CAPÍTULO 5

APLICACIÓN HEURÍSTICA

El presente capítulo muestra la aplicación de los conceptos teóricos mencionados en el capítulo anterior con el fin de obtener una solución inicial al problema de la clasificación de alelos HLA. También se aplica una heurística para determinar qué atributos serán seleccionados que contribuirán a la solución inicial.

5.1 APLICACIÓN DE LA ENTROPÍA DE SHANNON A LA CLASIFICACIÓN

La aplicación de la entropía de Shannon al problema de la clasificación de alelos HLA con las bases teóricas mencionadas anteriormente se muestra a continuación con un pequeño ejemplo diferente al que se venía manejando al del catálogo "X". Esto se debe a las dimensiones que se manejan al obtener las biclases y bialelos en el catálogo "X", lo cual permite explicar de manera clara la obtención de los datos que se utilizarán y por lo tanto es más sencillo explicarlo con el nuevo ejemplo, y posteriormente presentar los resultados del catálogo "X" en el Capítulo 6 aplicando los pasos vistos en este capítulo.

Se desea obtener la entropía de Shannon de un sistema dado, así que partiendo de la matriz de respuestas ***R*** (Tabla 5.1.1) se toman como nuevo ejemplo un sistema con 4 biclases BC= {BC_1, BC_2, BC_3, BC_4} sumando en total 15 bialelos (elementos). Cada uno

de los elementos de las biclases responde de manera afirmativa o negativa (0 ó 1) a 3 preguntas (atributos). En el caso de la clasificación de los alelos HLA el atributo es la pregunta P_i.

Biclase	Bialelo	P0	P1	P2
BC_1	1	1	1	0
	2	1	1	1
BC_2	3	0	1	1
	4	1	1	0
	5	1	1	0
	6	1	0	1
	7	0	1	1
	8	0	1	0
BC_3	9	0	1	0
	10	0	1	1
	11	0	0	1
BC_4	12	1	0	1
	13	0	0	1
	14	0	1	1
	15	0	0	0

Tabla 5.1.1 Representación de la matriz ***R*** para el ejemplo de la Entropía de Shannon

A la Tabla 5.1.1 se le aplica la fórmula de la entropía de Shannon, ecuación (5.1.1) únicamente cambiando la nomenclatura para los datos que se manejan en la clasificación.

$$H(C) = -\sum_{i=1}^{N} P(b)_i \log_2 P(b)_i \qquad (5.1.1)$$

En donde:

$H(C)$ = Entropía total de la clasificación

N = El número total de bialelos (elementos) en la matriz de respuestas ***R***

$P(b)_i$ = Probabilidad de que bialelo b pertenezca a la biclase BC_i

$i = \{1, 2, 3, \ldots, N\}$

El siguiente paso es calcular las probabilidades y logaritmos de cada una de las biclases, que van a depender del número de elementos que cada una contenga con el fin de obtener el valor de la entropía y así conocer el desorden que existe en este sistema. Los datos se presentan en la Tabla 5.1.2.

Biclase	Número de Bialelos / Biclase	P_i	$\log_2$	$P_i \log_2 P_i$
BC_1	2	0.133	-2.907	-0.388
BC_2	6	0.400	-1.322	-0.529
BC_3	3	0.200	-2.322	-0.464
BC_4	4	0.267	-1.907	-0.509
Total	**15**	**1**		

Tabla 5.1.2 Datos que se utilizarán para el cálculo de la Entropía

Tomando los datos de la Tabla 5.1.2 es posible calcular la entropía de éste sistema, como se muestra a continuación:

$$H(C) = -\sum_{i \in I_0} P(b)_i \log_2 P(b)_i = -(-0.388 - 0.529 - 0.464 - 0.509) = 1.889$$

Donde $I_0 = \{i \in \{1,2,3,4\} \mid P(C_i \mid b_0 \neq 0)\}$. Es importante mencionar que es necesario colocar

la condición $P(Ci|b_0)$ lo cual significa que no se toman en cuenta las probabilidades de los elementos con valor "0" como se presenta en la siguiente demostración:

Demostración de Límites: Se cuenta con la función *r(p)* que se muestra en la ecuación (5.1.2) a la cual se le aplica la función de límite (5.1.3) para encontrar una solución:

$$r(p) = \begin{cases} p \log p & p > 0 \\ 0 & p = 0 \end{cases} \qquad (5.1.2)$$

$$\lim_{p \to 0} p \log p = \lim_{p \to 0} \frac{p}{\frac{1}{\log p}} \begin{array}{l} \longrightarrow \\ \longrightarrow \end{array} \frac{0}{0} \qquad (5.1.3)$$

Teorema L'Hôpital: Dadas f y g como diferenciables y con las funciones $f(a) = 0$ y $g(a) = 0$ se tiene:

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)} \qquad (5.1.4)$$

Donde *f '(x)* y *g '(x)* significan las derivadas de las funciones *f(x)* y *g(x)* respectivamente.

