Vol. 17, n° 2, pp. 75-86, 2018
Revista UIS Ingenierías
Página de la revista: revistas.uis.edu.co/index.php/revistauisingenierias

Predicción estructural de proteínas usando técnicas de clasificación

Protein structure prediction using classification techniques

Christian Charry-Ceballos[1], Óscar Bedoya-Leiva[2]

[1] Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle. Email: chriseth@gmail.com
[2] Bioinformática y Biocomputación, Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Colombia. Email: oscar.bedoya@correounivalle.edu.co

Recibido: junio 22, 2017. Aceptado: diciembre 27, 2017. Versión final: febrero 27, 2018.

Resumen

En este artículo se presenta un nuevo enfoque para la predicción estructural de proteínas. A diferencia de los métodos que existen actualmente, en este trabajo se propone un enfoque en el que se utilizan algoritmos de clasificación basados en aprendizaje supervisado para efectuar la predicción estructural. El desempeño del método propuesto se compara con el de métodos tradicionales como *local feature frequency* (LFF) usando el conjunto de datos Scop 2,05. Los resultados obtenidos indican que hay una diferencia favorable para el método propuesto y alcanzan porcentajes de proteínas correctamente clasificadas de 92,13 %, en cuanto a clase; 96,32 %, pliegue; 93,05 %, en superfamilia, y 76,35 %, en familia. Así se superan métodos tradicionales como LFF, que obtiene porcentajes de 85,90 %, 90,54 %, 79,85 % y 67,38 % para los mismos niveles estructurales.

Palabras clave: bioinformática; clasificadores; predicción estructural; proteínas; Scop.

Abstract

In this paper, a new protein structure prediction method is presented. Unlike current methods, this work introduces an approach based on supervised classification algorithms during the protein structure prediction. The accuracy of the proposed method was compared to traditional methods such as LFF (Local Feature Frequency) when using the Scop 2,05 dataset. The results indicate that there is a significant difference between these two methods. The proposed method reaches accuracy values of 92.13 %, 96.32 %, 93.05 %, and 76.35 %, at class, fold, superfamily, and family levels, respectively, and the LFF method reaches accuracy values of 85.90 %, 90.54 %, 79.85 % and 67.38 %, for the same structural levels.

Keywords: bioinformatics; classifiers; structural prediction; proteins; Scop.

1. Introducción

Las proteínas son cadenas complejas de aminoácidos que cumplen diversos papeles dentro del funcionamiento de los seres vivos. Estas a su vez interactúan con otras biomoléculas, con el fin de efectuar funciones de control, regulación, transporte, recepción, defensa, entre otras. Con el desarrollo de nuevas tecnologías, métodos y técnicas en el campo de la biología molecular y estructural, se ha venido presentando el problema de tener grandes volúmenes de información de estructuras proteicas que se han almacenado en diversas bases de datos. Tal es el caso del Banco Mundial de Proteínas (PDB), que hace seis años contaba con aproximadamente 36.000 estructuras almacenadas y a la fecha cuenta con más de 130.365 (mayo de 2017). Además del PDB, se

ISSN impreso: 1657 – 4583, ISSN en línea: 2145 – 8456

tienen bases de datos estructurales que permiten organizar las proteínas de acuerdo con criterios como la similitud en función, la identidad en cuanto a secuencias, la similitud en estructuras secundarias, entre otros criterios. Actualmente existen bases de datos estructurales como Scop (Structural Classification of Proteins), Cath (Class, Architecture, Topology and Homologous superfamily) y FSSP (Families of Structurally Similar Proteins) que son construidas utilizando diferentes aproximaciones respecto al método empleado para la clasificación estructural. Scop cuenta con un método de clasificación que es completamente manual y en el que un experto se encarga de asignar las proteínas en las clases estructurales. Cath es una base de datos que combina una parte manual y otra realizada de manera automática por algoritmos de alineamiento. Finalmente, FSSP utiliza un método completamente automatizado.

La base de datos Scop está organizada mediante una jerarquía de cuatro niveles (clase, pliegue, superfamilia, y familia). Cada proteína tiene asignado un código Scop que indica la clase, el pliegue, la superfamilia y la familia a la que pertenece. Por ejemplo, en la etiqueta a.1.2.1, la letra 'a' corresponde a la clase, 'a.1' al pliegue, 'a.1.2' a la superfamilia y 'a.1.2.1' a la familia. En el caso de las dos etiquetas a.1.1.1 y a.1.1.14, sus códigos indican que comparten la misma clase (p. e. clase a), pliegue (p. e. pliegue a.1) y superfamilia (p. e. superfamilia a.1.1), pero que pertenecen a familias diferentes. Cada nivel en la jerarquía Scop tiene unas condiciones específicas [1]. Por ejemplo, dos proteínas que pertenecen a la misma superfamilia-Scop tienen la misma función, mientras que dos proteínas en la misma familia-Scop tienen la misma función y una alta identidad en cuanto a la secuencia. En este trabajo se aborda el problema de la predicción estructural de proteínas. Este problema consiste en tomar una proteína y asignar de forma automática la clasificación estructural a la que pertenece en una base de datos determinada. Por ejemplo, en el caso de la jerarquía Scop, se intenta asignar una proteína en los cuatro niveles que la conforman (clase, pliegue, superfamilia, y familia). La importancia de abordar este problema está en reemplazar un proceso que en el caso de Scop se hace de forma manual y que al hacerse de forma automática permitiría tener un proceso más rápido de asignación, al tiempo que se intenta conservar la exactitud del experto.

