Algunas aplicaciones de la estructura booleana del código genético

Some applications of the Boolean structure of the Genetic Code

R. Grau, Deborah Galpert, María del C. Chávez,[1] R. Sánchez, Gladys Casas, E. Morgado

Universidad Central "Marta Abreu" de Las Villas (UCLV).
[1] Instituto Nacional de Investigaciones en Viandas Tropicales (INIVIT).
{rgrau, deborah, mchavez, robersy, gladita, morgado}@uclv.edu.cu

Resumen

Las estructuras booleanas del código genético constituyen modelos matemáticos mínimos y muy simplificados que nos ayudan a comprender mejor la lógica subyacente del código genético. Más específicamente, estas estructuras reflejan una fuerte conexión entre los órdenes del código genético y las propiedades físico-químicas de los aminoácidos. En este artículo presentamos dos aplicaciones de esta estructura algebraica en problemas típicos de Bioinformática. El primer problema trata de la clasificación de las mutaciones de una proteína dada. El siguiente es un caso particular del problema de predicción de estructura secundaria. Usamos además técnicas estadísticas y de inteligencia artificial en la solución de ellos.

Palabras clave: Código genético, álgebra de Boole, propiedades de aminoácidos, redes bayesianas, información mutua, clasificación de mutaciones, estructura secundaria

Abstract

The Boolean structures of the genetic code are very simplified minimal mathematical models that help us to comprehend better the logic underlying the genetic code.

Specifically, these structures reflect a strong connection between genetic code orders and the physical-chemical properties of amino acids. In this paper we present two applications of this algebraic structure in bioinformatics problems. The first problem deals with the classification of mutations in a protein. The other one is a particular prediction problem of secondary structures of proteins. Statistical and artificial intelligence techniques are also used to solve the problems.

Keywords: Genetic code, Boolean algebra, amino acid properties, bayesian networks, mutual information, mutation classification, secondary structure

Introducción

La evolución molecular tiene lugar en la inmensa mayoría de los genes encontrados en todas las especies vivas. Somos testigos, por ejemplo, de la gran variabilidad genética mostrada por los virus y microorganismos patógenos. Las **estrategias evolutivas** de estos los hacen capaces de evadir la acción de los fármacos utilizados por el hombre, por lo cual, constituye hoy día un reto para las ciencias biológicas. Muchos han sido los intentos para obtener una descripción matemática formal que explique satisfactoriamente este proceso, pero aún no ha sido

biológica. Para muchos investigadores, la clave para analizar este proceso parte de la estructura y organización del código genético. El Grupo de Bioinformática de la UCLV ha propuesto y publicado varias de estas posibles formalizaciones en términos de estructuras algebraicas de dicho código, entre ellas una estructura de Álgebra de Boole que pretende ser más cercana a las propiedades físico-químicas y biológicas de los aminoácidos que conforman las proteínas. En el presente trabajo se muestran dos aplicaciones de dicha estructura.

Sobre la estructura booleana del código genético y su interpretación biológica

El código genético es el sistema bioquímico que permite establecer las reglas a través de las cuales la secuencia de nucleótidos de un gen es transcripta en la secuencia de codones del ARN mensajero y luego traducida en la secuencia de aminoácidos de la proteína correspondiente. El conjunto de codones es una extensión del alfabeto de cuatro letras de la molécula de ADN. Estas letras son las moléculas básicas del ADN: Adenina, Guanina, Citosina y Tiamina, usualmente denotadas por A, G, C y T (en el ARNm, T se cambia por U, Uracilo). Ellas se parean en la doble hélice del ADN de acuerdo con los puentes de hidrógeno: G con C (tres puentes) y A con T (dos puentes). La organización no-aleatoria del código genético ha sido subrayada ya por trabajos anteriores y se han propuesto hipótesis para explicar el orden observado (Crick 1968; Alf-Steinberger 1969; Swanson 1984; Freeland y Hurst 1998), pero el origen de este orden permanece siendo un enigma.

Han existido varios intentos para introducir una caracterización formal (incluso algebraica del código genético (Bertman y Jungck 1979; Siemion *et al* 1995; Jiménez-Montaño 1996; Bashford, Tsohantjis y Jarvis 1998; Jiménez-Montaño 1999; Bashford y Jarvis 2000; Karasev and Stefanov 2001; Balakrishhnan 2002). Las caracterizaciones más relevantes y cercanas a nuestro modelo involucran la representación binaria de las cuatro bases nucleótidas. Han sido propuestos varios de estos modelos con una representación binaria diferente. Incluso, en la misma revista en que fueron publicados por primera vez los resultados de la UCLV (Sánchez, Morgado y Grau 2004a), se publican también otros resultados basados en una representación binaria de las cuatro bases, siguiendo un orden diferente (Bouton, De Oliveira, Campello de Souza, Santos Magalhaes *et al 2004*). Sin embargo, el orden del código o del retículo correspondiente, no tuvo hasta ahora una interpretación físico-química consecuente, o la codificación se limitó a los aspectos formales y no se aprovecharon las operaciones booleanas para producir resultados con una interpretación físico-química o biológica clara. El objetivo del presente trabajo es describir dos aplicaciones de la nueva estructura booleana propuesta, una al estilo clásico como sistema de codificación y otra donde se hace uso de las consecuencias algebraicas de dicha estructura y para así profundizar en la correspondencia entre las consideraciones algebraicas y los datos experimentales.

