CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR DE ENSENADA, BAJA CALIFORNIA

PROGRAMA DE POSGRADO EN CIENCIAS EN CIENCIAS DE LA COMPUTACIÓN

Identificación *in silico* de sitios catalíticos en estructuras tridimensionales de proteínas

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Doctor en Ciencias

Presenta:

David Israel Flores Granados

Ensenada, Baja California, México

2015

Tesis defendida por

David Israel Flores Granados

y aprobada por el siguiente comité

Dr. Carlos Alberto Brizuela Rodríguez
Director del Comité

Dr. José Alberto Fernández Zepeda
Miembro del Comité

Dr. Hugo Homero Hidalgo Silva
Miembro del Comité

Dr. Israel Marck Martínez Pérez
Miembro del Comité

Dr. Rogerio Rafael Sotelo Mundo
Miembro del Comité

Dra. Ana Isabel Martínez García
Coordinador del Programa de Posgrado en Ciencias de la Computación

Dr. Jesús Favela Vara
Director de Estudios de Posgrado

Febrero, 2015

Resumen de la tesis que presenta David Israel Flores Granados como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la Computación.

Identificación *in silico* de sitios catalíticos en estructuras tridimensionales de proteínas

Resumen elaborado por:

David Israel Flores Granados

Las proteínas son unas de las macromoléculas más abundantes en los organismos, las funciones que realizan son muy diversas: catálisis, reconocimiento inmune, adherencia celular, transducción de señales, transporte, movimiento y organización celular. La inferencia automática de funciones para proteínas es un reto vigente en la bioinformática estructural. Una línea de investigación para tal propósito es la identificación de residuos catalíticos. Los métodos desarrollados para estas tareas y que están basados en secuencias son la elección natural cuando las proteínas de consulta tienen un alto porcentaje de identidad con enzimas bien anotadas. Sin embargo, cuando la homología no es aparente, lo cual ocurre con un gran número de estructuras de la iniciativa del genoma estructural, entonces la información basada sólo en la secuencia es insuficiente, haciéndose necesario el uso de información proveniente de la estructura tridimensional de las proteínas. En esta investigación se analizó un método para la predicción de residuos catalíticos denominado *CMASA*, el cuál se basa en comparaciones de estructuras locales. El método alcanza valores altos en: exactitud, precisión y coeficiente de correlación de Matthews; sin embargo, ante la presencia de mutaciones puntuales o ausencia de datos relevantes, dichas medidas de desempeño se ven afectadas fuertemente. Para atacar dicho deterioro del desempeño de CMASA, este trabajo propone extender la biblioteca de plantillas de CMASA, de tal forma que se incluyan patrones de menor tamaño con la capacidad de absorber las mutaciones o datos faltantes. Numerosos experimentos computacionales se realizaron con casos como pruebas de concepto, al igual que para un par de casos reales. Los resultados obtenidos indican que la extensión se desempeña adecuadamente cuando un sitio catalítico potencial contiene un residuo mutado o cuando un residuo o sus átomos no están presentes. La extensión propuesta pudo predecir correctamente los residuos de un mutante de timidilato sintetasa, 1EVF. También predice correctamente los residuos catalíticos para la glutaredoxina de plantas (3RHC) a pesar de que carece de información relevante en un átomo de la cadena lateral en el archivo PDB.
Otra parte importante de esta investigación se centró en el análisis para la identificación funcional de un grupo de proteínas pertenecientes a la clase de las cinasas; el estudio evaluó distintas representaciones estructurales con enfoques tanto globales como locales. Los resultados preliminares de esta evaluación muestran el alto grado de complejidad que implica encontrar una representación estructural capaz de discriminar a los elementos de la superfamilia de las cinasas.

Palabras Clave: **Sitios catalíticos, Estructuras terciarias, Mutación, Enzimas, Funciones de proteínas, Cavidades, Cinasas, Triangulación de Delaunay.**

Abstract of the thesis presented by David Israel Flores Granados as a partial requirement to obtain the Master of Science degree in Doctor in Sciences in Computer Science.

***in silico* Identification of catalytic sites in three-dimensional structures of proteins**

Abstract by:

David Israel Flores Granados

Proteins are one of the most abundant macromolecules in organisms, the functions they perform are very diverse: catalysis, immune recognition, adhesion cellular, signal transduction, transport, movement and cell organization. Automatic inference of function protein is a current challenge in structural bioinformatics. A line of research for this purpose is the identification of residues catalytic. The methods developed for this task and based on sequences, are the natural choice when the proteins for queries with a high percentage of identity with enzymes well annotated. However when the homology is not apparent, which occurs with many structures from the structural genome initiative, structural information should be exploited. This research was analyzed a method for prediction of catalytic residues, called CMASA. This method is based on comparisons of local structures, reaching high values: accuracy, precision and Matthews correlation coefficient. However, in the presence of point mutations or absence of relevant data, these performance measures are affected strongly. To overcome performance degradation of CMASA, this work proposes to extend the template library CMASA, so that smaller patterns are included, with the ability to absorb mutations or missing data. Extensive computational experiments are shown as proof of concept instances, as well as for a few real cases. The results show that the extension performs well when the catalytic site contains mutated residues or when some residues are missing. The proposed modification could correctly predict the catalytic residues of a mutant thymidylate synthase, 1EVF. It also successfully predicted the catalytic residues for 3HRC despite the lack of information for a relevant side chain atom in the PDB file.
Another important part of this research focused on the analysis for the identification a functional group of proteins belonging to the class of kinases; the study evaluated different structural representations with both global and local approaches. Preliminary results of this evaluation show the high degree of complexity for finding a structural representation capable to discriminate elements belonging to the kinase superfamily.

Keywords: **Catalytic sites, tertiary structures, punctual mutation, Enzymes, protein functions, Pockets, Protein kinases, Delaunay triangulation.**

Dedicatoria

A la memoria de mi madre, Rosario Granados Vique.

Cuyo amor, valor y perseverancia son la mejor guía de mis pasos.

Agradecimientos

A mi director de tesis, Dr. Carlos Alberto Brizuela Rodríguez por su admirable dedicación profesional. Bajo su dirección he adquirido conocimientos valiosos, pero sobre todo, una actitud hacia el trabajo científico y al trato con calidad humana.

A los miembros del comité de tesis: Dr. Rogerio Sotelo, Dr. José Alberto Fernández, Dr. Israel Martínez y Dr. Hugo Hidalgo; por sus valiosa contribución científica en la conducción de este trabajo de investigación.

Al Centro de Investigación Científica y de Educación Superior de Ensenada por brindarme la oportunidad de tener una formación científica en un ambiente cordial.

A la Universidad del Caribe y sus directivos por el apoyo en todo momento para llevar a cabo este proyecto. Les extiendo también un agradecimiento a mis compañeros: Nancy Aguas, Rocío Giner, Mónica René, José Enrique Álvarez, Fernando Gómez y Francisco López; por su valiosa ayuda durante este periodo.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de doctorado.

A mis compañeros de CICESE por su solidaridad y ánimo para no detenerme.

Finalmente, mi mayor reconocimiento a mi esposa Anitza por su amor, paciencia y apoyo incondicional junto con mis hijos Arantza y Diego Leonel, quienes con su presencia le dan un mayor sentido a este logro; junto con mis hermanos: Hilda, Rafael y Abisaí a quienes agradezco con cariño cada una de sus enseñanzas y vivencias juntos.

Tabla de contenido

Página

Tabla de contenido (continuación)

Lista de figuras

Figura Página

Lista de figuras (continuación)

Lista de figuras (continuación)

Lista de tablas

Capítulo 1. Introducción

1.1. Motivación

Las funciones de las proteínas son esenciales en varios procesos biológicos debido a que actúan como catalizadoras de reacciones químicas, transportadoras de moléculas o reguladoras del crecimiento celular, entre otras (Shulman-Peleg, 2008). El dogma central de la biología estructural enuncia que los plegamientos de las estructuras de las proteínas determinan su funcionalidad biológica (Wright y Dyson, 1999). Las regiones superficiales que se forman en estos plegamientos son importantes por que ahí se efectúan las interacciones moleculares, como las que se presentan en los sitios enzimáticos. Por la importancia de estas regiones, la identificación de sus características claves como sus rasgos geométricos, físico-químicos, electrostáticos, entre otros, se mantiene como una área activa de investigación (Binkowski y Joachimiak, 2008).

El tamaño del universo de plegamientos de proteínas descubiertas se expande rápidamente y, por consecuencia, en las últimas cuatro décadas ha habido un crecimiento acelerado de las bases de datos que contienen su información tridimensional. Este hecho ha impulsado el desarrollo de métodos computacionales que permitan inferir las funciones de nuevas proteínas a partir de de nuevos genomas y transcriptomas. Dado que muchas de ellas se pueden clasificar correctamente dentro de un sistema jerárquico de grupos de proteínas conocidas y caracterizadas que comparten la misma función, el número de nuevas proteínas que requieren un mayor análisis se reduce considerablemente (Cai *et al.*, 2011). Este último grupo presenta retos computacionales para identificar características subyacentes que permitan su clasificación correcta.

1.2. Antecedentes

La identificación automática de elementos pertenecientes a las proteínas importantes para su función, como es el caso de las enzimas, se ha estudiado desde varios contextos. Cuando hay una semejanza considerable entre la secuencia de una proteína cuya función es desconocida y alguna enzima conocida (homología), los métodos de comparación basados en secuencias como BLAST (Altschul *et al.*, 1990), PSI-BLAST (Altschul *et al.*, 1997) o PROSITE (Sigrist *et al.*, 2002), tienen una gran eficiencia para detectar los

elementos semejantes. Sin embargo, los métodos basados en la información tridimensional de las proteínas se deben usar cuando la relación evolutiva entre los elementos a comparar es distante y por lo general sus secuencias son significativamente distintas aunque su función sea la misma (parología).

El proyecto Structural Genomics Initiative (Chandonia y Brenner, 2006) ha contribuido con más de 12000 nuevas estructuras al Protein Data Bank (PDB) y ha acelerado el desarrollo de métodos para la predicción de funciones en proteínas basados en sus conformaciones tridimensionales (Tseng *et al.*, 2009) y en particular para actividades catalíticas de las enzimas (Yahalom *et al.*, 2011; Volkamer *et al.*, 2013; Rahimi *et al.*, 2013; Nilmeier *et al.*, 2013). Estos métodos suponen que las funciones enzimáticas están determinadas por sólo unos cuantos elementos esenciales en la parte superficial del sitio catalítico; entonces, dos proteínas con funciones similares deberían tener patrones locales similares en el correspondiente sitio activo, sin considerar sus secuencias o estructuras globales. Estos métodos también toman un patrón predefinido asociado a una función particular y buscan ese patrón en estructuras locales de la proteína de consulta.

La mayoría de los métodos que usan el enfoque mencionado previamente tienen bases de datos de patrones, los que se suponen son responsables de determinadas funciones. Esta clase de métodos realizan una búsqueda para detectar alguno de estos patrones y transfieren la función de aquel que tenga una mejor coincidencia con la proteína de consulta. Los patrones en la base de datos se pueden hallar por métodos existentes diseñados para este propósito como se reportó en (Milik *et al.*, 2003; Wangikar *et al.*, 2003; Jambon *et al.*, 2003; Ausiello *et al.*, 2008), o mediante la información de sitios activos obtenidos experimentalmente como los del Atlas de Sitios Catalíticos (CSA) (Furnham *et al.*, 2014).

Uno de los primeros métodos en utilizar patrones con caracteríticas de las estructuras 3D de las proteínas fue ASSAM (Artymiuk *et al.*, 1994). En este método se construyó un grafo para la conformación tridimensional de la proteína de consulta y otro para el patrón, posteriormente se usó el algoritmo de isomorfismos de subgrafos de Ullman (Ullmann, 1976) para determinar si el subgrafo del patrón estaba contenido en el grafo de la proteína de consulta. Para generar los grafos, cada bloque estructural, en el patrón y en la proteína,

se representó por dos pseudo-átomos (S y E), los cuales indican el inicio y el fin de una característica del grupo funcional. Estos átomos se convierten en nodos de un grafo tridimensional, de tal modo que cada nodo se etiquetó con el tipo de bloque y de pseudo-centro (S o E), lo que implica que el método considera la posición y orientación. Sin embargo, una versión más reciente (Spriggs *et al.*, 2003) consideró la accesibilidad al solvente, las representaciones burdas de los plegamientos y los enlaces disulfuro, entre otras mejoras. Este método pudo encontrar fácilmente patrones de las triadas catalíticas de algunas serina proteasas y otros patrones comunes en enzimas.

Utilizando el algoritmo TESS, Wallace *et al.* (1997) realizaron una búsqueda de patrones estructurales comunes. Con el método de hashing geométrico detectaron algunos patrones responsables de funciones en todas las enzimas registradas en el PDB; algunos de estos patrones fueron triadas catalíticas y sitios activos en ribonucleasas y lisozimas. TESS se dividió en dos fases: preprocesamiento y búsqueda de plantillas. En la primera fase todas las estructuras de los PDBs se preprocesan para contribuir a la búsqueda de patrones. Donde cada uno de los aminoácidos que conforman la proteína se representan por tres átomos y una llave hash obtenida con base a la distancia entre un par de átomos y cuya distancia no sea mayor a un umbral. En la primera fase se le asigna a cada entrada de la tabla las posiciones relativas de los átomos más cercanos. Posteriormente, en la fase de búsqueda por plantillas, todos los patrones se verifican con aquellos que están en la tabla. Las entradas con el número más grande de coincidencias indican la presencia del patrón en la proteína.

Como una mejora a TESS, se propuso un método conocido como JESS (Barker y Thornton, 2003). Este procedimiento permitió incluir restricciones químicas y geométricas arbitrariamente en la definición de las plantillas obtenidas de la base de datos. Con esto, se obtienen modelos más flexibles que los generados con TESS.

THEMATICS (Ondrechen *et al.*, 2001) es un método desarrollado para encontrar sitios catalíticos en proteínas con estructuras 3D conocidas cuando no hay secuencias o estructuras similares de otras proteínas; por lo tanto, este método no requiere homólogos o información adicional aparte de la estructura 3D de la proteína de consulta, tomando como base las propiedades físico-químicas de cada bloque que forma parte de las

proteínas, así como de sus vecinos.

eF-Site (Kinoshita y Nakamura, 2003) es un método que considera el potencial electrostático, la geometría y la hidropatía sobre la superficie de la proteína como criterio para una medida de similitud entre regiones de las estructuras locales. eF-Site mantiene una base de datos con información concerniente a la superficie molecular de los sitios activos con función conocida. Su algoritmo crea dos grafos, uno para la estructura 3D de consulta y otro para el patrón con el que se esté comparando, donde los vértices de cada grafo son puntos en la superficie, etiquetados con el potencial electrostático y la hidrofobicidad entre otras características. A partir de los dos grafos descritos, se crea un tercero, donde sus nodos son los pares de vértices cuyas etiquetas son similares tanto en el grafo del patrón como en el grafo asociado a la estructura 3D de la proteína de consulta. Los nodos resultantes están conectados mediante una arista si la distancia correspondiente es similar en ambos grafos. En el último paso se busca un cliqué máximo con el algoritmo de Bron y Kerbosch (1973). El cliqué representa el patrón estructural común más grande entre la proteína y la estructura local. El método fue capaz de encontrar mejores similitudes en estructuras locales que en las estructuras globales.

SiteEngine (Shulman-Peleg *et al.*, 2004) también se diseñó para encontrar regiones sobre la superficie de las proteínas similares a los sitios específicos de enlace de una proteína dada. La representación es la que se propuso en Cavbase (Schmitt *et al.*, 2002). Bajo este método, cada residuo se representa por un pseudo-centro genérico, el cual codifica propiedades físico-químicas importantes en las interacciones moleculares.

Query3d (Ausiello *et al.*, 2005) es un método para encontrar patrones de estructuras locales que son comunes entre proteínas. También tiene un sistema manejador de bases de datos, este puede encontrar regiones similares entre dos estructuras de proteínas, una estructura y todas las estructuras en el PDB, y entre subconjuntos de los bloques constitutivos de las proteínas.

Un método rápido y preciso para la predicción de sitios catalíticos es el denominado como CMASA (Li y Huang, 2010). CMASA implementa un algoritmo que compara la estructura 3D de una proteína de consulta contra las plantillas de una base de datos. Las

plantillas contienen información de la estructura tridimensional que conforman los sitios catalíticos. Las plantillas se construyeron a partir de las bases de datos del Atlas de Sitios Catalíticos (Furnham *et al.*, 2014) y el Protein Data Bank (Berman *et al.*, 2000). Los patrones se modelan como dos matrices de distancia entre dos elementos que caracterizan el sitio catalítico.

Otro método con un enfoque similar es el de CatSid (Nilmeier *et al.*, 2013); el punto común en ambos métodos es que representan los patrones mediante matrices de distancia, mientras en CMASA se usan dos matrices, CatSid usa sólo la matriz de distancia relacionada con uno de los elementos. Sin embargo, la diferencia principal es que CMASA utiliza un enfoque basado en similitud y CatSid usa uno basado en aprendizaje de máquina.

1.3. Definición de la problemática a tratar

La predicción automática de funciones de proteínas es un problema en la bioinformática estructural aún no resuelto completamente. Desafortunadamente, la definición ambigua de lo que se considera como función hace que el problema sea aún más complicado, ya que las funciones consideran varios aspectos como los bioquímicos, fisiológicos, celulares y médicos (Rost *et al.*, 2003; Copley, 2012). Debido a la complejidad de los factores implicados en las funciones de las proteínas, es prácticamente imposible desarrollar modelos predictivos para el caso general, ya sea que esté basado en sus secuencias, estructuras 3D o cualquier otra característica descriptiva (descriptores).

Los métodos predictores de funciones en proteínas toman como base las características que parecen más prometedoras para la detección de la función, por ejemplo: propiedades físico-químicas, geométricas, electrostáticas, entre otras. Por lo tanto, la representación de características que extraigan la información esencial de las funciones subyacentes de las proteínas y su posterior clasificación automática es un reto computacional vigente para asignar a las nuevas estructuras en sus grupos funcionales correspondientes. Sin embargo, el desempeño de muchos de ellos es sensible ante mutaciones o ante la pérdida de información relevante.

En este trabajo de investigación se estudian dos problemas fundamentales que se

presentan en el desarrollo de métodos computacionales para la identificación de una región funcional en proteínas (sitios catalíticos), dichos problemas se presentan en forma de preguntas:

1. ¿Cómo se pueden detectar los rasgos pertenecientes a un sitio catalítico potencial, incluso con la presencia de mutaciones o datos faltantes?
2. ¿Qué rasgos estructurales en las proteínas permiten extraer información relevante para la identificación de funciones específicas, especialmente en proteínas con un bajo porcentaje de identidad en sus secuencias, como las pertenecientes a grupos de la superfamilia de las cinasas?. En este mismo sentido, ¿Cómo se debe decidir entre elegir rasgos a nivel de toda la proteína (enfoque global), de sólo algunas regiones (enfoque local) o de una combinación de ambas (enfoque mixto)?

1.4. Objetivos generales

Un problema significativo con el método CMASA (Li y Huang, 2010) es la dificultad para predecir sitios catalíticos cuando uno de sus residuos ha sufrido una mutación o le faltan datos relevantes y las plantillas no son capaces de identificar los residuos restantes del sitio. Superar esta problemática se planteó como un objetivo principal en esta tesis.

Un segundo objetivo fue analizar la efectividad en la identificación de sitios catalíticos en las distintas representaciones de proteínas (estructuras de dimensión 0, primarias, secundarias y terciarias) y sus medidas de similitud asociadas.

1.5. Objetivos específicos

- Se enuncian a continuación los objetivos intermedios para la identificación de sitios catalíticos con alguno de sus residuos mutados,:
    - Establecer un conjunto de prueba cuyos sitios catalíticos estén registrados en el Atlas de Sitios Catalíticos (Furnham *et al.*, 2014).
    - Establecer un conjunto de prueba que no tenga sitios catalíticos, como grupo de control.

- Realizar mutaciones puntuales *in silico* en el primer conjunto de prueba.
- Reproducir y verificar los resultados reportados en el método CMASA y su desempeño con el conjunto de prueba mutado.
- Proponer una modificación al método CMASA para superar los problemas que se presentan con mutaciones puntuales.

- Los objetivos específicos para la identificación de sitios catalíticos en cavidades de un grupo representativo de la superfamilia de las cinasas se conformó de la siguiente forma:
  - Definir un conjunto de prueba de proteínas representativas de varios grupos pertenecientes a las cinasas.
  - Búsqueda de regularidades en representaciones más simples de estructuras de proteínas : dimensión 0, primarias, secundarias para corroborar la necesidad de analizar la estructura terciaria.
  - Propuesta de un método de identificación que extraiga características geométricas y de interacción molecular de cavidades sobre la estructura terciaria.
  - Comprobar que las medidas de similitud hayan permitido la identificación de sitios funcionales.

1.6. Justificación

Diversos estudios muestran que los métodos para inferencia de funciones basados en la identificación de motivos de estructuras locales son los más adecuados para capturar la esencia de las funciones bioquímicas (Tramontano, 2005; Binkowski *et al.*, 2004; Laskowski *et al.*, 2005), especialmente aquellos que usan la comparación de estructuras *ab initio* (estructuras bien conocidas y correctamente anotadas).

En el caso de los sitios activos de enzimas, su conformación está compuesta por sólo algunos residuos conservados, los cuales a menudo no los detectan los métodos de comparación basados en secuencias, sin embargo, los métodos que utilizan rasgos de la estructura tienen mayor potencial de identificarlos aún en proteínas no homólogas (Shulman-Peleg, 2008).

1.7. Contribuciones

Las principales contribuciones de esta investigación al estado del arte se pueden enmarcar en los siguientes aspectos:

- Un esquema simple y de rápida ejecución para la predicción de sitios catalíticos que realiza CMASA ante la presencia de mutaciones y datos faltantes. CMASA es un método que tiene las mejores medidas de desempeño entre los métodos de su tipo, como JESS, FFF y SPASM (Li y Huang, 2010). Sin embargo, ante situaciones como la descrita previamente (mutación o datos faltantes); su efectividad de predicción puede descender significativamente.

- Se incorpora un nuevo descriptor para caracterizar cavidades de proteínas mediante una triangulación en 3D. Los descriptores son tetraedros cuyos vértices indican alguna propiedad físico-química de un punto asociado a un aminoácido en la superficie de la cavidad; mientras que las aristas de la triangulación representan distancias entre estos puntos. Las cavidades son regiones importantes porque ahí se llevan a cabo interaciones moleculares que definen un número considerable de funciones.

- Se aporta evidencia experimental (*in silico*) de la dificultad de separar funcionalmente a la superfamilia de las cinasas, tal como se indica en Thompson *et al.* (2009); Scheeff y Bourne (2005).

1.8. Organización del documento de tesis

El contenido de la tesis está organizado de la siguiente manera:
En el Capítulo 1 se expone la motivación que dió pie a este proyecto de investigación, así como la problemática asociada y los métodos más representativos que se han propuesto para darle solución.

En la primera parte del Capítulo 2 se definen los conceptos biológicos básicos concernientes a las proteínas, las diversas representaciones para describir sus estructuras, y las funciones que realizan, enfatizando la importancia de las proteínas que pertenecen al

grupo de las enzimas. Mientras que la segunda parte de los fundamentos se enfoca a los aspectos computacionales utilizados en los métodos de caracterización y comparación de proteínas con base en sus estructuras.

El Capítulo 3 contiene la propuesta de una extensión a un método para la predicción de sitios catalíticos denominado CMASA (Li y Huang, 2010), este tiene una gran precisión para la predicción de sitios catalíticos, la cual se altera significativamente cuando se presentan mutaciones en algunos de los elementos del sitio catalítico de la proteína de consulta; la extensión propuesta supera estas dificultades usando mejoras en las plantillas de patrones de comparación.

En el Capítulo 4 se describe un caso de estudio para la caracterización y clasificación de un grupo de proteínas enzimáticas pertenecientes a la clase de las cinasas, derivado del problema de inferencia de funciones para proteínas a partir de las estructuras 3D.

En el Capítulo 5 se enuncian las conclusiones y el trabajo a futuro de esta investigación.

En los apéndices A, B y C se detallan los experimentos para el reconocimiento funcional con las representaciones estructurales para las dimensiones 0, 1 y 2, respectivamente.

Capítulo 2. Fundamentos

2.1. Fundamentos biológicos

2.1.1. Composición de las proteínas

Las proteínas son cadenas de heteropolímeros constituidas por aminoácidos, a menudo referidos como *residuos* (Koehl, 2006). Por lo común, sólo hay 20 aminoácidos naturales que se combinan para formar las proteínas. Cada uno de ellos consta de una columna vertebral formada por un átomo central de carbono (C_α) al cual están unidos un átomo de hidrógeno, un grupo amino (NH_2) y un grupo carboxilo ($COOH$); unido también al C_α hay un conjunto de átomos denominado cadena lateral (ver Figura 1(A)). En las moléculas proteícas los residuos se hallan unidos covalentemente por enlaces peptídicos (ver Figura 1(B)), formando cadenas largas no ramificadas, generalmente entre 100 y 500 residuos (Lehninger, 1995).

Un aminoácido se diferencía de otro por los átomos que constituyen sus cadena laterales, la Figura 2 muestra las cadenas laterales para los 20 aminoácidos que pueden ser parte de las moléculas proteícas.

Una característica implícita en cada cadena lateral es el tipo de carga eléctrica que adquiere, las cuales pueden estar cargadas o contener sólo hidrocarburos saturados no polares. Los aminoácidos no-polares no tienen carga eléctrica y por lo general, no son solubles en agua. Por el contrario, los aminoácidos polares tienen concentraciones locales de cargas y pueden ser globalmente neutros, cargados negativamente (ácidos), o cargados positivamente (*bases*) (Branden y Tooze, 1998). A los aminoácidos de tipo base se les denomina comúnmente *receptores* (*AC*), mientras que a los ácidos se les denomina *donadores* (*DO*) (Koehl, 2006). La clasificación por carga eléctrica para cada uno de los veinte aminoácidos se muestra en la Tabla 1.

La distribución de los grupos polares (cargados o no) y no-polares (hidrófobos) de la cadena lateral determina la solubilidad de las proteínas en el agua. La hidrofobicidad de los aminoácidos, lo mismo que la de los péptidos y las proteínas, puede determinarse a partir de las solubilidades respectivas en el agua y un solvente menos polar (como el

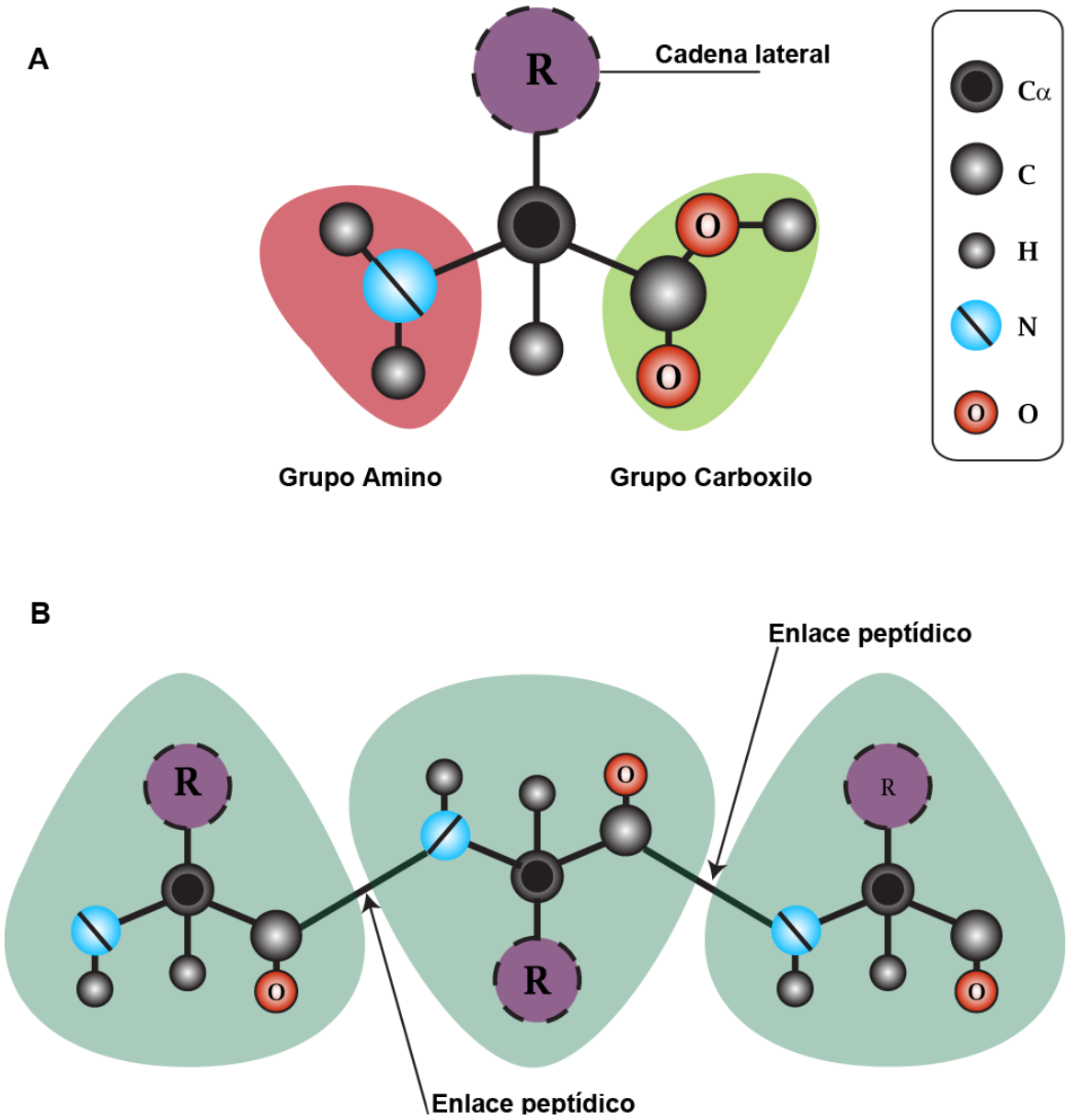


Figura 1: Elementos constitutivos de las proteínas. (A) Composición general de un aminoácido. Cada aminoácido se distingue de los demás por los átomos que conforman sus cadenas laterales. (B) Cadena de aminoácidos unidos por enlaces peptídicos.

Tabla 1: Clasificación por carga eléctrica de los aminoácidos que constituyen las proteínas

Clasificación	Aminoácido
No-polares	Glicina(Gly), Alanina(Ala), Valina (Val), Leucina (Leu), Isoleucina (Ile), Prolina (Pro), Metionina (Met), Phe (Fenilananina), Trp (Triftófano)
Polares	Serina (Ser), Treonina (Thr), Aspargina (Asn), Glutamina (Gln), Cisteína (Cys), Tirosina (Tyr)
Ácidos	Aspartato (Asp), Glutamato (Glu)
Bases	Lisina (Lys), Arginina (Arg), Histidina (His)

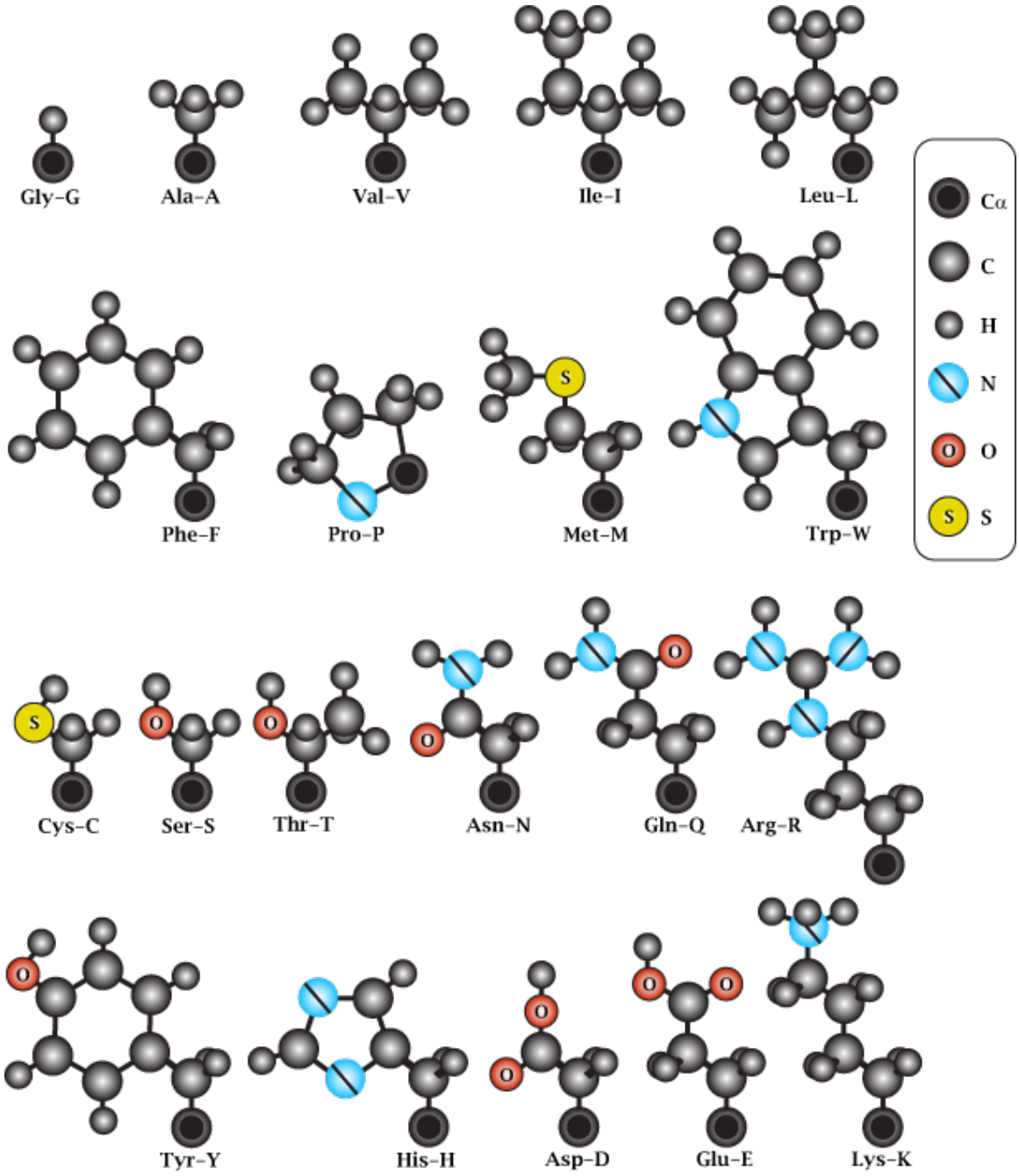


Figura 2: Representaciones de las cadenas laterales para los 20 aminoácidos naturales que componen las proteínas.

etanol) (Cheftel *et al.*, 1989).

2.1.2. Representaciones estructurales

El orden en el cual aparecen enlazados los aminoácidos define lo que se conoce como *secuencia*, *estructura primaria* o *estructura en dimensión 1*. Este tipo de representacio-

nes es el más utilizado para detectar homología entre proteínas, es decir una relación evolutiva que implique un ancestro común (Tramontano, 2005). Los aminoácidos de las secuencias están conectados mediante sus columnas vertebrales de tal forma que constituyen la *cadena principal* de la proteína; esta cadena adopta conformaciones locales canónicas, las cuales se denominan *estructuras secundarias* o de *dimensión 2*. Las conformaciones locales se pueden clasificar en tres tipos: *hélices* α, *láminas* β y *lazos*. Las hélices α tienen una forma helicoidal con 3.6 aminoácidos por vuelta, este tipo de conformaciones a menudo se empacan juntas para formar núcleos hidrofóbicos; por otra parte, las β láminas asemejan una capa casi plana sobre la cadena principal; y las conformaciones tipo lazo generalmente conectan las regiones α con la β (Koehl, 2006).

En su ambiente nativo, las secuencias grandes de aminoácidos se pliegan en una estructura 3D específica, donde las diferentes estructuras son adecuadas para realizar diferentes funciones (Silla y Freitas, 2011). La forma que adoptan estas estructuras es única para cada proteína y se refiere a ellas como estructuras *terciarias*, *nativas* o de *dimension 3* (Crippen, 1978). Para ilustrar las diversas formas de representar una proteína, en la Figura 3(A) se muestra la estructura primaria de la proteína de la insulina humana (2C8R, sólo su cadena A). Se hace énfasis en la conformación de un segmento de su cadena principal, la cual es de tipo hélice α, como se ve en la Figura 3(B) y el lugar que ésta ocupa en la estructura tridimensional (ver Figura 3(C)).

Por su forma y tamaño, las proteínas pueden dividirse en tres grandes grupos: proteínas fibrosas, proteínas de membrana y proteínas globulares (Voet D, 1990).