Para el problema de la predicción estructural de proteínas se han propuesto diversas estrategias [2-11]. El método LFF [2] usa submatrices de las matrices de distancias como fragmentos estructurales. En una matriz de distancias se calculan las distancias euclidianas entre los carbonos alfa de los aminoácidos que forman la proteína. Un total de 100 submatrices son usadas para describir las interacciones que ocurren en la representación 3D de la proteína, para encontrar similitudes entre proteínas. El método SSEF [5] usa tripletas de elementos de estructura secundaria como fragmento estructural. Estos fragmentos se representan como un conjunto de ángulos y distancias entre los vectores SSE (*secondary structure elements*) correspondientes. El método SEGF [9] usa segmentos contiguos de 32 residuos a lo largo de la proteína como fragmento estructural. Sobre cada fragmento se obtiene un conjunto de 14 descriptores de forma [12]. Los descriptores utilizados usan invariantes geométricas y conceptos de teoría de nudos.

En este artículo se propone un nuevo método para la predicción estructural de proteínas que usa clasificadores. Los métodos clásicos de predicción estructural de proteínas se basan en utilizar el algoritmo de k vecinos más cercanos. De esta forma, la clasificación estructural de una proteína objetivo p es predicha al compararla con un conjunto de proteínas cuya clasificación Scop sí es conocida. En este trabajo se propone construir un modelo de clasificación que reemplace el algoritmo de k vecinos más cercanos. La hipótesis que se plantea es que se puede mejorar la exactitud de la predicción estructural de proteínas al usar algoritmos de clasificación que se basan en aprendizaje supervisado. El resto de este artículo está dividido en secciones. En la sección 2 se presenta el método propuesto. En la sección 3 se presentan y se discuten los resultados de las pruebas. En la sección 4 se muestran las conclusiones. Finalmente, en la sección 5 se presentan las recomendaciones.

2. Metodología

El método propuesto permite realizar la predicción estructural de proteínas mediante técnicas de clasificación, es decir, permite conocer de forma automática la clasificación Scop para una proteína objetivo. Inicialmente se recupera la información estructural de los archivos del PDB y se representa cada proteína como una matriz de distancias. Luego, se obtiene un conjunto de modelos. Cada modelo corresponde a una submatriz típica que puede ocurrir en las matrices de distancias. Estos modelos o submatrices se obtienen usando un algoritmo de *clustering*. Finalmente, se utilizan diferentes técnicas de clasificación para efectuar la predicción estructural de las proteínas. A continuación, se describen cada uno de los pasos del método propuesto.

2.1. Representación de proteínas usando matrices de distancias

La información estructural de cada proteína es procesada y transformada para crear matrices de distancia. Una matriz de distancia contiene las distancias euclidianas

entre cada par de átomos que componen una proteína. Dado que las distancias que son consideradas más relevantes son las de átomos cercanos, se hace un truncamiento de distancia máxima de 20 angstroms entre cada par de átomos. Una forma de representar gráficamente las matrices de distancias es a través de los mapas de calor. En estos mapas los valores numéricos se representan por medio de colores. Los mapas de calor permiten diferenciar visualmente algunos elementos típicos de la estructura tridimensional de las proteínas, por ejemplo, las hélices se pueden identificar como tiras adyacentes a la diagonal principal; las hojas beta, como diagonales cruzadas o paralelas a ella, y las espirales, como saltos en la diagonal entre las estructuras de las diagonales cruzadas.

En la figura 1 se muestra la matriz de distancia de la proteína 1auga que cuenta con una secuencia de 212 aminoácidos. En este caso se tiene una matriz de 212x212 valores. En esta figura se puede ver la presencia de los espacios vacíos (en azul), que indica la falta de contacto entre átomos, y la información correspondiente al esqueleto de la proteína (naranja oscuro). A continuación, se explica el proceso de representación de proteínas usando matrices de distancias.

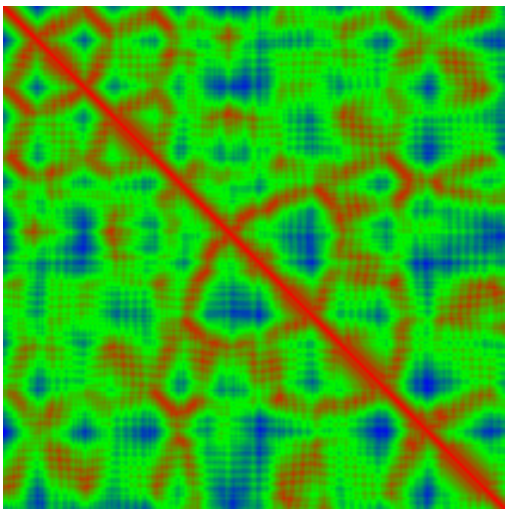

Figura 1. Matriz de distancia de la proteína 1auga visualizada como escala de color. **Fuente**: elaboración propia.

