Patógenos

Informe final

Determinación de bacterias patógenas/no patógenas humanas por presencia/ausencia de genes

Estudiantes:
Juan Braga
Martin Etchart

Tutores:
Pablo Cancela
Federico Lecumberry

Contenidos

- Introducción
  - Descripción del problema
  - Objetivos
- Exploración de la base
  - Características de los datos
  - Resultados primarios para patogenicidad
  - Resultados primarios para taxonomía
  - Reducción de la dimensionalidad
    - Diffusion maps
- Clasificación por taxonomía
  - Selección de características
  - Selección de características por nivel jerárquico de taxonomía
  - Medida de la correlación entre atributos seleccionados a partir de un distinto nivel jerárquico.
- Clasificación por patogenicidad
  - Depuración de la base
    - Filtrado por patogenicidad
    - Selección de características
  - Resultados de clasificación
    - Resultados de trabajos previos
    - Resultados para la base depurada
  - Incorporación de taxonomía a la clasificación
    - Taxonomía como atributo
    - Clasificadores por taxonomía o grupos taxonómicos
- Conclusiones
  - Resultados
  - Trabajo a futuro
- Bibliografía

1. Introducción

1.1. Descripción del problema

Aunque la mayoría de las bacterias son inocuas o beneficiosas para sus huéspedes, otras son agresivas y pueden causar serias enfermedades y hasta la muerte. Dado que los costos de secuenciamiento de genes son cada vez más económicos y veloces, existen secuenciamientos completos de genomas de bacterias patógenas y no-patógenas para humanos. Datos de presencia o ausencia de una o más genes en el genoma de una bacteria pueden permitir la predicción de si ésta será o no patógena.

1.2. Objetivos

Diseñar un sistema restringido a los datos de patógenos humanos y no patógenos que minimice el error de clasificación, basándose en la presencia o ausencia de un conjunto a determinar de genes.

Diseñar un sistema restringido a los datos de patógenos humanos y no patógenos que minimice el error de clasificación utilizando un conjunto de 120 características (genes).

Incorporación de la información de taxonomía como un atributo o etapa de pre-clasificación para clasificación de patógenos/no patógenos humanos.

2. Exploración de la base

2.1. Características de los datos

La base de datos con la que trabajamos consiste en una lista de 767 bacterias, como se muestra en la Tabla 2.1, donde se tiene para cada una de ellas Id, Nombre, Taxonomía, presencia o ausencia de 814 genes y Patogenicidad.

Id	*Nombre de bacteria*	*Taxonomía*	*Presencia/ausencia de genes*				*Patogenicidad*
			gen 1	*gen 2*	...	*gen 814*	
1	Riemere...	Bacteroide...	1	0	...	0	Bird
2	Mycopla...	Firmicutes...	1	1	...	0	No
...	...						
767	Escheri...	Gammapro...	0	1	...	1	Human

Tabla 2.1 - Ejemplo de la base de datos utilizada para el proyecto.

En la Figura 2.1 se muestra las distribución de las muestras de la base según su patogenícidad en donde se puede ver que la base consiste principalmente en bacterias patógenas humanas y no patógenas humanas.

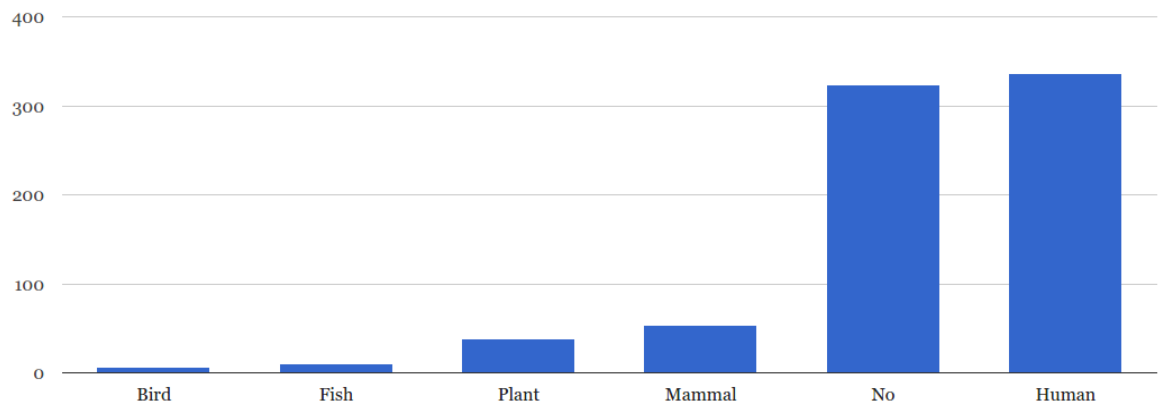

Figura 2.1 - Cantidad de muestras según patogenicidad.

La taxonomía, familia o grupo taxonómico, puede ser una de los diez opciones que se muestran en la Figura 2.2. En esta figura se puede ver la distribución de los datos según la taxonomía y se indica con color, para cada familia, la cantidad de las bacterias que son patógenas para cada una de las especies contempladas en la base: Ave, Pez, Humano, Mamífero, No humano, Planta.

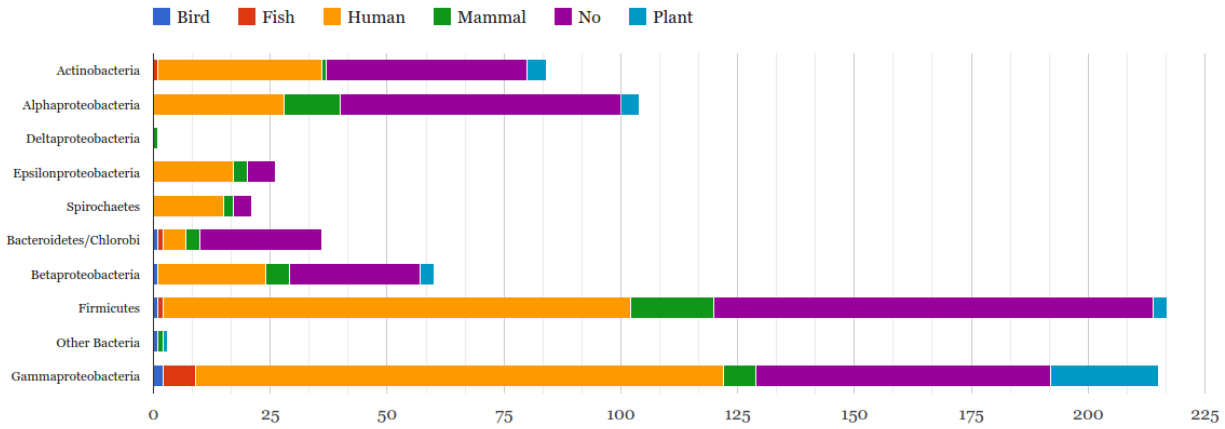

Figura 2.2 - Distribución de los datos por taxonomía coloreadas patogenicidad.

2.2. Resultados primarios para patogenicidad

El objetivo principal de la pruebas primarias para patogenicidad realizadas sobre la base era el de lograr una primera aproximación al problema que se planteaba en términos de resultados de clasificación.

