Capítulo 1

Introducción al problema y objetivos

En este capítulo describimos los problemas que esta tesis aborda. El primero, la optimización de la genotipificación de ADN, y en específico la genotipificación de muestras del virus del papiloma humano, es el problema del cual conocemos la respuesta óptima y nos servirá como guía de desempeño. Los problemas como éste pueden traducirse al segundo tópico de este capítulo, la selección de características, que es el problema general que también estudiamos en este documento.

1.1. Genotipificación de ADN, el problema de la selección de compuestos

Genotipificar o tipificar Ácido Desoxiribonucléico (ADN) es el acto de identificar el genotipo de una muestra de éste. Existen diferentes formas de hacerlo, entre ellas está el método conocido como Polimorfismo en la Longitud de los Fragmentos de Restricción o RFLP por sus siglas en ingles. La mayor parte de los datos con los que se trabajará en esta tesis fueron obtenidos a través de RFLP.

El método de RFLP utiliza proteínas llamados *enzimas de restricción* para iden-

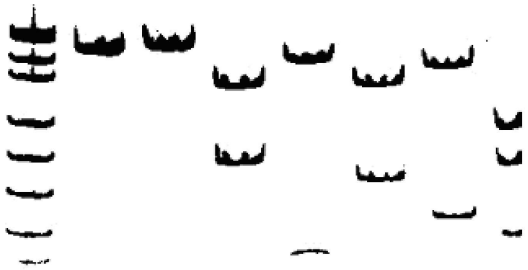

Figura 1.1: Patrones de restricción para diferentes tipos de VPH, generados con la enzima de restricción DdeI [18]

tificar el ADN. Éstas reconocen una subcadena de ADN y cortan la cadena de ADN original en los puntos donde ocurra dicha subcadena. Se obtienen fragmentos de diferentes tamaños y son el número de fragmentos y el tamaño de éstos lo que nos permite identificar una muestra [18].

Para obtener el número y el tamaño de los fragmentos se utiliza una técnica llamada *electroforesis*. Dicha técnica genera patrones de restricción como el mostrado en la figura 1.1, donde cada columna representa un patrón de restricción. Éstos se generan del desplazamiento de los fragmentos de ADN en un gel después de aplicarles una diferencia de potencial. El desplazamiento está en función del tamaño de los fragmentos, por lo que podemos basarnos en esta medición y en la cantidad de fragmentos para identificar parcialmente la muestra.

Disponemos de esta información en la forma de una matriz como se muestra en la tabla 1.1, donde cada renglón corresponde a la salida de una enzima para cierto virus. La salida corresponde a la información sobre el número de fragmentos generados y la distancia recorrida por cada uno de ellos durante la electroforesis.

Ésta contiene la información de la distancia recorrida por cada fragmento generado por cada enzima para cada virus.

Debido a que los patrones de restricción se forman con base solamente en el tamaño

Tabla 1.1: Matriz de distancias

Virus	Enzima	No. Fragmentos	D1	D2	D3	D4
1	1	1	31.1	0	0	0
1	2	1	31.7	4.5	0	0
1	3	1	32.7	0	0	0
1	4	1	30.1	0	0	0
1	5	1	21.1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	205	1	31.2	24.2	20.5	0

de las moléculas, es posible tener patrones de restricción iguales para virus diferentes utilizando la misma enzima. De esta manera tenemos que si dos enzimas A y B generan el mismo patrón para cierto virus, ubicándolo en un genotipo *G1* o *G2*, es necesario utilizar una enzima C que sabemos que genera un patrón diferente para los genotipos *G1* y *G2* para poder identificar correctamente la muestra.

Existen más de 200 enzimas disponibles para la tipificación. En el caso del Virus del Papiloma Humano (HPV) existen más de 50 variantes del virus. Escoger manualmente, como se hace actualmente, un grupo pequeño de enzimas con el que sea posible identificar con certeza todos los tipos virales del HPV resulta muy difícil. Los grupos resultantes de hacer esta selección manualmente tienden a ser más grandes de lo necesarios y tener mucha redundancia, desperdiciando tiempo y dinero.

Es posible utilizar diferentes enfoques para obtener grupos más pequeños que los seleccionados anteriormente. La presente investigación se centra en el estudio de este problema como un problema de selección de características. La figura 1.2 muestra el punto donde entra esta investigación dentro del la tipificación de ADN.