Resumen del modelo teórico usado en el presente trabajo

Sánchez *et al* (2004a, 2004b y 2005a) describieron en detalle el modelo teórico de Álgebra de Boole propuesto por la UCLV para el código genético y sus implicaciones. Primeramente se dotó al conjunto de las 4 bases {U,C,G,A} de un orden y una estructura booleana y luego se extendió al conjunto de tripletas (codones) del código. Se parte de que en cualquier álgebra booleana $(B(X), \vee, \wedge)$, construida a partir de un conjunto X, dados 2 elementos $\alpha, \beta \varepsilon X$ se tiene $\alpha \leq \beta$, si y solo si $\neg\alpha \vee \beta = 1$ (1 es el elemento máximo del álgebra). Además, si $\alpha \leq \beta$ o $\alpha \geq \beta$ se dice que los elementos α y β son

comparables. Dos elementos $\alpha, \beta \varepsilon X$, se dicen complementarios si y solo si $\alpha \cdot \beta=1$ y $\alpha + \beta=0$ (0 es el elemento mínimo del álgebra). Es conocido además que, en cualquier retículo booleano con 4 elementos todos los elementos son comparables, excepto 2 de ellos, que son complementarios. Nuestro retículo booleano de las 4 bases nucleótidas X={U, C, G, A} se construye asumiendo que las bases complementarias en el retículo son también complementarias (apareadas) en la molécula del ADN. Este retículo de 4 bases debe tener un máximo, un mínimo y 2 elementos no comparables. Se asume que el elemento máximo en el código genético de 64 codones debe ser la tripleta del elemento máximo del retículo de 4 bases y por tanto UUU, CCC, GGG o AAA. Para obtener un retículo booleano con significación biológica se tuvieron en cuenta las propiedades físico-químicas de estos codones y sus aminoácidos respectivos, en particular:

1. Ambos codones, GGG y CCC tienen el mismo número máximo de puentes de hidrógeno. Esta propiedad debía estar reflejada en el retículo de manera que GGG sea complementario a CCC. Además, ambos codifican para cadenas aminoácidos pequeñas con poca diferencia en la polaridad: Glicina y Prolina. Esta propiedad de **similaridad** determinó que estos elementos debían ser comparables.

2. Los codones UUU y AAA tienen el mismo número mínimo de puentes de hidrógenos entre ellos, y por tanto, el elemento complementario de UUU en el retículo debía ser AAA. Pero estas tripletas codifican respectivamente para cadenas aminoácidos con polaridades opuestas extremas (Leucina y Lisina). Consecuentemente esta propiedad **opuesta** determinó que estos elementos no fueran comparables.

Así, existen solamente dos posibilidades para el retículo de cuatro letras. De inciso1 resulta que el mínimo elemento tiene que ser C y el máximo G (o a la inversa) y de 2) resulta que U y A no deben ser comparables, y en particular no pueden ser mínimo o máximo si queremos tal significado biológico. Por tanto, en la tercera potencia se obtienen 2 retículos (álgebras) posibles, llamados convencionalmente primal y dual: $(B(X),\vee,\wedge)$ (primal) y $(B'(X),\wedge,\vee)$ (dual). Los diagramas de Hasse de los retículos de cuatro bases y de los de 64 codones son mostrados en los artículos (Sánchez, Morgado y Grau, 2004a, 2004b y 2005a).

En estos mismos artículos se hacen evidentes los isomorfismos de los retículos de las 4 bases con los retículos booleanos $((Z_2)^2,\vee,\wedge)$ y $((Z_2)^2,\wedge,\vee)$, donde $Z_2=\{0,1\}$. Entonces es posible representar el retículo primal por la correspondencia: G↔00; A↔01; U↔10; C↔11 y para el retículo dual usar: C↔00; U↔01; A↔10; G↔11. Ya que las álgebras de Boole del código genético son obtenidas como la tercera potencia directa de las álgebras de Boole de las cuatro bases, explícitamente: $C(X)=B(X)\mathbf{x}B(X)\mathbf{x}B(X)$, estas álgebras son isomorfas a $((Z_2)^6,\vee,\wedge)$ y $((Z_2)^6,\wedge,\vee)$, isomorfismo inducido por el isomorfismo básico $\Phi: B(X)\rightarrow(Z_2)^2$. Ello permite realizar operaciones lógicas entre nucleótidos y codones, término a término, y que están implementadas en *packages* originales del software *Mathematica*, para el estudio de deducciones, que como ha sido probado, y comentaremos después, están íntimamente asociadas a las mutaciones (Sánchez *et al* 2004a).

Interpretaciones biológicas del modelo demostradas

En el diagrama de Hasse de ambas álgebras del código genético se reflejan las conexiones entre las propiedades algebraicas y las propiedades físico-químicas de los aminoácidos (Sánchez *et al* 2004a). Por ejemplo, la imagen (por la negación booleana) de un codón que codifica para un aminoácido hidrofóbico es siempre la de un codón

que codifica para un aminoácido hidrofílico. En el diagrama de Hasse se puede ver también que los codones que codifican para aminoácidos con diferencias hidrofóbicas extremas se encuentran en cadenas de longitud máxima diferentes. Como resultado, no es posible deducir un codón 5'X1AX33' que codifica para un aminoácido hidrofílico de un codón 5'X'1UX'3 3' que codifica para un aminoácido hidrofóbico y viceversa.

Sánchez *et al* (2004a y 2004b) consideran la función de distancia de Hamming entre codones vistos como sextetos binarios, como el número de dígitos diferentes entre ellos. Por ejemplo: d(CGU, AUC) = d(110010, 011011) = 3. Además, calculan la distancia de Hamming entre aminoácidos como la distancia media entre los codones que codifican para ellos. Muestran cómo la distancia de Hamming entre 2 codones en el diagrama de Hasse refleja las diferencias entre las propiedades físico-químicas de los aminoácidos. Las mayores distancias se corresponden con transversiones en la segunda base de codones que frecuentemente alteran las propiedades hidrofóbicas y las funciones biológicas de las proteínas. Se presentan altos valores de la distancia de Hamming entre codones que codifican para aminoácidos hidrofílicos e hidrofóbicos. Los autores demuestran la tendencia al incremento de esta distancia entre aminoácidos, conjuntamente con las diferencias físico-químicas vistas en función de la polarizabilidad, polaridad y volumen normalizado Van der Waals.

Desde el punto de vista de las deducciones, ambas álgebras booleanas reflejan el resultado experimental bien conocido de que sustituciones en una base simple son fuertemente conservativas en lo que se refiere a cambios en la polaridad de aminoácidos (Friedman y Weinstein 1964; Parker 1989). Particularmente, a partir de aminoácidos polares cuyos codones tienen A en la segunda posición es imposible, por medio de deducciones, obtener aminoácidos provenientes de codones que tengan U en la segunda base. Sánchez *et a*l (2004a y 2005a) demostraron también, con datos experimentales de varias proteínas, que las mutaciones más frecuentemente observadas y que minimizan el efecto posterior en ellas corresponden a deducciones desde sus respectivos tipos salvajes (*wild types*).