Se utilizó la base completa de 767 bacterias como instancias o muestras etiquetados en patogenicidad según la especie a la que afecta y 814 genes como atributos. No se utilizó en ninguna forma la información de taxonomía para esta parte

Se eligieron dos clasificadores para realizar las pruebas basándonos en la experiencia de los prácticos realizados en el curso, k-vecinos más cercanos k-NN y árboles de decisión C4.5. En el caso de k-NN se utilizó el método para selección automática del número de vecinos resultando en la selección del primer vecino más cercano por lo que finalmente se utilizó el clasificador 1-NN. En el trabajo realizado en [1] se utiliza un clasificador para patógenos humanos y no patógenos humanos basado en un SVM con núcleo lineal por lo que también se utilizó este clasificador para las pruebas de patogenicidad con la diferencia que en este caso es un clasificador multiclase. Todos los clasificadores fueron utilizados con parámetros por defecto, excepto en los casos en los que se indica, en un esquema de validación cruzada de 10 particiones. Para todas las tareas de entrenamiento y clasificación se utilizó el software Weka.

En la Tabla 2.2 se muestran los resultados de clasificación para cada clasificador según su patogenicidad. El clasificador SVM supera en promedio a C4.5 y a 1-NN aun que 1-NN también logra resultados similares. La clasificación en Bird y Fish es muy pobre mientras que para Human y No los resultados son razonablemente buenos. Hay una clara relación entre los resultados de clasificación y la cantidad de muestras de cada clase disponibles en la base.

El problema como está planteado no cumple con el criterio de dimensiones usual ya que no se cumple la proporción 10:1 de muestras de entrenamiento por característica siendo esta proporción inferior a 1.

	Bird	*Fish*	*Human*	*Mammal*	*No*	*Plant*	*Weighted Average*
1-NN	0.000	0.300	0.860	0.434	0.861	0.737	**81.1%**
C4.5	0.000	0.100	0.804	0.377	0.824	0.658	**76.0%**
SVM	0.000	0.200	0.872	0.528	0.864	0.737	**82.3%**

Tabla 2.2 - Resultados primarios de clasificación según patogenicidad para cada uno de los clasificadores probados.

2.3. Resultados primarios para taxonomía

Debido a que uno de los objetivos era el de analizar la posibilidad de la incorporación de la información de taxonomía para mejorar los resultados de clasificación, se comenzó por analizar que tan discriminativos eran los genes como atributos para clasificar por taxonomía, en este caso asumiendo que no se conoce nada sobre la patogenicidad.

Para esto, al igual que en la parte anterior, se utilizaron los mismos clasificadores con los mismos parámetros sobre la base completa, 767 bacterias, 814 genes, y etiquetados por grupo taxonómico. En la Tabla 2.3 se muestran los resultados para la clasificación en donde se puede ver que que el clasificador 1-NN supera a SVM y a C4.5 en este caso.

	Errores	*Aciertos*	*Weighted average*
1-NN	34	733	**95.57%**
C4.5	41	726	**94.66%**
SVM	63	704	**91.79%**

Tabla 2.3 - Resultados primarios de clasificación según taxonomía para cada uno de los clasificadores probados.

Debido a los buenos resultados de clasificación para cualquiera de los 3 clasificadores se puede inferir que la información de taxonomía están fuertemente representadas por el conjunto de genes de la base. La matriz de confusión para la clasificación con 1-NN se muestran a continuación en donde se puede ver que las confusiones son pocas y las principales confusiones se dan entre *Betaproteobacteria* y *Gammaproteobacteria* que pertenecen a un mismo subgrupo taxonómico.

a	b	c	d	e	f	g	h	i	j	<-- classified as
33	0	2	0	0	1	0	0	0	0	a = Bacteroidetes/Chlorobi
0	53	1	4	0	0	1	1	0	0	b = Betaproteobacteria
2	0	214	0	1	0	0	0	0	0	c = Firmicutes
2	4	1	206	0	0	1	1	0	0	d = Gammaproteobacteria
0	0	3	0	0	0	0	0	0	0	e = Other Bacteria
1	0	2	0	0	81	0	0	0	0	f = Actinobacteria
0	0	0	0	0	0	26	0	0	0	g = Epsilonproteobacteria
1	0	1	2	0	0	0	100	0	0	h = Alphaproteobacteria
0	0	1	0	0	0	0	0	20	0	i = Spirochaetes
0	0	0	0	0	0	0	0	1	0	j = Deltaproteobacteria

2.4. Reducción de la dimensionalidad

Debido a las características de alta dimensión del problema en el que se tienen 814 características binarias por muestra se planteó la necesidad de utilizar técnicas de reducción de la dimensionalidad como forma de visualizar y entender mejor los datos. Para esto se utilizó el toolbox *drtoolbox* para Matlab basado en [3][4] disponible para descarga en la web[1]. Se probó con *isomaps*, *diffusion maps* entre otras técnicas y se eligió *diffusion maps* para realizar un análisis de los resultados que sigue.

2.4.1. Diffusion maps

Diffusion maps es una técnica de reducción de la dimensionalidad no lineal y no supervisada que se centra en descubrir una variedad o manifold oculto de donde los datos fueron muestreados. Se utilizaron los datos del espacio de 814 características o genes como entrada al algoritmo de donde se obtiene una representación de cada muestra o bacteria en un espacio tridimensional para facilitar la visualización.

En la Figura 2.3 se muestran diferentes vistas de reducción de la dimensionalidad con diffusion maps en donde las muestras se representan en un espacio de 3 dimensiones y se colorean según su patogenicidad. A simple vista no parece haber una separación clara entre clases, se puede ver cierta estructura y agrupamientos entre los puntos pero la patogenicidad no se ve bien representada en en este espacio.

En la Figura 2.4 se muestran la misma reducción de la dimensionalidad realizada pero en este caso se colorean según las muestras según su taxonomía.En este caso lo agrupamiento son más claros y se puede decir que la taxonomía se ve bien representada en este espacio reducido. Esta visualización de los datos permite explicar los resultados primarios obtenidos para clasificación por taxonomía y como un clasificador tan simple como 1-NN logra esos resultados.

