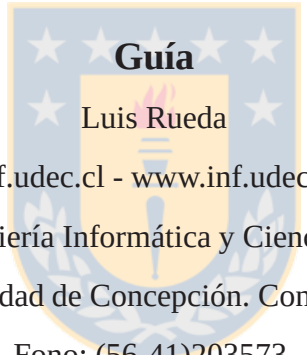
# Estudio de características energéticas en zonas de interacción proteína-proteína, para identificación de interacciones transitorias y permanentes

(Study of the energetic characteristics in protein-protein interaction interfaces to identifying transient and obligate protein-protein complexes)

Tatiana Gutiérrez Bunster
tgutierrez@udec.cl



**Guía**

Luis Rueda

lrueda@inf.udec.cl - www.inf.udec.cl/~lrueda

Departamento de Ingeniería Informática y Ciencias de la Computación.

Universidad de Concepción. Concepción.

Fono: (56-41)203573.

**Co-Guía**

José Martínez-Oyanedel

jmartine@udec.cl

Departamento de Bioquímica y Biología Molecular. Facultad de Ciencias Biológicas.

Universidad de Concepción. Concepción.

Fono: (56-41)203812 - (56-41)203822.

16 de junio de 2008

# Índice general

# Índice de figuras

# Indice de Tablas

## Resumen

La mayoría de los procesos celulares que sostienen la vida involucran interacciones entre moléculas, tales como interacciones ADN-proteína, interacciones entre proteínas, o entre proteínas y moléculas pequeñas. Por esta razón han sido extensamente estudiadas para intentar descubrir, como se producen estas interacciones, además se han propuesto métodos para lograr predecirlas. Esto se traduce en que la biología molecular ha producido gran cantidad de información funcional y estructural sobre estos complejos. Numerosas estructuras de complejos proteicos se encuentran disponibles en bancos de datos, sin embargo aún no es posible predecir exitosamente con un método establecido, las posibles zonas de interacción de una proteína con otra, así mismo no es posible predecir la estabilidad de la interacción.

La Informática constituye una herramienta poderosa para el estudio de la información funcional y estructural, además de las características que surgen del análisis de complejos de interacción Proteína-Proteína. Uno de los problemas cruciales para entender y clasificar interacciones de proteínas es caracterizar y discriminar la superficie de la interacción. Por esto se propone que existen algunas características de las superficies de interacción que permiten discriminar entre complejos proteína-proteína permanentes y complejos proteína-proteína transitorios.

Para esto, en el presente proyecto se estudian las características energéticas de interfaces de interacción proteína proteína - área de interacción - de estructura tridimensional conocida, clasificadas como interacciones transitorias y permanentes, que caracterizan el comportamiento de un complejo de proteína y sus interacciones.

Para complementar estudios anteriores, es utilizado un algoritmo de selección de características -forward search- en conjunto con la distancia de Chernoff en una base de datos de 296 complejos permanentes y transitorios. Se utilizó el programa FastContact para obtener la contribución energética de cada complejo en su área de interacción, obteniéndose 642 características por complejo.

Para estudiar la precisión en la clasificación, se utilizaron métodos de reducción lineal de dimensiones -Heteroscedástico HDA, Homoscedástico FDA, Chernoff CDA- combinados con un clasificador cuadrático y lineal, además de utilizar máquina de soporte de vectores.Lo que permitió generar un ranking de las características más influyentes para encontrar un factor

discriminante que distinga la potencialidad de la superficie de interacción de pertenecer a una interface de interacción -mayor separabilidad entre las clases- o tipos de interacción. Los mejores resultados estuvieron cercanos al 81 % de precisión. Los resultados obtenidos con el clasificador más preciso se pueden aplicar en el futuro a otros complejos no-clasificados como herramienta predictiva. La selección realizada utilizando en conjunto el algoritmo foward search y la distancia de Chernoff alcanzó una alta precisión. El análisis de las mejores características discriminantes muestra que las energías de desolvatación contribuyen por sobre la energía electrostática a la separación de clases.

# Capítulo 1

# Introducción

## 1.1. Planteamiento

La inteligencia artificial (IA) se presenta como una ciencia que desarrolla software y hardware que permita simular el comportamiento humano [90]. Muchas ciencias han participado para el desarrollo de la IA, una de ellas es la matemática, que entrega la teoría formal relacionada con la lógica, la probabilidad, la teoría de las decisiones y la computación. Otra ciencia es la psicología que proporciona el lenguaje científico para expresar las teorías.

El reconocimiento de patrones es una área de la IA, conocida como aprendizaje automático (Machine Learning) . El propósito de esta área es el clasificar un grupo de patrones conocido como conjunto de pruebas en dos o más clases de categorías. Esto se logra al comparar el conjunto de prueba con el conjunto de entrenamiento (previo) o training set. Por ejemplo un clasificador como K-NN (el vecino más cercano) mide la distancia entre varios puntos dados (compara), para saber que puntos son mas cercanos a la meta en un modelo parametrizado [31]. Para lograr estos objetivos, el reconocimiento de patrones involucra una gran variedad de subdisciplinas, como el análisis de discriminantes, la extracción de características, la estimación del error, el análisis de regiones, la inferencia gramatical, entre otros [90].

Las áreas de aplicación de reconocimiento de patrones son muy variadas en el procesamiento de la información, por ejemplo, se puede mencionar el área biomédica, en la cual se puede indicar el estudio y clasificación de cromosomas, reconocimientos de genes en secuencias, además de las áreas agrícolas, industriales, astronómicas, en las cuales se realizan procesamiento de imágenes, reconocimiento de la voz, escritura [31], señales sísmicas, radar, diagnóstico de enfermedades, fallos en maquinarias y procesos industriales entre otros.

Debido a la necesidad de contar con sistemas informáticos, se ha mantenido un fuerte desarrollo en un conjunto de métodos aplicados para el logro de procesos de aprendizaje automático. Actualmente estos métodos se han consolidado como una línea establecida del estudio en reconocimiento de patrones [31].

Si se enfoca la aplicación del reconocimiento de patrones en el área de la Bioinformática, se puede observar que la mayoría de los procesos celulares que sostienen la vida involucran interacciones entre moléculas, tales como interacciones ADN-proteína, interacciones entre proteínas, o entre proteínas y moléculas pequeñas. Por esta razón son extensamente estudiadas para intentar descubrir, como se producen estas interacciones, además se han propuesto métodos para lograr predecirlas. Esto se traduce en que la biología molecular ha producido gran cantidad de información funcional y estructural sobre estos complejos [9, 10]. Numerosas estructuras de complejos proteicos se encuentran disponibles en bancos de datos, sin embargo aún no es posible predecir exitosamente con un método establecido, las posibles zonas de interacción de una proteína con otra, así mismo no es posible predecir la estabilidad de la interacción.

La informática constituye una herramienta poderosa para el estudio de la información funcional y estructural, además de las características que surgen del análisis de complejos de interacción Proteína-Proteína (IPP). Uno de los problemas cruciales para entender y clasificar interacciones de proteínas es caracterizar y discriminar la superficie de la interacción. Por esto se propone que existen algunas características de las superficies de interacción que permiten discriminar entre complejos proteína-proteína permanentes y complejos proteína-proteína transitorios.

Para esto, en el presente proyecto, se intenta estudiar las características energéticas de la interfaces de interacción proteína proteína - área de interacción - de complejos de estructura tridimensional conocida, clasificados como interacciones transitorias y permanentes, a través del área de reconocimiento de patrones, que permite que se caracterice el comportamiento de un complejo de proteína y sus interacciones.

El propósito de clasificar estas características es encontrar un factor discriminante que permita distinguir la potencialidad de la superficie de pertenecer a una interface de interacción, o tipos de interacción. Como resultado se espera identificar las características discriminantes en una proteína que caracterice los tipos de complejos de interacción proteína-proteína transitorias y permanentes.

## 1.2. Definición del Problema

Existen variadas investigaciones sobre características relevantes para las interacciones proteína-proteína (IPP), alguna de ellas son los parches hidrofóbicos (áreas de interacción ricas en residuos hidrofóbicos con rechazo al agua), las moléculas de agua presentes en la interfaz, la presencia de cavidades en la superficie. Esto muestra que existen muchos factores que participan de una interacción, pero a su vez, no se ha identificado cual de ellos es de mayor relevancia para una interacción.

Los estudios sobre IPP, se han focalizado en el área de unión entre ambas proteínas, es decir, en la superficie de interacción. La cual también ha sido ampliamente estudiada, pero dentro de estos estudios, no existen investigaciones específicas de las características energéticas como factores relevantes en la diferenciación entre tipos de interacciones.

Las características energéticas de la superficie de interacción, han sido muy poco utilizadas, para caracterizar la superficie de interacción como criterio de análisis en la clasificación de complejos proteína-proteína transitorias y permanentes.

Una interacción depende mayormente de los cambios energéticos que se producen en y entre las proteínas. Esto tiene relación directa con la problemática de este proyecto, debido a que las características energéticas son las que determinan la estabilidad de dichas interacciones. Si una interacción se mantiene o no en el tiempo, se traduce directamente en una IPP permanente o transitoria.

Por esta razón, al saber que las características energéticas son un tópico relevante en el estudios de IPP, surge el interés por trabajar con ellas para el presente proyecto, enfocándose en las IPP transitorias y permanentes.

## 1.3. Hipótesis

La hipótesis en la que se sustenta este proyecto es que existen algunas características energéticas de las superficies de interacción que permiten discriminar entre complejos proteína-proteína permanentes y complejos proteína-proteína transitorios.

## 1.4. Objetivos

Este trabajo de investigación tiene como objetivo identificar características energéticas que discriminen complejos IPP transitorios de complejos IPP permanentes, enfocándose, solo en la zona de interacción. Para esto, se desarrolla la metodología de trabajo que mediante la utilización de métodos de discriminación, se proporcionan alternativas de solución al problema, utilizando el reconocimiento de datos.

Para esto se deben profundizar diferentes métodos de selección de característica, criterios de evaluación y clasificadores. Además de estudiar los diferentes aspectos que involucran el tema de Interacción proteína-proteína, todo esto descrito en los capítulos 2 y 3 respectivamente.

También se debe crear una base de datos de complejos proteicos ya clasificados en transitorios y permanentes.

El paso a seguir es lograr seleccionar las características adecuadas que permitan una buena clasificación. Se buscan métodos de selección de características que se puedan aplicar a diferentes bases de datos, en el marco del aprendizaje supervisado, específicamente para la tarea de clasificación. El proceso de selección de características se entiende como el recorrido de un espacio hasta encontrar una combinación de características que optimice alguna función definida sobre un conjunto de características. En general, un algoritmo de selección consta de dos componentes básicos: criterio de selección y método de búsqueda. En este trabajo se propone trabajar con el método de búsqueda secuencial hacia adelante y distancia de Chernoff como criterio de selección. Esto permite ordenar las características por orden de importancia en la determinación de la clase.

Luego, para el estudio de la precisión en la clasificación, se utilizan métodos como la máquina de vectores soporte y reducción lineal de dimensiones -Heteroscedástico HDA, Homoscedástico FDA, Chernoff CDA- combinados con un clasificador cuadrático y lineal. Además paralelamente se utiliza . Los métodos de selección y clasificación se describen en el capítulo 4.

Para comprobar la bondad de la selección y clasificación propuesta, se presentan en el capítulo 5 los resultados obtenidos, realizando comparaciones entre los métodos de selección y clasificación, todo esto sobre el mismo conjunto de datos. Además se debe interpretar desde un punto de vista biológico los resultados obtenidos.

## 1.5. Organización

El presente documento de Tesis de Magíster consta de 6 Capítulos, cada uno de ellos integrado por un conjunto de secciones. Además, se incluyen diversos apéndices con información complementaria.

**Capítulo 1.** Se realiza una introducción al área de reconocimiento de patrones, además se presenta como surgió el tema a trabajar.

**Capítulo 2.** Se entregan los principios básicos de reconocimiento de patrones, los criterios de evaluación más utilizados y las etapas que la componen.

**Capítulo 3.** Introducción al área de la Bioinformática, también se presentan las principales estrategias propuestas en la literatura para resolver diferentes problemáticas del área de la Biología Molecular, específicamente en aquellas áreas de interés para el presente proyecto y se finaliza con las propuestas para entregar un aporte a lo ya desarrollado. Para esto se presenta una hipótesis y la definición de los objetivos a cumplir para comprobar la hipótesis.

**Capítulo 4.** En este capítulo, se detalla la metodología a realizar, el detalle de los datos utilizados y los métodos para su validación.

**Capítulo 5.** Detalle de los resultados obtenidos, con los parámetros que fueron utilizados, además de la validación de la metodología propuesta, y la posterior interpretación de los resultados.

**Conclusiones.** Finalmente, se presentan las conclusiones obtenidas durante esta investigación, con base en la evidencia descrita en el Capítulo 5, para finalizar con el cumplimiento de la hipótesis y objetivos, lo que lleva a algunas alternativas para obtener mayor rendimiento en propuestas futuras.

# Capítulo 2

# Reconocimiento de patrones

El reconocimiento de patrones tiene variadas áreas de aplicación y algunas de ellas fueron mencionadas en el capítulo anterior. El presente informe se centra en el área de la Biología, en la cual se han realizado amplios estudios en el reconocimiento de patrones biológicos de proteínas y genes, además de predicciones de estructuras de proteínas e interacciones de proteínas.

El estudio se focaliza en las proteínas y específicamente en interacciones entre ellas, para esto se deben clasificar patrones con base en un conocimiento a priori o información estadística extraída de los datos. Los patrones a clasificar suelen ser grupos de medidas u observaciones, definiendo puntos en un espacio. La diferencia entre procesos de reconocimiento de patrones radica en el método de extracción, salida de los datos u obtención de las características del elemento a reconocer (de acuerdo al medio y las técnicas apropiadas) .

## 2.1. Enfoques de Reconocimiento de patrones

En el reconocimiento de patrones se tienen diferentes lineas de trabajo, de acuerdo al enfoque que posea el estudio a desarrollar. A continuación se presentan tres enfoques de los cuales uno se utilizará en la problemática del proyecto. Están los que se basan en la teoría de probabilidad y estadística, los que utilizan funciones discriminantes y los que utilizan algoritmos de búsqueda o métodos de optimización basados en heurísticas.

### 2.1.1. Reconocimiento estadístico de patrones

El reconocimiento estadístico es el procedimiento de clasificar los objetos en clases y con el conocimiento a priori adecuado (conocer las clases). El reconocimiento estadístico se subdivide en dos enfoques de acuerdo a su distribución: reconocimiento paramétrico y no paramétrico.

**Reconocimiento paramétrico,** parte con el supuesto de una distribución, ya sea normal, exponencial, etc., y luego se debe utilizar la muestra de entrenamiento para estimar los parámetros de esa distribución. Se busca el valor del parámetro más adaptado a los datos muestrales. La muestra de entrenamiento que se tiene de las situaciones supervisadas, es un conjunto de patrones perfectamente identificados a priori y que representan a todas las clases de interés.

**Reconocimiento no paramétrico,** parte con el supuesto de una distribución libre, que reciben este nombre porque su diseño no depende de ningún tipo de supuesto sobre el modelo probabilístico. Se utilizan funciones discriminantes para establecer regiones de estudio donde se encuentran las clases y se determina si un nuevo patrón pertenece a una clase, solo haciendo uso de la información entregada por el conjunto de patrones de entrenamiento.

### 2.1.2. Reconocimiento sintáctico de patrones

Este enfoque busca encontrar aquellas relaciones estructurales que guardan los objetos de estudio con respecto a otros objetos y como estos pueden ser capaces de describirlos. Hay aplicaciones con este tipo de reconocimiento usadas en la biología molecular para el análisis de secuencias de proteínas, evaluar eficiencias y desarrollar nuevas metodologías de investigación. El objetivo es poder construir una gramática que clasifique la estructura desde el universo de objetos.

### 2.1.3. Reconocimiento lógico combinatorio de patrones

Una de sus tareas esenciales es tratar cuidadosamente las características que describen a los objetos en estudio. Esto se realiza a través de la ayuda de formalismos matemáticos (deducción matemática) que permiten derivar nuevos conocimientos a partir de los existentes.

## 2.2. Etapas

Un sistema de reconocimiento de patrones se compone de varios pasos que van desde un sensor que recoge las observaciones a clasificar (u otro método de recolección), hasta un sistema de clasificación, basado en las características extraídas. No siempre se utilizan los mismos pasos debido a que no están claramente separados. Para el proyecto se seguirán los pasos, que se visualizan en la figura 2.1[31].

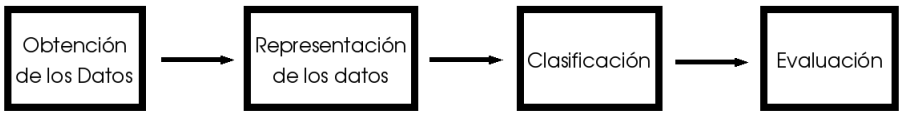



Figura 2.1: Etapas del proceso de reconocimiento de patrones.

### 2.2.1. Obtención de los datos

Esta etapa años atrás era la más difícil de procesar, debido a que era realizada de manera manual. En la actualidad, el proceso se efectúa de manera digital, lo que implica una mejor automatización en extracción de información. Alguna de las herramientas que hacen posible esta extracción son: la cámara, el micrófono, termómetros, entre otros.

### 2.2.2. Representación de los datos

La finalidad de esta etapa es encontrar aquellas características que representan de mejor manera a cada tipo de objeto. Estas características deben ser entregadas de una manera clara para ser utilizadas por el computador.

#### 2.2.2.1. Extracción de características

De acuerdo a las definiciones obtenidas desde Theodoridis [90], los patrones u objetos son estructuras que son descritas por características del mundo real (también llamadas atributos o

variables), que les permiten a los objetos ser ubicados en una clase específica. Se representan mediante un vector de longitud fija $x$ de $d$-dimensiones (número de variables requeridas) y se denota en la Ecuación 2.1

$$x = [x_1, x_2, ..., x_n]^t \tag{2.1}$$

donde $t$ es el vector transpuesto y $x_i$ son las características ($i = 1, 2, ..., n$), es decir, son variables específicas que identifican a un único objeto, las cuales pueden ser de tipo numéricas (discretos o continuos) o simbólicas.

Si se conocen las categorías, la clasificación de $c$ clases, $\omega_1, \omega_2, ..., \omega_c$ asociadas a cada objeto $x$, se incluye en el conjunto final descrito por una matriz de $m$ objetos por $n$ características, y la columna adicional que indica la clase a la que pertenecen. La matriz final es de orden $m \times n$ representada en la Ecuación 2.2.

$$M = \begin{bmatrix} \omega_1 & x_{1,1} & x_{1,2} & ... & x_{1,n} \\ \omega_2 & x_{2,1} & x_{2,2} & ... & x_{2,n} \\ ... & ... & ... & ... & ... \\ \omega_c & x_{m,1} & x_{m,2} & ... & x_{m,n} \end{bmatrix} \tag{2.2}$$

#### 2.2.2.2. Selección de características.

La selección de características busca encontrar las características que sean más especificas y representativas de cada clase, es decir, discriminantes entre clases, donde se transforma la información observada en valores numéricos o simbólicos. Las técnicas de selección de características han sido aplicadas a la predicción de estructuras de proteínas [62]. Pero el tener un alto número de características, como es el caso de las proteínas, no es un indicador de que todas sean útiles para efectuar clasificaciones. Cuando se trabaja con una selección de ellas (generalmente un menor número), entregan un mayor criterio de relevancia y representan aquellas funciones y estructuras de importancia en una interacción entre proteínas. Los

selectores utilizados para minimizar el error de clasificación de los clasificadores particulares utilizan características superiores que disminuyen los costos computacionales [90].

Los pasos a considerar en una selección de características [47] son:

- Buscar un subconjunto mínimo de características que no introduzca confusión entre clases. Debido a que las proteínas están formadas por aminoácidos, las características de aquellos que estén en la superficie así como su posición relativa, y la geometría de la superficie (forma) definirán finalmente algunas propiedades que caracterizan su modo de acción o las capacidades de interacción con otras proteínas o moléculas.
- Seleccionar un subconjunto de características que permita discriminar aquellos objetos que pertenecen a una u otra clase. Para esto se identifica cual de estas tiene mayor relevancia (mayor peso), lo que influye en la clasificación de un objeto en una clase específica [75].

#### 2.2.2.3. Normalización

Al tener un conjunto de características seleccionadas, pueden presentarse casos en que los rangos de valores de las características sean muy variados. Por esta razón se aplican normalizaciones, en las cuales los datos se transforman para mantener estabilidad en los cálculos realizados. Es decir, consiste en la transformación de un conjunto de datos ($x_n$) a otro ($\hat{x_n}$), con media cero y desviación estándar uno.

### 2.2.3. Métodos de Clasificación

La clasificación se puede interpretar como la división de un espacio compuesto por características en regiones mutuamente excluyentes. De esta manera cada región esta asociada una clase $\omega$ ($\omega_1, \omega_2, ..., \omega_c$) de acuerdo a un patrón particular y las características se distribuyen entre las $c$ clases disponibles.

Una vez que se selecciona un subconjunto de características relevantes utilizando algún criterio en particular, la calidad de las características representativas se debe evaluar mediante un método de clasificación y una posterior técnica de evaluación de este último. El objetivo de la clasificación es utilizar el vector $x$ con las características disponibles para asignar al objeto la entrada a una clase. En muchos casos, este paso involucra determinar la probabilidad de cada

una de las clases. Para una posterior clasificación se utiliza el método RLD (reducción lineal de dimensiones), que permite encontrar la transformación lineal que reduzca la dimensión de un espectro de $n-$dimensiones a uno de $d-$dimensiones ($d < n$), preservando la maximalidad (mantener su calidad, es decir, que sea notoria la diferencia entre las clases) de la información discriminatoria para varias clases dentro de un modelo.

Por otro lado, la kernelización de métodos de clasificación ha permitido proyectar o transformar los datos a dimensiones mayores para la derivación de un clasificador simple y eficiente. Por ejemplo, el lineal. Las criterios más conocidas que han sido kernelizados son las Máquinas de vectores soporte, el criterio de Fisher [64]y el de Bayes. Los kernels más usados y que han mostrado eficiencia de clasificación en datos empíricos son los polinomiales, el sigmoideo y los radiales es [31, 90]. El uso de kernels permite trabajar en casos complejos, además de permitir trabajar no solamente con distribuciones lineales.

#### 2.2.3.1. Aprendizaje

El aprendizaje se refiere al algoritmo que permitirá reducir la cantidad de errores en la información para el posterior entrenamiento que se puede realizar de dos formas:

**Aprendizaje Supervisado.** La clasificación supervisada, consiste en clasificar nuevos objetos basándose en la información de un conjunto ya clasificado (conjunto de entrenamiento, del cual ya se conoce la clase de los datos a priori) que se usa para entrenar al sistema;

**Aprendizaje no Supervisado.** La clasificación no supervisada, consiste en dado un conjunto no clasificado (no se tiene un conjunto para aprender a clasificar la información a priori) encontrar la clasificación de la misma (identificar las clases) a través de cálculos estadísticos o no estadísticos.

#### 2.2.3.2. Métodos de Clasificación

El mecanismo para decidir que método de clasificación utilizar, se basa en el diseño de reglas y algoritmos de clasificación de los cuales existen varios tipos: usando distribuciones probabilísticas, de contenido de información espacial y algebraica. Se han propuesto diversos algoritmos siendo éstos una combinación de las características antes mencionadas con diferentes métodos de selección y clasificadores. A continuación se presentan algunos de los métodos utilizados para la predicción de interacciones:

- **Naïve Bayes (NB)**

  Naïve Bayes es una técnica de clasificación descriptiva y predictiva basada en la teoría de la probabilidad del análisis del Teorema de Bayes [89]. Esta teoría supone un tamaño de la muestra asintóticamente infinito y una independencia estadística entre variables independientes, refiriéndose en este caso a las características, no a la clase. Además se trabaja con un conjunto finito de clases y se pueden calcular las distribuciones de probabilidad de cada clase para establecer la relación entre las características (variables independientes) y la clase (variable dependiente).
  En ocasiones también se puede asumir que los conjuntos tienen distribuciones a priori $\pi_1, \pi_2, ..., \pi_g$. En tal caso se define la regla discriminante de Bayes, como aquella que asigna una observación $g$ a la población $\pi_j$, si $\pi_j f_j(g)$, es máximo ($j = 1, 2, ..., g$). Esta expresión es la verosimilitud a posteriori de $\pi_j$. Existen algunas formas estándar de asignar probabilidades a priori: una es de manera uniforme, que equivale a darles el mismo valor $\pi_j = \frac{1}{g}$ a cada una; otra forma es dar valores proporcionales a los tamaños de las submuestras de cada población presentes en la matriz de datos, $\pi_j = \frac{n_j}{n}$. Esto es válido si se asume que la muestra de entrenamiento se formó seleccionando $h$ individuos al azar. De éstas, $h_j$ resultaron de la población $a_{ij}$, $(j = 1, 2, ..., g)$. No es válido si inicialmente se decidieron los tamaños $h_j$ de las diferentes poblaciones [31]. Con el clasificador Bayesiano puede demostrarse teóricamente que minimiza el error de manera óptima. Sin embargo, la suposición de independencia estadística de las variables es una limitación.

- **Regla de los vecinos más cercano (k-NN)**

  Corresponde a un método de aprendizaje basado en ejemplos, que utiliza una función de distancia para determinar que elementos del conjunto de datos de entrenamiento esta más cerca de las clases identificadas [90, 31]. La idea básica considera la utilización de un conjunto de datos de entrenamiento que constituye todo el conocimiento a priori del sistema. El supuesto es considerar a los elementos cercanos, como aquellos que tienen la mayor probabilidad de pertenecer a la misma clase. Por esto, cuando se desea clasificar un nuevo caso $y$, se debe obtener la distancia entre $x$ y los casos contenidos en el conjunto de entrenamiento, se asigna a la clase correspondiente de acuerdo al elemento que obtuvo la menor distancia (es decir, el más cercano a $y$). Para cada asignación a una clase, se requiere calcular las distancias existentes entre el elemento a clasificar y el total de los elementos de entrenamiento para encontrar los $k$ vecinos.

- **Análisis de Discriminantes lineales (LDA)**

  Esta técnica se busca una proyección de los datos (direcciones que sean eficientes para la discriminación) en un espacio de menor dimensión que los datos de entrada, con el fin de que la separabilidad de las clases sea la mayor posible [31]. El objetivo es la reducción de dimensionalidad a la vez que se mantenga la máxima discriminación posible (que resalte la discriminación entre las clases). Se dice supervisada, ya que para poder buscar la proyección se debe entrenar el sistema con patrones ya identificados en las diferentes clases. A diferencia de PCA, LDA no busca minimizar el error de representación obtenido, que tiene por objetivo, la reducción de dimensión que preserva la variabilidad en el espacio origen al máximo. Es decir, se debe representar de una manera precisa las muestras del espacio de mayor dimensión y conservar la representatividad. Esto no indica que haya garantía de que las direcciones de máxima variación contengan buena capacidad de discriminación.

- **Redes neuronales**

  El objetivo es simular las propiedades observadas en los sistemas neuronales naturales a través de modelos matemáticos recreados mediante mecanismos artificiales [31]. Esto entrega muchas ventajas sobre los sistemas convencionales, algunas de ellas son: Aprendizaje adaptativo, auto-organización, tolerancia a fallos, operación en tiempo real y fácil inserción dentro de la tecnología existente.

