Diversidad y complejidad de la resistencia a medicamentos del HIV-1: Clasificación de mutaciones para predecir susceptibilidad o resistencia

Alma Ríos[1], Jesús González[1], Rigoberto Fonseca[1]

Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla México
{jagonzalez, rfonseca}@inaoep.mx

Resumen. El tratamiento del virus de inmunodeficiencia humana tipo 1 (HIV-1 por sus siglas en inglés) requiere entender las bases genéticas de la resistencia a medicamentos. El entendimiento es esencial para el desarrollo de nuevos antirretrovirales y mejorar el uso de las drogas existentes. En el presente trabajo, se analiza un amplio grupo de patrones de mutación asociados con la resistencia a medicamentos. Particularmente, se trabajó con los inhibidores de las encimas proteasa y transcriptasea inversa. Las cuales son fundamentales en la replicación del HIV-1. Se trabajó con un árbol de decisión J48, un clasificador bayesiano simple (Naïve Bayes) y un bayesiano simple TAN (Tree Augmented Naïve Bayes), para reconocer la resistencia a cada medicamento a partir de las mutaciones del virus. Los datos utilizados son parte del proyecto DYNAMO. Primeramente se agrupo los datos por medicamento, y se probó los clasificadores. En una segunda etapa se eliminó las posiciones sin mutación. Y finalmente, se seleccionó las posiciones más importantes en función de los resultados del árbol de decisión. En este trabajo se muestran los diferentes resultados obtenidos al realizar un "10 fold cross validation". Adicionalmente, se implementó una versión para la generación de la estructura de un Naïve Bayes aumentado a árbol (TAN) y se compara las redes generadas con el TAN del software Weka[1].

1 Introducción

Las pruebas de resistencia a medicamentos han mostrado ser beneficiales para el manejo clínico de los pacientes infectados con el HIV tipo 1. De un lado, el genoma nos asigna a una especie determinada, y en gran medida influye sobre los aspectos distintivos que nos hace únicos. Por otra parte, la manifestación visible de un organismo es su fenotipo. Ésta puede ser el color de piel, cabello, resistencia a medicamentos, etc. Los ensayos con fenotipos miden directamente la resistencia al medicamento. Los ensayos con genotipos proveen una evidencia indirecta de resistencia (son las comúnmente usadas).

[1] http://www.cs.waikato.ac.nz/ml/weka/

En la replicación del HIV tipo 1 participan tres enzimas esenciales que son: la Integrasa, la Proteasa, y la Transcriptasa Inversa. En la base de datos del proyecto DYNAMO, las mutaciones son representadas con letras o "sin mutació". Para el proyecto se considerarón cuatro medicamentos para "Nucleoside inhibitors of the reverse transcriptase (NRTI)", tres medicamentos para "Nonnucleoside reverse transcriptase inhibitors (NNRTI)", y cinco medicamentos para "Protease inhibitors (PI)". Cada grupo de medicamentos tiene asociadas 240, 240 y 99 posiciones de posible mutación, respectivamente. Lo que implica muchas posibles combinaciones de mutación a ser consideradas (Beerenwinkel, y otros, 2002).

Una manera de analizar la resistencia a medicamentos es por medio del factor "Cutoff". Cuando en un ensayo, este valor supera un límite se dice que la variedad de virus es resistente. Las variaciones de virus se determinan analizando las mutaciones en los genotipos. Se desea probar la capacidad de clasificación de un árbol de decisión, un bayesiano simple y un TAN.

Para la construcción del TAN particular se basó en el trabajo de Friedman (Friedman, y otros, 1997). En el que presentan un algoritmo denominado Tree Augmented Network (TAN) el cual consiste básicamente en una adaptación del algoritmo de Chow-Liu.

1.1 Naïve Bayes Aumentado a Árbol (Tree Augmented Network (TAN))

Para obtener este tipo de estructura se comienza por una estructura de árbol con las variables predictoras, para posteriormente conectar la variable clase con cada una de las variables predictoras. La Fig. 1 ilustra un ejemplo de estructura Naïve Bayes aumentada a árbol.

