Facultad de Ingeniería

Escuela de Ingeniería en Bioinformática

Predicción de función enzimática mediante las propiedades de los sitios de unión a ligando utilizando *Random Forest*.

Raúl Rubén Arias Carrasco

Profesor Tutor: José Antonio Reyes

Profesor Co-Tutor: Mauricio Arenas

Profesor Informante: Jans Alzate

Memoria para optar al título de Ingeniero en Bioinformática

Talca-Chile

25 de febrero del 2013

AGRADECIMIENTOS

En primer lugar quiero agradecer a mi pareja Paulina Castillo ya que sin su apoyo incondicional, comprensión y muchas veces ayuda, me hubiese tardado mucho más en terminar esta memoria.

Por otro lado, quiero agradecer a mi familia que me ayudó a seguir mi vocación y siempre ha estado en todo momento apoyándome en mis decisiones.

Una de las mejores soluciones al estrés es la alegría y los buenos momentos que tuve con mis amigos, los cuales fueron de gran ayuda cuando la monotonía de trabajar todo el día en lo mismo me desmotivaba.

Por último, sin olvidar a mis tutores que gracias a su ayuda y orientación, esta memoria logró cumplir todos los objetivos.

A la Ciencia**.**

INDICE DE CONTENIDOS

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

Las enzimas desempeñan un papel importante en los organismos vivos. Son conocidas para catalizar reacciones bioquímicas, pero su función específica es variable dependiendo de sus propiedades bioquímicas. Las enzimas son generalmente categorizadas en una clase funcional determinada, de acuerdo con el esquema de la Comisión de Enzimas (*EC*). El esquema *EC* es una clasificación funcional jerárquica de cuatro niveles, donde las clases de enzimas son asignadas en 4 números. El primer número representa una de las 6 clases principales de las reacciones bioquímicas que las enzimas catalizan. Estas son: (1) las oxidorreductasas, transferasas (2), hidrolasas (3), liasas (4), isomerasas (5) y ligasas (6).

Recientes avances en las tecnologías de secuenciación han mostrado un crecimiento exponencial en las secuencias de proteínas. Además, en años recientes, el número de estructuras de proteínas cristalizadas también está aumentando en una tasa alta. Más de 3.100 proteínas con función desconocida han sido depositadas en la base de dato *Protein Data Bank* (*PDB*) y este número está aumentando exponencialmente. La mayoría de las enzimas son proteínas y consecuentemente para nuevas proteínas encontradas es importante identificar su función biológica, por ejemplo, uno de los 6 números *EC*. La clasificación automática de la función enzimática ha ganado mucho interés en los últimos años, principalmente utilizando algoritmos de *Machine Learning*. Los principales enfoques se basan en la similitud de secuencia, la semejanza estructural o la combinación de ambos tipos de características. Sin embargo, se ha demostrado recientemente que la función enzimática es menos conservadas que las expectativas anteriores.

Esta investigación se centra en la predicción de las seis clases funcionales de enzimas definidas por el primer nivel del esquema *EC*, basado únicamente en las estructuras de proteínas cristalizadas. Pero esta investigación se centra en las propiedades conocidas de sitios de unión a ligando de proteínas cristalizadas. Con esta información fue posible obtener un modelo basado en *Random Forest* que

alcanza una exactitud de 90% en las 10 validaciones cruzadas y tasas de falsos positivos que no exceden el 4%. Por lo tanto, este estudio muestra que las propiedades de los sitios de unión a ligandos de proteínas están directamente relacionados con la función enzimática.

ABSTRACT

Enzymes play an important role on living organisms. They are known to catalyze biochemical reactions, but their specific function is variable depending on their biochemical properties. Enzymes are usually categorized into a certain functional class, according to the Enzyme Commission (*EC*) scheme. The *EC* scheme is a four level hierarchical functional classification, where enzyme classes are assigned in 4 numbers. The first number represents one of the 6 main classes of the biochemical reactions that the enzymes catalyze. These are: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases and (6) ligases.

Recent advances in sequencing technologies have seen an exponential growth in protein sequences. In addition, in recent years, the number of crystallized protein structures is also increasing on a high rate. More than 3,100 proteins with unknown function are been deposited on the *Protein Data Bank* (*PDB*) database and this number is increasing exponentially. Most of enzymes are proteins and consequently for new found proteins is important to identify their biological function, for instance, one of the 6 EC numbers. The automatic classification of enzymatic function has gained much attention on recent years, mainly utilizing Machine Learning algorithms. The main approaches are based on sequence similarity, structural similarity or the combination of both of them types of features. However, it has been recently demonstrated that the enzymatic function is less conserved than previous expectations.

This research is focused on the prediction of the six enzyme functional classes defined by the first level of the EC scheme, based only on crystallized protein structures. But this research is focused on properties of known ligand-binding sites of crystallized proteins. With this information was possible obtain a model based on Random Forest which reaches an accuracy of 90% in the 10 cross-validations and false positive rates that do not exceed the 4%. Therefore, this study shows that the properties of the ligand binding sites of proteins are directly related to the enzymatic function.

INTRODUCCIÓN

El concepto de enzima (del griego: *en*, en; *zyme*, levadura) fue introducido en 1878 para describir que existía algo en las levaduras que permitía a éstas realizar la fermentación alcohólica. Biológicamente hablando, una enzima es definida por la función que desempeña dentro de los seres vivos. Esta función se denomina catalizador biológico (Voet & Voet, 2010), es decir que son capaces de aumentar la velocidad de reacciones bioquímicas respetando las leyes de la termodinámica, aunque difieren de las reacciones ordinarias en aspectos importantes, tales como:

- Mayores tasas de reacción: Las tasas de reacciones enzimáticamente catalizadas son típicamente factores de 10^6 a 10^{12} veces mayor que la misma reacción sin la acción enzimática.
- Condiciones de reacción más favorables: La mayoría de las enzimas realizan su función en condiciones fisiológicas. Por el contrario, en la naturaleza para que ocurra una reacción poco favorable se necesitan altas temperaturas o una presión muy elevada.
- Alta especificidad de reacción: Las enzimas tienen un alto grado de especificidad al identificar tanto sus sustratos, como sus productos.
- Capacidad regulatoria: La actividad catalítica de muchas enzimas varía en respuesta a otras moléculas, distintas a los sustratos.

A cada enzima, se le asigna un nombre acorde a la reacción que cataliza con su sustrato. La *Comisión de Enzimas* (*EC*) (Karlson y col., 1992) organiza todas las enzimas dentro de 6 grandes familias. Estas son: (1) *oxidorreductasas*, que catalizan reacciones de oxido-reducción; (2) *transferasas*, que transfieren grupos químicos; (3) *hidrolasas*, que hidrolizan varios tipos de enlaces; (4) *liasas*, que rompen enlaces por medios distintos a la hidrólisis; (5) *isomerasas*, que realizan cambios geométricos o estructurales en moléculas; (6) *ligasas*, que catalizan la unión entre moléculas.

La clasificación jerárquica *EC* asigna un único código de 4 campos para diferenciar la actividad enzimática. Como se muestra en la Figura 1 el primer dígito (en celeste) indica una de los 6 clases enzimáticas anteriormente expuestas. El

segundo número (en café) indica el grupo donador de electrones conocido como la subclase enzimática. El tercer número (en burdeo) indica el grupo receptor de electrones denominado como la sub-subclase enzimática. Por último, el cuarto dígito (en morado) indica el sustrato específico de la reacción enzimática.

Figura 1. Representación del código de clasificación *EC.* En la parte inferior se muestra un ejemplo de un código para la clase de las transferasas.

Esta nomenclatura entrega una base fundamental para la clasificación enzimática. Cabe destacar que este trabajo se enfocó en la predicción del primer dígito de la clasificación *EC* que está asociado a las 6 principales clases enzimáticas anteriormente descritas.

1. Predicción de función enzimática.

1.1. ¿Por qué predecir función enzimática?

La función de una enzima está dada por las composición aminoacídica y la disposición de estos residuos en el espacio, en donde se define la zona responsable de llevar a cabo la reacción química, conocida como el sitio catalítico (Cilia & Passerini, 2010). Más aún, es de vital importancia conocer los sitios de unión a moléculas de interés para lograr comprender la función que la proteina desempeña (Capra y col., 2009).

Toda proteína cristalizada es depositada en un gran repositorio mundial de proteínas llamado *Protein Data Bank* (*PDB*). Creado en los comienzos de este milenio por Berman y col. (Berman y col., 2000). En la actualidad, esta base de datos dispone de más de 86000 proteínas. Como se puede apreciar en la Figura 2 un

53,5% del total de proteínas pertenecen a la clasificación de enzimas. Estas pueden ser utilizadas como fuente de información para asignar función a nuevas proteínas cristalizadas que no la posean.

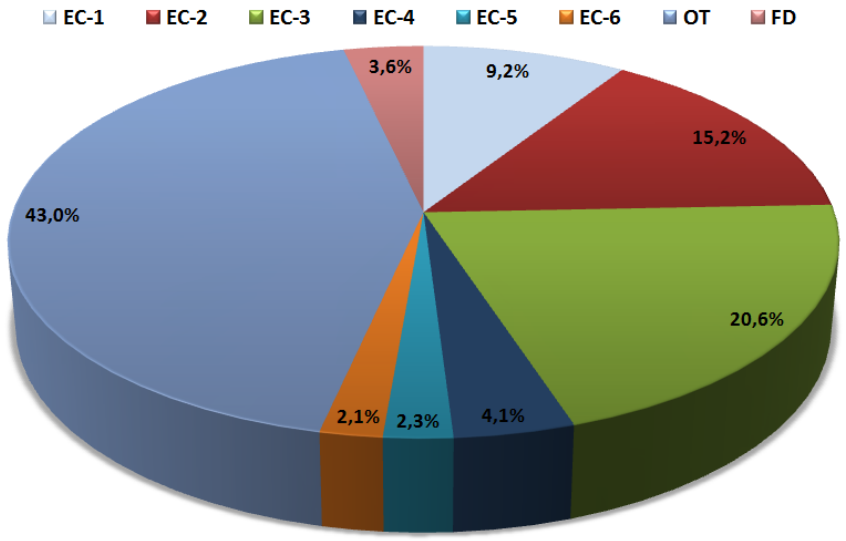

Figura 2. Distribución porcentual de las proteínas disponibles en *PDB*, de un total de 86.008. **EC-1**: Oxidorreductasas; **EC-2**: Transferasas; **EC-3**: Hidrolasas; **EC-4**: Liasas; **EC-5**: Isomerasas; **EC-6**: Ligasas; **OT**: Otras; **FD**: Función Desconocida. Datos obtenidos desde http://www.rcsb.org/pdb el 12 de noviembre del 2012 a las 9:26 am.

El poder asignar posibles funciones a estas proteínas podría entregar conocimiento útil para comprender procesos biológicos sin descifrar, así como también en el estudio de enfermedades o tener utilidades desde el punto de vista industrial en áreas como la biotecnología. Por ejemplo, el caso del desarrollo de biocombustibles, que necesitó años de investigación para generar toda una cascada de reacciones enzimáticas con el objetivo de ir mejorando el producto final (Jegannathan, 2011). Para lo cual, cada una de las enzimas involucradas en el desarrollo del biocombustible, fue primero simulada computacionalmente bajo una función enzimática predicha, para poder tener una aproximación del desempeño y funcionalidad de ésta y luego fue llevada a cabo la reacción en el laboratorio. Otra área de gran utilidad del predecir la función enzimática de las proteínas, es el diseño de drogas que puedan aumentar o disminuir su actividad biológica. El reciente hallazgo de enzimas involucradas en el cáncer (Griffith y col., 2010), es una aplicación directa del predecir la función enzimática. Por ejemplo, se pueden

desarrollar compuestos que inhiban las enzimas de interés frenando el aumento del crecimiento celular cancerígeno.

Una de las primeras metodologías para asignar la función enzimática es el traspaso de la clasificación *EC* por medio de la homología en la secuencia aminoacídica con una base de datos de enzimas correctamente clasificadas. Este método permite obtener rápidamente varios candidatos de función para la proteína de entrada. Utilizando una comparación de secuencias mediante un método de alineamiento como una variante de *BLAST* o *PSI-BLAST* (Altschul y col., 1997). En su estudio, (Audit y col., 2007) sugieren que las herramientas basadas en búsqueda por homología son suficientes para predecir el número *EC* más probable para una secuencia proteica de entrada. Sin embargo, si la secuencia no tiene un buen homólogo en la base de datos, es decir que tenga sobre el 50% de identidad en secuencia, se logra un resultado con incertidumbre acerca de la clasificación, ya que la predicción en la mayoría de los casos será errónea (Rost, 2002). Un año después, (Tian & Skolnick, 2003) concluyeron que con un 60% de identidad de secuencia se logra un 90% de exactitud al asignar la clasificación enzimática completa, siempre y cuando la base de datos fuera previamente limpiada de errores y redundancia. Esta última consideración es de relevancia ya que en las bases de datos públicas de secuencias existe un alto porcentaje de errores de anotación, producidos al usar sistemas autónomos que realizan una sobre anotación, es decir que transfieren más información desde una proteína modelo de lo que la evidencia dicta. A consecuencia de esto, se produce en el tiempo una propagación de errores en anotación, tal como lo demuestran en su estudio (Schnoes y col., 2009).

Considerando los antecedentes previos es que nace la necesidad de buscar otro método que permita determinar la función de una proteína. Lo que se propone en este estudio es utilizar la información de la interacción de las proteínas con sus ligandos (sustratos), logrando con esto comprender la función que podría desempeñar una proteína no caracterizada. En relación a estas últimas, se puede observar en la Figura 3 que desde la creación del repositorio de archivos de proteínas *PDB*, ha aumentando cada año el número de proteínas con función desconocida, llegando en la actualidad a un total de 3112.

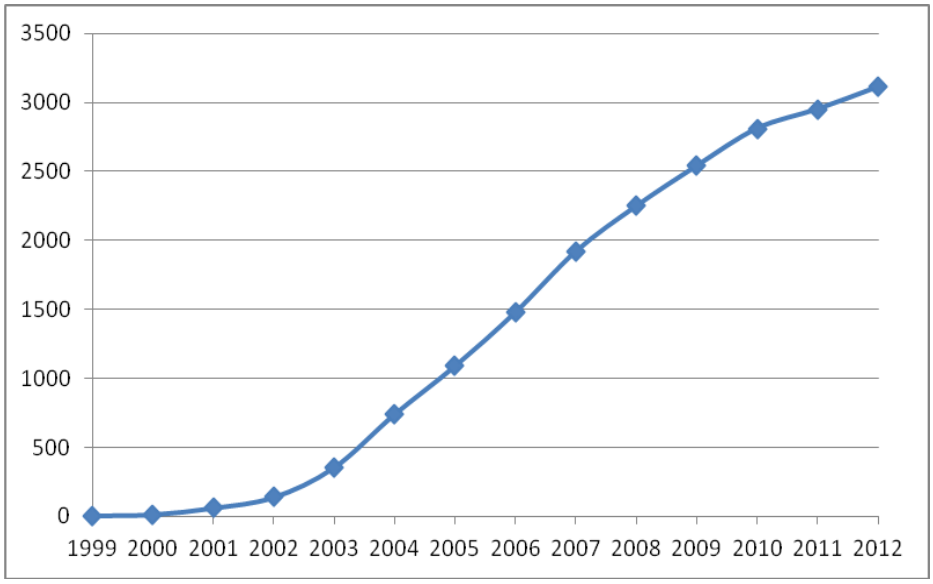

Figura 2. Aumento de la cantidad de proteínas con función desconocida. Datos obtenidos desde http://www.rcsb.org/pdb el 3 de enero del 2013 a las 11:13 am.

1.2. Predicción de función enzimática utilizando *Machine Learning*

1.2.1. *Machine Learning*

En el año 1959, Arthur L. Samuel introdujo por primera vez el concepto de *Machine Learning* (Bishop, 2006) como "Campo de estudio que otorga a las computadoras la capacidad de aprender sin ser explícitamente programadas". Décadas después, Mitchell (Mitchel & Hill, 1997) define el concepto como "un proceso en el cual un computador aprende de una experiencia E con respecto a una tarea de aprendizaje T, la cual es medida por una performance P y se busca que la performance P, mejore con la experiencia E". Por lo tanto, se puede decir en palabras computacionales que *Machine Learning* consiste en la programación de algoritmos para optimizar un criterio de aprendizaje, mediante el uso de datos (Larranaga, 2006). Toda tarea de clasificación consiste en primera instancia de obtener información de los ejemplos, caracterizados por medio de atributos. Con el objetivo de lograr con esta información generar modelos que permitan separar entre las distintas clases los nuevos ejemplos. A los cuales se les calculó previamente el mismo tipo de atributos.

Según el objetivo del modelo de *Machine Learning* se pueden distinguir dos grandes categorías:

- Modelos de aprendizaje supervisado: En este caso, el objetivo es predecir la clase de un ejemplo en base a sus atributos calculados, los que son recibidos como entrada para generar el modelo. Previamente conociendo la clase a la cual pertenece el ejemplo (Hastie y col., 2009). Las tareas principales son la clasificación y la regresión.
- Modelos de aprendizaje no-supervisado: En este caso, no hay una clase asociada a los ejemplos y el objetivo es describir una cantidad de patrones entre los atributos de entrada calculados (Hastie y col., 2009). La tarea principal es el *clustering*.

El presente estudio se ha enfocado en una tarea de clasificación supervisada, la cual consiste en predecir la función enzimática de las proteínas, utilizando las propiedades de los sitios de unión a ligando de éstas, mediante el método de *Machine Learning* llamado *Random Forest* (*RF*). El algoritmo consiste en la combinación de varios *árboles de decisión* entregando como salida la clase predicha, que es la moda de las clases entregadas por los árboles de decisión individuales. El método ha sido ampliamente utilizado en tareas pertenecientes a Bioinformática (Qi, 2012), teniendo excelentes resultados en tareas que utilizan variada información para realizar su clasificación. Además, existen algunas aproximaciones en el área de predicción de función enzimática. Una de ellas, en el año 2009 (Latino & Aires-de-Sousa, 2009) caracterizaron una enzima por las diferencias fisicoquímicas entre sus productos y reactantes, logrando un 78% de exactitud en los datos de testeo para predecir el primer dígito de la clasificación *EC*. Otro estudio realizado en el año 2012 (Kumar & Choudhary, 2012), consideraron 97 características de la secuencia de aminoácidos para predecir los 2 primeros números del código *EC*. El primer nivel de clasificación de dicho modelo es comparable con el presente estudio, ya que predice el primer dígito de la clasificación *EC* conocido como clase principal, con un 87,7% de exactitud.

Machine Learning tiene muchas aplicaciones en la Bioinformática (Larranaga, 2006), tales como redes genéticas involucrando las áreas de Biología de Sistemas y Microarreglos, construcción de arboles filogenéticos en el área de Evolución, anotación de proteínas involucrando las áreas de minería de textos y proteómica, entre muchas otras. Como se puede observar en la Figura 4, el área de la Bioinformática que envuelve este estudio, es la proteómica. Aunque el análisis de datos cristalográficos desde archivos de proteínas se considera un estudio perteneciente también al área de minería de textos.

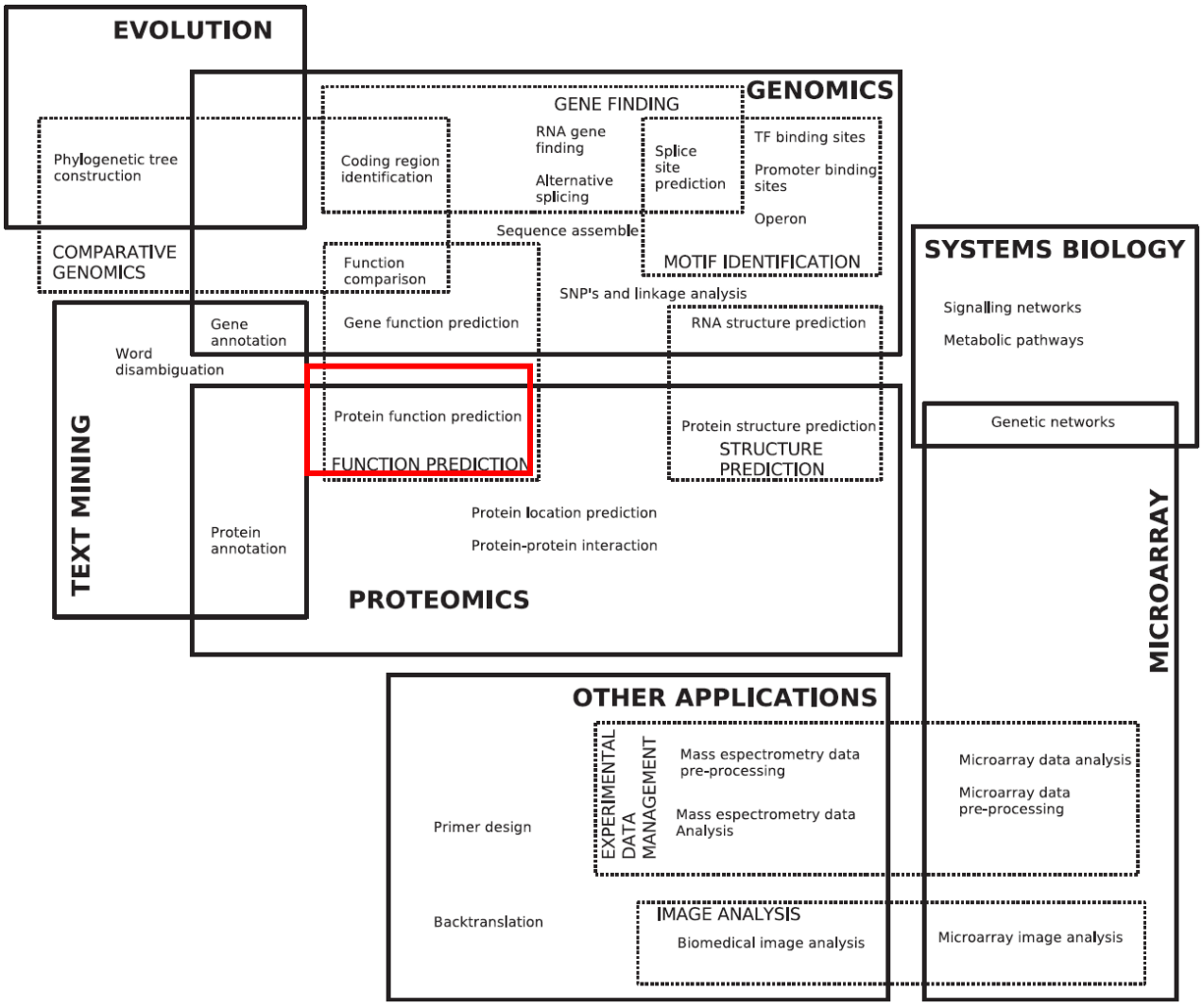

Figura 4. Tópicos de Bioinformática abordados con *Machine Learning*. En rojo se observa el área abordada en este estudio. Adaptada de (Larranaga, 2006).

