

**METAHEURÍSTICA HÍBRIDA PARA LA IDENTIFICACIÓN DE MODELOS DE
AUTÓMATA CELULAR EN TRAYECTORIAS DE SIMULACIÓN DE PLEGAMIENTO DE
PROTEÍNA**

ANEXOS



Trabajo de Grado

Francisco Javier Obando Vidal

José Ricardo Gallego Garcés

Director: Mag. Néstor Milciades Díaz Mariño

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Departamento de Sistemas

Grupo de I+D en Tecnologías de la Información

Modelado y Simulación de Sistemas Complejos

Popayán, Agosto de 2012

TABLA DE CONTENIDO

1	ANEXO 1 - BUSQUEDA TABU PARA LA IDENTIFICACIÓN DE MODELOS DE AUTÓMATA CELULAR EN TRAYECTORIAS DE SIMULACIÓN DE PLEGAMIENTO DE PROTEÍNA.....	1
1.1	INTRODUCCIÓN.....	1
1.2	BÚSQUEDA TABÚ (TS)	1
1.2.1	<i>Solución inicial</i>	2
1.2.2	<i>Función de Evaluación:</i>	2
1.2.3	<i>Vecindario</i>	3
1.2.4	<i>Restricciones Tabú</i>	5
1.2.5	<i>Lista tabú</i>	5
1.2.6	<i>Criterio de aspiración</i>	6
1.2.7	<i>Diversificación</i>	6
1.2.8	<i>Intensificación</i>	7
1.2.9	<i>Condición de parada</i>	7
1.3	PRUEBA DE TS.....	7
1.4	CONSTRUCCIÓN DE MODELOS AC CON BÚSQUEDA TABÚ	9
1.5	ANÁLISIS DE MODELOS.....	13
1.6	CONCLUSIONES	15
1.7	REFERENCIAS Y BIBLIOGRAFIA.....	16
2	ANEXO 2 - RESULTADOS OBTENIDOS POR LAS METAHEURÍSTICAS HÍBRIDAS.....	17
2.1	RESULTADOS DE EVALUACIÓN DE LOS MODELOS OBTENIDOS POR LAS DIFERENTES TECNICAS HÍBRIDAS.....	17
2.2	EVALUACIÓN DE MODELOS OBTENIDOS MEDIANTE HÍBRIDACION POR MEJORA DE HIJOS.....	18
2.2.1	<i>Modelos individuo 13</i>	18
2.2.2	<i>Modelos individuo 32</i>	19
2.2.3	<i>Modelos individuo 33</i>	20
2.2.4	<i>Modelos individuo 50</i>	21
2.2.5	<i>Modelos individuo 53</i>	22
2.3	EVALUACIÓN DE MODELOS OBTENIDOS MEDIANTE HÍBRIDACION POR INTENSIFICACIÓN	23
2.3.1	<i>Modelos individuo 13</i>	24
2.3.2	<i>Modelos individuo 33</i>	25
2.3.3	<i>Modelos individuo 53</i>	26
2.4	OTROS RESULTADOS OBTENIDOS	27
	<i>Modelos individuo 13 por mejora de hijos</i>	27

2.4.1	<i>Modelos individuo 33 por mejora de hijos</i>	30
2.4.2	<i>Modelos individuo 53 por mejora de hijos</i>	32
2.5	REFERENCIAS Y BIBLIOGRAFIA.....	35

ÍNDICE DE TABLAS

TABLA 1. MEDIDAS DE CALIDAD DE LA EVALUACIÓN DE TS.....	14
TABLA 2. MEDIDAS DEL INDIVIDUO 13 DATASET 1.	18
TABLA 3. MATRIZ DE CONFUSIÓN INDIVIDUO 13 DATASET 1.....	19
TABLA 4. MEDIDAS DEL INDIVIDUO 32 DATASET 1.	19
TABLA 5. MATRIZ DE CONFUSIÓN INDIVIDUO 32 DATASET 1.....	20
TABLA 6. MEDIDAS DEL INDIVIDUO 33 DATASET 1.	21
TABLA 7. MATRIZ DE CONFUSIÓN INDIVIDUO 33 DATASET 1.....	21
TABLA 8. MEDIDAS DEL INDIVIDUO 50 DATASET 1.	22
TABLA 9. MATRIZ DE CONFUSIÓN INDIVIDUO 50 DATASET 1.....	22
TABLA 10. MEDIDAS DEL INDIVIDUO 53 DATASET 1.	23
TABLA 11. MATRIZ DE CONFUSIÓN INDIVIDUO 53 DATASET 1.	23
TABLA 12. MEDIDAS DEL INDIVIDUO 13 DATASET 1.	24
TABLA 13. MATRIZ DE CONFUSIÓN INDIVIDUO 13 DATASET 1.	24
TABLA 14. MEDIDAS DEL INDIVIDUO 33 DATASET 1.	25
TABLA 15. MATRIZ DE CONFUSIÓN INDIVIDUO 33 DATASET 1.	25
TABLA 16. MEDIDAS DEL INDIVIDUO 53 DATASET 1.	26
TABLA 17. MATRIZ DE CONFUSIÓN INDIVIDUO 53 DATASET 1.	26
TABLA 18. MEDIDAS DEL INDIVIDUO 13 CAMBIO EN V18.	28
TABLA 19. MATRIZ DE CONFUSIÓN INDIVIDUO13 CAMBIO EN V18.....	28
TABLA 20. MEDIDAS DEL INDIVIDUO 13 CAMBIO EN V49.	29
TABLA 21. MATRIZ DE CONFUSIÓN INDIVIDUO13 CAMBIO EN V49.....	29
TABLA 22. MEDIDAS DEL INDIVIDUO 33 CAMBIO EN V18.	30
TABLA 22. MATRIZ DE CONFUSIÓN INDIVIDUO 33 CAMBIO EN V18.....	30
TABLA 24. MEDIDAS DEL INDIVIDUO 33 CAMBIO EN V42.	31
TABLA 25. MATRIZ DE CONFUSIÓN INDIVIDUO 33 CAMBIO EN V42.....	31
TABLA 26. MEDIDAS DEL INDIVIDUO 53 CAMBIO EN V18.	32
TABLA 27. MATRIZ DE CONFUSIÓN INDIVIDUO 53 CAMBIO EN V18.....	33
TABLA 28. MEDIDAS DEL INDIVIDUO 53 CAMBIO EN V42.	33
TABLA 29. MATRIZ DE CONFUSIÓN INDIVIDUO 53 CAMBIO EN V42.....	34

ÍNDICE DE FIGURAS ANEXO 1

FIGURA 1. VECINDAD DE MOORE (ADAPTADO DE [1])	2
FIGURA 2. REPRESENTACIÓN DE LA SOLUCIÓN INICIAL.....	2
FIGURA 3. ECUACIÓN PARA EL CÁLCULO MCC (TOMADO DE [1]).	3
FIGURA 4. VECINDARIO EN AMPLITUD.....	4
FIGURA 5. VECINDARIO POR PROFUNDIDAD.	5
FIGURA 6.ESQUEMA ILAS EN UN CONJUNTO DE DATOS(TOMADO DE [1]).	8
FIGURA 7.GRAFICA ROC(ADAPTADO DE [1]).	8
FIGURA 8.COMPORTAMIENTO DE TS 60 ITERACIONES.....	10
FIGURA 9. COMPORTAMIENTO DE SOLUCIONES GLOBALES DE TS HASTA LA ITERACIÓN 60.....	10
FIGURA 10.COMPORTAMIENTO DE TS 128 ITERACIONES (ILAS 1, 3 Y 6).....	11
FIGURA 11.FUNCIÓN DE EVALUACIÓN Vs PROMEDIO.....	12
FIGURA 12.SOLUCIÓN ITERACIÓN 52 A) REPRESENTACIÓN SOLUCIÓN B) REPRESENTACIÓN DEL VECINDARIO OBTENIDO POR TS C) VECINDARIO DE MOORE DE LA SOLUCIÓN.....	13
FIGURA 13 ÁRBOL DE DECISIÓN MODELO AC.....	15
FIGURA 14.EJEMPLO DE PLANTILLA DE MEDIDAS DE CALIDAD DE UN MODELO AC.....	17
FIGURA 15.FENOTIPO INDIVIDUO 13 POR MEJORA DE HIJOS.....	18
FIGURA 16.FENOTIPO INDIVIDUO 32 POR MEJORA DE HIJOS.....	19
FIGURA 17.FENOTIPO INDIVIDUO 33 POR MEJORA DE HIJOS.....	20
FIGURA 18.FENOTIPO INDIVIDUO 50 POR MEJORA DE HIJOS.....	21
FIGURA 19.FENOTIPO INDIVIDUO 53 POR MEJORA DE HIJOS.....	22
FIGURA 20.FENOTIPO INDIVIDUO 13 POR INTENSIFICACIÓN.....	24
FIGURA 21.FENOTIPO INDIVIDUO 33 POR INTENSIFICACIÓN.....	25
FIGURA 22.FENOTIPO INDIVIDUO 53 POR INTENSIFICACIÓN.....	26
FIGURA 23. VECINDAD DE MOORE (ADAPTADO DE [3]).....	27
FIGURA 24. FENOTIPO DEL INDIVIDUO 13 CON CAMBIO EN LA CELDA V18.....	28
FIGURA 25. FENOTIPO DEL INDIVIDUO 13 CON CAMBIO EN LA CELDA V49.....	29
FIGURA 26. FENOTIPO DEL INDIVIDUO 33 CON CAMBIO EN LA CELDA V18.....	30
FIGURA 27. FENOTIPO DEL INDIVIDUO 33 CON CAMBIO EN LA CELDA V42.....	31
FIGURA 28. FENOTIPO DEL INDIVIDUO 53 CON CAMBIO EN LA CELDA V18.....	32
FIGURA 29. FENOTIPO DEL INDIVIDUO 53 CON CAMBIO EN LA CELDA V42.....	33
FIGURA 30. VECINDAD POR FIJA O POR PRIORIDAD.....	34

