Journal of Data Mining in Genomics & Proteomics

Garcia et al., J Data Mining Genomics Proteomics 2012, 3:3
DOI: 10.4172/2153-0602.1000120

Research Article Open Access

Interactive Network Exploration in the KDD Process, Contributions in the Study of Population Variability of a Corn *Fijivirus*

Mario Alejandro Garcia[1]*, Maria de la Paz Gimenez Pecci[2], Juan Bautista Cabral[1], Adrian Nieto Castillo[1] and Irma Graciela Laguna[2,3]

[1]National Technological University Cordoba Regional Faculty (UTN FRC), Cordoba, Argentina
[2]National Institute of Agricultural Technology (INTA), Cordoba, Argentina
[3]National Research Council Scientific and Technical (CONICET), Córdoba, Argentina

Abstract

The genetic variability of individuals of the same species can be studied through networks that represent the genetic distances between them. We studied the case of *Mal de Rio Cuarto virus* (MRCV), defining distance measures between genome profiles of different individuals and creating a network of haplotypes. Topological properties of the network were analyzed and this was examined in two dimensions, forming space-time environments. The examination led to the observation that, in the first crop years tested, the number of haplotypes and the distance between them was greater than in subsequent crops. A variability indicator was calculated for each environment and compared with its expected value, confirming the observation made during the examination and concluding that virus variability decreased after an epidemic occurred during the crop year 1996-97. An analysis of variability of MRCV through haplotype networks is presented. We propose the use of this tool, which is unusual in KDD processes, bringing a new approach that affects the concepts of knowledge representation, structured data modeling, visualization, exploration and interactive discovery. The main contribution of this case to the KDD process is the proposal of interactive exploration of networks, which turned out to be intuitive and easy to apply for analysis.

Keywords: KDD; Haplotype networks; *Mal de Rio Cuarto virus*; Data mining

Introduction

Mal de Rio Cuarto is a corn disease that is endemic in the Rio Cuarto Department, Cordoba, Argentina. Severe epidemics occur periodically, due to which the vectors, alternate hosts, environmental and climatic conditions that lead to epidemics are being studied, tolerance is sought in high-yielding hybrids and a "typical" virus from the endemic area has been sequenced in order to obtain transgenic plants that express resistance [1]. There is little knowledge of the role of the pathogen in such epidemics.

The causative agent *Mal de Rio Cuarto virus* (MRCV) has gradually been recorded in new areas, plant hosts and vectors, suggesting its ability to produce adaptive leaps and subsequent adjustment to new environments [1]. To understand the epidemiology of the disease requires knowledge of the variability and genetic structure of the causative virus populations [2].

To study genetic diversity, there is a database containing the results of electrophoretic analysis of the ten viral dsRNA genome segments, characteristic of the *Fijivirus* genus, performed on samples from 8 host species, in 13 locations, over 13 seasons [3].

The electrophoretic profile of MRCV is represented by a binary string of length 18, which contains the ten known segments of the virus, some of which can be placed in different positions [4], and two extra genomic bands [5]. A "0" in an electrophoretic profile position indicates the absence of the corresponding virus dsRNA segment/band, while a "1" indicates its presence [2]. The different electrophoretic profiles will be called haplotypes (haploid genotypes).

Distance between profiles

Developing profiles of entities based on a set of attributes and then comparing these entities through their profiles is a common and often effective paradigm in machine learning. For these profiles, often represented as vectors of binary or real numbers, the comparison is done by measuring "distances" between each pair of profiles. The effectiveness of the learning depends on an accurate measurement of distances. In general, given a set A of N attributes, $A = \{a_i | i = 1\ldots, N\}$, the profile of entity x in A gives an N vector of real numbers. If all the attributes of A can take only the discrete values 0 and 1, then $p(x) \rightarrow \Re^N$ will be a binary profile. The distance between a pair of profiles x and y is a function $D(x,y) \rightarrow \Re$. The Hamming distance [6] is the simplest, and also one of the most commonly used, measures of distance for binary profiles; this is the simple sum of the differences between each individual attribute:

$$D(x,y) = \sum_{i}^{n} d(i)$$

Where

$$d(i) = | x_i - y_i |$$

While Hamming distance and Euclidean distance are the commonly adopted measures of profile similarity, both of them imply an underlying assumption that the attributes are independent and contribute equally in describing the profile. Therefore, the distance between two profiles is simply a sum of distance (i.e., difference) between them at each attribute. These measures become inappropriate when the attributes are not equally contributing, or not independent, but rather correlated to one another. This is often the case in the real-world biological

