KDD a través redes de haplotipos: Análisis exploratorio de datos del genoma Mal de Río Cuarto virus (MRCV)

García, Mario Alejandro (1); Gimenez Pecci, María de la Paz (2); Steiner, Guillermo Max (1); Laguna, Irma Graciela (2); Gordillo, Romina Noemí (1)

1- UTN FRC (Universidad Tecnológica Nacional, Facultad Regional Córdoba)

2- INTA IFFIVE (Instituto Nacional de Tecnología Agropecuaria, Instituto de Fitopatología y Fisiología)

Abstract

El Mal de Rio Cuarto virus causa epidemias en cultivos de maíz en la Argentina, generando cuantiosas pérdidas. El análisis de la variabilidad genética es fundamental en la creación del modelo de comportamiento del virus. Este modelo tiene una gran importancia en los intentos de minimizar el impacto de la enfermedad. En este trabajo se aborda el estudio de la variabilidad genética desde la minería de datos, tomando una herramienta del área de la biología, como son las redes de haplotipos, y adaptándolas al proceso de KDD. El método resultante permite estudiar las relaciones de cada variante del virus en forma dinámica e interactiva, analizando los datos desde distintas perspectivas. También permite al experto definir distintas medidas para las distancias entre haplotipos en función del conocimiento del tema y comprobar los efectos que causan sobre la red. Estos cambios, además de ofrecer más información para el análisis, permiten una integración natural de la técnica con el proceso de KDD estándar CRISP.

Introducción

El virus del Mal de Río Cuarto (Mal de Río Cuarto virus, MRCV) es un miembro del género *Fijivirus* que causa epidemias en cultivos de maíz en Argentina, generando cuantiosas pérdidas. En Argentina este cultivo significa un ingreso anual muy importante en exportaciones, siendo uno de los principales productores a nivel mundial. La enfermedad causada por este virus fue detectada en la década del 60 y se ha dispersado progresivamente, infectando en la actualidad una amplia gama de gramíneas silvestres y cultivadas. Es transmitido por *Dephacodes kuscheli* y *D. haywardi* en forma persistente y propagativa. El genoma está representado por diez segmentos de RNA de doble cadena (Giménez P. 2004).

El análisis de la variabilidad genética es fundamental en la creación del modelo de comportamiento del virus. Este modelo tiene una gran importancia en los intentos de minimizar el impacto de la enfermedad (Giménez P. 2008).

Para estudiar la diversidad genética, se cuenta con una base de datos que contiene los resultados de análisis de electroforesis (Giménez P. 2005) realizados sobre muestras de 8 especies hospedantes en 13 localidades durante 13 campañas.

Perfil electroforético

El perfil electroforético se representa a través de una cadena binaria de longitud 18, que contiene los diez segmentos conocidos del virus, algunos de los cuales se pueden ubicar en distintas posiciones, y dos bandas extra genómicas (Giménez P. 2008). Un “0” en una posición del perfil electroforético indica la ausencia del segmento correspondiente, mientras que un “1” indica la presencia.

El MRCV presenta 21 perfiles electroforéticos distintos, a los que llamaremos haplotipos (genotipos haploides). Estos haplotipos son el objeto principal de nuestro trabajo. La herramienta presentada más adelante permite estudiar la relación entre los haplotipos según los demás atributos de la base de datos.

KDD

KDD (Knowledge discovery in database), muchas veces llamado Minería de Datos aunque esta sea solo una etapa, es un proceso que intenta encontrar información útil y novedosa (que pueda influir en la toma de decisiones) y que permanece oculta en una base de datos (Fayyad 1996).

El KDD y la Bioinformática ofrecen nuevas y promisorias áreas de aplicación de las ciencias de la computación.

Bioinformática es la disciplina que administra e interpreta información biológica mediante métodos y herramientas informáticas.

Los avances en genética, cómo por ejemplo la finalización del Proyecto Genoma Humano, el desarrollo de los microarrays de expresión génica, entre otros, indujeron a los investigadores de Biología a comenzar a utilizar bases de datos para almacenar los grandes volúmenes de información que generan. Además, existen bases de datos con información biológica de público acceso para la comunidad investigadora.

Nuevas ramas de la medicina, como la Medicina Personalizada y la Farmacogenética prometen grandes avances en el ámbito de la salud, pero el problema es ahora la dificultad que encuentran los investigadores para interpretar y lograr conclusiones importantes de los grandes volúmenes de datos que generan.