[1] http://lvdmaaten.github.io/drtoolbox/

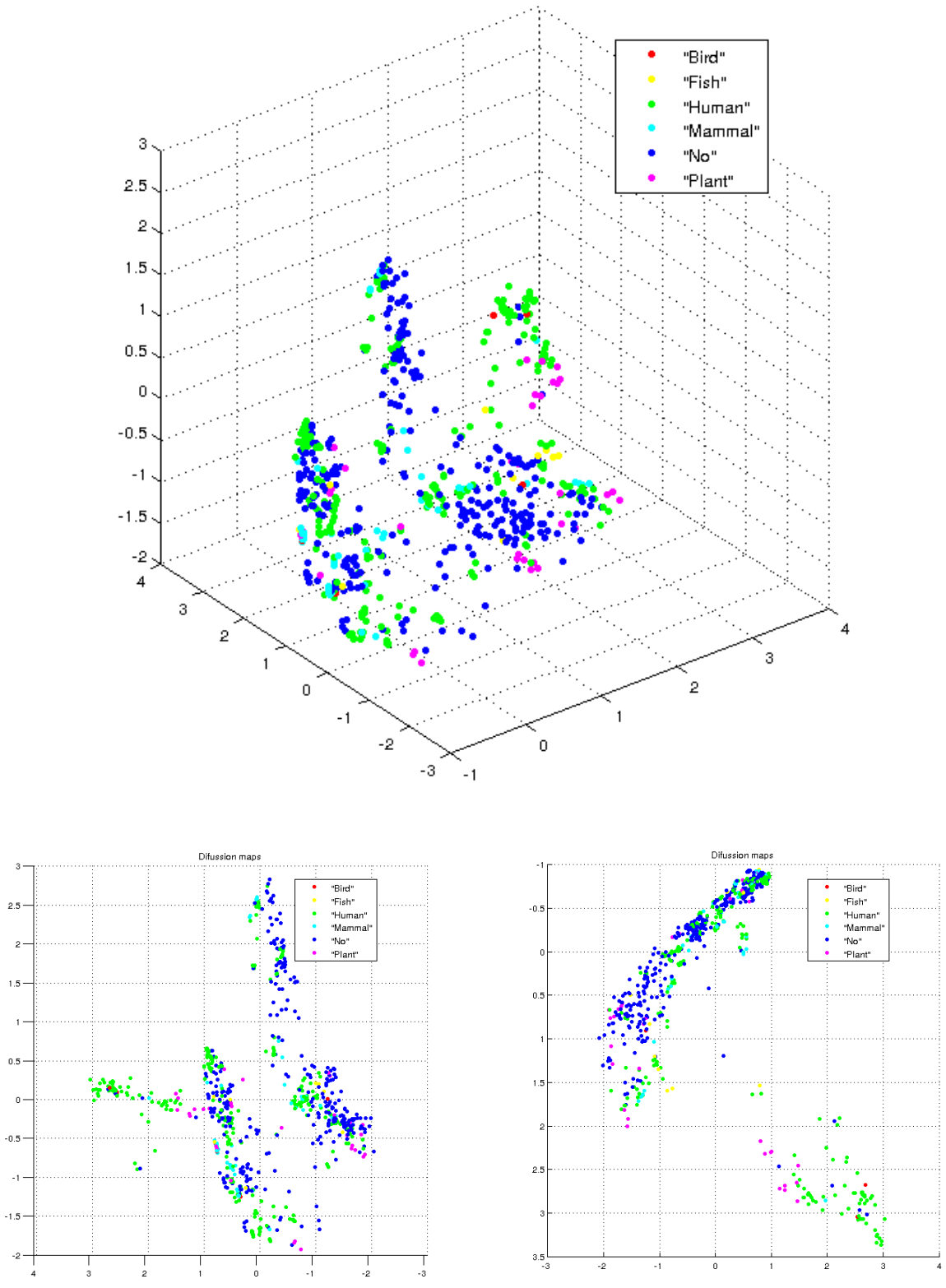

Figura 2.3 - Proyecciones de diffusion maps para reducción de dimensionalidad a 3d indicando patogenicidad.

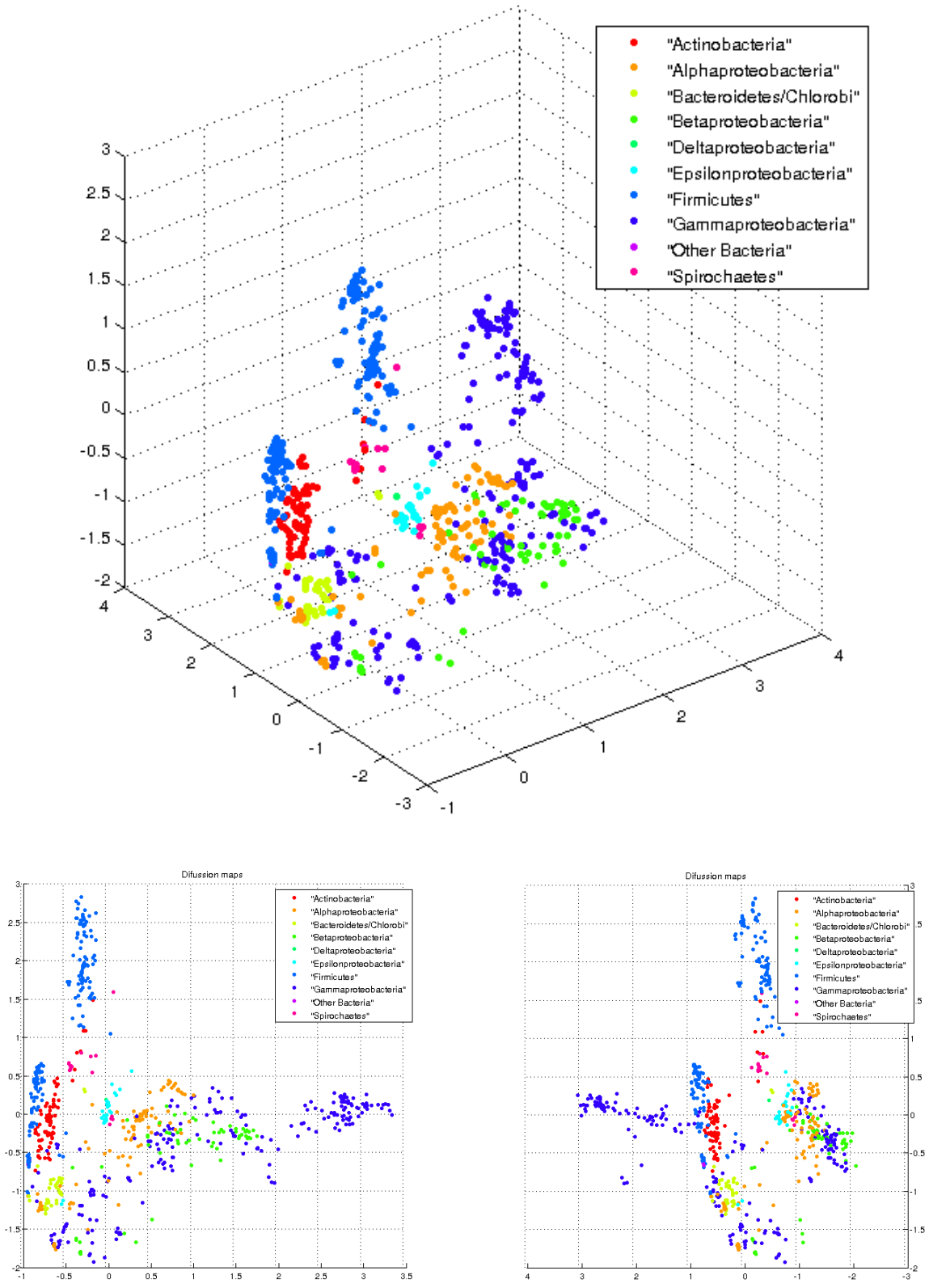

Figura 2.4 - Proyecciones de diffusion maps para reducción de dimensionalidad a 3d indicando taxonomía.

Para profundizar en las observaciones realizadas y confirmar numéricamente estas conclusiones se marcaron los errores de clasificación de 1-NN para taxonomía como se muestra en la Figura 2.5. Para cada uno de estos errores indicados con un cuadrado negro se indica la familia en la que se clasificó. Se puede ver que los errores señalados son coherentes a la ubicación espacial de las muestras en el espacio 3d.

Para verificar esto se calcularon las distancias en 814d y en 3d de cada muestra clasificada incorrectamente al vecino más cercano de la clase real y al vecino más cercano del resto de las clases. En todos los casos y para el cálculo en los dos espacios la distancia al vecino más cercano del resto de las clases fue inferior a la distancia al vecino mas cercano de la clase real.