***Corresponding author:** Mario Alejandro Garcia, National Technological University Cordoba Regional Faculty (UTN FRC), Cordoba, Argentina, E-mail: mgarcia@sistemas.frc.utn.edu.ar

Received May 03, 2012; **Accepted** September 28, 2012; **Published** October 07, 2012

Citation: Garcia MA, Gimenez Pecci MP, Cabral JB, Castillo AN, Laguna IG (2012) Interactive Network Exploration in the KDD Process, Contributions in the Study of Population Variability of a Corn *Fijivirus*. J Data Mining Genomics Proteomics 3:120. doi:10.4172/2153-0602.1000120

Copyright: © 2012 Garcia MA, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

problems [7]. Due to this, Hamming distance was adopted only as a basic calculation of distance and was adapted through knowledge about the domain of the problem.

Haplotype networks

Genetic distances are usually represented graphically by a dendrogram, which is a tree of profiles calculated from the genetic distance between species or between different individuals of the same species. In data mining, this diagram is related with the algorithms of hierarchical clustering.

The option of representing the distances between profiles with a network structure instead of a tree offers the following advantages:

a) A network can represent all the distances between haplotypes and not only some of them, as occurs with a tree. Haplotype networks can represent cycles and multiple connections between profiles [8].

b) Relations between individuals of the same species should not always be represented by a tree, because they are not hierarchical. This is the result of sexual reproduction, of smaller numbers of relatively recent mutations and frequently of recombination [9].

c) Networks offer a new way to categorize systems from different sources under a single framework. This approach has uncovered unexpected similarities between the organization of different complex systems (social, economic, technological, physical and biological), indicating that the networks they describe are governed by general principles and mechanisms of organization [10].

Objective

The objective of this KDD process is to analyze the variability of the virus over time and of the geographical distribution of the samples found in the database available.

Materials and Methods

Database

The database consists of a fact table with 25 attributes, among which is the geographical area where the sample of the plant infected by MRCV was collected, the crop year and the segments of the electrophoretic profile recorded. 21 different electrophoretic profiles of MRCV were determined, which are shown in Table 1. The most common haplotype is number 9, which is present in 56% of the samples [4].

The geographic attribute used was Region from endemic, which divides the area where the virus was found in 7 regions according to their location relative to the endemic area. These are "Endemic", "North of endemic", "High Valley", "Core", "South of endemic", "Northeast of endemic" and "East of endemic" [2]. The attribute used to analyze the time dimension was Crop year, whose values vary from "1989/1990" to "2003/2004".

Distance calculation

A matrix of distances $D(d_{ij})$ was calculated in which d_{ij} is the distance between haplotype i and haplotype j. D is symmetrical and its main diagonal is 0.

Each d_{ij} is equal to the Hamming distance between haplotypes i and j, with three exceptions:

Exception 1: The subscript letter used to distinguish electrophoretic migration positions identities the same genomic segment (B3, B9 or B10) with different electrophoretic distance (in a different position in the gel) e.g.: $B3_a$ and $B3_b$ are the same segment which in one case migrated to position a, and in the other migrated to position b. Under this rule, the migration of a band, although it is seen in the haplotype table as two changes because it is no longer in one position and begins to exist in another, is considered a single change.

Exception 2: The extragenomic band E10 could by produced by the restriction of the genomic band $B3_a$ [5], which accounts for its appearance together with band $B3_b$. (E10 = $B3_a$–$B3_b$), so if there is a change in the band E10 and in band B3, it is considered a single mutation.

Exception 3: The extragenomic band E5 may be the result of the restriction of genomic band B5. As in the previous case, when there is a change in both bands, it is considered to be due to the same mutation and only one distance unit is added.

