CONSTRUCCIÓN DE ALFABETOS REDUCIDOS DE AMINOÁCIDOS USANDO DESCRIPTORES MOLECULARES

TATIANA DEL PILAR SUÁREZ JÁUREGUI

UNIVERSIDAD DE PAMPLONA
FACULTAD DE CIENCIAS BÁSICAS
DEPARTAMENTO DE BIOLOGÍA Y QUÍMICA
PAMPLONA
2016

CONSTRUCCIÓN DE ALFABETOS REDUCIDOS DE AMINOÁCIDOS USANDO DESCRIPTORES MOLECULARES

TATIANA DEL PILAR SUÁREZ JAUREGUI

Trabajo de grado para optar al título de magíster en química

Director
GUILLERMO RESTREPO RUBIO
Químico, MSc., Dr. rer. nat.

UNIVERSIDAD DE PAMPLONA
FACULTAD DE CIENCIAS BÁSICAS
DEPARTAMENTO DE BIOLOGÍA Y QUÍMICA
PAMPLONA
2016

Nota de aceptación

_____
_____
_____
_____
_____
_____

_____
Jurado

_____
Jurado

_____
Director

Pamplona, 2016

DEDICATORIA

A los niños, niñas, hombres y mujeres con cáncer en Colombia.

«A este niño le falta mano dura». Pero mi papá le respondía: «Si le hace falta, para eso está la vida, que acaba dándonos duro a todos; para sufrir, la vida es más que suficiente, y yo no le voy a ayudar»
Ahora pienso que la única receta para poder soportar lo dura que es la vida al cabo de los años, es haber recibido en la infancia mucho amor de los padres.”

Hector Abad

AGRADECIMIENTOS

“A Dios, el cimiento de mi vida y a mi madre, fortaleza en mi debilidad y apoyo incondicional”

Doctor Guillermo Restrepo, por su formación, paciencia, comprensión y correcciones oportunas.

Doctora Diana Alexandra Torres, por su apoyo incondicional con nuestra cohorte y su diligente trabajo como directora de la maestría en Química.

Doctor Jorge Madrid, Cirujano Oncólogo, por su asesoría frente al cáncer de seno.

Doctor Luis Fernando Arbeláez, por haber aceptado evaluar este trabajo de maestría y por sus valiosas correcciones en el área de Bioquímica.

Doctor Aldo Combariza, por haber aceptado evaluar este trabajo de maestría y por sus valiosas correcciones en el área computacional.

Magister Daniel Barrera, por sus valiosas ideas sobre proteínas.

Wilmer Leal, por compartir sus conocimientos en las áreas de matemáticas y programación.

Eugenio Llanos, por su aporte en el campo de la programación.

Magister Nancy Quintero, por todo su servicio y ayuda incondicional.

A mis queridos compañeros y amigos: **Rosana Suárez, Yaneth Cardona y Fernando Pinzón**, por su apoyo y palabras de ánimo en los momentos difíciles.

A toda mi **familia y amigos** a quienes les resté tiempo y me perdonaron porque han comprendido mis sueños y metas. A ellos, gracias por su amor incondicional.

ÍNDICE GENERAL

CAPÍTULO 1 ..... **1**
INTRODUCCIÓN 1
1.1 Justificación 1
1.2 Objetivos 2

CAPÍTULO 2 ..... **3**
MARCO REFERENCIAL 3
2.1 Plegamiento proteico 3
2.2 Semejanza entre aminoácidos 4
2.3 Descriptores moleculares 5
2.4 Métodos de reducción y clasificación 6
2.5 Mutaciones no sinónimas 7

CAPÍTULO 3 ..... **11**
METODOLOGÍA 11
3.1 Caracterización de aminoácidos 11
3.2 Selección de descriptores 11
3.3 Construcción de alfabetos reducidos de aa usando HCA y explorando *ties in proximity* 11
3.4 Selección del conjunto de mutaciones y análisis a partir de descriptores 11

CAPÍTULO 4 ..... **14**
RESULTADOS Y ANÁLISIS 14
4.1 Caracterización de aa y selección de descriptores moleculares 14
4.2 Construcción de alfabetos reducidos de aa usando HCA y explorando *ties in proximity* 14
4.3 Análisis del conjunto de mutaciones a partir de descriptores 15

CAPÍTULO 5 ..... **31**
CONCLUSIONES 31
RECOMENDACIONES 33
BIBLIOGRAFÍA 34
ANEXOS .....

CAPÍTULO 1

INTRODUCCIÓN

Justificación

Las proteínas cumplen funciones biológicas primordiales que dependen de su estructura tridimensional; no obstante, las dificultades para cristalizarlas hacen que el número de estructuras conocidas sea muy bajo comparado con la gran cantidad de información que hay sobre las secuencias de aminoácidos (aa)[1]. Como consecuencia, un gran número de estudios bioinformáticos se fundamentan en el análisis de los aa a través del alineamiento de proteínas[2]. De esta manera, si al comparar dos secuencias, sus aa no coinciden, pero existe entre ellas un antepasado común, las no coincidencias logran interpretarse como sustituciones o mutaciones puntuales no sinónimas que pueden o no modificar la función de la proteína. Los aa intercambiados dentro de la secuencia muestran una relación de semejanza; un ejemplo de esto son las matrices de sustitución, que comparan y asignan pesos a pares de aa de acuerdo a cambios evolutivos entre proteínas, que resultan ser claves para la evaluación de los alineamientos.[3] Sin embargo, esta no es la única forma en la que se ha estudiado la semejanza entre aa; las clasificaciones más comunes se han conseguido con ayuda de propiedades fisicoquímicas y bioquímicas, representadas a través de un valor numérico conocido como descriptor.[4]

Los descriptores pueden tener origen experimental o teórico; dentro de estos últimos están los topológicos y cuánticos que no son usados con frecuencia para tratar aa y no existe para ellos una base de datos que los compile, como sí la hay para descriptores experimentales, tales como Amino Acid Index[5] y ProtScale[6].

En la actualidad las clasificaciones para aa no involucran todos los descriptores experimentales y teóricos que existen, sino que se basan en selecciones subjetivas, con poco soporte matemático. Una aplicación de estas clasificaciones son los alfabetos reducidos de aa que han sido usados como base para el alineamiento de proteínas, el desarrollo de fármacos y la predicción del plegamiento proteico. Puesto que los alfabetos dependen de las propiedades usadas para su construcción, una mala elección de los descriptores puede llevar a alfabetos erróneos.

Con base en estas experiencias, se observó la necesidad de clasificar aa proteicos para la construcción de alfabetos reducidos, teniendo en cuenta la mayor cantidad de descriptores de aa posibles y seleccionando dichas propiedades sólo a través de

herramientas matemáticas. En este trabajo, a partir de técnicas como el análisis de agrupamientos jerárquico, se realizaron tantas clasificaciones de aa como propiedades fueron seleccionadas; cada clasificación basada en semejanza fue usada como un alfabeto reducido de aa. Finalmente y a modo de ejemplo, se relacionaron la colección de alfabetos reducidos y sus descriptores con las sustituciones de aa producidas en las secuencias de tres proteínas silvestres (N-carbamilasa, Lucifersas y PI3K), esto se realizó para analizar el cambio de cada par de aa (silvestre y mutado) a través de sus propiedades.

Objetivos

- Objetivo general

Construir alfabetos reducidos de aminoácidos usando descriptores moleculares y analizar algunas mutaciones puntuales no sinónimas.

- Objetivos específicos

- Reunir el mayor número de descriptores de aminoácidos posibles a través de bases de datos, algoritmos y literatura científica, para seleccionar aquellos que den mayor información sobre estas moléculas.
- Clasificar los aminoácidos con base en cada uno de los descriptores seleccionados para construir una colección de alfabetos reducidos, teniendo en cuenta los algoritmos que exploran a fondo el efecto de los *ties in proximity* sobre los resultados de las clasificaciones.
- Usar la colección de alfabetos reducidos, para analizar su relación con las mutaciones no sinónimas de N-carbamilasa, luciferasa y PI3K.

CAPÍTULO 2

MARCO REFERENCIAL

2.1 Plegamiento proteico

Los aminoácidos (aa) son compuestos constituidos por un grupo amino, un grupo carboxilo y una cadena lateral que da lugar a una gran cantidad de aa; sin embargo sólo 20 de estos compuestos llamados α,L-aminoácidos (Figura 1), codificados por 64 codones en el código genético y dos modificaciones postraduccionales (pirrolisina y selenocisteína)[a], hacen parte de las proteínas; macromoléculas que tienen un papel muy importante en todas las formas de vida[7].

Las proteínas están formadas por la combinación de aa unidos a través de enlaces químicos que les permiten adoptar estructuras y funciones específicas. Se han estudiado cuatro tipos de estructuras en las proteínas, siendo la más sencilla la estructura primaria: una secuencia de aa, que se pliega como resultado de la formación de enlaces de hidrógeno entre los grupos amino y carboxilo, generando la estructura secundaria. Un nivel superior es la estructura terciaria: causada por el el arreglo tridimensional de las estructura secundaria que trae como consecuencia la formación de puentes de disulfuro entre las cadenas laterales de algunas cisteínas (uno de los 20 aa), que al ser unidas por atracciones no covalentes pueden llegar a formar complejos proteicos conocidos como estructuras cuaternarias.[8,9]

Debido a que las estructuras de último nivel están altamente relacionadas con la función biológica de las proteínas, el plegamiento se ha convertido desde hace más de cuatro décadas[10,11,12,13] en objeto de estudio y es un tema de gran importancia para la bioquímica y la bioinformática actual.[9,14,15]

[a] El código genético involucra un conjunto de reglas que permite, dentro del proceso de traducción, el cambio de tres nucleótidos (un codón) por un aminoácido; en este proceso se construye la proteína. Algunas veces, y dependiendo de la función de una proteína, en la etapa posterior a la traducción (postraduccional), los aminoácidos lisina y cisteína cambian su estructura.

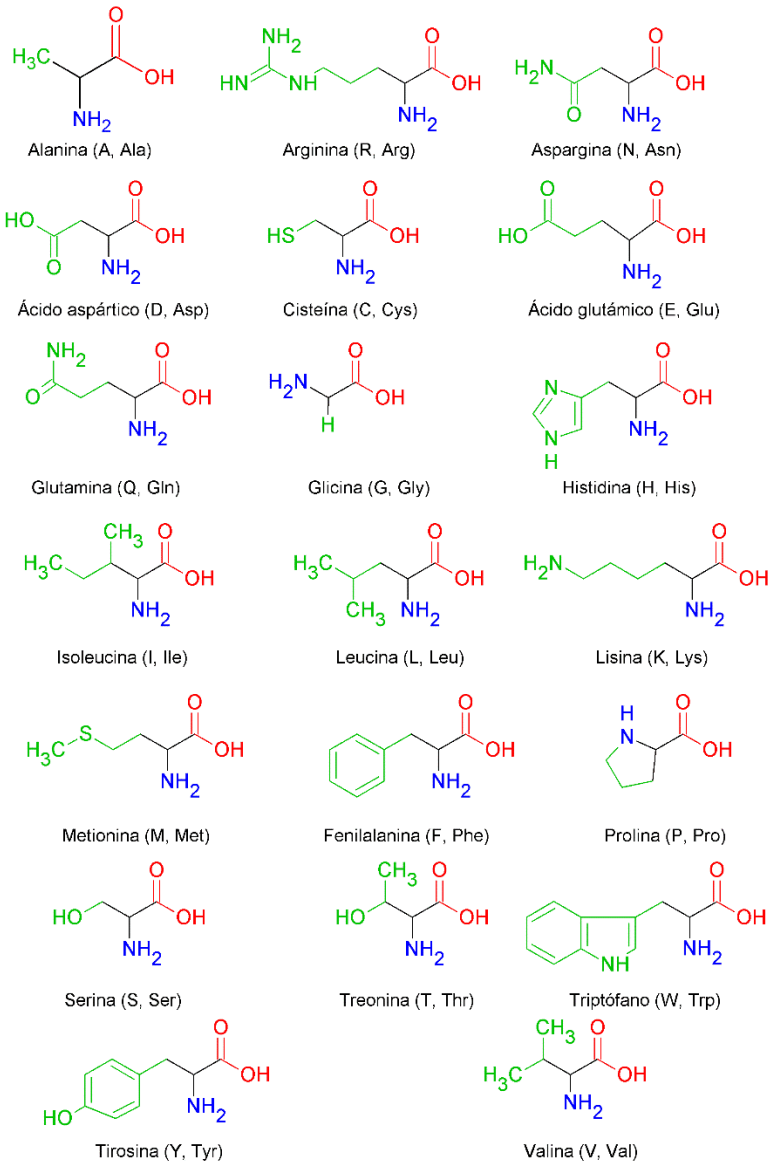

Figura 1: Estructuras moleculares, nombres y códigos de los 20 aa proteicos estándar (paréntesis). El grupo amino es mostrado en azul, el carboxilo en rojo y la cadena lateral en verde.

2.2 Semejanza entre aminoácidos

Una representación de la semejanza entre aa son las matrices de sustitución BLOSUM y PAM que relacionan dichas moléculas de acuerdo a cambios evolutivos entre proteínas asignando pesos a pares de aa. Otro tipo de aproximación a la semejanza es la caracterización numérica de las moléculas a través de descriptores y su posterior clasificación en el espacio métrico que los descriptores definen. La primera clasificación de aa fue realizada por Ramachandran *et al.*[16] en 1963,

quienes usaron como descriptor los ángulos diedros ψ y φ de la primera estructura proteica cristalizada; sobre esta misma década, Sneath[17] (1966) realizó un análisis de agrupamientos jerárquico tomando como base descriptores fisicoquímicos y bioquímicos y en 1982 Kyte y Dolittle[18] clasificaron aa a través de características hidrofílicas e hidrofóbicas. En la actualidad la clasificación de aa tiene diversos fines que dependen del objetivo que persiga la investigación [9,19]; muestra de ello son los trabajos de Etchebest[20] quien en 2007 clasificó aa a partir de propiedades de la estructura local y los comparó con mutaciones de proteínas presentes en dos tipos de microorganismos para relacionar las sustituciones en las proteínas objeto de estudio con la estructura local de los aa. Suyu y Wang[21] en 2010 modelaron la localización de proteínas subnucleares a partir de diferentes clasificaciones de aa hechas con información de sus secuencias.

Un alfabeto reducido[b] es también una clasificación de aa que busca reemplazar un aa por otro de la misma clase dentro de una secuencia proteica, sin que se vea alterada la función biológica de la proteína. Los trabajos desarrollados usando alfabetos reducidos involucran: la disminución del tiempo de cálculo en métodos *ab-initio* para predecir la estructura tridimensional de proteínas[22]; estudios de semejanza entre aa para generar la secuencia consenso en alineamientos múltiples[23,24]; la construcción de alfabetos por Cannata *et al* en 2002, basados en las matrices de sustitución BLOSUM y PAM para predecir el plegamiento de proteínas[25]. Otros trabajos involucran el diseño de nuevos péptidos antimicrobianos para mejorar su actividad biológica[26]; alfabetos reducidos generados por inferencia filogenética para disminuir errores en los modelos Markov[27]; la reducción del alfabeto a cinco letras, lo que permitió la predicción de proteínas de manera semejante a la combinación de 20 aa, pero con una gasto computacional mucho más bajo[28]. Otros estudios son también referenciados en Peterson *et al* (2009)[29] y Huang *et al.* (2015)[30]

Los trabajos mencionados basan sus clasificaciones en propiedades de los aa que tienen sesgos o han sido seleccionadas subjetivamente y pese a que en la actualidad existe una gran cantidad de descriptores para caracterizar estas moléculas, sólo se hace uso de un grupo restringido de estos. A continuación discutimos una forma de clasificar aa haciendo uso de un gran grupo de descriptores informativos, que ofrecen todas las formas posibles de caracterizar aa que serán útiles para la construcción de los alfabetos reducidos.

[b] Definición 1: $Sea\ A\ un\ alfabeto.\ Diremos\ que\ A'\ es\ un\ alfabeto\ reducido\ de\ A\ sii\ \exists f: A \rightarrow A' sobreyectiva\ y\ \exists \propto_1, \propto_2 \in A: f(\propto_1) = f(\propto_2).$

2.3 Descriptores moleculares

Un descriptor molecular es una función matemática que asigna un valor numérico a una molécula para indicar algún aspecto particular de ella o, en su defecto, de la sustancia asociada a dicha molécula[31]; estos descriptores pueden ser de origen experimental o teórico. En áreas como bioquímica y bioinformática, los aa a menudo se caracterizan por descriptores fisicoquímicos y bioquímicos, que están compilados principalmente en dos bases de datos: Amino Acid Index[5] y ProtScale[6]; a la fecha (Febrero 1 de 2016) con 544 y 57 descriptores, respectivamente. Sin embargo, existen otros descriptores como los químico-cuánticos y los topológicos,[4,32] que no se encuentran normalmente en bases de datos, pero que sí pueden calcularse[33,34,35,36,37,38] a partir de algoritmos que operan sobre la estructura molecular (grafo)[c]. Actualmente la cantidad de descriptores experimentales y teóricos asciende a más de 2.000. Sin embargo, las ventajas de la plétora de descriptores traen consigo un inconveniente, la necesidad de filtrar los más relevantes para la caracterización molecular. De esta forma han surgido métodos matemáticos para realizar tareas de selección y clasificación[4]; algunos de estos métodos serán mencionados y fueron usados en esta investigación.

2.4 Métodos de reducción y clasificación

2.4.1 *Contenido medio de información ($\bar{I}$)*

$\bar{I}$, también llamado entropía de Shannon, es una medida del grado de diversidad de los valores o elementos de un conjunto. En nuestro caso indica el grado de diversidad de un descriptor. El contenido medio de información fue definido por Shannon y Weaver como[39]:

$$\bar{I} = \frac{-\sum_{i=1}^{n} P_i log P_i}{\bar{I}_{max}}$$

Ecuación 1

Donde n es el número de elementos del conjunto, P_i es la probabilidad de seleccionar aleatoriamente un elemento de la clase $iésima$ e $\bar{I}_{max}$ es el máximo valor de diversidad obtenido con n elementos diferentes. El rango de $\bar{I}$ es [0,1], 0 indica

c
$Definición$: $Un\ grafo\ G\ es\ un\ par\ ordenado\ \big(V(G), E(G)\big), donde\ V\ es\ un\ conjunto\ de\ elementos,$ $llamados\ vértices\ y\ E\ es\ un\ conjunto\ de\ pares\ de\ vértices, llamados\ aristas.$

que todos los valores del descriptor son equivalentes y 1 que todos son diferentes. En este sentido, si lo que se desea son propiedades que distingan a los aa, $\bar{I}$ debería ser cercano a 1 o 1.

2.4.2 Análisis jerárquico de agrupamientos y *Ties in proximity*

El análisis jerárquico de agrupamientos (HCA, de sus siglas en inglés) es un método de clasificación no supervisado que permite agrupar los elementos de un conjunto con base en sus propiedades.[40] Para hacerlo se debe tener en cuenta una función que permita cuantificar la semejanza entre los objetos y una metodología para agruparlos. El resultado final es un dendrograma, es decir una representación gráfica de los agrupamientos.[41]. Sin embargo, aun siendo fijados un conjunto de elementos (X), unos atributos (a_i) que caracterizan a cada $x_i \in X$, una función de semejanza (sf) y una metodología de agrupamiento (gm), el algoritmo HCA puede encontrarse con más de una distancia mínima entre los elementos de X, tal que al proseguir el agrupamiento por uno de los mínimos, el dendrograma puede llegar a ser totalmente diferente del dendrograma obtenido si el otro mínimo hubiera sido escogido. Este inconveniente, conocido como *ties in proximity*[42] está presente a lo largo del proceso de agrupamiento y es objeto de diversos estudios[43,44].

Recientemente Leal *et al*.[45] propusieron cuatro metodologías para seleccionar él o los dendrogramas más probables productos de *ties*. Las dos primeras metodologías, *graph* y *relaxed graph cluster contrast*, consideran al dendrograma como grafo, mientras que las dos últimas, *set* y *relaxed set cluster contrast*, como conjunto. En la presente investigación, se hizo uso del *relaxed cluster contrast* que se explica detalladamente en la metodología.

2.5 Mutaciones no sinónimas

Aquellas proteínas donde se dan cambios puntuales de aa dan lugar a las mutaciones puntuales no sinónimas. Debido a los diferentes usos de los alfabetos en proteínas y a la capacidad de cuantificar la frecuencia de los agrupamientos a través de la metodología de Leal *et al.*, es posible evaluar la efectividad de un alfabeto reducido en la medida en que el aa_i sustituido sea reemplazado por algún aa de la clase a la que pertenece aa_i.

Basados en los trabajos de Oh *et al.*[46], Law *et al.*[47] y Etchebest *et al.*[48], se escogieron las enzimas N-carbamilasa desde *Agrobacterium Tumefaciens* y luciferasa desde *Phonitus Pylaris*. De manera semejante, la investigación desarrollada por Hart *et al.* en 2015[49] permitió seleccionar la proteína PI3K (Fosfatidil inositol 4,5 bifosfato 3 quinasa, subunidad catalítica alfa), cuyos cambios de aa fueron tomados de la base de datos Catalogue of Somatic Mutation in Cancer (COSMIC)[50].

A continuación se describen las proteínas seleccionadas y las variaciones ocurridas en la función biológica como consecuencia de las mutaciones en sus secuencias.

2.5.1 N-carbamilasa

La N-carbamilo-D aminoácido amidohidrolasa o N-carbamilasa, es una enzima de 304 aa tipo hidrolasa que cataliza la hidrólisis del grupo N-carbamilo a partir de N-carbamilo D-amino y es empleada industrialmente para la formación de D-aminoácidos.[51] Estos aa son usados con frecuencia para sintetizar antibióticos, anti-fúngicos, pesticidas y edulcorantes.[52] Pese a las ventajas de esta enzima, especificidad y no afectación ambiental, presenta limitantes industriales que incluyen un bajo poder oxidativo y térmico. Oh *et al.* en 2002 usaron *DNA shuffling* y evolución directa para crear una librería de 10.000 clones de la enzima nativa; después de los ensayos previos, sólo dos enzimas mutadas (1S15 y 2S3) fueron seleccionadas y analizadas. En este estudio escogimos el mutante 2S3 por contener seis mutaciones que involucran las dos únicas presentes en la enzima mutante 1S15. La tabla 1 muestra las mutaciones y las posiciones en que ocurrieron dichos cambios.

Tabla 1: Sustituciones de aa y posiciones en la enzima nativa N-carbamilasa del mutante 2S3.

Aminoácido nativo	Aminoácido mutado	Posición
Q	L	23
V	A	40
H	Y	58
G	S	75
M	L	184*
T	A	262*

*Sustituciones presentes en el mutante 1S15.

