Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

Tipo de artículo: Artículo original
Temática: Bioinformática
Recibido: 20/05/2017 | Aceptado: 25/06/2017

Predictor de interacciones entre estructuras secundarias de proteínas

Predictor of interactions between secondary protein structures

Julio César Quintana-Zaez[1,2,*], Nicolás Quintana-Bernabé[1], Reinaldo Molina-Ruiz[3], Cosme E. Santiesteban-Toca [4,*]

[1]Facultad de Informática, Universidad "Máximo Gómez Báez" de Ciego de Ávila, Cuba. {jcquintana, nquintana}@unica.cu

[2]Centro de Investigaciones de la Informática, Universidad Central "Marta Abreu" de las Villas, Cuba.

[3]Centro de Bioactivos Químicos, Universidad Central "Marta Abreu de las Villas", Cuba. reymolina@uclv.edu.cu

[4]Departamento de Informática, Centro de Bioplantas, Ciego de Ávila, Cuba. cosme@bioplantas.cu

[*] Autor para correspondencia: jcquintana@unica.cu, cosme@bioplantas.cu

Resumen

Los métodos de predicción de mapas de contacto son un paso intermedio para la predicción de estructuras de proteínas. A pesar de los avances logrados la precisión de las predicciones continúa por debajo del umbral deseado. Una vía mediante la cual el desempeño de estos métodos puede ser elevado es realizando la predicción de las interacciones entre estructuras secundarias. En este artículo se realiza un estudio de la influencia de las interacciones en el plegamiento de las proteínas. Donde, se propone un novedoso meta multiclasificador basado en árboles de decisión para predecir dichas interacciones. El método consiste en un esquema que combina el resultado de diferentes multiclasificadores especializados en las interacciones en el mapa de contacto final. El conjunto de proteínas empleado para validar el modelo contó con 2020 elementos y fue dividido en cuatro particiones, con respecto a su tamaño. La capacidad de generalización promedio alcanzada para los cuatro grupos de proteínas es de 51% de precisión, con una sensibilidad de 74%. El mejor desempeño del algoritmo se logró en proteínas de tamaño medio donde se alcanzó un 55% de precisión.

Palabras clave: Mapa de contacto, multiclasificadores, árboles de decisión, interacciones, estructuras secundarias

Abstract

The methods for the prediction of contact maps are an intermediate step for the prediction of protein structures. Despite the progress made, the accuracy of the predictions continues below the desired threshold. One way in which the

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

performance of these methods can be heightened is by predicting the interactions between secondary structures. In this paper, we study the influence of interactions in the folding of proteins where we propose a novel multi-class goal based on decision trees to predict such interactions. The method consists of a scheme that combines the result of different specialized multiclassifiers into the interactions in the final contact map. The set of proteins used to validate the model counted on 2020 elements and were divided into four partitions, with respect to their size. The average generalization capacity achieved for all four protein groups is 51% accuracy, with a sensitivity of 74%. The best performance of the algorithm was attained in medium-sized proteins where 55% accuracy was achieved.

Keywords: *Contact map, multiclassifiers, decision trees, interactions, secondary structures*

Introducción

Conocer la estructura de las proteínas es relevante para la Bioinformática debido a que puede servir de base para el diseño de nuevos fármacos, la obtención de productos naturales, entre otros. Además, algunas de las enfermedades más agresivas que afectan la salud humana como el Alzhéimer están estrechamente relacionadas a problemas en el plegamiento de las proteínas (Cohen, 2004). Por otro lado, se conoce que la estructura de la proteína determina su función. Por tal razón, predecir la estructura de las proteínas se ha convertido en uno de los problemas más interesantes dentro de la Bioinformática y la Biología Computacional (Mitra, & Hayashi 2006).

Una de las vías estudiadas para dilucidar la estructura de las proteínas se basa la predicción de los contactos entre residuos o mapas de contacto (Márquez-Chamorro et al., 2015). Los mapas de contacto son una representación matricial de la estructura terciaria de las proteínas (Xie et al., 2015). Son considerados únicos para cada proteína, por lo que pueden ser empleados para identificar, agrupar o comparar proteínas (Andonov, Malod-Dognin, & Yanev, 2011). En las últimas décadas, se han desarrollado diversos métodos para la predicción de contactos entre residuos, entre los cuales se encuentran métodos basados en redes neuronales (Ding et al., 2013; Tegge et al., 2009), máquinas de soporte de vectores (Cheng y Baldi, 2007; Howe y Mohamad, 2011), modelos ocultos de *Markov* (Ashkenazy, Unger, y Kliger, 2011), Árboles de decisión (Santiesteban-Toca et al., 2012), aprendizaje profundo (o *Deep learning*) (P. Di Lena, Nagata, y Baldi, 2012) y algoritmos genéticos.

