Revista EIA, ISSN 1794-1237 / Año XIV / Volumen 14 / Edición N.27 / Enero-Junio 2017 / pp. 111-125
Publicación semestral de carácter técnico-científico / Universidad EIA, Envigado (Colombia)

DETECCIÓN DE HOMÓLOGOS REMOTOS USANDO PROPIEDADES FISICOQUÍMICAS

Óscar Bedoya[1]

RESUMEN

En este artículo se presenta un nuevo método para la detección de homólogos remotos en proteínas llamado CDA (Análisis de Distribución de Característica). El método CDA utiliza distribuciones de las propiedades fisicoquímicas de los aminoácidos para cada proteína. Dadas las secuencias de entrenamiento de una familia SCOP (Clasificación Estructural de Proteínas), se calcula su correspondiente distribución característica promediando los valores de las distribuciones para las proteínas que la componen. La hipótesis en esta investigación es que cada familia de proteínas F tiene una distribución característica que separa sus secuencias del resto de las proteínas en un conjunto de datos. Se seleccionó un conjunto de 72 propiedades fisicoquímicas para crear diferentes distribuciones características de la misma familia. Cada distribución característica se usa como un clasificador de familias SCOP. Por último, se utiliza una clasificador Bayesiano para combinar la información de los clasificadores individuales y obtener una mejor decisión. Encontramos que cada familia tiene un conjunto de propiedades fisicoquímicas que permiten una mejor discriminación de sus secuencias. El método CDA alcanza una tasa de aciertos positivos de 0,793, una tasa de falsos positivos de 0,005 y un puntaje ROC de 0,918. El método propuesto mejora la precisión de algunas de las estrategias existentes tales como SVM-PCD y SVM-RQA.

PALABRAS CLAVE: detección de homólogos remotos, familia SCOP, propiedades fisicoquímicas.

REMOTE PROTEIN HOMOLOGY DETECTION USING PHYSICOCHEMICAL PROPERTIES

ABSTRACT

A new method for remote protein homology detection, called CDA (Characteristic Distribution Analysis), is presented. The CDA method uses the distributions of physicochemical properties of amino acids for each protein. Given the training sequences of a SCOP (Structural Classification Of Proteins) family, a characteristic distribution is

[1] Doctorado en Ingeniería. Maestría en Ingeniería. Profesor asociado. Universidad del Valle, Cali, Colombia.

Autor de correspondencia: Bedoya, Ó. (Óscar): Escuela de Ingeniería de Sistemas y Computación. Edificio 331 - espacio 2103. Ciudad Universitaria Meléndez - Universidad del Valle, Cali, Colombia. Tel.: 3212100 - Ext: 2781.
Correo electrónico: oscar.bedoya@correounivalle.edu.co

Historia del artículo:
Artículo recibido: 05-VIII-2013 / Aprobado: 03-X-2017
Disponible online: 30 de agosto de 2017
Discusión abierta hasta octubre de 2018

DOI: https://doi.org/10.24050/reia.v14i27.1161

achieved by averaging the values of the distributions of its proteins. The hypothesis in this research is that each protein family F has a characteristic distribution that separates its sequences from the rest of the proteins in a dataset. A set of 72 physicochemical properties was selected to create different characteristic distributions of the same family. Each characteristic distribution is used as a classifier. Finally, a Naive Bayes classifier is trained to combine the information of the individual classifiers and obtain a better decision. We found that each family has a set of physicochemical properties that allow the discrimination of their sequences better. CDA achieves a True Positive (TP) rate of 0,793, a False Positive (FP) rate of 0,005, and a Receiver Operating Characteristic (ROC) area of 0,918. The CDA method outperforms some of the current strategies such as SVM-PCD and SVM-RQA.

KEYWORDS: Remote Homology Detection, Physicochemical Properties, SCOP Family.

DETECÇÃO DE HOMÓLOGOS REMOTOS USANDO PROPRIEDADES FISICOQUÍMICAS

RESUMO

Neste artigo apresenta-se um novo método para a detecção de homólogos remotos em proteínas chamado CDA (Análises de Distribuição Característica). O método utiliza distribuições das propriedades fisicoquímicas dos aminoácidos. Dada uma família SCOP calcula-se sua correspondente distribuição característica promediando os valores das distribuições para as proteínas que a compõem. A hipótese nesta investigação é que cada família F tem uma distribuição característica que permite diferenciar as sequências em F do resto de proteínas. Ao existir muitas propriedades, ao redor de 554 no AAindex, selecionou-se um conjunto de 72 índices para criar as distribuições. Cada distribuição característica usa-se como um classificador de famílias SCOP. Por último, utiliza-se um classificador Bayesiano para combinar a informação dos classificadores individuais criados a partir das distribuições. O método CDA atinge uma taxa de acertos positivos de 0,793, uma taxa de falsos positivos de 0,005 e uma pontuação ROC de 0,918. O método proposto melhora a exatidão de algumas das estratégias existentes tais como SVM-PCD e SVM-RQA.

PALAVRAS-CHAVE: detecção de homólogos remotos, família SCOP, propriedades fisicoquímicas.

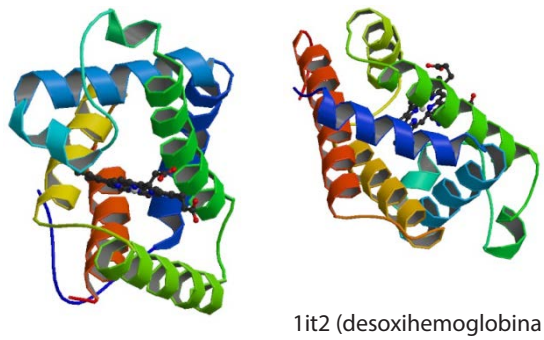
1. INTRODUCCIÓN

La detección de homólogos remotos identifica la homología estructural en proteínas relacionadas evolutivamente que presentan baja similitud de secuencia. Puede ser definido como un proceso que toma una proteína diana y recupera proteínas que son similares en función pero distantes en secuencia. La detección de homología puede ser una tarea difícil porque las proteínas en el espacio de búsqueda comparten similitudes de secuencia baja con el dominio de destino y la relación debe medirse a niveles estructurales y/o funcionales en 3D (Bedoya & Tischer, 2014). La función y la estructura generalmente se conservan más durante la evolución que la secuencia de aminoácidos. Por lo tanto, las proteínas que no exhiben alta similitud de secuencia aún podrían estar relacionadas funcional y estructuralmente (Yang et al., 2008).

La definición formal de homólogos remotos se refiere a las secuencias de proteínas con menos del 25% de identidad de secuencia que exhiben una función similar (Homaeian et al., 2007; Huang & Bystroff, 2006). Sin embargo, la detección de homólogos remotos también puede definirse como el problema

de tomar una proteína P diana y recuperar proteínas en la misma súper familia de P que pertenecen a una familia diferente. La **Figura 1** muestra un ejemplo de homólogos remotos. Se comparó la identidad de secuencia y la similitud estructural de los dominios 1emy (Elephas maximus) y 1it2 (desoxihemoglobina Hagfish). La identidad de secuencia es el número de residuos coincidentes en una alineación de secuencia entre dos dominios. El alineamiento estructural establece la homología entre dos estructuras poliméricas en función de su conformación tridimensional. El SSAP (Programa de Alineación de la Estructura Secuencial) (Orengo & Taylor, 1996) se usó para calcular el alineamiento estructural SSAP proporciona a la RMSF (Desviación Cuadrática Media) y la puntuación SSAP como salidas. El RMSD es una medida de la divergencias de dos estructuras alineadas y el puntaje SSAP mide el alineamiento estructural, donde 100 es la similitud estructural más alta. Se obtuvo una identidad de secuencia de 16%, un puntaje SSAP de 84,97 y un RMSD de 2,13. Los resultados muestran que estos dos dominios comparten alta similitud estructural y una baja identidad de secuencia y, por lo tanto, pueden considerarse homólogos remotos.

