ANALYSIS OF MAPPED HUMAN GENOME FOR DISEASE PREVENTION AND CURE

MANASA KAUSHIK
Dept. of Information Sciences and Engineering
PESIT-BSC
Hosur Road, Bangalore, India
manasakaushik1@gmail.com

NAMRATHA MANIKONDA
Dept. of Information Sciences and Engineering
PESIT-BSC
Hosur Road, Bangalore, India
nammy.manikonda@gmail.com

MANASA M
Dept. of Information Sciences and Engineering
PESIT-BSC
Hosur Road, Bangalore, India
way2manasa@gmail.com

SAMANTHA MINNIE WILLIAMS
Dept. of Information Sciences and Engineering
PESIT-BSC
Hosur Road, Bangalore, India
samanthaminnie.22@gmail.com

***Abstract* — In the recent years, a lot of focus has been given on determining the entire DNA sequence of humans so that they can be mapped and analyzed. This knowledge can then be put to beneficial use of disease prevention and cure. Here, we introduce a modern statistical machine learning approach that can establish a statistical relationship between clinical, genomic and environmental variables. We also introduce the concepts of well-known regression analyses such as linear and logistic regressions for clinical analyses and Bayesian networks to analyze more complicated data. The human genome however, would require huge data storage spaces and hence, we introduce big data in this aspect i.e., in terms of clinical data, gene expression studies and their interactions with the environment.**

We conclude this with a modern approach called Bayesian network that is promising in its capabilities of analyzing big data sets to provide a more comprehensive understanding of human physiology and disease, thus enabling us to intervene and cure human genome defects that would otherwise remain out of reach.

***Keywords* – Bayesian Networks, Hierarchical Bayesian Networks, Linear Regression, Logistic Regression, Polynomial Regression**

1. INTRODUCTION

The study of human genomics, like other major sources such as astronomy, YouTube and Twitter has led to a serious issue in terms of data storage and data analysis. The implementation of big data analysis has become a necessity in terms of data storage as it would help sort relevant data and consequently enhance space optimization. Human genomics is the study of the complete genetic material of humans. The typical human DNA would contain approximately 3.2 million bases pairs, (bases are those that primarily constitute the DNA) which would mean 6.4 million bases. The human genome, if mapped, would prove to be of astounding help as it would provide sufficient detail about its structure and help us understand how we could modify it for the greater good.

The present applications of human genome are already hitting new roofs. Applications such as *personalized medicine, creation of new drugs, preventative medicine* turned out to be exemplary. However, storing the genome is no simple task. By 2025, around a 100 million to 2 billion genomes could have been mapped, according to an article published in the journal PLoS Biology. On estimation, the data storage demands a minimum of 2 Exabytes, over which a genomic analysis would cause an explosion of data driving it to occupy

space up to a whooping 40 Exabytes. The problem would not end here, as the computing requirements for acquiring and distributing would bring a greater overhead. This is because the data that must be stored for a single genome are 30 times larger than the size of the genome itself, to make up for errors incurred during sequencing and preliminary analysis. This huge amount of data cannot be sorted and analyzed in given short period of time. This hindrance hasn't yet been overcome successfully.

Multiple approaches have been proposed and among them, the use of graphical models and machine learning have proven to be most efficient. We present this paper with an intention of presenting a simple and statistical view for obtaining a clear picture of the present situation in the field of genomics and proposing a theoretically better approach. This would open new doors and establish a greater horizon for research and development in the upcoming years.

2. EXISTING SYSTEMS

The present implementation of genome mapping involves a few notable approaches.

2.1. HAPLOTYPE METHODS

The BLADE algorithm, is one such algorithm that uses the Bayesian method. It was proposed by Liu *et al*, and is in direct association with the paper 'Bayesian Analysis of Haplotypes for Linkage Disequilibrium Mapping'. This model covers multiple possibilities of observing and segregating defects, and models the haplotypes in the ancestral genes, the ones over which the disease-causing mutations occurred. Thus, it can model historical recombinations and mutation events based on a primary set of founder genes that given initially.

2.2. SPATIAL-BASED HAPLOTYPE METHODS

The research by Molitor *et al* and Thomas *et al* make use of special mapping techniques, and provide the haplotypes with similarities as inputs.

This is done under the notion that haplotypes that are similar are most likely to carry the same disease-causing variants. The spatial mapping is done by providing inputs in the form of a weighted matrix, where the haplotypes with the 'closest' attributes are given the highest weights. The dependencies among the haplotypes are represented by using a Conditional Autoregressive Prior (CAR), which depicts how similar haplotypes would most likely show similar risk factors. Thus this method would provide us with results that contains multiple regions, where each of them contain multiple haplotypes and the regions with high risks are often clustered together.