Otra implicación de la estructura booleana que se utilizará en las aplicaciones es el concepto de *valor de la información mutua*. Esta fue definida por Sánchez et al (2005b), partiendo de las deducciones conjuntas n_{ij} (en las álgebras primal y dual) de todos los codones del código genético hacia los codones que codifican para los aminoácidos *i* y *j*. El valor de la información mutua entre dos aminoácidos: *i, j*

está dado por $V(i,j)=\log_4 \left[\frac{nij}{ni.nj} N \right]$ donde n_i es el número de deducciones posibles al aminoácido *i* (lo mismo para *j*) y N es el número total de deducciones posibles entre aminoácidos. El valor de la Información Mutua entre dos secuencias de aminoácidos de la misma longitud *n* se define entonces como la suma de valores de la información mutua de los aminoácidos en cada posición, esto es:

$$V((X_1,...,X_n),(Y_1,...,Y_n))=\sum_k V(X_k,Y_k)$$

El valor de la información mutua entre dos secuencias, así definido, pretende ser una medida de "cuanto dos secuencias tienen información importante común" y es por tanto una "medida de similaridad entre mutaciones (ya interpretadas como deducciones)"

Construcción de árboles filogenéticos a partir de secuencias de ADN y su integración en una Red Bayesiana

Se ha preparado una base de datos con secuencias de mutaciones del gen de la proteasa del VIH, formado por 99 codones. Se considerarán como **variables predictivas** los elementos de la secuencia codificados en forma de pareja binaria, como se estableció antes por Sánchez et al (2004a): G↔00; A↔01; U↔10; C↔11 para un total de 99 * 6 = 594 variables. La variable dependiente identifica tres **familias** de mutaciones dentro de la base de datos y que fueron obtenidas en este trabajo por técnicas clásicas de *clustering* pero que pueden ser preliminarmente obtenidas por cualquier otro criterio estadístico o biológico. Se han aplicado técnicas de inteligencia artificial y de estadística que permiten obtener **árboles filogenéticos** a partir de las interacciones entre las variables y las familias obtenidas. Mediante la unión de dichos árboles, se obtiene el modelo que permite el pronóstico de la familia a la que deben pertenecer nuevas mutaciones o recíprocamente el comportamiento de las posiciones en la secuencia de ADN de una nueva mutación en una familia dada.

Construcción de los árboles filogenéticos

Se utiliza la técnica del CHAID (*Chi-Squared Automatic Interaction Detector*) como herramienta intermedia para obtener los **árboles filogenéticos**. Se utiliza genéricamente el término de **árboles filogenéticos** porque inicialmente las clases fueron denominadas **familias** y por tanto su desglose puede interpretarse como especies, **sub/especies**, etc., como en una taxonomía. Pero más generalmente, la variable dependiente que representa la **familia** puede hablar de **nivel de resistencia antiviral**, **nivel de actividad fungicida**, etc. Lo esencial es que la técnica de CHAID permite reducir considerablemente el modelo probabilístico, pues tiene en cuenta solamente las interacciones fundamentales de las variables predictivas con las familias o clases conformadas. En nuestro ejemplo: para conformar la red se consideran solo las 7 variables más predictivas y sus interacciones (en total 19 variables), ellas son las que recomienda sucesivamente el CHAID por tener la mayor significación el test $\chi 2$ de asociación de ellas con la familia (SPSS 1994a). En la figura 1 se pueden ver 2 de los 7 árboles construidos. Ellos involucran 4 de las 19 variables finalmente consideradas. Por ejemplo, el primer árbol (a la izquierda de la figura 1) demuestra que las familias pueden identificarse completamente por las posiciones 546 y 323 que corresponden respectivamente al segundo número binario de la tercera base del codón 91 (546=91*6) y al primer número binario de la tercera base del codón 54 (323=53*6+5/6): si en la posición 546 hay un 1, definitivamente la secuencia pertenece a la familia 2. En caso contrario, las familias 1 y 3 se distinguen por lo que aparezca en la posición 323: un 0 o un 1 respectivamente. Análogamente el árbol de la derecha de la figura 1 demuestra que las familias pueden identificarse completamente también por las posiciones 544 y 324.

