Universidad Central "Marta Abreu" de las Villas

Facultad de Matemática, Física y Computación

Centro de Estudios de Informática

Procedimiento de extracción de rasgos 3D-proteicos basado en Álgebra Lineal: Aplicaciones en estudios bioinformáticos

Tesis presentada en opción al Título Académico de

Máster en Bioinformática y Biología Computacional

por:

Ing. Ernesto Contreras Torres

Tutores:

Prof. Tit, Yovani Marrero Ponce, **Dr.C.**

Prof. Asist, César Raúl García Jacas, **Dr.C.**

Prof. Asist, Andrea Samper Ortega, **M.Sc.**

Santa Clara, **2016**

Agradecimientos

Agradezco en primera instancia, a mi “GRAN FAMILIA”, especialmente, a mis abuelas, abuelos, padres y hermano por la conducción excepcional de mi formación. A mis familiares de La Habana, por abrirme las puertas de sus hogares. A todos mis amigos de estudio, trabajo y fusil, desde el círculo infantil hasta el presente. A todos mis maestros y personal auxiliar en todos los niveles de enseñanza. Debo destacar, a mi tutor “Trinchet”, que me inició en el apasionante mundo de la Inteligencia Artificial por su guía acertada y confianza depositada. A mis tutores, los Doctores en Ciencias: Yovani Marrero Ponce y César Raúl García Jacas por todas las enseñanzas, paciencia, guía certera y motivación. Es el momento preciso para reconocer de forma ESPECIAL a la OBRA de toda la VIDA a la MAESTRA en CIENCIAS Andrea Samper Ortega, por sus demostraciones de amistad, integridad, profesionalidad, honestidad, altruismo y aportes invaluables a mi formación personal y profesional. A TonySé de la Rosa, por “plantarse” en tres y dos para que yo comenzara esta Maestría, cuando algunos dudaron de mis capacidades. A esos, también les agradezco, pues me incentivaron a seguir adelante. A todo el Comité Académico de la Maestría (de forma especial a los profesores Gladyta y Grau, a quienes admiro y respeto por su sabiduría y humildad); y a esta legendaria UNIVERSIDAD por ofrecerme la posibilidad de continuar mi formación postgraduada. A quien considero un gran amigo: Reisel González. Asimismo, es justo reconocer a Mario Pupo por todos los aportes al desarrollo de esta investigación. También, agradezco a Rigo por las sugerencias en los inicios de la implementación del software y su visión humorística de esta “jungla de cemento”. Además, debo agradecer a Roberto Téllez por sus oportunas contribuciones en el desarrollo del “Front-End”. Finalmente, y no por ello menos importante, es justo reconocer a los Doctores en Ciencias Médicas: Alicita y Blas por su apoyo en momentos difíciles. A TODOS, sinceramente les digo: MUCHAS GRACIAS.

Si se aspira a obtener victoria, se debe lograr el balance entre dos aspectos contrapuestos*: exploración y explotación*.

Contreras-Torres, E.

Dedicatoria

A mis tres reyes: mi ***Madre****, mi* ***Padre*** *y* ***Dios****.*

A mi ***Hermano****.*

A toda mi ***Familia****.*

Resumen

En el presente trabajo, se propone un nuevo procedimiento para la extracción de rasgos tridimensionales (3D) proteicos basado en las formas algebraicas 2-lineales utilizando la k^{th} matriz multi-métrica bidimensional de similitud-disimilitud para codificar información relativa a las interacciones no covalentes de estos biopolímeros. Se proponen además esquemas de generalización para el cálculo de las distancias inter-atómicas mediante el empleo de varias métricas. Se usaron las matrices simple-estocástica y de probabilidad mutua para normalizar la matriz multi-métrica bidimensional de similitud-disimilitud no estocástica. Asimismo, se generaliza la obtención de índices totales y locales por medio de varios operadores de agregación. Con el objetivo de discriminar entre las diferentes interacciones no covalentes entre las cadenas laterales de los aminoácidos, se definen procedimientos de cortes macromoleculares geométricos y topológicos. Además, se desarrolló un software denominado *ToMoCoMD-CAMPS MuLiMs-MCoMPAs* que automatiza el cálculo de los descriptores propuestos. Se realizaron estudios de variabilidad basado en entropía de Shannon y análisis de componentes principales. Adicionalmente, se creó una métrica denominada *Entropía Promedio de Shannon Estandarizada* y una nueva representación gráfica, de utilidad en los análisis de variabilidad. Además, los descriptores propuestos se aplicaron satisfactoriamente en la clasificación estructural de proteínas, así como en la predicción de la velocidad de plegamiento de cadenas polipeptídicas. En ambos estudios se obtuvieron modelos robustos y de buena capacidad predictiva. Finalmente, se anticipa la potencial aplicación de los descriptores propuestos en la modelación de otras propiedades biológicas y/o funciones de interés en ciencia de proteínas.

Palabras Clave: *descriptor 3D-proteico, formas algebraicas, matriz multi-métrica de similitud-disimilitud, métricas, operadores de agregación*

Índice

AGRADECIMIENTOS I

DEDICATORIA III

RESUMEN IV

ÍNDICE 5

INTRODUCCIÓN 8

CAPÍTULO 1: CONSIDERACIONES DE LA REVISIÓN BIBLIOGRÁFICA 14

1.1. Descriptores 3D-proteicos 14
1.2. Funciones para modelar el comportamiento de un fenómeno 16
1.3. Métricas de distancia para representar relaciones inter-atómicas 17
1.4. Operadores de agregación para generalizar las contribuciones atómicas 19
1.5. Métodos estadísticos y de aprendizaje automático utilizados en los estudios realizados 20
1.6. Conclusiones del capítulo 22

CAPÍTULO 2: PROCEDIMIENTO DE EXTRACCIÓN DE RASGOS 3D-PROTEICOS 23

2.1 Representaciones 3D-proteicas 23
2.2. Vector macromolecular basado en propiedades de la cadena lateral de aminoácidos para la representación de cadenas polipeptídicas 24
2.3. Matriz multi-métrica de similitud-disimilitud: nueva representación de la estructura tridimensional de proteínas 24
2.4. Procedimientos de normalización de la matriz multi-métrica bidimensional de similitud-disimilitud 25
2.5. Definición matemática de los descriptores 3D-proteicos totales y locales 26
2.6. Operadores de agregación de las contribuciones atómicas 28
2.7. Procedimientos de cortes macromoleculares: geométricos y topológicos 29
2.8. Descripción del procedimiento de extracción de rasgos 3D-proteicos 31
2.9. Conclusiones del capítulo 32

ÍNDICE

CAPÍTULO 3: SOFTWARE PARA EL CÁLCULO DE LOS DESCRIPTORES 3D-PROTEICOS 33

3.1. Biblioteca Chemical Development Kit (CDK) 33
3.2. Biblioteca Jmol 33
3.3. Diseño del software ToMoCoMD-CAMPS MuLiMs-MCoMPAs 33
3.3.1 Front-end: Interfaz gráfica de usuario 34
3.3.2. Back-end: Biblioteca de clases para calcular los descriptores 3D-proteicos 35
3.4. Cálculo multi-núcleo de los descriptores 3D-proteicos 36
3.5. Evaluación del rendimiento del software ToMoCoMD-CAMPS 37
3.6. Conclusiones del capítulo 39

CAPÍTULO 4: EVALUACIÓN DE LOS NUEVOS DESCRIPTORES 3D-PROTEICOS 41

4.1. Análisis de variabilidad basado en Entropía de Shannon de los 3D-DMs 41
4.1.1. Análisis comparativo de los 3D-DMs acorde a las métricas utilizadas para el cálculo de distancias inter-atómicas 42
4.1.2. Análisis comparativo de los 3D-DMs conforme al uso de propiedades de la cadena lateral de aminoácidos 43
4.1.3. Análisis comparativo de los 3D-DMs acorde al operador de agregación aplicado 44
4.2. Análisis de ortogonalidad de los 3D-DMs 45
4.2.1. Independencia lineal de los 3D-DMs acorde a la métrica utilizada para calcular la distancia inter-atómica 46
4.2.2. Independencia lineal de los 3D-DMs acorde a la propiedad de aminoácido utilizada 46
4.2.3. Independencia lineal de los 3D-DMs acorde al operador de agregación utilizado 46
4.3. Aplicación de los 3D-DMs a la predicción de las clases estructurales de proteínas 47
4.3.1. Descripción del conjunto de datos 47
4.3.2. Medidas de evaluación del desempeño de los modelos de clasificación 47
4.3.3. Desarrollo de los modelos de clasificación 47
4.3.4. Evaluación del desempeño de los modelos de clasificación 49

ÍNDICE

Tabla 1.1: Parámetros estadísticos del modelo obtenido usando Random Forest 50

Tabla 1.2: Parámetros estadísticos del modelo obtenido usando K-NN 50

Tabla 1.3: Parámetros estadísticos del modelo obtenido usando MLP 50

4.4. Aplicación de los 3D-DMs a la predicción de la velocidad de plegamiento de cadenas polipeptídicas 51

4.4.1. Descripción de los conjuntos de entrenamiento y prueba 51

4.4.2. Medidas de evaluación de los modelos de regresión 51

4.4.4. Desarrollo de los modelos de regresión 53

4.4.3. Evaluación del desempeño de los modelos de regresión 54

4.5. Conclusiones del capítulo 55

CONCLUSIONES 56

RECOMENDACIONES 57

REFERENCIAS BIBLIOGRÁFICAS 58

PRODUCCIÓN CIENTÍFICA DEL AUTOR 65

Introducción

Como se argumenta en [1] el notable crecimiento de las bases de datos de secuencia y estructura de proteínas ha ampliado la brecha existente entre el número de proteínas de estructura conocida y el número de estas con propiedades y/o funciones conocidas. Particularmente importante resulta la descripción de su estructura tridimensional (terciaria o 3D), pues como es bien conocido, existe una estrecha relación entre la conformación espacial de estos biopolímeros y la(s) función(es) biológica(s) que desempeñan en los organismos vivos [2]. Como se describe en [3], el desarrollo de una terapia para determinada patología es un proceso comúnmente constituido por tres pasos. El primer paso es la identificación de la diana biológica o terapéutica, es decir, la identificación de una molécula biológica, principalmente proteínas, involucrada en algún mecanismo implicado en algún proceso patológico. Un estudio relativamente reciente desarrollado por el Boston Consulting Group y que implicó a 50 compañías e instituciones académicas, reveló que el proceso de desarrollo de un nuevo medicamento hasta su uso autorizado en terapéutica requiere, en promedio, la inversión de 880 millones de dólares (USD) y 15 años de investigación [3]. Lo anterior evidencia la extrema complejidad asociada a la tarea de desarrollar "un nuevo medicamento", pero también muy valorada por la sensibilidad que genera el impacto negativo de las enfermedades en la sociedad moderna.

Los estudios QSAR (acrónimo de Quantitative Structure-Activity Relationships) permiten estimar, con aceptable grado de precisión, la actividad/propiedad de nuevos compuestos por lo que pueden aplicarse como estrategia de tamizaje virtual como alternativa a los costosos procesos de síntesis y bioensayos. El desarrollo de métodos para la caracterización numérica de proteínas, también conocidos como [descriptores biomacromoleculares (DMs)] y su aplicación combinada con técnicas estadísticas y/o de aprendizaje automático ha demostrado ser efectiva en la predicción de propiedades biológicas de interés [4-7]. Resulta conveniente señalar que la codificación de la

estructura química de pequeñas moléculas orgánicas ha sido un área intensamente abordada, evidenciado en el amplio número de descriptores moleculares (DMs) reportados hasta el presente [8-12]. Sin embargo, en el contexto de proteínas, se reportan pocos parámetros para caracterizar su secuencia [13] y en mucho menor grado para considerar su estructura tridimensional [14-17]. Además, resulta altamente difícil representar toda la complejidad molecular y modelar todas las interacciones biológicas empleando un único descriptor o un pequeño número de estos, pues el (los) mismo(s) solo codifica(n) una porción de la información química contenida en representaciones específicas de la estructura molecular [9].

En este contexto, el desarrollo de estrategias de caracterización cuantitativa de proteínas, constituye, sin lugar a dudas, un área emergente, debido a la imperiosa necesidad de métodos que permitan estudiar la relación estructura-(propiedad y/o función) en las bases de datos mencionadas anteriormente [3, 18]. Por lo tanto, el desarrollo de nuevos descriptores y nuevas representaciones de proteínas capaces de extraer información relevante y por tanto, proveer una mejor caracterización de su estructura macromolecular, constituye un área de creciente importancia [19, 20].

Como se describe en [18], la extensión "natural" de los descriptores propuestos para caracterizar pequeñas moléculas orgánicas ha sido empleada como estrategia general en la definición de descriptores topológicos (2D) proteicos. Esta idea intuitiva también se utilizó en la definición de los índices 2D-proteicos basados en las formas algebraicas: cuadráticas [5], lineales [21] y bilineales [22], los cuales se aplicaron de forma satisfactoria en estudios de predicción de estabilidad biológica. Recientemente fueron introducidos los índices geométricos (3D) QuBiLS-MIDAS (acrónimo de Quadratic, Bilinear and N-Linear Algebraic Maps based on n-Tuple (Dis)-Similarity Matrix and Atomic WeightingS) para la codificación de la estructura química de moléculas orgánicas [23]. Estos 3D-DMs tienen los siguientes rasgos: 1) están basados en las formas algebraicas 2-lineales, 2)

emplean métricas diferentes a la Euclidiana para el cálculo de las distancias inter-atómicas, 3) utilizan operadores de agregación distintos a la suma sobre los índices atómicos para generalizar la obtención de DMs totales y locales. Estos DMs se aplicaron satisfactoriamente en diferentes estudios quimio-informáticos [23, 24], en los cuales se obtuvieron modelos de alto poder predictivo, lo que sugiere que estos DMs codifican información relevante de la estructura molecular.

Basados en el buen desempeño de los índices bilineales (IB)-2D-proteicos y los QuBiLS-MIDAS se introdujeron en los IB-3D-proteicos sobre la *matriz coulómbica* [25]. Estos nuevos DMs utilizan tres ($1 \leq p \leq 3$) métricas de Minkowksi para el cálculo de las distancias entre las posiciones espaciales de los átomos C_α de los aminoácidos, la suma como operador de agregación (OA) de las contribuciones atómicas, y se aplicaron satisfactoriamente en la discriminación estructural de proteínas. El buen desempeño de los índices QuBiLS-MIDAS y los IB-3D-proteicos, sugiere la exploración de otros objetos matemáticos como las formas algebraicas: lineales y cuadráticas, el uso de otras métricas de distancia para el establecimiento de relaciones inter-atómicas y la aplicación de OA diferentes a la suma para generalizar la obtención de índices totales y locales. Lo anterior pudiera conducir a alcanzar resultados similares o superiores respecto a los obtenidos por los IB-3D-proteicos.

Por otra parte, el lineamiento 131 de la política económica y social del Partido y la Revolución plantea lo siguiente: “Sostener y desarrollar los resultados alcanzados en el campo de la biotecnología, la producción médico-farmacéutica,…, las ciencias básicas, las ciencias naturales,…, y los servicios científicos y tecnológicos de alto valor agregado”.