La información común empleada en el proceso de aprendizaje de los métodos para la predicción de mapas de contacto proviene de diferentes fuentes (Xie et al., 2015). De manera general, esta se relaciona con la secuencia de aminoácidos, la estructura secundaria, múltiples alineamientos, mutaciones correlacionadas, información evolutiva, entre otras. Donde, son empleados los patrones físicos y químicos de los residuos dentro de las internaciones entre estructuras

(Tegge et al., 2009), o se especializan en predecir contactos específicos (Randall et al., 2008; Karakaş, Woetzel, & Meiler, 2010; Di Lena, Nagata, & Baldi, 2012).

Debido a la diversidad de métodos desarrollados y esquemas de experimentación empleados para probar dichos métodos, es difícil establecer una línea que permita una comparación directa. Sin embargo, a pesar de los avances alcanzados por los métodos para predecir mapas de contacto, la precisión lograda por los algoritmos es relativamente baja con un promedio cerca del 50% para toda la proteína (Márquez-Chamorro et al., 2015). Y según la competición bi-anual, Evaluación Crítica para la Predicción de Contactos (*Critical Assesement for Contact Prediction or CASP*), la precisión para los contactos a largo rango, los cuales constituyen una tarea sumamente compleja por sus características (más de 24 residuos de separación en la secuencia), es alrededor de 20%-25% (Monastyrskyy et al., 2014).

Recientemente, se han desarrollado técnicas que emplean una representación espacial de la vecindad entre estructuras secundarias, residuos e incluso fragmentos de la proteína. Las cuales predicen los mapas de contacto entre estructuras secundarias (*coarse contact maps*), en lugar de los contactos entre residuos, (Wang, Zhu, & Cai, 2009; Di Lena et al. 2008). Es por ello que, en el presente artículo se propone un predictor, basado en la combinación de múltiples clasificadores, sustentado en la hipótesis de que las estructuras secundarias y sus vecindades son las responsables del 90% de los contactos entre residuos. Por tanto, si se predicen dichas interacciones, se puede elevar el nivel de precisión de la predicción de los mapas de contactos de proteínas.

En la primera sección del artículo se hace un estudio de la importancia de las interacciones entre estructuras secundarias dentro del plegamiento de las proteínas y se plantea la hipótesis que sustenta el modelo propuesto. Seguido, se describe el modelo propuesto y finalmente se analizan los resultados. Dónde: (1) se selecciona el mejor clasificador para predecir las interacciones entre estructuras secundarias, (2) se evalúa el desempeño del método propuesto para un dominio de aplicación y (3) se propone un mecanismo de interpretación. Por último, se presentan las conclusiones y trabajos futuros.

Materiales y métodos

El algoritmo propuesto en esta investigación se basa en la predicción de las interacciones entre estructuras secundarias. Donde para cada interacción se emplea un esquema especializado en los patrones físicos, químicos y espaciales de estas interacciones. Las interacciones entre regiones no estabilizadas (*coils-coils*), son excluidas.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

Base biológica de la propuesta

Las interacciones no-covalentes a lo largo de la secuencia juegan un papel importante en el plegamiento y la estabilidad de la proteína, (Gromiha, 2009). Donde, las interacciones entre estructuras secundarias contienen suficiente información acerca del plegamiento final de la proteína (Zaki, Shan Jin, & Bystroff, 2003). Por tal razón, se realizó un análisis del papel que desempeñan dichas interacciones, con el empleo de 2019 proteínas no-redundantes extraídas del banco de datos de proteínas (*Protein Data Bank, o PDB*) (Rose et al., 2013).

La Figura 1, muestra la representación tridimensional de una proteína (izquierda) y su mapa de contacto (derecha). Donde, las flechas de color amarillo simbolizan las láminas, las rosa a las hélices y los alambres de color blanco a las estructuras *coils*. Las fechas negras representan las interacciones entre las estructuras secundarias. A la derecha, en el mapa de contacto, se observan cómo están distribuidos los contactos entre residuos dentro de las interacciones. Donde, las bandas de color amarillo, rosa y blanco representan las estructuras secundarias. Los grupos de puntos azules son los contactos entre residuos dentro de las interacciones.