Formalmente, la matriz de distancias para una proteína p de n residuos, se calcula como $D_p=\{d_p(i,j): i,j = 1,\ldots,n_p\}$, donde cada posición de la matriz $d_p(i,j)$ es la distancia euclidiana entre los carbonos alfa ($C\alpha$-$C\alpha$) entre cada par de residuos i y j con base en sus coordenadas atómicas.

2.2. Obtención de los modelos

A pesar de que las proteínas pueden tener diferentes formas 3D y, por consiguiente, matrices de distancia diferentes, según [2], existen submatrices de las matrices de distancias que son típicas y que, por lo tanto, se pueden encontrar en diferentes proteínas. Estas submatrices típicas se usan como modelos para representar las proteínas. Una vez se cuenta con la información estructural de las proteínas en forma de matrices de distancia, se extraen submatrices solapadas de ellas y se aplica un algoritmo de *clustering*.

Para obtener los modelos que se usan en la predicción estructural se utiliza un conjunto de proteínas con sus respectivas submatrices y se aplica el algoritmo de *clustering* Clara (*clustering for large applications*) [13], seguido del análisis PAM (*partitioning around medoids*) para así obtener los l medoides de cada proteína (l = 50 en este caso). En particular, se utiliza un algoritmo de *clustering* porque este es capaz de dividir un conjunto de datos en grupos, de tal forma que los datos de cada grupo sean similares. Al usar un algoritmo de *clustering* se obtiene, además de los grupos, un conjunto de medoides. Cada grupo tiene asignado un medoide que representa la configuración típica de los datos del grupo al que pertenece. En el método propuesto se utilizan como modelos los medoides que resultan del algoritmo de *clustering*.

Para obtener los modelos se inicia seleccionando al azar un conjunto no redundante de proteínas P. Luego, para cada proteína *p* del conjunto P se calcula su matriz de distancia D_p y de esta se obtienen submatrices solapadas de tamaño *mxm*. Las submatrices solapadas se pueden representar como la colección $\delta p^{(m)}=\{\delta p^{(m)}[i,j]:i,j=1,\ldots,n_p-m+1\}$, de matrices m x m descritas por $\delta p^{(m)}[i,j]= \{d_p(i',j'):i' = i:(i+m -1), j' = j:(j+m-1)\}$. En este trabajo se utilizan 100 proteínas y matrices de tamaño 10x10, esto es $P=100$ y m=10. Estas cantidades son las mismas que se usaron en [2], ya que según los resultados reportados permiten obtener modelos representativos a partir de las matrices de distancia. Luego, se agrupan todas las submatrices de las P proteínas (en este caso 5.000 submatrices) y se aplica nuevamente el algoritmo Clara, junto con el análisis PAM para *k* grupos, donde cada grupo es representado por un medoide, y así se obtienen los *k* modelos. Estos modelos se usan como referencia para la comparación de similitud a nivel de estructuras secundarias o terciarias presentes en cada submatriz.

En la figura 2 se ilustra como ejemplo los modelos generados con k=10 y m=10, donde cada uno corresponde a una submatriz que representa conformaciones 3D típicas que ocurren en diferentes proteínas. Los modelos representados con mapas de calor permiten identificar visualmente algunas estructuras tridimensionales típicas de las proteínas. Por ejemplo, las líneas rojas indican el esqueleto de la proteína y las áreas naranja y amarilla indican las hojas y las hélices.

Finalmente, los espacios en verde corresponden a las áreas sin estructuras. A pesar de que en la figura se muestran solamente 10 modelos, para efectuar la predicción estructural de proteínas que se propone en este trabajo, se utiliza un total de 100 submatrices tal como se hace en [2].

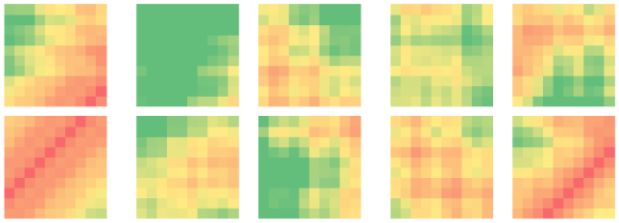

Figura 2. Representación de 10 modelos de tamaño 10x10. **Fuente**: elaboración propia.

2.3. Generación de los perfiles

Para generar los perfiles se utilizan los *k* modelos obtenidos con el proceso de *clustering*. En este trabajo se utilizan k=100 modelos. Para cada proteína *p* se obtiene su mapa de distancia D_p y las submatrices solapadas. Luego, se toma cada submatriz s y se calcula la distancia euclidiana entre s y los k modelos. Una submatriz s se asigna al modelo cuya distancia sea menor. Un modelo intenta representar las distancias que ocurren en cada submatriz. Finalmente, se obtiene un vector de tamaño k que indica la cantidad de veces que ocurre cada modelo en una proteína p. Este vector se conoce como el perfil de una proteína. Ya que los k modelos que se usan para calcular el perfil de cada proteína son conformaciones 3D típicas, se espera que dos proteínas que tengan similitud en su forma 3D tengan conteos que sean similares, es decir, que los modelos ocurran en las mismas cantidades. En los modelos obtenidos se tiene uno que aparece muchas más veces que los demás. Este modelo es el modelo vacío, esto es, aquel que no tiene contactos.

