A Hybrid Model for Pattern Discovery in HCV

Muhammad Naeem and Sohail Asghar

Department of Computer Science, Mohammad Ali Jinnah University Islamabad Pakistan
University Institute of IT PMAS-Arid Agriculture University, Rawalpindi Pakistan
E-mail: naeems.naeem@gmail.com, sohail.asg@gmail.com

Abstract

Identification of patterns in nucleotide sequence of HCV is useful in devising any strategy to foster fatal disease. Numerous mathematical models have been presented in literature. For the purpose of most coherent nucleotide sequence in Hepatitis C Virus (HCV), we have proposed a hybrid model to gene expression data mining which employs a series of unsupervised learning techniques to figure out the useful patterns within data. The proposed methodology involves data pre-processing, exploratory clustering followed by clique modeling with outcome of pattern identification and model visualization. We have evaluated our technique using data from Hepatitis C Virus nucleotide sequence of three regional countries Iran, Pakistan and Azerbaijan. The discovered patterns are related to structural similarity relation in bioinformatics data. The experimental results have indicated that out of large number of nucleotide bases, we can identify a small number of useful patterns suitable for further classification purposes. The proposed hybrid modeling approach delivers potential for diagnostic as well as virology applications. Methodology discussed in this study is an approach to converge huge knowledge related to DNA into a very small number of useful information. The described methodology is an approach for knowledge diversification to integration to deliver an insight for analysis in research networks related to virology and medicine.

Keywords: *Patterns Identification, HCV Genotype, DNA. Clustering, Clique detection*

1. Introduction

The advances in data growth and acquisition facilities have enabled the collaboration of many disciplines such as bioinformatics and data mining delivering a number of diverse research areas as described [1]. One of the underlying reasons triggering such collaboration is inability of trivial tools to analyze complex large amount of data, a serious and growing concern from many research communities. Clustering is an unsupervised technique to group large number of observations. A large number of clustering techniques have been developed during the past three decades; however there are situations when analysis of clusters is only a useful starting point for other applications. The possible reason behind this is that the outputs of the clusters deliver a blurred picture with overwhelming data. This results in the problem of extraction of useful patterns or labeling a cluster by selection from the overwhelming number of observations from a single cluster. Figure 1 and Figure 2 are showing the visual representation of observation points, the experiments performed in our study on HCV dataset of 56953 observations. In these clusters, retrieval of the most coherent structurally similar observations still remain an open problem due to the existence of many inherent uncertain factors. Such problems have motivated our study for an intensive effort to dig out the useful patterns.

A partitioning with too many clusters may lead to more complications making it hard to interpret and analyze, whereas a division with few clusters results in misleading the final decision due to the possible loss of information [2]. The problem of determining the number of clusters was termed as "the fundamental problem of cluster validity". In literature it has been concluded that there is no clustering algorithm capable of being universally used to deliver the solution of all grouping problems [2]. It is required that the techniques need to be designed with certain assumptions to favor some kind of biases in perspective of underlying data [2].

