CAPÍTULO 4

METODOLOGÍA Y RESULTADOS

En este cuarto capítulo se describirá el procedimiento para obtener el valor de la entropía de Shannon para el problema de clasificación de bialelos, se desarrollará la heurística de selección de atributos para la ganancia de la información con la cual se va a proporcionar como solución inicial un catálogo de preguntas. Después se aplicará un algoritmo para mejorar esta solución y finalmente se da una idea de cómo se pueden aplicar los resultados.

4.1 ENTROPIA DE SHANNON

En la sección 2.3 se describió como se creó la matriz de evaluación M_E de 38781 bialelos por 1117 preguntas y se debe recordar que la información que proporciona es para indicar que ningún bialelo presenta la pregunta, se identifica con el digito 0 y si para al menos un bialelo o para los dos la pregunta es afirmativa se identifica con el digito 1. Esta matriz M_E es la base para obtener el valor de la entropía y para desarrollar la heurística de selección de atributos.

El primer paso para el desarrollo de la entropía es obtener la probabilidad de cada microestado del sistema, es decir, la probabilidad de cada biclase. Teniendo como referencia que el número de biclases es 231 y el total de alelos son 38781, los alelos que existen en cada biclase se pueden observar en la matriz de evaluación M_E. La entropía de microestados no equiprobables para este problema de acuerdo a la expresión 3.4 se representa:

$$H(C) = -\sum_{i=1}^{38871} p_i \log_2 p_i \tag{4.1}$$

Desarrollando la ecuación:

$$H(C) = p_{1/1} \log_2 p_{1/1} + p_{1/2} \log_2 p_{1/2} + p_{1/3} \log_2 p_{1/3} + p_{1/11} \log_2 p_{1/11} + ... + p_{80/80} \log_2 p_{80/80} \tag{4.2}$$

Para conocer el valor de la entropía se utilizó un programa computacional en lenguaje C y el valor que se obtuvo es:

$$H(C) = 6.472276$$

Con este dato se va a desarrollar la heurística para la selección de atributos, los cuales corresponde al catálogo de preguntas $A = \{A_1, A_{2,} A_3, ..., A_{1117}\}$ y maximizar la ganancia de la información. A continuación se presenta el código del programa que fue utilizado para obtener el valor de la entropía.

Programa para obtener el valor de la entropía de Shanon H(C)

```
// Determinar HC
if ((fp=fopen("bialelos.txt", "rt"))!=NULL){
while (fgets (linea, 1024, fp)){
i=0;
while((linea[i]!='\0') && (nuevo==0)) if (linea[i]=='*') nuevo=1; else i++;
if(nuevo==1){
nuevo=0;
if (band!=0) {
hc=hc+((p/elementos)*(log(p/elementos)/log(2)));
p=0;
}else band=1;
fgets (linea, 1024, fp);
fgets (linea, 1024, fp);
}else{
fgets (linea, 1024, fp);
fgets (linea, 1024, fp);
p++;
```

```
}
}
hc=-1*(hc+((p/elementos)*(log(p/elementos)/log(2))));
fclose (fp); }
```

4.2 HEURISTICA PARA LA SELECCIÓN DE PREGUNTAS

En esta sección se desarrolla la heurística para la selección de atributos descrita en el apartado 3.2. Como ya se describió esta heurística se basa en la ganancia de la información la cual permite cuantificar la información proporcionada por un atributo y permite resolver el problema de clasificación. La definición de la ganancia de la información esta dada por la ecuación 3.5.

Para obtener el valor de $H(C \mid A_k)$ se calcula la media ponderada de la entropía de Shannon en cada subgrupo (ver ecuación 3.7).Y para calcular la entropía en cada subgrupo es necesario hacer referencia a la ecuación 3.8.