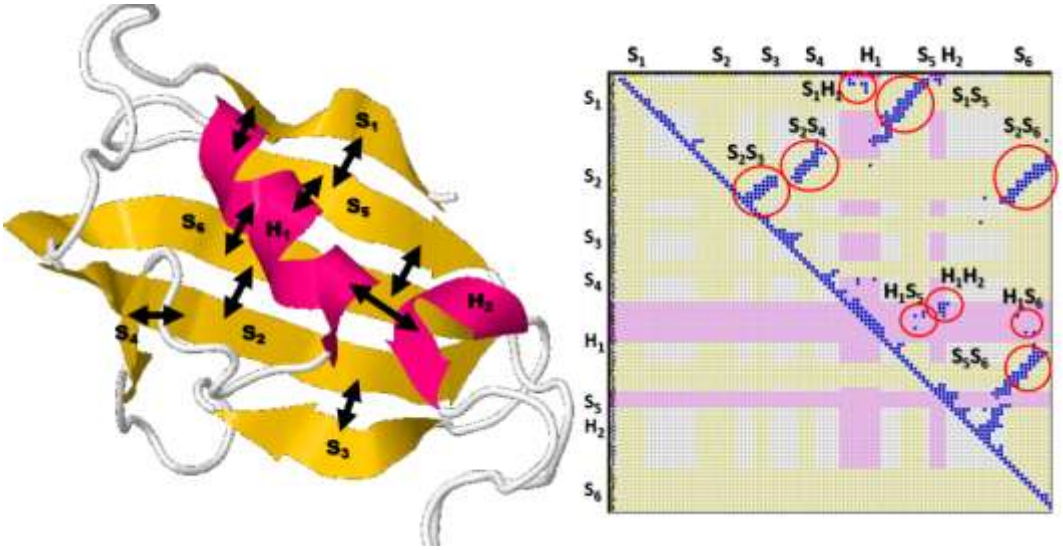

Figura 1. Representación de la proteína 1c9h. A la izquierda, se muestra su esquema tridimensional, a la derecha el mapa de contacto asociado.

Como puede observarse en la Figura 1, un gran porcentaje de las los contactos está relacionado con las interacciones entre estructuras secundarias. En esta investigación, cada estructura es tratada como una entidad única y como separación se toma en cuenta el número de estructuras intermedias, Figura 2.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

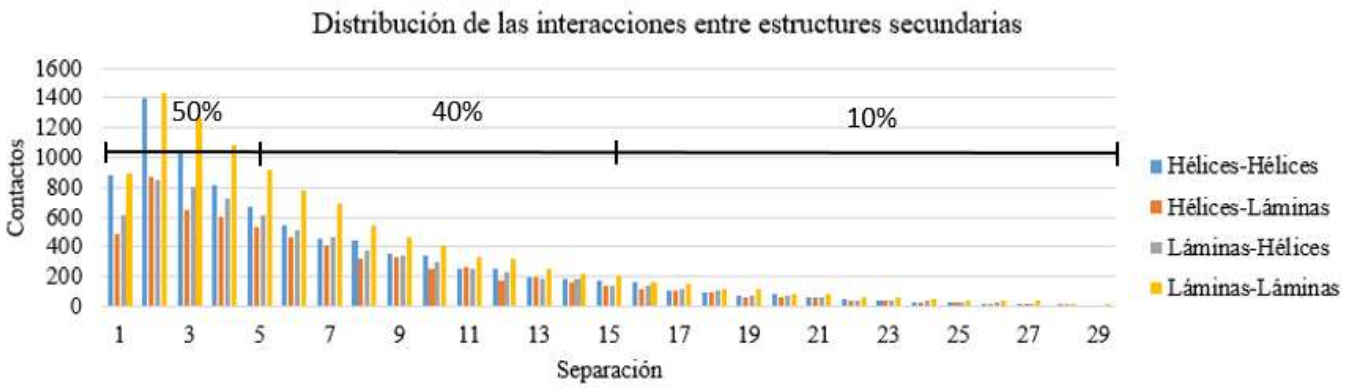

Figura 2. Distribución de las interacciones entre estructuras secundarias. Segmentos horizontales representan el porciento de las interacciones Hélices-Hélices, Hélices-Láminas, Láminas-Hélices, Láminas-Láminas (α–α, α–β, β–α, and β–β) para cada intervalo de separación.

La Figura 2, muestra la distribución de las interacciones entre estructuras secundarias (α–α, α–β, β–α, and β–β), en función de la separación en la secuencia. Como se puede observar cerca de 50% de las interacciones tiene lugar a menos de cinco estructuras de separación. El 90% de las interacciones ocurre entre las estructuras α–α, α–β, β–α, β–β, α–coil, β–coil, a un máximo de 15 estructuras de separación. Además, ocurren con más frecuencia las interacciones entre estructuras α–α and β–β, llegando a superar en aproximadamente un 30% al resto de las interacciones. En cambio, solo un 10% de los contactos ocurre en las interacciones coil-coil.

Modelo propuesto

Del análisis anterior se desprende que, si se predicen las interacciones entre estructuras secundarias, se puede inferir cerca del 90% de los contactos entre residuos. Sin embargo, el bajo nivel de precisión alcanzado por los algoritmos propuestos hasta la fecha podría estar relacionado con: (a) el error de respuesta de los clasificadores frente a datos que aún no le han sido presentados; (b) elementos en los clasificadores que los hacen caer en soluciones locales; (c) la incapacidad de dichos clasificadores en representar adecuadamente el problema. Todo lo cual justifica la combinación de varios clasificadores (Kuncheva, 2004; Francia, & García, 2006).

