UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Nuevos Descriptores 3D Macromoleculares Algebraicos para QSA(F)R / Similitud en Proteínas

Julio Enrique Terán Zavala

Yovani Marrero Ponce, Ph.D.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Magíster en Química
Mención Físico-Química/Química-Física

Quito, 15 de noviembre de 2018

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Nuevos Descriptores 3D Macromoleculares Algebraicos para QSA(F)R / Similitud en Proteínas

Julio Enrique Terán Zavala

Firmas

Yovani Marrero Ponce, Ph.D.

Director del Trabajo de Titulación _____

F. Javier Torres, Ph.D.

Director del Programa de Maestría en Química _____

César Zambrano Semblantes, Ph.D.

Decano del Colegio de Ciencias e Ingenierías _____

Hugo Burgos Yánez, Ph.D.

Decano del Colegio de Posgrados _____

Quito, 15 de noviembre de 2018

© Derechos de Autor

Firma del estudiante _____

Nombre: Julio Enrique Terán Zavala

Código: 00141000

C.I.: 1714578091

Lugar y fecha: Quito, 15 de noviembre de 2018

AGRADECIMIENTOS

A mi querida familia, quienes siempre han estado a mi lado, para ayudarme en mi desarrollo profesional, académico y, sobre todo, personal.

A mi tutor, Yovani Marrero-Ponce, quien no solo ha sido una guía respecto al ámbito académico, sino también ha sido un gran consejero y maestro.

A la Universidad San Francisco, en especial a sus autoridades, Carlos Montúfar, Ximena Córdova, César Zambrano y Daniela Almeida, por permitirme el continuar mi formación académica y profesional.

Al Departamento de Ingeniería Química de la USFQ, en donde grandes amigos y colegas siempre estuvieron dispuestos a ayudarme en cualquier momento.

Al Grupo de Química Computacional y Teórica (QCT) quienes fueron un gran apoyo, un ejemplo desde el punto de vista científico y humano.

A Cristina, por cambiarlo todo.

RESUMEN

En el presente trabajo se propone una nueva familia de descriptores moleculares tridimensionales geométricos para proteínas, basados en conceptos del álgebra lineal y multi-lineal, mediante el uso de formas algebraicas 2-lineales (lineal, bilineal y cuadrática) y 3-lineales (trilineal, trilineal cúbica, trilineal cuadrática bilineal, trilineal lineal y trilineal bilineal) como casos específicos de las formas algebraicas N-lineales. Para este fin se definen las k-ésimas matrices espaciales 2-tuplas y 3-tuplas de similitud-disimilitud, para representar la información química existente en las relaciones entre dos y tres aminoácidos de una proteína, respectivamente.

Además, se proponen varias métricas y multi-métricas para extraer la información existente por la interacción entre los aminoácidos presentes, así como transformaciones probabilísticas para normalizar las representaciones matriciales definidas. También, se introducen procedimientos de cortes moleculares topológicos y geométricos con el propósito de visualizar las interacciones no covalentes dependientes de la región. Por último, se define un procedimiento que generaliza el cálculo de los descriptores a través de la obtención de índices de orden aminoacídico que pueden ser fusionados mediante el uso de un conjunto de operadores matemáticos.

Se desarrolló en java, el *software* multi-plataforma MuLiMS-MCoMPAs, que aprovecha toda la capacidad computacional mediante un esquema de cálculo de memoria compartida, para calcular estos descriptores.

Los estudios realizados demostraron que los descriptores propuestos discriminan de mejor manera a proteínas estructuralmente distintas y codifican información ortogonal respecto a otros enfoques propuestos en la literatura. Además, con respecto a una aplicación clásica de las proteínas como la velocidad de plegamiento, se observó que los descriptores propuestos extraen información que permite un ajuste superior respecto a la variable respuesta y respecto a otros descriptores reportados en la literatura. Consecuentemente, se puede concluir que esta propuesta teórica facilita la definición de descriptores moleculares tridimensionales que tienen información que permite el modelado de propiedades y funciones para proteínas, y que pueden ser generados mediante el uso de un software validado, multiplataforma y amigable con el usuario que está disponible para el público.

Palabras clave: Descriptores 3D, proteínas, formas algebraicas multilineales, multimétricas, operadores de agregación, transformación probabilística, velocidad de plegamiento, clasificación estructural proteica.

ABSTRACT

In this work, a new type of tridimensional geometrical molecular descriptors for proteins is proposed. These descriptors make use of various linear and multilinear algebra concepts by applying 2-linear (linear, quadratic and bilinear) and 3-linear (trilinear, trilinear cubic, trilinear quadratic bilinear, trilinear linear and trilinear bilinear) forms as specific cases of the N-linear algebraic forms. The definition of the k^{th} 2-tuple and 3-tuple similarity-dissimilarity spatial matrices are required for the transformation and for the representation of the existing chemical information available in the relationships between two and three amino acids of a protein, respectively.

Additionally, several metrics and multi-metrics are proposed for interaction information extraction as well as probabilistic transformations, which serve for matrix normalization. Moreover, topological and geometrical cut-offs are introduced as a strategy for non-covalent interaction regions. Finally, a generalizing procedure is proposed considering amino acid-based indices that can be fused together by using mathematical operators for molecular descriptors calculations is proposed.

These defined molecular descriptors can be calculated using a java based multiplatform software called MuLiMS-MCoMPAs. This software can use all computational capacity by the parallelization of the code.

The performed experiments demonstrated that the proposed 3D descriptors perform better than other approaches in the discrimination of structurally different proteins and codify orthogonal information. Moreover, based on a classical protein application such as folding rate, the proposed descriptors extract information that adjust better to the responsible variable and it ranks better than other reported descriptors. Consequently, it can be concluded that this theoretical proposal allows the definition of tridimensional molecular descriptors that possess information that permit to generate models for protein properties or function purposes. These indices can be calculated using a multiplatform, validated and user-friendly software that is available on the internet.

Keywords: 3D descriptors, proteins, multilinear algebraic forms, aggregation operators, probabilistic transformations, folding rate, protein structural classification.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Las proteínas son responsables de todas las funciones en los organismos vivos. Entre los ejemplos más comunes podemos citar: la replicación del genoma, la catálisis enzimática y las respuestas de varios sistemas como el inmunológico, nervioso y respiratorio [1]. La característica más representativa de las proteínas con respecto a otros sistemas químicamente replicativos es su estabilidad estructural (estructura terciaria). Esta estructura da a las proteínas su actividad funcional en el organismo. La relación entre la secuencia (estructura primaria), la estructura geométrica y la función de una proteína se conoce como el "paradigma de la biología molecular" [2]. La relación estructura-función puede evidenciarse al estudiar anomalías en procesos biológicos que resultan en enfermedades. Como ejemplos se pueden indicar el enfisema pulmonar, el Alzheimer y la fibrosis quística.

A pesar de la complejidad de las proteínas debido a su conformación, gran cantidad de átomos e interacciones, las proteínas tienen la habilidad de encontrar un estado de equilibrio mediante su plegamiento. Este proceso se ha estudiado *in vitro*, para una gran cantidad de proteínas. Es importante mencionar, que esta estabilidad y el proceso de plegamiento, tienen una base termodinámica más que biológica, debido a que la proteína busca el estado de menor energía [3]. El mecanismo de cómo las proteínas encuentran este estado de menor energía se conoce como el problema del plegamiento de proteínas o el segundo código genético [4]. Varios campos de la ciencia buscan tratar de encontrar una solución para esta temática, con el objetivo de entender mejor varios procesos en los que las proteínas se encuentran involucradas.

Para proteínas, existen dos grandes tipos de repositorios de información: los que contienen información estructural como Unitprot [5] y Genbank [6] donde se encuentran millones de secuencias de proteínas almacenadas; y los que contienen información

geométrica de las proteínas como el Protein Data Bank [7], donde se contiene un limitado número de estructuras. La limitación en la cantidad de estructuras tridimensionales disponibles se debe a la gran cantidad de recursos, equipos y tiempo para su determinación. Este particular nos hace observar la marcada diferencia entre la información existente en el campo de las proteínas y la necesidad de buscar nuevos métodos que permiten la predicción estructural de las proteínas, para así, poder explicar de mejor forma la relación estructura-función.

Los métodos computacionales de predicción estructural de proteínas han sido un campo activo en la ciencia en los últimos 30 años. A pesar de ser teóricamente muy robustos, los métodos *ab initio* y el uso de potenciales energéticos se van descartando como una herramienta para la predicción estructural en proteínas debido a los altos costos computacionales, a la complejidad, desde el punto de vista químico, de las proteínas, y que los resultados obtenidos de estos métodos no logran esclarecer la relación entre estabilidad y estructura.

Por otro lado, los estudios de relaciones estructura-actividad/propiedad ha sido un campo de investigación que ha aportado datos y relaciones relevantes en el campo de la química juntamente con el uso de técnicas estadísticas y de aprendizaje automático [8]–[11]; estos estudios se realizan gracias a la gran cantidad de información relacionada a estructuras y moléculas orgánicas. Los descriptores moleculares se plantean como actores principales de esta estrategia de modelación. Estos descriptores son herramientas que permiten la extracción de información estructural de los sistemas y la expresan mediante un número, que puede ser utilizado con el fin de modelar propiedades o relaciones fisicoquímicas de interés [12]. Cabe mencionar que la cantidad de descriptores para proteínas con respecto a los descriptores existentes para moléculas orgánicas es de 1 a 1000 [12].

Como consecuencia de esto, es importante desarrollar nuevos descriptores que permitan la extracción de información valiosa para lograr una mejor caracterización de la estructura de las proteínas [13], [14].

En los trabajos de Marrero *et al*. [15]–[19], el uso de descriptores moleculares topológicos para moléculas orgánicas se ha aplicado en la definición de índices N-lineales utilizando formas algebraicas como herramienta matemática. En varias aplicaciones, se obtuvieron modelos que permiten una alta predicción de propiedades fisicoquímicas para sistemas orgánicos.

Con base a lo anterior, se plantea que la **problemática científica** es la definición de nuevos descriptores tridimensionales para proteínas que permitan la estimación de propiedades o funciones fisicoquímicas, con contenido de información diferente a descriptores ya reportados.

A partir de esto, se plantea como **objetivo general** de este trabajo generar nuevos descriptores tridimensionales para proteínas para predicción de propiedades estructurales y fisicoquímicas mediante el uso de formas algebraicas multilineales, que a su vez permita cumplir los siguientes **objetivos específicos**:

1. Definir descriptores 3D para proteínas mediante formas algebraicas usando varias herramientas matemáticas adicionales como son: multimétricas, métricas, operadores de agregación, transformaciones probabilísticas, cut-off y partición de tensores y matrices de transformación.
2. Implementar y validar el funcionamiento del software ToMoCoMD-CAMPS MuLiMs-MCoMPAs (acrónimo de Topological Molecular Computer-Aided Modelling in Protein Science Multi-Linear Maps based on N-Metric & Contact Matrices of 3D-Protein and Amino acid Weightings) generado para automatizar el

cálculo de los descriptores propuestos mediante la generación de gráficos de rendimiento computacional.

3. Reducir la dimensionalidad del espacio de descriptores que el software puede generar mediante un análisis de componentes principales (ACP), un análisis de redundancia de información y modelado.
4. Evaluar la utilidad de los descriptores moleculares propuestos en la predicción de las clases estructurales de las proteínas y en la predicción de su velocidad de plegamiento mediante su rendimiento en modelos de clasificación y regresión, respectivamente, para un conjunto de proteínas con respecto a los mejores métodos reportados en la literatura.

Para dar respuesta a estos objetivos, se trabaja sobre la siguiente **hipótesis**: es posible definir una nueva familia de descriptores tridimensionales para proteínas utilizando conceptos de álgebra lineal y multi-lineal con información ortogonal a los índices ya reportados en la literatura, y que permitan obtener modelos de estructura-actividad que predigan de mejor manera diferentes propiedades fisicoquímicas de interés para el campo de las proteínas.

La presente tesis tiene como **novedad científica** la generación de un nuevo tipo de descriptores tridimensionales para proteínas utilizando formas algebraicas bilineales y trilineales, utilizando varias herramientas matemáticas de generalización como métricas, multimétricas, transformaciones probabilísticas, operadores de agregación, cut-offs y generación de descriptores locales de aminoácido. Adicionalmente, se realizan análisis exploratorios (de variabilidad y componentes principales) y de selección de atributos mediante herramientas de teoría de información para reducir la dimensionalidad y optimizar la herramienta de cálculo generada.

Como **aporte práctico** de este trabajo se tiene el software ToMoCoMD-CAMPS MuLiMs-MCoMPAs, el cual permite el cálculo de los descriptores moleculares propuestos, y la evaluación de los índices para dos aplicaciones representativas en el campo de las proteínas como son el plegamiento de proteínas y la clasificación SCOP de proteínas.

La tesis está estructurada en cuatro capítulos: en el capítulo 1, se realiza una revisión bibliográfica sobre el estado actual de los descriptores para proteínas, así como las diferentes herramientas matemáticas y estadísticas que se utilizarán. En el capítulo 2 se presentarán las bases de datos utilizadas para la validación de los descriptores propuestos y las herramientas computacionales utilizadas en este trabajo. En el capítulo 3 se presentan los resultados que validan el presente trabajo, así como un análisis comparativo de los modelos generados utilizando los índices calculados mediante el software ToMoCoMD-CAMPS MuLiMs-MCoMPAs con respecto a los resultados reportados en varios artículos científicos. En el capítulo 4 se exponen las conclusiones y recomendaciones de este trabajo.

1. REVISIÓN DE LA LITERATURA

1.1. Formas algebraicas, métricas y multimétricas

Las formas algebraicas son funciones matemáticas que realizan una transformación desde un espacio n-dimensional a un espacio 1-dimensional (campo escalar) [20], [21]. Las formas algebraicas están estrechamente relacionadas con conceptos de álgebra lineal y las más comunes son las formas lineales, bilineales y cuadráticas [22]. Considerando el álgebra tensorial (o multi-lineal), estas formas algebraicas se generalizan denominándose formas algebraicas N-lineales, donde N es el número de parámetros que recibe la función. Las formas algebraicas requieren de una matriz o un tensor de orden N para realizar la transformación de un espacio a otro. A continuación, se realizará la descripción matemática de dichas formas.

1.1.1. Formas algebraicas lineales, bilineales y cuadráticas.

Una aplicación ($f : A \to B$), se denomina lineal entre los espacios vectoriales A y B sobre un mismo campo escalar K, si cumple los siguientes axiomas:

$$f(\bar{v} + \bar{w}) = f(\bar{v}) + f(\bar{w}) \quad (1)$$

$$f(\overline{\lambda v}) = \lambda\, f(\bar{v}) \quad (2)$$

Para todo $\bar{v}, \bar{w} \in A \; y \; \lambda \in K$. Esto significa que si se toman dos elementos $\bar{v}, \bar{w} \in A$ y se determina su suma, esto corresponde en B a buscar las imágenes de los elementos y aplicar la misma operación; lo mismo sucede con el producto escalar. Por tanto, las formas algebraicas lineales constituyen un *homomorfismo* entre los espacios vectoriales A y B (e.g. $f : \mathbb{R}^4 \to \mathbb{R}^2$), o un *endomorfismo* si A= B (e.g., $f : \mathbb{R}^4 \to \mathbb{R}^4$).

Formalmente, sean A y B espacios vectoriales de dimensión finita n y m respectivamente, entonces una aplicación lineal $f: A \to B$ se define a partir de los elementos de las bases de los respectivos espacios vectoriales. De esta manera, si $E = \{\overline{e_1}, \overline{e_2}, \dots, \overline{e_n}\}$ es una base de A entonces cualquier vector $\bar{x} \in A$ está determinado por:

$$\bar{x} = \sum_{i=1}^{n} x^i \overline{e_\iota} \quad (3)$$

y aplicando (2),

$$f(\bar{x}) = f\left(\sum_{i=1}^{n} x^i \overline{e_\iota}\right) = \sum_{i=1}^{n} x^i f(\overline{e_\iota}) \quad (4)$$

donde, $(x^1, x^2, \dots, x^n) \in \mathbb{R}^n$ son las coordenadas del vector $\bar{x}$ en la base E, y $f(\overline{e_\iota})$ puede expresarse como una combinación lineal respecto a la base $C = \{\overline{b_1}, \overline{b_2}, \dots, \overline{b_m}\}$ del espacio vectorial B como sigue:

$$f(\overline{e_\iota}) = a_{1_i}\overline{b_1} + a_{2_i}\overline{b_2} + \cdots + a_{m_i}\overline{b_m} \quad (5)$$

donde, a_{ji} son las coordenadas de los vectores $\overline{e_i}$ en la base C. Por lo tanto, la aplicación lineal f queda determinada por una matriz $V_{m\,x\,n}$ compuesta por los vectores columna de las coordenadas de los vectores $\overline{e_i}$ en la base C, es decir:

$$f(\bar{x}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x^1 \\ x^2 \\ \cdots \\ x^n \end{bmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} x^i \qquad (6)$$

Sin embargo, en ocasiones, una aplicación lineal realiza transformaciones desde un espacio vectorial al campo de escalares ($f : A \to \mathbb{R}$). Esta transformación se conoce como *forma o función lineal*, es decir, la transformación se realiza como combinación lineal de las coordenadas de cualquier vector $\bar{v} \in A$ de dimensión n:

$$f'(\bar{x}) = \sum_{i=1}^{n} v^i = [u^1 \quad u^2 \quad \cdots \quad u^n] \begin{bmatrix} v^1 \\ v^2 \\ \cdots \\ v^n \end{bmatrix} \qquad (7)$$

Donde $\bar{u}$ puede ser tomado como el vector unitario.

Por otro lado, se define la *forma bilineal* como una aplicación $b : T\,x\,T \to \mathbb{R}$ definida sobre el mismo campo escalar K, la cual es lineal en todos sus argumentos tomados separadamente. Es decir, que esta función satisface las condiciones de linealidad mostradas en la ecuación (8), para cualquier escalar $\lambda \in K$ y cualquier vector $\bar{v}_1, \bar{v}_2, \bar{w}_1, \bar{w}_2 \in A$.

$$\left.\begin{aligned} b(\lambda\overline{v_1}, \overline{w_1}) &= b(\overline{v_1}, \lambda\overline{w_1}) = \lambda\, b(\overline{v_1}, \overline{w_1}) \\ b(\overline{v_1} + \overline{v_2}, \overline{w_1}) &= b(\overline{v_1}, \overline{w_1}) + b(\overline{v_2}, \overline{w_1}) \\ b(\overline{v_1}, \overline{w_1} + \overline{w_2}) &= b(\overline{v_1}, \overline{w_1}) + b(\overline{v_1}, \overline{w_2}) \end{aligned}\right\} \qquad (8)$$

Consecuentemente, si A es un espacio vectorial sobre los números reales, y el conjunto de vectores $E = \{\overline{e_1}, \overline{e_2}, \ldots, \overline{e_n}\}$, es el sistema base de dimensión n, entonces se pueden definir formas no ambiguas para cualquier vector $\bar{x}, \bar{y} \in A$, siendo sus coordenadas $(x^1, x^2, \ldots, x^n)\ \epsilon\ \mathbb{R}^n$ y $(y^1, y^2, \ldots, y^n)\ \epsilon\ \mathbb{R}^n$, respectivamente. Esto significa que los vectores

$\bar{x}, \bar{y}$ son expresados como combinaciones lineales respecto a la base correspondiente como se muestra a continuación:

$$\bar{x} = \sum_{i=1}^{n} x^i \bar{e_i} \qquad \bar{y} = \sum_{j=1}^{n} y^i \bar{e_j} \tag{9}$$

Por consiguiente:

$$b(\bar{x}, \bar{y}) = b\left(\sum_{i=1}^{n} x^i \bar{e_i}, \sum_{j=1}^{n} y^i \bar{e_j}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} x^i y^j b\left(\bar{e_i}, \bar{e_j}\right) \tag{10}$$

Y si son tomados los coeficientes a_{ij} como n x n escalares de $b\left(\bar{e_i}, \bar{e_j}\right)$, es decir,

$$a_{ij} = b\left(\bar{e_i}, \bar{e_j}\right) = \bar{e_i} \otimes \bar{e_j} \tag{11}$$

entonces,

$$b(\bar{x}, \bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x^i y^j = [X]^T V [Y] = \begin{bmatrix} x^1 & \cdots & x^n \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} y^1 \\ \cdots \\ y^n \end{bmatrix} \tag{12}$$

De esta manera, la definición matemática de la forma bilineal es escrita como una ecuación matricial donde $[Y]$ es un vector columna de las coordenadas de $\bar{y}$, y $[X]^T$ es la transpuesta del vector columna $[X]$ de las coordenadas de $\bar{x}$, ambos sobre un sistema base de $\mathbb{R}^n$.