- **Máquinas de vectores soporte**

  Este tipo de técnica es una extensión de los modelos lineales que se aplican a problemas con dos clases, es decir, el espacio inicial se transforma en un nuevo espacio con nuevos atributos obtenidos por una combinación no lineal de los atributos originales. El modelo lineal obtenido en el nuevo espacio representa un límite de decisión no lineal en el espacio [87, 64, 26]. Se busca construir un hiperplano como frontera de decisión entre las clases de tal forma que se maximize el margen de separación entre los patrones de las demás clases. El margen es la distancia perpendicular entre el hiperplano separador y el hiperplano que pasa sobre los puntos más cercanos (los vectores de soporte).

- **Árboles de decisión**

  Un árbol de decisión es un modelo de predicción utilizado para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de

un problema. Su estructura permite seleccionar una y otra vez diferentes opciones para explorar las diferentes alternativas posibles de decisión. A partir de un conjunto de entrenamiento puede expresarse de manera recursiva, donde se escoge un atributo como nodo raíz y se divide el conjunto de entrenamiento según sea el valor del atributo elegido y se repite el proceso para cada uno de ellos [31]. Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión. El árbol de decisión suele contener nodos internos (contiene un test sobre algún valor de una de las propiedades), nodos de probabilidad (debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema), nodos hojas (representa el valor que devolverá el árbol de decisión) y arcos (las ramas definen posibles caminos que se toman según la decisión tomada).

- **Redes Bayesianas**

  Una red Bayesiana es un modelo probabilístico multivariado que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente las influencias causales [31]. Gracias a su motor de actualización de probabilidades, el Teorema de Bayes, las redes bayesianas son una herramienta extremadamente útil en la estimación de probabilidades ante nuevas evidencias. Relaciona un conjunto de variables aleatorias mediante un grafo dirigido (acíclico) que indica explícitamente la relación, en la cual cada nodo representa una variable aleatoria y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres [31]. Alguno de los tipos de Redes Bayesianas usuales incluyen: Redes Bayesianas Dinámicas (DBNs), Redes Gaussianas y Cadenas de Markov. Las redes Bayesianas se caracterizan por tener una sola de las variables de la base de datos (clasificador) que se desea predecir, mientras que el resto corresponde a los datos propios del caso que se desea clasificar. Pueden existir una gran cantidad de variables en la base de datos, algunas de las cuales estarán directamente relacionadas con la variable clasificadora que se quiere predecir pero también pueden existir variables que no son influyentes sobre dicha clase.

- **Modelos de Markov ocultos**

  El modelo oculto de Markov (HMM) es un caso particular de una red Bayesiana que el sistema a modelar es un proceso de Markov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos de dicha cadena a partir de los parámetros observables [90].

Se debe determinar los parámetros del modelo - la probabilidad de transición $a_{ij}$ y $b_{jk}$ - desde un conjunto de muestras de entrenamiento. No se conoce ningún método para obtener el conjunto óptimo o más cercano de parámetros desde los datos, pero se puede determinar casi siempre una buena solución por una técnica más directo.

Las reglas para realizar una clasificación, es utilizar características extraídas de los objetos, donde la mayoría de los modelos usan características numéricas. Cada objeto a clasificar, por ejemplo, peces o animales, se representa como un vector en el espacio Euclidiano multidimensional, y cada componente del vector representa una característica, es decir, un rasgo que se espera sea significativo para la clasificación, como se puede observar en la Figura 2.2. El conjunto de vectores forman la matriz de trabajo que permite realizar procesos de selección y clasificación de los datos.

El uso de cualquiera de estos métodos dependerá de la cantidad de clases, características, dimensiones e información a utilizar en el estudio [11, 12, 18, 38]. Cuando las clases no son conocidas de antemano, existe la necesidad de identificarlas y para ello existen varios algoritmos, como son: *k*-medias, *k*-medias difuso, método de la máxima verosimilitud y otras de clustering jerárquico [31].

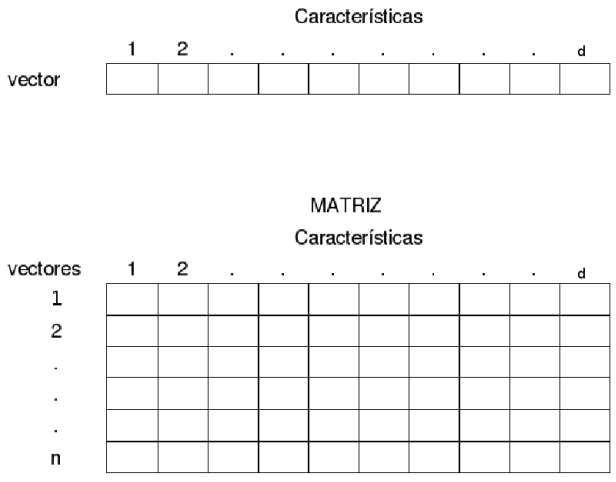



Figura 2.2: Vector y matriz de trabajo.

Para la resolución de problemas y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología

correspondiente y que dependerá de esa clasificación. Los métodos de clasificación pueden dividirse en distintos grupos según el procedimiento que se lleve a cabo.

- Los métodos estadísticos, como el Discriminante Lineal y K-Vecinos, en los que la clasificación se realiza teniendo en cuenta características estadísticas de los datos, tales como media, mediana, funciones de densidad de probabilidad, entre otras.
- Los métodos basados en árboles de decisión, en los que la clasificación se realiza mediante una serie de preguntas sobre los atributos de los individuos.
- Los métodos de agrupamiento, los cuales buscan generar centros de clases, ya sea de forma determinística o difusa, y agrupar los datos alrededor de dichos centroides. Los métodos de agrupamiento determinístico, como K-means, asignan a cada individuo una sola clase; y los métodos de agrupamiento difusos que asocian a cada individuo un grado de pertenencia a cada una de las clases.

En el presente trabajo se utilizan dos clasificadores que pertenecerían a la primera agrupación mencionada, un clasificador lineal y un clasificador cuadrático ambos basados en el teorema de Bayes. Los cuales son utilizados con diferentes métodos que permiten identificar cuales características son las que entregan una mejor separabilidad entre las clases.

### 2.2.4. Evaluación

Es una etapa muy importante en el diseño, ya que permite predecir el comportamiento de este a futuro, es decir, cuando deba clasificar objetos desconocidos. En la etapa de aprendizaje como en la etapa de resultados (reconocimiento de los patrones), es necesario contar con distintos elementos que permitan evaluar la efectividad de los procesos. Existen varios métodos dirigidos a la evaluación de clasificadores y otros a la evaluación de la clasificación en la etapa de resultados. Algunos de estos métodos serán detallados a continuación:

#### 2.2.4.1. Evaluación del clasificador.

En esta fase se utilizan los resultados obtenidos por el clasificador a través del análisis de la tasa de aciertos o errores y califica al clasificador para recomendar decisiones y acciones que dependen de un costo o riesgo particular. Por lo tanto y en este sentido el objetivo de la

evaluación es buscar un mínimo de errores y fallas. Para poder efectuar el análisis, se utilizan métodos conocidos como estimadores del error, los cuales determinan la proporción de patrones mal clasificados. Si se tiene un conjunto de datos $R$, que posee $m$ patrones distribuidos en $c$ clases, con un objeto $y$ que se presenta a todo objeto de entrenamiento $x_i(i = 1,2,...,m)$, se logra la siguiente función $E$ que muestra el acierto o error. Existe un error cuando la clase asignada a $y$ es distinta a la clase verdadera de $y$, y es un acierto cuando las clases asignada y verdadera son iguales. Las clases verdaderas están preasignadas, para ser utilizadas en la búsqueda de errores.

Luego la tasa de error (el error de clasificación) se mide como la probabilidad de error, la cual entrega una buena visión acerca de la calidad de un clasificador. Para esto existen diferentes medidas de la tasa de error (aparente, real, esperada y Bayes). A continuación se mencionan o se describen algunos métodos que permiten realizar la evaluación de los clasificadores, estos son: Restitución o reclasificación, partición mediante un conjunto de pruebas, validación cruzada con $v$ conjuntos o rotación y validación cruzada con $v = n$ (leave-one-out) [31, 90].

- Validación cruzada. Dado un conjunto de datos $D = D_1 U D_2 U ... U D_c$ donde cada $D_i$ tiene $n_i$ etiquetas. El método de validación consiste en dividir $D_i$ en dos subconjuntos $D_{i1}, D_{i2}$ (no es necesaria la misma cardinalidad), para luego usar uno de los conjuntos de datos para entrenamiento y el otro conjunto para validación. Por cada $i$ se tiene un $D_{i1}$(conjunto de entrenamiento) y $D_{i2}$ (conjunto de test). Los pasos son: 1) realizar el entrenamiento, 2) realizar test y validación, 3) ajustar modelos, 4) volver a 1 si es necesario. La meta es obtener la menor clasificación de error.

- Validación cruzada de $v$-pliegues. Es una generalización de la validación cruzada, donde $D_i$ es dividida en $v$ diferentes grupos de igual tamaño $\frac{n_i}{v}$, si $n_i \geq v$. El clasificador es entrenado $v$ veces. Es decir, que existen $v$ tasas de error y una tasa de error general. La más utilizada es la de *diez* pliegues, o la de $n$ pliegues

- Validación cruzada con $v = n$ (validación leave-one-out). Se tiene $T_{(n_i-1)}$ y $T_s$ es el estimado que queda afuera, esto se debe repetir $n_i$ veces por cada ejemplo. Para este método se necesitan altos recursos computacionales.

#### 2.2.4.2. Evaluación de la clasificación.

En este tipo de evaluación es común utilizar la precisión general para evaluar el desempeño del clasificador, pero no es adecuado cuando existe un desequilibrio en el conjunto de los

datos [7]. Por la tendencia a clasificar de manera incorrecta patrones de la clase minoritaria, debido a la influencia de la clase mayoritaria. Es decir, que la mayor cantidad de elementos se encuentran en la clase mayoritaria, esto hace que el resto de los elementos pertenecientes a la minoritaria se clasifiquen de manera errónea. Se debe mencionar que para esta etapa no se hace uso de el mismo conjunto de patrones que en la etapa de aprendizaje. Algunos de los criterios de evaluación utilizados incluyen:

- La precisión general y media geométrica.

  La precisión general utiliza la media aritmética de $p$ valores de manera que es la suma de patrones clasificados correctamente divididos por el total de patrones presentados para su clasificación. Es un criterios sencillo en implementar, pero poco adecuado por considerar de manera conjunta las precisiones individuales de la clase.

  La media geométrica es aquella que considera separadamente las precisiones observadas por cada una de las clases. Es útil solo si todos los números son positivos (Si uno de ellos es 0, el resultado es 0, si hay números negativos, la media es negativa o inexistente. También puede expresarse como la raíz $n-$ésima estima del producto de un conjunto de $n$ valores observados.

- Varianza y desviación estándar.

  La varianza es la media de los cuadrados de las desviaciones de sus $m$ valores respecto a su media. La varianza proporciona la dispersión de los datos en torno a la media, calculando la media de las diferencias de los valores. Como habrá valores que estén por sobre o por debajo de la media, el ajuste para compensar la situación consiste en calcular el cuadrado de las diferencias.

  La desviación estándar es el criterio de evaluación utilizado para analizar la dispersión de los datos y/o resultados obtenidos en los experimentos mediante la obtención de los cuadrados de las desviaciones de los valores de la variable respecto a su media. Es una medida de distancia promedio de los valores observados a su media. La distancia de cada valor a la media se mide tomando el cuadrado de la diferencia entre ese valor y la media. Luego de obtener el promedio de esos cuadrados, tomamos la raíz cuadrada. La desviación estándar es la raíz cuadrada de la varianza. Cuanto mayor sea la dispersión, mayor es la desviación estándar. Si no hubiera ninguna variación en los datos, es decir, si fueran todas iguales, la desviación estándar sería cero.

- Matriz de confusión y coeficiente Kappa.

La matriz de confusión es una herramienta utilizada para la presentación y el análisis del resultado de una clasificación debido a su capacidad de plasmar los conflictos entre las clases. La matriz de confusión se puede ver como una matriz cuadrada de orden *axa* con varias filas y columnas auxiliares para contabilizar diversos parámetros estadísticos. Esta matriz es construida utilizando las clases del conjunto de entrenamiento; las filas representan los datos de referencia, los valores marginales indican el número de patrones que, perteneciendo a una determinada clase, no fueron incluidos en ella y las celdas no diagonales de las columnas señalan los resultados de clasificar patrones que se incluyeron en una determinada clase cuando realmente pertenecían a otra. A lo largo de la diagonal principal, se indican los patrones que fueron clasificados de manera correcta. La columna "Asignación por clase" muestra el total de patrones que fueron asignados a una determinada clase, mientras que la columna "Asignación correcta" indica el porcentaje de patrones asignados de forma correcta de esa clase. La fila "% Bien clasificados" proporciona el porcentaje de patrones que fueron clasificados de forma correcta en una determinada clase. Uno de los análisis que se pueden realizar a la matriz es la precisión general, la cual se obtiene dividiendo el total de patrones bien clasificados entre los mal clasificados.

El coeficiente Kappa es otra forma de medir la exactitud de la clasificación. Kappa es un indicador (de rango entre -1 y 1) indicando falta de concordancia (cercano a 0) o concordancia total (cercano a 1). Este método puede medir la exactitud de manera más precisa que la matriz de confusión por calcular, no solamente los valores de sus columnas de los extremos, sino también los contenidos en el interior de la matriz [34, 58].

# Capítulo 3

# Bioinformática y sus Aplicaciones

Hoy en día la informática se encuentra muy ligada a la biología, principalmente por su gran apoyo en sus distintas ramas. Los apoyos son a través de la creación de métodos de análisis e interpretación, que ayudan a comprender de mejor manera el área de la biología, como son las secuencias de nucleótidos y aminoácidos, dominios de proteínas y estructura de proteínas a través del análisis e interpretación de los datos. También permite crear y desarrollar herramientas que puedan usar, manejar y acceder la información, además de algoritmos que permitan relacionar diferentes partes de un conjunto de datos [4].

Se puede decir que la modelización y simulación de sistemas biológicos, el desarrollo y aplicación de algoritmos orientados al análisis de datos en distintas áreas de conocimiento biológico es conocido como Bioinformática. Una definición más formal según el Centro Nacional para la Información Biotecnológica (National Center for Biotechnology Information - NCBI, 2001): Bioinformática es un campo de la ciencia en el cual confluyen tres disciplinas: biología (donde se originan los datos a analizar); computación (que proporciona el hardware y las vías de comunicación de los resultados entre investigadores); y tecnología de la información (que entrega los programas y resultados a analizar).

El estudio de las diferentes áreas a llevado a un alto crecimiento de datos disponibles que producen un gran volumen de información, debido a los proyectos de investigación, principalmente en el área de la genómica y la genética. Un ejemplo de esto se puede ver en el Anexo H. Además se intenta organizar las diferentes bases de datos disponibles, es decir, lograr trabajar con ellas de manera conjunta, mediante el desarrollo de interfaces comunes tanto a nivel de usuario final como a nivel de plataformas de desarrollo, para extraer mayor cantidad de información.Como por ejemplo,el sequence retrieval system (SRS) [33], es

un gestor de bases de datos desarrollado específicamente para trabajar con bases de datos biológicas que proporciona un acceso eficiente a las bases de datos de contenido biológico, permitiendo acceder a las mismas en cualquier formato en que estén disponibles los datos, utilizando mecanismos complejos de búsqueda. SRS fue desarrollado en el Instituto Europeo de Bioinformática (EBI) de Hinxton, Inglaterra.

A pesar de la gran cantidad de datos en secuencias y de los avances en las técnicas experimentales para proporcionar modelos aproximados de la estructura y dinámica de las proteínas (cristalografía de rayos X o resonancia magnética nuclear), cada día aumenta la diferencia entre el número de secuencias y el número de estructuras conocidas. Desde 1996, se emplea la Bioinformática para la predicción de la estructura de las proteínas a partir de su secuencia de aminoácidos por medio de técnicas como modelamiento comparativo o predicción de plegamiento. Los métodos de predicción de estructuras tienen por objetivo proporcionar un modelo para poder dirigir los estudios biológicos y dar una base estructural para la interpretación de los fenómenos biológicos cuando no se dispone de la estructura determinada experimentalmente.

## 3.1. Proteínas

Para enfocarse en la problemática abarcada por este proyecto, nos centraremos en las proteínas y sus interacciones, para esto se realiza una descripción de su importancia para la bilogía, como para el presente proyecto.

Las proteínas determinan la forma y la estructura de las células y dirigen la mayoría de los procesos vitales. Las funciones de las proteínas son específicas de cada una de ellas y permiten a las células mantener su integridad, defenderse de agentes externos, reparar daños, controlar y regular funciones, entre otras. Todo depende de su correcto plegamiento, si una proteína no se pliega correctamente será no funcional y, por lo tanto, no será capaz de cumplir su función biológica. El estudio de la función biológica de las proteínas o sus interacciones, se encuentra íntimamente relacionada con su estructura nativa[1], la cual está determinada por las múltiples interacciones que se suceden entre los aminoácidos que forman la cadena polipeptídica. Los aminoácidos naturales son moléculas fundamentales cuya polimerización lineal forma las proteínas, las cuales interaccionan a través de sus propiedades como la hidrofobici-

[1]Estructura tridimensional de una proteína en condiciones fisiológicas, considerada la más estable de las estructuras posibles.

dad, capacidad de formación de puentes de hidrógeno y la presencia de cargas parciales (ver anexo C).

## 3.2. Interacciones

Las interacciones entre los átomos de los aminoácidos están sujetas a restricciones topológicas impuestas por la conectividad de la cadena, por lo que el perfecto equilibrio entre las interacciones estabilizantes (que mantienen la formación de la estructura nativa) y las interacciones desestabilizantes (que interrumpen la formación de la estructura nativa y que impiden que las proteínas adquieren una estructura incompatible con su función biológica) dan como resultado el plegamiento nativo de la proteína, el cual generalmente corresponde a un mínimo global de energía potencial[2]. En presencia de moléculas adicionales de otras proteínas se formarán también interacciones atractivas y repulsivas con los aminoácidos de las otras cadenas que pueden llevar a la formación de agrupaciones intermoleculares o agregados, como son las interacciones: proteína-proteína, proteínas-ADN, proteínas-moléculas pequeñas [1], proteína-ligando entre otras. Estas interacciones dependen del estado o circunstancias (temperatura, pH, fuerza iónica, entre otras) en que se encuentran las partes y del entorno.

Al estudiar las interacciones entre proteínas se puede observar que se encuentran involucrados en múltiples procesos celulares tales como transducción de señales (retransmitir una señal con cambios de mensajeros), interacción antígeno-anticuerpo, regulación de la expresión de genes y en el funcionamiento de un gran variedad de otros procesos (complejos multi-moleculares) cuya funcionalidad es la formación de multímeros para crear el estado biológicamente activo [61]. La unión de dos o más cadenas de proteínas, a través de una interacción se le denomina Complejo.

La información de la estructura de algunos de estos complejos, se encuentra almacenada en la base de datos de estructuras tridimensionales de proteínas determinadas experimentalmente “Proteína Data Bank” - PDB [10]. Esta contiene las coordenadas cartesianas de cada átomo que forma la estructura atómica de una proteína. Esto permite utilizar herramientas de visualización y análisis especialmente de su superficie de interacción. Actualmente contiene 48.235 estructuras acumuladas hasta la fecha de las cuales 70 son las identificadas el presente año[3][10].

[2]Energía conformacional dada por las contribuciones de las energías electrostáticas, polares, de Van der Waals, etc.

[3]PDB, última actualización del sitio, 8 de Enero de 2008.

Cuando se analizan las estructuras de proteínas almacenadas en las bases de datos, se pueden reconocer similitudes y por lo tanto, encontrar proteínas de estructura similar (plegamientos semejantes) pero con distinta funcionalidad y proteínas con igual función, pero con diferente estructura (plegamientos diferentes)[43]. La identificación de estas similitudes permite clasificar familias que comparten plegamientos similares (Anexo C).

A pesar de conocer la estructuras de muchas proteínas, no existen métodos para predecir la interacción proteína-proteína (IPP) de alta precisión, que indique que proteínas interactuarán entre ellas y de que manera. Por esta razón, tampoco es posible predecir la estabilidad de la interacción todo esto, debido a que no se puede determinar la función de una proteína, conociendo sólo la secuencia o estructura en forma aislada (se debe tener en cuenta que todavía no existe un entendimiento completo del plegamiento de una proteína). La información estructural de las proteínas no ha avanzado tan rápidamente como la información sobre secuencias y funciones (48.235 proteínas v/s 17.960.667 secuencias) [10]. Como consecuencia la biología molecular ha producido gran cantidad de información funcional y estructural sobre estos complejos. Es aquí donde la informática contribuye, ya que las evidencias acumuladas [3] señalan una estrecha relación entre secuencias, estructuras y función.

La informática ha hecho posible en algunos casos estudiar las interacciones entre proteínas mediante experimentos a gran escala. Sin embargo, la información disponible sobre interacciones entre proteínas proviene del estudio de las estructuras tridimensionales de complejos proteicos por experimentos de interacción *in vivo* (técnicas que se realizan en un organismo vivo). Actualmente se conocen más de 12.000 estructuras de complejos que involucran más de 2 cadenas polipeptídicas (ver anexo C), pero sólo se conoce la estructura atómica de 1.943 estructuras, acumuladas a la fecha (última actualización PDB) (ver anexo B).

Durante los últimos 10 años, se han realizado estudios e investigaciones sobre las funciones e interacciones de las proteínas. En estos estudios han ido cambiando los métodos (características, geometría, probabilísticos) y las tecnologías (computacionales, biológicas) utilizadas, pero aún no existe un método establecido que permita conocer con un solo procedimiento una proteína en su totalidad y así permitir el manejo de las interacciones proteína-proteína de forma real (*in vivo*). Los métodos que se utilizan actualmente son artificiales, que se identifican como: *in silico,* técnica realizada por computador o vía simulación computacional e *in vitro,*técnica a desarrollar en un tubo de ensayo, o ambiente controlado fuera de un organismo vivo. Por esas razones se han realizado y se realizan esfuerzos en investigaciones para entender a nivel atómico (en detalle) los principales responsables de estas interacciones [48, 51, 82, 91]. Los esfuerzos han sido dirigidos a la caracterización de la geometría (forma)

y las características fisicoquímicas (energéticas de la interfaces de interacción) [22, 30], de modo de obtener el siguiente tipo de información:

- la preferencia de los residuos por aparecer en la superficie [36],
- los parámetros geométricos y complementariedad de forma entre las cadenas que interactúan [60],
- el rol de los puentes de hidrógeno, puentes salinos e interacciones hidrofóbicas y polares en la superficie de las proteínas [93],
- la pérdida de superficie accesible al solvente como resultado de la interacción [86] y
- el análisis de la conservación de los residuos en la superficie de interacción [65].

## 3.3. Zona de interacción

Debido a que las proteínas están formadas por aminoácidos, las características de aquellos que estén en la superficie así como su posición relativa, y la geometría de la superficie (forma) definen finalmente algunas propiedades que caracterizan su modo de acción o las capacidades de interacción con otras proteínas o moléculas. Cuando una proteína participa en una interacción proteína-proteína (IPP), la realiza con una o varias parte de su superficie.

En la Figura 3.1, se muestra un complejo de interacción proteína-proteína (interacción entre dos proteínas, las cuales son representadas por colores diferentes y el área marcada con rojo, corresponde a la interacción), que se unen por una pequeña área, la cual se denomina **zona de interacción** (interface), descripción general dada a los sitios de unión. Esta zona de interacción tiene propiedades diferentes al resto de la superficie, lo que les permite que interactúe específicamente con una o más proteínas.

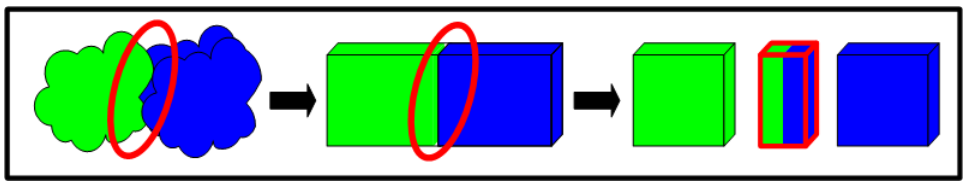


Figura 3.1: Esquema de una Zona de IPP (representación).

### 3.3.1. Características presentes en IPP.

Varios estudios [3, 12, 49, 66, 93] señalan que para encontrar las características que determinen de mejor manera una interacción, es necesario evaluar los criterios físico-químicos presentes en la mayoría de las proteínas. Esto quiere decir, que se deben evaluar las características que participan en la interacción como tal, y las que se encuentran en el resto de la proteína. Alguna de estas características que han sido utilizadas para la determinación de las zonas de interacción incluyen:

- Área de la superficie de interacción: corresponde al área que ambas proteínas tienen contacto y se obtiene a través de la suma de las áreas individuales de los átomos que no están expuestos al solvente, cuando se forma parte del complejo.
- Porcentaje de área hidrofóbica: corresponde al porcentaje del área total que es aportada por los residuos hidrofóbicos.
- Porcentaje de área hidrofílica: corresponde al porcentaje del área total que es aportada por los residuos hidrofílicos.
- Número de aminoácidos polares: es el número de residuos polares que forman parte del área de interacción.
- Número aminoácidos apolares: es el número de residuos apolares que forman parte del área de interacción.
- Número de puentes salinos: corresponde a las interacciones entre grupos cargados que se establecen entre los residuos que forman parte de las superficies de interacción.
- Número de puentes hidrógeno: corresponde al número de interacciones puente hidrógeno (C=O——H-N) que se forman entre los residuos que forman parte de las superficies de interacción.

Algunas de estas características han sido estudiadas de manera independiente, como por ejemplo estudios de:

- los puentes de hidrógeno y puentes salinos. En los cuales se analiza su participación en el plegamiento propio de la proteína (folding), es decir, sin interactuar con otra proteína y luego su posible participación en interacción (binding) [93] con otra proteína.

- las cargas iónicas, hidrofobicidad y cantidad de residuos en secuencia, para identificar potenciales interacciones, con ayuda del método MVS [13].

Jones y su grupo [49] definieron los inicios de las interacciones proteína-proteína en base a estudios realizados con complejos de estructura conocida. Las investigaciones se enfocaron en las características físico-químicas, tales como la hidrofobicidad. También se evaluó la forma y el tamaño de los complejos, además de su superficie y cambios conformacionales. Posteriormente se evaluó las interacciones desde el punto de vista de los dominios y la interface (área de interacción), tomando como características para análisis: el área superficial accesible (ASA - $Å^2$), la planaridad (desviación de los átomos de un plano óptimo que describe la superficie de interacción), la segmentación (número de posiciones o zonas de la proteína que interaccionan con la otra), el porcentaje de residuos polares y los puentes de hidrógeno. Realizaron análisis de los parámetros antes descritos en monomeros y oligomeros de proteína, para luego ser comparados con los de la superficie intra-cadenas (contacto entre residuos de la misma estructura de dominio), de complejos permanentes y transitorios, lo que resultó con una alta similaridad entre la superficie de dominios intra-cadenas y las interfaces inter-dominios (contacto de residuos de diferentes estructuras de dominio) [52].

Otros estudios comparan 6 tipos de interfaces de proteínas [72], de las cuales dos son internas (interfaces intra-cadenas y inter-dominios) y cuatro externas (homo-oligomero permanente, homo-oligomero transitorio, hetero-oligomero permanente, hetero-oligomero transitorio). Los resultados muestran que los diferentes tipos de interfaces difieren por la composición de los aminoácidos y las preferencias de los residuos de contacto. En el estudio también se sugiere que en el pasado existían más tipos de interacciones y cada una de ellas estaba basada en diferentes mecanismos físico-químicos.