Las proteínas fibrosas son moléculas elongadas e insolubles en agua, tienen un papel estructural importante en los músculos del cuerpo; a menudo, este tipo de proteínas tienen estructuras repetitivas de forma regular. Por otra parte, las proteínas de membrana están restringidas a los fosfolípidos con membrana de bicapa que rodean a las células y muchos de sus organelos. La amplitud en el tamaño y formas de las proteínas de membrana es muy amplio pero se han clasificado en dos grandes categorías: todas aquellas con formas helicoidales, como la bacteriodopsina, y las que tienen primordialmente estructuras de forma laminar β, como las porinas. Por su parte, las proteínas globulares tienen secuencias no repetitivas. Por el número de residuos que las componen, adoptan

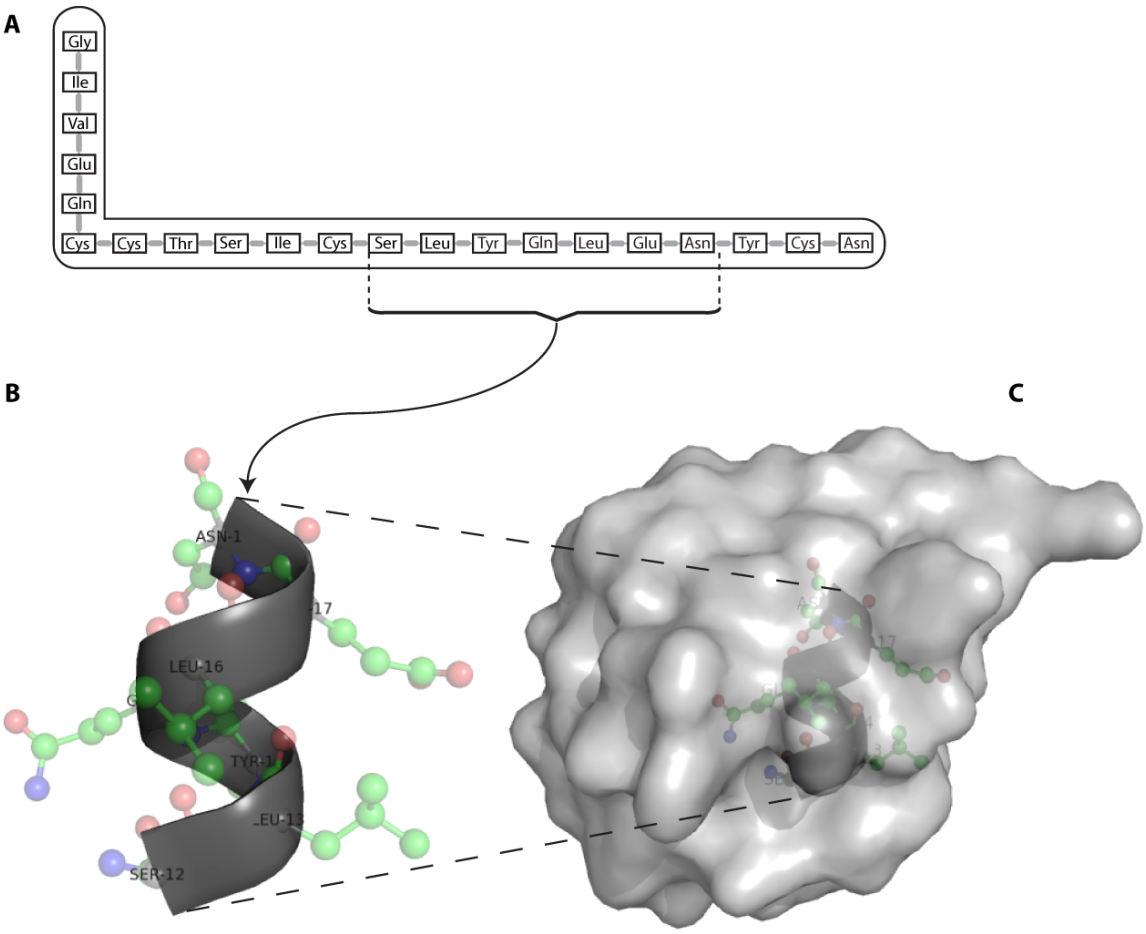


Figura 3: Diversas estructuras para la proteína de la hormona insulina. (A) Secuencia de aminoácidos para la estructura primaria de la cadena A. (B) Conformación tipo hélice alfa de la estructura secundaria para un segmento de los residuos de la cadena A. (C) Superficie que cubre los átomos de la estructura terciaria

una estructura compacta única. En este tipo de proteínas, los residuos cuyas cadenas laterales son no-polares tienden a agruparse para formar el interior del núcleo hidrofóbico, mientras los residuos con cadena lateral hidrofílica se mantienen accesibles al solvente en el exterior del “globo”.

Además de las clasificaciones por rasgos estructurales, se pueden agrupar las proteínas por el papel biológico que desempeñen en los organismos. En la siguiente sección se presenta una síntesis de los tipos de funciones en los que se clasifican a las proteínas.

2.1.3. Funciones biológicas de las proteínas

Las proteínas ejecutan la mayoría de las funciones que realizan las células vivas (Vesely, 2004) y constituyen aproximadamente el 25 % de su peso (Tramontano, 2005). La

Tabla 2 describe algunos ejemplos representativos de proteínas y su clasificación funcional.

Un grupo notable de proteínas son las enzimas, debido a que se especializan en las catálisis de las reacciones biológicas, es decir, su función bioquímica es catalizar la conversión de uno o más sustratos a productos. Muchas de estas reacciones requieren de altas temperaturas o condiciones químicas extremas para ocurrir en ausencia de catalizadores. Aunque actualmente se han identificado la mayor parte de las enzimas relacionadas con el metabolismo básico de las células, quedan por resolver muchos problemas importantes, entre ellos el control genético de la síntesis enzimática (Lehninger, 1995).

La función de una nueva proteína se puede asignar con base en su estructura terciaria. Si la estructura de la nueva proteína tiene plegamientos que se pueden relacionar con los de una proteína cuya función ya es conocida, entonces existen bases para inferir que la función es compartida con la nueva proteína. Sin embargo, la similitud estructural no implica necesariamente homología, y la homología no implica necesariamente equivalencia en las funciones (Pevsner, 2005). En la siguiente sección se detallan algunos aspectos importantes para la clasificación estructural de las proteínas y en especial con las de tipo enzimáticas.

2.1.4. Clasificación jerárquica de las proteínas

Para facilitar la comprensión y el acceso a la información de las estructuras obtenidas por diversos métodos, los investigadores hacen uso de esquemas y bases de datos. Uno de los más importantes es la base de datos SCOP (Hubbard *et al.*, 1997), debido a la precisión de las herramientas automatizadas y al curado de los datos usados en la construcción de su repositorio. Algunas revisiones más extensas de otras bases de datos se encuentran en Holm (1996), Brown *et al.* (1996) y Holm y Sander (1994). La clasificación que realiza SCOP y otras bases de datos como CATH (Sillitoe *et al.*, 2013), DALI, FSSP o Pfam (sólo usa estructuras primarias) (Finn *et al.*, 2014), tienen como base las relaciones evolutivas y los principios que rigen las estructuras terciarias, las cuales se agrupan en los siguientes niveles jerárquicos (Hubbard *et al.*, 1997):

- **Familia**. Proteínas que se agrupan con base en uno de los dos criterios siguientes:

Tabla 2: Clasificación general de proteínas con base en su función y algunos de sus elementos representativos (Lehninger,1995).

Tipo/Ejemplo	Localización o función
Enzimas	
Cinasa A	Conduce a la descomposición de glucógeno en el hígado y los músculos
ADN-polimerasa	Replica y repara ADN
Reserva	
Caseína	Proteína de la leche
Ferritina	Reserva de hierro en el bazo
Transportadoras	
Hemoglobina	Transporta O_2 en la sangre de los vertebrados
Ceruplasmina	Transporta *Cu* en la sangre.
Contráctiles	
Actina	Filamentos móviles en las miofibrillas
Dineína	Cilios y flagelos
Protectoras	
Anticuerpos	Forman complejos con proteínas extrañas
Trombina	Componente del mecanismo de coagulación
Toxinas	
Gosipina	Proteína tóxica de la semilla del algodón
Ricina	Proteína tóxica de la semilla del ricino
Hormonas	
Insulina	Regula el metabolismo de la glucosa
Hormona del crecimiento	Estimula el crecimiento de los huesos
Estructurales	
α-*Queratina*	Piel, plumas, uñas, pezuñas
Colágeno	Tejido conectivo fibroso

1. Aquellas cuyas estructuras primarias tienen 30 % o más de porcentaje de identidad.
2. Aquellas con porcentajes de identidades menores al 30 %, pero cuyas funciones y estructuras sean muy similares; por ejemplo las globinas tienen elementos con porcentajes de identidad en secuencias igual al 15 %.

- **Superfamilia**. Familias cuyas proteínas tienen un bajo porcentaje de identidad pero sus estructuras, y en muchos casos sus funciones, sugieren que tienen un origen evolutivo común; por ejemplo, la actina y la hexocinasa.

- **Plegamiento común**. Superfamilias y familias se consideran que tienen un plegamiento común si sus proteínas tienen la misma estructura secundaria mayor con las

mismas conexiones topológicas.

- **Clases**. Agrupamiento de plegamientos tomando como base las estructuras secundarias dominantes: (1) todas alfas (para proteínas que están formadas esencialmente por hélices α), (2) todas betas (para proteínas que están formadas esencialmente por láminas β), (3) alfa y beta (para proteínas que están formadas por hélices α y láminas β intercaladas a lo largo de toda su estructura), (4) alfa más beta (para proteínas que están formadas por hélices α y láminas β aisladas a lo largo de toda su estructura) y (5) multidominio (para proteínas para las cuales no se conozcan homólogos).

2.1.4.1. Nomenclatura y clasificación de las enzimas

2.1.5. Visualizadores interactivos para estructuras de proteínas

Derivado del rápido crecimiento en el descubrimiento de enzimas, se ha adoptado una clasificación sistemática de acuerdo a las recomendaciones de una comisión internacional creada para tal propósito. Este sistema divide a las enzimas en seis clases principales, cada una de las cuales se divide a su vez en subclases hasta tres niveles de especificidad, de acuerdo con el tipo de reacción catalizada. En la Tabla 3 se muestra la clasificación EC con dos niveles de especificidad para algunos de los grupos más representativos.

El crecimiento acelerado en la adquisición de datos pertenecientes a las estructuras terciarias, se debe en gran medida a los proyectos *Protein Structure Initiative* y *Structural Genomics Initiative* (Chandonia y Brenner, 2006). Para disponer de toda la información generada al respecto, se han construido múltiples bases de datos y esquemas de referencia; una breve revisión para algunos de estos recursos informáticos se presenta en la siguiente sección.

2.1.6. Bases de datos para estructuras de proteínas

- **Protein Data Bank**. Una vez que se determina la secuencia de la estructura primaria de una proteína y que se han recolectado los datos de sus coordenadas atómicas, estructuras químicas de cofactores y descripciones de la estructura cristalizada, se deposita su estructura terciaria en un repositorio principal: El Banco de

Datos de Proteínas (PDB por sus siglas en inglés) (Berman:2000). Posteriormente, se evalúa la calidad de los modelos depositados y se valida que los datos experimentales concuerden. Actualmente el PDB cuenta con más de 100,000 estructuras y se puede acceder a través de la dirección http://www.pdb.org. Otras bases de datos relevantes para estructuras de proteínas se describen a continuación, algunas de ellas operan sobre las entradas que se extraen del PDB. Por la importancia de la información que contiene, también se describe una base de datos que sólo contiene información de secuencias de proteínas (Uniprot).

- **Uniprot**. En los procesos de investigación que involucran el estudio de proteínas, es recurrente la necesidad de adquirir las secuencias de sus estructuras primarias desde las fuentes más actualizadas y confiables. El Universal Protein Resource es el catálogo de secuencias de proteínas más completo creado por Consortium (2007) (al momento de escribir este documento había más de 500 mil secuencias verificadas y un poco más de 86 millones y medio que aún no se verificaban). Este recurso web centralizado es resultado de la colaboración del European Bioinformatics Institute (EBI), el Protein Information Resource (PIR) y el Swiss Institute of Bioinformatics (SIB). Las principales actividades de Uniprot incluyen la curación manual de las secuencias de proteínas auxiliada por análisis computacionales, almacenamiento de secuencias y la disposición de información adicional a través de referencias cruzadas a otras bases de datos. Se puede acceder a esta base de datos mediante la dirección web http://www.uniprot.org.

- **Catalytic Site Atlas (CSA)**. Furnham *et al.* (2014) crearon esta base de datos para proveer sitios catalíticos resueltos y curados manualmente, estos sitios tienen pocos residuos altamente conservados y están directamente relacionados en la actividad catalítica de las enzimas cuyas estructuras se han depositado en el PDB. El núcleo de la base de datos del *CSA* está compuesto por dos fuentes de información: la primera proveniente de literatura primaria y verificada manualmente mediante criterios bien definidos; la segunda fuente se conforma por alineamientos de secuencia aplicando el algoritmo de *PSIBLAST* a homólogos filogenéticamente distantes.

  Otros métodos suelen usar las entradas del CSA para inferir residuos catalíticos

en otras proteínas mediante homologías, como por ejemplo CMASA (Li y Huang, 2010). El acceso a la base de datos en línea del CSA es: http://www.ebi.ac.uk/thornton-srv/databases/CSA/.

- **Computed Atlas of Surface Topography of proteins-CASTp**. CASTp (Dundas *et al.*, 2006) es un servidor web que incorpora información funcional acerca de un gran número de residuos superficiales anotados en las estructuras del PDB, Uniprot y Online Mendelian Inheritance in Man (OMIM). Esta base de datos se especializa en mapear los átomos de las superficies de cavidades expuestas y enterradas de las proteínas. El servidor web de CASTp se aloja en: http://sts.bioengr.uic.edu/castp/index.php.

- **PDBSum**. de Beer *et al.* (2014) desarrollaron una base de datos cuyas entradas contienen análisis pictóricos, determinados experimentalmente, para modelos de estructuras de macromoléculas obtenidos del PDB. Esta base de datos representa las móleculas de cada registro y muestra esquemáticamente las interacciones entre ellas. PDBSum está disponible como servidor web en la dirección: http://www.ebi.ac.uk/pdbsum/ con un poco más de cien mil estructuras.

Una vez que una estructura terciaria se construye, se refina y se pone a disposición para su consulta, el reto que se presenta posteriormente es obtener información útil a partir de ella. En muchos casos la representación de la estructura secundaria es invaluable, pero un modelo con resolución atómica contiene significativamente mayor y mejor información (Bourne y Weissig, 2003). Para extraer esta información detallada desde un modelo se requieren herramientas que permitan la manipulación interactiva para: el cálculo de distancias y ángulos, la consulta de coordenadas de átomos, alternar entre diversas representaciones entre otras operaciones. En la Tabla 4 se listan algunos de los visualizadores más utilizados para dichas tareas.

En el siguiente capítulo se describen los métodos computacionales que se utilizan comúnmente en la inferencia de funciones de proteínas tomando como base las diferentes representaciones de sus estructuras, especialmente las de tipo terciaria.

2.2. Fundamentos computacionales

La biología computacional tiene como meta la comprensión de los sistemas vivientes mediante cálculos a nivel molecular. En particular, su objetivo es predecir cómo interactúan las moléculas para establecer sus múltiples conexiones y regulaciones dentro de las células (Wolfson *et al.*, 2005). En especial, el estudio de las estructuras de las proteínas ha contribuido de forma importante en la comprensión de las relaciones estructura-función. De acuerdo a la Ontología de Genes (Consortium, 2004), la clasificación de las funciones de las proteínas se representa jerárquicamente por tres niveles:

1. La función molecular de la proteína.
2. Los procesos biológicos en los que participan.
3. El componente celular en el cual trabajan.

En esta investigación se aborda únicamente el primer nivel, el cual describe actividades biólogicas como, por ejemplo, las reacciones catalíticas a nivel molecular. Los algoritmos de comparación y clasificación de proteínas que operan en este nivel y cuyo enfoque tiene como base las estructuras terciarias de las proteínas, buscan detectar similitudes que compensen las limitaciones que se presentan cuando se utilizan las estructuras primarias (Shulman-Peleg, 2008), lo que conduce a definir el significado de similitud.

En Bartlett *et al.* (2005) se realiza una revisión de los principales criterios que se han utilizado para determinar la similitud entre dos proteínas, los cuales se listan a continuación:

- Si sus secuencias están bien alineadas (medido por herramientas como PSI-BLAST (Altschul *et al.*, 1990).
- Si sus estructuras tridimensionales tienen una fuerte coincidencia (por ejemplo, en el método DALI (Liisa y Chris, 1995) se identifica las coincidencias mediante alineamientos de las cadenas principales).

- Si ambas contienen cavidades o sitios de enlace comunes (las cavidades se pueden obtener por métodos como CASTp (Dundas *et al.*, 2006) o Surfnet (Laskowski, 1995)).
- Si ambas contienen determinados motivos, los cuales están compuestos por aminoácidos (para detectar motivos pueden usarse métodos como Evolutionary Trace (Yao *et al.*, 2003).
- Si tienen interacciones moleculares comunes para determinadas regiones (por ejemplo, en el método DIP (Xenarios *et al.*, 2002) se utilizan las propiedades físico-químicas en cavidades).

Además de la complejidad que implica establecer un criterio único para determinar la similitud, también debe considerarse la ambigüedad de lo que puede considerarse una función realizada por una proteína. Ambos aspectos hacen necesario establecer un marco contextual que permita identificar cúales son las herramientas computacionales claves para caracterizar y clasificar las proteínas, así como en qué etapas del proceso de inferencia de funciones éstas se utilizan.

2.2.1. Marco contextual

Aunque la terminología usada en la literatura para describir la similitud entre secuencias o estructuras tridimensionales resulte a menudo confusa, se puede considerar que la palabra *patrón* es adecuada para describir las propiedades de una secuencia o estructura, y con las cuales se puede decidir un grado de coincidencia con otras proteínas. También se puede considerar que esta denominación abarca términos como: *motivo*, *plantilla*, *huella*, *fragmento*, *núcleo* y *sitio* (Eidhammer *et al.*, 2000).

Uno de los objetivos que se propuso en este proyecto de investigación, fue detectar patrones en las estructuras terciarias de las proteínas, a través de las comparaciones por pares, de tal forma que tuvieran un significado biológico para describir características comunes entre proteínas biológicamente relacionadas y proponer un método que permita inferir sus funciones. En Eidhammer *et al.* (2000) se propone un modelo con los módulos que componen la mayoría de los predictores de funciones para proteínas. La Figura 4

muestra este modelo en un esquema donde el módulo de clasificación se presenta con mayor detalle. En las siguientes subsecciones se definen los componentes principales de los métodos de clasificación, enfatizando aquellos que se usaron para el desarrollo del método propuesto en esta investigación.

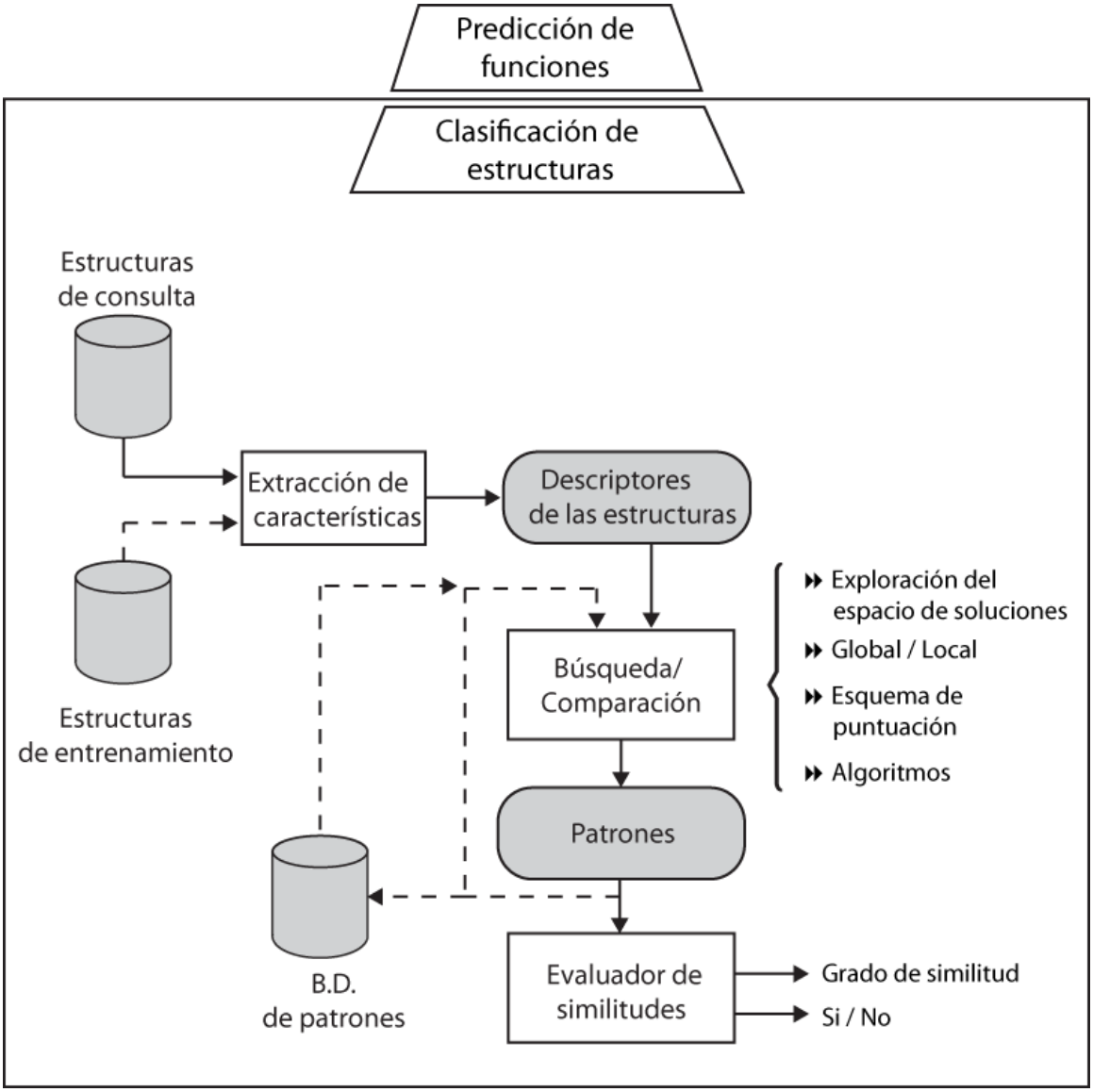


Figura 4: Diagrama de los elementos que conforman los métodos de clasificación de estructuras de proteínas para la predicción de funciones. Los rectángulos sin sombreado indican procedimientos computacionales, mientras que las figuras sombreadas representan las entradas o salidas para cada procedimiento. Las líneas punteadas indican flujo opcional.

2.2.2. Extracción de características

La tarea de extraer características, que sean relevantes para los procesos de comparación, requiere identificar aspectos claves de los elementos seleccionados en la estructura,

tales como: la **geometría**, por ejemplo, las coordenadas de los átomos en la estructuras terciarias obtenidas del Protein Data Bank; su **topología**, es decir su orden a lo largo de la cadena principal; sus **propiedades**, por ejemplo, propiedades físico-químicas.

Una vez establecidos los aspectos claves de las características a extraer, un camino natural consiste en descomponer las proteínas en unidades bien identificables o *descriptores*, de tal forma que estas se puedan analizar individualmente así como su relación con las demás unidades. Eidhammer *et al.* (2000) establecen dos enfoques de selección para unidades de estructuras terciarias:

1. *Basados en elementos*. Este tipo de descriptores son los más utilizados, sus unidades básicas son elementos naturales de las estructuras, por ejemplo: átomos, residuos, porciones de la cadena lateral y elementos de la estructura secundaria.
2. *Basados en espacios*. El espacio en el cual la estructura está localizada se divide en celdas definidas geométricamente usando rejillas o envolturas alrededor de puntos de referencia.

Los descriptores para secuencias de residuos o de estructuras secundarias son triviales debido a que sólo indican para cada residuo el tipo de aminoácido o conformación de sus carbonos α. Mientras que para los descriptores de estructuras terciarias Brown *et al.* (1996) definen las siguientes propiedades ideales: **invariante** a cambios pequeños, tales como la translación o la rotación; **robusto** ante cambios menores en las estructuras, de tal forma que no se manifiesten en cambios drásticos en el descriptor; y **discriminar adecuadamente** cuando las estructuras sean diferentes.

Al conjunto de descriptores de la estructura terciaria se le denomina como la *representación de la estructura terciaria* o simplemente *representación estructural*, y se almacena en una estructura de datos para realizar comparaciones automatizadas (Silla y Freitas, 2011). Diversos autores, como Matsuda *et al.* (1997); Taylor y Orengo (1989); Jonassen *et al.* (1999); Chew *et al.* (1999); Eidhammer *et al.* (2000) clasifican las representaciones de estructuras en alguna de las cinco categorías siguientes:

- *Cadenas*. Se utiliza una secuencia donde el *i-ésimo* residuo se determina por una

letra y la posición relativa de su C_α con respecto a las posiciones de los C_α de los residuos *i-2, i-1, i+1*.

- *Conjunto de descriptores*. En esta representación se extraen una o varias coordenadas por residuo: las coordenadas del átomo C_α, el promedio de las coordenadas de la cadena lateral u otros pseudoátomos. Además de las coordenadas, cada unidad puede tener asociada una propiedad adicional, tipo de aminoácido, propiedades físico-químicas o de interacción molecular, grado de exposición al solvente, estructura secundaria asociada, etc.

- *Grafos*. Un grafo etiquetado se puede usar para representar cada unidad como un vértice y las aristas como sus relaciones. Por ejemplo, en (Grindley *et al.*, 1993) las estructuras secundarias forman los vértices del grafo y las aristas se etiquetan con la distancia entre las unidades y las relaciones angulares entre los nodos.

- *Vectores de características*. Las unidades se almacenan en un arreglo de tamaño fijo, esta representación se ocupa tanto por descriptores basados en elementos como los basados en espacio.

Conforme al esquema de la Figura 4, una vez obtenidos los descriptores y almacenados en las representaciones correspondientes de la proteína, el siguiente paso es realizar búsquedas o comparaciones con otras representaciones o patrones que permitan la detección de nuevos patrones o para evaluar un grado de similitud. La siguiente sección describe de forma breve los métodos de comparación que se utilizaron en esta investigación.

2.2.3. Comparaciones entre representaciones

Shulman-Peleg (2008) agrupa los métodos de comparación para la predicción potencial de funciones en: *comparación de secuencias*, *comparación de estructuras completas* y *comparación de patrones de enlace*. Eidhammer *et al.* (2000) describen estas comparaciones de forma general y con mayor formalidad: ”para dos objetos *A* y *B* teniendo los elementos $a_1, a_2, ..., a_m$ y $b_1, b_2, ..., b_n$, respectivamente, se define una *equivalencia* como un conjunto de pares $L(A, B) = (a_{i_1}, b_{j_1}), (a_{i_2}, b_{j_2}), (a_{i_3}, b_{j_3})$. La equivalencia se llama

alineamiento si los elementos de *A* y *B* están ordenados y si los pares en *L(A,B)* son colineales, es decir, si $i_1 < i_2 < ... < i_l$ y $j_1 < j_2 < ... < j_l$".

Si se puede identificar un ancestro común en una relación evolutiva entre dos proteínas de diferentes especies, entonces se dice que sus secuencias son homólogas. Las proteínas con estas características casi siempre comparten una estructura tridimensional significativa (Pevsner, 2005). Si además, estas proteínas cuentan con una alta tasa de conservación de aminoácidos (es decir, aminoácidos que no se podrían reemplazar en el proceso evolutivo sin afectar negativamente la función de la proteína), se puede suponer que estas proteínas comparten los mismos mecanismos funcionales (Tramontano, 2005).

Dado que la homología es una inferencia de tipo cualitativa (existe o no), también se han desarrollado métodos que sirven para determinar el grado de relación entre las secuencias. El propósito de alinear dos secuencias es alcanzar el máximo nivel de identidad (residuos cuyos aminoácidos coinciden) (Pevsner, 2005). Por tal motivo, los algoritmos de alineamiento de secuencias deben ser capaces de identificar mutaciones que han ocurrido en su trayectoria evolutiva y que han causado divergencia en las proteínas a comparar. Existen tres tipos de mutaciones en los aminoácidos de una proteína: *sustituciones*, *inserciones* y *eliminaciones* (Baxevanis y Oullette, 2001). Las inserciones y eliminaciones ocurren cuando los residuos se añaden o remueven, a ambas operaciones se les denomina genéricamente como *huecos*, y su uso se penaliza en los esquemas de puntuación de los algoritmos de alineamiento de secuencias. En un esquema típico se contemplan dos tipos de penalizaciones: una por la *apertura de un hueco* o *gap* y otra por cada residuo adicional que extiende el hueco (ver Figura 5).

Por otra parte, los esquemas de puntuación califican las mutaciones por sustitución con *matrices de sustitución* para todos los posibles intercambios de un aminoácido por otro. Uno de los modelos más utilizado es BLOSUM (Henikoff y Henikoff, 1992). Para su construcción se uso la base de datos BLOCKS, la cual contenía 500 grupos de alineamientos locales múltiples (bloques) para secuencias de proteínas cuya relación evolutiva es distante. En especial, la matriz BLOSUM50 se formó con la reunión de todas las proteínas de la base de datos que en su alineamiento tuvieron 50 % o más de identidad en los aminoácidos de una de las secuencias. La Figura 6 contiene los valores de la puntuación

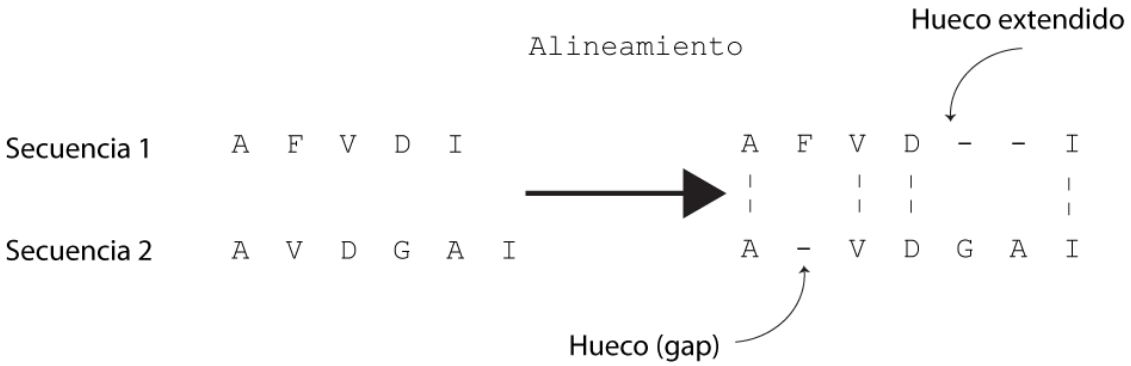


Figura 5: Representación de un alineamiento de dos secuencias hipotéticas. Los huecos introducidos en el alineamiento se marcan con guiones (-) y se penalizan en los esquemas de puntuación.

para esta matriz.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-2	7																		
N	-1	-1	7																	
D	-2	-2	2	8																
C	-1	-4	-2	-4	13															
Q	-1	1	0	0	-3	7														
E	-1	0	0	2	-3	2	6													
G	0	-3	0	-1	-3	-2	-3	8												
H	-2	0	1	-1	-3	1	0	-2	10											
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5										
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6								
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5			
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-4	15		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

Figura 6: Matriz de sustitución de aminoácidos BLOSUM50. Las matrices de sustitución son parte de los esquemas de puntuación para algoritmos de alineamientos globales o locales como Needleman-Wunsch o Smith-Waterman respectivamente.

A continuación, se da una breve explicación de dos algoritmos para alineamientos de

secuencias, y cuyos enfoques son representativos para la mayoría de los tipos de alineamientos de este tipo: uno considera las secuencias completas de las proteínas a comparar (*Alineamiento global*) y otro sólo en las regiones con mayor identidad (*Alineamiento local*). Un aspecto a resaltar del enfoque local, es que permite detectar regiones con un mayor grado de conservación ante la presión evolutiva, por lo que pueden representar con mayor probabilidad unidades funcionales (Tramontano, 2005). Muchos métodos de búsqueda en bases de datos de secuencias, como BLAST (Altschul *et al.*, 1997) o FASTA (Altschul *et al.*, 1990) usan este tipo de alineamientos debido a su velocidad y accesibilidad.

2.2.3.1. Algoritmo de Needleman-Wunsch para el alineamiento global de secuencias

Needleman y Christian (1970) propusieron un algoritmo que produce un alineamiento óptimo de dos secuencias, aún cuando existan mutaciones de inserción y borrado. Este algoritmo es un ejemplo de programación dinámica donde el alineamiento óptimo se obtiene mediante la reducción del problema a una serie de alineamientos más pequeños.

De forma sintetizada, el algoritmo de Needleman-Wunsch consta de tres pasos:

1. **Configuración de una matriz de comparación**. Una matriz bidimensional de m filas $\times$ n columnas se crea para comparar dos secuencias A y B, la primera de m elementos y la segunda de n. Los identificadores de los residuos en las secuencias A y B se colocan como etiquetas en los ejes x e y, respectivamente. En cada celda de la matriz hay cuatro posibles ocurrencias: *i)* los dos residuos pueden coincidir, *ii)* los dos residuos pueden no coincidir, *iii)* un hueco puede introducirse desde la primera secuencia y *iv)* un hueco puede introducirse desde la segunda secuencia (Pevsner, 2005). En la Figura 7(A) se visualizan estas ocurrencias en la matriz de comparación.

   La posibilidad de insertar huecos al inicio de las secuencias se representa con la inserción de una fila en la parte superior de la matriz y una columna en el extremo izquierdo, los valores de sus celdas van acumulando las penalidades (ver Figura 7(B))

2. **Cálculo de puntuaciones para la matriz de comparación.** El esquema de puntuaciones propuesto en Needleman y Christian (1970) le asigna un valor de **+1** a las celdas de las matrices donde las dos proteínas comparten el mismo aminoácido (posiciones sombreadas en la Figura 7(B)). Sin embargo, a menudo se obtienen mejores resultados si inicialmente se le asignan los valores de acuerdo a una matriz de sustitución (Pevsner, 2005), como puede ser BLOSUM50 para proteínas con relaciones distantes (ver Figura 7(C)).

   La forma en cómo se calcula el valor de cada celda restante es mediante una función de recurrencia, tal como se se muestra en la Figura 7(D) y 7(E).

3. **Identificación del alineamiento óptimo**. Una vez que la matriz de comparación se ha llenado de acuerdo al paso anterior, se utiliza un procedimiento de rastreo (trace-back) para construir el alineamiento óptimo.

2.2.3.2. Algoritmo de Smith-Waterman para el alineamiento local de secuencias

Smith y Waterman (1981) propusieron un algoritmo para detectar alineamientos significativos locales se propuso en . El algoritmo usa una matriz de comparación semejante a la del enfoque global (con una fila y una columna adicional) y busca el camino óptimo donde haya una mayor presencia de diagonales (coincidencia de aminoácidos entre el par de secuencias). La regla para definir cada posición de la matriz difiere de la usada en el algoritmo de Needleman-Wunsch, aunque la puntuación en cada celda se calcula como el máximo entre la diagonal precedente y la puntuación obtenida por la inserción de un hueco, el puntaje no puede ser negativo, por lo que la puntuación *s(i,j)* está determinado como el máximo de cuatro posibles valores:

1. El valor de la celda en la posición *i-1*, *j-1* (el vecino superior izquierdo). A este valor se le suma el puntaje específico para la posición *i,j* (*s(i,j)*), que depende de la existencia de coincidencia o no del tipo de aminoácidos para los residuos que comparten la celda.

2. *s(i,j-1)* (puntuación de la celda izquierda) menos una penalidad por hueco.

3. *s(i-1,j)* (puntuación de la celda superior) menos una penalidad por hueco.

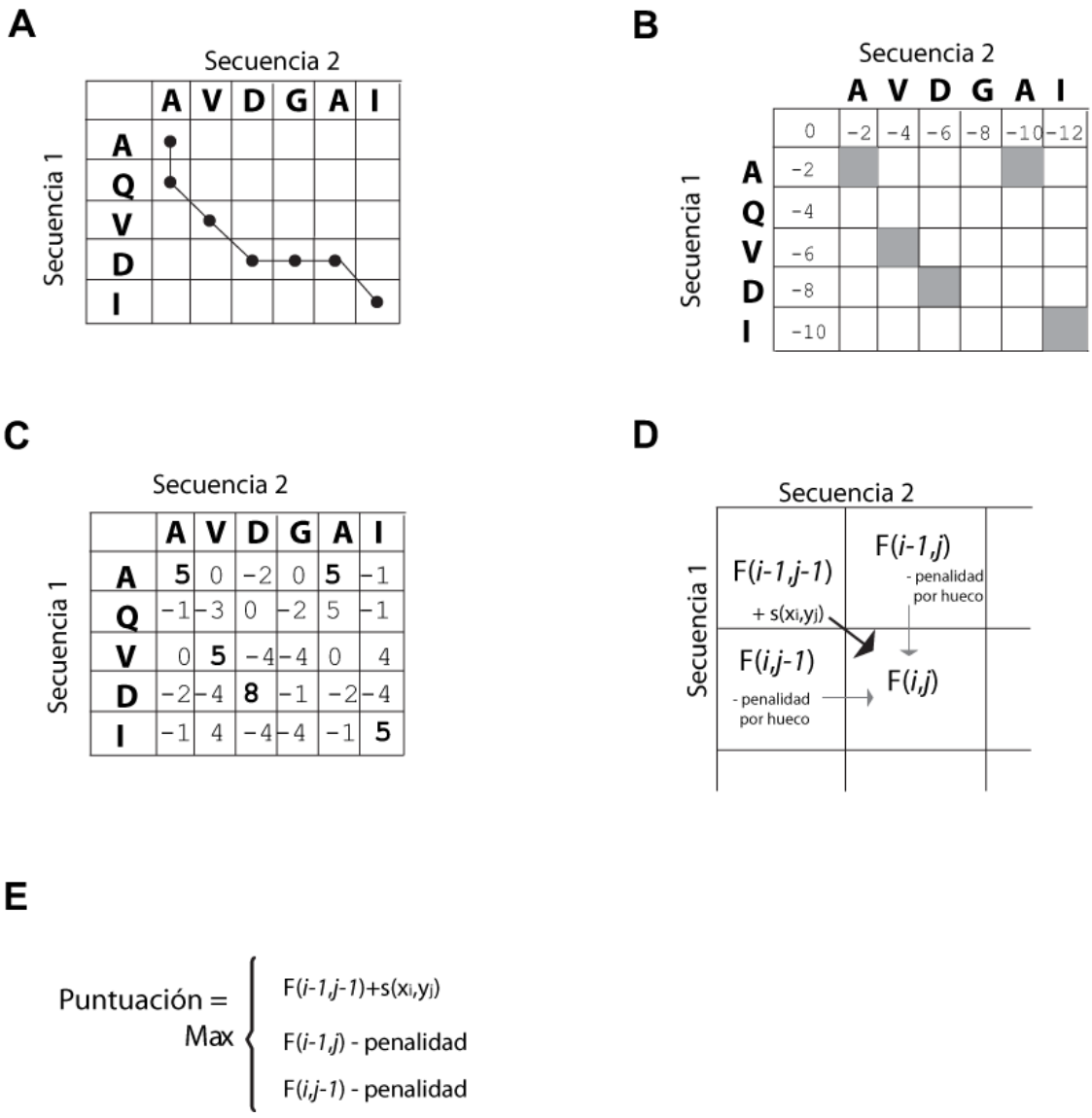


Figura 7: Alineamiento de dos secuencias usando el algoritmo de Needleman-Wunsch. (A). Representación de un alineamiento con los huecos introducidos de la Figura 5, los huecos de la secuencia 1 se representan con trayectorías horizontales y los de la secuencia 2 de forma vertical. (B) Valores de la matriz de sustitución para los aminoácidos de las secuencias a alinear. (C) La configuración inicial de la matriz de comparación, la fila *m+1* y columna *n+1* se les asigna una penalidad al representar huecos al inicio del alineamiento, para este caso la penalidad es de -2. (D) Para cada celda *F(i,j)* de la matriz de comparación se calcula su puntaje a partir del valor máximo entre las celdas superior más su penalidad asociada, celda izquierda más su penalidad asociada y celda superior izquierda más el valor de la coincidencia de la misma celda $s(x_i, y_j)$. (E) Función de recurrencia del algoritmo Needleman-Wunsch y cálculo para la celda $F(x_2, y_2)$ de las secuencias 1 y 2.