Por otro lado, se puede definir una forma bilineal b como simétrica si $b(\bar{x}, \bar{y}) = b(\bar{y}, \bar{x})$ para todo $\bar{x}, \bar{y} \in A$. Por tanto, a partir de este concepto, se define como *forma cuadrática* a la función $q : A \to \mathbb{R}$, dada por: $q(\bar{x}) = b(\bar{x}, \bar{x})$

1.1.2. Formas algebraicas N-lineales.

Las formas N-lineales son conceptos generalizadores de las formas bilineales, lineales y cuadráticas explicadas en la sección anterior. Sea $A_1, \cdots, A_n$ espacios vectoriales sobre el

campo de los números reales $\mathbb{R}$. Se puede definir como forma N-lineal a una función $w = A_1 \times \cdots \times A_n \to \mathbb{R}$ que cumple las siguientes propiedades:

$$\begin{gathered} w(\lambda\bar{v}_{11}, \dots, \bar{v}_{m1}) = w(\bar{v}_{11}, \dots, \lambda\bar{v}_{m1}) = \cdots = \lambda w(\bar{v}_{11}, \dots, \bar{v}_{m1}) \\ w(\bar{v}_{11} + \bar{v}_{12}, \dots, \bar{v}_{m1}) = w(\bar{v}_{11}, \dots, \bar{v}_{m1}) + w(\bar{v}_{12}, \dots, \bar{v}_{m1}) \\ \vdots \\ w(\bar{v}_{11}, \dots, \bar{v}_{m1} + \bar{v}_{m2}) = w(\bar{v}_{11}, \dots, \bar{v}_{m1}) + w(\bar{v}_{11}, \dots, \bar{v}_{m2}) \end{gathered} \quad (13)$$

donde $\bar{v}_{11}, \bar{v}_{11} \in A \; y \; \lambda \in K$.

Por tanto, si $A_1, \cdots, A_n \in \mathbb{R}^n$ y los sistemas de vectores $\mathrm{E} = \{\overline{e_1}, \overline{e_2}, \dots, \overline{e_n}\}$,…, $Z = \{\overline{z_1}, \overline{z_2}, \dots, \overline{z_l}\}$ son las bases de los respectivos m espacios vectoriales de dimensiones n, …, l, respectivamente, entonces pueden definirse vectores $\bar{v}_1 \in A_1, \dots, \bar{v}_m \in A_m$ como combinaciones lineales respecto a las bases E, …, Z, respectivamente.

De esta forma se tiene:

$$\bar{v}_1 = \sum_{i=1}^{n} v_1^i \, \overline{e_i} \dots \; \bar{v}_m = \sum_{k=1}^{l} v_m^k \, \overline{z_k} \quad (14)$$

Por consiguiente,

$$w = (\overline{v_1}, \dots, \bar{v}_m) = w\left(\sum_{i=1}^{n} v_1^i \, \overline{e_i} \dots \; \bar{v}_m \sum_{k=1}^{l} v_m^k \, \overline{z_k}\right) = \sum_{i=1}^{n} \dots \sum_{k=1}^{l} \left(v_1^i \dots v_m^k\right) w(\overline{e_i} \dots \overline{z_k}) \quad (15)$$

Y si se consideran los elementos $a_{i \dots k}$como $n \times \dots \times l$ escalares de $w(\overline{e_i} \dots \overline{z_k})$ es decir,

$$a_{i \dots k} = \; b(\overline{e_i}, \dots, \overline{z_k}) = \; \overline{e_i} \otimes \dots \otimes \overline{z_k} \quad (16)$$

Entonces,

$$w = (\overline{v_1}, \dots, \bar{v}_m) = \sum_{i=1}^{n} \dots \sum_{k=1}^{l} (a_{i \dots k}) \left(v_1^i \dots v_m^k\right) \quad (17)$$

1.1.3. Métrica.

Una métrica o una función de distancia es una función que define una distancia entre un par de elementos (a,b) de un conjunto. Estas deben cumplir las siguientes condiciones [23]:

i.) d(a,b) ≥ 0 (debe ser positiva)

ii.) d(a,b) = d(b,a) (debe ser simétrica)

iii.) d (a,b) ≤ d(a,c)+d(c,b). (debe cumplir la desigualdad del triángulo)

iv.) d (a,b) = 0 si a=b (debe cumplir el axioma de la identidad)

Las métricas son esenciales en algunas áreas de la ciencia tal como teoría de grafos, biología molecular, entre otras [24], [25].

La Tabla 1, presenta algunos ejemplos de métricas que se pueden utilizar para relaciones entre dos elementos.

1.1.4. Multimétrica

La generalización de la noción de métrica como relación entre *dos objetos* o *binaria* se conoce como multimétrica, y se expresa como la relación entre *varios objetos.* Las definiciones matemáticas para proponer un elemento como multimétrica se hacen a continuación, considerando la notación propuesta por Warrens [26]:

Sea, $x_{1,k} = (x_1, x_2, x_3, \dots, x_k)$, una k-upla y sea, $x_{1,k}^{-i} = (x_1, x_2, x_{i-1}, x_{i+1} \dots, x_k)$, la ($k$-1)-upla donde el menos en el supraíndice de $x_{1,k}^{-i}$ se usa para indicar que el objeto x_i ha sido retirado de la k-upla. A partir de aquí, se define una multimétrica como una medida de disimilitud que satisface:

Tabla 1. Métricas para definición de distancia entre dos elementos

Nombre	Fórmula	Rango
Mannhattan/ City-Block	$d_{XY} = \sum_{j=1}^{n} \lvert x_j - y_j \rvert$	$[0, \infty)$
Euclídea	$d_{XY} = \sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^2}$	$[0, \infty)$
Chebyshev/Lagrange	$d_{XY} = \max\{\lvert x_j - y_j \rvert\}$	$[0, \infty)$
Bhattacharyya	$d_{XY} = \sqrt{\sum_{j=1}^{n} \left(\sqrt{x_j} - \sqrt{y_j}\right)^2}$	$[0, \infty)$
Mahalanobis	$d_{XY} = \sqrt{(X - Y)^t S^{-1} (X - Y)}$	$[0, \infty)$
Correlación	$d_{XY} = 1 - \frac{\sum_{j=1}^{n} (x_j - \overline{X})(y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n} (x_j - \overline{X}) \sum_{j=1}^{n} (y_j - \overline{Y})}}$	$[0,2]$
Separación Angular	$d_{XY} = 1 - \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \ \sum_{j=1}^{n} y_j^2}}$	$[0,2]$
Camberra	$d_{XY} = \sum_{j=1}^{n} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$	$[0, n]$
Soergel	$d_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} \max\{x_j, y_j\}}$	$[0,2]$
Lance-Williams/Bray-Curtis I	$d_{XY} = \frac{\sum_{j=1}^{n} \lvert x_j - y_j \rvert}{\sum_{j=1}^{n} (\lvert x_j \rvert + \lvert y_j \rvert)}$	$[0,1]$
"Wave-Edges"	$d_{XY} = \sum_{j=1}^{n} \left(1 - \frac{\min\{x_j, y_j\}}{\max\{x_j, y_j\}}\right)$	$[0,2n]$

i.) $(k-1) d_k(x_{1,k}) \leq \sum_{i=1}^{k} d_k(x_{1,k+1}^{-i})$

(Desigualdad poliédrica o desigualdad triangular generalizada para métricas fuertes)

ii.) $d_k(x_1, x_{1,k-1}) = d_k(x_{1,2}, x_{2,k-1},) = \cdots = d_k(x_{1,k-1}, x_{k-1})$

(Requerimiento de que si dos objetos son iguales d_k debe permanecer invariantes)

iii.) $d_{k-1}(x_{1,k-1}) = \frac{1}{p} d_k(x_1, x_{1,k-1})$

(Expresa que d_k y d_{k-1} son iguales hasta un factor de multiplicación *p* cuando dos objetos son idénticos)

iv.) $d_k(x_1, x_{1,k-1}) \leq d_k(x_1, x_{2,k})$

(Expresa que d_k sin objetos idénticos siempre es mayor o igual que d_k con objetos idénticos)

La Tabla 2, presenta algunos ejemplos de Multimétricas que se pueden utilizar para relaciones entre 3 elementos.

Tabla 2. Multimétricas basadas en relaciones geométricas.

Medida	Fórmula	Simetría
Área del Triángulo	$T_{XYZ} = \sqrt{s(s-d_{XY})(s-d_{YZ})(s-d_{ZX})}$ $s = \frac{d_{XY} + d_{YZ} + d_{ZX}}{2}$	S
Área Inscrita del Triángulo	$T_{XYZ} = \pi \left(\frac{2\sqrt{s(s-d_{XY})(s-d_{YZ})(s-d_{ZX})}}{d_{XY} + d_{YZ} + d_{ZX}} \right)^2$	S
Suma de lados	$T_{XYZ} = d_{XY} + d_{YZ}$	A
Ángulo de Enlace (ángulo entre los lados)	$A_X, A_Y, A_Z \; coordenadas \; de \; tres \; elementos$ $U = A_X - A_Y, V = A_Z - A_Y$ $T_{XYZ} = \alpha = arccos\left(\frac{U * V}{\lvert U \rvert * \lvert V \rvert}\right)$	A

1.2. Representaciones gráficas y descriptores moleculares

1.2.1. Teoría de Grafos

1.2.1.1. Conceptos generales

El concepto matemático de grafo se puede definir de la siguiente forma: Sea V un conjunto finito de vértices y E el conjunto de aristas que unen pares no ordenados de los elementos de V. En términos matemáticos un grafo es definido como G = (V, E) [27], [28]. Si $a_k = \{v_i, v_j\}$ es una arista, entonces los vértices v_i, v_j, se llaman extremos de la arista a_k. Los

vértices v_i, v_j se llaman adyacentes si existe una arista a_k tal que $a_k = \{v_i, v_j\} \in V$, (si existe una arista que los une). Dos aristas se denominan adyacentes si ellas tienen un vértice en común [29]. El número de vértices en un grafo es designado como *n* y el número de aristas por *m*.

Un grafo ponderado o "etiquetado", se refiere a que a cada vértice v_i del grafo G = (V, E) se le adiciona en correspondencia el peso w_i de un conjunto de pesos $W = \{w_i / i = 1, 2, \ldots\}$. Como resultado, se obtiene un conjunto de los vértices ponderados $\{(v_i, w_i) / i = 1, 2, \ldots, n\}$; este procedimiento se puede aplicar de igual forma para las aristas [30].

Si en G hay aristas múltiples (vértices que están unidos por más de una arista), entonces el grafo G se llama multigrafo. Las aristas de la forma $\{v_i, v_i\}c$, se denominan lazos o bucles. Si en G hay lazos (pueden también existir aristas múltiples), entonces el grafo G se llama pseudografo (ver Figura 1).

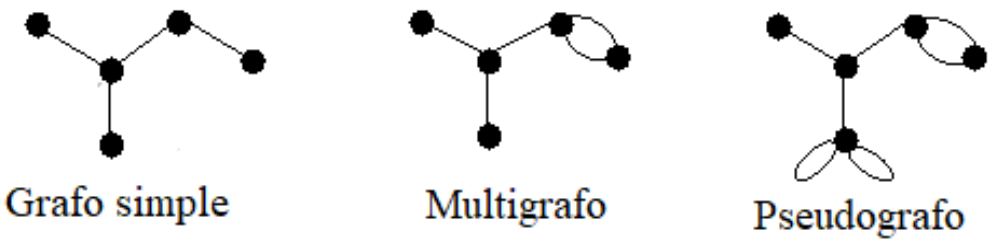


Figura 1. Tipos de Grafos

En este trabajo se empleará el concepto de pseudografo, pudiendo este ser orientado (vértices ordenados) o no orientados (consideración más general) (será la consideración que se utilizará para este estudio). En un pseudografo, el grado del vértice v_i es igual al número total de aristas (que no sean lazos) incidentes a este vértice, más el número de lazos incidentes a él [29].

1.2.1.2. Grafo químico

En la química grafo-teórica, los objetos del grafo pueden representar orbitales, átomos (o sus núcleos), enlaces, grupos de átomos, moléculas, o colecciones de moléculas. Las aristas de un grafo químico simbolizan las interacciones entre objetos químicos y se usan para definir enlaces químicos, reacciones, mecanismos de reacciones, modelos cinéticos, u otra relación [28], [31]–[33]. En la literatura existente sobre la química grafo-teórica y sus aplicaciones se puede encontrar que la principal aplicación de los grafos químicos es para la generación de índices estructurales para los estudios de estructura-actividad [34]–[40].

1.2.1.3. Empleo de matrices para la representación de grafos moleculares

Los grafos moleculares son una representación no numérica de la estructura química, por este motivo, es importante realizar una transformación de estos mediante el uso de matrices [41], [42]; esta transformación es importante ya que así, se puede realizar la manipulación computacional de las proteínas. Existen varios tipos de matrices que pueden estar asociadas a un grafo molecular como son las matrices de adyacencia y la matriz de distancia; a continuación, se presentaran sus respectivas definiciones.

La matriz de adyacencia A = A(G) del grafo G con *n* vértices, es la matriz cuadrada simétrica *n* x *n* y los elementos $[A(G)]_{ij}$ se definen de la siguiente forma:

$$[A(G)]_{ij} = 1 \text{ si } i \neq j \text{ y } e_{ij} \in E$$
$$= 0 \text{ si } i=j \text{ o } e_{ij} \notin E \qquad (18)$$

donde E representa el conjunto de las aristas de G. En la matriz de adyacencia A(G) la fila i y columna i corresponden al vértice v_i de G.

La matriz de adyacencia $A(G_w)$ del grafo molecular G con vértices y aristas ponderadas (con *n* vértices) es la matriz simétrica *n* x *n* (cuadrada) y los elementos $[A(w)]_{ij}$ se definen de la siguiente forma [42]:

$$[A(Gw)]_{ij} = V(w)w_i \text{ si } i=j$$
$$= E(w)w_{ij} \text{ si } e_{ij} \in E \quad (19)$$
$$= 0 \text{ si } e_{ij} \notin E$$

donde $V(w)w_i$ es el peso del vértice v_i, $E(w)w_{ij}$ es el peso de la arista e_{ij}, y w es un determinado peso o etiqueta que se utilice para computar V_w y E_w.

La matriz de distancia D = D(G) de un grafo G con n vértices, es la matriz simétrica n-dimensional (cuadrada) y los elementos [D]ij…n se definen de la siguiente forma [37], [41]:

$$[D]ij\ldots n = dij\ldots n \text{ si } i \neq j \neq \ldots \neq n$$
$$= 0 \text{ si } i=j=\ldots=n \quad (20)$$

donde dij…n es la métrica o la multimétrica definida entre los vértices de G. Si se considera el centro de masa de proteína como referencia para definir las distancias, los elementos i=j=…=n serían diferentes de cero.

1.2.2. Descriptores moleculares

1.2.2.1. Definición

Los descriptores moleculares se pueden definir como el número resultante de un procedimiento lógico-matemático que transforma la información contenida en una representación molecular de una sustancia para utilizarla como herramienta para la predicción de una propiedad fisicoquímica o una función de interés [12]. Se define como descriptores moleculares geométricos o tridimensionales a los descriptores resultantes de la extracción de información de una representación espacial que contiene coordenadas cartesianas (x,y,z) para cada átomo constituyente de la estructura [12]. Existe un alto número de descriptores para moléculas orgánicas, tanto de tipo topológico como geométrico y mixto; sin embargo, el desarrollo de descriptores moleculares para biomoléculas es un campo novedoso y poco explorado [43].

La posibilidad de describir las diferentes interacciones químicas existentes en moléculas biológicas mediante el uso de herramientas matemáticas es una opción viable debido a los recursos disponibles a la fecha [44], [45].

1.2.2.2. Clasificación de los descriptores moleculares por su dimensión

Los descriptores moleculares se pueden clasificar de forma general según las dimensiones que abarcan: Descriptores Constitucionales (0D), que se obtienen directamente de la fórmula molecular y son independientes de la estructura; Descriptores Unidimensionales (1D) que dan información acerca de los fragmentos estructurales de la molécula; Descriptores Bidimensionales o Invariantes de Grafos (2D) que se basan en la representación topológica (pseudografo) y Descriptores Tridimensionales (3D), que consideran la configuración espacial de la molécula además de la conectividad [46].

1.2.3. Índices algebraicos para moléculas orgánicas y su software

Marrero-Ponce *et al.* introdujeron nuevos conjuntos de descriptores moleculares para estudios de estructura-actividad considerando características relacionadas a la topología (2D) de moléculas orgánicas [47]–[51]. Estos descriptores se obtienen codificando información de la estructura mediante el uso de formas algebraicas y utilizando matrices de densidad electrónica grafo-teórica. Basado en su desempeño, y buscando generalizar la teoría matemática utilizada (formas algebraicas N-lineales, relacionadas al algebra tensorial), se planteó realizar la definición de descriptores moleculares geométricos (3D) para moléculas orgánicas, que permitan la utilización de formas algebraicas N-lineales así como ciertas consideraciones matemáticas adicionales tal como el uso de métricas y operadores de agregación, para poder aportar más información a los índices resultantes [18], [52], [53]. Se generó una herramienta computacional denominada QuBiLs MIDAS (Quadratic, Bilinear and N-Linear Maps based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings), que permitió el cálculo de los índices utilizando la teoría propuesta. Este

software se desarrolló en Java, convirtiéndolo en un software multiplataforma; cabe mencionar que este software está disponible en la red [54].

1.3. QSAR y Métodos estadísticos y de aprendizaje automático utilizados

1.3.1. Modelos QSAR

Las relaciones cuantitativas estructura-actividad (QSAR) son el resultado final de un proceso que comienza con la descripción adecuada de estructuras moleculares y termina con hipótesis y predicciones del comportamiento de las moléculas en sistemas biológicos, fisicoquímicos y ambientales. Los modelos QSAR se basan asumiendo que la estructura de una molécula debe contener las variables responsables de sus propiedades físicas, químicas y biológicas, y que esta información se extrae mediante descriptores moleculares [55].

En estos estudios, los descriptores moleculares (**X**) se correlacionan con una variable respuesta (**Y**). Es decir, este análisis puede definirse como una aplicación de métodos matemáticos y estadísticos al problema de encontrar una ecuación empírica de la forma $\mathbf{Y}_i = \mathbf{f}_i(X_1, X_2, ...X_n)$, donde $\mathbf{Y}_i$ son las propiedades y/o actividades biológicas de la molécula, y X_1, X_2, $...X_n$ son propiedades estructurales experimentales o calculadas (descriptores moleculares) de los compuestos.

1.3.1.1. Metodología general empleada en los estudios QSAR.

Los principios de la metodología QSAR pueden describirse mediante los siguientes pasos comunes: 1) Formulación del problema, en la cual se determina el objeto de análisis y el nivel de información requerido; 2) Parametrización cuantitativa de la estructura molecular de los compuestos químicos orgánicos/secuencia de biopolímeros; 3) Medición de la propiedad de interés (efectos biológicos u otros); 4) Elección del tipo de modelo QSAR que se va a desarrollar; 5) Selección de los compuestos (diseño estadístico de la serie); 6) Análisis matemático de los datos y validación interna y externa de los modelos obtenidos; 7) Interpretación de los resultados y aplicación de los modelos desarrollados al

diseño/descubrimiento de un nuevo compuesto líder, desarrollando procedimientos de cribado virtual. Sin embargo, el desarrollo de cualquier estudio QSPR/QSAR, es un ciclo iterativo [56].

1.3.1.2. Regresión lineal múltiple

Es la forma más general del análisis de regresión lineal; esta estrategia busca explicar la relación que existe entre una variable independiente (que debe ser continua) y dos o más variables independientes [57]. Una de las relaciones más importantes que se buscan en este análisis es la correlación múltiple, que se representa por **R** [58].

La expresión matemática que expresa la relación entre las variables para el caso de la regresión lineal múltiple es:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (21)$$

Donde:

Y es la variable respuesta o variable dependiente

X_n son las variables independientes

b_n son los coeficientes de regresión del modelo.

1.3.1.2.1. Principio de la parsimonia para seleccionar el número óptimo de variables

El principio de la parsimonia indica que un fenómeno debe ser descrito con un número mínimo de elementos posibles. El coeficiente de correlación aumenta en medida en que se añaden más términos independientes a la ecuación, sin embargo, existe un número de variables en el que, a pesar de que se agreguen más variables, el coeficiente de correlación no aumentará significativamente.