En este contexto, el KDD, que hasta ahora era principalmente un aporte de la Ciencia a los Negocios (Business Intelligence), ha encontrado un vasto campo de acción, impulsado algunas veces, por los mismos laboratorios que proveen los microarrays para poder brindar a los investigadores una solución más completa.

Fases de un proceso de KDD

Las fases del proceso de KDD son las siguientes (Pérez 2007):

- Selección: Recopilación e integración de las fuentes de datos, identificación de las variables relevantes y muestreo.
- Exploración: Utilización de técnicas de análisis exploratorio, deducción de las distribuciones y análisis de correlaciones.
- Limpieza: Detección de valores atípicos, tratamiento de datos faltantes y eliminación de datos erróneos.
- Transformación: Utilización de técnicas de reducción y aumento de la dimensión, discretización, escalado simpe y multidimensional.
- Minería de datos: Aplicación de técnicas predictivas (como métodos bayesianos , árboles de decisión, redes neuronales, etc.) y técnicas descriptivas (por ejemplo, clustering, reglas de asociación, etc.)
- Evaluación e interpretación de los resultados: Determinación de intervalos de confianza y análisis ROC.
- Difusión y uso de modelos: Visualización y simulación.

De acuerdo a esta organización del proceso de KDD, la herramienta propuesta se debe utilizar principalmente en la fase de Exploración. Asimismo, algunas de sus características la ubican dentro de la fase de Minería de Datos, lo que no implica una contradicción, ya que en el proceso de KDD es común que se compartan herramientas entre las fases y que las fases no sigan un orden estricto. Muestra de esto último es el estándar CRISP-DM.

CRISP-DM

CRISP-DM (CRoss-Industry Standard Process for Data Mining) propone seis etapas (fig.1) para el ciclo de vida de un proyecto de minería de datos (http://www.crisp-dm.org/Overview/index.htm)

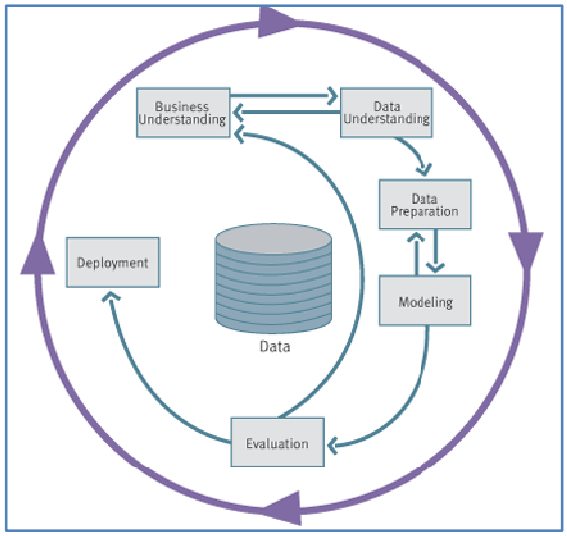


Figura 1. Fases del modelo de procesos CRISP-DM

Las etapas CRISP-DM no son incompatibles con las mencionadas en el punto anterior. El estándar especifica que la secuencia de las fases no es estricta. Siempre es necesario avanzar y retroceder. Depende de la salida de una fase qué fase o tarea debe llevarse a cabo en el paso siguiente. Las flechas indican las más frecuentes e importantes dependencias entre las fases.

Es importante tener en cuenta estas características del estándar porque es una de las cualidades que se pretende mantener con la herramienta planteada. En este enfoque, el investigador tiene la posibilidad de modificar un modelo o hipótesis en cualquier etapa del análisis sin salir del procedimiento. Además, el círculo externo indica que el proceso de minería de datos puede continuar aún después de haber obtenido un resultado.

Redes de haplotipos

La evolución genética es generalmente representada mediante un dendograma (fig. 2). Un dendograma es un

árbol calculado a partir de las distancias genéticas entre especies o entre individuos distintos de la misma especie. En minería de datos, esta representación está relacionada con los algoritmos de clustering jerárquico.