Dado que la función de límite (5.1.3) no se puede resolver debido a que $\frac{0}{0}$ no existe, es necesario aplicar el teorema de L´Hôpital dando la ecuación (5.1.4).

$$\lim_{p\to 0} p\log p = \lim_{p\to 0} \frac{p}{\frac{1}{\log p}} = \lim_{p\to 0} \frac{\log p}{\frac{1}{p}} = \lim_{p\to 0} \frac{\frac{1}{p}}{-\frac{1}{p^2}} = \tag{5.1.4}$$

$$\lim_{p\to 0} \frac{1}{-\frac{1}{p}} = \lim_{p\to 0} -p = 0$$

Al obtener el valor de la entropía del sistema, como lo menciona Peña [P2004], "éste corresponde al error que se comete al signar un bialelo aleatoriamente a una biclase"

El programa en lenguaje Java realiza todo el cálculo de los algoritmos y probabilidades para obtener la entropía de Shannon del sistema.

Cabe mencionar que a partir de la aplicación de la entropía de Shannon en adelante ya no es posible obtener los resultados que el programa calcula para los catálogos mostrados en la Tabla 3.1.2, debido a que la capacidad del equipo utilizado no cuenta con la memoria RAM necesaria para procesar tal gran cantidad de datos. Por lo que para comprobar los resultados que se vayan obteniendo en el transcurso del proyecto, se realizan pequeños ejemplos en la herramienta Excel manualmente para después obtener los resultados en el programa en lenguaje Java con el fin de comparar y confirmar los resultados en ambas herramientas. Uno de ellos se muestra en el Apéndice **D** en Excel y Apéndice **E** en Java.

El siguiente paso es obtener la entropía que cada uno de los atributos tiene, por lo que es necesario aplicar la heurística para elegir los atributos.

5.2 APLICACIÓN DE LA HEURÍSTICA PARA ELEGIR ATRIBUTOS

Esta heurística consiste en calcular la entropía de cada uno de los atributos, señalando como atributo a cada una de las preguntas que forman el conjunto $\boldsymbol{P}$ = {P_1, P_2, P_3, …, P_k } mencionado en el Capítulo 3.

Así que un atributo P_k va a subdividir los bialelos (elementos) en subgrupos S_j donde j=1, …, *c*). Donde el valor ***c*** se obtiene de $\mathbf{2^n}$ y significa el número de combinaciones que se pueden realizar con (0,1) con *n* dígitos, cabe mencionar que este valor va a depender del número de iteración en la que se encuentre, la cual será definida más adelante. Por ejemplo, si n= 3 el valor de c= 2^3 = 8 combinaciones con (0,1) dando como resultado S= {S_{111}, S_{000}, S_{100}, S_{110}, S_{101}, S_{001}, S_{010}, S_{011}}.

Con la información de los subgrupos se puede calcular la proporción de los elementos que se encuentran en S_j el cual se va a representar con $W_i = \frac{S_j}{N}$.

Con estos datos es posible calcular la entropía de los atributos P_k seleccionados H(C|P_k), definida como "la medida ponderada de la entropía de Shannon en cada S_j" como lo describe Peña [P2004]. En la ecuación (5.2.1) se presenta la fórmula para calcular la heurística para elegir los atributos.

$$H(C \mid \mathrm{P_k}) = \sum_{j=1}^{c} W_j \times H(C \mid \mathrm{S_j}) \qquad (5.2.1)$$

El componente correspondiente a $H(C \mid \mathrm{S_j})$ se calcula a través de la ecuación (5.2.2),

tomando como dato la probabilidad de que un elemento forme parte de la biclase BC_i en el caso de que el bialelo pertenezca al subgrupo S_j, representándose por medio de $p(BC_i \mid S_j)$.

$$H(C \mid S_j) = -\sum_{i=1}^{N} p(C_i \mid S_j) \log_2 p(BC_i \mid S_j) \qquad (5.2.2)$$