También se debe recordar que la función $p(C_i \mid S_j)$ es la probabilidad que un elemento pertenezca a la clase C_i si el elemento pertenece al subgrupo S_j. Aplicando estos conceptos y definiciones para los datos del problema y para la matriz de evaluación M_E de 38781x1117 se definen las variables de la siguiente forma:

E = conjunto de preguntas; $\{A_1, A_2, A_3, ..., A_{1117}\}$

v_k = número de valores que puede presentar el atributo o pregunta {0,1}

Sj = conjunto de subgrupos que se pueden presentar de acuerdo a los valores y las combinaciones de las preguntas; $S_j\,(j = 1, ..., v_k)$

W_i = la proporción de elementos en cada S_j

C_i = biclase; $c_i = \{1,2,3,...231\}$

N = bialelos; $b_n = \{1,2,3,...,38781\}$

El objetivo de esta heurística es ir seleccionando la pregunta que de una mayor ganancia de la información para formar un catálogo de preguntas y obtener una solución inicial. El procedimiento para la selección de cada pregunta es como se describió en la sección 3.2. Y los criterios de la ganancia de la información que se deben cumplir para obtener el catálogo de preguntas son:

- $I_G = H(C) = 6.472276$
- $I_G R = 100\%$

Con este catálogo de preguntas se va a proporcionar una combinación única de 1y 0 que permita identificar cada biclase y resolver el problema de clasificación de bialelos. Cabe mencionar que la dimensión de los datos de la matriz M_E es muy extensa y para procesar estos datos y obtener el catálogo de preguntas inicial se realizó un programa computacional en lenguaje C. Pero debido a la dimensión de los datos, el programa tardaba mucho tiempo en ejecutarse, generando una iteración por día, cada iteración proporciona la pregunta con mayor ganancia de la información y la añade al catálogo. Después de analizar esta situación se decidió ir eliminando las preguntas que no proporcionaban una ganancia de la información en cada iteración para acortar el tiempo de proceso del programa, un ejemplo se describe a continuación.

Tabla 4.1 Ejemplo de los resultados del programa para la heurística de selección de preguntas

Iteración	Pregunta	$H(C\|A_k)$	I_G	$I_G R$
4	A798	2.480699	3.991577	0.616719
5	A798	2.480699	3.991577	0.616719

Como se puede observar en la tabla 4.1 la pregunta A798 en la iteración 4 y en la iteración 5 tiene el mismo valor de I_G por lo que no proporciona ganancia de la información a partir de la iteración 5. Por lo que se puede eliminar con toda seguridad de que no dará ganancia de la información a partir de la iteración 6. Con esta condición el programa se logró ejecutar para obtener los resultados del catálogo de preguntas en 192 horas. A continuación se incluye el código del programa y en la sección siguiente los resultados obtenidos.