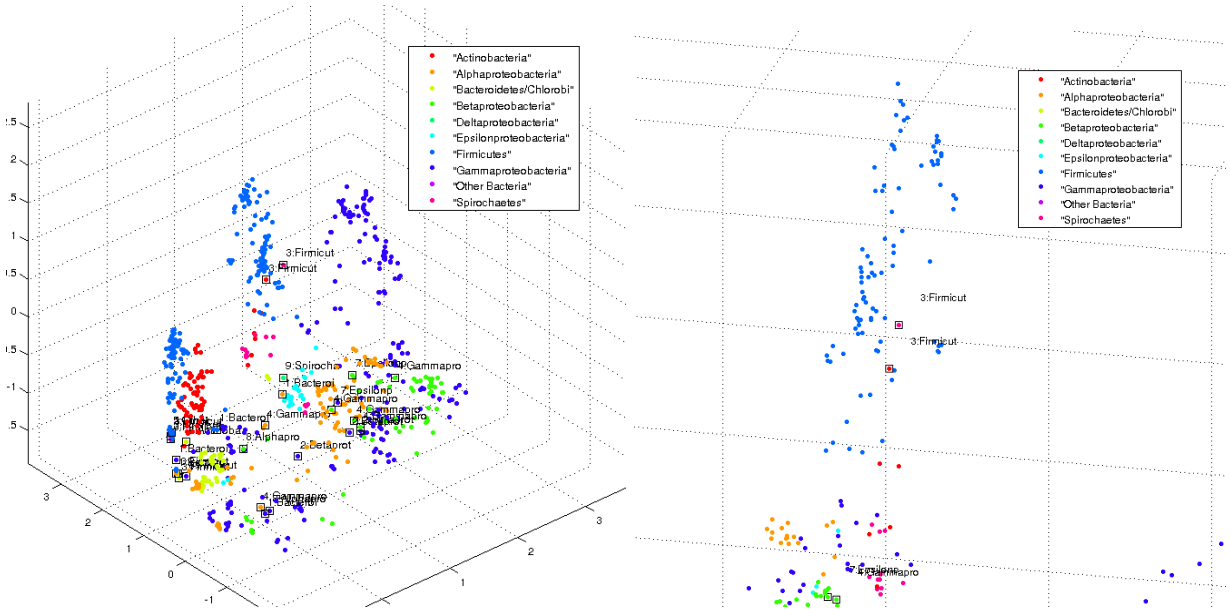

Figura 2.5 - Errores de clasificación de 1-NN por taxonomía en el espacio reducido.

3. Clasificación por taxonomía

Como se pudo ver en la Sección 2.3, los resultados primarios para clasificación por taxonomía dieron valores de acierto alentadores con el objetivo de agregar esta información a la clasificación por patogenicidad. En lo que sigue se detalla el estudio realizado sobre la base de datos y las relaciones entre atributos que esconde la misma.

3.1. Selección de características

Se comenzó aplicando algunos de los métodos de selección de características vistos en el curso a los datos, donde en este caso el label es la taxonomía. Se trabajó con le herramienta Weka.

Taxonomía	*CFeatureSelect + BestFirst*	*GainRatioEval + Ranker (50)*	*GainRatioEval + Ranker (100)*
Atributos	36	50	100
Acierto	95.8%	92.4%	95.4%

Tabla 3.1 - Selección de características primarias para taxonomía.

La mejor selección la hizo *CrossFeatureSelection + BestFirst* por lo que es la que usaremos en lo que sigue.

3.2. Selección de características por nivel jerárquico de taxonomía

Dada la estructura jerárquica que presenta la taxonomía (ver Figura 3.1 [1]) del problema se propuso un abordaje que la contemple con el objetivo de entender la relación inherente de los datos.

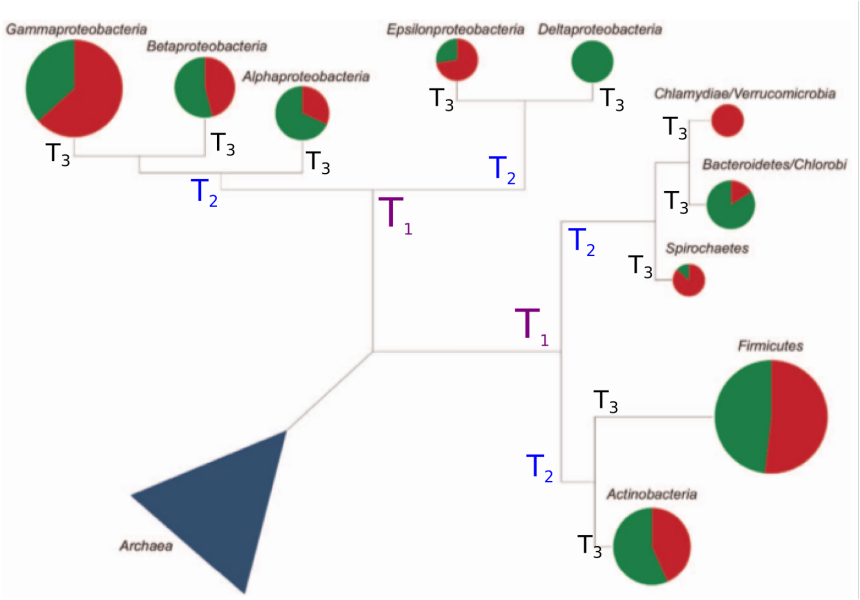

Figura 3.1 - Árbol jerárquico de taxonomías extraído de [1].

Se generaron atributos nominales del tipo binario que expresan la pertenencia o no, a una familia o agrupación de familias, en distintos niveles del árbol. Representando de esta forma el árbol de la Figura 3.1. Se utilizaron tres niveles de profundidad que se detallan a continuación (cada letra es la inicial del nombre de la familia):

1. T1ABG-DE / T1FA-SBCCV
2. T2ABG / T2ED / T2FA / T2SBCCV
3. T3A / T3B / T3G / T3E / T3D / T3F / T3A / T3S / T3BC / T3CV

Para cada uno de estos nuevos atributos se hizo una selección de características utilizando *CfsSubsetEval + BestFirst* de Weka. De esta forma se mantiene una relación jerárquica entre los atributos seleccionados para nodos relacionados por parentezsco. En la planilla del Anexo 1 se resume los atributos seleccionados para los distintos niveles. Algunas aclaraciones importantes:

- El árbol comienza en el centro y profundiza hacia los laterales.
- Está subdividido en las dos ramas de más alto nivel, representadas por un cuadro separado.
- Los colores de la fila superior representan parentesco entre columnas.
- Las coincidencias de genes se representan en las filas. Entre cuadros distintos sólo se respeta la coincidencia para las columnas centrales (las de más alto nivel).

Cabe destacar que a priori era razonable esperar mayor nivel de coincidencias de atributos seleccionados para un nivel y sus inferiores (por ejemplo los seleccionados en T1ABG-DE y los que salen de T2ABG y T2DE).

Se pasó entonces a medir el desempeño de los clasificadores 1-NN y C4.5 para los distintos subconjutos de características seleccionadas:

1-NN	*T1*	*T2*	*T3*
ABG-DE	98.6%	98.4%	98.6%
FA-SBCCV	97.0%	98.0%	98.7%

1-NN	*T2*	*T3*
ABG	98.9%	98.9%
DE	99.1%	99.2%

1-NN	*T2*	*T3*
FA	98.8%	98.4%
SBCCV	93.9%	98.3%

C4.5	*T1*	*T2*	*T3*
ABG-DE	98.9%	99.1%	97.9%
FA-SBCVV	98.4%	97.9%	98.7%

C4.5	*T2*	*T3*
ABG	99.1%	97.8%
DE	98.8%	98.7%

C4.5	*T2*	*T3*
FA	98.8%	98.4%
SBCCV	94.0%	96.7%

Tablas 3.2 - Resultados de clasificación por taxonomía para subgrupos seleccionado de características.