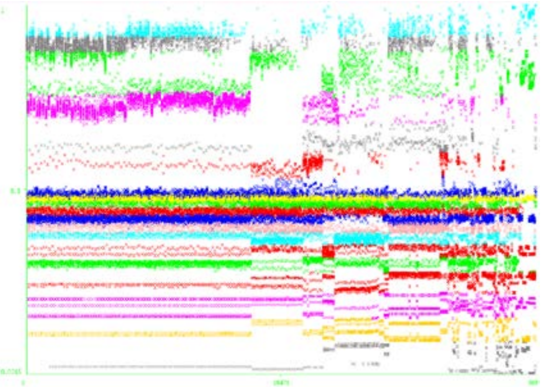


Figure 1. Visual Results of 20 Clusters

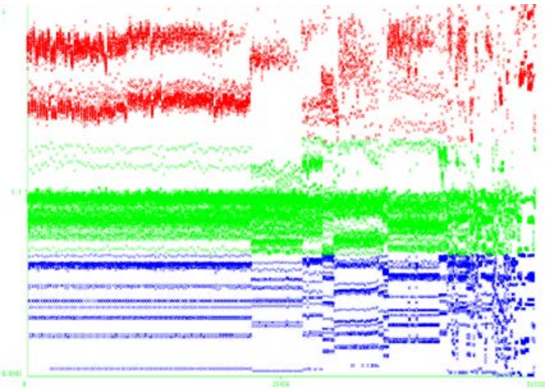


Figure 2. Visual Results of 3 Clusters

We know that nitrogen base sequences are represented by letters of four nucleotide bases {A, C, G, T}. The research on the phylogenetic taxonomy of HCV nucleotide sequences has pointed out that there are six HCV genotypes numbered 1–6 [3]. Every genotype is classified into many subtypes. It was shown that every genotype vary in its geographical distribution as well as in its mode of transmission [3]. Among all of these genotypes, the first and second type has been observed with the broadest distribution in the USA, Far East, Europe and partially in African territories. Recently it was reported that Genotype 1b is widespread in Rondônia State of Brazil [4]. Genotypes 3 and 4 both are mostly rich in considerable number of subtypes as reported [3]. Genotype 3 was also found in a broad distribution observed in Thailand, India, Europe, USA and to some extent in Japan. Genotype 4 has been distinguished as the dominant genotype amongst infected individuals from the Middle East, North Africa particularly Egypt where it was observed with a high population prevalence [5-7]. In fact it is reported that HCV Genotypes are mostly prevalent in Asian and African underdeveloped nations [8]**.** On the other side, genotypes 5, 6 indicate a limited geographical spread. The genotype 5 has been reported in South Africa while the genotype 6 was found in Macau, Vietnam and Hong Kong [3]. However, severity level is mostly reported in some of countries [9-11]. Investigation of the epidemiology of HCV infections plays a large role in the schemes of its prevention [12, 13].

Study of genotypes is clinically important because different genotypes are relevant to vaccine development, epidemiological questions as well as for the clinical management of HCV infection [14-16]. Motivated from the discussion about the blurred results from clustering analysis, we present specific challenges pertinent to clustering category, introducing a new hybrid approach. The broader genetic variability of the virus genome has motivated to raise the research question. The problem of identification of structurally similar pattern can be reduced to the determination of the cliques of hepatitis nucleotide sequence dataset using various distance measures in clustering?

2. Related Work

The nature of the experiment performed in our study is multi-stage. Importance of multi-stages system using data mining techniques in the domain of bioinformatics has already been exploited [17]. It was reported that two-stage system gives more Sensitivity, Specificity with high selectivity [17]. In literature, HCV has been targeted from various perspectives and directions. Some include divulging information on the statistical analysis [9, 10]. There are diverse applications for models of molecular sequences ranging from phylogenetics to genome annotation and comparative genomics [18]. Early models were claimed to have good results but only with assumption that nucleotide sequences evolve independently, however more realistic models were introduced with consideration of a wide range of different sources of context dependency [19]. There is a remarkable importance of similarity score for the conserved sequences like mitochondrial DNA sequences [20]. For the analysis of the DNA sequence, it was shown that the DNA data was signalized into one-dimensional or multi-dimensional discrete complex sequences [20]. A comparison of structural properties and sequence similarity was introduced by employing Euclidean distance similarity measure [21]. The parameters for structural similarity were information on stability, the minimum energy conformation and flexibility showing that combining structural and sequence similarity can improve promoter recall [21]. A relative similarity distance measure to analyze the local or global similarity of DNA sequences was introduced [22]. Given S,Q be the DNA sequence: The relative distance measure based on Lempl-Zive complexity was formally defined as:

$$rd(S,Q) = \frac{\sqrt{(c(SQ)-c(S))^2 + (c(QS)-c(Q))^2}}{c(SQ)+c(QS)} \text{ . } \quad (1)$$

It is not straightforward for researchers to dig out information directly from primary sequences resulting in the motivation for the presentation of new techniques to analyze them [23]. The importance of deducing knowledge out of DNA sequence was already reported in which classification of living birds of 1,058 species was carried out [24]. The notion of statistical significance of detected similarities to identify the local similarities between moderately small parts of sequences was discussed in detail [25]. It was shown that the probability Q_L of the longest similarity being of length L is: Q_L=Prob(longest similarity is of length L) = Prob($L_{max} \leq L$) – Prob(L_{max} – 1) = $(1-P_L+1)^{nL}-(1-P^L)^{nL-1}$. Whether the presence of clustered DNA sequences in promoters can predicts the expression properties of its corresponding mRNA [26]. The clustering sequences were found in the promoters of genes with a related function [26]. The integrated probability indicating the observed value (*m**) to be greater than expected frequency such that *m** is greater than the most probable value of *m:* $I = 2\sum_{m^*}^{m(\max)} P_m$

where I is integrated probability [26]. In bioinformatics domain cliques identification in graphical models was exploited [27]. Clique detection algorithm was applied to formulate the molecular surface database for functional sites [2]. The importance of cliques in clustering has been mentioned in literature various times [28]. This has motivated our work in this study to demonstrate that the detection of maximal fully connected subgraph from clusters can discover the patterns with features of high structural similarity.

