromosomal disorder identification during pregnancy which will provide in-depth critical decision making envision to medical experts to plan possible diagnosis, treatments, prognosis and future support to the patient.

1.2 Motivation:

The study of human DNA is itself very complex and challenging and requires extensive attention to draw results on analysis. Maternal blood samples are gone through substantial pathology procedures in medical lab and come up with unsupervised data of DNA that contains unique genetic code, structure of chromosomes and count of total number of chromosomes in the pregnant woman and the fetus. This unstructured data is in the form of banding patterns, chromosome map (Karyogram), protein structure and cytogenetic mapping. With the advent of new analytical technique (Graph data mining) in DNA analysis, it is quite faster, convenient and accurate to discover hidden wanted patterns and structure of 23 chromosomes in pair (together 46) present in every DNA cell. Due to its extreme sensitivity, time constraints are applied on blood sample drawn from pregnant women (at least 10 weeks pregnancy) and multiple protocols and conditions should be checked, such as not carrying multiple fetuses, obese, Pregnant via an egg donor or gestational carrier, consuming specific blood thinner medicines).After investigating these all parameters, procedure can take place for the retrieval of correct blood sample, which provides unstructured (data) for further analysis. The main challenges for the chromosomal disorder identification tasks are: rare occurrence of anomalous chromosomal structure (shape, count,

protein levels etc.), attainment of all medical conditions (for blood sample) and traditional analyzing techniques for investigation, make identification of outcome result with suspecting of false alarm rate sometimes false positive and false negative. With this new prenatal (cfDNA) testing, still there are cases, coming up with inaccurate predictions on chromosomal disorders due to somewhere lacking in data sample (bias data) and nonfunctional and inefficient statistical analytical approaches which are not meant to be handled this kind of unique (each and every DNA is different from another) data where in depth understanding and reading of human DNA structure network and pattern matching with normal standards. With general classification and clustering techniques are not compensating the complete exploration of this structure and pattern mapping type data in order to provide highly accurate descriptive and predictive analysis on DNA study related to chromosomal disorder identification. Genetic analysis of DNA and its components (chromosomes, protein, strands etc.) study is very similar like protein and chemical graph structures. In this scenario the optimal approach is to implement graph data mining techniques on DNA datasets which will be able to search hidden patterns, perform pattern matching, obtained subgraphs structures to compare with standard DNA structures and derived highly efficient and accurate impressive results. This advanced approach will eradicate the chances of false positive and false negative results, which are being occurred with traditional statistical analysis methodologies.

2 Related Work: [2]

For the study of DNA chromosomal disorder identification, the best strategy in graph data mining is frequent sub graph mining and in this DNA screening case, the input graph datasets (transactional settings means multiple small graphs) completely transformed in to encoded format so it will be easy to find out subgraphs, after that for implementation of frequent subgraphs, systematically generated of a set of candidate subgraphs(candidate generation), these candidate subgraphs are the collection of subgraphs which have high frequency of occurrence and require to be checked or tested(used to count ,number of instances are present in the graph database). There are many existing relatable methods in the literature for implementation of Frequent Graph mining with transactional settings and here are mentioning some of them.

2.1 AGM (Apriori Graph Mining algorithm:

The main aim of this algorithm is to present graphs by an adjacency matrix (represent unweighted graphs as an n x n binary matrix and for weighted graph adjacency matrix assigns weights to the edges) and execute Breadth First Search (BFS) technique for retrieval of frequent subgraphs, these frequent subgraphs generates candidate subgraphs and when any two candidate subgraphs join by considering the size of the subgraphs, and it is identified with the number of vertices in both subgraphs would be same and after matching of number of vertices, both can be merged together to form an outcome graph(or resultant graph, its size is one vertex more than the two merged subgraphs) and on each and every iteration there is increment (addition) of one vertex in the subgraphs and it

is keep recurring till the groups of candidates (resultant graphs) can be tested against given standard data. It is vertex based candidate generation algorithm. The best part in this algorithm, there is less complexity in computational time of directed subgraph patterns due to the directions in edges, there number of occurrences are minimum and containing more subgraphs patterns but if sub graphs are undirected (edges are not having any direction) than the complexity of computational time has increased. AGM Algorithm performance is lacking with large size graphs in the datasets.