One other notable approach is the "Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine (Changwon Yoo, Luis Ramirez, Juan Liuzzi)" as published in the *Int Neurounol Journal 2014;18:50-57,* where they introduce the concepts of regression analyses that include linear and logistical approaches and the statistical model of Bayesian networks. This approach is largely suitable at analyzing big datasets that consists of extremely large and diverse data. We will discuss this particular approach at large, as we find it to be most suitable, keeping in mind the similarities in the datasets being provided and the modelling of the system.

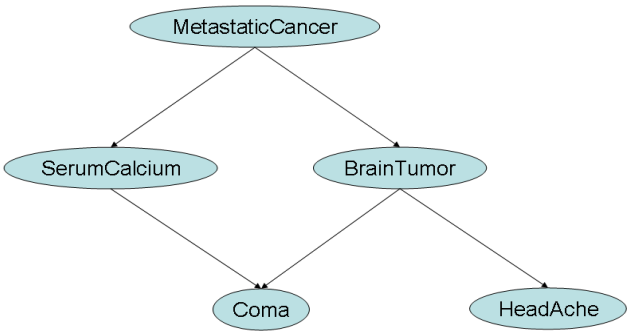

Fig. 1 A causal Bayesian Network depicting a simple probabilistic pathway.

As a compendium, this uses *Linear Regression* in the sense that it is a statistical method that assumes an output based on the weightage provided by its input variables. It simply models the relationships between a dependent variable and one or more independent variables. A simple linear regression equation is:

$$Y=b0+\sum(biXi)+\epsilon \qquad (1)$$

where Y is a continuous dependent variable,Xi are independent variables that are usually continuous, ϵ depicts some "error".

Logistic Regression is similar on most aspects, except that it uses a 'logit', basically a function that transforms this weighted value at output by using a mathematical function. This result is mapped to a value between 0 and 1, which can thus be considered as a probabilistic outcome. Both these data mining models are used for just clinical data.

Bayesian networks are used to learn from data. It's used here to learn from much more complex data than clinical data. The Bayesian network is plainly causal in nature, and depicts how the data at one layer can influence the next and then the one after it, thus falling into a chain like sequential events. This goes by the very popular Markov Chain, which describes this causal series of events.

The drawbacks of these models can be listed as follows:

Firstly, Linear Regression can assume only a straight-line relationship between the dependent and independent variables. This, in some aspects, could turn out to be incorrect.

Secondly, Bayesian Networks, though being inferential models, have a drawback as inferencing within a large number of variables, is not feasible practically. Adding overhead, they also have issues dealing with non-propositional domains. Aggregate data types are also difficult to represent, as issues such as loss of expressivity or a tendency to ignore possible probabilistic dependencies among the structure components.

And lastly, the major issue of storage of such extensive data remains unsolved completely.

3. PROPOSED SYSTEM AND APPROACH

So in the light of these drawbacks,we wish to propose another approach, which utilizes two primary components. Both these components run by utilising only the key parts of the human genome, thus averting the necessity of using the complete genome. These key factors can be markers, or mostly loci, which all occur in the same certain chromosome. These loci are bunches of sequences of genes. The components are as follows:

3.1. POLYNOMIAL REGRESSION MODEL:

The polynomial regression is a widely used as it assumes a non-linear relationship between the independent variable x and the dependent variable y, where the variable x can be modelled to an nth degree polynomial. And since the disease are always showcased when multiple loci on the same certain chromosome are set, it shows how some aspects influence the genetic flow with much greater impact. Such dominant traits in the chromosomes would be best represented with the use of polynomial regression. Besides, polynomial regressions have already been used to describe various non-linear phenomena such as the growth rate of tissues, the progression of disease epidemics, and the like.

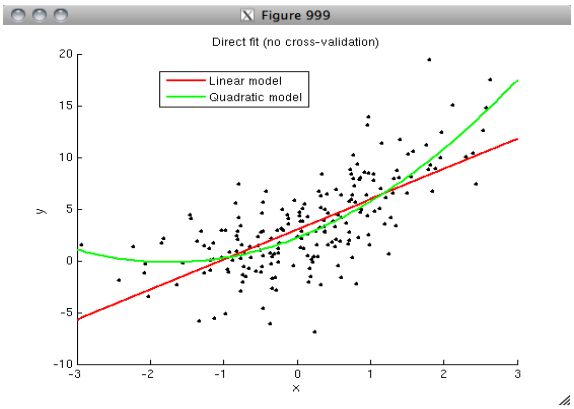

Fig. 2 A scatterplot depicting the non-linear relationship between variables x and y.