3. Materials and Methods

Let (Ɔ, Ƨ, Ď, Ꞇ) be a finite space where Ɔ represents cliques, Ƨ shows clusters, Ď indicates DNA set and Ꞇ represents the sequence similarity. Let Ʊ be a nonempty set. A set valued mapping Ʊ: Ƨ → Ď^{Ƭ} can be called a focal set. Using this focal set, our target is to find Ↄ whose every member is a clique. We shall expand the set valued mapping function Ʊ : Ƨ → Ď^{Ƭ} such that Ď is a set of finite number of DNA sequences obtained from US official site for Hepatitis dataset (HCV Sequence Database, 2005).The symbol Ƭ denotes various distance measures used in clustering. For the convenience of our shrewd reader, we have divided the whole methodology into steps.

3.1. Preprocessing

We applied the CLUSTAL W. [29]. CLUSTAL W is commonly used for progressive multiple sequence alignment of divergent protein sequence [29]. In this heuristic, individual weights in a particular sequence alignment are assigned to each sequence. This ensures the down weight near duplicate sequences as well as up weight the most divergent sequences. In the next stage, at different alignment stages, amino acid substitution matrices are altered in accordance with the divergence of the sequences which are to be aligned. In the third stage, residue-specific gap penalties are processed. In the last stage locally reduced gap penalties enable the opening up of new gaps at these positions. As a result of application of sequence alignment algorithm we achieved different sets of sequence alignment dataset Ʊ. Each dataset is related to the specific strain of HCV genotype.

3.2. Cluster Analysis

The clustering algorithms partition data into a definite yet not specified number of subsets, groups or categories. In literature, cluster has been defined as the internal homogeneity while keeping the external separation [2]. Both of the similarity and the dissimilarity in clustering are to be examinable in a sense of inter-similarity and intra-similarity quantification. Here, we provide some simple mathematical descriptions of clustering used in our approach.

Given a set of features $F = \{F_1, F_2, \ldots. F_j \ldots\ldots F_n\}$, where $F_j = [f_{j1}, f_{j2}, \ldots. f_{jd}]^T \in Я^d$ while every measure f_{ji} is known to be an attribute or dimension variable. K numbers of partitions of F are declared:

- $C = \{C_1, C_2, \ldots C_k\}$. Where $k \leq n$. Every cluster $\vee_{Ci} = \acute{Ø}$, $i \leq k$.
- $\bigcup_{i=1}^{k} C_i = F;$
- $C_i \cap C_j = \phi, i, j \leq k, \forall i \neq j$

The underlying distance measuring parameters are numerous. In our case, we used Euclidean, Manhattan, Minkowski and Tchebyschev distance measure [30]. Mathematical detail of which can be described as:

- $Euclidean\,(A, B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2}$ (3)
- $Manhat\tan(A, B) = \sum_{i=1}^{n} | A_i - B_i |$ (4)
- $\text{Tchebychev}\,(\text{A, B}) = \max(| \text{A}_2 \text{ - B}_2 |,| \text{A}_1 \text{ - B}_1 |)$ (5)
- $Minkowski\,(A, B) = \lim_{p->\infty} (\sum_{i=1}^{n} | A_i - B_i |^p)^{\frac{1}{p}} = \max_{i=1}^{n} | A_i - B_i |$ (6)