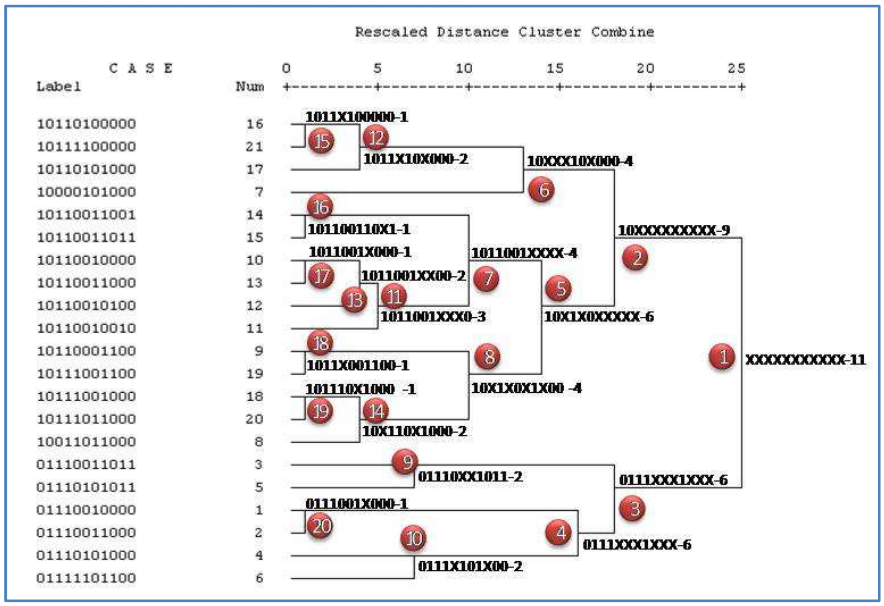


Figura 2. Dendograma de haplotipos de Mal de Río Cuarto virus (MRCV)

Un problema que tienen estos árboles es la imposibilidad de representar la distancia real entre algunos haplotipos. Por ejemplo, los haplotipos 18 y 20 de la figura 2 están a distancia 1, al igual que los haplotipos 20 y 8, pero en el árbol se unen primero los haplotipos 18 y 20, para agruparse en otro nivel con el 8. Este comportamiento genera una sensación de mayor distancia entre los haplotipos 20 y 8.

La genealogía a nivel de especies es distinta a la genealogía dentro de la misma especie.

Las relaciones entre las muestras de individuos de distintas especies son de naturaleza jerárquica. Esto se debe a que son producto del aislamiento reproductivo y divisiones de la población cada largos períodos de tiempo, durante la que la mutación combinada con la divergencia de la población condujo a la fijación de alelos diferentes.

En contraste, las relaciones entre individuos de la misma especie, no siempre pueden ser representadas por un árbol binario porque no son jerárquicas. Esto es porque son el resultado de la reproducción sexual y de una pequeña cantidad de relativamente recientes mutaciones. En el caso de los virus la reproducción no es sexual, pero este trabajo plantea una herramienta de uso general.

Para estudiar las relaciones filogenéticas dentro de la misma especie es necesario un método que permita representar ciclos y bifurcaciones múltiples. La solución es crear redes de haplotipos. Además de solucionar el problema de la representación de las distancias mencionado anteriormente, las redes permiten representar más información que los árboles binarios. Por ejemplo, la presencia de ciclos en la red puede indicar recombinación. En otros casos, los ciclos son producto de homoplasias (evolución convergente) e indican la ocurrencia de mutaciones reversas o paralelas. Otra ventaja de las redes, es que permiten hacer inferencias sobre las relaciones intra-especie, dando soporte a la teoría de la coalescencia. (Posada 2001)

Existen varios métodos para la creación de las redes de haplotipos. La mayoría de estos métodos están basados en la distancia, con la idea de minimizar las distancias (cantidad de mutaciones) entre haplotipos. Algunos de los métodos más conocidos son *Median networks*, *Median-joining networks*, *Molecular-variance parsimony* y *Likelihood network*.