Existen varios procedimientos para seleccionar el número óptimo de variables a incluir en un modelo de regresión lineal múltiple; entre estos se pueden mencionar los métodos *forward selection*, *backward elimination* y *stepwise selection* [59]. Este último método es normalmente el más usado ya que es una combinación de los anteriores.

1.3.1.2.2. Análisis de la varianza

El análisis de varianza o ANOVA (*ANalysis Of VAriance*) es una técnica estadística que se usa para verificar si las medias de dos o más grupos son significativamente diferentes entre ellas. La variabilidad total de la variable dependiente se divide entre la parte atribuible a la regresión y la parte residual. La distancia de un punto cualquiera Yi al promedio de los puntos predichos $\bar{Y}$, se subdivide en dos partes [58]:

$$Y_i - \bar{Y} = \left(Y_I - \bar{\bar{Y}}_i\right) + \left(\bar{\bar{Y}}_i - \bar{Y}\right) \quad (22)$$

Siendo $\bar{\bar{Y}}_i$ el valor predicho por la ecuación de predicción. El valor $\left(Y_I - \bar{\bar{Y}}_i\right)$, se denomina residual de la regresión. Este valor sería cero si la recta pasa exactamente por encima del punto Yi. El otro valor $\left(\bar{\bar{Y}}_i - \bar{Y}\right)$, corresponde a la estimación de Yi mediante la recta de regresión.

En ANOVA, F viene dado por:

$$F = \frac{MC_{regresion}}{MC_{residual}} \quad (23)$$

Donde MC es la media cuadrática que se obtiene dividiendo la estimación/residual al cuadrado para el número de grados de libertad.

Este factor (F) sirve para comprobar si el modelo de regresión ajusta a los datos y permite evaluar si se rechaza la hipótesis nula (R^2=0). Si el modelo se ajusta a los datos, el coeficiente de correlación (R^2) se puede calcular a partir de la suma de los cuadrados (SC) del ANOVA mediante la siguiente ecuación:

$$R^2 = 1 - \frac{SC_{residual}}{SC_{regresión}} \quad (24)$$

1.3.1.2.3. Importancia de la tolerancia en la regresión lineal múltiple

La tolerancia es una medida del grado de asociación lineal entre las variables independientes. Para la variable i, la tolerancia es igual a 1-R_i^2, donde R_i^2es el coeficiente de correlación al cuadrado entre la variable *i* considerada como variable dependiente y las demás variables independientes. Valores bajos en la tolerancia, indican que la variable *i* puede ser considerada como una combinación lineal de las otras variables independientes [60].

1.3.1.3. Análisis de Discriminante Lineal

El análisis de discriminante lineal (LDA) por sus siglas en inglés, es una técnica de clasificación que busca asegurar la máxima separabilidad posible de las variables (conocida sus características) mediante la asignación de estas a un grupo definido a priori, que contiene una serie de observaciones para cada individuo referidas a un conjunto de variables relevantes. En base a esta información se calcula una función discriminante (FD) que es una ecuación lineal con una variable dependiente que representa la pertenencia a un grupo; esta es la base para la clasificación de las variables en los grupos. [56], [61]–[63].

1.3.1.3.1. Estimación de los coeficientes

La información que contienen todas las variables independientes se analiza conjuntamente para obtener los coeficientes. Se trata de conseguir un promedio ponderado de las variables para obtener una puntuación que permita distinguir entre grupos. Dados dos grupos de compuestos, uno activo y otro pasivo, se obtienen dos funciones de clasificación:

$$D_1 = a_1X_1 + a_2X_2 + a_3X_3 + \cdots \quad (25)$$

$$D_2 = b_1X_1 + b_2X_2 + b_3X_3 + \cdots \quad (26)$$

Los coeficientes, a_i y b_i son los llamados pesos discriminantes. Estos coeficientes se obtienen por el procedimiento de regresión lineal múltiple. Esta función describe una linea entre los grupos [56], [61].

1.3.1.3.2. Matriz de clasificación de casos

Los resultados de la clasificación con la función discriminante se pueden expresar matricialmente y se conoce como matriz de clasificación o de confusión. Esta matriz presenta el porcentaje de casos bien clasificados para cada grupo y de forma total. Esta matriz consta de 4 clases importantes para su definición: Verdaderos Postivos (VP), Verdaderos Negativos (VN), Falsos Postivos (FP) y Falsos Negativos (FN). La evaluación de la función para todas las variables de interés en el análisis es un criterio importante para la validación de una función discriminante. Usualmente se exige que el porcentaje de casos bien clasificados no sea inferior a 75% para que el criterio de clasificación sea considerado aceptable. [12], [61].

1.3.1.3.3. Análisis de la hipótesis del LDA

La hipótesis nula en el LDA puede formularse de la siguiente forma: no existen diferencias significativas entre las medias de las puntuaciones discriminantes de los grupos. Existen varias pruebas para comprobar la hipótesis estadística presentada, entre algunos ejemplos se citan la lambda (λ) de Wilks, el cuadrado de la distancia de Mahalanobis (D2) y el factor de Fisher (F) [63].

La lambda de Wilks esta se puede definir como:

$$\lambda = \frac{Variabilidad_{intergrupos}}{Variabilidad_{total}} \quad (27)$$

Este valor representa la porción de la varianza total de las puntuaciones discriminantes que no ha sido explicada por la diferencia entre grupos. Este parámetro toma

valores entre 0 y 1; entre menor es lambda mayor es la diferencia entre las medias de las puntuaciones discriminantes de los grupos y esto indica el rechazo de la hipótesis nula [63].

El cuadrado de la distancia de Mahalanobis, D^2, mide la distancia al cuadrado entre los centroides de dos poblaciones. Sean p poblaciones de n individuos cada una. En cada población se conocen ciertas variables x. A cada población le corresponde una matriz de observaciones. A partir de estos datos y en notación matricial, Mahalanobis define la distancia entre los centroides de los grupos p y q mediante [63]:

$$D_{pq}^2 = (\mu_p - \mu_q)' \Sigma^{-1} (\mu_p - \mu_q) \quad (28)$$

donde:

μ_p, μ_q son los vectores columna que contienen las medias de las variables de los grupos respectivos.

Σ^{-1} es la matriz inversa de la varianza intragrupos.

(‘) indica la matriz transpuesta.

A partir de la D^2, se puede estimar la F de Fischer y utilizarla como prueba de contraste.

1.3.1.3.4. Criterios para la selección de variables en LDA

Los principales criterios para seleccionar variables se presentan a continuación: a) se selecciona a la variable que minimice la lambda de Wilks, b) se selecciona la variable que maximice la D^2 de Mahalanobis entre los grupos próximos, c) se selecciona la variable que maximice la menor F entre pares de grupos y d) se selecciona la variable que minimice la suma de la variación no explicada entre grupos.

1.3.1.4. Multicolinealidad entre variables con el uso RLM y LDA

La multicolinealidad es un término que describe que las variables que se escoge para un modelo de regresión lineal se aproximan a ser una combinación lineal de las otras, significando que existe demasiada correlación entre las variables [60], [63]. Esto tiene un efecto en la determinación de la importancia relativa de las variables. Uno de los métodos más utilizados es la tolerancia. El nivel de colinealidad que se ha reportado aceptable entre las variables utilizando el método de tolerancia está en el rango de 0.4 a 0.9 [64].

1.3.1.5. Compuestos outliers y técnicas para la selección de estos

Los *outliers* son variables que, al ser calculadas por el modelo de regresión, son pobremente predichos o no se ajustan de manera adecuada, generando una afectación en los parámetros estadísticos del modelo [65]. Existen varias técnicas para detectar la presencia de *outliers* tales como: análisis de los residuales estandarizados, el método de Leverage, y el método *leave one out* [66].

1.3.1.6. Procedimientos para validar modelos QSAR

Para que un modelo QSAR pueda considerarse como confiable, estos deben ser estadísticamente robustos tanto para la data con la que se entrenó el modelo como para la data que se utilice para la predicción externa. En esta sección se analizará diferentes procedimientos estadísticos utilizados para validar modelos QSAR.

1.3.1.6.1. Procedimiento de validación cruzada dejando un elemento fuera

La validación cruzada consiste en dividir un conjunto de variables denominadas de "entrenamiento" en n elementos de prueba no repetidos. Se escoge uno de estos elementos, se lo excluye del grupo inicial y se toman los elementos restantes como conjunto de entrenamiento. Cuando solo se deja un elemento en el grupo excluido, este procedimiento se conoce como *leave one out* y se calcula según la siguiente fórmula [66]:

$$Q^2_{loo} = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - Y_i^{"}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)} \quad (29)$$

donde:

Q^2_{loo}=coeficiente de validación cruzada “leave one out”

PRESS= Suma al cuadrado de los residuos de la predicción

TSS= Suma total de cuadrados

Y_i , $Y_i^{"}$ =es la respuesta observada y estimada del i-ésimo término

$\bar{Y}$, es el promedio de las respuestas del conjunto de entrenamiento.

1.3.1.6.2. Procedimiento de re-muestreo

En la técnica de re-muestreo o *bootstrapping* se determinan aleatoriamente conjuntos de entrenamiento y predicción, de manera que el conjunto de entrenamiento preserva el tamaño original del conjunto de datos (n) y se conforma mediante la selección de n objetos con repetición, mientras que el conjunto de prueba está constituido por los objetos no seleccionados. Este procedimiento se repite miles de veces y en cada iteración se calcula en PRESS, para finalmente determinar el poder predictivo promedio (Q^2_{boot}) [67], [68].

1.3.1.6.3. Procedimiento de intercambio de variables dependientes (*Y-Scrambling*)

El objetivo de este procedimiento es comprobar si un modelo posee variables que están aleatoriamente correlacionadas a las variables de respuesta. La prueba se realiza calculando de un nuevo modelo obtenido con las variables de entrenamiento del modelo original, pero modificando las variables respuesta aleatoriamente. Si el modelo original no se conformó considerando una correlación casual, la diferencia en la correlación entre el modelo nuevo y el original es cercana a cero [57], [66].

1.3.1.6.4. Procedimiento de validación externa

El conjunto de datos conocido como serie externa, busca evaluar si los modelos obtenidos con una serie de entrenamiento pueden ser generalizables para otras series de datos relacionadas. De esta forma, cuando se evalúa este conjunto de validación externa, obtenemos un coeficiente de correlación externo y una desviación estándar externa. Entre mayor sea el coeficiente de correlación externo, mayor poder predictivo tiene el modelo [66].

1.3.1.6.5. Parámetros estadísticos utilizados para la validación de modelos LDA

Existen varios parámetros que permiten la validación de los modelos de clasificación que utilizan LDA como estrategia de modelación. Entre esos tenemos los siguientes: la exactitud total (**Ac**), el coeficiente de correlación de Mattews (**MCC**), la sensibilidad (**Sens**), la especificidad (**Spec**) y la razón de falsos positivos (**$Fp_{razón}$**) [69]. Las ecuaciones de estos parámetros se indican a continuación:

$$\mathrm{Ac} = 100 * \frac{(\mathrm{VN} + \mathrm{VP})}{\mathrm{VP} + \mathrm{FP} + \mathrm{VN} + \mathrm{FN}} \quad (30)$$

$$\mathrm{MCC} = 100 * \frac{(\mathrm{VP} - \mathrm{VN}) - (\mathrm{FP} - \mathrm{FN})}{\sqrt{(\mathrm{VN} + \mathrm{FN}) * (\mathrm{VN} + \mathrm{FP})(\mathrm{VP} + \mathrm{FP})(\mathrm{VP} + \mathrm{FN})}} \quad (31)$$

$$\mathrm{Sens} = 100 * \frac{\mathrm{VP}}{(\mathrm{VP} + \mathrm{FN})} \quad (32)$$

$$Spec = 100 * \frac{VP}{(VP + FP)} \quad (33)$$

$$Fp_{razón} = 100 * \frac{FP}{(FP + VN)} \quad (34)$$

donde VP y VN son los verdaderos positivos y negativos y FP y FN son los falsos positivos y negativos.

1.3.2. Selección de variables y reducción de dimensionalidad

1.3.2.1. Métodos de reducción de variables no supervisados

1.3.2.1.1. Análisis de componentes principales

El análisis de componentes principales (ACP) es un procedimiento que permite la transformación de un conjunto de variables correlacionadas en un conjunto menor de variables no correlacionadas denominado componente principal. Este método permite definir el grado de redundancia que pueda existir entre dicho grupo de variables; de esta forma, se busca encontrar los componentes principales que expliquen la mayor cantidad de varianza posible de las variables originales. El primer componente extraído contiene la mayor varianza de todos los componentes escogidos; a partir del segundo componente, las variables que se encuentren en el componente explicarán la mayor varianza que el primer componente no consideró y las variables que contenga serán ortogonales a las del grupo anterior. Esta herramienta es muy útil para determinar *a priori* la posible existencia de colinealidad entre los índices [70], [71]. Cabe mencionar que una de las propiedades más deseables para los descriptores moleculares es que estos sean linealmente independientes con respecto a otros [72].

1.3.2.1.2. Análisis de variabilidad basado en entropía de Shannon

El método de Análisis de Variabilidad está basado en el cálculo de la entropía de Shannon (SE). La entropía de Shannon cuantifica el grado de incertidumbre de las variables en un conjunto y su expresión matemática se indica a continuación:

$$SE(X) = -\sum_{i=1}^{N} p_i \log_2 p_i \qquad (35)$$

donde: p_i es la probabilidad de que una variable escogida al azar pertenezca a un intervalo discreto i y N es el número de intervalos discretos [73].

Basado en esta definición, si una variable contiene mayor valor de entropía de Shannon, este contiene más contenido de información, lo que es una propiedad deseable cuando las variables se utilizan en el modelado de propiedades fisicoquímicas [72]. Además de evaluar las variables con la entropía de Shannon, existe otro procedimiento conocido para la reducción de la dimensionalidad basada en teoría de información que se conoce como el índice de Gini (gSE); este modelo evalúa la diversidad de la información contenida en una variable, y se caracteriza por un incremento en su valor cuando existe un incremento en la diversidad de las instancias. La expresión matemática del índice de Gini se indica a continuación:

$$gSE(X) = \sum_{i=1}^{N} p_i * p_{i+1} \qquad 0 \leq gSE \leq \frac{N-1}{2N} \qquad (36)$$

Estos procedimientos se han empleado en la literatura para evaluar el desempeño no supervisado de conjuntos de descriptores moleculares que se han generado mediante software especializado [74], [75].

1.3.2.2. Métodos de reducción de variables supervisados

1.3.2.2.1. Métodos de subconjunto (filtro)

La estrategia conocida como método de filtrado realiza una evaluación independiente basada en característica generales del conjunto de datos, filtrando las variables para generar los mejores subconjuntos. Los métodos de subconjunto toman un conjunto de atributos o variables y dan como resultado una medida numérica que guía la búsqueda para la reducción de la dimensionalidad. Este método evalúa la habilidad predictiva de cada atributo individualmente y el grado de redundancia que este tiene en su conjunto, prefiriendo conjunto que estén altamente correlacionados con la variable respuesta pero que tengan baja correlación entre los elementos. Existen varios tipos de método de búsqueda como *Best First*, *Greedy Stepwise*, *Evolutionary Search*, entre otros [76].

1.3.2.2.2. Métodos de subconjunto (*Wrapper*)

La estrategia conocida como método *wrapper* realiza una evaluación de un conjunto usando un algoritmo de aprendizaje automatizado que se utilizará para el modelado. Se denomina *wrapper* ya que el algoritmo de aprendizaje se "envuelve" en el proceso de selección de variables, pero requiere también del uso de un método de búsqueda al igual que en los métodos de filtro. Hacer una evaluación independiente de un grupo de variables fuera sencillo si existiera una buena forma de determinar cuándo un atributo es relevante para predecir la variable respuesta, por esto este tipo de método de subconjunto es importante para la selección de atributos.[76]

1.4. Descriptores para macromoléculas

La bioinformática es un campo reciente que busca estudiar la relación entre la secuencia de los varios polímeros biológicos (ADN, ARN, proteínas) y propiedades o funciones de interés. Entre las propiedades o funciones podemos encontrar plegamiento, unión a sustratos o inhibidores, resistencia a fármacos, defectos genéticos, clasificación de especies bioquímicas entre otras ([77], [78]. El primer estudio relacionado en este campo buscaba la representación cuantitativa la secuencia de aminoácidos; como resultado, Helleberg *et al*. describieron los 20 aminoácidos del código genético utilizando 29 propiedades. En este estudio se utilizó el análisis de componentes principales (ACP) para obtener 3 escalas de propiedades principales (z_1, z_2, z_3) para cada aminoácido [79]; los autores demostraron también que estas escalas pueden ser usadas para obtener representaciones cuantitativas de las familias de péptidos y proteínas [80], [81]. Este trabajo fue extendido por Sandberg *et al*., para un total de 87 aminoácidos (20 esenciales y otros 67), obteniendo un total de 5 escalas, donde las primeras 3 presentan una alta correlación con las 3 primeras escalas propuestas (z_1, z_2, z_3) [82]. En otro estudio, se realizó la transformación de una secuencia de aminoácidos a una representación numérica usando una escala definida por

Sandberg *et al*., permitiendo clasificar entre 2 clases, 190 secuencias de RNAt de dos cepas diferentes de bacterias [83].

Randic *et al.*, han propuesto representaciones tridimensionales y cuatridimensionales de la secuencia del ADN y su caracterización numérica [84], [85] para su transformación a representaciones matriciales, que permiten definir invariantes matriciales para facilitar la comparación de representaciones de las estructuras principales de aminoácidos. Así mismo, Marrero-Ponce *et al*., introdujo una nueva invariante grafo-teórica para la descripción de ácidos nucleicos y proteínas mediante el uso de la matriz de adyacencia del grafo biomolecular. Esta matriz permite la codificación de la información para las biomoléculas con el fin de generar índices o descriptores [86].

1.4.1. Descriptores para proteínas

Los descriptores moleculares para proteínas que consideran información topológica (2D) se fundamentan en el concepto de secuencia; estos descriptores buscan explicar las características fisicoquímicas de las proteínas y su función mediante la secuencia de los aminoácidos. Existen descriptores para proteínas basados en la composición de aminoácido utilizando herramientas estadísticas [87]–[91], la frecuencia de residuos y parámetros relacionados [92]–[94], mediante minimización de energía libre y métodos *ab-initio* ([95]–[98].

Se han propuesto varios descriptores moleculares 3D para proteínas que codifican la información espacial de estas para describir varias propiedades entre las cuales se pueden citar los descriptores basados en entropía de cadenas de Markov [99], los potenciales electrostáticos de Markov [100], descriptores basados en la estabilidad termodinámica [101], descriptores bilineales [19], entre otros. Adicionalmente a estos descriptores, en recientes publicaciones, se destacan las redes de contactos de proteínas como una estrategia para la descripción estructural de las proteínas [44], [102].

1.4.2. *Software* para el cálculo de descriptores de proteínas

El número de descriptores moleculares para proteínas, así como los *softwares* disponibles para su cálculo es limitado comparativamente con la cantidad de información disponible en el campo de las proteínas. Existen varios *softwares* para cálculo de descriptores de proteínas que son en base web como: PROFEAT[103], PseAAC [104] y existen *softwares* para cálculo de descriptores que son para descarga como: ProtDCal [105] y protr [106].

1.4.3. Aplicaciones utilizadas usando descriptores de proteínas

1.4.3.1. Velocidad de plegamiento de proteínas

En los años 50, Pauling enuncia que ciertas proteínas existen como una estructura helicoidal energéticamente estable [107]. En 1953, Watson y Crick probaron la existencia de una conformación helicoidal para una molécula de ácido desoxirribonucleico (ADN); esta molécula exhibía características similares a las descritas por Linus Pauling [108]. La conformación experimental del supuesto de Pauling se dio en 1958 utilizando una molécula de mioglobina [109].

Para 1961, los experimentos llevados a cabo por el científico Christian Anfinsen demostraron que la estructura de la ribonucleasa-pancreática-bovina-A (RNasa-A) se puede volver a plegar espontáneamente luego de retirar las condiciones que la desnaturalizaron [110] Este trabajo permitió observar que la estructura nativa de las proteínas puede ser estudiada experimentalmente, y que es un proceso reversible; además, esta investigación le confirió a Anfinsen el Premio Nobel en Química en 1973 [110], [111].

Esta serie de trabajos y experimentos definieron la formulación de lo que se define como el problema del plegamiento de las proteínas; la finalidad de este problema es encontrar la conformación geométrica del estado nativo de las proteínas a partir de su estructura primaria [112].