1.2.2. Métodos de predicción basados en *Machine Learning*

A continuación se exponen diversos estudios realizados para la predicción de función enzimática, que utilizan distintos métodos o técnicas de *Machine Learning*. La mayoría de los estudios realizan una clasificación de acuerdo al primer dígito de la clasificación *EC*. Algunos de ellos agregan etapas de clasificación mas especificas, intentando predecir los siguientes dígitos de la clasificación enzimática. Para un mejor entendimiento estos estudios se agruparon según el tipo de propiedades que utilizan para realizar la caracterización de la función enzimática, tales como la composición aminoacídica, composición de dominios funcionales y por último los que utilizan información estructural e información similar a este estudio, que son las propiedades de los sitios de unión a ligando. Los métodos más ampliamente utilizados de *Machine Learning* en estos estudios son *Vecinos Cercanos (KNN), Reglas de Asociación (AR), Maquina de Vectores de Soporte (SVM) y Random Forest (RF).*

1.2.2.1. Composición Aminoacídica

Uno de los primeros trabajos que utilizaba la composición de aminoácidos de las proteínas fue realizado en el año 2003 (Chou & Elrod, 2003). Su objetivo fue la predicción del segundo dígito de la clasificación *EC* solo para la clase de las oxidorreductasas, obteniendo una tasa de éxito del 63,6%. El concepto de composición aminoacídica (*CAA*) se definió como la proporción de los 20 aminoácidos presentes en la estructura primaria de la enzima. Aunque los métodos basados en *CAA* son unos de los más rápidos al momento de obtener las características de la secuencia, estos pierden información valiosa relacionada al orden de la secuencia.

Posteriormente, (Chou, 2005) extendió sus datos incorporando nuevas características derivadas de la secuencia proteica, que el autor llamó *composición amfifílica de pseudo-aminoácidos*, conservando el orden de la secuencia. Estas nuevas medidas consisten en separar la cadena aminoacídica en diferentes secuencias conservando el orden y midiendo los patrones de distribución *hidrofóbica*

e *hidrofílica* en cada posición de los fragmentos de la secuencia, esto aumentó la tasa de éxito a un 70,61%.

Dado que los estudios anteriormente mencionados solo clasificaban la subclase enzimática de las oxidorreductasas. En el año 2010 (Qiu y col., 2010) utilizaron la misma información para predecir el primer número de la clasificación *EC*. Lograron una tasa de éxito global del 91,6%, usando una *transformada discreta de onda* con el método *Maquina de Vectores de Soporte* (*SVM*).

Por último, (Kumar & Choudhary, 2012) utilizaron características de la secuencia aminoacídica de las proteínas obtenidas mediante las herramientas EMBOSS-PEPSTAT (Rice y col., 2000) y Expasy-ProtParams (Gasteiger y col., 2005) que entregan 61 y 36 propiedades respectivamente, tales como hidrofobicidad, proporción de los 20 aminoácidos, estructura secundaria, etc. Logrando mediante la utilización de *Random Forest*, 3 niveles de clasificación. El primer nivel discrimina si la secuencia de entrada es o no una enzima obteniendo una exactitud del 94.8%. Luego predice el primer dígito de la clasificación EC o la principal clase funcional con 87.7% de exactitud. Por último, predice el segundo número de la clasificación *EC* o también conocida como la subclase funcional obteniendo una exactitud global del 84.2%.

1.2.2.2. Composición de Dominios Funcionales

En el caso de predicción de función enzimática es necesaria la inclusión de información que pueda representar efectivamente una enzima, para aumentar la exactitud de los modelos predictivos. En base a esta última aseveración, es que en el estudio realizado en el año 2005 (Cai & Chou, 2005) cambiaron su enfoque de composición de aminoácidos por el de dominios funcionales. Considerando como dominio funcional aquella sección de secuencia aminoacídica que es altamente conservada y le otorga una función a la proteína. Representaron una proteína como un vector de 7.785 atributos, los cuales determinan la presencia o ausencia de un cierto dominio funcional para cada una de las secuencias de entrada, obtenidos desde la base de datos *interPro* (Apweiler y col., 2000). Este estudio, reportó una

tasa de éxito global del 85,35% para identificar el primer dígito de la clasificación enzimática utilizando la función de *discriminante covariante*. Posteriormente, en el año 2006, otro estudio (Chiu y col., 2006) que también obtuvieron los dominios desde *interPro*, pero su avance fue el separar por taxonomía las secuencias de entrenamiento antes de realizar la clasificación. Gracias a este último enfoque, concluyeron que en procariontes la tasa de clasificación alcanza un 80%, lo que les permitió, mediante *Reglas de Asociación,* comprender qué dominios procariontes se encontraban con mayor frecuencia y así entender la importancia biológica de dichos dominios. Por otro lado, en el año 2007 basándose en el método *SVM* (Lu y col., 2007) desarrollaron un método para predecir el primer dígito de la clasificación *EC*, utilizando la composición de 2.757 dominios funcionales obtenidos desde *Pfam* (Sonnhammer y col., 1997). Lograron un 91,3% de tasa de éxito para la tarea de discriminar entre las 6 principales clases de enzimas. Según su reporte, sus resultados utilizando la composición de dominios funcionales fueron superiores a las metodologías de anotación por homología. En relación a éstas últimas, los autores reportaron tasas de éxito global no superiores al 87% utilizando el mismo conjunto de datos de testeo.

Por último, la información evolutiva de las secuencias obtenida por medio de *matrices de puntaje de posición específica* (*PSSM*, por sus siglas en ingles), fue incluida a la información de dominios funcionales para el desarrollo de la herramienta conocida como *EzyPred* (Shen & Chou, 2007). Utilizando el método de *KNN* reportó una tasa de éxito global de 93,7%, prediciendo si una secuencia de entrada es o no una enzima. Luego, si la secuencia de entrada es una enzima, predice el primer dígito de la clasificación *EC*, reportando una tasa de éxito global del 91,3%.

1.2.2.3. Información Estructural y sitios de unión a ligando

La evidencia de los estudios previamente descritos ha demostrado que al incluir propiedades estructurales de las proteínas a los modelos predictivos, se obtiene información valiosa a partir de las predicciones. Por este motivo, en el año 2008 (Munteanu y col., 2008) caracterizaron una enzima por su fracción de residuos,

propiedades de superficie, información de estructura secundaria y ligandos. En base a esta información, desarrollaron un método para discriminar entre enzimas y no-enzimas. Reportaron una tasa de clasificación global del 76.17%.

Por otro lado, como las reacciones enzimáticas son directamente dependientes de la información estructural del sustrato que catalizan. En el año 2011 (Almonacid & Babbitt, 2011), propusieron una nueva métrica de evaluación de la clasificación *EC* a través de los mecanismos de las reacciones enzimáticas. Según su reporte, estos últimos son conservados dentro de la misma clase enzimática independientemente del ligando que este interactuando con la enzima.

Siguiendo el mismo enfoque anterior es que (Nath & Mitchell, 2012) utilizando 3 distintos métodos de *Machine Learning* tales como *SVM, KNN* y *RF* lograron predecir el primer dígito de la clasificación *EC*. Utilizaron la información del mecanismo de reacción obtenida desde MACiE (Holliday y col., 2005). Reportaron una exactitud variable dependiendo del tipo de descriptor y método utilizado, alcanzando un máximo de 90,7% para *Random Forest*.

Intentando clasificar las enzimas con el primer dígito de la clasificación enzimática, (Bray y col., 2009) se enfocaron en las características de los sitios activos de las enzimas. Lograron apenas un 33.1% de exactitud, lo que según su reporte es un 17% mejor que la clasificación por azar. Por otro lado, (Izrailev & Farnum, 2004) generaron un método para asignar el número *EC* completo utilizando la similaridad de ligandos entre las enzimas. La información de la interacción proteína-ligando fue obtenida desde BRENDA (Scheer y col*.,* 2011). Considerando como ligando a cualquier molécula reportada en la base de datos como reactante, inhibidor, cofactor o ión interactuando con la proteína. Esto les produjo excelentes resultados, reportando exactitudes superiores al 92% para las enzimas que se logró asignar una función. Concluyeron que la información de la interacción de la proteína con sus ligandos era una buena aproximación para validar los resultados de anotación automática de funciones enzimáticas.

En el transcurso de redacción de esta memoria se ha publicado un artículo (Volkamer y col. , 2012) que realiza una tarea similar a la planteada en este estudio.

En dicho artículo demuestran que solo con las características geométricas del bolsillo de la proteína, producido por la unión de un ligando, se puede predecir los 2 primeros dígitos de la clasificación *EC.* En sus datos se observa un claro desbalance con una mayor cantidad de ejemplos en las 3 primeras clases, destacando la clase mayoritaria de las hidrolasas que supera 13 veces la cantidad de la clase minoritaria (ligasas). En sus resultados reportan exactitudes para cada una de las clases, las cuales oscilan entre 62,8% y 80,9%. Una de las diferencias con este estudio es en las etapas previas a la obtención de características, ya que no realizan una reducción de redundancia en los datos. Además, la definición de un sitio de unión a ligando es sesgada ya que se considera como tal a toda aquella localización de un ligando que coincida con la predicción de DoGSite (Volkamer y col., 2010), el cual utiliza los mismos descriptores geométricos para predecir sitios activos en enzimas.

Teniendo en consideración estas últimas aproximaciones en la búsqueda de la función de las proteínas, es decir asignar una de las 6 clases enzimáticas del primer dígito de la clasificación *EC* a una proteína, surge el enfoque de esta memoria, el cual es utilizar las características fisicoquímicas, geométricas y evolutivas de la interacción de las enzimas con sus ligandos, ya que esta información es extremadamente beneficiosa para comprender su funcionamiento celular. Además, el caracterizar y analizar cómo se unen moléculas a las enzimas, puede ayudar a entender de mejor manera el mecanismo y el tipo de reacción que la enzima está realizando en el sitio de unión. Más aun, existe una amplia cantidad de datos disponibles para el análisis, ya que poco más de la mitad de las estructuras conocidas de proteínas pertenecen a la clasificación de enzimas.

HIPOTESIS DE TRABAJO

Un modelo de predicción basado en la información de los sitios de unión a ligando de una enzima es capaz de discriminar entre las distintas clases enzimáticas de manera eficiente.

OBJETIVOS

1. Objetivo General:

Desarrollar un modelo de predicción basado en *Random Forest* que permita predecir la función enzimática. Utilizando la información de los sitios de unión a ligando de cada enzima.

2. Objetivos Específicos:

1.- Generar un set no redundante de enzimas con estructura tridimensional conocida, perteneciente a los seis grupos primarios de clasificación *EC* (*Enzyme Commission numbers*).

2.- Generar y recopilar información de los sitios de unión a ligando de las enzimas, y consecuentemente construir un set de atributos basado en propiedades evolutivas, físico-químicas y geométricas.

3.- Entrenar y evaluar el desempeño del modelo basado en *Random Forest* utilizando el set de atributos generado anteriormente.

4.- Evaluar estrategias de aplicabilidad del modelo generado, para la predicción de función enzimática de nuevas estructuras cristalográficas.

METODOLOGÍA

1. Construcción del set de datos no redundante

La construcción del set de datos no redundante es el primer paso en todo proceso de minería de datos y consiste en seleccionar, limpiar y transformar los datos siguiendo criterios guiados por la problemática abordada y por los objetivos planteados (Larranaga, 2006). Lo más importante en esta etapa, es eliminar la mayor cantidad de datos redundantes, para que en etapas posteriores las predicciones no sean sesgadas.

1.1. Obtención del conjunto no redundante de enzimas

La obtención de un grupo representante de cada clase enzimática, es de importancia para este estudio. A la fecha aproximadamente 46.000 proteínas disponibles en *PDB* se encuentran clasificadas en alguno de los grupos de enzimas indicados por la *Enzyme Commission numbers (EC)* (Karlson y col., 1992).

En primer lugar, se procedió a descargar todos los archivos de proteínas desde la base de datos *PDB*. Una vez con los archivos disponibles de manera local se procedió a eliminar los que no cumplían con los requerimientos para este estudio (anexo 1.1). Esto se llevó a cabo mediante un script desarrollado en lenguaje *TCL* (*Tool Command Language*), el cual permite automatizar el uso del *software Visual Molecular Dynamics (VMD)* (Humphrey y col., 1996). Este *software* y sus diferentes módulos, facilitaron el análisis espacial de las macromoléculas. Primero, se sacó de la carpeta todos aquellos archivos que tuvieran como cabecera o primera línea las palabras *unknown function,* ya que estos no poseían alguna función asignada aunque algunos de éstos tenían *EC* asignado. Posteriormente, se eliminaron las proteínas que tenían más de un modelo proteico guardado en el archivo. Además, se sacaron de la carpeta original todas aquellas proteínas que no tuvieran ligandos. Finalmente, se guardó en carpetas separadas cada proteína según su primer dígito de la clasificación enzimática (*EC*).

Se ha reportado anteriormente (Chothia & Lesk, 1986; Wilson y col., 2000) que proteínas con baja identidad de secuencia entre sí pueden poseer funciones similares, por tal motivo se establece que un 25% de identidad de secuencia aminoacídica sería un buen límite para evitar la redundancia. Por lo tanto, se procedió a eliminar proteínas que tuvieran tal porcentaje de identidad. La eliminación de esta redundancia ayudará a evitar un sesgo hacia aquellas enzimas que poseen un mayor número de representantes. Esta etapa en todo proceso de *Machine Learning* se conoce como eliminación de redundancia en los datos. Para esto se utilizó la aplicación web PISCES (Wang & Dunbrack, 2003) la cual realiza un agrupamiento de las secuencias proteicas según el porcentaje máximo de identidad ingresado como parámetro. Se entregó, por separado, una lista por cada clase con todos los ID de 4 caracteres de las proteínas. Por otro lado, el mismo servidor permite hacer otros tipos de filtros, tales como:

- Largo de cadena: 60 - 10000
- Máxima resolución del cristal: 2.0 Å
- Máximo valor R: 3.0
- Omitir archivos distintos a rayos X: yes
- Omitir archivos de solo CA: yes

El criterio de largo de cadena fue escogido arbitrariamente, ya que el propio servidor recomienda entre 40 – 10.000, pero un largo de 40 aminoácidos es considerado corto para ser una proteína funcional, ya que en estudios (Keefe & Szostak, 2001) se ha reportado que con un mínimo de 40 aminoácidos se puede encontrar un dominio funcional, utilizando secuencias de largo 80. En cuanto al parámetro de resolución máxima se utilizó una resolución de 2.0 Å ,debido a que en reportes anteriores (Wei y col., 1999), se determinó que con dicha resolución en un modelo de proteína es suficiente para obtener una definición espacial correcta de la posición de los átomos. Por último, los otros parámetros se dejaron por defecto, ya que los autores del software recomiendan dichos valores para el análisis de proteínas.

1.2. Extracción de sitios de unión

Para la extracción de sitios de unión a ligando, se realizó un script en *TCL* (anexo 2.1). Mediante el script creado, se cargaron los archivos de proteínas en el programa *VMD*. Posteriormente, se realizó la selección de los aminoácidos que se encuentran dentro de un radio de 7 Å del centro de masa de cada ligando (Figura 5), ya que dentro de este rango se encuentran la mayoría de las interacciones (no enlazantes y enlazantes) que podrían estar involucradas en la unión proteína-ligando (Kumar & Nussinov, 2002).

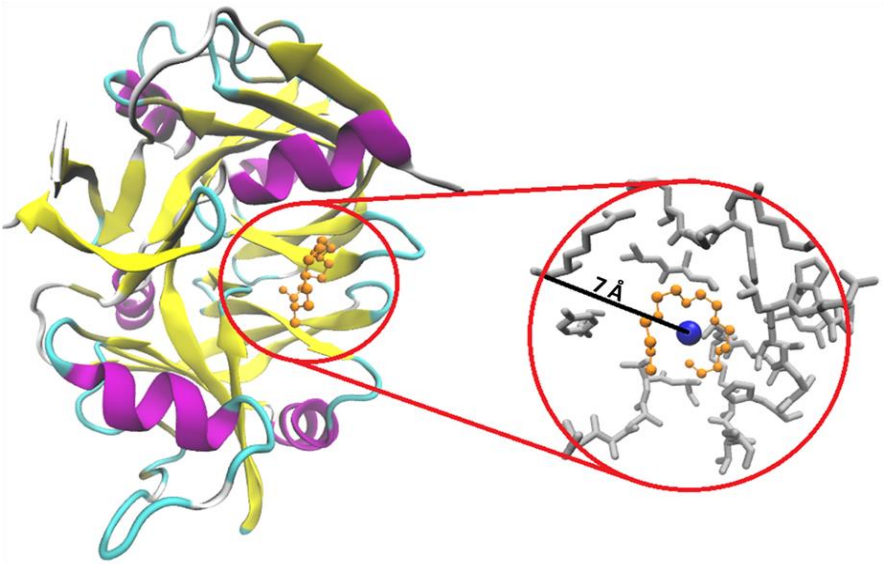

Figura 5. Ejemplo de sitio de unión a ligando. Todos los aminoácidos pertenecientes al sitio (en gris) están a 7 Å del centro de masa (en azul) del ligando (en naranjo).

Los aminoácidos que se encuentran a 7 Å de un ligando, deben cumplir al menos 3 características para ser considerados como sitios de unión a ligando y pasar a ser un ejemplo para clasificar.

- Primero, debe existir al menos un aminoácidos a ≤ 3 Å de distancia desde la periferia del ligando, esto debido a la distancia mínima de interacción por puentes de hidrógeno (Mildvan y col., 2002).

- Además, no deben existir aminoácidos distintos a los 20 comunes dentro del sitio, ya que existen modificaciones que complican el cálculo de propiedades.

- Por último, los aminoácidos a 7 Å del ligando deben tener una ocupancia igual a 1.0, ya que al tener menor ocupancia significaría que no se tiene certeza de la orientación de la cadena lateral de dichos aminoácidos y esto introduce errores.

Una vez obtenidos todos los sitios de unión a ligando, se procedió a realizar un análisis de la cantidad de aminoácidos presentes por sitio (anexo 2.2). El motivo de este análisis es corroborar si existe una distribución normal de las cantidades de aminoácidos por sitio y así no considerar sitios muy alejados de la media, es decir sitios con muy pocos o con demasiados aminoácidos dentro del radio de 7 Å.

2. Estimación de atributos

Para el cálculo de atributos, el tipo y nombre de atributos fueron obtenidos desde la tesis de pregrado (Pereira, 2012). Todos los cálculos fueron realizados con metodologías propias. Siendo las excepciones, el cálculo de energías y la composición aminoacídica por capas. Cabe destacar que se generaron nuevos atributos tales como los radios de sitio, las matrices de posición específica del sitio completo y la estructura secundaria presente en el sitio.

2.1. Atributos Geométricos

Considerando un mínimo de 4 aminoácidos presentes en el sitio, se calcularon las distancias y ángulos entre los primeros 4 aminoácidos más cercanos al centro de masa del ligando. El conjunto de puntos para los cálculos son: el centro de masa del ligando, los átomos más cercano de los 4 aminoácidos y sus carbonos alfa. Para esto, se realizó un procedimiento creado en el lenguaje *TCL* (anexo 2.3), el cual, por medio de instrucciones a *VMD* obtenía las coordenadas a analizar.