En la Tabla 3.1 se muestran algunas de las características utilizadas para estudios de interacción de complejos proteína-proteína, mencionadas previamente y obtenidas desde la base de datos Protein-Protein Interactions Server (PPIS).

| Parametros de interfaces de proteína | Valor |
|---|---|
| Área de la superficie accesible a la interfase | 789.53 |
| % Área de la superficie accesible a la interfase | 13.32 |
| Planaridad | 2.89 |
| Largo & amplitud | 33.51 & 20.83 |
| Largo/amplitud relación | 0.48 |
| Segmentos de residuo en la interface | 4 |
| % Interface en átomos polares | 47.61 |
| % Interface en átomos no polares | 52.30 |
| Enlaces de hidrógeno alpha/beta en Estructura secundaria | 6 |
| Puentes salinos | 0 |
| Enlaces de disulfuro | 0 |
| Volumen de Gap | 4356.51 |
| Índice Volumen de Gap | 2.76 |
| Puentes de moléculas de agua | 0 |
| Lista de residuos en la interface | sitio.html |

Tabla 3.1: Ejemplo de propiedades de las superficies de IPP, generados por PPIS .

### 3.3.2. Clasificación de interacciones entre proteínas

En el año 2003, Nooren [70], realizaron una clasificación de los diferentes tipos de interacciones , se definieron 3 grupos:

- complejos homo (interaccionan dos proteínas iguales) y hetero oligomeros (interaccionan dos proteínas diferentes);
- complejos permanentes (proteínas que normalmente no están solas, funcionan pero emparejadas) y no-permanentes (proteínas que pueden interaccionar solas o emparejadas);
- complejos transitorios (interacción de corta duración y sometidas a mayores cambios o mutaciones) y permanentes/obligados (interacción de más larga duración)[37].

En la tercera clasificación, las interacciones son diferenciadas basándose en el tiempo de vida del complejo. Las interacciones permanentes (duraderas en el tiempo) están integradas por múltiples subunidades (idénticas o diferentes), son más estables, debido a que no se producen

tantos cambios en sus partes. Mientras que las interacciones transitorias (temporales) pueden producir enlaces débiles o fuertes, pero de muy corta duración, lo que dificulta su análisis, por esto las proteínas que participan en la interacción sufren cambios que son complejos de reproducir para poder ser estudiados [49, 70, 84].

A continuación se presenta un resumen de algunas de las características que participan en la superficie de una IPP:

- El área de interfase (ASA) va desde 550 - 4900 $Å^2$ (5 - 30) de la superficie de las proteínas que interactúan.
- La interfase posee un mayor número de aminoácidos hidrofóbicos y aminoácidos como R, H, N, W, Y, S.
- Las estructuras secundarias que participan en interacciones son rc=47 %. H=36 %, E=17 %.
- El 10 % de las interfases corresponden a cavidades.
- Existe hasta 10 grupos cargados por unidad de superficie de interfase.
- Existe un puente de Hidrógeno cada 100 - 200 $Å^2$(10/superficie).
- kd<$10^{-4}$ - $10^{-14}$ M>6 - 19 (kcal/mol)

Existen características que diferencian las clases de IPP como la constante de afinidad (kd: uM. Transitorios y kd: nm. permanentes), el rol de los parches hidrofóbicos, la presencia o ausencia de agua en la interface y la presencia de cavidades.

## 3.4. Aplicación de Reconocimiento de Patrones en la Interacción de Proteínas

Otro intento por predecir las interacciones es la extracción de información (texto) desde literatura científica disponible, a través del reconocimiento de texto, donde se propone métodos de aprendizaje de patrones de manera automática. Se utilizan dos enfoques:

- Basado en algoritmos de alineamiento de secuencias, para generar patrones y buscar coincidencias con textos nuevos relacionados a interacciones proteína-proteína [77].

- Basado en probar y encontrar patrones en textos, utilizando autómatas de estados finitos y modelos de Markov [77].

Las búsquedas solo se realizan sobre proteínas conocidas. Se obtiene desde un 75 % a un 90 % de precisión en la extracción de información desde textos biológicos, lo que fue validado realizando una comparación con los resultados obtenidos de una validación cruzada de 10 pliegues. Los resultados indican que el aprendizaje de patrones es un método factible para la extracción de información de interacciones proteína-proteína desde textos libres e incluso sobre los patrones generados manualmente. Otro método propuesto para la extracción de información es de Ono [74] que se compone de varios pasos que van desde seleccionar la proteína a buscar y luego de diferentes reglas (pasos a continuación) se obtienen las coincidencias para extraer la información:

1. Identificar nombre o sinónimo de proteína a buscar. Cada investigador utiliza sus propias nomenclaturas, esto muestra la falta de estandarización en la información existente.

2. Componer sentencias complejas: De los resultados obtenidos en el paso uno, se generan reglas y sentencias complejas que permitan en el paso siguiente reconocer información de utilizada.

3. Los resultados obtenidos en el punto anterior son parseados y se elimina aquella información que genere ruido en los resultados.

4. La información se extrae, obteniendo el número de sentencias extraídas correctamente X o Y, con los cuales se evalúa la precisión del método X, con respecto a otro Y (X: número de sentencias de información que se contiene de IPP; Y: número total de sentencias recuperadas por otro método).

Las investigaciones realizadas para predecir IPP también se han centrado en el análisis de la estructura primaria de las proteínas, utilizando algoritmos rápidos y robustos, para comparar estructuras de proteínas con bases de datos distintas, con la finalidad de extraer familias de proteínas. Al tener representantes de las familias, se evaluó: su estructura, su alineamiento, parientes que deben tener un límite en el porcentaje de similitud. El límite de similitud se fija, debido a que si es muy alto, pasa a ser una cadena idéntica y no similar, en consecuencia quedaría en la categoría de redundancia de datos. Las cadenas que sobrepasan el límite son excluidas [29, 43, 56].

Existe un porcentaje de redundancia en las bases de datos disponibles, debido a variados estudios donde se repiten algunos factores y además existe un gran número de secuencias casi idénticas o idénticas. Para realizar un análisis estadístico de secuencias y estructuras, no debe existir redundancia, pero en estos casos es inevitable, aunque se hagan diferentes procesos de filtrado, siempre existirá un mínimo de redundancia. También se han identificado funciones desde la secuencia, revisando los enlaces (propiedades fisicoquímicas), para observar funciones similares entre las estructuras de superficie que están localizadas en las bases de datos. Esto permite que posteriormente sean analizadas y lograr obtener un umbral de similitud (a través de su funcionalidad, estructura u homología) [14, 13].

En el año 2003 se realizó un estudio que demostró que es posible la predicción de interacciones proteína-proteína desde la secuencia, pero que necesita de depuraciones para una mejor precisión [73]. Desde el año 2003 hasta el 2008, se ha investigado la estructura, función y evolución de proteínas, y aún no existe una forma de relacionar directamente estas clasificaciones a las interacciones proteína-proteína [20, 65, 67, 71, 88, 94, 98]. Existen estudios sobre interacciones proteína-proteína, pero no todos utilizan el mismo método, técnica o clasificador, además de las características propias de una proteína.

Una manera de trabajar para obtener buenos resultados es realizando una buena selección de la base de datos a utilizar para minimizar así la cantidad de información, por medio de una selección y además reducir la redundancia de datos existentes [42, 59, 74, 79]. Un aspecto importante es el ambiente en donde se encuentra la proteína, ya que esto puede determinar su estructura secundaria y sus posibles interacciones. Para esto se observan las características de los aminoácidos y así poder determinar la accesibilidad de los solventes en las cadenas laterales. También se observa que la interacción de éstos con otros grupos de residuos y el desarrollo de algoritmos que separan las estructuras alfa y beta, pueden indicar su formación en estructura secundaria [66].

Qi y su grupo [78]realizó una investigación sobre interacción proteína-proteína en la cual se divide la interacción en tres etapas: interacción física, relación co-complejo y co-miembro de la vía (es decir, son enzimas en que ambas participan de una vía enzimática o metabólica) y se intentó determinar la exactitud de predicciones de interacción. Para esto se seleccionaron seis métodos de clasificación que se aplicaron a un conjunto de datos biológicos y que fueron evaluados en diferentes bases de datos. Los clasificadores utilizados fueron: Random Forest (RF), RF basado en k-NN, Bayes, Árboles de decisión, Regresión logística, y MVS. El clasificador RF es el más robusto y favorable para las tres tareas antes mencionadas, entre los seis métodos para predecir interacciones proteína-proteína, debido a que entrega una mayor

precisión. Los resultados indican que la tarea relación co-complejo es la más fácil de predecir y las fuentes de datos biológicas que se utilizaron codificaron la tarea de la predicción co-compleja de mejor manera que las otras dos tareas. Se aclara, en el estudio que la situación puede cambiar cuando la cantidad de datos disponibles es mayor. La tarea más compleja fue sobre la interacción física, ya que son moduladas fuertemente por variaciones en los factores ambientales [78].

La característica más estudiada es la naturaleza de los aminoácidos que forman parte de las diferentes interfaces proteína-proteína. El estudio más completo fue realizado por Ofran y Rost [72], en el cual se estudian seis tipos de interfaces: intra/inter dominios, homo/hetero-olígomeros de complejos permanentes/transitorios. Se indica que de acuerdo a la composición de aminoácidos de las superficies estas son diferentes, ya que existe solo un 1,5 % de similaridad entre las superficies internas y externas, y un 0,2 % de similaridad entre hetero superficies que pertenecen a homo complejos permanentes y homo complejos transitorios, y entre complejos de formación transitoria la similitud es del 16,3 % entre homo y hetero complejos.

El estudio compara la composición de las diferentes superficies con la composición de aminoácidos de la base de datos de proteínas SwissProt [6] e identifican que ninguna de las superficies tiene una composición con una similitud mayor al 1 %, concluyendo que la composición aminoácidica para las diferentes superficies es diferente y en principio podría utilizarse esta característica con fines predictivos. Una gran cantidad de estudios han establecido que la composición de los residuos en la interface de interacción entre proteínas la diferencia generalmente de la del resto de la superficie [2, 5, 35, 50, 54, 63, 99].

La probabilidad de que un aminoácido forme parte de la interface, se define como el cuociente entre lo esperado de su frecuencia de participación en la superficie de la proteína y su frecuencia de participación en superficies de interacción. Existe un número de métodos de predicción de interfaces, en los cuales se han considerado las propensiones de las interfaces [16, 18, 27, 28]. Además, esta información se ha utilizado como una forma de jerarquizar los resultados obtenidos por docking (pruebas de acoplamiento) entre proteínas [22, 39, 68, 95].

También se han analizado influencias del tipo de residuo y de la estructura en accesibilidad al solvente [44]. Se definió una medida de relativa exposición a la hidrofobicidad, en conjunto con la estructura secundaria como parámetros de predicción.

Primero se describieron los análisis usados para determinar los parámetros para el algoritmo de la predicción de interfaz; éstos incluyen accesibilidad o información estructural tal como la

interacción entre estructuras beta plegadas o estructuras helicoidales. Con estos datos podría lograrse la identificación de las regiones superficiales implicadas en interacciones proteína-proteína, pero esto también depende del tiempo de ejecución del algoritmo utilizado para las predicciones, ya que varía según la proteína y el número total de residuos que contenga.

Trabajar con las superficies de las proteínas es un tema que ha sido abordado por varios investigadores han desarrollado, identificando y analizando las características energéticas de los "hot spots" (regiones de alta frecuencia de recombinación) en interfaces proteína-proteína usando variados enfoques detallados [53, 15, 57]. Haliloglu y su grupo [40] presentan un método para identificar "hot spots" en las interfaces de unión, a través de un proceso de similaridad de altas frecuencias de vibración entre los residuos que participan en el plegamiento de la proteína y aquellos residuos que participan en una interacción (en la superficie). Esto permite distinguir entre sitios de unión y el resto de la superficie de la proteína.

# Capítulo 4

# Metodología

En este capítulo se presenta la metodología desarrollada, que se compone de tres etapas. La primera es el proceso de Preparación de los datos que serán utilizados durante el proyecto. Luego se continúa con la etapa de Selección de las características más relevantes y se finaliza con la etapa de Evaluación (clasificación y evaluación), donde se aplican los diferentes métodos a los datos escogidos en la etapa anterior, estos se ven representados en la Figura 4.1.

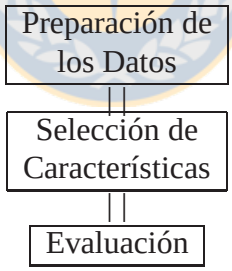



Figura 4.1: Metodología desarrollada.

Cada una de estas etapas se compone de varios pasos, los cuales serán guiados por las etapas del reconocimiento de patrones, definidos en el Capítulo 2, para ser descritos en las secciones siguientes.

Se presentan dos problemas en la metodología, la dimensión del espacio puede ser muy alta y no todas las características son esenciales o significativas para la clasificación, estos problemas se presentan en conjunto con las etapas de la metodología desarrollada.

## 4.1. Preparación de los Datos

Actualmente existen varias bases de datos disponibles sobre IPP [43, 56], cada una de las cuales se crea para una investigación específica. La mayoría de las bases de datos son para estudios e investigación, y no para uso comercial. Se revisaron bases de datos para identificar cuales son las de mayor accesibilidad y aptas para el trabajo a desarrollar.

Una manera de identificar las características y posterior clasificación (identificación de características discriminantes), es utilizar la clasificación de tipo de IPP por transitorias y permanentes [3, 23, 56, 49], definida por Thornton. Esto desde un enfoque macro hasta centrarse en la problemática que se encuentra en la zona de interacción de la proteína. La búsqueda de información de complejos de proteínas de estructura tridimensional conocida (estructura terciaria [92]) permitió una mejor validación de los datos obtenidos. La información fue seleccionada desde las bases de datos de complejos de interacción proteína-proteína y se escogieron características de la superficie y no redundantes.

Luego se debe hacer uso de estos complejos en una aplicación que permita obtener datos energéticos de la interacción, que sean útiles para ser analizados y permitan una futura clasificación.

## 4.2. Extracción de Características Esenciales

En esta etapa se selecciona un subconjunto de características manteniendo lo más alta posible la precisión en la clasificación. Esto no es una tarea simple, debido a que algunos de los clasificadores existentes utilizados para la selección de características (utilizados como criterios), no se conoce un algoritmo que pueda seleccionar un subconjunto de características en un tiempo de ejecución (computacional) de orden menor a una función exponencial expresada en términos de cantidad total de características, es decir, no se puede encontrar el óptimo, pero si soluciones subóptimas.

La predicción de la estructura tridimensional de las proteínas y de otros temas como el denominado docking (unión de moléculas a concavidades de proteínas) y el diseño de estructuras moleculares óptimas, se caracterizan por un denominador común: son problemas de alta complejidad, es decir, el conjunto de todos los problemas de decisión que pueden ser resueltos por algoritmos no deterministas de tiempo polinómico.

Un algoritmo de tiempo polinómico es aquel cuya complejidad en tiempo en el peor de los casos, está acotada superiormente por un polinomio. La teoría de la NP-completitud no proporciona un método para obtener algoritmos de tiempo polinómico, ni tampoco indica que estos algoritmos no existan. Esto muestra es que muchos de los problemas para los cuales no conocemos algoritmos polinómicos están relacionados computacionalmente. Para que un problema este en NP debe existir un algoritmo que haga la verificación en tiempo polinómico. Además de tener la propiedad de que puede ser resuelto en tiempo polinómico si y solo si todos los problemas NP-completos pueden ser resueltos en tiempo polinómico [25].

Varios algoritmos y criterios de optimización se han propuesto para la selección de características, siendo los más conocidos los de búsqueda secuencial, que serán detallados en las secciones siguientes.

## 4.3. Selección de Características

El próximo paso, después de obtener la matriz de trabajo, es lograr identificar aquellas características más relevantes para una clasificación entre clases. Por esto el objetivo principal de la selección de características es obtener el menor conjunto posible que represente con exactitud al conjunto original. Sin embargo, el nuevo conjunto de características debe capturar la variación del viejo conjunto lo mejor posible. Si se quiere desechar una característica, se debe escoger aquella que en orden de importancia para discriminar, ésta al último de todas las características. El proceso de obtención de características y su posterior clasificación se hará utilizando la aplicación Matlab para la programación de algoritmos, función objetivo y criterios de selección.

Una razón para reducir la cantidad de características, es la complejidad computacional, mientras mayor es el número de características, mayor es el tiempo necesario para la ejecución de un proceso. La relación más alta es entre el número de patrones de entrenamiento $n$ y el número de parámetros de clasificación libre. Existe un gran número de características que pasan directamente a ser parámetros de clasificación. La finalidad de esta sección es seleccionar las características más importantes a fin de reducir su número y al mismo tiempo mantener en lo posible la información discriminante de la característica. Esto se debe a que si se seleccionan características con poco poder de discriminación, el diseño de un posterior clasificador podría tener un menor rendimiento.

Se requiere obtener un factor relevante que influya en la IPP e identificar un subconjunto de

características discriminantes que permitan evaluar la posibilidad de separación de las clases a través de un ranking de las más influyentes en la interacción [75]. Esto se realiza a través de la implementación de un algoritmo de búsqueda secuencial, utilizando la distancia de Chernoff, como criterio de selección para finalmente obtener una lista de las características más relevantes. Paralelamente, se realizó un análisis de componentes principales, que permitirá buscar semejanzas, pero no se podrá comparar los resultados uno a uno, debido a que los resultados del método de búsqueda secuencial, entrega características y el ACP entrega componentes que representan a un conjunto de características.

Hay algoritmos y criterios de optimización de características relevantes que se han propuesto para lograr identificar aquellas características esenciales o significativas para una clasificación. Algunos métodos que se utilizan son los algoritmos secuenciales y los criterios de divergencia de Fisher [31], distancia de Chernoff [90], divergencia de Kulback-Leibler [90], e información mutua [55].

De los diferentes métodos y criterios de selección se optó por la distancia de chernoff como criterio, debido a que dos de los métodos de RLD que se utilizan estan basados en la distancia de Chernoff. En el caso del método propuesto por Rueda Herrera se optimiza la distancia de Chernoff en el espacio transformado. Loog Duin optimiza un criterio que utiliza matrices de distancia dirigida, las cuales incorporan la distancia de Chernoff pero en el espacio original. Y finalmente Fisher, que no utiliza el concepto de distancia de Chernoff, pero si se observan los valores obtenidos los resultados no son buenos, lo cual corrobora que la relación con la distancia de Chernoff es válida para la selección de características como para el método RLD.

### 4.3.1. Algoritmos para la Selección de Características

Los algoritmos de búsqueda secuenciales son aquellos donde se agregan o eliminan características iterativamente hasta satisfacer algún criterio de detención. Se pueden mencionar dos de ellos: los métodos hacia delante (forward) que comienzan con un vector vacio y se van agregando características; y los métodos de búsqueda hacia atrás (backward) comienzan con todas las características disponibles y se van eliminando una a una [31]. Los criterios de optimización, son aquellos que entregan un valor, el cual depende del criterio utilizado, y este permite escoger una característica (para una mejor selección), sobre las otras.

En el método de “Forward Search”, se definen dos pasos, los cuales serán explicados a través de un ejemplo [90]. Se define $l$ como las dimensiones a utilizar (posibles combinaciones de

características en un vector) y $d$ como las características disponibles inicialmente. El algoritmo se basa en los siguientes pasos:

1. El cálculo del *criterio objetivo* es realizado para cada característica o grupos de ellas y luego se seleccionan aquella(s) característica(s) que entregan el mejor resultado obtenido con el criterio, en el caso del ejemplo, es $X_1$ (el criterio será detallado en la sección siguiente).

2. Luego se forman todas las posibles combinaciones en vector de dos dimensiones, el cual contiene la característica ganadora del paso anterior y una de las características no seleccionadas en el paso anterior, que son $[X_1, X_2]^T$, $[X_1, X_3]^T$, $[X_1, X_4]^T$. Posteriormente se calcula para cada vector el criterio y se selecciona la mejor de ellas, en este caso $[X_1, X_3]^T$

   La combinaciones a realizar, dependen del valor de $l$. Si fuese 3 o más, se debe continuar con combinaciones de tres dimensiones, donde el vector contiene los dos ganadores del paso anterior, más una característica no seleccionada. Como es $[X_1, X_3, X_2]^T$, $[X_1, X_3, X_4]^T$ y seleccionar el mejor.

   Estos algoritmos sufren problemas debido a que las características descartadas en el método de búsqueda hacia atrás no pueden volver a seleccionarse. Del mismo modo, una vez seleccionadas las características en el método de búsqueda hacia delante, éstas no pueden ser descartadas posteriormente. Una solución que se propone es almacenar el segundo mejor resultado y realizar una búsqueda paralela con los dos valores, optando por la que presente mejor resultado.

#### 4.3.1.1. Criterio de Selección: Distancia de Chernoff.

El algoritmo de selección hace uso de una función objetivo o criterio que corresponde a la distancia de Chernoff, medida de similitud entre dos funciones de densidad de probabilidad. El criterio a utilizar se presenta en la Ecuación 4.1.

$$FOC = \frac{p_1 p_2}{2} (m_1 - m_2)^t S_w^{-1} (m_1 - m_2) + \frac{1}{p_1 p_2} log \left( \frac{| S_w |}{| S_1 |^{p_1} | S_2 |^{p_2}} \right) \tag{4.1}$$

Para obtener los valores de diferentes parámetros, se comenzó identificando: el número total de complejos ($n$), el número de complejos permanentes ($n_1$) y número de complejos transitorios ($n_2$), con lo cual se obtienen las probabilidades a priori de cada clase ($p_1$, $p_2$). También

se utilizan dos matrices ($X_1$ , $X_2$), permanentes (complejos de la clase 1 $\times$ número de características utilizadas en la iteración) y transitorias (complejos de la clase 2 $\times$ número de características utilizadas en la iteración) las cuales poseen las características que se han seleccionado en las iteraciones anteriores utilizando FOC (inicialmente se comienza con sólo una característica). El valor de la dimensión ($d$) depende de la iteración en que se encuentre el algoritmo de búsqueda. Además se obtiene la covarianza para cada clase, obteniendo $S_1$ y $S_2$. Además con las mismas matrices ($X_1$ , $X_2$) se obtiene la media por cada clase ($m_1$, $m_2$). Con esta información también se obtiene $S_w$, como $S_w = p_1 * S_1 + p_2 * S_2$. También se necesitan los determinantes de $S_1$, $S_2$ y $S_w$. Para esto, sea $S_1$ covarianza de $|S_1| = \prod_{i=1}^{d} \lambda i$,

$$S_1 = \phi \Lambda \phi^t \tag{4.2}$$

en la Ecuación 4.2es donde $\phi$, es la matriz con los vectores propios de $S_1$, $\Lambda = [\lambda_1, \lambda_2, ..., \lambda_d]$ la matriz con los valores propios de $S_1$, y $\phi^t$ la matriz transpuesta de $\phi$. Además se necesita obtener la inversa de *Sw*, esto se observa en la Ecuación **??**.

$$S_w^{-1} = \left(\phi \Lambda \phi^t\right)^{-1} = \left(\phi^{t^{-1}}\right) \Lambda^{-1} \phi^{-1} = \phi \Lambda^{-1} \phi^t$$

Cuando los componentes de la matriz de valores propios presenta ceros en la diagonal, se procede a establecer un umbral (Ejemplo: $1{,}0 \times 10^{-6}$) que permite reducir la posibilidad de encontrar componentes muy pequeños. Esto permite manejar la matriz para obtener la inversa de *Sw*, evitando la singularidad.

Para un manejo eficiente de la matriz con los valores propios, se procedió a reducir su tamaño, eliminando las columnas y filas que contienen el valor cero en la diagonal, debido a que no influyen en los resultados, pero facilita su utilización. Este corte de la matriz se produce en $k$. Por esta razón se presenta la multiplicación de las matrices y sus tamaños para obtener la matriz $S_w^{-1}$. Para esto se multiplica la matriz de vectores propios, de orden $k \times d$, con la inversa de la matriz de valores propios de orden $d \times d$, y todo esto multiplicado con la matriz transpuesta de los vectores propios de orden $d \times k$ (Ecuación 4.3 y 4.4),

$$\left(\phi_{(k\times d)}\Lambda^{-1}_{(d\times d)}\right)\phi^{t}_{(d\times k)} \tag{4.3}$$

$$((k\times d)(d\times d))(d\times k) = (k\times d)(d\times k) = (k\times k) \tag{4.4}$$

la matriz resultante será de tamaño $k\times d$ y esta se multiplica con la matriz de vectores propios transpuesta. Lo que entrega como matriz final es $S^{-1}_{w_{k\times k}}$ .

Con estas matrices se puede obtener el valor de FOC. Las fórmulas se debieron reducir a operaciones simples, para evitar que los resultados de las diferentes operaciones entregarán como resultado matrices singulares o de valores infinitos.

Para obtener el valor de FOC, de una combinación de las características, obtenidas de la búsqueda secuencial (en una de las iteraciones), se toma el valor máximo anterior obtenido y se compara dejando como criterio el valor máximo, lo que puede volver a cambiar cuando se haya agregado otra característica a la búsqueda. Finalmente se obtienen las características ordenadas de acuerdo al criterio FOC, es decir, un ranking de las características más influyentes en la interacción. Esta misma selección se realiza con las 642 características de la matriz, obteniendo 642 características ordenadas de acuerdo al criterio definido (valores máximos para FOC).

### 4.3.2. Análisis de Componentes principales

El Análisis de Componentes Principales - ACP es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. También se usa para determinar el número de factores subyacentes explicativos tras un conjunto de datos.

El ACP construye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual recaptura la varianza de mayor tamaño del conjunto de datos en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande en el segundo eje, y así sucesivamente. Para construir esta transformación lineal debe

construirse primero la matriz de covarianza o matriz de coeficientes de correlación. Debido a la simetría de esta matriz existe una base completa de vectores propios de la misma. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos. Además las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales.

Este método es una técnica clásica del reconocimiento estadístico de patrones [31]. El objetivo principal de ACP es la representación de las medidas numéricas de varias variables en un espacio de pocas dimensiones donde los sentidos puedan percibir relaciones que de otra manera permanecerían ocultas en dimensiones superiores. Dicha representación debe ser tal que al desechar dimensiones superiores (generalmente de la tercera o cuarta en adelante) la pérdida de información sea mínima. Si el valor de un componente puede ser predecido con otros valores, es claramente redundante y se puede desechar. Si se va a desechar un componente, la mejor a desechar será aquella que sea mejor predecida por las demás.

ACP entrega un conjunto de componentes lineales de una dimensión particular que representan lo mejor posible las varianzas de un conjunto de datos de mayor dimensión. Sin embargo, no existe garantía de que este conjunto de características sea bueno para clasificación debido a que una proyección puede suprimir detalles importantes (pequeñas varianzas de dirección pueden ser importantes); el método no toma en cuenta tareas de discriminación (típicamente se desea calcular características que permitan una buena discriminación, lo cual, no es lo mismo que varianzas grandes); o no se toma en cuenta que puede haber más de una clase en un conjunto de datos.

Al realizar esta reducción de dimensionalidad se pierde cierta cantidad de información, además se pierden las distancias perpendiculares a los ejes de coordenadas. Sin embargo, la pérdida de información se ve ampliamente compensada con la simplificación realizada, ya que muchas relaciones, como la vecindad entre puntos, es más evidente cuando éstos se dibujan sobre un plano que cuando se hace mediante una figura tridimensional que necesariamente debe ser dibujada en perspectiva. Además retiene aquellas características del conjunto de datos que contribuyen más a su varianza, manteniendo un orden de bajo nivel de los componentes principales e ignorando los de alto nivel. El objetivo es que esos componentes de bajo orden a veces contienen el más importante aspecto de esa información.