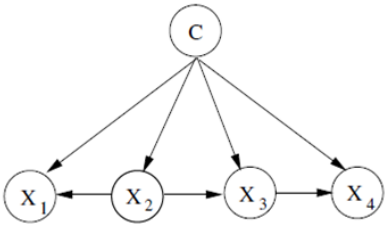

Fig. 1. Estructura Naïve Bayes aumentada a árbol

Friedman (Friedman, y otros, 1997) presentan un algoritmo denominado Tree Augmented Network (TAN). En el cual, se tiene en cuenta la cantidad de información mutua condicionada a la variable clase, en lugar de la cantidad de información mutua en la que se basa el algoritmo de Chow-Liu. La cantidad de información mutua entre las variables discretas X e Y condicionada a la variable C se define como:

$$I(X,Y|C) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r=1}^{w} p(x_i, y_j, c_r) \log \frac{p(x_i, y_j|c_r)}{p(x_i|c_r)p(y_j|c_r)} \quad \textbf{(1)}$$

Tal y como puede verse en el pseudocódigo del algoritmo TAN, éste consta de cinco pasos. En el primer paso se calculan las cantidades de información mutua para cada par de variables (X_i, X_j) condicionadas a la variable C. A continuación se debe construir un grafo no dirigido completo con n nodos, uno por cada una de las variables, en el cual el peso de cada arista viene dado por la cantidad de información mutua entre las dos variables unidas por la arista condicionada a la variable clase.

Pseudocódigo del algoritmo TAN

Paso 1. Calcular $I(X_i, X_j|C)$ con $i < j, i, j = 1, \ldots, n$
Paso 2. Construir un grafo no dirigido completo cuyos nodos corresponden a las variables predictoras: $X_1, \ldots, X_n$. Asignar a cada arista conectando las variables X_i y X_j un peso dado por $I(X_i, X_j|C)$.
Paso 3. A partir del grafo completo anterior y siguiendo el algoritmo de Kruskal construir un árbol expandido de máximo peso.
Paso 4. Transformar el árbol no dirigido resultante en uno dirigido, escogiendo una variable como raíz, para a continuación direccionar el resto de aristas.
Paso 5. Construir un modelo TAN añadiendo un nodo etiquetado como C y posteriormente un arco desde C a cada variable predictora X_i.

Se utiliza el algoritmo de Kruskal para construir el árbol expandido de máximo peso. El que parte de los $n(n-1)/2$ pesos obtenidos en el paso anterior. Éste se muestra en la **¡Error! No se encuentra el origen de la referencia.**. Las propiedades teóricas de este algoritmo de construcción de TAN son análogas a las del algoritmo de Chow-Liu.

Pseudocódigo del algoritmo de Kruskal

Paso 1. Asignar las dos aristas de mayor peso al árbol a construir.
Paso 2. Examinar la siguiente arista de mayor peso, y añadirla al árbol a no ser que forme un ciclo, en cuyo

```
caso se descarta y se examina la siguiente arista de
mayor peso.
Paso 3. Repetir el paso 2 hasta que se hayan
seleccionado n-1 aristas.
```

2 Trabajo relacionado

El primer trabajo relacionado es el presentado por Niko Beerenwinkel (Beerenwinkel, y otros, 2002). En el cual generan modelos, que aproximen la predicción de fenotipos a partir de genotipos, utilizando únicamente árboles de decisión. Utilizaron 14 antirretrovirales y trabajaron con 471 pruebas clínicas.

El siguiente trabajo es el presentado por Rhee (Rhee, y otros, 2006), es éste se utilizaron cinco métodos de aprendizaje estadísticos (árboles de decisión, redes neuronales, "suport vector regression", "least-squares regression", y "least angle regression"). Se consideraron 16 antirretrovirales y se realizó una validación cruzada de 5 pliegues (5-fold cross-validation) para cada método. El método que mejor resultado obtuvo fue el de "least angle regression".

