INGENIERÍA DE SISTEMAS

PESPAD: una nueva herramienta para la predicción de la estructura secundaria de la proteína basada en árboles de decisión

Claudia X. Mazo*, Oscar F. Bedoya* §

** Escuela de Ingeniería de Sistemas y Computación. Universidad del Valle*
§ *oscar.bedoya@correounivalle.edu.co*

(Recibido: Junio 24 de 2009 - Aceptado: Diciembre 06 de 2010)

Resumen

Una de las tareas que actualmente enfrentan los bioinformáticos es la predicción de la estructura secundaria de la proteína. Este problema consiste en que dada una secuencia de aminoácidos, se debe predecir la estructura de cada residuo, siendo hélices - y hojas- las más comunes. A pesar de los avances que se han logrado al plantear modelos para la predicción de la estructura secundaria, se intenta mejorar su exactitud predictiva. En este artículo se presenta una nueva herramienta para la predicción de la estructura secundaria de la proteína, llamada PESPAD, que supera significativamente la exactitud predictiva de los métodos individuales existentes. La herramienta se apoya en modelos construidos con base en árboles de decisión para la mezcla de expertos.

Palabras Claves: Predicción de proteínas, Clasificación, Árboles de decisión, Mezcla de expertos.

SYSTEMS ENGINEERING

PESPAD: a new tool for protein secondary structure prediction based on decision trees.

Abstract

Predicting the secondary structure of the protein is a central problem in bioinformatics. The problem is focused on try to predict the structure of each residue given the amino acid sequence. The most common secondary structures are - helix and - sheets. Despite the progress made in making models for predicting the secondary structure, there is still a need to improve their predictive accuracy. This paper presents a new tool for predicting the secondary structure of the protein, called PESPAD, which significantly exceeds the accuracy of predictive methods available now. The tool is based on models constructed by using decision trees for the mixture of experts.

Keywords: Protein prediction, Classification, Decision trees, Mixture of experts.

1. Introducción

Actualmente la bioinformática se ocupa de problemas que van desde la identificación y análisis de motivos, la predicción de genes y el modelamiento de sistemas biológicos hasta el análisis de las funciones asociadas a las secuencias de aminoácidos. Uno de los problemas que es de especial interés es la predicción de la estructura secundaria de la proteína. Este problema consiste en que dada una secuencia de aminoácidos, se debe predecir la estructura secundaria de cada residuo, siendo hélices - y hojas -β las más comunes. Para la predicción de estructuras secundarias se utilizan actualmente algoritmos que analizan la composición de la secuencia de aminoácidos, entre los que sobresalen: Chou-Fasman de Chou & Fasman (1974), GOR de Garnier et al. (1978), DSC de King & Sternber (1996), PREDATOR de Frishman & Argos (1997), SIMPA96 de Levin (1997), NNPredict de Kneller et al. (1990), PHD de Rost & Sander (1990) y PSIPRED de Jones (1999).

Estas estrategias presentan una exactitud reportada en la literatura que alcanzan actualmente el 75% para casos de prueba particulares, sin embargo, para algunos conjuntos de prueba logran tan solo un 55%. Según Allen et al. (2004) las decisiones de clasificación que se basan en un solo criterio tienen mayor probabilidad de errar en comparación a si se utilizara combinación de expertos. Es decir, tomar la decisión de cuál es la estructura secundaria con base en uno sólo de los algoritmos indicados anteriormente no es para nada confiable y puede tener un alto grado de incertidumbre. En Allen et al. (2004) se propone tomar las decisiones de clasificación con base en la mezcla de expertos.

Esta técnica consiste en integrar criterios, medidas o algoritmos, denominados expertos, para tomar una decisión más acertada. En el contexto de la bioinformática esta estrategia ha sido usada para resolver diferentes problemas, por ejemplo, Xu et al. (1994) plantea la mezcla por medio de árboles de decisión para predecir la ocurrencia de exones en secuencias de ADN. Así mismo, De Haan & Leunissen (2005) y Chen & Chaudhari (2007) realizan la integración de criterios o mezcla de expertos por medio de redes neuronales en el contexto de la predicción de estructuras secundarias. Por su parte, Birzele & Kramer (2006) proponen usar patrones frecuentes.

La estrategia de mezcla de expertos se puede realizar por diferentes métodos computacionales, entre los cuales se destacan los árboles de decisión y las redes neuronales. Sin embargo, el rango de algoritmos aplicables para esta tarea se puede ampliar a redes bayesianas, máquinas de vectores de apoyo y árboles de decisión oblicuos. La mezcla de expertos es aún un campo por explorar ya que es posible aplicar diferentes técnicas para llevar a cabo la integración de criterios de tal forma que se intente mejorar la exactitud de los modelos existentes.

En este artículo se realiza la integración de criterios para la predicción de la estructura secundaria de la proteína utilizando árboles de decisión. Los modelos obtenidos por medio de la mezcla de expertos fueron implementados en una herramienta, llamada PESPAD, cuya exactitud predictiva supera ampliamente la de los métodos individuales que existen actualmente, convirtiéndose en una nueva alternativa útil y de mayor exactitud para la predicción de la estructura secundaria de la proteína.

El resto de este artículo está organizado en secciones. En la sección 2, se presenta la metodología seguida en la investigación realizada. En la sección 3, se presentan y discuten los resultados. Finalmente, en la sección 4, se presentan las conclusiones.

1.1 Conceptos preliminares

1.1.1 Clasificación

La clasificación de datos es un proceso que involucra dos etapas: construcción y utilización de un modelo. En la primera, se construye un modelo que describe un conjunto predeterminado de clases analizando los atributos de las instancias de entrenamiento. Se asume que cada instancia pertenece a una clase predefinida, la cual está determinada por uno de los atributos, llamado etiqueta de clase. En el contexto de la clasificación, las tuplas se llaman también ejemplos u objetos.