4.2.1 Programa para seleccionar las preguntas para la solución inicial

```
// llenar matriz con 1 y 0
nclase=0;
strcpy(clase,"abcdefghij");
strcpy(claseant,"abcdefghij");
nlinea=1;
if ((fp=fopen("evaluacion.txt", "rt"))!=NULL){
fgets (lineaeva, 10000, fp);
while (fgets (lineaeva, 10000, fp)){
i=0;
while(lineaeva[i]!=':'){
clase[i]=lineaeva[i];
i++;
}
clase[i]='\0';
i=i+2;
if(strcmp(clase,claseant)!=0){
nclase++;
strcpy(claseant,clase);
}
tablahc[0][nlinea]=nclase;
for(j=1;j<=columnapre;j++)
```

```
if(lineaeva[i+((j-1)*2)]=='0')tablahc[j][nlinea]=2;
else tablahc[j][nlinea]=lineaeva[i+((j-1)*2)];
nlinea++;
}
fclose (fp);
} // fin llenado de la matriz con 1 y 0
do{
strcpy(nombre,"S");
sprintf(nombrei,"%i",iteraciones);
strcat(nombre,nombrei);
strcat(nombre,".txt\0");
if ((fs = fopen(nombre, "w")) == NULL) {
printf("error en la apertura del archivo!\n");
exit(0);
}

fprintf(fs, "H(c) = %f\n\n", hc);
fprintf(fs, "Combinacion : ");
for(l=0;l fprintf(fs, "\n\n");
for (preguntas=1;preguntas b_seleccion=0;
for(l=0;l if (tablahc[preguntas][0]=='x'){
b_seleccion=1;
fprintf(fs, "Pregunta eliminada %i \n", preguntas);
}
if(b_seleccion==0){
for(l=0;l strcpy(miclasec[l].atributos,"");
miclasec[l].cuantos=0;
miclasec[l].queclase=0;
miclasec[l].TOTAL=0;
}
k=0;
for(i=0;i for(j=0;j combinaciones[j]=tablahc[seleccionados[j]][i+1];
}
//combinaciones[j]=tablahc[preguntas+1][i+1];
combinaciones[j]=tablahc[preguntas][i+1];
combinaciones[j+1]='\0';
//fin de la creacion de la cadena de combinacion
band=0;
for(l=0;l if(strcmp(combinaciones,miclasec[l].atributos)==0) band=1;
if(strcmp(miclasec[l].atributos,combinaciones)==0) miclasec[l].cuantos++;
}
if (band==0){
strcpy(miclasec[k].atributos,combinaciones);
miclasec[k].cuantos=1;
strcpy(comb_ant,combinaciones);
k++;
}
```

```
}//elementos
clase[0]=tablahc[0][1];
for(i=0;i for(aux=0;aux combinaciones[aux]=tablahc[seleccionados[aux]][i+1];
}
combinaciones[aux]=tablahc[preguntas][i+1];
combinaciones[aux+1]='\0';
//fin de la creacion de la cadena de combinacion
if(tablahc[0][i+1]==clase[0]){
for(j=0;j if(strcmp(miclasec[j].atributos,combinaciones)==0) {
miclasec[j].queclase++;
j=k;
}
}else{
for(j=0;j if((miclasec[j].cuantos!=0) && (miclasec[j].queclase!=0))
miclasec[j].TOTAL=miclasec[j].TOTAL+((miclasec[j].queclase/miclasec[j].cuantos)*(log(
miclasec[j].queclase/miclasec[j].cuantos)/log(2)));
for(j=0;j if(strcmp(miclasec[j].atributos,combinaciones)==0) miclasec[j].queclase=1;
else miclasec[j].queclase=0;
clase[0]=tablahc[0][i+1];
}
}//elementos
for(j=0;j if((miclasec[j].cuantos!=0) && (miclasec[j].queclase!=0))
miclasec[j].TOTAL=miclasec[j].TOTAL+((miclasec[j].queclase/miclasec[j].cuantos)*(log(
miclasec[j].queclase/miclasec[j].cuantos)/log(2)));
for(j=0;j if(miclasec[j].TOTAL!=0){
miclasec[j].TOTAL=miclasec[j].TOTAL*(miclasec[j].cuantos/elementos);
T=T+miclasec[j].TOTAL;
}
}
T=T*-1;
fprintf(fs, "A%i\t", preguntas);
fprintf(fs, "%f\t", T);
fprintf(fs, "%f\t", hc-T);
for(l=0;l if ((hc-T)==singanancia[l])
tablahc[preguntas][0]='x';
if(masalto masalto=hc-T;
columnaalta=preguntas;
}
fprintf(fs, "%f\t\n", (hc-T)/hc);
printf("pregunta : %i -> %f\n",preguntas,hc-T);
if(T<=0)salir=1;
T=0;
}

}
seleccionados[iteraciones]=columnaalta;
iteraciones++;
```

```
fprintf(fs, "%i\t <- La pregunta con el valor %f es el mas alto\n", columnaalta,masalto);
singanancia[cont_sg]=masalto;
cont_sg++;
columnaalta=0;
masalto=0;
fclose(fs);
}while(salir=0);
}
```

4.2.2 Catálogo de preguntas

El programa proporcionó un catálogo de 65 preguntas, con el cual se puede obtener una solución inicial para el problema de clasificación de bialelos HLA. Es importante señalar que no se pudo obtener un valor de $I_G = 6.472276$ porque en la iteración final todas las preguntas se eliminaron ya que no proporcionaron ninguna ganancia de la información. Sin embargo los resultados obtenidos son aceptables porque el valor de la ganancia de la información relativa es de $I_G R = 99.9875\%$ y un $I_G = 6.471466$, es decir el 99.9875% de los bialelos pueden ser identificados y clasificados con este catálogo de preguntas.

