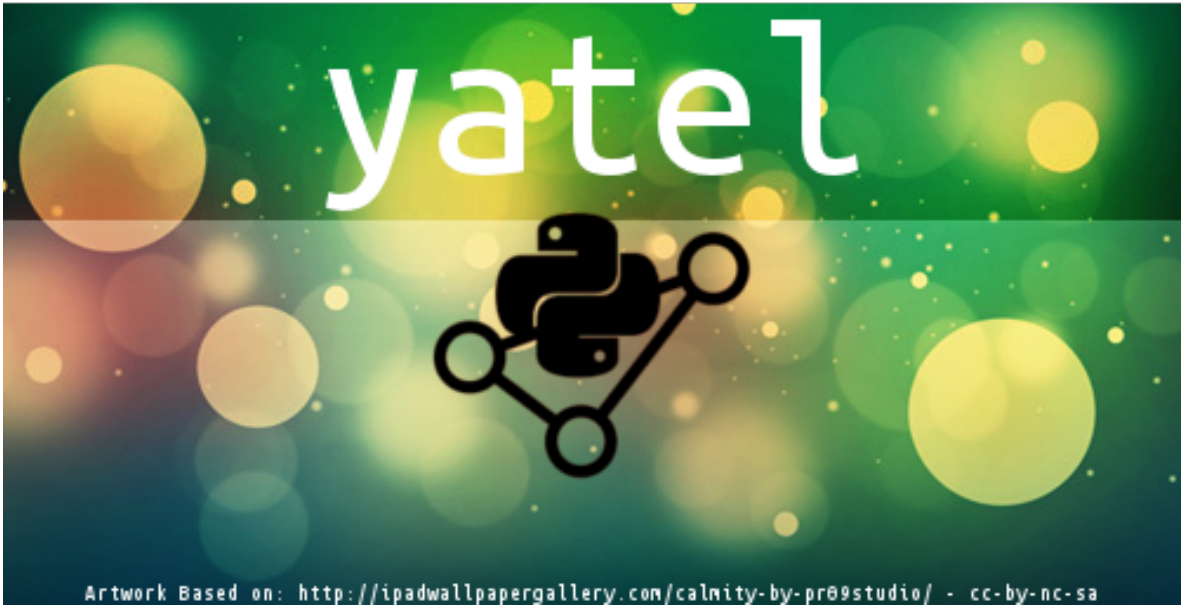
Caracterización del Sistema Mal de Río Cuarto del Maíz mediante Minería de Datos y Análisis de Redes



Integrantes

- - García, Mario Alejandro
- - Cabral, Juan Bautista
- - Gimenez Pecci, María de la Paz
- - Vera, Carlos
- - Liberal Rodrigo
- - Laguna, Irma Graciela
- - Bisonard, Eduardo Matías
- - Maurino, Fernanda
- - Vankeirsbilck, Inés
- - Cucco, Noelia del Valle
- - Nieto Castillo, Adrián L.

Paper

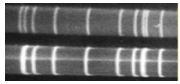
- **Título:** “Interactive network exploration in the KDD process, Contributions in the study of population variability of a Corn Fijivirus”
- **Autores:** M. A. García, M. P. Giménez Pecci, J. B. Cabral, A. Nieto, I. G. Laguna.
- **Publicación:** Journal of Data Mining in Genomics & Proteomics 2012 3:3
- **Editorial:** OMICS Publishing Group
- **ISSN:** 2153-0602. Año: 2012
- **URL:** http://goo.gl/pcjdG

Knowledge Discovery in Database (KDD)

- Es un proceso no trivial de identificación de información útil y desconocida que permanece oculta en una base de datos [Fayyad, 1996]
- Es un proceso centrado en la persona (human-centered) [Brachman, 1996]

Mal de Río Cuarto virus

- Análisis electroforético:



- Base de datos formada por

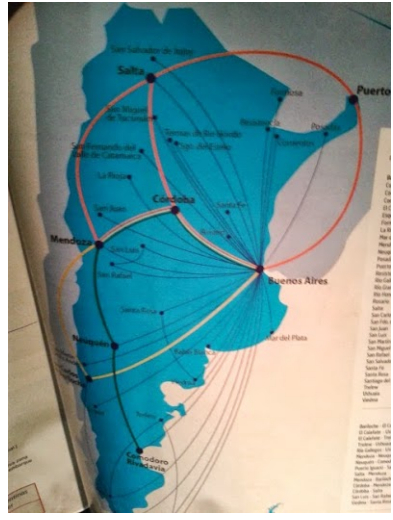
  perfiles electroforéticos + atributos que definen el ambiente de la planta

- Resultados de estudios anteriores:

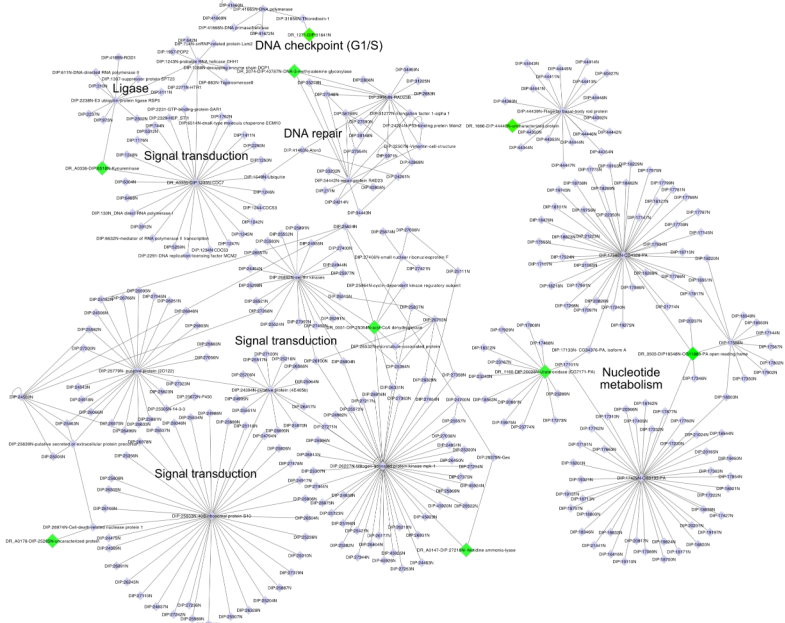
  Algunos segmentos electroforéticos dependen de otros

Network Science

- Es el estudio de las redes que representan fenómenos físicos, biológicos y sociales conduciendo a modelos predictivos de estos fenómenos.
- Topologías.
- Características comunes.