- Se observa que los niveles de acierto en clasificación por taxonomía siguen siendo altos, a pesar de haber bajado de aproximadamente 800 atributos a 20. Por lo que parece ser conveniente agregar la información de taxonomía a la clasificación por patogenicidad en caso que favorezca los aciertos.
- A pesar de la poca coincidencia entre los subgrupos de atributos seleccionados para todos ellos se logran buenos resultados de clasificación, por lo que existe redundancia en los datos.
- A pesar de lo anterior, los atributos que salen de la selección de carcaterísticas en el nivel T3 generalmente obtienen los mejores niveles de acierto.

3.3. Medida de la correlación entre atributos seleccionados a partir de un distinto nivel jerárquico.

Debido al bajo nivel de coincidencia entre atributos seleccionados para un nivel y sus inferiores y, además a los altos niveles de acierto para todos los subconjuntos de características, se procedió a medir la correlación entre los atributos de distintos grupos. Para esto se generó en Matlab matrices de correlación entre conjuntos comparando uno a uno todos los elementos.

A continuación se pueden ver las matrices desplegadas de forma gráfica donde por fila solamente se representan los tres atributos con mayor correlación y el valor correspondiente en intensidad de amarillo. Las matrices se muestran en las Figuras 3.2 a 3.7

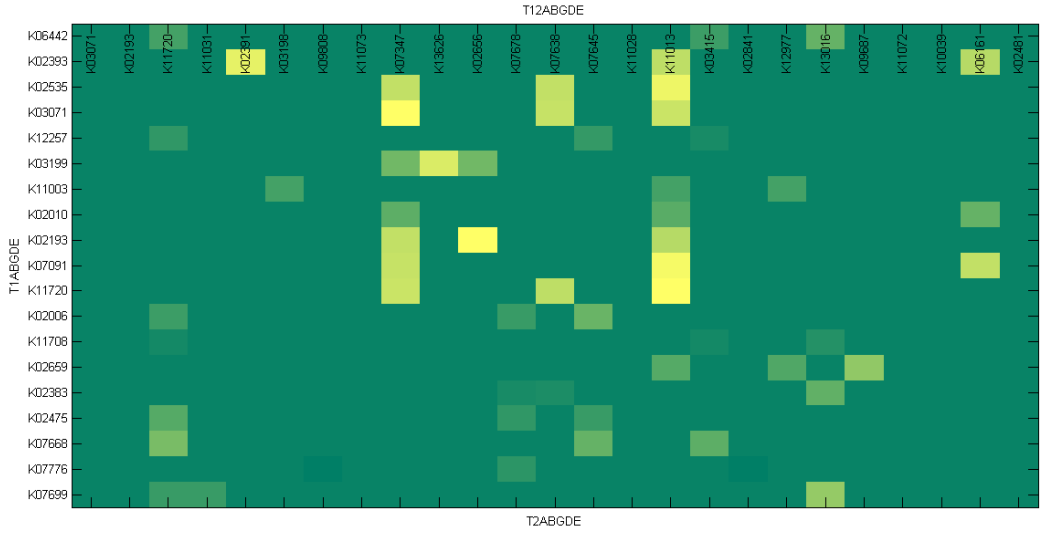

Figura 3.2 - Matriz de correlación T12ABGDE = T1ABGDE vs. T2ABGDE.

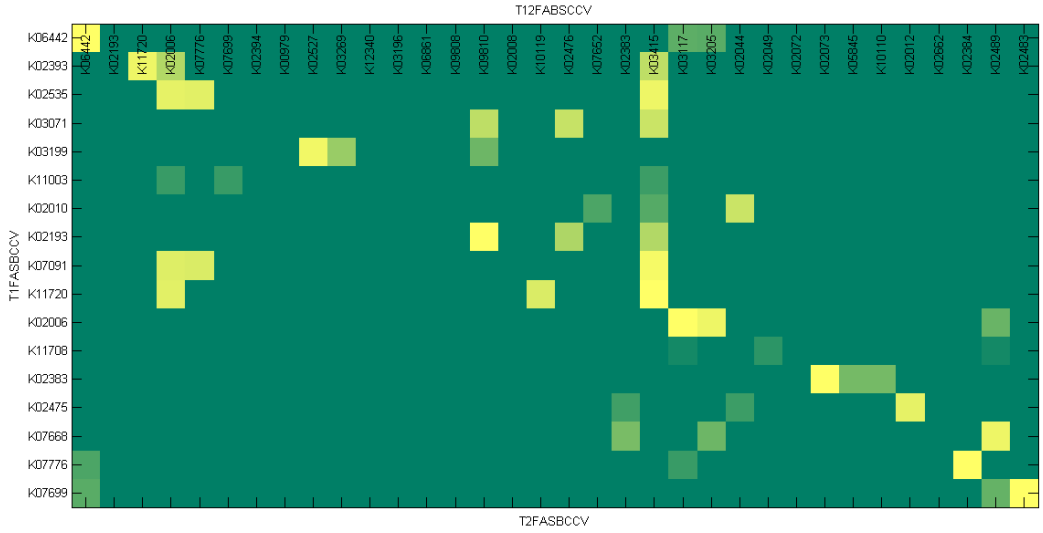

Figura 3.3 - Matriz de correlación T12FASBCCV = T1FASBCCV vs. T2FASBCCV.

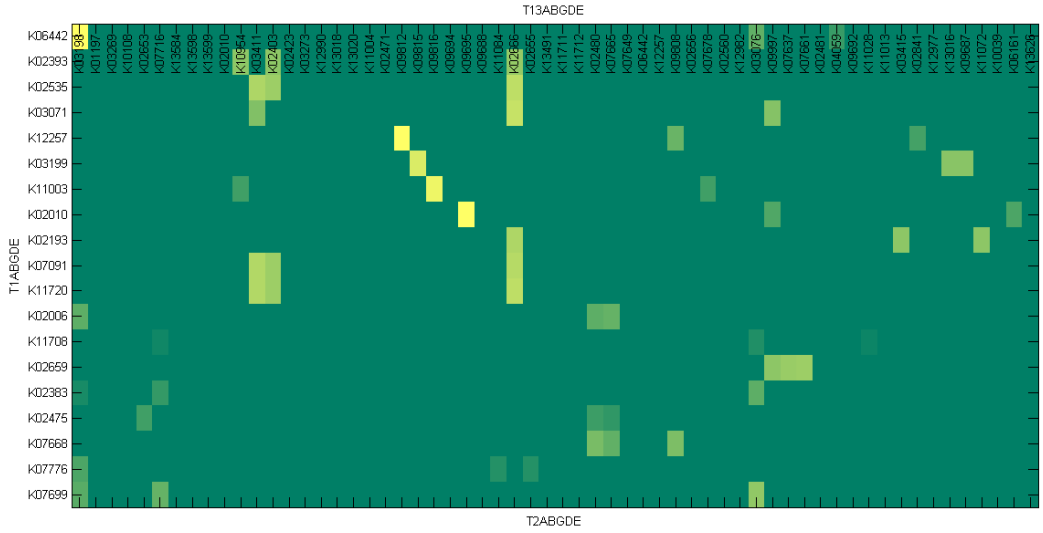

Figura 3.4 - Matriz de correlación T13ABGDE = T1ABGDE vs. T3ABGDE.