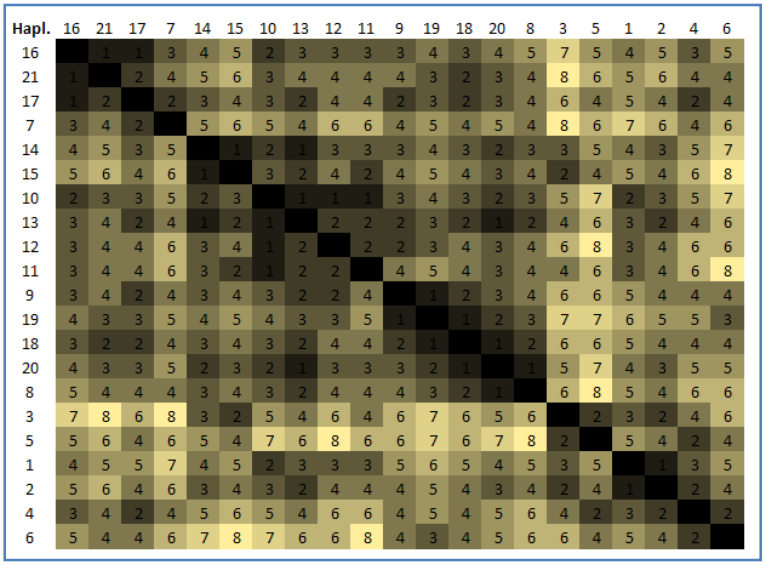
Creación de las redes de haplotipos

El primer paso crear una red de haplotipos es calcular la distancia entre los haplotipos. La forma clásica de calcular esta distancia es contar la cantidad de elementos distintos que hay entre cada haplotipo. Por ejemplo, en la *tabla 1* se puede observar cómo se determina la distancia entre los haplotipos 13 y 17 de MRCV. En este caso la distancia entre los haplotipos es 2, porque hay dos segmentos del perfil electroforético con valores distintos, el B9b y el B9c.

Hapl.	B3a	B3b	B5	B8	B9a	B9b	B9c	B10a	B10b	BE	B5a
13	1	0	1	1	0	0	1	1	0	0	0
17	1	0	1	1	0	1	0	1	0	0	0
Dif.	0	0	0	0	0	**1**	**1**	0	0	0	0

Tabla 1. Cálculo de la distancia entre los haplotipos 13 y 17 del MRCV

La siguiente figura (fig. 3) muestra todas las distancias entre los haplotipos de MRCV calculadas de la misma forma que en la tabla 1.



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

Figura 3. Distancias entre haplotipos de MRCV

Después de obtener las distancias, se debe armar la red correspondiente. En la (fig. 4) se puede ver una red de haplotipos que representa las distancias de la fig. 3.

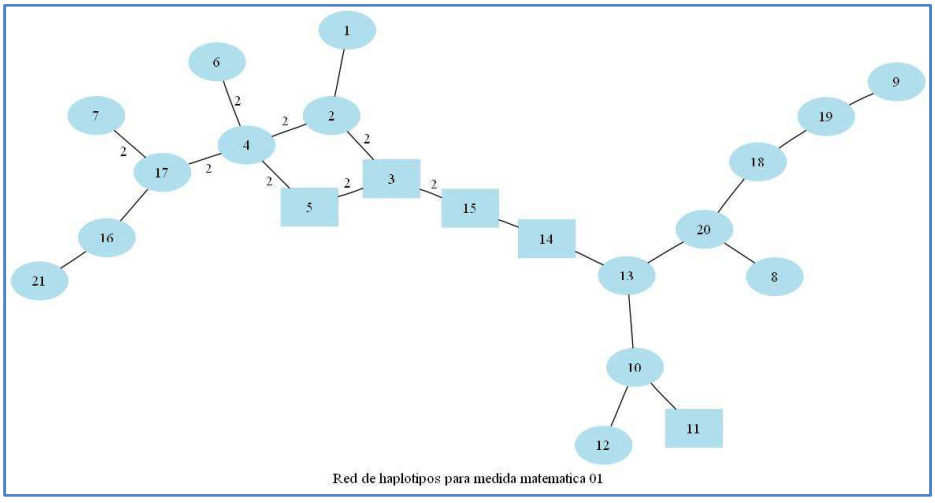


Figura 4: Red de haplotipos del MRCV

Para la creación de esta red, se calculó la mínima distancia de cada haplotipo a algún otro. Por ejemplo, para el haplotipo 10 se pueden calcular 20 distancias, una por cada uno de lo demás haplotipos. La mínima distancia del haplotipo 10 ($MínD_{10}$) es 1, porque está a una mutación de distancia de los haplotipos 11, 12 y 13.

En la generación del gráfico, se dibujan todos los arcos entre haplotipos que pertenecen al conjunto de las mínimas

distancias de cada uno. Por esa razón, en el gráfico el haplotipo 10 está conectado sólo a los haplotipos 11, 12 y 13, no con el 14 que está a distancia 2.

Para mejorar la visualización de la red, se decidió no etiquetar los arcos de valor uno, pero sí los que representan distancias mayores. Además, en este caso se intentó diferenciar a los haplotipos que presentan bandas extragenómicas mediante el uso de nodos rectangulares.