Por ejemplo, si existe una muestra con $n$ individuos para cada uno de los cuales se han medido $m$ variables (aleatorias) $F_j$. El ACP permite encontrar un número de factores subyacentes $p < m$ que explican aproximadamente el valor de las $m$ variables para cada individuo. El

hecho de que existan $p$ factores subyacentes puede interpretarse como una reducción de la dimensionalidad de los datos: donde antes se necesitaban $m$ valores para caracterizar a cada individuo ahora nos bastan $p$ valores. Cada uno de los $p$ encontrados se llama componente principal, de ahí el nombre del método.

PCA se utiliza para evitar que hayan matrices singulares en algunos métodos, y para ello se hace un preprocesamiento para evitar valores propios cercanos a cero, o bien matrices cercanas a singular lo que causa problemas computacionales en cálculos de inversa y logaritmos. Además se utiliza con SVM, debido a que intenta aumentar la eficiencia usando menos componentes o características de entrada en la SVM que mapea los datos a una dimension mayor. En este caso se esta utilizando ACP solo para la clasificación con todas las características.

## 4.4. Problemas para la Representación de los Datos en el Espacio Multidimensional

Puede ser compleja la convergencia de los algoritmos de entrenamiento del clasificador, o pueden existir componentes de alta dimensión que perturban el aprendizaje cuando hay pocos datos para el entrenamiento, problema conocido como overfitting [31]. Para resolver estos inconvenientes existen técnicas de reducción de dimensiones de distintas formas, siendo las lineales más utilizadas debido a su eficiencia y simplicidad. Las técnicas de reducción lineal de dimensiones (RLD) incluyen el ACP y análisis de componentes independientes (ICA) [90]. El objetivo es poder representar las medidas numéricas de varias variables en un espacio de pocas dimensiones donde se puedan percibir relaciones que de otra manera permanecerían ocultas en dimensiones superiores.

Una limitación que tienen PCA e ICA es que no consideran la discriminabilidad de los datos, debido a que pasan a ser un grupo representativo de los datos, por lo tanto, no son la mejor solución para problemas de clasificación supervisada, salvo que no pueda aplicarse otro tipo de técnica. Entre las técnicas lineales que consideran la discriminabilidad de los datos, los más conocidos son:

- el criterio de Fisher (Homocedástico) [31],
- el criterio Loog Duin (Heteroscedástico) [64, 46], y
- aquel que maximiza la distancia de Chernoff en el espacio reducido [80].

El objetivo de estas técnicas de reducción de dimensiones es transformar los datos para su posterior clasificación, siendo esta realizada por alguno de los algoritmos mencionados anteriormente: Bayes, k-NN, MVS, etc.

## 4.5. Métodos de Clasificación Utilizados

Con los valores obtenidos de la selección de características, se procede a aplicar diferentes clasificadores que permitan evaluar si la selección entrega una buena separabilidad de las clases. En esta etapa se desarrollan dos procesos paralelos, aplicación de reducción lineal de dimensiones (RLD) y Máquina de Vectores Soporte (MVS) para posteriormente efectuar comparaciones de los resultados obtenidos.

Se utiliza reducción lineal de dimensiones para las características obtenidas de FOC, en tres métodos: Heteroscedástico (Loog-Duin), Chernoff (Rueda-Herrera) y Homocedástico (Fisher). Para la aplicación de los componentes obtenidos del análisis de componentes principales se utiliza sólo el método de RLD Heteroscedástico. Finalmente a los datos obtenidos de los métodos, se aplican los clasificadores lineal y cuadrático. La etapa termina con la evaluación de los resultados (RLD y MVS) y una posterior interpretación de ellos.

### 4.5.1. Reducción Lineal de Dimensiones

El método de reducción lineal de dimensiones (RLD) busca la transformación lineal que permite reducir la dimensión de un modelo estadístico de $n-$dimensiones a uno de $d$-dimensiones ($d < n$), preservando la maximalidad de la información discriminatoria para varias clases dentro de un modelo. En este trabajo se consideran dos clases $\omega_1$ y $\omega_2$, donde se conocen las probabilidades a priori $p_1$ y $p_2$, representadas por dos distribuciones normales en vectores aleatorios de $n-dimensiones$. Para esto se debe encontrar una matriz de dimensión $d \times n$ llamada $A$ y los datos linealmente transformados $y = Ax$, para posteriormente ser clasificados.Los valores de las submatrices, depende de la cantidad de elementos que posea cada clase y de las características que se este utilizando.

#### 4.5.1.1. RLD Homocedástico – Criterio de Fisher

Para el caso de dos clases, $\omega_1$ y $\omega_2$, las mismas son representadas por dos distribuciones normales en vectores aleatorios de $n-dimensiones$, $x_1 \sim N(m_1, S_1)$ y $x_2 \sim N(m_2, S_2)$, y las

probabilidades a priori son $p1$ y $p2$ respectivamente. Para ayudar a la transformación lineal se tienen $x_1$ y $x_2$ en dos nuevos vectores aleatorios distribuidos normalmente $y_1$ e $y_2$ de dimensión $d$, $d < n$, utilizando una matriz $A$ de orden $d \times n$, de tal forma que el error de clasificación en el espacio transformado es el menor posible. Sean las matrices de dispersión $S_w = p_1 S_1 + p_2 S_2$ y $S_E = (m_1 - m_2)(m_1 - m_2)^t$. La finalidad es maximizar las distancias de las distribuciones transformadas. Encontrando una matriz $A$ que maximiza la Ecuación 4.5.

$$J_{FD}(A) = tr\{(AS_wA^t)^{-1}(AS_EA^t)\} \tag{4.5}$$

La matriz $A$ que maximiza la función anterior, se obtiene buscando los valores propios de la descomposición de la matriz, se muestra en la Ecuación 4.6.

$$S_{FD} = S_w^{-1} S_E \tag{4.6}$$

y tomando los $d$ (dimensiones) vectores propios cuyo valores propios son los más grandes. Debido a que la descomposición de los valores propios de la matriz anterior lleva a un solo valor propio distinto de 0, entonces $(m_1 - m_2), (m_1 - m_2)^t$, cuyo vector propio esta dado por $m_1, m_2$, se puede reducir a solo una dimensión [64].

#### 4.5.1.2. RLD basada en Distancia de Chernoff - Rueda-Herrera

Se propone en este método maximizar la separabilidad de la distribuciones en un espacio transformado. Se asume que la distribución original es normal y la distribución en el espacio transformado también es normal. Para el caso de dos clases, se maximiza la distancia de Chernoff entre los vectores randómicos transformados [81] (Ecuación 4.7):

$$k(\beta, A) = \frac{\beta(\beta - 1)}{2}(Am_1 - Am_2)^t \left[\beta AS_1A + (1-\beta)\beta AS_2A\right]^{-1}(Am_1 - Am_2) +$$

$$\frac{1}{2}\log\frac{|\,\beta AS_1A+(1-\beta)\,\beta AS_2A\,|}{|\,AS_1A\,|^{\beta}|\,AS_2A\,|^{1-\beta}} \tag{4.7}$$

Después de la transformación, se obtienen los nuevos vectores de la forma $y_1 \sim N(Am_1; AS_1A^t)$ e $y_2 \sim N(Am_2; AS_2A^t)$, que encuentran la matriz $A$ que maximiza (Ecuación 4.8):

$$J^*_{c12}(A) = tr\left\{p_1p_2(AS_wA^t)^{-1}(AS_EA^t)+\log\left(AS_wA^t\right)-p_1\log\left(AS_1A^t\right)-p_2\log\left(AS_2A^t\right)\right\} \tag{4.8}$$

Para maximizar $J^*_{c12}(A)$, los autores en Rueda-Herrera [81], proponen un algoritmo basado en el método gradiente, por lo que es necesario encontrar la matriz gradiente utilizando el operador, como se muestra en la Ecuación 4.9.

$$\nabla J^*_{c12}(A) = \frac{\partial J^*_{c12}}{\partial A} = 2p_1p_2\left[S_EA^t\left(AS_wA^t\right)^{-1}-S_wA^t\left(AS_wA^t\right)^{-1}\left(AS_EA^t\right)\left(AS_wA^t\right)\right]^t$$

$$+2\left[S_wA^t\left(AS_wA^t\right)^{-1}-p_1S_1A^t\left(AS_1A^t\right)^{-1}-p_2S_2A^t\left(AS_2A^t\right)^{-1}\right]^t \tag{4.9}$$

Para la obtención del logaritmo, se define la función $f$ como una función que se aplica a la matriz simétrica positiva definida por $A$ (dicha función puede ser la potencia o el logaritmo). Por medio de la descomposición entre sus valores entre $RVR^{-1}$con la matriz de valores entre $V = diag\,(v_1,...,v_n)$. Por lo tanto, sea $f(A)$es igual a $Rdiag\,(v_1,...,v_n)\,R^{-1} = R(f\,(V))\,R^{-1}$. Aunque generalmente, $A$ es no singular, determinar $f\,(A)$podría causar problemas numéricos si la matriz es cercana a singular. Se puede aliviar este problema computacional utilizando la descomposición de valor singular (svd) en vez de una descomposición de los valores entre o regularizando apropiadamente $A$ [64].

#### 4.5.1.3. RLD Heterocedástico – Loog-Duin

Para el caso de dos clases estas se encuentran distribuidas normalmente y se toma en consideración la distancia de Chernoff en el espacio original para minimizar la tasa de error en el espacio transformado. Se utiliza la matriz de distribución dirigida con la transformación lineal en el espacio original, para finalizar generalizando el criterio de Fisher en el espacio transformado, sustituyendo la matriz de dispersión dentro de la clase por la correspondiente matriz de distancia dirigida [81]. El criterio de Loog-Duin consiste en obtener la matriz $A$ que maximiza la función (Ecuación 4.10):

$$J_{LD2}(A) = tr\left\{ \left(AS_wA^t\right)^{-1} \left[ AS_EA^t - AS_w^{\frac{1}{2}} \frac{p_1 log\left(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}\right) + p_2 log\left(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}}\right)}{p_1 p_2} S_w^{\frac{1}{2}} A^t \right] \right\} \tag{4.10}$$

La solución de este criterio esta dado por la matriz $A$ que esta compuesta por los $d$-vectores propios (cuyo valores propios son los máximos) de la siguiente matriz (Ecuación 4.11).

$$S_{LD2} = S_w^{-1} \left[ S_E - S_w^{\frac{1}{2}} \frac{p_1 log\left(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}\right) + p_2 log\left(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}}\right)}{p_1 p_2} S_w^{\frac{1}{2}} \right] \tag{4.11}$$

### 4.5.2. Máquinas de Vectores Soporte

Las máquinas de vectores soporte (SVM por sus siglas en ingles, "Support Vector Machine"), fueron desarrolladas para el problema de clasificación. Algunas de las razones por las que este método ha tenido éxito es que no padece de mínimos locales y el modelo sólo depende de los datos con más información llamados vectores de soporte. MVS es un sistema para entrenar máquinas de aprendizaje lineal eficientemente tanto que para clasificación como para regresión. Las grandes ventajas que tiene MVS son la excelente capacidad de generalización,

debido a la minimización del riesgo estructurado y la estimación de los parámetros se realiza a través de la optimización de una función de costo convexa, lo cual evita la existencia de mínimos locales.

Para el aprendizaje no lineal con MVS se obtiene mediante el uso de las denominadas funciones kernel que permiten transformar el espacio de atributos de entrada en un espacio de trabajo de dimensionalidad mucho mayor. El planteamiento dual de las MVS permite trabajar con estas funciones kernel de manera eficiente sin calcular explícitamente la representación de los vectores en el espacio de trabajo. Entre los posibles hiperplanos separadores las MVS eligen el margen máximo. Este sesgo inductivo de aprendizaje complementa al habitual de minimizar el error en el conjunto de aprendizaje que contribuye a minimizar el riesgo sobre aprendizaje (overfitting) [41, 85].

Se desea construir un hiperplano que separe las dos clases, etiquetadas $y \in \{-1,+1\}$ de forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano -margen- sea máxima, con la intención de forzar la generalización de la máquina de aprendizaje.

La expansión del método MVS a funciones de decisión no lineales se realiza introduciendo el espacio de entrada $\chi \subseteq \mathbb{R}^d$ en otro espacio de mayor dimensión $\mathcal{F}$, denominado espacio de características, dotado de producto interno, vía una inyección no lineal, $\iota : \chi \subseteq R^d \rightarrow \mathcal{F}_{(.,.)}$, de forma que el hiperplano óptimo $f(x,\omega) = \langle \omega, x \rangle_{\mathcal{F}} + b = k(\omega,x) + b$, hallado linealmente en el espacio de características $\mathcal{F}$, con $k(a,b) = \langle a,b \rangle_{\mathcal{F}} = (i(a) \centerdot i(b))$, correspondiendo a un núcleo de Hilbert[1], permite definir como función decisión $h(x) = sign(f(x,\omega))$

El aprendizaje no lineal con MVS se consigue mediante el uso de las denominadas funciones kernel que permiten transformar el espacio de atributos de entrada en un espacio de trabajo de dimensionalidad mucho mayor (lo cual aumenta la capacidad computacional de la máquinas de aprendizaje lineal). Es por esto que las MVS pueden considerarse una caso particular de las algoritmos basados en kernel. Algunos de los kernel más nombrados son: Polinómico, red neural, lineal, base radial y sigmóideo.

En el caso de este proyecto se opto por trabajar con kernels polinomiales, pero solo con grado 2 y 3, debido a que con grados mayores la dimensión del problema daría una expansión a una $d$ muy alta.

El kernel utilizado es polinomial, se optó por éste debido a que entregó buenos resultados

[1]Cualquier función simétrica continua $k(a,b)$puede ser usada como un núcleo de Hilbert si satisface la condición de Mercer $\int\int k(a,b)g(a)g(b)\,da\,db \geq 0, \forall g$.

y además se utilizó los clasificadores lineal y cuadrático. Justamente, por un lado se tiene la relacion que existe entre el kernel utilizado en SVM, los otros métodos de RLD proponen una transformación lineal y además se tiene la aplicación de un clasificador lineal o cuadrático. Cabe destacar que habría sido interesante hacer uso de base radial, pero de acuerdo a las razones expresadas anteriormente no se utilizó, pero se propone como un trabajo futuro a realizar.

# Capítulo 5

# Experimentos, Resultados e Interpretaciones

En el presente capítulo se describe el proceso efectuado para el cumplimiento de las tres etapas definidas en el Capítulo anterior. Se detalla cada uno de los pasos realizados y sus respectivos resultados por cada una de las etapas. Además se propone una interpretación biológica de estos resultados.

En la Figura 5.1, se muestra en detalle cada etapa.

1. Preparación de los Datos, se procede a la recuperación de información sobre IPP. Luego se determina los datos a utilizar y sus posterior obtención, para la creación de la base de datos. Esta base de datos se toma en la segunda etapa.

2. Selección, se procede a obtener un factor discriminante que influya en la IPP, ya sea por la cantidad, presencia o ausencia de una o varias de estas características. Al tener estos discriminantes, se plantea identificar un subconjunto de características relevantes que ayuden a predecir comportamientos de estas interacciones [17, 29, 42, 69, 77, 97]. En ese caso, se ejecuto dos métodos diferentes, ACP y búsqueda secuencia con la distancia de Chernoff. Ambos métodos generaron dos listados uno de componentes y otro de características, que se utilizan en la tercera etapa.

3. Evaluación, se procede a realizar el trabajo de reconocimiento de patrones. Para este paso se deben seguir diferentes esquemas y técnicas de aprendizaje supervisado, las cuales permite identificar que características aportan a la separabilidad de las clases y

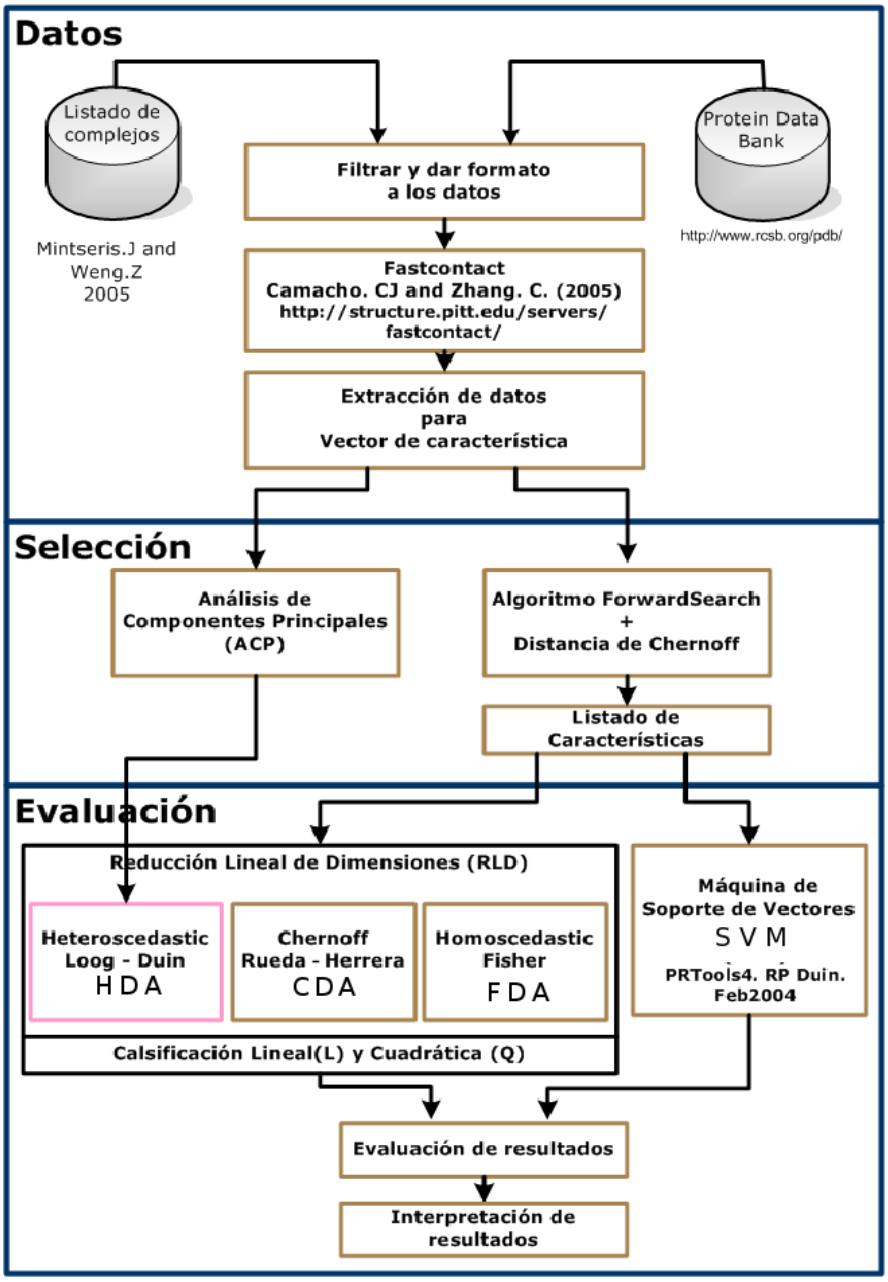



Figura 5.1: Metodología a seguir.

cuales no. Todo esto de acuerdo a la precisión de los resultados obtenidos del método propuesto y la evaluación posterior [45]. Aquí se aplican cuatro métodos a el listado de características (tres de RLD que son HDA, CDA, FDA y MVS), en cambio al listado de componentes solo se aplica a un método (HDA). Los métodos de RLD, se utilizan clasificador lineal y cuadrático, en cambio en MVS se utiliza kernel polinomial de grado 2 y 3. Todos estos resultados se someten a evaluación, para una posterior evaluación.

Para ejecutar las diferentes etapas, se utilizó un equipo del Laboratorio de Inteligencia Artificial[1] que cuentan un procesador Pentium 4, de 2.66GHz, con 1GB de memoria RAM con sistema operativo Windows XP. El proceso de selección de características, clasificación y evaluación se realizó con la aplicación Matlab.

## 5.1. Formación de la Base de datos de complejos

Se han efectuado numerosas investigaciones evaluando diferentes propiedades de complejos proteína-proteína [3, 12, 49, 66, 93], para ello se han utilizado las estructuras atómicas de complejos que se encuentran disponibles en el Protein Data Bank.

Basados en un trabajo donde se estudia las propiedades de las superficies de los complejos proteína-proteína tendiente a su caracterización de acuerdo a si el complejo es transitorio o permanente, realizado por Mintseris y Weng [67], donde utilizan un listado de complejos, que ya se encuentra clasificado como IPP transitorios y IPP permanentes.

La clasificación fue realizada de forma manual por el grupo de trabajo de Mintseris, de acuerdo a una de las clasificaciones de complejos de interacción proteína-proteína definidos por Thornton y sus colaboradores en 2003 [70], sobre complejos de estructura tridimensional conocida.

El listado se compone de 326 complejos, los cuales ya se encuentran clasificado como IPP transitorios y permanentes. El detalle de composición es la siguiente:

- Transitorios: 211 complejos.
- Permanentes: 115 complejos.

[1] Departamento de Ingeniería Informática. Facultad de Ingeniería. Universidad de Concepción.

A partir de esta lista, se obtuvo la información estructural de cada complejo (con un formato definido por el PDB), desde el PDB. De este grupo de 326 complejos, se toma una muestra de 296 complejos, clasificados en dos clases.

- Clase 1: IPP permanentes con 93 complejos ($\omega_1$), y
- Clase 2: IPP transitorias con 203 complejos ($\omega_2$).

Al tener el tamaño de la muestra y las clases definidas, se tienen conocimiento de las probabilidades de cada clase ( $p_1$ y $p_2$). Los 30 complejos restantes, que no son utilizados para las tres etapas, se dejan apartados para su posterior uso en la etapa de interpretación de los resultados desde el punto de vista biológico.

Para depurar los datos obtenidos desde el PDB, se revisó cada uno de los complejo para eliminar aquellos residuos duplicados, es decir, residuos modelados en dos conformaciones, donde se eliminó la conformación que poseía el menor valor de ocupancia y la otra conformación se consideró como ocupancia 1.00. De acuerdo a las necesidades de las diferentes aplicaciones utilizadas se modificó el archivo de datos del complejo. Para utilizar la aplicación FastContact, se necesitan los archivos de cada una de las cadenas que interaccionan por separado.

Para aquellos complejos en que la zona de interacción se forma con la participación de más de dos cadenas, la que aporta el mayor número de residuos se consideró la cadena 1 y las otras dos cadenas como la cadena 2. Para esto se escribió una aplicación en Perl[2] para separar y renombrar las cadenas. Cada complejo puede estar formado por varias cadenas, las cuales pueden pertenecer a una u otra proteína. Por ejemplo, si existe un complejo TT que lo conforman 5 cadenas (A, B, C, D y E), las cadenas ACD corresponden a la proteína 1 y BE a la proteína 2. Dado que en nuestro trabajo las cadenas relevantes son aquellas que interaccionan, sólo estas se analizaron. En el ejemplo anterior correspondería a las características de las cadenas A y B que son las que representan a las proteínas 1 y 2 respectivamente.

### 5.1.1. Aplicaciones para el estudio de interacción proteína-proteína

Existen variadas aplicaciones desarrolladas para trabajar con interacción entre proteínas específicamente con la caracterización de la superficies de interacción. Estas aplicaciones han

[2]Practical extraction and report language

sido el resultado de diferentes investigaciones y trabajos, donde se intenta estudiar las diferentes propiedades de las proteínas, como por ejemplo: peso, tamaño, forma, cadenas, energías entre otras [83].

Dado que la mayoría de los estudios han considerado propiedades de los residuos que forman la superficie o propiedades de la superficie, pero no han considerado propiedades energéticas de la interacción, en el presente trabajo se utilizan las propiedades energéticas como discriminantes de clasificación. Para obtener la propiedades energéticas de la superficie de interacción se utilizó el programa Fast Contact [21].

### 5.1.2. FastContact

Esta aplicación fue creada por Carlos Camacho y su Laboratorio[3] [21], la cual trabaja con las energías de la interface de interacción del complejo. Es una aplicación que entrega: para cada cadena que interacciona, los 20 residuos y los valores de las energías que más contribuyen en la interacción; y los 20 residuos y los valores de las energías que menos contribuyen en la interacción considerando energías de desolvatación y energía electrostática. Se evalúa además:

- la energía total electrostática,
- energía libre de desolvatación total,
- la energía libre de unión del complejo,

para pares de residuos que interaccionan:

- la energía electrostática y
- la energía libre de unión.

Es decir, para cada complejo entrega 1004 datos que serán considerados en el presente estudio. Se utiliza una función de puntuación de energía libre, es decir, se evalúan las energías que aportan los residuos del área de interacción, y se consideran sólo los residuos que pertenecen a la zona de interacción (Anexo F). Los 20 residuos que más contribuyen y los 20 que menos

[3] Associate Professor Department of Computational Biology Department of Molecular Genetics and Biochemistry University of Pittsburgh

contribuyen a cada una de los componentes de la energía libre de unión que se estima en la Ecuación 5.1.

$$\triangle G_{bind} = \triangle E_{elec} + \triangle G_{des} \tag{5.1}$$

donde, $\triangle E_{elec}$ corresponde al potencial electrostático Coulómbico de la dependencia de la constante dieléctrica con la distancia de $4r$ [76].

$\triangle G_{des}$corresponde a la energía libre de desolvatación que capta la mayoría de las características esenciales de la energía libre de desolvatación en proteínas, incluyendo las interacciones hidrofóbicas, el cambio de energía libre de grupos cargados y de átomos polares, y la pérdida de entropía en la cadena laterales. $G_{des}$ se calcula por un potencial de contacto empírico de la forma expresada en la Ecuación 5.2:

$$\triangle G_{des} = g(r) \sum \sum e_{ij} \tag{5.2}$$

donde $e_{ij}$ es el potencial atómico de contacto entre los átomos del receptor $i$ y los átomos del ligando $j$. La doble suma es tomada de todos los pares de átomos y $g(r)$ es cero para átomos que están más allá de 7Å y 1 para aquellos que están a menos de 5Å y entre estos dos valores hay una función suave que varía entre estos dos límites [96].

El programa FastContact utiliza las coordenadas atómicas de las proteínas considerando una, la primera ingresada como receptor y la segunda como ligando. Los archivos son modificados por la introducción de los átomos de hidrógenos de átomos polares y las cargas parciales de los átomos de acuerdo al tipo de átomo definidos en CHARMM 19 (archivo que define el formato para cada aminoácido, es decir se identifican los átomos que lo compone, todo esto es parte de la aplicación).

También se deben mantener los formatos de carácter químico que solicita la aplicación, estos se encuentran especificados en un archivo anexo al programa. Los resultados que entrega

FastContact son 4 archivos por consulta (por complejo), de los cuales uno es relevante para el proyecto, (la información que entrega se encuentra en detalle en el anexo F.

De acuerdo a los datos obtenidos en el archivo de salida (Inicialmente son más de 1000 datos) se creó una base de datos, para luego depurarla y trabajar con los datos más importantes. Los datos que se retiran, se debe a que son caracteres (texto) y en este caso no son de utilidad para trabajar.

En la Tabla 5.1, se presentan las estimaciones de la cantidad de características a trabajar.

| Cadena | Energía | min/max | A | B | C | | Totales |
|---|---|---|---|---|---|---|---|
| | Energía electroestática | | | | | | 1 |
| | Energía desolvatación | | | | | | 1 |
| | Energía libre de Unión | min | 20 | 40 | x 2 (caract) = 80 | | 80 |
| | | max | 20 | | | | |
| Ligando | Energía desolvatación | min | 20 | 40 | x 2 (caract) = 80 | + | 160 |
| | | max | 20 | | | | |
| | Energía electroestática | min | 20 | 40 | x 2 (caract) = 80 | | |
| | | max | 20 | | | | |
| Receptor | Energía desolvatación | min | 20 | 40 | x 2 (caract) = 80 | + | 160 |
| | | max | 20 | | | | |
| | Energía electroestática | min | 20 | 40 | x 2 (caract) = 80 | | |
| | | max | 20 | | | | |
| Ligando-Receptor | Energía desolvatación | min | 20 | 40 | x 3 (caract) = 120 | + | 240 |
| | | max | 20 | | | | |
| | Energía electroestática | min | 20 | 40 | x 3 (caract) = 120 | | |
| | | max | 20 | | | | |
| **Clases** | | | | | | | 1 |
| **Total de características a utilizar** | | | | | | | 643 |

Tabla 5.1: Distribución de características a utilizar en el proyecto.

