Data Mining

Seminar 1

Report - Critical Review

A critical Review : Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier

Rickard Sandberg, Gösta Winberg, Carl-Ivar Bränden, Alexander Kaske,

Ingemar Ernberg, and Joakim Cöster

Department of Computer Science
Kristianstad University
Course: DT584C
Master’s in computer science

Student: Hazrat Ali <react.dev.se@gmail.com>
Teacher: Dawit Mengistu

Introduction

A critical review of a research article evaluates the strengths and weaknesses of the article[1]. Article reviews plays a very important role for the authors and researchers because it gives insight and feedback as per the research moral.

We read the text carefully and presented our fair and reasonable evaluation of the article. More specifically we have tried to review this article[2] in terms of the following.

1. What claims are made in the article?
2. How are the claims supported ?
3. What conclusions are drawn ?
4. Are there explanations overlooked ?

Molecular Biology

Before diving into the critical review of the article we should develop the background for terminology of molecular biology which has been used in this article. This is our first step of understanding of human genome and the application of data mining techniques for analysis.

Nitrogenous base is a molecule that contains nitrogen and has the chemical properties of a base.

The nitrogenous bases of DNA(ATCG) contain adenine (A), guanine (G), thymine (T), and cytosine (C). The nitrogenous bases of RNA(AUCG) contains uracil (U) instead of Thymine(T) bases[3].

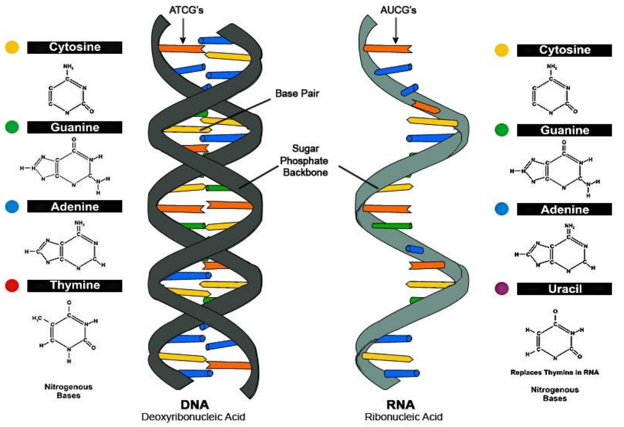

Fig 1: DNA vs RNA[4]

This is shown in Fig 1, along with composing molecular structures. So DNA(deoxyribonucleic acid) and RNA(ribonucleic acid) are **helix** (twisted) structures of two (non parallel) and one strands respectively. These segments (circle or semicircle structures in fig 1) are called **Genes** in some articles as the term gene is not correctly defined yet[5]. The combination of any of two molecules (ATCG, U) is termed as **nucleotides**[6]. Interestingly nucleotides has a specific pair structure containing two molecules (ATCG, U). Adenine only makes pair with Thymine and Guanine makes a pair with Cytosine[6] as shown in figure 2. The collection of these nucleotides makes a specific protein (fig 2).

The body of any living organism is composed of a **cell** and a portion of it is called **genome**. Genome in turn contains microorganisms called chromosomes.

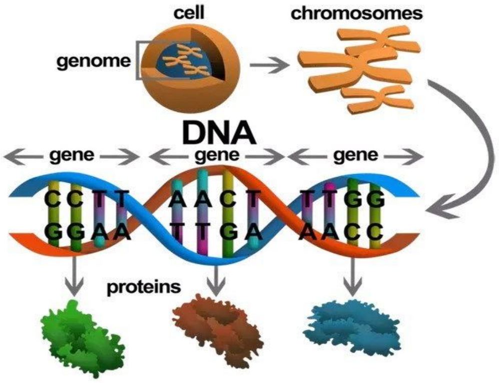

Fig 2. Difference between DNA and genes[7]

Each chromosome is composed of many DNA / RNA.

Prokaryotic are those organisms whose cells lack a nucleus eg. bacteria.

Eukaryotic are those organisms whose cells possess a membrane bound nucleus Fig3.

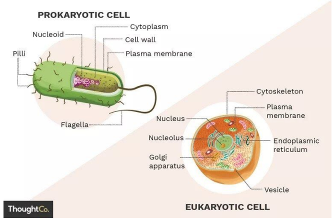

Fig 3 : Prokaryotic vs Eukaryotic

Gene transfer: The transfer of genetic information from parents to children is called **vertical gene transfers**. The transfer of genetic information by bacteria, virus, plasmids is termed as **horizontal gene** transfer[8]. This article has used horizontal gene transfer.

Microarray

The genome is first converted into a microarray which is a slide containing many genes usually between 500-20,000[6]. With advancement of technology researchers would be able to infer functions of new genes based on similarities in expression patterns of known genes[6]. The microarray technology explains the biological information flow in detail but here we will summarize for better understanding of the process which contains 4 steps[6].

1. Array fabrication

   This process involves preparation of glass slides which contains DNA sequences.

2. Sample preparation

   Only biological samples of interest are prepared and processed and others are isolated.

3. cDNA synthesis and labeling

   cDNA and mRNA are labelled with red or green fluorescent dye called Cy5 or Cy3 respectively.

4. Hybridization

   Each DNA and RNA is made of a collection of bases ATGCU. This is a process of binding two complementary DNA strands by base pairing. After hybridization the array is scanned at two wavelengths, once for Cy5 and another for Cy3 tagged sample. The fluorescent tags are excited by laser and fluorescence intensities are measured using a confocal microscope. For cDNA the raw data contains two gray scale images. Image processing tools are then used to extract numerical data from these images.