2.1.1. Ángulos

Se calcularon un total de 16 ángulos para caracterizar la interacción del ligando con los aminoácidos más cercanos al centro de masa al ligando. Como se puede ver

en la figura 6, se dividieron en 3 tipos de ángulos según las coordenadas que utilizan. Los del tipo A son aquellos ángulos formados por el centro de masa del ligando y el carbono alfa de uno de los 4 aminoácidos más cercanos, teniendo como vértice el átomo más cercano al ligando del mismo aminoácido. Los del tipo B son aquellos ángulos que se forman entre los átomos más cercanos al ligando de 2 aminoácidos distintos, teniendo como vértice el centro de masa del ligando. Por último, los del tipo C utilizan las coordenadas de los carbonos alfa de 2 de los 4 aminoácidos más cercanos, considerando el centro de masa del ligando como vértice del ángulo.

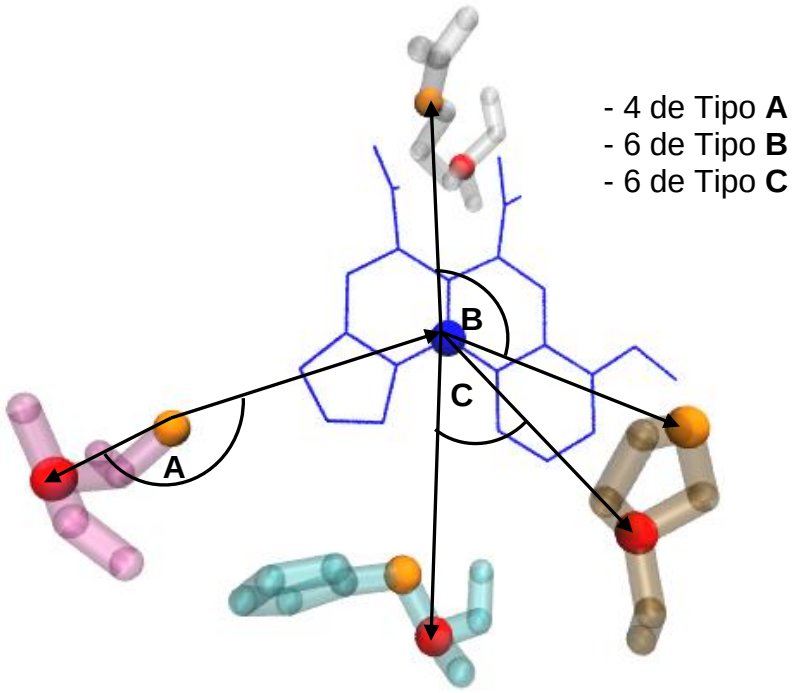

Figura 6. Ejemplo de ángulos calculados. En negro se observan los 3 tipos de ángulos calculados. En rojo se encuentran los carbonos alfa, en naranjo se encuentran los átomos más cercanos de cada aminoácido y en azul el ligando, con su respectivo centro de masa.

2.1.2. Distancias

Se calcularon un total de 20 distancias. Como se puede ver en la figura 7, se dividieron en 4 tipos de distancias según las coordenadas que utilizan. Las del tipo A son aquellas distancias formados por el centro de masa del ligando y el carbono alfa

de uno de los 4 aminoácidos más cercanos. Las del tipo B son aquellas distancias que se forman entre el centro de masa del ligando y el átomo más cercano de uno de los 4 aminoácidos. Las distancias de tipo C son aquellas que se forman entre los carbonos alfa de 2 de los 4 aminoácidos más cercanos. Por último, las distancias del tipo D utilizan las coordenadas de los átomos más cercanos de 2 de los 4 aminoácidos mas cercanos al centro de masa del ligando.

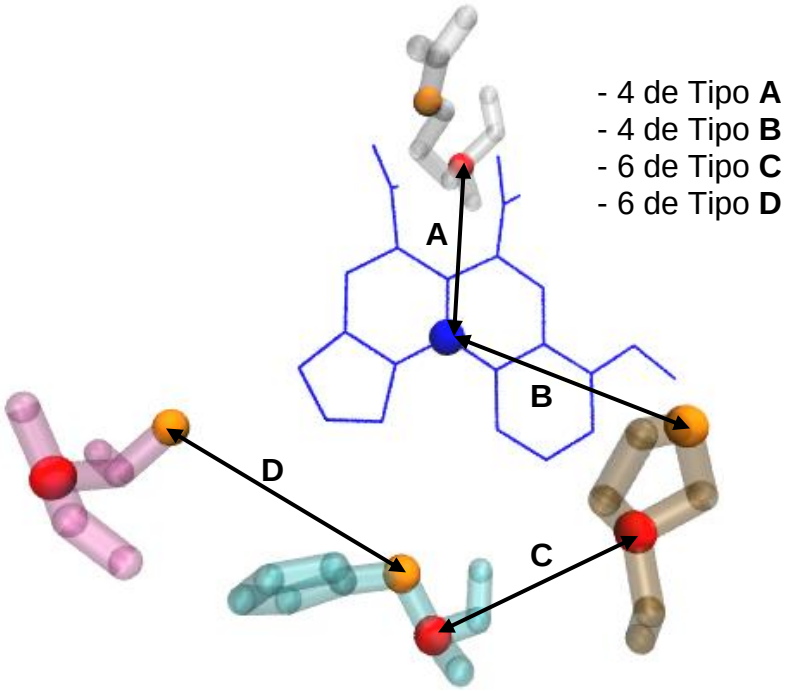

Figura 7. Ejemplo de distancias calculadas. las flechas en negro representan algunas distancias calculadas. En rojo se encuentran los carbonos alfa, en naranjo se encuentran los átomos más cercanos de cada aminoácido y en azul el ligando, con su centro de masa.

2.2. Energías

Se calcularon 14 energías a través de la función de energía *FOLDEF* (Guerois y col., 2002), implementada en el *software FoldX*. Este programa, entrega una serie de energías y propiedades fisicoquímicas de la molécula que se ingrese por medio de una lista (anexo 2.4). La función energética asigna por medio de datos experimentales, las contribuciones de las energías no enlazantes, reportando una

correlación del 90% de los datos predichos con los datos reales. Todas las energías son entregadas en Kcal/mol a una temperatura de 298 K y pH 7.

2.3. Hidrofobicidades y composición aminoacídica

La hidrofobicidad de cada aminoácido presente en una proteína, juega un rol esencial al momento de comprender el plegamiento de ésta (Biswas y col., 2003). Existen diversas escalas para la estimación de hidrofobicidades por aminoácidos. En este estudio, se utilizaron las 3 principales, (Hessa y col., 2005), (Kyte & Doolittle, 1982) y (Wimley & White, 1996), las cuales son presentadas en el anexo 2.5. Para las dos últimas, un aminoácido hidrofóbico tiene un valor positivo y para la primera un aminoácido hidrofílico tiene un valor positivo.

Para obtener la hidrofobicidad del sitio, se creó un script en *TCL* (anexo 2.6) en donde se suman las hidrofobicidades de cada aminoácido presente en el sitio en las distintas escalas y se obtuvieron 3 valores de hidrofobicidad total.

Por otro lado, para caracterizar la composición aminoacídica los aminoácidos se dividieron en 4 grandes categorías según su carga o polaridad, estas son no polares, polares neutros, polares negativos y polares positivos. Para la obtención de las proporciones de estas categorías presentes en los sitios se ocupó el mismo script anterior (anexo 2.6). Además, se agregó un total de aminoácidos en el sitio, para tener una relación entre el porcentaje de cada categoría y el total.

2.4. Composición atómica por capas

Para la obtención de capas del sitio, se generó un script en *TCL* (anexo 2.7), el cual para cada sitio a partir de 1 Å desde el centro de masa del ligando generó capas de tamaño 1 Å hasta llegar a los 7 Å, es decir se generaron 5 capas siendo la más grande la que va de 6 a 7 Å. Para cada capa, se calculó la proporción de cada tipo de átomo presente, siendo estos carbono, oxigeno, nitrógeno y azufre. También se calculó el total de átomos en cada capa y el total de átomos presentes en el sitio.

2.5. Radios de sitio

Para tener una aproximación del tamaño del bolsillo, se calcularon algunas distancias entre el centro de masa del sitio y los átomos más cercanos de cada aminoácido perteneciente al sitio, denominadas estas distancias como radios de sitio. Para tal objetivo, se creó un script en *TCL* (anexo 2.8), el cual calcula estas distancias. Esta información es procesada para determinar los mínimos, máximos, el promedio y la desviación estándar por cada sitio.

2.6. Estructura secundaria y área accesible al solvente

La estructura secundaria de una proteína ya ha sido utilizada para intentar predecir la función enzimática (Dobson & Doig, 2005). El nuevo enfoque propuesto en este estudio, fue utilizar la proporción de las 3 grandes categorías de estructura secundaria (alfa hélices, sabanas beta y random coils) presentes en el sitio de unión a ligando. La estructura secundaria se extrajo desde los archivos de las proteínas utilizando la aplicación *VMD* por medio de un script realizado en *TCL* (anexo 2.9).

El área accesible al solvente, está definida por el conjunto de puntos en la periferia de una molécula que está disponible para ser utilizada por cualquier tipo de solvente. Para realizar este cálculo, se añadió al script anterior (anexo 2.9) la obtención del área accesible al solvente del sitio completo, considerando la obstrucción que genera el resto de la proteína.

2.7. Puntajes evolutivos

Las matrices de puntuación por posición específica (*PSSM,* del inglés *Position-Specific Scoring Matrix*), son una metodología ampliamente utilizada en muchas de las áreas que involucran a las proteínas (Naik y col., 2007). En el presente estudio, se utilizaron dos nuevos enfoques para esta metodología. El primero es obtener la secuencia de los 12 primeros aminoácidos más cercanos al centro de masa del ligando (ver 2.9.1) y el otro enfoque es obtener la secuencia de todos los

aminoácidos presentes en el sitio, conservando el orden de aparición en la estructura primaria (ver 2.9.2).

Teniendo la secuencia del sitio en formato fasta, se procedió a generar una matriz de puntaje por cada clase enzimática, para luego escanear todas las secuencias de los sitios y obtener 6 puntajes, uno por cada función. Para realizar esto, se creó un script en lenguaje *PERL* (anexo 2.10.1), el cual invoca a los programas de *EMBOSS* (Rice y col., 2000), para parsear los resultados y generar un output en forma de columnas con los puntajes evolutivos para cada una de las clases.

2.7.1. Aminoácidos más cercanos

En este apartado, se generaron las secuencias aminoacídicas de cada sitio, a partir de los 12 primeros aminoácidos, ordenados según su cercanía al centro de masa del ligando. Esto se realizó por medio de un script en lenguaje *TCL* (anexo 2.10.2), el cual ordena los aminoácidos según su cercanía al ligando. Si el sitio tenía menos de 12 aminoácidos solo se escribió la secuencia de los presentes en el sitio.

2.7.2. Todos los aminoácidos del sitio

Para lograr tener una mejor aproximación de la distancia evolutiva que tiene un sitio con cada una de las clases enzimáticas, se generaron las secuencias de aminoácidos a partir del orden de aparición en estructura primaria, es decir conservando el orden de la secuencia de los aminoácidos presentes en el sitio. Para este motivo, se creó un script en *TLC* (anexo 2.10.3), en donde se carga cada sitio y se selecciona sus carbonos alfa, para escribir en un archivo en formato fasta, por cada clase enzimática, las secuencias aminoacídica de cada sitio.

3. Entrenar y evaluar el modelo

El algoritmo de *Machine Learning* que se utilizó para generar el modelo predictivo fue *Random Forest* (Breiman, 2001). Este algoritmo de aprendizaje supervisado,

consiste en generar N cantidad de *Arboles de Decisión* (Quinlan, 1986), utilizando la metodología de remuestreo *Bagging* (Breiman, 1996), considerando la clase predicha como la moda de las predicciones de todos los árboles generados. Se eligió este método ya que ha sido ampliamente utilizado en tareas pertenecientes a bioinformática (Qi, 2012), teniendo excelentes resultados en tareas que utilizan variada información para realizar su clasificación. Además, existen algunas aproximaciones en el área de predicción de función enzimática (Kumar & Choudhary, 2012; Latino & Aires-de-Sousa, 2009), que fueron presentadas en la introducción de este trabajo.

3.1. Árboles de Decisión

El método se define por ser un conjunto de condiciones organizadas en una estructura jerárquica. Las decisiones se toman siguiendo las condiciones que se cumplen desde la raíz, hasta alguna de las hojas del árbol, siendo la raíz el primer atributo seleccionado para tomar una decisión y todos los sucesivos pasan a ser hojas del árbol. Como se muestra en la figura 8, tras la evidencia de un set de datos climatológicos, se puede generar un árbol de decisión para predecir si lloverá o no, basándose en algunos atributos que pasan a llamarse nodos del árbol (en recuadros) y los posibles valores de cada uno que pasan a llamarse ramas del árbol. No existe un único árbol de decisión frente a algún problema con múltiples atributos, ya que la elección del atributo óptimo para particionar los ejemplos es una tarea a mejorar en este tipo de clasificadores y lo que se busca a final de cuentas es generar una rama pura, es decir que solo existan ejemplos de una sola clase en dicha rama. Para tal motivo, existe el sistema de selección de atributos que generen ramas con más ganancia de información o entropía, definida esta última por la formula:

$$-\sum p_i \times \log(p_j^i) \qquad (1)$$

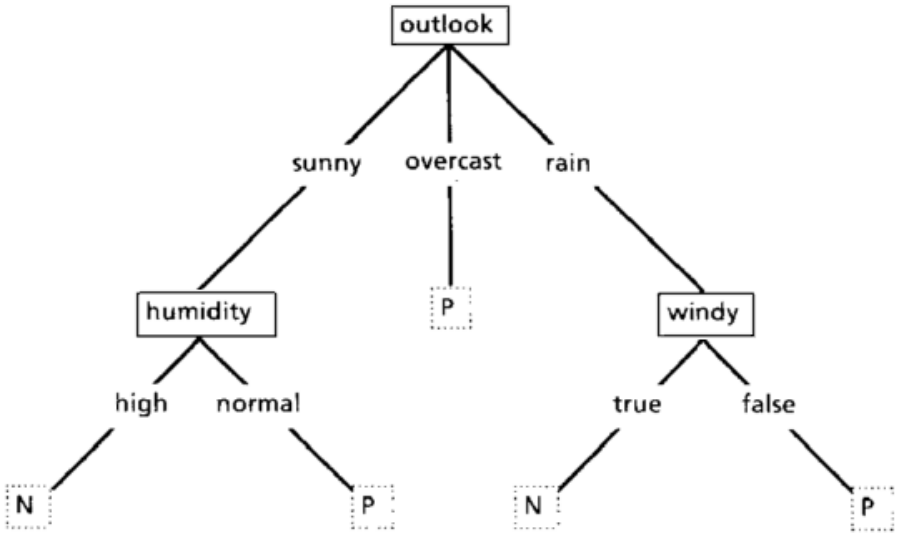

Figura 8. Un árbol de decisión simple. Esquema de condiciones simples frente a la pregunta si lloverá o no. Extraída de (Quinlan, 1986).

En donde, $\mathbf{p_i}$ es la probabilidad de la clase i; $\mathbf{p^i_j}$ es la probabilidad de la clase i en el nodo j. Se prueba con cada atributo en el nodo y luego se selecciona el que tenga mayor ganancia de información. Luego, se separan los ejemplos según el atributo seleccionado, si se logra tener solo una clase, se le denomina nodo impuro y se termina esa rama, si no, todavía quedan ejemplos de cada clase se procede recursivamente a seleccionar un atributo que mejor divida los ejemplos hasta que se consigan todas las ramas puras.

3.2. Random Forest

Partiendo de la idea de que existen muchos árboles de decisión que logren clasificar bien una tarea, es que se creó la metodología de *Random Forest* (Breiman, 2001). Generando muchos árboles de decisión con un subconjunto de atributos se obtienen mejores desempeños al momento de clasificar, ya que la clasificación pasa a ser un consenso de las clasificaciones anteriores.

En este estudio se utilizó *WEKA* (Witten y col., 1999), ya que es un *software* de código abierto que implementa variados algoritmos de *Machine Learning* en el

lenguaje de programación *JAVA*, el cual permite la ejecución del *software* independientemente del sistema operativo en uso. El algoritmo solo tiene 2 grandes parámetros a elegir, por un lado se encuentra la cantidad de árboles que se desean generar, y por otro, la cantidad de atributos a seleccionar para la generación del árbol. En primer lugar, se decidió tomar un valor de 1.000 árboles, ya que esto permite que las predicciones sean mejores y más estables. En cuanto al parámetro de cuantos atributos al azar considerar para la generación de los árboles, se dejó por defecto la cantidad de 7 atributos, ya que el mismo autor del algoritmo (Breiman, 2001) recomienda dicha cantidad como la óptima para los procesos de clasificación.

3.3. Validación y medidas de desempeño

Todo modelo de clasificación obtiene sus tasas de desempeño en relación a los datos de testeo. Para suplir esta necesidad de validación del entrenamiento del algoritmo, se utilizó la metodología de 10-validaciones cruzadas (Rodríguez y col., 2010). Esta metodología consiste en dividir el conjunto de datos de entrenamiento en 10 partes iguales, conservando las proporciones de cada clase. Utilizando nueve de los conjuntos anteriormente generados para entrenar el modelo se prueba dicho modelo en el conjunto restante, luego se utilizan otros nueve para generar un nuevo modelo y se prueba en un conjunto nuevo, así sucesivamente hasta completar las 10 validaciones, una con cada subconjunto de datos. Con dicha aproximación se obtienen valores fidedignos de desempeño del modelo, ya que en cada validación los datos de entrenamiento son totalmente distintos a los datos de prueba.

De la validación cruzada se obtuvieron diversas medidas de desempeño, tales como:

- TVP: VP/P (2)
- TFP: FP/N (3)
- Precisión: VP/ (VP+FP) (4)
- Recall: VP/ (VP+FN) (5)
- MCC: $(VP*VN+FP*FN) / \sqrt{(VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)}$ (6)

- ROC: Área bajo la curca ROC (7)

Donde, TVP es la tasa de verdaderos positivos, VP son los verdaderos positivos, P es la cantidad de casos positivos o el total de una clase, TFP es la tasa de falsos positivos, FP son los falsos positivos, N es el total de casos negativos, VN son los verdaderos negativos, FN son los falsos negativos y por último la sigla MCC proviene del inglés de Coeficiente de Correlación de Matthews.

Lo que se busca en todo proceso de clasificación es que todos estos indicadores, exceptuando la TFP, sean cercanos a 1, ya que esto indica una buena predicción. En el caso de la TFP, lo ideal es que sea lo más cercana a 0, esto quiere decir que el modelo no cometió muchos errores.

4. Evaluar estrategias de clasificación de estructuras

Existen 2 grupos de proteínas en donde se evaluó la estrategia de clasificarlas con el modelo generado.

Primero, proteínas con ligandos en su estructura pero que no fueron utilizadas en el entrenamiento. Para esto, se utilizaron todas aquellas proteínas que fueron liberadas en el servidor *PDB* entre el 1 de septiembre y el 31 de octubre, ya que las proteínas utilizadas en el entrenamiento del modelo fueron obtenidas los últimos días de agosto. Teniendo estas proteínas, se realizó un análisis de redundancia entre ellas y junto con el conjunto de proteínas de entrenamiento para evidenciar si las nuevas proteínas eran muy distantes a las utilizadas en el entrenamiento.

Por otro lado, se utilizaron proteínas que no tenían ligandos en su estructura y tenían un *EC* anotado para corroborar los resultados de predicción. Para analizar estas estructuras se necesitó primero conocer los sitios de unión a ligando. De acuerdo a una revisión de métodos que realizan dicha tarea (Leis y col., 2010), se eligió Q-siteFinder (Laurie & Jackson, 2005) ya que es uno de los que mejores resultados reporta. El servidor entrega una nube de puntos cercanos a donde predice que podría ubicarse un ligando por un cálculo disponibilidad energética. Tomando este output, se calculó el centro geométrico de dicha nube de puntos para cada una

de las 30 proteínas seleccionadas al azar, para ser considerado como la posición en donde se uniría un ligando.

Se utilizó la misma metodología descrita en los puntos 1 y 2 de este capítulo para ambos tipos de ejemplos. Luego, con el modelo generado anteriormente se precedió a clasificar los nuevos ejemplos utilizando el programa *WEKA* con la opción de re-evaluar el modelo con un set de testeo de entrada.

RESULTADOS

1. Set de datos no redundante

En relación a la obtención de un set de enzimas no redundante se siguió una metodología que se muestra en la Figura 9. En este esquema se explican los distintos filtros que se aplicaron a todas las proteínas disponibles en el servidor *PDB.* El total de proteínas analizadas fue 83.983 a la fecha del 26 de agosto del 2012. A partir de éstas se fueron sacando del conjunto (círculos en blanco del esquema) aquellas proteínas que no cumplían con requerimientos para realizar la clasificación. En primer lugar, se eliminaron 3.171 proteínas que tenían función desconocida asignada en la cabecera del archivo cristalográfico. Además, se eliminaron 41026 proteínas que no tenían un número *EC* asignado y 1861 enzimas que poseían más de un código *EC* anotado en su función. Por último, se sacaron del análisis 4698 enzimas que no tenían ligandos cristalizados en su estructura. Estos últimos se dejaron en una carpeta aparte para ser utilizados en la última etapa de este trabajo, explicada en el punto 4 de la metodología.