Por todo lo anterior, se plantea el siguiente ***problema científico***:

Resulta importante la definición de nuevos descriptores 3D-proteicos, con contenidos de información diferentes a los ya reportados, que permitan obtener modelos para estimar propiedades y/o funciones

biomacromoleculares de interés con un grado de precisión aceptable, de forma que resulten útiles en estudios bioinformáticos.

En tal sentido, se propone como **objetivo general**: Proponer nuevos descriptores basados en las formas algebraicas 2-lineales, que proporcionen a los investigadores nuevas herramientas para el establecimiento de relaciones cuantitativas de la estructura 3D de proteínas y propiedades y/o funciones, y como **objetivos específicos**:

- Definir nuevos descriptores 3D-proteicos (totales y locales) basados en las formas algebraicas 2-lineales en $\mathbb{R}^n$ mediante el uso de diversas métricas para el cálculo de las distancias inter-atómicas, y la aplicación de diferentes operadores de agregación para generalizar las contribuciones atómicas.
- Desarrollar el 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-Acids Weightings) para automatizar el cómputo de los descriptores propuestos.
- Evaluar el desempeño de los descriptores propuestos acorde a la información que estos codifican.
- Evaluar la utilidad de los índices propuestos en la caracterización de la estructura macromolecular en estudios de discriminación estructural, así como en estudios de predicción de la velocidad de plegamiento.

Luego de la realización de la revisión bibliográfica se formuló la siguiente hipótesis de investigación:

Es posible definir nuevas familias de descriptores 3D-proteicos basados en las formas algebraicas 2-lineales, utilizando diversas métricas para el cálculo de las distancias inter-atómicas y operadores de agregación diferentes a la suma, que permitan obtener modelos que correlacionen cuantitativamente la estructura tridimensional de proteínas con propiedades y/o funciones de interés.

Novedad, aportes y estructura del reporte de tesis:

La *novedad científica* de este trabajo se fundamenta en la propuesta de un nuevo procedimiento de extracción de rasgos 3D-proteicos basado en las formas algebraicas 2-lineales, utilizando diversas métricas para el cálculo de las distancias inter-atómicas y la aplicación de varios operadores de agregación sobre los índices de nivel atómico.

En este trabajo pueden destacarse los aportes siguientes:

Valor teórico:

1) La aplicación de diversas métricas de distancia para el cálculo de las distancias inter-atómicas.
2) Se crea una nueva métrica para cuantificar el contenido de información promedio de un conjunto de descriptores moleculares.

Valores prácticos:

1) Se desarrolla la aplicación informática denominada ToMoCoMD-CAMPS MuLiMs-MCoMPAs para el cómputo de los descriptores propuestos. ToMoCoMD-CAMPS permite además la generación y visualización de las representaciones 3D-proteicas propuestas.
2) Se crea una representación gráfica más sencilla e interpretable que las reportadas hasta la fecha para realizar análisis de variabilidad de descriptores moleculares.

3) Se aplican los descriptores propuestos en la clasificación estructural de proteínas, y en la predicción de la velocidad de plegamiento de cadenas polipeptídicas.

La *tesis* está *estructurada* en cuatro capítulos. En el Capítulo 1 se abordan los diferentes tópicos contemplados en la tesis. En el Capítulo 2 se definen los elementos teóricos que constituyen el fundamento del procedimiento propuesto para el cálculo de los descriptores 3D-proteicos. En el Capítulo 3 se aborda el diseño del software desarrollado y las pruebas realizadas al mismo. En el Capítulo 4 se presenta la evaluación el mérito de los descriptores propuestos: se evalúa su contenido de información y posible ortogonalidad, así como su utilidad en estudios QSAR. Finalmente, se presentan las Conclusiones, Recomendaciones, Referencias Bibliográficas, Producción Científica del Autor y Anexos.

Capítulo 1: Consideraciones de la Revisión Bibliográfica

En este capítulo se presentan los resultados de la revisión bibliográfica de los principales descriptores 3D-proteicos reportados hasta el presente. Además, se realiza un esbozo de los elementos de Álgebra Lineal empleados en la definición de los 3D-DMs que se proponen. Finalmente, se abordan las técnicas estadísticas y de aprendizaje automático utilizadas en los estudios realizados.

1.1. Descriptores 3D-proteicos

Los DMs pueden definirse como representaciones matemáticas de las moléculas que se obtienen al aplicar algoritmos específicos sobre una representación molecular definida, o a partir de procedimientos experimentales específicos [9]. También se podrían definir como el producto final de un procedimiento lógico-matemático que transforma la información química, codificada por una representación simbólica de una molécula, en un número útil o representan el resultado de algún experimento estandarizado [9]. La utilidad de un DM debe analizarse con doble sentido: el número puede brindar una interpretación más profunda en términos estructurales de la propiedad molecular y/o es capaz de tomar parte en un modelo para la predicción de propiedades moleculares de interés [9]. Los descriptores 3D o geométricos son aquellos que se calculan sobre representaciones del objeto rígido en el espacio definido por las coordenadas cartesianas (x,y,z) de los átomos de la molécula. En general, la búsqueda de nuevos DMs y su aplicación en el establecimiento de relaciones cuantitativas de estructura-actividad constituye actualmente un campo muy abordado, con un alto número de DMs moleculares reportados [9]. Sin embargo, como se explicó anteriormente, el desarrollo de descriptores biomacromoleculares se considera un área de creciente importancia.

En relación con lo anterior, una de las primeras aproximaciones para la descripción cuantitativa de proteínas es el radio de giro (RG) propuesto por Flory [26]. Otras de las medidas utilizadas para caracterizar la estructura 3D de proteínas es el coeficiente de compactación para la identificación de dominios proteicos [27]. Otros recursos para la descripción de proteínas son: el Orden de Contacto Relativo [28], la Distancia Total de Contactos [14], el Orden de Contactos de Amplio Rango [29], el Orden Efectivo de Contactos [30], el Parámetro Topológico de Cadenas [31], el Número de Contactos [32], el Grado de Aglomeración [33], la Longitud Efectiva [34] y el Orden Absoluto de Contactos [35], donde se obtuvo una buena correlación entre estos parámetros estructurales simples y la velocidad del plegamiento de proteínas. Otros estudios relacionados con la velocidad del plegamiento se encuentran en los siguientes reportes [15, 16].

Otra enfoque propuesto para la codificación de la estructura 3D de proteínas se basa en el uso de cadenas de Markov [17]. Por otra parte, la descripción de la conectividad de los sistemas biológicos utilizando redes químicas es una puerta directa para la introducción de herramientas matemáticas en la Proteómica [36, 37]. Las redes son objetos que se componen esencialmente por nodos y aristas. Los nodos representan las partes del sistema y las aristas las relaciones geométricas y/o funcionales entre las partes. Por ejemplo: en una proteína, los aminoácidos pueden desempeñar el papel de nodos y las aristas los enlaces covalentes y no covalentes. Estas redes pueden describirse numéricamente usando los llamados índices de Conectividad (CIs) [37]. La transformación de los grafos en CI (números) facilita la manipulación de la información y la búsqueda de relaciones estructura-función en Proteómica. El uso de enfoques gráficos para estudiar sistemas biológicos puede proveer una mejor caracterización y por lo tanto codificar información química relevante en estos sistemas [20, 37]. Ejemplos de estos índices son los basados en entropía de cadenas de Markov [38], los potenciales pseudo-electrostáticos de Markov [39] y los IB-3D-proteicos [25].

Otros parámetros geométricos han sido aplicados en estudios de predicción de la afinidad entre proteína-ligando (estudios QSBR) como son: ENTess [40] e I_3 [41]. Además se han utilizado descriptores simples en la predicción de la función de proteínas (clasificación de enzimas) [42]. Finalmente, en recientes trabajos se destacan las redes de contacto de proteínas (PCNs acrónimo de *Protein Contact Networks*) como enfoque gráfico para su descripción estructural [36, 43].

1.2. Funciones para modelar el comportamiento de un fenómeno

En diversas situaciones se asigna a cada elemento de un conjunto un elemento particular de un segundo conjunto (que puede ser el mismo que el primer conjunto). Por ejemplo, supongamos que a un grupo de personas se les asigna una letra del conjunto (A, B, C, D). Y supongamos que le asignamos a Ana la A, a Benito la B, a Chou la C, y a Daniel la D, la asignación anterior es un ejemplo de función. Las funciones se usan en definiciones de estructuras discretas tales como sucesiones o cadenas. También se utilizan para representar cuánto tiempo tarda un ordenador en resolver un problema de un tamaño determinado, entre otras aplicaciones [44].

Formalmente, una función se puede definir como sigue: Sean A y B conjuntos, una función *f* de A en B es una asignación de exactamente un elemento de B a cada elemento de A [44]. Se escribe f(a)=b es el único elemento de B asignado por la función *f* al elemento a de A [44]. Si *f* es una función de A en B, escribimos *f*: A→B. Ejemplos hay varios: f(x)=x (función lineal que corta el origen de coordenadas), $f(x)=x^2$ (función cuadrática), $f(x) = x^3$ (función cúbica), entre otras.

En el contexto del Álgebra Lineal se define un tipo particular de funciones que se denominan formas algebraicas, las cuales realizan una transformación entre dos conjuntos: el espacio vectorial n-dimensional y los números reales. En general, las formas algebraicas lineales (F), bilineales (B) y cuadráticas (Q) son las más empleadas para representar las transformaciones que ocurren en el decursar de un fenómeno. Aunque muchos de los fenómenos en la naturaleza y en el quehacer

científico no se comporten linealmente, el empleo de un modelo lineal es útil, aun cuando el fenómeno se comporte no linealmente, por cuanto es posible, en muchos casos, determinar los errores que se cometen al emplear un modelo lineal y con ello, se aprovechan las ventajas que ofrece dicho modelo [45]. Las formas algebraicas lineales, bilineales y cuadráticas en $\Re^n$ han sido empleadas satisfactoriamente como formalismo matemático en la definición de descriptores moleculares [5, 11, 21]. En el presente trabajo, estas formas algebraicas también son el fundamento matemático del procedimiento de obtención rasgos 3D-proteicos propuesto. Finalmente, para detalles sobre la definición matemática de dichas formas algebraicas, remitirse a [45].

1.3. Métricas de distancia para representar relaciones inter-atómicas

Las distancias son frecuentemente utilizadas para medir la similitud entre objetos representados por un determinado número de rasgos, lo cual es común en la química analítica donde los objetos son caracterizados por señales, parámetros, entre otros. La distancia es una descripción numérica de la separación entre dos entidades (variables u objetos) [46].

Según la definición propuesta en [46], sea X un conjunto, una función d: $X \times X \rightarrow \mathbb{R}$ es considerada una distancia (o disimilitud) en X si para todo x,y que pertenecen a X cumple las siguientes propiedades:

1. $d_{xy} \geq 0$, no negatividad.
2. $d_{xx}=0$, reflexividad.
3. $d_{xy}= d_{yx}$, simetría.

Asimismo, una función d: $X \times X \rightarrow \mathbb{R}$ es considerada una métrica en X si para todo x,y,z que pertenecen a X cumple las siguientes propiedades:

1. $d_{xy} \geq 0$, no negatividad.

2. d_{xy}=0, si x=y, reflexividad fuerte.
3. d_{xy}= d_{yx}, simetría.
4. $d_{xy} \leq d_{xz}+d_{zy}$, desigualdad triangular.

Dado un conjunto de datos, el problema fundamental radica en la selección de una distancia apropiada lo cual resulta difícil por el amplio número de distancias reportadas en la literatura [46]. Un conjunto de datos químico está usualmente constituido por un número de objetos y un número de parámetros que han sido medidos para cada objeto. Por lo tanto, los datos pueden representarse mediante una matriz numérica denotada como **X** *(n×p)*, donde *n* es la cantidad de filas (objetos) y *p* es el número de columnas (variables). Los elementos x_{ij} de la matriz X representan el valor del objeto *i* y la variable *j*. La conformación espacial de una molécula (ej. una proteína) puede ser representada como una matriz de *n×3*. En este caso, *n* es el número de átomos y 3 el número de variables (x,y,z), que son las coordenadas cartesianas de cada átomo.

Resulta importante señalar que, hasta el presente, la distancia Euclídea o Euclidiana ha sido considerada prácticamente como la métrica exclusiva para establecer una relación entre pares de átomos (distancia inter-atómica) en la definición de descriptores moleculares 3D. Lo anterior pudiera considerarse "natural", pues el camino más corto entre dos puntos es la diagonal (distancia Euclídea) y por lo tanto, a menor distancia la magnitud de la interacción inter-atómica es mayor. Sin embargo, el empleo de otras métricas de distancia pudiera proveer información complementaria, no codificada por la distancia Euclídea. Por ejemplo, si para resolver determinado problema (digamos fildear una pelota, lanzar un penalti o jugar ajedrez) siempre ejecutamos un conjunto finito de acciones (algoritmo) basado únicamente en el "camino más corto", estaríamos dejando de considerar otras variantes que pudieran ser más efectivas, lo cual está en correspondencia con el teorema Non-Free Lunch (http://www.no-free-lunch.org/) para algoritmos de optimización y búsqueda. Resulta

interesante el caso particular del llamado "juego ciencia" (ajedrez), donde el *alfil* tiene una alta movilidad (se mueve en diagonal) y constituye una pieza de gran importancia, por el contrario, el *peón* (jornalero o trabajador) solo puede dar "pasos cortos" (como máximo dos en la apertura) y "sin retroceso", la dama en cualquier dirección y sin limitación, la torre de forma horizontal y vertical. El punto interesante radica en que cada pieza una juega un rol importante en la partida. Si bien el *peón* da "pasos cortos" siempre hacia delante, cuando este llega a la octava casilla tiene la posibilidad de convertirse en torre, dama, alfil o caballo, es decir, cualquier pieza de su mismo color excepto en peón o rey [47].

Por lo expresado anteriormente, y considerando que existe un número considerable de distancias reportadas en la literatura [46], la utilización de diferentes métricas de distancia para el establecimiento de relaciones inter-atómicas pudiera proveer información adicional, de utilidad en la descripción cuantitativa de la geometría molecular, hipótesis que se corroboró en estudios quimio-bio-informáticos realizados [11, 24, 25].

Como se mencionó anteriormente, los índices IB-3D-proteicos utilizan tres métricas de Minkowski para el cálculo de la distancia entre los átomos C_α de los aminoácidos, y se aplicaron satisfactoriamente en la discriminación estructural de proteínas [25]. Con el objetivo de lograr una mejor descripción de la geometría biomacromolecular, en el presente trabajo se propone el uso de otras métricas reportadas en la literatura [46] de forma que sirva como esquema de generalización para la matriz de distancia geométrica de todos los pares de átomos *i, j* de una proteína.

1.4. Operadores de agregación para generalizar las contribuciones atómicas

En Química es común el empleo de la suma como operador de agregación para condensar numéricamente la contribución (aporte) de los componentes (átomos y/o enlaces) a la estructura y

funcionamiento de un sistema (molécula) [48]. Si establecemos una analogía con el empleo de diversas métricas para el cálculo de las distancias inter-atómicas, entonces, pudiera surgir la siguiente interrogante: ¿por qué no usar operadores de agregación diferentes a la suma para generalizar el "aporte" de cada atómo a la molécula como un todo para definir descriptores 3D? La aplicación de este mismo principio también procuró resultados promisorios [10, 11, 24].