Una limitación que surgió a partir del problema del plegamiento de las proteínas se conoce como la Paradoja de Levinthal; esta paradoja indica que normalmente los polímeros están compuestos por un conjunto de estructuras de equilibrio, mientras que las proteínas se caracterizan por tener una sola estructura de equilibrio. Este particular hace que un estudio completo de todas las posibles conformaciones tomaría una inmensa cantidad de tiempo, por lo que las proteínas nunca podrían encontrar dicha estructura. Levinthal propuso la existencia de una ruta auto guiada y definida que permitiría que la proteína alcance el estado de menor energía [113].

En 1998, Plaxco *et al*. presentaron una investigación que correlacionaba el logaritmo de la constante de velocidad de plegamiento y la estructura nativa de las proteínas expresada por un descriptor denominado *Contact Order* [114]. Este estudio fue el inicio para la búsqueda de nuevas explicaciones que relacionen a la cinética de plegamiento con la estructura geométrica de las proteínas. Descriptores como el *Long Range Order* (LRO) que considera los contactos de tipo geométrico como factores para el plegamiento [115], *Total Contact Distance* (TCD) que combina LRO y CO para considerar contactos cercanos de la proteína [116], *Chain Topological Parameter* que correlaciona el plegamiento de proteínas mejor que CO en un rango establecido [117], *Contact Number* que considera el número de aminoácidos en contacto con una distancia espacial menor a 6,5 A [118], *Folding Degree* que considera un modelo basado en la ley de Hooke para describir el plegamiento [101], entre otros.

1.4.3.2. Clasificación SCOP de proteínas

La clasificación SCOP (*Structural classification of proteins*) provee una descripción comprensiva y detallada de las relaciones estructurales de las proteínas que han sido determinada tridimensionalmente. El método usado para la clasificación en SCOP es esencialmente la inspección visual y comparación de las estructuras a través de herramientas

automáticas que hicieron la tarea manejable y ayudó en la generalización. La clasificación tiene niveles jerárquicos que encierran las relaciones estructurales y evolucionarias; estos niveles se detallan a continuación:

Familia: Las proteínas se aglomeran en base a dos criterios que implican que tienen un origen evolutivo similar; las proteínas que tienen identificados sus residuos en más de 30% y las proteínas que tienen caracterizados sus residuos con menos del 15%.

Plegamiento común: Super familias y familias están agrupadas como si tuvieran un mismo plegamiento si sus estructuras secundarias son similares y si sus conexiones topológicas son similares.

Clase: Para conveniencia del usuario, los diferentes tipos de plegamiento han sido agrupados en clases. La mayoría de los plegamientos han sido asignados a una de las 5 clases estructurales basadas en la estructura secundaria las cuales son: **todo alfa** (proteínas cuya estructura es esencialmente hélices alfa), **todo beta** (proteínas cuya estructura esencialmente son hojas beta), **alfa y beta** (proteínas que tienen estructuras alfa y beta interrelacionadas), **alfa mas beta** (proteínas que tienen estructuras alfa y beta en segmentos definidos) y **multidominio** [119].

1.5. Conclusiones parciales

Las herramientas algebraicas (lineales y N-lineales) permiten realizar transformaciones entre espacios vectoriales de dimensiones distintas; este concepto, aplicado a la ciencia de proteínas, puede generar nuevos descriptores geométricos que contengan más información respecto a la conformación y secuencia que otros descriptores reportados al momento. Se ha observado además que el número de descriptores moleculares para proteínas es limitado, por lo que es un campo que debe ser estudiado. La predicción de propiedades y funciones de las proteínas tiene muchos campos de aplicación en la actualidad, como

consecuencia, es importante generar herramientas teóricas que permitan el estudio de proteínas mediante el uso de descriptores moleculares, técnicas estadísticas y modelos de estructura-actividad.

2. MATERIALES Y MÉTODOS

2.1. Conjunto de datos utilizados para la generación de los índices 3D para proteínas

Con la finalidad de validar la teoría propuesta para el cálculo de descriptores macromoleculares 3D para proteínas, se debe evaluar la generación de estos descriptores mediante el uso de diferentes conjuntos de datos (conjunto de proteínas) que permitan la comparación de los resultados que se obtienen con esta propuesta contra los resultados reportados por otros autores.

2.1.1. Conjunto de proteínas para modelar el plegamiento de proteínas

Se tomó como referencia para la experimentación y modelado de la velocidad de plegamiento, una data de 80 proteínas (45 proteínas de plegamiento en dos etapas y 35 proteínas con plegamiento multietapa) debido a que este grupo de proteínas tiene velocidades de plegamiento sobre ocho ordenes de magnitud, lo que permite comparar las proteínas que tienen plegamiento en dos etapas con las de multietapa [118]. Además, se consideraron 17 proteínas que se extrajeron del KineticDB data base [120], que tenían las siguientes características: 1) la diferencia entre la longitud de la secuencia para cada estructura del PDB y la experimental no debe ser mayor a 10 residuos, 2) la secuencia estudiada y el pdb no deben tener diferencias y 3) las condiciones experimentales usadas para medir la velocidad de plegamiento deben corresponder a las condiciones propuestas por Maxwell et al.[101]

2.1.2. Conjunto de proteínas para realizar la clasificación estructural de proteínas

Según K.C. Chou, la clasificación SCOP es más natural, no solamente considera porcentajes de estructuras secundarias, y puede ser más confiable para su uso en la predicción de clases estructurales Por este motivo, Chou generó una conjunto de datos de 204 proteínas que contiene las 4 categorías estructurales de clase de la clasificación SCOP (52 todo alfa, 61 todo beta, 45 alfa/beta, 46 alfa + beta), con el fin de poder realizar estudios de predicción de estructura en proteínas [88]. Este set se divide en 155 proteínas que se utilizarán en la serie de entrenamiento y 49 proteínas en la serie de predicción.

2.1.3. Conjunto de proteínas para evaluar el rendimiento del software

Este conjunto de datos se compone de 152 proteínas. Está diseñado para realizar estudios supervisados y no supervisados de experimentos de velocidad de cálculo, análisis de cantidad de información mediante herramientas de teoría de información, Análisis de Componentes Principales.

2.2. *Software* que se utilizó para la generación de los nuevos descriptores 3D para proteínas

2.2.1. Generalidades

ToMoCoMD-CAMPS es una aplicación libre, multi plataforma, interactiva y amigable con el usuario, diseñada para calcular descriptores macromoleculares de proteínas y péptidos, con el objetivo de caracterizar dichas estructuras. Este *software* está compuesto de una suite denominada MuLiMs (acrónimo de *Multi-Linear Maps*) que incluye un módulo denominado MCoMPAs (acrónimo de *Multi-Linear Maps base don N-Metric and Contact Matrices of 3D-Protein and Amino acid Weightings*). Este *software* permite el cálculo de descriptores 3D para proteínas basado en algebraicas bilineales y trilineales. Consecuentemente, es el único software que permite el cálculo de este tipo de descriptores,

pudiendo utilizar relaciones entre dos y tres aminoácidos, aplicar varias métricas y multimétricas para la definición de la distancia entre aminoácidos, transformaciones probabilísticas (simple estocástica, probabilidad mutua), *cut-offs*, cálculo de descriptores por grupo de aminoácidos y el uso de operadores de agregación.

Este *software* está desarrollado utilizando lenguaje de programación Java y empleando la librería *Chemistry Development Kit* (CDK) para la manipulación de las estructuras químicas y el cálculo de propiedades atómicas. Además, el software está compuesto de una interfaz gráfica amigable con el usuario y una librería API (*Application Program Interface*). Adicionalmente, el *software* tiene la característica de realizar cálculos paralelos en todos los procesadores disponibles de la computadora.

2.2.2. Procedimiento seguido por el software para la generación de descriptores

Entrada: Archivo *Protein Data Bank* (PDB)

Salida: Archivo con los descriptores 3D-proteicos en formato (CSV, ARFF o TXT)

Aspectos externos

1. Generación de la representación 3D-proteica (Archivo(s) con extensión PDBX)

a. Selección del modelo.

b. Selección de la(s) cadena(s) polipeptídica(s).

c. Selección de la representación 3D-proteica.

d. Visualización (opcional) en (texto plano y en 3D) de la representación seleccionada.

2. Carga de los archivos que contienen la representación elegida.
3. Configuración de los parámetros para calcular los 3D-DMs

a. Selección de la(s) forma(s) algebraica(s).

b. Selección del(los) enfoque(s) matricial(es).

c. Selección de la(s) métrica(s) para el cálculo de la distancia inter-aminoácido.

d. Selección de los órdenes (parámetro *k*) de la matriz.

e. Selección o no de los procedimientos de cortes (geométrico y topológico) macromoleculares.

f. Selección del tipo de índices a calcular: totales (proteína como un todo), locales (fragmentos), o ambos.

g. Selección de la(s) propiedad(es) de la cadena lateral de aminoácidos.

h. Selección del(los) operador(es) de agregación de las contribuciones aminoacídicas.

Aspectos internos

4. Cálculo de la matriz de relaciones inter-aminoacídicas.

a. Uso (opcional) de cortes macromoleculares.

5. Cálculo de los vectores de propiedades.

6. Cálculo de los índices de nivel atómico.

7. Operadores de agregación de las contribuciones aminoacídicas.

8. Escritura a archivo en disco los descriptores.

2.2.3. Manejo de los descriptores obtenidos mediante el *software*

La configuración teórica asignada en la interfaz gráfica del *software* permite la definición de un proyecto. Cada proyecto genera un número definido de descriptores 3D para proteínas. Un proyecto tiene un formato de salida .xml.

Con un número de proyectos definidos y un set de proteínas, el *software* procede al cálculo de los descriptores, obteniendo un archivo de salida con los mismos de tipo .csv, .arff o .txt por cada proyecto calculado en el *software.*

Debido a la alta dimensionalidad de descriptores obtenidos, se deben efectuar procedimientos de reducción antes de utilizar los descriptores para la modelación (Ver sección 1.3.2).

2.3. *Software* que se utilizó para la reducción de dimensionalidad

2.3.1. *Software* para la reducción de la dimensionalidad utilizando teoría de información (IMMAN)

2.3.1.1. Generalidades

IMMAN (*Information Theory based Chemometric Analysis*) es una herramienta gratuita, amigable al usuario, diseñada para realizar selección de atributos supervisada y no supervisada, así como comparaciones entre conjuntos de datos utilizando medidas de teoría de información. Estas medidas de teoría de información, tanto supervisadas como no supervisadas, pueden ser usadas independientemente o utilizando relaciones multi-criterio. Además, el *software* permite el análisis gráfico de la cantidad de información contenido en un grupo de datos, utilizando varias representaciones como son los gráficos de correlación, distribución e importancia.

Este *software* está desarrollado utilizando lenguaje de programación Java y puede ser utilizado en una variedad de sistemas operativos y computadoras incluyendo *clusters* multi procesador, computadoras de escritorio o portátiles [121]. El *software* se encuentra disponible para la descarga en http://mobiosd-hub.com/imman-soft/.

2.3.1.2. Procedimiento utilizado para la reducción de la dimensionalidad

2.3.1.2.1. Reducción no supervisada de la dimensionalidad

Cada archivo obtenido del *software* MuLiMs-MCoMPAs (correspondiente a cada uno de los proyectos calculados), se unían en un solo archivo que contenía todos los descriptores. Este archivo se cargaba al software IMMAN, para realizar la reducción no supervisada de la

data. Dependiendo del número de descriptores, se buscaba reducir el 30% de la dimensionalidad utilizando la entropía de Shannon (SE) como medida. Posteriormente, se reducía un 50% de la dimensionalidad restante mediante el uso del índice de Gini (gSE).

2.3.1.2.2. Reducción supervisada de la dimensionalidad

Después de la reducción no supervisada realizada al archivo general, se procedía a agregarle la variable respuesta que se deseaba modelar para cada uno de los sets de datos [sea plegamiento de proteínas (función continua) o clasificación de proteínas (función discreta)]. Posteriormente, se cargaba el archivo nuevamente en IMMAN y se seleccionaba la reducción supervisada de la data. Dependiendo del número de descriptores, se buscaba reducir el 50% de la dimensionalidad remanente utilizando la medida *Symmetrical Uncertanty* (SU).

2.3.2. *Software* para la reducción de dimensionalidad utilizando métodos subconjunto (WEKA)

2.3.2.1. Generalidades

Weka fue desarrollado en la Universidad de Waikato en Nueva Zelanda, y el nombre representa *Waikato Environment for Knowledge Analysis*. El *software* está programado en Java y distribuido bajo los términos de GNU *General Public License*. Este *software* funciona en casi todas las plataformas y ha sido probado en los sistemas operativos Linux, Windows y Macintosh.

El *software* provee una interfaz uniforme para diferentes algoritmos de aprendizaje, junto con métodos de pre y post procesamiento para evaluar el resultado de los esquemas de aprendizaje en cualquier set de datos [76]. Weka está disponible en http://www.cs.waikato.ac.nz/ml/weka.

2.3.2.2. Procedimiento para la reducción de la dimensionalidad

Se ingresa al explorador de Weka y se carga el archivo reducido mediante procedimientos de teoría de información. Se escoge la pestaña *Select Atributes* y se selecciona la opción *CfsSubsetEval*. Se realiza la reducción supervisada utilizando un método de búsqueda que mejor se ajuste al método de modelado escogido (*Best First*, *Greedy Stepwise*, *Evolutionary Search*). Una vez que el software termina la selección de atributos, se debe reducir el archivo utilizando el filtro *Remove* y guardar este archivo para su utilización en la modelación.

2.3.3. *Software* para la reducción de la dimensionalidad utilizando Análisis de Componentes Principales (SPSS)

2.3.3.1. Generalidades

La plataforma IBM SPSS® ofrece la posibilidad de realizar análisis estadísticos avanzados, una gran librería de algoritmos de aprendizaje automatizado, análisis de texto, extensibilidad con código *open source* e integración con *big data*. La facilidad de uso, flexibilidad y escalabilidad de IBM SPSS lo hace accesible a todo tipo de usuarios, con diferentes tipos de habilidades, y conocimiento, para ayudar a procesar información en cualquier campo de la ciencia o la industria, con el fin de encontrar nuevas oportunidades de investigación y mejorar procesos actuales mediante dichas herramientas.

2.3.3.2. Procedimiento

Al archivo reducido a partir de los métodos supervisados y no supervisados basados en teoría de información, se lo separa dependiendo del tipo de herramienta matemática utilizada para la definición de estos y se carga en el software aplicando todos los procedimientos de importación para aplicar el análisis de componentes principales.

Se ingresa a la pestaña *Analyze*, se escoge la opción *Dimension Reduction* seguido de la opción *Factor* y se cargan todas las variables para el análisis. En la opción *Extraction*, se

escoge el método *Principal Components*, y que se extraiga todos los *eigenvalues* que tengan mayores a 1. En la opción *Rotation*, se escoge la herramienta *Quartimax* y se procede a realizar el análisis. Posterior a esto, se extrae los factores que permitan explicar el 85% de la varianza de todo el conjunto de datos, y se analiza factor por factor con el fin de encontrar las variables que expliquen la mayor cantidad de varianza con el fin de permitir definir configuraciones para los proyectos de cálculo del software.

2.4. *Software* que se utilizó para el modelamiento de las aplicaciones utilizando los descriptores propuestos

2.4.1. Modelamiento utilizando la técnica de regresión lineal múltiple (MOBYDIGS)

2.4.1.1. Generalidades

MobyDigs es un *software* para el cálculo de modelos de regresión usando algoritmos genéticos (AG) para la selección de variables con el fin de la obtención de un conjunto de modelos predictivos adecuados. Este *software* fue desarrollado por el "*Milano Chemometrics and QSAR Research Group*" y se ha desarrollado para Windows como una aplicación de 32 bits.

La primera versión de MobyDigs se utilizó para actividades dentro del grupo de investigación. Luego, debido a varios pedidos de grupos de investigación externo, se lo comenzó a comercializar desde 2004 [122]. La información acerca del *software* se puede encontrar en línea en://www.talete.mi.it/MobyDigs.htm

2.4.1.2. Procedimiento para el modelado utilizando MobyDigs

Para realizar el proceso de búsqueda, el procedimiento de AG se configuró con los siguientes parámetros [122]:

- Número de iteraciones igual a 100 000.
- Tamaño de la población igual a 100.

- Umbral de reproducción/mutación igual a 0.5. Este valor se varia en el rango de 0 a 1 a lo largo de la exploración para cambiar la probabilidad de estos operadores.
- Umbral de selección inicialmente igual a 0.5 (indica selección por el método de ruleta) hasta alcanzar el 80% de la cantidad de iteraciones. Posterior a esto, se varía con 0 (indica selección aleatoria) y después se establece en uno (indica selección por torneo) con el objetivo de potenciar la presión selectiva.

El conjunto de proteínas utilizado para esta aplicación fue el de velocidad de plegamiento. Se construyeron modelos de una a cuatro variables para la correspondiente aplicación, usando como función objetivo el procedimiento Q^2_{loo} (ver Sección 1.3.1.6.1). Debido a que el programa MobyDigs genera un conjunto de modelos en cada exploración, se llevaron a cabo los siguientes pasos para determinar los mejores modelos:

1. Se retuvieron desde la población generada los 4 mejores modelos de cada dimensión acorde a la función Q^2_{loo}.

2. Se le aplicó a cada modelo retenido las técnicas de validación *bootstrapping* (Q^2_{boot}) para evaluar el poder predictivo (ver Sección 1.3.1.6.2) y *Y-scrambling* (a (Q^2)) para evaluar la posible correlación aleatoria con respecto a la aplicación modelada (ver Sección 1.3.1.6.3).

3. Se seleccionó como mejores modelos a aquellos que tuviese el menor valor según la función: $f(x) = (Q^2_{loo} + Q^2_{boot}) + |a(Q^2)|$.

4. Se calculó a los mejores modelos el poder predictivo utilizando el conjunto de proteínas para predicción, con el fin de evaluar su capacidad de generalización (Q^2_{ext}) para seleccionar los modelos finales.

2.4.2. Modelamiento utilizando la técnica de Análisis discriminante lineal (WEKA)

2.4.2.1. Procedimiento para el modelado

El conjunto de proteínas utilizado para esta aplicación fue el utilizado para la clasificación estructural SCOP que contiene 204 proteínas. Se construyeron modelos con el menor número de variables, usando como función objetivo el porcentaje correcto de clasificación y el coeficiente de correlación de Mattews (ver sección 1.3.1.7.5).

Se ingresa al explorador de Weka y se carga el archivo reducido dimensionalmente tanto por métodos basados en teoría de información como mediante el uso de subconjuntos. Se escoge la opción *Select Attributes* y se selecciona la opción *WrapperSubsEvaluator*; donde se tiene que seleccionar un método de búsqueda. Posterior a esto se selecciona en clasificador, bajo la categoría *functions*, la herramienta LDA.

Una vez que el *software* obtiene los resultados de la reducción de variables, se corre la reducción y se utiliza el filtro *Remove*. Este archivo se guarda para su utilización en la modelación. Posterior a esto, se realiza el modelado en la pestaña *Classify*, donde se escoge el clasificador, que en este caso es LDA, y se selecciona un proceso de validación cruzada de 10 *folds*. Un modelo adecuado realiza una clasificación correcta de al menos el 75%, y tener el coeficiente de corrección de Mattews lo más cercano a 1. Una vez que se obtenga los resultados del entrenamiento, se debe evaluar el modelo mediante la serie de predicción. De igual manera, se debe tener en cuenta que un buen modelo de predicción debe tener un alto porcentaje de clasificación correcta y un MCC lo más cercano a uno. Los modelos escogidos tienen los porcentajes de clasificación y MCC más altos tanto en entrenamiento como en predicción. De este experimento se obtienen los resultados para el modelado de clasificación tanto en entrenamiento como en predicción.

3. RESULTADOS Y DISCUSIÓN

3.1. Definición de nuevos descriptores macromoleculares 3D y su generación

En esta sección se presenta la teoría matemática que permite la definición de los nuevos descriptores tridimensionales para proteínas y todas las herramientas matemáticas que permiten la extracción de la información de la representación geométrica de las proteínas tales como la codificación de interacciones de interés mediante cortes moleculares y descriptores de grupos de aminoácidos, procedimientos de normalización matricial y generalizaciones en la construcción de los tensores geométricos para la transformación de información y operaciones de fusión de descriptores para aminoácido.