1 ANEXO 1 - BUSQUEDA TABU PARA LA IDENTIFICACIÓN DE MODELOS DE AUTÓMATA CELULAR EN TRAYECTORIAS DE SIMULACIÓN DE PLEGAMIENTO DE PROTEÍNA

1.1 INTRODUCCIÓN

La búsqueda tabú (TS) aquí propuesta, busca obtener modelos simulación basados en Autómatas Celulares(AC) que repliquen la trayectoria de simulación de plegamiento de proteínas con una precisión del 90%, explorando con esta técnica una nueva alternativa a la planteada en [1], en donde se realiza este experimento con un Algoritmo Genético obteniendo modelos de AC de alta precisión.

1.2 BÚSQUEDA TABÚ (TS)

La búsqueda Tabú busca solucionar problemas de optimización combinatoria, TS está basado en el principio de que es mejor una mala elección basada en información a una buena elección basada en el azar, ya que una mala elección basada en una estrategia proporciona mejor información para la búsqueda, a diferencia de una buena elección al azar que no proporciona ninguna información [2-3].

TS busca superar el estancamiento del optimo local, a través del uso de memoria adaptativa la cual se conoce como lista tabú [4-6]. Éste uso de memoria evita a la técnica volver a soluciones previamente exploradas y de esta forma evitar ciclos. Los modelos de AC están constituidos por estados, lattice, condición de frontera, vecindad y reglas, Siendo las entradas para TS en donde éstas entradas están parcialmente definidas con los siguientes elementos:

- Lattice: $N \times N$ (N es numero de AA de la proteína).
- Condición de Frontera: Finita.
- Estados: (0, 1), 0 para no contacto y 1 para contacto.

TS completara los modelos parcialmente definidos, buscando vecindades para extraer las reglas de AC que complementaran los modelos, por lo cual se define una vecindad de Moore como la máxima vecindad correspondiente al espacio de búsqueda definido, esta vecindad es de 49 celdas, como muestra la figura 1.

V1	V2	V3	V4	V5	V6	V7
V8	V9	V10	V11	V12	V13	V14
V15	V16	V17	V18	V19	V20	V21
V22	V23	V24	V25	V26	V27	V28
V29	V30	V31	V32	V33	V34	V35
V36	V37	V38	V39	V40	V41	V42
V43	V44	V45	V46	V47	V48	V49

Figura 1. Vecindad de Moore (Adaptado de [1])

Para la implementación de TS que identifique las vecindades y reglas de los modelos de AC se tiene en cuenta las siguientes características:

1.2.1 Solución inicial

La solución inicial está representada por una cadena binaria de longitud 49, en donde cada posición de la cadena puede tomar un valor de 1 ó 0, indicando la inclusión o no inclusión de la celda en la vecindad de modelo AC. La solución inicial puede ser previamente dada o puede ser generada aleatoriamente, la figura 2 muestra un ejemplo de solución inicial.

0000101101000000101010100101001101001001101010011

Figura 2. Representación de la Solución Inicial.

Cada posición de la solución inicial se puede representar en la vecindad de more de la figura 1, en donde cada posición de la solución inicial equivale a una celda de la vecindad.

1.2.2 Función de Evaluación:

La solución inicial será evaluada de la misma forma que en [1], en donde se usa el Coeficiente de Correlación de Matthews(MCC), esta medida no se ve afectada por sesgos de clase tomando valores entre -1 y 1, donde los valores más cercanos a 1 indica buenas predicciones, es decir los verdaderos positivos y los verdaderos

negativos son clasificados correctamente. En la figura 3 muestra la ecuación del cálculo para MCC.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}}$$

Figura 3. Ecuación para el cálculo MCC (Tomado de [1]).

1.2.3 Vecindario

Un vecindario o entorno es un conjunto de soluciones vecinas que se generan a partir de la solución inicial o actual de cada iteración, realizando transformaciones locales llamadas movimiento, que permiten dirigir la búsqueda hacia soluciones prometedoras. Los movimientos son generalmente simples pueden ser guiados o al azar realizando intercambio de una o varias celdas.

$N(S) = \{\text{soluciones obtenidas mediante la aplicación de una sola transformación local de } S\}$.

Para nuestro problema se aplicaron dos tipos de generación de estructura:

- **Generación de entorno por anchura:** Dada una solución actual (cadena binaria de 49 bits) se realiza un cambio de cada bit de la solución actual, explorando todas las posibles combinaciones de esta, en donde si la celda contiene 1 se cambia por 0, o si es 0 se cambia por 1, esto quiere decir que el vecindario estaría compuesto por 49 posibles soluciones.

Para todo solución X, se intercambia x_i donde $i = 1, 2, 3 \dots 49$, donde Si $x_i = 1$ se cambia por 0, si $x_i = 0$ se cambia por 1.

La figura 4 muestra la generación de un vecindario a partir de una solución actual, en este caso la cadena que está marcada con 1, en donde una nueva solución se genera por el cambio de un bit(color más oscuro) y la línea punteada gruesa representa todo vecindario de la solución actual. El paso siguiente es escoger la mejor solución de vecindario para este ejemplo la que se encuentra debajo de numero 2, a partir de la cual se genera el nuevo vecindario. la línea que regresa (3) indica que se puede regresar a la solución anterior, pero en el concepto de TS ese movimiento es tabú, lo que evitaría regresar a esa solución anterior ya explorada y de esta forma prevenir los ciclos en la búsqueda.

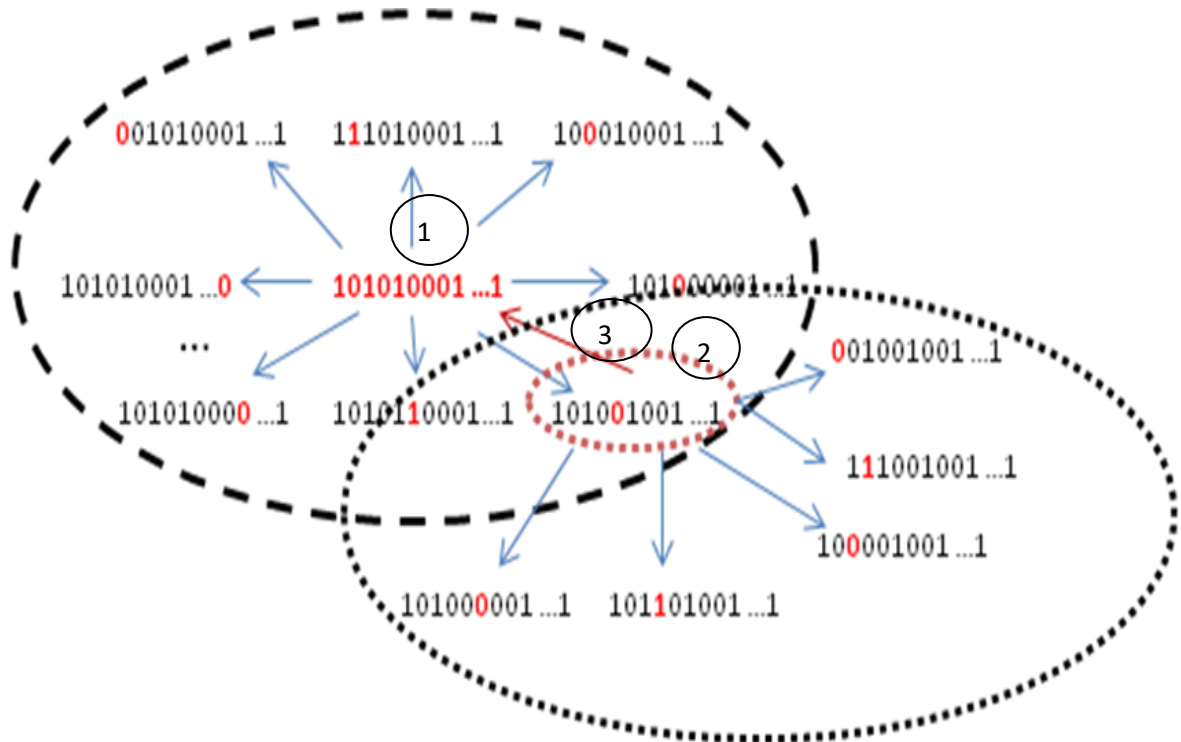


Figura 4. Vecindario en Amplitud.

La desventaja de este enfoque es que debo evaluar todas las posibles configuraciones, lo que provoca mayor tiempo de cómputo.