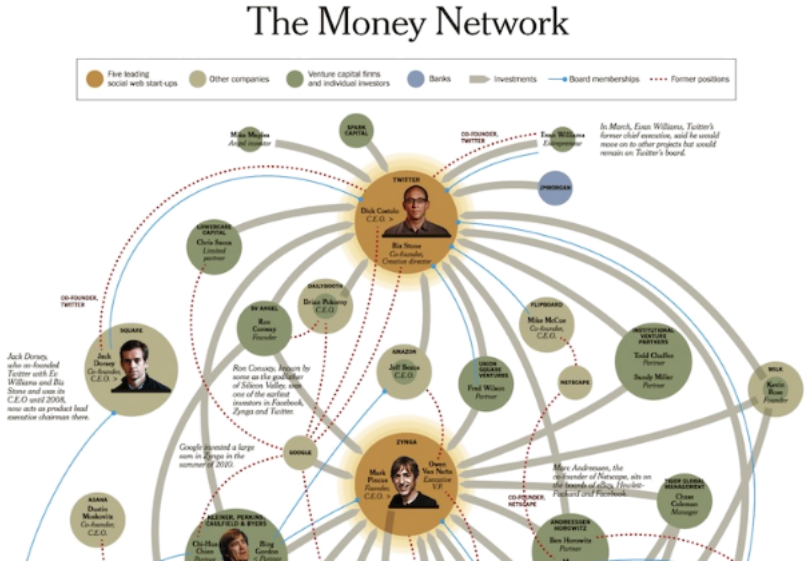


Networks 1



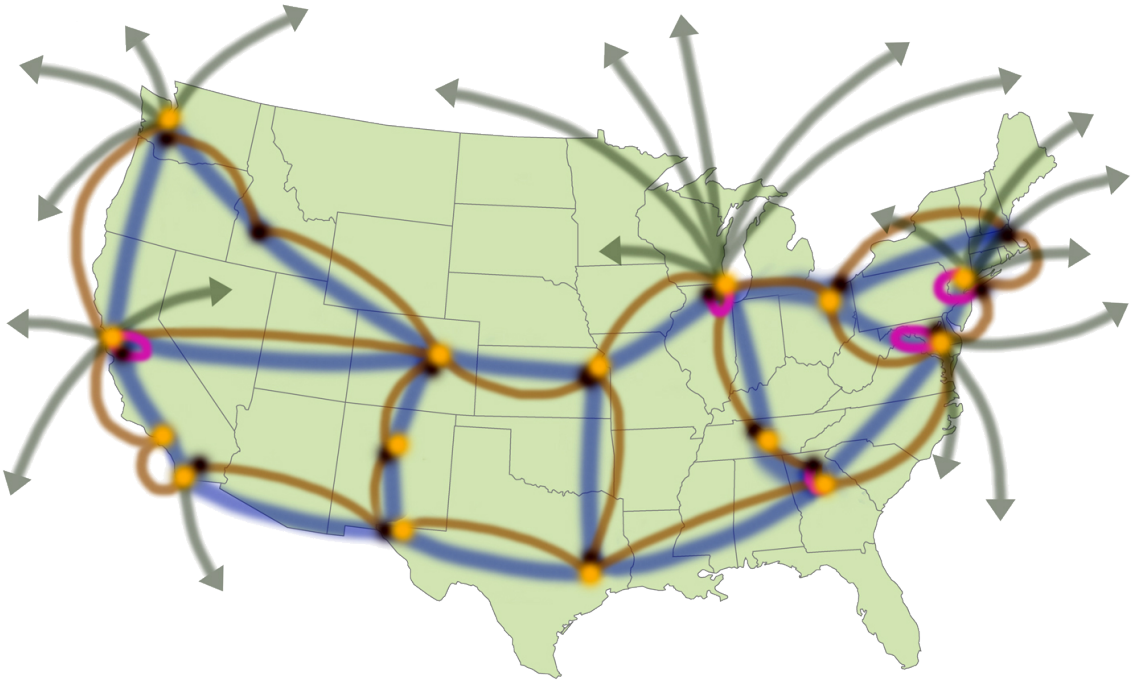
Interacción proteína-proteína

Networks 2



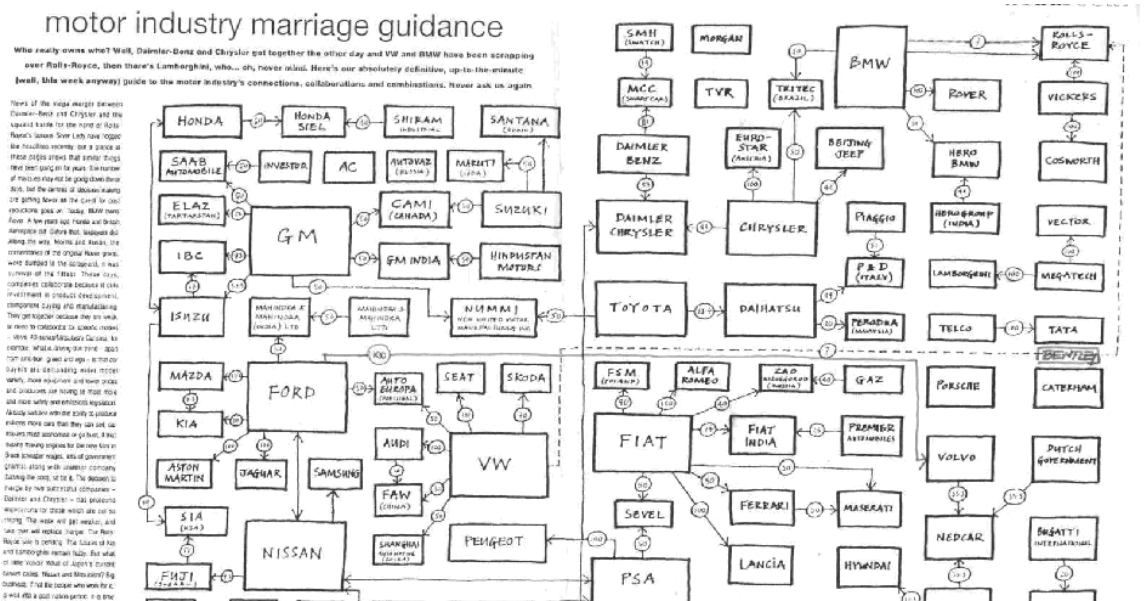
Redes sociales/económicas

Networks 3



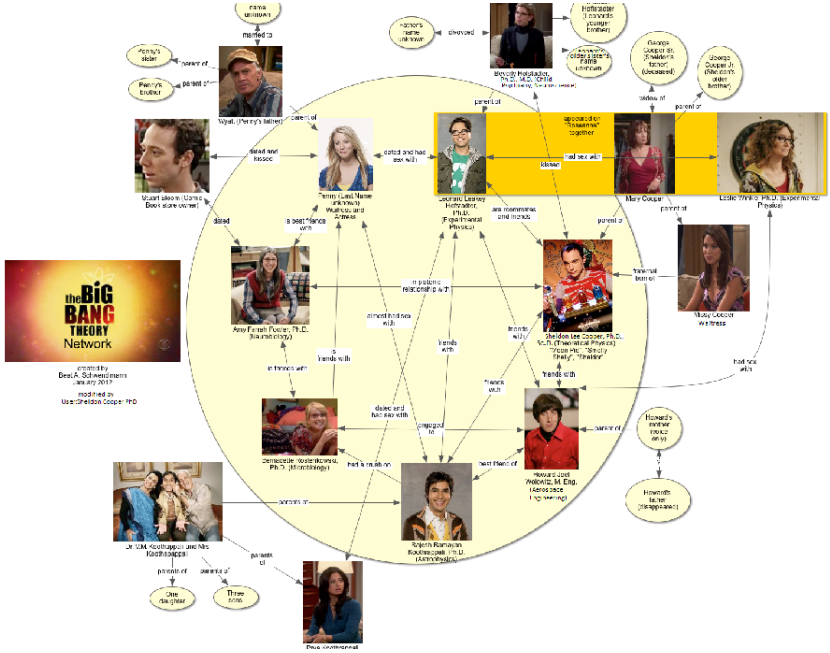
Red de distribución de energía

Networks 4