2.2 FSG (Frequent Sub Graph mining Algorithm):

This algorithm follows edge-based candidate generation method. This approach is also executing BFS strategy to look for frequent subgraphs, the number of edges define the size of the subgraph and two subgraphs of similar sizes can merged together and at every iteration new edge is adding in to merged subgraph creating (candidate generation) and every time its size is exactly one greater than the previous frequent subgraphs (edge wise). There are functionalities in FSG algorithm such as utilization of sparse graph representation (number of edges are less) which is directly minimize the storage and computation time of this algorithm, due to recursive addition of each edge on every iteration it increases the size of frequent subgraphs allows efficient candidate generation, FSG is also using canonical labelling (uniquely labelled graphs for identification, if two graphs are isomorphic to each other, then there is same code assign to both graphs for controlling graph redundancy),graph isomorphism(exact match between two subgraphs) and optimization of candidate generation and count the frequency of subgraphs occurrence. These supportive features of FGA algorithm bring promising results with small subgraphs and large graph database. The drawback of this algorithm is that, by executing canonical labels, system is getting same code of two similar subgraphs assuming that both are isomorphic (on the basis of same topological structure and similar labelling of edges and vertices, presentation of graphs is not important, they may look different) and this feature needs further evaluation for association with class P and NP-Complete problem (due to presence of subgraph isomorphism).

2.3 gSPAN (Graph-based substructure pattern mining algorithm)[3]):

This algorithm belongs to frequent subgraph mining and follows pattern growth approach technique and generates a tree like hierarchical structure. The searching criteria is based on DFS (depth first search) manner and generates frequent subgraphs structure overall possible patterns, using the right-most path extension and with every node contains DFS code (way of documenting the vertices and edges of a graph in tabular form, but with special rules).If the subgraphs are isomorphic then containing a same DFS code. In gSpan algorithm, it only carries the transactional list of discovered subgraph instead of setting of complete subgraph on every node of tree. It performs pruning by deleting nodes which are not satisfying the minimal DFS code (lexicographic order) and if subgraph attending the minimal DFS code, it is added to the transactional list and look

for next subgraph. This process is keep iterated till the DFS- code of subgraph is non-minimal. GSpan is also controlling and monitoring the duplication in discovered subgraphs. This algorithm is free from candidate generation and save space due to depth first search. The drawback of this algorithm is that, it is repeatedly loading graphs in to memory and checking for specific edges and end up with unrequired searching and loading those subgraphs which are not having those wanted edges.

3 Identified Methodologies:[4]

DNA chromosomal disorder identification is the process, where it is very important to understand DNA structure which is identical like any structure of molecules and inside the DNA there are 23 paired chromosomes (together 46) thread like bodies surrounded by proteins. The prime task is to convert all unstructured data obtained after lab work in to graphical format. So, it will convenient to understand the boundaries between all components inside DNA structure and the main target is to focus on chromosomes distributions and their structure. With the help of graph mining technique we are evaluating all possible subgraphs of the chromosome and match patterns with system defines graph patterns to do comparisons and make decisions on graph samples. For this identification, a very efficient graph mining pattern matching technique which is able to bring all possible outcomes without any data redundancy and explore/discover all possible nodes and edges or any part of graph sample which can add value in analysis and decision making on every outcome report.

3.1 Problem statement:

The target problem is that, to minimize or/and eradicate false positive and false negative chances from outcome results after analysis of DNA structure. For this we are implementing Diagonally Subgraphs Pattern Mining technique. In this process, first we extract information from unstructured files, transform it in to required format (graph datasets) and introduce diagonally subgraphs pattern matching approach to search frequent subgraph over a collection of graphs from the graph dataset.

Input Features	Output Features
• Database of graph transactions(multiple graphs) • Undirected simple graph (no loops, no multiples edges) • Each graph transaction has labels associated with its vertices and edges • Transactions may not be connected Minimum support threshold (min sup)	• Frequent subgraphs that satisfy the minimum support constraint • Each frequent subgraph is connected • All retrieved patterns of subgraphs from the graph data set will compare or perform pattern matching with standard given graph pattern of chromosome (Karyogram).

3.2 The Proposed Methods:

The Diagonally Subgraphs Pattern Mining (DSPM) is the combined technical approach of Apriori (breadth first search) and DFS (depth first search) and also carries some other key functionalities to make this algorithms exclusive and efficient. This multifunctional algorithm is possible best approach to handle this dynamic data and detailed investigation of each and every subgraphs and trees will be evaluated for final pattern matching with standard (or perfect) graph pattern of chromosome (Karyogram).Some important points related to DSPM are mentioned below. DSPM investigates search space in depth search manner and maintaining transaction ID lists (TID) with set of frequent subgraphs (present in canonical format and corresponding with tree search space). Another technique is going to apply on subgraphs is FAM pruning (which is monitoring for each generated candidate(breadth search manner), with k+1 edges and its subgraphs with K edges are also frequent and if not it means candidate is also not frequent so we can prune it) The reverse depth approach also contributes for flexible generation of candidates, application of FAM pruning of (k-1) edges of subgraphs and efficiently consume search space and also explores the son of each pattern in descending order.