Figura 1. Dos homólogos remotos de proteínas



1emy (Elephas maximus)

1it2 (desoxihemoglobina Hagfish)

Identidad de secuencia	Puntaje SSAP	RMSD
16,0	84,97	2,13

Se han propuesto varios métodos para la detección de homólogos remotos (Jaakkola et al., 2000; Hou et al., 2003; Goldstein, 2004; Dong et al., 2006; Gao, 2006; Yang et al., 2008; Webb-Robertson et al., 2010; Muda et al., 2011; Chitraranjan et al., 2011). Sin embargo, aún se necesita una estrategia efectiva. Los métodos existentes aún pueden confundirse por una baja similitud entre las secuencias de aminoácido, a pesar de que estén estrechamente relacionados en función (Huang & Bystroff, 2006). SVM I-sites (Hou et al., 2003) es un método de detección de homólogos remotos. Este utiliza la biblioteca del I-site para generar una puntuación al enviar cada sub-fragmento de una secuencia diana desconocida a la matriz de sustitución que representa cada I-site. Debido a que hay motivos de diferentes tamaños, las puntuaciones de similitud de los diferentes grupos de I-sites no son directamente comparables. En consecuencia, existe la necesidad de asignar cada puntuación a un rango de valores comparables. Hou et al. (2003) propuso utilizar una curva de confianza para cada grupo específico de I-sites. Una curva de confianza asigna puntajes de similitud a la probabilidad de la estructura local correcta basada en una prueba Jack-knife. La confianza de una predicción de fragmento es la probabilidad de que un segmento de secuencia con una puntuación dada tenga la estructura predicha por el motivo. Para predecir la estructura local de cualquier secuencia de proteínas desconocida, los patrones de secuencia (perfiles) para cada uno de los 263 grupos de la biblioteca de I-sites se utilizan para puntuar todos los sub-fragmentos de la secuencia diana desconocida. Un vector de características para una Proteína P en Huo et al. (2003) se calcula como la suma de los valores de confianza para 263 motivos en todos los sub-fragmentos de P.

Otro método de detección de homólogos remotos es presentado por Gao (2006). Utiliza la matriz-γ del modelo HMMSTR. La conocida matriz-γ (Rabiner & Biing-Hwangh, 1986) tiene 281 columnas que representan los estados de Markov de HMMSTR (Modelo de Markov oculto para la estructura proteínica) y N filas, donde N es la longitud de la proteína P presentada al modelo. Gao (2006) llama cada fila de la matriz-γ, un vector-γ. Luego, los vectores-γ- se agrupan para determinar los

vectores más representativos en un conjunto de datos de entrenamiento. El métodos k-means se usa como el algoritmo de agrupación y los centroides se toman como el conjunto de vectores-γ representativos. Finalmente, cada vector-γ de una proteína P en un conjunto de entrenamiento se mapea al clúster más cercano y por lo tanto, cada proteína se representa como una cadena de símbolos que indica la secuencia de los grupos mapeados. Todo el conjunto de proteínas de entrenamiento se indexa usando un árbol de sufijos para agilizar el proceso de consulta.

Muda et al. (2011) aborda la detección de homólogos remotos y los problemas de reconocimiento de pliegues. Se propone un clasificador de dos capas. En el primer nivel, se utiliza un clasificador SVM (máquina de soporte de vectores) para detectar los homólogos remotos. La clasificación se realiza con base en clasificadores binarios de uno contra todos. El vector de características utilizado para entrenar el SVM se basa en los valores numéricos del índice AA (Kawashima et al., 2008). La escala se realiza sobre datos numéricos para evitar dominio durante procesos de clasificación. SVM-PCD (Webb-Robertson et al., 2010) utiliza el concepto de distribuciones de propiedades fisicoquímicas para la detección de homólogos remotos. Cada proteína se representa como la distribución de sus 4-mers, el promedio de los valores fisicoquímicos en una ventana de 4 aminoácidos. Se obtiene una distribución de 18 valores para cada índice en la base de datos del índice AA. Webb-Robertson et al. (2010) proponen PCD(531), PCD(181), y PCD(61), que toman índices 531, 181, y 61 de las propiedades fisicoquímicas en el índice AA, respectivamente. Los valores considerados en cada caso se concatenan y se usan para entrenar una SVM para cada familia.

En este artículo, se presenta un nuevo método para la detección de homólogos remotos, denominado Análisis de Distribución de Características (CDA). El método CDA se basa en obtener una distribución de las propiedades fisicoquímicas de aminoácidos para cada proteína. Una distribución característica se construye con las secuencias de entrenamiento de cada familia SCOP. La hipótesis de esta investigación es que cada familia F tiene una distribución característica que separa sus secuencias del resto de las proteínas en un conjunto de datos. Hay 554 propiedades fisicoquímicas en el Índice AA. En esta investigación se utilizarán 72 propiedades fisicoquímicas comúnmente referidas. La metodología que se utiliza en esta investigación permite probar cada propiedad fisicoquímica independientemente de las demás y, por lo tanto, se puede obtener la propiedad fisicoquímica que mejor discrimina las secuencias de una familia de proteínas específica. Además, también se obtiene una decisión final cuando se utiliza el clasificador Bayesiano.

2. MÉTODOS

2.1. Ventana móvil con posición ponderada

El primer paso en el método CDA es transformar la secuencia de aminoácidos en valores fisicoquímicos definidos en un índice específico. Cada índice asigna un valor por cada uno de los 20 aminoácidos. Por ejemplo, el momento hidrofóbico basado en átomos se define por los 20 valores que se muestran en la **Tabla 1**. Como se puede observar, el momento hidrofóbico más alto pertenece al aminoácido Arginina (R) y el más bajo a la Alanina (A). Las propiedades fisicoquímicas son incluidas en la detección de homólogos remotos debido a la hipótesis de que se conservan en su mayoría durante la evolución (Grigoriev & Kim, 1999; Yang et al., 2008).