1.5. Métodos estadísticos y de aprendizaje automático utilizados en los estudios realizados

1.5.1. Análisis de Variabilidad

El método de Análisis de Variabilidad (AV) [49, 50] cuantifica el contenido de información y, por lo tanto, la variabilidad de los DMs. Este método no supervisado está basado en el cálculo de la Entropía de Shannon (ES) [51], bajo el principio de que DMs apropiados para estudios quimio-métricos pudieran poseer altos valores de entropía como un indicador de su tendencia a cambiar gradualmente con la modificación de la estructura molecular; mientras que DMs (casi constantes o constantes) pudieran tener valores bajos, siendo cero el límite para aquellos DMs que contienen el mismo valor para estructuras diferentes.

1.5.2. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es un procedimiento matemático que transforma un conjunto de variables correlacionadas de respuesta en un conjunto menor de variables no correlacionadas, llamadas *componentes principales* [52, 53]. Este procedimiento es útil cuando al trabajar con varias variables (posiblemente con un gran número de variables), se cree que existe cierto grado de redundancia en las mismas. En este caso, redundancia significa que algunas variables están correlacionadas con otras, tal vez porque están midiendo un mismo hecho. Debido a esta redundancia, puede reducirse las variables observadas en componentes principales (variables

ficticias), de forma tal, que en su totalidad expliquen la mayor varianza posible de las variables originales.

1.5.3. Random Forest

Random Forest (RF) es una técnica de clasificación, la cual consta de una combinación (ensamblado) de árboles predictores $\{h(\mathbf{x},\Theta_k), k=1, ...\}$, donde (Θ_k) son vectores aleatorios equidistribuidos de manera idéntica y cada árbol realiza el voto unitario a favor de la clase más frecuente de la entrada **x**. Para realizar la predicción de un nuevo caso, este realiza un recorrido hacia los nodos terminales de cada árbol. Luego se le asigna la etiqueta (clase) correspondiente al nodo terminal. Este proceso se repite en cada uno de los árboles del ensamblado, y la clase que obtenga el voto mayoritario es reportada como la predicción [54].

1.5.4. K-vecinos más Cercanos

Los K-vecinos más cercanos (K-NN) es una técnica de aprendizaje automático basada en instancias (casos). En este grupo de técnicas las instancias del conjunto de entrenamiento son almacenadas y se emplea una función de distancia para determinar qué caso(s) se encuentra(n) más *próximo*(s) al caso que se pretende clasificar. Resulta importante señalar que el parámetro *k* determina el número de vecinos. Frecuentemente, más de un vecino es usado (k>1) para realizar la clasificación, en estas situaciones la clase mayoritaria de los k-vecinos más cercanos (o la distancia ponderada promedio, si la variable respuesta es numérica) es asignada al nuevo caso [55].

1.5.5. Perceptrón Multicapa

El Perceptrón Multicapa (MLP) es una red neuronal artificial formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón simple [56]. El MLP puede ser totalmente o localmente conectado; en el primer caso, cada salida de una neurona de la capa *i* es entrada de todas las neuronas de la capa *i+1*,

mientras que en el segundo cada neurona de la capa *i* es entrada de una serie de neuronas (región) de la capa *i+1* [56].

1.5.6. Regresión Lineal Múltiple

La Regresión Lineal Múltiple (RLM) es una técnica estadística utilizada para estudiar las relaciones entre una única variable dependiente u objetivo y varias variables independientes (predictores) [57]. Este modelo puede ser expresado como: $Y=a+b_1X_1+b_2X_2+\ldots b_nX_n$, donde, Y es la variable dependiente o explicada, a es la intersección o término constante, X_1; X_2; X_n son las variables independientes o explicativas, y b_1; b_2;…; b_n son los parámetros que miden la influencia que las variables explicativas (X_i) tienen sobre la variable objetivo (Y) [57].

1.6. Conclusiones del capítulo

En este capítulo se presentaron las consideraciones de la revisión bibliográfica realizada. Como conclusiones se puede apuntar que:

1) En general, los descriptores 3D-proteicos definidos hasta la fecha se han aplicado en estudios relacionados con la velocidad del plegamiento de proteínas.
2) Se abordaron los elementos de Álgebra Lineal empleados para la definición de los descriptores 3D-proteicos (MuLiMs-MCoMPAs).
3) Se justificó la conveniencia de la introducción de diferentes métricas para el establecimiento de relaciones inter-atómicas, así como del uso de operadores de agregación distintos a la suma para generalizar las contribuciones atómicas.
4) Se abordaron las técnicas estadísticas y de aprendizaje automático empleadas en los estudios realizados.

Capítulo 2: Procedimiento de Extracción de Rasgos 3D-Proteicos

En este capítulo se define el sustento teórico de los descriptores 3D-proteicos (MuLiMs-MCoMPAs). Inicialmente, se definen diferentes opciones para representar la estructura tridimensional de proteínas. Luego, se generaliza la matriz geométrica por medio de diferentes métricas para el cálculo de las distancias inter-atómicas. Después, se presentan los esquemas de normalización aplicados a la matriz multi-métrica. Posteriormente, se enuncian los operadores de agregación utilizados para generalizar la obtención de índices totales y locales. A continuación, se especifican los procedimientos de cortes macromoleculares. Finalmente, se describe el proceso de ensamblaje de todos los elementos teóricos que conforman el procedimiento de extracción de rasgos 3D-proteicos.

2.1 Representaciones 3D-proteicas

El uso de las coordenadas espaciales de los C_α como alternativa para representar la posición de los aminoácidos de determinada proteína ha sido utilizada en la literatura para la definición de descriptores 3D-proteicos [9, 18, 28, 36]. En estudios recientes, se utilizó la denominada *Geometrical C_β constraint* de ciertos residuos aminoacídicos como rasgo en el desarrollo de modelos de predicción del parámetro Root-Mean-Square-Deviation (RMSD) el cual, como es bien conocido, constituye un índice comúnmente usado en la evaluación de algoritmos de predicción estructura de proteínas, obteniendo resultados prometedores [58]. Estos estudios sugieren que otras representaciones de la geometría proteica pudieran proveer nueva información, no codificada por la comúnmente utilizada, basada en las coordenadas espaciales de los C_α. Tomando en cuenta lo anterior, se proponen en el presente trabajo las siguientes opciones para representar la posición espacial de los aminoácidos: 1) C_α, 2) C_β, 3) átomo carbón del enlace amida y 4) media-aritmética [pseudo-átomo obtenido del cálculo del promedio de las coordenadas cartesianas de todos (se refiere a los átomos diferentes al hidrógeno) sus átomos].

2.2. Vector macromolecular basado en propiedades de la cadena lateral de aminoácidos para la representación de cadenas polipeptídicas

El empleo de vectores macromoleculares $\bar{x}_m$ basados en aminoácidos para la representación de cadenas polipeptídicas se describe en detalle en los siguientes reportes [5, 21, 22]. En estos estudios, se utilizan los C_α para representar los aminoácidos y propiedades de sus cadenas laterales como esquemas de ponderación. Las componentes (coordenadas) de los vectores macromoleculares son números, los cuales representan propiedades (descriptores) de la cadena lateral de los aminoácidos. Las propiedades usadas en el presente trabajo son: masa (MM) [59], volumen (MV) [60], escala de valores z [61], carga atómica (ECI) [62], área superficial (ISA) [62], índices de hidropatía: Hoop-Woods (HWS) [63] y Kyte-Doolittle (KDS) [64], punto isoeléctrico (PIE) [61]; parámetros de compatibilidad geométrica (GCP1 y GCP2); calor de formación (EPS) [65] y frecuencias relativas con que un aminoácido aparece formando hélices-α (PAH), hojas-β (PBS) y giros inversos (PTT), respectivamente [59].

Por lo tanto, una (proteína o péptido) compuesta(o) por *5,10,…, n* átomos puede representarse mediante vectores pertenecientes a los espacios R^5, R^{10}, R^n.

2.3. Matriz multi-métrica de similitud-disimilitud: nueva representación de la estructura tridimensional de proteínas

Con el objetivo de codificar información sobre las diferentes interacciones no covalentes entre los grupos *R* de los aminoácidos, las cuales tienen una alta relevancia en la formación y mantenimiento de la estructura macromolecular [66, 67], se generalizó la matriz geométrica [48] utilizando varias métricas para el cálculo de las distancias inter-atómicas. Esta matriz generalizada se denomina k^{th} matriz multi-métrica bidimensional de similitud-disimilitud (MMB-SDSM o MB^k), donde *k* es la potencia a la cual cada entrada (*mb*) de MB es elevada. Por su parte, el exponente *k* modela la relación funcional existente entre la distancia y las interacciones no covalentes entre las cadenas laterales de los aminoácidos. Los valores extremos del parámetro k=±12 están relacionados con la

forma funcional del potencial Lennard-Jones 6-12. Formalmente, los elementos *mb* de esta matriz se definen mediante la siguiente expresión:

$$mb_{ij} = \begin{cases} d_{ij}, para\ i \neq j \\ d_{io}, para\ i = j \end{cases} \quad \textbf{(1)}$$

donde, d_{ij} es la distancia entre los átomos *i* y *j*, d_{io} es la distancia al centro de la molécula de forma que los valores de la diagonal principal no son siempre ceros, con el propósito de lograr una mejor discriminación entre las estructuras biomacromoleculares. Finalmente, es importante destacar que cuando no se realizan transformaciones probabilísticas sobre MB^k, esta se denomina *k^{th} matriz multi-métrica bidimensional de similitud-disimilitud no estocástica* ($_{ns}MB^k$).

2.4.Procedimientos de normalización de la matriz multi-métrica bidimensional de similitud-disimilitud

El uso de transformaciones probabilísticas aparece reportado en la literatura en la definición de descriptores moleculares [10, 11, 39, 68]. Con el objetivo de normalizar la $_{ns}MB^k$ se utilizaron las *k^{th}* matrices simple-estocástica ($_{ss}MB^k$) y de probabilidad mutua ($_{mp}MB^k$). La $_{ss}MB^k$ es una matriz cuadrada de orden *n*, no simétrica y sus entradas ${}^{k}_{ss}mb_{ij}$ se definen como sigue:

$${}^{k}_{ss}mb_{ij} = \frac{{}^{k}_{ns}mb_{ij}}{\sum_{j=1}^{n} {}^{k}_{ns}mb_{ij}} \quad \textbf{(2)}$$

Como puede notarse en la **Ec**. **2**, la suma de los elementos de cada fila de esta matriz es utilizada como factor de escalado, obteniéndose una matriz no simétrica, donde sus respectivas entradas pueden ser interpretadas como el cambio en las probabilidades de interacción entre los grupos funcionales de aminoácidos de la proteína. Por su parte, la $_{mp}MB^k$ también es una matriz cuadrada de orden *n*, en la cual se emplea como factor de escalado la suma de todos los elementos de la matriz. Sus entradas ${}^{k}_{mp}mb_{ij}$ se calculan como sigue:

$$ {}^{k}_{mp}mb_{ij} = \frac{{}^{k}_{ns}mb_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}^{k}_{ns}mb_{ij}} \quad \textbf{(3)} $$

2.5. Definición matemática de los descriptores 3D-proteicos totales y locales

Si una proteína está compuesta por *n* átomos, entonces los k^{th} *índices bilineales, cuadráticos y lineales* para el átomo *"a"* se calculan como formas bilineales, cuadráticas y lineales, respectivamente, en la base canónica de $\mathbb{R}^n$:

$$ {}_{b}L_a = b^{a,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} mb_{ij}^{a,k} x^i y^j = [X]^T \,\mathrm{MB}^{a,k}[Y] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{4}) $$

$$ {}_{q}L_a = q^{a,k}(\bar{x},\bar{x}) = q^{a,k}(\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} mb_{ij}^{a,k} x^i x^j = [X]^T \,\mathrm{MB}^{a,k}[X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{5}) $$

$$ {}_{f}L_a = f^{a,k}(\bar{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} mb_{ij}^{a,k} u^i x^j = [U]^T \,\mathrm{MB}^{a,k}[X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{6}) $$

donde, (x_m^i, y_m^j) son las componentes de los vectores macromoleculares $(\bar{x}_m, \bar{y}_m)$ en la base canónica de $\mathbb{R}^n$. $[X]$ e $[Y]$ son vectores columnas (matrices de nx1) de las coordenadas de los vectores $\bar{x}_m, \bar{y}_m$, respectivamente; $[X]^T$es la transpuesta (matriz de 1xn) del vector de propiedades $[X]$.

Los coeficientes $mb_{ij}^{a,k}$ son los elementos de la k^{th} *matriz multi-métrica bidimensional de similitud-disimilitud de nivel atómico* ($MB^{a,k}$), los cuales se obtienen a partir de la k^{th} *matriz multi-métrica bidimensional de similitud-disimilitud total* (proteína como un todo) MB^k como sigue:

$$ \begin{aligned} mb_{ij}^{a,k} &= mb_{ij}^{k} \qquad \text{si } i = a \wedge j = a \\ &= \frac{1}{2} mb_{ij}^{k} \quad \text{si } i = a \vee j = a \\ &= 0 \qquad \text{en otro caso} \end{aligned} \quad \textbf{(7)} $$

Si una proteína es particionada entre "A" átomos, entonces la matriz MB^k puede ser particionada en "A" matrices de nivel atómico $MB^{a,k}$ y la matriz total MB^k es la suma de todas las matrices de nivel atómico. Por lo tanto, cada matriz $MB^{a,k}$ determina un índice de nivel atómico

[denotado aquí como LOVI acrónimo de LOcal Vertex Invariant] para el átomo*"a"* (L_a). De esta forma, los índices totales (proteína como un todo) se calculan como formas bilineales, cuadráticas y lineales y pueden ser representados como un vector $\bar{L}$ de longitud *n*, donde cada entrada corresponde al k^{th} índice bilineal, cuadrático o lineal de nivel atómico para el átomo*"a"*. Entonces, a partir de la definición anterior, los índices totales se calculan como combinaciones lineales de los índices de nivel atómico. Generalizaciones para este enfoque son presentadas en la sección siguiente. Como se aprecia en las (**Ecs**. 8, 9 y 10) los índices bilineales, cuadráticos y lineales se definen mediante la sumatoria de las entradas L_a del vector $\bar{L}$, lo cual es equivalente al producto entre el vector de [propiedad ($[X]^T$) o vector unidad ($[U]^T$)], la matriz MB^k y el vector de propiedad $[Y]$.

$$b^k(\bar{x},\bar{y}) = \sum_{a=1}^{n} {}_bL_a = [X]^T \mathrm{MB}^k [Y] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{8})$$

$$q^k(\bar{x}) = \sum_{a=1}^{n} {}_qL_a = [X]^T \mathrm{MB}^k [X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{9})$$

$$f^k(\bar{x}) = \sum_{a=1}^{n} {}_fL_a = [U]^T \mathrm{MB}^k [X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{10})$$