Las tuplas que se analizan para construir el modelo se denominan ejemplos de entrenamiento y forman colectivamente el conjunto de entrenamiento. Por otra parte, las tuplas que se utilizan para probar el modelo se conocen como ejemplos de prueba. El aprendizaje del modelo es supervisado, puesto que se conoce la clase a la cual pertenece cada ejemplo de entrenamiento, esto contrasta con el aprendizaje no supervisado en donde las etiquetas de clase de cada ejemplo de entrenamiento y el número de clases, no se conoce de antemano.

Por lo general, el modelo aprendido tras el proceso de clasificación se representa por medio de reglas de clasificación, árboles de decisión o de una formulación matemática. Por ejemplo, dada una base de datos con información de los créditos de clientes, las reglas de clasificación se pueden encontrar para identificar clientes con un pobre o un excelente crédito. Las reglas se pueden usar para categorizar ejemplos de datos futuros, así como para proporcionar un mejor entendimiento del contenido de la base de datos.

En la segunda etapa, el modelo se usa para clasificar ejemplos nuevos. De esta manera, es posible estimar su exactitud predictiva. Existen muchos métodos para estimar este valor y en general, hacen referencia al porcentaje de ejemplos del conjunto de prueba que son correctamente clasificados por el modelo. Para cada ejemplo de prueba, la etiqueta de clase conocida se compara con la clase predicha por el modelo para este ejemplo. Si la exactitud de un modelo se considera aceptable, éste se puede usar para clasificar futuras tuplas u objetos cuya etiqueta de clase no se conoce.

Además del conjunto el entrenamiento es necesario llevar a cabo la selección de datos para realizar la validación y la prueba del modelo. El conjunto de validación se compone de los datos usados para optimizar los parámetros en el clasificador y finalmente, el conjunto de prueba se forma con las instancias sobre las cuales se evalúa la exactitud.

Existen estrategias que orientan al experto en la forma como se debería dividir el conjunto de ejemplos para el entrenamiento, validación y prueba del clasificador. Una de las estrategias ampliamente utilizada se conoce como validación cruzada con k pliegues (k-fold cross validation), que consiste en dividir el conjunto de datos en k partes y calcular k veces la exactitud del clasificador. Para cada valor de k, se considera como conjunto de entrenamiento todos los datos que no pertenecen al grupo k. La fase de prueba se realiza utilizando los datos del grupo k. Este proceso se realiza k veces, de tal manera que se obtienen k estimaciones de la exactitud. Finalmente, se calcula el promedio usando los k valores.

1.1.2 Árboles de decisión

Un árbol de decisión es una estructura en la cual cada nodo interno denota una prueba sobre uno o varios atributos, cada rama representa una salida de la prueba y los nodos hoja representan clases. Cuando se consideran atributos numéricos, las pruebas son generalmente de la forma atributo≤valor, aunque se pueden encontrar árboles que tiene en sus pruebas combinaciones de sus atributos. Para atributos categóricos, las pruebas son de la forma atributo b, donde b es un subconjunto de valores posibles que puede tomar el atributo seleccionado. Para clasificar un ejemplo desconocido, los valores de los atributos del ejemplo se prueban en el árbol de decisión. Se traza un camino desde la raíz hasta el nodo hoja que indica la clase a la cual pertenece el ejemplo.

Uno de los algoritmos más usados para la construcción de árboles de decisión es C4.5 de Quinlan (1993). Se trata de un algoritmo voraz que construye árboles de decisión de arriba hacia abajo de manera recursiva. El atributo que hace parte de la prueba en cada nodo se selecciona según la ganancia de información, esto es, el atributo que minimice la información necesaria para clasificar los ejemplos en las particiones resultantes. La descripción completa del algoritmo se puede encontrar en Quinlan (1993).

La característica principal de los árboles de decisión es que son modelos de caja blanca en los cuales se puede ver directamente la frecuencia de aparición de cada atributo. Además, le permite al experto conocer el atributo con mayor poder de clasificación, es decir, aquel que se localice en el nodo raíz.

1.1.3 Estructura de la proteína

La proteína se forma en su estructura más básica por aminoácidos. Se tienen en total 20 aminoácidos, los cuales se determinan tras aplicar el código genético sobre cada codón identificado en la secuencia de ARN. Los codones se forman por la unión de tres nucleótidos consecutivos en la secuencia transcrita. Por ejemplo, dada la secuencia AUGGCCACU, se tendrían los codones AUG, GCC y ACU que aplicando la tabla del código genético se convierten en la secuencia de aminoácidos MAT. La secuencia de aminoácidos se conoce como la estructura primaria de la proteína.

Ahora bien, la relación que existe entre los diferentes aminoácidos provoca que la molécula de la proteína tome diferentes formas, éstas se pueden conocer por medios tales como la cristalografía de rayos X y la espectroscopía. Se trata de técnicas por medio de las cuales se puede conocer la forma asociada a la molécula, analizando la posición específica de sus átomos. Las formas más comunes que se han identificado son las hélices- y las hojas- . Estos nombres fueron asignados a propósito de la forma que toma la proteína, esto es, las hélices se caracterizan por su forma helicoidal mientras que las hojas por su forma aplanada. La forma que tiene asociada cada aminoácido se conoce como la estructura secundaria de la proteína y es el principal reto de este trabajo de investigación.

La proteína tiene además asociadas estructuras terciaria y cuaternaria. La estructura terciaria hace referencia a la disposición de todos los aminoácidos en el espacio y la cuaternaria a la unión de varias cadenas peptídicas que en conjunto conforman un ente.