Finally,

$$d_{ij} = dB3_{ij} + dB5_{ij} + dB8_{ij} + dB9_{ij} + dB10_{ij} + dBE5_{ij} + dBE10_{ij} \quad (1)$$

Where, with the Hamming distance criterion:

$$dB5_{ij} = \left| dB5_i - dB5_j \right|$$

$$dB8_{ij} = \left| dB8_i - dB8_j \right|$$

According to Exception 1:

$$dB3_{ij} = \frac{1}{2}\left(\left|B3_{ai} - B3_{aj}\right| + \left|B3_{bi} - B3_{bj}\right| + \left|B3_{ai} - B3_{aj} + B3_{bi} - B3_{bj}\right|\right)$$

$$dB9_{ij} = \frac{1}{2}\left(\left|B9_{ai} - B9_{aj}\right| + \left|B9_{bi} - B9_{bj}\right| + \left|B9_{ci} - B9_{cj}\right| + \left|B9_{ai} - B9_{aj} + B9_{bi} - B9_{bj} + B9_{ci} - B9_{cj}\right|\right)$$

$$dB10_{ij} = \frac{1}{2}\left(\left|B10_{ai} - B10_{aj}\right| + \left|B10_{bi} - B10_{bj}\right| + \left|B10_{ai} - B10_{aj} + B10_{bi} - B10_{bj}\right|\right)$$

Haplotype	$B3_a$	$B3_b$	B5	B8	$B9_a$	$B9_b$	$B9_c$	$B10_a$	$B10_b$	BE5	BE10
1	1	0	1	1	1	1	0	0	0	0	0
2	1	0	1	1	1	0	1	1	0	0	0
3	1	0	1	1	1	0	0	1	1	0	0
4	1	0	1	1	1	0	0	1	0	0	0
5	1	0	1	1	0	1	0	1	0	0	0
6	1	0	1	1	0	1	0	0	0	0	0
7	1	0	1	1	0	0	1	1	0	1	1
8	1	0	1	1	0	0	1	1	0	1	0
9	1	0	1	1	0	0	1	1	0	0	0
10	1	0	1	1	0	0	1	0	1	0	0
11	1	0	1	1	0	0	0	1	1	0	0
12	0	1	1	1	0	1	0	1	0	0	0
13	0	1	1	1	0	0	1	1	0	1	1
14	0	1	1	1	0	0	1	1	0	0	0
15	1	0	1	1	0	0	1	0	0	0	1
16	1	0	1	1	0	0	1	0	0	0	0
17	1	0	0	1	1	0	1	1	0	0	0
18	1	0	0	0	0	1	0	1	0	0	0
19	0	1	1	1	1	1	0	1	1	0	0
20	0	1	1	1	0	1	0	1	0	1	1
21	0	1	1	1	0	0	1	0	0	0	0

Table 1: Haplotypes of *Mal de Rio Cuarto virus* (MRCV) determined by the presence or absence of the bands formed by the genomic segments and two extra-genomic bands BE5 and BE10 recorded in the electrophoretic profiles determined by electrophoretic migration of viral dsRNA present in samples from naturally infected grasses.

According to Exception 2:

$$dBE10_{ij} = \left|BE10_i - BE10_j\right|\left(1-\left(\left|B3_{ai} - B3_{aj}\right| OR \left|B3_{bi} - B3_{bj}\right|\right)\right)$$

According to Exception 3:

$$dBE5_{ij} = \left|BE5_i - BE5_j\right|(1-\left|BE5_i - BE5_j\right|)$$

Haplotype networks

The network graph was formed by nodes representing the haplotypes and arcs that connect the haplotypes that are at distance = 1. Haplotypes that are at distance > 1 from any other profile are plotted connected to a virtual haplotype, identified with dotted line to the nearest existing haplotype. The most frequent haplotype was located in the center of the graph and the others on concentric circles whose level indicates the distance to the most frequent haplotype.

On the resulting network, the following 4 properties were calculated:

Diameter: Is the longest distance between any two nodes in the graph.

Average distance: Average distance between all nodes in the graph.

Distribution of degree of connectivity: The degree of connectivity k_i of a node i is the number of arcs that are incident on the node. In terms of the matrix of distances:

$$k_i = \sum_{j=1}^{N} d_{ij} \mid d_{ij} = 1$$

The list of degrees of connectivity of the nodes is called the degree sequence. The most basic topological characterization of a graph is obtained in terms of the distribution of degree of connectivity $P(k)$, defined as the probability that a node chosen at random will have degree k.