Rhee presenta la base de datos de la Universidad de Stanford2 (Rhee, y otros, 2002). La misma es una base de datos relacional en línea, que cataloga la evolución y la relación de los medicamentos con las variaciones de secuencias de las encimas transcriptasea inversa y proteasa. Adicionalmente, el trabajo muestra las posiciones del genotipo y sus posibles mutaciones.

3 Metodología y desarrollo

Los datos considerados consisten de un grupo de 639 ensayos para los medicamentos NRTI (Nucleoside inhibitors of the reverse transcriptase), 748 ensayos para los medicamentos NNRTI (Nonnucleoside reverse transcriptase inhibitors), y 848 ensayos para los PI (Protease inhibitors). Cada ensayo muestra sus posiciones de mutación. Siendo estas 240 para los NRTI, NNRTI, y 99 para los PI. Los ensayos además muestran su resultado del análisis de resistencia para cada medicamento. Se discretizó estos resultados en susceptible o resistente. Para esta tarea se utilizó la límite por medicamento reportados por Niko Beerenwinkel (Beerenwinkel, y otros, 2002). Los límites se pueden apreciar en la **Tabla 1**.

[2] http://hivdb.stanford.edu

Tabla 1. Límites por medicamento de susceptibilidad o resistencia (Beerenwinkel, y otros, 2002).

Medicamentos		Cutoff
NNRTI	**Nucleoside inhibitors of the reverse transcriptase**	
ddI	Didanosine	2.5
d4T	Stavudine	2.5
3TC	Lamivudine	8.5
ABC	Abacavir	2.5
NNRTI	**Nonnucleoside reverse transcriptase inhibitors**	
NVP	Nevirapine	8.5
DLV	Delavirdine	8.5
EFV	Efavirenz	8.5
PI	**Protease inhibitors**	
SQV	Saquinavir	3.5
IDV	Indinavir	3.5
RTV	Ritonavir	3.5
NFV	Nelfinavir	3.5
APV	Amprenavir	3.5

El procedimiento seguido inició con la división de los datos por medicamento. Se filtraron los registros. Primero dividiendo los ensayos en conjuntos de entrenamiento por cada medicamento, se removieron los registros que no tenían relación con el medicamento analizado. Se ejecutó una primera serie de pruebas en la que se entrenó una árbol de decisión J48, un bayesiano simple y un bayesiano simple TAN.

Se prosiguió con la eliminación de las columnas que no presentaban mutación y se ejecutó una segunda serie de pruebas. A continuación, se seleccionaron por cada medicamento del conjunto PI, solo las posiciones que se obtenían del árbol de decisión y se ejecutó una tercera serie de pruebas. Finalmente, se comparó los árboles obtenidos en la última serie con los generados por la implementación particular de TAN.

La implementación de TAN se realizó en Java siguiendo el algoritmo propuesto por Friedman (Friedman, y otros, 1997).

4 Experimentos y resultados

4.1 Primera serie de pruebas

Con los archivos de entrenamiento, considerando todas las variables por cada medicamento. Se entrenó un árbol de decisión, un bayesiano simple y un bayesiano simple TAN. La **Fig. 2** muestra la comparación del porcentaje de generalización de

cada método por medicamento. Para determinar el porcentaje de generalización se realizó un "10-Folds Cross Validation".

