https://www.uaeh.edu.mx/scige/boletin/icbi/historial.html

Boletín Científico del Instituto de Ciencias Básicas e Ingeniería

Pädi

Publicación Semetral Pädi No. 13 (2019) 84–89

Minería de patrones frecuentes para la comprobación de las reglas de Susumu Ohno y su uso en la identificación de evolución de organismos biológicos

Luis H. García-Islas[a], Anilu Franco-Arcega[a,*], J. Antonio Quiroz-Gutierrez[b], Kristell D. Franco-Sanchez[a]

[a] *Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo. Ciudad del Conocimiento, Kilómetro 4.5 carretera Pachuca - Tulancingo Col. Carboneras de Mineral de la Reforma, Hidalgo, México. C.P. 42184*

[b] *Universidad Autónoma del Carmen. C. 56 No. 4. Esq. Avenida Concordia Col. Benito Juarez C.P. 24180 Cd. del Carmen, Campeche, México*

Resumen

Con el desarrollo de la bioinformática, la búsqueda de patrones frecuentes en secuencias de ADN se ha vuelto un punto de atención crítico. Los patrones de secuencia biológicos, especialmente patrones que se repiten, usualmente reflejan alguna característica funcional o estructural importante. Como resultado de varios estudios, Susumu Ohno propuso una serie de reglas que las secuencias de ADN deben cumplir para la propia evolución de esos organismos biológicos. El presente trabajo realizó una validación de dichas reglas basándose en técnicas de minería de datos y cadenas de Markov evaluando 32,074 secuencias de ADN de diversos organismos biológicos obtenidos de la base de datos biológica del repositorio GenBank, perteneciente al Centro Nacional de Información de Bioinformática del departamento de salud de Estados Unidos, y con lo cual se pudo identificar organismos que poseen particularidades que pueden diferenciarlos en su evolución.

Palabras Clave: Minería de Datos, Patrones frecuentes, Cadenas de Markov, ADN, Bioinformática

1. Introducción

La biología puede definirse como el estudio integral de los seres vivos (Campbell et al., 2014). Todo ser vivo, también llamado organismo, comparte una serie de características o funciones, las cuales son:

1. Contar con una estructura organizada compleja.
2. Capacidad de adquirir energía y materiales del exterior y transformarlos.
3. Capacidad de autorregulación.
4. Capacidad de crecer y desarrollarse, siguiendo un programa genético.
5. Capacidad de responder a estímulos del medio ambiente.
6. Reproducirse utilizando una huella molecular llamada ADN.
7. Capacidad de evolucionar.

El ADN contiene información genética la cual es participe de dos procesos clave, la replicación y la expresión génica. Por medio de la replicación, el ADN se duplica, permitiendo la distribución equitativa del material genético a las células hijas como parte del proceso de división celular. Por otro lado, en la expresión génica, las enzimas leen a los genes, produciendo una síntesis de proteínas como resultado de este proceso. De manera general, se puede decir que el ADN se autorreplica, ya sea generando más ADN o transcribiéndose y generando ARN mensajero (transcripción), el cual experimenta un proceso de traducción durante la síntesis de proteínas. Francis Crick (1958) planteó que la información genética fluye del ADN al ARN y luego a las proteínas, nunca en sentido inverso, lo que se conoce como el dogma central de la biología molecular. La Figura 1 muestra este dogma, y en la cual se pueden observar los procesos que realizan los ácidos nucleicos (transcripción y traducción). Dichos procesos son de significante importancia en el área de la biología y en la vida misma (Calvo, 2015).

El ADN es una molécula muy grande pero compuesta sólo por pocas sustancias diferentes: Adenina (A), Citosina (C), Guanina (G) y Timina (T). Una secuencia está formada por unidades llamadas nucleótidos. EL ADN sirve como base para el proceso de transcripción, en donde se sintetiza un ARN usando como molde al ADN. En esta transcripción el ARN, a diferencia del ADN que tiene una forma de espiral de doble hélice, tiene una sola cadena de nucleótidos formada por 4 elementos: Adenina (A), Citosina (C), Guanina (G) y Uracilo (U). Después, a través del código genético, el ARN hace una traducción de éstas para obtener enzimas, las cuales están representadas por

*Autor en correspondencia.