Aplicando la heurística de elegir los atributos al ejemplo que se ha venido manejando de la Tabla 5.1.1, en este caso el atributo seleccionado es la pregunta "***P1***" con ***N***=15 bialelos en total, se calcula el valor de S_j , por lo cual el valor de n=1, así que para obtener el valor de las combinaciones se tiene que c= 2^1=2 dando un conjunto para S_j= {S_0, S_1}donde cada uno de los elementos de S_j están formados por:

S_0= { *bialelo **3**, bialelo **7**, bialelo **8**, bialelo **9**, bialelo **10**, bialelo **11**, bialelo **13**, bialelo **14**, bialelo **15*** } aquellos bialelos que tienen como combinación "0" para "P_1"

de donde se obtiene $W_0 = \frac{9}{15} = 0.6$

S_1= { *bialelo **1**, bialelo **2**, bialelo **4**, bialelo **5**, bialelo **6**, bialelo **12*** } aquellos bialelos que tienen como combinación "1" para "P_1" de donde se obtiene $W_1 = \frac{6}{15} = 0.4$

Como lo muestra la ecuación (5.2.1), es necesario calcular el valor de $H(C \mid S_j)$ pero para eso se aplica la ecuación (5.2.2) con el fin de obtener las probabilidades $p(BC_i \mid S_j)$ para cada biclase, quedando como se muestra a continuación.

- Para los valores de S_0, las probabilidades y sus respectivos logarítmos se presentan en la Tabla 5.2.1:

Biclase *i*	**S_0**	$p(BC_i \mid S_0)$	$\log p(BC_i \mid S_0)$
BC_1	0	3/9	-1.585
BC_2	3	3/9	-1.585
BC_3	3	0/9	0
BC_4	3	1/9	-1.585
Total S_0	9		

Tabla 5.2.1 Probabilidades y Logaritmos para S_0

- Con las probabilidades calculadas, se obtiene la entropía de la pregunta:

$$H(C \mid S_0) = -\sum_{i=5}^{4}(p(BC_i \mid S_0)\log_2(p(BC_i \mid S_0))) = -(\frac{3}{9}*(-1.585) + \frac{3}{9}*(-1.585) + (0) + \frac{3}{9}*(-1.585)) = 1.585$$

- Para los valores de S_1, las probabilidades y sus respectivos logarítmos se presentan en la Tabla 5.2.2:

Biclase *i*	S_1	$p(BC_i \mid S_1)$	$\log\ p(BC_i \mid S_1)$
BC_1	2	2/6	-1.585
BC_2	3	3/6	-1.000
BC_3	0	0/6	0
BC_4	1	1/6	-2.585
Total S_1	6		

Tabla 5.2.2 Probabilidades y Logaritmos para S_1

- Con las probabilidades calculadas, se obtiene la entropía de la pregunta:

$$H(C \mid \mathrm{S}_1) = -\sum_{i=5}^{4} (p(BC_i \mid S_1)\log_2(p(BC_i \mid S_1))) = -(\frac{2}{6} * (-1.585) + \frac{3}{6} * (-1)$$

$$+ (0) + \frac{1}{6} * (-2.585)) = 1.459$$

Por lo tanto ya se puede calcular el valor de la entropía de la pregunta "***P1***" (atributo) obtenidos los valores de W_i y de $H(C \mid \mathrm{S}_j)$ dando como resultado:

$$H(C \mid P_1) = W_0 \times H(C \mid S_0) + W_1 \times H(C \mid S_1) = (0.6 \times 1.585) + (0.4 \times 1.459) = 1.534$$

Este procedimiento se realiza para cada uno de los atributos para después obtener la ganancia de información y la ganancia de información relativa que cada pregunta

proporciona. En este caso las ganancias de información se calculan como sigue:

$$I_G = H(C) - H(C \mid P_i) = H(C) - H(C \mid P1) = 1.889 - 1.534 = 0.355$$

$$I_G R = \frac{I_G}{H(C)} \times 100 = \frac{0.355}{1.889} \times 100 = 18.793\%$$

Para este ejemplo se obtiene para cada unos de los atributos sus respectivas entropías, ganancias de información y la relativa. Estos resultados se ejemplifican en la Tabla 5.2.3.

Atributo *k*	$H(C\|P_k)$	I_G	$I_G R$
P_0	1.534	0.355	18.79%
P_1	1.631	0.285	13.67%
P_2	1.815	0.037	1.99%

Tabla 5.2.3 Resultados obtenidos para cada uno de los atributos

Al obtener la información de cada uno de los atributos se observa por medio de la Tabla 5.2.3 las ganancias relativas, éstas permiten definir cuál atributo proporciona la mayor ganancia de información la cual, en este caso, es la pregunta "$\boldsymbol{P_0}$" con una información del 18.79% en comparación a las otras "P_1"con 13.67% y "P_2" con solamente 1.99%.