1.1.4 Métodos de predicción de la estructura secundaria

Para la predicción de estructuras secundarias se utilizan algoritmos que analizan la composición de la secuencia de aminoácidos, entre los que sobresalen: Chou-Fasman, GOR, DSC, PREDATOR, SIMPA96, NNPredict, PHD y PSIPRED. A continuación se explican brevemente algunos de estos algoritmos.

El método Chou-Fasman se basa en una tabla que resultó de un análisis estadístico sobre la frecuencia de aparición de cada uno de los 20 aminoácidos en secuencias cuya estructura secundaria era conocida. Con base en este análisis presentado en Chou & Fasman (1974) se conoce la probabilidad de la ocurrencia de cada aminoácido en cada una las tres estructuras hélice, hoja y giro. El método se centra en detectar regiones de una secuencia de aminoácidos no caracterizada que tengan mayor probabilidad de pertenecer a una de las tres estructuras secundarias. Las probabilidades se obtienen de la tabla presentada en Chou & Fasman (1974).

El método GOR propone usar una ventana de observación deslizante de tamaño fijo de 17 aminoácidos. Si se desea predecir la estructura del aminoácido central de la ventana (posición 9), se deben sumar los puntajes asignados de sus 16 vecinos de acuerdo a dos tablas que acompañan el método, una para hélices y otra para hojas. Estas tablas se pueden encontrar en Garnier et al. (1978). Los vecinos tendrán los valores según el aminoácido y la posición en la que se encuentre. Ya que se tienen dos tablas, la asignación se hace a aquella estructura secundaria con cuya tabla se obtiene el mayor valor. El proceso continúa desplazando la ventana una posición a la derecha de la secuencia hasta llegar al final de la misma. Al igual que GOR, otros algoritmos que reportan valores altos de exactitud como SIMPA96 de Levin (1997), NNPredict de Kneller et al. (1990) y PSIPRED de Jones (1999) presentan tamaños de ventana que varía entre 15 y 18 aminoácidos. Esto muestra que la predicción de la estructura secundaria se vuelve más confiable al incorporar en el estudio de un aminoácido en particular la estructura de sus residuos vecinos.

En el método SIMPA96 se maneja la técnica del vecino más cercano utilizando una matriz de similitud. SIMPA96 maneja una ventana de observación deslizante de 13 a 17 aminoácidos y se basa en la matriz de puntuación Blosum62 para asignar los puntajes a cada aminoácido en la ventana. Se considera un valor de corte C, si la puntuación es menor que el valor C, el péptido es rechazado, de lo contrario su conformación observada se asigna a la secuencia de prueba. La descripción completa del método se puede encontrar en Levin (1997).

2. Metodología

En esta sección se presenta cada uno de los pasos que hicieron parte de la metodología seguida en la investigación realizada y que permitieron construir los modelos, realizar las pruebas y comparar los resultados obtenidos con los predictores individuales. Los modelos son el resultado de la mezcla de expertos que individualmente se encargan de la predicción de la estructura secundaria de la proteína dada una secuencia de aminoácidos. Se espera que al realizar la integración de los criterios existentes, se obtenga un modelo con una exactitud mayor a los métodos individuales. La integración se llevará a cabo usando árboles de decisión.

2.1 Selección de datos

El éxito en la construcción de un modelo se debe, en gran medida, en la selección del conjunto de entrenamiento. En el caso particular de la predicción de la estructura secundaria de la proteína, el conjunto de entrenamiento debe contener ejemplos que pertenezcan a las tres clases: hélice-α (H), hoja-β (E) y ninguna (N), esta última para aminoácidos que no presentan características propias de las hélices o de las hojas. El conjunto de entrenamiento se conformó por 3000 aminoácidos obtenidos de secuencias extraídas del PDB (Protein Data Bank). Para realizar la asignación de la estructura secundaria se utilizó el programa DSSP. De la misma fuente se obtuvo el conjunto de prueba.

2.2 Selección de los algoritmos clasificadores por árboles de decisión

Existen diferentes algoritmos para construir árboles de decisión. La elección de un algoritmo particular se basa en aspectos tales como la interpretabilidad, la exactitud predictiva y el tiempo de clasificación del árbol resultante. Además, hay algoritmos que se adaptan mejor a un contexto específico, por ejemplo, según Tjen-Sien et al. (2000) para el problema de la predicción de formas de ondas sonoras el algoritmo FACT propuesto por Loh & Vanichsetakul (1988) obtiene la exactitud predictiva más alta, mientras que en el problema de la clasificación de imágenes es el algoritmo IND de Buntine (1992).

Para seleccionar los algoritmos de clasificación por árboles de decisión se tomó como base el estudio realizado en Tjen-Sien et al. (2000). En el estudio se evaluaron 22 algoritmos de clasificación por árboles de decisión, 9 algoritmos estadísticos y 2 algoritmos de redes neuronales que fueron sometidos a 32 conjuntos de entrenamiento y pruebas para determinar aspectos como la precisión de la clasificación y el tiempo de construcción. Según los resultados obtenidos, los algoritmos más apropiados para problemas relacionados con el contexto de secuencias de datos biológicos, en este caso de aminoácidos, son: QUEST de Loh & Shih (1997), OC1 de Murthy et al. (1994), CRUISE de Kim & Loh (2001) y GUIDE de Loh (2002). Por lo tanto, se llevará a cabo la integración de expertos usando cada uno de estos cuatro algoritmos y se realizará la comparación, no sólo entre ellos, sino también con los métodos individuales.

2.3 Selección de los atributos para la construcción de los modelos

Cuando se realiza mezcla de expertos por medio de árboles de decisión, cada método individual es un atributo que se usará para la construcción del árbol. La exactitud predictiva del modelo dependerá de seleccionar aquellos métodos individuales sobre los que se basará la decisión de clasificación de la estructura secundaria. Como se mencionó anteriormente, cada nodo del árbol tiene una prueba sobre los atributos, en este caso, las pruebas se establecerán sobre los expertos que se seleccionen.