4. El número cero.

La predicción automática de funciones de proteínas basadas en secuencias de aminoácidos continúa siendo una área activa de investigación en el campo de la proteómica (Silla

y Freitas, 2011). Sin embargo, la baja conservación de regiones ante la presión evolutiva es una limitante importante que intenta superarse con los métodos de comparación con las estructuras terciarias. En la siguiente subsección se hace énfasis en un enfoque para comparar estructuras terciarias, el cual se usó en la propuesta de un método de clasificación del Capítulo 4 .

2.2.4. Comparación de estructuras tridimensionales

Uno de los enfoques tradicionales para comparar pares de estructuras tridimensionales de proteínas es la *superposición* de sus elementos equivalentes, de tal forma que estén lo más cercano posible. Las distancias que se calculan entre los elementos equivalentes sirven para cuantificar el grado de similitud (Mariani *et al.*, 2013). Si la geometría de las estructuras no cambia en el proceso, entonces se le denomina *superposición rígida*. Los algoritmos que usan superposiciones *basadas en RMSD* (desviación media cuadrática por sus siglas en inglés) se describen de forma general en la Sección 4.1.3.

Un esquema más amplio para clasificar los diversos enfoques de comparación fue propuesto por Eidhammer *et al.* (2000) y se muestra en la Figura 8. Los elementos que se remarcan son aquellos que se usan en el Capítulo 4 y sus fundamentos computacionales se describen a continuación.

2.2.4.1. Hashing geométrico

Para describir esta técnica de comparación, es necesario definir previamente las unidades a comparar y cómo se construyeron dichas unidades.

Dada una nube de puntos representativos en una estructura tridimensional, es posible calcular una triangulación que la caracterice. La triangulación de Delaunay es un algoritmo ampliamente utilizado en muchos campos para tal propósito, debido al número considerable de propiedades óptimas que posee (Aurenhammer, 1991).

Algunos términos indispensables para definir de manera precisa la triangulación de Delaunay se listan a continuación (Goodman y O'Rourke, 2004):

- **Símplex** es el envolvente convexo de *d+1* puntos independientes afines en $\mathbb{R}^d$.

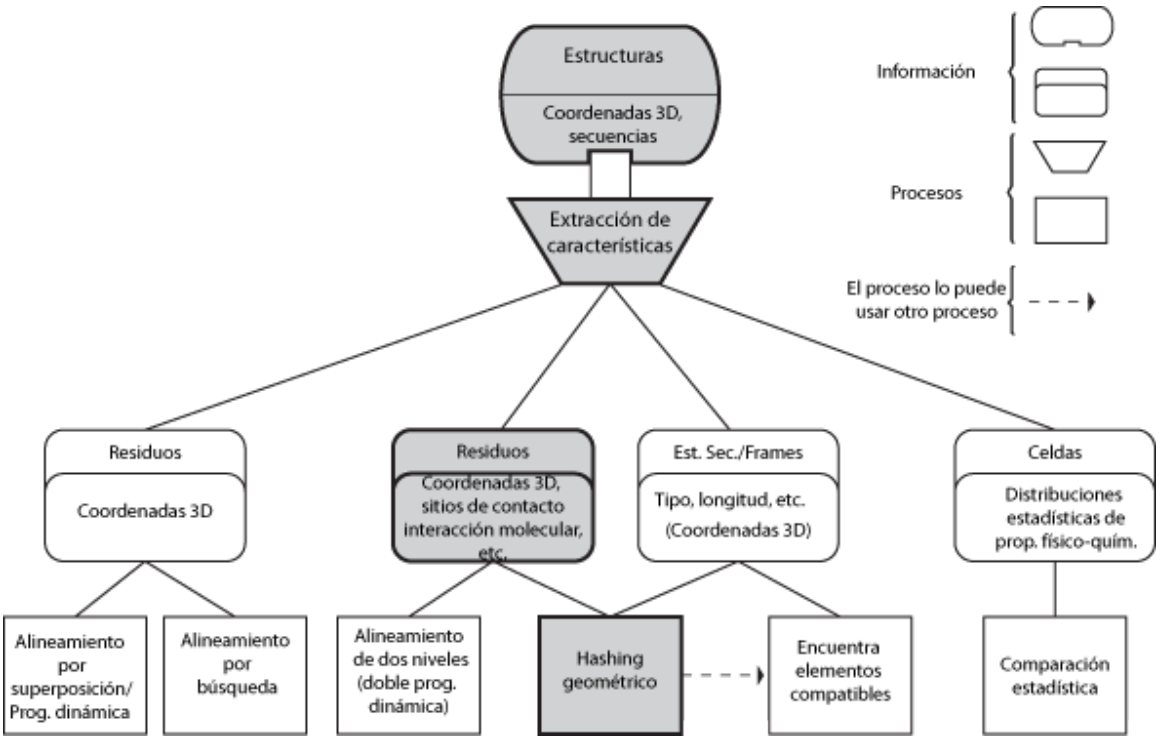


Figura 8: Esquema de los diferentes enfoques para la comparación de estructuras terciarias de proteínas. Los elementos sombreados indican la información y los procesos que se emplearon en el método propuesto de esta investigación.

- **Esfera circunscrita** es la esfera que contiene los vértices de un símplex.
- **Plano** es un sub-espacio afín de dimensión *k*, donde *k ¡ d*.

Una **Triangulación** para un conjunto *P* de *n* puntos en un espacio *d*-dimensional $\mathbb{R}^d$ es una descomposición simplicial del envolvente convexo de *P*. Para un espacio 3D, significa que el envolvente convexo de *P* se descompone en tetraedros tal que: los vértices del tetraedro pertenecen a *P* y la intersección de dos tetraedros da como resultado una arista o una cara o vacío.

La triangulación de Delaunay (TD) es una triangulación tal que, la esfera circunscrita para cada *d*-símplex está vacía, es decir, no contiene alguno de los puntos de *P* en su interior. Algunas de las propiedades más importantes de la triangulación de Delaunay en $\mathbb{R}^d$ son las siguientes (de Berg *et al.*, 2008):

- La TD es única.

- El número de símplices en la TD es a lo más $O(n^{\lceil \frac{d}{2} \rceil})$.
- La TD minimiza el radio máximo de una esfera inscrita al símplex.
- La TD maximiza el ángulo mínimo de los símplices, es decir, forma símplices regulares.
- Minimiza la rugosidad de la superficie.

El conjunto de tetraedros que se forman con la TD a partir de la nube de puntos de la estructura terciaria son unidades que se pueden usar en el hashing geométrico como elementos de comparación.

El hashing geométrico es una técnica originada en el área de visión computacional y que posteriormente se refinó para búsqueda de motivos en estructuras terciarias de proteínas (Nussinov y Wolfson, 1991). En Wallace *et al.* (1997) se describen dos fases principales del algoritmo:

- **Preprocesamiento**. En esta fase se construye un conjunto de tablas hash que almacenan las características con las que se comparan las proteínas de consulta (Wallace *et al*., 1997). La fase de preprocesamiento a su vez realiza secuencialmente las siguientes tareas:
  - *Adquisición del objeto a comparar*. En el contexto de esta investigación, los datos provienen de la base de datos de estructuras terciarias PDB.
  - *Extracción de características*. Puntos, líneas u otras características adecuadas, como los tetraedros o triángulos calculados con la triangulación de Delaunay.
  - *Transformación a invariantes*. Las características se transforman en medidas que describan sus propiedades geométricas (tuplas de *k* elementos) y además se enmarquen en un sistema coordenado estandarizado.
  - *Almacenamiento en la Tabla Hash*. Las medidas obtenidas del paso anterior sirven para situar la unidad descriptiva del objeto en una tabla *k*-dimensional. El método TESS (Wallace *et al*., 1997) almacena los invariantes de los átomos

circundantes (basados en distancias euclidianas) al del átomo que sirve como descriptor.

- **Reconocimiento**. En esta fase se lleva a cabo la comparación entre la tabla hash de una estructura de consulta y el conjunto de tablas hash almacenadas en la fase de preprocesamiento. Por cada celda de la tabla hash de consulta que este ocupada al igual que su correspondiente en una tabla de comparación se incrementa la puntuación (**+1**) de coincidencias, si ambas tienen el mismo contenido. Al finalizar esta fase, la medida de similitud entre la estructura de consulta y cada una de las estructuras contenidas en el grupo de comparación está dada por el número de coincidencias correspondientes.

La comparación de estructuras de proteínas por pares permite obtener medidas de similitud para clasificarlas en grupos de los que se puede inferir su funcionalidad ya sea por homología con otras proteínas previamente anotadas o por regiones altamente conservadas, como la identificación de tríadas catalíticas (Wallace *et al.*, 1997).

Una vez que se han establecido los fundamentos biológicos y computacionales para el desarrollo de de métodos que permitan la inferencia de funciones a partir de la estructura terciaria de las proteínas, se propone un método para la clasificación de proteínas basado en regiones específicas en este tipo de estructuras. El método implementa los procesos de extracción de características (caracterización), comparación por pares y evaluación de las similitudes obtenidas.

Tabla 3: Clasificación de las enzimas en su primer nivel jerárquico de acuerdo al EC. Se muestran también algunos grupos con dos niveles de especificidad. Información retomada de Lehninger (1995).

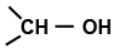
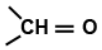
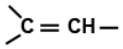
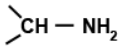
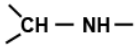
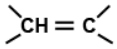
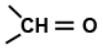
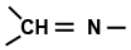
1. **Óxido-reductasas** (Reacciones de óxido-reducción)
    - 1.1 Actúan sobre 
    - 1.2 Actúan sobre 
    - 1.3 Actúan sobre 
    - 1.4 Actúan sobre 
    - 1.5 Actúan sobre 
    - 1.6 Actúan sobre NADH; NADPH
2. **Transferasas** (Transferencia de grupos funcionales)
    - 2.1 Grupos de un átomo de carbono
    - 2.2 Grupos aldehídicos o cetónicos
    - 2.3 Grupos acilos
    - 2.4 Grupos glucosilo
    - 2.7 Grupos fosfato
    - 2.7 Grupos que contienen azufre
3. **Hidrolasas** (Reacciones de hidrólisis)
    - 3.1 Ésteres
    - 3.2 Enlaces glucosídicos
    - 3.4 Enlaces peptídicos
    - 3.5 Otros enlaces **C - N**
    - 3.6 Anhídridos de ácido
4. **Liasas** (Adición a los dobles enlaces)
    - 4.1 
    - 4.2 
    - 4.3 
5. **Isomerasas** (Reacciones de isomerización)
    - 5.1 Racemasas
6. **Ligasas** (Formación de enlaces con escisión del ATP)
    - 6.1 **C - O**
    - 6.2 **C - S**
    - 6.3 **C - N**
    - 6.4 **C - C**

Tabla 4: Ejemplos de visualizadores moleculares comunes.

Herramienta	Comentario	URL
Chime	Complemento para navegador web	Instrucciones en el PDB
Cn3D	Obtiene información de la B.de D. NCBI	http://www.ncbi.nlm.nih.gov /Structure
Mage	Lee kinemages	http://kinemage.biochem.duke.edu
MICE Java applet		Instrucciones en el PDB
RasMol	El visualizador más popular	Instrucciones en el PDB
Swiss PDB viewer	Programa en plataforma	http://spdbv.vital-it.ch/
Visual Molecular Dynamics	Programa de la University of Illinois	http://ks.uiuc.edu/Research/vmd
Pymol	Crea animaciones e imágenes con alta definición	http://www.pymol.org/

Capítulo 3. xCMASA, una extensión simple al método CMASA para la predicción de residuos catalíticos en la presencia de una mutación puntual

En el análisis de estructuras tridimensionales de proteínas, los métodos automáticos para predicción de residuos catalíticos y los de predicción de funciones comparten técnicas de: caracterización, comparación y eventualmente clasificación.
CMASA es un método reciente para predicción de residuos catalíticos. El método logra valores muy altos para la exactitud y el coeficiente de correlación de Matthews. Sin embargo, cuando tiene que tratar con sustituciones puntuales o cuando datos relevantes están ausentes, la eficiencia del método se ve severamente afectada. En este capítulo se describe una extensión simple al método CMASA para superar esta dificultad. En la Sección 3.1 se realiza un análisis del funcionamiento del método CMASA y las partes que lo componen. En la Sección (3.2) se describen los experimentos computacionales que se realizaron sobre instancias artificiales. En la Sección 3.3 se presenta el análisis de dos casos reales. En la Sección 3.4 se presenta la discusión con respecto a los resultados obtenidos.

El aporte que se describe en este capítulo es una extensión a CMASA para hacer más eficiente su desempeño cuando se presentan mutaciones puntuales o hay datos ausentes en la entrada PDB, aunque CMASA se diseñó para manejar situaciones puntuales mediante una matriz de sustitución, se muestra que la implementación no trabaja adecuadamente en muchos casos. La idea propuesta es simple y consiste en extender la biblioteca de plantillas para considerar todas las *n-1* combinaciones de los *n* residuos catalíticos de cada sitio activo. Trabajando de esta forma, la extensión es capaz de recuperar los residuos no mutados de los sitios catalíticos sin degradar demasiado el desempeño en los casos cuando no presentan mutación.

3.1. Método

3.1.1. Extendiendo el método CMASA

El diagrama general del algoritmo de CMASA y la extensión propuesta se muestran en la Figura 9. Para una mejor comprensión de la contribución al método, es necesario revisar algunos aspectos básicos de CMASA.

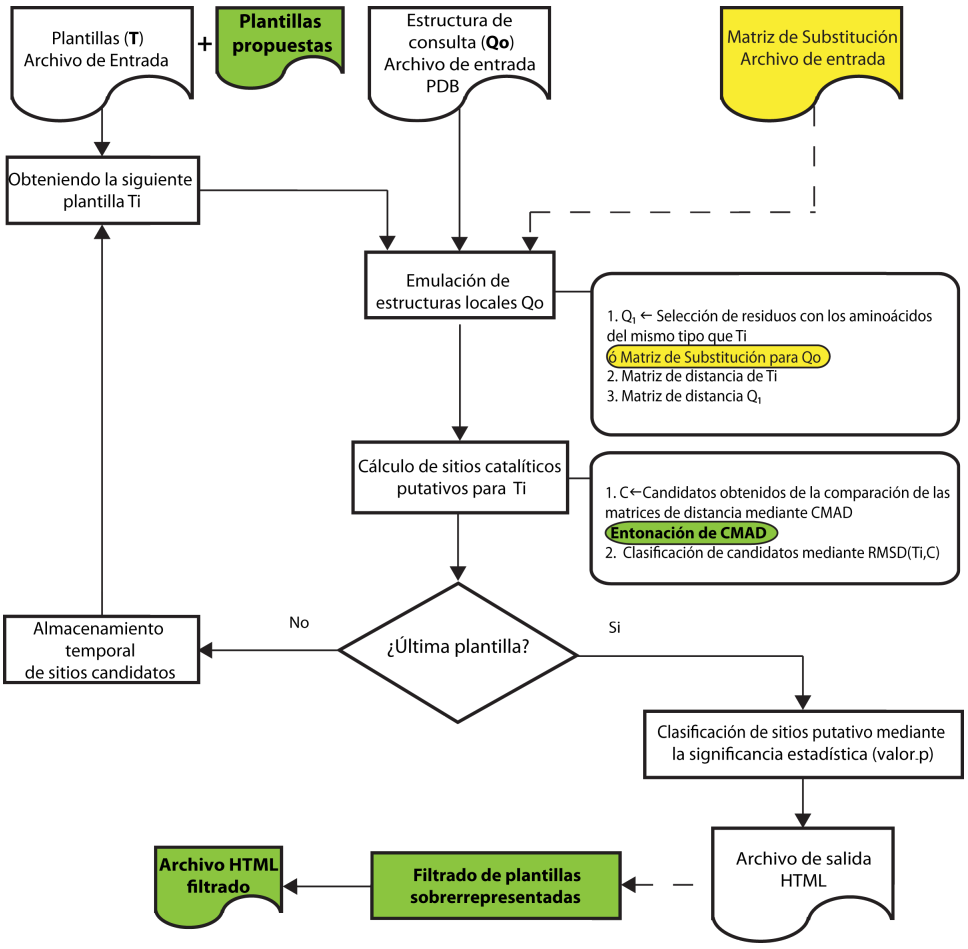
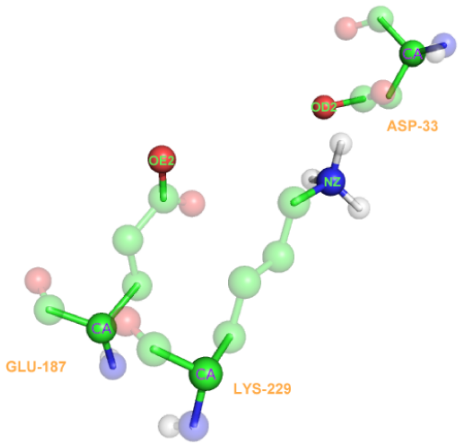


Figura 9: Diagrama de flujo del algoritmo del método CMASA y la extensión propuesta. Los diagramas con fondo verde indican las características agregadas, las figuras sombreadas de amarillo corresponden a los procesos asociados al utilizar SM, las flechas punteadas indican flujo opcional.

La Figura 10 ilustra cómo se representa una plantilla de residuos catalíticos en la base de datos del software CMASA y las matrices de distancia para sus átomos representativos.

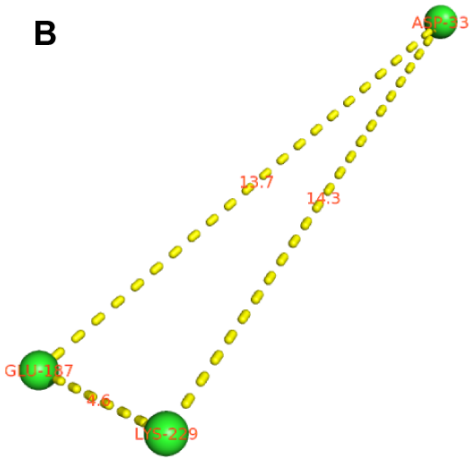
A



Id. AA	Residuo	CαX	CαY	CαZ	Fa. X	Fa. Y	Fa. Z
31	3	0	0	0	0	0	0
3	33	24.72	21.68	31.55	26.20	21.11	28.40
6	187	27.54	19.29	18.37	26.88	23.28	21.32
11	229	23.13	20.28	17.38	22.85	22.42	23.34

Coordenas 3D de los átomos representativos para los residuos catalíticos de la proteína 1ADO en la cadena A. CMASA ocupa estos datos como plantilla

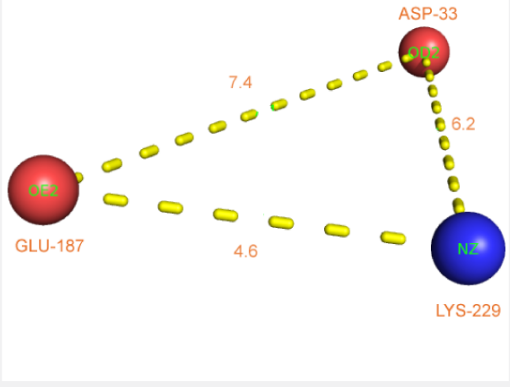
B



Residuo	ASP-33	GLU-187	LYS-229
ASP-33	0	13.7	14.3
GLU-187	13.7	0	4.6
LYS-229	14.3	4.6	0

Matriz de distancia para carbonos α

C



Residuo	ASP-33	GLU-187	LYS-229
ASP-33	0	7.4	6.2
GLU-187	7.4	0	4.6
LYS-229	6.2	4.6	0

Matriz de distancia para átomos más alejados

Figura 10: Ejemplo de una plantilla ocupada por CMASA para búsquedas de estructuras locales (proteína 1ADO). Todas las plantillas están compuestas por las coordenadas de los carbonos alfa (*CA*) y los átomos mas alejados (*fa*) de la cadena lateral correspondiente para cada residuo catalítico. (A) Visualización molecular de la plantilla y las coordenadas de los átomos significativos para sus correspondientes residuos catalíticos. La base de datos del software CMASA separa cada plantilla con un número mayor a 30 en la columna *Id.AA*, la columna *Residuo* de la misma línea contiene el número de residuos en la plantilla. (B) Matriz de distancia para los átomos *CA*. (C) Matriz de distancia para los átomos *fa*.

3.1.1.1. Emulación y comparación estructural en CMASA

Dada una plantilla estructural de residuos t_i y una estructura de consulta Q, CMASA selecciona los residuos en Q que coincidan con los de t_i. Entonces, las estructuras locales que se extraigan de Q estarán formadas exclusivamente por residuos filtrados equivalentes a los de t_i. Para ilustrar esto, el ejemplo de la Figura 11(A) presenta una proteína de consulta con los residuos $E_1, E_2, D_1, K_1, K_2, H_1, H_2$ y la plantilla t_i con los residuos E, D, K. Entonces el filtrado elimina los residuos H_1 y H_2 y mantiene el resto. Cada posible combinación de los residuos restantes se compara con los que están en t_i, para este caso hay un total de cuatro estructuras locales, lq_1 a lq_4, como se ilustra en la Figura 11(B). Si las distancias por pares son equivalentes en ambas, la estructura local en la consulta (lq_j) y en la plantilla t_i, se considera que se ha encontrado una posible coincidencia. Este procedimiento se repite con cada plantilla t_i en la base de datos para determinar cuáles son los residuos catalíticos putativos en Q. Las coincidencias están definidas por aquellas estructuras que son similares geométricamente a una plantilla determinada y la coincidencia tiene una probabilidad pequeña de producirse de forma aleatoria, es decir, tienen un *valor-p* menor al umbral.

Cuando la proteína de consulta tiene una plantilla subyacente t_i pero con una mutación puntual, CMASA propone el uso de una matriz de sustitución (SM por sus siglas en inglés); para futuras referencias en este trabajo el esquema de sustitución se denomina *CMASA-SM*. Entonces, S es una *CMASA-SM* y $S_{i,j}$ es 1 si el tipo del aminoácido i es intercambiable con el del tipo j, donde $i, j \in 0, 1, ..., 19$ representan los 20 aminoácidos.

Una de las desventajas del enfoque *CMASA-SM* es el número de comparaciones que debe realizar. Por ejemplo, si se tiene una estructura de consulta como la de la Figura 11, con los siguientes residuos después del proceso de filtrado: E_1, E_2, D_1, K_1, y K_2, y suponiendo que la SM (S) se construyó para la sustituciones del tipo: $H \rightleftarrows E$, $H \rightleftarrows D$, y $H \rightleftarrows K$. Entonces, en la primera sustitución, H_1 y H_2 se pueden reemplazar por E en la proteína consulta, por lo tanto, se tienen cuatro Es para combinar con un D y dos Ks. Agregando 8 combinaciones de esta forma. Si se continúa el mismo análisis, se mostrará que el reemplazo $H \rightleftarrows D$ añade 8 nuevas combinaciones, y $H \rightleftarrows K$ agrega 4 nuevas combinaciones. En el enfoque de SM se pueden considerar sustituciones simultaneas co-

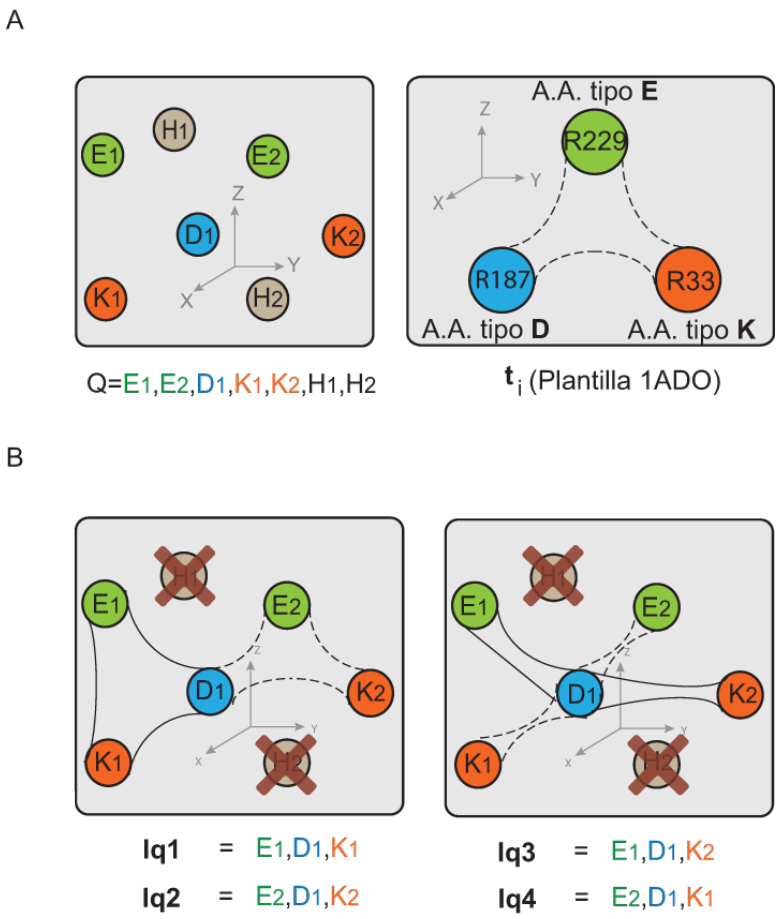


Figura 11: Emulación de estructuras locales en CMASA. La proteína de consulta Q es hipotética; la plantilla t_i tiene asociado el sitio catalítico de la proteína 1ADO para emular la estructura local. (A) Entrada del método: Q es la secuencia de residuos en la estructura de consulta; t_i (E,D,K) es la plantilla usada para emular las estructuras locales en Q. (B) Estructuras emuladas de t_i en Q, hay cuatro posibles combinaciones ($lq1, lq2, lq3, lq4$) que pueden coincidir con la plantilla t_i.

mo $H \rightleftarrows [E, D]$, $H \rightleftarrows [E, K]$ y $H \rightleftarrows [D, K]$, esto suma 2, 1 y 2 nuevas combinaciones, respectivamente. El número total de combinaciones, como se muestra en la Figura 12 (A) es 25. Por otra parte, cuando se usan las plantillas extendidas hay 4 combinaciones originales más 8 nuevas combinaciones agregadas por estas sub-plantillas. En la Figura 12 (B) cada combinación posible de dos residuos a ser buscada se indicada por medio de una arista que conecta los dos nodos.

3.1.1.2. CMASA y residuos mutados o ausentes

CMASA falla al predecir sitios catalíticos si: *i)* existe una diferencia en el número de residuos entre la plantilla y la consulta, es decir, cuando el número de residuos en la plantilla es más grande que en la consulta; *ii)* el átomo más alejado de la cadena lateral

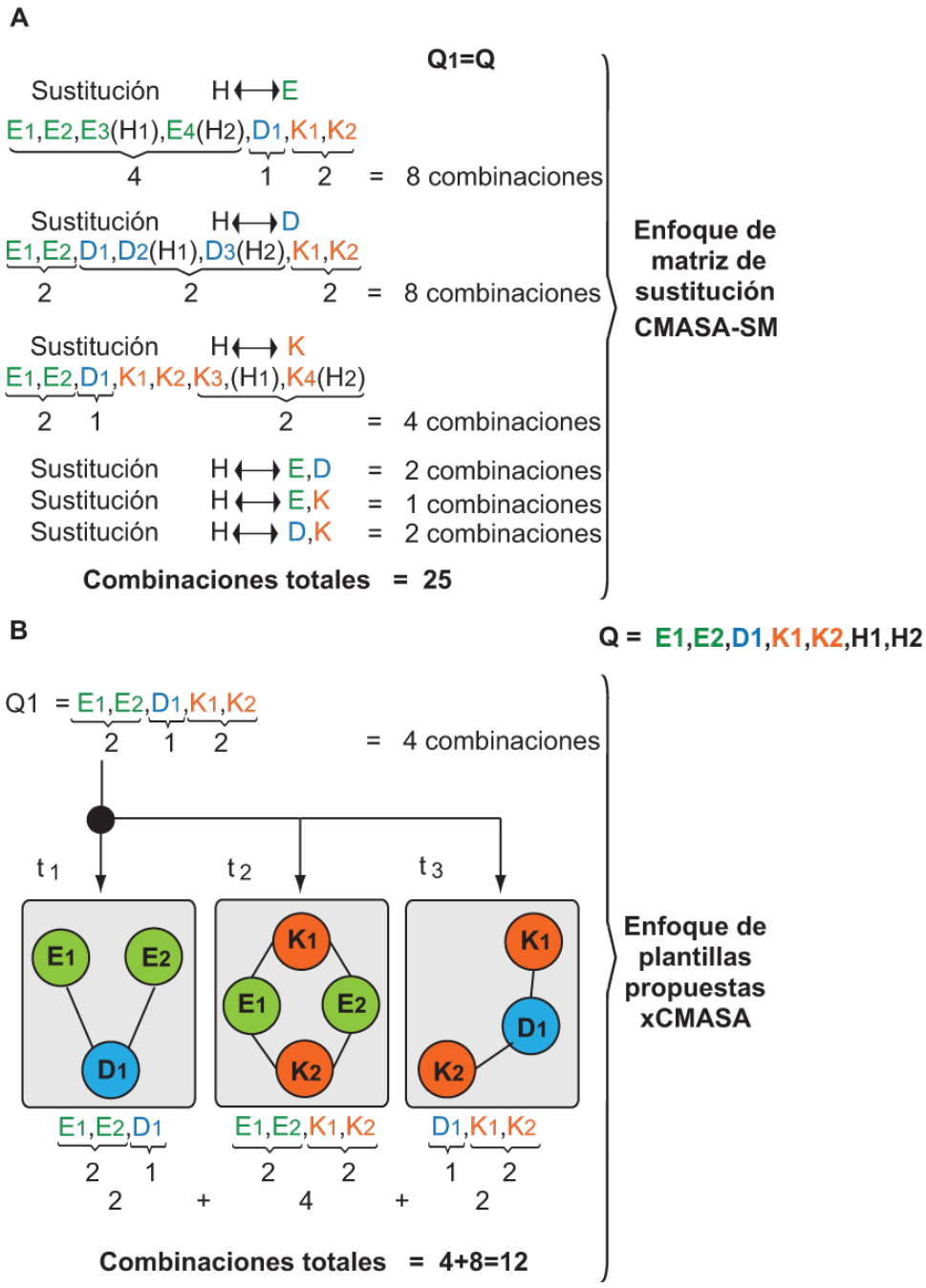


Figura 12: Calculando el número de comparaciones para manejar mutaciones puntuales. Usando la entrada mostrada en la Figura 11 y bajo el enfoque de matriz de sustitución, los residuos de tipo H son intercambiables por E,D,K. Esto genera las combinaciones que se muestran en (A). Por el contrario, xCMASA no requiere información adicional, dado que las sub-plantillas son derivados de t_i como se muestra en (B).

(ver Figura 10) de un posible residuo catalítico no está presente en una estructura local de consulta; *iii)* existe una diferencia en el tipo de residuos entre la plantilla y la estructura de la consulta. Esto último implica que CMASA no puede reconocer los sitios mutados, al

menos que ellos estén registrados en la base de datos. Cuando el número de residuos en la plantilla t_i es más pequeño que el número de residuos en el sitio catalítico de la consulta, CMASA produce una coincidencia correcta. Sin embargo, si el número de residuos en t_i es más grande que en el patrón de la consulta, entonces CMASA no proporciona una coincidencia correcta. La extensión propuesta en esta tesis se diseñó para mejorar la predicción cuando cualquiera de estas situaciones ocurren. Aunque CMASA-SM ofrece una solución al caso *iii)*, éste no puede manejar los casos *i)* ni *ii)*.

Vuori *et al.* (1992) realizaron una mutagénesis dirigida al sitio catalítico de proteínas pertenecientes a la familia disulfuro isomerasa (EC 5.3.4.1), el sitio catalítico está compuesto de C-G-H-C, cuando muta a S-G-H-C mantiene una de sus funciones. A los casos similares, donde uno de los residuos se somete a una mutación conservativa, CMASA no los reconoce con una SM identidad. Como prueba de concepto, para mostrar que CMASA falla en la presencia de una mutación puntual, se modificó un residuo catalítico en la estructura cristalográfica de 1MEK (EC 5.3.4.1), cuyos residuos catalíticos son C36-G37-H38-C39, la mutagénesis *in silico* de C36S se realizó con Swiss-PDB Viewer (Guex y Peitsch, 1997). Con esta variante, CMASA no es capaz de predecir los residuos catalíticos o aún asociar el mutante con cualquier clase enzimática. CMASA-SM puede recuperar el sitio con cualquier mutación posible de C a un residuo polar. Sin embargo, si la SM considera el intercambio de C a cualquier otro aminoácido, el software de CMASA no regresa algún sitio putativo significativo, debido a que la abundancia de residuos se incrementa y el valor-p supera el umbral. Por lo tanto, como se ve más adelante, puede haber casos donde aún si se le indica a CMASA el residuo específico a mutar, el algoritmo puede arrojar un resultado incorrecto.

Quizá la limitación más fuerte de CMASA es el mecanismo de la matriz de sustitución para manejar mutaciones. La razón que respalda esta afirmación es que de antemano no se conocen cuáles son las posibles sustituciones para una estructura de consulta, por lo tanto, la SM debería aplicarse a la plantilla y no a la consulta. Si se aplica la SM basada en plantillas, entonces cada sitio catalítico en t_i debería manejarlo por separado para minimizar el costo computacional. Sin embargo, aún si se considerarán los residuos de t_i uno por uno, el número de comparaciones sería muy grande. Sólo como ejemplo

ilustrativo: en la Figura 12 los residuos catalíticos E, D, K son capaces de mutar a H, es decir, un intercambio simple por residuo catalítico. Sin embargo, en un caso real, cada residuo catalítico debería ser sustituido por un conjunto de residuos equivalentes, es decir, cada residuo catalítico contribuiría con una SM particular, lo cual incrementaría el número de comparaciones. Otro punto débil de CMASA-SM es que la abundancia de los residuos incrementará el valor-p como puede verse en las Ecuaciones 3 y 4.

Vale la pena señalar que mientras CMASA es capaz de recuperar residuos mutados y no mutados, xCMASA recupera sólo los residuos no mutados.

3.1.1.3. CMASA extendido: xCMASA

Existen varias formas para superar las limitaciones asociadas con la SM. Una es generar por cada plantilla t_i de n_{ti} residuos, todas las sub-plantillas de $n_{ti} - 1$ residuos y usar estas sub-plantillas para extender la base de datos de plantillas. Esta idea puede ampliarse para generar todas las sub-plantillas de $n_{ti} - k$ residuos; sin embargo, cuando k se acerca a $n_{ti}/2$, el número de sub-plantillas incrementa exponencialmente. Afortunadamente, desde el punto de vista biológico, la mutación de un sólo punto ($k = 1$) es la más abundante. Otra opción para superar esta limitación puede ser la creación de sub-plantillas al mismo tiempo que la estructura local de la consulta se compara con cada plantilla. Debido a su simplicidad, se seleccionó la primera opción, es importante señalar que ambas son equivalentes. Los detalles para la opción seleccionada se presentan a continuación.

La extensión propuesta incluye tres componentes principales:

1. *Generación de nuevas plantillas a partir de las que hay en* T. Dada una plantilla t_i, el número de nuevas plantillas es $\binom{n_{t_i}}{n_{t_i}-1}$, donde cada nueva plantilla es una combinación de $n_{t_i} - 1$ residuos. Esto es, de cada plantilla t_i, n_{t_i} nuevas sub-plantillas se crean y agregan a la base de plantillas. La exclusión de sólo un residuo de t_i permite la recuperación de sus residuos no mutados. Si n_{max} es el número máximo de residuos que una plantilla puede tener y hay N_T plantillas, entonces el tamaño de la base de plantillas se incrementa en al menos $N_T \times n_{max}$ elementos. Para el caso

de la SM un simple reemplazo afecta el número de combinaciones con los residuos para formar estructuras locales, como se muestra en el ejemplo de la Figura 12.

2. *Entonación del parámetro CMAD.* El agregar plantillas más pequeñas (con menos residuos catalíticos) aumentan la probabilidad de coincidencias aleatorias entre las sub-plantillas y las estructuras locales de la consulta. Una forma de superar esta situación es mediante la entonación del umbral del CMAD. Esto se detalla en la Sección 3.2.

3. *Postprocesamiento.* Las coincidencias relacionadas con las plantillas extendidas se filtran cuando las plantilla originales de las cuales fueron derivadas también están presentes en la lista de coincidencias predichas. El único objetivo de esta parte es eliminar la información redundante en la salida y no afecta la eficiencia de la propuesta.

La implementación de xCMASA se basa en el código de CMASA, el cual es público a través de la dirección web:
http://159.226.149.45/other1/CMASA/CMASA.htm, bajo licencia GNU.