Clustering coefficient: Clustering coefficient is a measure of the number of nodes that form triangular graphs with their adjacent nodes [11]. It was introduced by Watts and Strogatz in [12] as a typical property of social networks, in which two people with a mutual friend have a high probability of knowing each other [13]. This property is important in the classification of networks. The clustering coefficient Cc_i of a node i is the relation between the number of existing arcs (not incident on i) in the sub-graph of the nodes connected to and the maximum number of possible connections [11].

$$Cc_i = \frac{2c_i}{k_i(k_i - 1)}$$

where c_i is the number of arcs existing between the neighbors of i and k_i is the degree of connectivity of i.

Lastly, the clustering coefficient of the network will be the average of the clustering coefficient of the nodes composing it [11].

$$CC = \frac{1}{n}\sum_{i=1}^{n} Cc_i$$

Exploring the network by environments

The proposed methodology for the exploration of network environments has two stages.

The first stage is to create a graphical representation of the network from the previously calculated distances between haplotypes. This graph should show all the haplotypes of the databases.

The second is the exploration stage. This defines on what dimensions and with how much detail the network will be analyzed. The dimensions are the attributes of the database, which in the case of MRCV refer to the geographic location, the crop year and the host species. The selection of attributes and levels of detail create a space for exploration, where each point of the space is called environment. For each environment, the entire network of haplotypes is visualized, but only the haplotypes that were found in that environment are highlighted on it. At each step of exploration, the analyst decides which direction to move and whether to increase or decrease the level of detail of any of the dimensions (Figure 1). This dynamic, interactive exploration helps the analyst in understanding the behavior of the virus and in generating new hypotheses.

Variability indicators

To associate values to the variability of each environment, the indicator sum of distances between haplotypes (SDH) was defined. This indicator is explained in Results section.

A valid alternative for this step is to perform an analysis of molecular variance (AMOVA), adapted from the analysis of variance (ANOVA) to molecular type data. AMOVA produces estimates of variance components and F-statistics analogs, reflecting the correlation of haplotypic diversity at different levels of hierarchical subdivision [14]. The method proposed in this paper facilitates the implementation of AMOVA because it uses common elements such as the distance matrix and the network of haplotypes.

Unlike AMOVA, the chosen indicator SDH is affected, not by the sampling frequency of each haplotype in the environment analyzed, but by the presence/absence of the haplotypes.

Results

Creation and analysis of the network

Table 2 shows the distances between haplotypes obtained.

Haplotype 9 was located in the center of the graph as it is the most frequent. It is useful to represent it in this way because it can be assumed

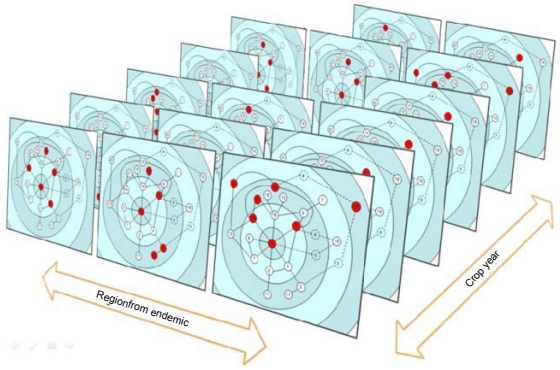

Figure 1: Exploration by environments of the haplotype networks of *Mal de Rio Cuarto virus* (MRCV) using the dimensions Region from endemic (space) and Crop year (time). Each plate represents an environment. The analyst visualizes only one environment at a time and chooses which direction of the environment space to move in order to change visualization. Each plate represents the entire network of haplotypes and the haplotypes existing in that environment are highlighted.