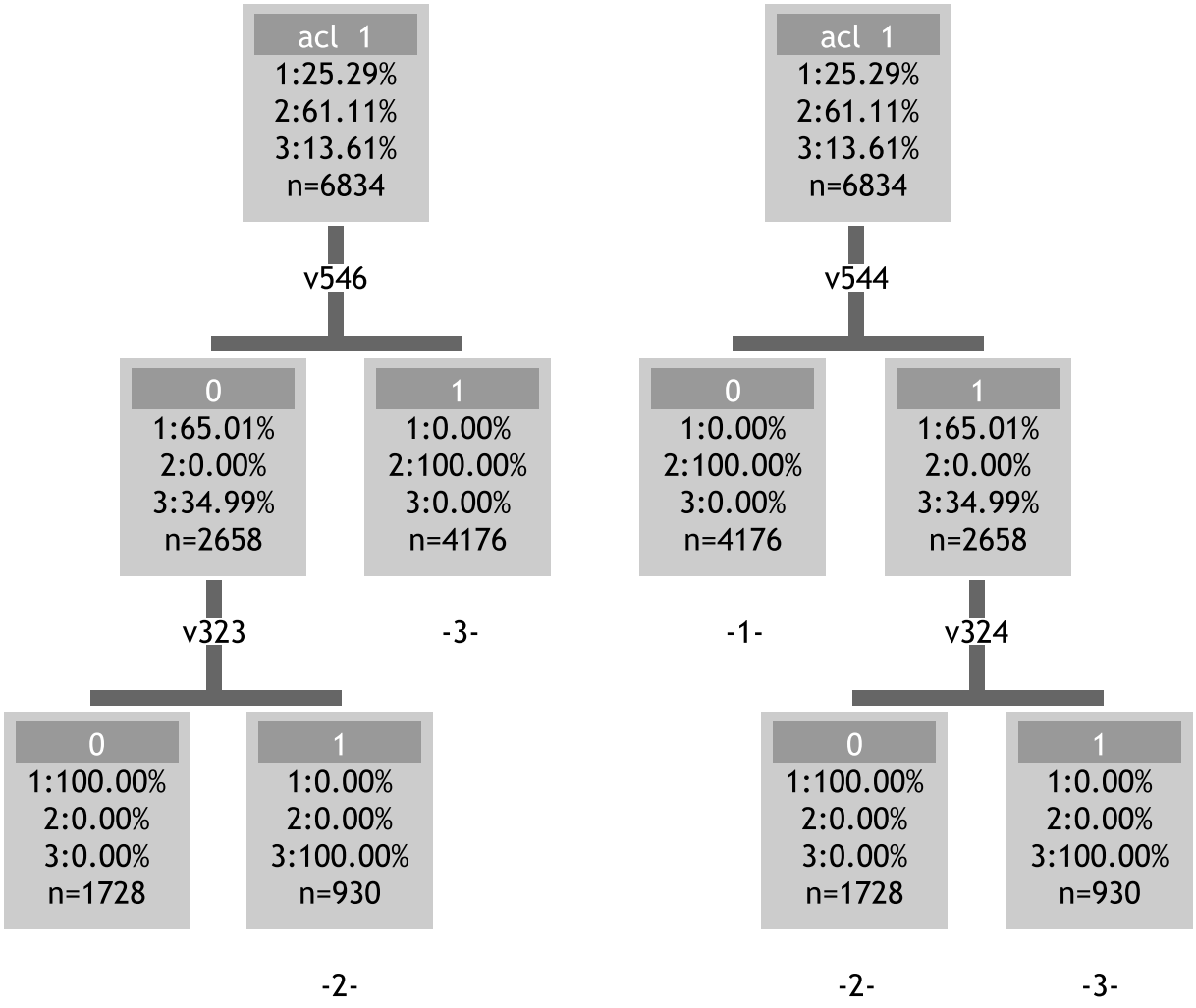

Fig. 1. Dos de los 7 ***árboles filogenéticos*** construidos. A la izquierda, el que rompe por la posición 546 y que interactúa con la posición 323. A la derecha el que rompe por la posición 544 interactuando con la 324. Los por cientos en cada nodo al lado de 1:2: y 3: definen la cantidad de mutaciones dentro de las familias 1, 2 y 3 respectivamente con los valores de ***n*** mencionados en cada nodo. Por ejemplo, en el árbol de la izquierda: el 100% de los que tienen v546=0 y v323=0 (1 728 mutaciones) están dentro de la familia 1. El 100% de los que tienen v546=0 y v323=1 (930 mutaciones) están en la familia 3, y el 100% de los que tienen v546=1 (4176 mutaciones) están en la familia 2. Conclusiones similares pueden establecerse a partir del árbol de la derecha y de los otros 5 árboles que pueden obtenerse por esta vía.

Con vistas a construir un gráfico general que represente la estructura de la Red Bayesiana adecuada, al utilizar la técnica de CHAID en la confección de los gráficos parciales, se sigue el criterio de que una vez que una variable pertenezca al modelo en cualquiera de los árboles ya construidos, la misma no se vuelve a utilizar. Esto permite reducir la complejidad del modelo y evitar que aparezcan ciclos no compatibles con la definición de red bayesiana, en particular garantiza que la estructura de la red sea un gráfico acíclico.

Construcción de la Red Bayesiana

Con la unión de todos los árboles creados se forma el modelo estructural de la Red Bayesiana, (ver Figura 2). Por el método de construcción, dicho modelo es un grafo acíclico dirigido, que expresa las dependencias esenciales, y por tanto las probabilidades condicionales, que tienen que ser calculadas. El grafo así obtenido, representa el modelo estructural de la red e informa del número mínimo de probabilidades condicionales que tienen que ser previamente

estimadas para hacer cualquier inferencia en ella. El complemento probabilístico de la estructura anterior es el cálculo de estas probabilidades condicionales (imprescindibles) y se logra siguiendo el método propuesto por Chávez, Grau y García (1999) que utiliza el SPSS (*Statistical Package for the Social Sciences* (SPSS 1994b)) como herramienta estadística. Esencialmente, las probabilidades necesarias se estiman por frecuencias condicionales a partir de la base de datos de aprendizaje con tablas de contingencia.