- **Generación de entorno por aleatorio:** Se define un número N de genes a cambiar del mejor individuo, se realizan un cambio aleatorio de genes dependiendo del tamaño N en donde si el gen es 1 se cambia por 0 y viceversa. Esto quiere decir que si el número N es 5 el vecindario estaría compuesto por 5 nuevos individuos. Además se define un número de rondas R , que indica el número de veces que se repite este proceso, evitando aquellos movimientos ya escogidos.

La Figura 5 muestra un ejemplo del vecindario en amplitud en donde se realizaran 5 rondas con 3 individuos a evaluar, el proceso inicia generando el vecindario del individuo escogido, se evalúa el vecindario y se escoge el mejor vecino (círculo rojo), esto me indica una ronda y se repite este proceso hasta llegar al máximo número de rondas a realizar.

Total de rondas = 5
 Rondas por nivel = 2
 Individuos evaluados por ronda = 3

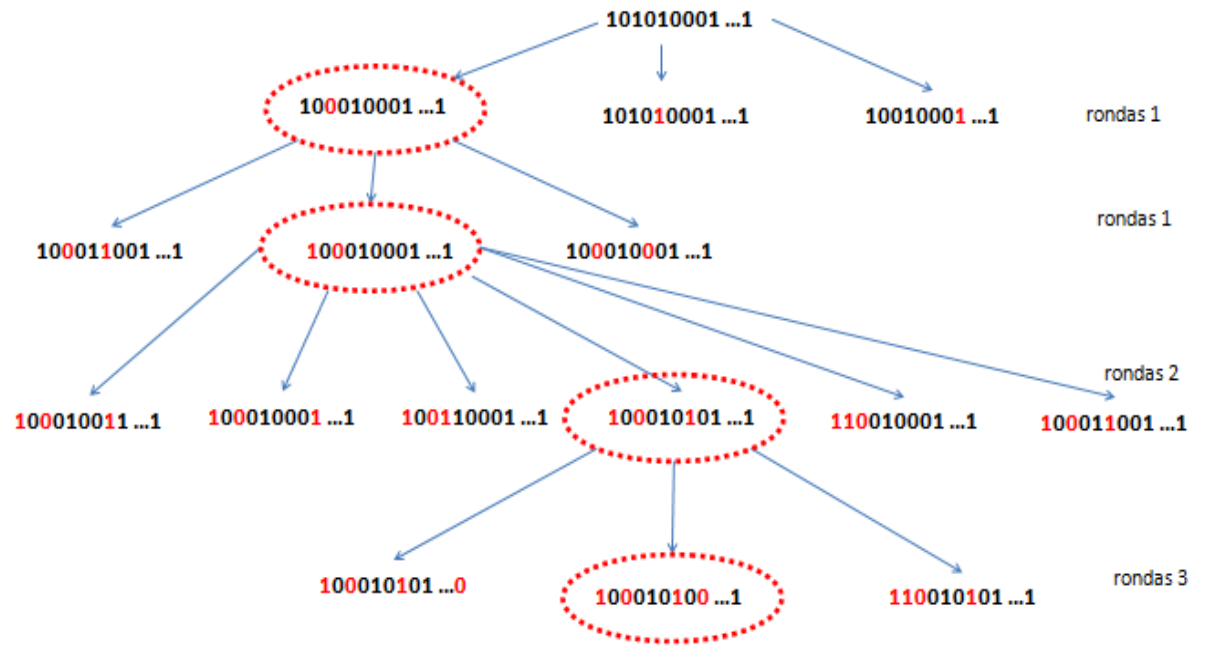


Figura 5. Vecindario por Profundidad.

1.2.4 Restricciones Tabú

El concepto de tabú (Usado comúnmente como marca de prohibido) es un distintivo de TS que se usa para prevenir ciclos en el instante que no hayan mejoras en la búsqueda, una restricción tabú indica que ya se buscó en ese sitio y no arrojó ninguna información prometedora en la búsqueda, por lo general estos tabúes se almacenan en la lista tabú, estos tabúes pueden ser movimientos o soluciones anteriormente visitadas [7].

1.2.5 Lista tabú

Es una lista donde se guardan estructuras individuos o movimientos a vetar evitando que se repitan y se caiga en ciclos dentro de la búsqueda, se debe tener cuidado al seleccionar el tamaño de la lista tabú, ya que un tamaño pequeño no cumpliría el objetivo principal de la lista tabú que es la de evitar los ciclos dentro

de la búsqueda y si es grande crearía podas que puedan dejar por fuera buenas soluciones.

Se definieron dos tipos de lista tabú

- **Lista tabú estática:** con un tamaño comprendido entre el 10% y el 20% del tamaño de la solución inicial, o la raíz cuadrada del tamaño de a solución actual.
- **Lista tabú dinámica:** inicialmente se empieza con un tamaño del 10% del tamaño de la solución actual, luego en cada iteración se incrementa el tamaño de la lista en uno hasta que la lista dinámica tengo un tamaño igual al 20% del tamaño de la solución ó cuando en el proceso de la búsqueda tabú se encuentre una solución mejor a la solución que se tiene hasta ese momento, en este caso la lista dinámica se volverá a asignar el valor del 10% del tamaño de la solución, la lista dinámica incrementara máximo hasta alcanzar el tamaño del 20% del tamaño de la solución.

1.2.6 Criterio de aspiración

Sirve para sacar elementos de la lista tabú a partir de un criterio establecido en el proceso de TS, en nuestro caso se estableció como criterio de aspiración los elementos de la lista tabú que al evaluarlos su evaluación sea superior a la evaluación de la mejor solución obtenida hasta ese momento, aquellos elementos que cumplen con esta condición son sacados de lista tabú.

1.2.7 Diversificación

Es el proceso por el cual TS trata de escapar de óptimos locales dándole prioridad a caminos poco explorados, este proceso debe ejecutarse después si la mejor solución encontrada no cambia en un número determinado de iteraciones (posible estancamiento), cuando esto sucede se dispara por un número pequeño de iteraciones (10) que penalice aquellos movimientos que han salido varias veces y premie aquellos movimientos que no han salido o que han salido pocas veces.

1.2.8 Intensificación

Es el proceso en el cual se intensifica o se explota mas la búsqueda en espacios que son muy prometedores, en nuestro caso de estudio después de terminada 20 iteraciones se verifica las tres soluciones con mejores función de evaluación y se intensifica en el estrato en el cual se encontró dicha soluciones.

Si al terminar la intensificación se encuentra una solución con una función de evaluación mejor a la solución global se hace el cambio y se sigue explorando mas caminos a partir de dicha solución, en caso contrario el proceso sigue su curso normal.

1.2.9 Condición de parada

Determina el momento en el cual ya no continuará ejecutando el TS. Como condición de parada se han definido los siguientes criterios:

- Máximo número de iteraciones = 300.
- Máximo número de iteraciones sin que la solución global sea cambiada = 100.
- Máximo valor en el score = 1.0

1.3 PRUEBA DE TS

Las características de diseño de prueba de [1], serán aplicadas de forma similar para TS. Se dividirán las evidencias en 10 conjuntos de 4.000 pasos de simulación por dinámica molecular (Una trayectoria tiene 40.000 pasos de simulación), donde un conjunto de datos será usado para la ejecución del TS. Al usar el enfoque de aprendizaje iterativo con estratos alternantes(ILAS)[8], se evita un sobre entrenamiento de las soluciones del TS al tener que enfrentarse a menor cantidad de datos y de esta forma reducir el costo computacional para cada iteración de TS. El enfoque ILAS se divide el conjunto de datos en S estratos de igual tamaño, es decir para un conjunto de datos (4.000 pasos de simulación), si el estrato tiene un tamaño de 200 se obtendrían 20 estratos, en donde el ultimo estrato de datos le sigue al primero como muestra la figura 6.

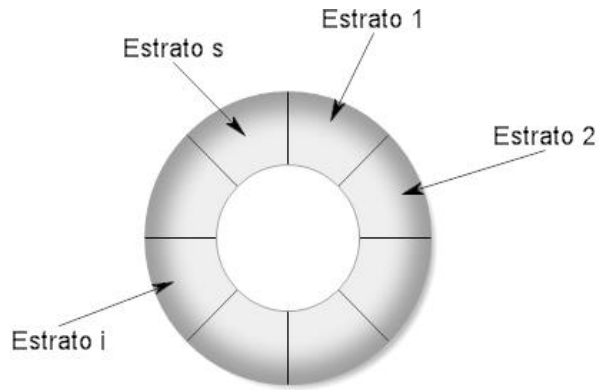


Figura 6. Esquema ILAS en un conjunto de datos (Tomado de [1]).

Para verificar el desempeño de los modelos obtenidos para TS, se usará un gráfico ROC [9], el cual se caracteriza por no verse afectado por sesgo de clase, este gráfico representa la tasa de falsos positivos (TFP)¹ de un predictor versus la tasa de verdaderos positivos (TVP)².

La figura 7 muestra el gráfico ROC que indica el mejor predictor (0,1) y el peor predictor (1,0) de los modelos obtenidos, en donde el valor cercano a 1 en TVP y cercano a 0 TFP indica una buena predicción y en el caso contrario, el valor cercano a 0 en TVP y cercano a 1 en TFP indica una mala predicción.