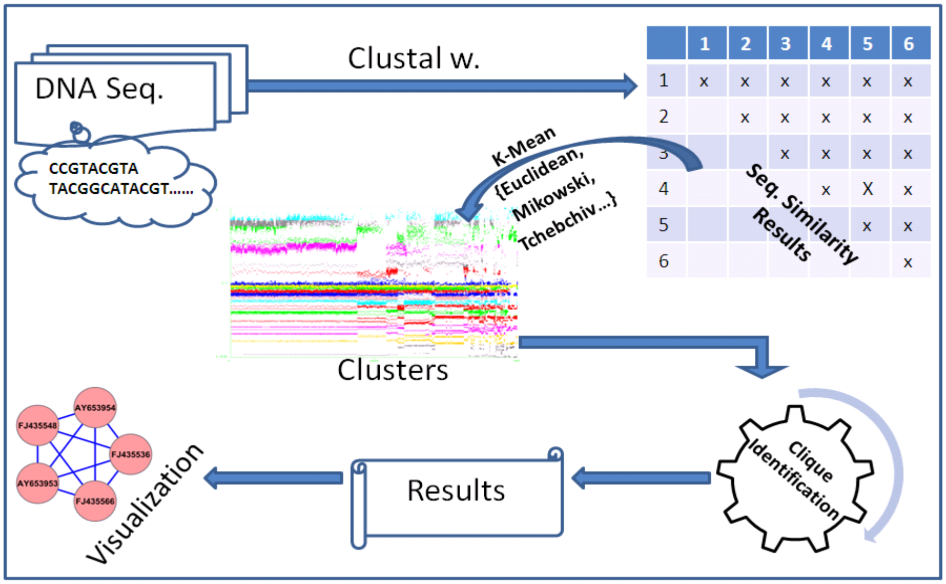
Where A and B denotes the member dimension of observation points for which clustering is under consideration. We provided a range of seeds from 3 to 20 for each distance measure described above.

3.3. Max. Clique Identification

Previously it was mentioned that Ɔ denotes cliques, now we shall explain it more in detail. Let Ɔ is a clique represented by a set $A = \{[(V_1, V_2),\ldots (V_1, V_n)]\ldots [(V_2, V_3),..,(V_2, V_n)],.. [(V_{n-1}, V_n)]\}$ Where V represents vertex such that for every pair of binaries (V_i, V_j) a relationship exists. We neglect the direction, *i.e.*, it doesn't matter whether V_i depends on V_j or vice versa. It denotes that both of vertices play a role of target and source for each other. We can map our hybrid modeling problem to clique problem such that: Given a simple graph $G = (V, E)$ and a numeral value $N \leq |V|$, is there a subset $S \subseteq V$ where $|S| = N$, the induced subgraph CLQ is the complete graph on S provided number of edges in the CQ = V_{cq} (V_{cq} -1)/2. Among all of CQ, We are interested in the CQ with highest value of N.

4. Experimental Setup and Data Preparation

This section includes details of data preparation process and experimental steps. We obtained our sample DNA files from HCV Databases which contains a very large number of dataset belonging to various countries [31]. We restricted ourselves to only three countries Iran, Pakistan and Azerbaijan. In our experiment, we have considered the strains 1, 1a, 1b, 3a. CLUSTAL W. sequence alignment software was used for the local and global similarity calculation reported in this study. Although a lot of online tools are available to provide the results for CLUSTAL W results [29]. However, to deliver a unified single framework used in our methodology, all of the components were written in house to ensure availability of data pre-processing and analysis functionality integrated into one environment. From the literature review, it was highlighted that large amount of genetic material data as well as the complexity of biological networks has escalated the demanding situation of apprehending and interpreting the hidden information lurking in the data. A preliminary step towards addressing such challenge is the use of clustering techniques. The clustering can be considered an essential step in the data mining process to discover natural structures while identifying useful patterns in the underlying data. The Figure 3 illustrates the steps performed in the methodology.