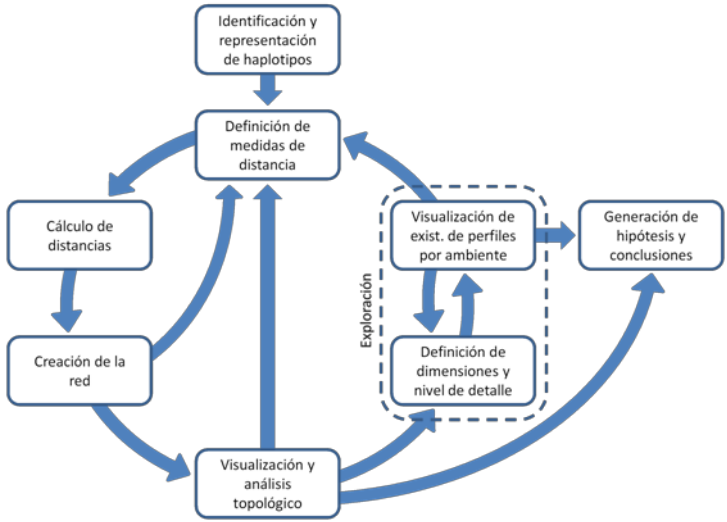
Relación entre automotrices

Networks 5

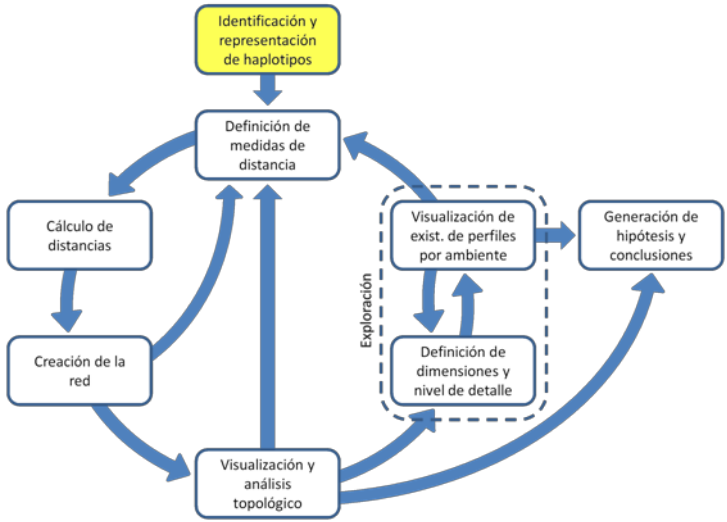


Red semántica TBBT

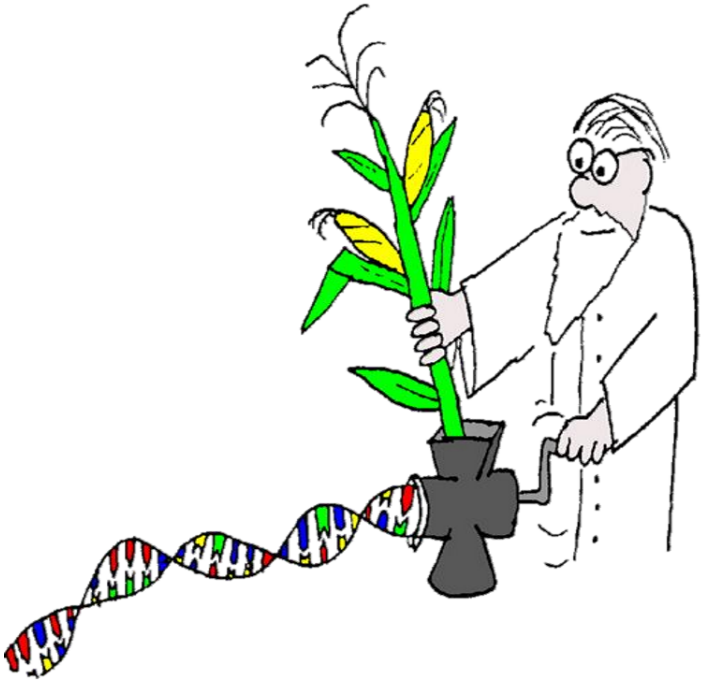
Proceso de análisis



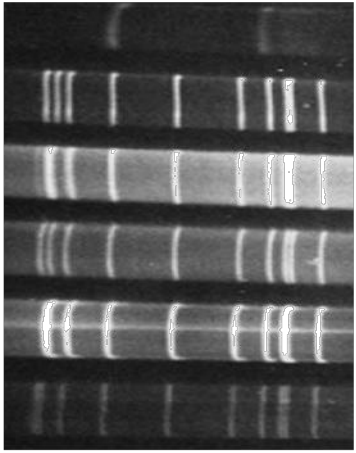
Identificación y representación de haplotipos