Hapl.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0	2	3	2	2	1	5	4	3	3	4	3	5	4	3	2	3	4	3	4	3
2	2	0	2	1	2	3	3	2	1	2	3	3	3	2	3	2	1	4	3	4	3
3	3	2	0	1	2	3	4	3	2	2	1	3	4	3	4	3	3	4	2	4	4
4	2	1	1	0	1	2	3	2	1	2	2	2	3	2	3	2	2	3	3	3	3
5	2	2	2	1	0	1	3	2	1	2	2	1	3	2	3	2	3	2	3	2	3
6	1	3	3	2	1	0	4	3	2	2	3	2	4	3	2	1	4	3	4	3	2
7	5	3	4	3	3	4	0	1	2	3	4	3	1	2	2	3	3	4	5	2	3
8	4	2	3	2	2	3	1	0	1	2	3	3	1	2	3	2	2	3	5	2	3
9	3	1	2	1	1	2	2	1	0	1	2	2	2	1	2	1	2	3	4	3	2
10	3	2	2	2	2	2	3	2	1	0	2	3	3	2	2	1	3	4	4	4	2
11	4	3	1	2	2	3	4	3	2	2	0	3	4	3	4	3	4	4	3	4	4
12	3	3	3	2	1	2	3	3	2	3	3	0	3	1	3	3	4	3	2	2	2
13	5	3	4	3	3	4	1	1	2	3	4	3	0	2	3	3	3	4	5	1	3
14	4	2	3	2	2	3	2	2	1	2	3	1	2	0	2	2	3	4	3	3	1
15	3	3	4	3	3	2	2	3	2	2	4	3	3	2	0	1	4	5	5	4	1
16	2	2	3	2	2	1	3	2	1	1	3	3	3	2	1	0	3	4	5	4	1
17	3	1	3	2	3	4	3	2	2	3	4	4	3	3	4	3	0	3	4	4	4
18	4	4	4	3	2	3	4	3	3	4	4	3	4	4	5	4	3	0	5	3	5
19	3	3	2	3	3	4	5	5	4	4	3	2	5	3	5	5	4	5	0	4	4
20	4	4	4	3	2	3	2	2	3	4	4	2	1	3	4	4	4	3	4	0	4
21	3	3	4	3	3	2	3	3	2	2	4	2	3	1	1	1	4	5	4	4	0

Table 2: Distances between haplotypes of *Mal de Rio Cuarto virus* calculated using Equation 1. The table header and first column contain the numbers of haplotypes.

that this haplotype is the most advantageous in the conditions of that environment and the distribution of the haplotypes of the population is centered around it, considering it the "master" haplotype. The distance of any profile to haplotype 9 (or number of changes necessary to convert any haplotype into haplotype 9) can be seen according to the number of circles that separate them. For haplotypes 18 and 19, the minimum distance to any other haplotype is greater than 1, and therefore in the graphic representation they connect to virtual profiles, as explained in sub section Haplotype Networks.

The network obtained has the following properties: diameter 5; clustering coefficient 0.246 and mean distance 2.767.

Figure 2 shows the distribution of the degree of connectivity (k) of the nodes in the network. Haplotype 9 also stands out in its degree of connectivity. It is the haplotype with largest number of connections (seven in total).

According to the calculated properties, the network topology resembles that of a small-world network [12], but it cannot be said to share all its features because the clustering coefficient is less than 0.5 [11]. A small-world network has the characteristic that, even though relatively few long-distance connections are present, the shortest path length (number of links/hops) between two nodes scales logarithmically or slower with the size of the network for fixed average degree. This means that even in small-world networks with many nodes, the shortest path length between two individual nodes is likely relatively small, hence the "small-world" designation [15]. This characteristic occurs in the haplotype network of MRCV, even though they are not being fully connected, there are certain transitions between haplotypes that have not been observed. It can be concluded that for any pair of randomly selected haplotypes the number of known mutations required to convert one into the other is relatively low.

Exploration by environments

In the exploration of the network by environments defined by geographic regions and the crop years through dynamic graphical representation of the existence of haplotypes, it was observed that in crop years prior to 1996/97, the haplotypes are generally more separated in the graph and that there is also a greater number of different profiles. This particular distribution of the profiles suggests at first glance the possibility that there was greater variability of the virus in the early periods studied.

Variability indicators

To validate the observation made during the network exploration by environments and confirm that this behavior is independent of the number of samples, a variability indicator was created, the sum of distances between haplotypes (SDH) for each environment.

$$SDH_A = \sum_{i=1}^{n_A-1} \sum_{j=i+1}^{n_A} d_{ij}$$

where:

SDH_A: sum of distances between haplotypes of environment A

n_A: number of haplotypes of the environment A

d_{ij}: distance between haplotype i and haplotype j

There is a relationship between the number of distinct haplotypes for each environment and genetic variability. If environment A has a greater number of haplotypes than environment B, then A has greater variability than B. However, this criterion cannot be used in isolation,