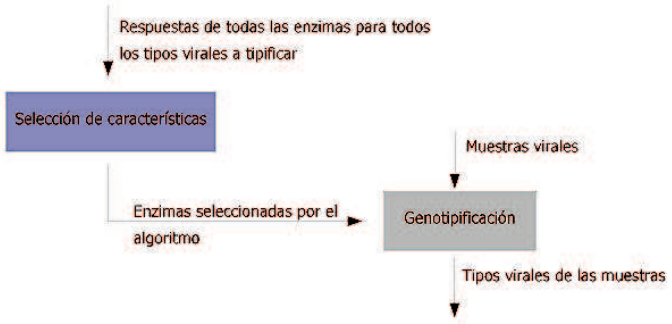

Figura 1.2: Esquema del proceso de optimización de la selección de enzimas

1.2. Selección de características

El problema de selección de características ha sido atacado fuertemente en los últimos años y está íntimamente ligado con el reconocimiento de patrones. Tsamardinos [31] señala que no existe una definición totalmente aceptada del problema pero propone una que creemos es muy adecuada para los propósitos de esta investigación:

Definición 1. [31] Un problema de selección de características o variables es una tupla $\langle X, \Phi, T, A, M \rangle$, donde X es una muestra de patrones de entrada definidos sobre un conjunto de características Φ, $T \in \Phi$ es una variable objetivo, A es un algoritmo de clasificación que produce un modelo de predicción para T dado T y X; y M es una métrica de desempeño del modelo del clasificador y de las características seleccionadas. Una solución al problema es un subconjunto de características $\phi \subseteq \Phi$ que maximiza $M(\phi, A(T, X \downarrow \phi))$, donde $X \downarrow \phi$ es la proyección de los datos X sobre las características pertenecientes a ϕ. En esta métrica M generalmente deseamos minimizar $|\phi|$ y maximizar el desempeño de T para el problema específico.

En palabras más simples, selección de características es el problema de encontrar, dado un conjunto de características, un subconjunto de estas tal que este subconjunto maximice cierta(s) propiedad(es) deseadas. Entre estas propiedades comúnmente se

encuentra que el nuevo conjunto permita clasificar correctamente todas las posibles clases a las que puede pertenecer una muestra del conjunto original.

1.2.1. Mapeo del problema de la selección de enzimas como un problema de selección de características

Hablando en los términos de la definición 1, el problema de optimización de la selección de enzimas como un problema de selección de características se traduce en encontrar un $\phi \subseteq \Phi$ que maximice M; donde M es la evaluación de 2 factores: el porcentaje de clasificación correcta de los virus a identificar y la cantidad de enzimas seleccionadas, que debe ser mínima. La puntuación obtenida por un subconjunto ϕ solo será positiva si el algoritmo de clasificación A es capaz de clasificar correctamente muestras para todos los tipos virales con los que contamos. Un subconjunto ϕ_1 tendrá mejor puntuación que un subconjunto ϕ_2 si $|\phi_1| > |\phi_2|$.

El espacio de características original Φ estará definido por la concatenación de la información de cada enzima. Recordemos que la salida de cada enzima esta definida por 4 valores. Sin embargo, el espacio de características esta definido de la siguiente manera:

$$\Phi = \{\{R_1\}, \{R_2\}, \ldots, \{R_k\}\} \tag{1.1}$$

donde R_k es el espacio reservado R para la información generada por la enzima k. Así pues necesitamos *comprimir* la información de los fragmentos generados por una cada enzima para que ésta pueda ser relacionada con un solo valor. Las técnicas para compresión utilizadas en esta investigación se detallan en la sección 2.4.1.

Continuado con el mapeo del problema a la definición 1, T es la relación uno a uno que existe entre las muestras contenidas en X y los genotipos que se desean clasificar.

1.3. Objetivos generales

Analizamos el problema de la optimización de la genotipificación del HPV como un problema de selección de características. Proponemos dos técnicas para atacar el problema y comparamos resultados. Introducimos ajustes al proceso estándar de los algoritmos genéticos para mejorar su desempeño en el caso específico de la optimización de la genotipificación de HPV. Exploramos de manera práctica los diferentes aspectos de un problema de selección de características.

1.4. Objetivos específicos

1. Mapear el problema de la optimización de la genotipificación de HPV a un problema de selección de características.

2. Proponer y aplicar una técnica basada en el análisis de componentes principales y redes neuronales artificiales para la selección de características.

3. Diseñar una algoritmo genético basado en el algoritmo canónico para la selección de características.

4. Implementar los puntos anteriormente mencionados y comparar su desempeño.

1.5. Descripción del documento

El presente documento está organizado de la siguiente manera: el capítulo 1 introduce al lector al problema y aporta las primeras definiciones necesarias para entender el documento. El capítulo 2 aborda el problema aplicando PCA como una guia para la selección apoyado con RNA's; se presenta un marco teórico para entender el procedimiento, se explica y se presentan los resultados de aplicarlo en nuestro caso de prueba,

el HPV. El capítulo 3 mantiene la estructura del capítulo 2 pero aplicado a algoritmos genéticos. El capítulo 4 presenta conclusiones generales y comparativas de los dos métodos aplicados, así como cambios que se observaron necesarios en ambos algoritmos para la mejora de éstos.