3.3 Main Procedure:

DSPM traverses in search space in reverse depth manner, search space contains all mined frequent patterns and they are in form of prefix based lattice (ascending ordered).It can generate group of candidates from every frequent pattern and checking FAM(frequency anti-monotone)property for each explored pattern using previous mined patterns. Now apply reverse lexicographic order, for each edge (level wise) and it is an iterative procedure, in a reverse DFS search of all subtree under subgraph, it means finding all frequent subgraphs with min DFS codes in search space. The utilization of search space in two steps one for large set of candidates that can be ordered in single support counting and second is for limited set of candidates obtaining from FAM pruning and handles in each iteration set limited set of candidates. DSPM validates and generates candidates by consuming these three features: Candidate Generation and FAM Pruning, Validating min DFS-Code and Support Counting (This is recursive process knowing, candidates are keep growing and transactional IDs are keep assigning then it updates the support counting of the candidates.) This DSPM behaves as hybrid algorithm, there are many techniques are being used in this algorithm for candidates generation (BFS breadth first search), candidates Pruning (BFS breadth first search), Search space exploration (DFS depth first search) and enumerating subgraph (DFS depth first search). This DSPM algorithm inherits two different techniques of pattern mining (DFS and BFS) and also apply reverse depth search and prefix base lattice which can look for the pattern (chromosome Karyogram), which we are looking for in each and every Graphs and subgraphs of DNA dataset, with all this functionality and algorithm never compromised over the running time and produce better results in comparison with other algorithms In Figure A, workflow of DSPM is clearly explained, where input data

is in form of DNA format and stays in graph database platform then explores all technique of DSPM algorithm, all crucial and most functional searches and pattern matchings and in the end all possible pattern of data is come up as outcome, after that comparison with standard distribution has been performed on extracted data patterns and produces results as per the match and visual information retrieved from this whole graph mining process. In Figure B, most efficient performed algorithm is DSPM on generation of frequent subgraph patterns, after that gSpan, following with FSG and in the last AGM.Accuracy score shows the managing and controlling power of algorithms and their functionality, they apply on data and bring accurate and precise results .As chosen most feasible algorithm DSPM resolve almost the issue of false positive and false negative cases in detection of chromosomal abnormality and represents that data has been searched and explore it's all components to bring error free patterns from input data.