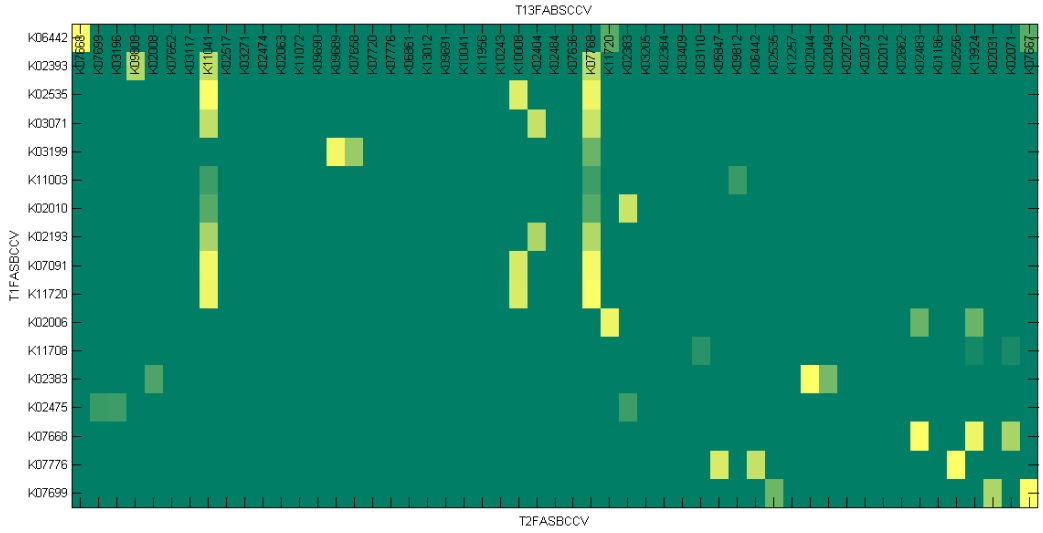

Figura 3.5 - Matriz de correlación T13FASBCCV = T1FASBCCV vs. T3FASBCCV.

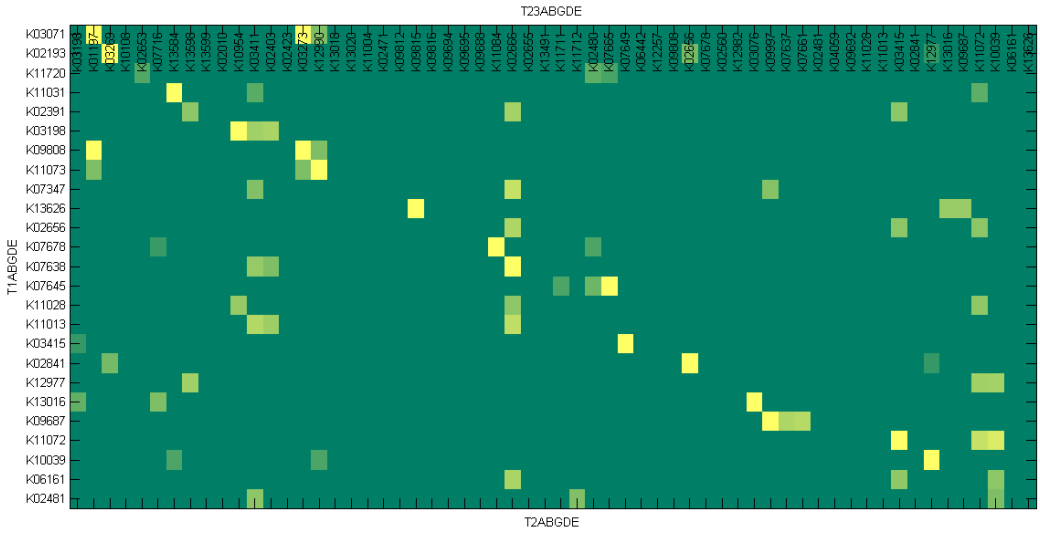

Figura 3.6 - Matriz de correlación T23ABGDE = T2ABGDE vs. T3ABGDE.

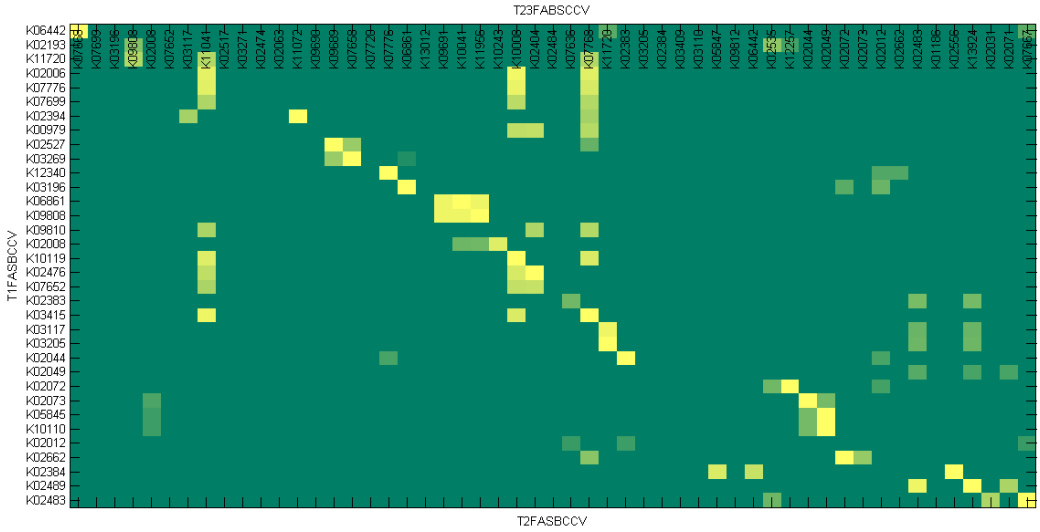

Figura 3.7 - Matriz de correlación T23FASBCCV = T2FASBCCV vs. T3FASBCCV.

4. Clasificación por patogenicidad

El objetivo planteado para clasificación por patogenicidad consistió en estudiar formas de incorporar la información de taxonomía en la clasificación y realizar una comparación con los resultados reportados en [1].

4.1. Depuración de la base

4.1.1. Filtrado por patogenicidad

La base original fue filtrada para utilizar solo las muestras para patógenos/no patógenos humanos. Para esta base se tiene 660 bacterias con una distribución como se muestra en la Figura 4.1. En esta base el número de familias se reduce de 10 a 8 y la patogenicidad no esta equilibrada para todas las familias.