Con el objetivo de que el modelo vacío no afecte el cálculo del perfil de una proteína, se debe efectuar un proceso de normalización. Para esto, se calcula el denominador de cada uno de los *k* modelos sumando los cuadrados de los conteos de cada proteína *p* para el mismo modelo y por último se calcula su raíz cuadrada. Formalmente, para normalizar el i-ésimo valor del vector de conteo de la proteína p se utiliza la ecuación (1).

$$A_{pi} = \frac{f(p,i)}{\sqrt{\sum_{p' in\ D} f^2(p',i)}} \qquad (1)$$

donde f(p,i) es el valor de conteo sin normalizar del i-ésimo modelo de la proteína p y p´ representa cada proteína en el conjunto de datos D.

2.4. Predicción estructural usando algoritmos de clasificación

La clasificación Scop usa diferentes identificadores, uno de ellos es el sccs (scop(e) *concise classification string*), que consiste en una notación separada por puntos que indica los niveles de la organización jerárquica de proteínas en orden descendente. Estos niveles son clase, pliegue, superfamilia y familia.

En esta investigación se propone construir un clasificador para cada nivel estructural de la jerarquía Scop. En particular, se toma el identificador sccs para conocer la etiqueta de clase que se quiere predecir. Por ejemplo, la proteína d3rofa_ tiene como identificador sccs c.44.1.0. Por lo tanto, se utilizan las etiquetas c.44.1.0, para el nivel de familia; c.44.1, para superfamilia; c.44, para pliegue, y c, para clase. Otro ejemplo es la proteína d1iiba_ que tiene como identificador sccs c.44.2.1. Estas dos proteínas pertenecen a la misma clase y al mismo pliegue, pero pertenecen a superfamilias y familias distintas, como lo muestran los identificadores.

En cada clasificador que se construye para cada uno de los cuatro niveles estructurales se utilizan los conteos normalizados. De esta manera, cada proteína se representa en el conjunto de datos por medio de 100 valores y una etiqueta de clase que corresponde al identificador sccs, según el nivel estructural en el que se esté trabajando. El objetivo de cada clasificador es predecir el identificador sccs solamente a partir de los 100 valores que representan a la proteína. Debido a que en este trabajo se propone utilizar una estrategia de aprendizaje supervisado, se debe tener un conjunto de entrenamiento por medio del cual se construya un modelo de clasificación. Es decir, se requiere tener un conjunto de entrenamiento que represente cada etiqueta de clase. Por esta razón es necesario usar un umbral que garantice una cantidad mínima de proteínas para cada nivel estructural. A nivel de clase-Scop se utilizan las siete categorías presentes en el conjunto Scop 2.05. Sin embargo, a nivel de pliegue-Scop se utiliza el umbral r=200, es decir, solamente se utilizan pliegues-Scop con 200 proteínas o más. Al realizar este filtro se obtienen ocho pliegues (p.e. c.47, c.37, a.4, d.58, b.1, c.23, c.2, y c.1). Para el nivel de superfamilia se utiliza el umbral r=50, y se obtiene un total de 42 superfamilias-Scop. Las superfamilias y la cantidad de proteínas se muestran en la tabla 1.

Tabla 1. Superfamilias seleccionadas y su cantidad de proteínas con r=50.

Superfamilia	Número de proteínas	Superfamilia	Número de proteínas
d.144.1	95	c.67.1	112
b.18.1	66	d.108.1	91
b.1.18	83	c.69.1	115
c.23.1	80	g.37.1	72
c.3.1	74	a.39.1	64
a.4.1	110	g.39.1	67
a.4.5	219	c.108.1	97
b.1.2	116	c.55.3	50
b.1.1	200	b.40.4	114
d.17.4	66	b.34.2	66
c.66.1	135	c.93.1	71
d.58.7	113	d.38.1	70
c.47.1	199	c.2.1	296
c.55.1	88	c.37.1	289
a.1.1	52	a.25.1	54
b.55.1	82	c.1.9	55
a.45.1	50	c.1.8	115
b.47.1	52	b.29.1	78
c.94.1	144	c.14.1	65
d.15.1	65	b.82.1	62
b.36.1	84	c.1.10	60

Fuente: elaboración propia.

Para el nivel de familia se utiliza un umbral de r=20 obteniendo un total de 41 familias-Scop. Las familias y la cantidad de proteínas se muestran en la tabla 2.

Para cada nivel de la jerarquía Scop se realiza una clasificación independiente. Por ejemplo, se debe obtener un clasificador para predecir la clase-Scop y otro para predecir la familia-Scop. En este trabajo se experimenta con diferentes técnicas de clasificación basadas en aprendizaje supervisado. En total se consideran 18 algoritmos de clasificación en cada uno de los cuatro niveles estructurales. Los algoritmos que se utilizan en las pruebas se dividen en las siguientes cinco categorías:

- Bayesianos: bayes *net*, *naive* bayes.
- Basados en funciones: *simple logistic*, *logistic R*, *multilayer perceptron*, SMO.
- Lazy learning: *kstar*, IBK.
- Meta clasificadores: ASC, *bagging*, *classification via regression*, *MultiClass classifier*, *random committee*, *random subspace*.
- Basados en árboles: *random forest*, LMT, J48, *random tree*.

Tabla 2. Familias seleccionadas y su cantidad de proteínas asociada con r=50.