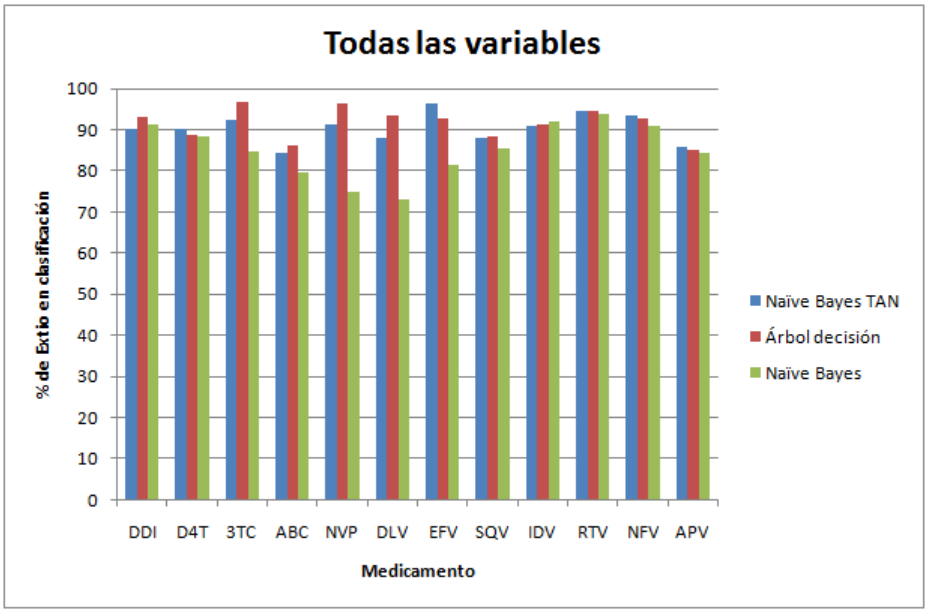

Fig. 2. Porcentaje de generalización del Naïve Bayes TAN, del árbol de decisión J48 y del Naïve Bayes. Obtenido con por cada medicamento, considerando todas las variables.

4.2 Segunda serie de pruebas

A cada archivo de entrenamiento se le eliminaron las columnas sin mutación. Y se ejecutó el grupo de experimentos con un Naïve Bayes TAN, un árbol de decisión J48 y un Naïve Bayes. Los resultados se resumen en la **Fig. 3**.

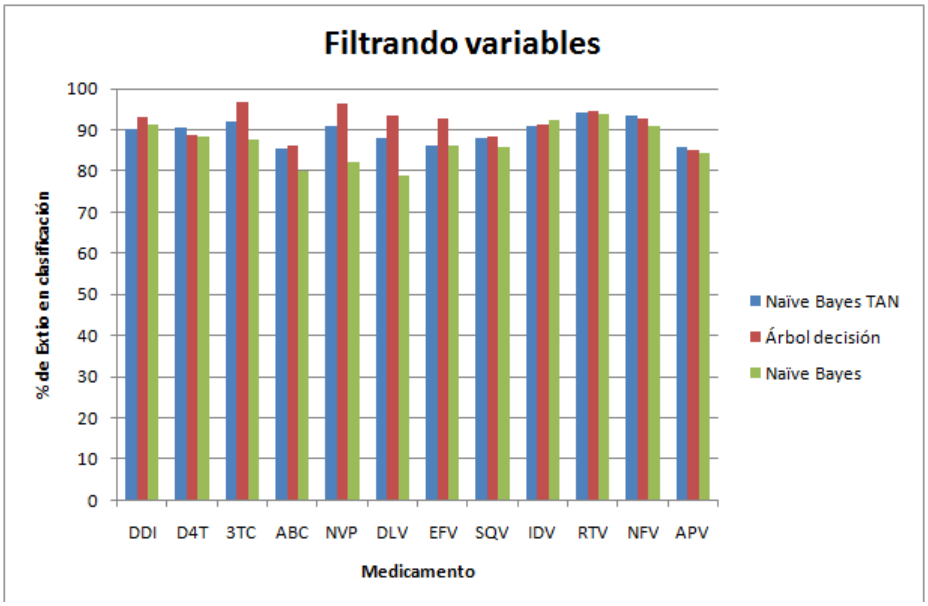

Fig. 3. Porcentaje de generalización del Naïve Bayes TAN, del árbol de decisión J48 y del Naïve Bayes. Obtenido con por cada medicamento, filtrando variables sin mutación.

4.3 Tercera serie de pruebas

A partir de los resultados de los árboles de decisión. Por cada medicamento de PI, se filtró las variables que aparecían en el respectivo árbol de decisión. A continuación, se ejecutó el grupo de experimentos con un Naïve Bayes TAN, un árbol de decisión J48 y un Naïve Bayes. Los resultados se muestran en la figura.