Construidas las listas para cada una de las 6 clases enzimáticas, se procedió a realizar la reducción de redundancia. Como se explicó en el punto 1.1 de la metodología, para este objetivo se utilizó el servidor PISCES, ya que permite el agrupamiento de las proteínas según un máximo de 25% de identidad. Del total de 32.926 enzimas encontradas anteriormente, solo el 6% de ellas quedó luego de realizada la reducción de redundancia. Como se puede observar en la Tabla 1 la cantidad total de enzimas no redundantes fue de 1.960 (Total PDBs-NR). En la segunda columna de esta tabla se puede observar un claro desbalance en los datos, ya que por ejemplo la cantidad de enzimas con función de hidrolasas supera casi 7 veces la cantidad de ligasas. Cada una de las 1.960 enzimas tiene en promedio 9 ligandos cristalizados en su estructura. Los aminoácidos cercanos a estos ligandos pasan a ser candidatos de sitios de unión si los aminoácidos que están a 7 Å del centro de masa del ligando cumplen con los requerimientos establecidos en el punto 1.2 de metodología.

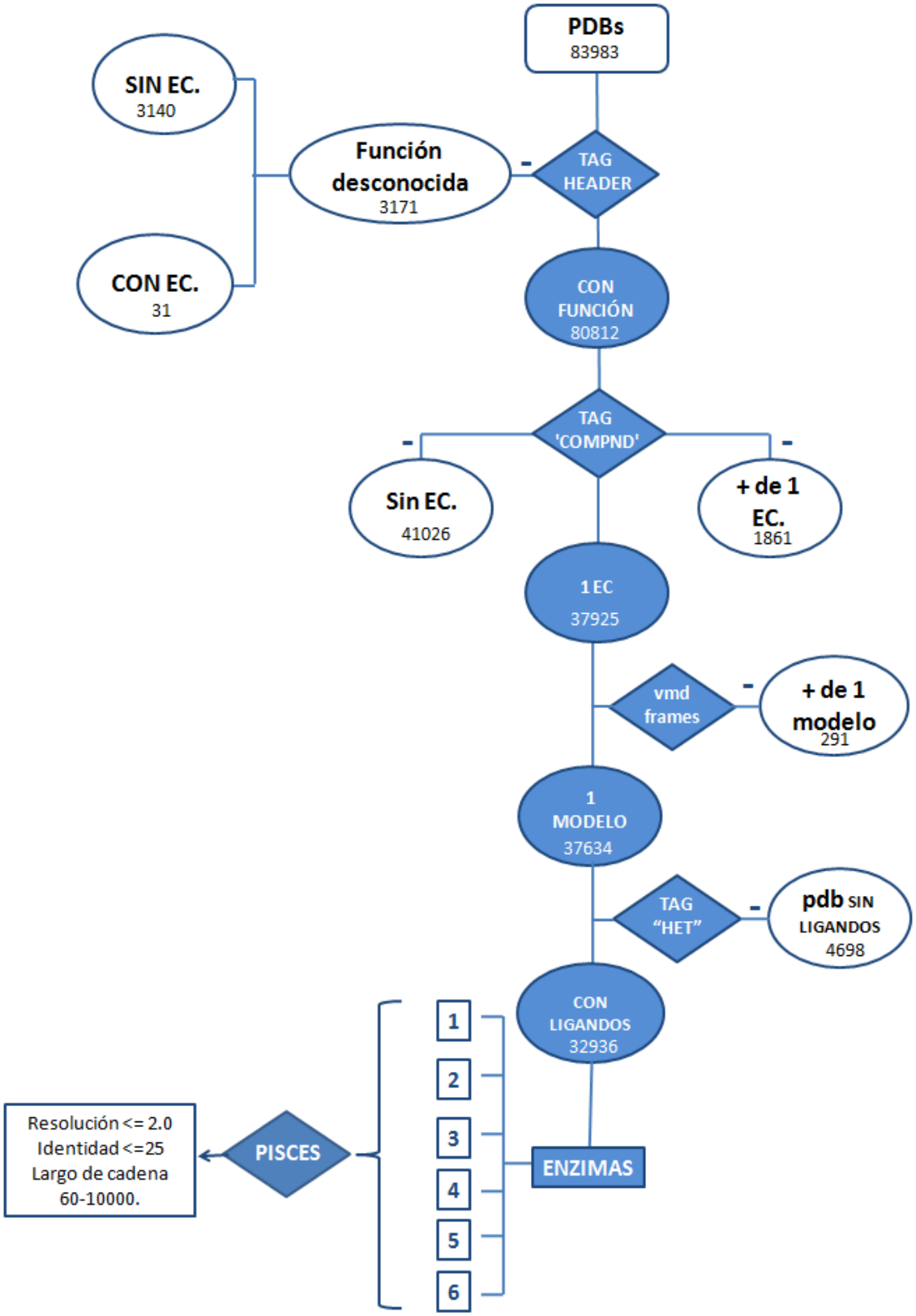

Figura 9. Esquema de filtros realizados a todas las proteínas. Partiendo de las proteínas totales en el servidor *PDB* a la fecha del 26 de agosto del 2012.

Como se puede observar en la Tabla 1, del total de 17.509 ligandos disponibles, solo quedaron 4.431 sitios de unión a ligando después de realizada su extracción desde las enzimas. Además, se puede observar en la quinta columna de la tabla que el número de proteínas consideradas en este estudio disminuyó a 1.253 al hacer la extracción de sitios, obteniéndose en promedio entre 3 y 4 sitios de unión a ligando por estructura cristalográfica. Por último, cabe destacar que el desbalance de los ejemplos aumentó en relación a las proporciones de proteínas iniciales, ya que la cantidad de sitios pertenecientes a la clase de las hidrolasas supera casi 10 veces a la cantidad de sitios perteneciente a la clasificación de isomerasas.

Tabla 1. Cantidades de proteínas y sitios por clase enzimática.

EC-Clase	PDBs-NR	Ligandos	Sitios de Unión	PDBs con sitios
1-Oxidoreductasa	314	3074	813	222
2-Transferasa	530	4524	1024	320
3-Hidrolasa	687	5839	1723	455
4-Liasa	212	1943	487	136
5-Isomerasa	118	1077	176	65
6-Ligasa	99	1052	208	55
Total	**1960**	**17509**	**4431**	**1253**

Luego de realizar la reducción de redundancia quedaron 1960 proteínas (PDBs-NR) con sus respectivos ligandos. La cantidad de proteínas que resultaron luego de la extracción de sitios (PDBs*) disminuyó.

Por último, en relación al análisis de la cantidad de aminoácidos por sitio cabe destacar que en promedio los sitios de unión a ligando tenían 12 aminoácidos. Con un mínimo de 1 aminoácido y un máximo de 23 aminoácidos por sitio, se procedió a seleccionar todos aquellos que tuvieran entre 4 y 20 aminoácidos, ya que con cantidades fuera de ese rango se dificultaría el cálculo de propiedades. Como se puede observar en la Figura 10, se sacaron del análisis solo el 4,7% del total de sitios de unión a ligando. Como resultado se obtuvieron sitios con cantidades similares de aminoácidos, que facilitan el cálculo de propiedades geométricas y evolutivas.

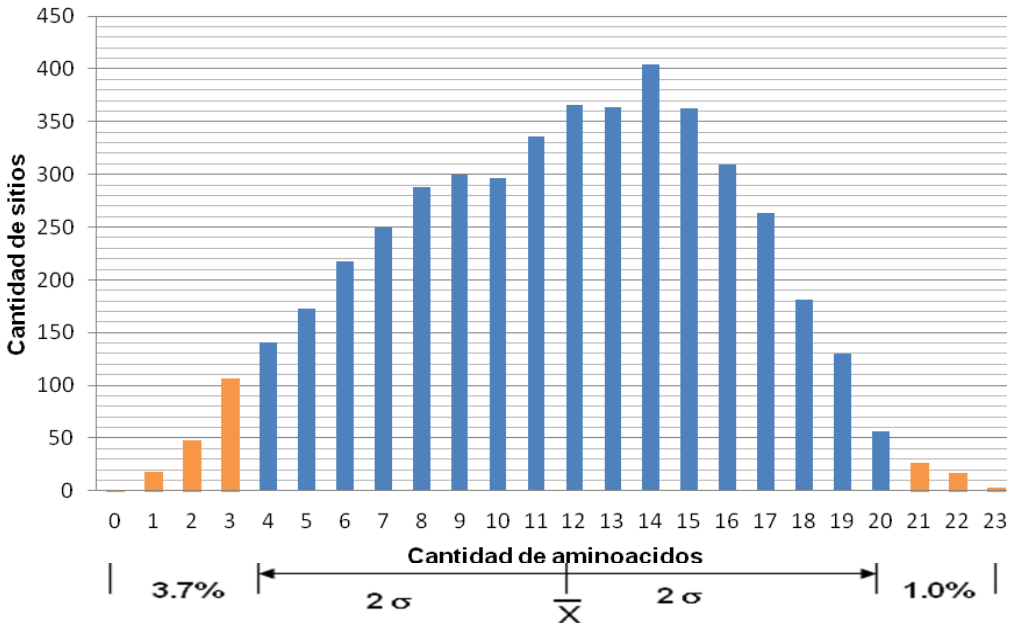

Figura 10. Distribución de cantidad de aminoácidos por sitio. Con promedio aproximado de 12 aminoácidos por sitio y con una desviación estándar de 4 aminoácidos. En naranjo se encuentran aquellos sitios que no se consideraron en este estudio (4,7%).

2. Estimación de atributos

A continuación se muestran todas las tablas resumiendo los resultados para cada uno de los tipos de atributos expuestos en el punto 2 de metodología. Todas las tablas fueron separadas por el nombre del atributo y por cada clase enzimática, para poder apreciar algunas diferencias en los valores para estas últimas.

2.1. Atributos geométricos

2.1.1. Ángulos

Respecto de los atributos relacionados a los ángulos formados por los átomos de los 4 aminoácidos más cercanos al centro de masa del ligando, se puede decir que no se encontraron grandes diferencias entre las distintas clases enzimáticas. Sin embargo, como se puede observar en la Tabla 2 los ángulos formados entre el carbono alfa, el átomo más cercano del mismo aminoácido y el centro de masa del

ligando (cX-aX-lig) presentaron un promedio más alto que los otros tipos de ángulos formados.

Por otro lado, todos los ángulos formados por coordenadas de distintos aminoácidos considerando como vértice el centro de masa del ligando, tienen un promedio aproximado de 90°, con una gran variabilidad de dichos ángulos.

Tabla 2. Ángulos formados por los 4 aminoácidos más cercanos al ligando.

	Oxidoreductasas	**Transferasas**	**Hidrolasas**	**Liasas**	**Isomerasas**	**Ligasas**
Ángulos	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
c1-a1-lig	129,3 ± 25,7	122,1 ± 24,2	129,5 ± 24,1	126,4 ± 24,9	130,6 ± 24,4	125,4 ± 21,7
c2-a2-lig	121,3 ± 28,2	119,6 ± 23,9	125,0 ± 24,6	123,8 ± 25,5	130,6 ± 25,3	126,1 ± 26,2
c3-a3-lig	123,7 ± 24,8	119,0 ± 26,0	122,9 ± 24,9	124,9 ± 25,5	124,5 ± 24,3	120,7 ± 26,2
c4-a4-lig	120,8 ± 24,9	118,1 ± 25,3	121,1 ± 25,7	122,9 ± 25,3	122,8 ± 26,0	118,7 ± 25,3
a1-lig-a2	94,3 ± 33,8	95,6 ± 33,4	96,3 ± 31,5	98,0 ± 34,9	102,2 ± 33,5	92,8 ± 30,7
c1-lig-c2	89,7 ± 36,0	89,5 ± 35,6	87,3 ± 34,2	93,3 ± 34,4	91,1 ± 33,9	85,8 ± 30,9
a1-lig-a3	92,1 ± 35,8	94,4 ± 35,5	90,2 ± 32,3	97,1 ± 34,8	96,3 ± 33,8	94,3 ± 34,3
c1-lig-c3	87,9 ± 37,5	89,1 ± 36,6	84,3 ± 33,6	88,9 ± 34,3	87,6 ± 35,8	94,0 ± 35,0
a1-lig-a4	89,1 ± 36,9	90,4 ± 38,4	87,7 ± 35,8	85,1 ± 35,1	92,0 ± 36,7	91,4 ± 34,0
c1-lig-c4	86,9 ± 38,3	86,6 ± 38,3	82,7 ± 36,3	83,8 ± 36,6	88,8 ± 38,6	86,7 ± 36,2
a2-lig-a3	90,4 ± 36,6	92,4 ± 36,3	91,1 ± 33,2	94,6 ± 35,0	95,1 ± 33,5	92,3 ± 34,6
c2-lig-c3	87,3 ± 36,7	87,3 ± 38,0	84,0 ± 33,9	88,5 ± 36,1	89,3 ± 34,1	91,3 ± 37,8
a2-lig-a4	92,1 ± 38,7	90,0 ± 35,8	89,3 ± 36,1	93,6 ± 35,6	86,6 ± 37,3	85,9 ± 34,8
c2-lig-c4	88,9 ± 39,2	86,7 ± 36,4	85,3 ± 35,8	89,3 ± 36,2	83,8 ± 35,7	84,1 ± 35,7
a3-lig-a4	91,3 ± 35,8	88,2 ± 35,2	91,2 ± 36,0	90,4 ± 37,1	93,9 ± 36,4	94,8 ± 35,6
c3-lig-c4	88,4 ± 37,4	84,6 ± 36,9	87,8 ± 37,4	85,6 ± 36,6	86,6 ± 36,4	96,9 ± 38,5

Todos los ángulos son calculados en grados (°). Se consideraron 2 coordenadas atómicas por aminoácido; **cX** es la coordenada del carbono alfa y **aX** es la coordenada del átomo más cercano del X-ésimo aminoácido más cercano al ligando. Por último, **lig** es la coordenada del centro de masa del ligando presente en el sitio.

2.1.2. Distancias

En relación a las distancias medidas entre los átomos de los 4 aminoácidos más cercanos al centro de masa del ligando, se puede observar en la Tabla 3 que la mayoría de las distancias tienen pequeñas variaciones, exceptuando las distancias entre los carbonos alfa de los aminoácido más cercanos. La principal diferencia visible entre las clases es que las isomerasas presentaron los mayores valores en todas las distancias entre los carbonos alfa de los aminoácidos más cercanos en comparación con las otras clases enzimáticas.

Además, todas las enzimas presentaron en promedio una distancia mínima de 3 Å con el átomo más cercano de los 4 aminoácidos, con la menor desviación estándar observada entre todas las distancias calculadas.

Tabla 3. Distancias formadas por los 4 aminoácidos más cercanos al ligando.

	Oxidorreductasas	**Transferasas**	**Hidrolasas**	**Liasas**	**Isomerasas**	**Ligasas**
Distancia	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
lig_a1	3,0 ± 0,7	3,1 ± 0,8	2,9 ± 0,8	3,1 ± 0,8	3,1 ± 0,8	2,9 ± 0,8
lig_c1	5,8 ± 1,5	5,6 ± 1,4	5,7 ± 1,5	5,8 ± 1,5	6,1 ± 1,5	5,7 ± 1,7
lig_a2	3,7 ± 0,9	3,7 ± 0,9	3,4 ± 1,0	3,6 ± 1,0	3,7 ± 0,9	3,3 ± 0,9
lig_c2	6,0 ± 1,6	6,0 ± 1,6	6,0 ± 1,6	6,2 ± 1,7	6,5 ± 1,7	5,7 ± 1,5
lig_a3	4,2 ± 1,0	4,2 ± 1,0	3,8 ± 1,2	4,0 ± 1,0	4,1 ± 1,0	3,8 ± 1,0
lig_c3	6,4 ± 1,6	6,3 ± 1,7	6,2 ± 1,7	6,4 ± 1,8	6,6 ± 1,6	6,1 ± 1,8
lig_a4	4,7 ± 0,9	4,6 ± 1,0	4,4 ± 1,1	4,5 ± 1,0	4,6 ± 0,9	4,2 ± 1,1
lig_c4	6,7 ± 1,6	6,5 ± 1,7	6,6 ± 1,8	6,7 ± 1,8	7,0 ± 1,8	6,2 ± 1,8
a1_a2	4,7 ± 1,5	4,9 ± 1,6	4,5 ± 1,5	4,8 ± 1,5	5,1 ± 1,6	4,2 ± 1,3
c1_c2	8,1 ± 3,1	7,9 ± 2,9	7,7 ± 2,8	8,4 ± 2,9	8,7 ± 3,1	7,5 ± 2,9
a1_a3	5,0 ± 1,6	5,2 ± 1,7	4,6 ± 1,5	5,1 ± 1,7	5,2 ± 1,6	4,7 ± 1,6
c1_c3	8,2 ± 3,2	8,1 ± 3,1	7,7 ± 2,9	8,4 ± 3,2	8,5 ± 3,4	8,3 ± 3,2
a1_a4	5,4 ± 1,6	5,4 ± 1,8	5,0 ± 1,6	5,0 ± 1,7	5,4 ± 1,7	4,9 ± 1,6
c1_c4	8,3 ± 3,2	8,1 ± 3,1	7,9 ± 3,1	8,1 ± 3,2	8,8 ± 3,6	7,9 ± 3,2
a2_a3	5,3 ± 2,0	5,4 ± 1,9	4,9 ± 1,8	5,3 ± 1,8	5,5 ± 2,0	4,9 ± 1,7
c2_c3	8,3 ± 3,5	8,1 ± 3,3	7,8 ± 3,0	8,4 ± 3,3	9,1 ± 3,7	8,0 ± 2,9
a2_a4	5,7 ± 2,0	5,6 ± 2,0	5,3 ± 1,9	5,6 ± 2,0	5,5 ± 2,0	4,8 ± 1,7
c2_c4	8,5 ± 3,4	8,3 ± 3,2	8,2 ± 3,1	8,8 ± 3,3	8,7 ± 3,6	7,7 ± 3,0
a3_a4	6,0 ± 2,0	5,8 ± 2,1	5,5 ± 2,0	5,6 ± 2,0	6,1 ± 2,1	5,5 ± 2,0
c3_c4	8,8 ± 3,4	8,3 ± 3,3	8,4 ± 3,2	8,5 ± 3,2	9,0 ± 3,7	8,9 ± 3,6

Todas las distancias fueron calculadas en Angstrom (Å). Se consideraron 2 coordenadas atómicas por aminoácido; **cX** es la coordenada del carbono alfa del X-ésimo aminoácido más cercano, **aX** es la coordenada del átomo más cercano del X-ésimo aminoácido más cercano al ligando. Por último, **lig** es la coordenada del centro de masa del ligando presente en el sitio.

2.2. Energías

De las 14 energías obtenidas por medio del *software FoldX*, se puede observar en la Tabla 4 que en general las oxidorreductasas presentaron una menor energía total en sus sitios de unión a ligando. Por otro lado, las ligasas son las enzimas que presentaron valores más altos en cuanto a la entropía de la cadena principal.

A pesar de estas pequeñas diferencias, los valores de las demás energías no tienen diferencias significativas que pueda ayudar a discriminar fácilmente entre cada

una de las clases enzimáticas

Tabla 4. Contribuciones energéticas calculadas con *FoldX.*

	Oxidorreductasas	**Transferasas**	**Hidrolasas**	**Liasas**	**Isomerasas**	**Ligasas**
Contribuciones	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Energía Total	5,5 ± 4,3	6,7 ± 4,8	7,8 ± 6,6	7,7 ± 5,9	7,3 ± 5,2	7,7 ± 4,9
Puentes de H cadena principal	-1,8 ± 1,6	-1,8 ± 1,6	-1,9 ± 1,8	-1,8 ± 1,5	-1,6 ± 1,5	-2,0 ± 1,5
Puentes de H cadena secundaria	-1,3 ± 1,4	-1,4 ± 1,4	-1,8 ± 1,8	-1,8 ± 1,7	-1,7 ± 1,6	-1,6 ± 1,5
Van der Waals	-4,0 ± 3,2	-3,8 ± 2,8	-4,8 ± 3,5	-4,7 ± 3,1	-4,7 ± 3,4	-5,0 ± 3,8
Electroestáticas	-0,2 ± 0,8	0,0 ± 1,4	0,3 ± 2,5	0,1 ± 1,7	-0,3 ± 1,4	-0,2 ± 0,6
Solvatación Polar	6,1 ± 5,1	6,5 ± 4,8	8,5 ± 6,6	8,1 ± 5,9	8,2 ± 6,7	7,9 ± 5,2
Solvatación Hidrofóbica	-4,9 ± 4,0	-4,5 ± 3,4	-5,5 ± 4,0	-5,5 ± 3,6	-5,5 ± 4,0	-5,9 ± 4,7
Choques de Van der Waals	0,8 ± 1,4	0,7 ± 0,9	1,0 ± 1,3	0,9 ± 1,0	0,7 ± 0,9	0,9 ± 0,9
Entropía cadena secundaria	3,6 ± 2,6	3,5 ± 2,4	4,2 ± 2,8	4,4 ± 3,0	4,6 ± 3,1	4,5 ± 2,7
Entropía cadena principal	6,0 ± 4,3	6,3 ± 4,4	6,5 ± 4,7	6,5 ± 4,2	5,7 ± 4,0	7,6 ± 5,2
Coques Torsionales	1,1 ± 0,9	1,1 ± 0,8	1,2 ± 0,9	1,4 ± 1,1	1,6 ± 1,1	1,5 ± 1,2
Choques cadena principal	2,3 ± 1,8	2,3 ± 1,8	2,1 ± 1,8	2,2 ± 1,6	1,8 ± 1,6	2,8 ± 2,3
Disulfuro	0,0 ± 0,2	0,0 ± 0,0	0,0 ± 0,3	0,0 ± 0,3	0,0 ± 0,0	0,0 ± 0,0
Electroestáticas grupos cargados	0,0 ± 0,1	0,0 ± 0,1	0,0 ± 0,1	0,0 ± 0,1	0,0 ± 0,2	0,0 ± 0,1
Energía Ionización	0,1 ± 0,3	0,0 ± 0,1	0,2 ± 0,3	0,1 ± 0,2	0,2 ± 0,3	0,1 ± 0,1
Número de residuos	179 ± 154	148 ± 114	162 ± 132	176 ± 122	186 ± 108	130 ± 105

Todas las contribuciones para el cálculo de la energía total (primera fila) se muestran a la izquierda medidas en Kcal/mol. Puentes de H se denomina a los puentes de hidrógeno.