2.5.2 Luciferasa

La luciferasa de luciérnagas es una enzima de 62 kDa y 523 aa en su secuencia, que cataliza la producción de luz a través de ATP-Mg^{2+}, luciferina de luciérnagas y oxígeno molecular; oxidando el sustrato y produciendo oxiluciferina. Esta enzima, ampliamente usada en bioluminiscencia[d], ha permitido el desarrollo de aspectos biológicos como: expresión génica, detección de patógenos y ensayos basados en la interacción proteína-proteína en células[53]. Pese al elevado uso de la proteína nativa o recombinante, la luciferasa se inactiva fácilmente a temperaturas elevadas y su espectro luminiscente varía a pH bajos (entre 4 y 5), dicha variación resulta

[d] La bioluminiscencia está definida como el número de fotones emitidos por molécula de luciferina consumida.

problemática para las aplicaciones que tiene la enzima. Buscando solucionar estos inconvenientes, Law *et al.* en 2006 reportaron un mutante de luciferasa de luciérnagas con resistencia a altas temperaturas y a bajos pH, a través de la sustitución de cinco aa cuyas posiciones y nombres se muestran en la tabla 2.

Tabla 2: Posiciones de los aa nativo y mutado de la enzima luciferasa de luciérnaga.

Aminoácido nativo	Aminoácido mutado	Posición
F	R	14
L	Q	35
V	K	182
I	K	232
F	R	465

2.5.3 PI3K

El gen fosfatidil inositol 4,5 bifosfato 3-quinasa (PIK3CA) codifica la proteína PI3K de 1.068 residuos de aa, compuesta por una subunidad reguladora tipo alfa de clase I de 85 kDa y una subunidad catalítica tipo alfa de clase I de 110 kDa con actividad quinasa; la subunidad catalítica p110α fosforila el OH 3' del fosfatidil inositol, fosfatidil inositol 4-fosfato y fosfatidil inositol 4,5-bifosfato para producir fosfatidil inositol 3,4,5-trifosfato o PIP3. La fosforilación catalizada por p110α es fundamental para las proteínas de señalización encargadas del crecimiento y división celular, el movimiento de las células y la producción de nuevas proteínas. Las mutaciones ocurridas en un sólo aminoácido se asocian a enfermedades que involucran anomalías capilares y a muchos tipos de cánceres, incluyendo cáncer de ovario, seno, pulmón, estómago y colorrectal; la mayoría de las mutaciones producidas en este gen son somáticas, es decir, se adquieren a lo largo de la vida de una persona. Los cambios en los aa (tabla 3) producen una sobreactivación de la enzima que no logra ser controlada por la subunidad reguladora, aumentando la señalización, lo que conduce a la proliferación anormal de las células.[49]

El cáncer es una de las tres enfermedades que causa mayor índice de mortalidad. El cáncer de mama, por ejemplo, es la enfermedad más común y la que más cobra vidas en mujeres[54]. Dados estos antecedentes, se empleó la base de datos COSMIC para seleccionar las mutaciones de PI3K presentes en cuatro subhistologías de cáncer de mama: Basal triple negativo, ER-PR positivo, HER positivo y Ductal.

Tabla 3: Sustituciones y posiciones de los aa en la proteína PI3K asociados a cáncer de mama.

AA	Mutación	Posición
H	R	1047
H	L	1047
E	K	545

La figura 2 ilustra la activación de PI3K que inicia cuando un factor de crecimiento o ligando se une a su receptor tirosina quinasa (RTK). Estos receptores incluyen a miembros del factor de crecimiento epidérmico humano (HER). Tras la activación del receptor, PI3K interactúa con su porción intracelular a través de su parte reguladora P85. Alternativamente una molécula adaptadora puede actuar como intermediario entre RTK y P85, de esta forma se inhabilita el efecto inhibidor de P85 en P110 y en ese momento se produce la activación completa de PI3K. Entonces la quinasa PI3K activada cataliza la fosforilación de bifosfatoinositol (PIP2) a trifosfato fosfatidilinositol (PIP3). PIP3 se acopla a AKT, una serina/treonina quinasa que es la mediadora central de la vía PI3K. Una vez localizada en la membrana de la célula, AKT se fosforila y es activada por mTOR, estimulando la síntesis de proteínas y el crecimiento celular.[55]

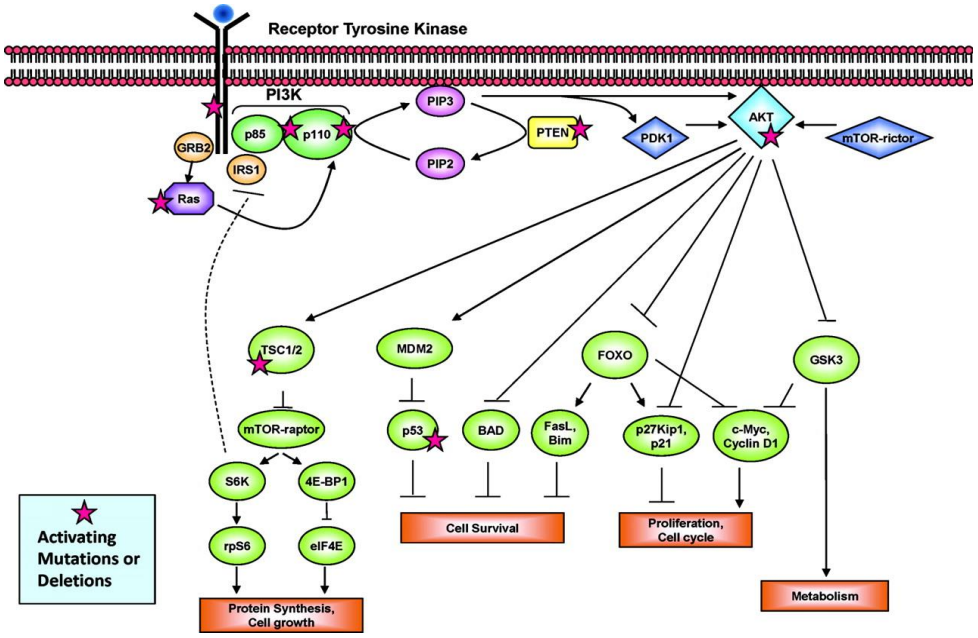

Figura 2: Esquema de la activación de la proteína PI3K.[55]

CAPÍTULO 3

METODOLOGÍA

3.1 Caracterización de aminoácidos

Desde las bases de datos Amino Acid Index[5] y ProtScale[6] se recolectaron 565 descriptores recopilados para los 20 aminoácidos (aa) proteicos. De manera semejante, se seleccionaron 651 descriptores topológicos y cuánticos a partir de una revisión en la literatura científica y finalmente, basados en la estructura molecular de cada aminoácido, se calcularon 1.401 descriptores usando PaDEL (Pharmaceutical Data Exploration Laboratory).[34] Un total de 2.717 descriptores fueron usados para iniciar este estudio.

3.2 Selección de descriptores moleculares

Sobre cada uno de los 2.717 descriptores se calculó el índice del contenido de información $(\bar{I})$ y se seleccionaron 1.729 descriptores que ofrecían por lo menos un 70% de información.

3.3 Construcción de alfabetos reducidos de aminoácidos usando HCA y explorando *ties in proximity*

Sobre cada uno de los 1.729 descriptores con alto contenido de información se corrió un HCA donde los objetos a clasificar fueron los 20 aa; esto llevó a una matriz de distancia para cada descriptor. A partir de cada matriz, empleando la metodología de agrupamiento unión promedio, se calcularon todos los dendrogramas (clasificaciones) productos de *ties*. Las clasificaciones generadas representan los alfabetos reducidos de aa sobre descriptores individuales de estas moléculas. Adicionalmente se analizó la distribución del número de dendrogramas (clasificaciones) generados como producto de *ties in proximity*.

3.4 Selección del conjunto de mutaciones y análisis a partir de descriptores

Con el fin de probar la utilidad de las clasificaciones construidas, se estudiaron los cambios de aa ocurridos en las proteínas N-carbamilasa, luciferasa y PI3K, cuyas mutaciones en la secuencia proteica generaron alteraciones en sus características biológicas.

Basados en las investigaciones de Oh *et al.* y Law *et al.* realizadas para N-carbamilasa y luciferasa, Etchebest *et al.* buscaron relacionar dichas sustituciones de aa con una clasificación realizada a partir de la estructura local de los aa, sin embargo dicho análisis no tuvo éxito, ya que no logró explicar totalmente estos cambios a partir de la estructura local.

Para obtener información sobre los genes asociados con mayor frecuencia al cáncer de seno, se usó la base de datos COSMIC y se hizo un filtro de aquellos genes cuya expresión en proteínas cumplieran con las siguientes características: 1) más del 90% de las mutaciones fueran puntuales no sinónimas, 2) el estado somático de la enfermedad estuviera confirmado y 3) el impacto de la mutación fuera patogénico. Finalmente dada la posición fija en las mutaciones se escogió el gen PIK3CA que se expresa en la proteína PI3K, cuyos cambios de aa son reportados en la tabla 3 y están relacionados con las cuatro subhistologías de cáncer de mama más frecuentes: Basal triple negativo, ER-PR positivo, HER positivo y Ductal.

Las mutaciones no sinónimas encontradas en cada proteína fueron comparadas con la colección de alfabetos reducidos. Para esto, sobre cada descriptor se calculó la frecuencia con la que aparecía el par de aa en los dendrogramas productos de *ties*. La frecuencia del par de aa se dedujo usando *relaxed set cluster contrast (*CC_{rs}*)*, ya que los agrupamientos de interés para las mutaciones no sinónimas estudiadas son de dos elementos (dos aa), lo que hace que la estructura del agrupamiento no sea relevante, debido a que es única (todo agrupamiento de dos elementos tiene asociado siempre un único grafo). De las dos metodologías de contraste basadas en conjuntos, se calculó el *relaxed cluster contrast* para tener en cuenta más detalles de la presencia parcial de la pareja de aa de cada mutación en los diferentes dendrogramas.

El *relaxed set cluster contrast* se define como:

$$CC_{rs}(C, D_i) = max \frac{|L(C) \cap S_j|}{|L(C) \cup S_j|}$$

donde C es el agrupamiento de la pareja de aa involucrados en la mutación, D_i es cada uno de los dendrogamas en el que se busca la presencia de C, $L(C)$ es la pareja de aa, S_j es el conjunto asociado al subárbol j del dendrograma estudiado. Un subárbol es una rama del dendrograma. Por lo tanto $CC_{rs}(C, D_i)$ busca el subárbol más pequeño del dendrogama que contiene a la pareja de aa. Un valor de $CC_{rs}(C, D_i) = 1$ indica que el subárbol más pequeño que contiene a la pareja tiene sólo a la pareja. $CC_{rs}(C, D_i) = 0{,}5$, por ejemplo, muestra que la mitad del cluster más pequeño que contiene a la pareja es ocupada por ella.

La frecuencia de un agrupamiento en los dendrogramas resultantes de *ties* se calcula mediante:

$$f(c) = \frac{1}{m} \sum_{i=1}^{m} cc(c, D_i)$$

$f(c)$ es el promedio de los valores de *cluster contrast* determinados para cada uno de los m dendrogramas producto de *ties*.

Finalmente, para buscar las propiedades asociadas a la mutación, se seleccionaron aquellos descriptores en los que el par de aa tuviera una $f(c) = 1$ y se compararon con los descriptores que generaran las más bajas frecuencias en un rango $f(c) = [0{,}1 - 0{,}2)$.

Para mejorar el entendimiento de la metodología desarrollada a través de HCA, *ties in prximity* y *relaxed set cluster constrast*, por favor diríjase al ejemplo propuesto en el material suplementario.

CAPÍTULO 4

RESULTADOS Y ANÁLISIS

4.1 Caracterización de aminoácidos y selección de descriptores

Los 2.717 descriptores recolectados incorporan características fisicoquímicas, cuánticas, topológicas y bioquímicas de los aa. A través de $\bar{I}$ se midió la diversidad de información aportada por cada descriptor. En la tabla I del material suplementario se muestran los 1.729 descriptores seleccionados con más del 70% de información para los aa. Los descriptores topológicos de PaDEL y los tomados de la literatura presentaron los valores de información más bajos, probablemente porque a nivel estructural los aa comparten muchas características dentro de la cadena lateral que se representan a través de valores booleanos.

4.2 Construcción de alfabetos reducidos de aminoácidos usando HCA y explorando *ties in proximity*

La figura 3 muestra la distribución del número de dendrogramas frente a los 1.729 descriptores usados para la clasificación.

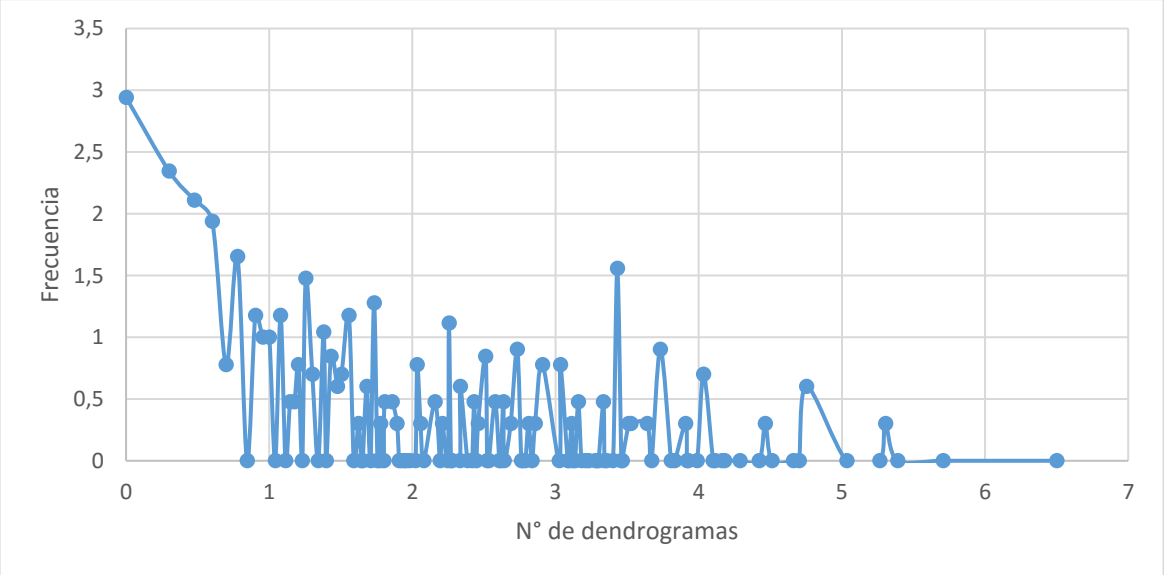

Figura 3: Distribución del número de dendrogramas productos de *ties in proximity* en 1.729 descriptores (ejes en escala logarítmica).

El número máximo de posibles dendrogramas que se pueden obtener producto de los *ties in proximity* en un conjunto de 20 elementos está determinado por el número de Felsenstein[e] [56] y es $8{,}2x10^{21}$, que indica el número máximo de clasificaciones posibles para los 20 aa dada una propiedad; sin embargo el máximo número de dendrogramas en este estudio fue 3.175.200, y estuvo presente en el descriptor *Average relative fractional occurrence in AL(i)* con código RACS820103 en la base de datos *Amino Acid Index*. La matriz de distancia de este descriptor muestra que ocho de los 20 aa tenían el mismo valor (cero), correspondiente al mínimo de los 20 datos, lo que implica que sólo se pudo medir la frecuencia en *AL(i)* para 12 aa, mientras los ocho aa restantes no se hicieron presentes.

Un total de 20.871.586 clasificaciones (dendrogramas) de los aa fueron obtenidas sobre los 1.729 descriptores; estas clasificaciones representan la colección de alfabetos reducidos de aa más grande que se ha realizado hasta el momento. La figura 3 muestra una distribución logarítmica de los resultados, en donde se observa que el 72% de los descriptores generan de 1 a 4 dendrogramas, esto indica que hay un máximo de cuatro clasificaciones diferentes para la mayor parte de los descriptores. El porcentaje de descriptores restantes (28%) contienen de 5 a 3.175.200 clasificaciones.

4.3 Análisis del conjunto de mutaciones a partir de descriptores

Las frecuencias de las mutaciones están ubicadas en un rango de cero a uno, un valor de cero indica que el par de aa (la mutación) no aparece junto en ninguna de las clasificaciones generadas para un descriptor específico y un valor de uno muestra que el par de aa (silvestre y mutado) aparece siempre unido, sin importar cuántos dendrogramas productos de *ties* hay para alguna propiedad. Las distribuciones de dichas frecuencias se muestran a continuación.

4.3.1 N-carbamilasa

Oh *et al.* mejoraron el poder oxidativo y térmico de la enzima N-carbamilasa cambiando los seis aa mostrados en la tabla 1. Los autores, usando mutación dirigida, sustituyeron cada aa individualmente sobre la enzima nativa y midieron el cambio térmico y oxidativo producido por cada mutación, los resultados de este trabajo son mostrados en la tabla 4. Tal como se observa, la mutación T262A

[e] El número de Felsenstein está definido como: $F(n) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$ donde *n* es el número de elementos a clasificar, en este caso los 20 aa.

resultó ser el cambio que más contribuyó a mejorar las propiedades de la enzima, seguido de las mutaciones M184L y H58Y quienes tuvieron una contribución positiva para la estabilidad térmica y oxidativa; mientras Q23L, G75S y V40A sólo aumentaron la resistencia a oxidarse.

Tabla 4: Estabilidad oxidativa y térmica de N-carbamilasa silvestre y cada uno de sus mutantes.

Enzima	**Termoestabilidad (%)[a]**	**Estabilidad oxidativa (%)[b]**
Silvestre	10,0 ± 0,8	5,0 ± 1,2
M184L	19,7 ± 2,0	20,0 ± 2,3
T262A	50,5 ± 3,4	40,9 ± 2,2
Q23L	19,1 ± 1,4	63,9 ± 2,3
V40A	7,0 ± 2,3	20,6 ± 1,2
H58Y	20,0 ± 1,9	20,6 ± 1,2
G75S	3,5 ± 1,1	42,5 ± 0,8
2S3	78,5 ± 3,1	79,3 ± 1,2

a. Tratamiento térmico a 70°C durante 30 minutos

b. Incubación con peróxido de hidrógeno 0,2mM durante 30 min a 25°C

*Valores tomados de Oh *et al.*[46]

La figura 4 muestra para cada mutación el número de descriptores que contienen los valores de frecuencia. Los histogramas de cada mutación exponen un mismo comportamiento: no se observan frecuencias iguales a cero, dado que no son posibles, pues esto implicaría que el par de aa no esté presente en ningún dendrograma. La frecuencia de 0,333 es la más común para un gran número de dendrogramas provenientes de diferentes descriptores. Esto significa que la pareja de aa en cuestión ocupa el 33% del tamaño del agrupamiento más pequeño que la contiene en cada dendrograma, o de otra manera, que el agrupamiento más pequeño y común para este estudio, que contiene a la pareja, es de seis elementos. Ya que el interés en este estudio se concentra en la identificación de los descriptores que hacen que determinadas parejas de aa sean semejantes, las frecuencias a estudiar en detalle fueron aquellas iguales a 1. Es decir, se determinaron aquellos descriptores donde la pareja de aa estuvo presente en todos los dendrogramas productos de *ties* y además la pareja siempre se dio en un agrupamiento de dos elementos.

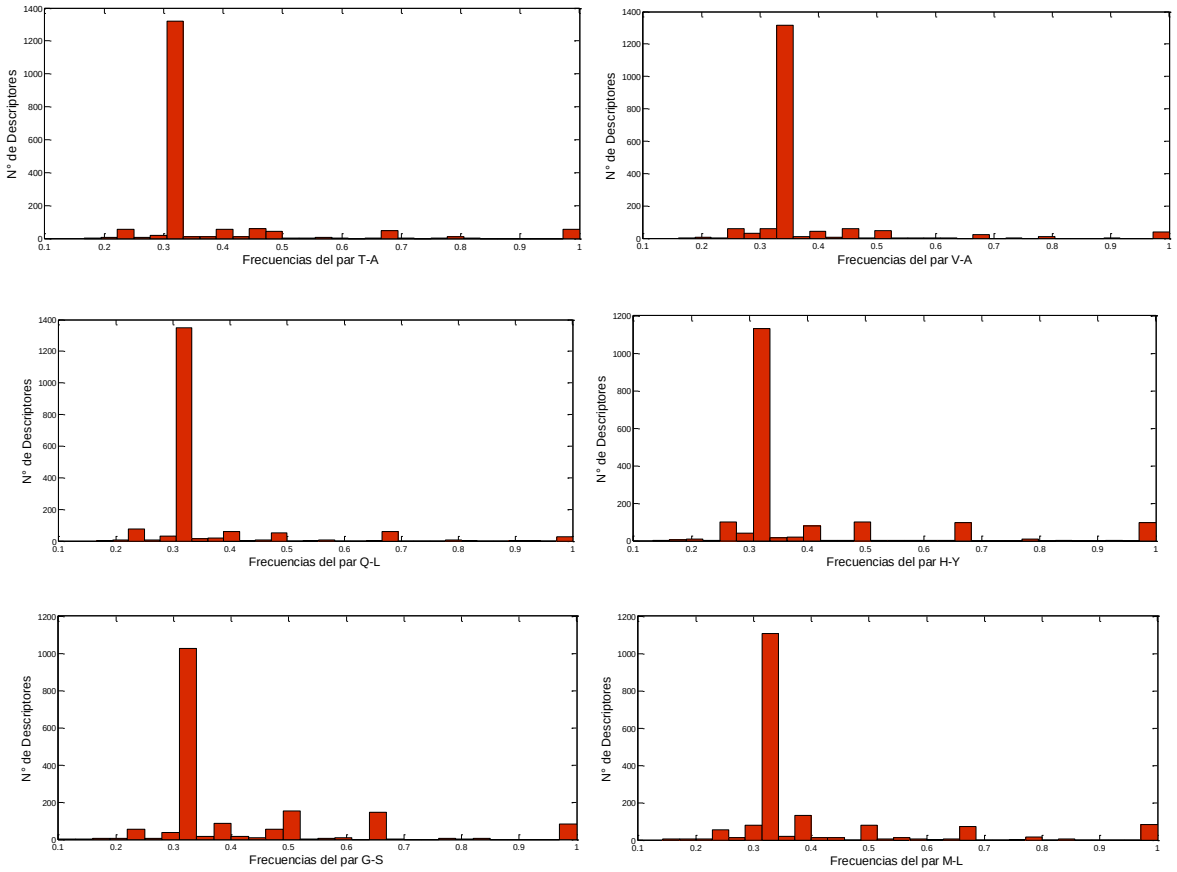

Figura 4: Distribución de las frecuencias en 1.729 descriptores para las seis mutaciones hechas sobre la enzima N-carbamilasa.