TABLA 1. ÍNDICE DE MOMENTO HIDROFÓBICO BASADO EN ÁTOMOS

A	R	N	D	C	Q	E	G	H	I
0,0	10,0	1,3	1,9	0,17	1,9	3,0	0,0	0,99	1,2
L	K	M	F	P	S	T	W	Y	V
1,0	5,7	1,9	1,1	0,18	0,73	1,5	1,6	1,8	0,48

En este artículo se usa una ventana móvil con una posición ponderada del tamaño 5 en lugar de promediar los valores de cada posición. Utilizamos la misma estrategia propuesta por e use Bedoya y Tischer (2014). De acuerdo con su estrategia, el peso en cada posición indica la contribución al valor representativo de la ventana y se asigna al aminoácido en su centro. Teniendo en cuenta los cinco valores (v_1,v_2,v_3,v_4,v_5) de una propiedad fisicoquímica para los aminoácidos (a_{i-2}, a_{i-1}, a_i, a_{i+1}, a_{i+2}), el valor de contribución c de la ventana asignada al aminoácido ai se calcula como en la **Ecuación (1)**.

$$c = v_{i-2}*0{,}05789 + v_{i-1}*0{,}24450 + v_i *0{,}39521 + v_{i+1}*0{,}24450 + v_{i+2}*0{,}05789 \quad (1)$$

El tamaño de la ventana trata de capturar interacciones locales entre aminoácidos que en realidad son vecinos cercanos. El tamaño de la ventana consideró que las relaciones 3D más importantes entre los aminoácidos se producen en un rango local. Dada una secuencia de aminoácidos de n residuos y la ventana móvil de tamaño 5, se obtiene un total de valores de contribución n-4. El conjunto de valores obtenidos de las ventanas móviles de la proteína completa se llama Vector de Contribución (CV) (Bedoya & Tischer, 2014).

2.2. Selección de propiedades fisicoquímicas

Hay 554 propiedades fisicoquímicas en el Índice AA. De acuerdo con Yang et al. (2008) y Webb-Robertson et al. (2010), hay índices que reflejan las características funcionales o estructurales de una familia de proteínas específica. Por ejemplo, en SVM-RQA (Yang et al., 2008) se encuentran los mejores índices para la familia 1.4.1.3 SCOP (Dominio de unión de ADN c-Myb). La familia 1.4.1.3 contiene cadenas laterales hidrófobas y proteínas helicoidales. Los índices más adecuados para esta familia son pK(-COOH), polaridad, propensión a la hélice alfa derivada de secuencias diseñadas y la frecuencia normalizada de la hélice alfa zurda. Los dos primeros índices están relacionados con la propiedad de hidrofobicidad y los dos últimos son índices relacionados con la estructura. En SVM-PCD (Webb-Robertson et al., 2010), se usaron los 531 índices en el Índice AA. También redujeron el número de índices con base en un análisis de correlación y mostraron que no se logra ganancia en precisión más allá de los 61 índices usados en SVM-PCD(61).

En este documento, se seleccionaron 72 índices considerando los resultados informados por Yang et al. (2008) y Webb-Robertson et al. (2010). La lista de los índices seleccionados se muestra en la **Tabla 2**. El objetivo de utilizar una cantidad considerable de índices es detectar cuáles son los más apropiados para usar el análisis CDA.

2.3. Obteniendo una distribución para cada secuencia de proteína

El siguiente paso en el análisis CDA es obtener la distribución de los vectores de contribución para cada proteína. La decisión de obtener una distribución está relacionada con la transformación de aminoácidos en un conjunto de valores de tamaño fijo. En este documento, 20 valores se utilizan para describir la distribución de los valores en el vector de contribución. Por lo tanto, las proteínas de diferentes tamaños se vuelven comparables porque todas se expresan como un conjunto de 20 valores.

Primero que todo, cada valor en el vector de contribución se normaliza con respecto a la media y la desviación estándar asociada con el índice que representa una propiedad fisicoquímica siguiendo la misma estrategia propuesta por Bedoya y Tischer (2014). La media y la desviación de un índice se calculan promediando 3200000 posibles valores de contribución que se pueden obtener de una ventana de tamaño 5. Las **Ecuaciones (2)** y **(3)** muestran el cálculo de la media y la desviación, respectivamente.

$$\mu = \frac{\sum_{i,j,k,l,m=1}^{20} (v_i*0{,}05789 + v_j*0{,}24450 + v_k*0{,}39521 + v_l*0{,}24450 + v_m*0{,}05789)}{3{,}2 \times 10^6} \quad (2)$$

$$\sigma = \sqrt{\frac{\sum_{i,j,k,l,m=1}^{20} \left((v_i*0{,}05789 + v_j*0{,}24450 + v_k*0{,}39521 + v_l*0{,}24450 + v_m*0{,}05789) - \right)^2}{3{,}2 \times 10^6 - 1}} \quad (3)$$

donde v_i, v_j, v_k, v_l, y v_m son los valores de un índice específico. Cada sumatoria va del uno al 20, indicando los 20 valores posibles de una propiedad fisicoquímica.

La media y la desviación se calculan para cada uno de los 72 índices. Por ejemplo, el índice de momento hidrofóbico basado en átomos tiene una media de 1,8221 y una desviación de 1,1948. Una vez calculadas la media y la desviación, cada valor en un vector de contribución se normaliza usando la **Ecuación (4)**. El conjunto de valores normalizados de una proteína se llama Vector de Contribución Normalizada (NCV).

$$NCVij = \frac{CVij - \mu_j}{\sigma_j} \quad (4)$$

donde CV_{ij} es el i-ésimo valor en el vector de contribución que usa el índice j-ésimo, μ_j es la media del índice j-ésimo y σ_j es la desviación estándar del índice j-ésimo. La normalización de la media y la desviación transforma los valores en CV en valores que están en su mayoría en el rango de -4σ a 4σ. El siguiente paso es tomar los valores normalizados y convertirlos en una distribución. Esto se hace mediante un proceso de *binning*. Utilizar este proceso en el rango de los valores normalizados consiste en calcular la frecuencia de cada bin, comenzando desde -1,8 y tomando intervalos de 0,3 hasta 3,9. El proceso de binning produce 20 valores de frecuencia. Finalmente, los valores de frecuencia se normalizan dividiendo cada valor por el número de valores en el vector de contribución normalizado.

La **Figura 2** muestra la distribución para dos secuencias. Las familias 1.27.1.1 y 1.36.1.5 se seleccionaron para observar la diferencia entre las distribuciones de dos proteínas cuando se usa el índice de momento hidrofóbico basado en átomos. Se espera que las distribuciones de secuencias que pertenecen a diferentes familias exhiban formas claramente diferentes.

2.4. Obteniendo una curva característica para cada familia

La hipótesis en este trabajo es que cada familia tiene una curva característica que representa los valores de las distribuciones de las secuencias en la familia. Una distribución característica para una familia F se obtiene tomando sus secuencias, calculando las distribuciones y promediando los valores en cada posición. Utilizamos el conjunto de datos propuesto por Liao y Noble (2003), que se ha convertido en el conjunto de datos estándar en la detección de homólogos remotos. El conjunto de datos está formado por 54 familias y cada familia tiene una cantidad diferente de secuencias y conjuntos específicos para el entrenamiento y las pruebas. Los detalles de las definiciones del conjunto de datos están disponibles en http://noble.gs.washington.edu/proj/svm-pairwise/. El conjunto de datos de entrenamiento disponible para cada familia se utilizó para obtener su distribución característica. Adicionalmente, las 857 secuencias referidas como las secuencias de prueba en (Liao y Noble, 2003) fueron usadas para calcular la precisión del método.