La obtención de toda esta información también se realiza a través del programa en Java, mencionando nuevamente que se puede aplicar a ejemplos pequeños debido a la falta de capacidad del equipo que se utilizó.

5.3 CÁLCULO DE ITERACIONES

La pregunta seleccionada de la sección anterior va a permitir realizar la siguiente iteración (*n*), proceso que se describe a continuación.

La primera iteración (*n*=1)se realiza obteniendo los datos que se presentan en la Tabla 5.2.3, de la cual se elige una sola pregunta, es decir, aquel atributo que proporcione mayor ganancia de información. Una vez seleccionada la pregunta, ésta se toma en cuenta para la siguiente iteración de otra manera. El atributo seleccionado va a combinar su contenido de la matriz de respuestas ***R*** con los demás atributos, formando una combinación de *n+1*.

A continuación se presentan los pasos de manera clara aplicados al ejemplo que se ha venido manejando.

- De la Tabla que proporcione la información de la entropía, ganancia de información y la ganancia de información relativa se elige la pregunta "$\boldsymbol{P_s}$" por contener la mayor ganancia de información. En este caso la pregunta "$\boldsymbol{P_0}$".
- La pregunta seleccionada se combina con las no seleccionadas para formar la iteración n+1. Para el ejemplo $n+1 = 1+1= 2$.
- Se combina la pregunta "Ps" con las preguntas no seleccionadas de acuerdo a la matriz ***R***. Es decir, "$\boldsymbol{P_1\ P_0}$" y "$\boldsymbol{P_2\ P_0}$" , como se muestra en la Tabla 5.2.4

Biclase	Bialelo	P1	P0	P2	P0
BC_1	1	1	1	0	1
	2	1	1	1	1
BC_2	3	1	0	1	0
	4	1	1	0	1
	5	1	1	0	1
	6	0	1	1	1
	7	1	0	1	0
	8	1	0	0	0
BC_3	9	1	0	0	0
	10	1	0	1	0
	11	0	0	1	0
BC_4	12	0	1	1	1
	13	0	0	1	0
	14	1	0	1	0
	15	0	0	0	0

Tabla 5.2.4 Combinación de Preguntas para la iteración $n = 2$

- Se aplica nuevamente el concepto de ***c*** para obtener el número de combinaciones que se tendrán para esta iteración. Así que $\boldsymbol{c} = 2^n = 2^2 = 4$ combinaciones con (0,1), dando como resultado el conjunto Sj = {S_{00}, S_{11}, S_{01}, S_{10}}.
- Se calcula para cada una de las S_j las proporciones W_j. En este caso W_{00}, W_{11}, W_{01}, W_{10} correspondientes.
- Se realiza todo el cálculo para obtener una nueva entropía para cada uno de los atributos $H(C \mid P_k)$. Para el ejemplo $H(C \mid P_k)$: { $H(C \mid P_{21})$, $H(C \mid P_{31})$ }
- Se calculan las ganancias de información y las ganancias de información relativas para cada atributo. En la Tabla 5.2.5 se muestran los valores correspondientes.

Atributo *k*	$H(C\|P_k)$	I_G	$I_G R$
P_1P_0	1.026	0.863	45.66%
P_2P_0	1.351	0.538	28.49%

Tabla 5.2.3 Resultados obtenidos para cada uno de los atributos para la n=2 iteración

- Se selecciona la pregunta que proporcione la mayor ganancia de información. En este caso el atributo que contiene "***P₁***"
- Se realiza nuevamente el segundo paso para realizar la siguiente iteración.

Biclase	Bialelo	P2	P1	P0
BC_1	1	0	1	1
	2	1	1	1
BC_2	3	1	1	0
	4	0	1	1
	5	0	1	1
	6	1	0	1
	7	1	1	0
	8	0	1	0
BC_3	9	0	1	0
	10	1	1	0
	11	1	0	0
BC_4	12	1	0	1
	13	1	0	0
	14	1	1	0
	15	0	0	0

Tabla 5.2.4 Combinación de Preguntas para la iteración $n = 2$

Es de gran importancia mencionar que el número de iteraciones se realiza hasta obtener un valor de $H(C \mid P_k)$ igual o aproximado al valor de la entropia de Shannon de *H(C)*. La aproximación mínima que debe cumplir el catálogo de preguntas seleccionadas es del 99.98% en la ganancia de información relativa.