Además de codificar información de una proteína como un todo, los tres enfoques matriciales propuestos ($_{ns}MB^k$, $_{ss}MB^k$, $_{mp}MB^k$) pueden ser utilizados para codificar información de ciertos fragmentos de interés. Por lo tanto, a partir de la k^{th} matriz multi-métrica bidimensional global MB^k se puede obtener la k^{th} *matriz multi-métrica bidimensional de fragmento-local* (MB_F^k). La matriz MB_F^k contiene información sobre distancias entre los átomos de los aminoácidos pertenecientes a determinados fragmentos polipeptídicos ***F*** y sus elementos ${}^k mb_{ijF}$ se calculan como sigue:

$${}^{,k}mb_{ijF} = {}^k mb_{ij}, \text{ si } (i \wedge j) \in F$$

$${}^{,k}mb_{ijF} = \frac{1}{2}{}^k mb_{ij}, \text{ si } (i \vee j) \in F \text{, pero no ambos} \quad \mathbf{(11)}$$

${}^{k}mb_{ijF} = 0$, en cualquier otro caso

Por lo tanto, los *k^{th} índices bilineales, cuadráticos y lineales de fragmento-local* se calculan mediante las expresiones:

$$b_F^k \ (\bar{x}, \bar{y}) = \sum_{a=1}^{n} {}_{bF}L_a = [X]^T \mathrm{MB}_F^k [Y] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{12})$$

$$q_F^k(\bar{x}) = \sum_{a=1}^{n} {}_{qF}L_a = [X]^T \mathrm{MB}_F^k [X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{13})$$

$$f_F^k(\bar{x}) = \sum_{a=1}^{n} {}_{fF}L_a = [U]^T \mathrm{MB}_F^k [X] \quad \forall\, a = 1,2,\dots,n \qquad (\mathbf{14})$$

El término fragmento (***F***) se refiere no solo a secuencias de aminoácidos, donde los aminoácidos "*i*" e "*i+1*"se encuentran unidos por un enlace peptídico, sino también a grupos de aminoácidos que se encuentran distantes en la estructura primaria. En el presente trabajo, los índices locales se calculan utilizando dos tipos de fragmentos: por *tipos de aminoácidos* y por *agrupaciones de aminoácidos*. En el primer caso, los aminoácidos se clasifican de acuerdo a la naturaleza de su cadena lateral, dando lugar a los siguientes fragmentos: apolar (RAP), polares con carga positiva (RPC), polares con carga negativa (RNC), polares no cargados (RPU), aromáticos (ARO) y alifáticos (ALG). También se definen las siguientes agrupaciones de aminoácidos: aminoácidos que no favorecen el plegamiento y/o no se encuentran con frecuencia formando hélices-α u hojas-β (UFG), aminoácidos que favorecen la formación de hélices-α (FAH), aminoácidos que favorecen la formación de hojas-β (FBS) y aminoácidos que favorecen la formación de giros-β (AFT). Finalmente, se definen agrupaciones que contienen los aminoácidos del mismo tipo (R grupo), es decir, 20 fragmentos, uno por cada aminoácido natural.

2.6.Operadores de agregación de las contribuciones atómicas

Como se explicó en el epígrafe 1.4, la noción de OA como esquema de generalización a la combinación lineal de las contribuciones atómicas para obtener los índices (globales y locales)

propuestos, se origina de la hipótesis que plantea que la definición global más apropiada de un sistema puede no ser necesariamente aditiva. Los OA empleados en el presente trabajo quedaron clasificados en tres grupos:

1) Normas: normas de Minkowski [p=1 (N1), p=2 (N2), p=3 (N3)]. Nótese que N1 se define como la suma de las componentes del vector de LOVIs.

2) Estadísticos de tendencia central: Media geométrica (GM), Media aritmética (AM), Media cuadrática (P2), Media potencial (P3) y Media armónica (HM).

3) Estadísticos de dispersión y forma: Varianza (V), Skewness (S), Kurtosis (K), Desviación estándar (SD), Coeficiente de variación (VC), Rango (RA), Percentil 25 (Q1), Percentil 50 (Q2), Percentil 75 (Q3), Q3-Q1 (i50), XMax (MX) y XMin (MN).

Finalmente, la definición matemática de estos operadores puede encontrarse en [23].

2.7.Procedimientos de cortes macromoleculares: geométricos y topológicos

Las interacciones no covalentes son responsables de la estructura final de las macromoléculas, de

la unión específica entre estas, de los procesos de autoorganización de estructuras macromoleculares y

celulares [66, 67]. Por tanto, estas juegan un papel fundamental en todos los aspectos estructurales y funcionales de los seres vivos. De la relación entre la distancia y la magnitud de las interacciones no covalentes de diversa naturaleza se comprueba que estas contribuyen en mayor o menor medida al mantenimiento de la estructura 3D de una macromolécula en dependencia de la distancia a la que se encuentren los grupos interactuantes [67]. De esta forma, algunas de estas interacciones solo son importantes cuando los grupos funcionales se encuentran muy cercanos entre sí.

Por otra parte, la relación entre la topología y el plegamiento de proteínas ha sido revelada en diversos estudios donde se ha encontrado correlación significativa entre diferentes parámetros estructurales simples, basados en rasgos topológicos del estado nativo, y la velocidad de plegamiento

[28, 69]. Entre estos parámetros se encuentran RCO y LRO (RCO, acrónimo de *Relative contact order*; LRO, acrónimo de *Long-range order*), propuestos por Plaxco y Gromiha quienes demostraron la existencia de correlación significativa entre estos parámetros y la velocidad de plegamiento de proteínas. De esta manera, podría ser útil construir matrices multi-métricas de similitud-disimilitud a partir de información sobre el contacto (interacción) entre aminoácidos que se encuentran alejados una distancia determinada (o rango de distancia) en la estructura primaria de la proteína con el objetivo de estudiar posibles relaciones entre una propiedad específica y las características topológicas del estado nativo de la proteína.

Teniendo en cuenta lo anterior y con el objetivo de discriminar entre las interacciones no covalentes de diferente naturaleza entre pares de aminoácidos se definen criterios de corte:

1) Corte geométrico basado en distancia Euclídea (lag l) denominado "length cut-off".
2) Corte topológico basado en distancia topológica (lag p) denominado "path cut-off".

Por ejemplo, el uso de corte geométrico (junto al exponente k) permite seleccionar aquellas interacciones no covalentes entre los grupos funcionales de los aminoácidos que contribuyen significativamente al mantenimiento de la estructura 3D de las proteínas. Por su parte, el parámetro *k* modela la relación funcional que existe entre la distancia y la fortaleza de la interacción entre los grupos funcionales de los aminoácidos *i y j*.

Por otra parte, el corte topológico permite seleccionar aquellas interacciones no covalentes entre aminoácidos que se hallan a una determinada distancia topológica. Debe notarse que la distancia topológica entre dos aminoácidos *i y j* está determinada por la longitud del camino más corto entre los vértices *i* y *j* del grafo cuyos vértices representan los carbonos alfa del esqueleto covalente del polipéptido y sus aristas se corresponden con los enlaces peptídicos entre los aminoácidos *i* y *j*.

Se debe aclarar que no es obligatoria la aplicación de los criterios *cut-off* en la construcción de la matriz multi-métrica, estos se aplican de acuerdo al problema en estudio.

2.8.Descripción del procedimiento de extracción de rasgos 3D-proteicos

Entrada: Archivo Protein Databank (PDB)

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

Aspectos externos

1. Generación de la representación 3D-proteica (Fichero(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 ficheros 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-atómica.
   d. Selección de los órdenes (parámetro *k*) de la matriz.
   e. Selección (opcional) de los procedimientos de cortes macromoleculares (geométrico y topológico).
   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 atómicas.

Aspectos internos

4. Cálculo de la matriz (no estocástica) de distancias inter-atómicas.
   a. Aplicación (opcional) de transformaciones probabilísticas (simple-estocástica y de probabilidad mutua).

b. Aplicación (opcional) de cortes macromoleculares.

5. Cálculo de los vectores de propiedades.
6. Cálculo de los índices de nivel atómico.
7. Aplicación del (los) operador(es) de agregación sobre el vector de índices de nivel atómico.
8. Escritura de los descriptores a fichero.

2.9.Conclusiones del capítulo

En este capítulo se definió el fundamento teórico de los nuevos descriptores 3D-proteicos. Como conclusiones se puede señalar que:

1) se propusieron representaciones 3D-proteicas diferentes a C_α.

2) se propusieron diversas métricas para el cálculo de las distancias inter-atómicas.

3) en la diagonal de la matriz se consideran valores diferentes a cero (distancia de cada átomo al centro de la proteína).

4) se utilizaron las matrices simple-estocástica y de probabilidad mutua como esquemas de normalización.

5) se aplicaron operadores de agregación distintos a la suma para generalizar la obtención de índices totales y locales a partir de un vector de índices atómicos.

6) se definieron procedimientos de cortes macromoleculares geométricos y topológicos.

7) se describió el procedimiento de extracción de rasgos 3D-proteicos.

Capítulo 3: Software para el Cálculo de los Descriptores 3D-Proteicos

En este capítulo se presenta el diseño del software ToMoCoMD-CAMPS MuLiMs-MCoMPAs. Inicialmente, se describen las herramientas utilizadas en su desarrollo. Finalmente, se presentan y discuten los resultados obtenidos en las pruebas de rendimiento realizadas al software.

3.1.Biblioteca Chemical Development Kit (CDK)

En el desarrollo del software ToMoCoMD-CAMPS se utilizó la biblioteca de código abierto implementada en Java Chemical Development Kit (CDK) [70]. De la misma se reutilizaron las clases que representan las estructuras biomacromoleculares. Además, se redefinieron acorde a las necesidades de la investigación los métodos (funcionalidades) que CDK proporciona para realizar los flujos de lectura y escritura de los archivos con formato Protein Databank (PDB).

3.2. Biblioteca Jmol

Con el propósito de visualizar las diferentes representaciones 3D-proteicas, se utilizó la biblioteca Jmol en su versión 13.0.9. Jmol es un visor Java de código abierto para estructuras químicas en tres dimensiones (http://www.jmol.org).

3.3. Diseño del software ToMoCoMD-CAMPS MuLiMs-MCoMPAs

El software MuLiMs MCoMPAs consta de dos componentes principales: el front-end y el back-end. En el front-end, se definen las interfaces gráficas de usuario para la configuración de los DMs y creación de proyectos para su posterior cómputo. Por otra parte, en el back-end están implementadas las clases encargadas de la realización del cálculo de los DMs. Este último, se desarrolló como una biblioteca [Application Programming Interface (API)] de forma tal que esta pueda ser empleada en el desarrollo de otras aplicaciones informáticas. Con estos dos componentes se logra independizar la capa externa del software de su lógica interna implementada en el back-end, y por lo tanto cualquier modificación en este último no provoca cambios en el front-end, y viceversa.

3.3.1 Front-end: Interfaz gráfica de usuario

Para facilitar el cómputo de los DMs propuestos, se diseñó una interfaz de escritorio amigable que viabiliza la configuración de sus parámetros: formas algebraicas, enfoques matriciales, métricas de distancia, órdenes matriciales, locales, las propiedades de la cadena lateral de aminoácidos, procedimientos de cortes macromoleculares y operadores de agregación.

Desde el punto de vista interno (funcionalidades), se desarrolló un módulo para generar las representaciones 3D-proteicas propuestas. Este módulo permite al usuario la selección, visualización en texto plano y en 3D, y la generación a un formato persistente (fichero en disco) de las representaciones de la(s) cadena(s) polipeptídica(s) con que desea realizar el estudio en cuestión. Además, el software permite utilizar o no la distancia al centro de la proteína, de forma que los elementos de la diagonal principal de la matriz no son siempre ceros, con el fin de lograr una mayor discriminación en las estructuras biomacromoleculares. Igualmente, este permite la generación de un reporte que incluye los vectores de propiedades utilizados y las matrices generadas en el cálculo de los 3D-DMs. Por otra parte, el usuario tiene la posibilidad de chequear, copiar y eliminar el historial de las acciones que va realizando.

Desde el punto de vista de diseño, es importante señalar que se emplearon los colores verde y blanco. Es bien conocido que el color verde es usado en diversos contextos como son: el vestuario de los cirujanos (confiere paz y esperanza), luz de semáforo (señal de avanzar) y el uniforme de nuestras Fuerzas Armadas (camuflaje). Por otra parte, el color blanco simboliza la paz y es la combinación de todos los colores básicos. Además se diseñó el *splash* del software el cual contiene en la parte superior izquierda un icono (pieza de rompecabezas o *puzzle*) en 3D con los mencionados colores. Este icono se corresponde con la esencia del presente trabajo (procedimiento de extracción de rasgos 3D-proteicos), en que este procedimiento puede verse como un puzzle, en el sentido de que se

combinan (ensamblan) diferentes partes (elementos teóricos) para lograr el resultado final lo cual sería en el primer caso el/los descriptor/es 3D-proteicos (número/s) y en el segundo caso el puzzle como tal. Este *splash* contiene al centro el nombre del software (*MuLiMs*) y en la esquina superior izquierda el nombre la suite (*ToMoCoMD-CAMPS*), en analogía con el software *Excel* de la suite *Office* de Microsoft. Finalmente, es conveniente acotar que se trabajó, según las capacidades del autor, en función de lograr la *simetría* en el diseño de las interfaces gráficas, lo cual constituye un *criterio* de belleza.

3.3.2. Back-end: Biblioteca de clases para calcular los descriptores 3D-proteicos

Las peticiones realizadas por los usuarios a través de la interfaz de escritorio son procesadas por la biblioteca MuLiMs-MCoMPAs. Este componente está estructurado en paquetes acorde a sus funcionalidades. El paquete principal es *tomocomd.camps.mulims*, el cual contiene los paquetes: descriptors, matrices, metrics, local, pdbfilter y workers, que implementan los conceptos empleados en la definición de los descriptores 3D-proteicos. El paquete *descriptors* contiene las clases relacionadas con el cómputo de los 3D-DMs. El paquete *matrices* incluye las clases responsables de construir los enfoques matriciales utilizadas para representar las interacciones no covalentes en proteínas. Este paquete también incluye el paquete *cut-off* que incluye las clases que realizan los cortes macromoleculares. El paquete *metrics* incluye las clases que representan las métricas de distancias inter-atómicas. El paquete *local* incluye las clases que representan los fragmentos locales predefinidos en el software. El paquete *pdbfilter* incluye las clases para generar las representaciones 3D-proteicas. Por último, el paquete *workers* contiene las clases necesarias para la realización y control del cálculo de los descriptores.