Los descriptores con frecuencia=1 para las seis mutaciones de N-carbamilasa son mostrados en la tabla II del material suplementario y son propiedades invariantes en la posición de la secuencia proteica donde se ubica la mutación. Un análisis de descriptores fue llevado a cabo sobre cada par de aa (silvestre y mutado) para entender el cambio térmico y oxidativo que se produjo en la enzima. El procedimiento realizado fue el siguiente: cada descriptor tiene asociado un número *m* de dendrogramas productos de ties, en esos m dendrogramas se buscó la pareja (*i,j*), siendo *i* y *j* dos aa diferentes. Esto llevó a m_{ij} dendrogramas, con $m_{ij} < m$. Posteriormente se buscaron los descriptores asociados a esos m_{ij} dendrogramas y se intentó dar una explicación de la razón por la cual *i* y *j* son similares teniendo en cuenta la posible interpretación de cada uno de los descriptores encontrados.

- Glutamina (Q) por Leucina (L)

26 descriptores están relacionados con el cambio de Q por L (tabla II). La frecuencia en estructuras α-hélice, la energía libre y el área superficial de los aa expuestos a un solvente, son, en general las propiedades fisicoquímicas que no varían al generarse dicha mutación en N-carbamilasa. Así mismo, las características estructurales de Q y L se asocian con ocho descriptores topológicos de Morse que combinan la estructura tridimensional de los aa basada en la difracción de electrones, midiendo la distancia euclidiana entre sus átomos y ponderándolos a partir de una propiedad atómica como: masa, volumen de Van der Waals, electronegatividad de Sanderson y polarizabilidad, las dos últimas propiedades se relacionan dentro de las distancias 4 a 8 $Å^{-1}$; otras propiedades topológicas semejantes fueron el descriptor geométrico MAXDP que permitió calcular variación eléctrica positiva usando la estructura tridimensional de los aa. Además se encontraron descriptores topológicos de dos dimensiones (2D), dentro de los que se incluyen: Kier & Hall que muestran características de conectividad de sus átomos y autocorrelación que involucra la estructura molecular (grafo) con propiedades fisicoquímicas. Finalmente, dos descriptores cuánticos mostraron semejanza entre el par de aa, el primero $S_{HF,}$ representa la blandura general de los aa, depende del potencial de ionización y la afinidad electrónica, y está relacionado con la función de Fukui, basada en la teoría funcional de la densidad (DFT) que indica las regiones donde una molécula cambia la densidad de carga durante una reacción; la segunda propiedad cuántica está relacionada con la población de Mulliken para el átomo de carbono y se calcula a partir de los métodos PM3 cuyo objetivo es caracterizar las propiedades electrónicas de los aa. Las propiedades electrónicas en la cadena lateral y la densidad de carga pueden afectar la estructura, la reactividad, la flexibilidad y la rotación de la proteína.

Es interesante resaltar que propiedades como la hidrofobicidad no mostraron relación de semejanza entre el par Q y L. Para el caso de N-carbamilasa se puede asociar el aumento de la hidrofobicidad con la resistencia a altas temperaturas, dado que las interacciones hidrofóbicas contribuyen a estabilizar la estructura tridimensional proteica.

- Valina (V) por Alanina (A)

Se encontraron 37 descriptores relacionados con la sustitución de este par de aa, la mayoría de ellos provenientes de las bases de datos de AA index y Protscale. Dentro de los descriptores más destacados se encuentra la mutabilidad relativa, que indica probabilidades de cambio muy parecidas para V y A dentro de la secuencia proteica; la misma tendencia a participar en estructuras secundarias específicas; el índice de carga eléctrica medido sobre la estructura 3D de los aa y la hidrofobicidad.

A diferencia de Q y L, el cambio de V por A relacionó la hidrofobicidad y la baja carga eléctrica, que pueden estar asociadas a la disminución de la termoestabilidad de la enzima. Otros descriptores basados en la estructura molecular de los aa son el descriptor cuántico $S_{HF,}$ antes descrito y los topológicos que estudian la estructura planar (2D) de los aa a través de grafos, cuyas aristas y/o vértices son pesados a través de propiedades que involucran aspectos como el número atómico, la masa, la electronegatividad y la polarizabilidad.

El volumen no resultó dentro del grupo de descriptores semejantes, por lo que el cambio de la cadena de V por A muestra una disminución en el tamaño y como consecuencia de impedimentos estéricos, estabilizando la estructura secundaria.

- Histidina (H) por tirosina (Y)

96 descriptores son semejantes para este par de aa, entre ellos el parámetro de Chou-Fasman que mide la frecuencia con la que cada aa se hace presente en estructuras secundarias como *coil* y α-hélice. Ambos aa son considerados polares a partir de su cadena lateral y como consecuencia levemente hidrófobos, sin embargo, la carga eléctrica no resultó semejante en este par de aa, por lo que podemos considerar que Histidina supera en polaridad a Tirosina; el aumento de la resistencia térmica de la enzima puede asociarse a una disminución en la polaridad al adicionar un aa aromático en la posición 58 de la secuencia. La semejanza a nivel estructural de estos dos aa se relaciona con descriptores topológicos de autocorrelación como Geary, Broto-Moreau y Moran que combinan las distancias entre los átomos a partir del grafo molecular con ponderaciones usando propiedades como: masa, polarizabilidad y electronegatividad. Además de Getaway, un tipo de descriptor relacionado con la geometría, la topología y el peso de los átomos presentes en los aa y Morse que ofrece información a partir de la estructura 3D y de propiedades como el volumen de Van der Waals, la polarizabilidad y la electronegatividad. Gran parte de estos descriptores que ponderan las características estructurales de los aa con alguna propiedad, ofrecen información sobre los impedimentos estéricos que puede generar el aa frente a la estructura de la proteína, sin embargo Tirosina e Histidina presentan una distribución de cargas semejante en la cadena lateral debido a su anillo de dobles enlaces con uno y dos heteroátomos, por lo que no pudo haber un cambio drástico en la estructura proteica. Otros descriptores que ratifican la importancia estructural del aa que se encuentre en la posición 58 son el estado electrotopológico del aa, calculado desde un grafo sin hidrógenos en el que se considera sólo información topológica y electrónica de los átomos pesados y sus enlaces; el índice extendido del átomo topoquímico (ETA) que considera el tamaño, la forma, la ramificación y la

funcionalidad del grafo molecular y finalmente el número de anillos (ring count) y las líneas que conectan los caminos entre las estructuras moleculares (chi path).

- Glicina (G) por Serina (S)

82 descriptores resultaron semejantes en esta mutación, entre ellos: la tendencia del par de aa a pertenecer a la estructura secundaria ß-plegada, la energía libre, un descriptor de polaridad, dos de hidrofobicidad e ISA (por sus siglas en inglés Isotropic Surface Area) que mide la porción del soluto accesible al solvente y también puede interpretarse como una medida de hidrofobicidad. Dentro de estas propiedades semejantes, la hidrofobicidad continúa desempeñando un papel primordial, no obstante existen una gran cantidad de descriptores que miden hidrofobicidad y no resultaron análogos, estos últimos son concordantes con las características estructurales de los dos aa, puesto que la presencia del grupo hidroxilo en la cadena lateral de la Serina permite la formación de puentes de hidrógeno con otras sustancias polares, mientras que el hidrogeno presente en la glicina, suele desestabilizar dichos enlaces; una disminución de las interacciones hidrofóbicas, aumenta la probabilidad de distorsión en la estructura de la enzima. Algunos descriptores topológicos de Broto y Morse, las matrices de Barisz y *chi path*, también resultaron semejantes, sin embargo, éstas analogías corresponden sólo a una parte de la estructura del aa que involucra pequeñas distancias y no distingue entre sus cadenas laterales. Finalmente la polarizabilidad atómica también resultó una propiedad semejante; la polarizabilidad está definida como la facilidad que presenta un átomo para ser distorsionado y está relacionada con los efectos estéricos en la proteína, por lo cual la presencia de glicina en la posición 75 impone el mínimo efecto inductivo en la proteína, lo que le permite mayor flexibilidad estructural, por otro lado, la Serina no es un aa voluminoso, pero su grupo hidroxilo puede generar algunas alteraciones en la estructura proteica.

El cambio de G por S produjo en la N-carbamilasa una disminución en la estabilidad térmica de la enzima, así como un aumento en la estabilidad oxidativa, de esta forma podemos asociar dichos cambios a un leve aumento en la polaridad del aa y la presencia del grupo hidroxilo de la Serina que siendo este de naturaleza electronegativa no tenderá a ceder electrones en presencia de algún oxidante.

- Metionina (M) por Leucina (L)

84 descriptores fueron encontrados para el cambio de M por L, entre ellos la energía libre de Gibbs, la hidrofobicidad y la frecuencia en estructuras α-hélice, propiedades fundamentales para mantener estables la proteína en su estado nativo.

Descriptores topológicos tridimensionales de Morse ponderados con propiedades fisicoquímicas como Volumen de Vander Waals, electronegatividad y polarizabilidad; el índice de forma de Peptitjean que combina la estructura 3D y el grafo molecular para caracterizar los aa, descriptores de autocorrelación ponderados con masa atómica y polarizabilidad, la matriz de Barisz, los valores propios modificados de carga, el estado electrotopológico del tipo de átomo y el átomo topoquímico extendido, muestran que en la posición 184 deben estar presentes aa hidrófobos, sin carga, soportando volúmenes medianos y cadenas laterales alifáticas.

- Treonina (T) por Alanina (A)

Finalmente, la mutación ubicada en la posición 262 está caracterizada por 56 descriptores. Dentro de las propiedades más sobresalientes están la hidrofobicidad, la energía libre y la tasa de migración (RF) medida a través de la interacción del soluto, el solvente y el absorbente. Los descriptores topológicos de Morse y los de autocorrelación con propiedades como masa atómica, polarizabilidad y volumen de Van der Waals. El estado electrotopológico del tipo de átomo, calculado desde un grafo sin hidrógenos en el que se considera sólo información topológica y electrónica de los átomos pesados y sus enlaces; la matriz de Barisz y el índice extendido del átomo topoquímico. Otros descriptores que estuvieron presentes sólo en esta mutación, incluyen el índice de grado de plegamiento, basado en los momentos espectrales calculados a través de una representación matricial de los ángulos diedros, relacionados con las frecuencias en α-hélice, y *asphericity* (ASP) que mide la desviación de la forma esférica, explicando el comportamiento de aa alifáticos como T y A. Este cambio de aa produjo un aumento en las estabilidades térmicas y oxidativas de la enzima, que como ya hemos dicho, pueden asociarse a un leve aumento en la hidrofobicidad y a la presencia de grupos que difícilmente ceden sus electrones en presencia de algún oxidante.

Para resumir, estas seis mutaciones comparten propiedades fisicoquímicas y topológicas relacionadas con la hidrofobicidad, la electronegatividad, la polarizabilidad, la masa y el volumen de Van der Waals. Como ya se ha aclarado estos descriptores pueden considerarse como esenciales para que la función biológica de la N-carbamilasa permanezca invariable. Aunque el cambio para Q23L, V40A, H58Y, G75S, M184L y T262A no varíe de manera considerable, existe diferenciación en los valores de sus descriptores que permite distinguirlos. Por ejemplo A, L y Y son más hidrofóbicos y menos polares que T, M y H; este leve aumento en la hidrofobicidad que no afectó la estructura secundaria de la N-carbamilasa permitió estabilizar la proteína a través de interacciones hidrofóbicas

que lograron hacerla más resistente a los aumentos de temperatura. Caso contrario sucedió con las sustituciones de G por S y V por A, donde los aa mutados pueden considerarse menos hidrofóbicos y más polares que los originales, situación que puede explicar por qué en estos cambios la enzima mutada se desestabilizó a temperaturas más bajas que la proteína silvestre. De manera semejante, el cambio por aa con volúmenes más pequeños evita impedimentos estéricos y aumenta la estabilidad de la proteína; esto se observó especialmente en descriptores topológicos encargados de medir tamaño, forma, ramificación y número de anillos. En la mutación Q23L la hidrofobicidad no resulta una propiedad semejante, pero hace parte de aquellos descriptores con frecuencia < 1; siendo L un aa más hidrofóbico, la estabilidad térmica se ve medianamente mejorada. Por otro lado, la electronegatividad se ve representada en la cadena lateral de Q a través del oxígeno y nitrógeno, dos átomos electronegativos y oxidantes; no obstante L no contiene átomos electronegativos en su estructura y como consecuencia de este cambio se mejoró notablemente la estabilidad oxidativa de la proteína. Una situación semejante ocurrió al cambiar A por T en la posición 262.

4.3.2 Luciferasa

La enzima mutante de luciferasa desde *Phonitus Pylaris* contiene cinco cambios en su secuencia: F14R, L35Q, V182K, I232K y F465R; teniendo en cuenta que en las posiciones 14 y 465 se repitió el cambio: F por R. De manera similar a las mutaciones de N-carbamilasa, las cuatro parejas de aa fueron asociadas por semejanza a diferentes descriptores de acuerdo a la colección de alfabetos reducidos. La figura 5 muestra la distribución de las frecuencias de cada par de aa en los 1.729 descriptores; allí más del 70% de los descriptores tienen frecuencias menores a 0,5, la mayoría con $CC_{rs} = 0{,}333$.

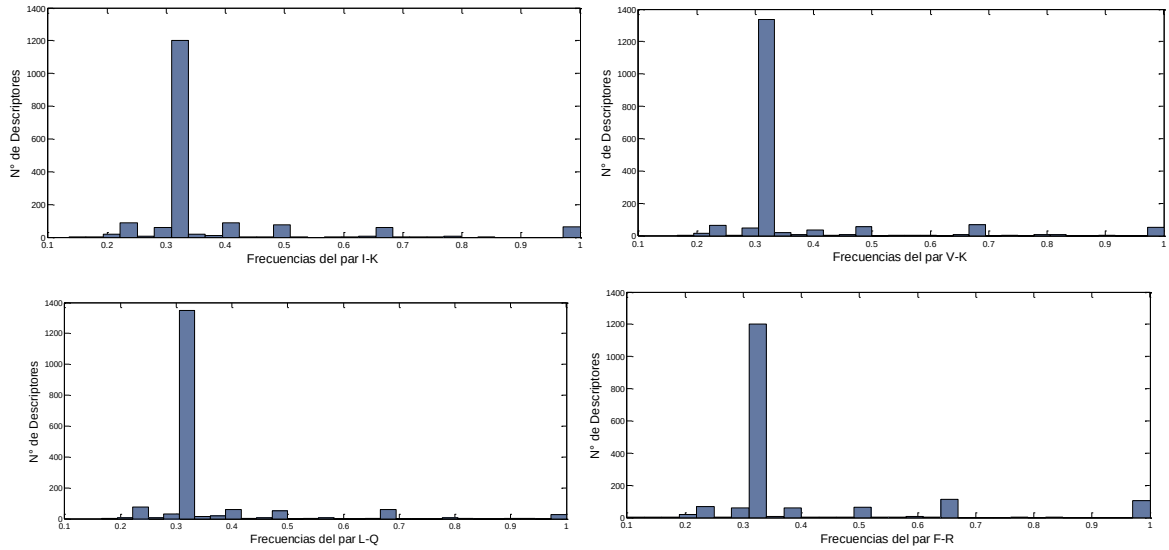

Figura 5: Distribución de las frecuencias en 1.729 descriptores para las cuatro mutaciones hechas sobre la enzima luciferasa.

- Fenilalanina (F) por Arginina (R)

106 descriptores divididos en propiedades fisicoquímicas, cuánticas y topológicas fueron asociados a la mutación F-R. Como se aprecia en la tabla II del material suplementario, descriptores como volumen y frecuencia en estructuras secundarias de tipo ß-plegada son invariantes en las posiciones 14 y 465 de la secuencia; efectivamente podemos considerar a F y R como aa de gran tamaño que pueden desestabilizar las estructuras secundarias de tipo α-hélice y que por esta misma razón son más frecuentes en ß-plegada. Es importante resaltar que descriptores como hidrofobicidad y carga eléctrica no hicieron parte del conjunto de propiedades semejantes en esta mutación; este resultado tiene sentido ya que la Fenilalanina es considerado un aminoácido hidrófobo debido al grupo aromático y a la no presencia de algún átomo electronegativo dentro de su cadena lateral, mientras que la Arginina posee un conjunto de átomos electronegativos que pueden formar puentes de hidrógeno; a esta característica se le suma la carga positiva debida a su capacidad para aceptar protones, es decir, a su comportamiento básico según la teoría de Brostend y Lowry. Por otra parte, las propiedades topológicas incluyen el contenido de información calculado a partir de características del grafo molecular, los descriptores de Morse y autocorrelación ponderados con volumen, masa, electronegatividad y polarizabilidad. Así mismo, descriptores relacionados con la forma y el grado de los vértices que incluyen el contenido de ramificaciones, atributos de forma, número de anillos, el índice de Pogliani y los índices moleculares

de Schultz y Gutman. Otros descriptores topológicos incluyen el número de caminos y el estado electrotpológico del tipo de átomo, el índice extendido del átomo topoquímico (ETA) que considera el tamaño, la forma, la ramificación y la funcionalidad del grafo molecular. Estas semejanzas pueden estar asociadas a la presencia de grupos voluminosos y ramificados en las estructuras de F y R, y a los dobles enlaces en los extremos de las cadenas laterales.

Adicionalmente, el par F-R tiene un amplio número de características cuánticas que no fueron halladas en las otras mutaciones trabajadas, estas incluyen el volumen molar parcial y la polarizabilidad calculada a partir de métodos: semiempíricos MP2, la blandura local, la teoría funcional de la densidad usando la cadena lateral y polarizabilidad para la estructura de Zwitterion.

- Leucina (L) por Glutamina (Q)

26 descriptores forman el conjunto de propiedades semejantes para el cambio de Q por L. Los descriptores fisicoquímicos, muestran como ambos aa tienen la misma probabilidad de pertenecer a estructuras α-hélice, áreas superficiales similares al ser expuestos a un solvente y altos valores de energía libre en regiones ß-plegada. Algunas propiedades no están relacionadas en el conjunto de descriptores equivalentes, sin embargo, también dan una idea de la trascendencia de la mutación. La cadena lateral de Leucina por ejemplo está formada por una serie de enlaces de carbonos que convierten al aa en no polar, mientras Glutamina tiene en el extremo de su cadena lateral un grupo amida, los átomos electronegativos en este aa confirman su polaridad; entonces al ser reemplazado L por Q hay un cambio importante en la hidrofobicidad de la proteína. Así mismo, las características estructurales de L y Q se asocian con ocho descriptores topológicos de Morse que combinan la estructura tridimensional de los aa basada en la difracción de electrones, midiendo la distancia euclidiana entre sus átomos y ponderándolos a partir de una propiedad atómica como: masa, volumen de Van der Waals, electronegatividad de Sanderson y polarizabilidad, las dos últimas propiedades se relacionan dentro de las distancias 4 a 8 $Å^{-1}$; otras propiedades topológicas semejantes fueron descritas anteriormente para N-carbamilisa con el cambio L por Q.

- Valina (V) por Lisina (K)

Se detectaron 52 descriptores que representan las propiedades semejantes entre el par V-K, tres de estas fisicoquímicas y 49 topológicas. Las tres primeras incluyen la composición de los aa, el parámetro de Zimm-Bragg y la escala de hidropatía

basada en valores propios de los aa; con excepción de esta última propiedad, calculada a partir del 50% de accesibilidad de los residuos en solvente, el número de descriptores relacionados con hidrofobicidad que no consideran al par de aa semejante es enorme. Dentro de los descriptores topológicos, la variación eléctrica negativa MAXDN encuentra semejanza entre ambos aa, esto tiene sentido dado que ninguno de los aa tiene cargas negativas dentro de su estructura. La carga eléctrica tampoco resultó una propiedad semejante, ya que Valina es considerado un aa alifático sin carga, mientras Lisina es un aa polar con carga positiva, dada su capacidad para aceptar un hidrógeno en el grupo amino. Cuando la cadena lateral de K no está ionizada la estructura secundaria que predomina es la α-hélice, de lo contrario hará parte de ß-plegada. Otros descriptores topológicos son los tridimensionales de Morse y los de autocorrelación ponderados con electronegatividad, polarizabilidad o volumen de Van der Waals. Los índices de conectividad como chi path cluster y Balaban, los valores propios modificados de carga, la matriz de Barisz, el estado electrotopológico del tipo de átomo y el átomo topoquímico extendido son propiedades semejantes para dicha mutación. Al hacer el cambio de V por K propiedades como el volumen y la polarizabilidad que no hacen parte de los descriptores semejantes, sufren un aumento relacionado con impedimentos estéricos, que pueden generar rigidez en la estructura proteica.

- Isoleucina (I) por Lisina (K)

I y K comparten 47 descriptores que incluyen el volumen de los residuos de aa y la tendencia a pertenecer en estructuras proteicas tipo α-hélice, siempre y cuando K no esté ionizado. Por otro lado, los descriptores topológicos involucran contenidos de información de enlaces, número de caminos moleculares, descriptores de Morse y autocorrelación pesados a través de volumen de Van der Waals, polarizabilidad y electronegatividad; la matriz de Barisz y los valores propios modificados de carga. De las cinco mutaciones producidas en luciferasa, la mutación I-K fue la única que encontró analogías a través de un descriptor cuántico; el Cα obtenido desde la forma de zwitterion de estos aa fue calculado usando métodos mecano cuánticos semiempíricos (PM3), es posible relacionar este descriptor con la tendencia a formar estructuras α-hélice.

En 2006 Law *et al.* usando mutación semialeatoria sobre la enzima luciferasa desde *Phonitus pylaris* lograron sustituir cinco aa, mejorando los cambios de pH que podrían afectar la bioluminiscencia y la resistencia a altas temperaturas que provocarían la desnaturalización de la proteína. Cabe agregar que los cambios realizados en esta enzima no afectaron su función y que las posiciones escogidas habían sido estudiadas en trabajos previos como susceptibles a mutaciones. Los

autores escogieron aa cargados e hidrofílicos para la mutación por considerar que estos ofrecerían las propiedades que deseaban para luciferasa. No obstante, en este estudio se mostró qué otras propiedades de los aa están involucradas en las variaciones ocurridas en la enzima, además de corroborar la importancia de la hidrofobicidad y la carga en los aa.