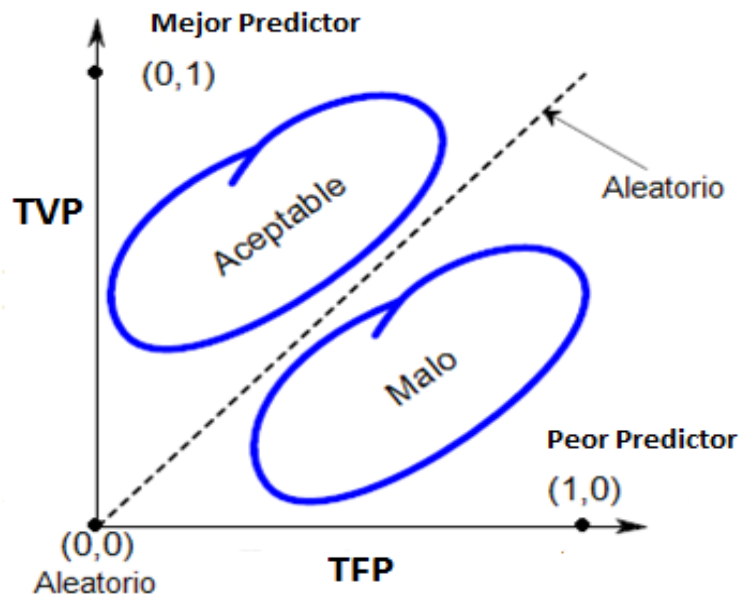


Figura 7. Gráfica ROC (Adaptado de [1]).

¹ Tasa de Falsos Positivos: ésta dada por $FP/(FP+VN)$; donde FP son los Falsos Positivos y VN los verdaderos Negativos, del predictor.

² Tasa de Verdaderos Positivos: ésta dada por $VP/(VP+FN)$; donde VP los Verdaderos Positivos y FN son los Falsos Negativos, del predictor.

Se generara un reporte de las medidas sensibilidad o VFP, especificidad o TFP y precisión³, para verificar la calidad de los modelos obtenidos por TS.

1.4 CONSTRUCCIÓN DE MODELOS AC CON BÚSQUEDA TABÚ

La selección de parámetros correspondiente para encontrar lo modelos de AC con Búsqueda Tabú se presentan y se describen a continuación.

- Longitud de la solución: 49
- Bits : {0,1}
- Generar vecindarios
- Movimiento: Inversión de bit.
- Tamaño del vecindario: 49
- Lista Tabú: 7.
- Diversificación Inicial: 20 (número máximo de iteraciones sin cambios en la solución global).
- Diversificación final: 7 (número máximo de iteraciones en el cual se aplica el criterio de diversificación)
- Intensificación: 10 (numero de iteraciones que se profundiza)
- Condición de parada forzosa: 40
- Máximo número de iteraciones: 200
- Función de Evaluación: Coeficiente de Correlación de Matthews – penalización por cantidad de reglas.
- Tamaño del estrato ILAS: 200 pasos de simulación.

La Figura 8 presenta un ejemplo del comportamiento que se obtiene para TS sobre un conjunto de datos de 4000 pasos de simulación, con un tamaño de estrato de 200 pasos, se observa una tendencia a obtener mejores soluciones a medida que pasa el tiempo de iteración. Además se observa cómo se aplica el criterio de aspiración mejorando la solución actual en los iteraciones 28 y 52, mejorando las soluciones.

El mejor valor de la función de evaluación se encontró en la iteración 52, con un MCC de 0.935078156; siendo esta solución la candidata a las medidas de calidad de evaluación.

³ Precisión: esta dada por $(VP + VN) / (P + N)$, en donde VP son lo verdaderos Positivos, VN los verdaderos Negativos, P las clases positivas y N las clases Negativas.

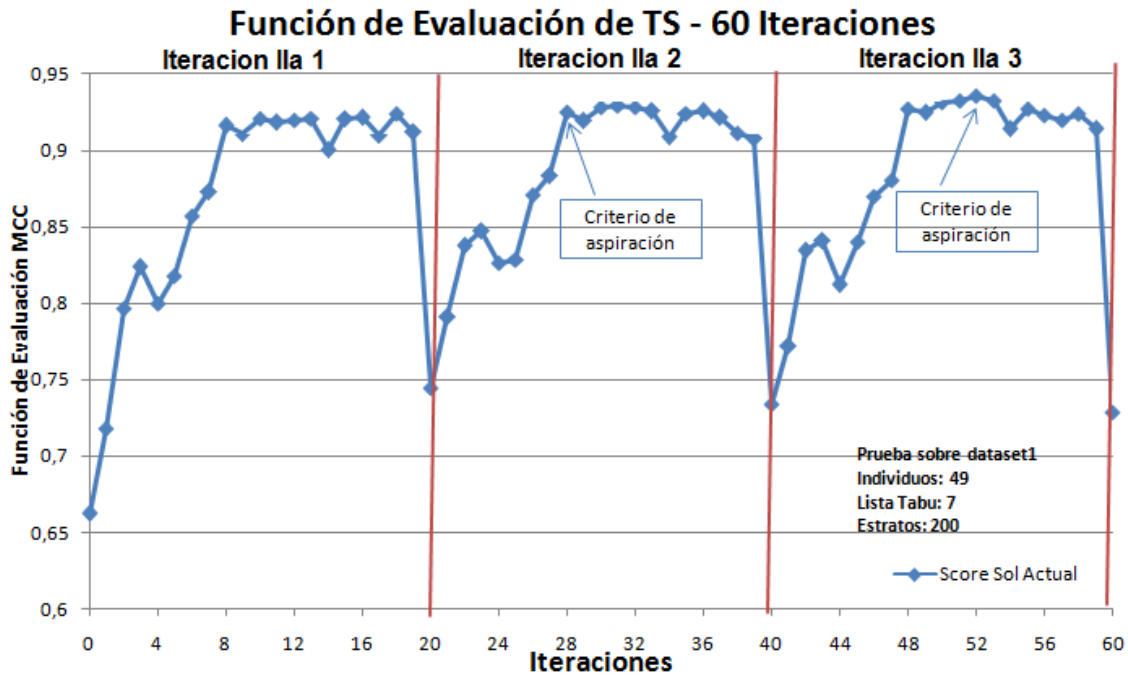


Figura 8. Comportamiento de TS 60 Iteraciones.

La figura 9 muestra el grafico del comportamiento de la función global de TS, en donde esta es reemplazada cada vez que se encuentra una mejor solución actual, en este caso se observa que a medida que pasa el tiempo se obtienen mejores resultados hasta la iteración 52 que es la mejor solución actual obtenida, reemplazando la solución global por esta.

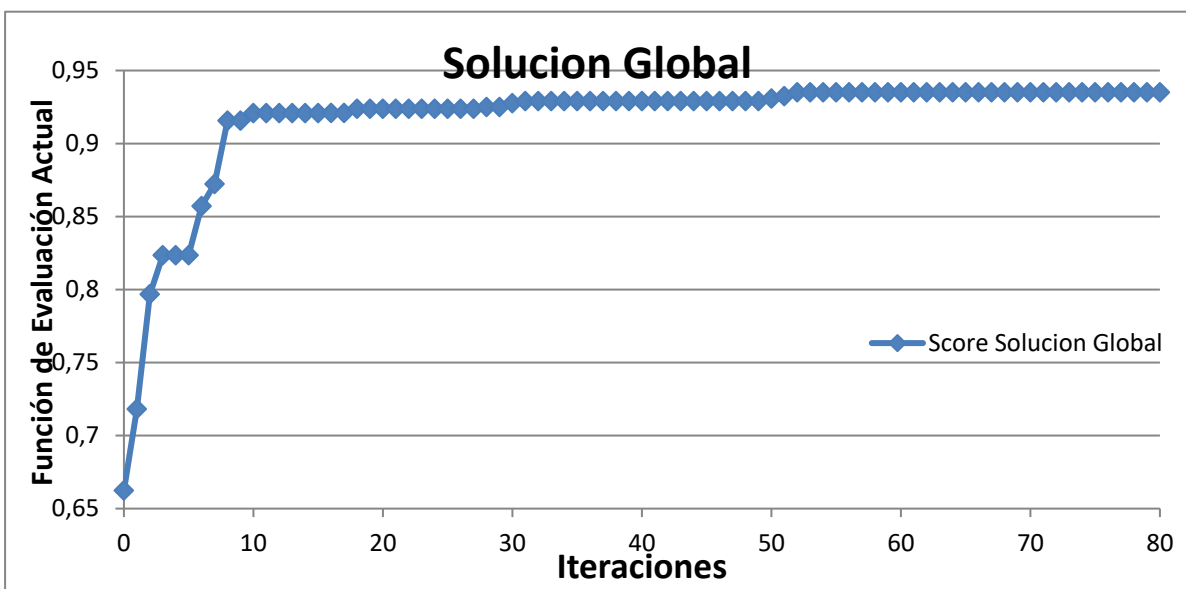


Figura 9. Comportamiento de soluciones globales de TS hasta la iteración 60.