TABLA 2. LISTA DE PROPIEDADES FISICOQUÍMICAS

Propiedad fisicoquímica	
Relación de reducción de accesibilidad	Frecuencia normalizada de hélice aislada
Índices aperiódicos para las proteínas beta	Frecuencia normalizada de hélice aislada
Volumen específico parcial aparente	Frecuencia normalizada de hélice alfa zurda
Momento hidrofóbico basado en átomos	Frecuencia normalizada de la hélice N-terminal
Energía media no ligada por átomo	Frecuencia normalizada de giro inverso, no ponderado
Energía media no ligada por residuo	Frecuencia normalizada de los residuos segundo y tercero a su vez
Promedio fraccionado de ocurrencia en AR(i)	Frecuencia de residuos normalizada en el terminal de hélice C1
Promedio fraccionado de ocurrencia en AR(i-1)	Frecuencia de residuos normalizada en el terminal de hélice C2
Ángulo promedio de orientación de la cadena lateral	Frecuencia de residuos normalizada en el terminal de hélice N1
Propensión media de giro en una hélice transmembranal	Frecuencia relativa normalizada de curvatura R
Propensión de hélice beta derivada de secuencias diseñadas	Volumen normalizado de van der Waals
Parámetro conformacional de la hélice interna	pK (-COOH)
Preferencia de conformación para todas las hebras beta	Polaridad
Valores de Delta G para los péptidos extrapolados a urea 0 M	Frecuencia relativa de ocurrencia
Dirección del momento hidrofóbico	Población relativa del estado conformacional C
Frecuencia de ocurrencia en curvatura beta	Valor de preferencia relativo en C'
Constante de equilibro helicoidal	Valor de preferencia relativo en C1
Potencial de hidratación	Valor de preferencia relativo en N3
Hidropatía	Parámetro de interacción de cadena lateral
Escala de hidropatía basada en valores de auto información	Tamaño
Parámetro hidrofóbico	Energía libre de solvatación
Punto isoeléctrico	Constantes de acoplamiento spin-spin 3JHalpha-NH
Pérdida media del área fraccional	Volúmenes de superficie e interiores en proteínas globulares
Parámetros de preferencia enterrados en la membrana	El parámetro Chou-Fasman de la conformación de la bobina
Peso molecular	Incrementos de la constante de Kerr
Carga negativa	Transferir energía, orgánico/agua solvente
Carga neta	Transferir energía libre de vap a chx
Escalas normalizadas de hidrofobicidad media	Regiones transmembranales de proteínas no mt
Parámetros de flexibilidad normalizados (valores B), promedio	Valor de theta(i-1)
Frecuencia normalizada de la hélice alfa	Parámetro de van der Waals R0
Frecuencia normalizada de la hélice alfa de LG	Pesos para el hélice alfa en la posición de ventana de -1
Frecuencia normalizada de la hélice alfa, no ponderada	Pesos para la lámina beta en la posición de ventana de 5
Frecuencia normalizada de la lámina beta de LG	Pesos para la lámina beta en la posición de ventana de -6
Frecuencia normalizada de la lámina beta en toda la clase beta	Pesos para la bobina en la posición de ventana de 3
Frecuencia normalizada de la lámina beta, no ponderada	Pesos para la bobina en la posición de ventana de 4
Frecuencia normalizada del giro beta	Pesos para la bobina en la posición de ventana de 6

Figura 2. Distribuciones de secuencias de las familias 1.27.1.1 y 1.36.1.5

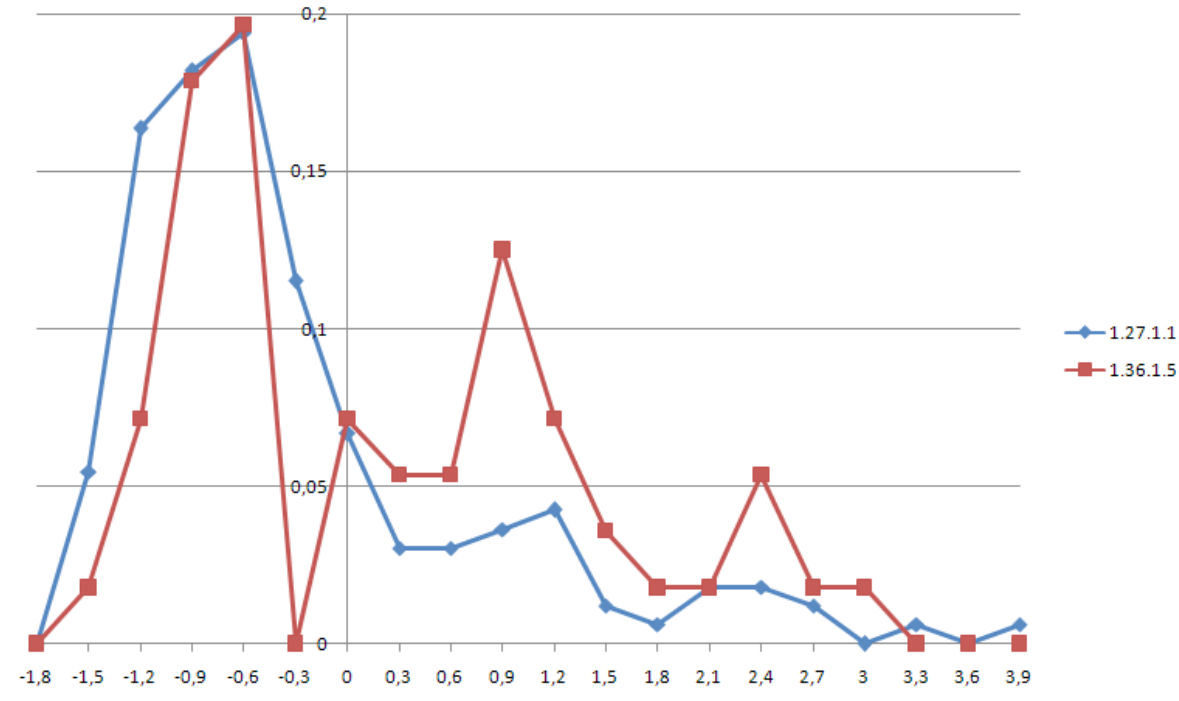


Figura 3. Distribuciones características para cuatro familias usando el índice de momento hidrofóbico basado en átomos