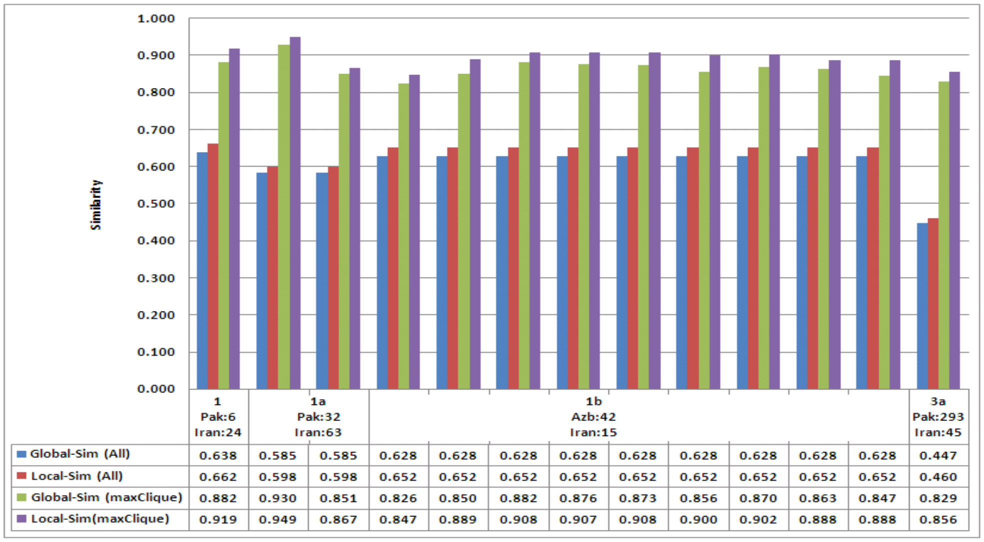


	1	2	3	4	5	6
1	x	x	x	x	x	x
2		x	x	x	x	x
3			x	x	x	x
4				x	X	x
5					x	x
6						x

Figure 3. Proposed Hybrid Model

5. Result and Discussion

We can describe the crux of this study such that all cliques resulted from the clustering runs resulted in a pattern of the most informative genes. In this section, we shall give its detail. In computational biology, sequence comparison is an important procedure which is directed to disclose similarity or dissimilarity relationships between molecular sequences. Nucleotide sequences are multiple-aligned using Clustal W. [29] which resulted in a matrix of HCV strains. K-mean clustering using four distance measures Euclidean, Manhattan, Minkowski and Tchebyschev distance measure were applied to identify the inherent group characteristics of the data [30]. A heuristic applied to identify the most informative groups of DNA without any associated threshold measure for clustering. Data visualization was applied to better elucidate and highlight the overall patterns of identified DNA sequences.