Para el dataset 1 (primeros 4.000 pasos de simulación de la trayectoria original) se realizó una ejecución con 128 iteraciones equivalentes a 6 iteraciones ILAS con un tamaño de estrato de 200 pasos de simulación. La Figura 10 muestra la grafica del comportamiento para las ILAS 1, 3 y 6 de las mejores soluciones de cada iteración de TS. El comportamiento oscilatorio en los valores de la función de evaluación en cada iteración se debe al hecho que cada en cada iteración se utiliza un estrato de datos diferente y que cada veinte generaciones se vuelve a repetir el mismo estrato.

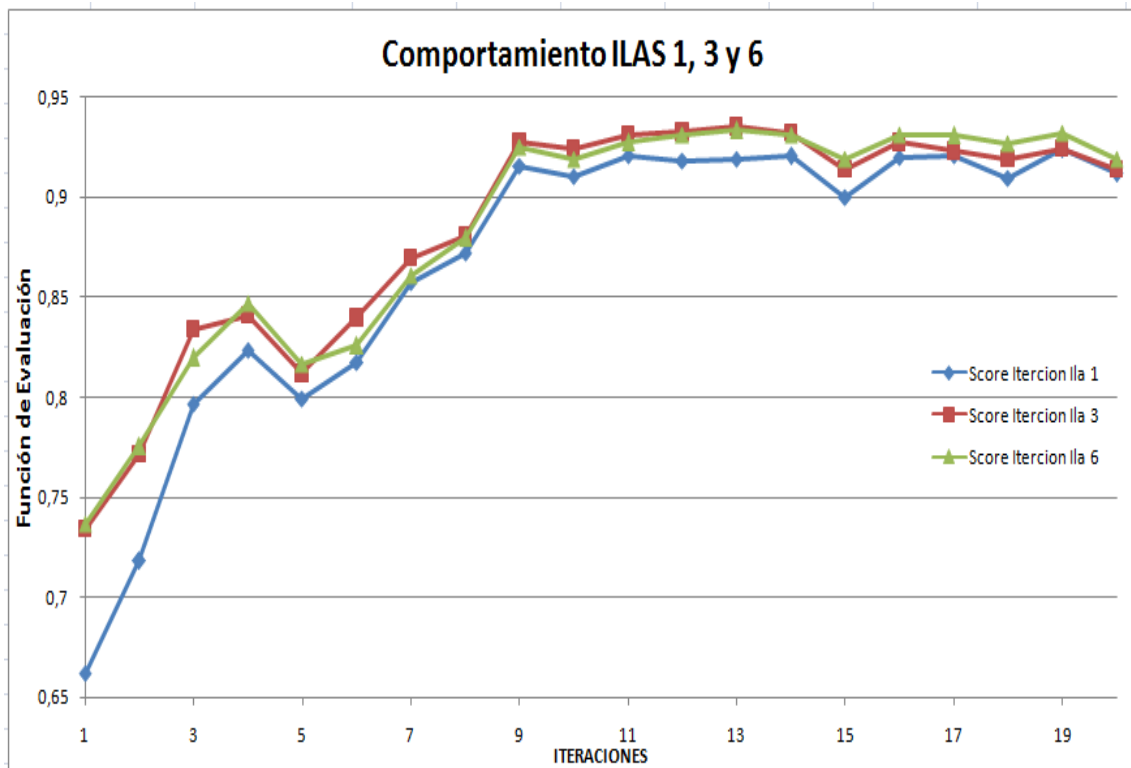


Figura 10. Comportamiento de TS 128 Iteraciones (ILAS 1, 3 y 6).

La figura 11 muestra la grafica de comportamiento para la función de evaluación y la media acumulada, en donde esta ultima va creciendo poco a poco a medida que pasa el tiempo. Se muestran 120 iteraciones equivalentes a 6 iteraciones ILAS.

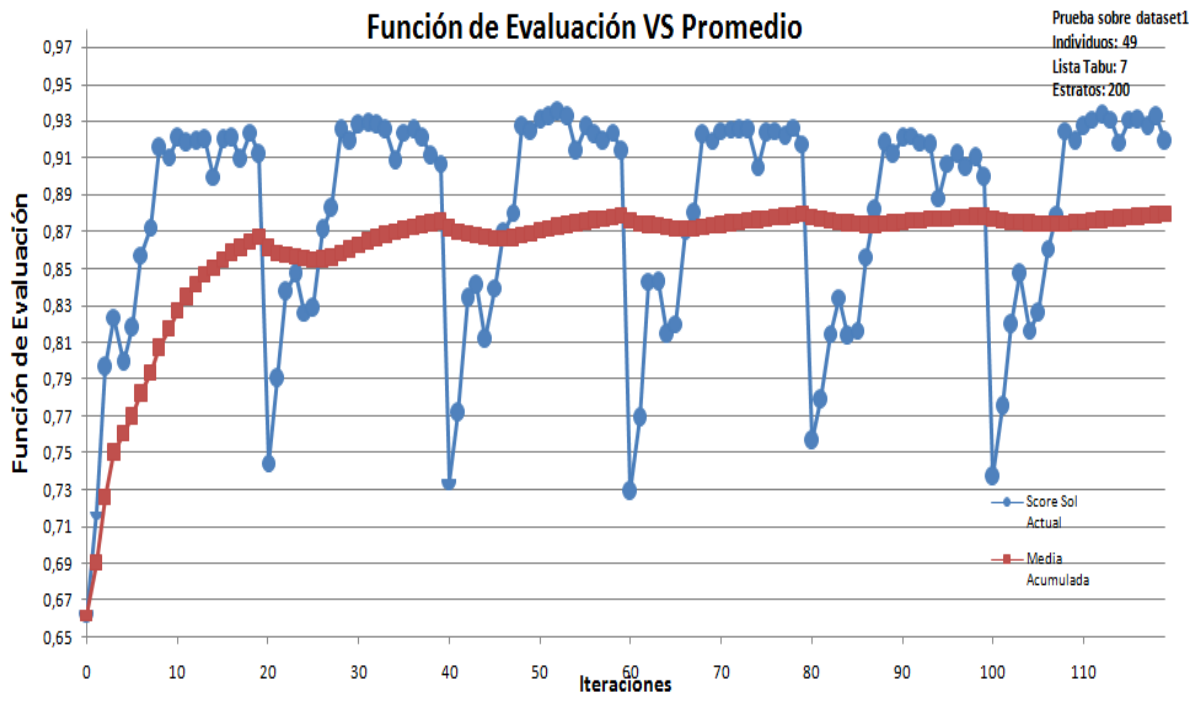


Figura 11. Función de evaluación Vs Promedio.

1.5 ANÁLISIS DE MODELOS

Para el conjunto de datos 1, que incluye los mapas de contacto correspondientes a los primeros 4.000 pasos de simulación, como se dijo anteriormente se destaca el comportamiento de la solución de la iteración 52, que obtiene el score más alto, siendo el candidato para la evaluación. La figura 12 muestra la solución y la representación de la vecindad de moore para el modelo de la iteración 52.

a) **0000001101001001110010000100001001000011101001011**

b)

0	0	0	0	0	0	1
1	0	1	0	0	1	0
0	1	1	1	0	0	1
0	0	0	0	1	0	0
0	0	1	0	0	1	0
0	0	0	1	1	1	0
1	0	0	1	0	1	1

c)

V1	V2	V3	V4	V5	V6	V7
V8	V9	V10	V11	V12	V13	V14
V15	V16	V17	V18	V19	V20	V21
V22	V23	V24	V25	V26	V27	V28
V29	V30	V31	V32	V33	V34	V35
V36	V37	V38	V39	V40	V41	V42
V43	V44	V45	V46	V47	V48	V49

Figura 12. Solución iteración 52 a) Representación Solución b) Representación del vecindario obtenido por TS c) Vecindario de Moore de la solución.

Para el dataset 1 los resultados obtenidos permiten concluir que el TS propuesto puede explicar la trayectoria de simulación de plegamiento de proteínas, ya que su evaluación de resultados son buenas:

- Precisión = 0.9282
- Coeficiente de Correlación de Matthews = 0.6817
- Especificidad = 0.97225
- Sensitividad = 0.68177

La tabla 1 muestra las medidas de calidad de la solución para 9 datasets, excluyendo el dataset 1.

Tabla 1. Medidas de calidad de la Evaluación de TS

DATASET	FPR o TFP	TPR o TVP	MCC	ACC
2	0,82378197	0,02032841	0,82352773	0,956894082
3	0,84875412	0,02592438	0,81914066	0,956788776
4	0,8401112	0,02212487	0,82790754	0,958022449
5	0,74863303	0,02938661	0,73727379	0,941549184
6	0,87705341	0,0204213	0,85688302	0,964903061
7	0,84197156	0,01549693	0,85046405	0,96396102
8	0,79951608	0,02077825	0,8053792	0,953333061
9	0,77455948	0,02319126	0,7809744	0,947627143
10	0,81290241	0,0244739	0,80296612	0,951938367

Se observa que este modelo obtiene una precisión superior al 94% destacándose el comportamiento de este modelo para los dataset 6 y 7.

A continuación se presentan las matrices de confusión de los datasets 6 y 7 los cuales obtuvieron los valores más altos de en las evaluaciones.