3.1.1. Representaciones 3D proteicas

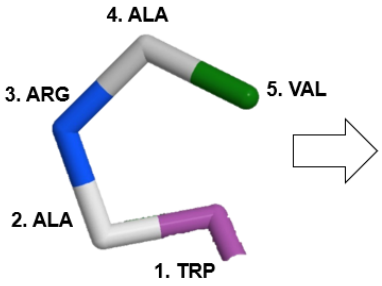
La definición de varios descriptores tridimensionales para proteínas se ha dado mediante el uso de la posición del carbono alfa (C_α) como estrategia para la representación de la posición de los aminoácidos de una proteína [12], [43], [44], [114]. En un estudio reciente, se utilizó las coordenadas del carbono beta (C_β) como estrategia para la extracción de información estructural en el desarrollo de modelos de predicción[123]. Este estudio indica que otras representaciones pueden proveer información complementaria que no es extraída mediante la representación de C_α. En base a lo anterior, en este trabajo se propone el uso de 4 representaciones espaciales de los aminoácidos: 1) C_α, 2) C_β, 3) Carbono del enlace amida (AB) y 4) pseudo-aminoácido (AVG) [obtenido del cálculo de la media aritmética de las coordenadas cartesianas de todos sus átomos (se refiere a los átomos diferentes al hidrógeno)].

3.1.2. Vector macromolecular

La representación de estructuras químicas mediante el uso de vectores moleculares ha sido explicada en detalle en otros trabajos [47], [49], [124], [125]. Consecuentemente, este concepto será adaptado a la estructura de las proteínas, considerando un aminoácido como un

elemento compuesto de la proteína. Los componentes del vector macromoleculares son valores numéricos obtenidos de las propiedades fisicoquímicas de cada aminoácido presente en la estructura de la proteína [15], [16]. En este trabajo, se utilizan varias propiedades fisicoquímicas y descriptores moleculares reportados en trabajos de investigación previos, que consideran 3 grandes grupos: propiedades estéricas, propiedades hidrofóbicas y propiedades electrónicas; se detallan algunos ejemplos a continuación: área de contacto isotrópica (ISA) [126], índice de hidropatía de Kyte-Doolittle (KDS) [127], índice de hidropatía de Hopp y Woods (HWS) [128], índice de carga electrónica (ECI) [126], punto isoeléctrico (PIE) [129], parámetros z (Z_1, Z_2, Z_3) [80], volumen molecular (MV) [130], probabilidad de hélice alfa y hoja beta (PAH y PBS) [131].

En la Figura 2 se realiza una representación del péptido 5WRX mediante tres tipos de vectores macromoleculares. Cada vector considera una propiedad aleatoria de uno de los grupos de propiedades fisicoquímicas para aminoácidos antes indicada. Como este péptido contiene 5 aminoácidos, se colocará la propiedad fisicoquímica de cada aminoácido en el orden que se encuentre en la secuencia.



Posición en la estructura Primaria	1	2	3	4	5
Amino ácidos (AA)	**TRP**	**ALA**	**ARG**	**ALA**	**VAL**
Peso Molecular (kg/kmol)	204	89	174	89	117
Índice de Hidropatía de Hoop-Woods	-3,04	-0,5	3,0	-0,5	-1.5
Índice de Carga Electrónica	1.08	0.05	1.69	0.05	0.07

Vector con propiedades estéricas $\mathbf{X} = [204 \quad 89 \quad 174 \quad 89 \quad 117]$

Vector con propiedades hidrofóbicas $\mathbf{Y} = [-3,04 \quad -0,5 \quad 3,0 \quad -0,5 \quad -1.5]$

Vector con propiedades electrónicas $\mathbf{P} = [1.08 \quad 0,05 \quad 1.69 \quad 0,05 \quad 0.07]$

Figura 2. Representación gráfica de la conformación del vector macromolecular para un péptido.

3.1.3. Matriz espacial N-tuplas de similitud-disimilitud: nueva representación geométrica de proteínas

El enfoque que se propone en el presente trabajo para la codificación de la información tridimensional (3D) de las estructuras de las proteínas es la **k-ésima matriz espacial N-tuplas de similitud-disimilitud**, como esquema tensorial general para representar la información capturada entre N aminoácidos, donde N indica el grado del tensor construido. Específicamente, son propuestas la **k-ésima matriz espacial total 2-tuplas de similitud-disimilitud** (ZB) y la **k-ésima matriz espacial total 3-tuplas de similitud-disimilitud** (ZT) para las relaciones entre dos y tres aminoácidos de una proteína, respectivamente.

El índice superior k, indica el exponente de las matrices ZB y ZT. De esta manera para $k = 0$ todas las entradas de las matrices ZB_0 y ZT_0 tienen valor 1; mientras que para $k = 1$ los coeficientes gb_{ij} y gt_{1ij} correspondientes a las matrices ZB^1 y ZT^1, que representan la información de las interacciones entre dos y tres aminoácidos, respectivamente. La definición matemática de los coeficientes zb_{ij} (ver Ecuación (37)) y zt_{1ij} (ver Ecuación (38)) se presentan a continuación.

$$zb_{ij}^{k} = \begin{cases} D_{ij} & \text{si } \boldsymbol{i} \wedge \boldsymbol{j} \textit{ no son iguales} \\ L_{ij} & \text{si } \boldsymbol{i,j} \textit{ son iguales} \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (37)$$

$$zt_{ijl}^{k} = \begin{cases} TT_{ijl} & \text{si } \boldsymbol{i} \wedge \boldsymbol{j} \wedge \boldsymbol{l} \textit{ no son iguales} \\ D_{ijl} & \text{si } \boldsymbol{i,j} \vee \boldsymbol{j,l} \vee \boldsymbol{i,l} \textit{ son iguales} \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (38)$$

donde, *i*, *j* y *l* representan aminoácidos de una proteína. D_{ij} es una métrica de similitud-disimilitud entre los aminoácidos *i* y *j* (ver Tabla 1), y TT_{ijl} son las reglas para codificar las relaciones ternarias entre los aminoácidos de una proteína.

A partir de las definiciones anteriores, se puede observar como la matriz espacial total 2-tuplas de similitud-disimilitud de orden 1 (ZB^1) constituye una generalización de la matriz de distancia geométrica debido a las múltiples métricas que se pueden emplear para su construcción (normalmente solo se utiliza la métrica euclídea).También, la matriz espacial total 3-tuplas de similitud-disimilitud de orden 1 (ZT^1) es una extensión de la matriz geométrica para representar la información tridimensional entre tres aminoácidos.

Considerando la ecuación (38) se observa que las reglas utilizadas para determinar las relaciones ternarias dependen de los aminoácidos que existen en la proteína, denotados por los subíndices *i*, *j* y *l*. Esto indica que solo se calculará una relación ternaria si los tres aminoácidos son distintos. Si no se pudiese determinar la relación ternaria, esta relación puede ser reducida a una medida inferior (relación binaria). Por este motivo, existen dos tipos de relaciones ternarias, completas y no completas; sus condiciones se indican a continuación:

$$\text{3nC (no completas)} = \begin{cases} TT_{ijl} & \text{tres aminoácidos distintos} \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (39)$$

$$\text{3C (completas)} = \begin{cases} TT_{ijl} & \text{tres aminoácidos distintos} \\ D_{ijl} & \text{dos aminoácidos distintos} \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (40)$$

donde 3C se refiere a las medidas ternarias que pueden ser reducidas (completas) y 3nC se refiere a las medidas ternarias que solamente consideran relaciones entre tres aminoácidos (No completas). T_{ijl} es la medida utilizada para establecer relaciones entre tres aminoácidos, D_{ij} es la métrica utilizada para definir la relación entre dos aminoácidos.

Es importante mencionar que, para algunas medidas ternarias, independientemente de que sean completas o no, se requiere escoger al menos una métrica para definir la distancia entre dos aminoácidos. Esto se debe a que la definición de algunas multimétricas requiere la determinación de la distancia entre dos aminoácidos (2 distancias para los 3 elementos de la relación). Para una mejor ilustración de esta operación, ver la Figura 3.

Hasta el momento solo se han definido las matrices ZB y ZT para los órdenes 0 y 1. Las matrices ZB^k y ZT^k para otros órdenes se calculan utilizando el producto de Haddamard.

El exponente k es un número entero cuyos valores pueden ser tanto positivos como negativos. Esto significa que cuando el parámetro *k* es un número negativo se calcula el recíproco para cada una de las entradas de cada elemento de las matrices de similitud-disimilitud para 2 y 3 tuplas. Este cálculo se puede aplicar para los elementos de la diagonal dependiendo del tipo de distancia que se considere (distancia al centro de masa de la proteína o distancia entre cada aminoácido). El uso de esta potencia en la matriz está inspirado en la naturaleza fisicoquímica de las diferentes interacciones no covalentes tales como las

interacciones coulombicas (*k=-2*), interacciones gravitacionales (*k=-1*) y la forma funcional del potencial del Lennard-Jones (k=*-6 a -12*).

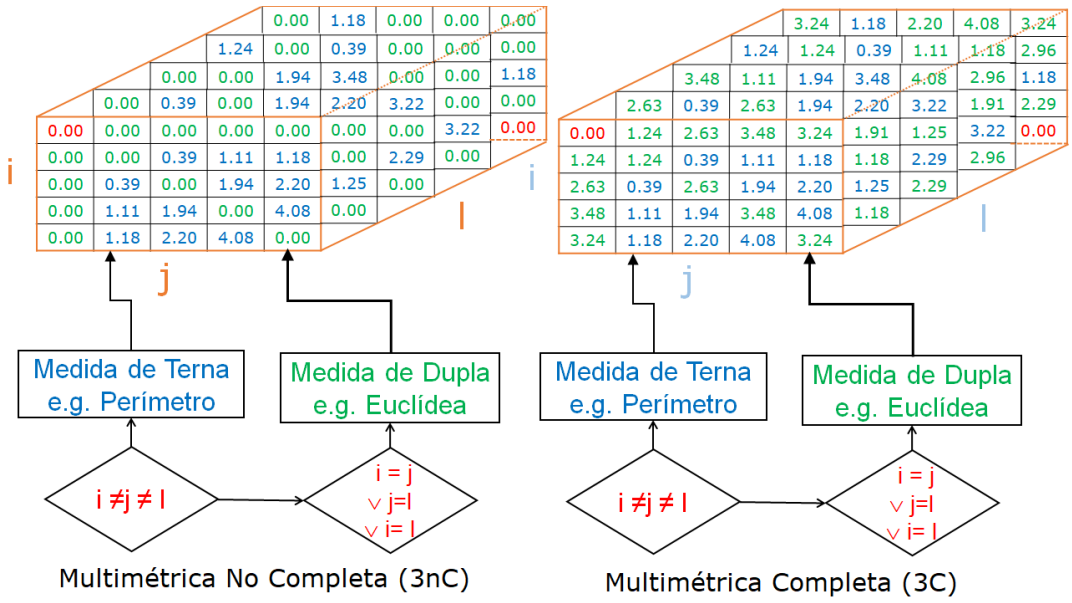


Figura 3. Esquema de ilustración para el cálculo de multimétricas completas y no completas.

3.1.4. Matriz espacial N-tuplas de similitud-disimilitud basada en tipos de aminoácidos o grupos locales.

La **k-ésima matriz espacial N-tuplas de similitud-disimilitud** propuesta para representar las relaciones entre dos (ZB^k) y tres (ZT^k) aminoácidos codifican la información correspondiente a la configuración geométrica de las proteínas independiente de los tipos de aminoácidos o grupos de aminoácidos que intervengan en las mismas. Sin embargo, es deseable considerar cuando se definen nuevos descriptores moleculares, la información correspondiente a diferentes tipos de aminoácidos, ya que muchas de las funciones de la proteína pueden estar caracterizadas por la configuración de una región en particular [72]. Con este antecedente, se propone **las k-ésimas matrices espaciales de grupo-local 2-tuplas, 3-tuplas de similitud-disimilitud**, representadas por ZB_G^k, ZT_G^k respectivamente, donde G es el grupo de aminoácidos en cuestión. Estas matrices se obtienen a partir de sus correspondientes definiciones totales como se muestra a continuación.

$$
\begin{cases}
zb_{ijG}^{k} = zb_{ij}^{k} & \text{si } \mathbf{i} \wedge \boldsymbol{j} \in \boldsymbol{G} \\
zb_{ijG}^{k} = \frac{1}{2} zb_{ij}^{k} & \text{si } \boldsymbol{i} \vee \boldsymbol{j} \in \boldsymbol{G} \\
zb_{ijG}^{k} = 0 & \text{en cualquier otro caso}
\end{cases} \qquad (41)
$$

$$
\begin{cases}
zt_{ijl-G}^{k} = zt_{ijl}^{k} & \text{si } \boldsymbol{i} \wedge \boldsymbol{j} \wedge \boldsymbol{l} \in \boldsymbol{G} \\
zt_{ijlG}^{k} = \frac{2}{3} zt_{ijl}^{k} & \text{si } \boldsymbol{i,j} \vee \boldsymbol{j,l} \vee \boldsymbol{i,l} \in \boldsymbol{G} \\
zt_{ijlG}^{k} = \frac{1}{3} zt_{ijl}^{k} & \text{si } \boldsymbol{i} \vee \boldsymbol{j} \vee \boldsymbol{l} \in \boldsymbol{G} \\
zt_{ijlG}^{k} = 0 & \text{en cualquier otro caso}
\end{cases} \qquad (42)
$$

donde, $zb_{ijG}^{k}, zt_{ijlG}^{k}$ son los coeficientes de las matrices ZB_{G}^{k}, ZT_{G}^{k} respectivamente, y $zb_{ij}^{k}, zt_{ijl}^{k}$ son las entradas de las matrices totales ZB^{k}, ZT^{k} respectivamente. Los tipos de grupos G, que se consideran en este trabajo se presentan en la Tabla 3. Cabe mencionar que las matrices locales se pueden definir considerando uno o más de los 20 aminoácidos esenciales, sin importar el grupo definido en la tabla mencionada.

Tabla 3. Grupos de Aminoácido considerados para la definición de los índices

Grupo	Aminoácidos
Apolar **(RAP)**	PRO, ILE, ALA, VAL, LEU, PHE, TRP, MET
Polar Cargado Positivamente **(RPC)**	LYS, HIS, ARG
Polar Cargado Negativamente **(RNC)**	ASP, GLU
Polar Sin Carga **(RPC)**	ASN, CYS, GLY, SER, THR, TYR, GLN
Aromático **(ARG)**	PHE, TRP, TYR
Alifático **(ALG)**	GLY, ALA, VAL, ILE, LEU, MET, PRO
Aminoácidos no plegables **(UFG)**	GLY, PRO
Aminoácidos que favorecen las Hélices Alfa **(FAH)**	ALA, CYS, LEU, MET, GLU, GLN, LYS, HIS
Aminoácidos que favorecen las Hojas Beta **(FBS)**	VAL, ILE, PHE, TYR, TRP, THR
Aminoácidos que favorecen los Giros Beta **(AFT)**	GLY, SER, ASP, ASN, PRO

3.1.5. Matriz espacial N-tuplas de similitud-disimilitud basada en cortes aminoacídicos

A pesar de que los enfoques matriciales propuestos codifican la información geométrica de las proteínas, estas no contienen información sobre la topología molecular y consideran todas las interacciones entre los aminoácidos presentes, sin poder distinguir entre ellas. Por este motivo se proponen dos tipos de cortes moleculares: uno basado en una distancia topológica y otro geométrico basado en a distancia euclídea, denotados como *lag p* y *lag l,* respectivamente. La aplicación de uno o ambos cortes moleculares permite la generación de las **k-ésimas matrices espaciales con cociente de vecindad totales o locales 2-tuplas y 3-tuplas de similitud-disimilitud**, representadas por ${}^{V}ZB^{k}_{(G)}$, ${}^{V}ZT^{k}_{(G)}$ respectivamente. Las entradas de estas matrices son los elementos correspondientes de las

matrices $ZB^k_{(G)}$, ${}^{V}ZT^k_{(G)}$ multiplicados por una razón que considera la cantidad de relaciones inter-aminoacídicas entre los N aminoácidos considerados, que poseen una distancia topológica o geométrica menor o igual a un umbral *p* y/o *l*, previamente definido. Los coeficientes de las matrices que consideran cortes moleculares (${}^{V}zb^k_{ij(G)}$, ${}^{V}zt^k_{ijl(G)}$) se calculan como se indica a continuación:

$$
\begin{aligned}
{}^{V}zb^k_{ij(G)} &= zb^k_{ij(G)} && \text{si } p_{min} \leq p_{ij} \leq p_{max} \mid l_{min} \leq l_{ij} \leq l_{max} \\
{}^{V}zb^k_{ij(G)} &= 0 && \text{en cualquier otro caso}
\end{aligned}
\qquad (43)
$$

$$
\begin{aligned}
{}^{V}zt^k_{ijl(G)} &= zt^k_{ijl(G)} && \text{si } p_{min} \leq p_{ij},\ p_{jl},\ p_{li} \leq p_{max} \mid l_{min} \leq l_{ij},\ l_{jl},\ l_{li} \leq l_{max} \\
{}^{V}zt^k_{ijl(G)} &= \tfrac{2}{3} zt^k_{ijl(G)} && \text{si } p_{min} \leq p_{ij},\ p_{jl(li)} \leq p_{max} \mid l_{min} \leq l_{ij},\ l_{jl(li)} \leq l_{max} \\
{}^{V}zt^k_{ijl(G)} &= \tfrac{1}{3} zt^k_{ijl(G)} && \text{si } p_{min} \leq p_{ij(jl)(li)} \leq p_{max} \mid l_{min} \leq l_{ij(jl)(li)} \leq l_{max} \\
{}^{V}zt^k_{ijl(G)} &= 0 && \text{en cualquier otro caso}
\end{aligned}
\qquad (44)
$$

donde, $zb^k_{ij(G)}, zt^k_{ijl(G)}$ son los coeficientes de las matrices totales o locales $ZB^k_{(G)}, ZT^k_{(G)}$ respectivamente, y que representan las relaciones binarias y ternarias entre los aminoácido *i,j* y *l*. El símbolo | significa que los cortes *lag p* y *lag l* pueden ser considerados solamente uno o ambos.

Los parámetros p_{xz} y l_{xz} constituyen la distancia topológica o geométrica entre los átomos x y z [x,z =(*i*,*j*,*l*)].

3.1.6. Matriz espacial N-tuplas de similitud-disimilitud basada en procedimientos de normalización

Todas las matrices anteriormente definidas (totales, locales, con corte molecular) se denominan no estocásticas (ns) ya que no presentan ningún procedimiento de normalización (la suma de los elementos en las filas o columnas no es igual a uno). Con el propósito de normalizar estas representaciones, se aplicarán los esquemas de normalización **simple estocástico (ss) y probabilidad mutua (mp)**. Por lo tanto, se proponen las **k-ésimas matrices espaciales total o local 2-tuplas, 3-tuplas simple estocásticas de similitud-disimilitud** y las **k-ésimas matrices espaciales total o local 2-tuplas, 3-tuplas con probabilidad mutua.** Estas matrices se calculan como se muestra a continuación:

$$ {}^{V}_{ss}zb^{k}_{ij} = \frac{{}^{V}_{ns}zb^{k}_{ij}}{S_i} = \frac{{}^{V}_{ns}zb^{k}_{ij}}{\sum_{j=1}^{n} {}^{V}_{ns}zb^{k}_{i}} \quad (45) $$

$$ {}^{V}_{mp}zb^{k}_{ij} = \frac{{}^{V}_{ns}zb^{k}_{ij}}{S_{ij}} = \frac{{}^{V}_{ns}zb^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}^{V}_{ns}z^{k}_{ij}} \quad (46) $$

$$ {}^{V}_{ss}zt^{k}_{ijl} = \frac{{}^{V}_{ns}zt^{k}_{ijl}}{S_{jl}} = \frac{{}^{V}_{ns}zt^{k}_{ijl}}{\sum_{j=1}^{n}\sum_{k=1}^{n} {}_{ns}z^{k}_{ijl}} \quad (47) $$

$$ {}^{V}_{mp}zt^{k}_{ijl} = \frac{{}^{V}_{ns}zt^{k}_{ijl}}{S_{ijl}} = \frac{{}^{V}_{ns}zt^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} {}^{V}_{ns}zt^{k}_{ijl}} \quad (48) $$

donde, n es el número de aminoácidos en la proteína, ${}^{V}_{ss}zb^{k}_{ij(G)}$, ${}^{V}_{ss}zt^{k}_{ijl(G)}$ son los elementos de las matrices simple estocásticas ${}^{V}_{ss}ZB^{k}_{(G)}$, ${}^{V}_{ss}ZT^{k}_{(G)}$.