2.3. Hidrofobicidades y composición aminoacídica

Los valores de hidrofobicidad dependen de la escala que se está utilizando. En este estudio se utilizaron las 3 escalas descritas en el punto 2.3 de la metodología. Como se puede observar en la Tabla 5 en general los sitios de unión a ligando son hidrofílicos, independientemente de la clase enzimática a la cual pertenezcan. Dado que para las 2 primeras escalas un valor negativo en un aminoácido indica que tiene carácter hidrofílico y para la tercera escala un valor positivo indica lo mismo. Sin embargo, se puede destacar que las oxidorreductasas son las enzimas que presentaron los menores valores en las 3 escalas.

En cuanto a las 4 proporciones de los distintos tipos de aminoácidos, se puede observar en la Tabla 5 que las oxidorreductasas presentaron mayor cantidad de aminoácidos apolares en comparación con las otras clases enzimáticas. Además, estas enzimas también poseen una menor cantidad de aminoácidos negativos.

Por último, en este tipo de propiedades de los sitios de unión a ligando, no se puede obtener una diferencia significativa entre las distintas clases

Tabla 5. Valores de hidrofobicidad y composición aminoacídica.

		Oxidorreductasas	Transferasas	Hidrolasas	Liasas	Isomerasas	Ligasas
		$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Escala	(1)	-4,1 ± 11,7	-8,4 ± 12,9	-11,0 ± 11,9	-9,6 ± 13,0	-12,2 ± 14,4	-10,8 ± 14,0
	(2)	-2,0 ± 3,2	-3,1 ± 3,2	-3,5 ± 3,7	-3,0 ± 3,9	-3,4 ± 3,8	-4,2 ± 3,9
	(3)	11,7 ± 5,8	14,1 ± 6,8	15,2 ± 7,7	15,0 ± 7,0	16,3 ± 8,2	15,6 ± 7,2
Aminoácidos	Apolares	0,37 ± 0,16	0,32 ± 0,17	0,32 ± 0,16	0,31 ± 0,17	0,32 ± 0,16	0,29 ± 0,16
	Polares	0,36 ± 0,18	0,36 ± 0,17	0,35 ± 0,16	0,38 ± 0,16	0,32 ± 0,17	0,36 ± 0,15
	Positivos	0,17 ± 0,13	0,17 ± 0,14	0,16 ± 0,12	0,18 ± 0,13	0,19 ± 0,12	0,20 ± 0,11
	Negativos	0,09 ± 0,11	0,15 ± 0,13	0,17 ± 0,13	0,13 ± 0,11	0,16 ± 0,13	0,15 ± 0,11
	Total	11,8 ± 3,9	11,6 ± 3,8	11,7 ± 4,3	12,6 ± 4,1	12,5 ± 3,8	12,4 ± 4,1

Escala 1: (Kyte & Doolittle, 1982); Escala 2: (Wimley & White, 1996); Escala 3: (Hessa y col., 2005). Todos los valores de composición aminoacídica se encuentran en proporciones de la cantidad de cada tipo de aminoácido divido en el total (última fila).

2.4. Composición atómica por capas

El análisis de los átomos presentes en distintas capas otorga mucha información respecto de las interacciones que pueden estar ocurriendo en el sitio de unión a ligando. Por lo general, como se puede observar en la Tabla 6, en la capa que va desde 1 a 2 Å no se encontraron átomos, esto se evidencia gracias a que sus promedios son muy cercanos a 0. En la siguiente capa (entre 2 a 3 Å) se observaron entre 1 a 2 átomos, siendo las ligasas las que presentaron la mayor presencia de azufre en dicha capa en relación a las otras clases enzimáticas. Por otro lado, en la misma capa las oxidorreductasas presentan mayor presencia de nitrógeno en relación a las demás clases. Además, esta misma clase en la capa que va de 3 a 4 Å es la clase que posee la mayor presencia de azufre. Por último, a pesar de estas pequeñas diferencias anteriormente mencionadas, no se pueden determinar tendencias en los valores que permitan llegar a patrones específicos en los tipos de

átomos para cada una de las distintas clases enzimáticas.

Tabla 6. Composición atómica por capas y total.

		Oxidorreductasas	Transferasas	Hidrolasas	Liasas	Isomerasas	Ligasas
	Átomo	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Capa 1-2 Å	C	0,0 ± 0,0	0,2 ± 4,4	0,1 ± 3,4	0,0 ± 0,0	0,0 ± 0,0	1,5 ± 11,3
	O	2,9 ± 16,5	2,1 ± 14,3	4,2 ± 19,9	3,0 ± 16,8	4,3 ± 19,1	2,4 ± 15,3
	N	2,2 ± 14,2	0,9 ± 9,1	1,5 ± 11,8	0,5 ± 6,8	3,7 ± 17,7	3,3 ± 17,2
	S	0,2 ± 5,0	0,2 ± 4,7	0,0 ± 0,0	0,0 ± 0,0	0,0 ± 0,0	0,5 ± 6,9
	Total	0,1 ± 0,3	0,0 ± 0,2	0,1 ± 0,3	0,0 ± 0,2	0,1 ± 0,3	0,1 ± 0,4
Capa 2-3 Å	C	11,6 ± 25,4	6,0 ± 19,2	7,0 ± 18,1	6,6 ± 19,4	9,4 ± 26,1	8,2 ± 20,5
	O	19,8 ± 36,9	27,0 ± 41,9	34,7 ± 43,6	27,9 ± 41,6	23,6 ± 39,8	34,8 ± 45,5
	N	17,8 ± 32,5	8,0 ± 24,9	11,7 ± 27,6	11,7 ± 28,7	7,3 ± 21,2	7,6 ± 21,6
	S	3,5 ± 16,8	3,9 ± 18,1	2,8 ± 15,1	0,9 ± 8,0	0,6 ± 7,5	11,9 ± 28,6
	Total	1,2 ± 1,6	1,2 ± 1,7	1,8 ± 2,1	1,3 ± 1,9	1,0 ± 1,5	1,9 ± 1,9
Capa 3-4 Å	C	49,8 ± 37,6	49,6 ± 39,1	50,3 ± 35,9	45,9 ± 37,0	44,6 ± 36,5	57,2 ± 35,0
	O	16,3 ± 26,0	17,6 ± 28,4	22,4 ± 28,7	22,2 ± 28,7	24,0 ± 30,1	15,2 ± 26,6
	N	16,5 ± 24,9	17,0 ± 26,7	15,7 ± 24,1	17,5 ± 25,8	18,1 ± 28,3	18,9 ± 25,9
	S	5,1 ± 20,7	0,6 ± 7,0	0,4 ± 4,6	1,9 ± 10,5	0,3 ± 2,4	0,5 ± 4,9
	Total	3,3 ± 2,5	3,1 ± 2,6	3,8 ± 3,0	3,5 ± 2,7	3,1 ± 2,7	3,8 ± 2,3
Capa 4-5 Å	C	62,2 ± 21,5	61,1 ± 20,4	60,4 ± 19,4	60,8 ± 20,9	63,2 ± 22,5	58,6 ± 18,7
	O	14,7 ± 18,8	17,6 ± 16,1	19,9 ± 18,6	16,7 ± 19,1	18,0 ± 16,0	17,6 ± 17,7
	N	21,6 ± 17,6	19,4 ± 15,5	19,1 ± 14,7	21,3 ± 17,8	17,4 ± 16,7	21,8 ± 14,2
	S	0,8 ± 3,6	0,6 ± 3,3	0,3 ± 1,9	0,8 ± 3,7	0,3 ± 2,0	1,0 ± 4,1
	Total	8,2 ± 4,9	9,2 ± 4,9	10,5 ± 5,6	10,0 ± 5,7	9,0 ± 4,7	9,8 ± 5,2
Capa 5-6 Å	C	66,3 ± 13,9	62,8 ± 14,9	65,2 ± 13,9	63,6 ± 13,1	66,1 ± 11,7	60,9 ± 14,7
	O	15,6 ± 11,1	17,7 ± 11,4	17,7 ± 11,3	17,6 ± 10,4	15,5 ± 9,7	18,1 ± 9,5
	N	17,3 ± 11,1	19,0 ± 11,6	16,7 ± 10,4	18,3 ± 11,7	17,9 ± 10,0	20,7 ± 11,3
	S	0,7 ± 2,7	0,5 ± 2,3	0,5 ± 1,9	0,5 ± 1,7	0,5 ± 1,7	0,4 ± 1,5
	Total	13,8 ± 6,2	13,8 ± 5,8	14,1 ± 6,1	14,8 ± 5,4	14,8 ± 5,4	14,8 ± 6,2
Capa 6-7 Å	C	63,2 ± 11,2	61,3 ± 12,3	61,5 ± 12,4	62,2 ± 11,4	62,9 ± 12,1	60,8 ± 13,1
	O	17,7 ± 8,9	19,7 ± 10,3	20,5 ± 10,6	18,9 ± 9,1	17,9 ± 8,9	20,3 ± 10,4
	N	18,5 ± 9,3	18,5 ± 9,4	17,6 ± 9,1	18,4 ± 8,8	18,8 ± 8,4	18,2 ± 9,6
	S	0,6 ± 1,9	0,5 ± 1,7	0,5 ± 1,8	0,5 ± 1,7	0,4 ± 1,3	0,6 ± 1,7
	Total	20,1 ± 7,5	18,7 ± 6,7	18,4 ± 7,3	20,0 ± 7,0	21,3 ± 6,8	20,4 ± 7,2
	Total	96,0 ± 36,4	93,1 ± 30,5	97,0 ± 37,6	102,4 ± 33,7	105,1 ± 32,8	104,2 ± 41,7

Todos los valores están expresados en porcentajes a excepción de las cantidades totales de átomos

2.5. Radios de sitio

Un aspecto importante para la predicción de función enzimática es caracterizar el tamaño del bolsillo formado por los aminoácidos cercanos al ligando. En este aspecto, como se puede observar en la Tabla 7 no se encontraron diferencias significativas entre las distintas clases enzimáticas. Aunque, si se puede observar

que las isomerasas y las hidrolasas son las que presentan mayores y menores radios de sitio respectivamente, teniendo en consideración las 4 medidas.

Tabla 7. Distancias al centro de masa del sitio de unión a ligando.

	Oxidorreductasas	Transferasas	Hidrolasas	Liasas	Isomerasas	Ligasas
Distancia	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Mínima	2,2 ± 0,9	2,2 ± 0,9	1,9 ± 0,8	2,3 ± 0,9	2,4 ± 0,9	2,0 ± 0,9
Máxima	6,9 ± 1,0	6,9 ± 1,0	6,7 ± 1,1	6,9 ± 0,9	7,0 ± 0,8	6,8 ± 1,0
Promedio	4,7 ± 0,7	4,7 ± 0,7	4,5 ± 0,7	4,7 ± 0,7	4,9 ± 0,5	4,6 ± 0,7
Desviación Estándar	1,4 ± 0,3	1,4 ± 0,3	1,5 ± 0,3	1,4 ± 0,3	1,4 ± 0,4	1,4 ± 0,3

Las distancias están expresadas en (Å). El promedio y desviación estándar de la izquierda son distancias por sitio.

2.6. Estructura secundaria y área accesible al solvente

En cuanto al cálculo de la proporción de estructura secundaria de los aminoácidos que conforman el sitio, se puede destacar que las hidrolasas y ligasas son las que presentaron menor cantidad de hélices. Por otro lado, las isomerasas y ligasas son las clases enzimas que mayor proporción de sabanas beta presentaron en comparación con las otras clases.

En relación al área accesible al solvente (SASA) de los sitios, que es calculada eliminado el ligando y considerando la obstrucción que genera el resto de la proteína para un sitio de unión a ligando. Se puede observar en la Tabla 8 que las hidrolasas son las que presentaron el menor área y las oxidorreductasas la mayor área de todas las clases enzimáticas.

Por último, la elevada desviación estándar en los datos no permitió obtener tendencias en las distintas clases enzimáticas para este tipo de información.

Tabla 8. Tipos de estructura secundaria y área accesible al solvente.

	Oxidorreductasas	Transferasas	Hidrolasas	Liasas	Isomerasas	Ligasas
Tipo	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Coil	0,44 ± 0,19	0,47 ± 0,19	0,51 ± 0,26	0,46 ± 0,22	0,44 ± 0,18	0,47 ± 0,19
Hélices	0,36 ± 0,25	0,32 ± 0,23	0,24 ± 0,25	0,31 ± 0,27	0,28 ± 0,21	0,25 ± 0,17
Sabanas	0,20 ± 0,20	0,22 ± 0,18	0,24 ± 0,23	0,23 ± 0,20	0,28 ± 0,22	0,28 ± 0,18
SASA	851 ± 561	768 ± 462	546 ± 310	608 ± 340	718 ± 470	694 ± 433

Los 3 valores de estructura secundaria se encuentran expresados en proporciones. El área accesible al solvente (SASA) está expresado en $Å^2$

2.7. Puntajes evolutivos

Las *PSSM* fueron obtenidas a partir de las secuencias de aminoácidos en dos tipos. Una con los aminoácidos más cercanos al centro de masa del ligando ordenados según su cercanía y otra con todos los aminoácidos pertenecientes al sitio conservando el orden de aparición en la estructura primaria. A partir de estas secuencias se generaron las secuencias consenso para cada una de las clases enzimáticas. Estas secuencias se muestran en la Tabla 9, en donde se puede evidenciar en el caso de las secuencias de los 12 primeros aminoácidos (segunda columna) una clara presencia de aspartato en casi todas las clases, exceptuando las oxidorreductasas y las ligasas. En el caso del consenso generado con todos los aminoácidos del sitio (tercera columna), se observó una mayor diferencia entre cada una de las clases, siendo lo más destacable la alta presencia de aspartato en hidrolasas y de glicina en las demás clases.

Tabla 9. Consensos de aminoácidos por clase enzimática.

Clase\Consenso	**Cercanía**	**Todo el sitio**
Oxidorreductasas	HHHGGGGGGGTG	GGGGGHVGGGGGGLGTAHHH
Transferasas	DDDDGGGGGGRG	RDGGGTGDGGGDGGGGYFFG
Hidrolasas	DDDDDDDGGGGG	DDDDDDDDDDDDDDDDGTHH
Liasas	DDDDGGSGGGAA	RDDDGDDGARGGAGYGRLES
Isomerasas	DDDERGGGLWGG	RGGKGGEDDGHHDDDDFIQM
Ligasas	HCCCTDEERGGA	RGEGGGEQCGNEEGDRRRYC

2.7.1. Aminoácidos más cercanos

Como se puede observar en la Tabla 10, los puntajes de las matrices fueron en la mayoría de los casos mayores para la clase con la cual fue generada la matriz, exceptuando el caso de las oxidorreductasas y transferasas que su mayor puntaje lo tuvieron con la matriz de las ligasas.

Por otro lado, en el caso de las oxidorreductasas, solo el puntaje generado con la matriz de las transferasas fue menor al generado con la misma matriz de la clase. Lo mismo ocurre con las transferasas, que solo el puntaje generado con la matriz de las oxidorreductasas fue menor al generado con la propia matriz de la clase de las transferasas. Esto sugiere que las *PSSM* a partir de los aminoácidos más cercanos no son útiles para las 2 primeras clases.

Tabla 10. Puntajes con secuencias de aminoácidos más cercanos al ligando.

	Oxidorreductasas	**Transferasas**	**Hidrolasas**	**Liasas**	**Isomerasas**	**Ligasas**
Matriz	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Oxidorreductasas	**11,8 ± 3,4**	11,4 ± 3,3	11,3 ± 3,7	11,9 ± 3,4	11,3 ± 3,8	11,5 ± 3,2
Transferasas	11,6 ± 3,8	**12,5 ± 4,1**	12,4 ± 4,5	12,7 ± 4,1	12,3 ± 4,5	12,4 ± 4,2
Hidrolasas	11,7 ± 4,0	12,5 ± 4,3	**13,4 ± 5,4**	13,0 ± 4,5	13,1 ± 5,1	12,7 ± 4,8
Liasas	12,0 ± 3,7	12,6 ± 4,0	12,8 ± 4,6	**13,4 ± 4,3**	12,7 ± 4,5	12,6 ± 4,3
Isomerasas	11,9 ± 3,5	12,7 ± 3,7	13,2 ± 4,7	13,1 ± 3,9	**13,8 ± 4,6**	12,9 ± 4,3
Ligasas	12,5 ± 3,9	13,1 ± 4,1	13,1 ± 4,7	13,4 ± 4,3	13,1 ± 4,7	**14,4 ± 4,8**

En el lado izquierdo de la tabla se encuentran las *PSSM* generadas con cada una de las secuencias de las distintas clases. Hacia abajo se observan los distintos puntajes obtenidos para una clase con las distintas matrices de todas las clases.

2.7.2. Todos los aminoácidos del sitio

En el caso de los puntajes obtenidos con las secuencias de aminoácidos del sitio completo. Se puede observar en la Tabla 11 que las oxidorreductasas, transferasas e hidrolasas en general no presentaron su máximo puntaje con la matriz de su propia clase. Además, estas tres clases tienen en general el máximo puntaje con la matriz de las ligasas. Todo esto sugiere que para las 3 primeras clases las *PSSM* generadas a partir de todos los aminoácidos perecientes al sitio no ayudarían a discriminarlas. Más aun, esta información ayudaría a clasificarlas erróneamente como ligasas

Tabla 11. Puntajes con secuencias de todos los aminoácidos del sitio.

	Oxidorreductasas	**Transferasas**	**Hidrolasas**	**Liasas**	**Isomerasas**	**Ligasas**
Matriz	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$	$\overline{X} \pm \sigma$
Oxidorreductasas	**11,2 ± 3,7**	11,2 ± 3,6	10,9 ± 3,9	11,7 ± 3,8	10,9 ± 3,9	11,2 ± 3,3
Transferasas	12,1 ± 4,3	**13,2 ± 4,5**	12,8 ± 4,9	13,6 ± 4,6	13,0 ± 4,9	13,2 ± 4,6
Hidrolasas	11,4 ± 4,0	12,4 ± 4,3	**12,8 ± 5,0**	13,0 ± 4,5	12,8 ± 5,0	12,6 ± 4,7
Liasas	12,4 ± 4,4	13,2 ± 4,6	13,2 ± 5,1	**14,2 ± 4,9**	13,3 ± 5,1	13,4 ± 4,7
Isomerasas	12,2 ± 3,9	13,2 ± 4,1	13,2 ± 4,8	13,8 ± 4,2	**14,4 ± 5,0**	13,5 ± 4,7
Ligasas	12,7 ± 4,4	13,5 ± 4,5	13,6 ± 5,2	14,1 ± 4,8	13,9 ± 5,1	**14,4 ± 5,5**

En el lado izquierdo de la tabla se encuentran las *PSSM* generadas con cada una de las secuencias de las distintas clases. Hacia abajo se observan los distintos puntajes obtenidos para una clase con las distintas matrices de todas las clases.

Después de analizar los distintos atributos, se puede concluir que en general estas propiedades no logran discriminar a simple vista entre cada una de las 6 clases

enzimáticas. Además, todas las propiedades presentaron alta variabilidad dentro de cada clase. Por lo tanto, esto justifica el uso de métodos de aprendizaje automático que permitan predecir a través de estos atributos la clasificación enzimática.

3. Entrenar y validar el modelo

A partir de todos los atributos previamente descritos se generó un modelo de clasificación basado en *Random Forest* y utilizando la metodología de 10 validaciones cruzadas para evaluar el desempeño del modelo generado en cada iteración.

Como se puede observar en la Tabla 12, se logró obtener una exactitud global del 70%. Esto representa un desempeño bajo en comparación con los estudios descritos anteriormente, que bordean el 80% de exactitud. Por otro lado, se observa para las hidrolasas una elevada tasa de 94% de verdaderos positivos, que es muy superior a las tasas de las demás clases que no superar el 60%. Además de la alta tasa de verdaderos positivos que presenta la clase 3, ésta es la clase que mayor cantidad de falsos positivos obtiene, con una tasa del 37%, muy superior a las otras clases que no superan el 7%.

Tabla 12. Desempeño del modelo de clasificación.