Matriz de confusión Dataset 6

VP: 615160	FP: 85741	P': 700901
FN: 86234	VN: 4112865	N': 4199099
P : 701394	N: 4198606	4900000

Matriz de confusión Dataset 7

VP: 594602	FP: 64991	P': 659593
FN: 111600	VN: 4128807	N': 4240407
P: 706202	N: 4193798	4900000

La figura 13 muestra el árbol de decisión de las reglas del modelo AC obtenido por TS, para la trayectoria. Con la celda de análisis V17.

Classification Tree	Class	P(Class)	P(Target)	# Inst
└─ <root>	0	0.731	0.731	167359
├─ V17 = 0	0	0.950	0.950	121784
│ └─ V26 = 0	0	0.986	0.986	112125
│ └─ V26 = 1	0	0.550	0.550	9427
│ │ └─ V43 = 0	0	0.703	0.703	6409
│ │ └─ V43 = 1	1	0.841	0.159	2074
│ │ └─ V43 = 3	1	0.541	0.459	736
│ │ │ └─ V31 = 0	0	0.994	0.994	173
│ │ │ └─ V31 = 1	1	0.705	0.295	563
│ │ │ │ └─ V34 = 0	1	1.000	0.000	304
│ │ │ │ └─ V34 = 1	0	0.641	0.641	259
│ │ │ └─ V43 = 4	1	0.938	0.062	208
│ │ └─ V26 = 3	1	0.853	0.147	232
├─ V17 = 1	1	0.771	0.229	28660
│ └─ V41 = 0	0	0.537	0.537	10814
│ └─ V18 = 0	0	0.802	0.802	3877
│ │ └─ V46 = 0	0	0.954	0.954	2238
│ │ │ └─ V31 = 0	0	0.993	0.993	1940
│ │ │ └─ V31 = 1	0	0.705	0.705	298
│ │ │ │ └─ V16 = 0	1	0.793	0.207	111
│ │ │ │ └─ V39 = 0	1	0.822	0.178	107
│ │ │ │ └─ V39 = 1	0	1.000	1.000	4
│ │ │ └─ V16 = 1	0	1.000	1.000	187
│ │ └─ V46 = 1	0	1.000	1.000	372
│ │ └─ V46 = 3	1	0.692	0.308	767
│ │ └─ V46 = 4	0	0.728	0.728	500
│ │ │ └─ V43 = 0	0	0.989	0.989	363
│ │ │ └─ V43 = 1	1	0.964	0.036	137
│ │ └─ V18 = 1	1	0.598	0.402	6705
│ │ └─ V18 = 3	1	1.000	0.000	232
│ └─ V41 = 1	1	0.958	0.042	17846
└─ V17 = 3	3	1.000	0.000	11144
└─ V17 = 4	4	1.000	0.000	5771

Figura 13 Árbol de decisión modelo AC.

1.6 CONCLUSIONES

Los modelos de AC obtenidos por las técnicas TS alcanzan una precisión superior al 92% para los primeros 4.000 pasos de simulación siendo este el segmento más pesado por la trayectoria de plegamiento de proteínas.

Las vecindades de los modelos AC, tienden a tener tamaños pequeños. El tamaño de la vecindad implica el numero de reglas a analizar de tal forma que a mayor cantidad de vecinos mayor el número de reglas y a menor cantidad de vecinos, menor es el numero de reglas.

Las técnicas de aprendizaje de maquina completan el modelo AC identificando la vecindad y las reglas del modelo AC.

La búsqueda Tabú utilizada es converge con rapidez, haciéndolo candidato para una futura hibridación con otra técnica metaheurística.

1.7 REFERENCIAS Y BIBLIOGRAFIA

- [1] N. Diaz, "Metodología y framework computacional para el diseño inverso de modelos de autómatas celulares de secuencias cortas de aminoácidos soportando en un proceso de minería de datos," *Tesis de Maestría, Universidad del Valle*, p. 56, 2010.
- [2] K. E. N. I. Maeda and C. Sakama, "Identifying cellular automata rules," *Journal of Cellular Automata*, vol. 2, pp. 1–20, 2007.
- [3] F. Glover, "Tabu Search-Part I,II," *Orsa Journal on Computing*, vol. 3, p. 17, 1989.
- [4] S. Luke, *Essentials of metaheuristics*, 2010.
- [5] J. Kari, "Theory of cellular automata: A survey," *Theoretical computer science*, vol. 334, pp. 3-33, 2005.
- [6] O. Lozano, "Búsqueda Tabú (tabu search)," <http://dis.unal.edu.co/~fgonza/courses/2003/pmge/present/tabusearch.pdf>:Accesado, p. 5, 2003.
- [7] M. Gendreau and J.-Y. Potvin, *Handbook of Metaheuristics* vol. 146: Springer, 2010.
- [8] J. Bacardit, *et al.*, "Speeding-up pittsburgh learning classifier systems: Modeling time and accuracy," 2004, pp. 1021-1031.
- [9] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clin Chem*, vol. 39, pp. 561-577, 1993.
- [10] J. C. Pete Chapman, Randy Kerber,, *et al.*, "CRISP-DM Step-by-step data mining guide," p. 78, 2000.

2 ANEXO 2 - RESULTADOS OBTENIDOS POR LAS METAHEURÍSTICAS HÍBRIDAS

En este apartado se describen y se analizan los resultados obtenidos por las metaheurísticas híbridas.

2.1 RESULTADOS DE EVALUACIÓN DE LOS MODELOS OBTENIDOS POR LAS DIFERENTES TÉCNICAS HÍBRIDAS

A continuación se presentan las diferentes tablas de resultados de las evaluaciones realizadas para cada uno de los modelos obtenidos por las técnicas híbridas.

La figura 14, muestra la plantilla que contiene las medidas de un modelo evaluado.

```
C:\models-ContactMaps\Dataset1\frame09.pdb\model0399.cm

Tamaño de las Evidencias: 4000
Datos de la Solucion a Evaluar
Nombre: 423304800
Solucion: 0000001101001001110010000100001001000011101001011 Sco
Representacion: [0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1,
Reglas: 1681
{'11010001111101300': {'1': 4}, '000000000001004100': {'0': 33}, '1--0
Vecindario: 18
[(-3, 3), (-2, -3), (-2, -1), (-2, 2), (-1, -2), (-1, -1), (-1, 0), (-1, 3)
Estados: ['0', '1', '3', '4']
Condicion de Frontera: finite
Colores: {'1': (187, 175, 122), '0': (142, 220, 123), '3': (212,

Datos Estadisticos
Verdaderos Positivos: 449063
Verdaderos Negativos: 4099096
Falsos Positivos: 116980
Falsos Negativos: 234861
sensitivity: 0.656597809113
specificity: 0.972253820851
Accuracy: 0.928195714286
PPV: 0.793337255297
MCC: 0.68177224549

Matriz de Confusion
VP = 449063 FP = 116980 P' = 566043
FN = 234861 VN = 4099096 N' = 4333957
P = 683924 N = 4216076 Total = 4900000
```

Figura 14. Ejemplo de plantilla de medidas de Calidad de un Modelo AC.

En donde se puede observar los datos estadísticos como Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP), Falsos Negativos (FN), Especificidad, Sensitividad, Precisión (ACC), Coeficiente de Correlación de Matthews (MCC), Valor predictivo Positivo (PPV) y la matriz de confusión [1-2]. Los modelos que se tienen en cuenta para el híbrido por mejora de hijos son los obtenidos en las generaciones 13, 32, 33, 50 y 53, que se destacan por sus valores de la función de evaluación.

2.2 EVALUACIÓN DE MODELOS OBTENIDOS MEDIANTE HÍBRIDACION POR MEJORA DE HIJOS

2.2.1 Modelos individuo 13

La figura 15 muestra el fenotipo para el individuo 13 del híbrido por mejora de hijos, el tamaño de la vecindad de este modelo AC es de 20.

0	0	0	0	0	0	1
0	1	1	0	0	1	0
0	0	1	0	1	0	1
1	1	0	0	1	0	0
0	0	1	1	0	1	1
0	1	0	0	0	1	1
1	0	0	1	0	1	0

Figura 15. Fenotipo Individuo 13 por mejora de hijos.

Las medidas de calidad para éste individuo para el dataset 1 se muestran a continuación en la tabla 2.

Tabla 2. Medidas del Individuo 13 dataset 1.

Dataset 1
Verdaderos Positivos: 430073
Verdaderos Negativos: 4157709
Falsos Positivos: 58367
Falsos Negativos: 253851
sensitivity: 0.62883156608
specificity: 0.986156084473

Accuracy: 0.936282040816
PPV: 0.880503234788
MCC: 0.711414685578

La tabla 3 muestra el cálculo de las medidas de evaluación desde la matriz de confusión.

Tabla 3. Matriz de confusión individuo 13 Dataset 1.

Matriz de confusión

430073	58367	488440
253851	4157709	4411560
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01384392	0,62883157	0,71141469	0,93628204	0,986156084	0,88050323	0,94245777	0,11949677

2.2.2 Modelos individuo 32

La figura 16 muestra el fenotipo para el individuo 32 del híbrido por mejora de hijos, el tamaño de la vecindad de este modelo AC es de 19.