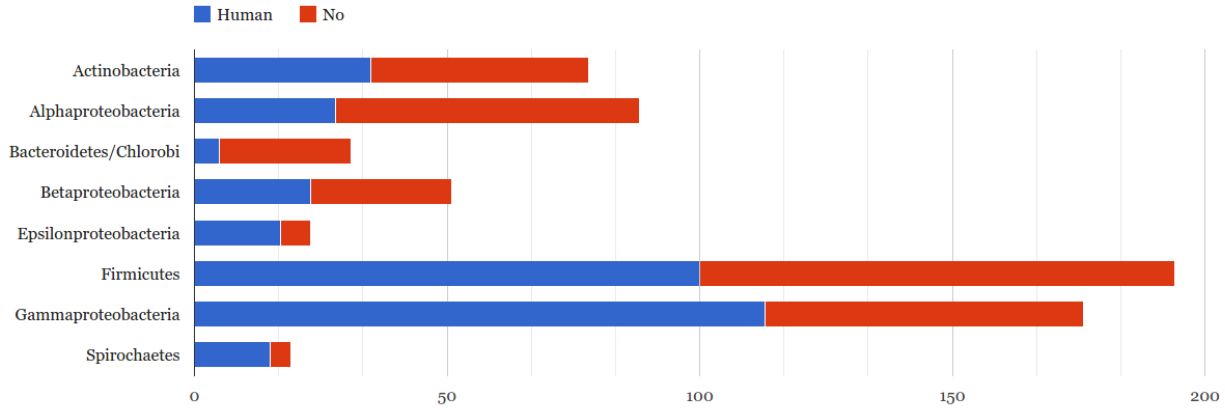

Figura 4.1 - Distribución de los datos por taxonomía para la base filtrada por patógenos/no patógenos humanos.

4.1.2. Selección de características

Las características seleccionadas para esta parte fueron tomadas del material suplementario de [1] en donde se listan los 120 genes seleccionados con el método allí descrito. Sobre la base filtrada por patogenicidad se generó una base final para las pruebas que siguen con únicamente esos 120 genes. Para este número de atributos la relación muestras por atributo se acerca a criterios más razonables quedando en 5.5:1.

4.2. Resultados de clasificación

4.2.1. Resultados de trabajos previos

En [1] se propone un clasificador C-SVM con núcleo lineal en un esquema de validación cruzada de 10 particiones. La selección de características se realiza utilizando el *SVMAttributeEval* de Weka en donde se realiza la selección de características para cada partición y finalmente mediante un esquema de votación se genera un Ranking de donde el número de características a utilizar puede ser elegido manualmente. Los autores encuentran mejor performance con un número de entre 120 y 150 características y finalmente reportan resultados para las 120 características que se encuentran en el material mencionado

anteriormente. Dentro del software *Bacfier*[2] que distribuyen como parte del trabajo se puede encontrar la base utilizada por los autores, `small120.arff`, la cual fue probada para contrastar lo que ellos reportan y lo que se obtiene con el clasificador como fue utilizado.

4.2.2. Resultados para la base depurada

Se utilizó tanto para la clasificación de la base depurada como para la base small120.arff un clasificador C-SVM con núcleo lineal y el resto de parámetros por defecto. También se utilizó selección automática de parámetros para seleccionar el parámetro de costo del clasificador variando C de 0.1 a 10.0 con un paso de 0.1 mediante *CVParameterSelection*.

En la Tabla 4.1 se muestran los resultados tanto para la base depurada como para la base small120.arff y los resultados reportados en el trabajo anterior. Para la base depurada los resultados superan ampliamente la clasificación utilizando los 814 genes completos aunque no alcanzan los niveles de acierto reportados. Un factor que puede estar influyendo en este incremento en los resultados puede ser la dimensionalidad del problema después de la selección de atributos.

C-SVM núcleo lineal	*Porcentaje de aciertos*	
	C=1.0	*CVParameterSelection*
Resultados reportados en [1]	95.40%	-
Resultados con `small120.arff`	95.14%	95.81% (C=1.5)
Resultados con base depurada	93.48%	93.93% (C=1.7)

Tabla 4.1 - Diferentes resultados de clasificación por patogenicidad con selección automática de parámetros.

Los resultados con la base `small120.arff` utilizando selección de parámetros superan los resultados reportados en [1]. Una aclaración importante sobre la diferencia entre los resultados logrados y los resultados reportados previamente reside en que existen diferencias en las bases de datos utilizadas. La base `small120.arff` tiene 714 muestras frente a las 660 de la base depurada.

4.3. Incorporación de taxonomía a la clasificación

Se estudió cómo incorporar la información de taxonomía a la clasificación por patogenicidad. Se propusieron dos enfoques, uno en el que se incorpora la información de taxonomía como nuevas características en la base de entrenamiento y otra, basada en las ideas en [2], en la cual para cada familia se entrena un clasificador independiente.

Para estas pruebas se utilizó la etiqueta real de taxonomía en la base aunque este conocimiento puede no ser conocido a priori. La idea planteada pretende predecir la familia

[2] https://code.google.com/p/bacfier/

de la bacteria para incorporar este conocimiento en la clasificación. Las pruebas son equivalentes a tener un clasificador de taxonomía perfecto con 100% de porcentaje de acierto.

4.3.1. Taxonomía como atributo

Se probaron tres formas de incorporar la taxonomía como atributo a la base:

- Columna de etiquetas de taxonomía (como atributos nominales)
- Una columna binaria por taxonomía de pertenencia/no pertenencia.
- Codificación jerárquica de la taxonomía, en forma de árbol como en la Figura 3.1.

En la Tabla 4.2 se muestran los resultados para cada una de estas variantes en donde en todos los casos se utilizó la base depurada con un clasificador SVM con núcleo lineal. Los resultados después de la selección de parámetros son iguales o muy similares entre sí. La codificación jerárquica es quien logra peores resultados, incluso para el caso en que no se incluye información de taxonomía.

C-SVM núcleo lineal	*Porcentaje de aciertos*	
	C=1.0	*CVParameterSelection*
Columna de taxonomía nominal	93.64%	93.94% (C=0.4)
Columnas binarias por taxonomía	93.18%	93.94% (C=1.3)
Codificación jerárquica	92.87%	93.64% (C=0.3)

Tabla 4.2 - Diferentes resultados de clasificación por patogenicidad con incorporación de taxonomía como atributo con y sin selección automática de parámetros.

Se puede concluir de estos resultados que el enfoque aquí planteado no aporta información relevante al problema e incluso puede llegar a deteriorar el comportamiento del clasificador. Otro clasificador o la fusión de clasificadores con este enfoque podría mejorar los resultados pero para SVM el enfoque no aporta.

4.3.2. Clasificadores por taxonomía o grupos taxonómicos

El enfoque planteado en esta parte consiste en entrenar un clasificador por taxonomía o grupo taxonómico. Cuando una nueva muestra se desea clasificar esta se clasifica en primer lugar por su taxonomía y luego se le aplica el clasificador correspondiente.

Se comenzó por entrenar un clasificador por taxonomía específica, es decir en el nivel T3 del árbol de la Figura 3.1. Los resultados de la clasificación se muestran en la Tabla 4.3 en donde se indica para cada familia el número de patógenos/no patógenos humanos, las instancias correctamente e incorrectamente clasificadas y el porcentaje de acierto. También se muestra un promedio de aciertos ponderado por la cantidad de muestras en cada familia. Para este nivel los resultados no mejoran los resultados de clasificación en promedio, la escasa cantidad de muestras en algunas familias pueden ser un problema en el entrenamiento, incluso en un esquema de validación cruzada de 10 particiones esto afecta aún más.

C-SVM núcleo lineal C=1.0	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de acierto*
	Humano	*No*		
Actinobacteria	35	43	70/8	89.74%
Alphaproteobacteria	28	60	82/6	93.18%
Bacteroidetes-Chlorobi	5	26	30/1	96.77%
Betaproteobacteria	23	28	47/4	92.16%
Epsilonproteobacteria	17	6	20/3	86.95%
Firmicutes	100	94	188/6	96.61%
Gammaproteobacteria	113	63	159/17	90.34%
Spirochaetes	15	4	18/1	94.74%
			Promedio ponderado de acierto	**92.94%**

Tabla 4.3 - Clasificadores de patogenicidad para cada taxonomía en el nivel T3.