3.2. Resultados y discusión

3.2.1. Conjuntos de prueba

Antes de describir los grupos de prueba, es necesario definir el concepto de plantilla y plantilla maestra. Una plantilla se define como un conjunto de coordenadas tridimensionales de cada carbono alfa (C_α) y del átomo más alejado de su cadena lateral por cada residuo catalítico de una enzima dada. Mientras que una plantilla maestra, para determinada familia, se define como la plantilla que minimiza la siguiente expresión (Li y Huang, 2010):

$$SumRMSD(i) = \sum_{j \neq i}^{n_r} RMSD(i,j) \tag{1}$$

Donde, $RMSD(i,j)$ es la desviación media cuadrática entre la *i-ésima* y la *j-ésima* familia y n_r es el número de plantillas en la familia.

Li y Huang (2010) utilizaron dos grupos de proteínas como conjuntos de prueba. Primero, están las proteínas del grupo positivo, cuyos residuos catalíticos están anotados en el CSA y pertenecen a alguna de 163 familias. Por otra parte, las proteínas en el grupo negativo no tienen asociada función catalítica alguna y por lo tanto no tienen registro en el CSA. En el material adicional de CMASA (Li y Huang, 2010) se tiene una lista de 886 proteínas de tipo positivo, con al menos tres residuos catalíticos, distribuidos en 164 familias. A partir de este conjunto, se consideraron 744 proteínas en 163 familias. Las proteínas excluidas presentaron alguna de las siguientes características:

- Diferencias en el número de residuos catalíticos entre miembros de la misma familia y su plantilla maestra.

  Por ejemplo, en el material adicional proporcionado por Li y Huang (2010), la plantilla maestra 1DXLA (estructura con identificador PDB 1DXL y cadena A) tiene cuatro residuos catalíticos C45-C50-H449-E454, mientras que en la base de datos del software CMASA (Li y Huang, 2010), las coordenadas tridimensionales constan de cinco residuos C45-C50-T215-H449-E454, es decir, se les agregó el residuo T215. Las siguientes enzimas pertenecen a la familia 1DXLA: 1EBDA, 1JEHA, 1LPFA, 1LVLA, 1ZMCA, 2A8XA y 3LADA; sin embargo, todas ellas tienen cuatro residuos catalíticos en la base de datos de CMASA, contrario a los cinco residuos en la plantilla maestra. Debido a la diferencia en el número de residuos en la plantilla maestra y los miembros de la familia, CMASA asocia cada una de éstas a una plantilla maestra diferente (1GERA) en lugar de asignarlas a 1DXLA, que es su plantilla maestra.

- Diferencia en los tipos de aminoácidos: La configuración estándar de CMASA realiza una predicción incorrecta cuando la plantilla maestra y la proteína de consulta pertenecen a la misma familia pero tienen al menos un residuo diferente. Por ejemplo, los residuos catalítico de la plantilla maestra 1E7PA son H257-E294-R301-H369-R404; mientras que la estructura 1KF6A, que pertenece a la misma familia, tiene los siguientes residuos H232-G269-R287-H355-R390. Dado que hay una diferencia entre los residuos E294 y G269, CMASA no es capaz de predecir el sitio catalítico. Resultados semejantes se obtienen con 1NEKA, miembro de la misma familia.

Un grupo de 744 estructuras proteícas forman parte del grupo de prueba positivo, las cuales pertenecen a una de 163 familias. El grupo negativo está compuesto por 10575 estructuras del nrPDB y fueron proporcionadas por los autores de CMASA. Se denominaron a estos dos grupos (el positivo y el negativo juntos) como el conjunto de prueba *A*. Un grupo de prueba adicional, el conjunto de prueba *B*, está compuesto por el mismo grupo negativo. Sin embargo, el grupo positivo es diferente. Este grupo está compuesto por un conjunto de proteínas mutadas a partir del grupo positivo en el conjunto *A* (744 proteínas). Después se filtraron todas las proteínas que tienen al menos tres residuos catalíticos, esto genera un grupo de 480 enzimas pertenecientes a 108 familias (de 163 iniciales). Posteriormente, una mutación puntual se aplicó a cada sitio catalítico. Los residuos seleccionados se sustituyeron por alanina (alanine scanning). La mutagénesis computacional de residuos a alanina está entre los métodos más rápidos para validar hipótesis y métodos relacionados a la función de las proteínas (Bromberg y Rost, 2008).

3.2.2. Medidas de similitud y criterios para evaluación del desempeño

- CMAD (Contact Matrix Average Deviation) (Li y Huang, 2010): Esta es una medida de similitud para las distancias de los n_{ti} átomos pertenecientes a una plantilla t_i y aquellos que pertenecen a la estructura local lq.

$$CMAD = \frac{1}{n_{ti}(n_{ti}-1)} \sum_{j=1}^{n_{ti}} \sum_{k=1}^{n_{ti}} \mid d(t_i[j], t_i[k]) - d(lq[j], lq[k]) \mid, \tag{2}$$

  donde $d(t_i[j], t_i[k])$ es la distancia entre los átomos j and k de la plantilla t_i; $d(lq[j], lq[k])$ es la distancia entre los átomos j y k de la estructura local de entrada lq. Para considerar una coincidencia entre la estructura de consulta y una plantilla dada, el CMAD de sus correspondientes C_α y sus átomos f_a debe ser igual o más pequeño a un umbral determinado ajustable por el usuario.

- Significancia estadística: Stark *et al.* (2003) proporcionaron una medida para la significancia estadística en la búsqueda de estructuras locales dependientes del RMSD:

$$P(RMSD \leq R_M) = 1 - e^{-EF(R_M)} \tag{3}$$

con

$$EF(R_M) = 473(\theta * 0.4^N R_M^{4.93N-5.88}), \quad (4)$$

donde EF es el valor esperado del número de estructuras coincidentes con la plantilla t_i, cuyo RMSD es menor o igual a R_M. N es el número de residuos en la estructura de consulta y θ es el producto de los porcentajes de abundancia de todos los residuos.

Los residuos catalíticos de las proteínas 1ADO y 1ALD se utilizaron para ilustrar el cálculo de las métricas descritas. En la Figura 13**(A)** se muestran las distancias entre los átomos más alejados de las cadenas laterales para los residuos ASP-33, GLU-187 y LYS-229 en cada proteína. Con estos valores se aplica la fórmula de CMAD (ver Ecuación 2), la cual establece una medida de semejanza entre los sitios mediante las distancias formadas entre sus propios átomos (indicadas por las líneas punteadas de colores). En el caso del valor-p dependiente del RMSD, en la Figura 13**(B)** se ilustra el alineamiento entre átomos que pertenecen al mismo tipo de aminoácidos y la aplicación de la fórmula del RMSD. Suponiendo que la proteína 1ALD fuera la estructura de consulta con 10 residuos, siete de los cuales tienen aminoácidos distintos a los de la figura, entonces la abundancia θ es igual a $\frac{1}{10} \times \frac{1}{10} \times \frac{8}{10}$ y su valor-p para un umbral de 0.950 y $N = 3$ es igual a 0.007 Si ahora, los 7 residuos iniciales se consideran con el mismo tipo de aminoácido que alguno de los tres residuos de la figura, entonces la abundancia θ es igual a $(1/10) \times (1/10) \times (8/10)$ y su valor-p para un umbral de 0.950 y $N = 3$ es igual a 0.059. Con este ejemplo se comprueba que cuando hay un tipo de aminoácido en los residuos de comparación cuya abundancia en los residuos de la estructura de consulta es considerable, entonces el valor-p se incrementa.

Los criterios utilizados para evaluar los resultados obtenidos por la extensión propuesta fueron el umbral de significancia estadística (valor-p = 1.0×10^{-4}) y la coincidencia exacta de todos los residuos catalíticos en t_i.

- Verdadero Positivo(TP): Es una proteína de consulta de tipo positivo que coincide exactamente con una plantilla maestra (t_i) y su valor-p está por debajo del umbral.

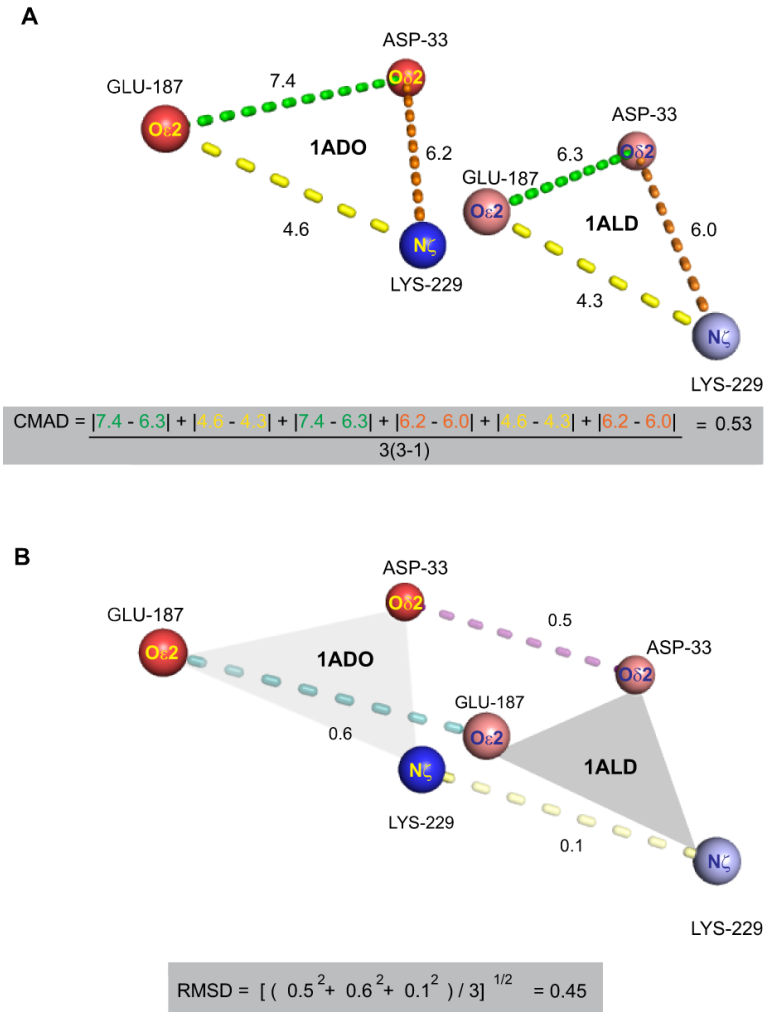


Figura 13: Cálculo de CMAD y RMSD para la comparación de dos estructuras locales pertenecientes a las proteínas 1ADO y 1ALD.

- Falso Negativo (FN): Dada una proteína de consulta de tipo positivo, se obtiene una predicción cuyos residuos catalíticos no coinciden exactamente con los residuos catalíticos de la proteína de consulta, o hay una coincidencia con un valor-p mayor al umbral.

- Falso Positivo (FP): Dada una proteína de consulta de tipo negativo, se obtiene una predicción cuyos residuos catalíticos coinciden con alguna plantilla maestra y su valor-p está por debajo del umbral.

- Verdadero Negativo (TN): Dada una proteína de consulta de tipo negativo, no se obtiene una predicción que coincida con cualquier plantilla maestra y su valor-p esté por debajo del umbral.

Basados en las definiciones previas, se calcularon los siguientes criterios de desempeño :

- Sensibilidad (Gart y Buck, 1966):

$$S_n = \frac{TP}{TP + FN}. \tag{5}$$

- Exactitud:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}. \tag{6}$$

- Precisión:

$$Pr = \frac{TP}{TP + FP}. \tag{7}$$

- MCC: Es el coeficiente de correlación de Matthews (Matthews, 1975):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{8}$$

3.2.3. Diseño de los experimentos computacionales

Los experimentos computacionales están divididos en cuatro escenarios. El primer escenario está destinado a reproducir los resultados reportados en (Li y Huang, 2010). El segundo consiste en evaluar las predicciones obtenidas por xCMASA en caso de que no se presenten mutaciones. El tercer escenario evalúa el desempeño de CMASA cuando una mutación puntual se aplica a los sitios catalíticos. El cuarto escenario analiza el desempeño de xCMASA cuando las proteínas mutadas se proporcionan como entradas al método. Todos los escenarios se evaluaron con el valor por omisión de CMAD de 1.2.

1. *Escenario NoM_MT (Sin Mutación y Plantillas Maestras).* En este escenario se usa el conjunto de prueba A, descrito en la Sección de 3.1. Se reprodujeron los experimentos presentados en Li y Huang (2010) y se muestran en la Tabla 5, donde puede observarse una mejora sobre los resultados reportados en (Li y Huang, 2010). Esta

diferencia se debe principalmente al filtrado que se realizó al conjunto de prueba. El cual elimina algunos casos donde CMASA falla.

Tabla 5: Reproducción de los resultados reportados en Li y Huang (2010) para plantillas maestras y consultas sin mutaciones.

Caso	Sn	Acc	Pr	MCC
CMASA Li y Huang (2010)	0.750	0.940	NA	0.820
Reproducción	0.795	0.981	0.907	0.840

2. *Escenario M_MT (Mutación y Plantillas Maestras)*. Este escenario usó el conjunto de prueba B y su objetivo fue evaluar el desempeño de CMASA cuando las proteínas mutadas se introducen como consultas al método. Como se muestra en la Tabla 5, con excepción de la exactitud, los otros criterios de desempeño disminuyen de forma considerable (M_MT) al compararlos con los resultados del escenario anterior (NoM_MT). Se observa que a pesar de que la exactitud no disminuye, el método sólo puede identificar 3 proteínas de 480, esas proteínas coincidieron con plantillas maestras de otras familias. CMASA tiene la opción de usar SM para hacer frente a las mutaciones. Para analizar las limitaciones de este enfoque, se seleccionó un conjunto de diez casos como se muestra en la Tabla 7. En todos los casos se usó una SM conociendo de antemano los tipos de residuos que se someterían a una mutación, aún en este caso, con CMASA-SM se obtuvieron predicciones con valores-p por encima del umbral; y en tres casos (1AHP, 1B4K, 1G8F) no se generaron coincidencias. Además de estos tres casos, todos los demás, obtuvieron la coincidencia con su plantilla maestra en la primera posición justo por debajo del umbral para el valor-p, con excepción de 1A3H. Se podría pensar que una solución a este problema sería cambiar el umbral para el valor-p; sin embargo, la coincidencia correcta para 1A3H aparece en la 6a. posición, y por lo tanto la entonación del valor-p no resuelve casos como este. Todavía más importante, aumentar el umbral del valor-p incrementa el número de FPs en los casos en donde no haya mutaciones.

3. *Escenario NoM_ET (Sin Mutación y Plantillas Extendidas)*: En este escenario se usó el conjunto de prueba A y la base de plantillas con el agregado de las plantillas

propuestas, es decir, 744 proteínas del tipo positivo y 10575 del tipo negativo. El escenario se propusó para evaluar xCMASA cuando las estructuras de consulta no se han mutado. La diferencia con respecto a los resultados obtenidos con el método original CMASA, escenario NoM_MT, tiene que ver con el número de FPs (ver Tabla 6). Mientras CMASA tiene 60 FPs (NoM_MT y M_MT), xCMASA genera 237 FPs (NoM_ET y M_ET), esto se debe principalmente a que las plantillas propuestas son más pequeñas e incrementan la probabilidad de generar coincidencias aleatorias que están por debajo del umbral de CMAD. Esta situación se puede mejorar entonando el valor del umbral CMAD como se muestra más adelante. Es interesante notar que xCMASA recuperó parcialmente varios de estos sitios catalíticos, como es el caso de la proteína 1AHP, aún cuando ésta fue un FN cuando se usó CMASA (ver Tabla 7).

4. *Escenario M_ET (Mutación y Plantillas Extendidas)* : Este escenario se usó para probar el grupo B y su objetivo es evaluar el desempeño de la extensión propuesta cuando las consultas están compuestas en su totalidad por proteínas mutadas, sólo se consideraron mutaciones puntuales. Los resultados obtenidos fueron: 0.967 de exactitud, 0.733 de sensibilidad y 0.646 de MCC, ver Tabla 6 (M_ET). La principal diferencia con el desempeño del conjunto de prueba A es que la proporción de FNs generados por xCMASA es mayor en este conjunto de prueba.

Tabla 6: Criterios de desempeño para CMASA (MT) y xCMASA (ET) con CMAD = 1.2. Consultas Mutadas (M) y No Mutadas (NoM).

Escenario	TP	FN	FP	TN	Sn	Acc	Pr	MCC
NoM_MT	592	152	60	10515	0.795	0.981	0.907	0.840
NoM_ET	592	152	237	10338	0.796	0.966	0.714	0.736
M_MT	3	477	60	10515	0.006	0.951	0.047	0.001
M_ET	357	123	237	10338	0.744	0.967	0.601	0.652

Como se puede ver en el segundo y tercer escenario, xCMASA produce un gran número de FPs en comparación a los de CMASA. Esto es inherente a las plantillas agregadas. El tamaño más pequeño de las plantillas incrementa las coincidencias con estructuras locales en las proteínas de consulta en el grupo negativo, donde no hay función catalíti-

Tabla 7: Ejemplos de predicciones para sitios catalíticos que fueron evaluados como FN por CMASA-SM y como TP por xCMASA.

Proteína	Plantilla maestra	Sitio catalítico	Mutación por alanine scanning	valor-p CMASA-SM	valor-p xCMASA
1AHP	1A8I	K533-R534-K539 -T641	R534A	NA	7.37E-06
2H12	1AL6	S242-H272-H313 -D371	H272A	1.24E-03	3.54E-09
1B4K	1AW5	D127-S175-K205 -K260	K205A	NA	8.00E-06
1CSN	1CKI	D131-K133-D135 -N136-T181	K133A	1.32E-03	1.46E-09
1BHE	1CZF	D202-D223-D224 -H251	H251A	3.08E-03	4.09E-06
1KKT	1DL2	E122-R126-D267 -E409	E122A	2.10E-03	7.52E-09
1A3H	1EDG	N138-E139-H200 -Y202-E228	Y202A	2.72E-01	2.68E-05
1F6D	1F6D	D95-E117-E131 -H213	E117A	6.30E-03	1.74E-05
1G8F	1I2D	R197-H201-H204 -R290	R290A	NA	4.56E-05
1BS4	1LME	G45-Q50-L91-E133	Q50A	1.06E-03	2.13E-08

ca. Una forma de disminuir este efecto es reduciendo el umbral de CMAD. Sin embargo, valores muy pequeños de CMAD incrementan el número de falsos negativos ya que las estructuras locales mutadas son discriminadas (vea la curva de sensibilidad para el escenario M_ET en la Figura 14). Un CMAD de 0.4 se seleccionó como un valor de compromiso que existe entre no obtener demasiados FPs para casos no mutadas y no obtener demasiados FNs para los que están mutados, también un valor de CMAD = 0.4 da la suma más alta sobre todos los criterios de desempeño, para xCMASA, en el conjunto de prueba B. La Tabla 8 muestra las medidas de desempeño de xCMASA con los casos mutados y no mutados con un valor de CMAD 0.4.

Tabla 8: Criterios de desempeño para xCMASA con CMAD = 0.4. Consultas mutadas (M) y no mutadas (NoM).

Escenario	TP	FN	FP	TN	Sn	Acc	Pr	MCC
NoM_ET	592	152	93	10482	0.795	0.978	0.864	0.818
M_ET	298	182	93	10482	0.621	0.975	0.762	0.675

Además del enfoque de cribado por alanina para evaluar la extensión, también se analizó el desempeño sobre un conjunto de casos reales . En este escenario se propusieron dos casos de estudio.

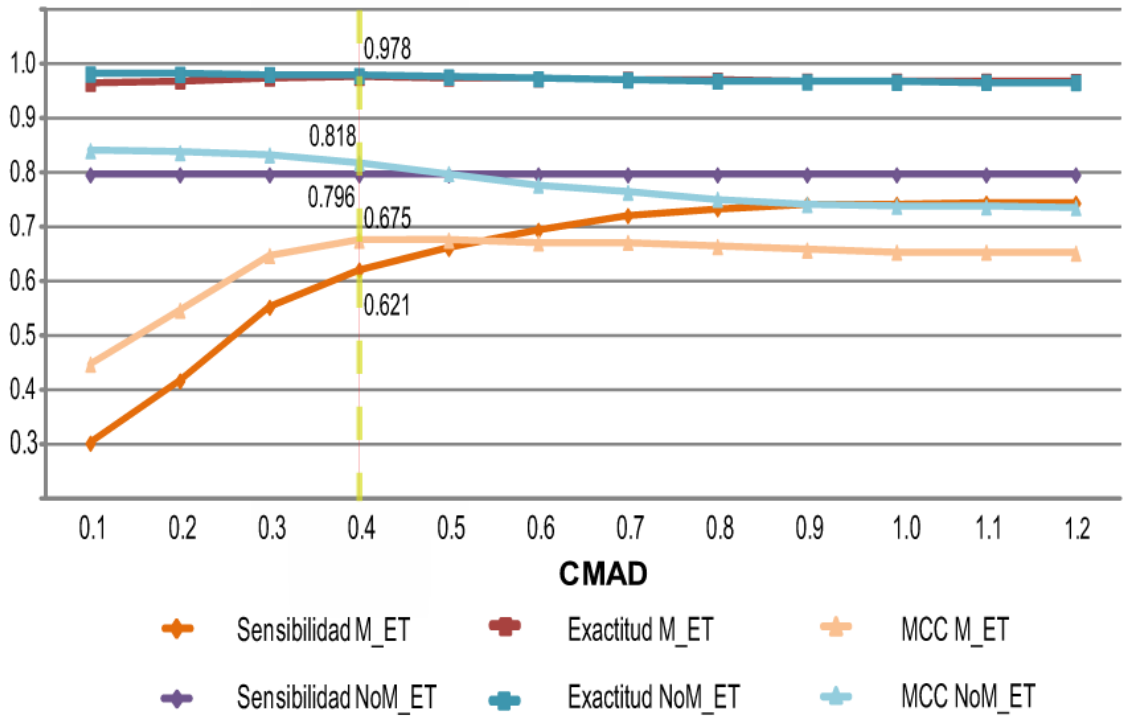


Figura 14: Medidas de desempeño en función de CMAD. La gráfica muestra las relaciones entre sensibililidad, exactitud y MCC de CMASA y xCMASA en escenarios con y sin mutaciones en un residuo.

3.3. Casos de estudio

Para seleccionar cada uno de estos casos, se realizaron los siguientes pasos: Primero, a partir de una revisión de la literatura se identificó una familia de proteínas mutadas con pocos representantes en la base de datos de plantillas de CMASA (CMASA incluye 15341 plantillas). Segundo, a partir de un análisis comparativo estructural con diferentes elementos de la familia, se identificaron algunas proteínas con mutación puntual en sitio catalítico de interés. Para realizar este análisis se utilizó el servicio web PDBeFold (Krissinel y Henrick, 2004). Los casos de estudio se diseñaron para evaluar a qué nivel CMASA original, CMASA-SM y xCMASA son capaces de identificar los residuos catalíticos en una proteína con mutaciones reales, o cuando cierta información relevante está asuente del archivo PDB.

- El papel del residuo S167 en la función catalítica del timilidato sintetasa (E.C. 2.1.1.45) en E. coli fue analizado en (Phan *et al.*, 2000). La estructura 1EVF contiene la variante S167T (ver Figura 15), los residuos catalíticos anotados en el CSA para esta proteína son E58-W80-Y94-C146-R166-D169-N177. Esta proteína no tiene una plantilla en la base de datos de CMASA, a pesar de que hay una serie de proteínas mutantes en la familia de las timilidato sintetasa mutantes en su residuo serina 167(1EV5 le corresponde S167A y a 1EVG le corresponde S167T). Para ambas proteínas los residuos catalíticos son Y94-C146-R166-D169-N177, al igual que en 1EVF. Sin embargo, el cambio en la serina 167 no le permite a CMASA detectar el resto de los residuos catalíticos que no se alteraron. Las salidas del programa no muestran coincidencias por abajo del umbral para el valor-p, y las que están por arriba del umbral no pertenecen a la familia de las enzimas. En el caso de CMASA-SM cuando la SM considera todas las posibles sustituciones de S por residuos polares, es decir, $S \rightleftarrows [T, Y, H, C, N, Q, W]$, el primer resultado de CMASA-SM es 1QQQ, con los residuos putativos E58-T167-D169-N177-H207.

  En este caso, T167 y H207 no son residuos anotados. Es importante notar que este es un escenario optimista para CMASA-SM, ya que dada una proteína de consulta, en este caso 1EVF, se desconoce de antemano que matriz de sustitución se debería utilizar. En contraste, xCMASA fue capaz de recuperar la similitud, es decir, sus salidas coincidieron con los residuos Y94-R166-D169-N177, ya que su primer resultado es una sub-plantilla de 1EV5, con un valor-p de 0.0 y CMAD de 0.1.

- Como un primer ejemplo de datos ausentes se analiza la proteína 3HRC, una proteína del dominio de las cinasas perteneciente a la familia de las transferasas, tiene a D205-K207-E209-N210-T245 como sus residuos catalíticos y su número E.C. es 2.7.11.1. La base de datos de CMASA tiene varias plantillas coincidentes con un patrón subyacente en 3HRC en términos de geometría, tipos de aminoácidos y clasificación enzimática, por ejemplo la estructura de la proteína 2OIC tiene los residuos catalíticos D311-K313-A315-N316-T351. Como en el caso de 3HRC, 2OIC está anotada como una transferasa con número E.C. 2.7.11.1 y difiere de 3HRC en los residuos E209-A315 (ver Figura 16). En este caso, aún si se le proporciona a CMASA-SM el residuo a mutar (E por A) no se obtiene una salida por debajo del

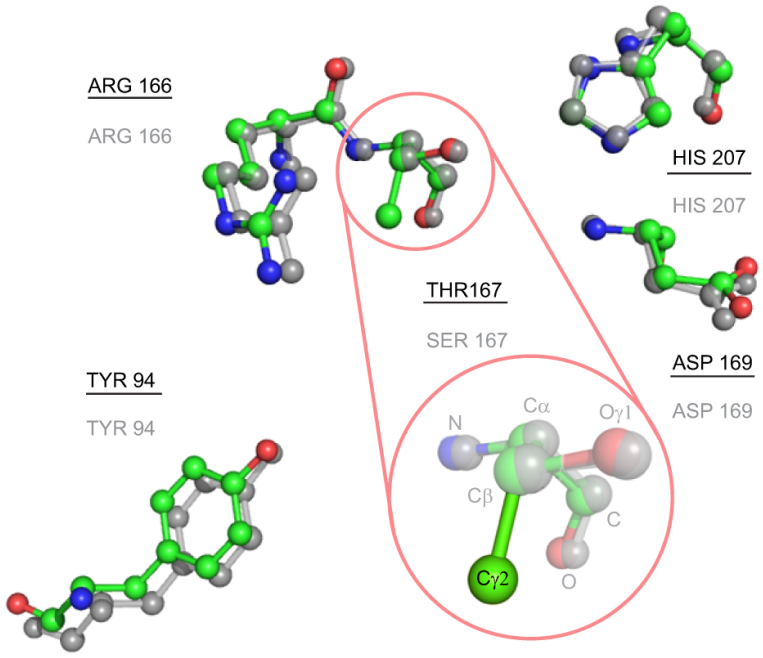


Figura 15: Comparación de estructuras locales catalíticas entre 1EVF y 1BQ1. El residuo número 167 de 1EVF mutó de SER a THR, la extensión propuesta puede detectar todos los residus catalíticos restantes a diferencia de CMASA o CMASA-SM. La figura muestra una superposición de los residuos catalíticos en la estructura de consulta 1EVF (con colores por elemento e identificadores subrayados) y los residuos de las plantilla asociada 1BQ1 (los elementos están en gris y los identificadores no están subrayados). En la figura se muestra la similitud entre las estructuras locales de los residuos no mutados.

umbral para el valor-p, la razón por la que no se logran identificar los residuos catalíticos se debe a que un átomo no está presente; xCMASA, por el contrario, fue capaz de identificar los siguientes residuos D205-K207-N210-T245 con un valor-p de 2.2×10^{-16} y CMAD de 0.114, donde las proteínas asociadas a la predicción pertenecen a la misma clase enzimática.

Se puede considerar nuevamente a 3HRC y la plantilla de 1UU9 como otro caso de datos faltantes, ya que 1UU9 tiene los residuos catalíticos D205-K207-E209-N210-T245, los mismos que tiene 3HRC. Sin embargo, CMASA no logra predecir todos los residuos catalíticos porque 3HRC no contiene el átomo $O_{\epsilon 2}$ del residuo E209 (ver Figura 17). Éste es el átomo más alejado en la cadena lateral para este residuo, entonces el método no puede representar el residuo y los procesos de emulación y comparación de la estructura local hacen caso omiso de todas las plantillas que puedan coincidir.

xCMASA determina a 1UU9 como su tercera coincidencia (después de 2OIC y 20IB)

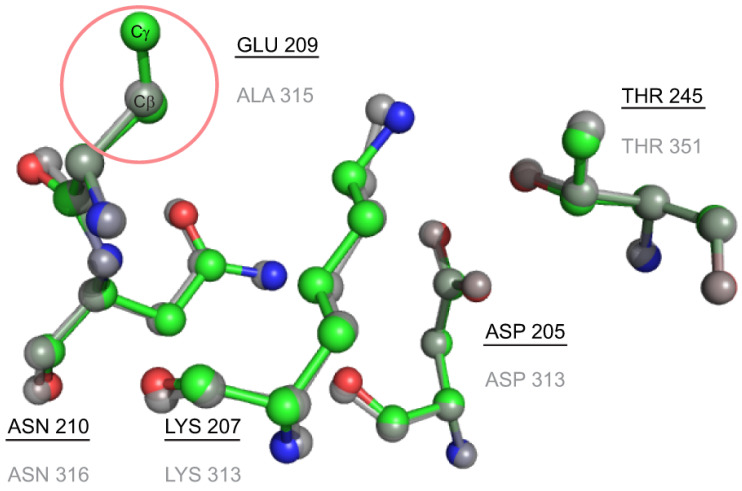


Figura 16: Comparación de estructuras locales catalíticas entre 3HRC y 2OIC. La plantilla de 2OIC tiene su estructura catalítica similar a la de 3HRC, con excepción del residuo A315. xCMASA es capaz de detectar los residuos catalíticos D205-K207-N210-T245. La figura muestra una superposición de los residuos catalíticos en la estructura de consulta 3HRC (con colores por elemento e identificadores subrayados) y los residuos de las plantilla asociada 2OIC (los elementos están en gris y los identificadores no están subrayados).

cuando 3HRC es la consulta. La información incompleta en el archivo PDB no es un evento ocasional y esto puede afectar la predicción de CMASA o CMASA-SM. Cabe resaltar, que cuando hay un átomo faltante, un preprocesamiento no trivial podría resolver el problema de completar el archivo PDB. En el caso de estudio expuesto, se uso una SM; sin embargo, CMASA se detiene después de algunos cálculos ya que no puede manejar la cantidad de candidatos que coinciden con las estructuras locales de la consulta. Este número llega a ser muy grande debido a que cada residuo de tipo E, en la proteína de consulta, se reempla por otros 19 aminoácidos y cada uno de los 19 aminoácidos en la consulta se reemplaza por E.

3.4. Discusión

Se propuso una extensión al método CMASA conocida como xCMASA. La extensión tiene como base la generación de un conjunto de sub-plantillas. La variante propuesta le permite a CMASA contar con una manera más eficiente de manejar las mutaciones puntuales. xCMASA preserva el poder de predicción de CMASA en caso de que no existan mutaciones y por un costo adicional acotado por el producto del máximo número de residuos en las plantillas por el número de plantillas. Esto permite superar las limitaciones

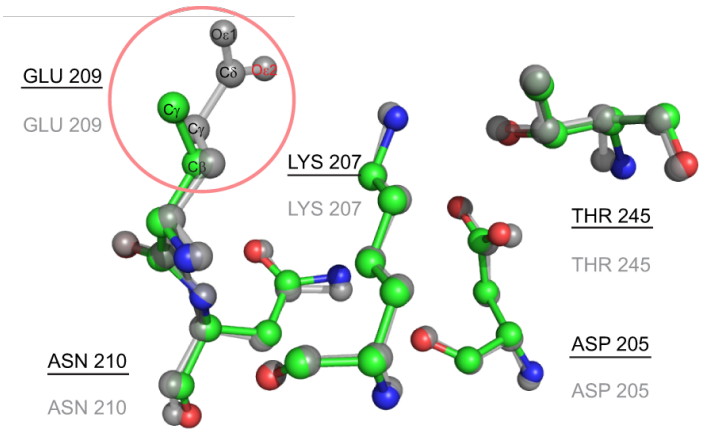


Figura 17: Comparación de estructuras locales catalíticas entre 3HRC y 1UU9. A pesar de que las proteínas 3HRC (con colores por elemento e identificadores subrayados) y 1UU9 tienen las variantes A209E (los elementos están en gris y los identificadores no están subrayados), en 3HRC hay un átomo faltante en la cadena lateral (O_2), esta situación ocasiona que CMASA falle. Por el contrario, xCMASA es capaz de detectar los residuos no mutados.

de CMASA-SM de tener que proveer una SM adecuada.

El método no sólo simplifica el proceso de ajustar la matriz de sustitución, éste también puede identificar sitios catalíticos aún cuando falte el átomo más alejado en un residuo catalítico.

Capítulo 4. Caracterización y clasificación de proteínas basadas en la estructura terciaria de sus cavidades: Un caso de estudio

En el Capítulo 3 se ve cómo una propuesta que se basa en patrones predefinidos puede ser exitosa para identificar estos patrones en una proteína de consulta. Una pregunta natural que surge al ver este resultado es: ¿Se pueden agrupar proteínas relacionadas funcionalmente usando información estructural que no incluya patrón alguno en forma explícita?. Como una respuesta aproximada a la pregunta planteada, en este capítulo se propone un método para la comparación de estructuras terciarias de proteínas mediante la caracterización tanto de rasgos geométricos como de interacciones moleculares en los residuos extraídos de las cavidades mayores de las proteínas, en especial, aquellas que alojen residuos catalíticos. Para analizar el desempeño del método propuesto se expone un caso de estudio compuesto por un conjunto particular de estructuras de proteínas de la superfamilia de las cinasas (Manning *et al.*, 2002).

Para enmarcar el desarrollo del método resulta adecuado dividir el contenido de este capítulo en tres partes: la primera consta de un marco de referencia para los experimentos que se describen en las secciones posteriores; en la segunda parte se muestran los resultados de algunos procedimientos de comparación entre estructuras de proteínas cuyas representaciones abarcan desde el nivel de dimensión 0 hasta el terciario; en la tercera parte se describe detalladamente el método propuesto. Tanto en las subsecciones de la segunda parte como en el método propuesto, se muestran los análisis de: las implementaciones computacionales, los resultados de los experimentos y la discusión de los mismos.

4.1. Marco de referencia

Las distintas representaciones de proteínas se obtienen por diversos métodos, de tal forma que cada una aporta información complementaria sobre sus estructuras. Mientras la estructura primaria se obtiene por métodos bioquímicos, las estructuras secundarias y terciarias requieren información atómica detallada de los aminoácidos que las componen.

Los rasgos considerados en el método propuesto de caracterización, comparación y clasificación en la Sección 4.2 tienen un enfoque local basado en las cavidades mayores de las proteínas. De forma complementaria se realizaron experimentos con los distintos tipos de representaciones estructurales, ya sea con un enfoque global o local, para evaluar la capacidad de discriminar grupos funcionales distintos.

4.1.1. Conjunto base de prueba

Se seleccionó un grupo de 31 proteínas pertenecientes a la superfamilia de las cinasas como base para los conjuntos de prueba de los experimentos descritos en ese capítulo (ver Tablas 9 y 10). El conjunto lo propuso originalmente Manning *et al.* (2002) como una guía para la comparación de estructuras terciarias. Gramada y Bourne (2006) retomaron los elementos de este conjunto para evaluar un método de clasificación basado en la caracterización multipolar de sus superficies (ver Figura 19). El procedimiento para la construcción del conjunto tuvo tres etapas: *i)* la agrupación de proteínas pertenecientes a la superfamilia de las cinasas cuyo porcentaje de identidad fuera mayor o igual a 45 % en sus secuencias, calculada mediante el programa BLASTCLUST (Altschul *et al.*, 1990); la selección de un elemento por cada de una de estas agrupaciones; y cómo último paso se realizó una depuración manual (curado) para conservar únicamente las proteínas que realizan interacciones con ATP (Adenosina trifosfato). En el conjunto de prueba base hay 25 proteínas que pertenecen a las cinasas típicas (TPK); mientras que 6 pertenecen a las cinasas atípicas (AK), es decir, fosforilizan a moléculas distintas a las proteínas. Las proteínas típicas, a su vez, se dividen en ocho grupos funcionales bien etiquetados y uno genérico para las proteínas que no pertenecen a los primeros.

Las cinasas catalizan la modificación covalente de un residuo Ser, Thr o Tyr con el grupo γ-fosfato del ATP. Esta interacción se realiza particularmente en el lóbulo c-terminal, por lo que usando el mismo mecanismo de transferencia, una variedad de sustratos se pueden fosforolizar, las cuales abarcan desde pequeñas moléculas como la colina-cinasa (ck) a regiones completas de tipo hélices (α-cinasas) o tipo giro (TPK) (Gramada y Bourne, 2006).