Además los resultados se mostraron al Dr. Javier Garcés, profesor del departamento de Química y Biología de la Universidad de las Américas, Puebla, quien esta a cargo de la investigación para la clasificación de bialelos HLA. Y para esta investigación es aceptable esta clasificación de bialelos con el 99% de probabilidad, por lo tanto, el catálogo solo podría incluir las primeras 22 preguntas (ver Tabla 4.2).

Tabla 4.2 Catálogo de 65 preguntas de la solución inicial

Iteración	Pregunta seleccionada	$H(C\|A_k)$	I_G	I_GR
1	A42	5.567057	0.905219	0.139861
2	A91	4.697415	1.774861	0.274225
3	A498	3.888624	2.583652	0.399188
4	A503	3.15231	3.319966	0.512952
5	A788	2.480699	3.991577	0.616719
6	A445	1.927597	4.544679	0.702176
7	A213	1.529739	4.942537	0.763647
8	A765	1.222245	5.250031	0.811157
9	A458	0.946711	5.525565	0.853728
10	A539	0.73019	5.742086	0.887182
11	A826	0.566947	5.905329	0.912404
12	A403	0.453741	6.018535	0.929895
13	A1099	0.35864	6.113637	0.944588
14	A8	0.282199	6.190077	0.956399
15	A333	0.227032	6.245244	0.964922
16	A701	0.187696	6.28458	0.971
17	A855	0.150694	6.321582	0.976717
18	A329	0.120521	6.351755	0.981379
19	A1048	0.098967	6.373309	0.984709
20	A166	0.080651	6.391625	0.987539
21	A466	0.065218	6.407059	0.989924
22	A1058	0.051522	6.420754	0.99204
23	A202	0.040142	6.432134	0.993798
24	A850	0.031052	6.441225	0.995202
25	A497	0.023026	6.44925	0.996442
26	A330	0.017325	6.454951	0.997323
27	A1079	0.012939	6.459337	0.998001
28	A368	0.009831	6.462445	0.998481
29	A540	0.007646	6.46463	0.998819
30	A442	0.006258	6.466018	0.999033
31	A346	0.005154	6.467122	0.999204
32	A7	0.004249	6.468027	0.999344
33	A842	0.003579	6.468697	0.999447
34	A987	0.003061	6.469215	0.999527
35	A1098	0.002613	6.469663	0.999596
36	A167	0.002273	6.470004	0.999649
37	A988	0.001988	6.470288	0.999693
38	A323	0.001728	6.470548	0.999733
39	A1020	0.001549	6.470727	0.999761
40	A635	0.00142	6.470856	0.999781
41	A857	0.001297	6.470979	0.9998
42	A561	0.001201	6.471075	0.999814
43	A10	0.001156	6.47112	0.999821
44	A865	0.001113	6.471164	0.999828
45	A375	0.001084	6.471192	0.999833
46	A23	0.001064	6.471212	0.999836

47	A642	0.00104	6.471236	0.999839
48	A9	0.001023	6.471253	0.999842
49	A738	0.001006	6.47127	0.999845
50	A1107	0.000989	6.471287	0.999847
51	A110	0.000974	6.471302	0.99985
52	A821	0.000955	6.471321	0.999852
53	A40	0.000942	6.471334	0.999854
54	A367	0.00093	6.471346	0.999856
55	A737	0.00091	6.471366	0.999859
56	A496	0.000898	6.471379	0.999861
57	A1027	0.000878	6.471398	0.999864
58	A473	0.000869	6.471407	0.999866
59	A1025	0.000856	6.47142	0.999868
60	A504	0.000847	6.471429	0.999869
61	A772	0.000834	6.471442	0.999871
62	A257	0.000826	6.47145	0.999872
63	A133	0.000821	6.471455	0.999873
64	A138	0.000816	6.47146	0.999874
65	**A763**	**0.00081**	**6.471466**	**0.999875**

4.3 ELIMINACIÓN DE PREGUNTAS

Ahora se tiene una solución inicial para el problema de clasificación de bialelos a partir de un catálogo de 65 preguntas, sin embargo, uno de los objetivos planteados para esta tesis es tratar de mejorar esta solución inicial y la heurística que se propone en este trabajo es:

1. Se tienen un catálogo inicial de preguntas $E_i = \{1,2,3,4,5,...65\}$
2. El valor de la entropía que ahora se define es el que proporcionó I_G en el catálogo de las 65 preguntas $H^*(C) = 6.471466$
3. De este catálogo inicial de preguntas, eliminar una pregunta y aplicar la heurística para la selección de preguntas para las 64 preguntas restantes, si se obtiene una ganancia de la información I_G igual al valor de H*(C) definido en el paso 2, entonces esta pregunta es eliminada del catálogo inicial E_i.
4. Si se eliminó la pregunta en el paso 3, se definirá un nuevo catálogo $E_i = \{1,2,3,4,5,...64\}$

5. Repetir el paso 3 y el paso 4 hasta que ya no se puedan seguir eliminando preguntas.
6. Obtener un catálogo con el menor número de preguntas

Para la realización de esta heurística también se elaboró un programa en lenguaje C, el código se presenta en el apéndice C, pero no se logró mejorar el catálogo inicial de preguntas porque al ir eliminando cada una de las preguntas del catálogo $A_i = \{1,2,3,4,5,...65\}$ no se pudo obtener un valor de $I_G = 6.471466$ por lo que se concluye que la combinación de las 65 preguntas son las que dan el valor de I_G obtenido en la heurística de selección de preguntas. Por lo tanto los resultados obtenidos en esta tesis se resumen con el catálogo de 65 preguntas obtenido en la fase de solución inicial. A continuación se va a dar una interpretación más extensa a estos resultados.

4.4 INTERPRETACIÓN DE LOS RESULTADOS

Esta sección puede servir como guía para aplicar los resultados que se obtuvieron en el presente trabajo de tesis. En primer lugar se explicará como debe interpretarse el catálogo de preguntas $E_i = \{1,2,3,4,5,...65\}$, es decir, se debe recordar que cada pregunta del catálogo inicial $E = \{A_1, A_{2,} A_3,..., A_{1117}\}$ contiene 20 posiciones y las combinaciones de letras (A, C. G. T) que se presentan en cada posible selección de 20 posiciones de los datos originales del problema que se presentan en la matriz de 178x129 (ver sección 2.1). Por lo tanto las 65 preguntas seleccionadas se presentan a continuación con 20 posiciones y las combinaciones de letras consecutivas.

Tabla 4.3 Catálogo de 65 preguntas con 20 posiciones y combinación de letras

	Pregunta seleccionada	Combinación de 20 posiciones	Combinación de letras
1	A42	25,29,33,40,	CAGG
2	A91	54,69,71,	GGA
3	A498	268,273,282,	CTA
4	A503	273,282,289,290,	TATA
5	A788	412,416,420,424,425,429,430,	TGCTCCA
6	A445	228,229,234,235,238,240,241,244,246,	AAGGCCTGG
7	A213	166,167,168,170,183,184,	GGCGCG
8	A765	404,407,412,416,420,	GGTAC
9	A458	234,235,238,240,241,244,246,251,	CGTGCTCC
10	A539	294,295,303,312,	TAGC
11	A826	429,430,433,444,446,448,	AAGGGC
12	A403	217,219,221,224,226,228,229,234,235,	CGCATGAGG
13	A1099	497,498,510,	CGT
14	A8	8,17,19,24,25,	CGATT
15	A333	197,198,202,205,209,216,	TGACGA
16	A701	343,345,346,347,350,353,	ATACCC
17	A855	444,446,448,450,451,453,454,457,	GGCCGGTC
18	A329	197,198,202,205,209,216,	TGACCA
19	A1048	482,484,486,487,488,491,497,498,	GGCGGCGT
20	A166	151,155,160,165,166,167,168,170,	GAAAGGCT
21	A466	238,240,241,244,246,251,	CCTGGC
22	A1058	484,486,487,488,491,497,498,	GACGCGT
23	A202	165,166,167,168,170,183,184,	GGGCGCT
24	A850	444,446,448,450,451,453,454,457,	GGCCAGTC
25	A497	268,273,282,	ATA
26	A330	197,198,202,205,209,216,	AGACCA
27	A1079	487,488,491,497,498,	GGCCG
28	A368	209,216,217,219,221,224,226,228,	CACGCAAA
29	A540	294,295,303,312,	TAGT
30	A442	226,228,229,234,235,238,240,241,244,	AAGCGTGCT
31	A346	202,205,209,216,217,219,221,	ACCACCC
32	A7	5,8,17,19,24,	CCGAA
33	A842	433,444,446,448,450,451,	GGGCCA
34	A987	457,465,466,472,	CTTC
35	A1098	491,497,498,510,	CGTC
36	A167	151,155,160,165,166,167,168,170,	GAAGGTCG
37	A988	457,465,466,472,	CTGC
38	A323	195,197,198,202,205,209,	AAGACG
39	A1020	472,480,482,484,486,487,488,491,	CGGGACGC
40	A635	331,338,339,340,341,343,345,346,347,350,	GCCAGAGACT
41	A857	444,446,448,450,451,453,454,457,	GGCCGCGC
42	A561	312,318,322,323,324,326,329,331,	TGGCTCCG
43	A10	8,17,19,24,25,	CGATA
44	A865	446,448,450,451,453,454,457,465,	GCCAGTCT
45	A375	209,216,217,219,221,224,226,228,	CACGCATG
46	A23	17,19,24,25,29,33,	GATATG