En el presente artículo se propone un esquema de combinación de clasificadores, basado en las interacciones entre las estructuras secundarias de las proteínas. El cual consiste en un grupo de múltiples clasificadores especializados que son seleccionados en dependencia del dato de entrada, por lo que pudiera ser considerado como un meta multiclasificador, Figura 3. Con este modelo se pretende obtener mayor provecho del manejo los datos de aprendizaje y de la combinación de los resultados de sus predicciones.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

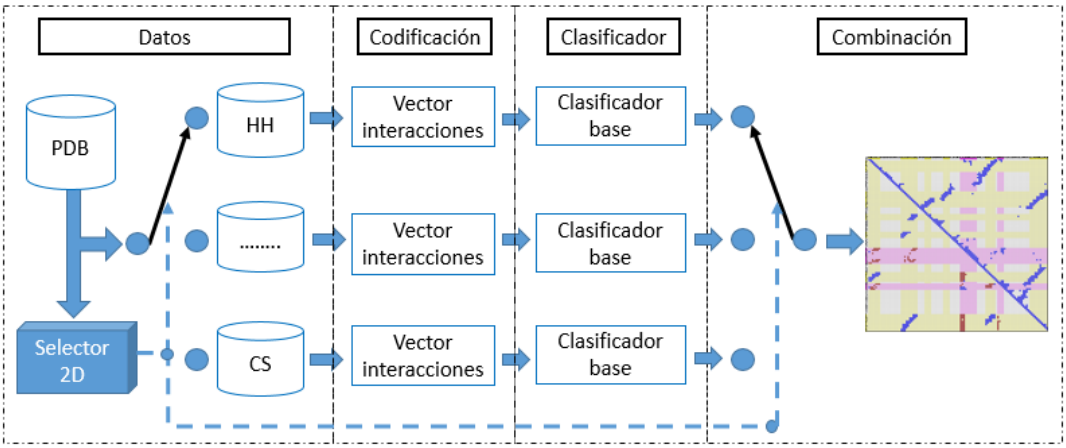

Figura 3. Modelo del meta multiclasificador propuesto. Cada clasificador base está compuesto por un multiclasificador especializado en el tipo de interacción en la que es entrenado.

Nivel de datos

La Figura 3, muestra la arquitectura del predictor propuesto, el cual consiste en cuatro niveles. Donde, a nivel de datos, se cuenta con un conjunto de proteínas extraídas del PDB. A cada proteína se le extraen las interacciones entre estructuras secundarias. Posteriormente, mediante el empleo de un criterio de selección, cada tipo de interacción es almacenada en conjuntos de datos específicos y disjuntos {α–α, α–β, α–coil, β–α, β–β, β–coil, coil–α, coil–β}. El objetivo de este nivel es dividir el problema en sub-problemas que permitan garantizar que cada modelo se especialice en un sub-espacio del problema general.

Nivel de vector

En el nivel de vector son creados los vectores de rasgos que describen las interacciones. En nuestra investigación las estructuras secundarias se representan como una entidad única con propiedades físicas y químicas (atributos), derivados de la frecuencia de los residuos que las integran. El modelo para describir dichas interacciones emplea un vector de 206 atributos de longitud, que incluye información sobre las estructuras objetivo y su vecindad (±1 estructura adyacente). La descripción de los atributos empleados en el vector de rasgos se muestra en la

Tabla **1**.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

Tabla 1. Descripción de los atributos empleados en el vector de rasgos. Cada estructura considerada cuenta con un conjunto de estos atributos.

Rasgo	Descripción	Tipo	Entradas
Hidrofobicidad	Distribución de hidrofobicidad (Hidrofóbicos, Hidrofílicos) (Márquez-Chamorro et al., 2014).	Numérico	2
Polaridad	Distribución de polaridad (Polares, No-Polares, Ácidos, Básicos) (Márquez-Chamorro et al., 2014).	Numérico	4
Átomos	Distribución de carga (Hidrogeno, Nitrógeno, Oxigeno, Carbón, Sulfuro) (Márquez-Chamorro et al., 2014).	Numérico	5
Frecuencia de residuos	Distribución de residuos de la estructura (Abu-Doleh, Al-Jarrah, & Alkhateeb, 2012)	Numérico	20
Tamaño de los residuos	Distribución del tamaño de los aminoácidos (Grandes, Pequeños)	Numérico	2
Tamaño de la estructura secundaria	Cantidad de residuos dentro de la estructura secundaria (Di Lena, Nagata, & Baldi, 2012)	Numérico	1
	Subtotal de rasgos para una estructura		**34**
	Total de rasgos para las dos estructuras objetivo y sus vecindades (6 x)		**204**
Separación	Numero de estructuras entre las estructuras objetivo	Numérico	1
Clase	Clase (Contacto o No-Contacto)	Nominal	1
	Total de Rasgos del vector		**206**