Cada valor min y max, entrega 20 características (A), que en conjunto son 40 (B). Además por cada una de estas 40 características existen 2 ó 3 características asociadas (C), lo que entrega el total de características por tipo de energía (D). Finalmente en la columna E se entrega el total de características obtenidas por cadena y por complejo.

Las 2 ó 3 características asociadas significa que se toman dos valores, uno es el valor de la energía, y otro valor que corresponde al número de aminoácido en la secuencia de la proteína, este corresponde a una característica que podría considerarse no importante, pero con este valor se puede rescatar posteriormente el tipo de aminoácido que si es importante para el área biológica.

En total son 642 características por cada uno de los 296 complejos, además se debe agregar la característica tipo de complejo. Se implementaron diferentes scripts Perl, que permitieron reunir y ordenar las características (Anexo G).

### 5.1.3. Preparación de los datos

Esta sección se finaliza la primera etapa de la metodología, en la cual se logra formar una matriz que contiene las características necesarias para realizar la etapa 2, que corresponde a la selección de estas.

En la Figura 5.2, se muestran los pasos que se realizaron:

- obtención del listado de complejos,
- recuperación de información desde servidores biológicos,
- filtrar y dar formato a los datos,
- ejecución de la aplicación Fast Contact y
- la extracción de datos para la creación de la matriz de trabajo.

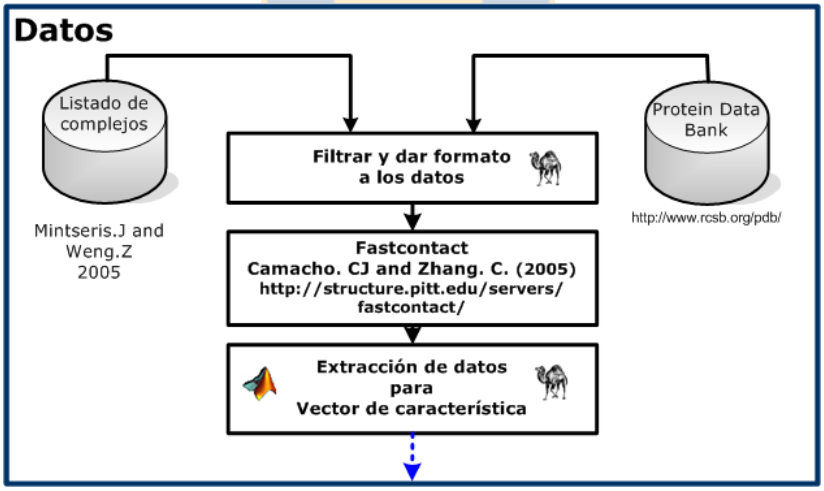



Figura 5.2: Colección de datos.

## 5.2. Selección de características energéticas.

La matriz de trabajo tiene 296 complejos con 642 características. Para la recopilación de la información de los complejos se utilizó lenguaje Shell y Perl, esto permitió un fácil manejo de los datos y un diseño para el formato a utilizar.

El algoritmo de búsqueda Forward search (secuencial) definido en la sección 4.3.1 (Algoritmos para la Selección de Características) fue implementado sin memoria, es decir, solo se fue almacenando la mejor característica de cada iteración y no las dos primeras. El criterio de selección a utilizar es el propuesto por Rueda-Herrera [81]. Se creo el algoritmo que obtiene el FOC, el cual permite seleccionar las características de acuerdo a su grado de importancia. Para esto se trabaja con los siguientes valores: la Matriz original,

$$Matriz_{296\times 643} = \begin{bmatrix} \omega_1 & x_1, x_2, \ldots x_{642} \\ \omega_1 & x_1, x_2, \ldots x_{642} \\ \ldots & \ldots \\ \omega_2 & x_1, x_2, \ldots x_{642} \\ \omega_2 & x_1, x_2, \ldots x_{642} \end{bmatrix},$$

la clase 1 esta referenciada a los complejos permanentes con $\omega_1 = 93$ y la clase 2 a los complejos transitorios con $\omega_2 = 203$. Se genera una matriz para cada clase:

$$X_{1_{(93\times d)}} = \begin{bmatrix} \omega_1 & x_1, \ldots x_d \\ \omega_1 & x_1, \ldots x_d \\ \ldots & \ldots \\ \omega_1 & x_1, \ldots x_d \end{bmatrix}$$

$$X_{2_{(203\times d)}} = \begin{bmatrix} \omega_2 & x_1, \ldots x_d \\ \omega_2 & x_1, \ldots x_d \\ \ldots & \ldots \\ \omega_2 & x_1, \ldots x_d \end{bmatrix}$$

El valor de la dimensión ($d$) depende de la iteración en que se encuentre el algoritmo de búsqueda y las características ya seleccionadas. Las probabilidades *apriori*: $p_1 = \frac{n_1}{n} = 0,3142$ y $p_2 = \frac{n_2}{n} = 0,6858$. Dado que $n = 296$, $n_1 = 93$, $n_2 = 203$. Además se debe indicar que para evitar situaciones de matrices singulares, debido a que los componentes de la matriz de

valores propios presenta ceros en la diagonal, se estableció un umbral de $1{,}0 \times 10^{-6}$, el cual permite evitar la singularidad y valores infinitos en el proceso se selección. De acuerdo a la ubicación del umbral en la matriz de valores propios, se establece el valor de $k$, para generar la inversa de la matriz de valores propios $S^{-1}_{w_{k\times k}}$ . Con estas matrices se puede obtener el valor de FOC.

De esto se obtuvo un listado de características ordenadas de mayor a menor, por su factor de relevancia para discriminar entre las clases. De los diferentes procesos ejecutados con diferentes dimensiones, se generaron submatrices que permiten una mejor visualización de los resultados entre los métodos. A continuación en la Tabla 5.2 se presentan las primeras 20 características más discriminantes para la separabilidad de las clases, utilizando la distancia de Chernoff, en conjunto con la búsqueda secuencial hacia delante. El resto de las características se encuentran en el Anexo G.2.

| Posición | Característica | Posición | Característica | Posición | Característica | Posición | Característica |
|---|---|---|---|---|---|---|---|
| 1 | 121 | 6 | 314 | 11 | 342 | 16 | 348 |
| 2 | 282 | 7 | 346 | 12 | 330 | 17 | 626 |
| 3 | 100 | 8 | 407 | 13 | 360 | 18 | 326 |
| 4 | 470 | 9 | 288 | 14 | 290 | 19 | 292 |
| 5 | 354 | 10 | 298 | 15 | 557 | 20 | 338 |

Tabla 5.2: Listado de las primeras 20 características de la selección.

Los tamaños de las submatrices generadas son:

- matriz de $296 \times 642$ (Total de características obtenidas en FOC).
- matriz de $296 \times 20$ (20 características más discriminantes),
- matriz de $296 \times 75$ (75 características más discriminantes) y
- matriz de $296 \times 277$ (El tiempo necesario para la ejecución de RLD es extenso -con las 642 características obtenidas por FOC- por esta razón se detuvo el proceso a los 25 días de iniciado, con lo cual se obtuvo la clasificación de sólo 277 características).

Después de observar los resultados, se procedió a analizar las características de acuerdo a su aparición dentro de las 20 más influyentes en la separabilidad de las clases. Se presenta la Tabla 5.3 que recopila esta información.

| Cadena | Energía | min/max | Frecuencia de Aparición | | |
|---|---|---|---|---|---|
| | Energía electroestática | | | | 0 |
| | Energía desolvatación | | | | 0 |
| | Energía libre de Unión | min | 0 | 0 | 0 |
| | | max | 0 | | |
| Ligando | Energía desolvatación | min | 2 | 2 | 2 |
| | | max | 0 | | |
| | Energía electroestática | min | 0 | 0 | |
| | | max | 0 | | |
| Receptor | Energía desolvatación | min | 1 | 6 | 14 |
| | | max | 5 | | |
| | Energía electroestática | min | 8 | 8 | |
| | | max | 0 | | |
| Ligando-Receptor | Energía desolvatación | min | 1 | 2 | 4 |
| | | max | 1 | | |
| | Energía electroestática | min | 1 | 2 | |
| | | max | 1 | | |

| | | |
|---|---|---|
| Energía libre de desolvatación | 2 + 6 | = 8 |
| Energía electroestática | 0 + 8 | = 8 |

Tabla 5.3: Distribución de las primeras 20 características del proceso de selección.

Se puede observar de manera detallada aquellas energías relevantes para la clasificación de las clases, como son las cadenas que conforman al complejo y los aminoácidos que las componen. El análisis de las mejores características discriminantes muestra que la energía electrostática contribuye tanto como la energía de desolvatación a la separación de clases. El análisis más detallado muestra que en el caso de la energía electrostática los aminoácidos que menos (min) contribuyen son los discriminantes y a el caso de la energía de desolvatación son las que más (max) contribuyen a los discriminantes.

## 5.3. Análisis de componentes principales

Paralelamente se efectuó una reducción de dimensiones con análisis de componentes principales que permite tener datos referenciales, para los resultados obtenidos con el método anterior. Estos componentes ya no son comparables con los resultados obtenidos anteriormente, ya que un componente principal representa a 1 o varias características, pero no es una característica, por esto se dice que sirve como marco referencial.

Para la aplicación del método se utilizó la matriz original ($296 \times 642$). El primer paso que se efectuó fue eliminar las características que no entregan valores relevantes para la obtención de

componentes representativos. Se utilizó como umbral de selección todos los valores mayores de $1{,}0 \times 10^{-6}$, que se encuentran en la matriz de valores propios. Para todo esto se utilizó código ya creado por L. Rueda.

Como resultado se obtienen 294 componentes representativos de las 642 características. Este valor corresponde al número total de valores propios, menos los valores propios que que tienen un valor por debajo del umbral definido. También se crearon submatrices (20, 75, 277 componentes) para tener un marco referencial con respecto al método propuesto FOC.

Debido a que se identificaron 294 componentes representativos, se decidió crear una matriz adicional a las ya definidas (sección 6.1.2.2.) con el método FOC, con este nuevo valor (matriz de $296 \times 294$). Aunque no permite realizar comparaciones, es para mantener el marco referencial propuesto al inicio.

A cada una de las submatrices definidas se realizó validación cruzada de 10 pliegues, para la posterior utilización del método de RLD Loog-Duin con dos clasificaciones: lineal y cuadrática.

| **Loog-Duin (PCA)** | **Componentes** | | | |
|---|---|---|---|---|
| Clasificador | 20 comp. | 75 comp. | 277 comp. | 294 comp. |
| Cuadrático | **25,9 %** | 29,0 % | 32,5 % | 32,5 % |
| Lineal | **25,3 %** | 26,0 % | 28,5 % | 28,8 % |

Tabla 5.4: Porcentajes de error al aplicar RLD con el método de Loog-Duin.

En la Tabla 5.4, se observan los errores obtenidos de la aplicación de RLD con el método de Loog-Duin (Heterocedástico) para las diferentes submatrices de componentes obtenidos con ACP. Desde el punto de vista de las submatrices, la de 20 componentes son las que obtienen los menores porcentajes de error, con ambos clasificadores.

Al comparar los resultados obtenidos por cada clasificador, se puede destacar que el clasificador lineal obtiene de las cuatro agrupaciones los menores porcentajes de error, con respecto a los obtenidos con el clasificador cuadrático.

Aunque no sean comparativos, permite tener un resultado de referencia para el método que se propone. Este desarrollo paralelo se puede observar en la Figura 5.3.

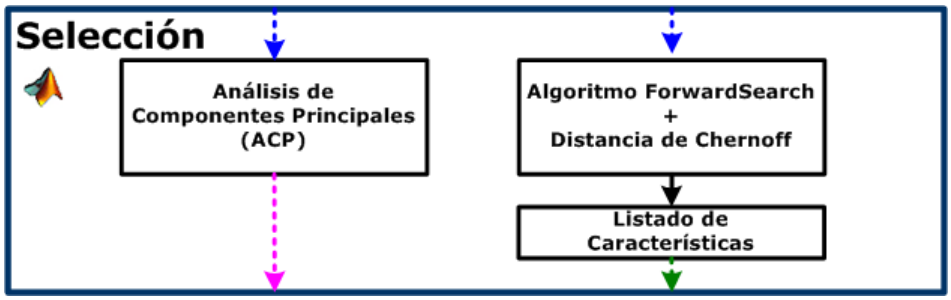



Figura 5.3: Selección de características.

## 5.4. Clasificación

Para estudiar la precisión en la clasificación, se utilizaron métodos de reducción lineal de dimensiones (Loog-Duin, Rueda-Herrera, Fisher) combinados con un clasificador cuadrático y otro lineal, además de un clasificador (MVS) basado en kernels.

- Cada uno de los métodos utilizados, aplica validación cruzada para entrenamiento utilizando los parámetros definidos.
- Se utilizan tres esquemas para trabajar: Homocedástico - Fisher's Discriminant Analysis (FDA); RLD basada en la Distancia de Chernoff - Rueda-Herrera, 2006 y Heterocedástico - Loog-Duin, 2004.
- Encontrar transformación de la matriz $A$ para cada dimensión $d = 1, \ldots, n-1$ con clasificación lineal y cuadrática aplicada a datos transformados.
- Realizar clasificación paralela, para una posterior evaluación de los clasificadores utilizados.

En la Figura 5.4, se representa los pasos a seguir, en la cual se desarrollan dos procesos paralelos, aplicación de reducción lineal de dimensiones (RLD) y Máquina de Vectores Soporte (MVS)

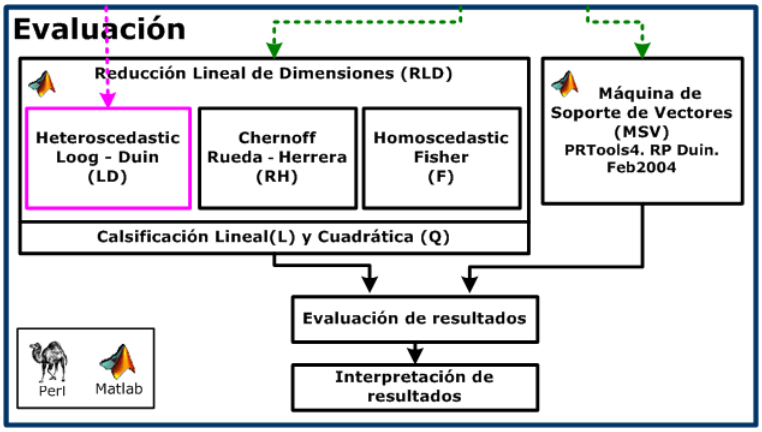



Figura 5.4: Etapa de Clasificación y evaluación.

### 5.4.1. Validación cruzada

En los métodos de RLD y MVS se aplicó validación cruzada de 10 pliegues, para el entrenamiento de los datos. El cual consiste en dividir el conjunto de complejos en 10 grupos equitativos y disjuntos, donde se inicia con un el primer grupo para test y los nueve restantes para entrenamiento. La segunda iteración se toma el grupo dos para test y los nueve restantes para entrenamiento, y así sucesivamente hasta que cada grupo haya sido utilizado como conjunto de test.

### 5.4.2. RLD

Para los tres métodos aplicados se obtuvieron los resultados, que se dividen según el tamaño de las matrices utilizadas. Las cuales fueron obtenidas con FOC. Los códigos para la implementación de las tres técnicas de RLD (Fisher, Loog-Duin y Rueda-Herrera) y la validación cruzada de 10 pliegues fueron reutilizados, efectuando modificaciones para su uso con los datos del proyecto, a partir de la implementación de Mohammed Liakat Ali, ex-alumno del profesor L. Rueda.

En la Tabla 5.5 se muestran los resultados obtenidos con el uso de los tres métodos de RLD utilizando la matriz de 20 dimensiones. Se muestra que el menor porcentaje de error se produce con el método de Loog-Duin con un 21,28 % con clasificación lineal. Si se compara

entre clasificadores, el lineal (L) obtiene mejores resultados que el clasificador cuadrático (C). Además se puede indicar que el método Rueda-Herrera entrega alguno de los menores porcentajes de error (C= 21,99 % y L= 21,32 %), además de Loog-Duin que obtiene el mejor resultado.

| Clasificador | Cuadrático | | | Lineal | | |
|---|---|---|---|---|---|---|
| Métodos | Fisher | Loog-Duin | Rueda-Herrera | Fisher | Loog-Duin | Rueda-Herrera |
| ERROR MÍNIMO | 24,25 % | 23,97 % | 21,99 % | 24,29 % | 21,28 % | 21,32 % |

Tabla 5.5: Porcentajes de error para clasificación con matriz de 20 dimensiones.

En la Tabla 5.6 se muestran los resultados obtenidos con el uso de los tres métodos de RLD utilizando la matriz de 75 dimensiones. El error más pequeño se obtiene con el método de Rueda-Herrera de 19,23 % con clasificador lineal. Además se destaca que obtiene un mejor resultado que los obtenidos con la matriz de 20 dimensiones.

| Clasificador | Cuadrático | | | Lineal | | |
|---|---|---|---|---|---|---|
| Métodos | Fisher | Loog-Duin | Rueda-Herrera | Fisher | Loog-Duin | Rueda-Herrera |
| ERROR MÍNIMO | 27,08 | 23,95 | 23,03 | 26,43 | 19,56 | 19,23 |

Tabla 5.6: Porcentaje de error para clasificación con matriz de 75 dimensiones.

En el caso de los procesos con las matrices de 277, 294 y 642 características, surgieron problemas de singularidad y valores infinitos, además por motivos de tiempo se decidió continuar con el resto de los procesos y trabajar las tres técnicas con sólo las matrices de 20 y 75 características.

### 5.4.3. Máquina de Vectores Soporte

En el caso de MVS se utilizó la matriz original, al igual que en ACP. Además se realizó validación cruzada de 10 pliegues, donde el total de complejos por clase es dividido en 10 partes iguales (particiones disjuntas), para entrenar el método.

En este caso se implementó un algoritmo que permite generar validación cruzada de 10 pliegues, además el código utilizado para la aplicación de Máquinas de vectores soporte se tomó de un paquete de funciones para Matlab para reconocimiento de patrones llamado PRTools4 [32]. El clasificador de soporte de vectores que se utiliza esta basado en diferentes kernels, en este caso se utilizó kernels polinomiales de orden 2 y 3. Para cada iteración de la validación cruzada se realiza la clasificación, obteniendo 10 errores, los cuales se promedian y se obtiene un porcentaje de precisión promedio de la clasificación a través de MVS. en las Tablas que se presentan a continuación entregan los resultados obtenidos en los diferentes grados utilizados, para las diferentes matrices.

La Tabla 5.7 presenta los errores promedio obtenidos con MVS utilizando Kernel polinomial de grado 2. El error promedio más pequeño se presenta en la sub-matriz de 20 dimensiones. En este caso, se aplica MVS a todas las submatrices, incluyendo la nueva creada por ACP. Además se puede visualizar la variación de los resultados de las diferentes sub-matrices. El error inicial es bajo (20), después se incrementa (75 - 277) y luego baja nuevamente (294 - 643). Se podría concluir que no es necesario tener la totalidad de las características para obtener una buena clasificación de las clases.

| **Máquina de vectores soporte, kernel polinomial de grado 2.** | | | | | |
|---|---|---|---|---|---|
| | **20** | **75** | **277** | **294** | **643** |
| **Error Promedio** | **24,29 %** | 33,06 % | 52,61 % | 42,26 % | 33,04 % |

Tabla 5.7: Errores promedio (MVS), más validación cruzada, con Kernel polinomial de grado 2.

La Tabla 5.8 muestra los errores promedios obtenidos con MVS utilizando Kernel polinomial de grado 3. El error más pequeño se presenta también en la sub-matriz de 20 dimensiones. Realizando una comparación entre los resultados de ambos kernels, se puede indicar que el kernel polinomial grado 2 entrega el error más pequeño.

| **Máquina de vectores soporte, kernel polinomial de grado 3.** | | | | | |
|---|---|---|---|---|---|
| | **20** | **75** | **277** | **294** | **643** |
| **Error Promedio** | **27,35 %** | 66,85 % | 32,16 % | 30,41 % | 47,12 % |

Tabla 5.8: Errores promedio (MVS) más validación cruzada, con Kernel polinomial de grado 3.

## 5.5. Resumen de Resultados

En la Tabla 5.9 se muestra un resumen de todos los resultados obtenidos por los diferentes métodos y clasificadores. Se encuentran separados por criterio de selección de las características o componentes a utilizar, los métodos y clasificadores que fueron aplicados en el proyecto, los tamaños de las matrices utilizadas y los porcentajes de error obtenidos de los diferentes grupos de datos utilizados. Hay que destacar que ACP trabaja con componentes y no con características. A continuación se comenzará a trabajar en términos de la precisión de los resultados, es decir, el porcentaje total (100 %) menos el porcentaje de error obtenido. Por ejemplo, con un error del 23,9 %, se puede indicar que posee un 76,1 % (100 % - 23,9 %) de precisión.

| Matriz de características | Métodos | Clasificador | 20 | 75 | 277 | 294 | 642 |
|---|---|---|---|---|---|---|---|
| Forward Search + FOC (20, 75, 277, 296, 642 características) | Loog-Duin (RLD) | C | 23,9 % | 23,9 % | - | - | - |
| | | L | 21,2 % | 19,5 % | - | - | - |
| | Fisher (RLD) | C | 24,2 % | 27,0 % | - | - | - |
| | | L | 24,2 % | 26,4 % | - | - | - |
| | Chernoff (RLD) | C | 21,9 % | 23,0 % | - | - | - |
| | | L | 21,3 % | 19,2 % | - | - | - |
| Matriz original (20, 75, 277, 296, 642 caract.) | MVS | polinomial 2 | 24,2 % | 33,0 % | 52,6 % | 42,2 % | 33,0 % |
| | | polinomial 3 | 27,4 % | 66,9 % | 32,2 % | 30,4 % | 47,1 % |
| ACP (20, 75, 277 y 294 componentes) | RLD Loog-Duin | ACP - C | 25,9 % | 29,0 % | 32,5 % | 32,5 % | - |
| | | ACP - L | 25,3 % | 26,0 % | 28,5 % | 28,8 % | - |

Tabla 5.9: Detalle de los errores promedio en la separabilidad de las clases.

Desde los resultados de la Tabla 5.9 se pueden realizar las siguientes observaciones:

- Los resultados obtenidos en la selección de características que se efectuó con el algoritmo Foward Search y la distancia de Chernoff, estos alcanzan la precisión más alta en la separabilidad de las clases. Los valores más altos son obtenidos con los métodos de Loog-Duin con un 80,5 % y Rueda-Herrera 80,8 %. Los resultados se producen en la misma dimensión que corresponde a la matriz de 75 dimensiones. En ambos casos los resultados se obtienen con el clasificador lineal.

- También se muestra que MVS (con Kernel polinomiales de grado 2 y 3) con validación cruzada de 10-pliegues entrega una precisión del 75,8 %, utilizando sólo 20 características de un total de 642. Pero los resultados obtenidos siguen siendo de menor precisión que los obtenidos con los métodos de RLD.

- En el caso de ACP, combinado con el método Loog-Duin y el clasificador lineal, hay una alta precisión del 74,7 % con 20 componentes. Usando 75 componentes la precisión se incrementa en uno por ciento. Por esto, no se justificaría agregar 55 características más, para sólo un aumento de un 1 %. En este caso ACP presenta menor precisión que los resultados obtenidos con RLD, pero obtiene mayor precisión que los obtenidos con MVS. No se debe olvidar que en este método se utilizan componentes que representan a un grupo de características y no las características.

- Se puede concluir que el método RLD, basado en el criterio de Chernoff (Rueda-Herrera), utilizando validación cruzada de 10 pliegues y clasificador lineal, es aquel que entrega el mejor porcentaje con un 80.8 % de precisión respecto a todos los otros métodos y los diferentes clasificadores utilizados.

A continuación se presentan cinco medidas para evaluar la clasificación, complementando los porcentajes de error y precisión presentados anteriormente. Las medidas son: valor predictivo positivo (PPV), valor predictivo negativo (NPV), sensibilidad, especificidad y precisión.

- $PPV = TP/(TP+FP)$
- $NPV = TN/(TN+FN)$
- $Sensibilidad = TP/(TP+FN)$
- $Especificidad = TN/(TN+FP)$
- $Precisin = (TP+TN)/n$

Para la obtención de estas valores se supone que existen dos clases Positivo (permanente - P) y Negativo (transitoria - N). Además se debe conocer las siguientes variables: $TP$ =verdaderos permanentes (permanentes clasificado como permanente); $TN$ =verdaderos transitorios (transitorio clasificado como transitorio); $FP$ =falsos permanentes (transitorio clasificado como permanente); $FN$ =falsos transitorios (permanentes clasificado como transitorio), TP+TN= número total de sentencias extraídas correctamente, TP+FP= número total de sentencias de la clase permanente y $n$ =número total de complejos utilizados.

En la Tabla 5.10 se muestran los valores obtenidos con los cinco métodos de evaluación de clasificadores.

| Métodos | Dim. | Clasif. | NPV | PPV | Especificidad | Sensibilidad | Precisión |
|---|---|---|---|---|---|---|---|
| Ficher (RLD) | 20 | C | 0,759 % | 0,757 % | 0,870 % | 0,588 % | 0,757 % |
| | | L | 0,729 % | 0,817 % | 0,896 % | 0,580 % | 0,757 % |
| | 75 | C | 0,798 % | 0,580 % | 0,806 % | 0,568 % | 0,729 % |
| | | L | 0,754 % | 0,699 % | 0,845 % | 0,565 % | 0,736 % |
| LoogDuin (RLD) | 20 | C | 0,788 % | 0,699 % | 0,851 % | 0,609 % | 0,760 % |
| | | L | 0,754 % | 0,860 % | 0,922 % | 0,615 % | 0,787 % |
| | 75 | C | 0,798 % | 0,677 % | 0,844 % | 0,606 % | 0,760 % |
| | | L | 0,733 % | 0,871 % | 0,929 % | 0,638 % | 0,804 % |
| Chernoff (RLD) | 20 | C | 0,793 % | 0,752 % | 0,875 % | 0,625 % | 0,780 % |
| | | L | 0,743 % | 0,881 % | 0,932 % | 0,612 % | 0,787 % |
| | 75 | C | 0,798 % | 0,709 % | 0,857 % | 0,617 % | 0,770 % |
| | | L | 0,768 % | 0,894 % | 0,939 % | 0,638 % | 0,807 % |
| Loog-Duin (ACP) | 20 | C | 0,704 % | 0,817 % | 0,894 % | 0,559 % | 0,739 % |
| | | L | 0,669 % | 0,914 % | 0,944 % | 0,559 % | 0,747 % |
| | 75 | C | 0,665 % | 0,806 % | 0,802 % | 0,524 % | 0,709 % |
| | | L | 0,547 % | 0,968 % | 0,974 % | 0,494 % | 0,679 % |
| | 277 | C | 0,788 % | 0,430 % | 0,751 % | 0,482 % | 0,767 % |
| | | L | 0,709 % | 0,731 % | 0,852 % | 0,535 % | 0,716 % |
| | 294 | C | 0,778 % | 0,333 % | 0,718 % | 0,408 % | 0,638 % |
| | | L | 0,660 % | 0,710 % | 0,832 % | 0,489 % | 0,767 % |

Tabla 5.10: Medidas.