47	A642	338,339,340,341,343,345,346,347,350,353,	CCGGAGACCC
48	A9	8,17,19,24,25,	CGATC
49	A738	372,375,380,383,390,	GCAGC
50	A1107	510,522,	TC
51	A110	90,98,103,107,	ACGG
52	A821	425,429,430,433,444,	CCAGG
53	A40	24,25,29,33,40,	TTAGC
54	A367	209,216,217,219,221,224,226,228,	CACGCACA
55	A737	372,375,380,383,390,	GCCGC
56	A496	258,268,273,	CAA
57	A1027	472,480,482,484,486,487,488,491,	CGGGCTGC
58	A473	240,241,244,246,251,258,	GCTCCG
59	A1025	472,480,482,484,486,487,488,491,	CGGGACGA
60	A504	273,282,289,290,	TATG
61	A772	407,412,416,420,424,425,	GTACTC
62	A257	183,184,186,188,192,195,197,198,202,	GAGGGAAGA
63	A133	121,127,130,139,	CAGT
64	A138	127,130,139,141,146,	AGCTG
65	A763	395,404,407,412,	TGGG

Se debe recordar que estas preguntas corresponden a las posiciones en las que se van a aplicar los reactivos para saber a qué biclase pertenece el órgano del receptor o del donador. Y si se presenta en por lo menos uno de los dos alelos la combinación de letras la respuesta afirmativa, se identificará con el digito 1 y si no se presenta la combinación de letras en el par de alelos analizados la respuesta negativa y se representa con el digito 0.

Ahora es importante señalar que con el catálogo de 65 preguntas se obtuvo una caracterización única de 0`s y 1`s para poder identificar a qué biclase pertenece los alelos HLA-I-A de un ser humano. Se obtuvieron específicamente 16058 combinaciones diferentes y cada una tiene una cadena de 65 posiciones de 0`s y 1`s. Cada posición corresponde a cada una de las 65 preguntas; con estas 16058 combinaciones se pueden identificar a las 231 biclases. El archivo donde se encuentran estas 16058 combinaciones se encuentra en el apéndice E.

Además se realizó un programa computacional en el cual se pide introducir una cadena de 65 caracteres de las 16058 combinaciones de 0`s y 1`s. Y este programa puede decir a que biclase pertenece la cadena introducida. Por lo tanto si una persona desea aplicar este método para analizar un órgano que contenga los alelos HLA-I-A solo tiene que seguir los siguientes pasos:

1. Aplicar los reactivos en las posiciones correspondientes al catálogo de 65 preguntas seleccionadas.
2. Obtener una combinación de 65 dígitos de 0`s y 1`s
3. Introducir al programa la combinación de 0`s y 1`s obtenidas y el programa dice a que biclase pertenece esta combinación

El programa se presenta dentro del apéndice E, el cual es un disco compacto.