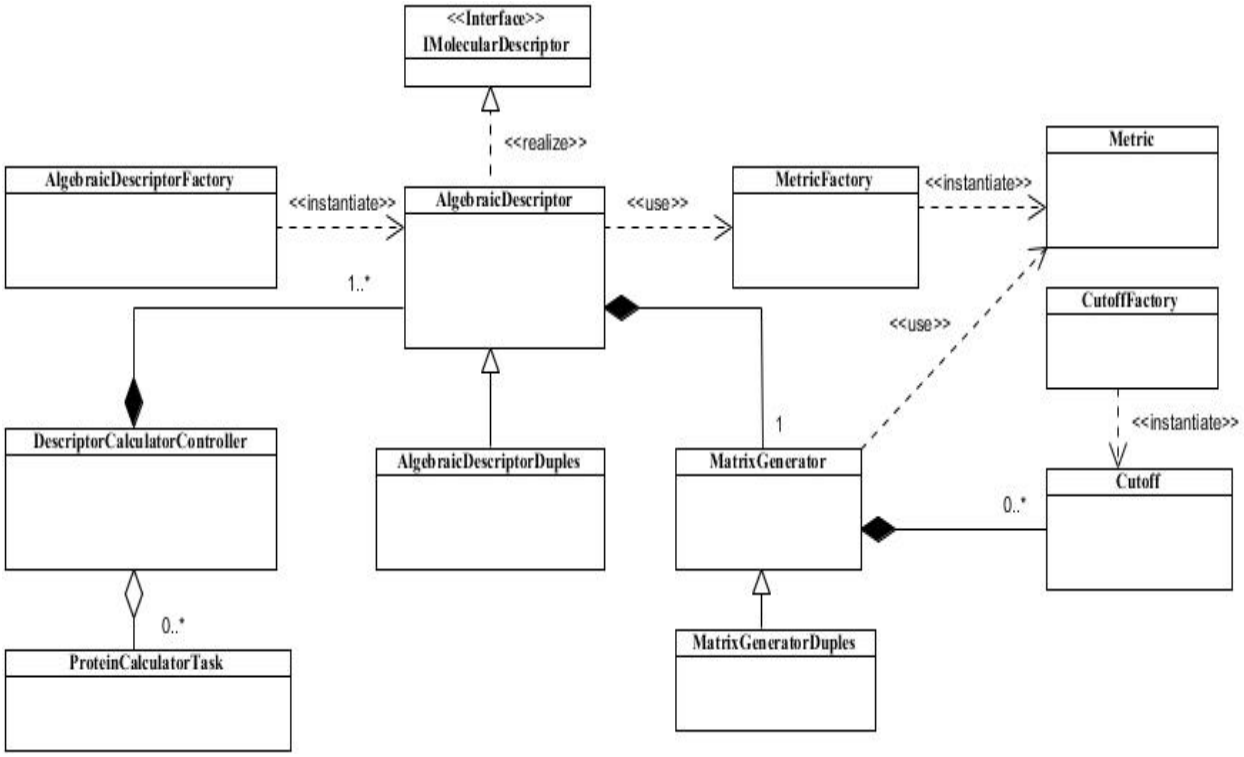

Figura 1 Diagrama UML de las clases principales para el cálculo de los descriptores 3D-proteicos

3.4.Cálculo multi-núcleo de los descriptores 3D-proteicos

Con el objetivo de reducir el tiempo requerido para calcular los descriptores 3D-proteicos fue necesario, en primera instancia, valorar una posible estrategia de paralelización en la cual debían considerarse dos aspectos fundamentales: el tiempo y el espacio (memoria). En el reporte [71] se introduce el software QuBiLS-MIDAS, el cual se desarrolló para calcular parámetros geométricos en pequeñas moléculas orgánicas. Este software utiliza un esquema de paralelización basado en el marco de trabajo (framework) Fork/Join [72]. En este framework existe un pool de hilos acorde al número de procesadores disponibles en el sistema, donde cada hilo tiene su propia cola de planificación. Una tarea recursiva (instancia de MoleculeCalculatorTask) contiene el conjunto total de descriptores a calcular y es responsable de dividir este cálculo en pequeñas subtareas (subinstancias de MoleculeCalculatorTask) en dependencia de la cantidad de hilos. Nótese que este esquema clasifica

dentro de los bien conocidos algoritmos *divide y vencerás* [73]. Como se describe en [72], una de las ventajas del framework Fork/Join es que aprovecha las arquitecturas multi-núcleo por medio de la estrategia de planificación *work-stealing*, esto significa que cuando un hilo no tiene otras tareas que atender entonces este toma tareas desde la cola de otro hilo seleccionado de forma aleatoria. En relación al factor tiempo, esta estrategia se consideró como apropiada para el cálculo de los índices MuLiMs-MCoMPAs. Sin embargo, al analizar la representación matricial implementada en QuBiLS-MIDAS quedó evidenciado que no era apropiado utilizarla para el cómputo de los descriptores 3D-proteicos, puesto que como norma, en el proceso de cálculo de índices que consideran relaciones entre pares (duplas) de átomos, en cada instante se cargan en memoria tres matrices de orden n (número de átomos de la molécula) por cada instancia de MoleculeCalculatorTask, lo cual en pequeñas moléculas orgánicas no tiene impacto significativo en el consumo de memoria RAM. Sin embargo, el dominio de aplicación del presente software es el cálculo de índices 3D en péptidos y proteínas. En este caso, el uso racional de la memoria es un factor crítico debido al alto número de átomos presentes en estos compuestos. Por lo tanto, se concibió una representación matricial de forma que en cada instante de tiempo solo exista en memoria una matriz por cada instancia de la clase ProteinCalculatorTask, lo cual representa una reducción (en dos tercios) del consumo de memoria respecto a la representación matricial usada en el software QuBiLS-MIDAS.

3.5.Evaluación del rendimiento del software ToMoCoMD-CAMPS

En este epígrafe, se presentan y discuten los resultados derivados en las pruebas de rendimiento realizadas al software ToMoCoMD-CAMPS según los parámetros: ganancia de velocidad y eficiencia en el cálculo de los 3D-DMs. La ganancia de velocidad o *Speed-up* (Sp) para *p* procesadores es el cociente entre el tiempo de ejecución de un programa secuencial y el tiempo de ejecución de la versión paralela de dicho programa en *p* procesadores. Por su parte, la eficiencia (E)

es el cociente entre el Speed-up y el número de procesadores. Esta métrica cuantifica el grado de aprovechamiento de los procesadores en la resolución del problema.

3.5.1. Análisis de las pruebas de rendimiento al software ToMoCoMD-CAMPS

Las pruebas de rendimiento se realizaron en una Laptop DELL INSPIRON N5010 con procesador Intel(R) Core(TM) i3 M370 a 2.40 GHz y 3GB de RAM con sistema operativo Windows 7 Ultimate x64. Es prudente señalar que solo se asignó 1 Gigabyte (GB) de memoria RAM a la Máquina Virtual de Java. Los experimentos se realizaron con una proteína de 500 aminoácidos y se utilizaron los Cα como representación 3D-proteica, por lo tanto, la dimensión de los vectores de propiedades y el orden de las matrices involucrados en el cálculo de los 3D-DMs es 500.

En la Figura 1 se ilustra el desempeño de la solución propuesta en términos de ganancia de velocidad, en la misma se evidencia un aumento en la ganancia de velocidad en tanto se incrementa el número de procesadores. El mejor desempeño (Sp=2.6) se alcanza con 4 procesadores, esto significa que utilizando 4 núcleos se logra reducir el tiempo de ejecución en 2 veces y fracción (de 190 a 73 segundos aproximadamente). Por su parte, en la Figura 2 se ilustra el comportamiento de la eficiencia, en la misma se observa que los valores de este parámetro oscilan entre 0.78 (usando 2 procesadores) y 0.65 (usando 4 procesadores). Estos resultados indican que usando (dos y cuatro) procesadores, se logra que estos trabajen en función de calcular los 3D-DMs, como mínimo el 65% del tiempo, lo que evidencia un aprovechamiento razonablemente bueno de su capacidad de cómputo.

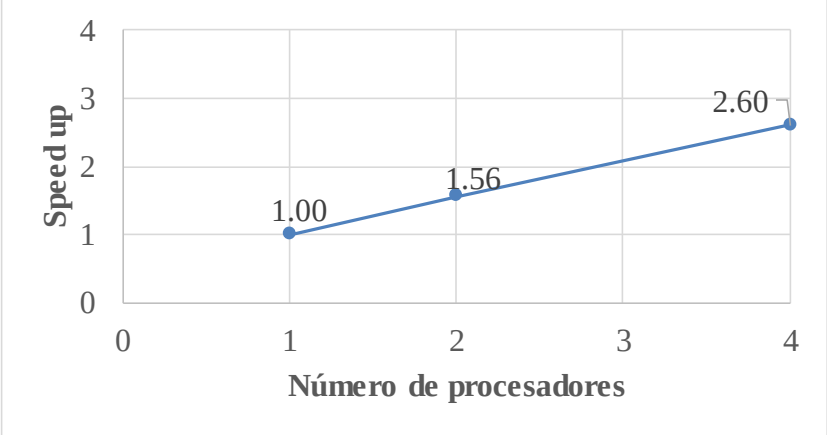

Figura 2 Speed up en función de la cantidad de procesadores

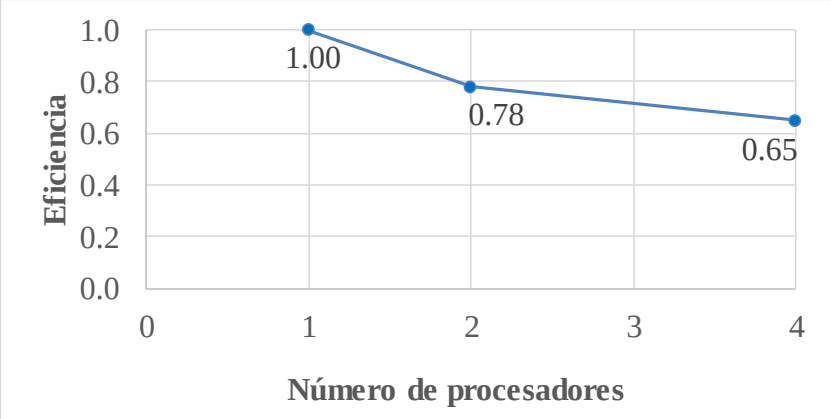

Figura 3 Eficiencia en función de la cantidad de procesadores

3.6. Conclusiones del capítulo

En este capítulo se presentó el diseño de la aplicación informática desarrollada para calcular los descriptores 3D-proteicos. Como conclusiones se puede apuntar que:

1) El software desarrollado se estructuró en dos componentes: el front-end (interfaz de usuario) y el back-end (biblioteca de clases).
2) La interfaz de usuario desarrollada posee un diseño racionalmente bueno desde el punto de vista estético.

3) La biblioteca de clases se implementó de forma que pueda ser utilizada en el desarrollo de otras aplicaciones informáticas y además de potencial empleo en *clusters* de computadoras para el procesamiento de bases de proteínas de mayor envergadura.
4) Se realizaron pruebas de rendimiento que evidencian que la aproximación paralela es capaz de reducir el tiempo necesario para obtener los 3D-DMs, respecto a la ejecución secuencial, en una proteína de dimensión cercana al promedio, obteniendo un buen desempeño en términos de ganancia de velocidad y eficiencia.
5) Se demostró que con la representación matricial concebida e implementada es posible procesar estructuras proteicas (de tamaño próximo al promedio) en estaciones de trabajo con una cantidad de memoria RAM razonable.

Capítulo 4: Evaluación de los Nuevos Descriptores 3D-Proteicos

En este capítulo se presenta la validación de los descriptores propuestos mediante estudios de Análisis de Variabilidad (AV) y Componentes Principales (ACP); así como la modelación de propiedades biológicas en proteínas. En los estudios de AV y ACP se evalúa el comportamiento de los parámetros esenciales de los DMs propuestos; dígase métricas de distancia inter-atómicas (MDIA), propiedades de la cadena lateral de aminoácidos (PCLA) y los operadores de agregación de las contribuciones atómicas. Resulta importante resaltar que el producto de los dos primeros (MDIA x PCLA) constituyen el *núcleo* de los 3D-DMs propuestos, así como las Repúblicas de Cuba y Bolivariana de Venezuela lo son para la Alternativa Bolivariana para las Américas (ALBA). Además, es importante señalar que uno de los propósitos de estos estudios es acotar en cierta medida el espacio de alta dimensión de DMs que genera el software ToMoCoMD-CAMPS, lo cual constituye el propósito de la heurística en el contexto de la Inteligencia Artificial. Finalmente, se aplican los 3D-DMs en estudios de predicción de las clases estructurales de proteínas predicción de las 4 clases estructurales de proteínas y en la modelación de la velocidad de plegamiento de cadenas polipeptídicas.

4.1. Análisis de variabilidad basado en Entropía de Shannon de los 3D-DMs

En este epígrafe, se realiza un estudio de la variabilidad de los 3D-DMs basado en Entropía de Shannon (ES), acorde al empleo de diferentes métricas para el cálculo de distancias inter-atómicas, el uso de propiedades de la cadena lateral de los aminoácidos, así como la aplicación de diferentes operadores de agregación de las contribuciones atómicas. Este estudio no supervisado se llevó a cabo con una base compuesta por 152 proteínas, la cual se empleó previamente en la literatura [74]. Para realizar este estudio se hizo un análisis de la representación gráfica utilizada en estudios precedentes [10, 11, 75]. De dicho análisis se concluye que: es frecuente la ocurrencia de *solapamiento* (sobre

todo cuando existen muchos parámetros a evaluar, como en el presente trabajo) entre las diferentes distribuciones de ES, hecho que dificulta la realización de análisis objetivos de la información presentada. Por tal motivo, el autor creó una nueva métrica denominada *Entropía Promedio de Shannon Estandarizada* (EPSE) que tiene como objetivo cuantificar el contenido de información promedio de un conjunto de descriptores moleculares. Esta métrica se calcula como sigue:

$$EPSE = \frac{\sum_{i=1}^{n} ES_i \Big/ n}{MAXES} \quad \textbf{(15)}$$

donde, *n* es el número descriptores, *ES* es la *Entropía de Shannon* [76] para el *i*-ésimo descriptor y *MAXES* es la entropía máxima y se obtiene mediante la siguiente expresión:

$$MAXES = \frac{\log_{10}(n)}{\log_{10}(2)} \quad \textbf{(16)}$$

El uso de esta métrica y un gráfico de barras permitió crear una representación de la variabilidad de los DMs, la cual es más sencilla e interpretable que la usada precedentemente [10, 11, 75]. Todo lo anterior constituye un "aporte" teórico-práctico del presente trabajo.

4.1.1. Análisis comparativo de los 3D-DMs acorde a las métricas utilizadas para el cálculo de distancias inter-atómicas

El objetivo de este estudio es evaluar la variabilidad de los 3D-DMs MuLiMs-MCoMPAs según la métrica utilizada para el cálculo de las distancias inter-atómicas. Como puede observarse en la figura 4, los mayores grados de variabilidad (superior a 0.6 de EPSE), son obtenidos por DMs que usan las métricas: Squared Euclidean (M19), Minkowski p=2.5 (M06), Minkowski p=3 (M07), Euclídea (M05), Euclidiana Promedio (M18), Minkowski p=3.5 (M09), Chebyshev (M08), Minkowski p=1.5 (M04), Manhattan (M03), SL_Like (M17) y Minkowski p=0.5 (M02). Luego aparece un grupo de DMs basados en las medidas: Lance-Williams (M11), Clark (M12), Soergel

(M13), Canberra (M10) y [Minkowski (p=0.25)] (M01). La menor variabilidad es obtenida por los DMs que emplean las métricas m15 (Wave-Edges) y m16 (Separación Angular). Este análisis sugiere que DMs con buen comportamiento de variabilidad es obtenido con las medidas: m19, m3 hasta m9, m18 y m17.