Para este ejemplo se llega a obtener un valor de entropía del atributo final aproximado al valor de la entropía de Shannon del sistema, pero no siempre se presenta esto por lo mencionado en el párrafo anterior.

En este caso el catálogo de preguntas seleccionadas fueron las 3 preguntas iniciales en el

mismo orden en que se aplicaron. Para cada catálogo a resolver el orden de las preguntas variará dependiendo de la ganancia de información que proporcione.

Para la tercera iteración y última (ya que no hay más preguntas que aplicar) proporciona la información que se presenta en la Tabla 5.2.5.

Atributo *k*	**$H(C\|P_k)$**	**I_G**	**$I_G R$**
P2P1P0	1.8892	1	100%

Tabla 5.2.5 Combinación de Preguntas para la iteración $n = 3$

Así que la solución para este ejemplo pequeño queda con el catálogo de preguntas en el orden y con el número de preguntas de [P0, P1 y P2], proporcionando una ganancia de información del 100% debido a que se aplicaron todas las preguntas iniciales. Este catálogo que se obtiene de preguntas se colocará en la sonda para poder realizar la clasificación en el orden y contenido que se obtenga al resolver el problema.

Por lo tanto, es importante recordar que al realizar todos los pasos anteriores hasta llegar a obtener una entropía de un atributo igual o aproximado a la entropía del sistema es posible crear una combinación de preguntas que contengan los atributos seleccionados en el transcurso de cada iteración, es decir, formar un catálogo de preguntas seleccionadas que darán como respuesta una combinación única de “0” y “1” que serán aplicadas en una sola unidad de tiempo para así poder identificar la biclase con el fin de clasificar los alelos HLA.

Otro punto importante que es necesario mencionar es que al aplicar la heurística a los

catálogos reales mencionados en la Tabla 3.1.2 y que se han resulto conforme se fue avanzando en este proyecto ya no fue posible resolverlos en esta etapa debido a la falta de memoria requerida para poder encontrar la solución de cada uno de ellos. Por lo tanto, retomando el ejemplo del catálogo "X" descrito en los capítulos anteriores se presentarán los resultados que formarán el catálogo de preguntas para clasificar los alelos con el fin de dar validez de este proyecto.

5.4 CATÁLOGO DE PREGUNTAS DEL CATÁLOGO "X"

Al realizar la limpieza del catálogo de alelos HLA, obtener las combinaciones de bialelos, preguntas y respuestas es posible realizar la entropía de Shannon del sistema.

Partiendo del catálogo inicial de preguntas que se obtuvo en la sección 3.2.1 y que se muestra en la Tabla 3.2.1.2 junto con la matriz de respuestas generada por el catálogo de preguntas representada tan solo por una pequeña parte en la Tabla 3.2.3.1, se calcula la entropía de Shannon para todo el sistema con los pasos mostrados en el Capítulo 5 en la sección 5.1.

En la Tabla 5.4.1 se presenta solo un pequeña parte de la generación de los datos de Excel para obtener la entropía de Shannon del sistema para el catálogo "X" (ver Apéndice **D**).

Biclase	Bialelos	P0	P1	P2	P3	P4	P5	P6	No.Bialelos/ Biclase	Pi (%)	log2 Pi	Pi log2 Pi
BC1	A*010101	1	0	1	0	0	1	0	11	0.166667	-2.58496	-0.430827
	A*010101											
	A*010101	1	0	1	0	0	1	0				
	A*010102											
	A*010101	1	0	1	1	0	1	0				
	A*0102											
	A*010101	1	1	1	0	0	1	0				
	A*020101											
	A*010101	1	0	1	0	0	1	0				
	A*020102											
	A*010101	1	1	1	0	0	1	0				
	A*020103											
	A*010101	1	1	1	0	0	1	0				
	A*020104											
	A*010101	1	1	1	0	0	1	0				
	A*020105											
	A*010101	1	0	1	0	0	1	0				
	A*03010101											
	A*010101	1	0	1	0	0	1	0				
	A*03010102N											
	A*010101	1	0	1	0	1	1	1				
	A*4301											
BC2	A*010102	1	0	1	0	0	1	0	10	0.151515	-2.72246	-0.412495
	A*010102											
	A*010102	1	0	1	1	0	1	0				
	A*0102											
	A*010102	1	1	1	0	0	1	0				
	A*020101											
	A*010102	1	0	1	0	0	1	0				
	A*020102											
	A*010102	1	1	1	0	0	1	0				
	A*020103											
	A*010102	1	1	1	0	0	1	0				
	A*020104											
	A*010102	1	1	1	0	0	1	0				
	A*020105											
	A*010102	1	0	1	0	0	1	0				
	A*03010101											
	A*010102	1	0	1	0	0	1	0				
	A*03010102N											
	A*010102	1	0	1	0	1	1	1				
	A*4301											

...