${}^{V}_{mp}zb^{k}_{ij(G)}$, ${}^{V}_{mp}zt^{k}_{ijl(G)}$ son los elementos de las matrices aplicando la probabilidad mutua ${}^{V}_{mp}ZB^{k}_{(G)}$, ${}^{V}_{mp}ZT^{k}_{(G)}$.

Los coeficientes ${}^{V}_{ns}zb^{k}_{ij(G)}$, ${}^{V}_{ns}zt^{k}_{ijl(G)}$ corresponden a las representaciones no estocásticas ${}^{V}_{ns}ZB^{k}_{(G)}$, ${}^{V}_{ns}ZT^{k}_{(G)}$.

Sij es la suma de todos los elementos de la matriz bidimensional y S_{ijl} la suma de todos los elementos de la matriz tridimensional.

3.1.7. Descriptores moleculares 3D N-Lineales para proteínas

Los nuevos descriptores moleculares tridimensionales para proteínas se pueden definir a partir de la representación tensorial de la estructura de las proteínas (ver sección 3.1) y la definición matemática de las formas algebraicas (ver sección 1.1). Por tanto, si una proteína tiene n aminoácidos, los k-ésimos descriptores bilineales (relaciones entre dos aminoácidos) y trilineales (relaciones entre tres aminoácidos) se calculan como formas algebraicas N-lineales en $\mathbb{R}^n$ sobre un conjunto de vectores base canónico y se expresan por las siguientes ecuaciones respectivamente:

$$ {}^{(V)}_{ns[ss,mp]}b^{k}_{(G)}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{(V)}_{ns[ss,mp]}zb^{,k}_{ij(G)}, x^i y^j = [X]^T\, ZB^k [Y] \quad (49) $$

$$ {}^{(V)}_{ns[ss,mp]}tr^{k}_{(G)}(\bar{x}\,,\bar{y},\bar{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}^{(V)}_{ns[ss,mp]}zt^{k}_{ijl(G)} x^i y^j p^l = [Y]^T [X]^T Z T^k [P] \quad (50) $$

donde (49) y (50) es el valor del descriptor calculado para todas las relaciones entre dos y tres aminoácidos según la representación matricial utilizada (ver sección 3.1).

Por otro lado, $x^1, \dots, x^n, y^1, \dots, y^n, z^1, \dots, z^n$ son las componentes de los vectores macromoleculares $(\bar{x},\bar{y},\bar{p})$ que se pueden definir según las propiedades indicadas en la sección 3.1.1.

De esta manera, se pueden conformar varias combinaciones permitiendo calcular distintos tipos de descriptores según puede observarse en la Tabla 4.

Tabla 4. Descriptores moleculares calculados según la ponderación de los vectores moleculares usados en las formas algebraicas N-lineales.

Descriptores Bilineales		Descriptores Trilineales	
Forma Bilineal	B = (x ≠ y)	Forma Trilineal	Tr = (x ≠ y ≠ p)
Forma Cuadrática	Q = (x=y)	Forma Trilinear Lineal	TrF = (y=I, p=I)
Forma Lineal	F = (y = I)	Forma Trilineal Bilineal	TrB = (x = y, p=I)
		Forma Trilineal Cuadrática Bilineal	TrQB = (x = y ≠ p)
		Forma Trilineal Cúbica	TrC = (x = y = p)

Sin embargo, el cálculo de los descriptores propuestos acorde a las ecuaciones (49) y (50), se realiza considerando la proteína como un todo. Considerando la definición de LAI (*Local Amino acid Invariant*) generada a partir del concepto de LOVI (*Local Vertex Invariant*) [132], [133], estos descriptores pueden generarse a nivel de aminoácido para así caracterizar sus componentes en su estructura más básica. Por lo tanto y a partir de las ecuaciones presentadas para el cálculo de descriptores, en las Ecuaciones (42) y (43), se definen los k-ésimos descriptores 2-lineales y 3-lineales para cada aminoácido de una proteína:

$$ {}_{b}L_{aa} = {}^{(V)}_{ns[ss,mp]}b^{aa,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{(V)}_{ns[ss,mp]}zb^{aa,k}_{ij(G)}, x^i y^j = [X]^T\, ZB^{aa,k}[Y] \quad (51) $$

$$ {}_{tr}L_{aa} = {}^{(V)}_{ns[ss,mp]}tr^{aa,k}(\bar{x},\bar{y},\bar{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} {}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)} x^i y^j p^l = [Y]^T[X]^T Z \quad (52) $$

donde, n es el número de aminoácidos en la proteína, “aa” indica el aminoácido analizado y los coeficientes ${}^{(V)}_{ns[ss,mp]}zb^{aa,k}_{ij(G)}$, ${}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)}$ son las entradas pertenecientes a las matrices de nivel aminoacídico ${}^{(V)}_{ns[ss,mp]}ZB^{aa,k}_{ij(G)}$, ${}^{(V)}_{ns[ss,mp]}ZT^{aa,k}_{ijl(G)}$ respectivamente.

De esta forma, si una proteína es particionada en “aa” aminoácidos, entonces las matrices ${}^{(V)}_{ns[ss,mp]}ZB^{k}_{ij(G)}$, ${}^{(V)}_{ns[ss,mp]}ZT^{k}_{ijl(G)}$ pueden ser particionadas en “aa” matrices de nivel aminoacídico ${}^{(V)}_{ns[ss,mp]}ZB^{aa,k}_{ij(G)}$, ${}^{(V)}_{ns[ss,mp]}ZT^{aa,k}_{ijl(G)}$ respectivamente. Esta descomposición debe cumplir con la propiedad de que la suma de todas las matrices de aminoácido sea igual a la matriz total inicial; para lograr esta descomposición, los valores de las matrices de aminoácido se obtienen como se indica a continuación:

$$
\begin{aligned}
{}^{(V)}_{ns[ss,mp]}zb^{aa,k}_{ij(G)} &= zb^{k}_{ij(G)} && \text{si } \boldsymbol{i} \wedge \boldsymbol{j} = \mathbf{aa} \\
{}^{(V)}_{ns[ss,mp]}zb^{aa,k}_{ij(G)} &= \frac{1}{2}zb^{k}_{ij(G)} && \text{si } \boldsymbol{i} \vee \boldsymbol{j} = \mathbf{aa} \\
{}^{(V)}_{ns[ss,mp]}zb^{aa,k}_{ij(G)} &= 0 && \text{en cualquier otro caso}
\end{aligned}
\qquad (53)
$$

$$
\begin{aligned}
{}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)} &= zt^{k}_{ijl} && \text{si } \boldsymbol{i} \wedge \boldsymbol{j} \wedge \boldsymbol{l} = \mathbf{aa} \\
{}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)} &= \frac{2}{3}zt^{k}_{ijl} && \text{si } \boldsymbol{i,j} \vee \boldsymbol{j,l} \vee \boldsymbol{i,l} = \mathbf{aa} \\
{}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)} &= \frac{1}{3}zt^{k}_{ijl} && \text{si } \boldsymbol{i} \vee \boldsymbol{j} \vee \boldsymbol{l} = \mathbf{aa} \\
{}^{(V)}_{ns[ss,mp]}zt^{aa,k}_{ijl(G)} &= 0 && \text{en cualquier otro caso}
\end{aligned}
\qquad (54)
$$

Si todos los elementos que conforman el LAI se suman, se obtendría el descriptor total de la proteína como si la separación no se hubiera llevado a cabo. Sin embargo, se puede generalizar la fusión de los descriptores locales de aminoácido mediante el uso de operadores de agregación.

La noción de operadores de agregación como un esquema de generalización para la combinación lineal de descriptores de aminoácido a los descriptores totales se deriva de la hipótesis que la definición global de un sistema no necesariamente debe ser aditiva. En este

sentido, la obtención de índices totales a partir de los índices de aminoácido es generalizada como se muestra en las siguientes ecuaciones:

$$ {}^{(V)}_{ns[ss,mp]}b^{aa,k}(\bar{x},\bar{y}) = OPER({}_{b}L_{aa}) \quad (55) $$

$$ {}^{(V)}_{ns[ss,mp]}tr^{aa,k}(\bar{x},\bar{y},\bar{p}) = OPER({}_{tr}L_{aa}) \quad (56) $$

Donde, OPER es el operador de agregación utilizado sobre el LAI de descriptores de aminoácido. Estos operadores de agregación fueron clasificados en cuatro grupos como se muestra a continuación [17]:

1. Normas: Normas de Minkowski (N1, N2, N3). Puede notarse que N1 es equivalente a la suma de los componentes del LAI.
2. Estadísticos de tendencia central: Media Geométrica (G), Media Aritmética (M), Media Cuadrática (P2), Media Potencial (P3), Media Armónica (A).
3. Estadísticos de dispersión y forma: Varianza (V), Skewness (S), Kurtosis (K), Desviación Estándar (SD), Coeficiente de Variación (CV), Rango (R), Percentil 25 (Q1), Percentil 50 (Q2), Percentil 75 (Q3), Rango Inter-quartil (I50), Li máximo (MX) y Li mínimo (MN).
4. Algoritmos clásicos: Autocorrelación (AC), Gravitacional (GV), Contenido Total de Información (TIC), Contenido Promedio de Información (MIC), Contenido de Información Estandarizado (SIC), Suma Total (TS), Estado Electro-topológico (ES), Ivanciuc – Balaban (IB) y conectividad de Kier-Hall (KH).

3.2. Evaluación del Rendimiento del *software* ToMoCoMD-CAMPS MuLiMS-MCoMPAs

En esta sección, se presentan y discuten los resultados obtenidos de las pruebas realizadas al *software* ToMoCoMD-CAMPS con respecto a determinación del tiempo requerido para el cálculo de descriptores bilineales y trilineales, los gráficos de rendimiento

respecto a la ganancia de velocidad respecto a la paralelización y la escalabilidad del *software*.

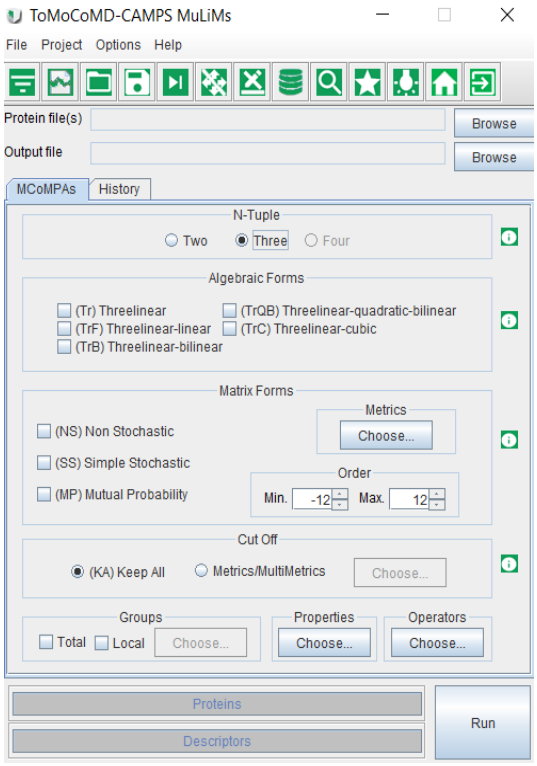
3.2.1. Interfaz gráfica del *software*

Como se presentó en la sección 2.2., el *software* MuLiMS MCoMPAs presenta una interfaz gráfica que permite la generación de los descriptores en una manera amigable para el usuario. Dicha interfaz gráfica se indica a en la Figura 4.

3.2.2. Cálculo multiprocesador de los descriptores propuestos

Las pruebas de rendimiento se realizaron en un sistema de cómputo de alto rendimiento (HPC-USFQ) que contiene 16 cores. Las características del core que se utilizó para las pruebas son las siguientes: Intel (R) Xeon (R) E5-2630 v3, 2.4 GHz de velocidad y 64 GB de RAM disponible. En este caso, solo se asignó 30 GB a la Máquina Virtual de Java. Los experimentos se realizaron con una data de 152 proteínas y se utilizó la representación Carbono alfa (Cα) para la generación de los índices en varios de los experimentos.

Figura 4. Interfaz gráfica del software MuLiMs-MCoMPAs



3.2.2.1. Determinación del tiempo requerido para el cálculo de los descriptores

Se realizó el cálculo de los descriptores bilineales y trilineales utilizando 50 y 37 proyectos respectivamente considerado las cuatro representaciones propuestas para evaluar si existía alguna diferencia significativa en el tiempo de cálculo entre ellas. En la Tabla 5, se presentan los resultados obtenidos para este experimento.

Tabla 5. Resultados del tiempo de cálculo obtenido en índices bilineales y trilineales para cada representación

Índice	Proc.	Protein	Repr.	Proyecto	Descriptores calculados	Tiempo total, s	Tiempo por proyecto, s	Tiempo promedio por descriptor, s
Terna	32	152	CA	37	63	180456	4877	77.39
			CB	37	63	180530	4879	
			AVG	37	63	180272	4872	
			AB	37	63	180346	4874	
Dupla	32	152	CA	50	6107	23294	466	0.076
			CB	50	6107	23286	466	
			AVG	50	6107	23366	467	
			AB	50	6107	23100	462	

De esta tabla se puede observar que los tiempos de cálculo de los índices en las cuatro representaciones para el caso trilineal es muy similar y las diferencias en el tiempo pueden haber ocurrido por diferencias en el tiempo de repartición de tareas en el procesador. Sin embargo, no existe una diferencia estadísticamente significativa entre dichos tiempos. Este particular nos indica que el cálculo de los descriptores es estable para cualquiera de las cuatro representaciones escogidas. Similares observaciones se pueden destacar para el caso de los índices bilineales.

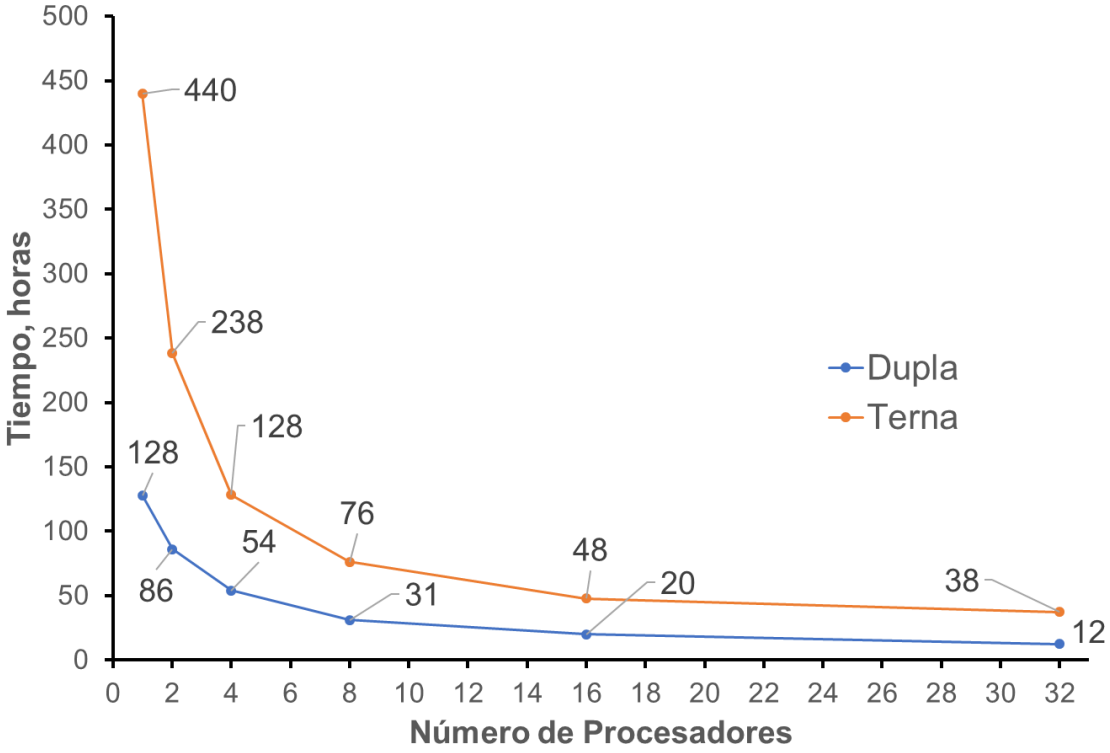
Así mismo se puede observar que la diferencia entre el tiempo promedio de cálculo para índices trilineales es 1000 veces mayor que para el caso de los índices bilineales. Este particular es consecuencia de la mayor complejidad matemática con la que se definen los índices trilineales en el cálculo.

3.2.2.2. Soft Speed Up

La ganancia de velocidad o *Speed-up* (Sp) para *p* procesadores es el cociente entre el tiempo de ejecución de un programa y el tiempo de ejecución de la versión paralela de dicho programa en p procesadores. Este experimento se realiza para la evaluación de la paralelización del software y definir el número de procesadores ideal para el cálculo de una data determinada en condiciones de cómputo establecidas. El *soft speed up* mantiene la cantidad de proyectos y de proteínas constante mientras se disminuye la cantidad de procesadores disponibles para el cálculo. En la Figura 5 se presentan los resultados del *soft speed up* tanto para el cálculo de los índices bilineales como para los índices trilineales.

De la Figura 5 se puede observar que el comportamiento del programa tiene una tendencia normal (ya que el tiempo de cálculo aumenta al disminuir la cantidad de procesadores disponibles) tanto para el cálculo de los índices trilineales como para los índices bilineales. De este gráfico se puede determinar que la cantidad de procesadores óptima para la realización de este cálculo tanto para índices bilineales como para índices trilineales es de 16 threads ya que, a partir de ese punto, la ganancia de tiempo al duplicar los procesadores ya no tiene un efecto significativo en la reducción del tiempo de cálculo (la repartición de tareas paralelizada del software ya no puede ser optimizada más).

Figura 5. Evaluación del Soft Speed up para índices bilineales y trilineales.



3.3. Predicción de estructura y velocidad de plegamiento utilizando los descriptores 3D para proteínas propuestos

3.3.1. Comparación interna de los índices propuestos para la definición de nuevos proyectos de cálculo en el *software*

Considerando la totalidad de configuraciones teóricas disponibles propuestas para el cálculo de descriptores, la cantidad de descriptores generada tanto para los índices trilineales como para los bilineales se presenta en la Tabla 6.

Tabla 6. Número de descriptores totales utilizando todas las configuraciones teóricas disponibles

Tipo de descriptor	Espacio total de descriptores
Trilineal	1,056,388,608
Bilineal	647,685,104
Total	1,704,073,712

Es claro que muchos de estos índices pueden contener información redundante, también que puedan ser colineales entre sí, o pueden modelar propiedades de las proteínas en el mismo grado. Por este motivo, se realizó la reducción de la dimensionalidad de los índices que se pueden calcular mediante tres procedimientos: cantidad de información de cada familia de índices considerando teoría de información, análisis de componentes principales y modelado de actividad-estructura (QSAR) utilizando regresión lineal múltiple como estrategia.

Como resultado de este análisis, se generaron nuevos proyectos para el cálculo de índices bilineales y trilineales, considerando las combinaciones que permitan la extracción de la mayor cantidad de información. Para el caso de índices bilineales, se observó que las representaciones de Carbono Beta (Cβ) y Carbono Amino (AB) obtienen los mejores índices; mientras que en el caso de los índices trilineales, se observó que las representaciones de Carbono Beta (Cβ) y Pseudo-Aminoácido (AVG) obtienen los mejores índices. Los archivos con las configuraciones de los nuevos proyectos constan en el material suplementario adjunto en CD.

Además, estos proyectos presentan una cantidad definida de índices que es representativa del espacio total y permite una exploración adecuada. La Tabla 7 indica la cantidad de descriptores que se pueden calcular a partir de estos nuevos proyectos y la reducción respecto al espacio inicial.

Tabla 7. Cantidad de descriptores generados con los nuevos proyectos y su comparación contra el espacio inicial.

Tipo de descriptor	Espacio total de descriptores	Índices representativos del espacio	Relación de reducción
Trilineal	1,056,388,608	20,263	52,134
Bilineal	647,685,104	13,648	47,456
Total	1,704,073,712	33,911	

3.3.2. Modelado de propiedades de proteínas y comparación con otros métodos

Con la finalidad de evaluar el rendimiento de los descriptores propuestos con respecto a otros descriptores reportados en la literatura, se realizaron modelos matemáticos para la predicción de la velocidad de plegamiento utilizando regresión lineal múltiple como estrategia y modelos de clasificación estructural basados en la base de datos SCOP utilizando análisis de discriminante lineal como estrategia. Los resultados obtenidos para cada aplicación y su comparación frente a las otras estrategias se indican a continuación.