Análisis exploratorio

En la mayoría de las redes de haplotipos, el tamaño del nodo representa la frecuencia de ocurrencia del haplotipo. También se suele insertar dentro de cada nodo un gráfico de torta mostrando las proporciones de algún atributo importante para el análisis, pero el gráfico continúa siendo estático.

La herramienta que se propone en este trabajo permite un análisis dinámico e interactivo de la red de haplotipos, brindándole al usuario la posibilidad de analizar la información desde distintos puntos de vista.

Mediante la selección de atributos de la base de datos, como por ejemplo *año* y *localidad*, se puede generar un espacio de redes de haplotipos donde cada una contenga solo los haplotipos que existen para los valores de los atributos en un punto determinado de ese espacio. Por ejemplo, para el año 1991 en la localidad de Rio Cuarto, sólo se encontraron muestras de los haplotipos 3, 13, 14 y 15 (fig. 5).

La interface de la herramienta permite navegar por el espacio de redes, mostrando sobre la estructura de la red general, calculada para el universo de las muestras, los haplotipos existentes en cada punto del espacio resaltando los nodos con un color especial.

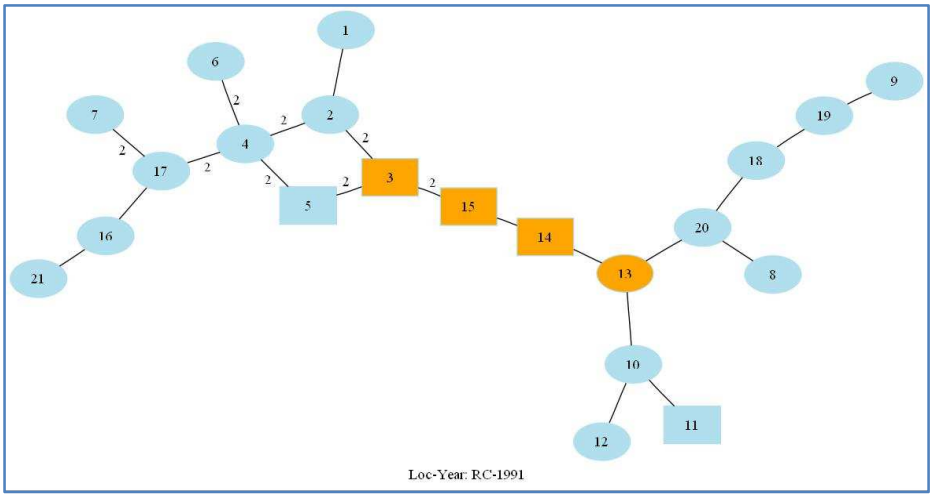


Figura 5: Haplotipos existentes para *Localidad* = “RC” y *año* = 1991

En la fig. 6 se muestra un ejemplo de navegación a través del espacio de redes que se forma entre los atributos *año* y *localidad*.

La información que provee esta exploración es mayor (o por lo menos complementaria) a la que brindan las redes con el gráfico de torta sobre alguno de los atributos.

Definición de nuevas medidas de la distancia.

Otra mejora que se propone con respecto a las herramientas tradicionales de análisis de redes de haplotipos, es la libertad de redefinir distancias.

Por ejemplo, en el caso de la (tabla 1), donde se calculaba la distancia entre los haplotipos 13 y 17, el experto en biología podría decidir un cambio en la distancia y asignarle 1 en lugar de 2. Este cambio tiene sentido, dado que los segmentos B9b y B9c son mutuamente excluyentes y el experto podría suponer que el cambio depende de una mutación y no de dos mutaciones simultaneas.

Un cabio en las distancias de los haplotipos generan una nueva topología de la red, comenzando así un nuevo ciclo de exploración.

Estos cambios en las distancias que puede generar el usuario experto, están en línea con el proceso propuesto

por el estándar CRISP-DM, ya que se pueden plantear como cambios en el modelo de análisis, que a su vez generan nuevo conocimiento y nuevas hipótesis que pueden volver a generar cambios en el modelo.