Scheeff y Bourne (2005) realizaron un estudio filogenético sobre la evolución estructu-

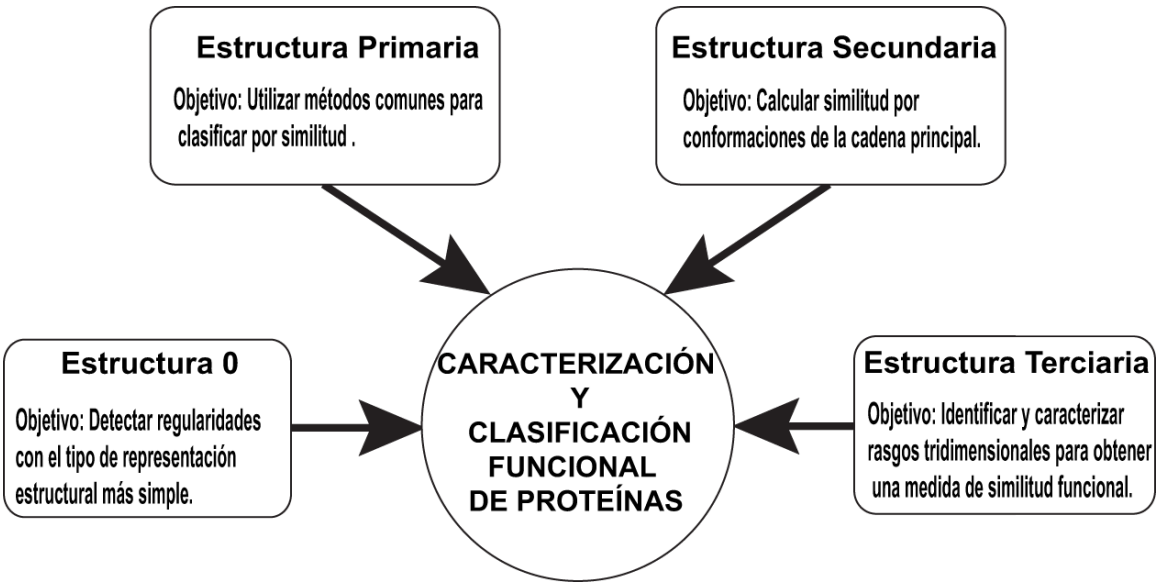


Figura 18: Tipos de representaciones estructurales para detectar semejanzas a distintos niveles.

ral de la superfamilia de las cinasas a partir de un ancestro común, donde se muestra que las proteínas de esta superfamilia han sufrido cambios considerables en sus secuencias de aminoácidos y por lo tanto en sus estructuras tridimensionales. En general, los porcentajes de identidad entre las secuencias de sus miembros representan una semejanza débil, por lo que no sería extraordinario que otras proteínas se unieran a esta superfamilia una vez que sus estructuras sean resueltas. Por otra parte, Thompson *et al.* (2009) publicaron un análisis comparativo de diez núcleos catalíticos localizados en las cavidades mayores de proteínas de la superfamilia de las cinasas, donde clasificaron por clusters los alineamientos de sus estructuras cristalográficas y cuyos objetivos de interacción fueron residuos Ser/Thr y Tyr.

A este conjunto de proteínas, pertenecientes al dominio de las cinasas, se les considera como un excelente grupo de prueba para evaluar métodos de clasificación basados en estructuras tridimensionales, debido a que estas proteínas se pueden rastrear a un ancestro común a pesar de que muchas de ellas tienen un porcentaje de identidad menor al 15 %. Las características de este grupo de prueba presentan un desafío para los métodos de clasificación funcional de proteínas en las diversas representaciones de las proteínas. Para corroborarlo, en este proyecto de investigación se llevaron a cabo experimentos para calcular similitudes en cada una de las representaciones que se muestran

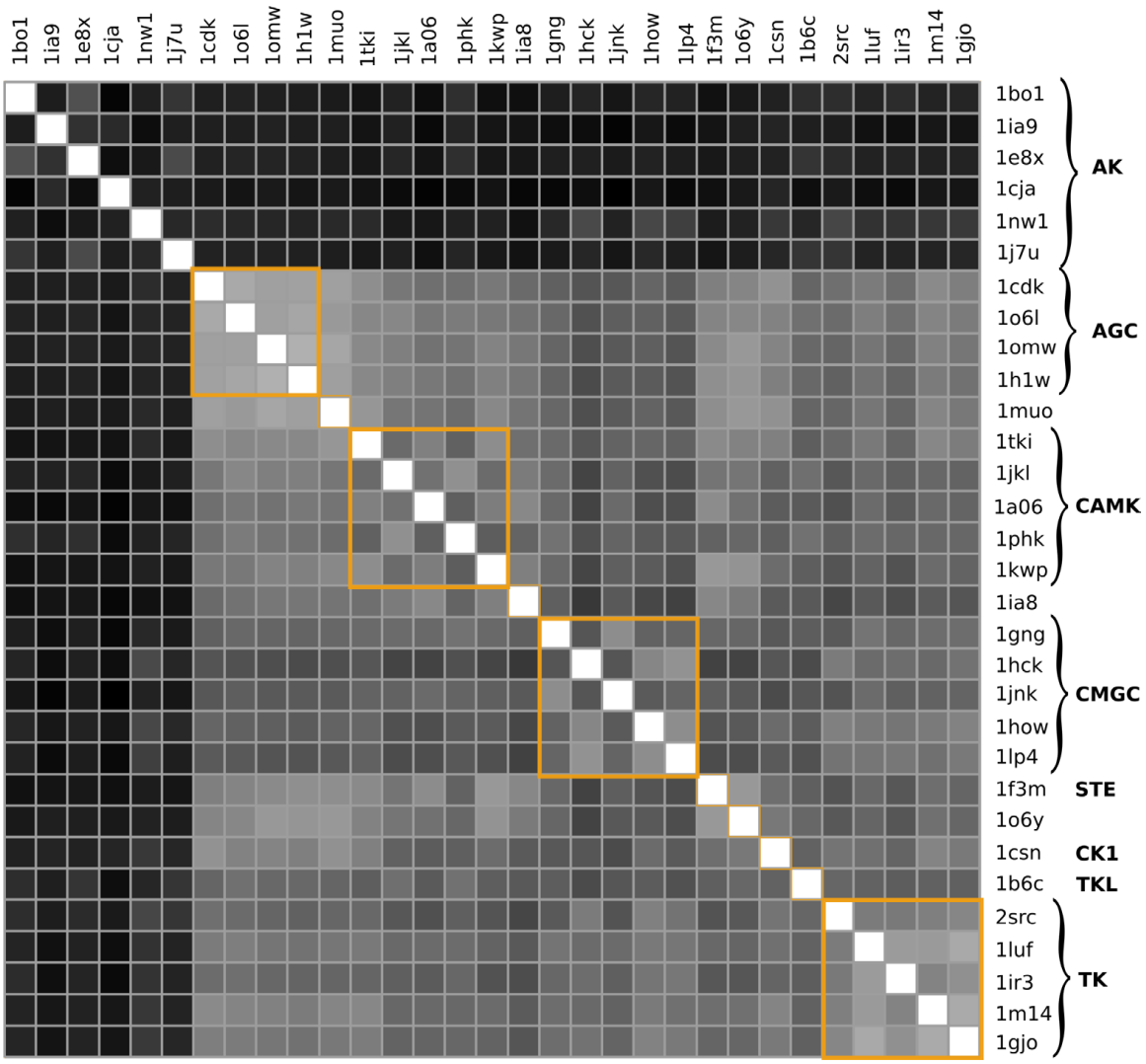


Figura 19: Matriz de similitud para proteínas de la superfamilia de las cinasas obtenida con el método de multipolos. La imagen se tomó de Gramada y Bourne (2006).

en la Figura 18. En la Sección 4.1.2 se especifican los experimentos para el cálculo de similitudes con las estructuras terciarias y los aspectos a considerar para el desarrollo del método propuesto.

4.1.2. Representaciones de estructuras de menor dimensionalidad y sus comparaciones

Uno de los objetivos persistentes en los métodos de comparación de proteínas es la búsqueda de descriptores que tengan la mínima cantidad de información pero que permitan describir significativamente a las proteínas. La representación estructural en *dimensión 0* cuenta determinadas características de las proteínas completas o de sólo

Tabla 9: Proteínas de la superfamilia de las cinasas utilizadas como conjunto base de prueba para los experimentos de los métodos propuestos.

PDB ID	Grupo	E.C.C.	No. de cavidades
1BO1	Atípica	2.7.1.68	107
1IA9	Atípica	2.7.1.68	88
1E8X	Atípica	2.7.1.153	163
1CJA	Atípica	2.7.11.1	126
1NW1	Atípica	2.7.1.32	104
1J7U	Atípica	2.7.1.95	77
1CDK	AGC	2.7.1.37	105
1O6L	AGC	2.7.1.37	47
1OMW	AGC	2.7.1.126	133
1H1W	AGC	2.7.1.37	32
1MUO	Otro	2.7.1.37	43
1TKI	CAMK	2.7.11.1	71
1JKL	CAMK	2.7.11.37	43
1A06	CAMK	2.7.11.17	43
1PHK	CAMK	2.7.1.38	25
1KWP	CAMK	2.7.11.1	96
1IA8	CAMK	2.7.11.1	41
1GNG	CMGC	2.7.1.37	99
1HCK	CMGC	2.7.11.22	43
1JNK	CMGC	2.7.1.37	45
1HOW	CMGC	2.7.11.1	40
1LP4	Otro	2.7.11.1	48
1F3M	STE	2.7.1.37	106
1O6Y	Otro	2.7.11.1	33
1CSN	CK1	2.7.11.1	50
1B6C	TKL	2.7.1.37	204
2SRC	TK	2.7.1.112	53
1LUF	TK	2.7.1.112	32
1IR3	TK	2.7.10.1	44
1M14	TK	2.7.1.112	51
1GJO	TK	2.7.1.112	44

una región, para almacenarlas en vectores que posteriormente se comparan. Dos tipos de características se usaron en los experimentos para esta representación: la primera es la cantidad de residuos de un mismo tipo y para la segunda característica se agruparon los aminoácidos por la propensión de sus cadenas laterales al contacto con solventes polares. Las dos características mencionadas se extrajeron de las cavidades mayores de las proteínas a comparar, donde se supone se realizan las interacciones moleculares que definen las funciones, y los resultados se muestran en el Apéndice A. Las matrices

Tabla 10: Porcentaje de residuos catalíticos en la cavidad mayor. Los residuos catalíticos se obtuvieron con el servidor web CSA V. 2.2.1 y las cavidades con el servidor web CASTP utilizando una esfera de prueba de 1.4 Å.

PDB ID	Grupo	Residuos catalíticos	% de residuos catalíticos en la cavidad mayor	No. de residuos en cavidad mayor
1BO1	Atípica	K150-D278	100 %	61
1IA9	Atípica	Sin registro	NA	NA
1E8X	Atípica	Sin registro	NA	NA
1CJA	Atípica	Sin registro	NA	NA
1NW1	Atípica	Sin registro	NA	NA
1J7U	Atípica	K44-D190	0 %	102
1CDK	AGC	D166-K168-E170-N171-T201	0	41
1O6L	AGC	D275-K277-E279-N280-T313	80 %	48
1OMW	AGC	D317-K319-A321-N322-T353	0 %	64
1H1W	AGC	D205-K207-E209-N210-T245	100 %	38
1MUO	Otro	D256-K258-E260-N261-T292	100 %	64
1TKI	CAMK	D144-R146-E148-N149-T182	100 %	39
1JKL	CAMK	D139-K141-E143-N144-T180	100 %	48
1A06	CAMK	D141-K143-E145-N146	50 %	40
1PHK	CAMK	D149-K151-E153-N154-T186	100 %	43
1KWP	CAMK	Sin registro	NA	NA
1IA8	CAMK	D130-E132-K134-N135-T170	40 %	26
1GNG	CMGC	D181-K183-Q185-N186-S219	0 %	97
1HCK	CMGC	D127-K129-Q131-N132-T165	100 %	25
1JNK	CMGC	D189-K191-S193-N194	80 %	73
1HOW	CMGC	D294-K296-E298-N299	80 %	50
1LP4	Otro	D156-K158-H160-N161-S194	100 %	49
1F3M	STE	D389-K391-D393-N394-T427	100 %	105
1O6Y	Otro	D138-K140-A142-N143	100 %	31
1CSN	CK1	D131-K133-D135-N136	80 %	43
1B6C	TKL	K335-K337-N338-T375	100 %	Cad. B 229
2SRC	TK	D386-R388-A390-N391	0 %	42
1LUF	TK	D724-A726-R728-N729	50 %	47
1IR3	TK	D1132-A1134-R1136-N1137	75 %	51
1M14	TK	D813-A815-R817-N818	50 %	49
1GJO	TK	D626-A628-R630-N631-T661	60 %	33

de similitud obtenidas con esta representación no agrupan adecuadamente las proteínas de acuerdo a su grupo funcional. La simplicidad en la representación de las proteínas de este conjunto de prueba no fue suficiente para detectar los rasgos que conserven la información relevante para discriminar los grupos (ver Apéndice A.2).

A menudo, el primer método que se utiliza para la comparación de dos proteínas es

Tabla 11: Relación entre los identificadores PDB y UniProtKB/Swiss-Prot para las proteínas de la superfamilia de las cinasas utilizadas como grupo de prueba.

PDB ID	Grupo	ID UniProtKB/Swiss-Prot	Formato FASTA Accesión Swiss-Prot
1BO1	Atípica	PI52B_HUMAN	P78356
1IA9	Atípica	TRPM7_MOUSE	Q923J1
1E8X	Atípica	P11G_PIG	O02697
1CJA	Atípica	Q94706_PHYPO	P80197
1NW1	Atípica	Q22942_CAEEL	Q22942
1J7U	Atípica	KKA3_ENTFA	P0A3Y5
1CDK	AGC	KAPCA_BOVIN	P00517
1O6L	AGC	AKT2_HUMAN	P31751
1OMW	AGC	ARBK1_BOVIN	P21146
1H1W	AGC	PDPK1_HUMAN	O15530
1MUO	Otro	STK6_HUMAN	O14965
1TKI	CAMK	Q10466_HUMAN	Q8WZ42
1JKL	CAMK	DAPK1_HUMAN	P53355
1A06	CAMK	KCC1A_RAT	Q63450
1PHK	CAMK	PHKG1_RABIT	P00518
1KWP	CAMK	MAPK2_HUMAN	P49137
1IA8	CAMK	CHK1_HUMAN	O14757
1GNG	CMGC	KG3B_HUMAN	P49841
1HCK	CMGC	CDK2_HUMAN	P24941
1JNK	CMGC	MK10_HUMAN	P53779
1HOW	CMGC	SKY1_YEAST	Q03656
1LP4	CMGC	CSK2A_MAIZE	P28523
1F3M	STE	PAK1_HUMAN	Q13153
1O6Y	Otro	PKNB_MYCTU	P0A5S4
1CSN	CK1	CKI1_SCHPO	P40233
1B6C	TKL	TGFR1_HUMAN	P36897
2SRC	TK	SRC_HUMAN	P12931
1LUF	TK	MUSK_RAT	Q62838
1IR3	TK	INSR_HUMAN	P06213
1M14	TK	EGFR_HUMAN	P00533
1GJO	TK	FGR2_HUMAN	P21802

el alineamiento de sus estructuras primarias o de dimensión 1. Los algoritmos de alineamiento para este tipo de estructuras producen buenos resultados, pero tienen un desempeño deficiente cuando las proteínas presentan 30 % o menos de porcentaje de identidad. Esta limitante sería poco significativa si no es por el hecho de que existe un amplia variedad de proteínas que comparten la misma función a pesar de tener estos niveles de identidad o menores. En el Apéndice B se describen con detalle los alineamientos por pares del conjunto de prueba base y sus resultados. Se puede resumir que con este ti-

po de representación no se obtuvieron agrupaciones evidentes entre los elementos de los grupos funcionales pero sí fueron mejores que los reportados por Gramada y Bourne (2006).

En este esquema de experimentos por capas, en donde cada una de ellas agrega información más detallada a las estructuras de las proteínas, se realizaron comparaciones con las conformaciones rígidas de la cadena principal (cadena de carbonos α), a este tipo de representaciones se les conoce como estructuras secundarias. El objetivo de las comparaciones por alineamientos a las estructuras secundarias de las proteínas del conjunto de prueba, fue detectar la existencia de regiones conservadas comunes que fueran indicativas de la misma función que se realiza por grupos. Sin embargo, al igual que en el par de representaciones previas, los resultados no muestran que este tipo de representaciones discrimine apropiadamente los grupos funcionales (ver Apéndice C).

La comparación de estructuras en dimensión 3, en la siguiente subsección, merece un análisis más detallado por la semejanza de sus descriptores con el método que se propone en la Sección 4.2 y para comparar los resultados sobresalientes de todas las representaciones previas.

4.1.3. Representaciones de dimensión 3

El volumen de información de las estructuras terciarias se ha incrementado significativamente durante las últimas cuatro décadas (Berman *et al.*, 2000), por ejemplo, la base de datos PDB tenía hasta el 2015 más de 100,000 estructuras registradas. El principal método para resolver estructuras terciarias ha sido la cristalografía de rayos X, pero está siendo reemplazada por las estructuras determinadas por resonancia magnética al ofrecer una mayor definición (Branden y Tooze, 1998). El acopio de esta información permite analizar con mayor certidumbre aspectos como las relaciones evolutivas entre las proteínas aún cuando su identidad a nivel de secuencia sea débil. Un procedimiento computacional, ampliamente utilizado para calcular similitudes entre un par de estructuras terciarias, *A* y *B*, se basa en el alineamiento de sus átomos equivalentes, ya sea de toda la estructura (enfoque global) o sólo de una región (enfoque local). El alineamiento consiste en resolver el problema de minimización definido en la Expresión 9. El valor ob-

tenido en la función de minimización se le conoce como la desviación media cuadrática de las distancias entre las estructuras (RMSD por sus siglas en inglés) (Pan *et al.*, 2014).

$$\text{mín} \sum_{i=1}^{N_B} \sum_{j=1}^{N_A} |T + RB_i - A_j|^2, \tag{9}$$

donde N_A es el número de puntos en la estructura *A* y N_B es el número de puntos en la estructura *B*. A_j y B_i son las coordenadas de los carbonos α de las estructuras *A* y *B*, respectivamente. *R* es una matriz de rotación y *T* es un vector de traslación.

Usando el modelo de la Expresión 9, Perutz *et al.* (1960) mostraron que las mioglobinas y hemoglobinas tienen estructuras semejantes a pesar de diferir en sus secuencias. A partir de ese estudio, se han realizado múltiples investigaciones para detectar similitudes estructurales entre proteínas con base en sus alineamientos y de esta forma inferir la funcionalidad de nuevas proteínas (Kolodny *et al.*, 2005). El interés en esta área de investigación se ve reflejado en la cantidad de nuevos métodos publicados durante las últimas cuatro décadas; para ejemplificar: en la Web of Knowledge ISI, el número de artículos que tenían en sus títulos las palabras claves: “Alineamiento de estructuras” o “Comparación de estructuras”, se duplicó cada cinco años desde el año 1980 al 2010. Sin embargo, varios métodos obtienen métricas de similitud contrastantes para un mismo conjunto de proteínas, por lo que el alineamiento de estructuras terciarias aún permanece como un campo de investigación activo (Hasegawa y Holm, 2009; Slater *et al.*, 2013; Collier *et al.*, 2014).

El problema de la comparación de estructuras terciarias es NP-difícil, por lo que no se conoce una solución exacta al alineamiento de estructuras de proteínas (Godzik, 1996). Como consecuencia de esto, los métodos de alineamiento de estructuras terciarias incorporan heurísticas para la minimización de la Expresión 9, por ejemplo, Aghili *et al.* (2005) y Standley *et al.* (2004) utilizan programación dinámica para detectar regiones similares, de tal forma que las superposiciones parciales incrementan la calidad de los alineamientos de tipo global o local. Szustakowski y Weng (2000) seleccionan el mejor alineamiento mediante un algoritmo genético, mientras que Teyra *et al.* (2008) proponen una medida de similitud a partir de una clasificación jerárquica por *clusters* y Veeramalai y Gilbert

(2008) proponen un modelo de alineamiento que considera información de las estructuras primarias y secundarias al igual que algunas propiedades bio-químicas.

En las siguientes subsecciones se describen los experimentos y resultados obtenidos con un método de alineamiento entre pares de estructuras terciarias cuya métrica tiene como base el RMSD; posteriormente se analizan dos casos donde se resalta el aporte de cada nivel estructural para determinar la similitud entre proteínas. El objetivo de estos experimentos fue identificar relaciones de semejanza entre proteínas del conjunto de prueba al mismo nivel estructural del método propuesto de la Sección 4.2, así como con las representaciones previas de la Sección 4.1.2.

4.1.3.1. Materiales y método

Las proteínas del conjunto de prueba descrito en la Sección 4.1.1 fueron utilizadas para realizar el alineamiento de estructuras terciarias por pares, este mismo grupo fue utilizado en el método de clasificación de Gramada y Bourne (2006) y sus estructuras terciarias se obtuvieron del Protein Data Bank (PDB).

Los alineamientos tridimensionales y sus medidas respectivas se calcularon con el visualizador molecular Pymol (Schrödinger, 2010); el cual emplea el algoritmo BLAST y una matriz de puntuación para alineamientos de secuencias como alineamiento inicial y posteriormente realiza un ciclo de refinamientos para ajustar el resultado. La medida de similitud asociada a este alineamiento fue la desviación cuadrática media (RMSD), cuya expresión se encuentra en la Ecuación 10, donde d_{ii} es la distancia euclidiana entre los c_α de los i-ésimos residuos alineados (Hasegawa y Holm, 2009).

$$RMSD = \sqrt{\frac{\sum d_{ii}^2}{N}}. \tag{10}$$

Se usó una matriz de puntuación BLOSUM62, una penalidad por apertura igual a -10.0, una penalidad por extensión igual a -0.5, una apertura máxima de 50 y cinco ciclos de ajuste.

El resultado de los alineamientos entre todos los pares se representa con una matriz de semejanza (*MST*), cuya posición *MST(i,j)* contiene un tono de gris asociado a la des-

viación en angstroms entre la proteína *i*-ésima y la *j*-ésima. Los tonos más claros indican mayor similitud al tener un RMSD más pequeño. La métrica no está acotada superiormente, pero se dividió en tres categorías: la primera abarca de 0 a 2Å, los valores en este intervalo indican una similitud significativa en algoritmos como CE (Bourne y Weissig, 2003); la segunda de los 2Å a los 8Å, y la tercera de los 8Å a los 15Å.

4.1.3.2. Resultados y discusión

La Figura 20 muestra la matriz de semejanza *MS3* para los alineamientos por pares de las estructuras terciarias, resaltan los valores que indican una mayor similitud entre las proteínas pertenecientes al grupo **TK**, al igual que las del grupo **AGC**. En contraste, las proteínas atípicas (grupo **AK**) se diferenciaron de los otros grupos al tener los valores más bajos de similitud, aún entre ellas mismas. A pesar de que algunos grupos tuvieron valores altos de similitud entre sus elementos, también lo tuvieron con un número significativo de proteínas pertenecientes a otros grupos (**AGC**, **CAMK** y **CMGC**).

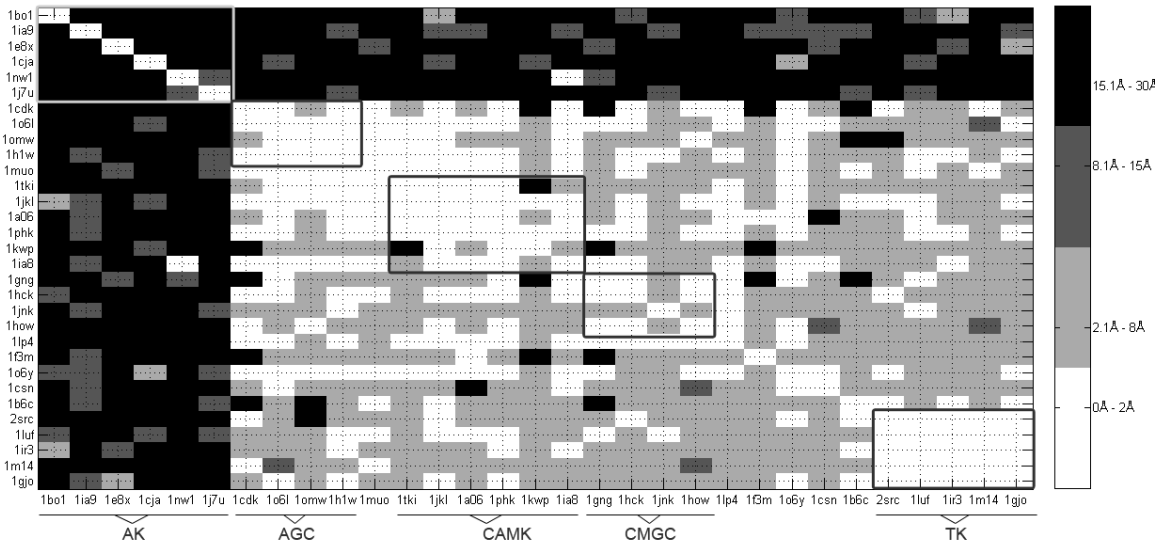


Figura 20: Matriz de semejanza *MS3* para los alineamientos globales entre estructuras terciarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia de las cinasas. Los valores de la medida de similitud están basados en la RMSD en Amstrongs.

Si se consideran los resultados de los alineamientos para los distintos niveles de representación estructural de forma conjunta, destacan algunas relaciones que hubieran sido difíciles de detectar con sólo uno de ellos. Por ejemplo, en el grupo **AGC**, el alineamiento de estructuras terciarias entre las proteínas 1H1W y 1O6L tuvo un RMSD menor a 2Å (ver

Figura 20 en el Apéndice B), como se menciona en la Subsección 4.1.3.1, este valor se considera como indicativo de una fuerte similitud en algunos métodos de comparación. Un resultado equivalente al anterior se observa en los alineamientos de las estructuras secundarias, donde las medidas de similitud fueron superiores a 60/100 (ver matrices de semejanza en las Figuras 40 y 41 del Apéndice C). El alineamiento global de las estructuras secundarias expresado en secuencias de caracteres se muestra en la Figura 21, un aspecto notable se observa en las inserciones/borrados (gaps) contiguas de la proteína 1O6L, debido a que las más grandes se encuentran en zonas donde sus pares alineados de la proteína 1H1W tienen conformaciones de tipo lazo (C $\rightarrow$ coil) (Líneas 1, 2 y 7).

Un resultado que se contrapone a los anteriores, al indicar una similitud débil, se obtuvo con el alineamiento global de sus estructuras primarias, ya que el porcentaje de identidad obtenido fue menor al 20 % (ver *MS* en la Figura 37); el alineamiento para las secuencias de aminoácidos se muestra en la Figura 22; en él se distinguen dos características que contribuyeron al porcentaje de identidad bajo: 1) la diferencia de tamaños en las secuencias y 2) el número considerable de sustituciones conservativas. Una deducción factible, a partir de los resultados obtenidos, es que existe cierto grado de convergencia evolutiva entre las proteínas 1O6L y 1H1W; es decir, son proteínas con estructuras tridimensionales parecidas, pero no así en sus secuencias de aminoácidos. Para corroborarla, se muestra en la Figura 23 la superposición de sus estructuras tridimensionales y sus alineamientos a niveles de estructuras secundarias y terciarias.

Un caso con características distintas se presenta entre las proteínas 1KWP del grupo **CAMK** y 1GNG del grupo **CMGC**. En la *MST* de la Figura 20 se puede observar que el RMSD calculado del alineamiento entre sus estructuras terciarias fue mayor a 15 Å, a diferencia del caso previo, este valor no es significativo para determinar una similitud fuerte. Sin embargo, en el alineamiento global de sus estructuras secundarias, la medida de similitud fue parecida a la del caso anterior (ver Figura 40 para la MS y la Figura 24 para el alineamiento explícito). Mientras que en el alineamiento global de sus estructuras primarias se calculó un 25.64 % de identidad (ver Figura 37 para la MS y la Figura 25 para el alineamiento explícito), este valor se encuentra ligeramente por encima del intervalo de 20 a 25 %, conocido como ”zona de penumbra”, a causa de la incertidumbre para

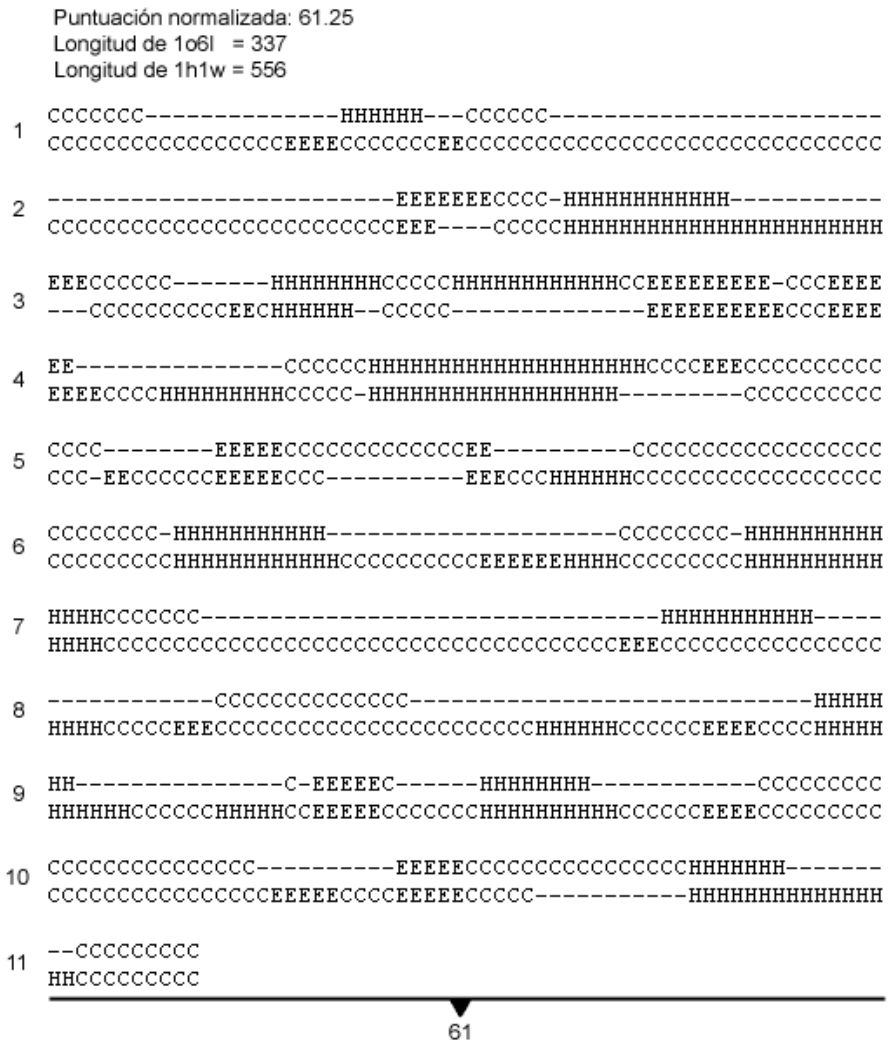


Figura 21: Alineamiento global de estructuras secundarias entre las proteínas 1o6l y 1h1w, obtenido del servidor web SSEA. La secuencia de la proteína 1o6l se muestra en la fila superior y en la fila inferior pertenece a la proteína 1h1w. Los guiones indican (-) inserción/borrado.

determinar la homología entre proteínas con este nivel de identidad (Chung y Subbiah, 1996). Dado que las medidas de los alineamientos en las representaciones estructurales terciarias y secundarias son discordantes, y en el alineamiento de las estructuras primarias no arroja un valor con información relevante, entonces la inspección visual de las superposiciones es la que permite identificar la diferencia evidente entre las estructuras secundarias y terciarias (ver Figura 26).

A partir de los dos casos descritos, resulta lógico suponer que los métodos con procesos de alineamientos en estructuras terciarias son la mejor opción para calcular medidas

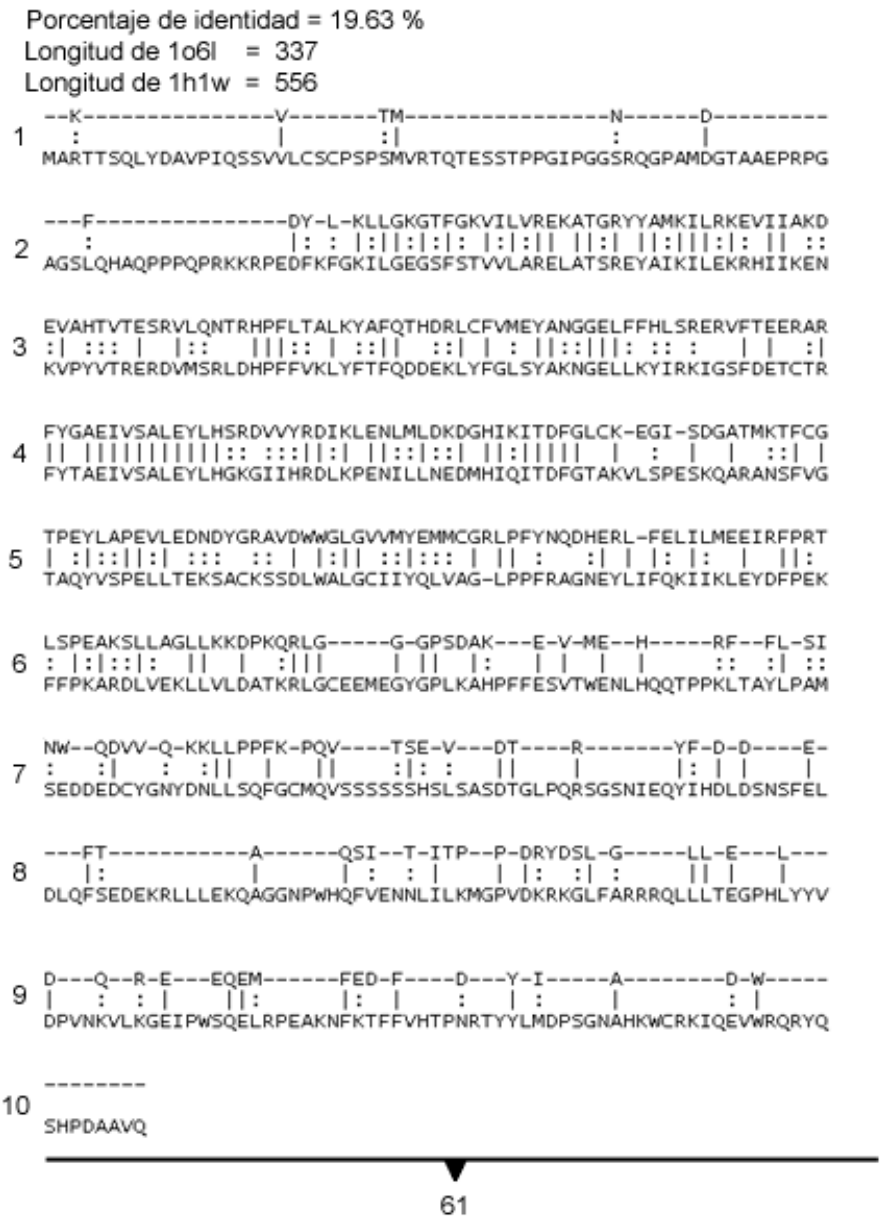


Figura 22: Alineamiento global entre estructuras primarias de las proteínas 1o6l y 1h1w. La secuencia de la proteína 1o6l se muestra en la fila superior y en la fila inferior pertenece a la proteína 1h1w. Las barras (—) indican coincidencia, los dos puntos (:) sustitución por aminoácidos conservativos y los guiones (-) inserción/borrado.

de semejanza; sin embargo, a continuación se listan algunas de las principales desventajas de los alineamientos de estructuras terciarias:

- Las medidas basadas en RMSD se calculan con la suposición de que hay una correspondencia única entre los átomos de las cadenas principales pertenecientes a las proteínas a comparar, aunque realmente no hay una formulación exclusiva para

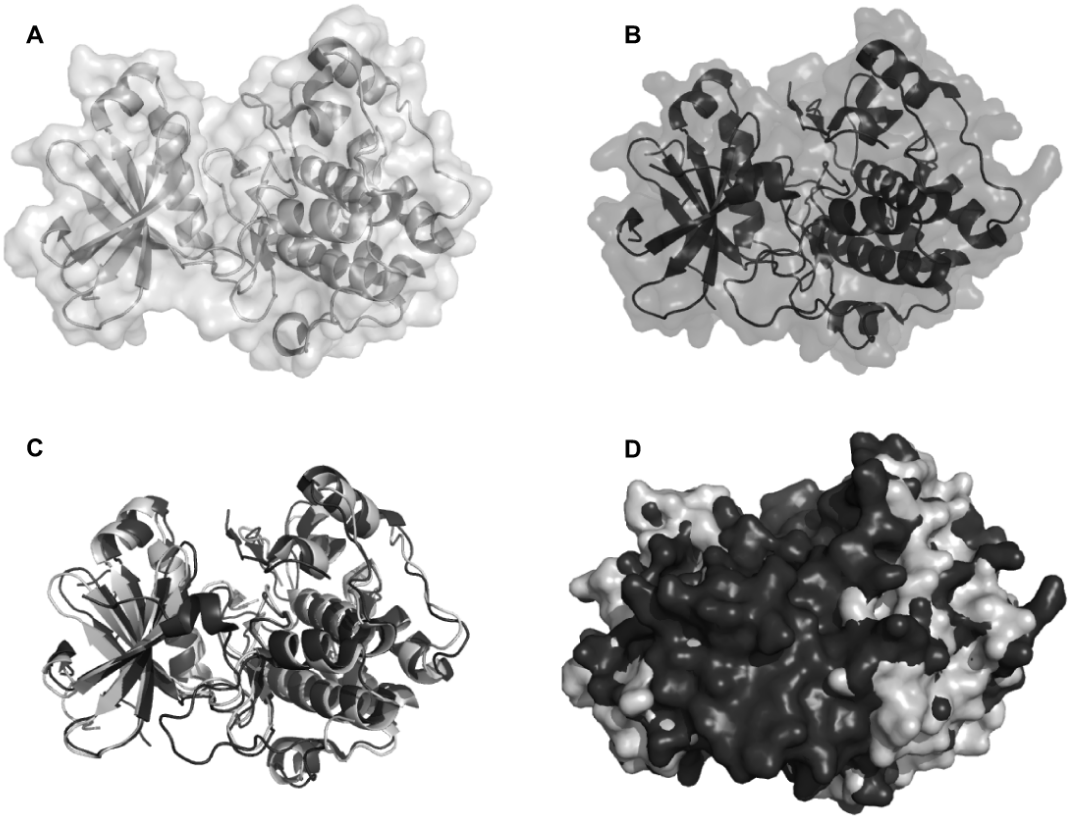


Figura 23: Superposición de las estructuras secundarias y terciarias de las proteínas 1H1W y 1O6L. (A) Superposición de las representaciones secundarias y terciarias de la proteína 1H1W. (B) Superposición de las representaciones secundarias y terciarias de la proteína 1O6L. (C) Superposición de estructuras secundarias. (D) Superposición de estructuras terciarias.

realizar este proceso, por lo que no es extraño que los métodos existentes proporcionen distintas medidas de similitud para un mismo par de estructuras (Gramada y Bourne, 2006). Este hecho hace que los investigadores descubran y redescubran semejanzas entre las estructuras terciarias (Godzik, 1996).