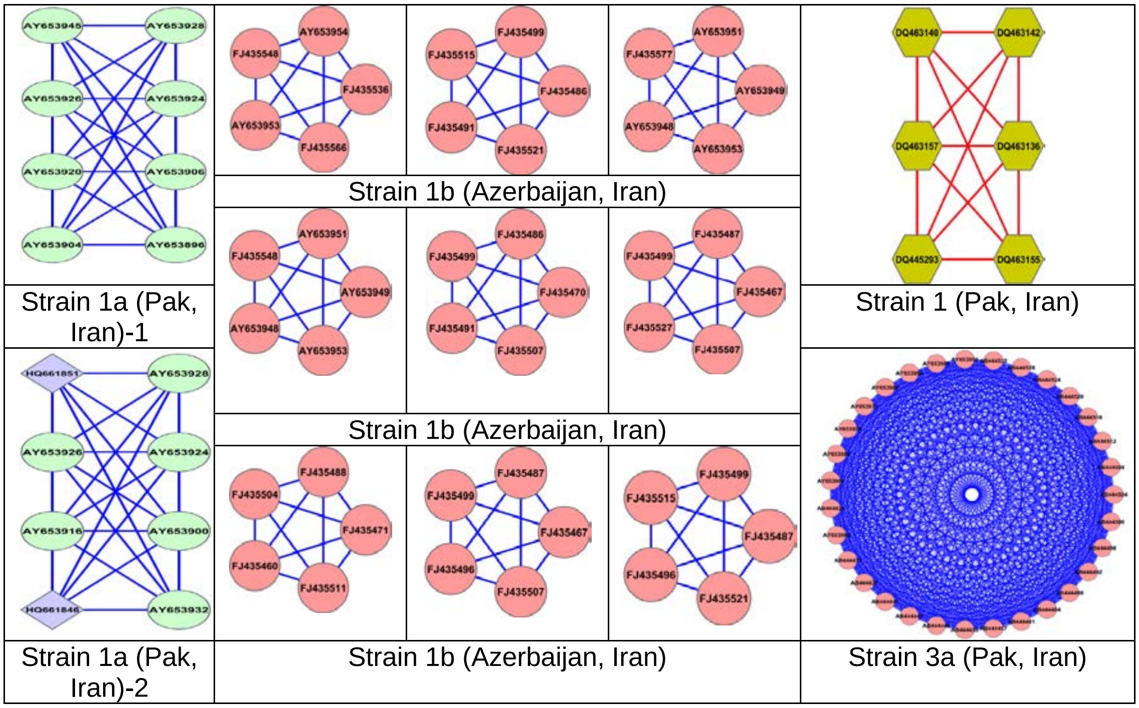


	1 Pak:6 Iran:24	1a Pak:32 Iran:63		1b Azb:42 Iran:15									3a Pak:293 Iran:45
🟦 Global-Sim (All)	0.638	0.585	0.585	0.628	0.628	0.628	0.628	0.628	0.628	0.628	0.628	0.628	0.447
🟥 Local-Sim (All)	0.662	0.598	0.598	0.652	0.652	0.652	0.652	0.652	0.652	0.652	0.652	0.652	0.460
🟩 Global-Sim (maxClique)	0.882	0.930	0.851	0.826	0.850	0.882	0.876	0.873	0.856	0.870	0.863	0.847	0.829
🟪 Local-Sim(maxClique)	0.919	0.949	0.867	0.847	0.889	0.908	0.907	0.908	0.900	0.902	0.888	0.888	0.856

Figure 4. Comparison of Sequence Similarity of Cliques to Whole DNA Dataset

Figure 4 shows the identified pattern of DNA genes in each of the four strains of HCV dataset obtained from the clique modeling followed by clustering runs. 18 clustering runs for each of the dataset were experimented with seeds range from 3 to 20. Interesting facts we observed from this study is that all of the max-cliques were of high local and global sequence similarity. As a result of each experiment we obtained a large number of cliques from each cluster run. However, the max clique nodes were those DNA sequences which have high sequence similarity to each other. This validated the fact that the sequence of operations performed in our study has the ability to detect a peculiar pattern in all of the strains. We shall explain it by the statistical facts mentioned in the Figure 4. The HCV strains 3a consisted of total 338 DNA sequence, 293 DNA from Pakistan and 45 DNA from Iran. Genotype 3a, 1a, 1b and 2 are most frequent in descending order in Isfahan province of Iran [32]. All DNA sequences from this set were cross examined against each other resulting in 56953 comparisons. Average global and local similarity was of 0.447 and 0.460. K-mean clustering was applied to these 56953 comparisons with seeds range of 3 to 20. This results in 207 clusters. We dig out cliques from these 207 clusters. The max. Cliques found among all of the cliques consisted of 31 nodes. The average global and local similarity for these nodes was 0.829 and 0.856 respectively. This strong similarity corroborates a previous research in which it was stated that over expression of Core gene of HCV 3a genotype indicate stronger effect in regulating RNA and protein levels of Cox-2, iNOS, VEGF, p-Akt in comparison to HCV-1a Core in hepatocellular carcinoma cell line Huh-7 accompanied by heightened PGE2 release and cell proliferation [33]. These procedures were repeated on other three strains 1, 1a and 1b. It is evident from the Figure-4 and Figure-5 that in each of the max cliques, the average similarity was much higher than that of the entire DNA in the dataset of each strain. From the preliminary result obtained, we can conclude that the proposed model has the potential to mine the interesting patterns from the dataset.

The Figure-5 is the visual representation of the cliques, we obtained from the experiment. It provides users with results which can simplify the future analysis. In the Figure strain 1a (Pak, Iran)-2 is showing nodes of two shapes. The diamond nodes [HQ661846 and HQ661851] are both from Pakistan whereas all of the rest belong to the dataset from Iran. One other important evidence we observed was that same cliques were retrieved from the cluster using all four distance measures mentioned in the proposed methodology. However two cliques of strain 1b were not retrieved only from manhattans and Murkowski distance measure.



Strain 1b (Azerbaijan, Iran)

Strain 1a (Pak, Iran)-1

Strain 1 (Pak, Iran)

Strain 1b (Azerbaijan, Iran)

Strain 1a (Pak, Iran)-2

Strain 1b (Azerbaijan, Iran)

Strain 3a (Pak, Iran)

Figure 5. Visual Representation of max. Cliques obtained from various HCV Strains

6. Conclusions

As the genome project has been completed, a lot of genomic sequences have been made available. This motivates the research community to analyze these sequences, and find out useful patterns with respect to their sequence structure. Under this motivation, this paper describes an approach to mine globally significant maximum clique patterns out of large amount of DNA gene data. The objective was to search for meaningful patterns related to identify highly correlated structurally similar sequences and their discrimination characteristics. One characteristic revealed in our study was that the DNA genes which have high global or local sequence similarity, they tend to form a maximum clique when clustered.