## 5.6. Interpretación Biológica de los Resultados

La idea de esta sección es corroborar que lo que se obtuvo automáticamente a través de la informática, tiene un sentido biológico para los usuarios que presentaron este problema inicialmente.

Se comparan los resultados del método y la validación. Lo que permitirá interpretarlos y obtener características discriminantes. Para esto se deben realizar los siguientes pasos:

1. Tomar las primeras 20 características, las cuales pueden ser el valor de la energía o el número de aminoácido en la secuencia de la proteína, los cuales están asociados a un aminoácido. Se debió identificar que aminoácido esta asociado a cada uno de los 20 valores (características). Esto se detalla en la Sección 5.1.2.

2. Crear con los aminoácidos obtenidos, una tabla con las frecuencias de aparición de los aminoácidos en las primeras 20 características, de todos los complejos. Por ejemplo, en la Tabla 5.11, se muestra que en la característica 121, la GLI (Glicina) aparece 10 veces en los 296 complejos y la VAL (Valina) 3 veces.

3. Obtener los promedios de las frecuencias de aminoácidos por energías, separadas por los valores máximos y mínimos, de cada clase. En la Tabla 5.12 se puede observar la frecuencia promedio que tiene un aminoácido con respecto al total. Por ejemplo, la GLI (Glicina) tiene un promedio de 7,67. Además se puede indicar que su aparición se encuentra en las energías de desolvatación y entre los valores mínimos.

4. Seleccionar los aminoácidos con las 3 más altas frecuencias promedio de aparición por energía. Se obtiene la Tabla 5.13, donde están separados por valores máximos/mínimos y por clases.

5. Luego se identifican aquellos aminoácidos que no se repiten entre las clases, los que se encuentran en negrita, en la Tabla 5.13. Por ejemplo, en las energías electrostáticas (mínimos) la PRO (Prolina) aparece en ambas clases con diferentes valores (24.25 y 30.38), por lo tanto no es útil para discriminar. En cambio ALA, VAL, GLI y LEU no se repiten, por lo que muestran diferenciación entre clases.

6. Tomar sólo aquellos aminoácidos que muestran diferenciación. Se crea la Tabla 5.14 el cual permitirá evaluar a otro grupo de complejos, e indicar si estos aminoácidos permiten discriminar entre las clases.

| | | MATRIZ ENERGIAS DESOLVATACION | | | | | | | | | | | | | | | | MATRIZ ENERGIAS ELECTROSTATICA | | | | | | | | | | | | | | | | ELECT CONTACTO | | | | E. LIBRE CONTACTO | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Permanentes | | | | | | | | Transitorios | | | | | | | | 20x8 | | | | | | | | | | | | | | | | 20x2 | | | | 20x2 | | | |
| | Canact | 121 | 100 | 282 | 314 | 288 | 298 | 290 | 292 | 121 | 100 | 282 | 314 | 288 | 298 | 290 | 292 | 354 | 346 | 342 | 330 | 360 | 348 | 326 | 338 | 354 | 346 | 342 | 330 | 360 | 348 | 326 | 338 | 407 | 470 | 407 | 470 | 557 | 626 | 557 | 626 |
| | Min/max | MIN | | | | MAX | | | | MIN | | | | MAX | | | | MIN | | | | | | | | MIN | | | | | | | | MIN | MAX | MIN | MAX | MIN | MAX | MIN | MAX |
| GLI | | 10 | 2 | 11 | 2 | 7 | 8 | 2 | 7 | 16 | 21 | 19 | 4 | 11 | 17 | 14 | 16 | 7 | 8 | 8 | 5 | 9 | 10 | 7 | 13 | 11 | 16 | 17 | 14 | 24 | 18 | 23 | 8 | 8 | 2 | 16 | 21 | 2 | 10 | 14 | 18 |
| VAL | | 3 | 12 | 9 | 0 | 4 | 2 | 4 | 1 | 15 | 12 | 15 | 0 | 13 | 10 | 9 | 7 | 10 | 6 | 14 | 7 | 6 | 9 | 3 | 8 | 14 | 19 | 19 | 29 | 12 | 19 | 15 | 25 | 6 | 12 | 19 | 12 | 4 | 9 | 9 | 19 |
| ALA | | 9 | 6 | 9 | 1 | 4 | 2 | 4 | 5 | 12 | 13 | 14 | 1 | 11 | 8 | 9 | 8 | 7 | 11 | 9 | 6 | 7 | 14 | 5 | 11 | 23 | 18 | 15 | 12 | 17 | 11 | 13 | 21 | 11 | 6 | 18 | 13 | 4 | 14 | 9 | 11 |
| LEU | | 10 | 13 | 8 | 0 | 2 | 2 | 4 | 5 | 23 | 23 | 22 | 1 | 15 | 9 | 14 | 8 | 9 | 10 | 9 | 4 | 5 | 6 | 4 | 4 | 17 | 17 | 22 | 17 | 13 | 17 | 15 | 25 | 10 | 13 | 17 | 23 | 4 | 6 | 14 | 17 |
| ILE | | 5 | 11 | 2 | 0 | 3 | 4 | 1 | 2 | 11 | 14 | 11 | 1 | 5 | 1 | 11 | 3 | 11 | 5 | 6 | 5 | 3 | 7 | 1 | 7 | 10 | 14 | 13 | 19 | 8 | 6 | 9 | 16 | 5 | 11 | 14 | 14 | 1 | 7 | 11 | 6 |
| MET | | 5 | 9 | 0 | 2 | 2 | 2 | 1 | 1 | 3 | 8 | 5 | 4 | 4 | 3 | 1 | 5 | 2 | 5 | 3 | 1 | 2 | 4 | 3 | 4 | 4 | 8 | 4 | 7 | 2 | 5 | 6 | 3 | 5 | 9 | 8 | 8 | 1 | 4 | 1 | 5 |
| PRO | | 5 | 5 | 5 | 0 | 0 | 5 | 2 | 6 | 13 | 13 | 13 | 3 | 10 | 9 | 9 | 7 | 11 | 7 | 16 | 46 | 12 | 15 | 63 | 24 | 15 | 16 | 13 | 57 | 17 | 22 | 79 | 24 | 7 | 5 | 16 | 13 | 2 | 15 | 9 | 22 |
| FEN | | 4 | 10 | 5 | 0 | 1 | 1 | 5 | 2 | 5 | 9 | 11 | 2 | 4 | 6 | 7 | 6 | 3 | 3 | 6 | 3 | 4 | 3 | 1 | 2 | 7 | 10 | 14 | 10 | 5 | 9 | 8 | 10 | 3 | 10 | 10 | 9 | 5 | 3 | 7 | 9 |
| TRP | | 2 | 3 | 2 | 1 | 2 | 1 | 0 | 1 | 3 | 8 | 2 | 3 | 0 | 2 | 1 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 7 | 3 | 0 | 2 | 3 | 0 | 2 | 2 | 3 | 7 | 8 | 0 | 0 | 1 | 3 |
| SER | | 4 | 2 | 7 | 3 | 5 | 5 | 3 | 3 | 18 | 10 | 16 | 18 | 18 | 27 | 15 | 15 | 3 | 2 | 3 | 2 | 7 | 3 | 0 | 1 | 14 | 9 | 15 | 2 | 16 | 12 | 4 | 8 | 2 | 2 | 9 | 10 | 3 | 3 | 15 | 12 |
| CIS | | 3 | 3 | 3 | 1 | 0 | 3 | 2 | 1 | 8 | 7 | 2 | 1 | 3 | 1 | 2 | 5 | 1 | 3 | 1 | 1 | 3 | 2 | 1 | 0 | 7 | 6 | 4 | 9 | 3 | 5 | 10 | 6 | 3 | 3 | 6 | 7 | 2 | 2 | 2 | 5 |
| THR | | 10 | 9 | 3 | 1 | 10 | 6 | 2 | 3 | 11 | 17 | 14 | 5 | 12 | 8 | 11 | 11 | 4 | 8 | 1 | 0 | 5 | 2 | 1 | 3 | 13 | 8 | 6 | 2 | 11 | 9 | 1 | 14 | 8 | 9 | 8 | 17 | 2 | 2 | 11 | 9 |
| TIR | | 8 | 3 | 4 | 2 | 5 | 1 | 8 | 5 | 4 | 16 | 8 | 4 | 3 | 9 | 7 | 7 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 5 | 3 | 4 | 7 | 5 | 7 | 0 | 2 | 2 | 3 | 3 | 16 | 8 | 1 | 7 | 7 |
| HIS | | 4 | 2 | 2 | 5 | 4 | 3 | 1 | 3 | 3 | 4 | 6 | 7 | 4 | 7 | 6 | 7 | 2 | 4 | 1 | 0 | 2 | 1 | 0 | 1 | 7 | 6 | 2 | 0 | 3 | 3 | 0 | 2 | 4 | 2 | 6 | 4 | 1 | 1 | 6 | 3 |
| ASN | | 2 | 0 | 4 | 9 | 8 | 3 | 12 | 4 | 6 | 4 | 5 | 24 | 8 | 18 | 9 | 13 | 4 | 3 | 1 | 1 | 6 | 1 | 0 | 0 | 7 | 3 | 7 | 2 | 7 | 10 | 1 | 3 | 3 | 0 | 3 | 4 | 12 | 1 | 9 | 10 |
| GLN | | 0 | 1 | 0 | 5 | 7 | 8 | 4 | 8 | 7 | 5 | 6 | 20 | 15 | 11 | 17 | 17 | 1 | 2 | 2 | 2 | 3 | 5 | 0 | 1 | 6 | 5 | 6 | 3 | 9 | 6 | 2 | 6 | 2 | 1 | 5 | 5 | 4 | 5 | 17 | 6 |
| GLU | | 4 | 0 | 6 | 12 | 9 | 8 | 11 | 2 | 16 | 4 | 11 | 31 | 21 | 21 | 18 | 15 | 1 | 1 | 3 | 2 | 7 | 3 | 1 | 2 | 16 | 9 | 13 | 1 | 10 | 9 | 1 | 6 | 1 | 0 | 9 | 4 | 11 | 3 | 18 | 9 |
| LIS | | 4 | 1 | 5 | 11 | 4 | 12 | 10 | 10 | 12 | 3 | 8 | 32 | 13 | 14 | 16 | 13 | 7 | 6 | 1 | 7 | 5 | 2 | 1 | 6 | 9 | 8 | 16 | 7 | 16 | 18 | 11 | 14 | 6 | 1 | 8 | 3 | 10 | 2 | 16 | 18 |
| ASP | | 0 | 0 | 3 | 17 | 6 | 8 | 8 | 12 | 8 | 9 | 9 | 25 | 18 | 14 | 17 | 22 | 2 | 2 | 3 | 0 | 1 | 3 | 1 | 0 | 7 | 12 | 5 | 2 | 14 | 9 | 2 | 1 | 2 | 0 | 12 | 9 | 8 | 3 | 17 | 9 |
| ARG | | 1 | 1 | 5 | 21 | 10 | 9 | 9 | 12 | 9 | 3 | 6 | 17 | 15 | 8 | 10 | 14 | 6 | 3 | 5 | 0 | 5 | 2 | 1 | 4 | 8 | 9 | 5 | 3 | 9 | 5 | 3 | 7 | 3 | 1 | 9 | 3 | 9 | 2 | 10 | 5 |

Tabla 5.11: Frecuencia de aparición de los aminoácidos en las 20 características.

| | desolvatacion | | | | electrostatica | | Econtacto | | | | Elibrecontacto | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | p | | t | | p | t | p | p | t | t | p | p | t | t |
| | min | max | min | max | min | min | min | max | min | max | min | max | min | max |
| GLI | 7,67 | 5,2 | 18,67 | 12,4 | 8,38 | 16,38 | 8 | 2 | 16 | 21 | 2 | 10 | 14 | 18 |
| VAL | 8,00 | 2,2 | 14,00 | 7,8 | 7,88 | 19,00 | 6 | 12 | 19 | 12 | 4 | 9 | 9 | 19 |
| ALA | 8,00 | 3,2 | 13,00 | 7,4 | 8,75 | 16,25 | 11 | 6 | 18 | 13 | 4 | 14 | 9 | 11 |
| LEU | 10,33 | 2,6 | 22,67 | 9,4 | 6,38 | 17,88 | 10 | 13 | 17 | 23 | 4 | 6 | 14 | 17 |
| ILE | 6,00 | 2,0 | 12,00 | 4,2 | 5,63 | 11,88 | 5 | 11 | 14 | 14 | 1 | 7 | 11 | 6 |
| MET | 4,67 | 1,6 | 5,33 | 3,4 | 3,00 | 4,88 | 5 | 9 | 8 | 8 | 1 | 4 | 1 | 5 |
| PRO | 5,00 | 2,6 | 13,00 | 7,6 | 24,25 | 30,38 | 7 | 5 | 16 | 13 | 2 | 15 | 9 | 22 |
| FEN | 6,33 | 1,8 | 8,33 | 5,0 | 3,13 | 9,13 | 3 | 10 | 10 | 9 | 5 | 3 | 7 | 9 |
| TRP | 2,33 | 1,0 | 4,33 | 2,0 | 0,38 | 2,50 | 2 | 3 | 7 | 8 | 0 | 0 | 1 | 3 |
| SER | 4,33 | 3,8 | 14,67 | 18,6 | 2,63 | 10,00 | 2 | 2 | 9 | 10 | 3 | 3 | 15 | 12 |
| CIS | 3,00 | 1,4 | 5,67 | 2,4 | 1,50 | 6,25 | 3 | 3 | 6 | 7 | 2 | 2 | 2 | 5 |
| THR | 7,33 | 4,4 | 14,00 | 9,4 | 3,00 | 8,00 | 8 | 9 | 8 | 17 | 2 | 2 | 11 | 9 |
| TIR | 5,00 | 4,2 | 9,33 | 6,0 | 1,13 | 4,13 | 2 | 3 | 3 | 16 | 8 | 1 | 7 | 7 |
| HIS | 2,67 | 3,2 | 4,33 | 6,2 | 1,38 | 2,88 | 4 | 2 | 6 | 4 | 1 | 1 | 6 | 3 |
| ASN | 2,00 | 7,2 | 5,00 | 14,4 | 2,00 | 5,00 | 3 | 0 | 3 | 4 | 12 | 1 | 9 | 10 |
| GLN | 0,33 | 6,4 | 6,00 | 16,0 | 2,00 | 5,38 | 2 | 1 | 5 | 5 | 4 | 5 | 17 | 6 |
| GLU | 3,33 | 8,4 | 10,33 | 21,2 | 2,50 | 8,13 | 1 | 0 | 9 | 4 | 11 | 3 | 18 | 9 |
| LIS | 3,33 | 9,4 | 7,67 | 17,6 | 4,38 | 12,38 | 6 | 1 | 8 | 3 | 10 | 2 | 16 | 18 |
| ASP | 1,00 | 10,2 | 8,67 | 19,2 | 1,50 | 6,50 | 2 | 0 | 12 | 9 | 8 | 3 | 17 | 9 |
| ARG | 2,33 | 12,2 | 6,00 | 12,8 | 3,25 | 6,13 | 3 | 1 | 9 | 3 | 9 | 2 | 10 | 5 |

Tabla 5.12: Frecuencia promedio de aparición de los aa, por máximos y mínimos.

| | **MÁXIMOS** | | | | **MÍNIMOS** | | | |
|---|---|---|---|---|---|---|---|---|
| | **PERMANENTE** | | **TRANSITORIO** | | **PERMANENTE** | | **TRANSITORIO** | |
| **ENERGÍA** | **aa** | **VALOR** | **aa** | **VALOR** | **aa** | **VALOR** | **aa** | **VALOR** |
| **Electrostática (8)** | | | | | PRO | 24.25 | PRO | 30.38 |
| | | | | | **ALA** | **8.75** | **VAL** | **19** |
| | | | | | **GLI** | **8.39** | **LEU** | **17.88** |
| **Desolvatación(8)** | **ARG** | **12.20** | **GLU** | **21.20** | LEU | 10.33 | LEU | 22.67 |
| | ASP | 10.20 | ASP | 19.20 | **ALA** | **8.0** | **GLI** | **18.67** |
| | LIS | 9.40 | LIS | 17.60 | VAL | 8.0 | VAL | 14 |
| **Interacción electrostática(2)** | LEU | 13 | LEU | 23 | ALA | 11 | VAL | 19 |
| | **VAL** | **12** | **GLI** | **21** | LEU | 10 | ALA | 18 |
| | **ILE** | **11** | **THR** | **17** | **THR** | **8** | **LEU** | **17** |
| | | | | | **GLI** | **8** | | |
| **Libre de interacción(2)** | PRO | 15 | PRO | 22 | **ASN** | **12** | **GLU** | **18** |
| | **ALA** | **14** | **VAL** | **19** | GLU | 11 | ASP | 17 |
| | GLI | 10 | GLI | 18 | **LYS** | **10** | **GLN** | **17** |

aa = aminoácidos

Tabla 5.13: Promedio de frecuencias por energías.

| Diferencias | | | | | |
|---|---|---|---|---|---|
| | | Máximos | | Mínimos | |
| | | Permanentes | Transitorios | Permanentes | Transitorios |
| **Energías electrostáticas** | (8) | - | - | ALA | VAL |
| | | | | GLI | LEU |
| **Energía Desolvatación** | (8) | ARG | GLU | ALA | GLI |
| **Energía de interacción electrostática** | (2) | VAL | GLI | THR | VAL |
| | | ILE | THR | GLI | |
| **Energía libre de interacción** | (2) | ALA | VAL | ASN | ASP |
| | | | | LIS | GLN |

Tabla 5.14: Aminoácidos que presentan diferencias entre clases

## 5.7. Validación de la Interpretación Biológica

Al inicio del proyecto, se apartó un conjunto de 30 complejos (apartados al inicio del proyecto, de manera aleatoria), con el fin de ser analizados desde un punto de vista biológico. Se trabajó con un grupo de 28, debido a que los dos restante presentaron problemas de formato, quedando inutilizables para ser usados con el programa Fast Contact.

Se utilizó un conjunto de 28 complejos , ya clasificados en 22 permanentes y 6 transitorios. Luego se les dió el formato adecuado para ejecutar la aplicación Fast Contact, esto se hizo reutilizando los algoritmos creados en la primera etapa del proyecto.

Luego con un nuevo algoritmo (Perl) se tomaron los archivos de salida que genera Fast Contact y se recopiló la información de las 20 características más discriminantes que se seleccionaron en la segunda etapa.

Al tener todos los datos, es decir, valores y su aminoácido asociado, se procede a identificar aquellos aminoácidos relevantes para la diferencia entre las clases. Se utiliza la Tabla 5.14, creada en el punto anterior.

La forma de clasificación utilizada, es sólo sumando la cantidad de aciertos de aminoácidos que se encuentran en una clase, versus la otra clase. Por ejemplo, se encuentran en Energía de

Desolvatación los siguientes valores: 2 ARG y 1 GLU entre los valores máximos, 3 ALA y 2 GLI entre los valores mínimos. esto indica que hay 5 aminoácidos en la categoría de permanentes (ARG, ALA) y 3 en la categoría de transitorios (GLU,GLI). De estas comparaciones se obtiene la Tabla 5.15.

| | **Permanentes (P)** | **Transitorios (T)** | **Total (P y T)** |
|---|---|---|---|
| **Aciertos** | 14 | 3 | 17 (61 %) |
| **Errores** | 2 | 3 | 5 |
| **Igualdad** | 6 | 0 | 6 |
| **Total** | 22 | 6 | 28 |

Tabla 5.15: Resultados por clase.

Del la Tabla 5.15 se puede concluir, que existe una precisión del 61 % y que se podría clasificar un conjunto de complejos en transitorios y permanentes. Aunque el valor entregado por los diferentes métodos y clasificadores es mucho mayor, cercano al 80 % de precisión, indica que esta interpretación puede mejorar, re-evaluando el paso 4 de la sección anterior. Esto puede ser a través de:

- un análisis más detallado de las frecuencias de los aminoácidos, no sólo evaluar su aparición sino su relación con los valores energéticos que tiene asociado.
- tomar como medida de selección más de tres valores (máximos de frecuencias promedios), o tomar todos los valores presentados, para ampliar el rango y aumentar la precisión.
- dar una puntuación a cada aminoácido de acuerdo a su frecuencia, en relación a aminoácidos que aparecen menos.

Sólo queda indicar que los resultados no son tan categóricos (menor precisión) como los obtenidos automáticamente a través de la informática (clasificación), pero son resultados que permiten obtener un factor de diferenciación entre las clases estudiadas. Es comprensible obtener una precisión menor en esta clasificación, porque sólo se esta evaluando la frecuencia de aparición de los aminoácidos en las características y no en los valores energéticos en detalle de cada residuo y sin evaluar las propiedades de los aminoácidos. Lo que indica que

hay trabajo que se puede realizar a futuro, ampliando el análisis de los resultados obtenidos tanto informáticos, como biológicos.

Se puede proponer como trabajo futuro, el automatizar el proceso de análisis desde el punto biológico, para así obtener reglas, que permitan un análisis más preciso de los datos.

# Capítulo 6

# Conclusiones

Muchos de los procesos biológicos dependen de interacciones entre proteínas, sean estas permanentes o de corta duración. Las interacciones entre proteínas han sido extensamente estudiadas pero aún no es posible predecir exitosamente las posibles zonas de interacción de una proteína con otra. Asimismo, no es posible predecir la estabilidad de la interacción.

En el presente proyecto se identificaron aquellas características de interfaces de IPP, que permitan diferenciar entre interacciones transitorias y permanentes.

Se trabajó con una base de datos de 296 complejos para los cuales se conoce su estructura tridimensional y clasificación en interacciones transitorias y permanentes. Para cada una de las superficies se determinó la contribución energética utilizando la aplicación Fast Contact, obteniéndose 642 características por complejo, las cuales fueron procesadas con un algoritmo de selección de características (búsqueda secuencial hacia adelante - Forward Search) en conjunto con la distancia de Chernoff, para evaluar la posibilidad de separación de las clases a través de de un ranking de las características más influyentes en la interacción.

Para estudiar la precisión en la clasificación se utilizaron métodos de reducción lineal de dimensiones, específicamente criterio de Loog-Duin, criterio de Fisher y criterio Rueda-Herrera, en conjunto con clasificadores lineal y cuadrático. Paralelamente se utilizó una Máquina de vectores soporte, para comparar el proceso completo propuesto en este proyecto y para validar la etapa de selección de características se aplicó análisis de componentes principales.

La selección de características realizadas utilizando el algoritmo Forward Search en conjunto con la distancia de Chernoff alcanzó la más alta precisión en la separabilidad de las clases.

Los valores más altos son obtenidos con los clasificadores Loog-Duin (20 características) y Rueda-Herrera (75 características) con un porcentaje de precisión del 80,8 % al 80,5 % respectivamente.

La precisión obtenida con el resto de los clasificadores no es menor, ya que los resultados alcanzan precisiones cercanas al 75 % . Algunos resultados más detallados se pueden observar en la utilización de MVS con Kernels polinomiales (de grado 2 y 3) y validación cruzada de 10-pliegues que entrega una precisión del 72 %, utilizando sólo 20 características de un total de 642. En el caso del análisis de componentes principales en conjunto con el método Loog-Duin y el clasificador lineal, se presenta una precisión del 74,7 %, con 20 componentes. Al utilizar 75 componentes la precisión se incrementa en un 1 %. Lo cual debe analizarse, para indicar si realmente se justificaría el agregar 55 características más, para sólo lograr un incremento de uno por ciento.

Desde el punto de vista biológico, los resultados obtenidos permiten especificar que energías contribuyen más a la separabilidad entre las clases, además de lograr identificar aquellos aminoácidos relevantes en estas energías y específicamente en que parte de la superficie de interacción. En este caso, las energías de desolvatación y electrostática de las cadenas, son las que contribuyen más a la separabilidad de las clases, que las energías de interacción electrostática y las energías libres de interacción. Estos resultados fueron validados realizando una nueva selección sobre complejos no utilizados inicialmente, lo que entregó una precisión cercana al 61 % (de acuerdo a la interpretación biológica), la cual puede ser mejorada, realizando nuevos análisis. Es comprensible obtener una precisión menor en esta clasificación, porque sólo se esta evaluando la frecuencia de aparición de los aminoácidos en las características y no en los valores energéticos en detalle de cada residuo (sin evaluar las propiedades de los aminoácidos).

Los resultados obtenidos pueden ser mejorados, aplicando otros criterios de selección y de clasificación. Además se puede ampliar el análisis de los resultados obtenidos desde la interpretación biológica, a través de la extracción de características más discriminantes, haciendo uso de métodos automáticos que permitan extraer reglas desde los resultados obtenidos con las primeras tres etapas.

Al finalizar este proyecto, se puede concluir que existen características energéticas en la zona de interacción entre proteínas, que permiten discriminar entre interacciones transitorias e interacciones permanentes.

# Bibliografía


[1] B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–4, 1998.

[2] S. Ansari and V. Helms. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, 61(2):344–55, 2005.

[3] A.I. Archakov, V.M. Govorun, A.V. Dubanov, Y.D. Ivanov, A.V. Veselovsky, P. Lewi, and P. Janssen. Protein-protein interactions as a target for drugs in proteomics. *Proteomics*, 3(4):380–91, 2003.

[4] T. K. Attwood and D. J. Parry-Smith. *Introducción a la Bioinformática*. Prentice-Hall, 2002.

[5] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3):708–19, 2003.

[6] A. Bairoch and B. Boeckmann. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res*, 19 Suppl:2247–9, 1991. (http://www.expasy.org/sprot/).

[7] R. Barandela, J. S. Sánchez, V. Garcia, and E. Rangel. Fusion of techniques for handling the imbalanced training sample problem. *Proceedings of 6th Simposium Iberoamericano de Reconocimiento de Pstrones. Brazil*, 2001.

[8] D. Benson, D. J. Lipman, and J. Ostell. GenBank. *Nucleic Acids Res*, 21(13):2963–5, 1993. (http://www.psc.edu/general/software/packages/genbank/genbank).

[9] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank: update. *Nucleic Acids Res*, 32(Database issue):D23–6, 2004.

[10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, (28):235–242, 2000. (http://www.rcsb.org/pdb/Welcome.do).

[11] M. Berrera, H. Molinari, and F. Fogolari. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*, 4:8, 2003.

[12] J.C. Biro. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model*, 3:15, 2006.

[13] J.R. Bock and D.A. Gough. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

[14] J.R. Bock and D.A. Gough. Whole-proteome interaction mining. *Bioinformatics*, 19(1):125–34, 2003.

[15] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9, 1998.

[16] A. J. Bordner and R. Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60(3):353–66, 2005.

[17] Philip E. Bourne and Helge Weissig. *Structural Bioinformatics*. Wiley-Liss, 2003.

[18] J.R. Bradford and D.R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94, 2005.

[19] Thomas P. Brock and Michael T. Madigan. *Microbiología*. 6ta edition, 1993.

[20] N.J. Burgoyne and R.M. Jackson. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, 22(11):1335–42, 2006.

[21] Camacho CJ and Zhang C. FastContact: rapid estimate of contact and binding free energies. *Bioinformatics*, 21(10):2534–6, 2005.

[22] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–43, 2002.

[23] M. Claeyssens and B. Henrissat. Specificity mapping of cellulolytic enzymes: classification into families of structurally related proteins confirmed by biochemical analysis. *Protein Sci*, 1(10):1293–7, 1992.

[24] Morgan Collins, Francis. The Human Genome Project: Lessons from Large-Scale Biology. *Science*, 300(5617):pp. 286 – 290, 11 April 2003. (www.ornl.gov).