Correos electrónicos: luishg@uaeh.edu.mx (Luis H. García-Islas), afranco@uaeh.edu.mx (Anilu Franco-Arcega), zerreit@yahoo.com (J. Antonio Quiroz-Gutierrez), kristell_franco@uaeh.edu.mx (Kristell D. Franco-Sanchez)

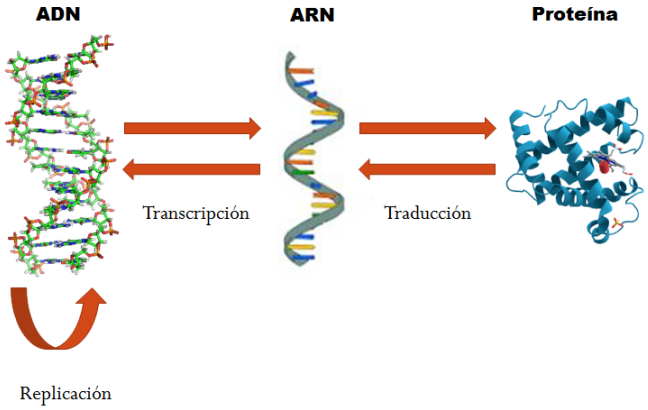

Figura 1: Dogma central de la Biología Molecular

Segunda letra

Primer letra	U	C	A	G	Tercer letra
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA STOP UAG STOP	UGU Cys UGC UGA STOP UGG Trp	U C A G
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG	U C A G

Figura 2: Código genético

proteínas. El código genético está representado por tripletas de nucleótidos (también llamados trímeros o codones), los cuales producen 64 diferentes combinaciones de los 4 nucleótidos de ARN (por ejemplo, ACG,CGU, etc.). La Figura 2 muestra la tabla de tripletas que se usan en la traducción de ARN a proteínas.

Un aspecto importante a estudiar dentro de la biología es la evolución de los organismos, y es precisamente la biología evolutiva el área que establece las relaciones filogénicas por medio del estudio de cómo se desarrollan los organismos biológicos. Además, esta disciplina permite identificar los mecanismos del desarrollo que permiten los cambios evolutivos de los individuos (Hall, 2003).

En 1970, Susumu Ohno, en su libro Evolution by Gene Duplication, supone que genes esenciales producidos son indispensables en la evolución para la supervivencia de los miembros de cualquier especie. Estos genes no pueden producir de manera potencial a nuevos genes por medio de mutaciones que alteren su función primaria. Sin embargo, al duplicar en una línea germina un gen esencial, en la copia adicional se permitirían mutaciones que proporcionarían la información genética para su función esencial.

Estos estudios dieron pié a una serie de reglas que Susumu Ohno propuso basado en el estudio de dímeros (pares de bases nucléicas).

Como se describe en la definición de una secuencia de ADN, cada secuencia codificada de ADN es una combinación infinita a partir de cuatro tipos de bases (ACGT). Ohno sugiere que la frecuencia excesiva de un elemento base produce una notable subrepresentación de los otros (Ohno, 1988). Con base en esto, Ohno propone una serie de reglas canónicas para todas las secuencias de ADN, las cuales son:

1. Sin importar la composición de los dímeros (conjunto de dos monomeros), las secuencias de código lo suficientemente largas muestran una notable deficiencia de dos dímeros base: TA Y CG.
2. Entre los cuatro dímeros base que poseen la base T como su primer elemento, esta deficiencia de TA siempre es compensada por un exceso de TG. Debido a que el dímero TG es también uno de los cuatro dímeros base teniendo G como segunda base, el exceso de este dímero se compensa con una deficiencia de CG.
3. Entre los cuatro dímeros compartiendo C como primer base, es el dímero CT el que compensa la deficiencia de CG.
4. En el caso de las cuatro bases que tienen el elemento A como su segunda base, la deficiencia de TA puede ser compensada por el exceso de cualquiera de las tres bases de dímero restantes (CA, GA o AA).

De acuerdo a los diversos estudios realizados por Ohno (como se citó en Osawa y Honjo, 2015) se ha podido comprobar que todas las secuencias de ADN base, sin importar su origen o función son mensajes escritos en palíndromos. De igual forma, las reglas de Ohno han sido de gran utilidad para comprobar la influencia de dichos patrones en la selección natural y en la evolución de los genes.