It can be concluded that the knowledge discovery experiments in this study leads to classification models with most informative genes. We have tested our method on HCV database initially on three regional countries [31]. The methodology has resulted in identification of a small number of the most coherent genetic material out of a total of 520 DNA sequences. In knowledge discovery system, success is measured in terms of achievement of discrimination of versatile groups from large dataset. We believe that the approach described in this study has potential for achieving this objective to a reasonable extent and can be further used in disease classification, diagnostic and virology applications. In the Middle East, Networking for overcoming viral hepatitis has been established [20]. The research carried out in this study may be helpful in analysis of such networks to foster HCV.

We can identify the future work in two dimensions. The first dimension is related to perform exhaustive experimentation on large dataset publicly available at HCV database [31]. The second dimension is related to the complete evaluation of the result set and computation of the recall values of the information gain. The overall result will be evaluated by the domain expert in order to validate the results and generalize them.

References

[1] L. J. Lancashire, C. Lemetre and G. R. Ball, “An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies”, Brief Bioinform, vol. 10, no. 3, **(2009)**, pp. 315-329.

[2] R. Xu and D. Wunsch, “Survey of Clustering Algorithm”, IEEE Transactions on Neural Networks, vol. 16, **(2005)**, pp. 645-678.

[3] B. Robertson, G. Myers and C. Howard, “Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization”, Arch Virol., vol. 143, **(1998)**, pp. 2493-2503.

[4] D. S. Vieira, M. V. Alvarado-Mora, L. Botelho, F. J. Carrilho, J. R. R. Pinho and J. M. Salcedo, “Distribution of hepatitis c virus (hcv) genotypes in patients with chronic infection from Rondônia, Brazil”, Virology Journal, vol. 8, **(2011)**, pp. 165.

[5] S. Akhtar and T. Moatter, “Intra-household clustering of hepatitis C virus infection in Karachi, Pakistan”, Trans R Soc Trop Med Hyg, vol. 98, **(2004)**, pp. 535-539.

[6] F. R. Ponziani, A. Milani and A. Gasbarrini, “Treatment of recurrent genotype 4 hepatitis C after liver transplantation: early virological response is predictive of sustained virological response”, An AISF RECOLT-C group study. Ann Hepatol., vol. 11, **(2012)**, pp. 338-342.

[7] E-D. M. Abdalla, A. E-D. M.S. Hosny and Y. I. M. Shamikh, “Efficacy of Hepatitis B Virus Vaccination on the Incidence of Hepatitis C Virus Infection”, Australian Journal of Basic and Applied Sciences, vol. 6, no. 3, **(2012)**, pp. 849-851.

[8] W. Aman, S. Mousa, G. Shiha and S. A. Mousa, “Current status and future directions in the management of chronic hepatitis C”, Virology Journal, vol. 9, **(2012)**, pp. 57.

[9] N. Méndez-Sánchez, E. García-Villegas, B. Merino-Zeferino, S. Ochoa-Cruz, A. R. Villa, H. Madrigal and R. A. Kobashi-Margáin, “Liver diseases in Mexico and their associated mortality trends from 2000 to 2007: a retrospective study of the nation and the federal states”, Ann of Hepatol, vol. 9, **(2010)**, pp. 428-438.

[10] N. Méndez-Sánchez, A. R. Villa, G. Vazquez-Elizondo, G. Ponicano-Rodriguez and M. Uribe, “Mortality trends for liver cancer in Mexico from 2000 to 2006”, Ann of Hepatol., vol. 7, **(2008)**, pp. 226-229.

[11] N. C. Chávez-Tapia, T. Barrientos-Gutiérrez, C. M. Guerrero-López, J. J. Santiago-Hernández, N. Méndez-Sánchez and M. Uribe, “Increased mortality from acute liver failure in Mexico”, Ann Hepatol., vol. 11, **(2012)**, pp. 257-262.

[12] D. B. Strader, T. Wright, D. L. Thomas and L. B. Seeff, “Diagnosis, management, and treatment of hepatitis C”, Hepatology, vol. 39, **(2004)**, pp. 1147-1171.