Se puede considerar que las cinco mutaciones comparten algunas propiedades, entre ellas la tendencia a formar estructuras α-hélice; en este caso se hace necesario que el aa mutado muestre semejanza en la formación de estructuras secundarias específicas, dado que un cambio brusco en la estructura de la enzima podría afectar negativamente la actividad biológica de luciferasa. Otros descriptores comunes son las propiedades topológicas de Morse y autocorrelación que ofrecen información de la estructura molecular y están relacionados con características fisicoquímicas como electronegatividad, polarizabilidad y volumen de van der Waals. De igual manera, se puede considerar que la mayoría de mutantes estudiados tenían formas, estructuras y volúmenes semejantes que permitieron que la proteína se mantuviera estable. Para F-R, L-Q e I-K, se presentaron una variedad de descriptores cuánticos asociados a la polarizabilidad y la densidad de carga, dadas las estructuras resonantes entre F-R y volúmenes semejantes entre L-Q e I-K que ofrecen una idea de los impedimentos estéricos en la proteína.

Es interesante resaltar que no estuvieron dentro del grupo de descriptores análogos, propiedades como hidrofilia y carga eléctrica. Como consecuencia de la sustitución de aa hidrofóbicos y sin carga por aa hidrofílicos y cargados, se produjo una mayor afinidad con el solvente y un aumento en la carga neta de la enzima, esto permitió que a pH bajos los iones H^+ no afectaran la envoltura acuosa de la proteína y las interacciones electrostáticas entre aa, evitando su desnaturalización. En este mismo sentido, altas temperaturas también son causa de desnaturalización debido a que aumenta la energía cinética de las moléculas y desorganiza la envoltura acuosa, es probable que al mejorar la afinidad con el solvente a causa de los grupos cargados y polares se haya generado mayor resistencia a los cambios de temperatura en la proteína mutada.

4.3.3 PI3K

Como ya se ha citado, PI3K es una proteína de gran interés oncológico que sufre una única mutación en su secuencia que afecta el comportamiento normal de las células. Con ayuda de la base de datos COSMIC, se seleccionaron las tres mutaciones puntuales no sinónimas más frecuentes, asociadas a cuatro subhistologías de cáncer de seno con base en el estado somático confirmado de la enfermedad y el impacto patogénico de la mutación. En la figura 6 se observa la distribución de las mutaciones puntuales no sinónimas en 14 muestras de

carcinoma Basal triple negativo, 30 muestras de carcinoma ER-PR positivo, 15 muestras de carcinoma HER positivo y 152 muestras de carcinoma Ductal. En todas las subhistologías la mutación H1047R resultó ser la más frecuente, seguido de E45K y H1047L.

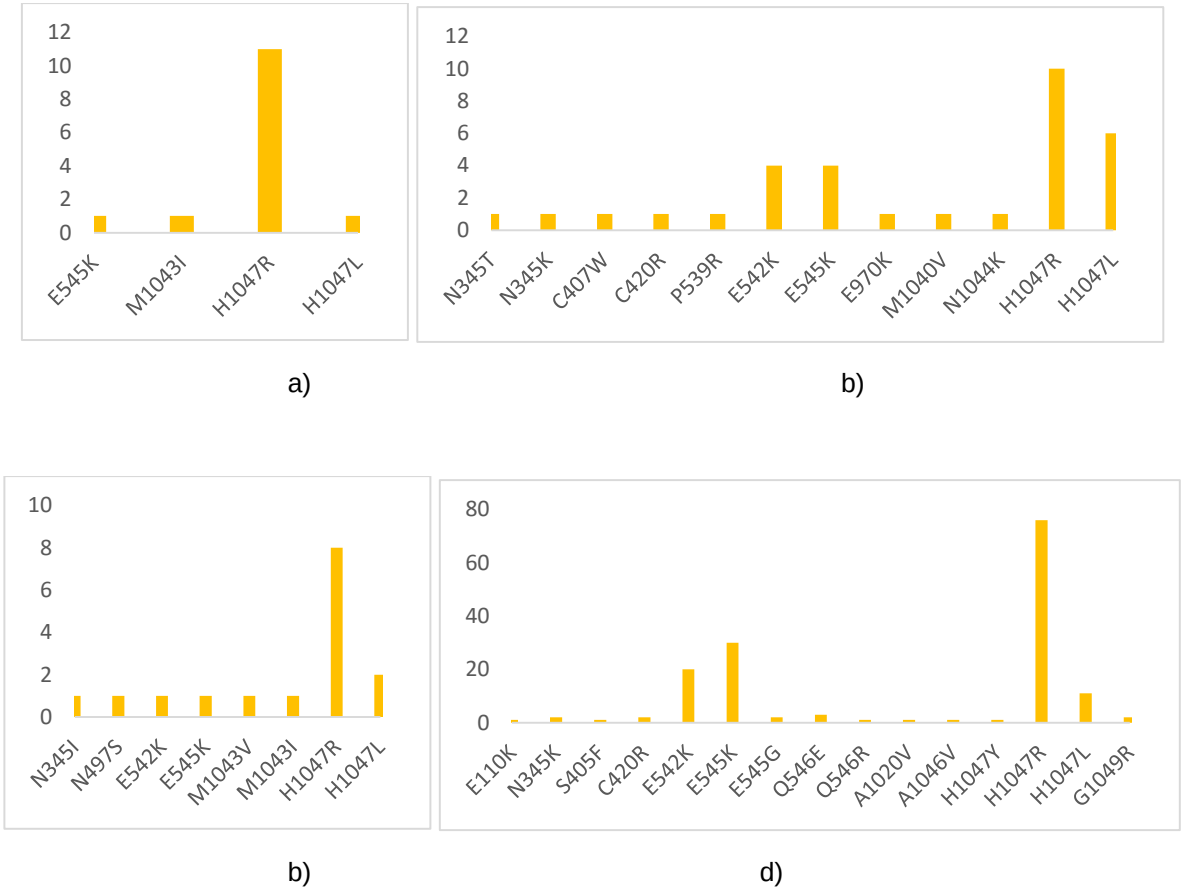

Figura 6: Mutaciones en PI3K para a) carcinoma Basal triple negativo, b) carcinoma ER-PR positivo, c) carcinoma HER positivo y d) carcinoma ductal.

Los mutantes E545K, H1047R y H1047L se analizaron con base en la colección de alfabetos reducidos y sus descriptores. En la figura 7 se muestra la distribución de las frecuencias para cada cambio de aa en los 1.729 descriptores; al comparar estas distribuciones con las encontradas para las mutaciones de N-carbamilasa y

luciferasa se observa el mismo comportamiento, la mayor parte de los descriptores no demuestra relaciones de semejanza entre el par de aa, no obstante el pequeño porcentaje que si lo hace es objeto de análisis en esta investigación.

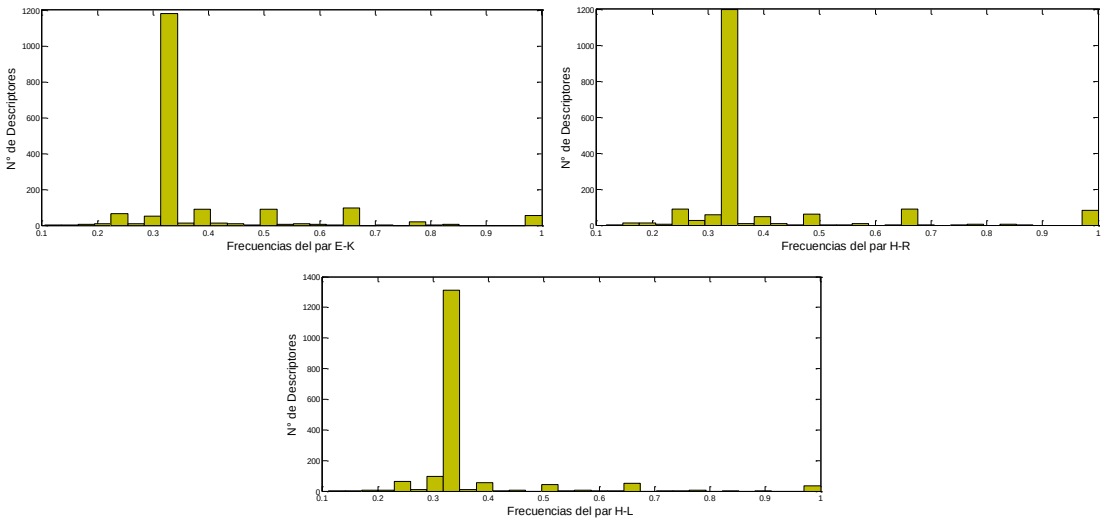

Figura 7: Distribución de las frecuencias en 1.729 descriptores para las tres mutaciones más frecuentes en la proteína PI3K.

Como se ha discutido en párrafos anteriores los descriptores examinados corresponden a aquellos con frecuencia=1. Un análisis de semejanza entre los pares de aa por propiedad fue llevado a cabo.

- Ácido glutámico (E) por Lisina (K)

54 descriptores resultaron semejantes entre E y K. Los más destacados dentro de las bases de datos incluyen frecuencia en giros, mutabilidad relativa, hidrofilicidad y RF (tasa de migración) medida a través de la interacción del soluto y el solvente. Los descriptores topológicos de dos dimensiones involucran los perfiles moleculares basados en el índice de Randic que se derivan de los momentos de distribución de distancia de la matriz geométrica y los descriptores de Morse, que como ya se ha explicado combinan la estructura tridimensional de los aa con ponderaciones a partir de masa y volumen de Van der Waals. Otro descriptor molecular es la matriz de Barisz. Finalmente dos descriptores cuánticos también resultaron análogos; la polarizabilidad de carga a través de la blandura local de los aa optimizada con el método semiempírico MP2 y las propiedades electrónicas para el átomo de carbono relacionada con la población de Mulliken.

- Histidina (H) por Arginina (R)

81 descriptores resultaron semejantes para este par de aa, entre ellos, frecuencias en estructuras secundarias tipo giros y ß-plegada. Otros descriptores fisicoquímicos fueron: energía libre, pK, hidrofobicidad y polaridad. Adicionalmente resultaron semejantes descriptores topológicos como Morse y autocorrelación ponderados con masa atómica, volumen de Van der Waals, electronegatividad y polarizabilidad, matrices ponderadas de Barisz y *detour* que tiene en cuenta el camino más largo del grafo para llegar de un vértice a otro. Otros descriptores que relacionan la forma del grafo y la carga son el estado electrotopológico del tipo de átomo, la carga topológica, el número de anillos, el contenido de información, el índice de flexibilidad de Kier y de Harary relacionado con la conectividad y el tamaño molecular; el índice extendido del átomo topoquímico (ETA) que considera el tamaño, la forma, la ramificación y la funcionalidad del grafo molecular y los valores propios modificados de carga. También se hicieron presentes dos descriptores cuánticos que se relacionan con la población de Mulliken para el átomo de carbono, calculada a partir de los métodos PM3 cuyo objetivo es caracterizar las propiedades electrónicas de los aa y el índice de separación de las cargas correspondientes (CSI_R) que está relacionado con la hidrofobicidad y la carga.

- Histidina (H) por Leucina (L)

33 descriptores que muestran semejanza entre el par de aa están basados en el volumen de los residuos y la frecuencia de los aa en estructuras α-hélice. Dentro de los descriptores topológicos se incluyen la matriz de Braisz, Morse y autocorrelación pesados con electronegatividad y polarizabilidad.

Es importante destacar que las mutaciones estudiadas no producen pérdida funcional o estructural de PI3K, sino que generan una sobreactivación de su subunidad catalítica P110α, aumentando los elementos de señalización.

El dominio helicoidal de la subunidad catalítica P110α, ubicado en las posiciones 517 a 694, interactúa con los dominios SH2 de la subunidad reguladora P85. En la posición 545 ocurre la mutación de E por K, es probable que la carga negativa inicial del ácido glutámico interactuara con las cargas positivas presentes en los dominios SH2, produciendo una atracción entre ambas proteínas diméricas que permitiera inhibir la función catalítica de P110α. No obstante, el cambio por un aa con carga positiva pudo producir una repulsión entre ambas partes que disminuyó la interacción entre los dominios y afectó la regulación de la actividad catalítica. Lo interesante aquí es que la función no se pierde, sino que la proteína se sobreactiva.

Una de las posibles razones a nivel biológico para que la proteína no perdiera su actividad puede ser explicada a través de los descriptores semejantes entre el par E-K, que como se ha descrito anteriormente poseen similitud frente a la presencia en estructuras secundarias, la hidrofílicidad que les permite mantener la interacción con el solvente y la masa molecular; a nivel topológico tienen formas, estructuras, enlaces y propiedades electrónicas semejantes.

Por otro lado los mutantes H-R y H-L ubicados en la posición 1047 de la secuencia hacen parte del dominio quinasa, encargado de la fosforilación de los lípidos. La unión a lípidos involucra una cantidad de interacciones electrostáticas e hidrofóbicas. Se ha sugerido que el cambio por un aa más positivo facilita las interacciones electrostáticas con los grupos fosfatos de los fosfolípidos, sin embargo el comportamiento de la actividad catalítica resulta semejante al cambiar H por L, este último no es un aa cargado. Hon *et al.* mostraron que a pesar de las propiedades químicas y estructurales diferentes de R y L, ambos cambios exhiben la misma medida de unión hidrofóbica y electrostática que activa los fosfopéptidos y es superior a la enzima silvestre. En efecto, según lo expuesto anteriormente, la pareja H-R comparte un mayor número de propiedades semejantes que la pareja H-L. La primera pareja de mutantes no sólo está relacionada con la carga positiva, también lo hacen con descriptores como hidrofobicidad, polaridad, topológicos de forma con ponderaciones y cuánticos de polarizabilidad y carga. El aumento de la actividad catalítica de la quinasa puede relacionarse con un leve aumento en la polaridad y la carga positiva de la Arginina, ya que el grupo fosfato que se va a transferir al inositol es de gran tamaño y tiene una alta densidad de carga negativa que puede verse disminuida con el aumento de la carga positiva. No obstante se observó un aumento en el volumen y las ramificaciones de la cadena lateral, propiedades que no hacen parte del conjunto de descriptores semejantes. Estas características parecieran no cumplirse para el par H y L que sí comparten propiedades de forma y tamaño como el volumen y las aristas que consideran ramificaciones y enlaces, mientras que descriptores como hidrofobicidad, carga y polaridad no resultaron semejantes. Esto llevaría a pensar que cuando no hay semejanza relacionada con la electronegatividad o carga eléctrica, los descriptores topológicos que incluyen la forma de la molécula y el volumen tienen un rol importante en la hiperactividad de la quinasa.

CAPÍTULO 5

CONCLUSIONES

Las clasificaciones de aa realizadas en este trabajo usando HCA y considerando *ties in proximity* representan la colección más grande y completa de alfabetos reducidos reportada hasta este momento, en la que además se involucra la mayor cantidad de propiedades fisicoquímicas, bioquímicas, topológicas y cuánticas de los aa.

Teniendo en cuenta que el número de dendrogramas posibles para un sólo descriptor es una cifra enorme ($8{,}2008 \times 10^{21}$), resulta interesante que el 72% de los descriptores tuviera entre uno y cuatro dendrogramas. Esto da una idea de la robustez de los agrupamientos de aa para la mayor parte de las propiedades estudiadas. El porcentaje de descriptores restantes presenta una gran cantidad de dendrogramas producto de la naturaleza de los datos que incluyen valores semejantes para los aa.

La colección de alfabetos resultó ser útil para analizar los cambios de aa en las secuencias de las enzimas: N-carbamilasa, luciferasa y PI3K. Las frecuencias asociadas a las mutaciones y sus descriptores mostraron un mismo comportamiento, en ellas, la mayor parte de las propiedades diferenciaban los aa y sólo un pequeño porcentaje mostró semejanza. No obstante los descriptores semejantes deben considerarse como propiedades invariantes y fundamentales en la estructura y la función de la proteína.

Las posiciones 58, 262 y 184 en la enzima N-carbamilasa, aumentaron la resistencia a oxidarse y la estabilidad térmica de manera conjunta, probablemente por un leve aumento en la hidrofobicidad y disminución de la carga eléctrica de los aa mutados. Sin embargo, la semejanza estructural, relacionada con la forma y el volumen de los aa silvestre y mutado, permitieron que la estructura secundaria no se desestabilizara. El cambio Q23L, también favoreció la estabilidad térmica y oxidativa de N-carbamilasa; esto puede asociarse a un aumento en la hidrofobicidad y a la semejanza entre propiedades como la tendencia a pertenecer a estructuras secundarias α-hélice, volúmenes medianos y la no presencia de carga eléctrica.

Por otro lado, los cambios de G75S y V40A en N-carbamilasa, disminuyeron la termoestabilidad de la enzima y aumentaron su resistencia a oxidarse. En la posición 75 hubo un crecimiento considerable de la estabilidad oxidativa, probablemente por la presencia del grupo hidroxilo en la cadena lateral del aa mutado que aumentó la polaridad del medio; puesto que el oxígeno es un átomo con alta electronegatividad y afinidad de enlace, en un medio oxidante este no

cederá sus electrones. El aumento en la polaridad causó una disminución en las interacciones hidrofóbicas que desestabilizaron la estructura a medida que la temperatura del medio aumentaba. La presencia de A en la posición 40, no tuvo un efecto fuerte en la hidrofobicidad, ni en su carga eléctrica, no obstante se observó una disminución del tamaño de la cadena lateral que puede asociarse a la resistencia a oxidarse.

Las mutaciones F14R, L35Q, V182K, I232K y F465R en luciferasa encontraron semejanza en descriptores topológicos relacionados con la forma y el volumen, que al ser semejantes evitaron desestabilizar la proteína; otras propiedades cuánticas relacionadas con polarizabilidad y densidad de carga, resaltan la importancia de estructuras resonantes en la posición 14 y 465 de la secuencia, además de tamaños semejantes en el resto de las posiciones mutadas. Descriptores como hidrofilia y carga eléctrica positiva no hicieron parte de las propiedades con frecuencia de uno, pero son importantes para explicar cómo el cambio de aa hidrofóbicos y sin carga por aa hidrofílicos y cargados mejoraron en la proteína mutada la resistencia a pH bajos y la estabilidad térmica.

Para la proteína PI3K que muestra una hiperactividad en la subunidad catalítica p110α, el cambio E-K está asociado a descriptores semejantes como hidrofilia, masa, forma y a propiedades disímiles como volumen, polaridad y carga eléctrica que produjeron efectos sobre la subunidad reguladora evitando el control sobre la actividad catalítica. En el cambio H-R, el aumento de la actividad catalítica en el dominio quinasa de p110α puede relacionarse con un leve aumento en la polaridad y la carga positiva de la R, mientras en los mutantes H-L, parece deberse a propiedades como forma y tamaño, además del aumento de interacciones hidrofóbicas.

Finalmente es importante resaltar que una mutación puntual puede ser representativa no sólo por las propiedades semejantes o diferentes del par de aa, sino por la posición en que ocurra dicho cambio.

RECOMENDACIONES

Construir una base de datos con interfaz gráfico y acceso libre para aquellos investigadores que deseen hacer uso de la colección de alfabetos reducidos.

Actualizar los descriptores periódicamente y considerar el trabajo realizado por Xiao *et al.*[36] para comparar los descriptores calculados allí con los usados en este trabajo, en caso de que no esté presente alguno, deberá incluirse en la colección de alfabetos reducidos.

Debido a que en este trabajo los descriptores no semejantes entre el par de aa (silvestre y mutado) fueron brevemente considerados y dada su importancia para explicar los cambios de las propiedades de N-carbamilasa, luciferasa y PI3K, resulta fundamental realizar un análisis de aquellos descriptores con frecuencias inferiores a 1.

Teniendo en cuenta la gran utilidad de los alfabetos reducidos, sería interesante usar la colección obtenida en este trabajo en campos como: alineamiento de proteínas, desarrollo de péptidos de aplicación farmacológica y predicción del plegamiento proteico.

Ampliar el estudio de mutaciones puntuales no sinónimas del gen PIK3CA o de otras proteínas de interés industrial o médico.

BIBLIOGRAFÍA

1 UniProtKB/Swiss-Prot protein knowledgebase release 2015_08 statistics [http://web.expasy.org/docs/relnotes/relstat.html].

2 POLANSKI, A y KIMMEL, M. Bioinformatics. Berlin: Springer-Verlag, 2007. 1-3p.

3 HENIKOFF, S y HENIKOFF, J . Amino acid substitution matrices from protein blocks. En: Proc. Natl Acad. Sci. USA. No. 89 (1992), p.10915–10919.

4 TODESCHINI, R y CONSONNI, V. Handbook of molecular descriptor for chemoinformatics. Primera Edición. Weinheim. Wiley-VCH Verlag GmbH, 2002.

5 aaindex. List of 544 Amino Acid Indices ver.9.1 [En línea] <http://www.genome.jp/aa.index/AA.index/list_of_indices> [Agosto de 2015].

6 UniProtKB/Swiss-Prot. Protein knowledgebase release statistics [En línea] <http://web.expasy.org/docs/relnotes/relstat.html> [Agosto de 2015]

7 NELSON, D y COX, M. Lehninger Principles of Biochemistry. Cuarta Edición. Hardcover: Ediciones W. H. Freeman Company, 2004. 78 p.

8 BRANDEN, C y TOOZE, J. Introduction to Protein Structure. Primera Edición. New York and London: Ediciones Garland Publishing inc, 1991, 5 p.

9 AHNERT, S., MARSH, J., HERNÁNDEZ, H. y ROBINSON, S. Teichmann. Principles of assembly reveal a periodic table of protein complexes. En: Science No. 6266(2015).

10 SANGER, F y THOMPSON, E. The amino acid sequence in the glycyl chain of insulin. En: Biochem J. No. 52 (1952); p. 3.

11 LEVINTHAL, C. Are there pathways for protein folding?. En: Journal de Chimie Physique et de Physico-Chimie Biologique. No.65 (1968), p. 44–45.

12 ANFINSEN, C. Principles that govern the folding of protein chains. En: Science. No. 181 (1973), p. 223-230.

13 ROOMAN, M., KOCHER, J., WODAK, S. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary information. En: Biochemistry. No. 31 (1992), p. 10226-10238.

14 MURPHY, L, WALLQVIST, A, LEVY, R. Simplified amino acid alphabets for protein fold recognition and implications for folding. En: Protein Eng. No. 13 (2000), p. 149-152.

15 KISTER, A. Amino Acid distribution rules predict protein fold: protein grammar for beta-strand sandwich-like structures. En: Biomolecules. No. 5 (2015), p. 41-59

16 RAMACHANDRAN, G, RAMAKRISHNAN, C, SAISEKHARAN, V. Stereochemistry of polypeptide chain configurations. En: Journal of Molecular Biology. No. 7 (1963), p. 95–99.

17 SNEATH, P: Relations between chemical structure and biological activity in peptides. En: J Theoret Biol. No. 12 (1966), p. 157-195.