Los métodos que existen para la predicción de la estructura secundaria de la proteína que se usan comúnmente son: Chou-Fasman, GOR, DSC, PREDATOR, SIMPA96, NNPredict, PHD, PSIPRED, entre otros. De estos métodos se seleccionaron 5 teniendo como criterio principal la exactitud predictiva reportada. Los métodos seleccionados son: Chou-Fasman, GOR, PREDATOR, SIMPA96 y DSC. Con los modelos que resultan de integrar estos criterios, se espera superar la exactitud de los modelos individuales.

2.4 Construcción de los modelos de mezcla de expertos

El siguiente paso en la investigación consistió en aplicar cada uno de los algoritmos de clasificación por árboles de decisión para construir en total 4 modelos. Inicialmente se preparó el conjunto de entrenamiento para que se adaptara a la entrada de cada algoritmo. Se realizó un proceso que consistió en tomar cada aminoácido del conjunto de entrenamiento y se calculó la salida de los 5 métodos individuales, conociendo de antemano la asignación correcta.

De esta forma, el conjunto de entrenamiento se compone de ejemplos que tienen, por un lado la asignación de cada método y por otro la asignación real. Estos datos se pasan a los algoritmos que construyen los árboles de decisión para que al combinar las salidas de los métodos individuales intente realizar una predicción más acertada.

2.4.1 Construcción del modelo con CRUISE

Para la construcción del modelo de predicción aplicando el algoritmo CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) se utilizó la implementación disponible en http://www.stat.wisc.edu/~loh/cruise.html. CRUISE es un algoritmo que construye árboles de decisión paralelos a los ejes, esto es, solo están permitidas pruebas en los nodos del árbol de la forma atributoS, donde S es un subconjunto de los valores que puede tomar el atributo. El algoritmo se basa en el análisis LDA (Linear Discriminant Analysis) para realizar la partición del rango de valores numéricos y se caracteriza porque permite crear árboles n-arios. La clasificación de un dato de prueba se realiza siguiendo la rama del árbol que cumpla las condiciones planteadas en cada uno de los nodos y se clasifica de acuerdo con la etiqueta que se tenga como hoja al final del recorrido. El árbol resultante se muestra en la Figura 1.

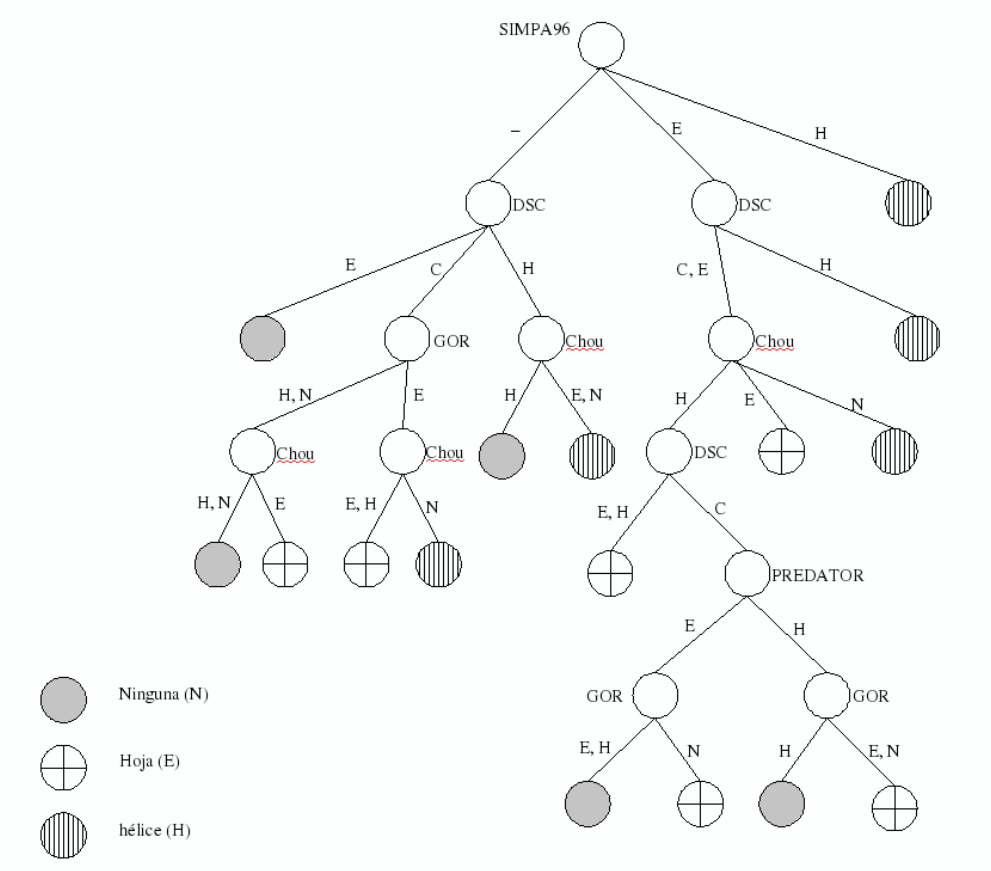

Figura 1. Árbol CRUISE

2.4.2 Construcción del modelo con OC1

Para la construcción del modelo de predicción aplicando el algoritmo OC1 (Oblique Classifier 1) se utilizó la implementación disponible en http://www.cs.jhu.edu/~salzberg/announce-oc1.html. El algoritmo OC1 construye árboles de decisión oblicuos, esto son, aquellos en los que se permiten pruebas en los nodos del árbol que son combinación lineal de los atributos. A diferencia de los árboles paralelos a los ejes, se pueden tener pruebas sobre varios atributos en un mismo nodo. Las pruebas en un árbol oblicuo regularmente se expresan en forma de hiperplanos, es por esto que en el árbol obtenido con este método, que se presenta en la Figura 2, aparecen combinaciones lineales de los atributos. En la nomenclatura usada en el árbol, x[1] corresponde a la salida dada por el algoritmo Chou-Fasman, x[2] para GOR, x[3] para PREDATOR, x[4] para SIMPA96 y x[5] para DSC.