...

BC10	A*03010102N	1	0	1	0	0	1	0	2	0.030303	-5.04439	-0.15286
	A*03010102N											
	A*03010102N	1	0	1	0	1	1	1				
	A*4301											
BC11	A*4301	1	0	0	0	1	0	1	1	0.015152	-6.04439	-0.09158
	A*4301											

Tabla 5.4.1 Datos que permitirán el cálculo de la entropía total del sistema

En base a los datos que se encuentran localizados en el Apéndice **D** se calcula la entropía del sistema con la información contendida en cada uno de los elementos de cada biclase:

$$H(C) = -\sum_{i=1}^{11} P(b)_i \log_2 P(b)_i = -(-0.4308 - 0.4124 - ... - 0.1528 - 0.0975) = 3.2363$$

Con el cálculo de la entropía total del sistema, se realizan las iteraciones con el fin de obtener las ganancias de información y así formar el catálogo de preguntas que permitirán la clasificación.

5.4.1 RESULTADOS DE LAS ITERACIONES

En esta sección se muestran los resultados que se generaron para el ejemplo del catálogo "X" representando cada una de las iteraciones que se obtuvieron al aplicar la entropía de Shannon a las preguntas y seleccionando aquellas que formarán el catálogo final de acuerdo a aquellas que proporcionen la mayor ganancia de información.

Los resultados de la primera iteración se muestran en la Tabla 5.4.1.1 en donde se coloca el número de pregunta, la entropía que proporciona dicha pregunta, la ganancia de información y la ganancia de información relativa.

Atributo *k*	$H(C\|P_k)$	I_G	$I_G R$
P_0	2.958	0.279	8.6140%
P_1	2.797	0.439	13.5650%
P_2	3.074	0.163	5.0000%
P_3	**2.731**	**0.506**	**15.6260%**
P_4	3.163	0.073	2.2700%
P_5	3.123	0.113	3.5000%
P_6	3.163	0.073	2.2700%

Tabla 5.4.1.1 Ganancias de información de las preguntas para la iteración n=1

Se revisan los valores generados por la herramienta en Excel para la primera iteración dando como resultado la selección de la pregunta “P3” por proporcionar la mayor ganancia de información con un 15.6260%. Dado que aún no se llega a igualar o aproximar el valor de la entropía total del sistema y aún hay preguntas que aplicar se realiza la siguiente iteración. Los resultados se muestran en la Tabla 5.4.1.2

Atributo *k*	$H(C\|P_k)$	I_G	$I_G R$
P_0P_3	2.4954	0.7409	22.8927%
P_1P_3	**2.2955**	**0.9409**	**29.0710%**
P_2P_3	2.6116	0.6248	19.3044%
P_4P_3	2.6593	0.577	17.8299%
P_5P_3	2.6214	0.615	19.0017%
P_6P_3	2.6593	0.577	17.8299%

Tabla 5.4.1.2 Ganancias de información de las preguntas para la iteración n=2

Como se puede observar, la ganancia de información va a ir aumentando conforme se vayan agregando más preguntas al catálogo, pero hay que recordar que el objetivo es obtener un número de preguntas seleccionadas que permitan clasificar pero sin llegar a seleccionar todas.

Para el ejemplo, la siguiente pregunta a seleccionar para incluirla al catálogo es “P1” ya que, como la iteración anterior, es la pregunta que da la mayor ganancia de información.
El contenido del catálogo de preguntas hasta este momento es de “P3P1” las cuales se van a combinar con las preguntas que no se han seleccionado en la siguiente iteración.

Dado que son varias iteraciones, los resultados de éstas se muestran en la Tabla 5.4.1.3 marcando en negritas las preguntas que se fueron seleccionando en cada iteración y combinando la pregunta seleccionada con las no seleccionadas en la iteración que sigue.