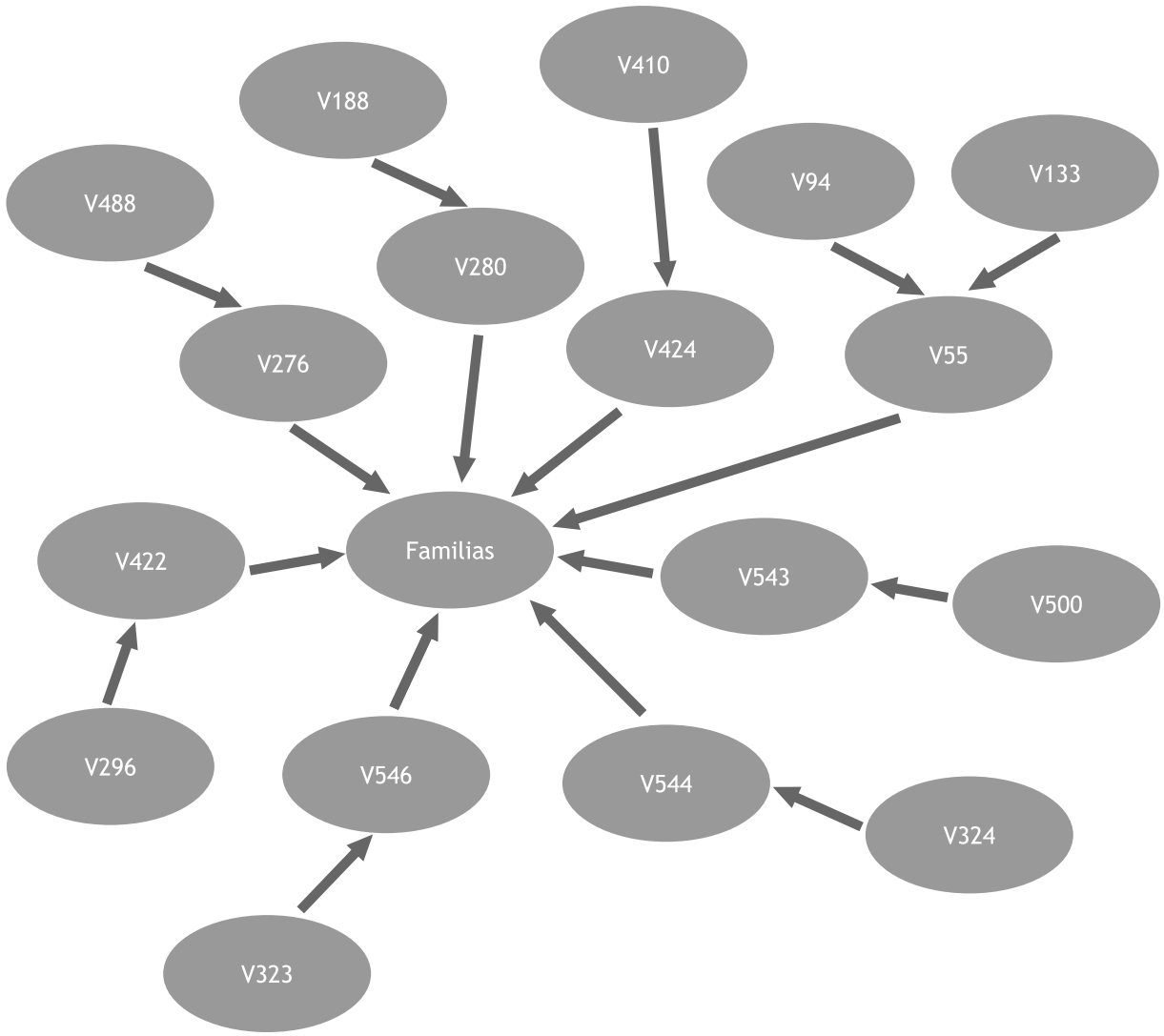

Fig. 2. Red Bayesiana obtenida por la unión de los distintos árboles filogenéticos. Ver en particular la representación esquemática de la dependencia de la familia respecto a v546 (en interacción con v323) o de v544 (en interacción con v324).

Propagación sobre la Red Bayesiana

Se utilizó una primera versión del software ByShell (Chávez y Rodríguez, 2002) para propagar evidencias en Redes Bayesianas. Este software tiene implementado el algoritmo de propagación en redes múltiplemente conexas, específicamente el de árboles de conglomerados el cual se reporta como uno de los menos complejos de los algoritmos de propagación exacta. Actualmente se trabaja en propuestas en forma distribuida para estos algoritmos y el uso de algoritmos aproximados utilizando técnicas de simulación (Castillo, Gutiérrez y Hadi, 1996; Stuart y Norvig 1996; Williams, Wilson y Hancock 1998; Hunter 2003).

Con la propagación sobre la red se puede dar un pronóstico de la familia a la que deben pertenecer nuevas mutaciones. Dando evidencias de valores de varias (no necesariamente todas) de las posiciones claves, la red predice la familia a la cual debe pertenecer y la probabilidad de la misma. Por ejemplo, de la figura 1 se evidencia que si se informan los datos de las posiciones 546 y 323, se clasifica la secuencia con seguridad. Recíprocamente se puede deducir el comportamiento de las posiciones en la secuencia de ADN de una nueva mutación en una familia dada. Por ejemplo, si se tiene identificada la familia como la 2, y se quiere investigar la posición 296 de la secuencia, que corresponde a la segunda posición binaria de la primera base del codón 50 pues 296=49,6+2/6, la red bayesiana pronosticará con probabilidad del 97%, que dicha base es una A (01) o una C (11), pues su segunda posición tiene esa probabilidad de ser un 1. Analizando la primera posición de ese nucleótido entre los datos es posible identificar que ella es siempre un 0 y por tanto la base es A (Adenina). Así, se tiene la posibilidad de trabajar con el nivel de información más elemental posible (números binarios que en pareja identifican bases, y después integrar en sextetos: codones, y finalmente codificando en aminoácidos en la proteína). Integrando esta información elemental se puede formular un conjunto de reglas para posiciones claves dentro de la secuencia, las cuales pueden ser interpretadas por los especialistas en bioquímica o biología. En la Tabla 1 se muestran, por ejemplo, las conclusiones que resultan del análisis de los codones vinculados a las posiciones esenciales detectadas en la red bayesiana y evidentes de la Figura 1 (variables 323 y 546). Ello supuso analizar las posiciones 319-324 correspondientes al codón 54 (319=53*6+1) y las posiciones 541-546 (541=90*6+1) del codón 91. Resulta por ejemplo, que la familia 2 tiene como codón 91: AGC codificando para la Serina, mientras que las familias 1 y 3, tienen ambas ACU, codificando para la Threonina. Estas últimas familias entre sí se diferencian por el codón 54, que en la familia 1 es AUG (codificador de la Metionina) y en la familia 3 es AUC (que codifica para Ilenina).

Tabla 1. Integración de conclusiones desde información elemental. La segunda familia de mutaciones se identifica porque tiene en las posiciones 541-546 el codón AGC (codificante para la Serina). Las otras familias (1 y 3) tienen en estas posiciones el codón ACU (que codifica para la Threonina); ellas entre sí, se diferencian por el codón 54 (posiciones 319-324): en la familia 1 siempre aparece AUG (que codifica para Metionina) mientras que en la familia 3 aparece siempre AUC (codificante para Ilenina).

Codón 54 (posiciones 319-324)

Posiciones	Familia 1	Familia 2	Familia 3
319 - 320	01	01	01
321 -322	10	10	10
323 - 324	00	11	11
	AUG (Met)	AUC(Ile)	AUC (Ile)

Codón 91 (posiciones 541-546)

Posiciones	Familia 1	Familia 2	Familia 3
541 - 542	01	01	01
543 - 544	11	00	11
545 - 546	10	11	10
	ACU (Thr)	AGC (Ser)	ACU (Thr)

Se consideran resultados de esta aplicación los **árboles filogenéticos** obtenidos a partir de datos de secuencias de ADN con árboles de decisión que permiten establecer relaciones entre las familias previamente definidas, sobre la base de la interacción de posiciones claves en las secuencias, y la construcción de una red bayesiana que permite pronosticar las familias o clases a partir de evidencias de presencia de nucleótidos en determinadas posiciones, o recíprocamente, caracterizar cada una de las clases identificando los nucleótidos que aparecen en algunas posiciones con mayor probabilidad. Se obtienen de inmediato, resultados satisfactorios, considerando apenas la estructura booleana como un sistema de codificación, pues este permite extraer conclusiones como las anteriores, sobre la base del nivel de información más elemental posible (números binarios, por debajo incluso de bases nucleótidas) e integrarlas posteriormente. Es presumible que puedan obtenerse mejores resultados si se utiliza más información algebraica del código genético, por ejemplo para la formación preliminar de familias o la construcción de **árboles filogenéticos** basadas en **disimilaridades o energías** como se muestra en la siguiente aplicación.