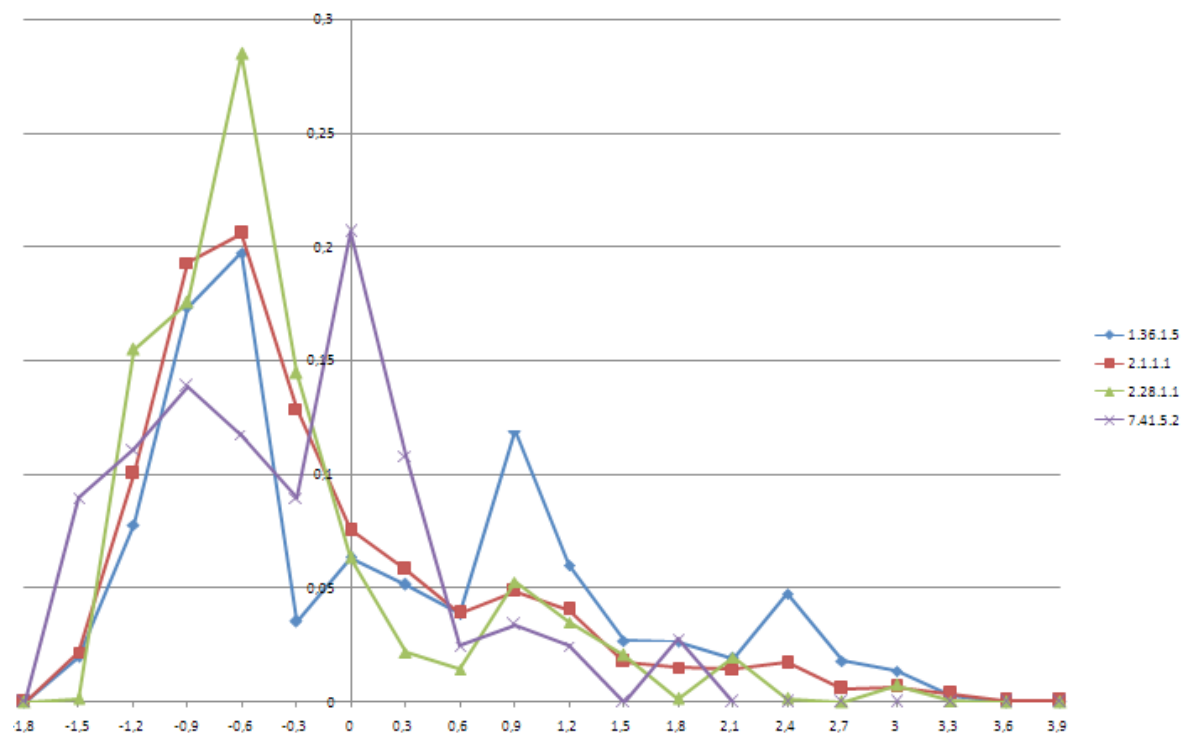
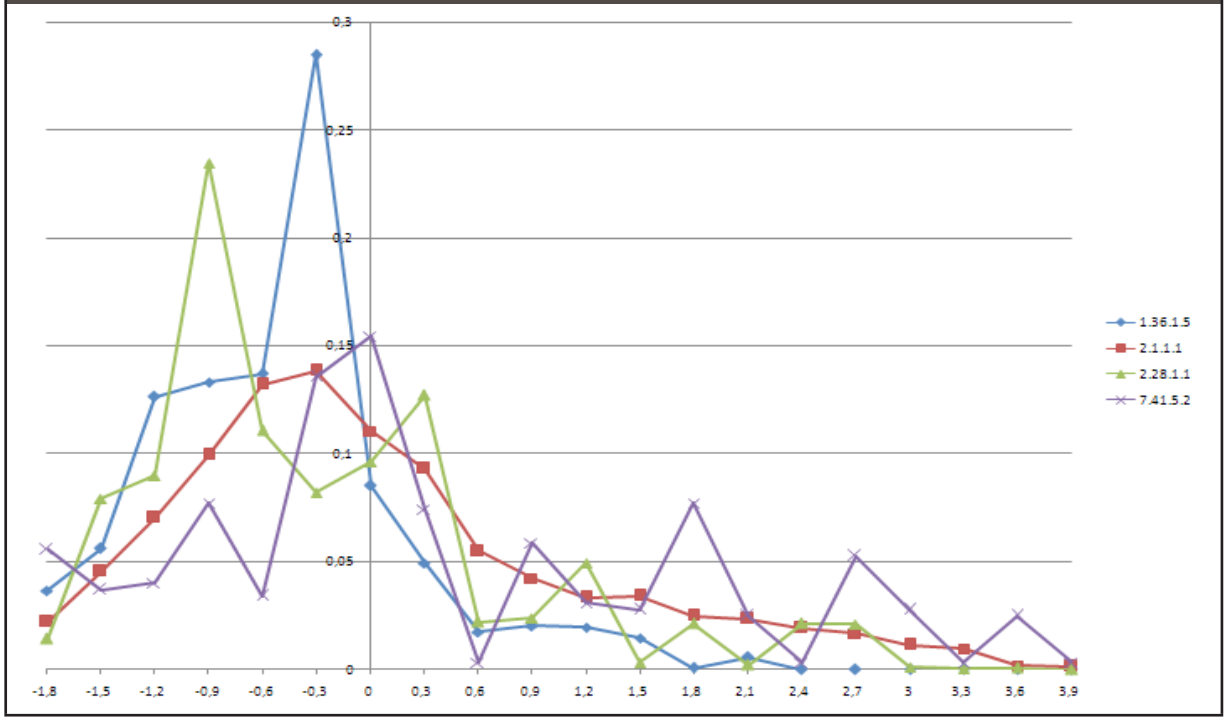


Figura 4. Distribuciones características para cuatro familias que usan la propensión a la hélice alfa derivada del índice de secuencias diseñadas



La **Figura 3** muestra las distribuciones características para las familias 1.36.1.5, 2.1.1.1, 2.28.1.1, y 7.41.5.2. Se utilizó el índice de momento hidrofóbico basado en átomos. Cada distribución característica se obtiene al promediar los valores de las distribuciones en la misma familia. Hay 72 distribuciones características para cada familia (es decir, se obtiene una distribución para cada propiedad fisicoquímica). Se observó que algunos índices discriminan las 54 familias mejor que otros. Además, hay familias que son difíciles de representar y solo unos pocos índices pueden discriminarlas.

La **Figura 4** muestra las distribuciones característica de las familias 1.36.1.5, 2.1.1.1, 2.28.1.1, y 7.41.5.2 cuando se utiliza la propensión a la hélice alfa derivada del índice de secuencias diseñadas. Cada propiedad fisicoquímica específica ofrece una visión diferente de la misma familia. Como se puede observar a partir de las Figuras 3 y 4, la familia 1.36.1.5 exhibe valores medios de momentos hidrofóbicos y está formada por secuencias que muestran una alta propensión a la hélice alfa.

2.5. Detectando homólogos remotos

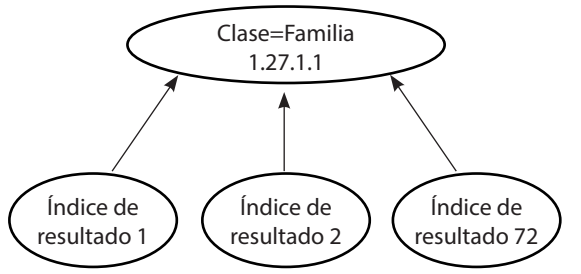
El método CDA crea un modelo de clasificación para cada familia. Aunque no todas las propiedades fisicoquímicas son apropiadas para discriminar a las 54 familias, esperamos que al menos uno de los índices represente a cada familia. La clasificación de una proteína P en una familia SCOP se hace transformando P en un vector de contribución normalizado y comparándolo con las distribuciones características de las 54 familias. La comparación se realiza utilizando la distancia de Manhattan. Formalmente, dado el vector de contribución normalizado de P, $NCV_p = (v_1, v_2, ..., v_{20})$, la distancia entre P y la i-ésima distribución característica usando la j-ésima propiedad fisicoquímica se calcula de la siguiente manera:

$$d(P,NCV_{ij}) = \sum_{k=1}^{20} |v_k - NCV_{ijk}| \qquad (5)$$

donde NCV_{ijk} es el k-ésimo valor en la i-ésima distribución característica usando la j-ésima propiedad fisicoquímica. Una vez las distancias hacia las distribuciones características son calculadas, la proteína P es asignada a la familia con la menor distancia (es decir, la distribución más cercana).