Nivel de clasificadores base

En el nivel de clasificadores base son empleados multiclasificadores idénticos, pero especializados en cada tipo de interacción (Figura 4). La razón fundamental por la que se decidió el empleo de nuevos multiclasificadores fue crear un esquema de entrenamiento que permitiese incrementar el espacio de búsqueda a la vez que se redujera el nivel de desbalance y solapamiento de los datos.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

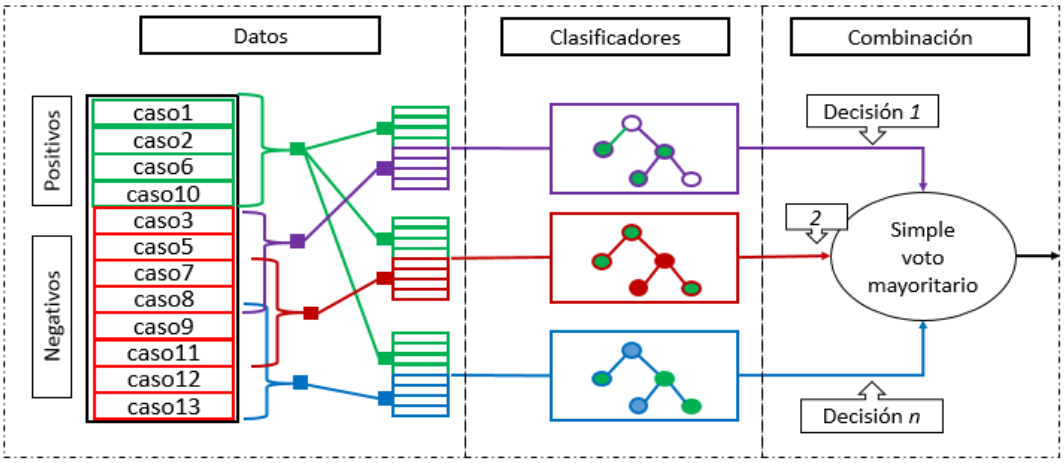

Figura 4. Modelo del multiclasificador base empleado.

En la Figura 4, se muestra el proceso de creación de los multiclasificadores base. Donde, debido al alto nivel de desbalance y solapamiento que existe, el conjunto de vectores inicial es dividido en *K* particiones impares mediante el empleo de un muestreo con remplazo. Cada subconjunto mantiene todos los elementos de la clase contactos y agrega elementos de la clase no-contacto, de forma que se logre una razón de desbalance de 1:2. Este proceso garantiza la diversidad de los multiclasificadores base propuestos. La combinación sus decisiones se realizan mediante el voto mayoritario simple.

Nivel de combinación

Finalmente, corresponde el nivel de combinación de los resultados del meta multiclasificador. A pesar de existir varios métodos de combinación de la clasificación, ya sean a nivel abstracto, de ranking o de medición (Kuncheva, 2004), se decidió por una combinación a nivel de medición. Esto se debe a que se conoce previamente cuál es la efectividad de cada multiclasificador clasificador para cada tipo de interacción (1).

$$D = \sum_{i=1}^{I} w_i * e_i \quad (1)$$

Donde, $\boldsymbol{D}$ es la decisión final, $\boldsymbol{w_i}$ el peso asignado por el selector, $\boldsymbol{e_i}$, la estimación realizada por los multiclasificadores e $\boldsymbol{I}$, el tipo de interacción. Donde, $\boldsymbol{i \in I}$ = **{**HH, HS, SH, SS, HC, CH, SC, CS**}**. Si se toma en cuenta que el selector devuelve como resultado ***1*** si la interacción es de un tipo y ***0*** para el resto. Entonces el resultado de la combinación sería equivalente a la selección del clasificador adecuado. Lo que da lugar a la construcción en el mapa de contacto de la proteína.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

Resultados y discusión

En este epígrafe se seleccionan los clasificadores base empleados, se analiza el desempeño del algoritmo en cuanto a su robustez y capacidad de generalización y la comparación con los clasificadores del estado-del-arte. Para evaluar los resultados del esquema experimental fueron empleadas la precisión, la sensibilidad y la media armónica entre ambas (García et al., 2010).

Selección de clasificadores base

Debido a que se desea obtener modelos que permitan explicar qué sucede en el proceso de plegamiento de las proteínas, solo se consideraron algoritmos basados en árboles de decisión (Wu, & Kumar, 2009). Los algoritmos empleados son las implementaciones de Weka de J48, *LADTree*, *RandomForest*, *RandomTree*, *REPTree*, *ADTree*, *BFTree*, *LMT*, *FT*, *NBTree* (Kalmegh, 2015). Para validar dichos algoritmos se utilizó un conjunto de 46 proteínas mezcladas con representatividad de Alfa, A+B, A/B y Beta.

A cada uno de los clasificadores se aplicó un proceso de selección de parámetros óptimos. Donde, los algoritmos se entrenaron y probaron con todo el conjunto de proteínas. Finalmente, para cada algoritmo se seleccionó la mejor combinación de parámetros, los mejores resultados obtenidos para esta combinación se muestran en la Tabla 2.