Summary

The article is about identifying the genome of origin. Since the data of many genomes are available, the task is feasible to analyse some genomes and compare it with already known ecoli, meningitis, influenza. Since we have known species for the comparison therefore the analysis could be done using classification technique of data mining.

The article observed that there exist non random patterns in the genomes of bacteria (Prokrates). Therefore they investigated the frequency of genome signatures (oligonucleotides) in different bacteria and found their origin. The Bayesian classifier was then applied to horizontal gene transfer and it was found that sodC and bioC genes were transferred from H Influenza to N Meningitis which validates the results.

This article used Naive Bayesian classifier to predict the origin in a sample 400 length sequences of 28 genomes with 85% accuracy.

Several studies related to this research were mentioned where proteins and other genes iin both prokaryotic and eukaryotic organisms have been studied.

In recent research different methods of classification have been mentioned like Dinucleotide composition and chaos game representation (CGR).

The article found that the performance depends on the sequence length. They trained classifier using specific genome sequences to distinguish between classes using 90% of data, the rest of the 10% data was used for testing purposes.

Critique

The strengths of the article:

- Comparatively a short and concise abstract is given which presents the overall article.
- The article starts with a very good and promising introduction with full reference support. It gives a comprehension and background of what is needed to be done and what has already been done in this area of biodiversity.
- Some methods for this kind of research have also been mentioned. A little more explanation and comparison of these research and analysis methods would be more useful though.
- The article concretely explains the number of base (400) taken from 28 eukaryotic organisms and the results in a round figure (85%)
- The article validates its results in horizontal gene transfer.
- Clearly explained that 90% dataset was used for training and the rest 10% for testing.

Some weaknesses are given below:

- Some claims like norandom sequence found in genome signatures are neither referenced nor researched. It would be nice to mention these papers who worked with norandom sequences.
- They analysed the 28 genomes which seems a very small dataset and could have been done on more data or in other words they could not clearly explained the data set.
- Their results are 85% correct which is a great achievement, However it would be more interesting to mention the reasons of failure and elaborate on it.
- The article mentions ‘overlapping motifs are extracted, .. the naïve Bayesian classifier uses the extracted motifs to predict their most probable genomic origin by comparing the frequencies of the extracted motifs with the motif frequencies of the different genomes’. However it has not been mentioned why are the motifs overlapping or what articles show this phenomena.
- Several of the abbreviations (GC…) are not explained. Therefore a table of these terms would help the reader to understand it more easily.

- The distinction between sequence and motifs are a bit confusing. The use of terminology genomic signatures, genomic sequence, frequency profiles, oligonucleotides, genomes for some kind of genes are confusing. It would be easy to call them patterns and the specific one as oligonucleotides pattern for example. But I have tried to explain all this in section ‘Molecular biology’ above in this article.
- There are several variations of a single term which is hard for data miners to distinguish at first glance. Therefore the model of research should be kept simple and in succinct language form.
- The article mentions two overlapping terms which could have been explained better, i.e ‘Genom of origin’ and ‘Horizontal gene transfer’
- 4 n possible motifs are mentioned but the article does not discuss about the number 4. May be because there are four bases (ATCG, U) ?
- Abstract says 28 bacteria while discussion sections says 25 bacterias.
- Visualizing the Genomic Signature Concept - a little more explanation required.

Conclusion

In the article under review, Bayesian algorithm has been used for identifying genome of origin.

This a good article with great results, however the documentation could have been done better.

A better explanation and visualization of the test bed would be highly appreciated and would enable to reproduce the research work.

More emphasis on introduction and results but method and experimentation lacks details.

References

[1] H. T. Coutts, “Critical Reviews of Journal Articles.” Sep-2014.

[2] Rickard Sandberg, Gösta Winberg, Carl-Ivar Bränden, 1 Alexander Kaske, 2 and Ingemar Ernberg, Joakim Cöster, “Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier,” *Cold Spring Harb. Lab. Press*.

[3] D. V. Nguyen, A. B. Arpat, N. Wang, and R. J. Carroll, “DNA Microarray Experiments: Biological and Technological Aspects,” *Biometrics*, vol. 58, no. 4, pp. 701–717, 2002.

[4] https://www.flickr.com/photos/102642344@N02/10082811755, “DNA vs RNA.” [Online]. Available: https://www.flickr.com/photos/102642344@N02/10082811755. [Accessed: 28-Dec-2019].

[5] T. Bohn, “Difference between DNA and Genes.” [Online]. Available: https://www.quora.com/What-is-the-difference-between-DNA-and-genes. [Accessed: 28-Dec-2019].

[6] V. Baladandayuthapani, S. Ray, and B. K. Mallick, “Bayesian Methods for DNA Microarray Data Analysis,” in *Bayesian Thinking*, vol. 25, D. K. Dey and C. R. Rao, Eds. Elsevier, 2005, pp. 713–742.

[7] P. Deb, “Difference between DNA and Genes,” *What-is-the-difference-between-DNA-and-genes*. [Online]. Available: https://www.quora.com/What-is-the-difference-between-DNA-and-genes. [Accessed: 28-Dec-2019].

[8] D. P. C. David Nanette J. Pazdernik, *Molecular Biology (Second Edition), 2013*, Second. .