Dadas unas proteínas P diana y 72 distribuciones características, cada propiedad fisicoquímica da un resultado (es decir, una decisión de clasificación). Cada clasificación intenta asignar P a su familia real mediante el uso de una propiedad fisicoquímica diferente. Debido a que algunas de las decisiones de clasificación podrían coincidir con la familiar correcta y algunas otras podrían estar equivocadas, se entrenó a un clasificador Bayesiano para aprender la familia correcta dados los 72 resultados. La **Figura 5** muestra la estructura del clasificador Bayesiano utilizado para la familia 1.27.1.1. La decisión de clasificación se toma con base en los resultados de los 72 índices. Se podría esperar que, aunque no todos los resultados coinciden con la clasificación real, al menos algunos de ellos son correctos y el clasificador Bayesiano los identifica. El clasificador Bayesiano sigue el modelo condicional $p(C|F_1, F_2, ..., F_n)$ donde C es el número de clases (es decir, las 54 familias) y F_1 a F_n son las variables características (es decir los resultados del índice 1 al índice 72).

Figura 5. Clasificador Bayesiano



Una vez que se determina la familia SCOP para una proteína P, los homólogos remotos se identifican al devolver las secuencias en la misma súper familia pero fuera de la familia predicha.

3. RESULTADOS Y DISCUSIÓN

En esta sección se describen los resultados obtenidos en los experimentos. En primer lugar, calculamos la precisión de tener una distribución característica para cada familia. Luego, el clasificador Bayesiano fue probado. Los experimentos se llevaron a cabo mediante el uso de dos scripts. El cálculo de la distribución de cada secuencia de proteína en el conjunto de datos y la curva característica para cada familia se realizó utilizando el lenguaje de programación Bio-Python. El clasificador Bayesiano se obtuvo utilizando la herramienta de minería de datos WEKA (Hall et al., 2009). Los parámetros se mantuvieron por defecto.

3.1. Precisión de clasificación

Primero que todo, se probó la potencia discriminativa de las distribuciones características. Utilizamos los mismos conjuntos de datos de entrenamiento y prueba propuestos por Liao y Noble (2003). Las distribuciones características para cada familia se obtuvieron mediante el uso del conjunto de datos de entrenamiento y luego se calculó la precisión del método en el conjunto de datos de prueba. El uso de todas las propiedades fisicoquímicas para clasificar las secuencias de proteínas en su familia correcta mostró el siguiente top 5 de índices ordenados por la cantidad de coincidencias. La **Tabla 3** muestra los índices que clasifican la mayoría de las proteínas dado un total de 857 secuencias en el set de prueba.

TABLA 3. LISTA DEL TOP 5 DE ÍNDICES

Índice	Porción de secuencias correctas
Propensión a la hélice alfa derivada de secuencias diseñadas	428/857
El parámetro Chou-Fasman de la conformación de la bobina	408/857
Regiones transmembranales de proteínas no mt	403/857
Volumen específico parcial aparente	401/857
Frecuencia normalizada de giro inverso	401/857

TABLA 4. LOS MEJORES ÍNDICES POR FAMILIA

Familia SCOP	Mejor índice	Familia SCOP	Mejor índice
1.27.1.1	Parámetros de preferencia enterrados en la membrana	2.9.1.4	Frecuencia relativa normalizada de curvatura R
1.27.1.2	Promedio fraccionado de ocurrencia en AR(i-1)	3.1.8.1	Volumen específico parcial aparente
1.36.1.2	Valor de preferencia relativo en C'	3.1.8.3	Momento hidrofóbico basado en átomos
1.36.1.5	Propensión a la hélice alfa derivada de secuencias diseñadas	3.2.1.2	Valor de preferencia relativo en C'
1.4.1.1	Momento hidrofóbico basado en átomos	3.2.1.3	Frecuencia normalizada de giro inverso, no ponderado
1.4.1.2	Propensión a la lámina beta derivada de secuencias diseñadas	3.2.1.4	Relación de reducción de accesibilidad
1.4.1.3	Hidropatía	3.2.1.5	Ángulo promedio de orientación de la cadena lateral
1.41.1.2	Valor de preferencia relativo en C1	3.2.1.6	pK (-COOH)
1.41.1.5	Carga neta	3.2.1.7	Frecuencia normalizada de hélice aislada
1.45.1.2	Preferencia de conformación para todas las hebras beta	3.3.1.2	Frecuencia normalizada de giro inverso, no ponderado
2.1.1.1	Frecuencia normalizada de la hélice N-terminal	3.3.1.5	Frecuencia relativa de ocurrencia
2.1.1.2	Propensión a la hélice alfa derivada de secuencias diseñadas	3.32.1.1	Constante de equilibrio helicoidal
2.1.1.3	Pérdida media del área fraccional	3.32.1.11	Incrementos de la constante de Kerr
2.1.1.4	Relación de reducción de accesibilidad	3.32.1.13	Propensión a la hélice alfa derivada de secuencias diseñadas
2.1.1.5	Frecuencia relativa de ocurrencia	3.32.1.8	Propensión media de giro en una hélice transmembranal
2.28.1.1	Momento hidrofóbico basado en átomos	3.42.1.1	Parámetro de interacción de cadena lateral
2.28.1.3	Transferir energía, orgánico/agua solvente	3.42.1.5	Pesos para la bobina en la posición de ventana de 4
2.38.4.1	Momento hidrofóbico basado en átomos	3.42.1.8	Peso molecular
2.38.4.3	Momento hidrofóbico basado en átomos	7.3.10.1	Valores de Delta G para los péptidos extrapolados a urea 0 M
2.38.4.5	Polaridad	7.3.5.2	Frecuencia normalizada del giro beta
2.44.1.2	Parámetros de flexibilidad normalizados (valores B), promedio	7.3.6.1	Transferir energía, orgánico/agua solvente
2.5.1.1	Incrementos de la constante de Kerr	7.3.6.2	Valor de preferencia relativo en N3
2.5.1.3	Valor de theta(i-1)	7.3.6.4	Frecuencia relativa de ocurrencia
2.52.1.2	Valor de preferencia relativo en C1	7.39.1.2	Pesos para la bobina en la posición de ventana de 3
2.56.1.2	Parámetro hidrofóbico	7.39.1.3	Frecuencia normalizada de hélice alfa zurda
2.9.1.2	Parámetro hidrofóbico	7.41.5.1	pK (-COOH)
2.9.1.3	Promedio fraccionado de ocurrencia en AR(i)	7.41.5.2	pK (-COOH)

El índice de propensión a la hélice alfa derivado de secuencias diseñadas detecta la familia correcta 428 de las 857 secuencias. Fue el mejor índice teniendo en cuenta la cantidad de coincidencias. Además, encontramos que aunque dos índices tienen el mismo número de coincidencias correctas, no necesariamente significa que clasifiquen las mismas secuencias (es decir, las 401 secuencias del índice de volumen específico parcial aparente no son necesariamente las mismas que las 401 secuencias de la frecuencia normalizada del índice de giro inverso). Contando el número de secuencias que tienen al

menos un índice que permite identificar a su familia correcta, da un total de 840 secuencias. Esto muestra que, aunque el mejor índice identifica 428 familias correctas (49,975%), el conjunto de 72 propiedades fisicoquímicas permite detectar el 98,01% de las secuencias completas. Hay 17 secuencias que ninguna de las propiedades fisicoquímicas utilizadas en esta investigación puede identificar (es decir, dos secuencias en la familia 2.1.1.4, tres secuencias en la familia 2.28.1.1, y 12 secuencias en la familia 2.44.1.2).