3.3.2.1. Velocidad de plegamiento

Para el modelado de la velocidad de plegamiento, se utilizó el conjunto de proteínas (80 proteínas) propuesto por Ouyang [118] como serie de entrenamiento, retirando el caso "2BLM" debido a que solo contenía una representación de carbono alfa; para la serie de predicción se utilizó un conjunto de 17 proteínas propuesto en el artículo de Ruiz-Blanco [101]. La modelación se realizó utilizando el *software* MOBYDIGS [122], que combina la técnica de regresión lineal múltiple con un método de subconjunto y *wrapper* basado en Algoritmo Genético. Se realizaron varios experimentos de exploración que incluían la generación de modelos considerando solo índices bilineales, solamente índices trilineales y la combinación de índices bilineales con índices trilineales. Los modelos que mejor se ajustaron a la variable respuesta mediante el análisis de los parámetros estadísticos para predecir la

velocidad de plegamiento con los descriptores obtenidos en el *software* MuLiMs-MCoMPAs constan el material suplementario adjunto en CD.

Del conjunto de proteínas considerado como serie de predicción, basados en el error de predicción mostrado en todos los modelos obtenidos, se procedió a excluir 4 proteínas del conjunto de predicción (*outliers*). Para el caso de los modelos obtenidos con índices bilineales se encontró que los *outliers* son: pdb1a6n, pdb1spr_A, pdb1t8j, pdb2vik. Para el caso de los modelos obtenidos con índices trilineales se encontró que los *outliers* son: pdb1jo8, pdb1spr_A, pdb1t8j, pdb2vik. Se observa que 3 de los 4 *outliers* son comunes para ambos índices.

En esta sección se presentarán los valores estadísticos obtenidos por los dos mejores modelos obtenidos utilizando los índices bilineales, los 2 mejores modelos utilizando índices trilineales y los 2 mejores modelos utilizando la combinación de los dos tipos de índices en la Tabla 8; además, las ecuaciones de regresión obtenidas para cada uno de estos modelos se presentan a continuación:

Modelo 1:

ln(k)= -68463.2*A + 0.07262*B + 2.4516*C + 0.7509*D + 13.9890 (48)

donde:

k= velocidad de plegamiento

A= CB_Q2_B_M19_NS-3_T_LGP[+12.0]_LGL[4-11]_PAH-PBS_MCoMPAs (49)

B= AVG_N1_B_M2_SS-5_T_KA_ECI-KDS_MCoMPAs

C= AB_Q2_B_M32_SS5_ALA_KA_Z2-PBS_MCoMPAs

D= AB_S_B_M14_NS-6_FBS_LGP[8-11]_PBS-EPS_MCoMPAs

Modelo 2

ln(k)= 2.2566*A – 68047.7*B + 0.07381*C + 0.78523*D + 14.2208 (50)

donde:

k= velocidad de plegamiento

A= CB_Q2_B_M32_SS2_FBS_KA_Z2-PBS_MCoMPAs

B= CB_Q2_B_M19_NS-3_T_LGP[+12.0]_LGL[4-11]_PAH-PBS_MCoMPAs

C= AVG_N1_B_M2_SS-5_T_KA_ECI-KDS_MCoMPAs

D= AB_S_B_M14_NS-6_FBS_LGP[8-11]_PBS-EPS_MCoMPAs

Modelo 3

$$\ln(k) = -0.0323*A + 20.3011*B + 7205.01*C - 1.7572 \quad (51)$$

donde:

k= velocidad de plegamiento

A= AVG_N3_TrQB_M55(M15)_SS-2_T_KA_PAH-ISA_MCoMPAs

B= AVG_Q1_TrC_M58(M15)_SS0_T_KA_PAH_MCoMPAs

C= AVG_GV[5]_MX_TrF_M41(M5)_MP7_o_T_KA_PBS_MCoMPAs

Modelo 4

$$\ln(k) = -0.03241*A + 19.9437*B + 156.728*C - 1.42882 \quad (52)$$

donde:

k= velocidad de plegamiento

A= AVG_N3_TrQB_M55(M15)_SS-2_T_KA_PAH-ISA_MCoMPAs

B= AVG_Q1_TrC_M58(M15)_SS0_T_KA_PAH_MCoMPAs

C= AVG_GV[5]_MX_TrF_M41(M5)_MP8_o_T_KA_PIE_MCoMPAs

Modelo 13

$$\ln(k) = -44766.6*A - 0.96157*B + 0.20729*C - 3.25903*D + 25.4265 \quad (53)$$

donde:

k= velocidad de plegamiento

A= CB_Q2_B_M19_NS-3_T_LGP[+12.0]_LGL[4-11]_PAH-PBS_MCoMPAs

B= CB_K_Q_M5_NS-1_T_LGP[1-3]_KDS_MCoMPAs

C= CB_K_B_M2_SS-1_FBS_KA_MM-ECI_MCoMPAs

D= CB_MIC_N1_TrQB_M45(M8)_SS2_o_T_KA_PAH-Z3_MCoMPAs

Modelo 14

$\ln(k) = -42920.3*A + 0.17709*B - 3.22386*C + 26.0880$ (54)

donde:

k= velocidad de plegamiento

A= CB_Q2_B_M19_NS-3_T_LGP[+12.0]_LGL[4-11]_PAH-PBS_MCoMPAs

B= CB_K_B_M2_SS-1_FBS_KA_MM-ECI_MCoMPAs

C= CB_MIC_N1_TrQB_M41(M5)_SS2_o_T_KA_PAH-Z3_MCoMPAs

Tabla 8. Mejores modelos obtenidos para la predicción de la velocidad de plegamiento de 96 proteínas cuyos descriptores se calcularon con el software MuLiMs-MCoMPAs.

Modelo	Q^2_{LOO}	Q^2_{BOOT}	SDEP	Q^2_{EXT} (C.O.)	SDEPext (C.O.)	Q^2_{EXT} (S.O.)	SDEPext (S.O.)
Modelos que utilizaron índices bilineales							
1	75.33	74.85	2.145	72.77	2.663	87.52	2.057
2	75.27	74.70	2.147	72.43	2.635	86.67	2.019
Modelos que utilizaron índices trilineales							
3	74.8	73.83	2.167	32.28	3.170	85.75	2.786
4	74.4	74.01	2.172	35.25	3.287	84.92	2.983
Modelos que utilizaron índices bilineales y trilineales							
13	77.69	77.62	2.039	60.87	2.387	79.57	2.964
14	79.70	79.26	1.945	55.57	2.556	78.19	2.606

El análisis de la Tabla 8 puede realizarse considerando los mejores modelos obtenidos en el entrenamiento y los mejores modelos obtenidos en la predicción. Considerando los mejores modelos respecto al entrenamiento, se puede observar que los modelos obtenidos mediante la combinación de índices trilineales y bilineales e índices trilineales son los que

mejor modelan la serie de entrenamiento, entendiendo que la velocidad de plegamiento depende de la estructura tridimensional y varios sitios de contacto específicos, se puede fundamentar que los índices trilineales extraen mayor información geométrica ya que las herramientas matemáticas utilizadas que abarcan mayor cantidad de posibilidades e interacciones, que permite un mejor ajuste con respecto a la variable respuesta y que los índices bilineales contienen información ortogonal con respecto a los índices trilineales, lo que mejora el ajuste. Con respecto a la serie de predicción, se observa que los índices bilineales y trilineales logran obtener modelos más robustos que logran explicar la velocidad de plegamiento de un conjunto de proteínas no utilizado para el entrenamiento. Esto se puede entender debido a la cantidad de información que los índices trilineales aportan desde el punto de vista estructural, y las que las configuraciones propuestas para los índices bilineales permiten extraer información relevante para el modelado de esta propiedad.

Con respecto a otros índices que se han utilizado para modelar la velocidad de plegamiento, en la Tabla 9 se indica el valor de predicción más alto de este trabajo con respecto a los valores de predicción obtenidos en otros trabajos. Se puede observar que los descriptores obtenidos utilizando nuestra teoría son superiores a los otros valores reportados. En la Tabla 10 se presenta una comparación del rendimiento de los descriptores respecto al modelado de la serie de entrenamiento para la velocidad de plegamiento. Se puede observar que los descriptores de este trabajo presentan una diferencia significativa respecto a los parámetros estadísticos del otro trabajo.

Tabla 9. Comparación de valores de predicción de la velocidad de plegamiento de los descriptores 3D para proteínas de este trabajo con respecto a otros.

Descriptores	Longitud Cutoff	Q^2 (*training*)	SDEP
Folding degree [101]	-	73.96	2.20
Long Range Order [115]	4	72.25	2.28
Contact order [114]	2	73.96	2.19
Total Contact Distance [116]	2	73.96	2.21
Este trabajo (Modelo 14)	-	79.7	1.95

Tabla 10. Comparación de los valores de predicción de la serie externa para la velocidad de plegamiento con respecto a otro trabajo.

Descriptores	Q^2 (*test*)	SDEP
Folding degree [101]	54.76	2.03
Este trabajo (Modelo 1)	87.52	2.06

3.3.2.2. Clasificación estructural SCOP

Para la predicción de las clases estructurales se utilizó el conjunto de proteínas (204 proteínas) propuesto por K.C. Chou basado en la clasificación SCOP (52 todo alfa, 61 todo beta, 45 alfa/beta, 46 alfa + beta) [88]. Este conjunto se dividió en dos grupos, 149 proteínas se utilizaron para el entrenamiento y 55 proteínas utilizadas para la serie de predicción. La modelación se realizó utilizando el *software* WEKA [76], que combina la técnica de análisis de discriminante lineal con un método de subconjunto que utiliza dos métodos de búsqueda como *Best First* y *Greedy Stepwise* y un método *wrapper*. Se realizaron varios experimentos de exploración que incluían la generación de modelos considerando solo índices bilineales, solamente índices trilineales y la combinación de índices bilineales con índices trilineales. Los mejores modelos obtenidos para la clasificación estructural con los descriptores obtenidos en el *software* MuLiMs-MCoMPAs constan el material suplementario adjunto en CD

En esta sección se presentarán los valores estadísticos obtenidos por los dos mejores modelos obtenidos utilizando los índices bilineales, los 2 mejores modelos utilizando índices trilineales y los 2 mejores modelos utilizando la combinación de los dos tipos de índices en la Tabla 11.

El análisis de la Tabla 11 puede realizarse considerando los mejores modelos obtenidos en el entrenamiento y los mejores modelos obtenidos en la predicción. Considerando los mejores modelos respecto al entrenamiento, se puede observar que los modelos obtenidos utilizando los índices bilineales y la combinación de índices bilineales y trilineales son los que mejor modelan la serie de entrenamiento; se puede observar la particularidad de que los índices bilineales pueden generar mejores modelos que los que se pueden generar utilizando los índices trilineales, a pesar que los últimos sean constituidos considerando características geométricas más amplias. El principio de *No Free Lunch* [134] indica que una sola estrategia de cálculo o de optimización no puede realizar el modelado en forma acertada para todos los casos. De todas formas, se puede observar que la diferencia estadística entre los modelos con descriptores bilineales y trilineales no es grande, por lo que esto sugiere que tanto la teoría para la generación de índices bilineales y trilineales es robusta y puede generar resultados que puedan ser usados en otras aplicaciones. Con respecto a la serie de predicción, se observa el mismo comportamiento descrito para el análisis de con respecto al conjunto de predicción.

Tabla 11. Mejores modelos obtenidos para la clasificación estructural SCOP de 204 proteínas cuyos descriptores se calcularon con el software MuLiMs-MCoMPAs

Modelo	Número de Variables	Clasificación correcta (%) *Training* (149)	MCC *Training*	Clasificación correcta (%) *Test* (55)	MCC *Test*
Modelos que utilizaron índices bilineales					
3	5	100.00	1.000	96.36	0.955
7	3	99.33	0.992	96.36	0.955
Modelos que utilizaron índices trilineales					
9	16	98.65	0.962	92.59	0.777
12	17	97.99	0.942	90.74	0.717
Modelos que utilizaron índices trilineales y bilineales					
15	13	99.33	0.981	96.36	0.893
16	9	99.33	0.981	98.18	0.943

Con respecto a otros índices que se han utilizado para la clasificación SCOP de proteínas utilizando el mismo conjunto de datos, en la Tabla 12 se indica el valor de predicción más alto de este trabajo con respecto a los valores de predicción obtenidos en otros trabajos. Se puede observar que los descriptores obtenidos utilizando nuestra teoría son ampliamente superiores a los otros valores reportados. En la Tabla 13 se presenta una comparación del rendimiento de los descriptores respecto al modelado de la serie de predicción para la clasificación SCOP de las proteínas. Se puede observar que los descriptores de este trabajo presentan una diferencia significativa respecto a los parámetros estadísticos presentados en el otro trabajo

Tabla 12. Comparación de valores de predicción de la clasificación SCOP de los descriptores 3D para proteínas de este trabajo con respecto a otros.

Descriptores	Clasificación correcta (%) *Training*
AA composition [135]	83.80
Pseudo AA composition [136]	91.20
Pair coupled AA composition [137]	74.50
PSI-BLAST [138]	94.10
Bilinear descriptors [19]	92.60
Este trabajo	99.33

Tabla 13. Comparación de los valores de predicción de la serie externa para la clasificación SCOP de proteínas con respecto a otro trabajo.

Descriptores	Clasificación correcta (%) *Test*
Bilinear descriptors [19]	92.70
Este trabajo	96.33

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

Se definieron nuevos descriptores tridimensionales basados en formas algebraicas multilineales que permiten codificar la información geométrica y topológica entre dos y tres aminoácidos de una proteína; demostrándose con los distintos estudios realizados en este trabajo, que estos descriptores poseen mayor variabilidad y cantidad de información codificada con respecto otros descriptores definidos en la literatura.

Se demostró mediante el estudio de dos aplicaciones representativas (velocidad de plegamiento y clasificación estructural) que los nuevos descriptores permitan la extracción de información estructural de relevancia de las proteínas basado en la comparación de los parámetros estadísticos obtenidos. Por lo tanto, estos descriptores pueden constituir una

nueva alternativa para la generación de modelos predictivos de propiedades fisicoquímicas de interés y predicción de las funciones de una proteína.

Se desarrolló y validó el *software* denominado ToMoCoMD-CAMPS MuLiMs MCoMPAs que permite la generación de los descriptores moleculares 3D propuestos aprovechando las prestaciones computacionales disponibles de la mejor manera y brindando al usuario una herramienta amigable que puede ser utilizada en cualquier plataforma disponible.

Se realizó la generación de nuevos proyectos de cálculo para el software MuLiMs MCoMPAs (15 proyectos para la definición de índices bilineales y 10 proyectos para la generación de índices trilineales) reduciendo el espacio de descriptores aproximadamente en 50.000 veces mediante procedimientos estadísticos, de teoría de información y modelado, que brindan al usuario una exploración representativa del espacio total en el menor tiempo posible.

Se evaluaron cuatro tipos de representaciones de aminoácidos para la extracción de información espacial de las proteínas y basados en los diferentes análisis realizados en este trabajo, se observó que la mayor cantidad de información extraída de las proteínas cuando se calculan los índices bilineales fue utilizando las representaciones de Carbono Beta (Cβ) y Carbono Amida (AB) mientras que para los índices trilineales fue utilizando las representaciones de Carbono Beta (Cβ) y Pseudo Aminoácido (AVG).

4.2. Recomendaciones

Utilizar métodos de truncaje esférico y operadores de agregación generalizados como aporte teórico adicional en la generación de los descriptores 3D para proteínas propuestos con el fin de considerar generalizaciones adicionales que permitan extraer mayor cantidad de información no considerada para los índices actuales.

Evaluar el desempeño de los descriptores moleculares 3D generados en estudios multi referencia (varias aplicaciones representativas para el campo de proteínas) con diferentes conjuntos de datos de proteínas con el fin de demostrar en que tipos de aplicaciones estos índices se desempeñan mejor que otros y cuanta información ortogonal a los otros descriptores se genera mediante el cálculo de nuestros descriptores.

5. REFERENCIAS

[1] T. N. Bui and G. Sundarraj, "An efficient genetic algorithm for predicting protein tertiary structures in the 2D HP model," in *Proceedings of the 2005 conference on Genetic and evolutionary computation - GECCO '05*, 2005, p. 385.

[2] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm," *J. Mol. Biol.*, vol. 293, no. 2, pp. 321–331, 1999.

[3] E. Shakhnovich, "Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry and Biology Meet," *Chem. Rev.*, vol. 106, no. 5, pp. 1559–1588, 2009.

[4] C. de Duve, "The second genetic code," *Nature*, vol. 333, p. 117, May 1988.

[5] R. Apweiler, "The universal protein resource (UniProt) in 2010," *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 190–195, 2009.

[6] D. A. Benson, "GenBank," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D34–D38, 2004.

[7] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, p. 980, Dec. 2003.

[8] H. González-Díaz, E. Uriarte, and R. de Armas, "Predicting stability of Arc repressor mutants with protein stochastic moments," *Bioorg. Med. Chem.*, vol. 13, no. 2, p. 323—331, Jan. 2005.

[9] Y. Marrero Ponce *et al.*, "Protein quadratic indices of the 'macromolecular pseudograph's α-carbon atom adjacency matrix'. 1. Prediction of Arc repressor alanine-mutant's stability," *Molecules*, vol. 9, no. 12, pp. 1124–1147, 2004.

[10] H. González-Díaz, R. Ramos de Armas, and R. Molina, "Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Ψ-RNA packaging region with drugs," *Bioinformatics*, vol. 19, no. 16, pp. 2079–2087, Nov. 2003.

[11] H. González-Díaz, Y. Pérez-castillo, G. Podda, and E. Uriarte, "Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices," *J. Comput. Chem.*, vol. 28, no. 12, pp. 1990–1995, Apr. 2007.

[12] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, vol. 2. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2009.

[13] M. Randić, J. Zupan, A. Balaban, D. Vikić-Topić, and D. Plavšić, "Graphical Representation of Proteins †," *Chem. Rev.*, vol. 111, no. 2, pp. 790–862, Feb. 2011.

[14] M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić-Topić, and D. Plavšić, "Graphical representation of proteins as four-color maps and their numerical characterization," *J. Mol. Graph. Model.*, vol. 27, no. 5, pp. 637–641, 2009.

[15] Y. Marrero-Ponce, R. Medina-Marrero, J. A. Castillo-Garit, V. Romero-Zaldivar, F. Torrens, and E. A. Castro, "Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor," *Bioorg. Med. Chem.*, vol. 13, no. 8, pp. 3003–3015, 2005.

[16] S. E. Ortega-Broche, Y. Marrero Ponce, Y. E. Díaz, F. Torrens, and F. Pérez-Giménez, "tomocomd-camps and protein bilinear indices - novel bio-macromolecular descriptors for protein research: I. Predicting protein stability effects of a complete set of alanine substitutions in the Arc repressor," *FEBS J.*, vol. 277, no. 15, pp. 3118–3146, Aug. 2010.

[17] Y. Marrero Ponce *et al.*, "Optimum search strategies or novel 3D molecular descriptors: Is there a stalemate?," *Curr. Bioinform.*, vol. 10, no. 5, 2015.

[18] C. García-Jacas, E. Contreras-Torres, Y. Marrero-Ponce, M. Pupo-Meriño, S. J. Barigye, and L. Cabrera-Leyva, "Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets," *J. Cheminform.*, vol. 8, no. 1, pp. 1–16, 2016.

[19] Y. Marrero Ponce, E. Contreras-Torres, C. García-Jacas, S. J. Barigye, N. Cubillán, and Y. J. Alvarado, "Novel 3D bio-macromolecular bilinear descriptors for protein science: Predicting protein structural classes," *J. Theor. Biol.*, vol. 374, pp. 125–137, Jun. 2015.

[20] R. W. Johnson, J. R. Johnson, and C.-H. Huang, "Multilinear Algebra and Parallel Programming," *J. Supercomput.*, vol. 5, no. 2–3, pp. 189–217, Oct. 1991.

[21] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus*. 1987.

[22] C. García-Jacas, "Nueva codificación tridimensional de la estructura química de moléculas orgánicas," Universidad Central "Marta Abreu" de Las Villas, 2013.

[23] E. Deza and M.-M. Deza, "Chapter 1 - General Definitions," in *Dictionary of Distances*, E. Deza and M.-M. B. T.-D. of D. Deza, Eds. Amsterdam: Elsevier, 2006, pp. 2–30.

[24] E. Deza and M.-M. Deza, “Chapter 4 - Metric Transforms,” E. Deza and M.-M. B. T.-D. of D. Deza, Eds. Amsterdam: Elsevier, 2006, pp. 44–49.