Iteración *n=3*

Atributo *k*	$H(C\|P_k)$	I_G	I_GR
$P_0P_1P_3$	**2.1801**	**1.0562**	**32.6363%**
$P_2P_1P_3$	2.1896	1.0467	32.3421%
$P_4P_1P_3$	2.2074	1.0289	31.7917%
$P_5P_1P_3$	2.2076	1.0287	31.7864%
$P_6P_1P_3$	2.2074	1.0289	31.7917%

Iteración *n=4*

Atributo *k*	$H(C\|P_k)$	I_G	I_GR
$P_2P_0P_1P_3$	2.0742	1.1621	35.9074%
$P_4P_0P_1P_3$	**2.0704**	**1.1659**	**36.0251%**
$P_5P_0P_1P_3$	2.0922	1.1441	35.3516%
$P_6P_0P_1P_3$	2.0704	1.1659	36.0251%

Iteración *n=5*

Atributo *k*	$H(C\|P_k)$	I_G	I_GR
$P_2P_4P_0P_1P_3$	**2.0015**	**1.2349**	**38.1570%**
$P_5P_4P_0P_1P_3$	2.0114	1.225	37.8510%
$P_6P_4P_0P_1P_3$	2.1037	1.1327	34.9996%

Iteración *n=5*

Atributo *k*	$H(C\|P_k)$	I_G	I_GR
$P_5P_2P_4P_0P_1P_3$	2.0014	1.2349	38.1576%
$P_6P_2P_4P_0P_1P_3$	2.0014	1.2349	38.1576%

Tabla 5.4.1.3 Resultados de las iteraciones realizadas al catálogo “X”

Hay que destacar que conforme se vaya avanzando en las iteraciones, las ganancias de información de cada una de las preguntas se van pareciendo más. Esto se debe a que las preguntas que se van agregando al catálogo llegan a ser, en cierto punto, iguales o muy aproximadas en la ganancia de información y puede ser que se llegue a un punto en el que escoger cualquier pregunta ya no importe.

Como se observa en la Tabla 5.4.1.3 se llega a escoger todas las preguntas, esto se debe a que existen alelos que se repiten ya sea en dos o más subclases dando lugar a que no se pueda definir a qué subclase pertenece. Por esta razón, aún seleccionando todas las preguntas iniciales no se llega a obtener el 100% de la información. Por lo cual el catálogo de preguntas para el catálogo de alelos de "X" queda formado por **[P3 P1 P0 P4 P2 P5 P6]**. Esta combinación la tendrá la sonda para clasificar los alelos.

5.4.2 INTERPRETACIÓN DEL CATÁLOGO DE PREGUNTAS

En la Tabla 5.4.2.1 se muestra la información contenida en cada pregunta que será la que se localizará en la sonda biológica que permitirá la clasificación de los alelos a una subclase. En dicha Tabla se muestra en la primera columna el orden en que será aplicada la pregunta, en la segunda columna se coloca la pregunta que se aplicará, en la tercera se presenta las posiciones que abarca la pregunta y en las siguientes se muestra la entropía, ganancia de información y ganancia de información relativa que la pregunta seleccionada genera.

Orden para Aplicar	No. de Pregunta	Posición que abarca	$H(C\|P_i)$	I_G	I_GR
1	P0	5	-	-	-
2	P1	5		-	-
3	P2	25, 29	T	T	T
4	P3	25, 29	C	A	A
5	P4	25, 29	A	C	C
6	P5	29	-	A	A
7	P6	29	-	C	C

Tabla 5.4.2.1 Información que genera el catálogo de preguntas

Este conjunto de preguntas seleccionado en el orden en que van definiéndose en el catálogo de preguntas al aplicarlas al órgano o receptor permitirán clasificar los alelos, como se mencionó. El tiempo utilizado para obtener los resultados por medio del programa en Java para el catálogo “X” fue de 0.07135 minutos.

Estos resultados fueron obtenidos en la herramienta Excel dando el mismo resultado para el programa realizado en el lenguaje Java, con lo cual se comprueba el funcionamiento correcto del programa.

Otros ejemplos pequeños fueron aplicados para ser clasificados en el programa en Java proporcionando los resultados para formar un catálogo final de preguntas que proporcionen una ganancia de información sin la necesidad de elegir todas las preguntas.

En la Tabla 5.4.2.2 se presentan las dimensiones de cada uno de los catálogos que se corrieron en el programa junto con el tiempo de corrida, el número de preguntas iniciales y el número de preguntas que conforman el catálogo de preguntas final.