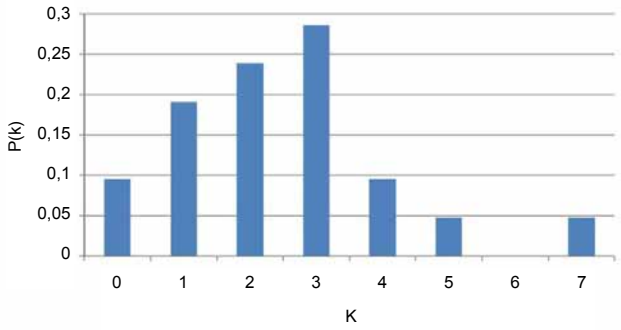

Figure 2: Distribution of degree of connectivity (*k*) of the haplotype network of *Mal de Rio Cuarto virus* (MRCV). The degree of connectivity k of a particular node is the number of arcs incident to the node.*P*(*k*) is the probability that a randomly chosen node has degree *k*.

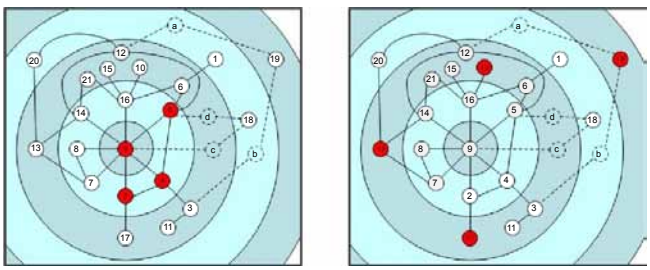

Figure 3: Graphic representation of the haplotype network of *Mal de Rio Cuarto virus* (MRCV) according to the distances in Table 2 in two different environments. The arcs drawn in solid lines indicate distance = 1 between connected haplotypes. The concentric circles in the background represent the distance between the nodes contained and the core haplotype, where each circle from the center increases one unit of distance. The same network is shown for two different environments, where the nodes painted indicate the haplotypes present in each environment.

because it does not differentiate cases such as that of Figure 3, where two networks have the same number of haplotypes, but they are more widely spaced in one than in the other. In the environment on the left, there are four haplotypes that are within short distances from each other. In the environment on the right, the number of haplotypes is the same, but the distance between them is greater.

The greater separation between haplotypes within the same network also shows greater variability, because to increase the distance of haplotypes requires a greater number of changes.

The sum of the distances of the paths between existing haplotypes (SDH) in each environment represents both the distance between haplotypes as well as the number of haplotypes of the environment, since the number of paths depends on the number of haplotypes.

SDH grows rapidly with the number of samples taken in the environment studied. For this reason the results are displayed in relation to the expected value of SDH, making it possible to compare environments with different amounts of samples.

$$E(SDH_A) = \sum_{i=1}^{n_{A-1}} \sum_{j=i+1}^{n_A} (1-(1-P(h_i))^{nA})(1-1-P(h_j))^{nA})d_{ij}$$

where:

$E(SDH_A)$: expected value of SDH for the environment A

n_A: number of haplotypes of the environment A

d_{ij} : distance between haplotype i and haplotype j

$P(h_i)$: probability of existence of haplotype i

The calculation of the SDH indicator was performed for each environment in relation to the expected value of SDH. The environments are categorized into two classes, "Until 1996/97" and "Since 1997/98," which highlight the division in behavior of the virus observed during the examination (Figure 4). Thus, the indicator SDH shows that the variability was high until the crop year of the epidemic in 1996/97 [16], and then variability markedly decreased.

Discussion and Conclusions

According to the calculated index, the variability of *Mal de Rio Cuarto virus* has decreased over time, with a clear change of the indicator in the crop year after the epidemic of crop year 1996/97, which divides the crop years into until 1996/97 and Since 1997/98.

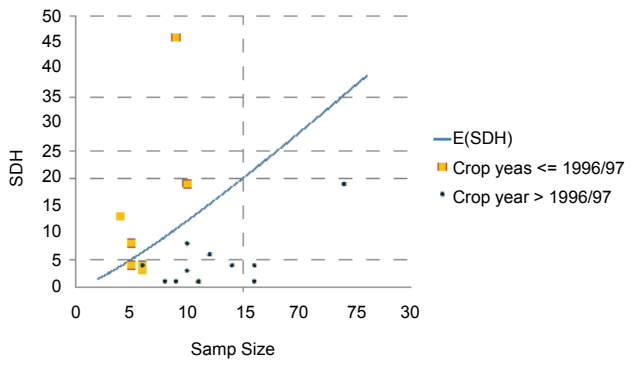

Figure 4: Sum of distance between haplotypes (SDH) of *Mal de Rio Cuarto virus* (MRCV) vs. the expected value of SDH. SDH values were calculated for each environment defined by combinations of the attributes Region from endemic and Crop year and were divided into the categories "Until 1996/97" and "Since 1997/98" by the Crop year attribute.