Clasificación de la estructura secundaria de secuencias de aminoácidos utilizando la distancia de Hamming y la información mutua entre aminoácidos

En la clasificación de la estructura secundaria de las proteínas se definen las clases: Principalmente-H Alpha helix, principalmente-E Beta Sheet, combinada-H-E e irregular (Maccallum 1997.). Los programas como Structural Alignment Algorithm (SSAP) realizan la puntuación de similaridad entre las interacciones de residuos **par a par** (ambientes estructurales) alrededor de los aminoácidos. Según los reportes consultados, varios de estos programas fallan en la clasificación de proteínas que tienen un plegamiento similar, pero diferentes funciones.

El algoritmo propuesto puede aprender de la información de estructuras secundarias, tanto de secuencias de proteínas como de codones. En el prototipo construido, el aprendizaje se realiza a partir de la base de datos de estructuras secundarias de codones *Integrated Sequence-Structure Database* (ISSD) (Adzhubei y Adzhubei 1999). El algoritmo que se propone utiliza los resultados de la estructura booleana del código genético conjuntamente con la técnica Metrópolis para efectuar la clasificación. La pertenencia a una clase de estructura es controlada por una prueba estadística clásica de Mann-Whitney (ver por ejemplo, Grau 1994). Los resultados de las corridas con secuencias de estructuras conocidas extraídas de la base de datos (test) de estructuras secundarias DSSP (Kabsch y Sander 1983) han mostrado un alto por ciento de precisión tanto para estructuras E como H.

Reseña sobre el uso del algoritmo de Metrópolis en esta aplicación

El algoritmo de Metrópolis permite en general, la generación aleatoria de vectores de variables posiblemente correlacionadas. Se utiliza aquí para generar secuencias de aminoácidos como cadenas de Markov con función de distribución estacionaria *P*. Desde el punto de vista de la teoría de la información es equivalente a decir que las secuencias de aminoácidos generadas son consideradas mensajes emitidos por una fuente de Markov con función de distribución estacionaria *P* (Kabsch y Sander, 1983). La técnica Metrópolis genera de manera aleatoria perturbaciones del estado actual y las acepta o rechaza en dependencia de cómo la probabilidad del estado es afectada. Usualmente *P* es expresada

en términos de una función de energía utilizando la distribución de Boltzmann-Gibbs (Baldi y Brunak 2001).

$$P(s)=\frac{e^{-\frac{E(s)}{KT}}}{Z} \quad Z=\sum_{s} e^{-\frac{E(s)}{KT}}$$

donde; k es la constante de Boltzmann y T la temperatura absoluta. Como se plantea en (Jiménez-Montaño1996), los aminoácidos en los mensajes emitidos por las proteínas se distribuyen de acuerdo con sus energías E_i (i = 1,2,…,20), de modo que la población de aminoácidos en las proteínas siguen una distribución Boltzmann-Gibbs, siendo:

$$P_i=\frac{e^{-\frac{E_i}{KT}}}{Z} \quad Z=\sum_{i} e^{-\frac{E_i}{KT}}$$

la fracción de la población del aminoácido i en las proteínas.

Sánchez *et al* (2005b) demostraron que la energía Ei del aminoácido i en las proteínas es proporcional al valor de la información $E_i \approx \varepsilon V_i$, donde
$e = kTln(4) = 1.38629\ kT$ es la mínima disipación de energía. Ellos calcularon el valor de la información del aminoácido i a partir del número nij de deducciones conjuntas de todos los codones del código genético hacia los codones que codifican para los aminoácidos i y j obteniendo:

$$P_i = e^{-ln(20)V_i} \qquad V_i = -log_4\left[\frac{1}{N}\sum_{j=1}^{20} n_{ij}\right] \qquad N=\sum_{i,j} n_{ij}$$

Al considerar el código genético como un sistema de información booleano (Sánchez *et al* 2005b), tenemos que $E_i \approx \varepsilon\ I_i$ donde I_i es la información "cargada" por el aminoácido en el mensaje de la estructura secundaria. Fue probado además en este trabajo se probó que se cumple la proporcionalidad $Ei \approx \varepsilon\ V_i$ para los tipos de estructuras E y H. Finalmente, utilizando la ecuación de Shannon I_i =***-log(frecuencia)*** el algoritmo se pudo basar en las frecuencias observadas para una clase en una base de datos de estructuras secundarias como ISSD.

1. Mientras i =1 hasta n
2. Repetir
3. Generar un valor aleatorio uniforme u entre 0 y 1
4. Si i =1 entonces Seleccionar A1 de modo que $fr(A_1) \geq u$
5. De lo contrario Seleccionar A_i de modo que $fr(A_i \wedge A_j) \geq u$ // donde $j = i$ -1
6. Calcular r_{ij} =min (1, (fr (A_i | A_j) / fr (A_j | A_i))* (fr (A_i) / fr (A_j)))
7. Generar un valor aleatorio uniforme u entre 0 y 1
8. Si u < r_{ij} aceptar A_i
9. Hasta u < r_{ij}
10. Fin Mientras
11. Guardar secuencia k

Clasificación

En el proceso de clasificación se realizan 10 000 corridas del algoritmo Metrópolis descrito anteriormente para cada una de las clases E y H. Se obtienen 10 000 secuencias de aminoácidos de cada clase. La clasificación se puede realizar por una de las dos vías siguientes:

Se obtiene la distribución G de cada una de las secuencias generadas y de la secuencia Incógnita a partir de la función *d* distancia de Hamming.

$$g_i = \left\{ \frac{\sum_{i=1}^{n} d(X_i, X_j)}{\sum_{k=1}^{n} \sum_{j=1}^{n} d(X_k, X_j)} \right\}$$

Se calcula la entropía relativa de las distribuciones de las secuencias generadas en relación con la distribución de la secuencia incógnita. Este cálculo se realiza tanto para la clase E como para la H y luego la pertenencia a una de estas clases o a ninguna de ellas se define por una prueba Mann-Whitney de comparación de rangos medios de entropías relativas. Si la prueba resulta significativa la secuencia se clasifica en E o H y la menor media se corresponde con la clase a la que debe pertenecer la secuencia incógnita. En caso de no ser significativa la prueba, la secuencia no debe pertenecer a ninguna de estas clases. El algoritmo es el siguiente:

1. Calcular la matriz de distancias de Hamming
2. Calcular el vector de probabilidad {Ui} de la secuencia incógnita
3. Para cada secuencia generada
   Calcular el vector de probabilidad {Ki}
   Calcular la entropia relativa =Sj Ki*logbase20(Ki/Ui))
   // Ejecutar este ciclo para todas las secuencias generadas E y H
4. Ejecutar la prueba estadística Mann-Whitney para comparar las medias entropías Relativas calculadas para muestras E y H.