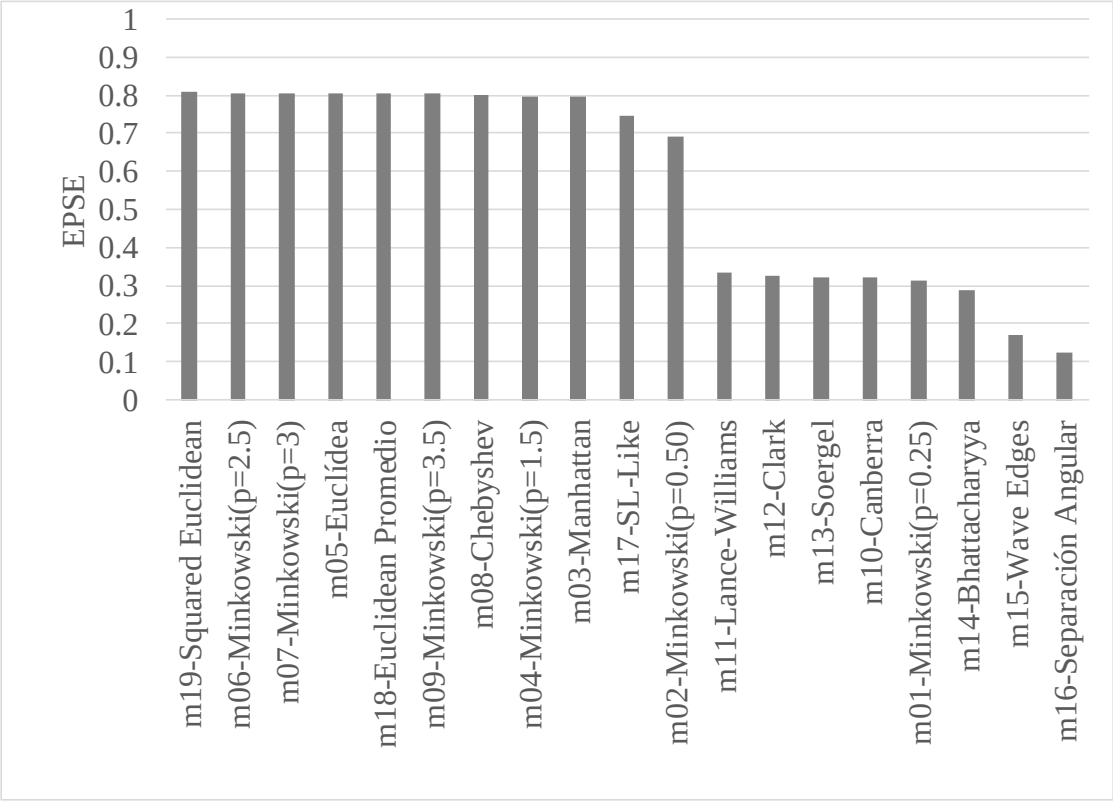

Figura 4. Entropía de Shannon estandarizada promedio de los 3D-DMs según la métrica utilizada para el cálculo de la distancia inter-atómica.

4.1.2. Análisis comparativo de los 3D-DMs conforme al uso de propiedades de la cadena lateral de aminoácidos

El objetivo de este estudio es evaluar la variabilidad de los 3D-DMs acorde a la propiedad de aminoácido utilizada. Como puede observarse en la figura 5, al variar las diferentes propiedades de aminoácidos se obtienen descriptores de variabilidad muy similar (por encima de 0.6 de EPSE). Este análisis sugiere que todas las propiedades son apropiadas para ser utilizadas en estudios QSAR.

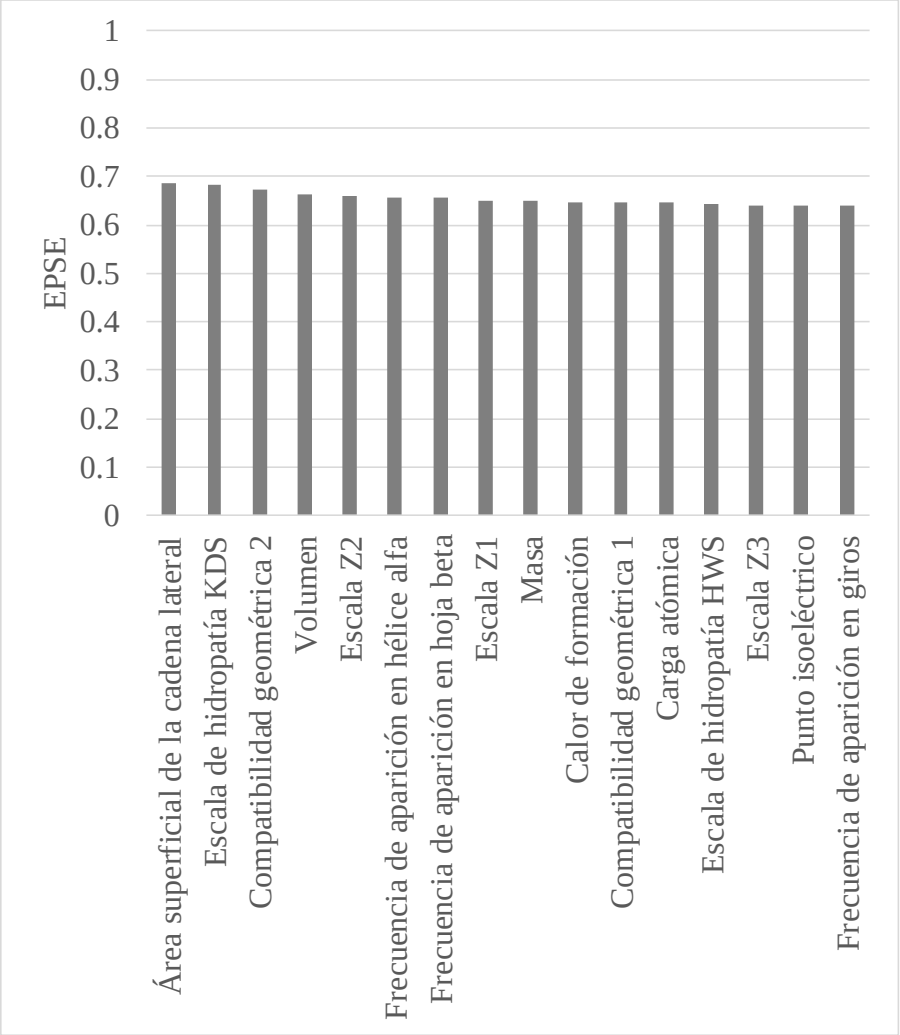

Figura 5. Entropía de Shannon estandarizada promedio de los 3D-DMs según la propiedad de aminoácido utilizada.

4.1.3. Análisis comparativo de los 3D-DMs acorde al operador de agregación aplicado

El objetivo de este estudio es analizar la variabilidad de los 3D-DMs según el operador de agregación utilizado. La mayor variabilidad (por encima de 0.6 de EPSE) es alcanzada por los DMs basados en: Skewness (S), Percentil 50 (Q2), Q3-Q1 (i50), Percentil 75 (Q3), Percentil 25 (Q1), Coeficiente de variación (VC) y Kurtosis (K). Luego aparecen los DMs basados en Media Armónica (HM) y XMin (MN). Posteriormente, se presenta un grupo de DMs de variabilidad similar, los cuales emplean los operadores: Rango (RA), Desviación Estándar (SD), XMax (MX), las normas (N1, N2 y

N3), Media Cuadrática (P2), Media Potencial (P3) y Media Aritmética (AM). Por último, los DMs de menor variabilidad son los basados en Varianza (V) y Media Geométrica (GM). Este análisis indica que DMs con buen comportamiento de variabilidad es obtenido con los operadores: S, Q2, i50, Q3, Q1, VC y K.

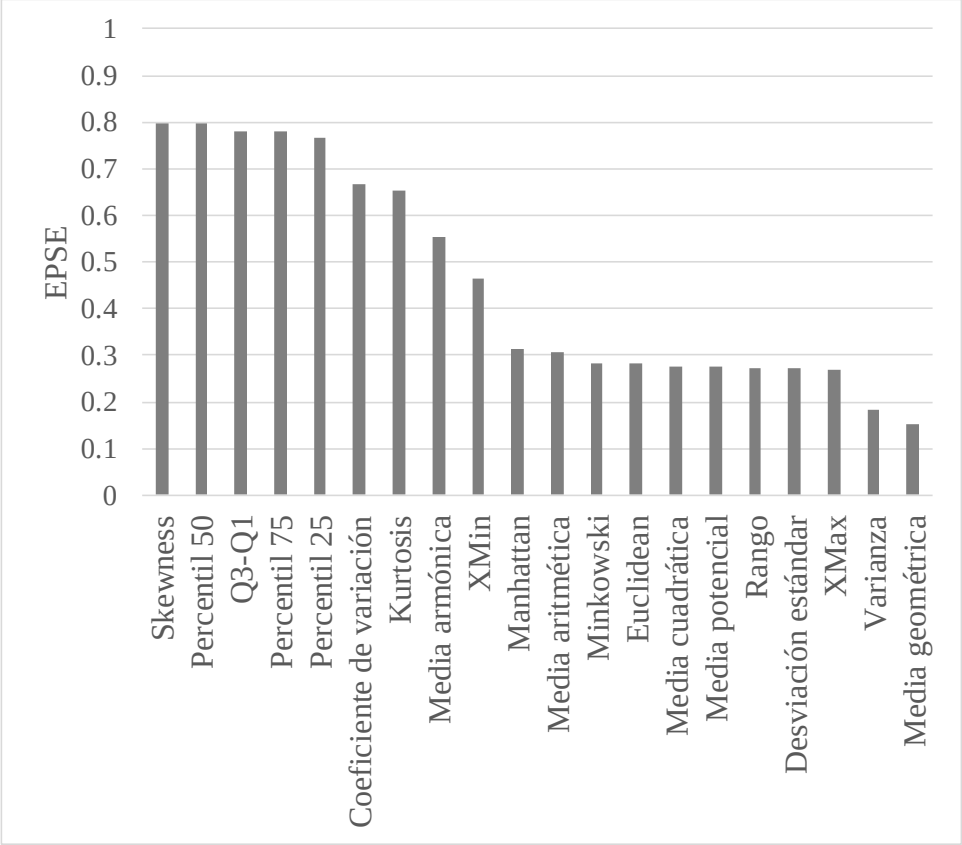

Figura 6 Entropía de Shannon estandarizada promedio de los 3D-DMs acorde al operador de agregación utilizado (basados en Minkowski, estadísticos de tendencia central, dispersión y forma).

4.2. Análisis de ortogonalidad de los 3D-DMs

En el presente acápite se realiza un estudio para evaluar la posible ortogonalidad de los 3D-DMs mediante el método ACP descrito en el epígrafe 1.6.1. Para realizar el estudio se empleó el software STATISTICA 8.0 (http://www.statsoft.com) y la misma base empleada en los AV.

4.2.1. Independencia lineal de los 3D-DMs acorde a la métrica utilizada para calcular la distancia inter-atómica

Para realizar el estudio se calcularon 988 DMs usando las diferentes métricas de distancia. En este estudio se determinaron 7 componentes las cuales explican el 94.97% de la varianza total. Al examinar esas componentes es posible concluir que: 1) los DMs que usan las métricas Bhattacharyya (M14) y Separación Angular (M16) codifican información ortogonal, 2) los DMs que usan las métricas M14 y Wave-Edges (M15), [Minkowski (p=0.25)] (M01) y [Minkowski p=0.5 (M02)]; están correlacionados en los factores 3 y 4, respectivamente.

4.2.2. Independencia lineal de los 3D-DMs acorde a la propiedad de aminoácido utilizada

Para realizar el estudio fueron calculados 624 DMs, obteniéndose 10 componentes que explican aproximadamente el 92.68% de la varianza acumulada. El análisis de esas componentes revela que: 1) los DMs que utilizan la masa (MM) y el volumen (MV) presentan saturaciones en el factor 4, 2) los DMs que emplean GCP2 y KDS aparecen correlacionados en el factor 5, 3) los DMs que utilizan frecuencia relativa de aparición en hélice alfa (PAH) están exclusivamente cargados en el factor 6, 3) los descriptores basados en escala Z3 (Z3) y punto isoeléctrico (PIE) están correlacionados en el factor 9, 4) Los DMs que emplean EPS están exclusivamente cargados en el factor 10.

4.2.3. Independencia lineal de los 3D-DMs acorde al operador de agregación utilizado

Para realizar el estudio fueron calculadas 1040 variables, obteniéndose 10 componentes las cuales explican aproximadamente el 87.85% de la varianza acumulada. Al analizar estas componentes se puede señalar que: 1) los DMs que usan los operadores Q3-Q1 (i50) y Percentil 75 (Q3) presentan valores correlacionados en el factor 3, 2) los DMs que emplean el operador Kurtosis (K) presentan cargas exclusivas en el factor 5, 3) los DMs que usan los operadores Percentil 25 (Q1) y XMin (MN) están cargados exclusivamente en los factores 9 y 10, respectivamente.

4.3. Aplicación de los 3D-DMs a la predicción de las clases estructurales de proteínas

En este epígrafe se evalúa el desempeño de los 3D-DMs en la clasificación estructural de proteínas. En primer lugar, se describe el conjunto de datos utilizado. Seguidamente, se describen las medidas de evaluación de desempeño de los modelos de clasificación. Luego, se definen los procedimientos empleados para desarrollar y evaluar los modelos de clasificación. Por último, se presentan y discuten los resultados obtenidos.

4.3.1. Descripción del conjunto de datos

Para el desarrollo de los modelos de clasificación se utilizó el conjunto de datos propuesto en [77] el cual está compuesto por 204 proteínas de las cuales 52 son *All-α*, 61 *All-β*, 45 *α/β* y 46 *α+β*. Con el propósito de obtener los modelos de clasificación y evaluar su capacidad predictiva se utilizó una división realizada previamente; para detalles sobre el procedimiento utilizado para realizar dicha división y la composición cuantitativa y cualitativa de los conjuntos resultantes remitirse al material suplementario de [78]. De esta forma, el conjunto original quedó dividido en series de entrenamiento (SE) y predicción (SP).

4.3.2. Medidas de evaluación del desempeño de los modelos de clasificación

Exactitud o precisión (Q): es el porciento de casos (proteínas) clasificados correctamente[79].

Sensibilidad (SEN): es la probabilidad de clasificar correctamente un caso positivo [79].

Especificidad (ESP): es la probabilidad de que una predicción positiva sea correcta [79].

4.3.3. Desarrollo de los modelos de clasificación

Los modelos de clasificación se obtuvieron por medio de las técnicas: Random Forest, K-NN y MLP, las cuales están implementadas en el software WEKA 3.7.10 (http://www.cs.waikato.ac.nz/ml/weka) utilizando descriptores 3D-proteicos calculados con el

software ToMoCoMD-CAMPS. Este software genera un alto número de DMs por la posibilidad de combinar sus diferentes parámetros (formas algebraicas, enfoques matriciales, propiedades, operadores de agregación, etc.). Como se mencionó anteriormente, para "acotar" en alguna medida este espacio de alta dimensión se generó un lote de proyectos basado en los estudios de AV y ACP. No obstante, el cálculo de los proyectos anteriores genera un número considerable de DMs, por esta razón, se desarrolló un procedimiento de selección de rasgos, el cual se describe a continuación:

1. Seleccionar de cada fichero generado 1000 DMs de mayor variabilidad por ES con el software IMMAN 1.0 [80].
2. Adicionar a cada fichero generado en el paso 1, la variable respuesta (clase estructural).
3. Aplicar a cada fichero obtenido en el paso 2, la técnica de selección de atributos *Correlation Feature Selection* disponible en el software WEKA 3.7.10.
4. Mezclar todos los ficheros obtenidos en el paso 3.