2.4.3 Construcción del modelo con GUIDE

Para la construcción del modelo de predicción aplicando el algoritmo GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) se utilizó la implementación disponible en http://www.stat.wisc.edu/~loh/guide.html. El algoritmo GUIDE se caracteriza porque es capaz de detectar relaciones entre los atributos que forman el árbol. Su importancia para esta investigación está en que permitirá conocer si hay métodos de predicción que al usarse en forma conjunta con otros métodos pierden relevancia en la tarea de predicción. El árbol resultante se muestra en la Figura 3.

2.4.4 Construcción del modelo con QUEST

Para la construcción del modelo de predicción aplicando el algoritmo QUEST (Quick, Unbiased and Efficient Statistical Tree) se utilizó la implementación disponible en http://www.stat.wisc.edu/~loh/quest.html. QUEST se caracteriza porque sólo construye árboles binarios, lo que permite que dicho proceso sea más rápido que por ejemplo con CRUISE, el cual construye árboles n-arios o que OC1, que tiene particiones oblicuas. Lo que se quiere probar con este algoritmo, además de la exactitud predictiva, es la diferencia en el tiempo de construcción que tiene con respecto a los otros algoritmos utilizados. El árbol resultante se muestra en la Figura 4.

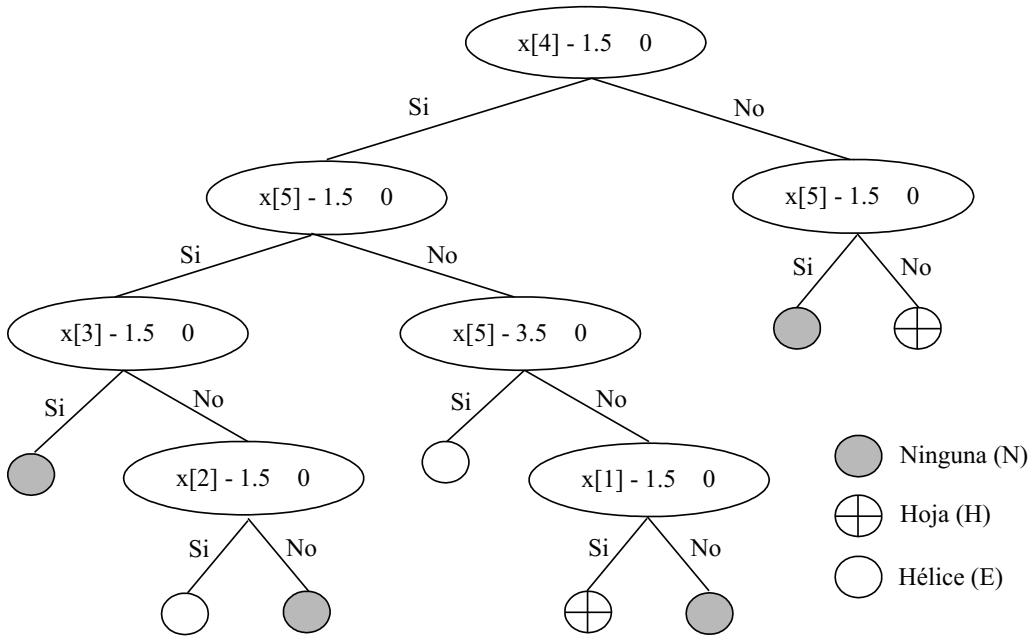

Figura 2. Árbol Oc1

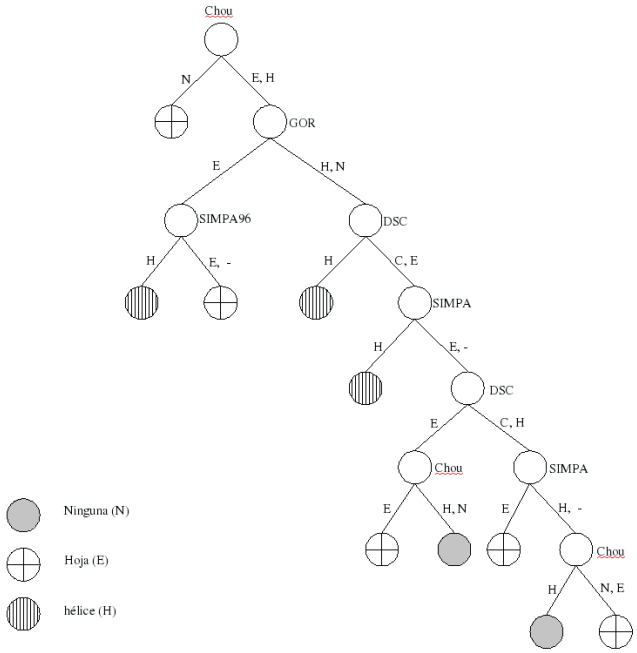

Figura 3. Árbol GUIDE

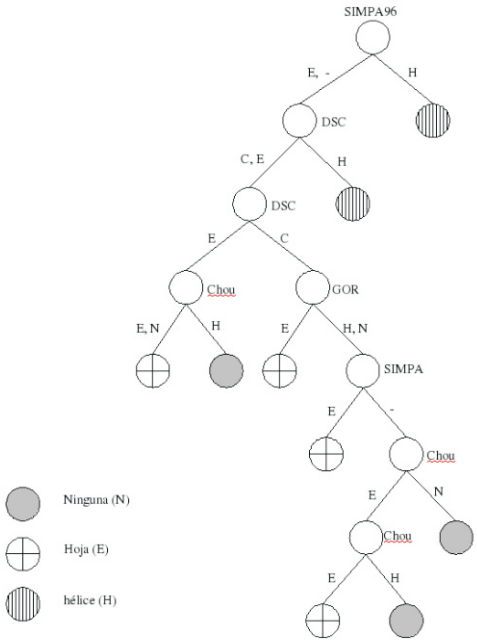

Figura 4. Árbol QUEST

2.5 Implementación de los modelos

Se desarrolló una herramienta llamada PESPAD (Predicción de la Estructura Secundaria de la Proteína usando Árboles de Decisión) que implementa los modelos obtenidos en esta investigación. Además de que permite aplicar los métodos individuales, presenta estadísticas comparativas entre dichos predictores y los modelos propuestos. La herramienta fue desarrollada en C++ como una aplicación local. En la Figura 5 se muestra una captura de pantalla de la aplicación.