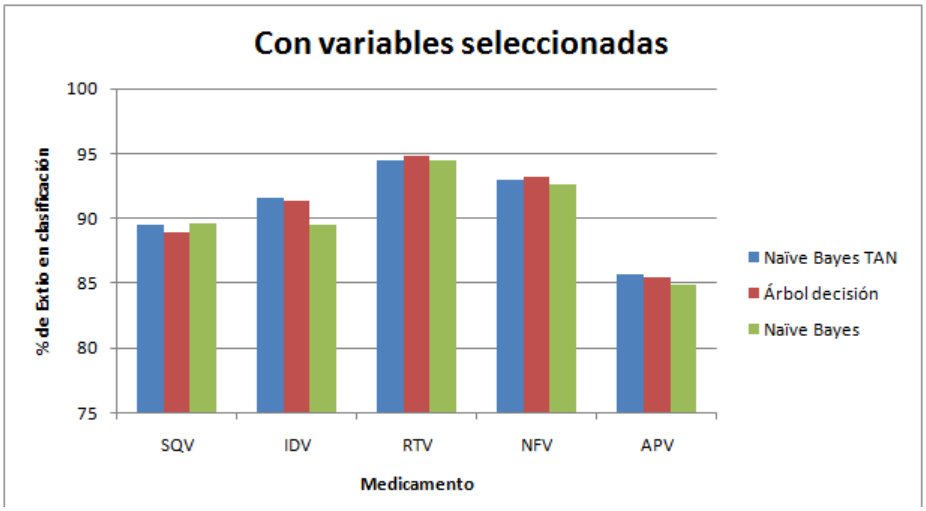

Fig. 4. Porcentaje de generalización del Naïve Bayes TAN, del árbol de decisión J48 y del Naïve Bayes. Obtenido con por cada medicamento, filtrando variables seleccionadas de la serie de pruebas anterior.

De esta serie de experimentos se obtuvo mejores resultados con el Naïve Bayes TAN en general.

4.4 Generación de estructuras TAN utilizando la implementación particular

Finalmente, se ejecutó el generador de estructuras TAN desarrollado con los datos de la tercera serie de pruebas. Al comparar las estructuras generadas con las estructuras TAN producidas por el software Weka, se observaron diferentes enlaces. La razón es por la variante del algoritmo TAN implementada.

5 Conclusiones y trabajo futuro

De los experimentos realizados se puede concluir que en una primera iteración el mejor modelo es un árbol de decisión. Sin embargo sus resultados se pueden mejorar, si a partir de las variables del árbol se construye un Naïve Bayes TAN.

El trabajo a futuro se propone probar con un método de clustering para asociar las posiciones de mutación y la susceptibilidad. Y separando los ensayos susceptibles de los resistentes, ejecutar nuevamente el método de clustering. Con el objetivo de determinar que tienen en común las mutaciones que se mantienen susceptibles y las que se vuelven resistentes.

El desarrollo de un método bayesiano sin la utilización de librerías especializadas es muy complejo. Sin embargo, se tiene una implementación de la generación de estructuras TAN, se debe extender para poder realizar la propagación de probabilidades con la estructura generada.

Referencias

1. Beerenwinkel Niko [y otros] Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype [Publicación periódica] // Proceedings of the National Academy of Sciences of the United States of America. - 2002. - págs. 8271-8276.
2. Friedman Nir, Geiger Dan y Goldszmidt Moises Bayesian Network Classifiers [Publicación periódica]. - The Netherlands : Kluwer Academic Publishers, 1997. - págs. 1-37.
3. Rhee Soo-Yon [y otros] Genotypic predictors of human immunodeficiency virus type 1 drug resistance [Publicación periódica] // Proceedings of the National Academy of Sciences of the United States of America. - 2006. - págs. 17355-17360.
4. Rhee Soo-Yon [y otros] Human immunodeficiency virus reverse transcriptase and protease sequense database [Publicación periódica]. - [s.l.] : Oxford University Press, 2002. - 1 : Vol. 31.