Aprovechando los resultados de clasificación y selección de características en los niveles superiores del árbol de familias se realizó lo mismo para los niveles T2 y T1 del árbol. En las Tablas 4.4 y 4.5 se muestran los resultados. Se puede ver que en estos casos se mejora la clasificación de patogenicidad promedio incluyendo la información de taxonomía con parámetro de costo por defecto para todas las familias.

Para los clasificadores en estos niveles se realizó el procedimiento usual de selección del parámetro de costo obteniendo los resultados que se muestran en las Tablas 4.6 y 4.7. Se puede ver que el ajuste de este parámetro mejoran significativamente los resultados de clasificación.

C-SVM núcleo lineal C=1.0	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de aciertos*
	Humano	*No*		
ABG	164	151	297/18	94.28%
DE	17	6	20/3	86.95%
FA	135	137	258/14	94.85%
SBCCV	20	30	46/4	92.0%
			Promedio ponderado de acierto	**94.08%**

Tabla 4.4 - Clasificadores de patogenicidad para cada taxonomía en el nivel T2.

C-SVM linear kernel C=1.0	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de aciertos*
	Humano	*No*		
ABG-DE	181	157	312/26	92.31%
FA-SBCCV	155	167	308/14	95.65%
			Promedio ponderado de acierto	**93.93%**

Tabla 4.5 - Clasificadores de patogenicidad para cada taxonomía en el nivel T1.

C-SVM linear kernel CVParameterSelection	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de aciertos*
	Humano	*No*		
ABG (C=0.4)	164	151	299/16	94.92%
DE (C=0.6)	17	6	20/3	86.95%
FA (C=2.5)	135	137	260/12	95.58%
SBCCV (C=0.2)	20	30	46/4	92.0%
			Promedio ponderado de acierto	**94.69%**

Tabla 4.6 - Clasificadores de patogenicidad para cada taxonomía en el nivel T2 con ajuste del parámetro de costo.

C-SVM linear kernel CVParameterSelection	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de aciertos*
	Humano	*No*		
ABG-DE (C=0.3)	181	157	317/21	93.787%
FA-SBCCV (C=2.0)	155	167	311/11	96.58%
			Promedio ponderado de acierto	**95.15%**

Tabla 4.7 - Clasificadores de patogenicidad para cada taxonomía en el nivel T1 con ajuste del parámetro de costo.

Las mejoras en los resultados obtenidos para los clasificadores de patogenicidad en los niveles T1 y T2 con el ajuste del parámetro de costo superan a la clasificación sin información de taxonomía. En particular en T2 se observa que los subgrupos que peores resultados aportan corresponden con los que tienen menor número de muestras.

Se realizó una prueba adicional en este sentido definiendo un nivel de agrupamiento en el cual no se respeta la jerarquía del árbol y se agrupa DE con SBCCV. Se realizó el ajuste de parámetro para este nuevo grupo y los resultados obtenidos superan a la clasificación en T1 con ajuste. En la Tabla 4.8 se muestran los resultados para el nivel nombrado como T1.5. Este

agrupamiento definido por un criterio práctico aporta resultados que podrían modificar el foco del tipo de agrupamiento.

C-SVM linear kernel CVParameterSelection	*Patógeno*		*Correctamente/ Incorrectamente clasificados*	*Porcentaje de aciertos*
	Humano	*No*		
ABG (C=0.4)	164	151	299/16	94.92%
DE-SBCCV (C=1.0)	37	36	70/3	95.89%
FA (C=2.5)	135	137	260/12	95.58%
			Promedio ponderado de acierto	**95.30%**

Tabla 4.8 - Clasificadores de patogenicidad para cada taxonomía en el nivel T1.5 con ajuste del parámetro de costo.

La taxonomía demostró aportar información que mejora la clasificación pero se podría pensar en realizar algún tipo de *clustering* automático de los datos que no necesariamente tenga una interpretación biológica clara como lo es la taxonomía.

5. Conclusiones

5.1. Resultados

Se realizó una exploración de los datos de la base que permitió entender mejor un problema de alta dimensión y datos esparsos y principalmente binarios.

Se utilizaron técnicas de reducción de la dimensionalidad para permitir la visualización de los mismos y se encontró una fuerte relación entre los resultados de clasificación para taxonomía y los agrupamientos en el espacio reducido. Esto se verificó numéricamente analizando las distancias de los errores de clasificación por 1-NN a través de las distancias entre los vecinos más cercanos de la clase real y la clase más cercana, tanto en el espacio 814d y en el 3d.

Mediante el estudio "artesanal" de la naturaleza de las muestras, desde el punto de vista taxonómico, se lograron conclusiones y resultados que aportan a las técnicas clásicas de reconocmiento de patrones. Se encontró un grupo seleccionado de características que mejora la técnica de mayor desempeño sobre los datos. Se validó la utilidad de la información taxonómica para la clasificación por patogenicidad. Se demostró la redundancia en los datos desde el punto de vista taxonómico. Tener en cuenta que se trata de un problema con mayor

Quedá planteado un posible proceso generalizable para problemas con espacios muestrales similares. Datos esparsos, con mayor dimensionalidad y taxonomía con relación jerárquica.

Relativo a la clasificación por patogenicidad/no patogenicidad para humanos sin información de taxonomía, se logró mejorar los resultados reportados en [1] (94.5%) mediante el ajuste de parámetros del clasificador (95.81% con C=1.5) utilizando la misma base de entrenamiento. Se encontraron diferencias entre la base de entrenamiento que nos fue provista y la base utilizada por los autores por lo que no fue posible realizar una comparación directa de los resultados. Para nuestra base se alcanzó en este caso un porcentaje de acierto del 93.93%.

La incorporación de la información de taxonomía para clasificación de patógenos/no patógenos humanos demostró ser de utilidad en uno de los enfoques adoptados para el cual se entrena un clasificador por subgrupo taxonómico. Los mejores resultados se produjeron para la agrupación en el nivel T1 alcanzando un 95.15% en promedio.

Se generó un nivel artificial T1.5 basado en la observación de los resultados en T2 para el que la clasificación mejoró aún más a 95.30%.

5.2. Trabajo a futuro

Los buenos resultados en el agrupamiento artificial en T1.5 abren un nuevo camino para pensar en una preclasificación en un esquema de *clustering* automático en el entrenamiento no basado en taxonomía.

Sería interesante realizar la pre-clasificación de taxonomía en nuestra base y la base utilizada en [1] para realizar una validación final del procedimiento ya que la utilización de las etiquetas reales de taxonomía no forma parte del caso de uso real del método que se propone.

Validar si el proceso realizado para la selección jerárquica de características es generalizable y desarrollar la idea.

6. Bibliografía

[1] Reduced Set of Virulence Genes Allows High Accuracy Prediction of Bacterial Pathogenicity in Humans. Gregorio Iraola, Gustavo Vázquez, Lucía Spangenberg, Hugo Naya. PLoS ONE (2013).

[2] PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. Salvatore Cosentino, Mette Voldby Larsen, Frank Møller Aarestrup, Ole Lund. PLoS ONE (2013).

[3] Dimensionality Reduction: A Comparative Review. L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Tilburg University Technical Report, TiCC-TR 2009-005, (2009).

[4] Visualizing High-Dimensional Data Using t-SNE. L.J.P. van der Maaten and G.E. Hinton.. Journal of Machine Learning Research 9(Nov): 2579-2605 (2008).