18 KYTE, J y DOOLITTLE, R. A Simple method for displaying the hydropathic character of a protein. En: J. Mol. Biol. No. 157 (1982), p. 105-132.

19 CÁRDENAS, C., OBREGÓN, M., LLANOS, E., MACHADO, E., BOHÓRQUEZ, H., VILLAVECES, J., PATARROYO, M. Constructing a useful tool for characterizing amino acid conformers by means of quantum chemical and graph theory indices. En: Comput Chem. No- 6 (2002), p. 667-682.

20 ETCHEBEST, C. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. En: Eur Biophys J. No. 36 (2007), p. 1059–1069.

21 SUYU, M. y WANG, F. Amino acid classification based spectrum kernel fusion for protein subnuclear localization. En: BMC Bioinformatics. No. 11 (2010), p. 1-8.

22 KAMTEKAR, S., SCHIFFER, J., XIONG, H., BAHIK, J. y HECHT, M. Protein design by binary patterning of polar and nonpolar amino acids. En: Science. No. 262 (1993); p. 1680-1685.

23 KARLIN, S. y GHANDOUR, G. Multiple alphabet amino acid sequence comparison of the Ig-kappa chain constant domain. En: Proc.Natl Acad.Sci. No. 87 (1985), p. 8597-8601.

24 SAGOT, M., VIARI, A. y SOLDANO, H. Multiple sequence comparison a peptide matching approach. En: Theor.Comp.Sci. No. 180 (1997), p. 115-137

25 CANNATA N, TOPPO, S, ROMUALDI, C Y VALLE, G. Simplifying amino acids alphabet by means of a branch and bound algorithm and substitution matrices. En: Bioinformatics. No.18 (2002), p. 1102-1108.

26 LOOSE, C., JENSEN, K., RIGOUTSOS, I. y STEPHANOPOULOS, G. A linguistic model for the rational design of antimicrobial peptides. En: nature Letter. No. 19 (2006); p. 867-869.

27 SUSKO, E. y ROGER, A. On Reduced Amino Acid Alphabets for Phylogenetic Inference. En: Mol. Biol. Evol. No. 24 (2007), p. 2139–2150.

28 BACARDIT, J., STOUT, M., HIRST, J., VALENCIA, A., SMITH, R. y KRASNOGOR, N. Automated Alphabet Reduction for Protein Datasets. En: BMC Bioinformatics. No.10 (2009), p.1-16.

29 PETERSON, E., KONDEV, P., THERIOT, J. y PHILLIPS, R. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. En: Bioinformatics. No. 11 (2009), p. 1356–1362

30 HUANG, J., WANG, T., HUANG, S. y Li, X. Reduced alphabet for protein prediction. En: Proteins. No. 83 (2015), p. 631-639.

31 RESTREPO, G.; HARRÉ, R. Mereology of quantitative structure-activity relationships models. En: HYLE – Int. J. Phil. Chem. No. 1 (2015), p.19-38.

32 GIULIO, M. A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. En: Gene. No.346 (2005), p. 1–6.

33 YAP, C. PaDEL-descriptor an open source software to calculate molecular descriptors and fingerprints. En: J Comput Chem. No. 32 (2011). p.1466-1474.

34 The Official Gaussian Website. [En línea] <http://www.gaussian.com/index.htm> [Agosto de 2015].

35 XIAO, N., CAO, D.S., ZHU, M.F. y XU, Q.S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. En: Bioinformatics. No. 31 (2015). p. 1857-1859.

36 CODESSAPRO. Topological Descriptors. [En línea] <http://www.codessa-pro.com/descriptors/topo/index.htm> [Agosto de 2015].

37 Molecular Structure Generation. [En línea] <http://www.molgen.de/> [Agosto de 2015].

38 TETKO, I., GASTEIGER J., TODESCHINI, R., MAURI, A., LIVINGSTONE, D., ERTL, P., PALYULIN, V., RADCHENKO, E., ZEFIROV, N., MAKARENKO, A., TANCHUK, V. y PROKOPENKO, V. Virtual computational chemistry laboratory - design and description. En: J. Comput. Aid. Mol. Des. No. 19 (2005), p. 453-463.

39 SHANNON, C. y WEAVER, W. The mathematical theory of communication. Urbana. Edición: University of Illinois press, 1968. 8-17p.

40 QUINTERO, N.; COHEN, I.; RESTREPO, G. Chemotopological study of positron emitters radionuclides used in PET diagnostic imaging: physical, physico-chemical, dosimetric, quantum and nuclear properties. En: J. Radioanal. Nucl. Ch. No. 295 (2013), p. 823-833.

41 ALDENDERFER, M. y BLASHFIELD, R. Cluster analysis. Series: Quantitative application in social sciences. Primera Edición. London. SAGE publication, 1984, 7-74 p.

42 MACCUISH, J., NICOLAOU, C. y MACCUISH, N.E. Ties in proximity and clustering compounds. En: J. Chem. Inf. Comput. Sci. No.41 (2001), p. 134-146.

43 ZHOU, X. y MAO, K. The ties problem resulting from counting-based error estimators and its impact on gene selection algorithms. En: Bioinformatics. No. 22 (2006), p. 2507-2515.

44 ARNAU, V., MARS, S. y MARÍN, I. Iterative Cluster Analysis of Protein Interaction Data. En: Bioinformatics. No. 21 (2005), p. 364-378.

45 LEAL, W., LLANOS, E., RESTREPO, G., SUAREZ, C. y. PATARROYO, M. How frequently do clusters occur in hierarchical clustering analysis? A graph theoretical approach to studying ties in proximity. En: J Cheminform. No. 8 (2016), p. 1-16.

46 OH, K.H., NAM, S.H. y KIM, H.S. Improvement of Oxidative and Thermostability of N-Carbamyl-DAmino Acid Amidohydrolase by Directed Evolution. En: Protein Engineering. No. 15 (2002), p. 689-695.

47 LAW, E., GANDELMAN, O., TISI, L., LOWE, C. y MURRAY, J. Mutagenesis of solvent-exposed amino acids in Photinus pyralis luciferase improves thermostability and pH-tolerance. En: Biochem J. No. 397 (2006), p. 305-312.

48 ETCHEBEST, C., BENROS, E.C., BORNOT E.A. CAMPROUX, BREVERN G.A. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. En: Eur Biophys J. No. 36 (2007), p. 1059–1069.

49 HART, J., ZHANGB, Y., LIAOB, L., UENOA, l., DUA, L., JONKERSA, M., YATES, L. y VOGTA, P. The butterfly effect in cancer: A single base mutation can remodel the cell. En: PNAS. No. 112 (2014), p. 1131-1136.

50 Catalogue of Somatic Mutation in Cancer [En línea] <http://cancer.sanger.ac.uk/cosmic> [Enero de 2016].

51 RODRÍGUEZ, S., GARCÍA, A., LAS HERAS, F., CLEMENTE, J., RODRÍGUEZ, F., GARCÍA, J., LORIS, R. y GAVIRA, J. Mutational and Structural Analysis of L-N-Carbamoylase Reveals New Insights into a Peptidase M20/M25/M40 Family Member. En: J. Bacteriol. No. 21 (2012), p. 5759-5768.

52 Protein Data Bank: N-carbamoyl-D-amino acid hydrolase P60327 (DCAS_AGRSK) [En línea] <http://www.rcsb.org/pdb/protein/P60327?evtc=Suggest&evta=UniProtAccession&evtl=autosearch_SearchBar_querySuggest> [Febrero 2016]

53 GIGUÈRE, V. Application of the firefly luciferase reporter gene. En: Methods Mol Biol. No. 7 (1991), p. 237-241.

54 TORRE, L., SIEGEL, R. y JEMAL, A. Global Cancer: Facts & Figures. Tercera Edición. American Cancer Society, 2015.

55 BASELGA, J. Targeting the Phosphoinositide-3 (PI3) Kinase Pathway in Breast Cancer. En: The Oncologist. No. 16 (2011), p. 12-19.

56 FELSENSTEIN, J. The Number of Evolutionary Trees. En: Syst. Zool. No. 27 (1978), p.27-33.

ANEXOS

Tabla I: 1729 descriptores de aminoácidos seleccionado con ayuda del índice del contenido de información.

Descriptores de las bases de datos AA index y Expasy	
Identidad	**Nombre del descriptor**
1	ANDN920101 alpha-CH chemical shifts (Andersen et al., 1992)
2	ARGP820101 Hydrophobicity index (Argos et al., 1982)
3	ARGP820102 Signal sequence helical potential (Argos et al., 1982)
4	ARGP820103 Membrane-buried preference parameters (Argos et al., 1982)
5	BEGF750101 Conformational parameter of inner helix (Beghin-Dirkx, 1975)
6	BEGF750102 Conformational parameter of beta-structure (Beghin-Dirkx, 1975)
7	BEGF750103 Conformational parameter of beta-turn (Beghin-Dirkx, 1975)
8	BIGC670101 Residue volume (Bigelow, 1967)
9	BIOV880101 Information value for accessibility; average fraction 35% (Biou et al., 1988)
10	BIOV880102 Information value for accessibility; average fraction 23% (Biou et al., 1988)
11	BULH740102 Apparent partial specific volume (Bull-Breese, 1974)
12	BUNA790101 alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
13	BUNA790102 alpha-CH chemical shifts (Bundi-Wuthrich, 1979)
14	BUNA790103 Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
15	BURA740101 Normalized frequency of alpha-helix (Burgess et al., 1974)
16	BURA740102 Normalized frequency of extended structure (Burgess et al., 1974)
17	CHAM810101 Steric parameter (Charton, 1981)
18	CHAM820101 Polarizability parameter (Charton-Charton, 1982)
19	CHAM820102 Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
20	CHAM830101 The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)
21	CHAM830102 A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of B-sheet (Charton-Charton, 1983)
22	CHAM830106 The number of bonds in the longest chain (Charton-Charton, 1983)
23	CHOC750101 Average volume of buried residue (Chothia, 1975)

24	CHOC760101 Residue accessible surface area in tripeptide (Chothia, 1976)
25	CHOC760102 Residue accessible surface area in folded protein (Chothia, 1976)
26	CHOC760104 Proportion of residues 100% buried (Chothia, 1976)
27	CHOP780101 Normalized frequency of beta-turn (Chou-Fasman, 1978a)
28	CHOP780201 Normalized frequency of alpha-helix (Chou-Fasman, 1978b)
29	CHOP780202 Normalized frequency of beta-sheet (Chou-Fasman, 1978b)
30	CHOP780203 Normalized frequency of beta-turn (Chou-Fasman, 1978b)
31	CHOP780204 Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)
32	CHOP780205 Normalized frequency of C-terminal helix (Chou-Fasman, 1978b)
33	CHOP780206 Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
34	CHOP780207 Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)
35	CHOP780208 Normalized frequency of N-terminal beta-sheet (Chou-Fasman, 1978b)
36	CHOP780209 Normalized frequency of C-terminal beta-sheet (Chou-Fasman, 1978b)
37	CHOP780210 Normalized frequency of N-terminal non beta region (Chou-Fasman, 1978b)
38	CHOP780211 Normalized frequency of C-terminal non beta region (Chou-Fasman, 1978b)
39	CHOP780212 Frequency of the 1st residue in turn (Chou-Fasman, 1978b)
40	CHOP780213 Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)
41	CHOP780214 Frequency of the 3rd residue in turn (Chou-Fasman, 1978b)
42	CHOP780215 Frequency of the 4th residue in turn (Chou-Fasman, 1978b)
43	CHOP780216 Normalized frequency of the 2nd and 3rd residues in turn (Chou-Fasman, 1978b)
44	CIDH920101 Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)
45	CIDH920102 Normalized hydrophobicity scales for beta-proteins (Cid et al., 1992)
46	CIDH920103 Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)
47	CIDH920104 Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)
48	CIDH920105 Normalized average hydrophobicity scales (Cid et al., 1992)
49	COHE430101 Partial specific volume (Cohn-Edsall, 1943)
50	CRAJ730101 Normalized frequency of middle helix (Crawford et al., 1973)
51	CRAJ730102 Normalized frequency of beta-sheet (Crawford et al., 1973)
52	CRAJ730103 Normalized frequency of turn (Crawford et al., 1973)
53	DAWD720101 Size (Dawson, 1972)
54	DAYM780101 Amino acid composition (Dayhoff et al., 1978a)

55	DESM900101 Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)
56	DESM900102 Average membrane preference: AMP07 (Degli Esposti et al., 1990)
57	EISD860101 Solvation free energy (Eisenberg-McLachlan, 1986)
58	EISD860102 Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
59	EISD860103 Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)
60	FASG760101 Molecular weight (Fasman, 1976)
61	FASG760102 Melting point (Fasman, 1976)
62	FASG760103 Optical rotation (Fasman, 1976)
63	FASG760104 pK-N (Fasman, 1976)
64	FASG760105 pK-C (Fasman, 1976)
65	FAUJ880101 Graph shape index (Fauchere et al., 1988)
66	FAUJ880102 Smoothed upsilon steric parameter (Fauchere et al., 1988)
67	FAUJ880103 Normalized van der Waals volume (Fauchere et al., 1988)
68	FAUJ880104 STERIMOL length of the side chain (Fauchere et al., 1988)
69	FAUJ880106 STERIMOL maximum width of the side chain (Fauchere et al., 1988)
70	FAUJ880107 N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)
71	FAUJ880108 Localized electrical effect (Fauchere et al., 1988)
72	FAUJ880113 pK-a(RCOOH) (Fauchere et al., 1988)
73	FINA770101 Helix-coil equilibrium constant (Finkelstein-Ptitsyn, 1977)
74	GARJ730101 Partition coefficient (Garel et al., 1973)
75	GEIM800101 Alpha-helix indices (Geisow-Roberts, 1980)
76	GEIM800102 Alpha-helix indices for alpha-proteins (Geisow-Roberts, 1980)
77	GEIM800103 Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
78	GEIM800104 Alpha-helix indices for alpha/beta-proteins (Geisow-Roberts, 1980)
79	GEIM800105 Beta-strand indices (Geisow-Roberts, 1980)
80	GEIM800106 Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)
81	GEIM800107 Beta-strand indices for alpha/beta-proteins (Geisow-Roberts, 1980)
82	GEIM800108 Aperiodic indices (Geisow-Roberts, 1980)
83	GEIM800109 Aperiodic indices for alpha-proteins (Geisow-Roberts, 1980)
84	GEIM800110 Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)
85	GEIM800111 Aperiodic indices for alpha/beta-proteins (Geisow-Roberts, 1980)
86	GOLD730101 Hydrophobicity factor (Goldsack-Chalifoux, 1973)
87	GOLD730102 Residue volume (Goldsack-Chalifoux, 1973)
88	GRAR740101 Composition (Grantham, 1974)
89	GRAR740103 Volume (Grantham, 1974)
90	GUYH850101 Partition energy (Guy, 1985)

91	HOPA770101 Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)
92	HUTJ700101 Heat capacity (Hutchens, 1970)
93	HUTJ700102 Absolute entropy (Hutchens, 1970)
94	HUTJ700103 Entropy of formation (Hutchens, 1970)
95	ISOY800101 Normalized relative frequency of alpha-helix (Isogai et al., 1980)
96	ISOY800102 Normalized relative frequency of extended structure (Isogai et al., 1980)
97	ISOY800103 Normalized relative frequency of bend (Isogai et al., 1980)
98	ISOY800104 Normalized relative frequency of bend R (Isogai et al., 1980)
99	ISOY800105 Normalized relative frequency of bend S (Isogai et al., 1980)
100	ISOY800106 Normalized relative frequency of helix end (Isogai et al., 1980)
101	ISOY800107 Normalized relative frequency of double bend (Isogai et al., 1980)
102	ISOY800108 Normalized relative frequency of coil (Isogai et al., 1980)
103	JANJ780101 Average accessible surface area (Janin et al., 1978)
104	JANJ780102 Percentage of buried residues (Janin et al., 1978)
105	JANJ780103 Percentage of exposed residues (Janin et al., 1978)
106	JOND750101 Hydrophobicity (Jones, 1975)
107	JOND750102 pK (-COOH) (Jones, 1975)
108	JOND920101 Relative frequency of occurrence (Jones et al., 1992)
109	JOND920102 Relative mutability (Jones et al., 1992)
110	JUKT750101 Amino acid distribution (Jukes et al., 1975)
111	JUNJ780101 Sequence frequency (Jungck, 1978)
112	KANM800101 Average relative probability of helix (Kanehisa-Tsong, 1980)
113	KANM800102 Average relative probability of beta-sheet (Kanehisa-Tsong, 1980)
114	KANM800103 Average relative probability of inner helix (Kanehisa-Tsong, 1980)
115	KANM800104 Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)
116	KARP850101 Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)
117	KARP850102 Flexibility parameter for one rigid neighbor (Karplus-Schulz, 1985)
118	KARP850103 Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)
119	KHAG800101 The Kerr-constant increments (Khanarian-Moore, 1980)
120	KRIW710101 Side chain interaction parameter (Krigbaum-Rubin, 1971)
121	KRIW790101 Side chain interaction parameter (Krigbaum-Komoriya, 1979)
122	KRIW790102 Fraction of site occupied by water (Krigbaum-Komoriya, 1979)
123	KRIW790103 Side chain volume (Krigbaum-Komoriya, 1979)
124	LAWE840101 Transfer free energy, CHP/water (Lawson et al., 1984)

125	LEVM760101 Hydrophobic parameter (Levitt, 1976)
126	LEVM760102 Distance between C-alpha and centroid of side chain (Levitt, 1976)
127	LEVM760103 Side chain angle theta(AAR) (Levitt, 1976)
128	LEVM760104 Side chain torsion angle phi(AAAR) (Levitt, 1976)
129	LEVM760105 Radius of gyration of side chain (Levitt, 1976)
130	LEVM760106 van der Waals parameter R0 (Levitt, 1976)
131	LEVM760107 van der Waals parameter epsilon (Levitt, 1976)
132	LEVM780101 Normalized frequency of alpha-helix, with weights (Levitt, 1978)
133	LEVM780102 Normalized frequency of beta-sheet, with weights (Levitt, 1978)
134	LEVM780103 Normalized frequency of reverse turn, with weights (Levitt, 1978)
135	LEVM780104 Normalized frequency of alpha-helix, unweighted (Levitt, 1978)
136	LEVM780105 Normalized frequency of beta-sheet, unweighted (Levitt, 1978)
137	LEVM780106 Normalized frequency of reverse turn, unweighted (Levitt, 1978)
138	LEWP710101 Frequency of occurrence in beta-bends (Lewis et al., 1971)
139	MAXF760101 Normalized frequency of alpha-helix (Maxfield-Scheraga, 1976)
140	MAXF760102 Normalized frequency of extended structure (Maxfield-Scheraga, 1976)
141	MAXF760103 Normalized frequency of zeta R (Maxfield-Scheraga, 1976)
142	MAXF760104 Normalized frequency of left-handed alpha-helix (Maxfield-Scheraga, 1976)
143	MAXF760105 Normalized frequency of zeta L (Maxfield-Scheraga, 1976)
144	MAXF760106 Normalized frequency of alpha region (Maxfield-Scheraga, 1976)
145	MEEJ810101 Retention coefficient in NaClO4 (Meek-Rossetti, 1981)
146	MEEJ810102 Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)
147	MEIH800101 Average reduced distance for C-alpha (Meirovitch et al., 1980)
148	MEIH800102 Average reduced distance for side chain (Meirovitch et al., 1980)
149	MEIH800103 Average side chain orientation angle (Meirovitch et al., 1980)
150	MIYS850101 Effective partition energy (Miyazawa-Jernigan, 1985)
151	NAGK730101 Normalized frequency of alpha-helix (Nagano, 1973)
152	NAGK730102 Normalized frequency of bata-structure (Nagano, 1973)
153	NAGK730103 Normalized frequency of coil (Nagano, 1973)
154	NAKH900101 AA composition of total proteins (Nakashima et al., 1990)
155	NAKH900102 SD of AA composition of total proteins (Nakashima et al., 1990)
156	NAKH900103 AA composition of mt-proteins (Nakashima et al., 1990)
157	NAKH900104 Normalized composition of mt-proteins (Nakashima et al., 1990)
158	NAKH900105 AA composition of mt-proteins from animal (Nakashima et al., 1990)
159	NAKH900106 Normalized composition from animal (Nakashima et al., 1990)
160	NAKH900107 AA composition of mt-proteins from fungi and plant (Nakashima et al., 1990)

161	NAKH900108 Normalized composition from fungi and plant (Nakashima et al., 1990)
162	NAKH900109 AA composition of membrane proteins (Nakashima et al., 1990)
163	NAKH900110 Normalized composition of membrane proteins (Nakashima et al., 1990)
164	NAKH900111 Transmembrane regions of non-mt-proteins (Nakashima et al., 1990)
165	NAKH900112 Transmembrane regions of mt-proteins (Nakashima et al., 1990)
166	NAKH900113 Ratio of average and computed composition (Nakashima et al., 1990)
167	NAKH920101 AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)
168	NAKH920102 AA composition of CYT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)
169	NAKH920103 AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa, 1992)
170	NAKH920104 AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)
171	NAKH920105 AA composition of MEM of single-spanning proteins (Nakashima-Nishikawa, 1992)
172	NAKH920106 AA composition of CYT of multi-spanning proteins (Nakashima-Nishikawa, 1992)
173	NAKH920107 AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992)
174	NAKH920108 AA composition of MEM of multi-spanning proteins (Nakashima-Nishikawa, 1992)
175	NISK800101 8 A contact number (Nishikawa-Ooi, 1980)
176	NISK860101 14 A contact number (Nishikawa-Ooi, 1986)
177	OOBM770101 Average non-bonded energy per atom (Oobatake-Ooi, 1977)
178	OOBM770102 Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
179	OOBM770103 Long range non-bonded energy per atom (Oobatake-Ooi, 1977)
180	OOBM770104 Average non-bonded energy per residue (Oobatake-Ooi, 1977)
181	OOBM770105 Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)
182	OOBM850101 Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)
183	OOBM850102 Optimized propensity to form reverse turn (Oobatake et al., 1985)
184	OOBM850103 Optimized transfer energy parameter (Oobatake et al., 1985)
185	OOBM850104 Optimized average non-bonded energy per atom (Oobatake et al., 1985)
186	OOBM850105 Optimized side chain interaction parameter (Oobatake et al., 1985)
187	PALJ810101 Normalized frequency of alpha-helix from LG (Palau et al., 1981)