La herramienta permite realizar la predicción de la estructura secundaria para una secuencia de aminoácidos particular o un conjunto de éstas en un archivo de texto. Se presenta un área para graficar la secuencia de aminoácidos y las predicciones de los 5 métodos individuales y los 4 modelos propuestos, esto le permite al usuario contrastar los resultados obtenidos de forma gráfica.

2.6 Pruebas de los modelos propuestos

El siguiente paso en la metodología consiste en calcular la exactitud predictiva de los modelos propuestos y los métodos individuales para llevar a cabo su comparación. Para esto, inicialmente se seleccionaron los criterios de comparación y luego se ejecutó un plan de pruebas.

2.6.1 Parámetros de comparación

La calidad de los modelos de predicción de la estructura secundaria de la proteína se puede calcular por medio de diferentes parámetros. Sin embargo, se utilizan principalmente dos indicadores para la comparación con los métodos de predicción individuales, estos son, la exactitud de residuos Q_3 presentada en la Ec.(1) y el coeficiente de correlación de Matthew presentado en la Ec.(2). A continuación se presentan sus definiciones.

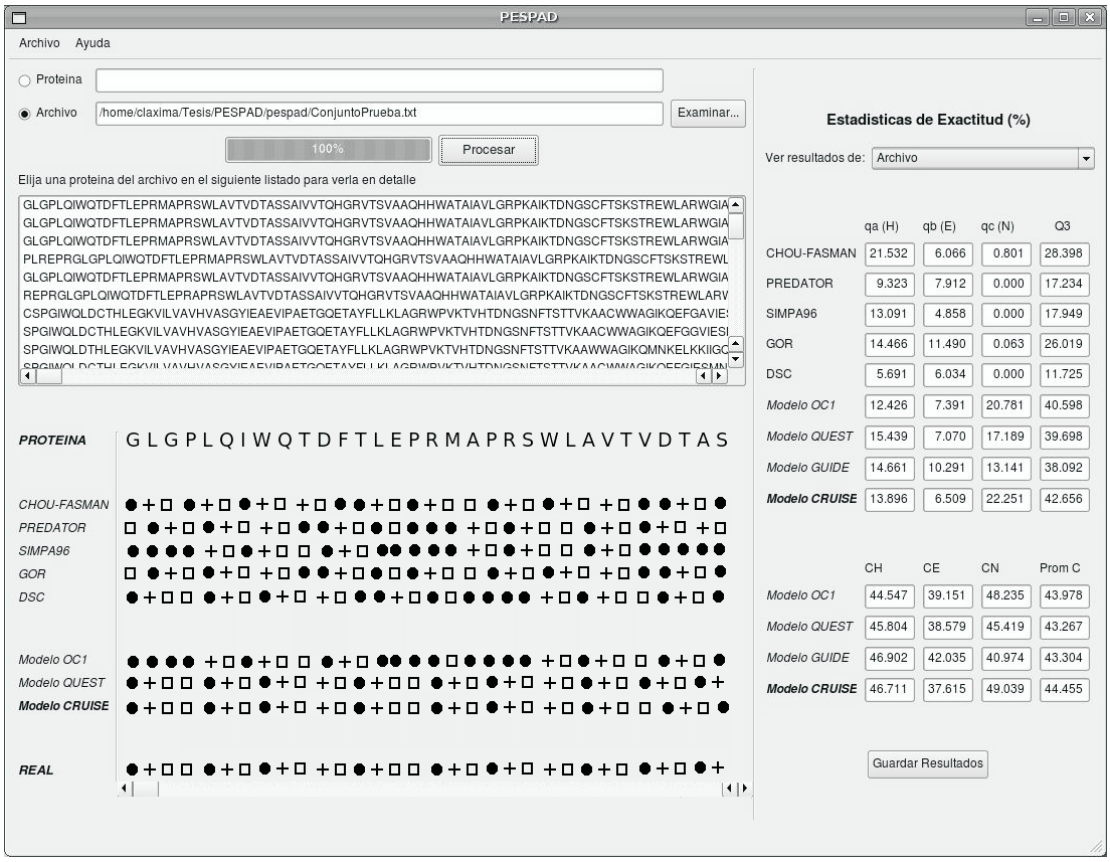

Figura 5. Ejecución de PESPAD

• Exactitud de residuos Q_3:

$$Q_3 \quad \frac{(q_a \quad q_b \quad q_c)\ 100}{N} \qquad (1)$$

donde q_a es la cantidad de hélices correctamente clasificadas, q_b es la cantidad de hojas correctamente clasificadas, q_c es la cantidad de aminoácidos que no son hélices ni hojas correctamente clasificadas y *N* corresponde al tamaño del conjunto de prueba.