3.4 An Overview of implementation of Diagonally Subgraphs Pattern Mining (DSPM) algorithm Figure A:

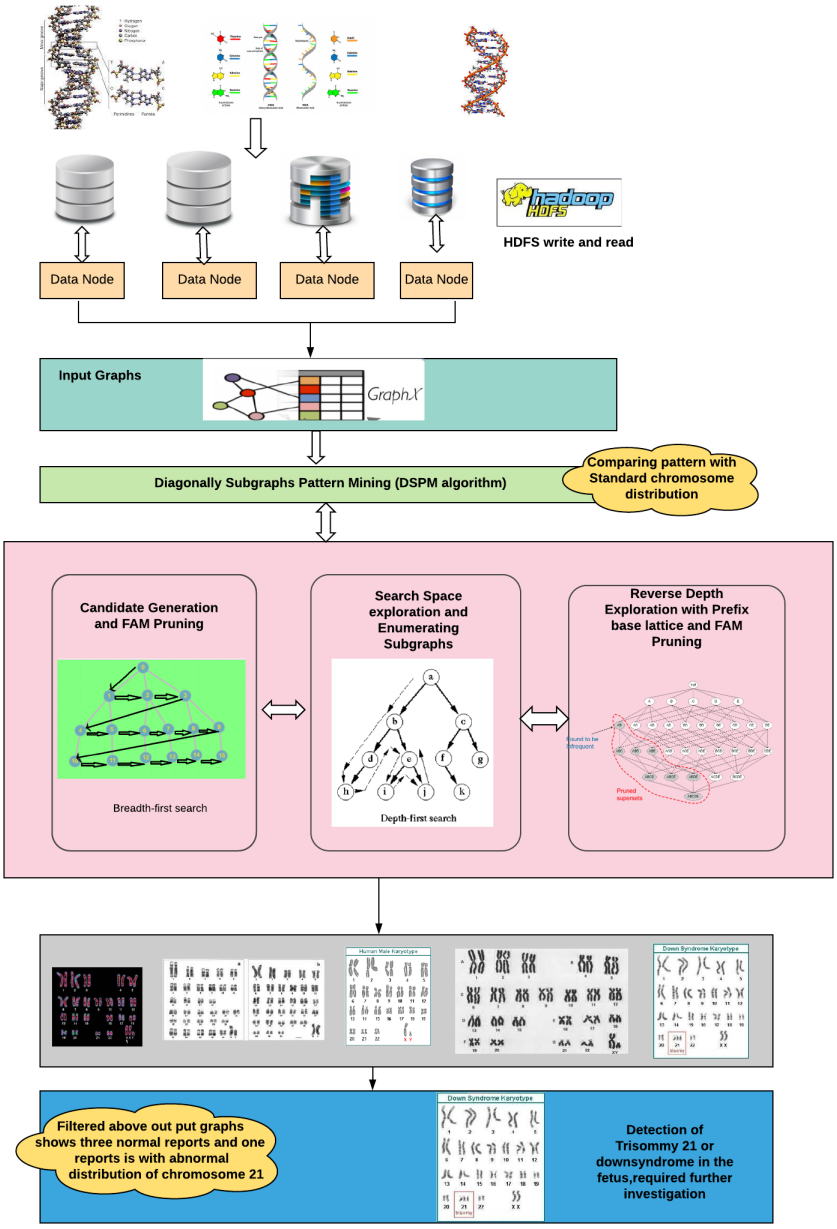
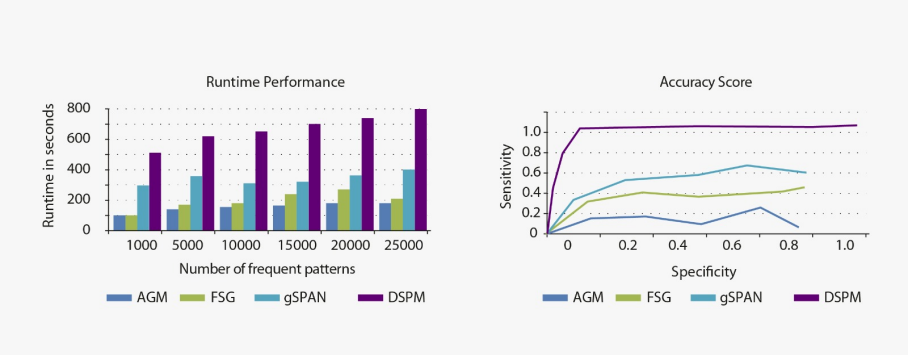


Figure A

3.5 An overview of Performances and Accuracy of all above discussed Algorithms:



4 Conclusion and Future Work:

Hence, the customized and formulated algorithm (DSPM) performs and evaluates all aspects of provided data and by adapting and consuming new techniques due to the complexity and longevity of data and outcome results are satisfying without compromising the execution time of number of frequent patterns and accuracy (which was the target problem of the system) is also higher than other comparative algorithms .In future there must be more progressive and robust algorithms of graph mining should introduce to the system and there are many subgraph mining tools available such as Mfinder, MAVisto, FANMOD, Kavosh, NetMODE and many more which are participating in graph mining and they have their own algorithms and limitation to deal with different diversity of data. There is possibility of Study of DNA structure data, with implementation of any latest graph mining tool will bring more informative results from the graph datasets. There is broad space available to investigate as Future work related DNA chromosomal disorder identification for instance ,study of obtained pattern for more chromosomal ailments such as Trisomy 16,Trisomy 22,Triploidy,Sex chromosome aneuploidy and certain disorders which are causing due to chromosomal deletion(microdeletion syndrome, Prader-Willi syndrome etc.),single gene disorder and many more. DNA structure is most amazing source of data and utmost information we can extract from it, Experts will be able to analyze, understand and provide productive decision making ability to medical teams in current and research studies .For this all ,capturing this complex data in the decent format where during transformation not too much data will lost and for modelling purposes the most simulated, appropriate and functional statistical analytical algorithms are required to bring information from data in most efficient and accurate form.

References

1. Mayo Clinic,(2020). Prenatal cell-free DNA screening. Mayo Clinic Website

2. Aida Mrzic, P. M. (n.d.). Grasping frequent subgraph mining for bioinformatics applications. Retrieved from BioData Mining webpage Link to website Grasping frequent subgraph mining for bioinformatics applications
3. T.Ramraja, R.Prabhakarb (2017). Frequent Subgraph Mining Algorithms – A Survey,Location India,Publisher: ScienceDirect. Click here
4. Moti Cohen Ehud Gudes(2004, January). Retrieved from Reserach Gate. Diagonally Subgraphs Pattern Mining Click here