[13] M. Sherman, S. Shafran, K. Burak, K. Doucette, W. Wong, N. Girgrah and E. Yoshida, “Management of chronic hepatitis C: consensus guidelines”, Can J Gastroenterol., vol. 21(Suppl C), **(2007)**, pp. 25C-34C.

[14] M. Liew, M. Erali, S. Page, D. Hillyard and C. Wittwer, “Hepatitis C genotyping by denaturing high-performance liquid chromatography”, J Clin Microbiol., vol. 42, **(2004)**, pp. 158-163.

[15] N. N. Zein, “Clinical significance of hepatitis C virus genotypes”, Clin Microbiol Rev., vol. 13, **(2000)**, pp. 223-235.

[16] N. N. Zein and D. H. Persing, “Hepatitis C genotypes: current trends and future implications”, Mayo Clin Proc., vol. 71, **(1996)**, pp. 458-462.

[17] Z. H. `Inan and M. Kuntalp, "A study on fuzzy C-means clustering based systems in automatic spike detection", Computers in Biology and Medicine, vol. 37, **(2007)**, pp. 1160-1166.

[18] W. Delport, K. Scheffler and C. Seoighe, "Models of coding sequence evolution", Brief in Bioinfo, vol. 10, **(2008)**, pp. 97-109.

[19] A. Siepel and D. Haussler, "Phylogenetic estimation of contextdependent substitution rates by maximum likelihood", J Mol Biol Evol, vol. 21, **(2004)**, pp. 468-88.

[20] J. Zhang, R. Wang and F. Bai, "A Quasi-MQ EMD method for similarity analysis of DNA sequences", Applied Mathematics Letters, vol. 24, **(2011)**, pp. 2052-58.

[21] E. J. Gardiner, C. A. Hunter, X. J. Lu and P. Willett, "A Structural Similarity Analysis of Double-helical DNA", J. Mol. Biol., vol. 343, **(2004)**, pp. 879-889.

[22] L. Na and W. Tian-ming, "A relative similarity measure for the similarity analysis of DNA sequences", Chemical Physics Letters, vol. 408, **(2005)**, pp. 307-311.

[23] G. Ying and W. Tian-ming, "A new method to analyze the similarity of the DNA sequences", Journal of Molecular Structure: THEOCHEM, vol. 853, **(2008)**, pp. 62-67.

[24] G. S. Charles, E. A. Jon and L. M. Burt, "A Classification of the living birds of the world based on DNA-DNA hybridization studies", American Ornithologists' Union, vol. 105, **(1998)**, pp. 409-423.

[25] L. D. Brooks, B. S. Weira and H. E. Schaffer, "The Probabilities of Similarities in DNA Sequence Comparisons", GENOMICS, vol. 3, **(1988)**, pp. 207-216.

[26] P. C. FitzGerald1, A. Shlyakhtenko, A. A. Mir and C. Vinson, "Clustering of DNA sequences in human promoters", Genome Res., vol. 14, **(2004)**, pp. 1562-1574.

[27] P. Keunwan and K. Dongsup, "A Method to Detect Important Residues Using Protein Binding Site Comparison", Genome Informatics, vol. 17, **(2006)**, pp. 216-25.

[28] M. Pavan and M. Pelillo, "Dominant Sets and Pairwise Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, **(2007)**, pp. 167-172.

[29] J. D. Thompson, D. G. Higgins and T. G. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", Nucleic Acids Res, vol. 22, **(1994)**, pp. 4673-80.

[30] E. Deza and M. M. Deza, "Encyclopedia of Distances", Springer, DOI: 10.1007/978-3-642-00234-2_1, **(2009)**, pp. 94.

[31] HCV Sequence Database, wesite **(2005)** [cited 2012 September 25]; Available from: URL: http://hcv.lanl.gov/content/index.

[32] S. H. Zarkesh-Esfahani, M. T. Kardi and M. Edalati, "Hepatitis C virus genotype frequency in Isfahan province of Iran: a descriptive cross-sectional study", Virology Journal, vol. 7, **(2010)**, pp. 69.

[33] S. Jahan, S. Khaliq, B. Ijaz, W. Ahmad and S. Hassan, "Role of HCV Core gene of genotype 1a and 3a and host gene Cox-2 in HCV-induced pathogenesis", Virology Journal, vol. 8, **(2011)**, pp. 155.