Familia	Número de proteínas	Familia	Número de proteínas
b.34.2.1	35	c.66.1.0	34
b.34.2.0	31	c.108.1.0	42
c.3.1.5	38	c.47.1.0	102
d.108.1.1	51	a.121.1.1	34
a.104.1.0	31	c.23.1.0	52
c.67.1.0	59	d.58.7.0	41
c.93.1.0	55	d.58.7.1	69
c.94.1.1	51	b.1.2.1	97
c.94.1.0	90	b.55.1.1	42
c.2.1.0	119	g.37.1.1	52
c.2.1.3	30	c.37.1.8	44
c.2.1.2	63	b.1.1.1	43
d.15.1.1	30	c.69.1.0	36
b.36.1.0	33	a.4.5.0	44
b.36.1.1	41	a.4.1.9	33
d.144.1.0	52	g.39.1.3	40
d.144.1.7	33	c.37.1.0	63
b.40.4.5	43	c.37.1.1	30
c.1.11.0	31	b.1.1.0	80
c.37.1.19	38	b.1.1.4	53
c.14.1.0	45		

Fuente: elaboración propia.

El método propuesto realiza los pasos 2.1, 2.2 y 2.3 de la misma forma que el método LFF, es decir, ambos son métodos basados en proyección en los que una proteína se representa como un vector de conteo que indica la cantidad de veces que un conjunto de modelos ocurren en su matriz de distancia. La diferencia entre el método propuesto y el método LFF está en el paso 2.4, en el cual se realiza la predicción estructural. Mientras que el método LFF utiliza el algoritmo de k vecinos más cercanos, en el método propuesto se utilizan algoritmos de clasificación supervisada. En este caso, para cada nivel estructural de la jerarquía Scop se construye un modelo que permite conocer la etiqueta de clase con base en un conjunto de entrenamiento. Por otra parte, el

método LFF no construye un clasificador usando aprendizaje automático, sino que la predicción para una proteína se hace calculando la similitud a nivel de los vectores conteo.

3. Resultados

Las pruebas incluyen el uso de diferentes clasificadores en cada uno de los niveles de la jerarquía Scop así como una comparación con el método LFF, el cual ha sido ampliamente usado para el problema de la predicción estructural de proteínas. Para la comparación de resultados se utiliza como medida de exactitud el porcentaje de proteínas correctamente clasificadas. Los valores reportados en esta sección se obtuvieron usando el mismo conjunto de prueba tanto para el método LFF como para el método propuesto. El método propuesto requiere un conjunto de entrenamiento mientras que el método LFF no utiliza un proceso de aprendizaje.

3.1. Conjunto de datos

En esta investigación se utiliza la base de datos estructural Scop. En particular, se usa la versión 2.05 que cuenta con 13.270 proteínas divididas en 7 clases, 1.200 pliegues, 1.973 superfamilias y 4.438 familias. Para realizar las pruebas se creó una serie de *scripts* escritos en BioPython, los cuales implementan el método descrito en este trabajo, desde el preprocesamiento hasta la creación de los perfiles que se usan para representar las proteínas. Posteriormente se hace uso de la implementación disponible en Weka [14] para los 18 algoritmos de clasificación.

Se utiliza como medida de exactitud el porcentaje de proteínas correctamente clasificadas. Una proteína se considera correctamente clasificada si se acierta en el nivel estructural correspondiente. El porcentaje de proteínas correctamente clasificadas se obtiene al dividir la cantidad de proteínas, cuyo identificador sccs se predice de manera correcta (para cada nivel de la jerarquía Scop), entre la cantidad de proteínas en el conjunto de prueba.

3.2. Resultados a nivel de clase-Scop

A nivel de clase-Scop los datos son fácilmente seleccionados para clasificación con siete etiquetas posibles debido al énfasis que hace Scop en usar únicamente las primeras siete clases. Esto permite una adecuada representación de los datos en el nivel más general de la jerarquía y genera condiciones equitativas para la comparación de resultados entre las técnicas de clasificación y el método LFF.

Para realizar las pruebas se utilizaron las 13.270 proteínas del Scop 2.05 y la técnica de validación cruzada (CVF) con k=5. Para la validación cruzada se crearon 5 conjuntos diferentes con los datos seleccionados y sobre cada uno se obtuvo el porcentaje de proteínas correctamente clasificadas. Con los resultados de estos conjuntos de datos se calculó el porcentaje promedio de proteínas correctamente clasificadas.

En la figura 3 se muestra uno de los modelos de clasificación obtenidos a nivel de clase-Scop. En particular, se muestra el árbol de decisión que se obtiene por medio del clasificador LMT. Como se puede ver, los nodos internos presentan comparaciones con algunos de los 100 valores que forman el vector de conteos normalizados. En las hojas se tienen las etiquetas de clase, que en este caso corresponden a las siete clases-Scop. La clasificación de una proteína cuya clase-Scop es desconocida ocurre al evaluar los valores que forman su vector de conteo normalizado y someterlo a las pruebas que hay en el árbol. De esta forma, la proteína tendrá la clase-Scop que alcance en la hoja correspondiente.