[25] Leiserson Charles E. Cormen Thomas H., Stein Clifford. *Introduction To Algorithms*. 2da edition, 2001.

[26] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, 2000.

[27] S. J. De Vries, A. D. Van Dijk, and A. M. Bonvin. WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins*, 63(3):479–89, 2006.

[28] S.J. De Vries and A.M. Bonvin. Intramolecular surface contacts contain information about protein-protein interface regions. *Bioinformatics*, 22(17):2094–8, 2006.

[29] Zoltán Dezso, Zoltán Oltvai, and Albert-Laszlo Barabási. Analysis of experimentally determined protein complexes in the yeast. *Bioinformatics*, 2003.

[30] Y. Duan, B. V. Reddy, and Y. N. Kaznessis. Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions. *Protein Sci*, 14(2):316–28, 2005.

[31] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and sons. New York, NY, second edition, 2000.

[32] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, and D. M.J. Tax. PRTools, a Matlab toolbox for pattern recognition,. Technical report, Delft Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, February 2004.

[33] Etzold T., Ulyanov A., and Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, (266):114–128, 1996. Sequence Retrival System, SRS. PubMedID: 8743681. European Molecular Biology Laboratory, Heidelberg, Germany. (ncbi.nlm.nih.go).

[34] Fleiss JL, Levin B, and Paik MC. *Statistical methods for rates and proportions*. John Wiley, New York, 3rd edition, 2004.

[35] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102, 2001.

[36] A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science*, 292(5523):1876–82, 2001.

[37] Cher-Sing Goh, Duncan Milburn, and Mark Gertein. Conformational changes associated with protein-protein interactions. 14:104–109, 2004.

[38] S.M. Gomez, W.S. Noble, and A. Rzhetsky. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics*, 19(15):1875–81, 2003.

[39] K. E. Gottschalk, H. Neuvirth, and G. Schreiber. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel*, 17(2):183–9, 2004.

[40] T. Haliloglu, O. Keskin, B. Ma, and R. Nussinov. How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. *Biophys J*, 88(3):1552–9, 2005.

[41] R. Hebrich. *Learning Kernel classifiers*. Mit.Press. Cambridge, MA, 2002.

[42] U. Hobohm, M. Scharf, and R. Schneider. Selection of representative protein data sets. *Protein Sci*, pages 409–417, 1992.

[43] L. Holm, C. Ouzounis, and C. Sander. A database of protein structure families with common folding motif. *Protein Sci*, pages 1691–1698, 1992.

[44] J. Hoskins, S. Lovell, and T.L. Blundell. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci*, 15(5):1017–29, 2006.

[45] P.F. Hsieh, D.S. Wang, and C.W. Hsu. A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information. *IEEE Trans Pattern Anal Mach Intell*, 28(2):223–35, 2006.

[46] International Workshops SSPR and SPR. *Non-Iterative Heteroscedastic Linear Dimension Reduction for Two-Class*, volume LNCS 2396. Springer, 2002.

[47] D. Jain, A. and Zongker. Feature Selection: Evaluation, Application and Sample Performance. *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, 19(2):153–158, February 1997.

[48] J. Janin. *Protein-Protein Recognition.*, chapter Kinetics and thermodynamics of protein-protein interactions from a structural perpective. Oxford University Press, 2000. 344pp.

[49] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, January 1996.

[50] S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–32, 1997.

[51] S. Jones and J. M Thornton. *Protein-Protein Recognition.*, chapter Analysis and classification of protein-protein interactions from a structural perspective. Oxford University Press, 2000.

[52] Susan Jones, Antoine Marin, and Janet Thornton. Protein domain interfaces: characterization and comparation with oligomeric protein interfaces. *Protein Engineering*, 13(2):77–82, 2000.

[53] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–94, 2005.

[54] O. Keskin, C. J. Tsai, H. Wolfson, and R. Nussinov. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, 13(4):1043–55, 2004.

[55] Magnani Kim, Seung-Jean. Optimal kernel selection in Kernel Fisher discriminant analysis. In *Proceedings of the 23rd international conference on Machine learning*, volume 148, pages 465 – 472, Pittsburgh, Pennsylvania, 2006. 23rd international conference on Machine learning, ACM International Conference Proceeding Series.

[56] K. Kinoshita and H. Nakamura. Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci*, 14(3):711–8, 2005.

[57] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–21, 2002.

[58] Kraemer HC. Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research*, 1(2):183–99, 1992.

[59] K. Kupas, A. Ultsch, and G Klebe. An algorithm for finding similarities in protein active sites. *Advances in bioinformatics and its applications.*, pages 373–379, December 2004.

[60] M. C. Lawrence and P. M. Colman. Shape complementarity at protein/protein interfaces. *J Mol Biol*, 234(4):946–50, 1993.

[61] Albert Lehningerr. *Principios de Bioquimica*, volume 1232 pp. OMEGA, 1 edition, 2005.

[62] D.F. Lin, K.-L.; Chun-Yuan Lin; Chuen-Der Huang; Hsiu-Ming Chang; Chiao-Yun Yang; Chin-Teng Lin; Chuan Yi Tang; Hsu. Feature Selection and Combination Criteria for Improving Accuracy in Protein Structure Prediction. *IEEE Transactions on Nanobioscience*, 6(2):186–196, June 2007.

[63] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–98, 1999.

[64] Marco Loog and Robert P.W. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 26(6), June 2004.

[65] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, 100(10):5772–7, 2003.

[66] J.R. Macdonald and . Johnson, Jr. Environmental features are important in determining protein secondary structure. *Protein Sci*, 10(6):1172–7, 2001.

[67] J. Mintseris and Z. Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, 102(31):10930–5, 2005. (http://zlab.bu.edu/julianm/MintserisWengPNAS05.html).

[68] G. Moont, H. A. Gabb, and M. J. Sternberg. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, 35(3):364–73, 1999.

[69] David Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, second edition, 2004.

[70] I. M. Nooren and J. M. Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–92, 2003.

[71] T.M. Nye, C. Berzuini, W.R. Gilks, M.M. Babu, and S. Teichmann. Predicting the strongest domain-domain contact in interacting protein pairs. *Stat Appl Genet Mol Biol*, 5(1):Article5, 2006.

[72] Y. Ofran and B. Rost. Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–87, 2003.

[73] Y. Ofran and B. Rost. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, 544(1-3):236–9, 2003.

[74] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–61, 2001.

[75] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancia. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), August 2005.

[76] R. W. Pickersgill. A rapid method of calculating charge-charge interaction energies in proteins. *Protein Eng*, 2(3):247–8, 1988.

[77] C. Plake, J. Hakenberg, and U. Leser. Learning patterns for information extraction from free text. Knowledge Management in Bioinformatics. Dept. Computer Science, Humboldt-Universitt zu Berlin, March 2005.

[78] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, 2006.

[79] D. Rajamani, S. Thiel, S. Vajda, and C.J. Camacho. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci U S A*, 101(31):11287–92, 2004.

[80] L. Rueda and M. Herrera. A New Approach to Multi-class Linear Dimensionality Reduction. pages 634–643. 11th Iberoamerican Congress on Pattern Recognition, 2006. Introduza el texto aquÃ.

[81] L. Rueda and M. Herrera. A New Linear Dimensionality Reduction Technique based on Chernoff Distance. pages 299–308. 10th Iberoamerican Conference on Artificial Intelligence. Ribeirao Preto, Brazil, October 2006.

[82] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–24, 2004.

[83] Saha R. P., Bahadur R. P., Pal A., Mandal S. and, and Chakrabarti P. ProFace: A server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Struct. Biol.*, 6: 11, 2006.

[84] H.A. Scheraga. Contribution of physical chemistry to an understanding of protein structure and function. *Protein Sci*, 1(5):691–3, 1992.

[85] Scholkopf and A. Samola. *Learning with kernels. Support vector machines, regulatization optimization and beyond*. Mit.Press. Cambridge, MA, 2002.

[86] H. P. Shanahan and J. M. Thornton. Amino acid architecture and the distribution of polar atoms on the surfaces of proteins. *Biopolymers*, 78(6):318–28, 2005.

[87] Shigeo Abe. *Support Vector Machines for Pattern Classification (Advances in Pattern Recognition)*. Springer, 2005.

[88] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright. Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17, 2008.

[89] T. Bayes. An essay towards solving a problem in the doctrine of chances. 53(370-418):1763.

[90] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier academic press, third edition, 2006.

[91] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–73, 2002.

[92] I. Xenarios and D. Eisenberg. Protein interaction databases. *Curr Opin Biotechnol*, 12(4):334–9, 2001.

[93] D. Xu, C.J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, 10(9):999–1012, 1997.

[94] J. Yu and F. Fotouhi. Computational approaches for predicting protein-protein interactions: a survey. *J Med Syst*, 30(1):39–44, 2006.

[95] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem*, 48(7):2325–35, 2005.

[96] C. Zhang, G. Vasmatzis, J. L. Cornette, and C. DeLisi. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol*, 267(3):707–26, 1997.

[97] Z. Zhang, S. Kochhar, and M.G. Grigorov. Descriptor-based protein remote homology identification. *Protein Sci*, 14(2):431–44, 2005.

[98] H. X. Zhou and S. Qin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–9, 2007.

[99] H. X. Zhou and Y. Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44(3):336–43, 2001.

# Apéndices

## A: PGH

La genómica y la genética, tuvieron apoyo de parte de la informática, por un tema esencial como es el volumen de los datos. Un ejemplo de esto es el Proyecto Genoma Humano (PGH) [24]. Se inició oficialmente en 1990 como un programa de quince años, con el cual se pretendía determinar la secuencia completa del genoma humano, localizando con exactitud los 100.000 genes aproximadamente y el resto del material hereditario de nuestra especie, responsables de las instrucciones genéticas de lo que somos desde el punto de vista biológico (Enrique Iañez Pareja). No era un solo proyecto, sino diversas iniciativas que poseen un fin común y donde se espero incrementar los conocimientos de los procesos biológicos y de la fisiología y patología de los seres humanos. Los rápidos avances tecnológicos permitieron reducir el tiempo pronosticado para este proyecto, logrando completar la secuenciación del genoma humano. Se publicó la secuencia del genoma humano, contrario a los intereses del sector privado. Actualmente se tiene acceso gratuito por internet a las páginas del consorcio del PGH [24] y se puede obtener secuencias de genes u otras secuencias de interés de regiones cromosómicas específicas [9].

## B: Bases de datos

El alto crecimiento de la base de datos PDB se puede observar en la Tabla 6.1, el crecimiento se puede observar más claramente en el Figura 6.1.

| Año | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 | 2002 | 2001 | 2000 | 1999 | 1998 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Anual | 143 | 7.305 | 6.536 | 5.419 | 5.227 | 4..187 | 3.017 | 2.839 | 2.632 | 2.360 | 2.064 |
| Total | 48.235 | 48.092 | 40.787 | 34..319 | 28.900 | 23.673 | 19.486 | 16.468 | 13.629 | 10.997 | 8.636 |
| Año | 1997 | 1996 | 1995 | 1994 | 1993 | 1992 | 1991 | 1990 | 1989 | 1988 | 1987 |
| Anual | 1.563 | 1.174 | 951 | 1.296 | 698 | 193 | 187 | 144 | 74 | 54 | 24 |
| Total | 6.572 | 5.009 | 3.835 | 2.884 | 1..588 | 890 | 697 | 510 | 366 | 292 | 238 |
| Año | 1986 | 1985 | 1984 | 1983 | 1982 | 1981 | 1980 | 1979 | 1978 | 1977 | 1976 |
| Anual | 19 | 20 | 22 | 35 | 32 | 16 | 16 | 10 | 7 | 24 | 13 |
| Total | 214 | 195 | 175 | 153 | 118 | 86 | 70 | 54 | 44 | 37 | 13 |

Tabla 6.1: Crecimiento de la Base de datos Protein Data Bank.
Cantidad de estructuras ingresadas en un Año y la cantidad Total de estructuras acumuladas hasta el 8 de Enero del 2008, ultima actualización del sitio PDB [10].

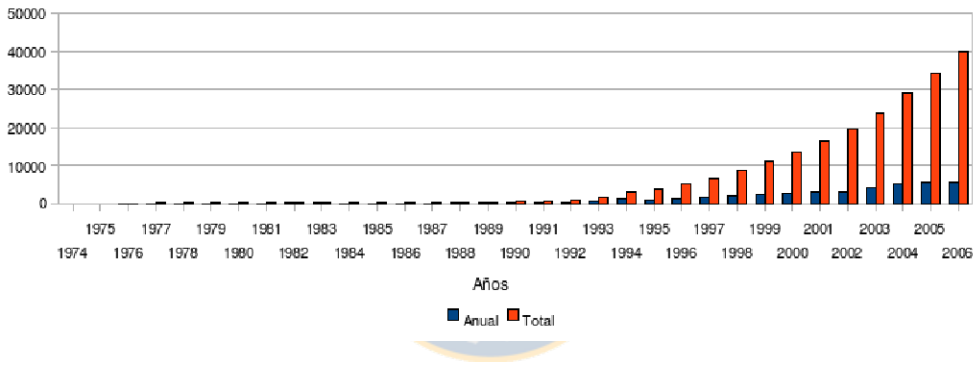



Figura 6.1: Gráfica del crecimiento de la Base de datos, Protein Data Bank .
El eje Y, muestra dos tipos de valores: el número de estructuras de proteínas descubiertas en el transcurso de un año (Anual) y el número Total de estructuras que se han acumulado a través de los años [10].

El crecimiento de las estructuras de complejos se puede observar en la Tabla 6.2 y ver gráficamente en la Figura 6.2.

| Año | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 | 2002 | 2001 | 2000 | 1999 | 1998 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Anual** | 5 | 259 | 291 | 190 | 238 | 177 | 125 | 111 | 95 | 113 | 97 |
| **Total** | 1943 | 1938 | 1679 | 1.388 | 1.198 | 960 | 783 | 658 | 547 | 452 | 339 |
| **Año** | **1997** | **1996** | **1995** | **1994** | **1993** | **1992** | **1991** | **1990** | **1989** | **1987** | **1986** |
| **Anual** | 43 | 107 | 37 | 32 | 11 | 5 | 3 | 1 | 3 | 0 | 0 |
| **Total** | 242 | 199 | 92 | 54 | 23 | 12 | 7 | 4 | 3 | 0 | 0 |

Tabla 6.2: Complejos la Base de datos, Protein Data Bank .
Cantidad de estructuras de complejos ingresadas en un Año y la cantidad Total acumulada de estructuras de complejos a la fecha [10].

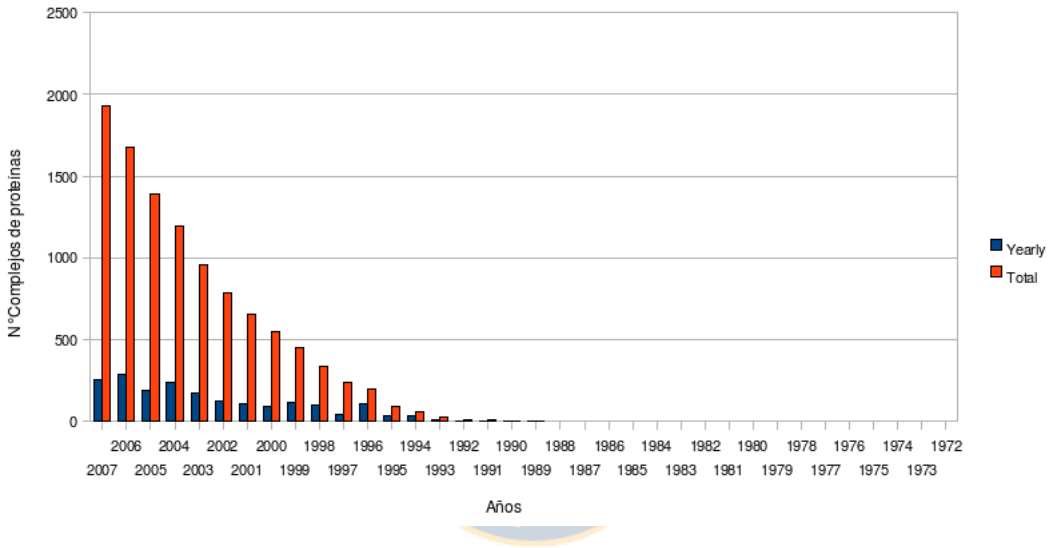



Figura 6.2: Número de Complejos en la Base de datos, Protein Data Bank .
El eje Y, muestra dos tipos de valores: el número de estructuras de complejos descubiertas en el transcurso de un año (Anual) y el número Total de estructuras de complejos que se han acumulado a través de los años [10].

# C: Proteínas

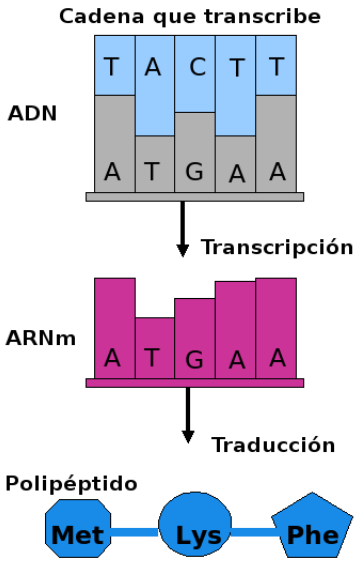



Figura 6.3: Proceso de ADN a proteínas

Las proteínas son moléculas formadas básicamente por carbono, hidrógeno, oxígeno y nitrógeno, además pueden contener azufre y en algunos tipos de proteínas, fósforo, hierro, magnesio y cobre entre otros elementos. Los aminoácidos se caracterizan por poseer un grupo carboxilo (-COOH) y un grupo amino (-NH2). Las otras dos partes del carbono se saturan con un átomo de H y con un grupo variable denominado radical R. Un enlace peptídico, es un enlace covalente que se establece entre el grupo carboxilo de un aminoácido y el grupo amino del siguiente, dando lugar al desprendimiento de una molécula de agua [60].

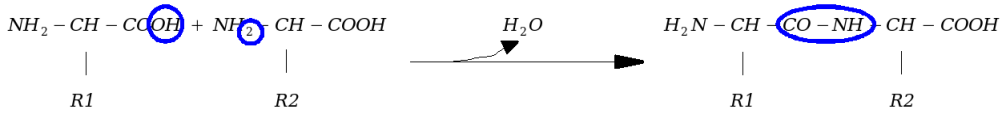



Figura 6.4: Enlace Peptídico

La unión entre dos aminoácidos da lugar a un péptido; si el número de aminoácidos que forma la molécula no es mayor de 10, se denomina oligopéptido, si es superior a 10 se llama polipéptido y si el número es superior a 100 aminoácidos (aprox.) se habla de proteína.

## C.1. Estructuras de las proteínas

La organización de una proteína viene definida por cuatro niveles estructurales denominados: estructura primaria, secundaria, terciaria y cuaternaria, las cuales se muestran en la Figura 6.5. Cada una de estas estructuras informa la disposición de la anterior en el espacio [19].

- La estructura primaria: Es la secuencia de aminoácidos de la proteína, indica qué aminoácidos componen la cadena polipeptídica y el orden en que dichos aminoácidos se encuentran. La función de una proteína depende de su secuencia y de la forma que ésta adopte.
- La estructura secundaria: Es la disposición de la secuencia de aminoácidos en el espacio. Los aminoácidos a medida que van siendo enlazados, adquieren una disposición espacial estable, la estructura secundaria. Existen dos tipos de estructura: la a(alfa) o hélice y la conformación beta.

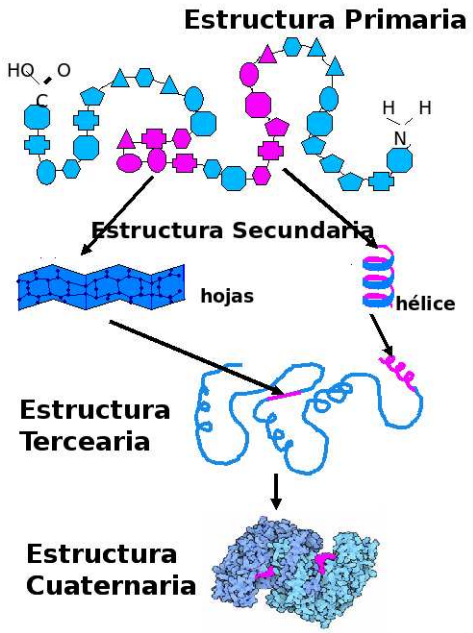



Figura 6.5: Estructuras de una proteína.

- La estructura terciaria: Informa sobre la disposición de la estructura secundaria de un polipéptido al plegarse sobre sí misma. Esta conformación se mantiene estable gracias a la existencia de enlaces (interacciones intramoleculares) entre los radicales R de los aminoácidos. Aquí aparecen varios tipos de enlaces:

- el puente disulfuro entre los radicales de aminoácidos que tiene azufre.
- los puentes de hidrógeno
- los puentes eléctricos
- las interacciones hidrófobas [66].

- Estructura cuaternaria: Informa de la unión, mediante enlaces débiles de varias cadenas polipeptídicas con estructura terciaria, para formar un complejo proteico.

## C.2. Propiedades de las proteínas

1. Especificidad: se refiere a que cada proteína lleva a cabo una función determinada y la realiza, debido a que posee una estructura primaria determinada y una conformación espacial propia; por lo que un cambio en la estructura de la proteína puede significar una pérdida de la función.

2. Desnaturalización: pérdida de la estructura terciaria, por romperse los puentes que forman dicha estructura. Una proteína soluble en agua cuando se desnaturaliza se hace insoluble en agua y precipita. La desnaturalización se puede producir por cambios de temperatura o variaciones del pH. En algunos casos, las proteínas desnaturalizadas pueden volver a su estado original a través de un proceso llamado renaturalización.

3. Se analiza la influencia del tipo de residuo y de la estructura en accesibilidad al solvente. Se define una medida de relativa exposición a la hidrofobicidad, en conjunto con la estructura secundaria como parámetros de predicción. Primero se describen los análisis usados para determinar los parámetros para el algoritmo de la predicción de interfaz; éstos incluyen accesibilidad o información estructural tal como la interacción entre estructuras beta plegadas o estructuras helicoidales. Con estos datos podría lograrse la identificación de las regiones superficiales implicadas en interacciones proteína-proteína.

4. Solubilidad: esta propiedad se mantiene siempre y cuando los enlaces fuertes y débiles estén presentes. Si se aumenta la temperatura y el pH, se pierde la solubilidad.

5. Capacidad Electrolítica: se determina a través de la electrólisis, en la cual si las proteínas se trasladan al polo positivo es porque su radical tiene carga negativa y viceversa.

### C.3. Funciones de las proteínas

Algunas de las funciones que podemos mencionar son:

- Estructural o de soporte, son proteínas que participan en la formación de la sangre, piel, uñas y constituyen la estructura de muchos tejidos de soporte del organismo, como los tendones y los huesos.
- Enzimática, son las más numerosas y especializadas. Actúan como biocatalizadores de las reacciones metabólicas, tener la capacidad de aumentar la velocidad de una reacción, en al menos un millón de veces.

  Son un elemento esencial para el crecimiento y el mantenimiento de los tejidos de todo el cuerpo.
- Hormonal, son proteínas que participan en la regulación de procesos metabólicos
- Represoras son elementos importantes dentro del proceso de transmisión de la información genética en la biosíntesis de otras moléculas.
- Anticuerpos, son proteínas altamente específicas que tienen la capacidad de identificar sustancias extrañas como: los virus, las bacterias y las células de otros organismos.
- Transporte, proteínas específicas transportan muchos iones y moléculas específicas. Un ejemplo es la hemoglobina que transporta el oxígeno y una porción del gas carbónico desdes y hacia los pulmones, respectivamente.
- Movimiento, tienen la capacidad de modificar su estructura en relación con cambios en el ambiente electroquímico que las rodea y producir a nivel macro el efecto de una contracción muscular.

## D: Ejemplo Flor

A continuación, para esclarecer el procedimiento a seguir en la selección de características discriminantes se presenta el siguiente ejemplo.

El problema a solucionar es identificar dos clases desde un conjunto de datos (divididos por características), los cuales corresponden a las Flores tipo Setosa y cuales a las del tipo Versicolor. Se tienen 50 muestras de la clase Setosa y 50 muestras de la clase Versicolor. Los datos

utilizados se encuentran disponibles en la base de datos Machine Learning repository[1] (UCI). Cada muestra contiene cuatro características que permiten la futura clasificación. Están son: longitud de sépalo y pétalo, ancho de sépalo y pétalo (las partes de la flor mencionadas se especifican en la Figura 6.6. Las características a representar son: longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo, todo esto en centímetros.

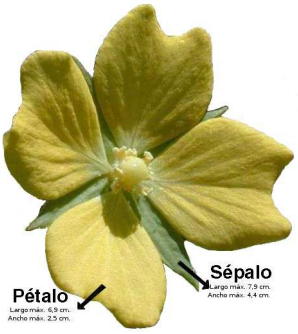



Figura 6.6: Partes de una Flor.

Si se tienen dos clases en el grupo, con 4 características, se pueden definir las diferencias entre las clases. Sin embargo, si se evalúan sólo dos características, también se pueden diferenciar las dos clases, más aún, con una sola característica también se podría distinguir diferencias entre clases. La finalidad, es encontrar un subconjunto de características que clasifican correctamente (o con un alto grado de precisión) futuras muestras.

En la figura 6.7, se muestran graficadas las características 2 (eje X) y 3 (eje Y) de las clases Setosa y Versicolor. Cada muestra esta representada por un punto de un color, según su clase: Setosa es azul y Versicolor es rojo.

En el gráfico 6.7(a) se representan ambas características, de lo cual se puede indicar que existe una alta precisión para clasificar las clases, con 100 % para los datos de entrenamiento. En el caso del gráfico 6.7(b) se usa solamente la característica 3, en cuyo caso, las muestras están ubicadas en el eje Y. Mientras que en esta representación la precisión de clasificación es alta, casi igual a la original (a), con 100 % para los datos de entrenamiento, en el caso 6.7(c) se usa solamente la característica 2, en cuyo caso, las muestras están ubicadas en el eje X, donde se confunden, por lo tanto, la precisión de clasificación no es tan alta como en el caso original (a), ni como en el caso 6.7(b), donde se usa una sola característica.

[1] http://www.ics.uci.edu/~mlearn/MLSummary.html

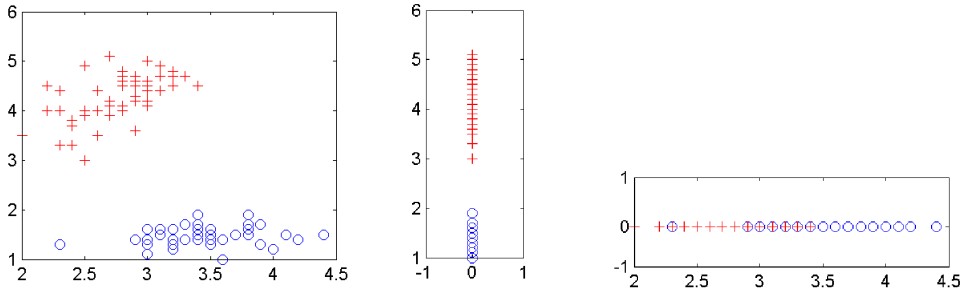


Figura 6.7: Representación gráfica de las clases .

Como conclusión, la característica 3 permite clasificar los datos sin perder precisión, mientras que, usando la característica 2 las clases se confunden y la precisión de clasificación es bastante baja. Se puede indicar que con un subconjunto de características se puede discriminar entre datos específicos sin perder precisión en la clasificación.