Tabla 2. Resultados experimentales de la selección de clasificadores base y los parámetros para los que muestran su mejor comportamiento.

	NBT	**ADT**	**DS**	**LMT**	**REPT**	**LADT**	**RT**	**J48**	**RF**	**SC**
Precisión	1	0,77	0,62	1	1	0,81	1	1	1	0,96
Sensibilidad	0,87	0,55	0,6	0,84	0,83	0,54	0,91	0,8	0,94	0,75
Fm	0,93	0,64	0,61	0,91	0,9	0,64	0,95	0,89	0,97	0,84

La **¡Error! No se encuentra el origen de la referencia.**, muestra los resultados de la selección de clasificadores base, donde se puede apreciar que los valores obtenidos para la precisión por los métodos empleados esta entre 62% y 100% con un promedio de 91%. En cuanto a sensibilidad los valores están entre 54% y 94% con promedio de 76%. Para la medida Fm los valores observados son entre 61% y 97% con promedio 82%. Un análisis particular de los resultados podemos notar que el método de peor desempeño en cuanto a la precisión es *DecisionStump* con un 62% de precisión. En cambio, para la sensibilidad, los métodos con menor capacidad de recuerdo para el conjunto de proteínas empleado son ADT y LADT con valores de 55% y 54% respectivamente. De manera general, el algoritmo de mejor desempeño fue *RandomForest* el cual supera al resto de los métodos empleados en el estudio.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 3, Julio-Septiembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 105-113

Robustez y capacidad de generalización

Para analizar la robustez y capacidad de generalización del método propuesto se empleó un conjunto de 2020 proteínas dividido por diferentes longitudes de secuencia (<100, 100-200, 200-300, >300). Se realizó una validación cruzada para 10. Los resultados se muestran en la Tabla 3.

Tabla 3. Resultado experimental del desempeño del algoritmo propuesto en proteínas con diferentes longitudes de secuencia.

	< 100	**100-200**	**200-300**	**300 >**
Precisión	0,49	0,55	0,51	0,49
Sensibilidad	0,57	0,78	0,8	0,81
Fm	0,51	0,63	0,61	0,61

La Tabla 3, muestra el desempeño alcanzado por el meta multiclasificador propuesto empleando un *RandomForest* como clasificador base. Donde, el promedio para las métricas empleadas es de 51%, 74% y 59%. Se puede observar que el mejor desempeño en cuanto a precisión se obtuvo en proteínas con una longitud de secuencia entre 100-200 aminoácidos. Para el resto de los grupos los valores para esta medida se comportaron de manera similar y cerca del 50% (1). Para la sensibilidad, se logró recordad un mayor número de interacciones para los conjuntos de proteínas con longitud superior a los 200 aminoácidos. Donde, se obtuvieron valores cercanos al 80% (1). Finalmente, para Fm, los valores fueron similares para los conjuntos de proteínas con longitud superior a 100 aminoácidos 62% (1). En cambio, para proteínas pequeñas, la media armónica entre la precisión y la sensibilidad fue cerca de 51%.

Dominio de aplicación

Para analizar el desempeño del método propuesto en diferentes dominios de aplicación se empleó un conjunto de 45 proteínas dividido por clases estructurales (Alfa, Beta, A+B y A/B, Todas). Se realizó una validación cruzada para 10 particiones con dos repeticiones. Los resultados se muestran en la Tabla 4.

Tabla 4. Resultado del algoritmo propuesto en proteínas con diferentes clases estructurales.

	Alfa	**A+B**	**A-B**	**Beta**	**Todas**
Precisión	0,35	0,28	0,29	0,25	0,41
Sensibilidad	0,77	0,72	0,67	0,85	0,82
Fm	0,47	0,39	0,38	0,37	0,53

La Tabla 4, muestra el desempeño alcanzado por el meta multiclasificador propuesto. Donde, para el conjunto de proteínas Alfa se alcanzó la mayor precisión con un 35%. En cambio, para el conjunto Beta, la precisión fue de 25%. En general el promedio para esta medida fue de 29%. Por otra parte la sensibilidad fue superior en el conjunto de proteínas Beta con un valor de 85%, para el resto de los conjuntos los valores fueron alrededor del 72% (5). En cuanto a la métrica Fm, al promedio fue de 40%. Donde para el conjunto de proteínas con mayoría Alfa la media armónica entre la precisión y la sensibilidad lograda fue de 47%. Por otro parte, un análisis sobre el conjunto de todos los grupos,

la precisión alcanzada supera la de los grupos por separado, con un 41%. Con una sensibilidad similar a la obtenida en proteínas Beta, donde la métrica Fm logra superar el 50%.

Conclusiones

Los métodos de predicción de mapas de contacto son un paso intermedio para la predicción de estructuras de proteínas. Una forma mediante la cual el desempeño de estos métodos puede ser elevado es realizando la predicción de las interacciones entre estructuras secundarias. Este proceso puede reducir el espacio de búsqueda de contactos entre residuos, además es suficientemente informativo con respecto al plegamiento de las proteínas.