Resulta prudente señalar que el principio de diseño que gobierna el procedimiento anterior está basado en experiencias precedentes aplicadas con éxito en diversos dominios de aplicación [81-83]. Ejemplos hay varios: la ciencia [equipos multidisciplinarios para abordar determinado(s) problema(s)], la construcción (mezcla de agua, arena, cemento, piedra y otros elementos), las metaheurísticas (donde se combinan heurísticas a un nivel superior), la nutrición (mezcla de alimentos con vitaminas), entre otros. Asimismo, debe señalarse el caso particular de la legendaria orquesta "Van Van" de Juan Formell, merecedora de numerosos galardones entre los que se destaca un premio Grammy en la categoría "salsa". Nótese que uno de los secretos del éxito de "Van Van" es la *fusión* del *son* tradicional con elementos del *rock* y *jazz* dando lugar al *songo* (http://www.ecured.cu/Van_Van). Otro ejemplo que ilustra que la *combinación* de las fuerzas, medios, inteligencia y tesón en el arte de la guerra conduce a victoria se puede encontrar en la Historia de Cuba, específicamente, en la Toma de Placetas, donde el *Comandante* Ernesto Guevara

de la Serna, colocó una emboscada en un punto intermedio en la carretera que conduce a Placetas, y cerró la vía Báez-Santa Clara con el *apoyo* de las fuerzas del Directorio Revolucionario del 13 de marzo. Como se refleja en (http://www.ecured.cu/Batalla_de_Santa_Clara), tal estrategia permitió tomar el poblado en pocos días de combate.

También, en estudios quimio-informáticos realizados en [83] queda demostrado que la "mezcla" de técnicas de selección (no supervisada y supervisada) de atributos proporciona mejores resultados que empleándolas por separado.

4.3.4. Evaluación del desempeño de los modelos de clasificación

En esta sección se evalúa el desempeño de los modelos de clasificación desarrollados con los descriptores 3D-proteicos. En la tabla 5 se muestran los parámetros estadísticos asociados a los modelos obtenidos. Como puede notarse, los tres modelos reproducen perfectamente (100% de exactitud global) los datos de la SE. Con el objetivo de evaluar la robustez de los modelos, se aplicó la estrategia de validación interna 10-fold cross validation sobre la SE (149 proteínas) obteniendo una exactitud global de 92.61% (138/149), 95.30% (142/149) y 94.63% (141/149) en los modelos basados en Random Forest, K-NN y MLP, respectivamente. Como en otros métodos de validación interna, altos valores de exactitud global es una demostración de la robustez de un modelo [84]. Por otra parte, respecto a la capacidad predictiva de los modelos, se observa que los desarrollados con Random Forest y K-NN clasifican correctamente el 92.73% (51/55) de los casos. Por su parte, los desarrollados con MLP exhiben una precisión global de 83.64% (49/55). Una evaluación más profunda se llevó a cabo mediante el cálculo de otras medidas como SEN y ESP. Como se refleja en la Tabla 1, se obtienen altos valores para ambos parámetros.

Tabla 1.1: Parámetros estadísticos del modelo obtenido usando Random Forest

Conjunto	**Exactitud global Q(%)**	**Sensibilidad (%)**		**Especificidad (%)**	
SE	100	**clases**		**clases**	
		α+β	100	α+β	100
		α/β	100	α/β	100
		All-β	100	All-β	100
		All-α	100	All-α	100
SP	92.73	**clases**		**clases**	
		α+β	75	α+β	97
		α/β	100	α/β	100
		All-β	93	All-β	93
		All-α	100	All-α	100

Tabla 1.2: Parámetros estadísticos del modelo obtenido usando K-NN

Conjunto	**Exactitud global Q(%)**	**Sensibilidad (%)**		**Especificidad (%)**	
SE	100	**clases**		**clases**	
		α+β	100	α+β	100
		α/β	100	α/β	100
		All-β	100	All-β	100
		All-α	100	All-α	100
SP	92.73	**clases**		**clases**	
		α+β	66.7	α+β	100
		α/β	100	α/β	100
		All-β	100	All-β	93.20
		All-α	100	All-α	97.66

Tabla 1.3: Parámetros estadísticos del modelo obtenido usando MLP

Conjunto	**Exactitud global Q(%)**	**Sensibilidad (%)**		**Especificidad (%)**	
SE	100	**clases**		**clases**	
		α+β	100	α+β	100
		α/β	100	α/β	100
		All-β	100	All-β	100
		All-α	100	All-α	100
SP	83.64	**clases**		**clases**	
		α+β	33.3	α+β	100
		α/β	100	α/β	97.47
		All-β	100	All-β	85.40
		All-α	92.3	All-α	97.47

4.4. Aplicación de los 3D-DMs a la predicción de la velocidad de plegamiento de cadenas polipeptídicas

En el siguiente estudio es evaluada la correlación de los 3D-DMs con el logaritmo (ln_kf) de la velocidad de plegamiento de proteínas. Inicialmente, se abordan los procedimientos y las medidas de evaluación de los modelos de regresión. Luego, se describen los conjuntos de datos así como los procedimientos utilizados para obtener y evaluar los modelos de regresión. Finalmente, se presentan y discuten los resultados obtenidos.

4.4.1. Descripción de los conjuntos de entrenamiento y prueba

Para el desarrollo y evaluación de los modelos de regresión, se emplearon conjuntos de 80 [85] y 16 proteínas [15], como SE y SP, respectivamente. Resulta importante señalar que de la SE fue excluida la proteína con identificador "2BLM" por contener únicamente las coordenadas espaciales de sus átomos C_α.

4.4.2. Medidas de evaluación de los modelos de regresión

En la evaluación de un modelo QSAR deben tomarse en consideración dos aspectos esenciales: el desempeño interno (bondad de ajuste y robustez) y externo (capacidad de predicción) como se sugiere en (http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf). Las medidas que a continuación se describen quedaron clasificadas en tres grupos: (ajuste, robustez y predicción externa), en correspondencia con lo expresado anteriormente.

Medida de ajuste: Coeficiente de correlación cuadrado o coeficiente de determinación (R^2): este parámetro estima la proporción de la variable respuesta explicada por la regresión. Es decir, si no existe relación lineal entre la variable dependiente (endpoint) y las variables independientes, entonces $R^2=0$; por otra parte, si existe un ajuste perfecto, entonces $R^2=1$.

Medidas de robustez:

Una de las técnicas empleadas para evaluar la robustez es la denominada validación cruzada (VC) por re-muestreo (bootstrapping) [86], en la cual se determinan aleatoriamente conjuntos de entrenamiento y de prueba. La premisa básica del re-muestreo es que los datos deben ser representativos de la población de que fueron extraídos. En este procedimiento el conjunto de entrenamiento preserva el tamaño (n) original del conjunto de datos y se conforma mediante la selección de n objetos con repetición, por su parte el conjunto de prueba está constituido por objetos no seleccionados para formar parte del entrenamiento. Este procedimiento se repite varias veces y en cada iteración se calcula el PRESS (ver **Ec.** 17), para finalmente determinar el poder predictivo promedio (Q^2_{boot}). Otra técnica ampliamente utilizada es la prueba de *permutación de las respuestas* o *Y-scrambling* para identificar modelos basados en correlación aleatoria, es decir, modelos cuyas variables independientes están correlacionadas aleatoriamente con la(s) variable(s) respuesta(s). La prueba se realiza determinado la calidad del modelo (comúnmente R^2 o Q^2) modificando aleatoriamente la secuencia del vector respuesta, es decir, asignando a cada compuesto una respuesta seleccionada aleatoriamente del verdadero conjunto de respuestas. Si el modelo original no presenta correlación aleatoria, entonces existe una diferencia notable entre la "calidad" del modelo original en comparación con el modelo obtenido con las variables respuestas permutadas.

Medidas de predicción externa: La validación cruzada dejando uno fuera *leave-one-out* (LOO) es el procedimiento de VC más sencillo. En este caso, dados *n* compuestos (casos), se calculan *n* modelos *reducidos*, es decir, obtenidos con *n-1* casos. Estos modelos se usan para predecir la respuesta del compuesto excluido. El poder predictivo se calcula como la suma al cuadrado de las diferencias entre la variable experimental y la respuesta estimada. Matemáticamente se define mediante la siguiente fórmula:

$$Q^2_{loo} = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^{nce} \left(Y_i - Y_i''\right)^2}{\sum_{i=1}^{nce} \left(Y_i - \tilde{Y}\right)^2} \quad \textbf{(17)}$$

donde, *PRESS* es la suma de los cuadrados de los residuos de la predicción, *TSS* es la suma total de cuadrados, *nce* es el número de elementos del conjunto de entrenamiento, Y_i, Y'' es la respuesta observada y estimada del i-ésimo elemento respectivamente $\tilde{Y}$ es el promedio de las respuestas del conjunto de entrenamiento. El parámetro Q^2_{loo} varía en el intervalo [0-1]. Los modelos con valores de Q^2_{loo} próximos a 0 presentan un bajo poder predictivo, mientras que los modelos con valores cercanos a 1 tienen una alta capacidad de predicción.

La validación externa permite evaluar si los modelos obtenidos son generalizables a nuevos compuestos químicos y de esta forma evaluar el verdadero poder predictivo de los mismos. Para esto se divide la base en dos conjuntos: la serie de entrenamiento para construir el modelo y la serie de predicción para evaluar el modelo obtenido. El estadístico utilizado con este fin se denomina Q^2_{ext} y se calcula mediante la expresión:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{nce} \left(Y_i - Y_i''\right)^2}{\sum_{i=1}^{nce} \left(Y_i - \tilde{Y}\right)} \quad \textbf{(18)}$$

4.4.4. Desarrollo de los modelos de regresión

Los modelos Regresión Lineal Múltiple (RLM) se obtuvieron con el software MobyDigs 1.0 [87] empleando descriptores 3D-proteicos calculados con el software ToMoCoMD-CAMPS MuLiMs-MCoMPAs. Para realizar el estudio, se empleó el conjunto de proyectos utilizado en la clasificación estructural de proteínas, por lo que fue necesario planear y aplicar un procedimiento para reducir la alta dimensión de DMs. A continuación se describe el procedimiento empleado:

1. Eliminar DMs con correlación (en valor absoluto) de Pearson respecto a la variable respuesta (ln_kf) inferior a 0.5.
2. Aplicar los siguientes filtros a los ficheros generados en el paso 1:
   a) Eliminar DMs con **correlación** mayor o igual a 0.95.
   b) Eliminar DMs con **kurtosis** mayor que 8.
   c) Eliminar DMs con **entropía estandarizada** menor que 0.3.
3. Mezclar los ficheros obtenidos en el paso 2.

4.4.3. Evaluación del desempeño de los modelos de regresión

En el Anexo 1, se encuentran las ecuaciones y los parámetros estadísticos de los modelos de RLM obtenidos para la predecir de la velocidad de plegamiento de proteínas. En general, se observa que los valores obtenidos por la técnica "bootstrapping" (Q^2_{boot}) superan el 65% de la varianza total, por lo tanto, los modelos desarrollados pueden considerarse como robustos o con buena capacidad de predicción interna. Asimismo los coeficientes determinados en el procedimiento "Y-scrambling" [$a(Q^2)$] presentan en todos los casos valores inferiores a 0.196 indicando que la correlación entre las variables independientes y la variable modelada presenta un ínfimo grado de aleatoriedad. Por otra parte, como el número de proteínas del conjunto de entrenamiento es pequeño (inferior a 100), el parámetro Q^2_{loo} puede ser considerado como un indicador de la capacidad predictiva externa de los modelos [9, 88]. En este sentido, se observa que los modelos obtenidos explican como mínimo el 65% de la varianza total. Adicionalmente, si se examina el desempeño en el conjunto externo se puede decir que los modelos obtenidos poseen un buen poder predictivo, puesto que los valores del estadístico Q^2_{ext} oscilan entre 0.501 y 0.657, lo que indica que los modelos obtenidos explican aproximadamente entre el 50% y 66% de la varianza total.

4.5. Conclusiones del capítulo.

En este capítulo se evaluó el poder predictivo de los descriptores 3D-proteicos aplicados de forma combinada con las técnicas de aprendizaje automático: Random Forest, K-NN y MLP, y estadística RLM en estudios de clasificación estructural y predicción de velocidad de plegamiento; respectivamente. Como conclusiones se puede señalar que:

1) Se logró una alta precisión en los estudios de discriminación estructural. Además los parámetros de evaluación complementarios exhiben valores favorables.
2) Se obtuvieron modelos de RLM con buen ajuste, robustez y capacidad de generalización.
3) Se desarrollaron y aplicaron satisfactoriamente dos procedimientos de selección atributos para estudios de clasificación y regresión.
4) Se demostró que los descriptores 3D-proteicos extraen información relevante de la estructura macromolecular.

Conclusiones

Con la realización del presente trabajo se arriban a los siguientes postulados:

- Se propuso un nuevo procedimiento de extracción de rasgos 3D-proteicos basado en las formas algebraicas 2-lineales.
- Se desarrolló un software nombrado ToMoCoMD-CAMPS MuLiMs que automatiza el cálculo de los descriptores 3D-proteicos, el cual permite una interacción amigable con el usuario y aprovecha las prestaciones de las arquitecturas multi-núcleo presentes en los equipos de cómputo actuales.
- Se realizaron estudios para evaluar la variabilidad y posible ortogonalidad de los 3D-DMs a fin de evaluar su contenido de información, independencia lineal y acotar en alguna medida el espacio de alta dimensionalidad de DMs.
- Se propuso una nueva métrica denominada *Entropía estandarizada promedio* para cuantificar el contenido de información promedio de un conjunto de DMs.
- Se creó una nueva representación gráfica más sencilla e interpretable que las reportadas hasta la fecha, de utilidad en futuros análisis de variabilidad de los descriptores moleculares.
- Se diseñaron y aplicaron de forma satisfactoria procedimientos de selección de rasgos para problemas de clasificación (discriminación estructural de proteínas) y regresión (predicción de la velocidad de plegamiento).
- Se evaluó el mérito de los nuevos descriptores 3D-proteicos en estudios de clasificación y regresión obteniendo modelos con muy buen ajuste, alta robustez y buena capacidad de predicción.

Recomendaciones

Al cierre de esta investigación quedaron ciertos aspectos que ameritan ser tratados posteriormente, en tal sentido se propone:

- Evaluar el desempeño de las restantes representaciones 3D-proteicas propuestas, así como la combinación de los índices totales, locales y cortes macromoleculares sobre índices totales y locales.
- Aplicar los 3D-DMs propuestos en la modelación de otras propiedades y/o funciones biológicas de interés.
- Aplicar los procedimientos de selección de atributos desarrollados en otros dominios de aplicación.
- Extender los 3D-DMs propuestos para considerar interacciones entre más de dos átomos.
- Valorar la posible utilización de la implementación realizada para extraer información en otros biopolímeros (ej. ácidos nucleicos) de interés.