Additionally, it was determined that the network of haplotypes of *Mal de Rio Cuarto virus* resembles a network of small-world type [12], but its clustering coefficient is lower than expected in this topology [11]. The use of networks in the KDD process was very successful and managed to highlight behavior of the object of study that had not been evident so far. Although an AMOVA analysis [2] and also a haplotype analysis by environments had been performed [4], the difference or distance between the profiles of each environment could be detected only with the implementation of the haplotype networks. Examination by network environments was instrumental in obtaining these results, since the hypothesis later demonstrated with indicator SDH arose after exploration work with different types and levels of clustering the information. In a human-centered process, where the creativity and experience of the analyst play a key role [17], the proposed tool was able to offer a fresh perspective, complementary to the other techniques of KDD process.

References

1. Laguna IG, Irma G, Remes-Lenicov A, Virla E, Avila A et al. (2002) Diffusion of *Mal de Rio Cuarto virus* (MRCV), its vector, associated delphacids and alternative hosts in Argentina. Journal of the Entomological Society of Argentina 61: 87-97.
2. Gimenez Pecci MP, Bruno C, Balzarini M, Laguna IG (2007) Application of molecular variance analysis on data from electrophoretic profiles of genomic segments of *Mal de Rio Cuarto virus* (MRCV) in Argentina. Proceedings of the National Academy of Sciences 13: 141-152.
3. Gimenez Pecci MP, Carpane P, Dagoberto E, Laguna IG (2005) Variability of electrophoretic profile of the genome segments of the causative virus of Rio Cuarto Corn Disease in Argentina. XIII Latin American Congress of Plant Pathology 4: 562.
4. Gimenez Pecci MP, Carpane P, Murua L, Bruno C, Balzarini M (2008) Variability of *Mal de Rio Cuarto virus* (MRCV) according haplotype frequency obtained from electrophoretic profiles of genomic segments. Proceedings of the National Academy of Sciences 14: 99-107.
5. Gimenez Pecci MP, Laguna IG, Garcia MA, Carpane P (2007) Extra-genomic bands on the electrophoretic profile of the dsRNA of *Mal de Rio Cuarto virus* (*Fijivirus*, Reoviridae). Argentine Journal of Microbiology 1: 108.
6. Hamming RW (1950) Error Detecting and Error Correcting Codes. Bell System Tech Journal 9: 147-160.
7. Liao L (2006) Hierarchical Profiling, Scoring and Applications in Bioinformatics. Hsu H. Advanced Data Mining Technologies in Bioinformatics, (ed), Idea Group Publishing: 13-31.
8. Junker BH, Schreiber F, Gemeinholzer B (2008) Phylogenetic Networks. Analysis of Biological Networks, (ed), Wiley-Interscience: 255-281.
9. Posada D, Crandall KA (2001) Intra-specific gene genealogies: trees grafting into networks. Trends in Ecology & Evolution 16: 37-45.
10. Wuchty S, Ravasz E, Barabasi AL (2003) The Architecture of Biological Networks, Complex Systems in Biomedicine, (eds), Kluwer Academic Publishing, New York.
11. Lewis TG (2008) Network Science, Theory and Applications. Wiley Publishing Group.
12. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393: 440-442.
13. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Physics Reports 424: 175-308.
14. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479-491.
15. Mitchell M (2006) Complex systems: Network thinking. Artificial Intelligence 170: 1194-1212.

16. Lenardon SL, March GJ, Nome SF, Ornaghi JA (1998) Recent outbreak of *Mal de Rio Cuarto virus* on corn in Argentina. Plant Disease 82: 448.

17. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy (1996) The Process of Knowledge Discovery in Databases: A Human-Centered Approach. Advances in Knowledge Discovery and Data Mining, MIT Press: 37-58.