En este artículo se realizó un estudio de la influencia de las interacciones en el plegamiento de las proteínas. Donde se demostró que las interacciones α-α, α-β, β-α, β-β, β-coil, coil-β, α-coil, coil-α tienen una responsabilidad directa sobre el 90% de los contactos entre aminoácidos.

Se propuso un meta multiclasificador basado en árboles de decisión para predecir las interacciones entre estructuras de proteínas. Se realizó una comparación entre diferentes algoritmos para seleccionar los clasificadores base para el meta multiclasificador, donde el mejor desempeño lo obtuvo *RandomForest*.

El método se validó en varios conjuntos de proteínas, donde el mejor desempeño en cuanto a precisión se obtuvo en las proteínas Alfa con un 79%. En cambio, la sensibilidad superior se logró en proteínas Beta. Además, en cuanto a robustez y capacidad de generalización el modelo propuesto se comporta mejor en proteínas con longitud superior a los 200 aminoácidos. De manera general el algoritmo propuesto obtuvo un promedio de 51% de precisión, con una sensibilidad de 74%.

Agradecimientos

Se les agradece a los profesores de la Facultad de Informática de la Universidad de Ciego de Ávila “Máximo Gómez Báez” por el tiempo dedicado. A los profesores del Centro de Investigaciones de la Informática (CII) y del Centro de Bioactivos Químicos (CBQ) de la Universidad Central “Marta Abreu de las Villas”, a los trabajadores e investigadores del Centro de Bioplantas de Ciego de Avila por la ayuda y los conocimientos brindados.

Referencias

ABU-DOLEH, A.A., AL-JARRAH, O.M. y ALKHATEEB, A., 2012. Protein contact map prediction using multi-stage hybrid intelligence inference systems. *Journal of Biomedical Informatics*, vol. 45, no. 1, pp. 173-183. ISSN 15320464. DOI 10.1016/j.jbi.2011.10.008.

ANDONOV, R., MALOD-DOGNIN, N. y YANEV, N., 2011. Maximum Contact Map Overlap Revisited. *Journal of Computational Biology*, vol. 18, no. 1, pp. 27-41. ISSN 1066-5277, 1557-8666. DOI 10.1089/cmb.2009.0196.

ASHKENAZY, H., UNGER, R. y KLIGER, Y., 2011. Hidden conformations in protein structures. *Bioinformatics*, vol. 27, no. 14, pp. 1941–1947.

CHAMORRO, A.E.M., DIVINA, F., AGUILAR-RUIZ, J.S. y CORTÉS, G.A., 2011. A multi-objective genetic algorithm for the Protein Structure Prediction. *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*. S.l.: IEEE, pp. 1086–1090.

CHEN, P. y LI, J., 2010. Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. *BMC structural biology*, vol. 10, no. 1, pp. 1.

CHENG, J. y BALDI, P., 2007. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, vol. 8, no. 1, pp. 113. ISSN 14712105. DOI 10.1186/1471-2105-8-113.

COHEN, J., 2004. Bioinformatics—an introduction for computer scientists. *ACM Computing Surveys (CSUR)*, vol. 36, no. 2, pp. 122–158.

DI LENA, P., MARGARA, L., VASSURA, M., FARISELLI, P. y CASADIO, R., 2008. A new protein representation based on fragment contacts: towards an improvement of contact maps predictions. *Computational Intelligence Methods for Bioinformatics and Biostatistics*. S.l.: Springer, pp. 210–221.

DI LENA, P., NAGATA, K. y BALDI, P., 2012. Deep architectures for protein contact map prediction. *Bioinformatics*, vol. 28, no. 19, pp. 2449-2457. ISSN 1367-4803, 1460-2059. DOI 10.1093/bioinformatics/bts475.

DING, W., XIE, J., DAI, D., ZHANG, H., XIE, H. y ZHANG, W., 2013. CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks. En: B. XUE (ed.), *PLoS ONE*, vol. 8, no. 4, pp. e61533. ISSN 1932-6203. DOI 10.1371/journal.pone.0061533.

FRANCIA, S.S. y GARCÍA, M.N.M., 2006. *Multiclasificadores: Métodos y Arquitecturas*. S.l.: Universidad de Salamanca. Departamento de Informática y Automática.

GARCÍA, S., FERNÁNDEZ, A., LUENGO, J. y HERRERA, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, vol. 180, no. 10, pp. 2044-2064.

GROMIHA, M.M., 2009. Multiple Contact Network Is a Key Determinant to Protein Folding Rates. *Journal of Chemical Information and Modeling*, vol. 49, no. 4, pp. 1130-1135. ISSN 1549-9596, 1549-960X. DOI 10.1021/ci800440x.