• Coeficiente de correlación de Matthew C:

$$C \quad \frac{(tp \quad tn) \quad (fp \quad fn)}{\sqrt{(tp \quad fp)\ (tp \quad fn)\ (tn \quad fp)\ (tn \quad fn)}} \qquad (2)$$

donde *tp* es la cantidad de verdaderos positivos, *tn* es la cantidad de verdaderos negativos, *fp* es la cantidad de falsos positivos y *fn* es la cantidad falsos negativos.

2.6.2 Plan de pruebas y ejecución

Para llevar a cabo una comparación justa entre los métodos individuales y los modelos propuestos, se tomó un conjunto de prueba de 600 aminoácidos cuya estructura secundaria era conocida y se sometió a la predicción de los 5 métodos individuales y los 4 modelos propuestos. El consolidado de los resultados en términos del indicador Q_3 se muestra en la Tabla 1. Este indicador representa el porcentaje de aciertos de los métodos. Además, en la Tabla 2 se muestra el consolidado de los resultados en términos del indicador C. Esta medida tiene en cuenta no sólo el porcentaje de aciertos sino también los fallos de los modelos, indicados por los falsos positivos y los falsos negativos. Los resultados de las pruebas se conocieron utilizando la aplicación propuesta PESPAD.

Tabla 1. Consolidado de valores Q_3

Algoritmo	q_a	q_b	q_c	Q_3
Chou-Fasman	79	99	0	0.443
GOR	150	77	0	0.372
PREDATOR	94	97	0	0.318
SIMPA96	85	136	2	0.378
DSC	162	63	41	0.297
Modelo OC1	119	114	94	0.545
Modelo CRUISE	153	108	92	0.588
Modelo GUIDE	156	136	71	0.605
Modelo QUEST	150	107	86	0.572

Tabla 2. Consolidado de valores C

Algoritmo	Clase	TP	TN	FP	FN	C
Modelo OC1	H	119	348	52	81	0.596
	E	114	294	106	86	0.521
	N	94	285	115	106	0.482
Modelo CRUISE	H	153	326	74	47	0.649
	E	108	314	86	92	0.527
	N	92	313	87	108	0.492
Modelo GUIDE	H	156	304	96	44	0.617
	E	136	311	89	64	0.584
	N	71	348	52	129	0.459
Modelo QUEST	H	150	345	55	50	0.677
	E	107	291	109	93	0.507
	N	86	307	93	114	0.477

Tabla 3. Consolidado de los tiempos de construcción y uso

Modelo	Construcción (Seg)	Clasificación (Seg)
OC1	2.660	0.831
CRUISE	2.120	0.857
GUIDE	2.410	0.708
QUEST	2.340	0.741

Además de la exactitud de los modelos obtenidos, se realizaron pruebas para conocer los tiempos de construcción y uso de los árboles. El consolidado de los resultados se muestra en la Tabla 3.

3. Resultados y discusión

En esta sección se discuten los resultados de las pruebas realizadas y se lleva a cabo un análisis comparativo entre los modelos propuestos y los métodos de predicción individuales.

3.1 Análisis de los modelos propuestos

Con base en los valores Q_3 obtenidos en las pruebas realizadas se puede observar que los modelos propuestos que han sido implementados en la herramienta PESPAD, logran una mayor exactitud predictiva que los métodos existentes. Su éxito se logra al integrar los diferentes métodos de tal forma que se pueda tomar una decisión de clasificación confrontando las salidas de algunos de los algoritmos más exactos que se han propuesto. Por lo tanto, continuando con la idea de Allen et al. (2004) de integrar criterios para tomar una mejor decisión de clasificación, se puede notar que dicha integración de criterios sí resulta efectiva para la predicción de estructura secundaria de la proteína usando árboles de decisión. La importancia de este resultado toma mayor valor por el hecho de que los modelos propuestos han sido implementados como una herramienta de software, la cual se convierte en una nueva alternativa que puede ser utilizada por la comunidad científica para intentar predecir de forma más exacta estructuras secundarias dadas las secuencias de aminoácidos.

Otro punto a destacar en los modelos propuestos es que además de que se logra aumentar la cantidad de aciertos, según el coeficiente de correlación C, los modelos mantienen la capacidad de detectar falsos positivos. Su importancia radica en la capacidad que mantienen los modelos de desechar este tipo de secuencias a la hora de hacer la clasificación.

A continuación se presenta un análisis detallado que compara los modelos obtenidos:

• El modelo propuesto que logra mayor cantidad de aciertos es GUIDE. Esto se verifica en la cantidad de hélices y hojas que logra detectar este árbol y que sobrepasa a los otros modelos construidos. Sin embargo, se debe aclarar que los árboles obtenidos con los algoritmos CRUISE y QUEST tienen una exactitud similar.

• Cuando se permiten particionamientos oblicuos se perjudica la exactitud predictiva del árbol en el contexto de la predicción de la estructura secundaria de la proteína. Los árboles que tienen particionamientos paralelos a los ejes presentan mayor exactitud. Este resultado se puede interpretar teniendo en cuenta que trabajar con hiperplanos sobre cada nodo representa un problema de mayor complejidad computacional que los particionamientos paralelos a los ejes. En el caso de la predicción de la estructura secundaria de la proteína, resulta más complejo detectar hiperplanos que separen de forma exacta los datos.

• El hecho de utilizar árboles con particiones binarias como QUEST o n-arias como CRUISE, no difieren significativamente en su exactitud predictiva. Los valores calculados en los indicadores C y Q_3 obtenidos con QUEST que construye árboles binarios, son similares a los obtenidos con CRUISE donde se generan árboles n-arios.

• Los árboles que se enfocan en detectar relaciones entre los atributos son potencialmente mejores en los sistemas por mezcla de expertos en el contexto de la predicción de la estructura secundaria de la proteína. Este resultado se evidencia por la exactitud obtenida con el algoritmo GUIDE que se caracteriza precisamente por detectar relaciones entre los métodos individuales usados como expertos.