0	0	0	0	0	0	1
0	0	1	0	0	1	0
0	0	1	0	1	0	0
0	1	1	0	1	0	1
0	0	1	0	0	1	0
1	0	0	0	1	1	1
1	0	0	1	0	1	1

Figura 16. Fenotipo Individuo 32 por mejora de hijos.

Tabla 4. Medidas del Individuo 32 dataset 1.

Dataset 1
Verdaderos Positivos: 454930
Verdaderos Negativos: 4144163

Falsos Positivos: 71913
Falsos Negativos: 228994
sensitivity: 0.665176247653
specificity: 0.982943144289
Accuracy: 0.938590408163
PPV: 0.863502030017
MCC: 0.725061778187

Tabla 5. Matriz de confusión individuo 32 Dataset 1.

Matriz de confusión

454930	71913	526843
228994	4144163	4373157
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,017056 856	0,665176 248	0,725061 778	0,938590 408	0,982943 144	0,86350 203	0,947636 456	0,13649 797

2.2.3 Modelos individuo 33

La figura 17 muestra el fenotipo para el individuo 33 del híbrido por mejora de hijos, el tamaño de la vecindad de este modelo AC es de 22 .

0	0	0	0	0	0	1
0	1	1	0	0	0	0
0	1	1	0	1	1	0
0	1	1	0	1	0	1
0	0	1	1	0	1	0
1	1	0	0	1	1	0
1	0	1	1	0	0	1

Figura 17.Fenotipo Individuo 33 por mejora de hijos.

Tabla 6. Medidas del Individuo 33 dataset 1.

Dataset 1
Verdaderos Positivos: 406051
Verdaderos Negativos: 4158028
Falsos Positivos: 58048
Falsos Negativos: 277873
sensitivity: 0.593707780397
specificity: 0.986231747246
Accuracy: 0.931444693878
PPV: 0.874923238361
MCC: 0.686348374056

Tabla 7. Matriz de confusión individuo 33 Dataset 1.

Matriz de confusión

406051	58048	464099
277873	4158028	4435901
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01376825	0,59370778	0,68634837	0,93144469	0,986231747	0,87492324	0,93735816	0,12507676

2.2.4 Modelos individuo 50

La figura 18 muestra el fenotipo para el individuo 50 del híbrido por mejora de hijos, el tamaño de la vecindad de este modelo AC es de 18.

0	0	0	0	0	0	1
1	0	1	0	0	0	0
0	1	1	0	1	0	0
0	1	1	0	1	0	1
0	0	0	0	0	1	0
0	1	0	0	1	1	1
1	0	1	0	0	0	1

Figura 18. Fenotipo Individuo 50 por mejora de hijos.

Tabla 8. Medidas del Individuo 50 dataset 1.

Dataset 1
Verdaderos Positivos: 455064
Verdaderos Negativos: 4068858
Falsos Positivos: 147218
Falsos Negativos: 228860
sensitivity: 0.665372175856
specificity: 0.965081749001
Accuracy: 0.923249387755
PPV: 0.755566329394
MCC: 0.665415006157

Tabla 9. Matriz de confusión individuo 50 Dataset 1.

Matriz de confusión

455064	147218	602282
228860	4068858	4297718
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,03491825	0,66537218	0,66541501	0,92324939	0,965081749	0,75556633	0,94674848	0,24443367

2.2.5 Modelos individuo 53

La figura 19 muestra el fenotipo para el individuo 53 del híbrido por mejora de hijos, el tamaño de la vecindad de este modelo AC es de 20.

0	0	0	0	1	0	1
1	0	1	0	0	0	0
0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	0	1	1	0	1	0
0	1	0	0	1	1	0
1	0	1	0	0	1	1

Figura 19. Fenotipo Individuo 53 por mejora de hijos.

Tabla 10. Medidas del Individuo 53 dataset 1.

Dataset 1
Verdaderos Positivos: 459867
Verdaderos Negativos: 4135123
Falsos Positivos: 80953
Falsos Negativos: 224057
sensitivity: 0.672394885981
specificity: 0.980798970417
Accuracy: 0.937753061224
PPV: 0.850314337488
MCC: 0.722389551471

Tabla 11. Matriz de confusión individuo 53 Dataset 1.

Matriz de confusión

459867	80953	540820
224057	4135123	4359180
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01920103	0,67239489	0,72238955	0,93775306	0,98079897	0,85031434	0,94860111	0,14968566

2.3 EVALUACIÓN DE MODELOS OBTENIDOS MEDIANTE HÍBRIDACION POR INTENSIFICACIÓN

2.3.1 Modelos individuo 13

La figura 20 muestra el fenotipo para el individuo 13 del híbrido por intensificación, el tamaño de la vecindad de este modelo AC es de 17.

0	0	0	0	0	0	1
0	1	1	0	0	0	0
0	1	0	1	1	0	0
0	0	0	0	1	0	1
1	0	1	1	1	1	0
0	0	0	0	1	0	1
1	0	1	0	0	0	0

Figura 20. Fenotipo Individuo 13 por Intensificación.

Las medidas de calidad para éste individuo en el dataset 1 se muestran a continuación en la tabla 12 .

Tabla 12. Medidas del Individuo 13 dataset 1.

Dataset 1
Verdaderos Positivos: 442511
Verdaderos Negativos: 4122601
Falsos Positivos: 93475
Falsos Negativos: 241413
sensitivity: 0.647017797299
specificity: 0.977828910105
Accuracy: 0.931655510204
PPV: 0.825601788106
MCC: 0.693764115702

La tabla 13 muestra el cálculo de las medidas de evaluación desde la matriz de confusión.

Tabla 13. Matriz de confusión individuo 13 Dataset 1.

Matriz de confusión

442511	93475	535986
241413	4122601	4364014
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,02217109	0,6470178	0,69376412	0,93165551	0,97782891	0,82560179	0,94468097	0,17439821

2.3.2 Modelos individuo 33

La figura 21 muestra el fenotipo para el individuo 33 del híbrido por Intensificación, el tamaño de la vecindad de este modelo AC es de 18.

0	0	0	0	0	0	1
1	1	1	0	0	1	0
0	0	0	1	0	0	1
0	1	0	0	1	0	0
0	0	1	1	1	1	0
0	1	0	0	0	0	1
1	0	1	0	0	1	0

Figura 21. Fenotipo Individuo 33 por Intensificación.

Tabla 14. Medidas del Individuo 33 dataset 1.

Dataset 1
Verdaderos Positivos: 475472
Verdaderos Negativos: 4049291
Falsos Positivos: 166785
Falsos Negativos: 208452
sensitivity: 0.695211748674
specificity: 0.960440703631
Accuracy: 0.923421020408
PPV: 0.740314235579
MCC: 0.673267549257

Tabla 15. Matriz de confusión individuo 33 Dataset 1.

Matriz de confusión

475472	166785	642257
208452	4049291	4257743
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,0395593	0,69521175	0,67326755	0,92342102	0,960440704	0,74031424	0,95104167	0,25968576

2.3.3 Modelos individuo 53

La figura 22 muestra el fenotipo para el individuo 53 del híbrido por Intensificación, el tamaño de la vecindad de este modelo AC es de 17.

0	0	0	0	0	0	1
0	0	1	0	0	0	0
0	1	0	1	0	0	1
0	0	0	0	1	0	1
0	0	0	1	1	1	0
0	1	0	0	1	0	1
1	0	1	0	0	1	1

Figura 22. Fenotipo Individuo 53 por Intensificación.

Tabla 16. Medidas del Individuo 53 dataset 1.

Dataset 1
Verdaderos Positivos: 459867
Verdaderos Negativos: 4135123
Falsos Positivos: 80953
Falsos Negativos: 224057
sensitivity: 0.672394885981
specificity: 0.980798970417
Accuracy: 0.937753061224
PPV: 0.850314337488
MCC: 0.722389551471

Tabla 17. Matriz de confusión individuo 53 Dataset 1.

Matriz de confusión

478211	176698	654909
205713	4039378	4245091
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,04191053	0,69921658	0,66940936	0,92195694	0,958089465	0,73019458	0,95154097	0,26980542

2.4 OTROS RESULTADOS OBTENIDOS

Al analizar las vecindades de los modelos AC obtenidos por la hibridación por mejora de hijos e intensificación, se identificaron algunas diferencias en las vecindades obtenidas que con ayuda de la figura 23, en donde se muestra la vecindad de Moore nos servirá de guía para identificar las diferencias encontradas.

V1	V2	V3	V4	V5	V6	V7
V8	V9	V10	V11	V12	V13	V14
V15	V16	V17	V18	V19	V20	V21
V22	V23	V24	V25	V26	V27	V28
V29	V30	V31	V32	V33	V34	V35
V36	V37	V38	V39	V40	V41	V42
V43	V44	V45	V46	V47	V48	V49

Figura 23. Vecindad de Moore (Adaptado de [3]).

En todos las vecindades obtenidas mediante hibridación por mejora de hijos por lo general no encontraba encendido el bit de la celda V18 y en ocasiones los de la celdas V42 y V49; de la misma forma en la hibridación por intensificación no se encendía el bit 49, por lo cual se experimenta modificando estos bits para ver cuál es el comportamiento de los modelos AC.