De manera general, un gen se puede definir como una partícula de material genético que, junto con otras, se halla dispuesta en un orden fijo a lo largo de un cromosoma, y que determina la aparición de los caracteres hereditarios en los seres vivos. Es la unidad funcional y física de la herencia que pasa de padres a hijos. Este gen es representado por un segmento de una secuencia de ADN. Para poder encontrar estos fragmentos, existen diversas técnicas, entre ellas la identificación de patrones frecuentes, que es parte de la minería de datos.

Los Patrones Frecuentes son conjuntos de elementos, subsecuencias o subestructuras que aparecen en un conjunto de datos cuya frecuencia no es menor que un umbral definido por el usuario. La minería de patrones frecuentes fue propuesta para análisis de mercado a través de reglas de asociación (Agrawal et al., 1993).

El uso de la Minería de Patrones Frecuentes (Frequent Pattern Mining por su traducción al inglés) ha sido ampliamente estudiado en la literatura debido a sus numerosas aplicaciones en una amplia variedad de problemas de minería de datos, como la Minería de Patrones Secuenciales (Aggarwal y Han, 2014). Esta área tiene múltiples aplicaciones en diversos entornos como datos espacio temporales, detección de errores en software y datos biológicos (Kumar et al., 2017; Medeiros et al., 2016; Mutakabbir et al., 2014).

En el caso particular de la minería de patrones secuenciales, se busca obtener patrones estadísticamente relevantes en conjuntos de datos que están representados en forma secuencial, tales como las secuencias de ADN. Los patrones frecuentes obtenidos son utilizados en tareas de detección de dependencias funcionales, predicción de tendencias, interpretación de fenómenos y como soporte de decisiones en estrategias de producción (Hossain et al., 2013; Liu et al., 2014; Tahir et al., 2017).

Considerando lo anterior, este trabajo realizó un estudio por medio de técnicas de minería de patrones frecuentes basadas en cadenas de Markov que permitiera comprobar las reglas de Susumu Ohno. El resultado fue contrastado con una referencia existente para comprobar la factibilidad del estudio realizado por Kumal et al. (2017).

2. Descripción del método y resultados

Para llevar a cabo este estudio, se tomaron 32,074 secuencias de ADN de diferentes organismos de la base de datos biológica del GenBank (2017) perteneciente al NCBI (National Center of Biotechnology Information - Centro Nacional de Información de Biotecnología). Los organismos obtenidos se muestran en la Tabla 2.

Tabla 1: Secuencias obtenidas del NCBI

Base	# Secuencias
AH1N1	83
Flu	163
Zika	725
HIV	1654
Cactaceae	1677
Chikungunya	4459
Diabetes	8313
Petroleum	15000

El método utilizado para la identificación de patrones frecuentes en grupos de secuencias fue el método de minería de patrones de secuencias basado en cadenas de Markov (Sequence Pattern Mining Based on Markov Chain - SPMM) propuesto por Junyan and Chenhui (2015), el cual es rápido y eficiente para secuencias de ADN.

El algoritmo implica que dada una base de datos de secuencias y un umbral de grado de soporte mínimo, la minería de patrones de secuencias es la búsqueda de todas las subsecuencias frecuentes de modo tal que la probabilidad de dichas subsecuencias de aparecer en la base de datos no es menor que dicho umbral. Los autores crean una matriz de probabilidades de transiciones para cada secuencia en la base de datos como se describe a continuación. Dada una secuencia de ADN, la probabilidad pXY representa la probabilidad de transición de un nucleótido X a un Y (probabilidades de dímeros). La Figura 3 muestra como se deben considerar las combinaciones de dos nucleótidos para construir la matriz de probabilidades, con base en la ocurrencia de los dímeros en la secuencia.

$$P_d = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

Figura 3: Matriz de probabilidades

De manera inicial, el algoritmo establece una matriz de probabilidades de dímeros para cada secuencia. Por ejemplo, teniendo dos secuencias $s_1 = GACTGCCTACTGAAC$ y $s_2 = TTCTGACACTGTTC$, las correspondientes matrices de probabilidades se muestran en las Figuras 4 y 5.

$$P_{s_1} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} \\ 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \end{bmatrix}$$

Figura 4: Matriz de probabilidades para la secuencia s_1

$$P_{s_2} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{2}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