# E: Pasos para la obtención de los datos

Al tener identificados los complejos y las clases a utilizar (transitorios y permanentes), además de la base de datos donde se recuperará la información tridimensional de estos. Se procede a identificar los datos de los cuales se compone el listado que se encuentra en un archivo de texto plano. La información por complejo se compone de: nombre del complejo, bajo la codificación utilizada en la base de datos Protein Data Bank (PDB) y las cadenas que participan en la interacción. Esto se puede observar en la Figura 6.8.

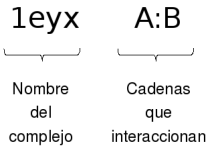



Figura 6.8: Estructura del listado de complejos, de Mintseris.

Con el nombre, se debe recuperar la información estructural de cada uno desde PDB. La forma tradicional es a través de 4 pasos, que se detallan y también se pueden observar en la Figura 6.9:

1. ingresar código del complejo,
2. ingresar a la sección de descargas,
3. escoger el archivo a descargar,
4. ubicar el sitio donde se almacenará el archivo.

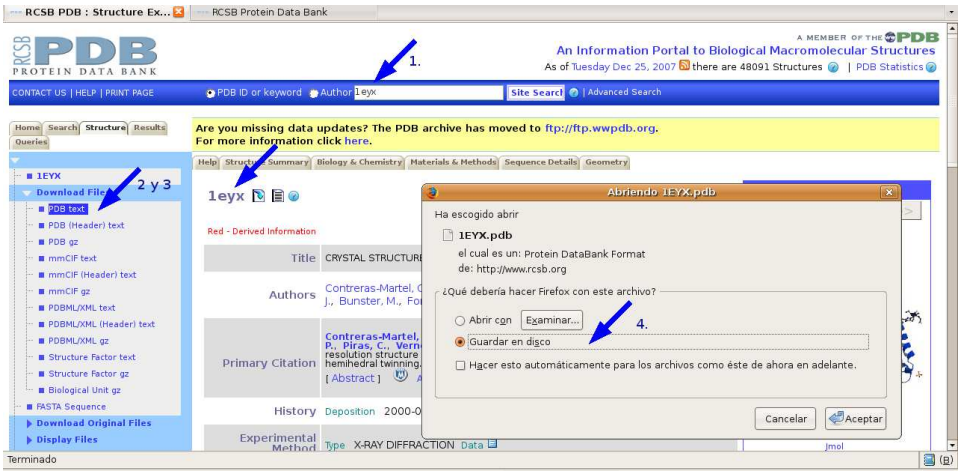



Figura 6.9: Descargar un complejo desde PDB.

El problema de realizar estos pasos, no es la complejidad de efectuarlos, sino el tiempo requerido (poco eficiente) para realizar cada uno de los complejos (326 x 4). Para esto se trabajo en ambiente Shell (línea de comandos), con un pequeño script iterativo para cada complejo, accesando directamente al servidor, esto redujo de manera considerable el tiempo requerido para esta de recuperación de información de los complejos.

# F: Fastcontact

La estructura de los archivos obtenidos desde el Protein data bank, son depurados, pero se debe mantener los formatos de carácter biológico que solicita la aplicación Fast contact, los cuales se encuentran especificados en un archivo adjunto (`charmm19.rtf`) al programa. La ejecución para el programa es a través de Shell:

```
/fastcontact.x charmm19.rtf cadena1.pdb cadena2.pdb 0 0 0 >salida.txt
```

La información entregada por la aplicación Fast Contact debe también ser recuperada y estructurada, para crear la matriz de trabajo. En la Tabla 6.3 se muestra los diferentes energías que entrega los archivos de salida por complejo, además de algunos ejemplos de valores.

| | |
|---|---|
| Geometrical Center | 11.7963324 44.4886757 -35.3821759 |
| Geometrical Center | 41.3511282 20.4744533 -34.5945198 |
| CONFORMATION NUMBER 0: | 1a6d-2.pdb |
| Desolvation Free Energy: | 1.31434864 |
| Electrostatic (4r) Energy: | -53.0412104 |
| Top **20** Min & **20** Max residues contributing to the binding free energy : | -10.622 496 ARG ... (Min) |
| | 0.247 168 ARG ... (Max) |
| Top 20 Min & Max ligand residues contributing to the desolvation free energy | -3.370 501 ILE ... (Min) |
| | 0.329 257 GLU ... (Max) |
| Top 20 Min & Max ligand residues contributing to the electrostatics energy | -14.688 496 ARG ... (Min) |
| | 0.228 240 ASP ... (Max) |
| Top 20 Min & Max receptor residues contributing to the desolvation free energy | -4.173 34 MET ... (Min) |
| | 0.455 285 TYR... (Max) |
| Top 20 Min & Max receptor residues contributing to the electrostatics energy | -11.528 33 LYS ... (Min) |
| | 0.357 248 ASP ... (Max) |
| Top 20 Min & Max receptor-ligand residue electrostatic contacts | -9.711 32 ASP 496 ARG ... (Min) |
| | 0.746 463 ASP 105 GLU ... (Max) |
| Top 20 Min & Max receptor-ligand residue free energy contacts | -8.885 253 GLU 228 LYS ... (Min) |
| | 0.975 237 LYS 229 LYS ... (Max) |

Tabla 6.3: Ejemplo de archivo de salida de la aplicación Fast Contact.

# G: Procesos utilizados en la Metodología

## G.1. Recopilación de datos

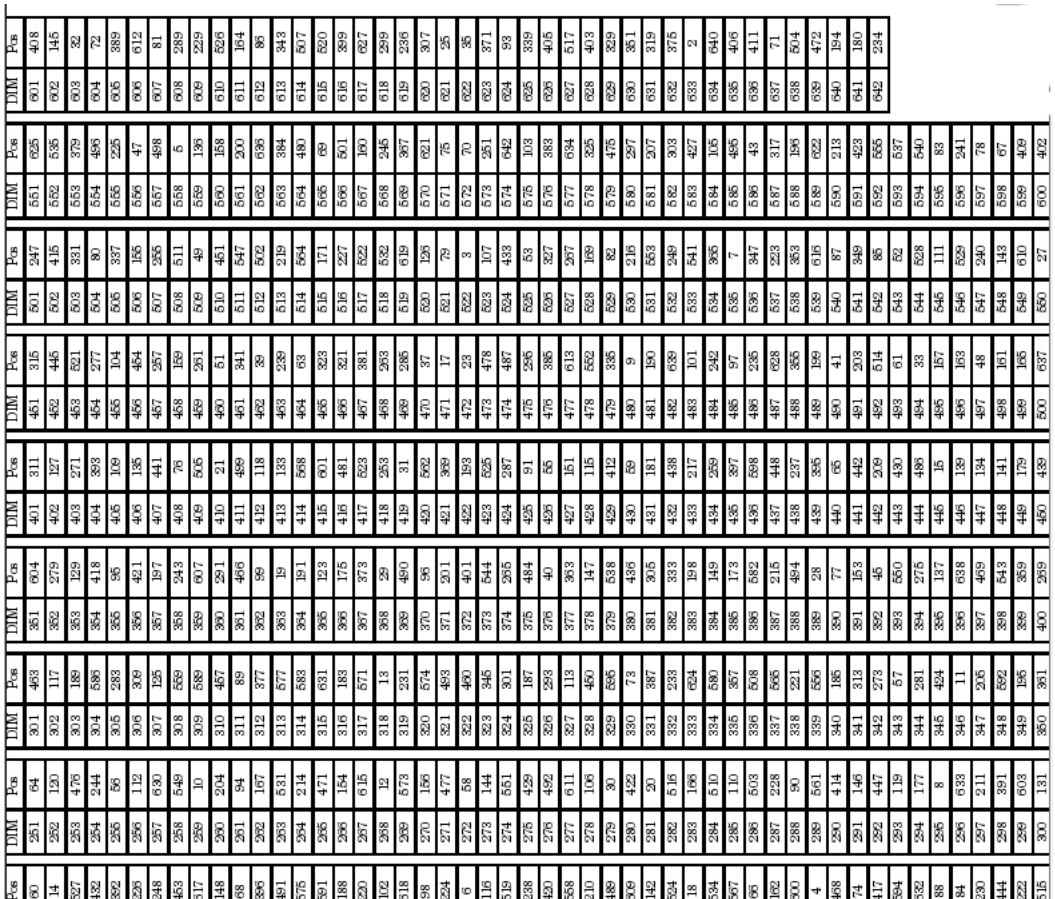
En la Figura 6.10, se muestran ejemplos de los valores de diferentes características que se utilizaron. Cada columna es una característica en la matriz final. Excepto las columnas que hacen referencia a el aminoácido que participa en la interacción, debido a que no es información de utilidad para la clasificación de las clases. Estos aminoácidos si son relevante para la etapa de interpretación de los resultados, por esta razón la información de aminoácidos de guarda.

| Formato | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Línea | Valor | Residuo | aa | Línea | Valor | Residuo | aa | Residuo | aa |
| 9 | -1549 | 266 | PRO | 9 | -1549 | 266 | PRO | 255 | PRO |

| Formato | | | | | | |
|---|---|---|---|---|---|---|
| Línea | Valor | Residuo | Línea | Valor | Residuo | Residuo |
| 9 | -1549 | 266 | 9 | -1549 | 266 | 255 |

Figura 6.10: Ejemplo de valores de características.

## G.2. Selección de características: Ranking del total de características.



| DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos | DIM | Pos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 121 | 51 | 370 | 101 | 170 | 151 | 500 | 201 | 60 | 251 | 64 | 301 | 463 | 351 | 604 | 401 | 311 | 451 | 315 | 501 | 247 | 551 | 625 | 601 | 408 |
| 2 | 282 | 52 | 270 | 102 | 456 | 152 | 542 | 202 | 14 | 252 | 120 | 302 | 117 | 352 | 279 | 402 | 127 | 452 | 445 | 502 | 415 | 552 | 535 | 602 | 145 |
| 3 | 100 | 53 | 465 | 103 | 250 | 153 | 232 | 203 | 627 | 253 | 476 | 303 | 189 | 353 | 129 | 403 | 271 | 453 | 621 | 503 | 331 | 553 | 379 | 603 | 32 |
| 4 | 470 | 54 | 678 | 104 | 388 | 154 | 602 | 204 | 432 | 254 | 244 | 304 | 586 | 354 | 418 | 404 | 393 | 454 | 277 | 504 | 80 | 554 | 486 | 604 | 72 |
| 5 | 364 | 55 | 368 | 105 | 108 | 155 | 140 | 205 | 382 | 255 | 66 | 305 | 283 | 355 | 95 | 405 | 109 | 455 | 104 | 505 | 337 | 555 | 225 | 605 | 389 |
| 6 | 314 | 56 | 186 | 106 | 276 | 156 | 606 | 206 | 225 | 256 | 112 | 306 | 309 | 356 | 421 | 406 | 135 | 456 | 454 | 506 | 155 | 556 | 47 | 606 | 612 |
| 7 | 346 | 57 | 334 | 107 | 382 | 157 | 479 | 207 | 248 | 257 | 630 | 307 | 125 | 357 | 197 | 407 | 441 | 457 | 257 | 507 | 255 | 557 | 498 | 607 | 81 |
| 8 | 407 | 58 | 305 | 108 | 584 | 158 | 426 | 208 | 463 | 258 | 549 | 308 | 569 | 358 | 243 | 408 | 76 | 458 | 159 | 508 | 511 | 558 | 5 | 608 | 289 |
| 9 | 288 | 59 | 382 | 109 | 318 | 159 | 22 | 209 | 617 | 259 | 10 | 309 | 589 | 359 | 607 | 409 | 505 | 459 | 261 | 509 | 49 | 559 | 136 | 609 | 229 |
| 10 | 298 | 60 | 458 | 110 | 587 | 160 | 641 | 210 | 148 | 260 | 204 | 310 | 467 | 360 | 291 | 410 | 21 | 460 | 61 | 510 | 461 | 560 | 158 | 610 | 626 |
| 11 | 342 | 61 | 300 | 111 | 597 | 161 | 92 | 211 | 68 | 261 | 94 | 311 | 89 | 361 | 466 | 411 | 469 | 461 | 941 | 511 | 547 | 561 | 200 | 611 | 164 |
| 12 | 330 | 62 | 635 | 112 | 576 | 162 | 218 | 212 | 396 | 262 | 167 | 312 | 377 | 362 | 99 | 412 | 118 | 462 | 39 | 512 | 502 | 562 | 636 | 612 | 86 |
| 13 | 360 | 63 | 328 | 113 | 128 | 163 | 467 | 213 | 491 | 263 | 531 | 313 | 577 | 363 | 19 | 413 | 133 | 463 | 239 | 513 | 219 | 563 | 384 | 613 | 343 |
| 14 | 290 | 64 | 378 | 114 | 599 | 164 | 404 | 214 | 575 | 264 | 214 | 314 | 583 | 364 | 191 | 414 | 568 | 464 | 63 | 514 | 564 | 564 | 480 | 614 | 507 |
| 15 | 557 | 65 | 629 | 115 | 608 | 165 | 122 | 215 | 691 | 265 | 471 | 315 | 631 | 365 | 123 | 415 | 601 | 465 | 823 | 515 | 171 | 565 | 69 | 615 | 620 |
| 16 | 348 | 66 | 246 | 116 | 563 | 166 | 530 | 216 | 188 | 266 | 164 | 316 | 183 | 366 | 175 | 416 | 481 | 466 | 321 | 516 | 227 | 566 | 501 | 616 | 399 |
| 17 | 625 | 67 | 16 | 117 | 465 | 167 | 44 | 217 | 220 | 267 | 615 | 317 | 671 | 367 | 373 | 417 | 623 | 467 | 381 | 517 | 622 | 567 | 160 | 617 | 627 |
| 18 | 325 | 68 | 294 | 118 | 518 | 168 | 192 | 218 | 102 | 268 | 12 | 318 | 13 | 368 | 29 | 418 | 253 | 468 | 263 | 518 | 532 | 568 | 245 | 618 | 299 |
| 19 | 292 | 69 | 358 | 119 | 380 | 169 | 485 | 219 | 618 | 269 | 573 | 319 | 231 | 369 | 490 | 419 | 31 | 469 | 285 | 519 | 619 | 569 | 367 | 619 | 236 |
| 20 | 338 | 70 | 296 | 120 | 461 | 170 | 80 | 220 | 98 | 270 | 156 | 320 | 574 | 370 | 96 | 420 | 562 | 470 | 37 | 520 | 126 | 570 | 621 | 620 | 307 |
| 21 | 308 | 71 | 434 | 121 | 536 | 171 | 564 | 221 | 224 | 271 | 477 | 321 | 483 | 371 | 201 | 421 | 369 | 471 | 17 | 521 | 79 | 571 | 76 | 621 | 25 |
| 22 | 304 | 72 | 620 | 122 | 24 | 172 | 581 | 222 | 6 | 272 | 58 | 322 | 460 | 372 | 401 | 422 | 193 | 472 | 23 | 522 | 3 | 572 | 70 | 622 | 35 |
| 23 | 124 | 73 | 42 | 123 | 316 | 173 | 449 | 223 | 116 | 273 | 144 | 323 | 345 | 373 | 544 | 423 | 625 | 473 | 478 | 523 | 107 | 573 | 251 | 623 | 371 |
| 24 | 260 | 74 | 428 | 124 | 446 | 174 | 569 | 224 | 619 | 274 | 561 | 324 | 301 | 374 | 265 | 424 | 287 | 474 | 487 | 524 | 433 | 574 | 642 | 624 | 93 |
| 25 | 362 | 75 | 138 | 125 | 366 | 175 | 176 | 225 | 238 | 275 | 429 | 325 | 187 | 375 | 484 | 425 | 91 | 475 | 245 | 525 | 53 | 575 | 103 | 625 | 339 |
| 26 | 278 | 76 | 437 | 126 | 376 | 176 | 588 | 226 | 420 | 276 | 492 | 326 | 293 | 376 | 40 | 426 | 85 | 476 | 385 | 526 | 327 | 576 | 383 | 626 | 405 |
| 27 | 36 | 77 | 114 | 127 | 62 | 177 | 45 | 227 | 568 | 277 | 611 | 327 | 113 | 377 | 363 | 427 | 151 | 477 | 613 | 527 | 297 | 577 | 694 | 627 | 517 |
| 28 | 372 | 78 | 280 | 128 | 435 | 178 | 560 | 228 | 210 | 278 | 106 | 328 | 450 | 378 | 147 | 428 | 115 | 478 | 552 | 528 | 169 | 578 | 325 | 628 | 403 |
| 29 | 419 | 79 | 258 | 129 | 178 | 179 | 398 | 229 | 489 | 279 | 30 | 329 | 695 | 379 | 538 | 429 | 412 | 479 | 335 | 529 | 82 | 579 | 475 | 629 | 329 |
| 30 | 416 | 80 | 440 | 130 | 208 | 180 | 488 | 230 | 609 | 280 | 422 | 330 | 73 | 380 | 436 | 430 | 69 | 480 | 9 | 530 | 216 | 580 | 297 | 630 | 361 |
| 31 | 340 | 81 | 362 | 131 | 254 | 181 | 546 | 231 | 142 | 281 | 20 | 331 | 387 | 381 | 305 | 431 | 181 | 481 | 190 | 531 | 553 | 581 | 207 | 631 | 319 |
| 32 | 605 | 82 | 150 | 132 | 497 | 182 | 570 | 232 | 624 | 282 | 516 | 332 | 233 | 382 | 333 | 432 | 438 | 482 | 639 | 532 | 249 | 582 | 303 | 632 | 375 |
| 33 | 396 | 83 | 168 | 133 | 473 | 183 | 539 | 233 | 18 | 283 | 166 | 333 | 624 | 383 | 198 | 433 | 217 | 483 | 101 | 533 | 541 | 583 | 427 | 633 | 2 |
| 34 | 606 | 84 | 623 | 134 | 548 | 184 | 509 | 234 | 594 | 284 | 510 | 334 | 580 | 384 | 149 | 434 | 259 | 484 | 242 | 534 | 365 | 584 | 105 | 634 | 640 |
| 35 | 344 | 85 | 312 | 135 | 94 | 185 | 590 | 235 | 567 | 285 | 110 | 335 | 357 | 385 | 173 | 435 | 397 | 485 | 97 | 535 | 7 | 585 | 495 | 635 | 406 |
| 36 | 302 | 86 | 566 | 136 | 512 | 186 | 533 | 236 | 66 | 286 | 503 | 336 | 508 | 386 | 582 | 436 | 698 | 486 | 235 | 536 | 347 | 586 | 43 | 636 | 411 |
| 37 | 262 | 87 | 324 | 137 | 64 | 187 | 443 | 237 | 162 | 287 | 228 | 337 | 665 | 387 | 215 | 437 | 448 | 487 | 628 | 537 | 223 | 587 | 317 | 637 | 71 |
| 38 | 310 | 88 | 298 | 138 | 38 | 188 | 212 | 238 | 600 | 288 | 90 | 338 | 221 | 388 | 494 | 438 | 237 | 488 | 365 | 538 | 353 | 588 | 196 | 638 | 504 |
| 39 | 464 | 89 | 184 | 139 | 254 | 189 | 394 | 239 | 4 | 289 | 561 | 339 | 556 | 389 | 28 | 439 | 335 | 489 | 129 | 539 | 616 | 589 | 622 | 639 | 472 |
| 40 | 332 | 90 | 394 | 140 | 182 | 190 | 132 | 240 | 468 | 290 | 414 | 340 | 185 | 390 | 77 | 440 | 65 | 490 | 41 | 540 | 87 | 590 | 213 | 640 | 194 |
| 41 | 266 | 91 | 410 | 141 | 614 | 191 | 579 | 241 | 74 | 291 | 146 | 341 | 313 | 391 | 153 | 441 | 442 | 491 | 203 | 541 | 349 | 591 | 423 | 641 | 180 |
| 42 | 272 | 92 | 462 | 142 | 400 | 192 | 322 | 242 | 417 | 292 | 447 | 342 | 273 | 392 | 45 | 442 | 209 | 492 | 514 | 542 | 85 | 592 | 565 | 642 | 234 |
| 43 | 425 | 93 | 172 | 143 | 202 | 193 | 585 | 243 | 694 | 293 | 119 | 343 | 57 | 393 | 550 | 443 | 430 | 493 | 61 | 543 | 82 | 593 | 537 | | |
| 44 | 386 | 94 | 284 | 144 | 274 | 194 | 513 | 244 | 632 | 294 | 177 | 344 | 281 | 394 | 275 | 444 | 496 | 494 | 33 | 544 | 628 | 594 | 540 | | |
| 45 | 296 | 95 | 374 | 145 | 474 | 195 | 482 | 245 | 88 | 295 | 8 | 345 | 424 | 395 | 137 | 445 | 15 | 495 | 157 | 545 | 111 | 595 | 83 | | |
| 46 | 356 | 96 | 320 | 146 | 26 | 196 | 596 | 246 | 84 | 296 | 633 | 346 | 11 | 396 | 638 | 446 | 139 | 496 | 163 | 546 | 629 | 596 | 241 | | |
| 47 | 413 | 97 | 380 | 147 | 459 | 197 | 572 | 247 | 230 | 297 | 211 | 347 | 205 | 397 | 469 | 447 | 134 | 497 | 48 | 547 | 240 | 597 | 78 | | |
| 48 | 174 | 98 | 545 | 148 | 431 | 198 | 206 | 248 | 444 | 298 | 391 | 348 | 692 | 398 | 543 | 448 | 141 | 498 | 161 | 548 | 143 | 598 | 67 | | |
| 49 | 350 | 99 | 693 | 149 | 483 | 199 | 1 | 249 | 222 | 299 | 603 | 349 | 195 | 399 | 359 | 449 | 179 | 499 | 165 | 549 | 610 | 599 | 409 | | |
| 50 | 286 | 100 | 130 | 150 | 452 | 200 | 162 | 250 | 515 | 300 | 131 | 350 | 361 | 400 | 269 | 450 | 439 | 500 | 637 | 550 | 27 | 600 | 402 | | |

Ranking de las 642 características, orden de acuerdo a la relevancia en la discriminación entre las clases, para esto se utilizó el criterio de selección, aplicado en la sección 5.3.2. De este listado se toman las primeras 20 características para los pasos siguientes en la clasificación y posterior interpretación.

# H: Generación de datos en el área de la Biología

El Proyecto Genoma Humano (PGH) [24] genera Terabytes de datos. Debido a esta gran cantidad de datos generados, fue necesario el desarrollo de nuevas metodologías de análisis con la utilización de herramientas que permitan acceder, manejar, distribuir el volumen y complejidad de estos datos.

Una forma de trabajar esta información es con la creación de base de datos que centralicen los datos y posibiliten un mejor manejo de ella. Un ejemplo de estas base de datos es GenBank [8, 9], base de datos pública de secuencias (con alrededor de con 65.369.091.950 bases en 61.132.599 secuencias almacenadas, con el método tradicional de Genbank y 80.369.977.826 bases en 17.960.667 secuencias almacenadas en el método de división WGS (85.759.586.764 bases, desde 82.853.685 secuencias reportadas al 15 Febrero de 2008).

GenBank es actualizada por los autores de cada proyecto que ha descubierto una secuencia nueva, es decir, ellos son los encargados de nutrir la base de datos con los resultados de sus investigaciones, las cuales si se obtienen avances, deben aplicarlos en la base de datos GenBank.

Por otra parte esto ha contribuido a disminuir la diferencia entre el número de secuencias conocidas y el de estructuras.

# I: Sitios disponibles para descargar información biológica.

Algunas de estas aplicaciones son programas y otras servidores, en la Tabla 6.4, se muestra un listado de algunos sitios disponibles en Internet para obtener información de interacción de proteínas y algunos específicamente (las tres primeras) de energías de interacción.

| **Aplicaciones** | **Sitio http://** |
|---|---|
| FastContact 2.0 | structure.pitt.edu/servers/fastcontact/ |
| PPI | www.bioinformatics.leeds.ac.uk/ppi_pred |
| Foldx | foldx.crg.es/ |
| InterProSurf | curie.utmb.edu/prosurf.html |
| STRING is a. | string.embl.de/ |
| Interface | 202.141.148.29/resources/bioinfo/interface/ |
| MIPS Mammalian PPI DB | mips.gsf.de/proj/ppi/ |
| Agile Protein Interaction DataAnalyzer | bioinfow.dep.usal.es/apid/index.htm |
| SCOWLP | www.scowlp.org |
| **Bases de datos** | **Sitio http://** |
| Human Protein Reference Database | www.hprd.org/ |
| IntAct Home | www.ebi.ac.uk/intact/site/index.jsf |
| JCB | www.imb-jena.de/jcb/ppi/ |
| Protein Interaction Database | www.proteinlounge.com/inter_home.asp |
| Database of Interacting Proteins | dip.doe-mbi.ucla.edu/ |
| BioGRID | www.thebiogrid.org/ |
| Database of protein-protein complexes. | www.ces.clemson.edu/compbio/databases/complexes/Search.htm |

Tabla 6.4: Sitios disponibles para trabajar con IPP.

# Glosario

**Biología molecular (BM).** Es parte de la Biología que estudia los seres vivos y los fenómenos vitales con arreglo a las propiedades de su estructura molecular. La BM está dedicada al estudio de los mecanismos moleculares y genéticos implicados en los procesos biológicos fundamentales en el desarrollo y fisiología de los organismos vivos. El polimorfismo de nucleótido único (SNPs o single nucleotide polymorphism) es un objetivo importante de la BM aplicada al estudio de la evolución. Se trata de puntos concretos de los genomas en los que un nucleótido puede ser diferente en varios individuos, dando lugar a caracteres diferentes, como el color de los ojos, de la piel, del pelo, la forma de la nariz, las forma en que metabolizamos sustancias, etc. Un buen ejemplo de SNP son los alelos de los grupos sanguíneos humanos [61].

**Enlace peptídico.** Tiene lugar mediante la pérdida de una molécula de agua entre el grupo amino de un aminoácido y el carboxilo de otro, el resultado es un enlace covalente CO-NH. Es decir, un enlace amida sustituido. Podemos seguir añadiendo aminoácidos al péptido, porque siempre hay un extremo NH2 terminal y un COOH terminal.

**Genómica.** Involucra el conocimiento y análisis de genomas completos. Permite establecer relaciones entre BD de genomas, evolutivas, funcional y estructural.

**Genoma.** Es el número total de cromosomas, todo el ADN (ácido desoxirribonucleico) de un organismo, incluídos sus genes, los cuales llevan la información para la elaboración de todas las proteínas requeridas por el organismo, y las que determinan el aspecto, el funcionamiento, el metabolismo, la resistencia a infecciones y otras enfermedades, y también algunos de sus procederes.

**Homología.** Decimos que hay una relación de homología si hay un parecido que demuestra un origen evolutivo común. Que la homología es transitiva quiere decir que si una proteína A es homóloga a otra B, y B es homóloga a C, entonces A es homóloga a C,

aunque no se parezcan. Sin embargo, las proteínas están constituidas por dominios que aparecen en múltiples combinaciones y esto puede hacer que la transitividad no sea aplicable; podemos matizar así: la homología es transitiva a nivel de dominios.

**Nucleótidos.** Molécula formada por una base nitrogenada, una pentosa y una molécula de ácido fosfórico. Constituye la base de los ácidos nucleicos, ADN y ARN. El nucleótido se considera un monómero, mientras que los ácidos nucleicos serían los polímeros.

**Proteómica.** Involucra el conocimiento y análisis de proteínas presentes en las células y corresponden a las proteínas producto de expresión en un estado fisiológico determinado. Establecer relaciones entre BD y secuencias - estructuras evolutivas, funciones y estructural.

**Proyecto Genoma Humano (PGH).** Investigación científica internacional que busca seleccionar un modelo de organismo humano por medio del mapeo de la secuencia de su DNA.