En la tabla 3 se muestra el porcentaje de aciertos a nivel de clase-Scop para los 18 algoritmos de clasificación y el método LFF. La tabla está organizada de mayor a menor de acuerdo con las exactitudes obtenidas. En este caso, métodos como IBK y *kstar* tienen exactitudes por encima del 90 %. Como se observa, los mejores resultados se alcanzan con las técnicas de *lazy learning* y de árboles, y los peores resultados, con las técnicas bayesianas. La comparación de los métodos permite observar la diferencia que existe entre el método LFF y los métodos de aprendizaje supervisado, que se proponen utilizar en este trabajo. Un total de 13 de los 18 algoritmos de clasificación alcanzan una exactitud mayor que el método LFF, y muestra que para la predicción estructural de proteínas a nivel de clase-Scop resulta más adecuado utilizar aprendizaje supervisado.

3.3. Resultados a nivel de pliegue-Scop

En la tabla 4 se muestra el porcentaje de aciertos a nivel de pliegue-Scop para los 18 clasificadores y el método LFF. Los porcentajes se calculan sobre los ocho pliegues (p.e. c.47, c.37, a.4, d.58, b.1, c.23, c.2, y c.1) que se obtienen al usar el umbral r=200. Como se puede observar, la posición que ocupa el método LFF se mantiene igual a la que obtuvo a nivel de clase-Scop. El desempeño entre técnicas indica que las mejores técnicas son *multilayer perceptron* y *simple logistic*, ambos basados en funciones, seguidos por el método LMT basado en árboles. Además, se puede observar que métodos como IBK y *kstar*, que alcanzaron las exactitudes más altas a nivel de clase-Scop, también

están entre las técnicas que obtienen los porcentajes más altos a nivel de pliegue-Scop.

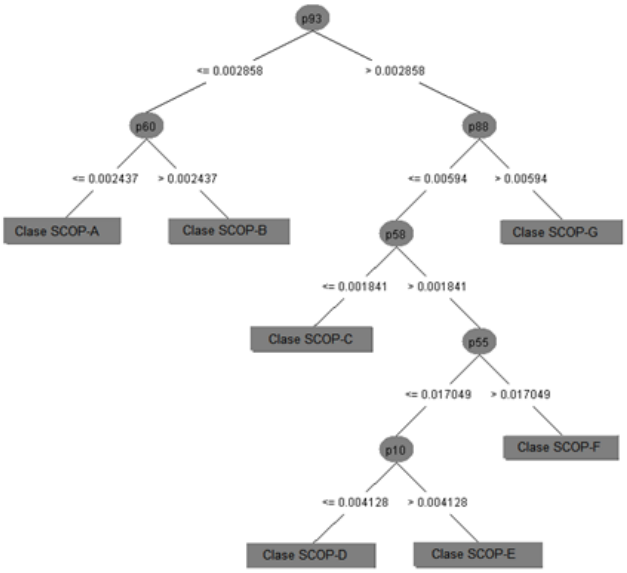

Figura 3. Visualización del modelo para LMT a nivel de clase-Scop. **Fuente**: elaboración propia.

Tabla 3. Algoritmos y porcentajes de proteínas correctamente clasificadas a nivel de clase-Scop.

Técnica	% Proteínas correctamente clasificadas
IBK	92,13 %
Kstar	92,11 %
Random forest	89,84 %
Random committee	89,09 %
LMT	88,40 %
Logistic R	87,58 %
Random subspace	87,58 %
Classification VIA *regression*	87,52 %
Bagging	87,21 %
Multiclass classifier	86,81 %
SMO	86,78 %
Simple logistic	86,75 %
Multilayer perceptron	85,82 %
LFF	85,60 %
ASC	83,49 %
J48	82,47 %
Random tree	80,94 %
Bayes net	61,86 %
Naive bayes	50,22 %

Fuente: elaboración propia.

3.4. Resultados a nivel de superfamilia-Scop

En la tabla 5 se muestra el porcentaje de aciertos a nivel de superfamilia-Scop. En la tabla se muestran los resultados obtenidos por los 18 clasificadores y el método LFF para el conjunto de prueba de las 42 superfamilias que se obtienen cuando se usa el umbral r=50. En las primeras cinco posiciones se encuentran los mismos algoritmos que alcanzaron los mejores resultados a nivel de pliegue-SCOP. De igual forma, el método LFF sigue siendo superado por 13 de los 18 algoritmos de clasificación. A nivel de superfamilia se nota una diferencia significativa entre las exactitudes de LFF y los mejores cinco algoritmos. Mientras que el método LFF tiene un porcentaje de aciertos de 79,85 %, el algoritmo *simple logistic* alcanza un exactitud de 93,05 %. Esta diferencia resulta importante para los biólogos quienes buscan detectar de forma automática las superfamilias de un conjunto de proteínas cuya función es desconocida.

3.5. Resultados a nivel de familia-Scop

En la tabla 6 se muestra el porcentaje de aciertos a nivel de familia-Scop para los 18 clasificadores y el método LFF en el conjunto de prueba de las 41 superfamilias que se obtienen cuando se usa el umbral r=30. La predicción a nivel de familia se vuelve más difícil que en los otros niveles estructurales. En este caso las exactitudes varían entre el 57,93 % y el 76,35 %.

Tabla 4. Algoritmos y porcentajes de proteínas correctamente clasificadas a nivel de pliegue-Scop.