- Las mayores diferencias entre las médidas de similitud basadas en RMSD se presentan entre los alineamientos globales y los que se enfocan en características locales, tales como los de empaquetamiento o los que usan patrones de interacción molecular (Godzik, 1996).

- Las proteínas pueden estar sujetas a pequeñas fluctuaciones estructurales de la cadena lateral o fluctuaciones *in vivo* de sus cadenas principales cuando éstas son muy grandes, por lo que una simple toma estática de la estructura cristalizada no puede reflejar este comportamiento, lo que implica una mayor susceptibilidad a erro-

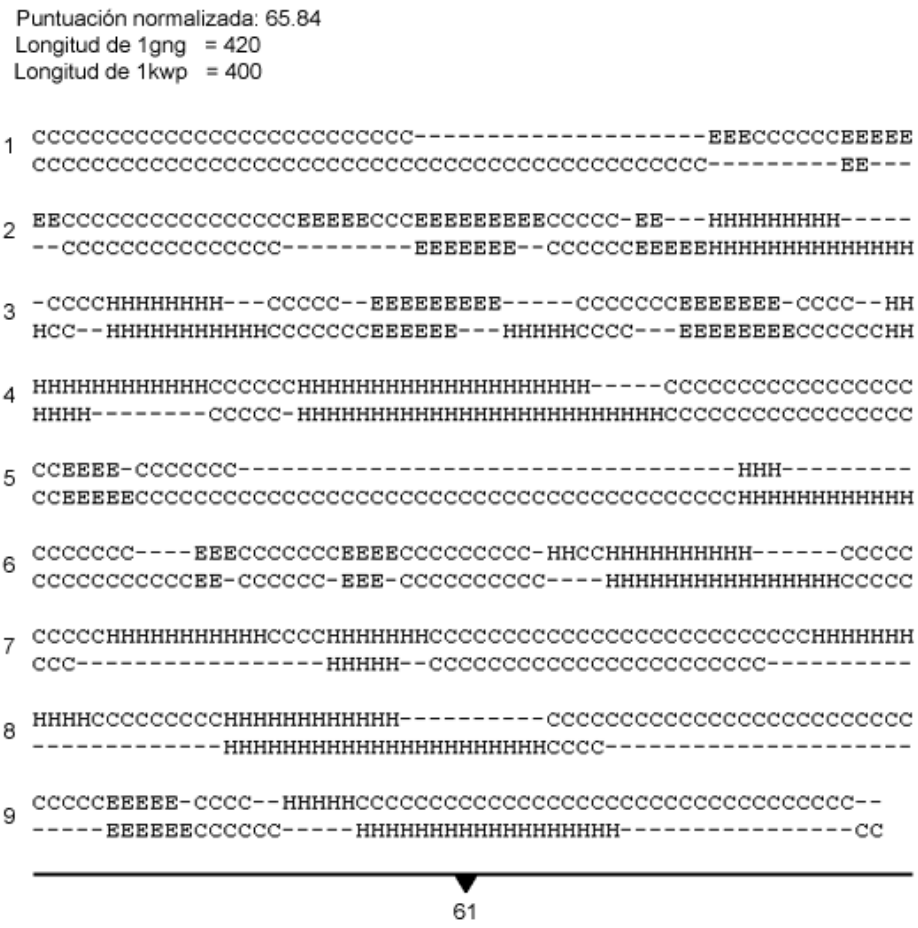


Figura 24: Alineamiento global de estructuras secundarias entre las proteínas 1gng y 1kwp, obtenido del servidor web SSEA. La secuencia de la proteína 1gng se muestra en la fila superior y en la fila inferior pertenece a la proteína 1kwp. Los guiones indican (-) inserción/borrado.

res de comparación con los métodos que usan modelos de cuerpos rígidos (Berman *et al.*, 2000).

Algunos métodos recientes están utilizando un enfoque mixto global-local en los carbonos α para combinar las ventajas de ambas caracterísiticas y superar las desventajas mencionadas previamente, como los que se proponen en Pan *et al.* (2014); Leif y Jinfeng (2012). Además, en el método de Pan *et al.* (2014), el alineamiento es independiente del orden en como se almacenaron las secuencias de las estructuras, esto permite una mayor precisión en el cálculo de similitudes para determinadas conformaciones complejas a nivel global.

Las medidas de similitud obtenidas y analizadas en esta sección no muestran una

Figura 25: Alineamiento global entre estructuras primarias de las proteínas 1gng y 1kwp. La secuencia de la proteína 1gng se muestra en la fila superior y en la fila inferior pertenece a la proteína 1kwp. Las barras (—) indican coincidencia, los dos puntos (:) sustitución por aminoácidos conservativos y los guiones (-) inserción/borrado.

discriminación clara entre grupos funcionales. Sin embargo, aún con métodos como el alineamiento de secuencias de estructuras primarias, se mejora el resultado publicado en Gramada y Bourne (2006). Este aspecto presenta un reto para proponer descriptores y métodos en estructuras terciarias de proteínas que extraigan los rasgos subyacentes para detectar similitudes en grupos funcionales. En la siguiente sección se describe una propuesta para clasificar grupos de proteínas mediante la caracterización de sus cavidades funcionales, incorporando propiedades geométricas y de interacción molecular. El método propuesto usa la comparación de triangulaciones como reemplazo de los alineamientos en cualquier nivel estructural con el objetivo de evitar las desventajas que cada

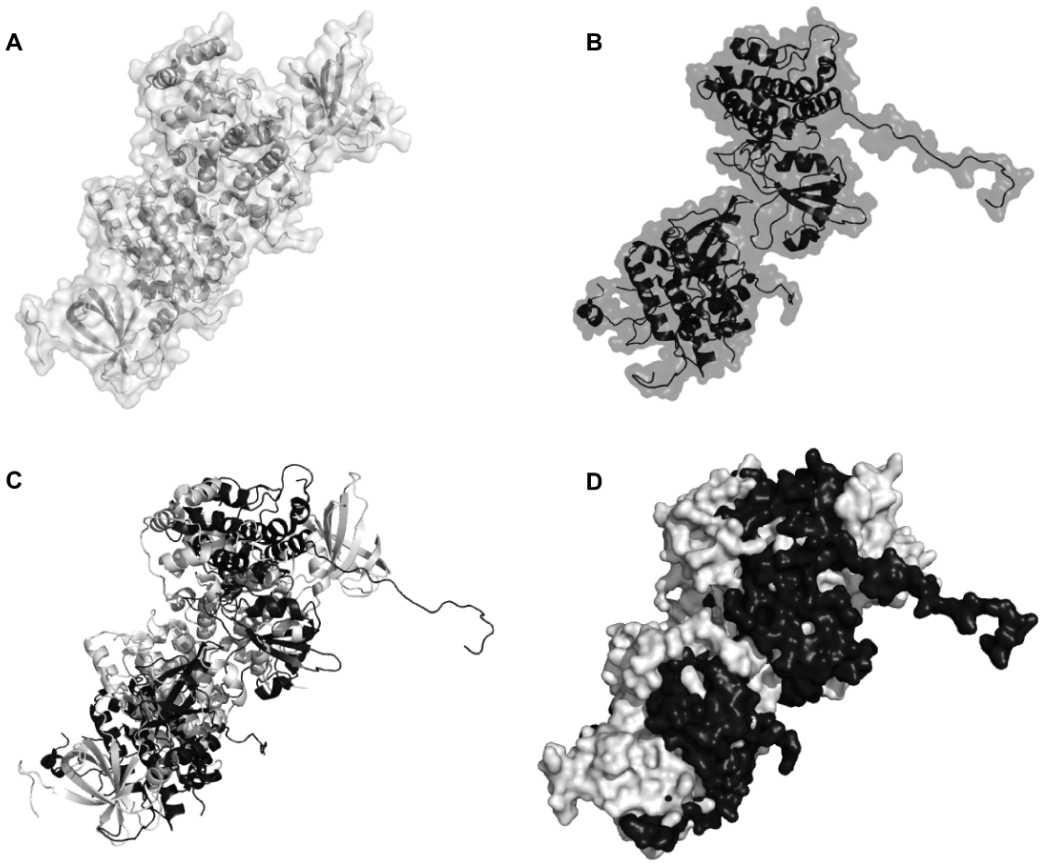


Figura 26: Superposición de las estructuras secundarias y terciarias de las proteínas 1gng y 1kwp. (A) Superposición de las representaciones secundarias y terciarias de la proteína 1gng. (B) Superposición de las representaciones secundarias y terciarias de la proteína 1kwp. (C) Superposición de estructuras secundarias. (D) Superposición de estructuras terciarias.

uno de ellos presenta.

4.2. Propuesta de un método de caracterización y clasificación de proteínas basado en la estructura terciaria de sus cavidades

En las estructuras de las proteínas, las regiones superficiales conservadas cobran mayor importancia debido a que en ellas se presentan a menudo las interacciones moleculares, tal como ocurre con los sitios catalíticos (Binkowski y Joachimiak, 2008). La identificación y representación de características claves en estas regiones se justifican bajo el principio de que la semejanza estructural implica funciones iguales, mientras que las diferencias de forma y localización pueden ser indicativos de diversidad en la misma familia estructural-funcional (Thompson *et al.*, 2009). Algunos estudios como los publicados por Laskowski *et al.* (1996) y Thompson *et al.* (2009) muestran que frecuentemente

las regiones con actividad catalítica, tienen una estrecha correspondencia con las superficies pertenecientes a las cavidades mayores de las proteínas. Además, estas contienen residuos claves en las interacciones moleculares con un alto grado de conservación ante la presión evolutiva.

Proponer y caracterizar los rasgos que se han conservado en las estructuras tridimensionales son tareas esenciales que requieren los métodos de comparación de estructuras para la predicción de sus funciones, y en las que han contribuido de forma importante las tecnologías de adquisición de datos, el desarrollo de algoritmos y la geometría computacional. En la introducción de este documento se presentó una revisión de caracterizaciones y de métodos que las usan para comparar de forma general estructuras tridimensionales. Dundas *et al.* (2006) proponen específicamente un método para detectar átomos estructuralmente conservados en las cavidades funcionales de las proteínas, a los que denominan: *firma de cavidades*. También incorporan un mecanismo que genera plantillas de forma automática para caracterizar familias mediante una clasificación jerárquica por *clusters*; las firmas obtenidas por nivel intentan capturar las fluctuaciones que tengan los átomos de las firmas de cavidades de las que se derivan y de esta forma representar rasgos evolutivos que existan entre ellas.

En la sección 4.1 se calcularon matrices de similitud como referentes para evaluar los resultados obtenidos con la propuesta de esta sección; por lo que a continuación se describe el método desarrollado en este proyecto de investigación para caracterizar y comparar cavidades de proteínas, la forma en la que se construyó este método le permite extenderse a otras regiones de la estructura con algunas modificaciones simples.

4.2.1. Materiales y método

Se utilizaron las estructuras terciarias de 26 proteínas como conjunto de prueba para los experimentos con el método propuesto. Se excluyeron las proteínas del conjunto base que no tienen registro en el Atlas de Sitios Catalíticos (CSA), tal como se muestra en la Tabla 10 con los registros que no tienen la etiqueta NA en la tercera columna. Las coordenadas tridimensionales de los átomos pertenecientes a las estructuras se obtuvieron del Protein Data Bank (PDB) y los residuos que conforman las cavidades se extrajeron

del servidor CASTp (Dundas *et al.*, 2006).

Se diseñó un método que constó de tres fases: *i)* Caracterización de las cavidades, *ii)* Almacenamiento de las características y *iii)* Comparación de cavidades mediante sus características. En la Figura 27 se ilustra un esquema del funcionamiento del método y las fases mencionadas.

i) Fase de caracterización. Esta fase del método cuenta a su vez de dos subfases:

1. Cálculo de pseudo-centros. En esta subfase se suministraron los archivos PDB al servidor CASTp para calcular los residuos superficiales que conforman sus cavidades. Posteriormente, por cada proteína se generó una nube de puntos en el espacio tridimensional a partir de los residuos que conformaron su cavidad mayor. Este procedimiento mapeó un conjunto de átomos representativos de los residuos superficiales a sus centros geométricos, los cuales se denominan pseudo-centros. Shulman-Peleg *et al.* (2005) describieron este modelo de relación residuo-pseudocentro y su interacción molecular asociada: enlace de hidrógeno donador (DO), enlace de hidrógeno receptor (AC), mixta receptor/donador (DA), alifática hidrofóbica y aromática (AL). En la Tabla 12 se muestran los pseudocentros correspondientes a cada tipo de aminoácido retomados del modelo de Shulman-Peleg (algunos tipos de aminoácidos tienen designados más de un pseudocentro).
2. Triangulación 3D. La teselación o seccionamiento de la nube de pseudocentros en tetraedros, y de triángulos en su envolvente convexo, se realizó con el algoritmo de Delaunay para 3D. El algoritmo de Delaunay ofrece una partición única con tetraedros adecuados, es decir, maximiza los ángulos internos más pequeños para obtener la mejor regularización en la construcción de los tetraedros (Berg *et al.*, 2008) (ver Sección 2.2.4.1).

   Concluida esta fase, la cavidad de una proteína se representa, ya sea por un conjunto de tetraedros en el que cada vértice es un pseudocentro con coordenadas en el espacio y un identificador que describe su tipo de interacción mole-

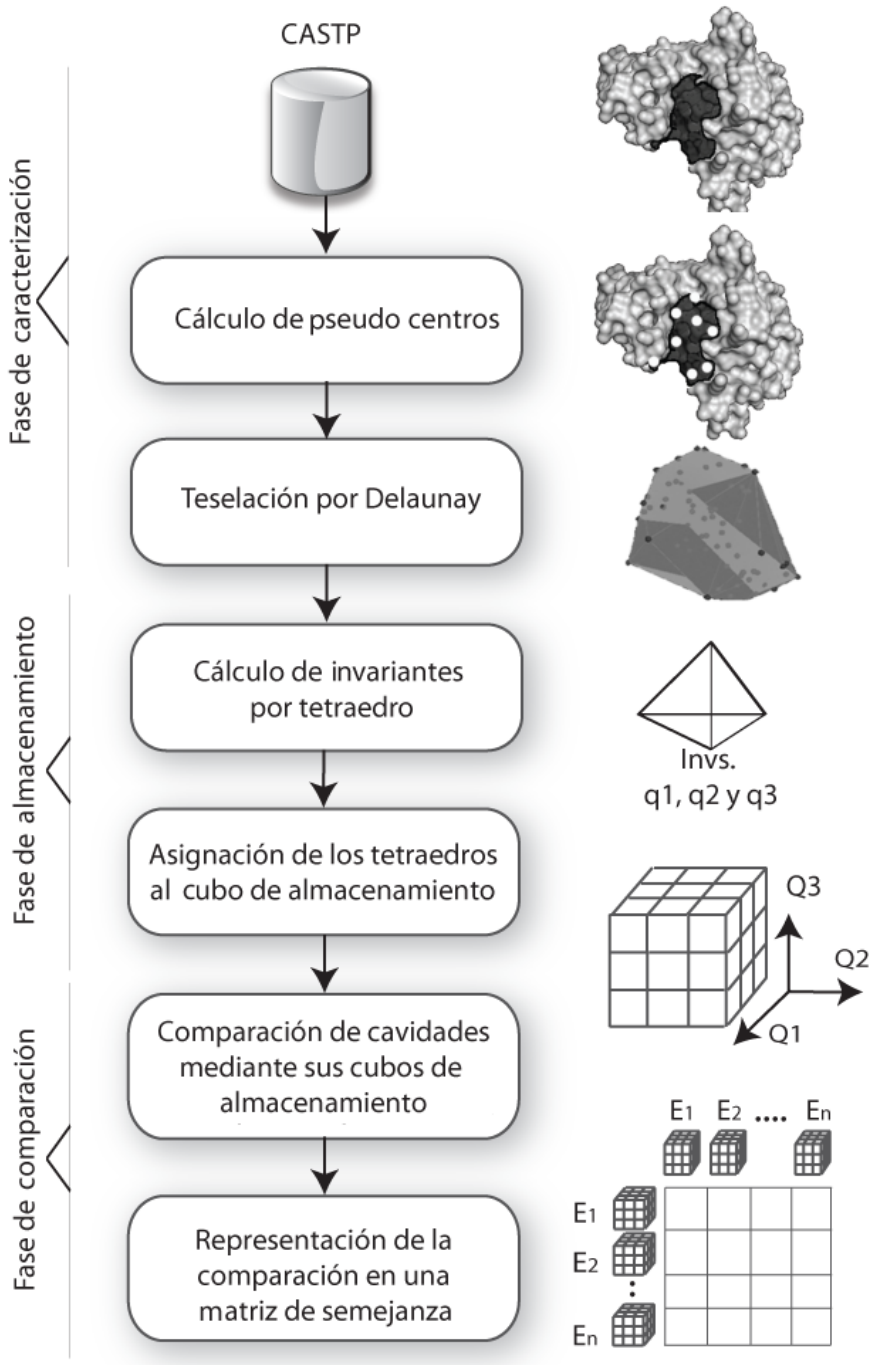


Figura 27: Diagrama del método de comparación de cavidades propuesto.

cular, o por una representación alterna constituida por el conjunto de triángulos que componen las caras de los tetraedros resultantes.

ii) Almacenamiento de las características. El proceso de almacenamiento de los tetraedros, o triángulos para el envolvente convexo, calculados en la fase anterior se

Tabla 12: Generación de pseudocentros a partir de la información tridimensional de los tipos de aminoácidos

Id. Aminoácido	Propiedad físico-química	Átomos representantes
ALA	AL	$C\beta$
ARG	AL	$C\beta$, $C\gamma$, $C\delta$
	DO	$N\epsilon$
	DO	$N\eta 1$
ASN	AC	$O\delta 1$
	DO	$N\delta 2$
ASP	AC	$O\delta 1$
	AC	$O\delta 2$
CYS	AL	$C\beta$, $S\gamma$
GLN	AC	$O\epsilon 1$
	DO	$N\epsilon 2$
GLU	AC	$O\epsilon 1$
	AC	$O\epsilon 2$
GLY		
HIS	PI	$C\gamma$, $N\delta 1$, $C\delta 2$, $C\epsilon 1$, $N\epsilon 2$
	DA	$N\delta 1$
	DA	$N\epsilon 2$
ILE	AL	$C\beta$, $C\gamma 1$, $C\gamma 2$, $C\delta 1$
LEU	AL	$C\beta$, $C\gamma$, $C\delta 1$, $C\delta 2$
LYS	AL	$C\beta$, $C\gamma$, $C\delta$, $C\epsilon$
	DO	$N\zeta 2$
MET	AL	$C\beta$, $C\gamma$, $S\delta$, $C\epsilon$
PHE	PI	$C\gamma$, $C\delta 1$, $C\delta' 2$, $C\epsilon 1$, $C\epsilon 2$, $C\zeta$
PRO	AL	$C\beta$, $C\gamma$, $C\delta$
SER	DA	$O\gamma$
THR	AL	$C\gamma 2$
	DA	$O\gamma 2$
TRP	PI	$C\gamma$, $C\delta 1$, $C\delta 2$, $N\epsilon 1$, $C\epsilon 2$, $C\epsilon 3$, $C\zeta$, $C\zeta 3$, $C\eta 2$
	DO	$N\epsilon 1$
TYR	PI	$C\gamma$, $C\delta' 1$, $C\delta' 2$, $C\epsilon 1$, $C\epsilon 2$, $C\zeta$
	DA	$O\eta$
VAL	AL	$C\beta$, $C\gamma 1$, $C\gamma 2$

llevó a cabo en dos partes:

1. Cálculo de invariantes. Por cada tetraedro de las cavidades a comparar se calculan tres medidas normalizadas entre [0,1], que se expresan en las siguientes ecuaciones (Knupp, 2001):

- $Q_1 = 3\frac{r}{R}$, r es el radio de la esfera inscrita y R el radio de la esfera circunscrita.

- $Q_2 = \frac{12(3V)^{\frac{2}{3}}}{S}$, que es el eigenvalor del tetraedro, donde V es su volumen y S es la suma de los cuadrados de las longitudes de sus aristas.

- $Q_3 = \alpha$, α es el ángulo sólido más pequeño.

Mientras que las medidas para los triángulos se calcularon con las siguientes ecuaciones:

- $QHC_1 = \frac{l}{L}$,

- $QHC_2 = \frac{lm}{L}$,

- $QHC_3 = \frac{-(l^2 - lm^2 - L^2)}{2l * lm}$,

donde L es el lado mayor, lm es el lado con valor intermedio y l es el lado más pequeño.

2. Asignación en el cubo hash. A cada tetraedro de la triangulación se le calculó una llave para determinar su ubicación en el cubo de almacenamiento asociado a la proteína a la que pertenece. El valor de la llave está dado por las invariantes del tetraedro respectivo, donde cada caso de las invariantes se mapea a uno de los D intervalos en los que se dividieron los ejes del cubo (ver Figura 28). El espacio designado almacenó la información correspondiente a los tres valores escalares que caracterizaron los rasgos geométricos del tetraedro y cuatro etiquetas de un alfabeto asociado a los tipos de interacción molecular de los pseudocentros que forman los vértices del tetraedro.

iii) Comparación de cavidades. En esta fase se realizaron las siguientes tareas:

1. Comparación de cubos hash. Dados dos cubos hash de almacenamiento E_i y E_j que se obtuvieron en la fase anterior, se contaron las coincidencias verda-

deras entre sus posiciones ocupadas. Una coincidencia verdadera es aquella en donde se encuentra ocupada la misma posición (x, y, z) en ambos cubos de almacenamiento y además las interacciones moleculares de $E_i(x, y, z)$ son iguales a las de $E_j(x, y, z)$, donde $x \in N$ y $1 \leq x \leq D$. El número total de coincidencias verdaderas entre E_i y E_j, para todos los valores posibles (x, y, z), se denominó $f_m(E_i, E_j)$.

Un enfoque semejante se usó en (Fischer *et al.*, 1995; Alesker *et al.*, 1996; Wallace *et al.*, 1997). En todos estos trabajos, se construyó un sistema coordenado local por cada residuo pero sólo las coordenadas de sus vecinos se transformaron para almacenarse en el hash geométrico. Entonces para un par de residuos (uno por cada estructura), sus respectivos vecinos espaciales se compararon y se hizo un conteo simple de aquellos que son similares en sus posiciones (en los respectivos sistemas coordenados locales).

2. Presentación de los resultados. La medida de semejanza entre dos cubos que contienen los tetraedros de la triangulación para una cavidad se expresó como:

$$S(E_i, E_j) = f_m(E_i, E_j)/min\{|E_i|, |E_j|\}, \tag{11}$$

donde $|E_i|$ y $|E_j|$, son los números de tetraedros en los cubos E_i y E_j, respectivamente. Todas las medidas de similitud entre los pares de las proteínas del conjunto de prueba se almacenaron en una matriz de similitud que se denomina *MSCav* y cuya representación gráfica asocia tonos de grises a los intervalos que discretizan sus valores. Al igual que en las matrices de similitud de las representaciones anteriores, los tonos más claros indican una mayor similitud entre los pares y los intervalos con los que se agrupan los valores no son regulares.

4.2.2. Prueba de concepto

Cuatro proteínas del grupo **AGC** (1CDK, 1O6L, 1OMW, 1H1W) se seleccionaron para modificar sus coordenadas y verificar la robustez del método con las representaciones

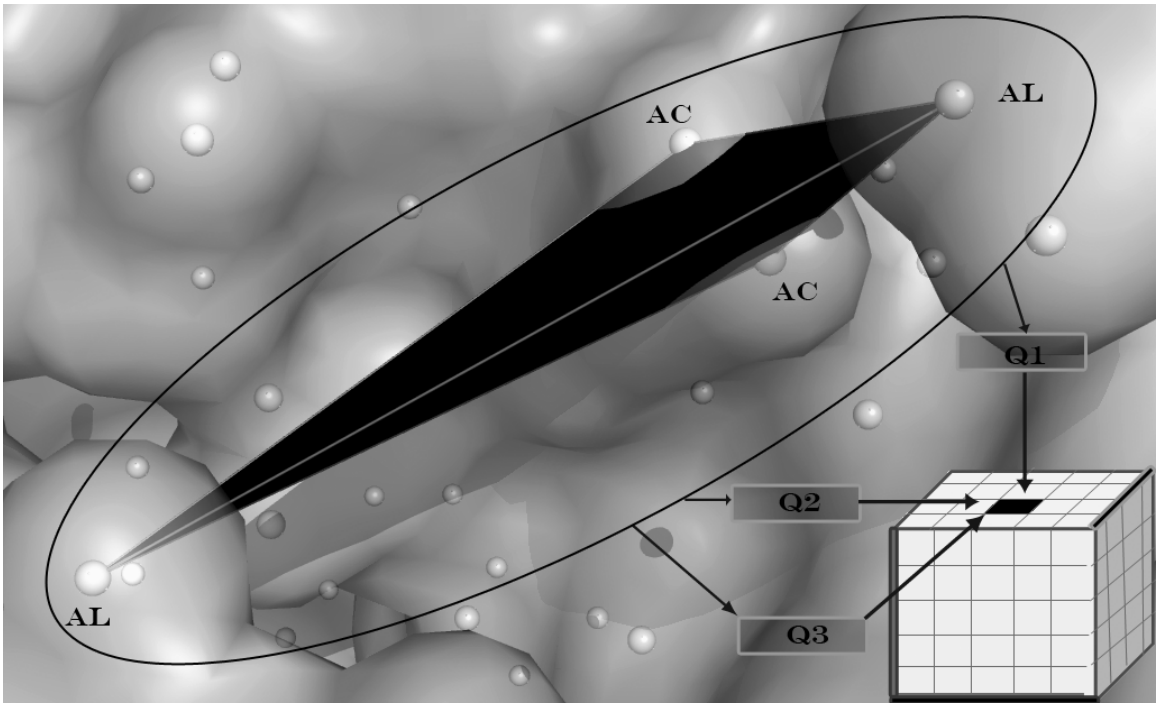


Figura 28: Extracción de características y almacenamiento de un tetraedro que conforma parte de la triangulación 3D de una cavidad. Los identificadores de las interacciones son: AC (enlace de hidrógeno receptor) y AL (alifática hidrofóbica); mientras que Q1, Q2 y Q3 representan los invariantes del tetraedro.

de tetraedros y triángulos del envolvente convexo. Cada una de estas proteínas se consideró como la estructura representativa de un grupo hipotético al que pertenecen otras dos estructuras derivadas de ellas. La información detallada se muestra en la Tabla 13.

Tabla 13: Conjunto de estructuras utilizadas en la prueba de concepto. Las proteínas derivadas se obtuvieron al aplicarles las transformaciones afines de rotación y traslación.

Id. Proteína	Nombre	Rotación (α, β, γ)	Traslación en Å ($\Delta x, \Delta y, \Delta z$)
1	1CDK	$(0, 0, 0)$	$(0, 0, 0)$
2	1CDK_R1	$(\pi/4, -\pi/2, 0)$	$(0, 0, 0)$
3	1CDK_R2	$(\pi/4, 0, -\pi/2)$	$(3, -4, 5)$
4	1O6L	$(0, 0, 0)$	$(0, 0, 0)$
5	1O6L_R1	$(0, -\pi/4, 0)$	$(-3, 12, -6)$
6	1O6L_R2	$(-\pi/4, 0, -\pi/2)$	$(0, 12, 0)$
7	1OMW	$(0, 0, 0)$	$(0, 0, 0)$
8	1OMW_R1	$(0, 0, 0)$	$(4, 12, -3)$
9	1OMW_R2	$(\pi/2, \pi/2, \pi/2)$	$(0, 0, -3)$
10	1H1W	$(0, 0, 0)$	$(0, 0, 0)$
11	1H1W_R1	$(0, \pi/3, 0)$	$(4, 12, -3)$
12	1H1W_R2	$(0, -\pi/3, 0)$	$(4, 12, -3)$

La división D de cada eje del cubo hash fue igual a 12, también se utilizó un factor de escalamiento ρ igual a 100 para minimizar los errores por redondeo en los cálculos de las invariantes. Dos niveles de ruido proveniente de una distribución uniforme se agregaron a las estructuras para evaluar la respuesta del método. El primero tuvo un intervalo de [-0.1 a 0.1] Å y el segundo de [-0.5 a 0.5] Å. Los resultados de estos experimentos mostraron una adecuada separación de grupos con el primer nivel de ruido añadido, pero casi nula con el segundo nivel, en especial con la representación de triángulos obtenidos de los envolventes convexos de los tetraedros (ver Figura 29).

4.2.3. Resultados y discusión

La Figura 30 muestra la matriz de similitud *MSCav*. Se distingue al grupo **CMGC** como aquel que presenta los valores más altos de similitud entre sus miembros, pero el que presenta valores más homogéneos es el grupo **AGC** al igual que el grupo **AK**, aunque este último sólo tiene dos elementos. En contraste, en el grupo **TK** no hay un predominio claro de valores en algún intervalo. Es notable que la proteína 1B6C tiene similitudes con valores altos con todas las demás proteínas, pero esta relación no es transitiva. La causa de este comportamiento radica en que la cantidad de residuos en la cavidad mayor de 1B6C es significativamente mayor que en todas las demás cavidades mayores de las otras proteínas (ver Tabla 10), y por lo tanto haya más tetraedros coincidentes y la métrica de la Ecuación 11 se incremente.
La Figura 31 muestra los diez tetraedros y triángulos con mayor frecuencia sobre todo el conjunto de prueba, de acuerdo a sus interacciones moleculares. Los pseudocentros con propiedades alifáticas (*AL*) y receptoras (*AC*) son predominantes en las combinaciones que forman las etiquetas de estos tetraedros y triángulos.

En esta sección se propuso un método de caracterización geométrica y físico-química de estructuras terciarias para comparar pares de cavidades mayores del grupo de prueba formado por proteínas de la superfamilia de las cinasas. En la representación de la matriz de similitud *MSCav* no se distingue una división clara entre las proteínas que pertenecen al mismo grupo funcional. Algunas de las razones por las que el método no logra los resultados esperados se enuncian a continuación:

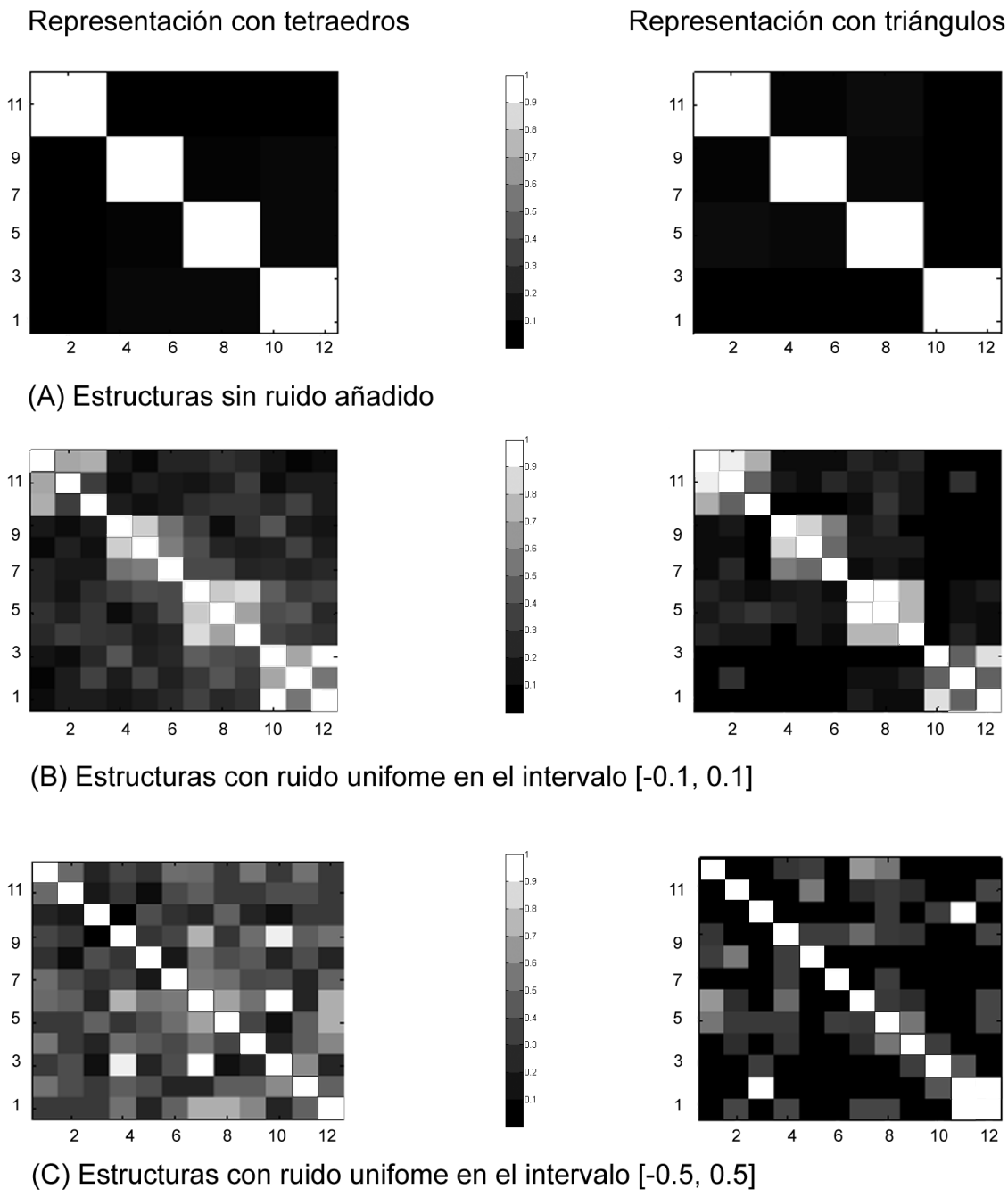


Figura 29: Matrices de semejanza de las estructuras de la Tabla 13 para prueba de concepto.

1. A menudo, los tetraedros y triángulos que conformaron los seccionamientos de las cavidades y sus envolventes convexos, respectivamente, no contienen todos los pseudocentros asociados a los residuos catalíticos, es decir, los residuos catalíticos quedan dispersos en diferentes tetraedros o quedan dentro del envolvente convexo

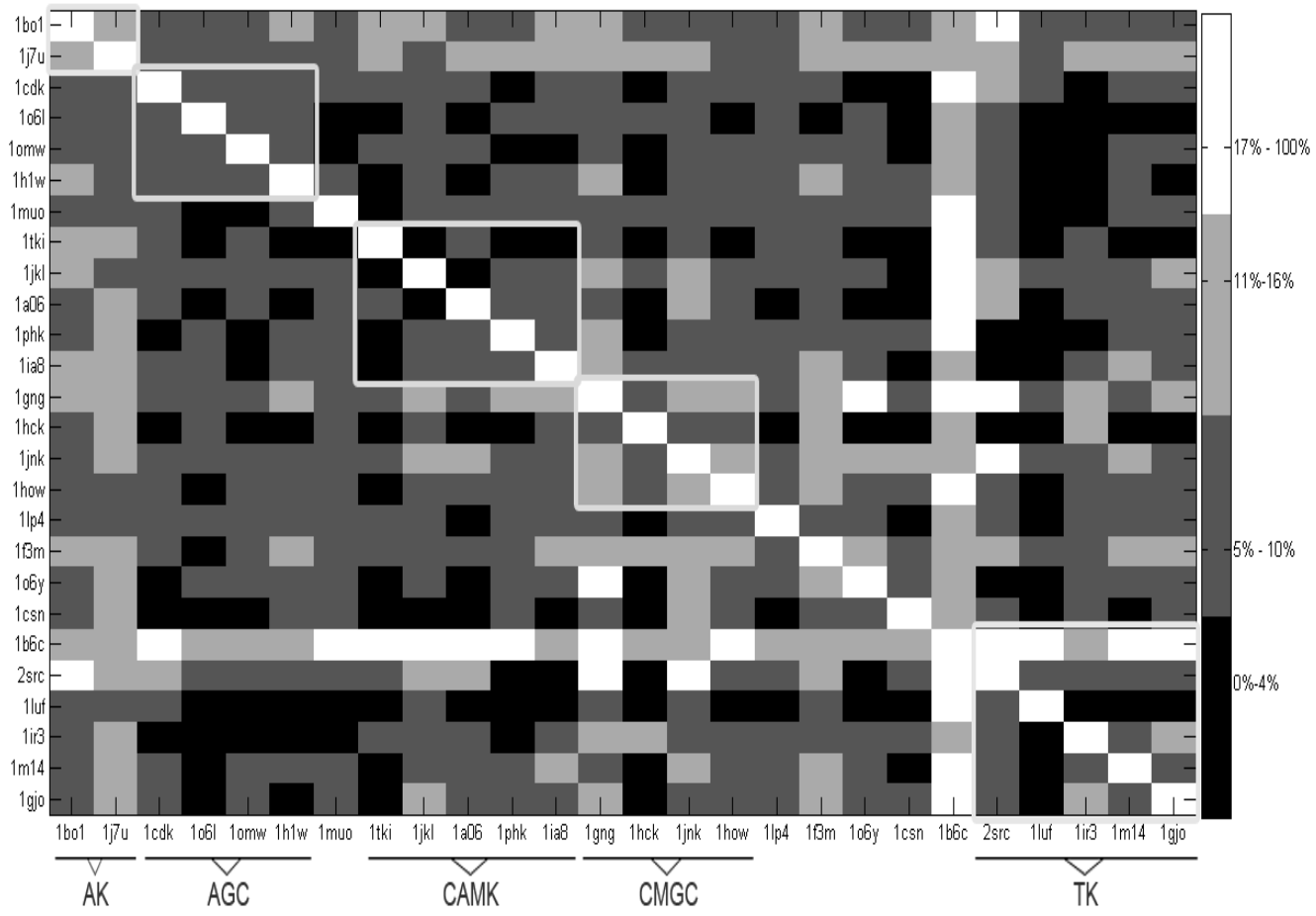


Figura 30: Matriz de similitud *MSCav* obtenida con el método de comparación de tetraedros para las cavidades mayores de las proteínas del grupo de prueba (ver Tabla 9).

de la cavidad, el cual está formado por triángulos. Aunque geométricamente la triangulación 3D de Delauny ofrece características deseables para la caracterización y comparación de cavidades, ésta pocas veces resultó en tetraedros que agruparon los puntos relacionados con la actividad catalítica de la cavidad. Además, frecuentemente el algoritmo creó tetraedros cuyos vértices tenían una separación mayor a la que tienen los residuos catalíticos.