[25] E. Deza and M.-M. Deza, “Chapter 3 - Generalizations of Metric Spaces,” E. Deza and M.-M. B. T.-D. of D. Deza, Eds. Amsterdam: Elsevier, 2006, pp. 36–43.

[26] M. Warrens, *Similarity coefficients for binary data: Properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. 2008.

[27] F. Harary, *Graph Theory*, First. Reading: Addison-Wesley, 1969.

[28] I. Gutman and O. Polansky, *Mathematical Concepts in Organic Chemistry*, First. 2012.

[29] R. Wilson, *Introduction to Graph Theory*, Fourth Edi. Edinburgh: Prentice Hall, 1996.

[30] V. A. Gorbátov, *Fundamentos de la Matematica discreta*. URSS: MIR MOSCU, 1988.

[31] A. Balaban, Ed., *Chemical Applications of Graph Theory*. London: Academic Press, 1976.

[32] D. H. Rouvray, *Computational chemical graph theory*. New York: Elsevier, 1990.

[33] Z. Mihalić and N. Trinajstić, “A graph-theoretical approach to structure-property relationships,” *J. Chem. Educ.*, vol. 69, no. 9, p. 701, Sep. 1992.

[34] L. Kier, *Molecular Connectivity in Chemistry and Drug Research*, First. London: Academic Press, 1976.

[35] L. Kier, *Molecular Connectivity in Structure-Activity Analysis*. Wiley, 1986.

[36] L. Kier and L. Hall, “An Electrotopological-State Index for Atoms in Molecules,” *Pharm. Res.*, vol. 7, no. 8, pp. 801–807, 1990.

[37] J. Devillers and A. Balaban, Eds., *Topological Indices and Related Descriptors in*

QSAR and QSPR. Gordon and Breach Science Publishers, 1999.

[38] D. Bonchev, *Information Theoretic Characterization of Chemical Structures*. 1983.

[39] M. V Diudea, *QSPR/QSAR studies by molecular descriptors*. Nova Science Publishers, 2001.

[40] A. Balaban, A. Chiriac, I. Motoc, and Z. Simon, "Steric and other Structural Parameters for QSAR BT - Steric Fit in Quantitative Structure-Activity Relations," A. T. Balaban, A. Chiriac, I. Motoc, and Z. Simon, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1980, pp. 2–22.

[41] N. Trinajstic, *Chemical graph theory*. Boca Raton: CRC Press, 1992.

[42] O. Ivanciuc, T. Ivanciuc, and A. Balaban, "Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules," *J. Chem. Inf. Comput. Sci.*, vol. 38, no. 3, pp. 395–401, May 1998.

[43] H. Gonzalez-Diaz, S. Vilar, L. Santana, and E. Uriarte, "Medicinal Chemistry and Bioinformatics - Current Trends in Drugs Discovery with Networks Topological Indices," *Curr. Top. Med. Chem.*, vol. 7, no. 10, pp. 1015–1029, 2007.

[44] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani, "Protein Contact Networks: An Emerging Paradigm in Chemistry," *Chem. Rev.*, vol. 113, no. 3, pp. 1598–1613, Mar. 2013.

[45] H. González-Díaz, Y. González-Díaz, L. Santana, F. M. Ubeira, and E. Uriarte, "Proteomics, networks and connectivity indices," *Proteomics*, vol. 8, no. 4, pp. 750–778, Feb. 2008.

[46] O. M. Rivera-Borroto, "Estrategias QSAR combinadas, TOMOCOMD-CARDD y

quimiométricas, para el Descubrimiento de candidatos a fármacos nuevos/novedosos frente a trichomonas vaginalis," Universidad Central "Marta Abreu" de Las Villas, 2008.

[47] Y. Marrero Ponce, F. Torrens, R. García-Domenech, S. E. Ortega-Broche, and V. R. Zaldivar, "Novel 2D TOMOCOMD-CARDD molecular descriptors: atom-based stochastic and non-stochastic bilinear indices and their QSPR applications," *J. Math. Chem.*, vol. 44, no. 3, pp. 650–673, 2008.

[48] Y. Marrero Ponce, "Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications," *Bioorg. Med. Chem.*, vol. 12, no. 24, pp. 6351–6369, 2004.

[49] J. A. Castillo-Garit, O. Martinez-Santiago, Y. Marrero Ponce, G. M. Casañola-Martín, and F. Torrens, "Atom-based non-stochastic and stochastic bilinear indices: Application to QSPR/QSAR studies of organic compounds," *Chem. Phys. Lett.*, vol. 464, no. 1–3, pp. 107–112, 2008.

[50] Y. Marrero Ponce, "Linear Indices of the 'Molecular Pseudograph's Atom Adjacency Matrix': Definition, Significance-Interpretation, and Application to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 6, pp. 2010–2026, Nov. 2004.

[51] Y. Marrero Ponce, F. Torrens, Y. J. Alvarado, and R. Rotondo, "Bond-based global and local (bond, group and bond-type) quadratic indices and their applications to computer-aided molecular design. 1. QSPR studies of diverse sets of organic chemicals," *J. Comput. Aided. Mol. Des.*, vol. 20, no. 10, pp. 685–701, 2006.

[52] C. Garcia-Jacas, Y. Marrero-Ponce, S. J. Barigye, J. R. Valdes-Martin, O. M. Rivera-

Borroto, and J. Olivero-Verbel, “N-Linear Algebraic Maps for Chemical Structure Codification: A Suitable Generalization for Atom-pair Approaches?,” *Curr. Drug Metab.*, vol. 15, pp. 441–469, 2014.

[53] C. García-Jacas, Y. Marrero Ponce, S. J. Barigye, T. Hernández-Ortega, L. Cabrera-Leyva, and A. Fernández-Castillo, “N-tuple topological/geometric cutoffs for 3D N-linear algebraic molecular codifications: variability, linear independence and QSAR analysis,” *SAR QSAR Environ. Res.*, vol. 27, no. 12, pp. 949–975, 2016.

[54] C. García-Jacas, Y. Marrero-Ponce, L. Acevedo-Martínez, S. J. Barigye, J. R. Valdés-Martiní, and E. Contreras-Torres, “QuBiLS-MIDAS: A parallel free-software for molecular descriptors computation based on multilinear algebraic maps,” *J. Comput. Chem.*, vol. 35, no. 18, pp. 1395–1409, 2014.

[55] R. Todeschini, V. Consonni, and P. Gramatica, “Chemometrics in QSAR,” in *Comprehensive Chemometrics*, Elsevier, 2009, pp. 129–172.

[56] S. Wold, L. Eriksson, and S. Clementi, “Statistical Validation of QSAR Results,” in *Chemometric Methods in Molecular Design*, 2008, pp. 309–338.

[57] A. S. H. Samprit Chatterjee, *Regression Analysis by Example*, 5th editio. Wiley, 2012.

[58] E. Suárez, C. Pérez, R. Rivera, and M. Martinez, “Matrix Representation of the Linear Regression Model,” in *Applications of Regression Models in Epidemiology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017, pp. 49–68.

[59] E. Suárez, C. Pérez, R. Rivera, and M. Martinez, “Selection of Variables in a Multiple Linear Regression Model,” in *Applications of Regression Models in Epidemiology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017, pp. 77–86.

[60] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics*. Hoboken, NJ, USA:

John Wiley & Sons, Inc., 1980.

[61] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” *AI Commun.*, vol. 30, no. 2, pp. 169–190, May 2017.

[62] F. Pan, G. Song, X. Gan, and Q. Gu, “Consistent feature selection and its application to face recognition,” *J. Intell. Inf. Syst.*, vol. 43, no. 2, pp. 307–321, Oct. 2014.

[63] R. Bisquerra, *Introducción conceptual al análisis multivariable: un enfoque informático con los paquetes SPSS-X, BMDP, LISREL y SPAD*. Barcelona: PPU, 1989.

[64] J. Dearden, M. T. D. Cronin, and K. Kaiser, *How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR)*, vol. 20. 2009.

[65] W. J. Egan and S. L. Morgan, “Outlier Detection in Multivariate Analytical Chemical Data,” *Anal. Chem.*, vol. 70, no. 11, pp. 2372–2379, Jun. 1998.

[66] A. Tropsha, P. Gramatica, and V. K. Gombar, “The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models,” *QSAR Comb. Sci.*, vol. 22, no. 1, pp. 69–77, 2003.

[67] C. Léger, D. N. Politis, and J. P. Romano, “Bootstrap Technology and Applications,” *Technometrics*, vol. 34, no. 4, pp. 378–398, 1992.

[68] J. Shao, “Bootstrap Model Selection,” *J. Am. Stat. Assoc.*, vol. 91, no. 434, pp. 655–665, Jun. 1996.

[69] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen, *Assessing the accuracy of prediction algorithms for classification: An overview*, vol. 16. 2000.

[70] W. F. Massy, “Principal Components Regression in Exploratory Statistical Research,” *J. Am. Stat. Assoc.*, vol. 60, no. 309, pp. 234–256, Mar. 1965.

[71] J. F. Hair and W. C. . Black, *Multivariate Data Analysis*, Seventh Ed. Edinburgh: Pearson Education, 2014.

[72] M. Randić, “Molecular Shape Profiles,” *J. Chem. Inf. Model.*, vol. 35, no. 3, pp. 373–382, May 1995.

[73] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[74] S. J. Barigye *et al.*, “Relations frequency hypermatrices in mutual, conditional, and joint entropy-based information indices,” *J. Comput. Chem.*, vol. 34, no. 4, pp. 259–274, 2012.

[75] J. W. Godden, F. L. Stahura, and J. Bajorath, “Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations,” *J. Chem. Inf. Comput. Sci.*, vol. 40, no. 3, pp. 796–800, May 2000.

[76] I. H. Witten, E. Frank, M. A. Hall, and C. J. B. T.-D. M. (Fourth E. Pal, Eds., “Appendix B - The WEKA workbench,” in *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2017, pp. 553–571.

[77] J. B. Grace, “Bioinformatics: Mathematical Challenges and Ecology,” *Science (80-.).*, vol. 275, no. 5308, p. 1861 LP-1865, Mar. 1997.

[78] E. Marshall, “Hot Property: Biologists Who Compute,” *Science (80-.).*, vol. 272, no. 5269, p. 1730 LP-1732, Jun. 1996.

[79] S. Hellberg *et al.*, “The Prediction of Bradykinin Potentiating Potency of Pentapeptides. An Example of a Peptide Quantitative Structure-activity Relationship.,” *Acta Chem. Scand.*, vol. 40b, pp. 135–140, 1986.

[80] S. Hellberg, M. Sjoestroem, B. Skagerberg, and S. Wold, “Peptide quantitative

structure-activity relationships, a multivariate approach," *J. Med. Chem.*, vol. 30, no. 7, pp. 1126–1135, Jul. 1987.

[81] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, and S. Wold, "Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids," *Quant. Struct. Relationships*, vol. 8, no. 3, pp. 204–209, Oct. 1989.

[82] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold, "New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids," *J. Med. Chem.*, vol. 41, no. 14, pp. 2481–2491, Jul. 1998.

[83] J. Jonsson, M. Sandberg, and S. Wold, "The evolutionary transition from uracil to thymine balances the genetic code," *J. Chemom.*, vol. 10, no. 2, pp. 163–170, Oct. 1996.

[84] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, "On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization," *J. Chem. Inf. Comput. Sci.*, vol. 40, no. 5, pp. 1235–1244, Sep. 2000.

[85] M. Randić and A. Balaban, "On A Four-Dimensional Representation of DNA Primary Sequences," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 532–539, Mar. 2003.

[86] Y. Marrero Ponce, H. González-Díaz, V. R. Zaldivar, F. Torrens, and E. A. Castro, "3D-Chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of σ-receptor antagonist activities," *Bioorg. Med. Chem.*, vol. 12, no. 20, pp. 5331–5342, 2004.

[87] K.-C. Chou, "A novel approach to predicting protein structural classes in a (20–1)-D

amino acid composition space," *Proteins Struct. Funct. Bioinforma.*, vol. 21, no. 4, pp. 319–344, Oct. 1995.

[88] K.-C. Chou, "A Key Driving Force in Determination of Protein Structural Classes," *Biochem. Biophys. Res. Commun.*, vol. 264, no. 1, pp. 216–224, 1999.

[89] K.-C. Chou, "Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect," *Biochem. Biophys. Res. Commun.*, vol. 278, no. 2, pp. 477–483, 2000.

[90] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins Struct. Funct. Bioinforma.*, vol. 43, no. 3, pp. 246–255, 2001.

[91] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[92] J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins," *J. Mol. Biol.*, vol. 120, no. 1, pp. 97–120, 1978.

[93] P. Y. Chou and G. D. Fasman, "Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins," *Biochemistry*, vol. 13, no. 2, pp. 211–222, Jan. 1974.

[94] V. Lim, "Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure," *J. Mol. Biol.*, vol. 88, no. 4, pp. 857–872, 1974.

[95] K. C. Chou, G. Nemethy, and H. A. Scheraga, "Energetics of interactions of regular structural elements in proteins," *Acc. Chem. Res.*, vol. 23, no. 5, pp. 134–141, May 1990.

[96] M. K. Gilson and B. H. Honig, “Energetics of charge–charge interactions in proteins,” *Proteins Struct. Funct. Bioinforma.*, vol. 3, no. 1, pp. 32–52, Oct. 1988.

[97] M. Levitt, “Protein folding by restrained energy minimization and molecular dynamics,” *J. Mol. Biol.*, vol. 170, no. 3, pp. 723–764, 1983.

[98] D. H. J. Mackay, A. J. Cross, and A. T. Hagler, “The Role of Energy Minimization in Simulation Strategies of Biomolecular Systems,” in *Prediction of Protein Structure and the Principles of Protein Conformation*, G. D. Fasman, Ed. Boston, MA: Springer US, 1989, pp. 317–358.

[99] R. Ramos de Armas, H. González Díaz, R. Molina, and E. Uriarte, “Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants,” *Proteins Struct. Funct. Bioinforma.*, vol. 56, no. 4, pp. 715–723, May 2004.

[100] H. González-Díaz and E. Uriarte, “Proteins QSAR with Markov average electrostatic potentials,” *Bioorg. Med. Chem. Lett.*, vol. 15, no. 22, p. 5088—5094, Nov. 2005.

[101] Y. Ruiz-Blanco, Y. Marrero Ponce, P. Prieto, J. Salgado, Y. Garcia, and C. Sotomayor Torres, “A Hooke’s law-based approach to protein folding rate,” *J. Theor. Biol.*, vol. 364, pp. 407–417, 2015.

[102] A. Giuliani, L. Di Paola, and R. Setola, “Proteins as Networks: A Mesoscopic Approach Using Haemoglobin Molecule as Case Study,” *Curr. Proteomics*, vol. 6, no. 4, pp. 235–245, 2009.

[103] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, “Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, vol. 39, no. suppl_2, pp.

W385–W390, Jul. 2011.

[104] P. Du, X. Wang, C. Xu, and Y. Gao, “PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou’s pseudo-amino acid compositions,” *Anal. Biochem.*, vol. 425, no. 2, pp. 117–119, 2012.

[105] Y. Ruiz-Blanco, W. Paz, J. Green, and Y. Marrero Ponce, “ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins,” *BMC Bioinformatics*, vol. 16, no. 1, p. 162, Dec. 2015.

[106] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, “protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences,” *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, Jun. 2015.

[107] L. Pauling and R. B. Corey, “Two hydrogen-bonded spiral configurations of the polypeptide chain,” *J. Am. Chem. Soc.*, vol. 72, no. 11, p. 5349, 1950.

[108] J. D. Watson and F. H. C. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, p. 737, Apr. 1953.

[109] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wykoff, and D. C. Phillips, “A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis,” *Nature*, vol. 181, p. 662, Mar. 1958.

[110] C. B. Anfinsen, E. Haber, M. Sela, and F. H. White Jr, “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 47, no. 9, pp. 1309–1314, Sep. 1961.

[111] C. B. Anfinsen, “Principles that Govern the Folding of Proteins Chains,” *Nobel Lect.*, vol. 181, no. 4096, pp. 223–230, 1973.

[112] K. Yue and K. A. Dill, “Sequence-structure relationships in proteins and copolymers,”

Phys. Rev. E, vol. 48, no. 3, pp. 2267–2278, Sep. 1993.

[113] M. Karplus, "The Levinthal paradox: yesterday and today," *Fold. Des.*, vol. 2, pp. S69–S75, 1997.

[114] K. W. Plaxco, K. T. Simons, and D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins," *J. Mol. Biol.*, vol. 277, no. 4, pp. 985–994, 1998.

[115] M. M. Gromiha and S. Selvaraj, "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction," *J. Mol. Biol.*, vol. 310, no. 1, pp. 27–32, 2001.

[116] H. Zhou and Y. Zhou, "Folding Rate Prediction Using Total Contact Distance," *Biophys. J.*, vol. 82, no. 1, pp. 458–463, Jan. 2002.

[117] B. Nölting *et al.*, "Structural determinants of the rate of protein folding," *J. Theor. Biol.*, vol. 223, no. 3, pp. 299–307, 2003.

[118] Z. Ouyang and J. Liang, "Predicting protein folding rates from geometric contact and amino acid sequence," *Protein Sci.*, vol. 17, no. 7, pp. 1256–1263, 2008.

[119] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *J. Mol. Biol.*, vol. 247, pp. 536–540, 1995.

[120] N. S. Bogatyreva, A. A. Osypov, and D. N. Ivankov, "KineticDB: A database of protein folding kinetics," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 1, pp. 342–346, 2009.

[121] R. W. Pino, S. J. Barigye, Y. Marrero Ponce, C. García-Jacas, J. R. Valdes-Martiní, and F. Perez-Gimenez, "IMMAN: free software for information theory-based

chemometric analysis," *Mol. Divers.*, vol. 19, no. 2, pp. 305–319, 2015.

[122] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, *MobyDigs: software for regression and classification models by genetic algorithms*, vol. 23. 2003.

[123] A. Mishra, P. S. Rana, A. Mittal, and B. Jayaram, "D2N: Distance to the native," *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1844, no. 10, pp. 1798–1807, 2014.

[124] J. A. Castillo-Garit, Y. Marrero Ponce, F. Torrens, and R. Rotondo, "Atom-based stochastic and non-stochastic 3D-chiral bilinear indices and their applications to central chirality codification," *J. Mol. Graph. Model.*, vol. 26, no. 1, pp. 32–47, 2007.

[125] J. A. Castillo-Garit, Y. Marrero Ponce, and F. Torrens, "Atom-based 3D-chiral quadratic indices. Part 2: Prediction of the corticosteroid-binding globulin binding affinity of the 31 benchmark steroids data set," *Bioorganic Med. Chem.*, vol. 14, no. 7, pp. 2398–2408, 2006.

[126] E. R. Collantes and W. J. Dunn, "Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues," *J. Med. Chem.*, vol. 38, no. 14, pp. 2705–2713, 1995.

[127] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, 1982.

[128] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proc. Natl. Acad. Sci. USA*, vol. 78, no. 6, pp. 3824–3828, 1981.

[129] A. Sillero and J. M. Ribeiro, "Isoelectric points of proteins: Theoretical determination," *Anal. Biochem.*, vol. 179, no. 2, pp. 319–325, 1989.

[130] A. A. Zamyatnin, "Protein volume in solution," *Prog. Biophys. Mol. Biol.*, vol. 24, no.

C, pp. 107–123, 1972.

[131] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, Seventh Ed. New York: Macmillan Learning, 2017.

[132] A. Balaban, “Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs,” *J. Chem. Inf. Comput. Sci.*, vol. 34, no. 2, pp. 398–402, Mar. 1994.

[133] R. Todeschini and V. Consonni, “New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees,” *MATCH - Commun. Math. Comput. Chem.*, vol. 64, pp. 359–372, 2010.

[134] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[135] Y.-D. Cai, K.-Y. Feng, W.-C. Lu, and K.-C. Chou, “Using LogitBoost classifier to predict protein structural classes,” *J. Theor. Biol.*, vol. 238, no. 1, pp. 172–176, 2006.

[136] T.-L. Zhang and Y.-S. Ding, “Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes,” *Amino Acids*, vol. 33, no. 4, pp. 623–629, 2007.

[137] Y.-D. Cai, X.-J. Liu, X. Xu, and K.-C. Chou, “Prediction of protein structural classes by support vector machines,” *Comput. Chem.*, vol. 26, no. 3, pp. 293–296, 2002.

[138] K. E. Chen, L. A. Kurgan, J. Ruan, and C. Tg, “Prediction of Protein Structural Class Using Novel Evolutionary Collocation-Based Sequence Representation,” *J. Comput. Chem.*, vol. 00000, pp. 1–9, 2008.