Identificación y representación de haplotipos

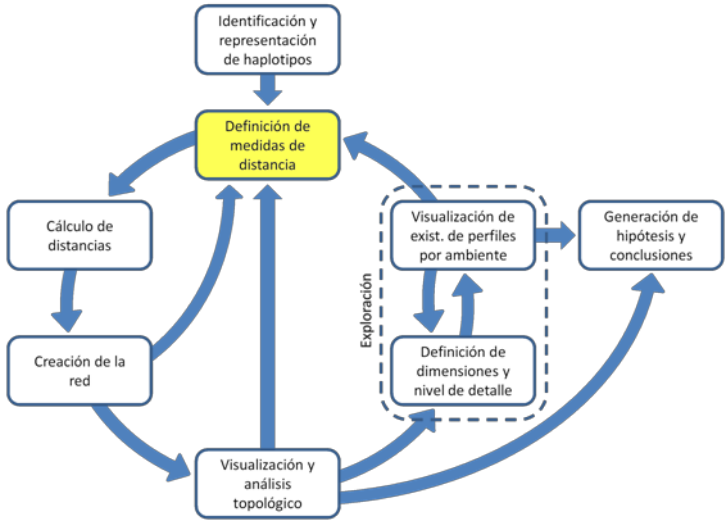


Identificación y representación de haplotipos



Hapl.	B3a	B3b	B5	B8	B9a	B9b	B9c	B10a	B10b	E5	E10
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
15	1	0	1	1	0	0	1	0	0	0	1
16	1	0	1	1	0	0	1	0	0	0	0
11	1	0	1	1	0	0	0	1	1	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
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
21	0	1	1	1	0	0	1	0	0	0	0

Definición de medidas de distancia



Definición de medidas de distancia

$$dij = dB3ij + dB5ij + dB8ij + dB9ij + dB10ij + dBE5ij + dBE10ij$$

donde:

$dB3ij = (|B3ai - B3aj| + |B3bi - B3bj| + |B3ai - B3aj + B3bi - B3bj|)/2$ *(excepción 1)*

$dB5ij = |B5i - B5j|$ *(dist. de Hamming)*

$dB8ij = |B8i - B8j|$ *(dist. de Hamming)*

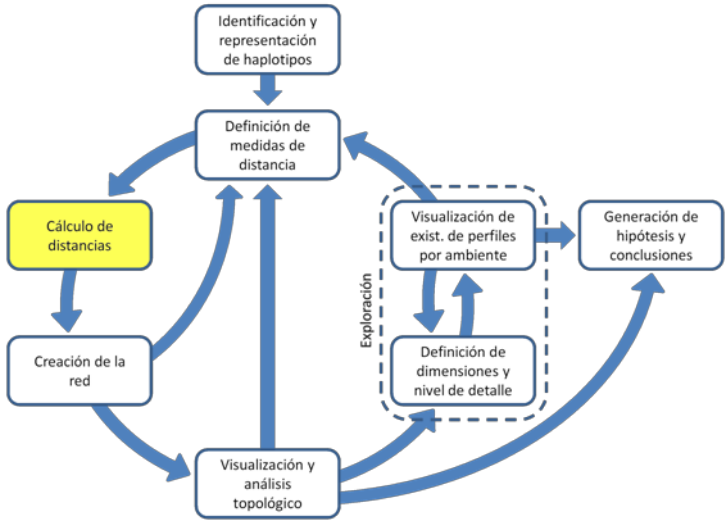
$dB9ij = (|B9ai - B9aj| + |B9bi - B9bj| + |B9ci - B9cj| + |B9ai - B9aj + B9bi - B9bj + B9ci - B9cj|)/2$ *(excepción 1)*

$dB10ij = (|B10ai - B10aj| + |B10bi - B10bj| + |B10ai - B10aj + B10bi - B10bj|)/2$ *(excepción 1)*

$dBE5ij = |BE5i - BE5j| (1 - |B5i - B5|)$ *(excepción 3)*

$dBE10ij = |BE10i - BE10j| (1 - (|B3ai - B3aj| \text{ OR } |B3bi - B3bj|))$ *(excepción 2)*