2. La diversidad en la forma de las cavidades hace que el método no detecte adecuadamente regularidades en los tetraedros que se adaptan a la cavidad.

3. La actividad catalítica de algunas proteínas se presentan a menudo en sus cavidades tanto superficiales como internas, en ellas se han identificado residuos con una interacción evidente como la tríada Ser/His/Asp; sin embargo, hay agrupaciones cu-

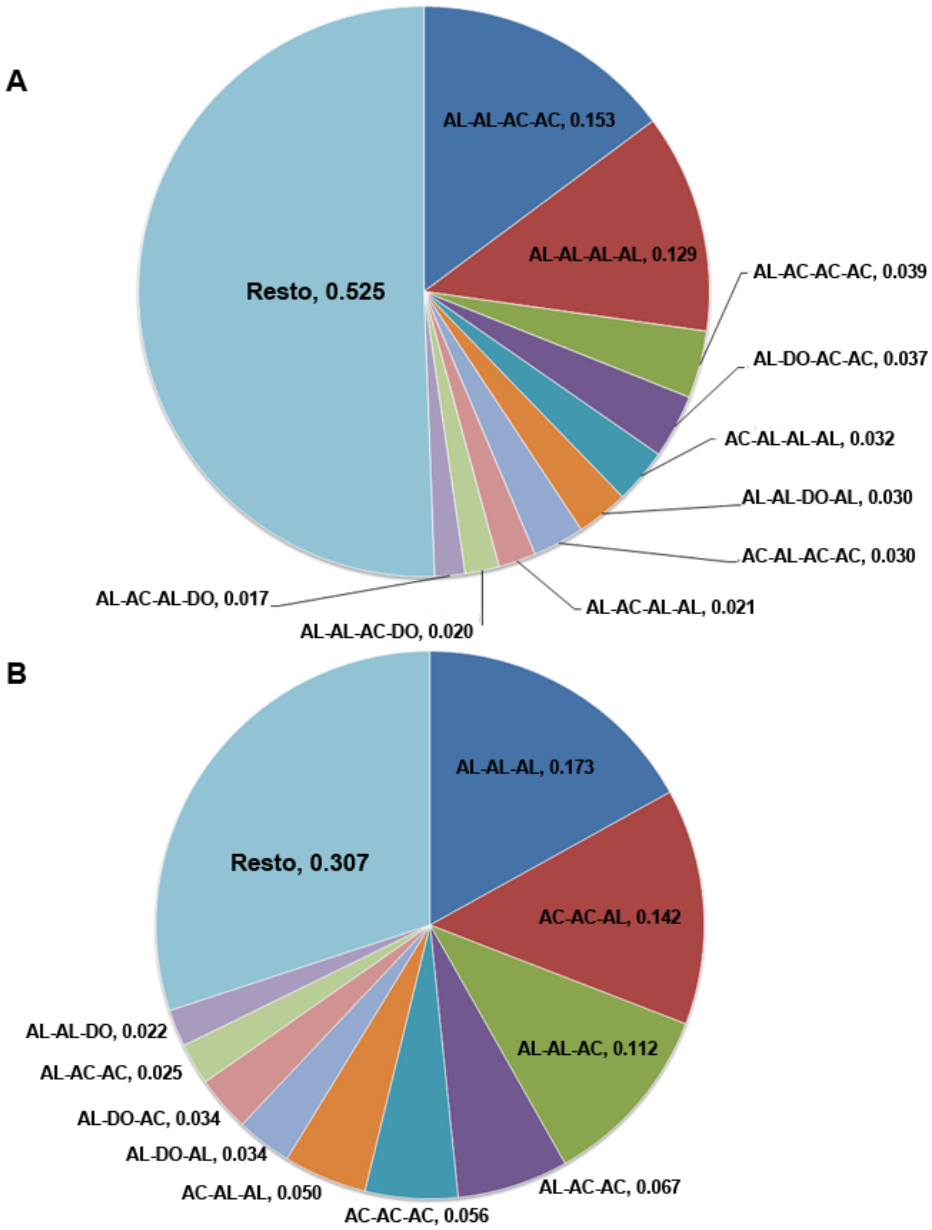


Figura 31: Proporción de tetraedros y triángulos en el conjunto de prueba utilizado en el método de caracterización y clasificación propuesto. La proporción está dada con base en sus propiedades de interacción molecular. (A) Proporción por triángulos. (B) Proporción por tetraedros.

yos residuos aparentemente no tienen un papel clave, pero las cavidades a las que pertenecen los requieren para sus actividades funcionales, tal es el caso de algunas aminopeptidasas o N-acetiltransferasas (Binkowski y Joachimiak, 2008).

4. A pesar de que se redujo la región a caracterizar, el conjunto de residuos claves en la función que realiza la proteína puede estar distribuido en más de una cavidad.

Pese a que el método propuesto no logró agrupar correctamente las proteínas del conjunto de prueba en sus grupos funcionales; el principio de conservación de residuos catalíticos ante la presión evolutiva, proporciona una base eficaz para seguir desarrollando métodos para la predecición de funciones. Un algoritmo reciente, denominado TIPSA (Triangulation-based Iterative-closest-point for Protein Surface Alignment) (Leif y Jinfeng, 2012), incorpora una búsqueda de tetraedros similares entre dos sitios catalíticos, caracterizados mediante una triangulación de Delaunay en 3D. Este método usa el algoritmo Húngaro para hallar átomos coincidentes que se almacenan como patrones para posteriormente hacer comparaciones con regiones de proteínas de consulta. El algoritmo alcanzó un rendimiento comparable al de los mejores métodos en la literatura, además de que proporciona el conjunto de átomos comunes de forma jerárquica. A diferencia del método propuesto, el algoritmo TIPSA primero hace un alineamiento de la estructura terciaria de forma global y posteriormente alinea iterativamente los átomos de las cavidades (enfoque local) para generar pseudocentros que representan los residuos con mayor conservación, esto le da una ventaja sustancial sobre el método que se desarrolló en el presente proyecto.

Capítulo 5. Conclusiones

Las tareas de caracterización, comparación y clasificación de proteínas que contribuyen a la inferencia de sus funciones, tomando como base sus estructuras terciarias, requiere un conocimiento de distintas disciplinas como: biología molecular, biología estructural, bioinformática, ciencias de la computación, por mencionar las más relevantes (Pevsner, 2005). La falta de estandarización en los datos de muchos repositorios agrega una mayor dificultad para la realización de experimentos *in silico*, los cuales a menudo están formados por una serie de procedimientos, en donde cada fase adquiere datos de distintos repositorios. En el presente capítulo se resaltan los aspectos más notables en este proyecto de investigación con respecto a los resultados obtenidos en las tareas mencionadas previamente, posteriormente se enuncian las conclusiones principales y finalmente una perspectiva del trabajo a futuro.

5.1. Sumario

El fundamento de la bioinformática estructural que establece una relación directa entre la forma de una proteína y la función que realiza, así como las estrategias de predicción utilizadas en los métodos computacionales para establecer esta relación de forma automatizada, en especial para la detección de residuos catalíticos, respaldaron la investigación realizada en esta tesis y la cual cubrió dos aspectos:

1. **Una ampliación al método CMASA**. Este método (Li y Huang, 2010), realiza predicciones de residuos catalíticos con un alto valor en sus medidas de desempeño, tales como: exactitud, precisión y coeficiente de correlación de Matthews. Sin embargo, CMASA se ve seriamente afectado cuando existen mutaciones en un residuo del sitio catalítico de la proteína de consulta. Para superar esta dificultad, se propuso una extensión donde por cada plantilla de su base de datos de sitios catalíticos, se derivaron nuevas plantillas realizando todas las combinaciones de *n-1* residuos para cada sitio de *n* residuos.

2. **Un método para la caracterización de estructuras terciarias de proteínas y su identificación funcional**. Se propuso un método usando descriptores con forma de tetraedros; cuyos vértices representaron propiedades geométricas y de interacción

molecular asociadas a los aminoácidos que componen la superficie de determinada cavidad perteneciente a una proteína; mientras que las aristas de los tetraedros indicaron la distancia euclidiana entre los átomos representativos de los aminoácidos. La comparación de cavidades se realizó agrupando a las proteínas con un mayor número de descriptores (tetraedros) similares mediante la técnica de *hashing geométrico*.

5.2. Conclusiones

A partir de los resultados obtenidos en este proyecto de investigación, se enuncian las principales conclusiones para cada uno de los objetivos planteados.

1. Con respecto a la extensión del método CMASA:

   - Es capaz de identificar los residuos no mutados de aquellos sitios catalíticos que presentan una mutación en alguno de sus aminoácidos o cuando faltan átomos relevantes en alguno de ellos.
   - Mantuvo el poder de predicción del método original cuando no hay mutaciones en los sitios catalíticos potenciales .
   - El número de comparaciones a realizar es significativamente menor al que se requiere con matrices de sustitución, mecanismo propuesto en CMASA para atacar las mutaciones.
   - Es fácilmente escalable para dobles o triples mutantes, aunque estas situaciones son excepcionales en casos reales y generan un costo computacional mayor.

   Una de las características a resaltar es que las combinaciones de n-1 residuos permite hacer el método más sensible, razón por la cual genera más falsos positivos, para corregir este comportamiento se requiere una mayor exigencia en las medidas de similitud, en especial con el CMAD (Desviación Media de la Matriz de Contacto).

2. Con respecto al método propuesto:

   - Las agrupaciones obtenidas para el caso de estudio (cavidades mayores pertenecientes a proteínas de la superfamilia de las cinasas) tuvieron una menor exactitud a las reportadas en Gramada y Bourne (2006) y aún menor a los calculados con alineamientos locales de sus estructuras primarias.
   - En las proteínas del grupo de prueba, las cavidades mayores tienen topologías marcadamente distintas, por lo que los descriptores que se propusieron para estructuras terciarias locales, cuya base son tetraedros, no detectaron regularidades subyacentes a pesar de las cualidades geométricas que ofrece el método de Delaunay para las triangulaciones que caracterizaron las cavidades.
   - Los residuos claves en los sitios catalíticos no siempre están contenidos en la cavidad mayor. Un aspecto que complica aún más la detección de sitios catalíticos, es la variación en la forma de las cavidades que se obtienen con el método CASTp (Dundas *et al*., 2006), debido a que pequeños cambios en el parámetro del radio de la esfera de prueba producen cambios significativos en la definición de la forma de las cavidades. La esfera de prueba representa una molécula de solvente que ayuda a determinar cuáles son los aminoácidos que forman la superficie de la cavidad.
   - No obstante que las interacciones moleculares preservan características importantes de los residuos catalíticos; en la clasificación implementada en esta investigación no lograron ser un factor preponderante para la separación adecuada de los grupos funcionales.

   Es importante mencionar que la caracterización a través de información geométrica, así como de interacciones moleculares de los aminoácidos pertenecientes a la superficie de las cavidades, parecen no ser suficientes para identificar automáticamente la función de una proteína si no se incorpora información previa que sirva como guía de búsqueda de patrones, un ejemplo es un grupo de plantillas de patrones.

5.3. Trabajo a futuro

Dado que el método propuesto no utilizó información *a priori* de patrones que permitieran afinar en las tareas de comparación y clasificación; una adecuación viable para incrementar la exactitud del método sería construir un conjunto de plantillas, cuya información permita compararlas con los descriptores equivalentes en las cavidades de consulta. Estas plantillas deberían construirse utilizando recursos como el Atlas de Sitios Catalíticos (CSA) (Furnham *et al.*, 2014) tal como lo hace CMASA para la selección de los residuos que conforman sus plantillas; otra información relevante, cuya incorporación debe considerarse, son las interacciones moleculares de los residuos catalíticos o de sus vecinos inmediatos que formen parte de la superficie.

Además de la construcción de una base de datos de plantillas, se pueden comparar por pares las estructuras terciarias de las proteínas utilizando un enfoque mixto, es decir, realizar un primer alineamiento burdo de las cadenas principales de sus estructuras globales y posteriormente identificar las cavidades con mayor semejanza para calcular los descriptores y una medida de similitud con las plantillas disponibles en la base de datos, de esta forma se agruparán aquellas proteínas cuyas cavidades (estructuras locales) tienen mayor semejanza con determinada plantilla. Plantillas de sitios catalíticos que incluyan información de sus diversas trayectorias evolutivas o que consideren fluctuaciones posibles en los procesos de captura también se pueden construir en esta fase, tal como se hizo en Leif y Jinfeng (2012).

La inferencia de funciones de proteínas aún tiene reservados muchos retos, algunos de los cuales se han ido resolviendo conforme surge nueva información biológica, auxiliada por algoritmos y herramientas computacionales.

Lista de referencias

Aghili, S. A., Agrawal, D., y Abbadi, A. E. (2005). Pads: Protein structure alignment using directional shape signatures. En: L. Zhou, B. C. Ooi, y X. Meng (eds.), *DASFAA*. Springer, Vol. 3453 de *Lecture Notes in Computer Science*, pp. 17–29.

Alesker, V., Nussinov, R., y Wolfson, H. J. (1996). Detection of non-topological motifs in protein structures. *Protein Engineering*, **9**(12): 1103–1119.

Altschul, S., Gish, W., Miller, W., Myers, E., y Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403–410.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., y Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17): 3389–3402.

Artymiuk, P., Poirrette, A., Grindley, H., Rice, D., y Willet, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structure. *Journal of Molecular Biology*, **243**(2): 327–344.

Aurenhammer, F. (1991). Voronoi diagrams - a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, **23**(3): 345–405.

Ausiello, G., Via, A., y Helmer-Citterich, M. (2005). Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6**(4): S5.

Ausiello, G., Gherardini, P., Marcatili, P., Tramontano, A., Via, A., y Helmer-Citterich, M. (2008). Funclust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics*, **9**(2).

Barker, J. y Thornton, J. (2003). An algorithm for constraint-based structural template matching: application to 3d templates with statistical analysis. *Bioinformatics*, **19**(13): 1644–1649.

Bartlett, G. J., Annabel E., T., y Thornton, J. M. (2005). *Inferring Protein Function from Structure*, pp. 387–407. John Wiley and sons, Inc.

Baxevanis, A. D. y Oullette, Francis, B. F. (2001). *Bioinformatics (A practical guide to the analysis of genes and proteins)*. John Wiley and Sons. 187 pp.

Berg, M., Cheong, O., van Kreveld, M., y M., O. (2008). *Computational Geometry - Algorithms and Applications*. Springer (3a. ed.). Berlin. 191 pp.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., y Bourne, P. E. (2000). The protein data bank. **28**(1): 235–242.

Binkowski, T. A. y Joachimiak, A. (2008). Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. *BMC structural biology*, **8**(1): 45.

Binkowski, T. A., Freeman, P., y Liang, J. (2004). pvsoar: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Research*, **32**(suppl 2): W555–W558.

Bourne, P. y Weissig, H. (2003). *Structural bioinformatics*. John Wiley and Sons. New Jersey. 354 pp.

Branden, C. y Tooze, J. (1998). *Introduction to Protein Structure*. Garland (2a. ed.). Estocolmo, Suecia. 373 pp.

Bromberg, Y. y Rost, B. (2008). Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, **24**(16): i207–i212.

Bron, C. y Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph. *Communications of the ACM*, **16**(9): 575–579.

Brown, N. P., Orengo, C. A., y Taylor, W. R. (1996). A protein structure comparison methodology. *Computers and Chemistry*, **20**(3): 359–380.

Budowski-Tal, I., Nov, Y., y Kolodny, R. (2010). Fragbag, an accurate representation of protein structure, retrieves structural neighbors from the entire pdb quickly and accurately. *Proceedings of the National Academy of Sciences*, **79**(8): 3481–3486.

Cai, X.-H., Jaroszewski, L., Wooley, J., y Godzik, A. (2011). Internal organization of large protein families: Relationship between the sequence, structure, and function-based clustering. *Proteins: Structure, Function, and Bioinformatics*, **79**(8).

Chandonia, J.-M. y Brenner, S. E. (2006). The impact of structural genomics: Expectations and outcomes. *Science*, **311**(5759): 347–351.

Cheftel, J.-C., Cuq, J.-L., y Lorient, D. (1989). *Proteínas alimentarias: bioquímica, propiedades funcionales, valor nutricional, modificaciones químicas*. Acribia.

Chew, L. P., Huttenlocher, D., Kedem, K., y Kleinberg, J. (1999). Fast detection of common geometric substructure in proteins. *Journal of Computational Biology*, **6**(3-4): 313–325.

Chung, S. Y. y Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**(10): 1123 – 1127.

Collier, J. H., Allison, L., Lesk, A. M., Garcia de la Banda, M., y Konagurthu, A. S. (2014). A new statistical framework to assess structural alignment quality using information compression. *Bioinformatics*, **30**(17): i512–i518.

Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, **32**(suppl 1): D258–D261.

Consortium, T. U. (2007). The universal protein resource (uniprot). *Nucleic Acids Research*, **35**(suppl 1): D193–D197.

Consortium, T. U. (2014). Activities at the universal protein resource (uniprot). *Nucleic Acids Research*, **42**(D1): D191–D198.

Copley, S. (2012). Moonlight is mainstream: Paradigm adjustment required. *BioEssays*, **34**: 578–588.

Crippen, G. M. (1978). The tree structural organization of proteins. *Journal of Molecular Biology*, **126**(3): 315–332.

de Beer, T. A. P., Berka, K., Thornton, J. M., y Laskowski, R. A. (2014). Pdbsum additions. *Nucleic Acids Research*, **42**(D1): D292–D296.

de Berg, M., Cheong, O., van Kreveld, M., y Overmars, M. (2008). *Computational Geometry (Algorithms and Applications)*. Springer (3a. ed.), tercera edición. 196 pp.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., y Liang, J. (2006). Castp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research*, **34**(suppl 2): W116–W118.

Dundas, J., Adamian, L., y Liang, J. (2011). Structural signatures of enzyme binding pockets from order-independent surface alignment: A study of metalloendopeptidase and nad binding proteins. *Journal of Molecular Biology*, **406**(5): 713–729.

Eidhammer, I., Jonassen, I., y Taylor, W. R. (2000). Structure comparison and structure patterns. *Journal of Computational Biology*, **7**(5): 685 – 716.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., y Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, **42**(D1): D222–D230.

Fischer, D., Bachar, O., Nussinov, R., y Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *Journal of Biomolecular Structure and Dynamics*, **9**(4): 769–789.

Fischer, D., Tsai, C.-J., Nussinov, R., y Wolfson, H. (1995). A 3d sequence-independent representation of the protein data bank. *Protein Engineering*, **8**(10): 981–997.

Fontana, P.and Bindewald, E., Toppo, S., Velasco, R., Valle, G., y Tosatto, S. (2005). The ssea server for protein secondary structure alignment. *Bioinformatics*, **21**(3): 393–395.

Furnham, N., Holliday, G. L., de Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., y Thornton, J. M. (2014). The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, **42**(D1): D485–D489.

Gart, J. y Buck, A. (1966). Comparison of a screening test and a reference test in epidemiologic studies: Ii. a probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, **83**(3): 593–602.

Godzik, A. (1996). The structural alignment between two proteins: Is there a unique answer? *Protein Science*, **5**(7): 1325–1338.

Gold, N. D. y Jackson, R. M. (2006). Fold independent structural comparisons of protein ligand binding sites for exploring functional relationships. *Journal of Molecular Biology*, **355**(5): 1112 – 1124.

Goodman, J. y O'Rourke, J. (2004). *Handbook of Discrete and Computational Geometry, Second Edition (Discrete and Combinatorial Mathematics Series)*. Chapman and Hall/CRC (2a. ed.), segunda edición. 48 pp.

Gramada, A. y Bourne, P. (2006). Multipolar representation of protein structure. *BMC bioinformatics*, **7**(10): 242.

Grindley, H. M., Artymiuk, P. J., Rice, D. W., y Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, **229**(3): 707–721.

Guex, N. y Peitsch, M. (1997). Swiss-model and the swiss-pdbviewer: An environment for comparative protein modeling. *Electrophoresis*, **18**(15): 2714–2723.

Hasegawa, H. y Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current opinion in structural biology*, **19**(3): 341–348.

Henikoff, S. y Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, **89**(22): 10915–10919.

Hershey, J. y Olsen, P. (2007). Approximating the kullback leibler divergence between gaussian mixture models. En: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4, pp. IV–317–IV–320.

Holm, L. y Sander, C. (1994). Searching protein structure databases has come of age. *Proteins: Structure, Function, and Bioinformatics*, **19**(3): 165–173.

Holm, L., S. C. (1996). Mapping the protein universe. *Science*, **273**(5275): 595–602.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E., y Chothia, C. (1997). Scop: a structural classification of proteins database. *Nucleic Acids Research*, **25**(1): 236–239.

Jambon, M., Imberty, A., Deléage, G., y Geourjon, C. (2003). A new bioinformatic approach to detect common 3d sites in protein structures. *Proteins: Structure and Function and and Genetics*, **52**(2): 137–145.

Jonassen, I., Eidhammer, I., y Taylor, W. R. (1999). Discovery of local packing motifs in protein structures. *Proteins: Structure, Function, and Bioinformatics*, **34**(2): 206–219.

Kinoshita, K. y Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. *Protein Science*, **12**(8): 1589–1595.

Knupp, P. (2001). Algebraic mesh quality metrics. *SIAM Journal on Scientific Computing*, **23**(1): 193–218.

Koehl, P. (2006). *Protein Structure Classification*, pp. 1–55. John Wiley and Sons, Inc.

Kolodny, R., Koehl, P., y Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of molecular biology*, **346**(4): 1173–88.

Konagurthu, A., Stuckey, P. J., y Lesk, A. M. (2008). Structural search and retrieval using a tableau representation of protein folding patterns. *Bioinformatics*, **24**(5): 645–651.

Krissinel, E. y Henrick, K. (2004). Secondary-structure matching (pdbefold) and a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica*, **D60**(12): 2256–2268.

Kullback, S. (1959). *Information theory and statistics*. John Wiley and Sons.

Laskowski, R., Luscombe, N. M., Swindells, M. B., y Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Science*, **5**(12): 2438–52.

Laskowski, R. A. (1995). Surfnet: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, **13**(5): 323 – 330.

Laskowski, R. A., Watson, J. D., y Thornton, J. M. (2005). Protein function prediction using local 3d templates. *Journal of Molecular Biology*, **351**(3): 614–626.

Lehninger, A. L. (1995). *Principles of Biochemistry*. Worth (2a. ed.).

Leif, E. y Jinfeng, Z. (2012). Protein surface matching by combining local and global geometric information. *PLoS ONE*, **7**(7): e40540.

Li, G. y Huang, J. (2010). Cmasa: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation. *BMC Bioinformatics*, **11**(439): 1–13.

Liisa, H. y Chris, S. (1995). Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, **20**(11): 478–480.

Ma, J. y Wang, S. (2014). *Chapter Five - Algorithms, Applications, and Challenges of Protein Structure Alignment*, Vol. 94, pp. 121–175. Academic Press.

Manning, G., Plowman, G., Hunter, T., y Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. *Current opinion in structural biology*, **27**(10): 514–20.

Mariani, V., Biasini, M., Barbato, A., y Schwede, T. (2013). lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21): 2722–2728.

MATLAB (2010). *version 7.10.0 (R2010a)*. The MathWorks Inc. Natick, Massachusetts.

Matsuda, H., Taniguchi, F., y Hashimoto, A. (1997). An approach to detection of protein structural motifs using an encoding scheme of backbone conformations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 280–291.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta - Protein Structure*, **405**: 442–451.

Milik, M., Szalma, S., y Olszewski, K. (2003). Common structural cliques: a tool for protein structure and function analysis. *Protein Engineering, Design and Selection*, **16**(8): 543–552.

Needleman, S. B. y Christian, W. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3): 443–453.

Nilmeier, J., Kirshner, D., Wong, S., y Lightstone, F. (2013). Rapid catalytic template searching as an enzyme function prediction procedure. *PLoS ONE*, **8**(5): e62535.

Nussinov, R. y Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences*, **88**(23): 10495–10499.

Ondrechen, M., Clifton, J., y Ringe, D. (2001). Thematics: A simple computational predictor of enzyme function from structure. *Proceedings of the National Academy of Sciences USA*, **98**(22): 12473–12478.

Pan, Y., Wang, J., y Li, M. (2014). *Algorithmic Methodologies for Discovery of Nonsequential Protein Structure Similarities*. Wiley-IEEE Press. 16 pp.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., , y North, A. C. T. (1960). Structure of myoglobi: a three-dimensional fourier synthesis at 5.5 angstrom resolution, obtaindes by x-ray analysis. *Nature*, **185**: 416–422.

Pevsner, J. (2005). *Bioinformatics and Functional Genomics*. Wiley-Blackwell (2a. ed.). 48 pp.

Phan, J., Mahdavian, E., Nivens, M., Berger, S., Spencer, H., Dunlap, R., y Lebioda, L. (2000). Catalytic cysteine of thymidylate synthase is activated upon substrate binding. *Biochemstry*, **39**(23): 6969–78.

Porter, C., Bartlett, G., y Thornton, J. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, **32**: D129–D133.

Rahimi, A., Madadkar-Sobhani, A., Touserkani, R., y Goliaei, B. (2013). Efficacy of function specific 3d-motifs in enzyme classification according to their ec-numbers. *Journal of Theoretical Biology*, **336**(0): 36–43.

Rost, B., Liu, J., Nair, R., Wrzszczymski, O., y Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, **60**: 2637–2650.

Russell, R. B., Copley, R. R., y Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *Journal of Molecular Biology*, **259**(3): 349–365.

Scheeff, E. y Bourne, P. (2005). Structural evolution of the protein kinase-like superfamily. *PLoS computational biology*, **1**(5): e49.

Schmitt, S., Kuhn, D., y Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology*, **323**(2): 387–406.

Schrödinger, L. (2010). The PyMOL molecular graphics system, version 1.3r1. there is no corresponding record for this reference.

Shulman-Peleg, A. (2008). *Algorithms for the Detection and Prediction of 3D Proteing Binding Patterns and Interactions*. Tesis de doctorado, Tel Aviv University, The Raymond and Beverly Sackler Faculty of Exact Sciences, The Blavatnik School of Computer Science.

Shulman-Peleg, A., Nussinov, R., y Wolfson, H. (2004). Recognition of functional sites in protein structures. *Journal of Molecular Biology*, **339**(3): 607–633.

Shulman-Peleg, A., Nussinov, R., y H., W. (2005). Siteengines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic acids research*, **33**(suppl. 2): W337–W341.

Sigrist, C., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., A, B., y P, B. (2002). Prosite: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, **3**(3): 265–274.

Silla, C. N. J. y Freitas, A. A. (2011). Selecting different protein representations and classification algorithms in hierarchical protein function prediction. *Intelligent Data Analysis*, **15**(6): 979–999.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M., y Orengo, C. A. (2013). New functional families (funfams) in cath to improve the mapping of conserved functional sites to 3d structures. *Nucleic Acids Research*, **41**(D1): D490–D498.

Slater, A. W., Castellanos, J. I., Sippl, M. J., y Melo, F. (2013). Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. *Bioinformatics*, **29**(1): 47–53.

Smith, T. y Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1): 195–97.

Spriggs, R., Artymiuk, P., y Willett, P. (2003). Searching for patterns of amino acids in 3d protein structures. *Journal of Chemical Information and Computer Sciences*, **43**(2): 412–421.

Standley, D. M., Toh, H., y Nakamura, H. (2004). Detecting local structural similarity in proteins by maximizing number of equivalent residues. *Proteins: Structure, Function, and Bioinformatics*, **57**(2): 381–391.

Stark, A. y Russell, R. (2003). Annotation in three dimensions. pints: Patterns in non-homologous tertiary structures. *Nucleic Acids Research*, **31**(13): 3341–3344.

Stark, A., Sunyaev, S., y Russel, R. (2003). A model for statistical significance of local similarities in structure. *Journal of Molecular Biology*, **326**(5): 1307–1316.

Szustakowski, J. D. y Weng, Z. (2000). Protein structure alignment using a genetic algorithm. *Proteins: Structure, Function, and Bioinformatics*, **38**(4): 428–440.

Taylor, W. R. y Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology*, **208**(1): 1–22.

Teyra, J., Paszkowski-Rogacz, M., Anders, G., y Pisabarro, M. T. (2008). Scowlp classification: Structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, **9**(1): 9.

Thompson, E. E., Kornev, A. P., Kannan, N., Kim, C., Ten Eyck, L., y Taylor, S. (2009). Comparative surface geometry of the protein kinase family. *Protein Science*, **18**(10): 2016–26.

Tramontano, A. (2005). *The Ten Most Wanted Solutions in Protein Bioinformatics*. CRC Press. 5 pp.

Tseng, Y., Dundas, J., y Liang, J. (2009). Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *Journal of Molecular Biology*, **387**(2): 451–464.

Ullmann, J. (1976). An algorithm for subgraph isomorphism. *Journal of the Association for Computing Machinery*, **23**(1): 31–42.

Veeramalai, M. y Gilbert, D. (2008). A novel method for comparing topological models of protein structures enhanced with ligand information. *Bioinformatics*, **24**(23): 2698–2705.

Vesely, P. (2004). Molecular biology of the cell. by bruce alberts, alexander johnson, julian lewis, martin raff, keith roberts and peter walter. *Scanning*, **26**(3): 153–153.

Voet D, V. J. (1990). *Biochemistry*. John Wiley.

Volkamer, A., Kuhn, D., Rippmann, F., , y Rarey, M. (2013). Predicting enzymatic function from global binding site descriptors. *Proteins: Structure and Function and and Bioinformatics*, **81**(3): 479–489.

Volkert, L. y Staffer, D. (2004). A comparison of sequence alignment algorithms for measuring secondary structure similarity. En: *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. pp. 182–189.

Vuori, K., Myllyla, R., Pihlajaniemi, T., y Kivirikko, K. (1992). Expresion and site-directed mutagenesis of human protein disulfide isomerase in escherichia coli. *The Journal of Biological Chemistry*, **267**(11): 7211–7214.

Wallace, A., Laskowski, R., y Thornton, J. (1996). Derivation of 3d coordinate templates for searching structural databases: Application to the ser-his-asp catalytic triads of the serine proteinases and lipases. *Protein Science*, **5**(11): 1001–1013.

Wallace, A., Borkakoti, N., y Thornton, J. (1997). Tess: a geometric hashing algorithm for deriving 3d coordinate templates for searching structural databases. application to enzyme active sites. *Protein Science*, **6**(11): 2308–2323.

Wangikar, P., Tendulkar, A., Ramya, S., Mali, D., y Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *Journal of Molecular Biology*, **326**(3): 955–978.

Wolfson, H. J., Shatsky, M., Schneidman-Duhovny, D., Dror, O., Shulman-Peleg, A., Ma, B., y Nussinov, R. (2005). From structure to function: methods and applications. *Current protein & peptide science*, **6**(2): 171–83.

Wright, P. E. y Dyson, H. (1999). Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *Journal of Molecular Biology*, **293**(2): 321 – 331.

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S.-M., y Eisenberg, D. (2002). Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, **30**(1): 303–305.

Yahalom, R., Reshef, D., Wiener, A., Frankel, S., Kalisman, N., Lerner, B., y Keasar, C. (2011). Structure-based identification of catalytic residues. *Proteins: Structure and Function and Bioinformatics*, **79**(6): 1952–1963.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., y Marek (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology*, **326**(1): 255–261.

Zhu, J. y Weng, Z. (2005). Fast: A novel protein structure alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, **58**(3): 618–627.

Apéndice A. Comparaciones con representaciones de estructuras de dimensión 0

La más simple representación estructural de proteínas tiene como base la frecuencia de determinadas características, las cuales se representan en vectores. Eventualmente, a estas representaciones se les denomina *estructuras en dimensión 0* o *huellas estructurales*, las cuales también se usan para búsquedas dentro de grandes bases de datos como el Protein Data Bank (Hasegawa y Holm, 2009).

Varios métodos que utilizan la representación de estructuras en dimensión 0 se han propuesto para clasificar proteínas en familias que comparten la misma función (Konagurthu *et al.*, 2008; Budowski-Tal *et al.*, 2010; Bourne y Weissig, 2003). Los resultados reportados en este apéndice se obtuvieron tomando como base la divergencia de Kullback-Liebler como una medida de similitud (Kullback, 1959). Aunque la divergencia Kullback-Liebler es una herramienta ampliamente utilizada en estadística y reconocimiento de patrones (Hershey y Olsen, 2007)(ver ecuación 12), no es estríctamente una medida debido a que carece de simetría. Esta divergencia se calcula de la siguiente forma:

$$DK(P\|Q) = \sum_i P_i ln(\frac{P_i}{Q_i}), \tag{12}$$

donde *P* y *Q* son los vectores a comparar y el subíndice *i* denota el *i*-ésimo elemento, es decir, Q_i es la frecuencia relativa para el *i*-ésimo residuo del vector *Q*.

Un aspecto a notar en la Ecuación 12, es que arroja resultados no válidos cuando P_i ó Q_i son iguales a 0, lo que es común cuando se comparan estructuras locales, debido a que algunos tipos de aminoácidos podrían no tener residuos representantes. Para resolverlo, se propuso sustituir la operación del logaritmo por una de valor absoluto de la siguiente manera:

$$DKMod(P\|Q) = \sum_i P_i|P_i - Q_i|. \tag{13}$$

El cambio propuesto siempre produce resultados válidos y además mantiene el senti-

do de penalizar proporcionalmente la diferencia existente entre dos vectores de frecuencias. En la Figura 32 se representa el proceso de comparación entre dos vectores con las frecuencias de los 20 tipos de aminoácidos que se encuentran en las cavidades a comparar.

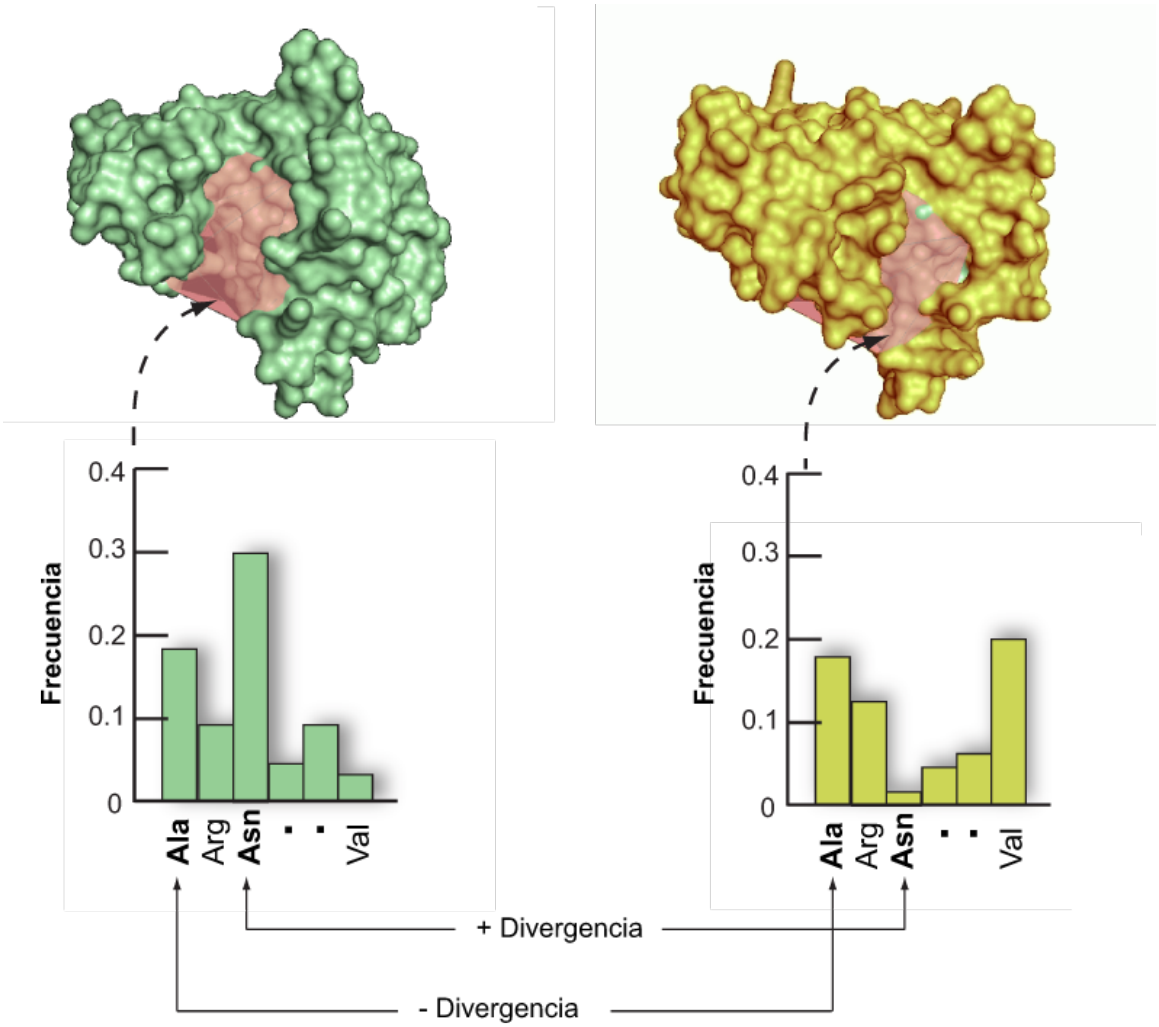


Figura 32: Esquema de comparación para histogramas usando una modificación propuesta a la medida de divergencia de Kullback-Leibler.

A.1. Tratamiento de los datos

Un subconjunto de 26 elementos a partir del grupo de prueba base, se utilizaron para los experimentos de comparación entre las representaciones de proteínas con dimensión 0. Se excluyeron las proteínas: 1IA9, 1E8X, 1CJA, 1NW1 y 1KWP del grupo base; debido a que no existen registros de sus residuos catalíticos en el servidor web CSA (Porter *et al.*, 2004). En la Tabla 10 se muestra cada elemento del conjunto de prueba, así como

el porcentaje de los residuos catalíticos que contienen su cavidad mayor.