Modelos individuo 13 por mejora de hijos

La figura 24 muestra la vecindad qued

0	0	0	0	0	0	1
0	1	1	0	0	1	0
0	0	1	1	1	0	1
1	1	0	0	1	0	0
0	0	1	1	0	1	1
0	1	0	0	0	1	1
1	0	0	1	0	1	0

la celda V18 en 1,

Figura 24. Fenotipo del individuo 13 con cambio en la celda V18.

Las medidas de calidad se muestran en la tablas 18 y 19.

Tabla 18. Medidas del Individuo 13 cambio en V18.

Dataset 1
Verdaderos Positivos: 428328
Verdaderos Negativos: 4132527
Falsos Positivos: 83549
Falsos Negativos: 255596
sensitivity: 0.626280112995
specificity: 0.980183231991
Accuracy: 0.930786734694
PPV: 0.836779148116
MCC: 0.687132500012

Tabla 19. Matriz de confusión individuo13 cambio en V18.

Matriz de confusión

428328	83549	511877
255596	4132527	4388123
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01981677	0,62628011	0,6871325	0,93078673	0,980183232	0,83677915	0,94175277	0,16322085

La figura 25 muestra la vecindad del individuo 13, con el bits de la celda V49 en 1, la vecindad queda con un tamaño de 21.

0	0	0	0	0	0	1
0	1	1	0	0	1	0
0	0	1	0	1	0	1
1	1	0	0	1	0	0
0	0	1	1	0	1	1
0	1	0	0	0	1	1
1	0	0	1	0	1	1

Figura 25. Fenotipo del individuo 13 con cambio en la celda V49.

Las medidas de calidad se muestran en la tablas 20 y 21.

Tabla 20. Medidas del Individuo 13 cambio en V49.

Dataset 1	
Verdaderos Positivos: 424023	
Verdaderos Negativos: 4142726	
Falsos Positivos: 73350	
Falsos Negativos: 259901	
sensitivity: 0.61998555395	
specificity: 0.982602306031	
Accuracy: 0.931989591837	
PPV: 0.852525167229	
MCC: 0.69148293426	

Tabla 21. Matriz de confusión individuo13 cambio en V49.

Matriz de confusión		
424023	73350	497373
259901	4142726	4402627
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01739769	0,61998555	0,69148293	0,93198959	0,982602306	0,85252517	0,94096684	0,14747483

En las medidas obtenidas se observa que al hacer el cambio del bits de la celda V49 se obtiene una precisión mejor que la obtenida al modificar el bits de la celda V18.

2.4.1 Modelos individuo 33 por mejora de hijos

La figura 26 muestra la vecindad del individuo 33, con el bits de la celda V18 en 1, la vecindad queda con un tamaño de 23.

0	0	0	0	0	0	1
0	1	1	0	0	0	0
0	1	1	1	1	1	0
0	1	1	0	1	0	1
0	0	1	1	0	1	0
1	1	0	0	1	1	0
1	0	1	1	0	0	1

Figura 26. Fenotipo del individuo 33 con cambio en la celda V18.

Las medidas de calidad se muestran en la tablas 22 y 23.

Tabla 22. Medidas del Individuo 33 cambio en V18.

Dataset 1	
Verdaderos Positivos: 425511	
Verdaderos Negativos: 4156377	
Falsos Positivos: 59699	
Falsos Negativos: 258413	
sensitivity: 0.622161234289	
specificity: 0.985840150889	
Accuracy: 0.935079183673	
PPV: 0.876962552297	
MCC: 0.705411945377	

Tabla 23. Matriz de confusión individuo 33 cambio en V18.

Matriz de confusión		
425511	59699	485210
258413	4156377	4414790
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01415985	0,62216123	0,70541195	0,93507918	0,985840151	0,87696255	0,94146653	0,12303745

La figura 27 muestra la vecindad del individuo 33, con el bits de la celda V42 en 1, la vecindad queda con un tamaño de 23.

0	0	0	0	0	0	1
0	1	1	0	0	0	0
0	1	1	0	1	1	0
0	1	1	0	1	0	1
0	0	1	1	0	1	0
1	1	0	0	1	1	1
1	0	1	1	0	0	1

Figura 27. Fenotipo del individuo 33 con cambio en la celda V42.

Las medidas de calidad se muestran en la tablas 24 y 25.

Tabla 24. Medidas del Individuo 33 cambio en V42.

Dataset 1
Verdaderos Positivos: 453067
Verdaderos Negativos: 4157109
Falsos Positivos: 58967
Falsos Negativos: 230857
sensitivity: 0.662452260778
specificity: 0.986013772048
Accuracy: 0.940852244898
PPV: 0.88483772562
MCC: 0.73462269831

Tabla 25. Matriz de confusión individuo 33 cambio en V42.

Matriz de confusion

453067	58967	512034
230857	4157109	4387966
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01398623	0,66245226	0,7346227	0,94085224	0,986013772	0,88483773	0,94738861	0,11516227

Al igual que lo observado en el modelo del individuo 13, la vecindad del modelo del individuo 33 al realizar el bit de la celda V42 se obtiene una precisión mejor que la obtenida al modificar el bits de la celda V18.

2.4.2 Modelos individuo 53 por mejora de hijos

La figura 28 muestra la vecindad del individuo 53, con el bits de la celda V18 en 1, la vecindad queda con un tamaño de 21.

0	0	0	0	1	0	1
1	0	1	0	0	0	0
0	0	1	1	1	0	1
0	1	0	0	1	0	1
0	0	1	1	0	1	0
0	1	0	0	1	1	0
1	0	1	0	0	1	1

Figura 28. Fenotipo del individuo 53 con cambio en la celda V18.

Las medidas de calidad se muestran en la tablas 26 y 27.

Tabla 26. Medidas del Individuo 53 cambio en V18.

Dataset 1
Verdaderos Positivos: 443800
Verdaderos Negativos: 4159603
Falsos Positivos: 56473
Falsos Negativos: 240124
sensitivity: 0.648902509636
specificity: 0.986605317361
Accuracy: 0.93947
PPV: 0.887115634863
MCC: 0.7273821584

Tabla 27. Matriz de confusión individuo 53 cambio en V18.

Matriz de confusión

425511	59699	485210
258413	4156377	4414790
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01415985	0,62216123	0,70541195	0,93507918	0,985840151	0,87696255	0,94146653	0,12303745

La figura 29 muestra la vecindad del individuo 33, con el bits de la celda V42 en 1, la vecindad queda con un tamaño de 21.

0	0	0	0	1	0	1
1	0	1	0	0	0	0
0	0	1	0	1	0	1
0	1	0	0	1	0	1
0	0	1	1	0	1	0
0	1	0	0	1	1	1
1	0	1	0	0	1	1

Figura 29. Fenotipo del individuo 53 con cambio en la celda V42.

Las medidas de calidad se muestran en la tablas 28 y 29.

Tabla 28. Medidas del Individuo 53 cambio en V42.

Dataset 1
Verdaderos Positivos: 471577
Verdaderos Negativos: 4157660
Falsos Positivos: 58416
Falsos Negativos: 212347
sensitivity: 0.689516671443
specificity: 0.986144462291
Accuracy: 0.944742244898
PPV: 0.889779676335
MCC: 0.753895137172

Tabla 29. Matriz de confusión individuo 53 cambio en V42.

Matriz de confusion

471577	58416	529993
212347	4157660	4370007
683924	4216076	4900000

TFP o FPR	TVP o TPR	MCC	ACC	Especificidad	PPV	NPV	FDR
0,01385554	0,68951667	0,75389514	0,94474224	0,986144462	0,88977968	0,95140809	0,11022032

El modelo del individuo 53 al realiza cambio de bit en la celda V42 obrítene una mejor precisión, que al hacer el cambio de la celda V18, lo cual nos indica en los modelos que las celdas V49 y V42 guardan información importante para el modelo. En vista de la experimentación realizada se identifican que bits solo los comunes en los modelos obtenidos, los cuales se muestran en la figura 30.

0	?	?	X	X	X	1
X	?	1	?	X	X	X
X	?	?	?	?	X	X
X	X	?	0	1	X	X
X	X	?	?	?	1	X
X	?	X	1	?	?	1
1	?	X	X	?	1	1

Figura 30. Vecindad por fija o por prioridad.

Los bits de las celdas V1, V25 son fijos en 0 y los bits de las celdas V7, V10, V26, V34, V39, V42, V43, V48 Y V49 son fijos en 1.

Según lo observado otros bits que se mueven constantemente son aquellos que están marcadas las celdas con signos de interrogación (?), y los bits marcados con la X una no se nota casi su cambio.

Con los bits que son fijos se propone un vecindario en prioridad en donde los bits con (?) tendrían una prioridad alta de cambio de estado de 1 a 0 o viceversa. Esto disminuiría la carga computacional ya que son menos los individuos a evaluar.

2.5 REFERENCIAS Y BIBLIOGRAFIA

- [1] J. Bacardit, *et al.*, "Speeding-up pittsburgh learning classifier systems: Modeling time and accuracy," 2004, pp. 1021-1031.
- [2] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clin Chem*, vol. 39, pp. 561-577, 1993.
- [3] N. Diaz, "Metodología y framework computacional para el diseño inverso de modelos de autómatas celulares de secuencias cortas de aminoácidos soportando en un proceso de minería de datos," *Tesis de Maestría, Universidad del Valle*, p. 56, 2010.