Se calcula la información mutua de las secuencias generadas E y H en relación con la Secuencia Incógnita. Se aplica la prueba Mann-Whitney para comparar los rangos medios de información mutua de cada una de las clases. Si la prueba resulta significativa la secuencia se clasifica en E o H y la mayor media se corresponde con la clase a la que debe pertenecer la secuencia incógnita. En caso de no ser significativa, la secuencia no debe pertenecer a ninguna de estas clases.

Prototipo construido

Se parte de 2 994 secuencias con estructura tipo E y 1977 secuencias con estructura tipo H de la base de datos ISSD. Se realiza un estudio de frecuencia de aminoácidos y de frecuencia de pares de aminoácidos en las secuencias. La Figura 3 muestra la arquitectura general de la solución.

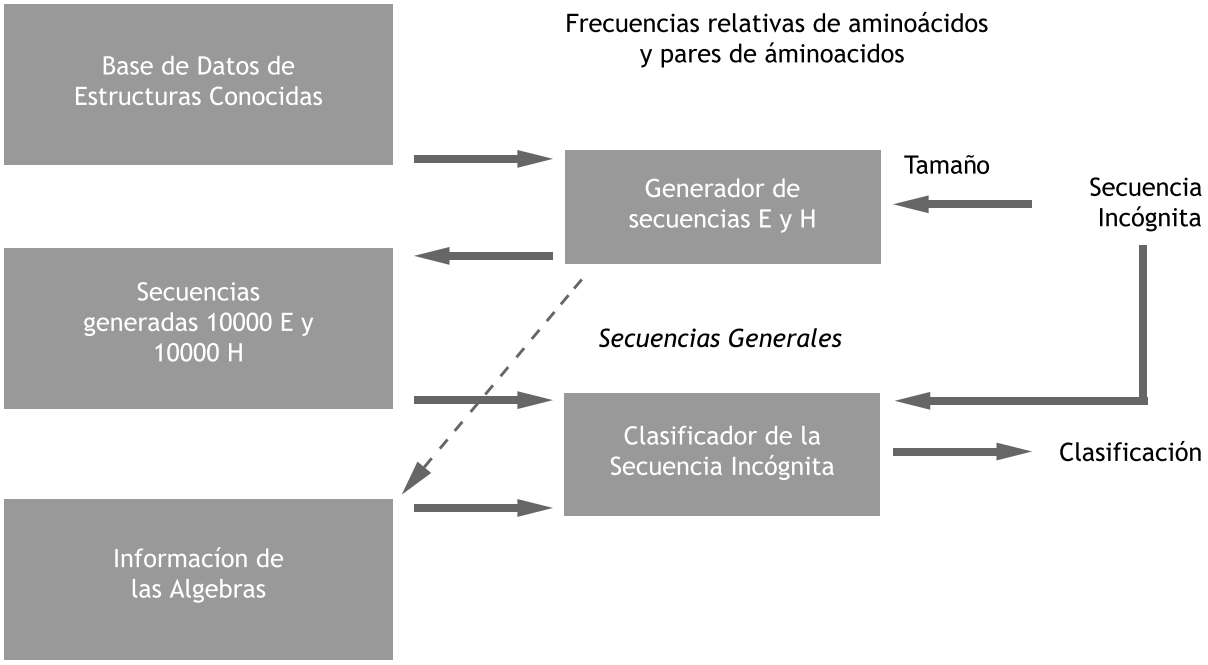

Fig. 3. Arquitectura general de la solución propuesta para la clasificación de estructura secundaria.

La información obtenida de la base de datos ISSD permite construir un algoritmo para la generación de secuencias de cada clase. Se calculan los vectores de probabilidad de la secuencia incógnita y de las secuencias generadas tomando en cuenta la distancia de Hamming. Luego la clasificación incluye la comparación de la entropía relativa de estos vectores de probabilidad, como se explica en la sección anterior. Se realiza otra variante de clasificación comparando la Información mutua entre la secuencia incógnita y las secuencias generadas. Los resultados se prueban con la base de datos (DSSP). Nótese que utilizando diferentes criterios se obtienen mejores o peores por cientos de buena clasificación pero con criterios adecuados, especialmente los de entropía relativa, se obtienen altos porcentajes de buena clasificación en secuencias con estructuras E de baja longitud y en secuencias "H" de más larga longitud. Estos resultados son seriamente comparables con los que aparecen en la literatura (alrededor del 80-85% de buena clasificación).

El algoritmo propuesto para la clasificación de estructuras secundarias de secuencias de aminoácidos permite clasificar las secuencias en tres tipos: beta-laminar, alfa hélice o coil. Este algoritmo evita la clasificación empleando la alineación de secuencias (ruta clásica para este tipo de análisis), constituye una alternativa en el análisis de las estructuras secundarias que se enfoca en la conservación de la estructura más que en la secuencia, y toma en cuenta además, las características físico-químicas de los aminoácidos a partir de las Álgebras de Boole del Código Genético. Aunque los resultados son satisfactorios, para su implementación final en el análisis de las estructuras secundarias de proteínas, el algoritmo debe ser complementado con otros algoritmos que permitan el análisis de la secuencia completa de aminoácidos en la proteína, sin su subdivisión por partes.

Tabla 2. Resultados del nuevo algoritmo de clasificación en la base de datos de prueba (DSSP) después de haber aprendido con la base de datos (ISSD). Nótese los altos por cientos de buena clasificación en secuencias con estructuras E de baja longitud y en secuencias H más largas.