188	PALJ810102 Normalized frequency of alpha-helix from CF (Palau et al., 1981)
189	PALJ810103 Normalized frequency of beta-sheet from LG (Palau et al., 1981)
190	PALJ810104 Normalized frequency of beta-sheet from CF (Palau et al., 1981)
191	PALJ810105 Normalized frequency of turn from LG (Palau et al., 1981)
192	PALJ810106 Normalized frequency of turn from CF (Palau et al., 1981)
193	PALJ810107 Normalized frequency of alpha-helix in all-alpha class (Palau et al., 1981)
194	PALJ810108 Normalized frequency of alpha-helix in alpha+beta class (Palau et al., 1981)
195	PALJ810109 Normalized frequency of alpha-helix in alpha/beta class (Palau et al., 1981)
196	PALJ810110 Normalized frequency of beta-sheet in all-beta class (Palau et al., 1981)
197	PALJ810111 Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981)
198	PALJ810112 Normalized frequency of beta-sheet in alpha/beta class (Palau et al., 1981)
199	PALJ810113 Normalized frequency of turn in all-alpha class (Palau et al., 1981)
200	PALJ810114 Normalized frequency of turn in all-beta class (Palau et al., 1981)
201	PALJ810115 Normalized frequency of turn in alpha+beta class (Palau et al., 1981)
202	PALJ810116 Normalized frequency of turn in alpha/beta class (Palau et al., 1981)
203	PARJ860101 HPLC parameter (Parker et al., 1986)
204	PLIV810101 Partition coefficient (Pliska et al., 1981)
205	PONP800101 Surrounding hydrophobicity in folded form (Ponnuswamy et al., 1980)
206	PONP800102 Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)
207	PONP800103 Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al., 1980)
208	PONP800104 Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)
209	PONP800105 Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al., 1980)
210	PONP800106 Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980)
211	PONP800107 Accessibility reduction ratio (Ponnuswamy et al., 1980)
212	PONP800108 Average number of surrounding residues (Ponnuswamy et al., 1980)
213	PRAM820101 Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
214	PRAM820102 Slope in regression analysis x 1.0E1 (Prabhakaran-Ponnuswamy, 1982)
215	PRAM820103 Correlation coefficient in regression analysis (Prabhakaran-Ponnuswamy, 1982)

216	PRAM900101 Hydrophobicity (Prabhakaran, 1990)
217	PRAM900102 Relative frequency in alpha-helix (Prabhakaran, 1990)
218	PRAM900103 Relative frequency in beta-sheet (Prabhakaran, 1990)
219	PRAM900104 Relative frequency in reverse-turn (Prabhakaran, 1990)
220	PTIO830101 Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)
221	PTIO830102 Beta-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)
222	QIAN880101 Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)
223	QIAN880102 Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)
224	QIAN880103 Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)
225	QIAN880104 Weights for alpha-helix at the window position of -3 (Qian-Sejnowski, 1988)
226	QIAN880105 Weights for alpha-helix at the window position of -2 (Qian-Sejnowski, 1988)
227	QIAN880106 Weights for alpha-helix at the window position of -1 (Qian-Sejnowski, 1988)
228	QIAN880107 Weights for alpha-helix at the window position of 0 (Qian-Sejnowski, 1988)
229	QIAN880108 Weights for alpha-helix at the window position of 1 (Qian-Sejnowski, 1988)
230	QIAN880109 Weights for alpha-helix at the window position of 2 (Qian-Sejnowski, 1988)
231	QIAN880110 Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)
232	QIAN880111 Weights for alpha-helix at the window position of 4 (Qian-Sejnowski, 1988)
233	QIAN880112 Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)
234	QIAN880113 Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)
235	QIAN880114 Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)
236	QIAN880115 Weights for beta-sheet at the window position of -5 (Qian-Sejnowski, 1988)
237	QIAN880116 Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)
238	QIAN880117 Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)
239	QIAN880118 Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)
240	QIAN880119 Weights for beta-sheet at the window position of -1 (Qian-Sejnowski, 1988)
241	QIAN880120 Weights for beta-sheet at the window position of 0 (Qian-Sejnowski, 1988)

242	QIAN880121 Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)
243	QIAN880122 Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
244	QIAN880123 Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)
245	QIAN880124 Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
246	QIAN880125 Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)
247	QIAN880126 Weights for beta-sheet at the window position of 6 (Qian-Sejnowski, 1988)
248	QIAN880127 Weights for coil at the window position of -6 (Qian-Sejnowski, 1988)
249	QIAN880128 Weights for coil at the window position of -5 (Qian-Sejnowski, 1988)
250	QIAN880129 Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
251	QIAN880130 Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)
252	QIAN880131 Weights for coil at the window position of -2 (Qian-Sejnowski, 1988)
253	QIAN880132 Weights for coil at the window position of -1 (Qian-Sejnowski, 1988)
254	QIAN880133 Weights for coil at the window position of 0 (Qian-Sejnowski, 1988)
255	QIAN880134 Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)
256	QIAN880135 Weights for coil at the window position of 2 (Qian-Sejnowski, 1988)
257	QIAN880136 Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)
258	QIAN880137 Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
259	QIAN880138 Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)
260	QIAN880139 Weights for coil at the window position of 6 (Qian-Sejnowski, 1988)
261	RACS770101 Average reduced distance for C-alpha (Rackovsky-Scheraga, 1977)
262	RACS770102 Average reduced distance for side chain (Rackovsky-Scheraga, 1977)
263	RACS770103 Side chain orientational preference (Rackovsky-Scheraga, 1977)
264	RACS820101 Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga, 1982)
265	RACS820102 Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga, 1982)

266	RACS820103 Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)
267	RACS820104 Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)
268	RACS820105 Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)
269	RACS820106 Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)
270	RACS820107 Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)
271	RACS820108 Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga, 1982)
272	RACS820110 Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)
273	RACS820111 Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)
274	RACS820112 Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
275	RACS820113 Value of theta(i) (Rackovsky-Scheraga, 1982)
276	RACS820114 Value of theta(i-1) (Rackovsky-Scheraga, 1982)
277	RADA880101 Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)
278	RADA880102 Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988)
279	RADA880103 Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)
280	RADA880104 Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
281	RADA880105 Transfer free energy from vap to oct (Radzicka-Wolfenden, 1988)
282	RADA880106 Accessible surface area (Radzicka-Wolfenden, 1988)
283	RADA880107 Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)
284	RADA880108 Mean polarity (Radzicka-Wolfenden, 1988)
285	RICJ880101 Relative preference value at N" (Richardson-Richardson, 1988)
286	RICJ880102 Relative preference value at N' (Richardson-Richardson, 1988)
287	RICJ880103 Relative preference value at N-cap (Richardson-Richardson, 1988)
288	RICJ880104 Relative preference value at N1 (Richardson-Richardson, 1988)
289	RICJ880105 Relative preference value at N2 (Richardson-Richardson, 1988)
290	RICJ880106 Relative preference value at N3 (Richardson-Richardson, 1988)
291	RICJ880107 Relative preference value at N4 (Richardson-Richardson, 1988)
292	RICJ880108 Relative preference value at N5 (Richardson-Richardson, 1988)
293	RICJ880109 Relative preference value at Mid (Richardson-Richardson, 1988)
294	RICJ880110 Relative preference value at C5 (Richardson-Richardson, 1988)

295	RICJ880111 Relative preference value at C4 (Richardson-Richardson, 1988)
296	RICJ880112 Relative preference value at C3 (Richardson-Richardson, 1988)
297	RICJ880113 Relative preference value at C2 (Richardson-Richardson, 1988)
298	RICJ880114 Relative preference value at C1 (Richardson-Richardson, 1988)
299	RICJ880115 Relative preference value at C-cap (Richardson-Richardson, 1988)
300	RICJ880116 Relative preference value at C' (Richardson-Richardson, 1988)
301	RICJ880117 Relative preference value at C" (Richardson-Richardson, 1988)
302	ROBB760101 Information measure for alpha-helix (Robson-Suzuki, 1976)
303	ROBB760102 Information measure for N-terminal helix (Robson-Suzuki, 1976)
304	ROBB760103 Information measure for middle helix (Robson-Suzuki, 1976)
305	ROBB760104 Information measure for C-terminal helix (Robson-Suzuki, 1976)
306	ROBB760105 Information measure for extended (Robson-Suzuki, 1976)
307	ROBB760106 Information measure for pleated-sheet (Robson-Suzuki, 1976)
308	ROBB760107 Information measure for extended without H-bond (Robson-Suzuki, 1976)
309	ROBB760108 Information measure for turn (Robson-Suzuki, 1976)
310	ROBB760109 Information measure for N-terminal turn (Robson-Suzuki, 1976)
311	ROBB760110 Information measure for middle turn (Robson-Suzuki, 1976)
312	ROBB760111 Information measure for C-terminal turn (Robson-Suzuki, 1976)
313	ROBB760112 Information measure for coil (Robson-Suzuki, 1976)
314	ROBB760113 Information measure for loop (Robson-Suzuki, 1976)
315	ROBB790101 Hydration free energy (Robson-Osguthorpe, 1979)
316	ROSG850101 Mean area buried on transfer (Rose et al., 1985)
317	ROSM880101 Side chain hydropathy, uncorrected for solvation (Roseman, 1988)
318	ROSM880102 Side chain hydropathy, corrected for solvation (Roseman, 1988)
319	ROSM880103 Loss of Side chain hydropathy by helix formation (Roseman, 1988)
320	SIMZ760101 Transfer free energy (Simon, 1976), Cited by Charton-Charton (1982)
321	SNEP660101 Principal component I (Sneath, 1966)
322	SNEP660102 Principal component II (Sneath, 1966)
323	SNEP660103 Principal component III (Sneath, 1966)
324	SNEP660104 Principal component IV (Sneath, 1966)
325	SUEM840101 Zimm-Bragg parameter s at 20 C (Sueki et al., 1984)
326	SUEM840102 Zimm-Bragg parameter sigma x 1.0E4 (Sueki et al., 1984)
327	SWER830101 Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)
328	TANS770101 Normalized frequency of alpha-helix (Tanaka-Scheraga, 1977)
329	TANS770102 Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)
330	TANS770103 Normalized frequency of extended structure (Tanaka-Scheraga, 1977)

331	TANS770104 Normalized frequency of chain reversal R (Tanaka-Scheraga, 1977)
332	TANS770105 Normalized frequency of chain reversal S (Tanaka-Scheraga, 1977)
333	TANS770106 Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)
334	TANS770107 Normalized frequency of left-handed helix (Tanaka-Scheraga, 1977)
335	TANS770108 Normalized frequency of zeta R (Tanaka-Scheraga, 1977)
336	TANS770109 Normalized frequency of coil (Tanaka-Scheraga, 1977)
337	TANS770110 Normalized frequency of chain reversal (Tanaka-Scheraga, 1977)
338	VASM830101 Relative population of conformational state A (Vasquez et al., 1983)
339	VASM830102 Relative population of conformational state C (Vasquez et al., 1983)
340	VASM830103 Relative population of conformational state E (Vasquez et al., 1983)
341	VELV850101 Electron-ion interaction potential (Veljkovic et al., 1985)
342	VHEG790101 Transfer free energy to lipophilic phase (von Heijne-Blomberg, 1979)
343	WARP780101 Average interactions per side chain atom (Warme-Morgan, 1978)
344	WEBA780101 RF value in high salt chromatography (Weber-Lacey, 1978)
345	WERD780101 Propensity to be buried inside (Wertz-Scheraga, 1978)
346	WERD780102 Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga, 1978)
347	WERD780103 Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
348	WERD780104 Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978)
349	WOEC730101 Polar requirement (Woese, 1973)
350	WOLS870101 Principal property value z1 (Wold et al., 1987)
351	WOLS870102 Principal property value z2 (Wold et al., 1987)
352	WOLS870103 Principal property value z3 (Wold et al., 1987)
353	YUTK870101 Unfolding Gibbs energy in water, pH7.0 (Yutani et al., 1987)
354	YUTK870102 Unfolding Gibbs energy in water, pH9.0 (Yutani et al., 1987)
355	YUTK870103 Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)
356	YUTK870104 Activation Gibbs energy of unfolding, pH9.0 (Yutani et al., 1987)
357	ZASB820101 Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982)
358	ZIMJ680101 Hydrophobicity (Zimmerman et al., 1968)
359	ZIMJ680104 Isoelectric point (Zimmerman et al., 1968)
360	ZIMJ680105 RF rank (Zimmerman et al., 1968)

361	AURR980101 Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998)
362	AURR980102 Normalized positional residue frequency at helix termini N''' (Aurora-Rose, 1998)
363	AURR980103 Normalized positional residue frequency at helix termini N'' (Aurora-Rose, 1998)
364	AURR980104 Normalized positional residue frequency at helix termini N'(Aurora-Rose, 1998)
365	AURR980105 Normalized positional residue frequency at helix termini Nc (Aurora-Rose, 1998)
366	AURR980106 Normalized positional residue frequency at helix termini N1 (Aurora-Rose, 1998)
367	AURR980107 Normalized positional residue frequency at helix termini N2 (Aurora-Rose, 1998)
368	AURR980108 Normalized positional residue frequency at helix termini N3 (Aurora-Rose, 1998)
369	AURR980109 Normalized positional residue frequency at helix termini N4 (Aurora-Rose, 1998)
370	AURR980110 Normalized positional residue frequency at helix termini N5 (Aurora-Rose, 1998)
371	AURR980111 Normalized positional residue frequency at helix termini C5 (Aurora-Rose, 1998)
372	AURR980112 Normalized positional residue frequency at helix termini C4 (Aurora-Rose, 1998)
373	AURR980113 Normalized positional residue frequency at helix termini C3 (Aurora-Rose, 1998)
374	AURR980114 Normalized positional residue frequency at helix termini C2 (Aurora-Rose, 1998)
375	AURR980115 Normalized positional residue frequency at helix termini C1 (Aurora-Rose, 1998)
376	AURR980116 Normalized positional residue frequency at helix termini Cc (Aurora-Rose, 1998)
377	AURR980117 Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)
378	AURR980118 Normalized positional residue frequency at helix termini C'' (Aurora-Rose, 1998)
379	AURR980119 Normalized positional residue frequency at helix termini C''' (Aurora-Rose, 1998)
380	AURR980120 Normalized positional residue frequency at helix termini C4' (Aurora-Rose, 1998)
381	ONEK900101 Delta G values for the peptides extrapolated to 0 M urea (O'Neil-DeGrado, 1990)
382	ONEK900102 Helix formation parameters (delta delta G) (O'Neil-DeGrado, 1990)
383	VINM940101 Normalized flexibility parameters (B-values), average (Vihinen et al., 1994)

384	VINM940102 Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours (Vihinen et al., 1994)
385	VINM940103 Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours (Vihinen et al., 1994)
386	VINM940104 Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)
387	MUNV940101 Free energy in alpha-helical conformation (Munoz-Serrano, 1994)
388	MUNV940102 Free energy in alpha-helical region (Munoz-Serrano, 1994)
389	MUNV940103 Free energy in beta-strand conformation (Munoz-Serrano, 1994)
390	MUNV940104 Free energy in beta-strand region (Munoz-Serrano, 1994)
391	MUNV940105 Free energy in beta-strand region (Munoz-Serrano, 1994)
392	WIMW960101 Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)
393	KIMC930101 Thermodynamic beta sheet propensity (Kim-Berg, 1993)
394	MONM990101 Turn propensity scale for transmembrane helices (Monne et al., 1999)
395	BLAM930101 Alpha helix propensity of position 44 in T4 lysozyme (Blaber et al., 1993)
396	PARS000101 p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000)
397	PARS000102 p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000)
398	KUMS000101 Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)
399	KUMS000102 Distribution of amino acid residues in the 18 non-redundant families of mesophilic proteins (Kumar et al., 2000)
400	KUMS000103 Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al., 2000)
401	KUMS000104 Distribution of amino acid residues in the alpha-helices in mesophilic proteins (Kumar et al., 2000)
402	TAKK010101 Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
403	FODM020101 Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)
404	NADH010101 Hydropathy scale based on self-information values in the two-state model (5% accessibility) (Naderi-Manesh et al., 2001)
405	NADH010102 Hydropathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)
406	NADH010103 Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)
407	NADH010104 Hydropathy scale based on self-information values in the two-state model (20% accessibility) (Naderi-Manesh et al., 2001)
408	NADH010105 Hydropathy scale based on self-information values in the two-state model (25% accessibility) (Naderi-Manesh et al., 2001)

409	NADH010106 Hydropathy scale based on self-information values in the two-state model (36% accessibility) (Naderi-Manesh et al., 2001)
410	NADH010107 Hydropathy scale based on self-information values in the two-state model (50% accessibility) (Naderi-Manesh et al., 2001)
411	MONM990201 Averaged turn propensities in a transmembrane helix (Monne et al., 1999)
412	KOEP990101 Alpha-helix propensity derived from designed sequences (Koehl-Levitt, 1999)
413	KOEP990102 Beta-sheet propensity derived from designed sequences (Koehl-Levitt, 1999)
414	CEDJ970101 Composition of amino acids in extracellular proteins (percent) (Cedano et al., 1997)
415	CEDJ970102 Composition of amino acids in anchored proteins (percent) (Cedano et al., 1997)
416	CEDJ970103 Composition of amino acids in membrane proteins (percent) (Cedano et al., 1997)
417	CEDJ970104 Composition of amino acids in intracellular proteins (percent) (Cedano et al., 1997)
418	CEDJ970105 Composition of amino acids in nuclear proteins (percent) (Cedano et al., 1997)
419	FUKS010101 Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)
420	FUKS010102 Surface composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
421	FUKS010103 Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
422	FUKS010104 Surface composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)
423	FUKS010105 Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)
424	FUKS010106 Interior composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
425	FUKS010107 Interior composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
426	FUKS010108 Interior composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)
427	FUKS010109 Entire chain composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)
428	FUKS010110 Entire chain composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
429	FUKS010111 Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
430	FUKS010112 Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)
431	TSAJ990101 Volumes including the crystallographic waters using the ProtOr (Tsai et al., 1999)

432	TSAJ990102 Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)
433	COSI940101 Electron-ion interaction potential values (Cosic, 1994)
434	PONP930101 Hydrophobicity scales (Ponnuswamy, 1993)
435	WILM950101 Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
436	WILM950102 Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
437	WILM950103 Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
438	WILM950104 Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce et al. 1995)
439	KUHL950101 Hydrophilicity scale (Kuhn et al., 1995)
440	GUOD860101 Retention coefficient at pH 2 (Guo et al., 1986)
441	JURD980101 Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)
442	BASU050101 Interactivity scale obtained from the contact matrix (Bastolla et al., 2005)
443	BASU050102 Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)
444	BASU050103 Interactivity scale obtained by maximizing the mean of correlation coefficient over pairs of sequences sharing the TIM barrel fold (Bastolla et al., 2005)
445	SUYM030101 Linker propensity index (Suyama-Ohara, 2003)
446	PUNT030101 Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases (Punta-Maritan, 2003)
447	PUNT030102 Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases (Punta-Maritan, 2003)
448	GEOR030101 Linker propensity from all dataset (George-Heringa, 2003)
449	GEOR030102 Linker propensity from 1-linker dataset (George-Heringa, 2003)
450	GEOR030103 Linker propensity from 2-linker dataset (George-Heringa, 2003)
451	GEOR030104 Linker propensity from 3-linker dataset (George-Heringa, 2003)
452	GEOR030105 Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)
453	GEOR030106 Linker propensity from medium dataset (linker length is between six and 14 residues) (George-Heringa, 2003)
454	GEOR030107 Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)
455	GEOR030108 Linker propensity from helical (annotated by DSSP) dataset (George-Heringa, 2003)
456	GEOR030109 Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
457	ZHOH040101 The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004)

458	ZHOH040102 The relative stability scale extracted from mutation experiments (Zhou-Zhou, 2004)
459	ZHOH040103 Buriability (Zhou-Zhou, 2004)
460	BAEK050101 Linker index (Bae et al., 2005)
461	HARY940101 Mean volumes of residues buried in protein interiors (Harpaz et al., 1994)
462	PONJ960101 Average volumes of residues (Pontius et al., 1996)
463	DIGM050101 Hydrostatic pressure asymmetry index, PAI (Di Giulio, 2005)
464	WOLR790101 Hydrophobicity index (Wolfenden et al., 1979)
465	OLSK800101 Average internal preferences (Olsen, 1980)
466	KIDA850101 Hydrophobicity-related index (Kidera et al., 1985)
467	GUYH850102 Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)
468	GUYH850104 Apparent partition energies calculated from Janin index (Guy, 1985)
469	GUYH850105 Apparent partition energies calculated from Chothia index (Guy, 1985)
470	JACR890101 Weights from the IFH scale (Jacobs-White, 1989)
471	BLAS910101 Scaled side chain hydrophobicity values (Black-Mould, 1991)
472	CASG920101 Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)
473	CORJ870101 NNEIG index (Cornette et al., 1987)
474	CORJ870102 SWEIG index (Cornette et al., 1987)
475	CORJ870103 PRIFT index (Cornette et al., 1987)
476	CORJ870104 PRILS index (Cornette et al., 1987)
477	CORJ870105 ALTFT index (Cornette et al., 1987)
478	CORJ870106 ALTLS index (Cornette et al., 1987)
479	CORJ870107 TOTFT index (Cornette et al., 1987)
480	CORJ870108 TOTLS index (Cornette et al., 1987)
481	MIYS990101 Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)
482	MIYS990102 Optimized relative partition energies - method A (Miyazawa-Jernigan, 1999)
483	MIYS990103 Optimized relative partition energies - method B (Miyazawa-Jernigan, 1999)
484	MIYS990104 Optimized relative partition energies - method C (Miyazawa-Jernigan, 1999)
485	MIYS990105 Optimized relative partition energies - method D (Miyazawa-Jernigan, 1999)
486	ENGD860101 Hydrophobicity index (Engelman et al., 1986)
487	FASG890101 Hydrophobicity index (Fasman, 1989)
488	Molecular weight
489	Bulkiness

490	Polarity/Grantham
491	Recognition factors
492	Hphob. OMH / Sweet et al.
493	Hphob. / Kyte & Doolittle
494	Hphob. / Abraham & Leo
495	Hphob. / Bull & Breese
496	Hphob. / Guy
497	Hphob. / Miyazawa et al.
498	Hphob. / Roseman
499	Hphob. / Wolfenden et al.
500	Hphob. HPLC / Wilson & al
501	Hphob. HPLC pH3.4 / Cowan
502	Hphob. / Rf mobility
503	HPLC / TFA retention
504	HPLC / retention pH 2.1
505	Molar fraction (%) of 2001 buried residues
506	Hphob. / Chothia
507	Ratio hetero end/side
508	Average flexibility
509	beta-sheet / Chou & Fasman
510	alpha-helix / Deleage & Roux
511	beta-turn / Deleage & Roux
512	alpha-helix / Levitt
513	beta-turn / Levitt
514	Antiparallel beta-strand
515	A.A. composition
516	Relative mutability
517	Polarity / Zimmerman
518	Refractivity
519	Hphob. / Eisenberg et al.
520	Hphob. / Hopp & Woods
521	Hphob. / Manavalan et al.
522	Hphob. / Black
523	Hphob. / Fauchere et al.
524	Hphob. / Janin
525	Hphob. / Rao & Argos
526	Hphob. / Tanford
527	Hphob. / Welling & al
528	Hphob. HPLC / Parker & al
529	Hphob. HPLC pH7.5 / Cowan