Algunas familias de proteínas son fáciles de representar (es decir, varios índices representan la mayoría de sus secuencias) y algunas otras familias son difíciles de representar (es decir, sólo unos pocos índices las representan). Para cada familia de proteínas hay un índice que representa la mayoría de sus secuencias. La **Tabla 4** muestra el mejor índice para cada una de las 54 familias.

TABLA 5. LOS MEJORES ÍNDICES POR FAMILIA

Familia SCOP	El mejor conjunto de índices
1.27.1.1	Parámetros de preferencia enterrados en la membrana Relación de reducción de accesibilidad
	Frecuencia normalizada de los residuos segundo y tercero en el giro
	Pesos para la bobina en la posición de ventana de 6
	Frecuencia normalizada de lámina beta, no ponderada
2.28.1.1	Momento hidrofóbico basado en átomos
	Frecuencia normalizada de giro inverso, no ponderado Frecuencia relativa normalizada de curvatura R
	Parámetro de interacción de cadena lateral
	Frecuencia de residuos normalizada en el terminal de hélice N1
7.41.5.2	pK (-COOH)
	Frecuencia normalizada de la hélice alfa
	Preferencia de conformación para todas las hebras beta
	Parámetro de van der Waals R0
	Relación de reducción de accesibilidad

Otro resultado importante se logró en esta investigación; cada familia tiene un conjunto de propiedades fisicoquímicas que exhiben el mayor potencial discriminativo en el método CDA. La **Tabla 5** muestra el mejor conjunto de índices para las familias 1.27.1.1, 2.28.1.1, y 7.41.5.2.

Un total de 840 proteínas de las 857 secuencias en el conjunto de prueba tienen al menos un índice que las discrimina por familias SCOP. Dada una proteína P con una familia desconocida, el vector de contribución normalizado de P tiene que ser comparado con las 54 distribuciones características. Adicionalmente, debido a que hay 72 distribuciones características para cada familia, cada propiedad fisicoquímica da un resultado (es decir, una decisión de clasificación). Los 72 resultados para P se envían al clasificador Bayesiano. Este calcula la probabilidad de que P pertenezca a cada una de las 54 clases, $p(C|F_1, F_2, ..., F_n)$ donde C es el número de clases (es decir, las 54 familias) dado F_1 a F_n (es decir, los 72 resultados obtenidos anteriormente). La clasificación obtenida por la técnica Bayesiana se toma como la familia SCOP predicha para P.

La construcción de un clasificador Bayesiano requiere un conjunto de datos adicional para su entrenamiento. Debido a que tenemos que mantener el conjunto de datos de prueba sin ser visto durante el entrenamiento, dividimos el conjunto de datos de entrenamiento propuesto por Liao y Noble (2003) en dos partes. El 70% de las secuencias en el conjunto de datos de entrenamiento se utilizó para obtener las distribuciones características para cada familia. El 30% restante se utilizó para entrenar el clasificador Bayesiano. Finalmente, el conjunto de datos de prueba se utilizó para obtener la precisión del clasificador Bayesiano. La **Tabla 6** muestra la tasa de TP (verdadero positivo), la tasa de FP (falso positivo), la Medida F y el área ROC (Característica de Funcionamiento del Receptor) para algunas familias.

TABLA 6. PRECISIÓN EN EL MÉTODO CDA

Familia	Tasa TP	Tasa FP	Medida F	Área ROC
1.27.1.2	1,000	0,007	0,727	1,000
1.36.1.5	1,000	0,000	1,000	1,000
1.4.1.3	1,000	0,000	1,000	1,000
1.41.1.5	0,840	0,005	0,840	0,996
2.1.1.5	0,370	0,013	0,417	0,943
2.38.4.3	0,364	0,004	0,444	0,919
2.5.1.3	0,600	0,002	0,667	0,952
3.2.1.3	0,333	0,001	0,462	0,897
3.32.1.1	0,444	0,006	0,444	0,890
7.3.5.2	0,556	0,007	0,614	0,843

Los valores promedio considerando las 54 familias para la Tasa TP, Tasa FP, Medida F y área ROC son 0,793, 0,005, 0,793, y 0,918, respectivamente. El área ROC se usa con frecuencia para comparar diferentes métodos. Se observó que para algunas familias (es decir, 1.36.1.5 y 1.4.1.3) la mayoría de los 72 resultados coinciden con la familia correcta. Estas familias son fáciles de representar con un clasificador Bayesiano y se obtiene una Tasa TP de 1,0 y una Tasa FP de 0,0. Por otro lado, hubo familias en las que sólo algunos pocos de los 72 resultados fueron correctos.

3.2. Reduciendo la dimensionalidad

Redujimos la dimensionalidad del método CDA en otro experimento. Se utilizó la misma metodología teniendo en cuenta sólo los mejores índices para las 54 familias. Debido a que encontramos que algunas familias comparten el mismo mejor índice, pudimos reducir el número de índices a 35. Estos índices se muestran en la **Tabla 7**.

Los valores medios del clasificador Bayesiano utilizando 35 índices fueron 0,754, 0,006, 0,753, y 0,901 para la Tasa TP, Tasa FP, Medida F y Área ROC, respectivamente. Aunque la Tasa de TP disminuye, el tiempo computacional del método se mejora porque sólo se calculan 35 índices.

TABLA 7. ÍNDICES UTILIZADOS PARA REDUCIR LA DIMENSIONALIDAD

Relación de reducción de accesibilidad	Frecuencia normalizada de la hélice N-terminal
Propensión de hélice alfa derivada de secuencias diseñadas	Frecuencia normalizada de giro inverso, no ponderado
Índices aperiódicos para las proteínas beta	Frecuencia de residuos normalizada en el terminal de hélice N1
Volumen específico parcial aparente	Frecuencia relativa normalizada de curvatura R
Momento hidrofóbico basado en átomos	pK (-COOH)
Energía media no ligada por átomo	Polaridad
Promedio fraccionado de ocurrencia en AR(i-1)	Frecuencia relativa de ocurrencia
Propensión media de giro en una hélice transmembranal	Población relativa del estado conformacional C
Parámetro conformacional de la hélice interna	Valor de preferencia relativo en C'
Hidropatía	Valor de preferencia relativo en C1
Parámetro hidrofóbico	Valor de preferencia relativo en N3
Parámetros de preferencia enterrados en la membrana	Energía libre de solvatación
Peso molecular	Incrementos de la constante de Kerr
Parámetros de flexibilidad normalizados (valores B), promedio	Valor de theta(i-1)
Frecuencia normalizada de la hélice alfa	Parámetro de van der Waals R0
Frecuencia normalizada del giro beta	Pesos para la lámina beta en la posición de ventana de 5
Frecuencia normalizada de hélice aislada	Pesos para la bobina en la posición de ventana de 4
Frecuencia normalizada de hélice alfa zurda	