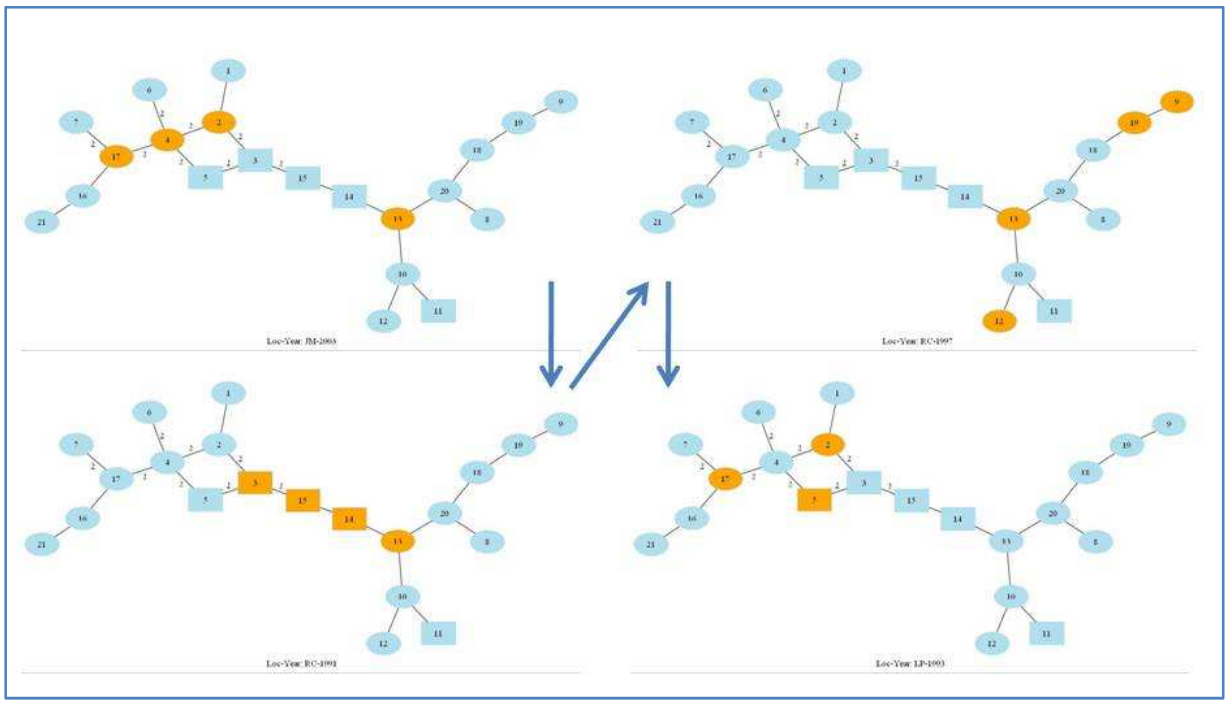


Figura 6: Exploración del espacio de redes del MRCV por año y localidad

Además de permitir al usuario la modificación de las distancias, se podría utilizar la información estadística que se genera durante el proceso de KDD y, mediante la inclusión de las leyes de la teoría de la coalescencia, realizar sugerencias automáticas de nuevas distancias entre haplotipos.

Financiamiento:

Este trabajo fue financiado por la Secretaría de Ciencia y Tecnología - U.T.N. en el marco del proyecto "Aplicación de técnicas de Data Mining al estudio del Mal de Río Cuarto del maíz", acreditado por la (Disp. 135/08, código EIPRCO752)

Referencias

Fayyad, U. M. 1996. **Data mining and knowledge discovery: making sense out of data.** *IEEE Expert, October, pages 20-25.*

Gimenez Pecci, M. P.; Lunello, P. y Ponz, F. 2004. **Desarrollo de metodologías de diagnóstico de alta sensibilidad para la detección del fijivirus del Mal de Río Cuarto.** *XII Congreso de la Sociedad Española de Fitopatología.*

Gimenez Pecci, M. P.; Carpane, P.; Dagoberto E. y Laguna, I. G. 2005. **Variabilidad del perfil electroforético de los segmentos genómicos del virus causal del Mal de Rio Cuarto del maíz en Argentina.** *XIII Congreso Latinoamericano de Fitopatología.*

Gimenez Pecci, M. P.; Laguna, I. G.; García, M. A. y Carpane, P. 2008. **Bandas extragenómicas en el perfil electroforético del DSRNA de Mal de Río Cuarto virus del maíz (fijivirus, reoviridae).** *IX Congreso Argentino de Virología.*

Pérez López, C.; González, D. S. 2007. **Minería de datos. Técnicas y herramientas.** *Ed. Thomson* ISBN: 978-84-9732-492-2

Posada, D. y K. A. Crandall. 2001. **Intraspecific gene genealogies: trees grafting into networks**. *Trends in Ecology and Evolution* 16:37–45. CrossRef, PubMed, CSA