530	HPLC / HFBA retention
531	Transmembrane tendency
532	HPLC / retention pH 7.4
533	% accessible residues
534	Hphob. / Rose & al
535	Average area buried
536	alpha-helix / Chou & Fasman
537	beta-turn / Chou & Fasman
538	beta-sheet / Deleage & Roux
539	Coil / Deleage & Roux
540	beta-sheet / Levitt
541	Total beta-strand
542	Parallel beta-strand
543	A.A. comp. in Swiss-Prot
544	ECI (Electronic Charge Index)
545	ISA (Isotropic Surface Area)
Descriptores topológicos de la literatura	
Identidad	**Nombre del descriptor**
546	IDE
547	IDM
548	IDDE
549	IDDM
550	IDMT
551	IDET
552	IC
553	TIC
554	SIC
555	CIC
556	BIC
557	TECC
558	AECC
559	DECC
560	UNIP
561	CENT
562	VAR
563	MSDI
564	TWC
565	MWC02
566	MWC03
567	MWC04

568	MWC05
569	MWC06
570	JGI1
571	JGT
572	ATS1m
573	ATS2m
574	ATS3m
575	ATS4m
576	ATS5m
577	ATS6m
578	ATS1v
579	ATS2v
580	ATS3v
581	ATS4v
582	ATS5v
583	ATS6v
584	ATS1e
585	ATS2e
586	ATS3e
587	ATS4e
588	ATS5e
589	ATS6e
590	ATS1p
591	ATS2p
592	ATS3p
593	ATS4p
594	ATS5p
595	ATS6p
596	MATS1m
597	MATS2m
598	MATS3m
599	MATS4m
600	MATS5m
601	MATS6m
602	MATS7m
603	MATS1v
604	MATS2v
605	MATS3v
606	MATS4v
607	MATS5v

608	MATS6v
609	MATS7v
610	MATS1e
611	MATS2e
612	MATS3e
613	MATS4e
614	MATS5e
615	MATS6e
616	MATS7e
617	MATS1p
618	MATS2p
619	MATS3p
620	MATS4p
621	MATS5p
622	MATS6p
623	MATS7p
624	GATS1m
625	GATS2m
626	GATS3m
627	GATS4m
628	GATS5m
629	GATS6m
630	GATS1v
631	GATS2v
632	GATS3v
633	GATS4v
634	GATS5v
635	GATS6v
636	GATS1e
637	GATS2e
638	GATS3e
639	GATS4e
640	GATS5e
641	GATS6e
642	GATS1p
643	GATS2p
644	GATS3p
645	GATS4p
646	GATS5p
647	GATS6p

648	DP01
649	DP02
650	DP03
651	DP04
652	DP05
653	DP06
654	DP07
655	DP08
656	DP09
657	DP10
658	DP11
659	DP12
660	DP13
661	DP14
662	DP15
663	SP01
664	SP02
665	SP03
666	SP04
667	SP05
668	SP06
669	SP07
670	SP08
671	SP09
672	SP10
673	SP11
674	SP12
675	SP13
676	SP14
677	SP15
678	W3D
679	AGDD
680	DDI
681	DDDA
682	MAXDN
683	MAXDP
684	DELS
685	RGm
686	PPm
687	ESTP

688	SPAN
689	SPAM
690	PJI3
691	SPH
692	ASP
693	FDI
694	SHP2
695	Mor01u
696	Mor02u
697	Mor03u
698	Mor04u
699	Mor05u
700	Mor06u
701	Mor07u
702	Mor08u
703	Mor09u
704	Mor10u
705	Mor11u
706	Mor12u
707	Mor13u
708	Mor14u
709	Mor15u
710	Mor16u
711	Mor17u
712	Mor18u
713	Mor19u
714	Mor20u
715	Mor21u
716	Mor22u
717	Mor23u
718	Mor24u
719	Mor25u
720	Mor26u
721	Mor27u
722	Mor28u
723	Mor29u
724	Mor30u
725	Mor31u
726	Mor32u
727	Mor01m

728	Mor02m
729	Mor03m
730	Mor04m
731	Mor05m
732	Mor06m
733	Mor07m
734	Mor08m
735	Mor09m
736	Mor10m
737	Mor11m
738	Mor12m
739	Mor13m
740	Mor14m
741	Mor15m
742	Mor16m
743	Mor17m
744	Mor18m
745	Mor19m
746	Mor20m
747	Mor21m
748	Mor22m
749	Mor23m
750	Mor24m
751	Mor25m
752	Mor26m
753	Mor27m
754	Mor28m
755	Mor29m
756	Mor30m
757	Mor31m
758	Mor32m
759	Mor01v
760	Mor02v
761	Mor03v
762	Mor04v
763	Mor05v
764	Mor06v
765	Mor07v
766	Mor08v
767	Mor09v

768	Mor10v
769	Mor11v
770	Mor12v
771	Mor13v
772	Mor14v
773	Mor15v
774	Mor16v
775	Mor17v
776	Mor18v
777	Mor19v
778	Mor20v
779	Mor21v
780	Mor22v
781	Mor23v
782	Mor24v
783	Mor25v
784	Mor26v
785	Mor27v
786	Mor28v
787	Mor29v
788	Mor30v
789	Mor31v
790	Mor32v
791	Mor01e
792	Mor02e
793	Mor03e
794	Mor04e
795	Mor05e
796	Mor06e
797	Mor07e
798	Mor08e
799	Mor09e
800	Mor10e
801	Mor11e
802	Mor12e
803	Mor13e
804	Mor14e
805	Mor15e
806	Mor16e
807	Mor17e

808	Mor18e
809	Mor19e
810	Mor20e
811	Mor21e
812	Mor22e
813	Mor23e
814	Mor24e
815	Mor25e
816	Mor26e
817	Mor27e
818	Mor28e
819	Mor29e
820	Mor30e
821	Mor31e
822	Mor32e
823	Mor01p
824	Mor02p
825	Mor03p
826	Mor04p
827	Mor05p
828	Mor06p
829	Mor07p
830	Mor08p
831	Mor09p
832	Mor10p
833	Mor12p
834	Mor13p
835	Mor14p
836	Mor15p
837	Mor16p
838	Mor17p
839	Mor18p
840	Mor19p
841	Mor20p
842	Mor21p
843	Mor22p
844	Mor23p
845	Mor24p
846	Mor25p
847	Mor26p

848	Mor27p
849	Mor28p
850	Mor29p
851	Mor30p
852	Mor31p
853	Mor32p
854	ITH
855	HC0
856	HC1
857	HIC
858	HGM
859	H1u
860	H2u
861	H3u
862	H4u
863	H5u
864	HTu
865	HATS0u
866	HATS2u
867	HATS3u
868	HATS4u
869	HATS5u
870	HATS6u
871	HATS7u
872	H0m
873	H1m
874	H2m
875	H3m
876	HTm
877	HATS0m
878	HATS2m
879	HATS3m
880	HATS4m
881	HATS5m
882	HATSm
883	H0v
884	H1v
885	H2v
886	H3v
887	H4v

888	HTv
889	HATS5v
890	HATS6v
891	HATS7v
892	HATSv
893	H0e
894	H1e
895	H2e
896	H3e
897	H4e
898	H5e
899	HTe
900	HATS0e
901	HATS1e
902	HATS2e
903	HATS3e
904	HATS4e
905	HATS5e
906	HATS6e
907	HATS7e
908	HATSe
909	H0p
910	H1p
911	H2p
912	H3p
913	H4p
914	H5p
915	HTp
916	HATS3p
917	HATS4p
918	HATS5p
919	HATS6p
920	HATS7p
921	HATSp
922	R1u
923	R2u
924	R3u
925	R4u
926	R5u
927	R6u

928	R7u
929	RTu
930	R1u+
931	R2u+
932	R3u+
933	R4u+
934	R5u+
935	R6u+
936	RTu+
937	R1m
938	R2m
939	R3m
940	R4m
941	R5m
942	R6m
943	RTm
944	R2m+
945	R3m+
946	R4m+
947	R5m+
948	RTm+
949	R1v
950	R2v
951	R3v
952	R4v
953	R5v
954	R6v
955	RTv
956	R1e
957	R2e
958	R3e
959	R4e
960	R5e
961	R6e
962	R7e
963	RTe
964	R1e+
965	R2e+
966	R3e+
967	R4e+

968	R5e+
969	R6e+
970	R7e+
971	RTe+
972	R1p
973	R2p
974	R3p
975	R4p
976	R5p
977	R6p
978	RTp
979	R1p+
980	RTp+
981	Hy
982	Balaban Index Chem3D 8.0 66127.977 162161.719
983	Cluster Count Chem3D 8.0 11.578 4.108
984	Molecular Topological Index Chem3D 8.0 1787.333 2525.753
985	Shape Attribute Chem3D 8.0 9.674 4.081
986	Kier & Hall Index 0 Codessa 2.7 6.659 2.392
987	Kier & Hall Index 1 Codessa 2.7 3.727 1.483
988	Kier & Hall Index 2 Codessa 2.7 2.794 1.392
989	Kier & Hall Index 3 Codessa 2.7 1.700 0.984
990	Kier Shape Index 1 Codessa 2.7 9.687 2.864
991	Kier Shape Index 2 Codessa 2.7 4.208 1.624
992	Kier Shape Index 3 Codessa 2.7 3.557 1.766
993	Kier Flexibility Index Codessa 2.7 3.612 1.544
994	Narumi Simple Topological Index (log) Dragon 5.2 6.383 3.068
995	Narumi Harmonic Topological Indx Dragon 5.2 1.678 1.497
996	Narumi Harmonic Geometric Indx Dragon 5.2 1.712 0.174
997	Total Structure Connectivity Index Dragon 5.2 0.431 0.102
998	Pogliani Index Dragon 5.2 26.485 7.901
999	Log of Product of Row Sum (PRS) Dragon 5.2 40.711 19.821
1000	Average Vertex Distance Degree Dragon 5.2 34.926 20.643
1001	Mean Square Distance Index Dragon 5.2 0.330 0.047
1002	Schultz Molecular Topological Index Dragon 5.2 941.852 970.611
1003	Gutman Molecular Topological Index Dragon 5.2 813.642 972.088
1004	Xu Index Dragon 5.2 11.551 3.795
1005	Superpendentic Index Dragon 5.2 91.663 280.189
1006	Harary H Index
1007	Square Reciprocal Distance Sum Index Dragon 5.2 28.068 14.874

1008	First Mohar Index TI1 Dragon 5.2 −2.991 16.266
1009	Second Mohar Index TI2 Dragon 5.2 2.666 0.913
1010	Hyper-Distance-Path Index Dragon 5.2 693.578 870.837
1011	Detour Index Dragon 5.2 348.444 435.631
1012	Balaban Distance Connectivity Index (J) Dragon 5.2 2.830 0.695
1013	Maximal Electrotopological Negative Variation Dragon 5.2 2.880 0.839
1014	Maximal Electrotopological Positive Variation Dragon 5.2 3.359 0.521
1015	Molecular Electrotopological Variation Dragon 5.2 18.892 8.083
1016	E-state Topological Parameter Dragon 5.2 221.514 1413.845
1017	Balaban Centric Index Dragon 5.2 26.948 15.951
1018	Lopping Centric Index
1019	Rk relative connectivity
1020	Rc Clustering coefficient
1021	Ro relative closeness
1022	Rb relative betweenness
Descriptores topológicos de PaDEL	
Identidad	**Nombre del descriptor**
1023	ALogP
1024	ALogp2
1025	AMR
1026	apol
1027	nAtom
1028	nHeavyAtom
1029	ATS0m
1030	ATS1m
1031	ATS2m
1032	ATS3m
1033	ATS4m
1034	ATS5m
1035	ATS6m
1036	ATS0v
1037	ATS1v
1038	ATS2v
1039	ATS3v
1040	ATS4v
1041	ATS5v
1042	ATS6v
1043	ATS0e
1044	ATS1e
1045	ATS2e

1046	ATS3e
1047	ATS4e
1048	ATS5e
1049	ATS6e
1050	ATS0p
1051	ATS1p
1052	ATS2p
1053	ATS3p
1054	ATS4p
1055	ATS5p
1056	ATS6p
1057	ATS0i
1058	ATS1i
1059	ATS2i
1060	ATS3i
1061	ATS4i
1062	ATS5i
1063	ATS6i
1064	ATS0s
1065	ATS1s
1066	ATS2s
1067	ATS3s
1068	ATS4s
1069	ATS5s
1070	ATS6s
1071	AATS0m
1072	AATS1m
1073	AATS2m
1074	AATS3m
1075	AATS4m
1076	AATS5m
1077	AATS0v
1078	AATS1v
1079	AATS2v
1080	AATS3v
1081	AATS4v
1082	AATS5v
1083	AATS0e
1084	AATS1e
1085	AATS2e

1086	AATS3e
1087	AATS4e
1088	AATS5e
1089	AATS0p
1090	AATS1p
1091	AATS2p
1092	AATS3p
1093	AATS4p
1094	AATS5p
1095	AATS0i
1096	AATS1i
1097	AATS2i
1098	AATS3i
1099	AATS4i
1100	AATS5i
1101	AATS0s
1102	AATS1s
1103	AATS2s
1104	AATS3s
1105	AATS4s
1106	AATS5s
1107	ATSC0c
1108	ATSC1c
1109	ATSC2c
1110	ATSC3c
1111	ATSC4c
1112	ATSC5c
1113	ATSC6c
1114	ATSC0m
1115	ATSC1m
1116	ATSC2m
1117	ATSC3m
1118	ATSC4m
1119	ATSC5m
1120	ATSC6m
1121	ATSC0v
1122	ATSC1v
1123	ATSC2v
1124	ATSC3v
1125	ATSC4v

1126	ATSC5v
1127	ATSC6v
1128	ATSC0e
1129	ATSC1e
1130	ATSC2e
1131	ATSC3e
1132	ATSC4e
1133	ATSC5e
1134	ATSC6e
1135	ATSC0p
1136	ATSC1p
1137	ATSC2p
1138	ATSC3p
1139	ATSC4p
1140	ATSC5p
1141	ATSC6p
1142	ATSC0i
1143	ATSC1i
1144	ATSC2i
1145	ATSC3i
1146	ATSC4i
1147	ATSC5i
1148	ATSC6i
1149	ATSC0s
1150	ATSC1s
1151	ATSC2s
1152	ATSC3s
1153	ATSC4s
1154	ATSC5s
1155	ATSC6s
1156	AATSC0c
1157	AATSC1c
1158	AATSC2c
1159	AATSC3c
1160	AATSC4c
1161	AATSC5c
1162	AATSC6c
1163	AATSC0m
1164	AATSC1m
1165	AATSC2m

1166	AATSC3m
1167	AATSC4m
1168	AATSC5m
1169	AATSC6m
1170	AATSC0v
1171	AATSC1v
1172	AATSC2v
1173	AATSC3v
1174	AATSC4v
1175	AATSC5v
1176	AATSC6v
1177	AATSC0e
1178	AATSC1e
1179	AATSC2e
1180	AATSC3e
1181	AATSC4e
1182	AATSC5e
1183	AATSC6e
1184	AATSC0p
1185	AATSC1p
1186	AATSC2p
1187	AATSC3p
1188	AATSC4p
1189	AATSC5p
1190	AATSC6p
1191	AATSC0i
1192	AATSC1i
1193	AATSC2i
1194	AATSC3i
1195	AATSC4i
1196	AATSC5i
1197	AATSC6i
1198	AATSC0s
1199	AATSC1s
1200	AATSC2s
1201	AATSC3s
1202	AATSC4s
1203	AATSC5s
1204	MATS1c
1205	MATS2c

1206	MATS3c
1207	MATS4c
1208	MATS5c
1209	MATS6c
1210	MATS1m
1211	MATS2m
1212	MATS3m
1213	MATS4m
1214	MATS5m
1215	MATS6m
1216	MATS1v
1217	MATS2v
1218	MATS3v
1219	MATS4v
1220	MATS5v
1221	MATS6v
1222	MATS1e
1223	MATS2e
1224	MATS3e
1225	MATS4e
1226	MATS5e
1227	MATS6e
1228	MATS1p
1229	MATS2p
1230	MATS3p
1231	MATS4p
1232	MATS5p
1233	MATS6p
1234	MATS1i
1235	MATS2i
1236	MATS3i
1237	MATS4i
1238	MATS5i
1239	MATS6i
1240	MATS1s
1241	MATS2s
1242	MATS3s
1243	MATS4s
1244	MATS5s
1245	GATS1c

1246	GATS2c
1247	GATS3c
1248	GATS4c
1249	GATS5c
1250	GATS6c
1251	GATS1m
1252	GATS2m
1253	GATS3m
1254	GATS4m
1255	GATS5m
1256	GATS6m
1257	GATS1v
1258	GATS2v
1259	GATS3v
1260	GATS4v
1261	GATS5v
1262	GATS6v
1263	GATS1e
1264	GATS2e
1265	GATS3e
1266	GATS4e
1267	GATS5e
1268	GATS6e
1269	GATS1p
1270	GATS2p
1271	GATS3p
1272	GATS4p
1273	GATS5p
1274	GATS6p
1275	GATS1i
1276	GATS2i
1277	GATS3i
1278	GATS4i
1279	GATS5i
1280	GATS6i
1281	GATS1s
1282	GATS2s
1283	GATS3s
1284	GATS4s
1285	GATS5s

1286	GATS6s
1287	SpAbs_DzZ
1288	SpMax_DzZ
1289	SpDiam_DzZ
1290	SpAD_DzZ
1291	SpMAD_DzZ
1292	EE_DzZ
1293	VE1_DzZ
1294	VE2_DzZ
1295	VE3_DzZ
1296	VR1_DzZ
1297	VR2_DzZ
1298	VR3_DzZ
1299	SpAbs_Dzm
1300	SpMax_Dzm
1301	SpDiam_Dzm
1302	SpAD_Dzm
1303	SpMAD_Dzm
1304	EE_Dzm
1305	VE1_Dzm
1306	VE2_Dzm
1307	VE3_Dzm
1308	VR1_Dzm
1309	VR2_Dzm
1310	VR3_Dzm
1311	SpAbs_Dzv
1312	SpMax_Dzv
1313	SpDiam_Dzv
1314	SpAD_Dzv
1315	SpMAD_Dzv
1316	EE_Dzv
1317	VE1_Dzv
1318	VE2_Dzv
1319	VE3_Dzv
1320	VR1_Dzv
1321	VR2_Dzv
1322	VR3_Dzv
1323	SpAbs_Dze
1324	SpMax_Dze
1325	SpDiam_Dze

1326	SpAD_Dze
1327	SpMAD_Dze
1328	EE_Dze
1329	VE1_Dze
1330	VE2_Dze
1331	VE3_Dze
1332	VR1_Dze
1333	VR2_Dze
1334	VR3_Dze
1335	SpAbs_Dzp
1336	SpMax_Dzp
1337	SpDiam_Dzp
1338	SpAD_Dzp
1339	SpMAD_Dzp
1340	EE_Dzp
1341	VE1_Dzp
1342	VE2_Dzp
1343	VE3_Dzp
1344	VR1_Dzp
1345	VR2_Dzp
1346	VR3_Dzp
1347	SpAbs_Dzi
1348	SpMax_Dzi
1349	SpDiam_Dzi
1350	SpAD_Dzi
1351	SpMAD_Dzi
1352	EE_Dzi
1353	VE1_Dzi
1354	VE2_Dzi
1355	VE3_Dzi
1356	VR1_Dzi
1357	VR2_Dzi
1358	VR3_Dzi
1359	SpAbs_Dzs
1360	SpMax_Dzs
1361	SpDiam_Dzs
1362	SpAD_Dzs
1363	SpMAD_Dzs
1364	EE_Dzs
1365	SM1_Dzs

1366	VE1_Dzs
1367	VE2_Dzs
1368	VE3_Dzs
1369	VR1_Dzs
1370	VR2_Dzs
1371	VR3_Dzs
1372	BCUTw-1l
1373	BCUTw-1h
1374	BCUTc-1l
1375	BCUTc-1h
1376	BCUTp-1l
1377	BCUTp-1h
1378	nBonds2
1379	nBondsS
1380	nBondsS2
1381	bpol
1382	SpMax1_Bhm
1383	SpMax2_Bhm
1384	SpMax3_Bhm
1385	SpMax4_Bhm
1386	SpMax5_Bhm
1387	SpMax6_Bhm
1388	SpMax7_Bhm
1389	SpMax8_Bhm
1390	SpMin1_Bhm
1391	SpMin2_Bhm
1392	SpMin3_Bhm
1393	SpMin4_Bhm
1394	SpMin5_Bhm
1395	SpMin6_Bhm
1396	SpMax1_Bhv
1397	SpMax2_Bhv
1398	SpMax3_Bhv
1399	SpMax4_Bhv
1400	SpMax5_Bhv
1401	SpMax6_Bhv
1402	SpMax7_Bhv
1403	SpMin1_Bhv
1404	SpMin2_Bhv
1405	SpMin3_Bhv

1406	SpMin4_Bhv
1407	SpMin5_Bhv
1408	SpMin6_Bhv
1409	SpMax1_Bhe
1410	SpMax2_Bhe
1411	SpMax3_Bhe
1412	SpMax4_Bhe
1413	SpMax5_Bhe
1414	SpMax6_Bhe
1415	SpMax7_Bhe
1416	SpMin1_Bhe
1417	SpMin2_Bhe
1418	SpMin3_Bhe
1419	SpMin4_Bhe
1420	SpMin5_Bhe
1421	SpMin6_Bhe
1422	SpMax1_Bhp
1423	SpMax2_Bhp
1424	SpMax3_Bhp
1425	SpMax4_Bhp
1426	SpMax5_Bhp
1427	SpMax6_Bhp
1428	SpMax7_Bhp
1429	SpMin1_Bhp
1430	SpMin2_Bhp
1431	SpMin3_Bhp
1432	SpMin4_Bhp
1433	SpMin5_Bhp
1434	SpMin6_Bhp
1435	SpMax1_Bhi
1436	SpMax2_Bhi
1437	SpMax3_Bhi
1438	SpMax4_Bhi
1439	SpMax5_Bhi
1440	SpMax6_Bhi
1441	SpMax7_Bhi
1442	SpMin1_Bhi
1443	SpMin2_Bhi
1444	SpMin3_Bhi
1445	SpMin4_Bhi