Técnica	% Proteínas correctamente clasificadas
Multilayer perceptron	96,32 %
Simple logistic	96,28 %
LMT	96,10 %
IBK	95,70 %
Kstar	95,38 %
SMO	95,12 %
Random forest	94,76 %
Logistic R	93,29 %
Random committee	93,03 %
Random subspace	91,88 %
Classification VIA *regression*	91,88 %
Multiclass classifier	91,84 %
Bagging	90,90 %
LFF	90,54 %
J48	87,51 %
ASC	86,07 %
Random tree	84,84 %
Bayes net	75,72 %
Naive bayes	67,82 %

Fuente: elaboración propia.

3.6. Análisis de resultados

Durante las pruebas realizadas se observaron algunas técnicas de clasificación que fueron mejores que otras en cada nivel estructural. A continuación, se listan los mejores cinco algoritmos de clasificación en cada nivel de la jerarquía Scop:

- Clase: IBK, *kstar*, *random forest*, *random committee*, LMT.
- Pliegue: *multilayer perceptron*, *simple logistic*, LMT, IBK, *kstar*.
- Superfamilia: *simple logistic*, LMT, *multilayer Perceptron*, *kstar*, IBK.
- Familia: LMT, *simple logistic*, *kstar* , *random forest*, SMO.

Como se puede observar, las técnicas de clasificación *kstar* y LMT estuvieron presentes en los primeros lugares para los cuatro niveles analizados, seguidas del algoritmo IBK que aparece en tres de los cuatro niveles. Esto permite concluir la notable ventaja de usar técnicas basadas en *lazy learning* para la clasificación estructural de proteínas, además del algoritmo LMT, basado en árboles. En la tabla 7 se muestran los consolidados de los porcentajes de aciertos.

Tabla 5. Algoritmos y porcentajes de proteínas correctamente clasificadas a nivel de superfamilia-Scop.

Técnica	% Proteínas correctamente clasificadas
Simple logistic	93,05 %
LMT	92,87 %
Multilayer perceptron	91,80 %
Kstar	90,81 %
IBK	90,67 %
SMO	90,10 %
Random forest	88,69 %
Logistic R	88,24 %
Multiclass classifier	87,15 %
Random committee	84,89 %
Classiffication VIA *regression*	83,52 %
Random subspace	81,56 %
Bagging	81,39 %
LFF	79,85 %
J48	72,42 %
ASC	70,86 %
Random tree	65,60 %
Bayes net	65,58 %
Naive bayes	60,62 %

Fuente: elaboración propia.

Los resultados de las pruebas realizadas también permiten observar que los cinco mejores clasificadores alcanzan exactitudes superiores al método LFF. Para efectos de comparación, se calcula el promedio obtenido por los algoritmos para cada nivel estructural. Se puede observar que a pesar de que el porcentaje de proteínas correctamente clasificadas del método LFF se mantiene por encima de la media, su exactitud es menor que la obtenida por las cinco mejores técnicas de clasificación en cada nivel estructural. Durante las pruebas realizadas se pudo comprobar que se obtienen exactitudes más altas cuando se usan clasificadores basados en aprendizaje supervisado que al usar el método LFF. La diferencia se vuelve más significativa en niveles estructurales que son más difíciles de predecir como los niveles de superfamilia y familia.

Tabla 6. Algoritmos y porcentajes de proteínas correctamente clasificadas a nivel de familia-Scop.

Técnica	% Proteínas correctamente clasificadas
LMT	76,35 %
Simple logistic	76,35 %
Kstar	76,35 %
Random forest	75,51 %
SMO	75,36 %
IBK	75,07 %
Multilayer perceptron	75,07 %
Classification VIA *regression*	72,16 %
Random committee	71,33 %
Random subspace	70,68 %
Logistic R	69,80 %
Bagging	69,80 %
Multiclass classifier	68,71 %
LFF	67,38 %
J48	64,08 %
Bayes net	63,44 %
ASC	63,39 %
Naive bayes	58,71 %
Random tree	57,93 %

Fuente: elaboración propia.

Tabla 7. Porcentajes de proteínas correctamente clasificadas en cada nivel Scop.

Porcentaje de proteínas correctamente clasificadas				
Técnica	**Familia**	**Superfamilia**	**Pliegue**	**Clase**
IBK	63,39	90,67	95,70	92,13
ASC	69,80	70,86	86,07	83,49
Bagging	63,44	81,39	90,90	87,21
Bayes net	72,16	65,58	75,72	61,86
Classification VIA *regression*	75,07	83,52	91,88	87,52
J48	64,08	72,42	87,51	82,47
Kstar	76,35	90,81	95,38	92,11
LFF	67,38	79,85	90,54	85,60
LMT	76,35	92,87	96,10	88,40
Logistic R	69,80	88,24	93,29	87,58
Multiclass classifier	68,71	87,15	91,84	86,87
Multilayer perceptron	75,07	91,80	96,32	85,82
Naive bayes	58,71	60,62	67,82	50,22
Random committee	71,33	84,89	93,03	89,09
Random forest	75,51	88,69	94,76	89,84
Random subspace	70,68	81,56	91,88	87,58
Random tree	57,93	65,60	84,84	80,94
Simple logistic	76,35	93,05	96,28	86,75
SMO	75,36	90,10	95,12	86,78
Promedio	69,87	82,08	90,26	83,80

Fuente: elaboración propia.