Secuencias de estructura conocida	Longitud Media	Longitud Media Distancia de Hamming Entropía Relativa	Información Mutua	Valor de la Información Entropía Relativa
100 secuencias E	5.48 5.42	98% 99%	61% 71%	64%
100 secuencias H	12.8	7%	72%	97%

Conclusiones

Se ha demostrado que la sola codificación binaria adecuada del código genético permite la construcción de **árboles filogenéticos** y su integración en una red Bayesiana que es útil para clasificar mutaciones de una proteína, particularmente la proteasa del VIH y luego facilitar la identificación de las familias o subfamilias de ellas. Se ha demostrado también cómo la misma codificación, usando además la distancia de Hamming y el concepto de Información mutua en el código genético, visto como sistema de información, permite nuevos enfoques para la solución de problemas más complejos, como el de la clasificación de estructura secundaria. Todo ello confirma que las estructuras booleanas del código genético nos ayudan a comprender la lógica subyacente en el mismo y brindan nuevas herramientas para la posible solución de problemas de la bioinformática.

Referencias Bibliográficas

Adzhubei, I. A. and A. A. Adzhubei. ISSD: Integrated Sequence -- Structure Database, Versión 2.0 Nucleic Acids Res, 1999, (27): 268-271, *Database in* http://www.protein.bio.msu.ru/issd/ 2001].

Alf-Steinberger, C. The Genetic Code and Error Transmission. Proc. Natl. Acad. Sci, 1969, 64(2): 584-591.

Balakrishhnan J. Symmetry Scheme for Amino Acid Codons. Phys. Rev. E, 2002, 65(2): 219-225.

Baldi, P. and S. Brunak, Bioinformatics The Machine Learning Approach. 2. MIT Press, 2001.

Bashford, J.D., I. Tsohantjis and P. D. Jarvis. A Supersymmetric Model for the Evolution of the Genetic Code Proc. Natl. Acad. Sci. USA, 1998, 95(3): 987-992.

Bashford, J.D. and P. D. Jarvis. The Genetic Code as a Periodic Table Biosystems, 2000, 57(3): 147-161.

Bertman, M.O. and J. R. JUNGCK. Group Graph of the Genetic Code J. Hered, 1979, 70(6): 379-384.

Bouton, E.A., H. M. de Oliveira, R. M. Campello de Souza, N. S. Santos-Magalhaes. Genomic Signal Analysis Based on Codongrams and a2grams, WSEAS Transactions Biology and Biomedicine, 2004, 1(2): 255-260.

Castillo E., J. M. Gutiérrez and A. S. Hadi. Expert Systems and Probabilistic Network Models, Springer-Verlag, 1996.

Chávez, M.C y L. O. Rodríguez. Bayshell, Software para crear redes Bayesianas e inferir evidencias en la misma, Registro de Software CENDA, mayo 2002.

Chávez, M. C., R. Grau y M. M. García. Un método para construir redes Bayesianas, Revista de Ingeniería de la Universidad de Antioquia, 1999, (19).

Crick, F.H.C. The Origin of the Genetic Code. J. Mol. Biol., 1968, 38(3): 367-379.

Freeland, S. and L. Hurst. The Genetic Code is One in a Million, Mol. Evol, 1998, 47(3): 238-248.

Friedman, S. M. and I. B. Weinstein. Lack of Fidelity in the Translation of Ribopolynucleotides, Proc. Natl. Acad. Sci. USA, 1964, 52: 988-996.

Grau, R.A. Estadística Aplicada con ayuda de paquetes de software, Universidad de Guadalajara, Editorial Universitaria, 1994.

Hunter, L. Planning to Learn About Protein Structure, in Hunter L. (ed) Artificial Intelligence and Molecular Biology, Cambridge, AAAI Press Book, 2003.

Jiménez-Montaño, M.A. The Hypercube Structure of the Genetic Code Explains Conservative and Non-Conservative Amino Acid Substitutions *in vivo* and *in vitro*, Biosystems, 1996, 39(2): 117-125.

.... Protein Evolution Drives the Evolution of the Genetic Code and Vice Versa, Biosystems, 1999, 64(2): 47-64.

Kabsch, W. and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features DSSP: Database of Secondary Structure in Proteins, 1983, 22: 2577-2637, *Database in* http://www.sander.ebi.ac.uk/dssp/ , 1998.

Karasev, V. A. and V. E. Stefanov. Topological Nature of the Genetic Code, J. Theor. Biol., 2001, 209(3): 303-317.

Maccallum, R. M. Computational Analysis of Protein Sequence and Structure. Thesis submitted to the University of London for the Degree of Doctor or Philosophy in the Faculty of Science, 1997.

Parker, J. Errors and Alternatives in Reading the Universal Genetic Code, Microbiol. Rev, 1989, 53(3): 273-298.

Sánchez, R., E. Morgado and R. Grau. A Genetic Code Boolean Algebras, WSEAS Transactions Biology and Biomedicine, 2004a, 1(2): 190-197.

.... The Genetic Code Boolean Lattice, MATCH Commun. Math. Comput. Chem, 2004b, (52): 29-46.

.... A Genetic Code Boolean Structure. I. Meaning of Boolean Deductions, Bulletin of Mathematical Biology, 2005a, (67): 1-14.

.... A Genetic Code Boolean Structure. II. The Genetic Information System as a Boolean Information System, Bulletin of Mathematical Biology, 2005b, Article in press, [http://www.sciencedirect.com/science].

Siemion I. Z., S. P. Siemion and P.J. Krajewski. Chou-Fasman Conformational Amino Acid Parameters and the Genetic Code, Biosystems, 1995, 36: 231-238.

SPSS Inc. SPSS for Windows CHAID. User Manual, 1994a.

.... SPSS for Windows. User Manual, 1994b.

Stuart, T. and P. Norvig. Inteligencia Artificial: Un enfoque Moderno, México, Prentice Hall, 1996.

Swanson, R. A. Unifying Concept for the Amino Acid Code. Bulletin of Mathematical Biology, 1984, 46(2): 187-203.

Williams W. L., R. C. Wilson and E. R. Hancock. Multiple Graph Matching with Bayesian inference, Pattern Recognition Lett., 1998, 38: 11-13.