1446	SpMin5_Bhi
1447	SpMin6_Bhi
1448	SpMin7_Bhi
1449	SpMax1_Bhs
1450	SpMax2_Bhs
1451	SpMax3_Bhs
1452	SpMax4_Bhs
1453	SpMax5_Bhs
1454	SpMax6_Bhs
1455	SpMax7_Bhs
1456	SpMax8_Bhs
1457	SpMin1_Bhs
1458	SpMin2_Bhs
1459	SpMin3_Bhs
1460	SpMin4_Bhs
1461	SpMin5_Bhs
1462	SpMin6_Bhs
1463	SpMin7_Bhs
1464	SpMin8_Bhs
1465	VC-3
1466	SPC-4
1467	SPC-5
1468	SPC-6
1469	VPC-4
1470	VPC-5
1471	VPC-6
1472	SP-0
1473	SP-1
1474	SP-2
1475	SP-3
1476	SP-4
1477	SP-5
1478	ASP-0
1479	ASP-1
1480	ASP-2
1481	ASP-3
1482	ASP-4
1483	ASP-5
1484	VP-0
1485	VP-1

1486	VP-2
1487	VP-3
1488	VP-4
1489	VP-5
1490	AVP-0
1491	AVP-1
1492	AVP-2
1493	AVP-3
1494	AVP-4
1495	AVP-5
1496	Sv
1497	Sse
1498	Spe
1499	Sare
1500	Sp
1501	Si
1502	Mv
1503	Mse
1504	Mpe
1505	Mare
1506	Mp
1507	Mi
1508	CrippenLogP
1509	CrippenMR
1510	SpMax_Dt
1511	SpDiam_Dt
1512	SpAD_Dt
1513	SpMAD_Dt
1514	EE_Dt
1515	VE1_Dt
1516	VE2_Dt
1517	VE3_Dt
1518	VR1_Dt
1519	VR2_Dt
1520	VR3_Dt
1521	ECCEN
1522	SHBd
1523	SHBa
1524	SwHBa
1525	SHBint3

1526	SHsNH2
1527	SHdsCH
1528	SHCsats
1529	SHCsatu
1530	SHother
1531	SssCH2
1532	SdsCH
1533	SsssCH
1534	SsNH2
1535	SdO
1536	minHBd
1537	minHBa
1538	minwHBa
1539	minHBint3
1540	minHsNH2
1541	minHdsCH
1542	minHCsats
1543	minHCsatu
1544	minHother
1545	minssCH2
1546	mindsCH
1547	minsssCH
1548	minsNH2
1549	mindO
1550	maxHBd
1551	maxHBa
1552	maxwHBa
1553	maxHBint3
1554	maxHsNH2
1555	maxHdsCH
1556	maxHCsats
1557	maxHCsatu
1558	maxHother
1559	maxssCH2
1560	maxdsCH
1561	maxsNH2
1562	maxdO
1563	sumI
1564	meanI
1565	hmax

1566	gmax
1567	hmin
1568	gmin
1569	LipoaffinityIndex
1570	MAXDN
1571	MAXDP
1572	DELS
1573	MAXDN2
1574	MAXDP2
1575	DELS2
1576	ETA_AlphaP
1577	ETA_dEpsilon_A
1578	ETA_dEpsilon_B
1579	ETA_dEpsilon_C
1580	ETA_dEpsilon_D
1581	ETA_Psi_1
1582	ETA_dPsi_A
1583	ETA_dPsi_B
1584	ETA_Shape_P
1585	ETA_Shape_Y
1586	ETA_Shape_X
1587	ETA_Beta
1588	ETA_BetaP
1589	ETA_Beta_s
1590	ETA_BetaP_s
1591	ETA_Beta_ns
1592	ETA_BetaP_ns
1593	ETA_dBeta
1594	ETA_dBetaP
1595	ETA_Beta_ns_d
1596	ETA_BetaP_ns_d
1597	ETA_Eta
1598	ETA_EtaP
1599	ETA_Eta_R
1600	ETA_Eta_F
1601	ETA_EtaP_F
1602	ETA_Eta_L
1603	ETA_EtaP_L
1604	ETA_Eta_R_L
1605	ETA_Eta_F_L

1606	ETA_EtaP_F_L
1607	ETA_Eta_B
1608	ETA_EtaP_B
1609	ETA_Eta_B_RC
1610	ETA_EtaP_B_RC
1611	FMF
1612	fragC
1613	nHBAcc
1614	nHBAcc2
1615	nHBAcc3
1616	nHBAcc_Lipinski
1617	nHBDon
1618	nHBDon_Lipinski
1619	HybRatio
1620	IC0
1621	IC1
1622	TIC0
1623	TIC5
1624	SIC0
1625	BIC3
1626	BIC4
1627	BIC5
1628	MIC0
1629	MIC1
1630	MIC2
1631	MIC3
1632	MIC4
1633	MIC5
1634	ZMIC5
1635	Kier1
1636	Kier2
1637	Kier3
1638	nAtomLC
1639	MDEC-13
1640	MDEC-14
1641	nT6Ring
1642	nT8Ring
1643	n11HeteroRing
1644	n12HeteroRing
1645	nG12HeteroRing

1646	nFHeteroRing
1647	nF10HeteroRing
1648	nF11HeteroRing
1649	nF12HeteroRing
1650	nFG12HeteroRing
1651	nTHeteroRing
1652	nT4HeteroRing
1653	nT5HeteroRing
1654	nT6HeteroRing
1655	nT7HeteroRing
1656	nT8HeteroRing
1657	nT9HeteroRing
1658	nT10HeteroRing
1659	nT11HeteroRing
1660	nT12HeteroRing
1661	nTG12HeteroRing
1662	nRotB
1663	RotBFrac
1664	nRotBt
1665	RotBtFrac
1666	LipinskiFailures
1667	topoRadius
1668	topoDiameter
1669	topoShape
1670	GGI1
1671	GGI2
1672	GGI5
1673	GGI7
1674	GGI9
1675	JGI1
1676	JGI2
1677	JGI3
1678	JGI4
1679	JGI5
1680	JGI6
1681	JGI7
1682	JGI8
1683	JGI9
1684	JGI10
1685	SpMax_D

1686	SpDiam_D
Descriptores cuánticos	
Identidad	**Nombre del descriptor**
1687	σ1
1688	HMΔPH
1689	Σr
1690	σalfa
1691	σF
1692	A1
1693	N
1694	Calfa
1695	C
1696	O
1697	A^3
1698	SII_HF
1699	S_HF
1700	SII_MP2
1701	S_MP2
1702	S
1703	S1
1704	S2
1705	alfa
1706	alfa1
1707	alfa2
1708	MP2: Static polarizability for conformer F1
1709	DFT: Static polarizability for conformer F1
1710	DFT(W): Dynamic polarizability for conformer F1
1711	DFT(Z): Static polarizability for the zwitterion structure
1712	EST: Estimated dynamic polarizability for zwitterion structure
1713	HE(19)
1714	HE(17)
1715	WXHX07
1716	BS02
1717	NCJS93
1718	M90
1719	KJM90
1720	VAA(Calculated): Volumen molar parcial
1721	V(vdW)R
1722	CSIR
1723	alfa DS: Global dipole-softness polarizability DS

1724	alfa1 DS: local dipole-softness polarizabilities of the backbone 1 DS
1725	alfa2 DS: side chain 2 DS
1726	−D2 /3S: Global dipole-softness polarizability
1727	−D1 2 /3S: dipole-softness polarizabilities of the backbone
1728	−D2 2 /3S: side chain
1729	−2D1D2 /3S: the interfragmental contribution

Tabla II: Descriptores seleccionados con $CC_{rs} = 1$ para cada mutación.

Identidad de los descriptores	**Mutación (proteína (s))**
128, 188, 270, 282, 326, 391, 1253,1409,1423,1443, 683, 700, 708, 747, 762, 765, 798, 826, 829, 857, 931, 949, 989, 1014	Q-L (N-carbamilasa, luciferasa)
3, 55, 109, 157, 163, 183, 275, 288, 310, 357, 365, 411, 424, 436, 445, 446, 448, 454, 464, 499, 525, 544, 597, 707, 709, 1073, 1117, 1206, 1293, 1305, 1329, 1341, 1353, 1456, 1579, 1609,1699	V-A (N-carbamilasa)
3, 9, 13, 20, 85, 102, 103, 170, 176, 197, 246, 256, 274, 279, 284, 296, 310, 318, 323, 337, 368, 369, 374, 387, 401, 477, 487, 539, 553, 577, 604, 612, 618, 622, 628, 629, 635, 643, 644, 691, 735, 737, 762, 767, 776, 794, 821, 826, 831, 869, 872, 879, 891, 892, 907, 921, 937, 942, 1073, 1082, 1102, 1109, 1117, 1129, 1131, 1133, 1158, 1180, 1189, 1205, 1217, 1256, 1258, 1261, 1276, 1296, 1308, 1332, 1341, 1356, 1382, 1409, 1480, 1481, 1495, 1524, 1536, 1545, 1552, 1559, 1580, 1581, 1582, 1612, 1629, 1642	H-Y (N-carbamilasa)
1, 12, 19, 29, 35, 45, 51, 57, 114, 118, 124, 138, 154, 174, 185, 196, 248,250, 257, 263, 285, 286, 291, 318, 349, 353, 369, 397, 414, 426, 440, 463, 478, 490, 509, 523, 545, 584, 585, 590, 610, 154, 713, 731, 753, 769, 785, 793, 795, 808, 833. 304, 1089, 1090, 1137, 1158, 1159, 1164, 1165, 1170, 1174, 1184, 1195, 1200, 1201, 1210, 1219, 1223, 1235, 1237, 1241, 1245, 1251, 1260, 1343, 1353, 1432, 1457, 1493, 1506, 1564	G-S (N-carbamilasa)
3,4 , 33, 45, 98, 159, 177, 180, 205, 228, 230, 231, 232, 237, 254, 262, 264, 306, 337, 344, 345, 353, 357, 364, 388, 456, 470, 539, 542, 646, 679, 684, 690, 704, 766, 808, 819, 843, 861, 941, 942, 984, 1015, 1056, 1110, 1146, 1151, 1167, 1196, 1200, 1237, 1238, 1241, 1255, 1337, 1344, 1345, 1346, 1428, 1458, 1459, 1522, 1525, 1526, 1534, 1535, 1536, 1539, 1540, 1548, 1549, 150, 1551, 1553, 1554, 1562, 1566, 1571, 1573, 1574, 1585, 1586	M-L(N-carbmilasa)
1, 13, 36 90, 179, 209, 257, 278, 296, 303, 308, 318, 327, 331, 360, 385, 422, 444, 453, 457, 474, 492, 496, 523, 545, 587, 590, 624, 692, 693, 725, 735, 769, 780, 826, 844, 921, 1022, 1089, 1090, 1152, 1159, 1160, 1164, 1186, 1229, 1343, 1419, 1506, 1526, 1536, 1540, 1554	T-A (M-carbamilasa)
18, 67, 139, 170, 239, 244, 273, 305, 364, 369, 377, 429, 431, 432, 448, 449, 453, 461, 462, 547, 549, 551, 553, 561, 609, 625,	F-R (Luciferasa)

710, 719, 747, 759, 806, 815, 823, 855, 877, 894, 912, 940, 971, 983, 984,985, 986, 999, 1002, 1003, 1004, 1023, 1026, 1028, 1030, 1032, 1033, 1040, 1041, 1054, 1055, 1142, 1166, 1236, 1322, 1344, 1346, 1360, 1361, 1362, 1364, 1371, 1387, 1414, 1430, 1473, 1474, 1448, 1490, 1496, 1500, 1509, 1521, 1533, 1540, 1547, 1568, 1570, 1611, 1613, 1628, 1658, 1661, 1679, 1690, 1705, 1708, 1709, 1710, 1711, 1712, 1713, 1714, 1715, 1717, 1718, 1719, 1720	
54, 325, 410, 572, 585, 682, 703, 734, 764, 799, 813, 849, 931, 1013, 1071, 1091, 1097, 1102, 1103, 1152, 1163, 1199, 1210, 1240, 1260, 1272, 1283, 1355, 1366, 1409, 1416, 1422, 1442, 1491, 1524, 1525, 1537, 1539, 1547, 1548, 1552, 1553, 1568, 1570, 1573, 1597, 1678	V-K (Luciferasa)
173, 185, 232, 246, 274, 303, 338, 370, 400, 412, 462, 543, 555, 556, 566, 575, 581, 713, 714, 715, 718, 740, 742, 772, 788, 796, 797, 809, 810, 836, 841, 859, 860, 1025, 1040, 1053, 1295, 1307, 1331, 1385, 1398, 1405, 1418, 1421, 1424, 1694	I-K (Luciferasa)
21, 27, 36, 41, 64, 76, 81, 83, 88, 93, 100, 213, 215, 361, 391, 434, 456, 457, 458, 495, 507, 516, 517, 537, 541, 565, 607, 634, 638, 646, 702, 718, 728, 798, 810, 840, 885, 917, 931, 943, 950, 954, 994, 997, 1006, 1007, 1031, 1034, 1101, 1103, 1116, 1160, 1163, 1207, 1369, 1383, 1386, 1405, 1410, 1416, 1431, 1457, 1510, 1511, 1512, 1514, 1518, 1519, 1537, 1576, 1579, 1616, 1634, 1635, 1662, 1671, 1672, 1683, 1686	H-R (PI3K)
23, 58, 70, 135, 258, 303, 338, 366, 379, 445, 454, 462, 701, 838, 859, 903, 968, 977, 1040, 1188, 1231, 1291, 1303, 1327, 1351, 1400, 1426, 1430, 1453, 1456, 1463, 1615, 1695, 1722	H-L (PI3K)
16, 39, 41, 97, 109, 114, 116, 144, 154, 158, 236, 241, 244, 249, 316, 360, 383, 401, 419, 422, 445, 458, 498, 535, 608, 648, 651, 652, 653, 664, 665, 666, 667, 668, 669, 685, 686, 698, 720, 735, 738, 763, 784, 816, 868, 948, 1022, 1184, 1338, 1345, 1346, 1470	E-K (PI3K)

Ejemplo

Construcción de los alfabetos de aminoácidos

Supongamos que 4 elementos (A, B, C y D) representan los aa proteicos y están caracterizados por una propiedad (d_1) a través de valores numéricos, tal y como se observa en la figura I-*a*. Sobre la matriz de la Figura I-*a*, se aplica HCA. El primer paso es calcular la semejanza entre los pares de aa y para ello se calcula una matriz de distancia (Figura I-*b*), usando la métrica de la distancia euclidiana. Lo valores de distancia más pequeños implican mayor grado de semejanza entre los pares de aa.

	d1
A	1
B	6
C	8
D	4

a

	A	**B**	**C**	**D**
A	0			
B	5	0		
C	7	2	0	
D	3	2	4	0

b

Figura I: *a*. Valores para los elementos A, B, C y D, a partir del descriptor d_1. *b*. Matriz de distancia generada aplicando la distancia euclidiana sobre los valores de la matriz *a*. En el recuadro rojo

A continuación, usando la metodología de agrupamientos unión promedio, se agrupa el par de aa que tenga la distancia mínima y se construye una nueva matriz de distancia reducida. Sin embargo el algoritmo de HCA puede encontrarse con más de un mínimo (*ties in proximity*) como se observa en los recuadros rojos para los pares C-B y D-B de la figura II-*a*. En este caso se construirán dos nuevas matrices de distancia reducidas (Figura II-*b*) (es importante tener en cuenta que muchos programas que calculan HCA, seleccionan al azar cualquiera de los mínimos que surgen de la clasificación y al final se quedan con un solo dendrograma de todos los dendrogramas probables; esto podría representar sesgo y pérdida de información). El proceso es repetido hasta que todos los elementos (aa) se hayan agrupado. En la figura II-*b* la presencia de dos mínimos en la matriz reducida de la izquierda da como resultado dos nuevas matrices reducidas (Figura II-*c*). Al final tres nuevas matrices de distancias reducidas generaran tres dendrogramas diferentes (Figura II-*d*), productos de un solo descriptor (d_1). Los dendrogramas productos de *ties in proximity* generados anteriormente, representan una colección de alfabetos para el descriptor 1 (d_1).

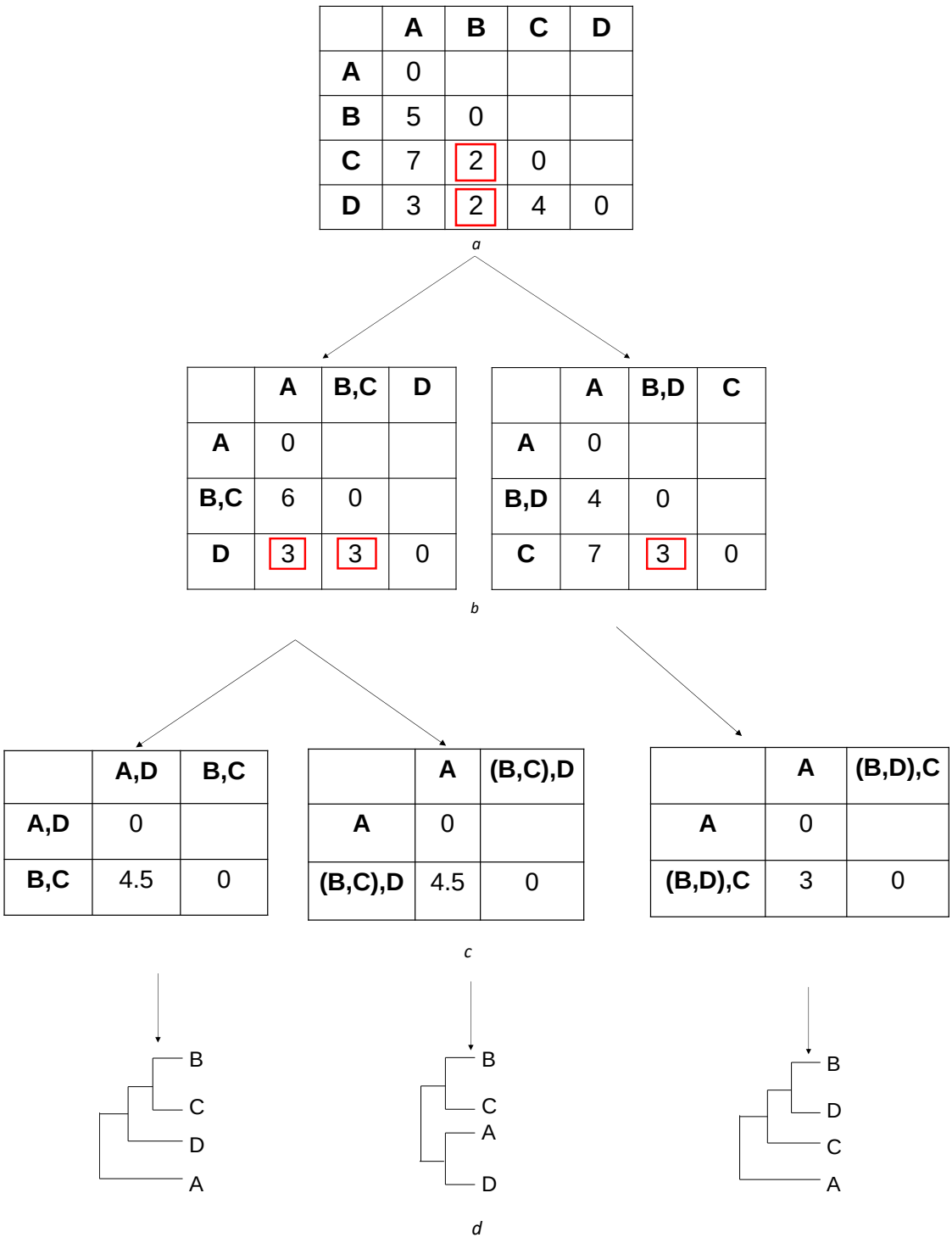

Figura II. Metodología para la construcción de alfabetos reducidos a partir de un descriptor (d_1). *a.* La matriz de distancia generada a partir de la métrica distancia euclidiana. *b.* Las matrices de distancia reducidas producto de los mínimos. *c.* Las matrices de distancias reducidas después de haber agrupado todos los elementos (aa). *d.* Los dendrogramas generados productos de los *ties in proximity*. Los cuadros en rojo, resaltan los mínimos presentes en cada una de las matrices de distancia.

Uso de relaxed set cluster contrast

Ahora supongamos que estamos interesados en calcular la frecuencia con que aparece el par B-C en estos alfabetos reducidos (dendrogramas); el par B-C representará una mutación ocurrida en la secuencia hipotética de alguna proteína. Esto lo haremos usando la metodología *relaxed set cluster contrast (*$CC_{rs}(C, D_i)$*)* de Leal *et al.*
La figura III muestra los sub-conjuntos generados en cada uno de los dendrogramas (D_i) producto de los *ties in proximity* generados en el ejemplo anterior. $CC_{rs}(C, D_i)$ buscará el sub-conjunto más pequeño en el que el par B-C aparecerá en cada uno de los D_i. Estos es: $CC_{rs}(C, D_1) = 2/3$, $CC_{rs}(C, D_2) = 2/2$ y $CC_{rs}(C, D_3) = 2/2$. Finalmente la frecuencia del par B-C será calculada promediando el CC_{rs} calculado en cada uno de los tres dendrogramas.

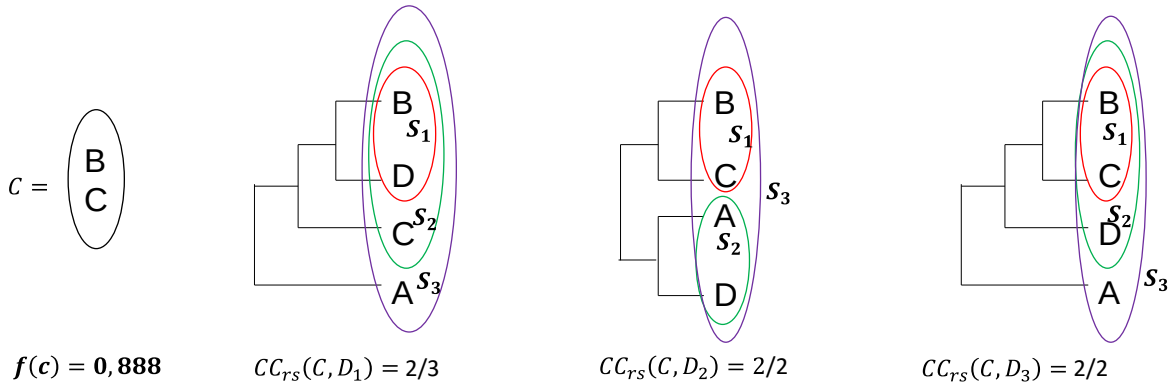

Figura III: Ejemplo en el que se ilustra la metodología *relaxed set cluster contrast* para calcular la frecuencia del par B-C (mutación) en los dendrogramas del descriptor d_1. Los óvalos coloreados representan los sub-conjuntos (S_1, S_2 y S_3).