4. Conclusiones

En este artículo se presentó un nuevo método para la predicción estructural de proteínas. El uso de aprendizaje supervisado para la predicción estructural de proteínas es un nuevo enfoque en los métodos automatizados que en este primer acercamiento ha demostrado mejores resultados que el enfoque tradicional por proyección. En promedio, se logra mejorar la predicción un 5,46 % para el nivel de clase; 5,41 % para pliegue; 11,99 % para superfamilia y 8,60 % para familia, en comparación con la predicción por proyección con el método LFF.

Entre las técnicas de clasificación que arrojaron los mejores resultados se encuentran, por un lado, el algoritmo *kstar* con 92,11 % a nivel de clase; 95,38 % para pliegue; 90,81 % para superfamilia y 76,35 % para familia. Y, por otro lado, se encuentra el algoritmo IBK con exactitudes de 92,13 % para clase; 95,7 % para pliegue; 90,67 % para superfamilia y 63,39 % para familia. Ambos algoritmos del tipo *lazy learning*. Según los resultados de las pruebas, los algoritmos *kstar* e IBK pueden ser utilizados para que los biólogos construyan clasificadores que permitan conocer la clasificación estructural de una proteína de acuerdo con la jerarquía Scop. De igual forma, los resultados de las pruebas permitieron probar la hipótesis que se planteó al inicio de la investigación, según la cual era posible mejorar la exactitud de los métodos de predicción estructural basados en el algoritmo k vecinos más cercanos al usar métodos de clasificación con aprendizaje supervisado.

5. Recomendaciones

En este trabajo se utilizó la base de datos estructural Scope. Actualmente, se encuentra en desarrollo una nueva versión de esta base de datos conocida como Scop2 [15], que se podría utilizar como conjunto de prueba. Esta versión de la base de datos incluye nuevas relaciones estructurales que se intentarían predecir con el método propuesto. Además, es posible extender el alcance de este trabajo a otras bases de datos estructurales como Cath y FSSP que son ampliamente usadas.

Además, se puede experimentar sobre otros parámetros que son relevantes para la predicción estructural de proteínas. Por ejemplo, el tamaño de las submatrices con las que se crean los modelos representativos de las matrices de distancia. Este tamaño representa cuánta información se captura de la matriz de distancia y, según [2], afecta la exactitud de los métodos de predicción estructural basados en modelos representativos.

Referencias

[1] A. Andreeva, D. Howorth, S. Brenner. T. Hubbard, C. Chothia, A. Murzin. " SCOP database in 2004: refinements integrate structure and sequence family data." Nucleic acids research, vol 32: D226-D229, 2004.

[2] I. Choi, J. Kwon, S. Kim. "Local Feature Frequency Profile: A Method to Measure Structural Similarity in Proteins." Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 101, no. 11, pp. 3797-3802, 2004.

[3] S. Gupta, J. Stamatoyannopoulos, T. Bailey, W. Stafford. "Quantifying similarity between motifs. Genome Biology", vol. 8, pg. R24, 2007.

[4] A. Narunsky, S. Nepomnyachiy, H. Ashkenazy, R. Kolodny, N. Ben-Tal. "ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure." Structure Volume 23, Issue 11, pp. 2162–2170, 2015.

[5] E. Zotenko, D. O'Leary, T. Przytycka. "Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification," BMC Structural Biology, pp. 1-12, 2006.

[6] X. Cui, S. Cheng, L. He, M. Li. "Fingerprinting protein structures effectively and efficiently." Bioinformatics 30 (7), pp. 949–955, 2014.

[7] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, A. Tramontano. "Critical assessment of methods of protein structure prediction Round VII." Proteins; 69(Suppl 8):3–9, 2007.

[8] A. Kryshtafovych, O. Krysko, P. Daniluk, Z. Dmytriv, K. Fidelis. "Protein structure prediction center in CASP8." Proteins;77(Suppl 9):5–9, 2009.

[9] E. Zotenko, R. Dogan, J. Wilbur, "Structural footprinting in protein structure comparison: the impact." BMC Structural Biology, pp. 1-14, 2007.

[10] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schewede. "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI." Proteins Volume 84, Issue S1, pp. 4–14, 2016.

[11] O. Bedoya, E. Satizábal. "NN-PRED: Un nuevo programa para la predicción de la estructura secundaria de la proteína usando redes neuronales," *Rev. UIS Ing*., vol. 12, no. 1, pp. 51-59, 2013.

[12] P. Rogen, B. Fain. “Automatic classification of protein structure by using Gauss integrals. “ Multidisciplinary journal covering the biological, physical, and social sciences, pp. 119-124, 2003.

[13] L. Kaufman, P. Rousseeuw. “Partitioning Around Medoids (Program PAM).” En L. Kaufman, & P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey, USA, 2008.

[14] I. Witten, F. Eibe, M. Hall, C. Jal. “Data mining. Practical machine learning tools and techniques.” Cambridge. MA: Morgan Kaufmann, 2017.

[15] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A. Murzin. “SCOP2 prototype: a new approach to protein structure mining.” Nucleic Acids Research, vol. 42, no. 18, pp. 310-314, 2014.