No. Columnas	No. Renglones	No. Inicial Preguntas	No. Final Preguntas	Tiempo para Obtener Respuesta (min)	Ganancia de Información	
10	10	27	6	0.2713	100%	
11	17	58	15	1.6754	100%	
15	20	170	17	5.0883	100%	
15	25	240	20	8.629	100%	
20	25	227	22	10.1243	100%	
60	20	184	184	102.43	95.86%	*
10	14	7	7	0.071	38.15%	*

Tabla 5.4.2.2 Información de las corridas en el programa (los que están marcados con * significan que no se llegó ni al 99.98% de la ganancia debido a que los alelos se encuentran repetidos en otras subclases)

5.4.3 SELECCIÓN DEL CATÁLOGO "X"

Es necesario mencionar que se colocó en específico este problema de clasificación basado en el catálogo "X", esto es con el fin de asegurar el buen funcionamiento del programa en situaciones "especiales". En este caso, y como se comentó anteriormente, no se llega a una ganancia de información del 100% debido a que el contenido genético (combinación de bases nitrogenadas) de uno o más alelos se encontraban presentes en 2 ó más subclases. Los alelos que contienen información repetida en otras subclases se localizan en la Tabla 5.4.3.1 marcando con una cruz el alelo en donde se repite esta información.

Alelos	A*010101	A*010102	A*0102	A*020101	A*020102	A*020103	A*020104	A*020105	A*03010101	A*03010102N	A*4301
A*010101	-	X	-	-	X	-	-	-	X	X	-
A*010102	-	-	-	-	X	-	-	-	X	X	-
A*0102	-	-	-	-	-	-	-	-	-	-	-
A*020101	-	-	-	-	-	X	X	X	-	-	-
A*020102	-	-	-	-	-		-	-	-	X	X
A*020103	-	-	-	-	-	-	X	X	-	-	-
A*020104	-	-	-	-	-	-	-	X	-	-	-
A*020105	-	-	-	-	-	-	-	-	-	-	-
A*03010101	-	-	-	-	-	-	-	-	-	X	-
A*03010102N	-	-	-	-	-	-	-	-	-	-	-
A*4301	-	-	-	-	-	-	-	-	-	-	-

Tabla 5.4.3.1 Tabla que muestra con un “x” los alelos que se presentan en otro u otros alelos

Es por eso que al obtener los resultados en el programa en Java se seleccionan todas las preguntas y se obtiene únicamente una ganancia de información del 38.1576% por no poder clasificar a un alelo a una sola subclase.

5.5 OTRO EJEMPLO

Se muestra a continuación otro ejemplo que parte de un catálogo “Y” que se muestra en la Tabla 5.4.4.1 para mostrar más adelante la información que proporciona cada una de las iteraciones y finalmente presentar el catálogo de preguntas que se obtuvo del catálogo “Y”.

Alelos	0001	0002	0003	0004	0005	0006	0007	0008	0009
A*010101	T	T	A	T	C	T	G	C	A
A*010102	T	A	A	A	C	T	T	C	A
A*020101	T	C	A	A	C	T	A	A	A
A*020103	T	C	A	T	C	T	G	G	A
A*03010101	T	T	A	T	C	T	G	C	T
A*03010102N	T	C	A	T	C	T	C	T	A
A*4301	T	A	A	G	C	T	T	C	A

Tabla 5.4.4.1 Catálogo “Y”

Los resultados generados en cada iteración se presentan en la Tabla 5.4.4.2 generando un catálogo de preguntas.

Iteración	Pregunta Seleccionada	Porcentaje de Información
1	P0	31.0833%
2	P1	55.886%
3	P2	74.7340%
4	P3	88.0201%
5	P4	96.2303%
6	P5	100%

Tabla 5.4.4.2 Resultados de cada iteración para el catálogo “Y”

Este ejemplo, dado que es muy pequeño proporciona una ganancia de información del 100% seleccionando únicamente 6 preguntas de 27 que se tenían originalmente. Es decir, este catálogo de 6 preguntas [P0, P1, P2, P3, P4, P5] permitirá clasificar los alelos en un 100%. El tiempo de corrida en el programa en Java para este catálogo es de 0.27136 minutos.

Se coloca este ejemplo del catálogo “Y” únicamente con el fin de mostrar otra forma en la que el programa muestra los resultados y que sí elige unas cuantas preguntas para clasificar los alelos HLA sin necesidad de seleccionar todas las preguntas.

En el siguiente capítulo se proporcionan los resultados obtenidos al realizar la aplicación de la heurística de la entropía y la ganancia de información ayudándose de la herramienta computacional que es la programación, en este caso: Java.