El método CDA alcanza una puntuación ROC de 0,918 usando 72 índices, y 0,901 usando 35 índices. SVM-PCD (Webb-Robertson et al., 2010), que es un método que también utiliza distribuciones de propiedades fisicoquímicas, informa una puntuación ROC de 0,902 en SVM-PCD(531) y 0,906 en SVM-PCD(61). SVM-PCD(531) usa 531 índices

y 18 valores en cada distribución y por lo tanto, se calcula un total de 9558 valores. SVM-PCD(61) usa sólo 61 índices y un total de 1098 valores. El método CDA calculo 1440 valores cuando se usan 72 índices y 700 valores cuando se consideran 35 propiedades fisicoquímicas. A diferencia de SVM-PCD, el método CDA no concatena los valores calculados para entrenar una SVM. El método CDA usa 72 valores para entrenar un clasificador Bayesiano. El método CDA usa menos valores que el método SVM-PCD para hacer una clasificación. SVM-RQA (Yang et al., 2008) exhibe una puntuación ROC de 0,912. Este mapea cada aminoácido a un valor numérico usando 480 propiedades fisicoquímicas. Las propiedades fisicoquímicas se agrupan en una matriz de inclusión que hace parte del análisis de cuantificación de recurrencia. Finalmente, se extraen 10 valores de cada matriz de inserción. Se usan un total de 4800 valores para representar cada proteína. El método CDA es comparable con SVM-RQA en precisión y usa menos valores para representar cada proteína. Ambos, los métodos SVM-PCD y SVM-RQA, se probaron en el mismo conjunto de datos que utilizamos en los experimentos.

4. CONCLUSIONES

En este trabajo se propuso un nuevo método para la detección de homólogos remotos de proteínas. Este es llamado método CDA (Análisis de Distribución Característica) y se basa en representar cada secuencia de proteína mediante una distribución de 20 valores obtenidos a partir de los valores fisicoquímicos de los aminoácidos. Probamos la hipótesis de que cada familia SCOP tiene una distribución típica de sus secuencias. El método CDA utiliza distribuciones características para separar las secuencias en cada familia del resto de las proteínas en un conjunto de datos. Encontramos que hay propiedades fisicoquímicas que discriminan mejor las secuencias de una familia de proteínas. La propensión a la hélice alfa derivada del índice de secuencias diseñadas, el momento hidrofóbico basado en átomos y el parámetro hidrofóbico lograron los mejores resultados para muchas familias. Además, se encontró que un conjunto específico de índices era más adecuado para cada familia. El método CDA logra una tasa TP de 0,793, una tasa FP de 0,005, y una puntuación ROC de 0,918. La reducción de la dimensionalidad también mostró resultados importantes. Un conjunto de 35 índices logró una tasa TP de 0,754, una tasa FP de 0,006, y una puntuación ROC de 0,901.

El método CDA requiere menos valores para representar una proteína que los métodos SVM-PCD y SVM-RQA y presenta valores de precisión comparables. El método CDA podría mejorarse al agregar información evolutiva de los perfiles de frecuencia. De acuerdo con Liu et al. (2012), el uso de una estrategia basada en perfiles aumenta la precisión en los métodos de detección de homólogos remotos.

REFERENCIAS

Bedoya, Oscar; Tischer, Irene (2014). Remote homology detection incorporating the context of physicochemical properties. Computers in Biology and Medicine, vol. 45, no. 1, pp. 43-50. ISSN: 0010-4825.

Chitraranjan, Charith; Alnemer, Loai; Al-Azzam, Omar; Salem, Saeed; Denton, Anne; Iqbal, Muhammad and Kianian, Shahryar (2011). Frequent Substring-Based Sequence Classification with an Ensemble of Support Vector Machines Trained using Reduced Amino Acid Alphabets. Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference, vol. 2, no. 1, pp.180-185.

Dong, Qi-Wen; Wang, Xiao-long; Lin, Lei (2006). Application of latent semantic analysis to protein remote homology detection. Bioinformatics vol. 22, no. 3, pp. 285–290.

Gao, Feng. Indexing methods for protein tertiary and predicted structures. PhD dissertation. 2006.

Goldstein, Richard and Qian, Bin (2004). Performance of an Iterated T-Hmm for Homology Detection. Bioinformatics, vol. 20, no. 14, pp. 2175-2180.

Grigoriev, Igor and Kim, Sung-Hou (1999). Detection of protein fold similarity based on correlation of amino acid properties. Proceedings of the National

Academy of Sciences of the United States of America (PNAS), vol. 96, no. 25, pp. 14318-14323;

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. The WEKA Data Mining Software: An update. SIGKDD Explorations, Vol. 11, Issue 1. 2009.

Homaeian, Leila; Kurgan, Lukasz; Ruan, Jishou; Cios, Krzysztof and Chen, Ke (2007). Prediction of protein secondary structure content for the twilight zone sequences. Proteins: Structure, Function, and Bioinformatics, vol. 69, no. 3, pp. 486–498.

Hou, Yuna; Hsu, Wynne; Lee, Mong Li and Bystroff, Christopher (2003). Efficient Remote Homology Detection Using Local Structure. Bioinformatics, vol. 19, no. 17, 2003, pp. 2294-2301.

Huang, Yao-ming and Bystroff, Christopher (2006). Improved pairwise alignments of proteins in the Twilight zone using local structure predictions. Bioinformatics, vol. 22, no. 4, pp. 413–422.

Jaakkola, Tommi; Diekhans, Mark and Haussler, David (2000) A discriminative framework for detecting remote protein homologies. Journal of Computational Biology, vol. 7, no. 1, pp. 95–114.

Kawashima, Shuichi; Pokarowski, Piotr; Pokarowska, Maria; Kolinski, Andrzej; Katayama, Toshiaki and Kanehisa, Minoru (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Research, vol. 36, no. 1, pp. 202-205.

Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X. Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection. PLoS ONE 7(9): e46633. (2012).

Liao, Li and Noble, William (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. Journal of Computational Biology, vol 10, no. 6, pp. 857-868.

Muda, Hilmi; Saad, Puteh; Othman, Razib (2011). Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. Computers in Biology and Medicine. vol. 41, no. 1, pp. 687–699.

Orengo, Christine and Taylor, Willie (1996): SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol vol. 266, no. 1, pp. 617-635.

Rabiner, L. and Biing-Hwang, J. An introduction to hidden Markov models. ASSP Magazine, IEEE 3.1 (1986): 4-16.

Webb-Robertson, Bobbie-Jo; Ratuiste, Kyle and Oehmen, Christopher (2010). Physicochemical property distributions for accurate and rapid pairwise protein homology detection. BMC Bioinformatics, vol. 11, no.1 pp. 145-183.

Yang, Yuchen; Tantoso, Erwin and Li, Kuo-Bin (2008). Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties. Journal of Theoretical Biology, vol. 252, no. 1, pp. 145–154.

PARA CITAR ESTE ARTÍCULO / TO REFERENCE THIS ARTICLE / PARA CITAR ESTE ARTIGO /

Bedoya, Ó. (2017). Detección de homólogos remotos usando propiedades fisicoquímicas. *Revista EIA*, 14(27), enero-junio, pp. 111-125. [Online]. Disponible en: https://doi.org/10.24050/reia.v14i27.1161