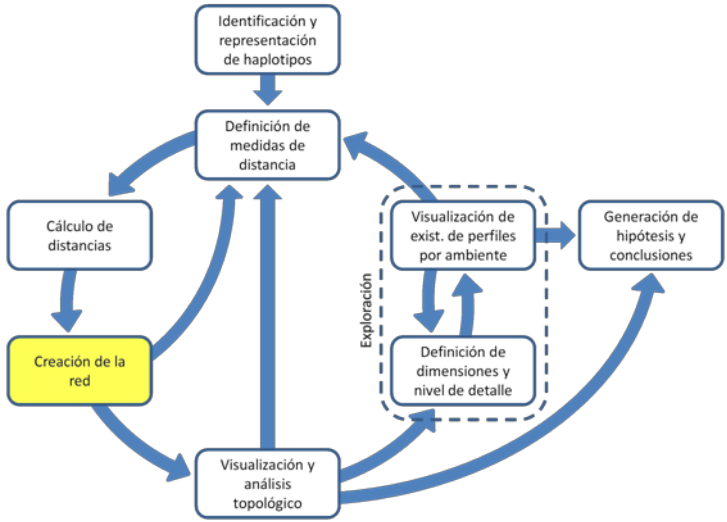
Cálculo de distancias



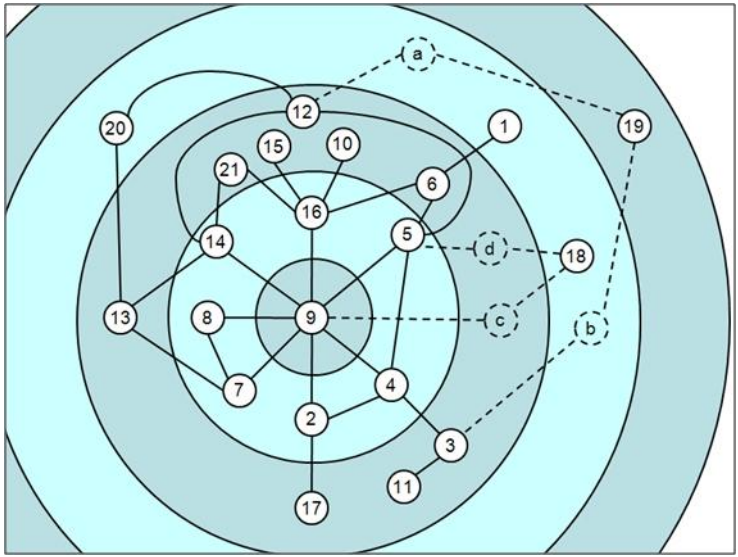
Cálculo de distancias

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

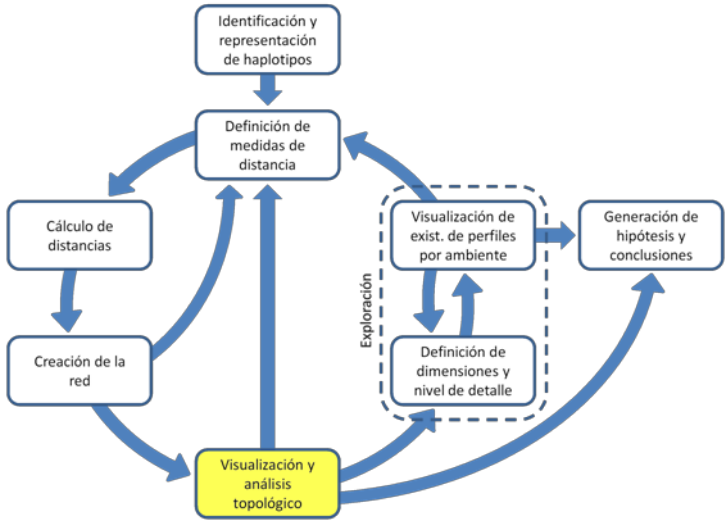
Creación de la red



Creación de la red



Visualización y análisis topológico



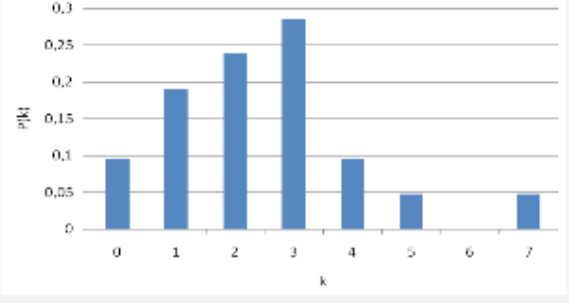
Visualización y análisis topológico

Clustering coefficient: $Cc_i = \frac{2c_i}{k_i(k_i \quad 1)}$ $CC = \frac{\sum_i^n{}_1 Cc_i}{n} = 0{,}246$

Diámetro = 5

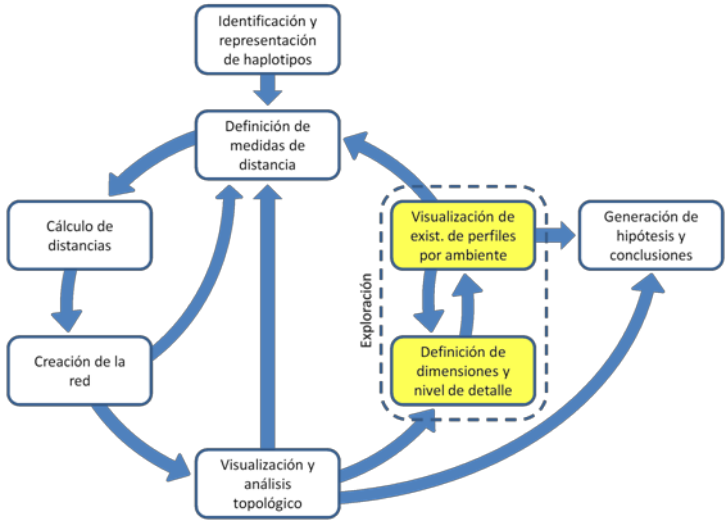
Distancia promedio = 2,767

Distribución de grado de conectividad:

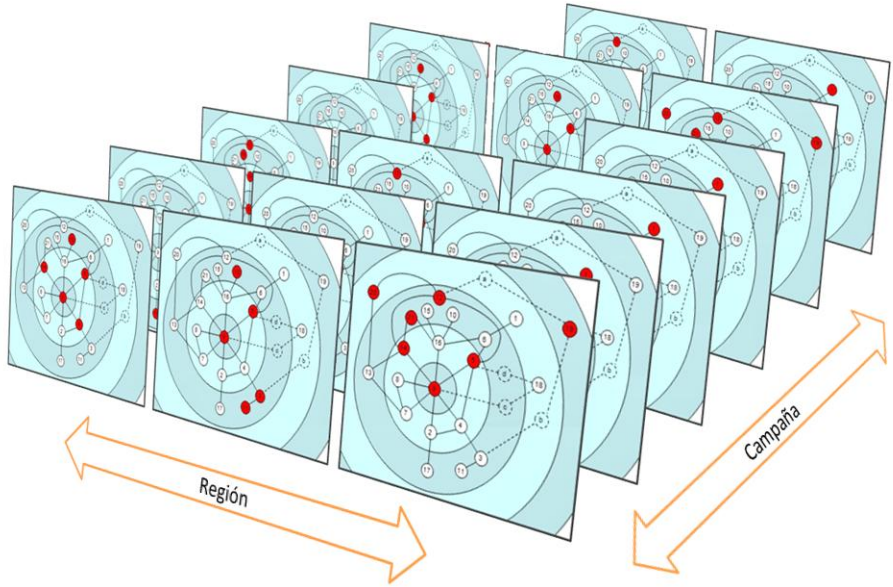


$$k_i = \sum_{j=1}^{N} d_{ij} \ | d_{ij} = 1$$

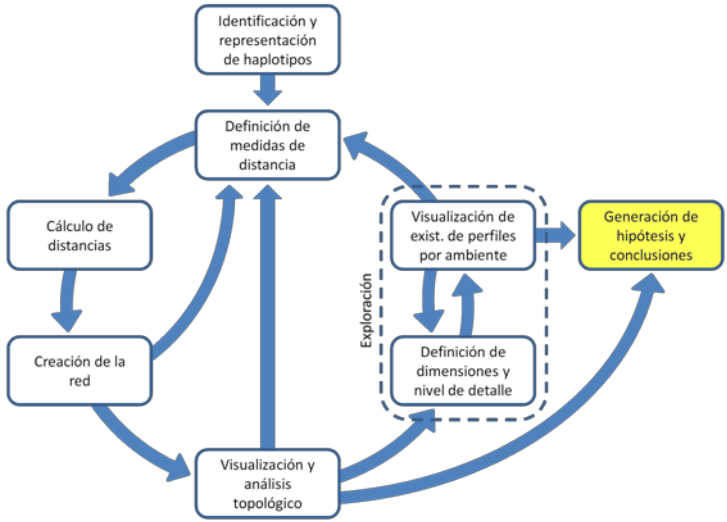
Exploración



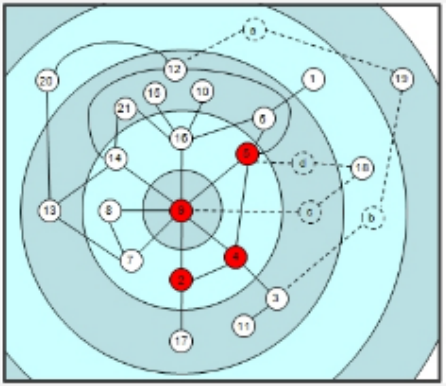
Exploración

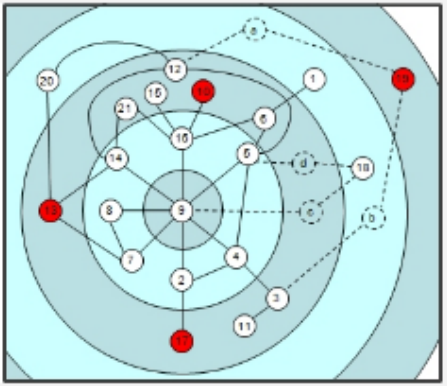


Generación de hipótesis y conclusiones



Generación de hipótesis y conclusiones





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

donde:

SDH_A: suma de distancias entre los haplotipos del ambiente A

n_A: cantidad de haplotipos del ambiente A

d_{ij}: distancia entre el haplotipo i y el haplotipo j

Generación de hipótesis y conclusiones

$$E(SDH_A) - \sum_{i\ 1}^{n_A-1} \sum_{j\ i+1}^{n_A} \left(1-\left(1-P(h_i)\right)^{n_A}\right)\left(1-\left(1-P(h_j)\right)^{n_A}\right) d_{ij}$$

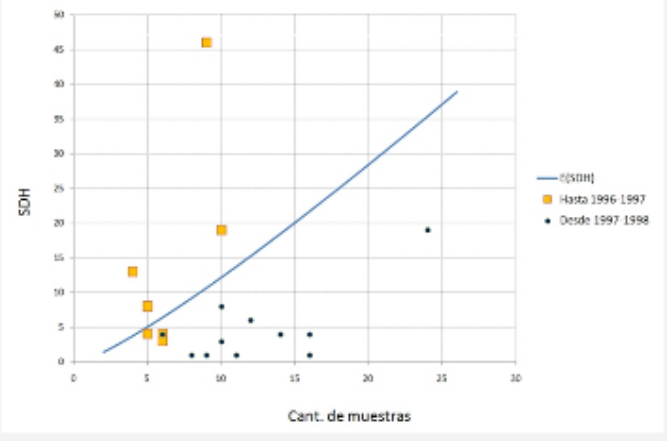
donde:

$E(SDH_A)$: *valor esperado de SDH del ambiente A*

n_A: *cantidad de haplotipos del ambiente A*

d_{ij}: *distancia entre el haplotipo i y el haplotipo j*

$P(h_i)$: *Probabilidad de existencia del haplotipo i*



Conclusiones del proyecto

- Según el índice calculado, la variabilidad del Mal de Río Cuarto virus, ha disminuido con el tiempo, habiendo una clara división del indicador en la campaña posterior a la epidemia de la campaña 1996/97.
- La utilización de redes en el proceso de KDD resultó muy satisfactoria y logró resaltar un comportamiento del objeto de estudio que no había sido evidente hasta el momento.
- En un proceso centrado en la persona (human-centered), donde la creatividad y experiencia del analista juega un rol fundamental, la herramienta propuesta es capaz de ofrecer una perspectiva novedosa y complementaria con las demás técnicas del proceso de KDD

Yatel

- Es la implementación en gran parte del proceso mencionado anteriormente.
- Falta trabajo (se aceptan colaboraciones)
- Esta implementado sobre: PyQt, Numpy, Ipython, Peewee, algo de Javascript y Pygments.
- Puede usarse como programa o como librería.
- Su version 0.2 es pip-instalable (`pip install yatel`) pero necesitan tener previamente numpy y pyqt.

Yatel - la app

En funcionamiento...

Yatel - la lib

```
1  from yatel import dom, weight, db
2
3  haps = [dom.Haplotype("hap0", b1=0, b2=1, b3=1),
4          dom.Haplotype("hap1", b1=0, b2=1, b3=0),
5          dom.Haplotype("hap2", b1=1, b2=1, b3=1)]
6
7  facts = [dom.Fact(haps[0].hap_id,
8                    date="25/5/10", place="Rio Cuarto", clima="lluvioso"),
9           dom.Fact(haps[1].hap_id,
10                   date="25/6/10", place="Rio Tercero",
11                   clima="soleado", estado_maiz="muerto"),
12          dom.Fact(haps[2].hap_id,
13                   date="25/5/11", place="Rio Cuarto", clima="nublado"),
14          dom.Fact(haps[0].hap_id,
15                   date="25/5/10", clima="soleado", estado_maiz="aislado")]
16
17 hamm = weight.Hamming()
18 edges = [dom.Edge(hamm.weight(haps[0], haps[1]),
19                   haps[0].hap_id, haps[1].hap_id),
20          dom.Edge(hamm.weight(haps[1], haps[2]),
21                   haps[1].hap_id, haps[2].hap_id)]
22
23 conn = db.YatelConnection("sqlite", "/home/juan/slides.db")
24 conn.init_with_values(haps, facts, edges)
25
26 # SECOND TIME!
27 conn.init_yatel_database()
```

Que le falta o problemas a resolver:

- Modulo de estadísticas.
- Minería de datos propiamente dicha.
- Exportar red a PNG.
- Posibilidad de navegar la red desde código.
- Biopython.
- Armar un exe.
- Testttttts!
- Integrar con excel.

¿Preguntas?

- **Charlas:**
  - http://bitbucket.org/leliel12/talks
- **Contacto:**
  - **Juan B Cabral**
    - Mail: jbc.develop@gmail.com
    - Twitter: @JuanBCabral
    - Blog: http://jbcabral.co