Anexos

Código fuente

El código fuente se adjunta en el archivo bayesianoTAN.zip. El proyecto entero se puede abrir utilizando Netbeans. En el adjunto se encuentran dos directorios principales. El primero es src que contiene el código fuente y dist donde se localiza el archivo BayesianoTAN.jar. Para ejecutarlo directamente se requiere tener instalado java y la sintaxis es el siguiente:

```
java -jar BayesianoTAN.jar nombre_archivo.csv
```

Resultados completos

Los resultados de todas las pruebas se adjuntan en el archivo resultados.zip, dentro de este archivo existe una carpeta “Resultados” con la siguiente estructura.

- Datos.- Directorio donde se encuentran los datos de entrenamiento. Que se filtraron para cada serie.
- Series 1, 2 y 3.- Contiene los resultados y modelos generados en los grupos de experimentos.

- T-Test para series 1 y 2.- Contiene los resultados de ejecutar una prueba T-test para analizar el porcentaje de acierto de los tres métodos, en las series de pruebas 1 y 2.
- Estructuras TAN.- Son las estructuras generadas con la implementación particular de TAN.

Estructuras TAN de los PI

Las estructuras TAN de los PI obtenidas en la tercera serie de experimentos se muestran en las figuras: **Fig. 1**, **Fig. 2**, **Fig. 3**, **Fig. 4**.

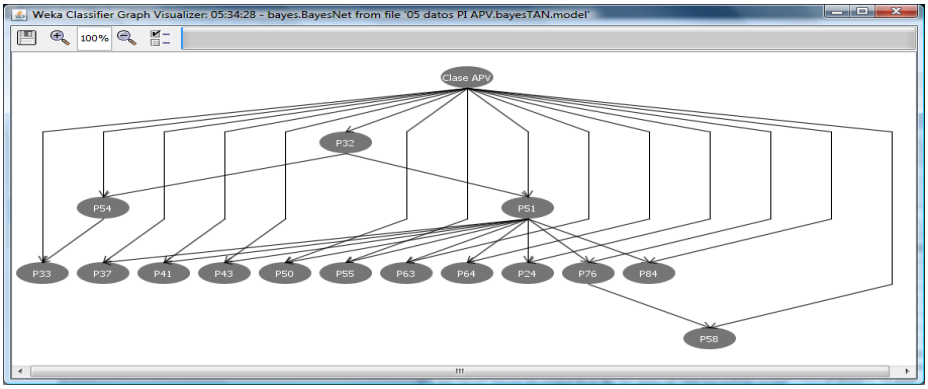

Fig. 1. Estructura TAN resultados de la serie 3 para el medicamento APV.

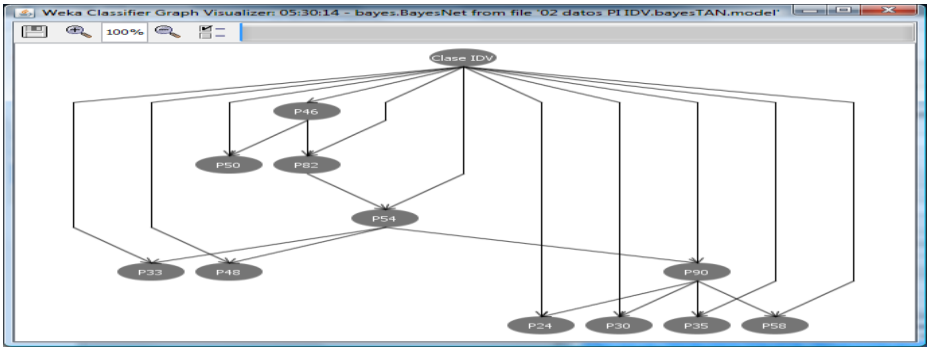

Fig. 2. Estructura TAN resultados de la serie 3 para el medicamento IDV.