• Para el problema específico de predecir hélices, el árbol construido con el algoritmo QUEST es el más exacto. Teniendo en cuenta el indicador C, que considera no solo los aciertos sino también los falsos positivos y negativos, se puede verificar que el árbol construido con QUEST obtiene en este tipo específico de estructura secundaria los valores más altos.

• Para el problema específico de predecir hojas, el árbol construido con el algoritmo GUIDE es el más exacto. Teniendo en cuenta el indicador C, que considera no solo los aciertos sino también los

falsos positivos y negativos, se puede verificar que el árbol construido con GUIDE obtiene en este tipo específico de estructura secundaria los valores más altos.

• De acuerdo a los tiempos registrados para la construcción de los modelos se verifica que toma más tiempo construir árboles oblicuos que paralelos a los ejes. En general, no se tiene una diferencia significativa en los tiempos de construcción y uso de los modelos. Esto permite afirmar que la integración de expertos, usando árboles de decisión, en el contexto de la predicción de la estructura secundaria no se ve afectada por los tiempos de construcción y uso de los modelos.

3.2 Comparación con los métodos individuales

Los resultados de las pruebas mostrados en la Tabla 1 son contundentes, los árboles propuestos superan significativamente a los métodos individuales. El mejor método individual fue Chou-Fasman que registra un valor Q_3 de 0.443 comparado con el mejor método propuesto que fue GUIDE, con un valor Q_3 de 0.605. Un valor mayor de Q_3 indica que tiene más aciertos, por ejemplo, un Q_3 de 0.5 indica que sobre un conjunto de prueba de N aminoácidos, en la mitad de éstos el modelo acierta en su predicción. En general, se cumple que cada uno de los modelos propuestos supera ampliamente las exactitudes de los métodos individuales.

4. Conclusiones

El desarrollo del presente trabajo permite afirmar que para el problema de la predicción de la estructura secundaria de la proteína, la solución de mezcla de expertos por medio de árboles de decisión es apropiada. Esta afirmación se basa principalmente en la exactitud predictiva de los modelos obtenidos que sobrepasaron significativamente a los métodos individuales. Por su parte, la herramienta desarrollada PESPAD representa una nueva alternativa útil y de mejor exactitud para la predicción de la estructura secundaria de la proteína que los métodos individuales existentes. Además, con el desarrollo del trabajo se obtuvieron algunos aportes valiosos en el contexto de la predicción de la estructura secundaria de la proteína, los cuales se mencionan a continuación:

• En general, para el problema de la predicción de la estructura secundaria de la proteína, se recomienda usar el árbol construido con el algoritmo GUIDE.

• Permitir particionamientos oblicuos perjudica la exactitud predictiva del árbol de decisión.

• Tener árboles con particiones binarias o n-arias no presentan diferencias significativas en cuanto a la exactitud predictiva de los modelos resultantes.

• Árboles que se enfocan en detectar relaciones entre los atributos son potencialmente mejores en los sistemas por mezcla de expertos en el contexto de la predicción de la estructura secundaria de la proteína.

• Para el problema específico de predecir hélices se recomienda usar modelo obtenido con QUEST.

• Para el problema específico de predecir hojas se recomienda usar el modelo obtenido con GUIDE.

5. Referencias bibliográficas

Allen, J., Pertea, M., & Salzberg, S. (2004). Computational Gene Prediction Using Multiple Sources of Evidence. *Genome Research* 14(2004), 142148.

Birzele, F., Kramer, S. (2006). A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics* 22(21), 2628-2634.

Buntine, W. (1992). Learning classication trees. *Statistics and Computing* 2(2), 63-73.

Chen, J., Chaudhari, N. (2007). Cascaded Bidirectional Recurrent Neural Networks for Protein Secondary Structure Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4(4), 572-582.

Chou, P., Fasman, G. (1974). Prediction of protein conformation. *Biochemistry* 13(2), 222-245.

De Haan, J., & Leunissen, J. (2005). *Protein secondary structure prediction. Comparison of ten common prediction algorithms using a neural network.* In: Essays in Bioinformatics. Moss DS, Jelaska S, Pongor S., editors. IOS Press, Amsterdam. p. 149-161.

Frishman, D., & Argos, P. (1997). 75% accuracy in protein secondary structure prediction. *Proteins* 27(3), 329-335.

Garnier, J., Osguthorpe, & D., Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* 120, 1, 97-120.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292(2), 195-202.

Kim, H., & Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association* 96(454), 589-604.

King, R., & Sternberg, M. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science* 5(11), 2298-2310.

Kneller, D., Cohen, F., & Langridge, L. (1990). Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *Journal of Molecular Biology* 214(1), 171-182.

Levin, J. (1997). Exploring the limits of nearest neighbour secondary structure prediction. *Protein Engineering* 10(7), 771-776.

Loh, W. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* 12(2002), 361-386.

Loh, W., & Shih, Y. (1997). Split selection methods for classication trees. *Statistica Sinica* 7(4), 815-840.

Loh, W., & Vanichsetakul, N. (1988). Tree-structured classication via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association* 83(403), 715-728.

Murthy, S., Kasif, S., & Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of Articial Intelligence Research* 2(1), 1-33.

Quinlan, J. (1993). C4.5: *Programs for Machine Learning.* San Francisco: Morgan Kaufmann Publisher.

Rost, B., & Sander, C. (1990). *Improved prediction of protein secondary structure by use of sequence profiles and neuronal networks.* Proceedings of the National Academy of Science U.S.A. 90(16), 7558-7562.

Tjen-Sien, L., Wei-yin, L., & Yu-shan, S. (2000). A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning* 40(3), 203-229.

Xu, Y., Einstein, J., Mural, R., Shah, M., & Uberbacher, C. (1994). *An improved system for exon recognition and gene modeling in human DNA sequences.* Proceedings of the second international conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, p. 376-384.