El procedimiento de comparación para representaciones estructurales de dimensión 0 constó de tres pasos: en el primero se seleccionaron los residuos de la cavidad mayor por cada proteína del grupo de prueba; en el segundo se construyeron los vectores de frecuencias para los tipos de aminoácidos presentes en los residuos obtenidos en el paso anterior; en el tercer paso se realizó la comparación por pares de los vectores representantivos de las cavidades, la medida que se utilizó fue la divergencia de Kullback-Leibler modificada, la cual permite procesar frecuencias iguales a cero (ver Ecuaciones 12 y 13). Para la selección de residuos pertenecientes a la cavidad mayor, se utilizó la herramienta en línea CASTp (Dundas *et al.*, 2006), la cual localiza y mide cavidades en la estructura tridimensional de las proteínas mediante la triangulación de sus superficies y una esfera de prueba (Dundas *et al.*, 2006). La esfera de prueba que se utilizó en los experimentos tenía un radio de 1.4 Å, aproximadamente el radio de una molécula de agua. Por otra parte, el cálculo de la medida de divergencia se implementó en el programa MATLAB© (MATLAB, 2010), y los resultados se muestran en la siguiente subsección.

A.2. Resultados

Las comparaciones entre pares de proteínas se representan en una matriz de divergencia (*MD0*). La posición *MD0(i,j)* contiene un tono de gris asociado al valor de divergencia resultante entre la *i*-ésima y la *j*-ésima proteína, donde los tonos más claros se asignaron a las divergencias más pequeñas. La Figura 33 representa la matriz de divergencia para todos los pares de proteínas de la Tabla 10. Es notable el predominio de valores en el intervalo $[0.025, 0.05]$, sin embargo, no se distinguen agrupaciones evidentes entre las proteínas de los mismos grupos funcionales; también resaltan los valores de mayor divergencia entre la proteína 1IA8 y todas las demás del grupo de prueba. Una revisión más detallada al vector de frecuencias de esta proteína mostró que casi la mitad de los 20 aminoácidos no tienen residuos representantes en su cavidad mayor, lo que produce una diferencia marcada con los demás vectores. En la Figura 34 se muestra una comparación entre los vectores de frecuencias para la proteína 1IA8 y las que presentan mayor (1BO1) y menor (1H1W) divergencia con ella.

Una de las características de los métodos que utilizan la representación estructural de

dimensión 0, es su dificultad para identificar regularidades en sub-estructuras (Hasegawa y Holm, 2009), como se pudo comprobar con los resultados obtenidos con el conjunto de prueba; sin embargo, este tipo de estructuras resultan útiles en métodos de aproximación rápida o como plantillas de comparación para estructuras locales previamente identificadas y clasificadas.

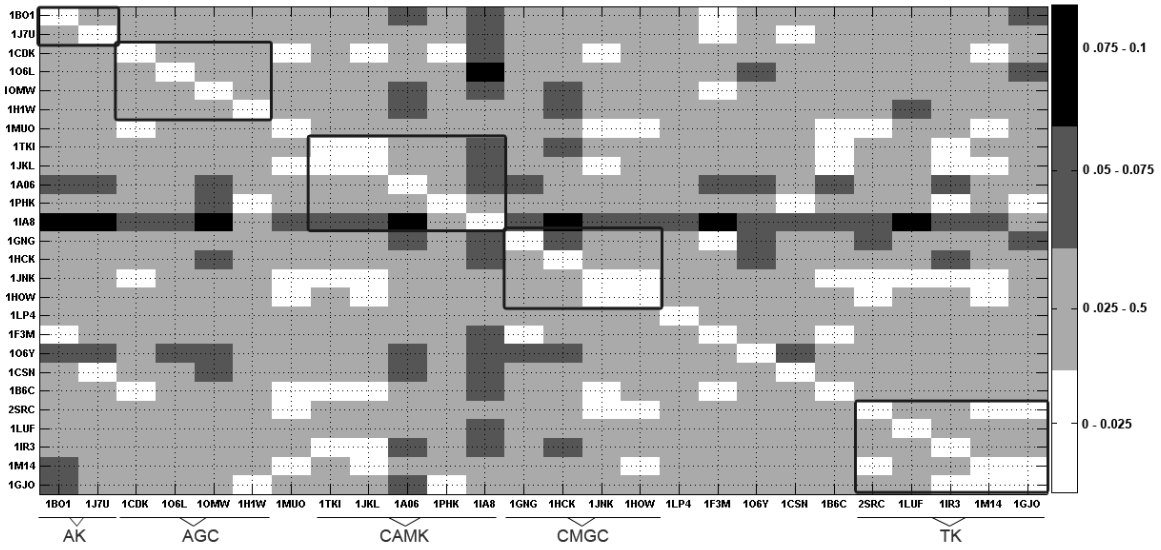
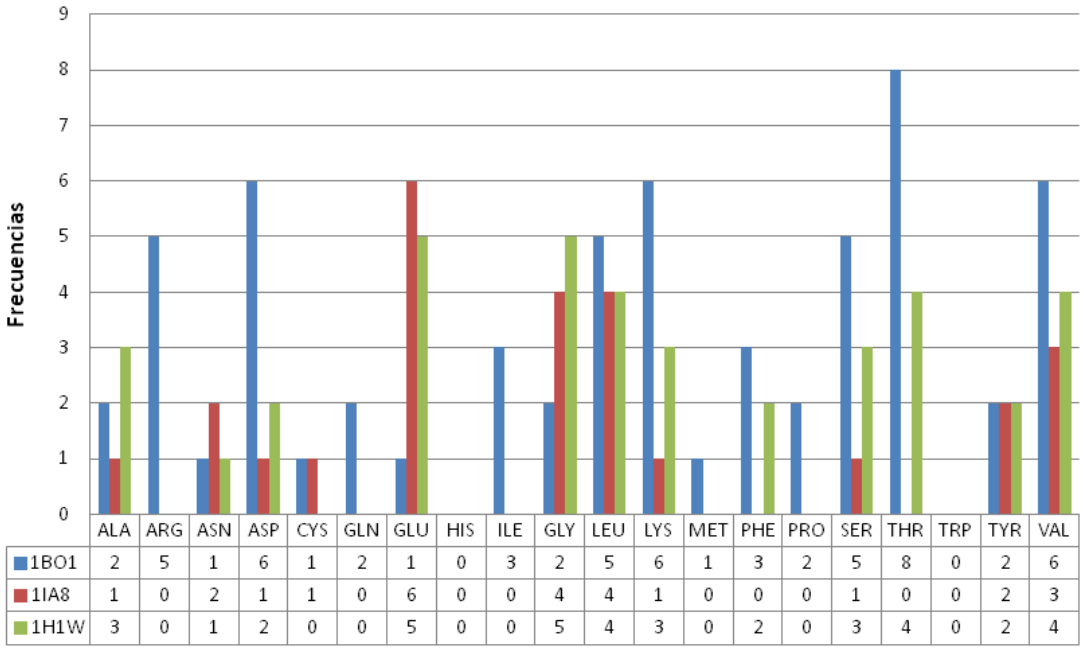


Figura 33: Matriz de divergencias $MD0$ para los vectores de frecuencias en las cavidades mayores de las proteínas del grupo de prueba.

Otra característica que puede aportar información significativa, es la proporción de los tipos de residuos de acuerdo a la propensión de sus cadenas laterales a estar en contacto con solventes polares, como las móleculas de agua. De esta manera, cada cavidad se representa con el número de residuos de tipo polar, hidrofóbicos y cargados. La matriz de divergencia *MDSolv* de esta caracterización para el grupo de prueba se muestra en la Figura 35; en ella se obtienen valores de divergencia pequeños en cada grupo funcional, sin embargo, también se ve el mismo comportamiento entre proteínas de grupos funcionales distintos, por lo que no es factible utilizarla como un clasificador de funciones para este grupo de prueba.



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	GLY	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
1BO1	2	5	1	6	1	2	1	0	3	2	5	6	1	3	2	5	8	0	2	6
1IA8	1	0	2	1	1	0	6	0	0	4	4	1	0	0	0	1	0	0	2	3
1H1W	3	0	1	2	0	0	5	0	0	5	4	3	0	2	0	3	4	0	2	4

Figura 34: Comparación entre vectores de frecuencias para las cavidades mayores de las proteínas 1IA8, 1BO1 y 1H1W

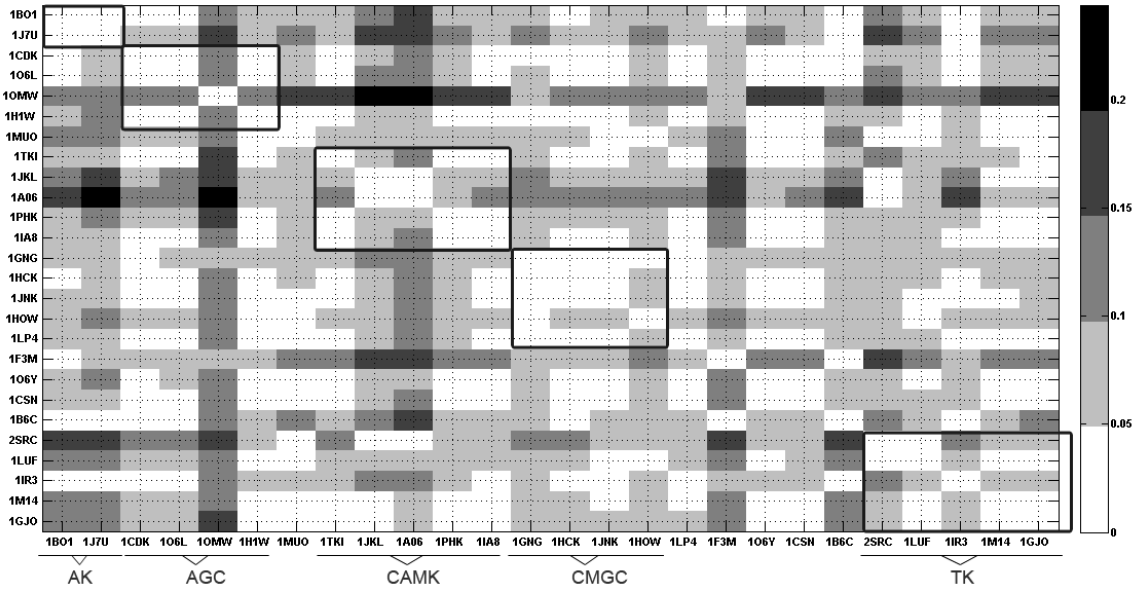


Figura 35: Matriz de divergencia *MDSolv* para las cavidades mayores de las proteínas del grupo de prueba. Los vectores de características contienen los números de residuos de acuerdo a sus tipos de contacto al solvente.

Apéndice B. Comparaciones con representaciones de estructuras de dimensión 1

Casi todas las inferencias funcionales o evolutivas de las proteínas obtenidas hasta la fecha se han realizado mediante la comparación de sus cadenas de aminoácidos y su posterior asociación con alguna familia. La cantidad de secuencias almacenadas en diversas bases de datos contribuyeron al desarrollo de métodos para hacer más eficientes y rápidos los procesos de comparación. Dichos procesos utilizan alineamientos ya sea con las secuencias completas de las proteínas a comparar (enfoque global) (Needleman y Christian, 1970) o con las regiones que tienen mayor coincidencias entre sus tipos de aminoácidos (enfoque local) (Smith y Waterman, 1981).

Varios estudios se han realizado para comparar la eficiencia de métodos de alineamiento secuencial; sin embargo, no es fácil determinar los valores de los parámetros en las funciones de puntuación que capturen mejor las similitudes entre los residuos de un par de secuencias (matrices de sustitución, penalidad por apertura y extensión de inserciones) (Kolodny *et al.*, 2005). Además, un ajuste óptimo para un par puede ser deficiente para otros pares de un mismo grupo de interés.

Para contrastar las agrupaciones con representaciones estructurales terciarias publicadas por Gramada y Bourne (2006), en esta investigación se calcularon alineamientos para representaciones estructurales de dimensión 1 con enfoques global y local.

B.1. Tratamiento de los datos

Las secuencias, en formato FASTA, del conjunto de prueba definido en 4.1.1 se utilizaron para calcular los alineamientos por pares. Dichas secuencias se obtuvieron mediante la base de datos *UniProtKB* (Consortium, 2014). En la Tabla 11 se muestran los identificadores utilizados en ésta base de datos y sus equivalentes en nomenclatura PDB.

Los alineamientos, sus enfoques y funciones de similitud empleados se dividieron en: *i)* Global con porcentaje de identidad, *ii)* Local con porcentaje de identidad y *iii)*Local con puntuación.

Para todos los alineamientos se usó una matriz de sustitución de aminoácidos BLOSUM50 (Henikoff y Henikoff, 1992), adecuada para comparar proteínas con una relación

distante, como es el caso de este grupo de prueba. También se usó una penalización por apertura (inserción y borrado) y por extensión igual a ocho (negativo). Los alineamientos con un enfoque global se calcularon con el algoritmo de Needlman-Wunsch, mientras que los alineamientos locales con el algoritmo de Smith-Waterman. Ambos algoritmos fueron ejecutados desde la caja de herramientas de software *Bioinformatics* de MATLAB © (MATLAB, 2010).

B.2. Resultados

Los resultados de los experimentos comprendidos en esta subsección se almacenaron en matrices de similitud cuya representación gráfica (*MS1*) asocia un tono de gris al valor resultante del alineamiento entre la *i*-ésima y la *j*-ésima proteína en la posición *MS1(i,j)*, donde los tonos más claros indican los valores con mayor similitud. En algunas de las representaciones para las matrices de similitud se utilizaron intervalos no uniformes para atenuar los problemas de escala y resaltar los valores semejantes. En los siguientes incisos se detallan los resultados para cada enfoque, indicando los hallazgos más significativos en cada uno de ellos.

i) Alineamientos globales con porcentajes de identidad. La Figura 36 representa la matriz de similitud *MS1_Id_G* con los porcentajes de identidad discretizados en 64 intervalos regulares. Dado que el tamaño de los intervalos no permitió distinguir algunos agrupamientos con bajos niveles de identidad, estos se redefinieron de forma no regular, tal como se muestra en la Figura 37. Con el ajuste previo, se nota que son pocos los porcentajes de identidad mayores al 25 %, excluyendo los valores de la diagonal; en contraste, los porcentajes de identidad con valores entre 10 % y 25 % fueron las más frecuentes, en esta categoría sobresalen las similitudes débiles (menores al 10 %) entre la proteína 1TKI contra todas las demás, debido a la gran cantidad de residuos que tiene su secuencia. Por otra parte, en los grupos **AGC** y **CMGC** se aprecia una mayoría de identidades con valores superiores al 25 %, y en menor medida con el grupo **CAMK**. El grupo **TK** tuvo un mayor número de identidades entre 20 % y 25 %, mientras que en el grupo **AK** la mayoría de los valores estuvieron entre 10 % y 20 %.

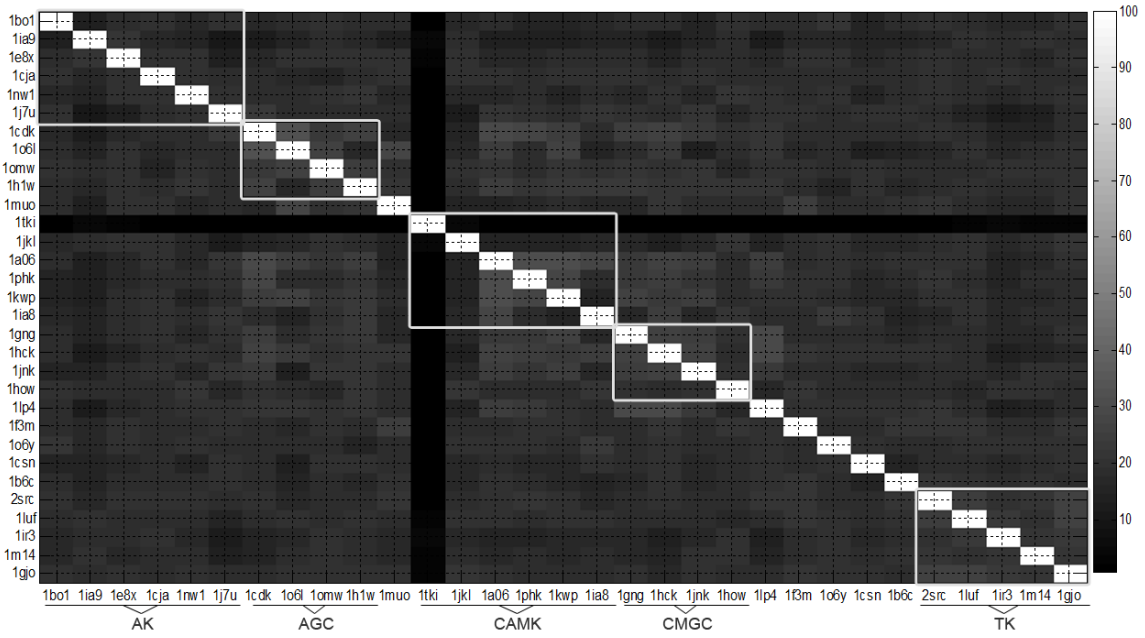


Figura 36: Matriz de semejanza *MS1_Id_G* de los alineamientos globales entre estructuras primarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia de las cinasas. Los valores representan el porcentaje de identidad y las categorías están en intervalos regulares

ii) Alineamientos locales con porcentajes de identidad. Dado que los alineamientos locales se enfocan en las secciones de las secuencias con mayor semejanza, los porcentajes de identidad fueron mayores a los obtenidos con el alineamiento global como se puede ver en la *MS1_Id_L* de la Figura 38. Cuatro intervalos no regulares se utilizaron para acentuar los valores de identidad más abundantes, en los resultados de este alineamiento se distinguen porcentajes superiores al 30 % entre las secuencias de todas las proteínas del grupo **TK**, mientras que la mayoría de las identidades entre las proteínas de los grupos **AGC**, **CAMK** y **CMGC** tomaron valores entre el 30 % y 40 %. Las proteínas del grupo **AK** tuvieron valores de identidad entre 20 % y 30 % en la mayoría de sus alineamientos.

iii) Alineamientos locales con puntuación. Cuando la salida del alineamiento local se expresa en términos de puntuación en lugar de porcentaje de identidad, como se muestra en la matriz de semejanza *MS1_Pt_L* en la Figura 39, se distingue cómo el grupo **AK** (grupo atípico) difiere con respecto a todos los demás por tener las puntuaciones más bajas. Por el contrario, en los grupos **TK** y **AGC** se tuvieron las puntuaciones más altas entre sus elementos con excepción del alineamiento entre

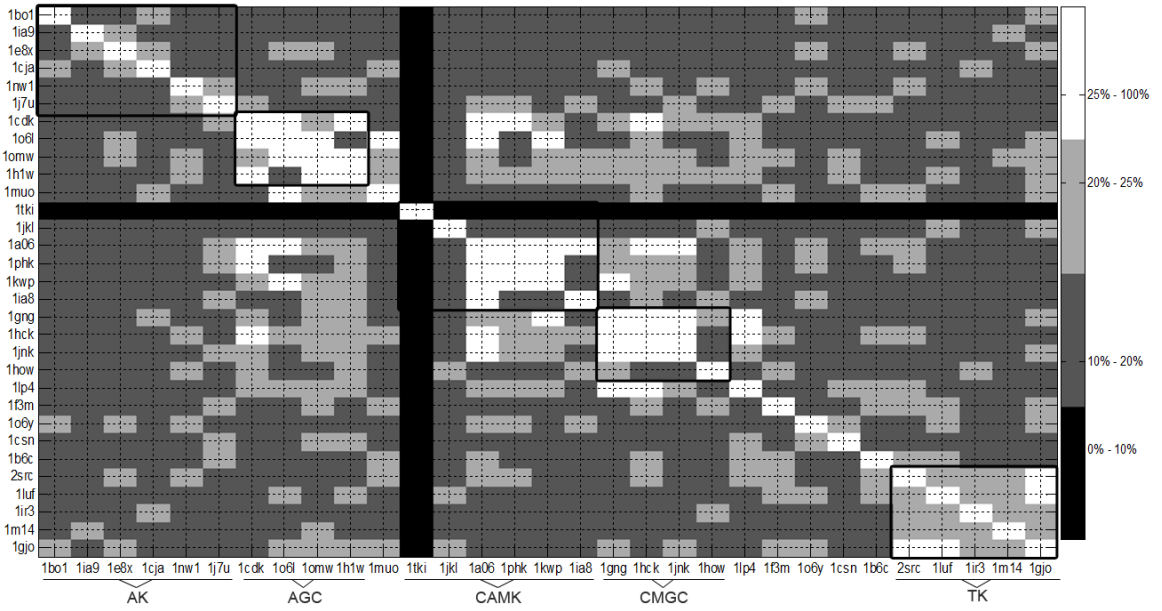


Figura 37: Matriz de semejanza *MS1_Id_G* para los alineamientos globales entre estructuras primarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia las cinasas. Los valores representan el porcentaje de identidad y las categorías están en intervalos irregulares para resaltar valores.

las proteínas 1CDK y 1O6L en el grupo **TK**, y el alineamiento de las proteínas 1LUF y 1IR3 en el grupo **AGC**. En comparación con los grupos **TK** y **AGC**, el grupo **CAMK** tuvo una proporción ligeramente menor de alineamientos con valores en la categoría superior de los intervalos pero mayor a la del grupo **CMGC**.

Los alineamientos locales, enfocados en partes significativas de las secuencias y no en su totalidad, mostraron los mejores agrupamientos entre las MS1 de los experimentos realizados, en especial los locales con puntuación (ver MS1_Pt_L de la Figura 39). A pesar de que con el enfoque local se distinguen mejor las similitudes entre los miembros que pertenecen a los mismos grupos funcionales, se presentaron valores altos de similitud entre secuencias de proteínas pertenecientes a distintas familias. Con los alineamientos locales también se puede observar con claridad cómo las secuencias de las proteínas atípicas (**AK**) tuvieron la semejanza más baja, inclusive entre las del mismo grupo, ésta característica se distingue en menor grado en las *MS* de los otros alineamientos.

Con los alineamientos globales, representados en las figuras 36 y 37, se comprobaron los porcentajes de identidad mencionados en (Gramada y Bourne, 2006), los cuales fueron tomados como referentes en la justificación del conjunto de prueba en la Sección 4.1.1.

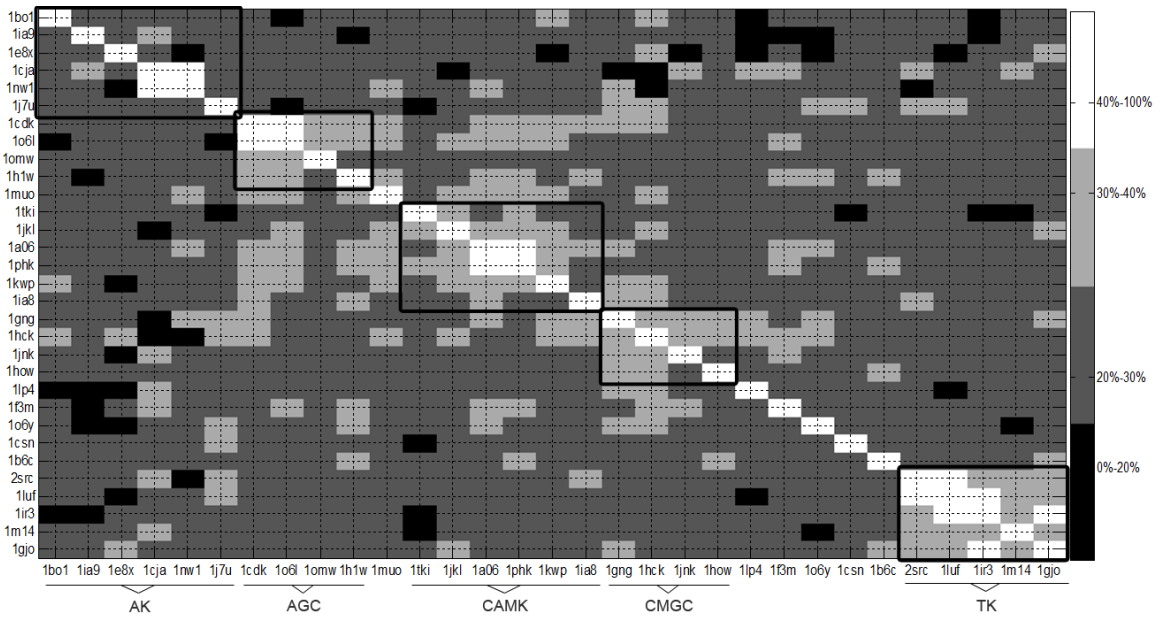


Figura 38: Matriz de semejanza *MS1_Id_L* para los alineamientos locales entre estructuras primarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia las cinasas. Los valores representan el porcentaje de identidad.

La mayoría de los alineamientos de esta *MS1* tuvieron menos del 25 % de identidad. Es importante resaltar que en los alineamientos globales y locales se distinguieron mayores semejanzas en las proteínas pertenecientes al mismo grupo que las semejanzas reportadas en el método de estructuras terciarias de (Gramada y Bourne, 2006).

Los resultados obtenidos con los alineamientos de la estructura de dimensión 1 proporcionaron un marco comparativo básico para evaluar los obtenidos en la propuesta del método con enfoque de estructuras terciarias. Sin embargo, en este tipo de similitudes no se considera información relevante de las conformaciones de las proteínas. En el apéndice C se aborda una medida de similitud que sí incorpora características conformacionales simples.

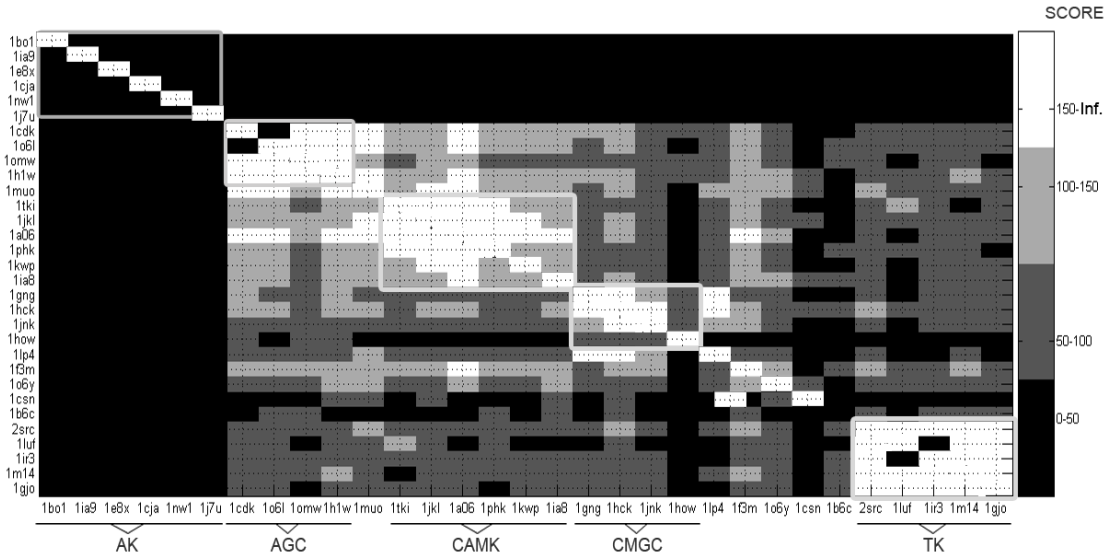


Figura 39: Matriz de semejanza *MS1_Pt_L* para los alineamientos locales entre estructuras primarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia las cinasas. Los valores representan las puntuaciones de los alineamientos.

Apéndice C. Comparaciones con representaciones de estructuras de dimensión 2

Los métodos de comparación entre pares de estructuras secundarias (dimensión 2) permiten detectar semejanzas en los rasgos conformacionales estables y rígidos dependientes de la cadena principal. A menudo, la información de estos rasgos, ya sea de forma global o local, sirve como auxiliar en los métodos de predicción de estructuras o de funciones para representaciones de estructuras terciarias (Russell *et al.*, 1996). Las estructuras secundarias se consideran descriptores de nivel grueso, y las similitudes se basan en los plegamientos en común. Las características conformacionales que se representan en las estructuras secundarias se clasifican en tres tipos principales: hélices α, hojas β y otras estructuras cíclicas γ (Volkert y Staffer, 2004). En los experimentos de esta sección se mapearon estos ocho estados a sólo tres: hélices (α), hojas (β) y otras estructuras (γ), tal como se muestra en la Tabla 14 (Volkert y Staffer, 2004). Por su forma de representación secuencial, análoga a las estructuras primarias, se aplicaron los mismos mismos enfoques de alineamiento y análisis de resultados descritos en el Apéndice A.

Los alineamientos de estructuras secundarias tuvieron como finalidad detectar patrones entre las relaciones de semejanza del conjunto de prueba, y de esta forma robustecer el método propuesto para la comparación con estructuras terciarias.

Tabla 14: Estados de las estructuras secundarias del modelo DSSP y su reducción simple (Extraída de Volkert y Staffer (2004)).

Símbolo en DSSP	Reducción simple
H	α
G	γ
I	γ
E	β
B	γ
T	γ
S	γ
.	γ

C.1. Tratamiento de los datos

Las secuencias de las estructuras primarias de las proteínas del grupo de prueba descrito en la Sección 4.1.1 y listadas en la Tabla 9, fueron utilizadas para predecir las

estructuras secundarias para cada uno de sus residuos. Con las nuevas secuencias se realizaron los alineamientos por pares con los mismos enfoques que se aplicaron a las estructuras de dimensión 1.

Los experimentos de esta subsección se realizaron en dos fases: en la primera se utilizó la base de datos UniProtKB (Consortium, 2014) para extraer la secuencia de las estructuras primarias de cada proteína en formato FASTA (ver Tabla 11); en la segunda fase se utilizó el predictor de estructuras secundarias para secuencias SSEA (Fontana *et al.*, 2005), el cual asignó uno de los tres tipos de conformaciones a cada residuo obtenido del primer paso; con el predictor SSEA también se realizaron los alineamientos con un criterio de puntuaciones con los enfoques globales y locales. Para ambos alineamientos los valores de similitud (en el sistema de puntuaciones) están normalizados en el intervalo [0,100] con un Z-score igual a 0, lo que equivale a una medida por porcentaje de identidad.

C.2. Resultados

La visualización de los alineamientos de las estructuras secundarias se representó con una matriz de similitud (*MS2*), cuya posición *MS2(i,j)* contiene un tono de gris asociado a la puntuación resultante del alineamiento entre la *i*-ésima y la *j*-ésima proteína, y donde los tonos más claros indican los valores con mayor similitud.

i) Alineamientos globales con puntuaciones. La Figura 40 representa la matriz de similitud *MS2_Pt_G* para este enfoque, los intervalos para agrupar sus valores son irregulares y se entonaron de forma empírica para capturar mejor las relaciones de semejanza. Es notable como en el grupo **AGC** los valores de similitud alternan entre los más altos de la escala de puntuación y los más bajos. Otro aspecto a resaltar son las puntuaciones altas que se obtuvieron entre la mayoría de las proteínas del grupo **AK** (Atípicas), situación que no se presentó con los alineamientos de las estructuras primarias ni en las comparaciones de las estructuras de dimensión 0.

ii) Alineamiento local con puntuaciones. Para identificar agrupaciones significativas en regiones de las estructuras secundarias y no en su totalidad, se realizaron alineamientos locales cuyos resultados se representaron en la matriz de similitud

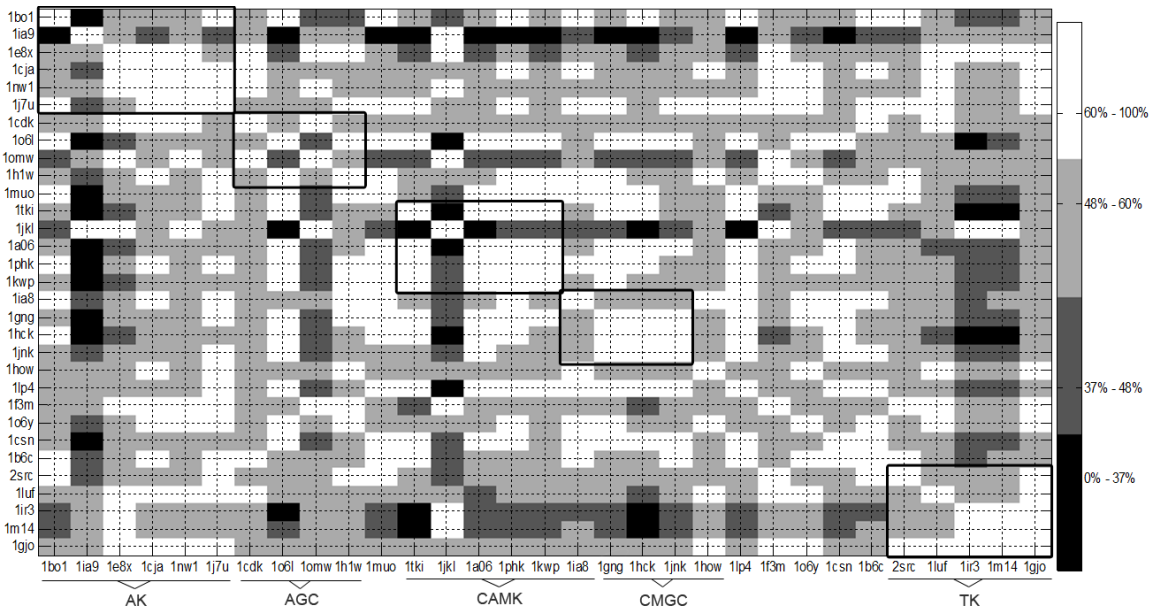


Figura 40: Matriz de semejanza *MS2_Pt_G* para los alineamientos globales entre estructuras secundarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia de las cinasas. Los valores representan una puntuación de similitud.

MS2_Pt_L de la Figura 41. Los alineamientos con este enfoque tuvieron, en general, las puntuaciones por arriba de 40/100. Sin embargo, muchas de estas se presentaron entre proteínas de distintos grupos funcionales. Sobresalió el grupo **CMGC** al tener más alineamientos con puntuaciones mayores a 60. Mientras que el grupo **AK** tuvo alineamientos con puntuaciones semejantes a las obtenidas en el alineamiento global. En el grupo **CAMK** se destacaron los alineamientos con las menores puntuaciones entre la proteína 1JKL y las demás proteínas de su mismo grupo.

Los alineamientos en las estructuras secundarias tuvieron como objetivo identificar entre las proteínas las similitudes en las conformaciones de la cadena principal y verificar si estas similitudes podrían superar la limitación que tienen las estructuras primarias para detectar rasgos indicativos de funciones. Sin embargo, el uso de un alfabeto con sólo tres letras, que representan conformaciones en la cadena principal de las proteínas, produjo en muchos de los alineamientos valores altos de similitud, inclusive entre elementos de grupos distintos; lo que ocasionó una pérdida de definición en los agrupamientos que se obtuvieron con los alineamientos de las estructuras primarias. El caso de mayor interés se presentó con las proteínas del grupo **AK**, sus puntuaciones de identidad para casi to-

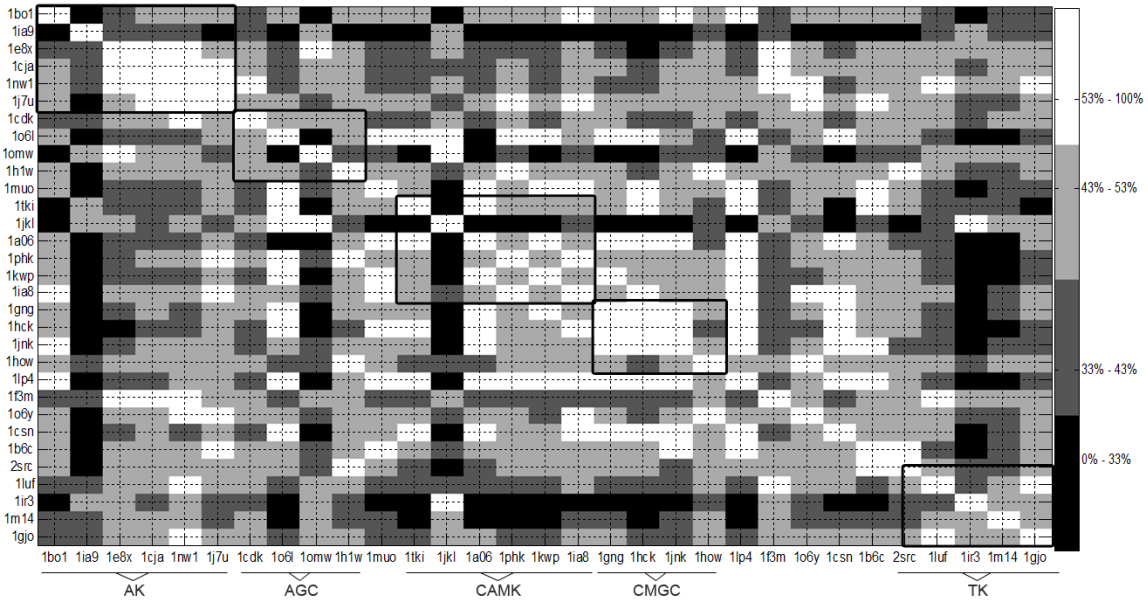


Figura 41: Matriz de semejanza *MS2_Pt_L* para los alineamientos locales entre estructuras secundarias de las proteínas pertenecientes al conjunto de prueba de la superfamilia de las cinasas. Los valores representan una similitud dada en puntos.

dos sus elementos, con excepción de 1BO1 y 1IA9, fueron los más altos en los intervalos propuestos, esta situación no se presentó en los alineamientos de las estructuras primarias debido a la diferencia de sus residuos pero no de sus conformaciones en la cadena principal.

La comparación de estructuras secundarias a nivel global o local (cavidades) es una estrategia de comparación con baja eficiencia para detectar similitudes, principalmente con aquellas proteínas que tienen una relación evolutiva es distante, como el grupo de prueba de las cinasas. La selección de información más detallada de la estructura secundaria, como la longitud, topología de las conexiones y eventualmente las orientaciones relativas; tienen el potencial de mejorar la agrupación de elementos de las mismas familias y discriminar las que no pertenezcan.