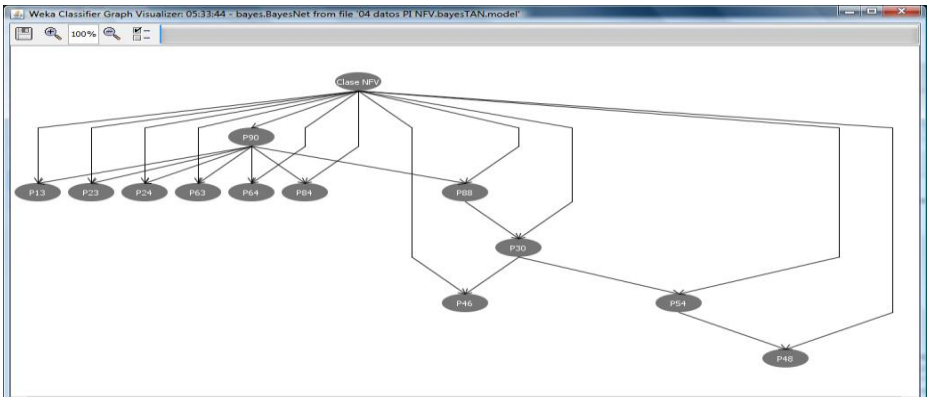

Fig. 3. Estructura TAN resultados de la serie 3 para el medicamento NFV.

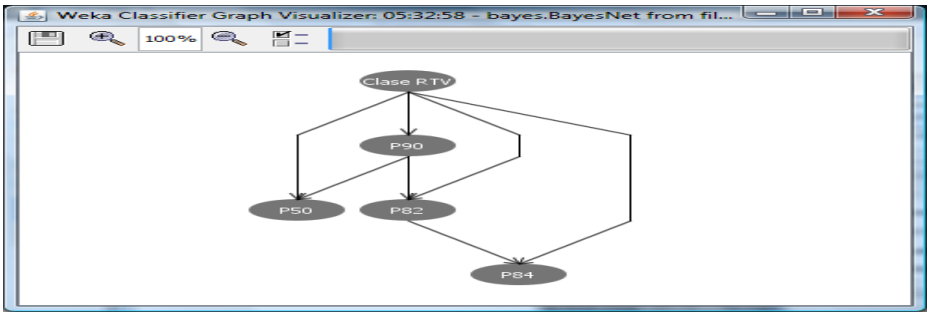

Fig. 4. Estructura TAN resultados de la serie 3 para el medicamento RTV.

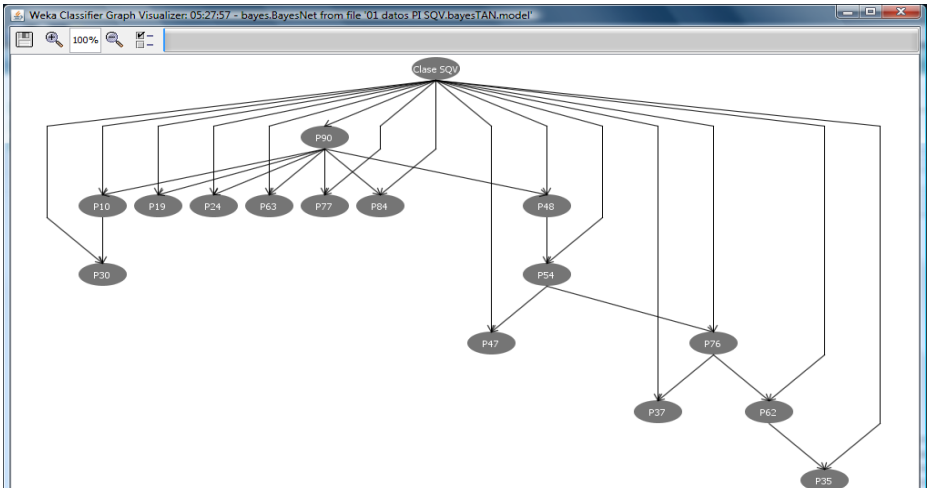

Fig. 5. Estructura TAN resultados de la serie 3 para el medicamento SQV.