EC-Clase	VP	FP	Precisión	Recall	MCC	ROC
1-Oxidorreductasa	0,60	0,02	0,88	0,60	0,65	0,92
2-Transferasa	0,57	0,07	0,71	0,57	0,60	0,88
3-Hidrolasa	0,94	0,37	0,62	0,94	0,64	0,90
4-Liasa	0,50	0,00	0,99	0,50	0,57	0,91
5-Isomerasa	0,49	0,00	1,00	0,49	0,57	0,89
6-Ligasa	0,46	0,00	0,94	0,46	0,55	0,92
Promedio	**0,70**	**0,17**	**0,76**	**0,70**	**0,64**	**0,90**

Las tasas de verdaderos positivos (TVP), las tasas de falsos positivos (TFP), la Precisión, el Recall, el Coeficiente de correlación de Matthews (MCC) y por último el área bajo la curva ROC para cada una de las distintas clases enzimáticas.

Dado todos estos indicadores, se evidencia que la mayor cantidad de errores que comete el modelo es al predecir cualquiera de las otras clases como una hidrolasa. Por otro lado, se puede destacar que la precisión para casi todas las clases es alta, exceptuando las hidrolasas que obtienen solo un 62%. Esto quiere decir que para las

demás clases no se están cometiendo muchos errores al predecir.

Todo esto demuestra que efectivamente existe un problema de desbalance de clases y que el método de *Random Forest* no es capaz de enfrentar y sobrepasar estas condiciones.

Considerando el alto desbalance de clases en los ejemplos presentado en el punto 1 de estos resultados y el claro sesgo en las predicciones hacia la clase de las hidrolasas (clase 3). Se decidió realizar la metodología de *oversampling* (Guo y col., 2008), para corroborar si el sesgo hacia la clase 3 es efectivamente a consecuencia de la mayor cantidad de ejemplos que ésta posee y consecuentemente lograr superar el problema de desbalance de clases.

Oversampling es un método ampliamente utilizado en casos de escases de datos de una clase, que consiste básicamente en repetir aleatoriamente los ejemplos de la clase minoritaria. Para este nuevo objetivo, se utilizó la rutina *SMOTE* (Chawla & Bowyer, 2011) dentro del *software Weka*, la cual implementa el método de *oversampling* recibiendo como parámetros la clase y el porcentaje de aumento de los ejemplos. Los porcentajes de aumento para las distintas clases fueron los siguientes:

- 100% para las clases 1
- 75% para la clase 2
- Nada para la clase 3
- 250% para la clase 4
- 900% para la clase 5
- 700% para la clase 6.

Todo esto se realizó con el objetivo final de balancear las clases, es decir dejar la cantidad de ejemplos de cada clase similares al total de ejemplos de las hidrolasas.

Luego de entrenar el modelo con los nuevos ejemplos generados aleatoriamente, en la Tabla 13 se puede observar que se logró obtener una exactitud global del 90%, que es muy superior a la obtenida anteriormente. Además, se observó un aumento considerable en las predicciones de cada clase, ya que todas superaron el 80% de

casos correctamente clasificados. Por otro lado, también se reduce significativamente las tasas de falsos positivos, que en general no superan el 4%. Sin embargo, las hidrolasas pasaron a ser la clase que tiene el menor desempeño. Esto se debe a que al realizar el *oversampling,* se obtuvo un equilibrio en cada uno de los indicadores de desempaño para las distintas clases enzimáticas.

Tabla 13. Desempeño del modelo luego de realizar oversampling.

EC-Clase	VP	FP	Precisión	Recall	MCC	ROC
1-Oxidorreductasa	0,84	0,02	0,91	0,84	0,84	0,98
2-Transferasa	0,85	0,04	0,81	0,85	0,83	0,98
3-Hidrolasa	0,80	0,04	0,80	0,80	0,79	0,96
4-Liasa	0,91	0,01	0,96	0,91	0,91	0,99
5-Isomerasa	0,99	0,01	0,96	0,99	0,98	1,00
6-Ligasa	0,98	0,01	0,95	0,98	0,97	1,00
Promedio	**0,90**	**0,02**	**0,90**	**0,90**	**0,88**	**0,98**

Las tasas de verdaderos positivos (TVP), las tasas de falsos positivos (TFP), la Precisión, el Recall, el Coeficiente de correlación de Matthews (MCC) y por último el área bajo la curva ROC para cada una de las distintas clases enzimáticas.

Finalmente, se puede decir que con la metodología de *oversampling* se mejoran todas las medidas de desempeño del modelo generado con el método de *Random Forest*, sobrepasando el problema de escases de ejemplos para las clases minoritarias.

4. Evaluar estrategias de clasificación

4.1. Nuevas proteínas con ligando

Comenzando con la evaluación de redundancia de los nuevos ejemplos (NE) junto con las proteínas utilizadas en entrenamiento (PDBs-NR) se puede observar en la Tabla 14 que se encontraron 20 nuevos grupos y que cada clase enzimática tiene nuevos datos que no se parecen a ningún ejemplo de entrenamiento, es decir que no tienen un porcentaje de identidad mayor a 25% con ninguna enzima utilizada anteriormente. Por otro lado, de las 781 nuevas proteínas encontradas en *PDB* se generaron 92 grupos al reducir redundancia entre las enzimas de la misma clase.

Tabla 14. Evaluación de reducción de redundancia incluyendo nuevos ejemplos.

EC-Clase	PDBs-NR	NE	PDBs+NE-NR	NE-NR
1-Oxidorreductasa	314	130	317	19
2-Transferasa	530	276	537	23
3-Hidrolasa	687	301	694	43
4-Liasa	212	24	213	3
5-Isomerasa	118	16	119	1
6-Ligasa	99	34	100	3
Promedio	**1960**	**781**	**1980**	**92**

Los nuevos ejemplos (NE) fueron incluidos con todos los archivos de proteínas utilizados en entrenamiento (PDBs), para realizar una reducción de redundancia por clase (PDBs+NE-NR). La reducción de redundancia entre los ejemplos nuevos (NE-NR) se muestra en la última columna.

Posteriormente, al conjunto de 781 proteínas nuevas se les extrajo los sitios de unión a ligando, con los mismos pasos mostrados en la sección 1.2 de la metodología. En la Tabla 15 se puede observar el total de 2.225 sitios para 583 enzimas. Dado que algunos sitios no cumplían los requerimientos expuestos en metodología, quedaron 198 proteínas fuera del análisis. Cabe destacar que en estos nuevos ejemplos también se encontró un claro desbalance en la cantidad de proteínas y más aun en la cantidad de sitios analizados, alcanzando una proporción de 20:1 entre la clase mayoritaria y minoritaria.

Tabla 15. Cantidad de proteínas y sitios de los nuevos ejemplos.

EC-Clase	PDBs	Ligandos	Sitios de Unión	PDBs con sitios
1-Oxidorreductasa	130	1339	114	512
2-Transferasa	276	2031	180	654
3-Hidrolasa	301	2779	233	846
4-Liasa	24	190	14	45
5-Isomerasa	16	171	12	42
6-Ligasa	34	657	30	126
Promedio	**781**	**7167**	**583**	**2225**

De un total de 781 proteínas (PDBs) con sus respectivos ligandos se extrajeron los sitios. La cantidad de proteínas que quedaron luego de la extracción de sitios (PDBs*) disminuyó.

Al clasificar los nuevos sitios con el modelo generado anteriormente se obtuvo una exactitud global de 51%. Por otro lado, también se obtuvo una alta tasa de falsos positivos para las hidrolasas y muy bajas tasas de verdaderos positivos para las demás clases. Esto puede ser a consecuencia de que los estos datos presentan nueva información para el modelo, que antes no tenía disponible para el

entrenamiento.

Considerando el bajo desempeño obtenido al clasificar estas nuevas estructuras, se decidió incluirlas en la generación del modelo predictivo, para evidenciar el impacto en el modelo al agregar esta nueva información. De este análisis se puede observar el resultado de la clasificación en la Tabla 16. En donde se obtuvieron resultados muy similares a los iniciales, incluyendo el mismo estudio de *oversampling* se puede observar apenas una diferencia del 0,4% de disminución en la exactitud global en relación al desempeño obtenido anteriormente sin la inclusión de los nuevos ejemplos, siendo esta nueva tasa de un 89%. Además. se conserva el equilibrio alcanzado entre las distintas clases. Destacando nuevamente que todas superan el 80% de casos correctamente clasificados y que todas obtienen baja cantidad de falsos positivos, no superiores al 5%

Tabla 16. Desempeño del modelo al incluir nuevos ejemplos.

EC-Clase	VP	FP	Precisión	Recall	MCC	ROC
1-Oxidorreductasa	0,83	0,02	0,91	0,83	0,83	0.98
2-Transferasa	0,86	0,05	0,79	0,86	0,83	0.97
3-Hidrolasa	0,80	0,04	0,79	0,80	0,78	0.96
4-Liasa	0,89	0,01	0,97	0,89	0,89	0.99
5-Isomerasa	0,99	0,01	0,97	0,99	0,98	0.99
6-Ligasa	0,98	0,01	0,96	0,98	0,97	0.99
Promedio	**0,89**	**0,02**	**0,90**	**0,89**	**0,88**	**0.98**

Las tasas de verdaderos positivos (VP), las tasas de falsos positivos (FP), la Precisión, el Recall, el Coeficiente de correlación de Matthews (MCC) y por último el área bajo la curva ROC para cada una de las distintas clases enzimáticas.

Todo lo anteriormente expuesto demuestra que el modelo predictivo presenta robustez, ya que es capaz de incluir nueva información y adaptarse para lograr un buen desempeño. Además, esta característica hace que el modelo sea actualizable en el tiempo ya que al surgir nuevas estructuras de enzimas, se puede incluir esa nueva información y no se verá afectado el desempeño global del modelo.

4.2. Proteínas sin ligando

En esta sección se procedió a clasificar 30 proteínas que no poseían ligandos co-cristalizados en su estructura. De las 30 proteínas seleccionadas para este análisis, se predijeron 10 sitios de unión a ligando por estructura, utilizando el servidor web *Q-siteFinder*. Posteriormente, se realizó la extracción de estos sitios, considerando como centro del ligando el centro geométrico de la nube de puntos en donde el servidor predecía la posición de un ligando. Se encontraron 265 sitios de unión a ligando que cumplían con los mismos requerimientos establecidos en el punto 1.2 de metodología.

Considerando cada uno de estos sitios como ejemplos separados de clasificación se obtuvo apenas un 22,8% de casos correctamente clasificados. En estos ejemplos la clase que mejor tasa de verdaderos positivos obtiene son las transferasas, con un 74% de correctamente clasificados. Aunque también son las enzimas que tienen la mayor tasa de falsos positivos, un 46,6%.

De estos resultados se puede decir que en este tipo de enfoque falta mucho por refinar en cuanto al método de predicción de la posición en donde se localizaría un ligando. Pero este tipo de estrategias logran al menos entregar una aproximación de la función que podría desempeñar la proteína en un posible sitio de unión a ligando.

4.3. Clasificar estructuras

Una proteína generalmente tiene más de un ligando o se puede predecir más de un sitio de unión a ligando utilizando métodos que realizan dicha tarea. Por tal motivo, para clasificar una nueva estructura proteica a una de las 6 clases enzimáticas se puede realizar un sistema de votación simple, en donde la clase asignada a la proteína de entrada es la moda o clase más repetida dentro de todas las predicciones individuales de sus sitios de unión a ligando. Como se puede observar en la Figura 11, una proteína perteneciente a la clase de transferasas (*EC*: 2; *PDB ID*: 3R6C) utilizada en entrenamiento con todas las predicciones de sus sitios de unión a ligando. Esta proteína posee 7 sitios, de los cuales 4 son predichos correctamente. Por lo tanto, al realizar una clasificación completa de la proteína, se

puede decir que se predice correctamente la función enzimática global de dicha proteína.

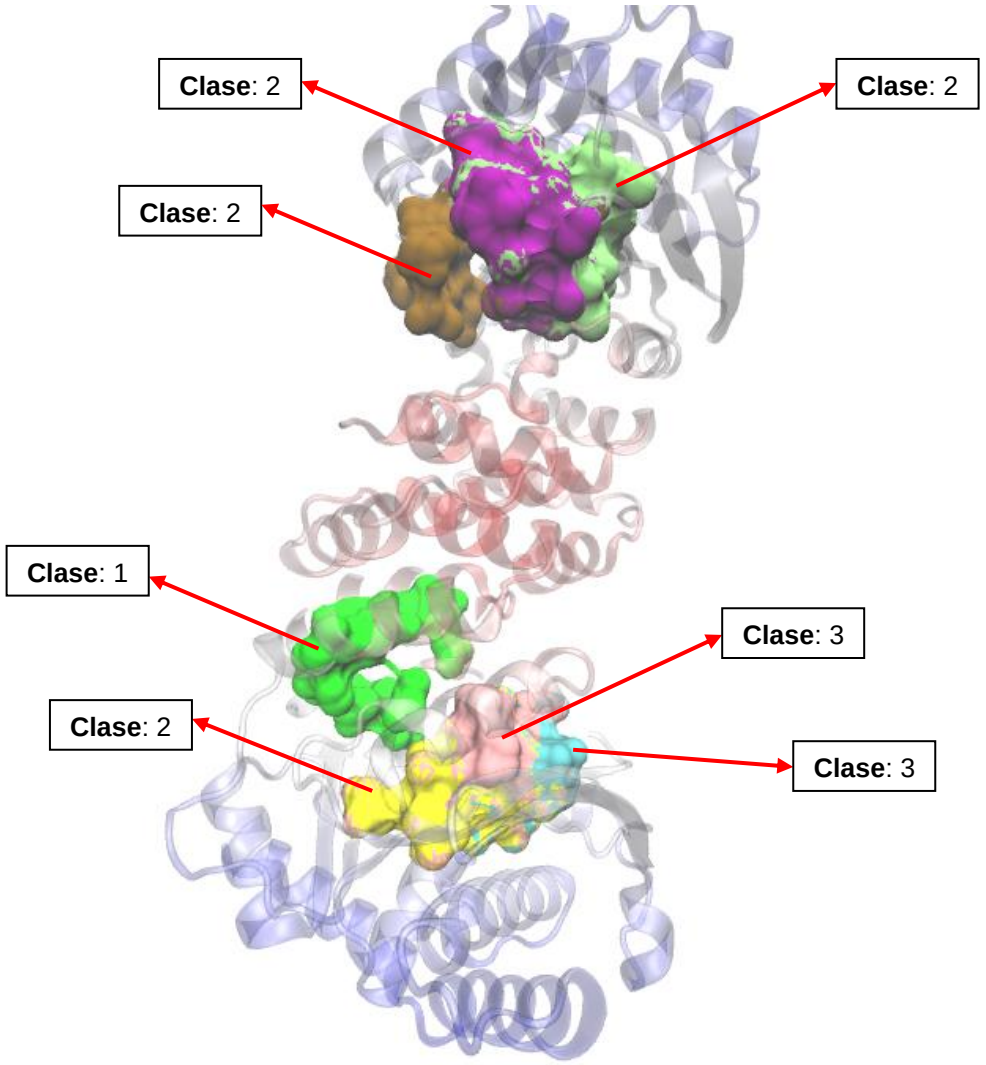

Figura 11. Ejemplo de esquema de clasificación de proteínas. En los cuadros de texto se observan las distintas predicciones para cada sitio de unión a ligando presentes en la proteína (PDB ID: 3R6C) de clasificación transferasa (EC: 2). Donde 4 de los 7 sitios de unión a ligando son correctamente clasificados.

DISCUSIÓN

En este trabajo se generó un modelo de clasificación que relaciona directamente las propiedades de los sitios de unión a ligando con las 6 clases pertenecientes al primer digito de clasificación enzimática (*EC*). Dicho modelo se generó utilizando el método de *Machine Learning* llamado *Random Forest*.

En primer lugar, de la obtención de un conjunto de estructuras no redundantes de enzimas se pudo observar una alta redundancia encontrada en las proteínas depositadas en el servidor *PDB*, ya que el 94% de las enzimas fueron eliminadas al realizar la reducción de redundancia junto con los criterios de calidad del cristal establecidos en el punto 1 de metodología. Por este motivo, se puede asegurar que la etapa de reducción de redundancia en los ejemplos es de vital importancia dentro de los procesos de *Machine Learning.* Por otro lado, se encontró un claro desbalance en los datos, llegando a proporciones de 10:1 entre la cantidad de sitios de unión a ligando de las clases mayoritaria y minoritaria. Posteriormente, se corroboró que este desbalance afectaba directamente el desempeño del modelo generado.

En relación a las propiedades de los sitios de unión a ligando calculadas, se pudo observar que por sí solas no logran discriminar entre las distintas clases enzimáticas. Aunque en algunas propiedades se observaron tendencias visibles para una clase en particular, de estas no se puede obtener una idea clara por su elevada desviación estándar.

Siguiendo con el análisis del desempeño obtenido del modelo de clasificación. En primera instancia, se logró obtener una exactitud global del 70%, obtenida mediante las 10 validaciones cruzadas. Además, se observó un claro sesgo de las predicciones hacia la clase de las hidrolasas. Dado que esta última fue la clase con mayor cantidad de ejemplos, se corroboró mediante el método de *oversampling* que el modelo mejora su desempeño al tener cantidades similares de ejemplos para cada una de las clases enzimáticas. Con esta metodología se logró una exactitud global del 90%, además de tasas de falsos positivos no superiores al 4% para todas las clases enzimáticas. Sin embargo, cabe destacar que la clase con menor desempeño

obtenido en relación a todos los indicadores reportados en la Tabla 13 de la sección 3 de resultados fue para la clase de las hidrolasas. Esto último es de esperar al realizar *oversampling* para las clases minoritarias, ya que estas últimas cobran mayor interés para el modelo, que anteriormente clasificaba la gran mayoría de los ejemplos hacia la clase mayoritaria (Japkowicz & Stephen, 2002).

De los resultados de la clasificación de nuevas enzimas con ligandos en sus estructuras, se puede decir que el bajo desempeño obtenido en esta tarea se debe a la inclusión de nueva información, que el modelo no disponía previamente con los ejemplos utilizados en entrenamiento. Sin embargo, el modelo fue capaz de incluir esta nueva información y no se vio mayormente afectado su desempeño. En este nuevo conjunto de datos, es decir los sitios de unión a ligando utilizados en el entrenamiento más los sitios de las nuevas estructuras depositadas en *PDB,* se logró obtener una exactitud global del 89%. Además, todas las tasas de verdaderos positivos exceden el 80% y los falsos positivos de cada una de las 6 clases no superan el 5%, que es muy similar a lo obtenido con el modelo inicial. Por otro lado, de los nuevos ejemplos de sitios de unión a ligando se pudo observar que el desbalance de clases puede aumentar en el tiempo, ya que la proporción entre la clase mayoritaria y la minoritaria en los nuevos ejemplos aumentó a 20:1.

A partir de las predicciones de los sitios de unión que se generaron con Q-siteFinder, se realizó una clasificación de estos sitios con el modelo obtenido inicialmente. En este nuevo conjunto de prueba la tasa de correctamente clasificados fue de apenas un 23%. Además, se pudo observar que la clase que obtiene mayor tasa de verdaderos positivos fueron las transferasas, con un 74%. Sin embargo, también es la clase que presenta la mayor cantidad de falsos positivos. Considerando el bajo rendimiento obtenido en este tipo de ejemplos, no se puede llegar a ninguna conclusión más que se requiere de mejores métodos para la localización del sitio de unión a ligando. En este aspecto queda mucho por realizar, pero sin duda que con estos resultados se dejan las bases para una nueva línea de investigación.

Por último, en cuanto a metodologías de clasificación de nuevas estructuras, se generó una estrategia que permita asignar el primer número de la clasificación EC a una proteína, a través de la moda entre las predicciones de cada uno de sus sitios de unión a ligando. Dicha estrategia puede ser utilizada para cualquier tipo de proteína, con o sin ligandos, para obtener un consenso de función enzimática a través de todas las predicciones de sus sitios de unión. Cabe destacar que esta es solo una propuesta de clasificación, porque puede ocurrir que para muchos investigadores sean más importantes las predicciones de cada sitio en particular ya que la enzima puede estar realizando más de una función.

CONCLUSIONES

A partir de este estudio se pudo generar un modelo de predicción de función enzimática que relaciona directamente las propiedades de los sitios de unión con esta última. Además, se logró combatir el problema del desbalance de clases obteniendo finalmente a una exactitud global del 90%. Con estos resultados demostró de manera efectiva que los sitios de unión a ligando presentan cierta conservación respecto de la función que se está realizando en ellos. Además, el presente estudio es un gran avance en este tipo de predicciones, ya que el único estudio de características similares (Volkamer y col., 2012), que utiliza la información geométrica de los bolsillos formados por la unión de un ligando reporta una exactitud global del 68,2% para predecir el primer número de la clasificación *EC*.

Se logró generar un modelo robusto, ya que soporta la inclusión de nuevos ejemplos sin ver afectado de manera significativa su desempeño. Gracias a esto, el modelo puede ser actualizable a medida que surgen nuevas estructuras de enzimas en la base de datos *PDB*.

Por último, se logró generar una estrategia de clasificación de enzimas utilizando el modelo generado. Este avance puede aportar en gran medida a las metodologías de anotación computacional de clasificación enzimática, ya que corrobora la función que podría desempeñar la enzima en cada uno de sus sitio de unión a ligando.

REFERENCIAS

Almonacid D, Babbitt P. Toward mechanistic classification of enzyme functions. Current opinion in chemical biology. 2011; *15*(3): 435–42.