A simple polynomial regression equation:

$$E(y|\beta) = \beta 0 + \beta 1 x_{add} + \beta 2 x^2_{add} \qquad (2)$$

where β denotes the vector of regression parameters and x_{add} is the variable that codes for the genotype data.

The above image captures exactly how the linear plot tends to overestimate the values of x and y at the middle portion of the graph and underestimate them at lower and higher values.

However, as it can be observed, the polynomial/non-linear (quadratic) plot is able to characterize the data better.

As an advantage, the polynomial regression is able to accommodate a much wider range of the fitted line, thus making it a lot more flexible and robust. Also, it is able to establish a satisfactory relationship between the variables.

3.2. HIERARCHICAL BAYESIAN NETWORKS (HBNs):

Hierarchical Bayesian Networks as a solution to overcome these problems. Hierarchical Bayesian networks are a representation for probabilistic independencies between variables that belong to structured domains. The inference mechanisms of Bayesian networks can also be extended for hierarchical Bayesian networks as well. In Hierarchical Bayesian networks, a node can be of an aggregate data type which permits random variables within the network to represent arbitrarily structured types. Just like Bayesian networks even Hierarchical Bayesian networks can encode conditional probability dependencies.

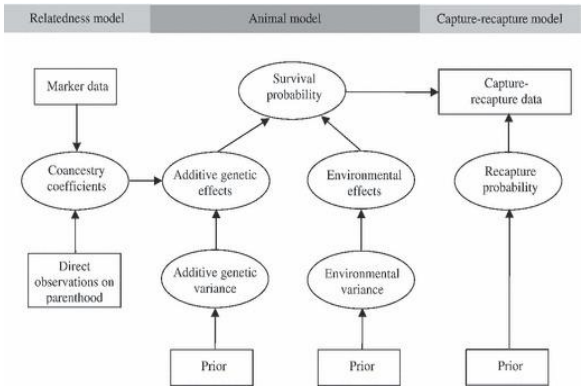

Fig. 3. A Hierarchical Bayesian Model represented graphically that depicts the additive effects of its structured domains

Hierarchical Bayesian networks can be divided into two parts, *structural* and *probabilistic*. The structural part contains the variables of the network and describes the probabilistic dependencies between them. The probabilistic part contain the conditional probability tables the help to quantify the links present in the structural part. The main advantages of this approach are:

1. In applications where the users provide multiple-observation data, the estimation provided by this approach tends to outbeat those provided by the traditional methods.
2. Estimation done by this approach is robust when compared to Bayesian networks as it takes all variables into accountant whether or not they have a significant impact on the outcome.

4. CONCLUSION

In all, we propose a much robust approach that operates on a completely non-linear method which can efficiently depict the dominant traits with the help of polynomial regression. It is an elegant way of incorporating the ideas of structured domains. The Hierarchical Bayesian network would also open multiple prospects of solving the unpredictability of atavism.

5. REFERENCES

[1] Bayesian Polynomial Regression Models to Fit Multiple Genetic Models for Quantitative Traits by Harold Bae, Thomas Perls, Martin Steinberg, and Paola Sebastiani

[2] Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine by Changwon Yoo, Luis Ramirez, Juan Liuzzi

[3] Disequilibrium Likelihoods for Fine-Scale Mapping of a Rare Allele by Jinko Graham and Elizabeth A. Thompson

[4] Inference of Gene Regulatory Network Based on Local Bayesian Networks by Fei Liu, Shao-Wu Zhang, Wei-Feng Guo, Ze-Gang Wei, Luonan Chen

[5] A survey of current Bayesian gene mapping methods by John Molitor, Paul Marjoram, David Conti and Duncan Thomas

[6] Bayesian gene set analysis for identifying significant biological pathways by Babak Shahbaba, Robert Tibshirani, Catherine M. Shachaf, and Sylvia K. Plevritis

[7] Hierarchical Bayes: Why All the Attention? by Bryan Orme, Sawtooth Software, Inc. 2000

[8] Bayesian Analysis of Genetic Differentiation Between Populations by Jukka Corander, Patrik Waldmann and Mikko J. Sillanpa ¨a

[9] Hierarchical Bayesian Networks:A Probabilistic Reasoning Model For Structured Domains by Elias Gyftodimos, Peter A. Falch

[10]ImageSource1:"https://www.w3.org/2004/09/13-Yoshio/PositionPaper.html" - Causal Bayesian Network.

[11] ImageSource2:"https://www.researchgate.net/figure/5478300_fig1_Figure-1-A-hypothetical-example-of-a-Bayesian-hierarchical-model-represented-as-a-graph"- Hierarchical Bayesian Network.

[12]ImageSource3:"http://randomanalyses.blogspot.in/2011/12/basics-of-regression-and-model-fitting.html" - Polynomial Regression vs Linear Regression.