Referencias Bibliográficas

1. Chou, K.-C., Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol., 2011. 273(1): p. 236-247.

2. Lehninger, A., D.L. Nelson, and M.M. Cox, Lehninger's Principles of Biochemistry. 2005, New York: WH Freeman and Company.

3. Óscar Miguel Rivera Borroto, Y.H.D., José Manuel García de la Vega, Ricardo Grau, Yovani Marrero Ponce, Maikel Cruz Monteagudo, Perspectiva general sobre el proceso de desarrollo de fármacos y las técnicas de cribado virtual basadas en la similitud molecular. Anales de la Real Academia de Farmacia, 2013.

4. González-Díaz, H., E. Uriarte, and R. Ramos de Armas, Predicting stability of Arc repressor mutants with protein stochastic moments. Bioorganic & medicinal chemistry, 2005. 13(2): p. 323-331.

5. Marrero-Ponce, Y., 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, 2004. 9(12): p. 1124-1147.

6. Díaz, H.G., R.R. 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, 2003. 19(16): p. 2079-2087.

7. González-Díaz, H., et al., Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. Journal of computational chemistry, 2007. 28(12): p. 1990-1995.

8. Todeschini, R. and V. Consonni, Handbook of Molecular Descriptors. 1st ed. Methods and Principles in Medicinal Chemistry, ed. R. Mannhold, H. Kubinyi, and H. Timmerman. Vol. 11. 2000, D-69469 Weinheim, Federal Republic of Germany: WILEY-VCH Verlag GmbH. 667.

9. Todeschini, R. and V. Consonni, Molecular Descriptors for Chemoinformatics. 1st ed, ed. R. Mannhold, H. Kubinyi, and G. Folkers. Vol. 1. 2009, Weinheim: WILEY-VCH. 667.

10. García-Jacas, C.R., N-linear algebraic maps to codify chemical structures: is a suitable generalization to the atom-pairs approaches? Curr Drug Metab, 2014. 15.

11. Marrero-Ponce, Y., Optimum search strategies or novel 3D molecular descriptors: is there a stalemate? Curr Bioinf, 2015. 10.

12. Barigye, S.J., et al., Trends in Information Theory Based Chemical Structure Codification. Chem. Rev., 2013.

13. Rao, H., et al., Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res., 2011. 39(suppl 2): p. W385-W390.

14. Zhou, H. and Y. Zhou, Folding rate prediction using total contact distance. Biophys. J., 2002. 82(1): p. 458-463.

15. Ruiz-Blanco, Y.B., et al., A Hooke′ s law-based approach to protein folding rate. Journal of theoretical biology, 2015. 364: p. 407-417.

16. Estrada, E., Characterization of the folding degree of proteins. Bioinformatics, 2002. 18(5): p. 697-704.

17. González, D., R.R. De Armas, and E. Uriarte, In silico Markovian bioinformatics for predicting 1Ha-NMR chemical shifts in mouse epidermis growth factor (mEGF). Online J. Bioinformatics, 2002. 1: p. 83-95.

18. González-Díaz, H., et al., Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr. Top. Med. Chem., 2007. 7(10): p. 1015-1029.

19. Randic, M., et al., Graphical Representation of Proteins†. Chem. Rev., 2010. 111(2): p. 790-862.

20. Randić, M., et al., Graphical representation of proteins as four-color maps and their numerical characterization. J. Mol. Graphics Modell., 2009. 27(5): p. 637-641.

21. Marrero-Ponce, Y., et al., 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., 2005. 13(8): p. 3003-3015.

22. Ortega-Broche, S.E., et al., 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., 2010. 277(15): p. 3118-3146.

23. Marrero-Ponce, Y., et al., Optimum Search Strategies or Novel 3D Molecular Descriptors: is there a Stalemate? Curr. Bioinf., 2015. 10(3).

24. García-Jacas, C.R., et al., Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets. Journal of Cheminformatics, 2016. 8(1): p. 1-16.

25. Marrero-Ponce, Y., et al., Novel 3D bio-macromolecular bilinear descriptors for protein science: Predicting protein structural classes. J. Theor. Biol., 2015. 374: p. 125-137.

26. Flory, P.J., Principles of Polymer Chemistry. 1953.

27. Zehfus, M.H. and G.D. Rose, Compact units in proteins. Biochemistry, 1986. 25(19): p. 5759-5765.

28. Plaxco, K.W., K.T. Simons, and D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol., 1998. 277(4): p. 985-994.

29. Gromiha, M.M. 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. Journal of molecular biology, 2001. 310(1): p. 27-32.

30. Weikl, T.R. and K.A. Dill, Folding rates and low-entropy-loss routes of two-state proteins. Journal of molecular biology, 2003. 329(3): p. 585-598.

31. Nölting, B., et al., Structural determinants of the rate of protein folding. Journal of theoretical biology, 2003. 223(3): p. 299-307.

32. Makarov, D.E., et al., How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proceedings of the National Academy of Sciences, 2002. 99(6): p. 3535-3539.

33. Micheletti, C., Prediction of folding rates and transition-state placement from native-state geometry. Proteins: Structure, Function, and Bioinformatics, 2003. 51(1): p. 74-84.

34. Ivankov, D.N. and A.V. Finkelstein, Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proceedings of the National Academy of Sciences of the United States of America, 2004. 101(24): p. 8942-8944.

35. Ivankov, D.N., et al., Contact order revisited: influence of protein size on the folding rate. Protein Science, 2003. 12(9): p. 2057-2062.

36. Di Paola, L., et al., Protein Contact Networks: An Emerging Paradigm in Chemistry. Chem. Rev., 2012. 113(3): p. 1598-1613.

37. González-Díaz, H., et al., Proteomics, networks and connectivity indices. Proteomics, 2008. 8(4): p. 750-778.

38. Ramos de Armas, R., et al., Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. Proteins: Struct., Funct., Bioinf., 2004. 56: p. 715–723.

39. González-Díaz, H. and E. Uriarte, Proteins QSAR with Markov average electrostatic potentials. Bioorg. Med. Chem. Lett., 2005. 15(22): p. 5088-5094.

40. Zhang, S., A. Golbraikh, and A. Tropsha, The Development of Quantitative Structure-Binding Affinity Relationship (QSBR) Models Based on Novel Geometrical Chemical Descriptors of the Protein-Ligand Interfaces. Journal of Medicinal Chemistry, 2006. 49(9): p. 2713-2724.

41. Estrada, E., Application of a novel graph-theoretic folding degree index to the study of steroid–DB3 antibody binding affinity. Computational Biology and Chemistry, 2003. 27(3): p. 305-313.

42. Dobson, P.D. and A.J. Doig, Predicting enzyme class from protein structure without alignments. Journal of molecular biology, 2005. 345(1): p. 187-199.

43. Giuliani, A., L. Di Paola, and R. Setola, Proteins as networks: a mesoscopic approach using haemoglobin molecule as case study. Current Proteomics, 2009. 6(4): p. 235-245.

44. Rosen, K.H., Matemática Discreta y sus aplicaciones. 2010.

45. Marcelo, M.V.V., Álgebra Lineal. 2008.

46. Deza, M.-M. and E. Deza, Dictionary of distances. 2006: Elsevier.

47. Berliner, H., The system: a world champion's approach to chess-Hans Berliner. 1999.

48. Todeschini, R. and V. Consonni, Molecular descriptors for chemoinformatics, in Methods and principles in medicinal chemistry, R. Mannhold, H. Kubinyi, and G. Folkers, Editors. 2009, Wiley-VCH: Weinheim.

49. Godden, J.W., F.L. Stahura, and J. Bajorath, Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations. J. Chem. Inf. Comput. Sci., 2000. 40: p. 796-800.

50. Godden, J.W. and J. Bajorath, Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. J. Chem. Inf. Comput. Sci., 2002. 42: p. 87-93.

51. Shannon, C.E., A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 2001. 5(1): p. 3-55.

52. Massey, W.F., Principal components regression in exploratory statistical research. J. Amer. Stat. Assoc, 1965. 60: p. 234–256.

53. Mardia, K.V., J.T. Kent, and J.M. Bibby, Multivariate Analysis. 1979, London: Academic Press.

54. Breiman, L., Random Forests. Machine Learning. 45(1): p. 5-32.

55. Witten, I.H. and E. Frank, Data Mining: Practical machine learning tools and techniques. 2005: Morgan Kaufmann.

56. Haykin, S. and S. Simon, Neural Networks: A Comprehensive Foundation, Prentice-Hall, 1998.

57. Hair, J., et al., Análisis multivariante. 5ª edición. editorial Prentice Hall. 1999, Madrid.

58. Mishra, A., et al., D2N: Distance to the native. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2014. 1844(10): p. 1798-1807.

59. Mathews, C.K., K.E. van Holde, and K.G. Ahern, Biochemistry. 2000, San Francisco: BenjaminCummings.

60. Zamyatnin, A., Protein volume in solution. Prog. Biophys. Mol. Biol., 1972. 24: p. 107-123.

61. Hellberg, S., et al., Peptide quantitative structure-activity relationships, a multivariate approach. J. Med. Chem., 1987. 30(7): p. 1126-1135.

62. Collantes, E.R. and W.J. Dunn III, Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogs. J. Med. Chem., 1995. 38(14): p. 2705-2713.

63. Hopp, T.P. and K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U. S. A., 1981. 78(6): p. 3824-3828.

64. Kyte, J. and R.F. Doolittle, A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 1982. 157(1): p. 105-132.

65. Sak, K., M. Karelson, and J. Järv, Modeling of the amino acid side chain effects on peptide conformation. Bioorg. Chem., 1999. 27(6): p. 434-442.

66. Nelson, D.L., A.L. Lehninger, and M.M. Cox, Lehninger principles of biochemistry. 2008: Macmillan.

67. Kar, A. Medicinal chemistry. 2007.

68. Carbo-Dorca, R., Stochastic Transformation of Quantum Similarity Matrixes and Their Use in Quantum QSAR (QQSAR) Models. Int. J. Quantum Chem. , 2000. 79: p. 163-177.

69. Gromiha, M.M., Importance of native-state topology for determining the folding rate of two-state proteins. J. Chem. Inf. Comput. Sci., 2003. 43(5): p. 1481-1485.

70. Steinbeck, C., et al., The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics J. Chem. Inf. Comput. Sci., 2003. 43 (2): p. 493–500.

71. García-Jacas, C.R., et al., QuBiLS-MIDAS: A Parallel Free-Software for Molecular Descriptors Computation based on Multi-Linear Algebraic Maps. J. Comput. Chem., 2014.

72. Lea, D. A Java fork/join framework. in Proceedings of the ACM 2000 conference on Java Grande. 2000: ACM.

73. Aho Alfred, V., et al., Data structures and algorithms. 1983, USA: Addison-Wesley.

74. Estrada, E., A Protein Folding Degree Measure and Its Dependence on Crystal Packing, Protein Size, Secondary Structure, and Domain Structural Class. J.Chem.Inf.Comput Sci, 2004. 44: p. 1238-1250.

75. Cubillán, N., Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: theory, diversity–variability analysis and QSPR applications. J Math Chem, 2015. 53.

76. Godden, J.W., F.L. Stahura, and J. Bajorath, Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci, 2000. 40.

77. Chou, K.-C., A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun., 1999. 264(1): p. 216-224.

78. Marrero-Ponce, Y., et al., Novel 3D bio-macromolecular bilinear descriptors for protein science: Predicting protein structural classes. Journal of theoretical biology, 2015. 374: p. 125-137.

79. Baldi, P., et al., Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 2000. 16(5): p. 412-424.

80. Urias, R.W.P., IMMAN: free software for information theory-based chemometric analysis. Mol Divers, 2015. 19.

81. De Bono, E., Lateral thinking for management. 1971.

82. Sternberg, R.J., Inteligencia exitosa: cómo una inteligencia práctica y creativa determina el éxito en la vida. 1997.

83. Urias, R.W.P., et al., IMMAN: free software for information theory-based chemometric analysis. Molecular Diversity, 2015. 19(2): p. 305-319.

84. Eriksson, L., et al., Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environ. Health Perspect., 2003. 111(10): p. 1361.

85. Ouyang, Z. and J. Liang, Predicting protein folding rates from geometric contact and amino acid sequence. Protein Science, 2008. 17(7): p. 1256-1263.

86. Wehrens, R., H. Putter, and L.M. Buydens, The bootstrap: a tutorial. Chemometrics and intelligent laboratory systems, 2000. 54(1): p. 35-52.

87. Todeschini, R., et al., MOBYDIGS version 1.0. 2005: Milano.

88. Tropsha, A., 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., 2003. 22: p. 69–77.

PRODUCCIÓN CIENTÍFICA DEL AUTOR

Publicaciones

1. García-Jacas, C. R.; Marrero-Ponce, Y.; Acevedo-Martínez, L.; Barigye, S. J.; Valdés-Martiní, J. R.; **Contreras-Torres, E**, **QuBiLS-MIDAS: A Parallel Free-Software for Molecular Descriptors Computation based on Multi-Linear Algebraic Maps**. *Journal of Computational Chemistry*. 2014; 35 (18):1395–1409. Factor de Impacto: 3.601. Indexada: Web of Science.
2. **Contreras-Torres, E**; Marrero-Ponce, Y.; García-Jacas, C. R., **Nueva codificación de la estructura tridimensional de proteínas**. Ponencia, Comisión 3, Memorias del *IV Encuentro Regional de Bioingeniería* BIOVC2014, ISBN: 978-959-312-023-4.
3. Marrero-Ponce, Y.; **Contreras-Torres, E**; García-Jacas, C. R.; Barigye, S.J; Cubillán N.; Alvarado, Y. J. **Novel 3D Bio-Macromolecular Bilinear Descriptors for Protein Science: Predicting Protein Structural Classes**. *Journal Theoretical Biology*, 2015, 374(125-137). Factor de Impacto: 2.12. Indexada: Web of Science.
4. García-Jacas, C. R; **Contreras-Torres, E**; Marrero-Ponce, Y.; Pupo-Meriño, M.; Barigye, S.J.; Cabrera-Leyva, L., **Examining the predictive accuracy of the novel 3D N-linear algebraic molecular codifications on benchmark datasets.** *Journal of Chemoinformatics*, 2016. Factor de Impacto: 4.55. Indexada: Web of Science.

Eventos Científicos

1. **Contreras-Torres, E**; Marrero-Ponce, Y.; García-Jacas, C. R., Nueva codificación de la estructura tridimensional de proteínas. Presentación de poster electrónico. *IV Encuentro Regional de Bioingeniería*, Santa Clara, Villa Clara, Cuba, 12 de diciembre de 2014.

Producción de Software

1. **Contreras-Torres, E**; García-Jacas, C.R.; Marrero-Ponce, Y.; ToMoCoMD-CAMPS MuLiMs-MCoMPAs, Versión 1.0, Departamento de Programación y Sistemas Digitales, Facultad 6, Universidad de las Ciencias Informáticas, La Habana, 2016.