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; *25*(17): 3389–402.

Apweiler R, Attwood T, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P. InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics (Oxford, England). 2000; *16*(12): 1145–50.

Audit B, Levy E, Gilks W, Goldovsky L, Ouzounis C. CORRIE: enzyme sequence annotation with confidence estimates. BMC bioinformatics, 2007; 8(4): S3.

Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I. The Protein Data Bank. Nucleic acids research. 2000; *28*(1): 235–42.

Bishop CM. Pattern Recognition and Machine Learning. Springer. 2006

Biswas KM, DeVido DR, Dorsey JG. Evaluation of methods for measuring amino acid hydrophobicities and interactions. Journal of Chromatography A. 2003; *1000*(1-2): 637–655.

Bray, T., Doig, A. J., & Warwicker, J. (2009). Sequence and structural features of enzymes and their active sites by EC class. *Journal of molecular biology*, *386*(5), 1423–36. doi:10.1016/j.jmb.2008.11.057

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, Leo. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cai, Y.-D., & Chou, K.-C. (2005). Using functional domain composition to predict enzyme family classes. *Journal of proteome research*, *4*(1), 109–11. doi:10.1021/pr049835p

Capra J, Laskowski R, Thornton J. M, Singh M, Funkhouser T. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology. 2009; *5*(12): 1-18

Chawla, N., & Bowyer, K. (2011). SMOTE: synthetic minority over-sampling technique. *Artificial Intelligence Research*, *16*, 321–357. Retrieved from http://arxiv.org/abs/1106.1813

Chiu, S.-H., Chen, C.-C., Yuan, G.-F., & Lin, T.-H. (2006). Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences. *BMC bioinformatics*, *7*, 304. doi:10.1186/1471-2105-7-304

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *the The European Molecular Biology Organization Journal*, *5*(4), 823–826.

Chou, K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics (Oxford, England)*, *21*(1), 10–9. doi:10.1093/bioinformatics/bth466

Chou, K.-C., & Elrod, D. W. (2003). Prediction of enzyme family classes. *Journal of proteome research*, *2*(2), 183–90.

Cilia, E., & Passerini, A. (2010). Automatic prediction of catalytic residues by modeling residue structural neighborhood. *BMC bioinformatics*, *11*, 115. doi:10.1186/1471-2105-11-115

Dobson, P. D., & Doig, A. J. (2005). Predicting enzyme class from protein structure without alignments. *Journal of molecular biology*, *345*(1), 187–99. doi:10.1016/j.jmb.2004.10.024

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook* (Vol. 112, pp. 571–607). Humana Press. doi:10.1385/1-59259-890-0:571

Griffith, D., Parker, J. P., & Marmion, C. J. (2010). Enzyme inhibition as a key target for the development of novel metal-based anti-cancer therapeutics. *Anti-cancer agents in medicinal chemistry*, *10*(5), 354–70.

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, *320*(2), 369–87. doi:10.1016/S0022-2836(02)00442-4

Guo, X. G. X., Yin, Y. Y. Y., Dong, C. D. C., Yang, G. Y. G., & Zhou, G. Z. G. On the Class Imbalance Problem. , 4 2008 Fourth International Conference on Natural Computation 192–201 (2008). Ieee. doi:10.1109/ICNC.2008.871

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. *The Mathematical Intelligencer* (Vol. 27, p. 745). Springer. doi:10.1177/001112877201800405

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., et al. (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, *433*(7024), 377–81. doi:10.1038/nature03216

Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P., Thornton, J. M., & Mitchell, J. B. O. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics (Oxford, England)*, *21*(23), 4315–6. doi:10.1093/bioinformatics/bti693

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33–8, 27–8.

Izrailev, S., & Farnum, M. a. (2004). Enzyme classification by ligand binding. *Proteins*, *57*(4), 711–24. doi:10.1002/prot.20277

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449.

Jegannathan, K. (2011). Biotechnology in Biofuels-A Cleaner Technology. *Journal of Applied …*, *11*(13), 2421–2425.

Karlson, P., Loening, K., & Webb, E. (1992). Enzyme nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. *Biochemical Education*, *21*(2), 102. doi:10.1016/0307-4412(93)90058-8

Keefe, A., & Szostak, J. (2001). Functional proteins from a random-sequence library. *Letters to Nature*, *410*(6829), 715–8. doi:10.1038/35070613

Kumar, C., & Choudhary, A. (2012). A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP journal on bioinformatics & systems biology*, *2012*(1), 1. doi:10.1186/1687-4153-2012-1

Kumar, S., & Nussinov, R. (2002). Close-range electrostatic interactions in proteins. *Chembiochem : a European journal of chemical biology*, *3*(7), 604–17. doi:10.1002/1439-7633(20020703)3:7<604::AID-CBIC604>3.0.CO;2-X

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132. doi:10.1016/0022-2836(82)90515-0

Larranaga, P. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. doi:10.1093/bib/bbk007

Latino, D. A. R. S., & Aires-de-Sousa, J. (2009). Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests.

Journal of chemical information and modeling, *49*(7), 1839–46. doi:10.1021/ci900104b

Laurie, A. T. R., & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)*, *21*(9), 1908–16. doi:10.1093/bioinformatics/bti315

Leis, S., Schneider, S., & Zacharias, M. (2010). In silico prediction of binding sites on proteins. *Current medicinal chemistry*, *17*(15), 1550–62.

Lu, L., Qian, Z., Cai, Y.-D., & Li, Y. (2007). ECS: an automatic enzyme classifier based on functional domain composition. *Computational biology and chemistry*, *31*(3), 226–32. doi:10.1016/j.compbiolchem.2007.03.008

Mildvan, A. S., Massiah, M. A., Harris, T. K., Marks, G. T., Harrison, D. H. T., Viragh, C., Reddy, P. M., et al. (2002). Short, strong hydrogen bonds on enzymes: NMR and mechanistic studies. *Journal of Molecular Structure*, *615*(1-3), 163–175. doi:10.1016/S0022-2860(02)00212-0

Mitchel, T., & Hill, M. (1997). *Machine Learning* (p. 414).

Munteanu, C. R., González-Díaz, H., & Magalhães, A. L. (2008). Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *Journal of theoretical biology*, *254*(2), 476–82. doi:10.1016/j.jtbi.2008.06.003

Naik, P. K., Mishra, V. S., Gupta, M., & Jaiswal, K. (2007). Prediction of enzymes and non-enzymes from protein sequences based on sequence derived features and PSSM matrix using artificial neural network. *Bioinformation*, *2*(3), 107–12.

Nath, N., & Mitchell, J. B. (2012). Is EC class predictable from reaction mechanism? *BMC Bioinformatics*, *13*(1), 60. doi:10.1186/1471-2105-13-60

Pereira, A. (2012). *Discriminación entre sitios de unión a metales mediante el uso de Máquinas de Vectores de Soporte y la combinación de diversos tipos de información*. Universidad de Talca.

Qi, Y. (2012). Random Forest for Bioinformatics. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning* (pp. 307–323). Boston, MA: Springer US. doi:10.1007/978-1-4419-9326-7

Qiu, J.-D., Huang, J.-H., Shi, S.-P., & Liang, R.-P. (2010). Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. *Protein and Peptide Letters*, *17*(6), 715–722.

Quinlan, J. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81–106.

Rice, P., Longden, I., & Bleasby, A. (2000a). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277.

Rice, P., Longden, I., & Bleasby, A. (2000b). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, *16*(6), 276–7.

Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569–575. doi:10.1109/TPAMI.2009.187

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of molecular biology*, *318*(2), 595–608. doi:10.1016/S0022-2836(02)00016-5

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., et al. (2011). BRENDA, the enzyme information system in 2011. *Nucleic acids research*, *39*(Database issue), D670–6. doi:10.1093/nar/gkq1089

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology*, *5*(12), e1000605. doi:10.1371/journal.pcbi.1000605

Shen, H.-B., & Chou, K.-C. (2007). EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochemical and biophysical research communications*, *364*(1), 53–9. doi:10.1016/j.bbrc.2007.09.098

Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*(3), 405–20.

Tian, W., & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of molecular biology*, *333*(4), 863–82.

Voet, D., & Voet, J. G. (2010). *Biochemistry* (4th ed., p. 1248). Wiley.

Volkamer, A., Griewel, A., Grombacher, T., & Rarey, M. (2010). Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of chemical information and modeling*, *50*(11), 2041–52. doi:10.1021/ci100241y

Volkamer, A., Kuhn, D., Rippmann, F., & Rarey, M. (2012). Predicting enzymatic function from global binding site descriptors. *Proteins*, (October), 1–11. doi:10.1002/prot.24205

Wang, G., & Dunbrack, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, *19*(12), 1589–1591. doi:10.1093/bioinformatics/btg224

Wei, L., Huang, E. S., & Altman, R. B. (1999). Are predicted structures good enough to preserve functional sites? *Structure (London, England : 1993)*, *7*(6), 643–50.

Wilson, C. A., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of molecular biology*, *297*(1), 233–49. doi:10.1006/jmbi.2000.3550

Wimley, W. C., & White, S. H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature structural biology*, *3*(10), 842–8.

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999, August 29). Weka: Practical machine learning tools and techniques with Java implementations. Hamilton, New Zealand.

ANEXOS

Anexo 1.1

```
#Lee de la carpeta de pdbs los ec y busca si tiene Ec y si pose función desconocida. Primero, si no posee función desconocida, pregunta si tiene más de un EC y mueve el archivo a la carpeta correspondiente de ser así. Luego, si no tiene EC mueve el pdb a otra carpeta. Si solo posee un EC, se procede a mover a su respectiva carpeta si es que tiene más de un frame el archivo. Si solo tiene un frame, se mueve a otra carpeta si no tiene ligandos. Si paso por todo lo anterior y no se movió el archivo, se agrega 1 al contador por cada clase y se mueve a una carpeta con el nombre de la clase. Por otro lado, si tiene función desconocida, se pregunta si posee EC para moverlo a la carpeta correspondiente. Finalmente, se escribe el archivo resultante con las cantidades.
set lista [exec ls PDB/PDBs/]
for {set i 1} {$i<7} { incr i} {
    set resultado($i) 0
    set functions($i) 0
}
set out [open "cantidades.csv" "w"]
foreach line $lista {
    set ec [exec awk {/^COMPND/ && $3=="EC:" {print $4" "$5" "$6" "$7}}
PDB/PDBs/$line]
    set function [exec awk {/^HEADER/ && /UNKNOWN FUNCTION/ {print}}
PDB/PDBs/$line]
    set ec [lsort -unique $ec]
    set enz [llength $ec]
    set func [llength $function]
    if {$func == 0} {
        if {$enz > 1} {
            exec mv PDB/PDBs/$line PDB/PDB_more_class/$line
        }
        if {$enz == 0} {
            exec mv PDB/PDBs/$line PDB/PDB_without_EC/$line
        }
        if {$enz == 1} {
            mol load pdb PDB/PDBs/$line
            set num_frames [molinfo top get numframes]
            if {$num_frames > 1} {
                exec mv PDB/PDBs/$line PDB/PDB_more_model/$line

            } else {
                set lig [exec awk {/^HET / {print $2}} PDB/PDBs/$line]
                set num_lig [llength $lig]
                if {$num_lig == 0} {
                    exec mv PDB/PDBs/$line PDB/PDB_without_lig/$line
                } else {
                    set  posicion [string index [lindex $ec 0] 0]
                    set resultado($posicion) [expr $resultado($posicion) +
1]
                    exec mv PDB/PDBs/$line PDB/PDBs_EC/$posicion/$line
                {
            }
            mol delete all
```

```
        }
    } else {
        if {$enz >= 1} {
            exec mv PDB/PDBs/$line PDB/PDB_UNK/EC/$line
            set  posicion [string index [lindex $ec 0] 0]
            set functions($posicion) [expr $functions($posicion) + 1]
        } else {
            exec mv PDB/PDBs/$line PDB/PDB_UNK/WITHOUT_EC/$line
        }
    }
}
puts $out "EC\tCantidad\tSin_EC"
for {set i 1} {$i<7} { incr i} {
    puts $out "$i\t$resultado($i)\t$functions($i)"
}
close $out
quit
```

Anexo 2.1

```
#Procedimiento para el cálculo del centro de masa de una selección de átomos, retorna una lista con las 3 coordenadas
proc center_of_mass {selection} {
    set com [veczero]
    set mass 0
    foreach coord [$selection get {x y z}] m [$selection get mass] {
        set mass [expr $mass + $m]
        set com [vecadd $com [vecscale $m $coord]]
    }
    return [vecscale [expr 1.0/$mass] $com]
}

#Parte leyendo todo el contenido de la carpeta de pdbs. Para cada archivo busca si tiene ligandos y almacena su EC. Posteriormente, carga el archivo en vmd y extrae los números de residuo de los ligandos. Para cada ligando, guarda cuantos aminoácidos tiene a 3 A desde la periferia del ligando (después pregunta que sea mayor que 1). Mueve todo el ligando a su centro de masa. Luego, guarda los aminoácidos no comunes dentro del radio del sitio (tienen que ser = 0). Si se cumplen las 2 condiciones anteriores, obtiene la ocupancia del todos los aminoácidos y tiene que ser = 1.0. Si se cumple la condición anterior guarda el sitio en su respectiva carpeta dependiendo del EC. Finalmente, escribe un archivo con la cantidad de pdbs analizados por clase y la cantidad de ligandos.
set lista [exec ls PDB/PDBs/]
set rad_site 7
for {set i 1} {$i<7} { incr i} {
    set resultado($i) 0
    set enzimas($i) 0
}
set out [open "listas/out.csv" "w"]
foreach line $lista {
    set ligando [exec awk {/^HET / {print $2}} PDB/PDBs/$line]
    set num_lig [llength $ligando]
    set ligando [lsort -unique $ligando]
    set cantidad [llength $ligando]
```

```
    set ec [exec awk {/^COMPND/ && $3=="EC:" {print $4}} PDB/PDBs/$line]
    set ec [string index [lindex $ec 0] 0]
    puts $line
    set resultado($ec) [expr $resultado($ec) + $num_lig]
    set enzimas($ec) [expr $enzimas($ec) + 1]
    set residues {}
    set names_lig {}
    mol load pdb PDB/PDBs/$line
    for {set i 0} {$i < $cantidad} {incr i} {
        if { [lindex $ligando $i] != "SO4" && [lindex $ligando $i] != "GOL"
&& [lindex $ligando $i] != "EDO"} {
            set temp_lig [atomselect top "resname [lindex $ligando $i] and
not protein"]
            set indexs [$temp_lig get residue]
            set indexs [lsort -unique $indexs]
            set count [llength $indexs]
            for {set j 0} {$j < $count} {incr j} {
                lappend names_lig [lindex $ligando $i]
                lappend residues [lindex $indexs $j]
            }
            $temp_lig delete
        }
    }
    set large [llength $residues]
    for {set i 0} {$i<$large} {incr i} {
        set temp [lindex $residues $i]
        set site_temp2 [atomselect top "alpha and (resname ALA ARG ASN ASP
CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL) and same
residue as within 3 of residue $temp"]
        set aa_num2 [$site_temp2 num]
        set site_res [atomselect top "residue $temp"]
        $site_res moveto [center_of_mass $site_res]
        set site_temp3 [atomselect top "protein and not (resname ALA ARG
ASN ASP CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL)
and same residue as within $rad_site of residue $temp"]
        set aa_num3 [$site_temp3 num]
        if {$aa_num2 >= 1 && $aa_num3 == 0} {
            set site [atomselect top "(protein and (resname ALA ARG ASN ASP
CYS GLN GLU GLY HIS ILE LEU LYS MET PHE PRO SER THR TRP TYR VAL) or residue
$temp) and same residue as within $rad_site of residue $temp"]
            set occup [$site get occupancy]
            set occup [lsort -unique $occup]
            if {[llength $occup] == 1} {
                if {[string range [lindex $occup 0] 0 2] == "1.0"} {
                    $site writepdb PDB/sites/$ec/[lindex $names_lig
$i].$temp.$line
                }
            }
            $site delete
        }
        $site_res delete
        $site_temp2 delete
        $site_temp3 delete
    }

    mol delete all
}
```

```
puts $out "EC\tPDB\tLigandos"
for {set i 1} {$i<7} { incr i} {
    puts $out "$i\t$enzimas($i)\t$resultado($i)"
}
close $out
quit
```

Anexo 2.2

```
#Para cada sitio de cada clase, se calculan la cantidad de aminoácidos presentes en el sitio. Con tales datos (la cantidad de aminoácidos presentes en el sitio) se genera una frecuencia por clase y una general y una frecuencia acumulada por clase y general.
for {set i 0} {$i<7} { incr i} {
    set enzimas($i) {0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0}
}
set sites(0) {}
set prom(0) 0
for {set class 1} {$class<7} { incr class} {
    set sites($class) {}
    set prom($class) 0
    set lista [exec ls PDB/sites/$class/]
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set alpha [atomselect top "alpha"]
        set resnum [$alpha num]
        $alpha delete
        set enzimas($class) [lreplace $enzimas($class) $resnum $resnum
[expr [lindex $enzimas($class) $resnum] +1]]
        set enzimas(0) [lreplace $enzimas(0) $resnum $resnum [expr [lindex
$enzimas(0) $resnum] +1]]
        mol delete all
        lappend sites($class) $resnum
        lappend sites(0) $resnum
        set prom($class) [expr $prom($class) + $resnum]
        set prom(0) [expr $prom(0) + $resnum]
    }
}
for {set class 0} {$class<7} { incr class} {
    set out [open "listas/histogram/num_aa_class$class.csv" "w"]
    set out2 [open "listas/histogram/acumula_aa_class$class.csv" "w"]
    puts $out "numero_aa,cantidad_sitios"
    puts $out2 "numero_aa,cantidad_acumulada_sitios"
    set largo [llength $sites($class)]
    set largo [expr $largo * 1.0]
    set prom($class) [expr $prom($class) / $largo]
    set desv 0.0
    for {set i 0} {$i<$largo} { incr i} {
        set desv [expr $desv + pow([expr [lindex $sites($class) $i] -
$prom($class)],2)]
    }
    set desv [expr $desv / $largo]
    set desv [expr sqrt($desv)]
    set acumula 0
```

```
    for {set i 4} {$i<100} { incr i} {
        puts $out "$i,[lindex $enzimas($class) $i]"
        set acumula [expr $acumula + [lindex $enzimas($class) $i]]
        puts $out2 "$i,$acumula"
    }
    puts $out "prom,$prom($class),desv,$desv"
    close $out
    close $out2
}
quit
```