HOWE, C.W. y MOHAMAD, M.S., 2011. Protein Residue Contact Prediction using Support Vector Machine. *World Academy of Science, Engineering and Technology*, vol. 60, pp. 1985–1990.

KALMEGH, S., 2015. Analysis of WEKA data mining algorithm REPTree, Simple CART and RandomTree for classification of Indian news. *International Journal of Innovative Science, Engineering and Technology*, vol. 2, no. 2, pp. 438-46.

KARAKAŞ, M., WOETZEL, N. y MEILER, J., 2010. BCL::Contact–Low Confidence Fold Recognition Hits Boost Protein Contact Prediction and *De Novo* Structure Determination. *Journal of Computational Biology*, vol. 17, no. 2, pp. 153-168. ISSN 1066-5277, 1557-8666. DOI 10.1089/cmb.2009.0030.

KUNCHEVA, L.I., 2004. *Combining pattern classifiers: methods and algorithms*. Hoboken, NJ: J. Wiley. ISBN 978-0-471-21078-8. TK7882.P3 K83 2004

MÁRQUEZ-CHAMORRO, A.E., ASENCIO-CORTES, G., DIVINA, F. y AGUILAR-RUIZ, J.S., 2014. Evolutionary decision rules for predicting protein contact maps. *Pattern Analysis and Applications*, vol. 17, no. 4, pp. 725–737.

MÁRQUEZ-CHAMORRO, A.E., ASENCIO-CORTÉS, G., SANTIESTEBAN-TOCA, C.E. y AGUILAR-RUIZ, J.S., 2015. Soft computing methods for the prediction of protein tertiary structures: A survey. *Applied Soft Computing*, vol. 35, pp. 398–410.

MITRA, S. y HAYASHI, Y., 2006. Bioinformatics with soft computing. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 5, pp. 616-635. ISSN 1094-6977. DOI 10.1109/TSMCC.2006.879384.

MONASTYRSKYY, B., D'ANDREA, D., FIDELIS, K., TRAMONTANO, A. y KRYSHTAFOVYCH, A., 2014. Evaluation of residue-residue contact prediction in CASP10: Contact Prediction in CASP10. *Proteins: Structure, Function, and Bioinformatics*, vol. 82, pp. 138-153. ISSN 08873585. DOI 10.1002/prot.24340.

RANDALL, A., CHENG, J., SWEREDOSKI, M. y BALDI, P., 2008. TMBpro: secondary structure, -contact and tertiary structure prediction of transmembrane -barrel proteins. *Bioinformatics*, vol. 24, no. 4, pp. 513-520. ISSN 1367-4803, 1460-2059. DOI 10.1093/bioinformatics/btm548.

ROSE, P.W., BI, C., BLUHM, W.F., CHRISTIE, C.H., DIMITROPOULOS, D., DUTTA, S., GREEN, R.K., GOODSELL, D.S., PRLIC, A., QUESADA, M., QUINN, G.B., RAMOS, A.G., WESTBROOK, J.D., YOUNG, J., ZARDECKI, C., BERMAN, H.M. y BOURNE, P.E., 2013. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research*, vol. 41, no. D1, pp. D475-D482. ISSN 0305-1048, 1362-4962. DOI 10.1093/nar/gks1200.

SANTIESTEBAN-TOCA, C.E., ASENCIO-CORTÉS, G., MÁRQUEZ-CHAMORRO, A.E. y AGUILAR-RUIZ, J.S., 2012. Short-Range interactions and decision tree-based protein contact map predictor. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* [en línea]. S.l.: Springer, pp. 224–233. [Consulta: 12 octubre 2015]. Disponible en: http://link.springer.com/chapter/10.1007/978-3-642-29066-4_20.

TEGGE, A.N., WANG, Z., EICKHOLT, J. y CHENG, J., 2009. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Research*, vol. 37, no. Web Server, pp. W515-W518. ISSN 0305-1048, 1362-4962. DOI 10.1093/nar/gkp305.

WANG, C.-Y., ZHU, H.-D. y CAI, L., 2009. A new prediction protein structure method based on genetic algorithm and coarse-grained protein model. *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on* [en línea]. S.l.: IEEE, pp. 1–5. [Consulta: 19 octubre 2015]. Disponible en: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5305230.

WU, X. y KUMAR, V., 2009. *The top ten algorithms in data mining* [en línea]. S.l.: CRC Press. [Consulta: 19 octubre 2015]. Disponible en: https://books.google.com/books.

XIE, J., DING, W., CHEN, L., GUO, Q. y ZHANG, W., 2015. Advances in Protein Contact Map Prediction Based on Machine Learning. *Medicinal Chemistry*, vol. 11, no. 3, pp. 265–270.

ZAKI, M.J., SHAN JIN y BYSTROFF, C., 2003. Mining residue contacts in proteins using local structure predictions. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 33, no. 5, pp. 789-801. ISSN 1083-4419. DOI 10.1109/TSMCB.2003.816916.