Anexo 2.3

```
#Procedimiento que calcula el ángulo formado entre 3 coordenadas, tomando
como vértice la coordenada central (H)
proc angle {D H A} {
    # cos = ( v1 * v2 ) / |v1| * |v2|
    set PI 3.14159265358979323846
    set hd [vecsub $D $H]
    set ha [vecsub $A $H]
    set cosine [expr [vecdot $hd $ha] / ( [veclength $hd] * [veclength
$ha])]
    #convertir cos en un angulo en grados
    return [expr acos($cosine)*(180.0/$PI)]
}
#Procedimiento para el cálculo del centro de masa de una selección de
átomos, retorna una lista con las 3 coordenadas
proc center_of_mass {selection} {
    set com [veczero]
    set mass 0
    foreach coord [$selection get {x y z}] m [$selection get mass] {
        set mass [expr $mass + $m]
        set com [vecadd $com [vecscale $m $coord]]
    }
    return [vecscale [expr 1.0/$mass] $com]
}
#A partir de la carpeta que contiene los sitios de cada clase carga los
sitios y primero encuentra los 4 primeros aminoácidos mas cercanos,
guardando la coordenada del carbono alfa y la del átomo mas cercano de cada
aminoácido. Luego, calcula todos los ángulos y distancias entre las
coordenadas guardadas y el centro de masa del ligando.
proc calculate_dist_ang {class} {
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/dist_ang_class$class.csv" "w"]
    if {$class == 1} {
        puts -nonewline $out
"id,d_lig_a1,d_lig_c1,an_c1_a1_lig,d_lig_a2,d_lig_c2,an_c2_a2_lig,d_lig_a3,
d_lig_c3,an_c3_a3_lig,d_lig_a4,d_lig_c4,an_c4_a4_lig,"
        puts -nonewline $out
"d_a1_a2,an_a1_lig_a2,d_c1_c2,an_c1_lig_c2,d_a1_a3,an_a1_lig_a3,d_c1_c3,an_
c1_lig_c3,d_a1_a4,an_a1_lig_a4,d_c1_c4,"
        puts -nonewline $out
"an_c1_lig_c4,d_a2_a3,an_a2_lig_a3,d_c2_c3,an_c2_lig_c3,d_a2_a4,an_a2_lig_a
4,d_c2_c4,an_c2_lig_c4,d_a3_a4,an_a3_lig_a4,"
        puts -nonewline $out "d_c3_c4,an_c3_lig_c4,"
```

```
        puts $out ""
    }
    foreach line $lista {
        puts -nonewline $out "$line,"
        mol load pdb PDB/sites/$class/$line
        set formats [split $line "."]
        set ligand [lindex $formats 0]
        set site [atomselect top "protein"]
        set res_site [lsort -unique -integer [$site get residue]]
        $site delete
        set lig [atomselect top "resname $ligand"]
        set cent_lig [center_of_mass $lig]
        $lig delete
        set long [llength $res_site]
        for {set i 0} {$i <$long} {incr i} {
            set dist_tem 1000
            set res_site2 [atomselect top "residue [lindex $res_site $i]
and not alpha"]
            set res_site_alpha [atomselect top "alpha and residue [lindex
$res_site $i]"]
            set coor_alpha [$res_site_alpha get {x y z}]
            set coor [$res_site2 get {x y z}]
            set cant [$res_site2 num]
            $res_site2 delete
            $res_site_alpha delete
            for {set j 0} {$j <$cant} {incr j} {
                set dist [vecdist $cent_lig [lindex $coor $j]]
                if {$dist < $dist_tem} {
                    set dist_tem $dist
                    set coor_res [lindex $coor $j]
                }
            }
            set dist_res($dist_tem) $coor_res
            set dist_res_alpha($dist_tem) [join $coor_alpha " "]
        }
        set dist_res_sort [lsort -real [array names dist_res]]
        for {set i 0} {$i < 4} {incr i} {
            puts -nonewline $out "[vecdist $cent_lig $dist_res([lindex
$dist_res_sort $i])],"
            puts -nonewline $out "[vecdist $cent_lig
$dist_res_alpha([lindex $dist_res_sort $i])],"
            puts -nonewline $out "[angle $dist_res_alpha([lindex
$dist_res_sort $i]) $dist_res([lindex $dist_res_sort $i]) $cent_lig],"

        }
        for {set i 0} {$i < 3} {incr i} {
            for {set j [expr $i + 1]} {$j <4} {incr j} {
                puts -nonewline $out "[vecdist $dist_res([lindex
$dist_res_sort $i]) $dist_res([lindex $dist_res_sort $j])],"
                puts -nonewline $out "[angle $dist_res([lindex
$dist_res_sort $i]) $cent_lig $dist_res([lindex $dist_res_sort $j])],"
                puts -nonewline $out "[vecdist $dist_res_alpha([lindex
$dist_res_sort $i]) $dist_res_alpha([lindex $dist_res_sort $j])],"
                puts -nonewline $out "[angle $dist_res_alpha([lindex
$dist_res_sort $i]) $cent_lig $dist_res_alpha([lindex $dist_res_sort
$j])],"
            }
```

```
        }
        puts $out ""
        unset dist_res
        unset dist_res_alpha
        mol delete all
    }
    close $out
}
for {set i 1} {$i <7} {incr i} {
    calculate_dist_ang $i
}
exec cat listas/dist_ang_class1.csv listas/dist_ang_class2.csv
listas/dist_ang_class3.csv listas/dist_ang_class4.csv
listas/dist_ang_class5.csv listas/dist_ang_class6.csv
>listas/all_dist_ang.csv
exec rm listas/dist_ang_class1.csv listas/dist_ang_class2.csv
listas/dist_ang_class3.csv listas/dist_ang_class4.csv
listas/dist_ang_class5.csv listas/dist_ang_class6.csv
quit
```

Anexo 2.4

```
<TITLE>FOLDX_runscript;
<JOBSTART>#;
<PDBS>#;
<BATCH>lista;
<COMMANDS>FOLDX_commandfile;
<Stability>resul;
<END>#;
<OPTIONS>FOLDX_optionfile;
<Temperature>298;
<R>#;
<pH>7;
<IonStrength>0.050;
<water>-CRYSTAL;
<metal>-CRYSTAL;
<VdWDesign>2;
<OutPDB>false;
<pdb_hydrogens>false;
<END>#;
<JOBEND>#;
<ENDFILE>#;
```

Anexo 2.5

Aminoácido	Hessa y col.	Kyte y col.	Wimley y col.
Ile	-0,6	4,5	0,31
Val	-0,31	4,2	-0,07
Leu	-0,55	3,8	0,56
Phe	-0,32	2,8	1,13
Cys	-0,13	2,5	0,24
Met	-0,1	1,9	0,23
Ala	0,11	1,8	-0,17
Gly	0,74	-0,4	-0,01
Thr	0,52	-0,7	-0,14
Ser	0,84	-0,8	-0,13
Trp	0,3	-0,9	1,85
Tyr	0,68	-1,3	0,94
Pro	2,23	-1,6	-0,45
His	2,06	-3,2	-0,96
Glu	2,68	-3,5	-2,02
Gln	2,36	-3,5	-0,58
Asp	3,49	-3,5	-1,23
Asn	2,05	-3,5	-0,42
Lys	2,71	-3,9	-0,99
Arg	2,58	-4,5	-0,81

Anexo 2.6

```
#A partir de la carpeta que contiene los sitios de cada clase carga los
sitios y suma las hidrofobicidades de cada aminoácido en 3 diferentes
escalas y las escribe en el archivo de salida. Posteriormente, calcula las
proporciones de cada tipo de aminoácidos presentes en el sitio (no polares,
polares neutros, positivos y negativos, mas el total de aminoácidos) y los
guarda en el archivo de salida.
proc calculate {class} {
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/hydro_class$class.csv" "w"]
    if {$class == 1} {
        puts $out
"hydro_1,hydro_2,hydro_3,no_pol,pol_neutros,positivos,negativos,total_aa"
    }
    array set scales1 {ILE 4.5 VAL 4.2 LEU 3.8 PHE 2.8 CYS 2.5 MET 1.9 ALA
1.8 GLY -0.4 THR -0.7 SER -0.8 TRP -0.9 TYR -1.3 PRO -1.6 HIS -3.2 GLU -3.5
GLN -3.5 ASP -3.5 ASN -3.5 LYS -3.9 ARG -4.5}
    array set scales2 {ILE 0.31 VAL -0.07 LEU 0.56 PHE 1.13 CYS 0.24 MET
0.23 ALA -0.17 GLY -0.01 THR -0.12 SER -0.13 TRP 1.85 TYR 0.94 PRO -0.45
HIS -0.96 GLU -2.02 GLN -0.58 ASP -1.23 ASN -0.42 LYS -0.99 ARG -0.81}
    array set scales3 {ILE -0.60 VAL -0.31 LEU -0.55 PHE -0.32 CYS -0.13
MET -0.10 ALA 0.11 GLY 0.74 THR 0.52 SER 0.84 TRP 0.30 TYR 0.68 PRO 2.23
HIS 2.06 GLU 2.68 GLN 2.36 ASP 3.49 ASN 2.05 LYS 2.71 ARG 2.58}
```

```
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set site [atomselect top "alpha"]
        set res_site [$site get resname]
        $site delete
        set largo [llength $res_site]
        set scale1 0.0
        set scale2 0.0
        set scale3 0.0
        for {set i 0} {$i <$largo} {incr i} {
            set scale1 [expr $scale1 + $scales1([lindex $res_site $i])]
            set scale2 [expr $scale2 + $scales2([lindex $res_site $i])]
            set scale3 [expr $scale3 + $scales3([lindex $res_site $i])]

        }
        puts -nonewline $out [format "%.2f,%.2f,%.2f," $scale1 $scale2
$scale3]
        set no_polares [atomselect top "alpha and resname VAL ALA TRP ILE
LEU PHE PRO MET"]
        set n_no_p [$no_polares num]
        $no_polares delete
        set polares_neutros [atomselect top "alpha and resname ASN CYS GLN
SER THR TYR GLY"]
        set n_p_neutros [$polares_neutros num]
        $polares_neutros delete
        set polares_positivos [atomselect top "alpha and resname HIS LYS
ARG"]
        set n_p_positivos [$polares_positivos num]
        $polares_positivos delete
        set polares_negativos [atomselect top "alpha and resname ASP GLU"]
        set n_p_negativos [$polares_negativos num]
        $polares_negativos delete
        set n_total [expr ($n_no_p + $n_p_neutros + $n_p_positivos +
$n_p_negativos) * 1.0]
        puts $out [format "%.2f,%.2f,%.2f,%.2f,%.2f" [expr $n_no_p /
$n_total] [expr $n_p_neutros / $n_total] [expr $n_p_positivos / $n_total]
[expr $n_p_negativos / $n_total] $n_total]
        mol delete all
    }
    close $out
}
for {set i 1} {$i <7} {incr i} {
    calculate $i
}
exec cat listas/hydro_class1.csv listas/hydro_class2.csv
listas/hydro_class3.csv listas/hydro_class4.csv listas/hydro_class5.csv
listas/hydro_class6.csv >listas/all_hydros.csv
exec rm listas/hydro_class1.csv listas/hydro_class2.csv
listas/hydro_class3.csv listas/hydro_class4.csv listas/hydro_class5.csv
listas/hydro_class6.csv
quit
```

Anexo 2.7

```
#Para cada sitio calcula la cantidad de Carbonos, Oxígenos, Nitrógenos y Azufres en cada capa, donde las capas comienzan a 1 Angstrom del ligando y tienen tamaño 1A, se generan 5 capas hasta la más grande que va desde los 6-7 A. También se calcula el total de átomos por capa y un total general de átomos en el sitio.
proc atoms_layers {class} {
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/layers_class$class.csv" "w"]
    if {$class == 1} {
        for {set i 2} {$i<8} {incr i} {
            puts -nonewline $out "C_$i,O_$i,N_$i,S_$i,Total_$i,"
        }
        puts $out "Total_atom"
    }
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set formats [split $line "."]
        set ligand [lindex $formats 0]
        for {set i 1} {$i<7} {incr i} {
            set suma 0.0
            set capa {}
            foreach element {C O N S} {
                set j [expr {$i+1}]
                set atoms [atomselect top "((exwithin $j of resname
$ligand) and not within $i of resname $ligand) and element $element"]
                set count [$atoms num]
                lappend capa $count
                set suma [expr {$suma+$count}]
                $atoms delete
            }
            set suma [expr $suma * 1.0]
            if {$suma != 0.0} {
                foreach n_atom {0 1 2 3} {
                    set porcent [expr {[lindex $capa $n_atom] * 100 /
$suma}]
                    puts -nonewline $out [format "%.2f," $porcent]
                }
                puts -nonewline $out "[expr int($suma)],"
            }
            if {$suma == 0.0} {
                foreach n_atom {0 1 2 3} {
                    puts -nonewline $out "0,"
                }
                puts -nonewline $out "$suma,"
            }
            unset capa
        }
        set all [atomselect top "protein"]
        set all_num [$all num]
        $all delete
        puts $out "$all_num"
        mol delete all
    }
    close $out
```

```
}
for {set i 1} {$i <7} {incr i} {
    atoms_layers $i
}
exec cat listas/layers_class1.csv listas/layers_class2.csv
listas/layers_class3.csv listas/layers_class4.csv listas/layers_class5.csv
listas/layers_class6.csv >listas/all_layers.csv
exec rm listas/layers_class1.csv listas/layers_class2.csv
listas/layers_class3.csv listas/layers_class4.csv listas/layers_class5.csv
listas/layers_class6.csv
quit
```

Anexo 2.8

```
#Procedimiento para el cálculo del centro de masa de una selección de
átomos, retorna una lista con las 3 coordenadas
proc center_of_mass {selection} {
    set com [veczero]
    set mass 0
    foreach coord [$selection get {x y z}] m [$selection get mass] {
        set mass [expr $mass + $m]
        set com [vecadd $com [vecscale $m $coord]]
    }
    return [vecscale [expr 1.0/$mass] $com]
}
#Calcula las distancias entre el centro de masa del sitio completo (todos
los aminoácidos) y el átomo mas cercano de cada aminoácido, luego escribe
en el archivo la distancia menor, la distancia mayor, el promedio y la
desviación estándar de estas distancias.
proc calculate {class} {
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/radius_class$class.csv" "w"]
    if {$class == 1} {
        puts $out "rad_small,rad_big,rad_prom,rad_desv"
    }
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set site [atomselect top "protein"]
        set res_site [lsort -unique -integer [$site get residue]]
        set cent_site [center_of_mass $site]
        $site delete
        set long [llength $res_site]
        set dist_res {}
        set prom 0.0
        for {set i 0} {$i <$long} {incr i} {
            set dist_tem 1000
            set res_site2 [atomselect top "residue [lindex $res_site $i]"]
            set name [lindex [$res_site2 get resname] 0]
            set coor [$res_site2 get {x y z}]
            set cant [$res_site2 num]
            $res_site2 delete
            for {set j 0} {$j <$cant} {incr j} {
                set dist [vecdist $cent_site [lindex $coor $j]]
                if {$dist < $dist_tem} {
                    set dist_tem $dist
```

```
                }
            }
            lappend dist_res $dist_tem
            set prom [expr $prom + $dist_tem]
        }
        set dist_res_sort [lsort -real $dist_res]
        set long2 [llength $dist_res_sort]
        set long2 [expr $long2 * 1.0]
        set prom [expr $prom / $long2]
        set desv 0.0
        for {set i 0} {$i <$long2} {incr i} {
            set desv [expr $desv + pow([expr [lindex $dist_res $i] -
$prom],2)]
        }
        set desv [expr $desv / $long2]
        set desv [expr sqrt($desv)]
        set long2 [expr $long2 - 1]
        puts $out [format "%.3f,%.3f,%.3f,%.3f" [lindex $dist_res_sort 0]
[lindex $dist_res_sort [expr int($long2)]] $prom $desv]
        unset dist_res
        unset dist_res_sort
        mol delete all
    }
    close $out
}
for {set i 1} {$i <7} {incr i} {
    calculate $i
}
exec cat listas/radius_class1.csv listas/radius_class2.csv
listas/radius_class3.csv listas/radius_class4.csv listas/radius_class5.csv
listas/radius_class6.csv >listas/all_radius.csv
exec rm listas/radius_class1.csv listas/radius_class2.csv
listas/radius_class3.csv listas/radius_class4.csv listas/radius_class5.csv
listas/radius_class6.csv
quit
```

Anexo 2.9

```
#Procedimiento para el cálculo del centro de masa de una selección de
átomos, retorna una lista con las 3 coordenadas

proc center_of_mass {selection} {
    set com [veczero]
    set mass 0
    foreach coord [$selection get {x y z}] m [$selection get mass] {
        set mass [expr $mass + $m]
        set com [vecadd $com [vecscale $m $coord]]
    }
    return [vecscale [expr 1.0/$mass] $com]
}
#Por cada sitio carga el pdb del cual fue extraído, mueve el ligando a su
centro de masa para calcular el sasa de los aminoácidos del sitio respecto
de la proteína completa. También se calcula la proporción de estructura
secundaria presente en el sitio (coils, hélices y sabanas).
proc secundary_structure {class} {
```

```
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/struct_class$class.csv" "w"]
    if {$class == 1} {
        puts $out "COIL,HELIX,SHEET,SASA"
    }
    foreach line $lista {
        set formats [split $line "."]
        mol load pdb "PDB/PDBs/[lindex $formats 2].pdb"
        set site_res [atomselect top "residue [lindex $formats 1]"]
        $site_res moveto [center_of_mass $site_res]
        set site [atomselect top "alpha and (same residue as within 7 of
residue [lindex $formats 1])"]
        set backbone [atomselect top "protein"]
        set all_site [atomselect top "protein and same residue as within 7
of residue [lindex $formats 1]"]
        set sasa [measure sasa 1.4 $backbone -restrict $all_site]
        $backbone delete
        $all_site delete
        set struct [$site get structure]
        set conteo(C) 0
        set conteo(T) 0
        set conteo(G) 0
        set conteo(H) 0
        set conteo(B) 0
        set conteo(E) 0
        $site delete
        set long [llength $struct]
        for {set i 0} {$i <$long} {incr i} {
            incr conteo([lindex $struct $i])
        }
        set long [expr $long * 1.0]
        puts $out "[expr ($conteo(C) + $conteo(T))/$long],[expr ($conteo(G)
+ $conteo(H))/$long],[expr ($conteo(B) + $conteo(E))/$long],$sasa"
        unset struct
        unset conteo
        unset formats
        $site_res delete
        mol delete all
    }
    close $out
}
for {set i 1} {$i <7} {incr i} {
    secundary_structure $i
}
exec cat listas/struct_class1.csv listas/struct_class2.csv
listas/struct_class3.csv listas/struct_class4.csv listas/struct_class5.csv
listas/struct_class6.csv >listas/all_struct.csv
exec rm listas/struct_class1.csv listas/struct_class2.csv
listas/struct_class3.csv listas/struct_class4.csv listas/struct_class5.csv
listas/struct_class6.csv
quit
```

Anexo 2.10.1

```
use warnings;
use strict;

#Primero crea el consenso o matriz de puntaje para cada clase, luego para cada uno de estos consensos escanea todas las secuencias. Se generan 6 columnas de puntajes pertenecientes a cada una de las clases.
my @nom_class = ();
$nom_class[1] = "one";
$nom_class[2] = "two";
$nom_class[3] = "tree";
$nom_class[4] = "four";
$nom_class[5] = "five";
$nom_class[6] = "six";
my $i = 0;
my $j = 0;
for ( $i= 1;$i < 7; $i++) {
    `prophecy -sequence listas/class$i.fasta -type G -datafile 'Epprofile' -name mymatrix -threshold 75 -open 3.0 -extension 0.3 -outfile test`;
    for ( $j =1;$j < 7; $j++) {
        `prophet -sequence listas/class$j.fasta -infile test -gapopen 1.0 -gapextend 1.0 -outfile test_fin`;
        if($i != 6) {
            `awk '\$2=="Score:" {print \$3}' test_fin >test$j`;
        } else {
            `awk '\$2=="Score:" {print \$3",$nom_class[$j]"}' test_fin >test$j`;
        }
        `rm test_fin`;
    }
    `cat test1 test2 test3 test4 test5 test6 >listas/pssm$i.csv`;
    `rm test1 test2 test3 test4 test5 test6 test`;
}
`paste -d "," listas/pssm1.csv listas/pssm2.csv listas/pssm3.csv listas/pssm4.csv listas/pssm5.csv listas/pssm6.csv >listas/all_pssm.csv`;
`rm listas/pssm1.csv listas/pssm2.csv listas/pssm3.csv listas/pssm4.csv listas/pssm5.csv listas/pssm6.csv`;
```

Anexo 2.10.2

```
#Procedimiento para el cálculo del centro de masa de una selección de átomos, retorna una lista con las 3 coordenadas
proc center_of_mass {selection} {
    set com [veczero]
    set mass 0
    foreach coord [$selection get {x y z}] m [$selection get mass] {
        set mass [expr $mass + $m]
        set com [vecadd $com [vecscale $m $coord]]
    }
    return [vecscale [expr 1.0/$mass] $com]
}
#Ordena los 12 primeros aminoácidos de acuerdo a la distancia que hay entre el centro de masa del ligando y el átomo mas cercano de cada aminoácido. Escribe un archivo en formato fasta para cada EC con la secuencia de aminoácidos encontrada de cada sitio.
proc calculate {class} {
```

```
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/class$class.fasta" "w"]
    array set translate {ILE I VAL V LEU L PHE F CYS C MET M ALA A GLY G
THR T SER S TRP W TYR Y PRO P HIS H GLU E GLN Q ASP D ASN N LYS K ARG R}
    set num_aminos 12
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set formats [split $line "."]
        set ligand [lindex $formats 0]
        set site [atomselect top "alpha"]
        set res_site [$site get residue]
        $site delete
        set lig [atomselect top "resname $ligand"]
        set cent_lig [center_of_mass $lig]
        $lig delete
        set long [llength $res_site]
        for {set i 0} {$i <$long} {incr i} {
            set dist_tem 1000
            set res_site2 [atomselect top "residue [lindex $res_site $i]"]
            set name [lindex [$res_site2 get resname] 0]
            set coor [$res_site2 get {x y z}]
            set cant [$res_site2 num]
            $res_site2 delete
            for {set j 0} {$j <$cant} {incr j} {
                set dist [vecdist $cent_lig [lindex $coor $j]]
                if {$dist < $dist_tem} {
                    set dist_tem $dist
                }
            }
            set dist_res($dist_tem) $name
        }
        set dist_res_sort [lsort -real [array names dist_res]]
        puts $out ">$line"
        if { $long > $num_aminos} {
            set long2 $num_aminos
        } else {
            set long2 $long
        }
        for {set i 0} {$i <$long2} {incr i} {
            puts -nonewline $out "$translate($dist_res([lindex
$dist_res_sort $i]))"
        }
        puts $out ""
        unset dist_res
        unset dist_res_sort
        mol delete all
    }
    close $out
}
for {set i 1} {$i <7} {incr i} {
    calculate $i
}
quit
```

Anexo 2.10.3

```
#Crea para cada clase un fasta que contiene la secuencia aminoacídica de
cada sitio en orden de aparición en la estructura
```

```
proc calculate_fasta {class} {
    set lista [exec ls PDB/sites/$class/]
    set out [open "listas/class$class.fasta" "w"]
    array set translate {ILE I VAL V LEU L PHE F CYS C MET M ALA A GLY G
THR T SER S TRP W TYR Y PRO P HIS H GLU E GLN Q ASP D ASN N LYS K ARG R}
    foreach line $lista {
        mol load pdb PDB/sites/$class/$line
        set site [atomselect top "alpha"]
        set fasta [$site get resname]
        $site delete
        set large [llength $fasta]
        puts $out ">$line"
        for {set i 0} {$i < $large} {incr i} {
            puts -nonewline $out "$translate([lindex $fasta $i])"
        }
        puts $out ""
        mol delete all
    }
    close $out
}
for {set i 1} {$i < 7} {incr i} {
    calculate_fasta $i
}
quit
```