Figura 5: Matriz de probabilidades para la secuencia s_2

Una vez hecho esto, se obtiene la probabilidad de la subsecuencia que necesita ser evaluada para determinar si es un patrón frecuente. Esto se hace evaluándola en cada matriz de probabilidad, obteniendo el promedio de todas ellas para así calcular la probabilidad de que la subcadena en cuestión aparezca dentro de la base de datos de secuencias. Finalmente,

el cálculo obtenido se evalúa con un umbral establecido por el usuario, con el fin de determinar si es un patrón frecuente. Por ejemplo, para la subsecuencia $\alpha = CTG$, será necesario calcular la probabilidad multiplicando la probabilidad del dimero CT y la del dimero TG. Utilizando cada una de las matrices correspondientes a cada secuencia, para la subsecuencia en cuestion, las probabilidades son $P_{s_1}(\alpha) = \frac{3}{4} * \frac{2}{3} = \frac{1}{2}$ y $P_{s_2}(\alpha) = \frac{2}{3} * \frac{1}{3} = \frac{2}{9}$, las cuales se promedian y obtienen como resultado $P_\alpha = \frac{1}{2}(\frac{1}{2} + \frac{2}{9}) = 0{,}36$. Si se determina un umbral arbitrario de 0.2, la subsecuencia CTG es un patrón frecuente del conjunto de secuencias procesadas. Todas aquellas secuencias candidatas cuya probabilidad sea mayor que el umbral establecido, por definición de las cadenas de Markov pueden ser consideradas como patrones frecuentes.

Considerando las 32,074 secuencias del presente estudio, la Tabla 2 muestra las probabilidades de los dímeros que son los utilizados para las comprobaciones de las reglas de Susumu Ohno, los cuales son, TA, CG, TG, CT, CA, GA y AA. Una vez que se obtuvo la matriz de probabilidades de dímeros, dichas probabilidades se evaluaron con las reglas universales de Susumu Ohno, cuyos resultados se pueden ver en la segunda parte de la Tabla 2.

Como puede observarse en la Tabla 2, todos los organismos cumplen con las reglas de Susumu Ohno, excepto el organismo Cactaceae, en particular con la regla de la deficiencia de CG es compensada por el exceso de CT. De igual forma, la Figura 6 muestra gráficamente las probabilidades para cada uno de los dímeros utilizados en las reglas de Susumu Ohno, donde puede apreciarse mejor la variación de las probabilidades entre dímeros y la diferencia de comportamiento del organismo Cactaceae.

Al detectar que para el organismo Cactaceae no se cumplió la regla que involucra a los dímeros CG y CT, se procedió a la obtención de sus aminoácidos haciendo uso del código genético, tal como lo dicta el dogma de la biología molecular (Calvo, 2015). Al realizar el proceso de traducción de ARN a aminoácidos por medio del código genético, todas las combinaciones del dímero CG (CGA, CGC, CGG, CGT), al realizar su replicación hacia ARN (CGA, CGC, CGG y CGU) y su respectiva traducción a aminoácidos, se produce en todos los casos Arginina, mientras que en las combinaciones respectivas para el dímero CT (CTA,CTC, CTG, CTT), su trascripción a ARN (CUA, CUC, CUG y CUU) produce Leucina en todas las combinaciones. Una vez que se obtuvieron los aminoacidos resultantes sobre los dímeros que no cumplen con una regla particular de Ohno, se contrastó con un estudio realizado por Kumar et al. (2017) el cual reporta el contenido relativo de los 21 aminoácidos que se pueden obtener de acuerdo al código genético en diversos tipos de organismos de plantas, incluyendo al organismo Cactaceae, este contenido se puede observar en la Tabla 3.

De acuerdo con el estudio de Kumar et al. (2017), al realizar una distribución de aminoácidos en esta familia de plantas, se observa que el porcentaje de arginina (1.8) es mayor que el porcentaje de leucina (1.7), lo cual coincide con la regla de Susumu Ohno que no se cumple para este organismo, es decir, la deficiencia de CG (arginina traducido en aminoácidos) se compensa con el exceso de CT (leucina en amionácidos). Esto coincide con el análisis realizado en este trabajo, en donde se presenta la validación de diferentes organismos con las reglas de Susumu Ohno.

3. Conclusión

El presente estudio permitió obtener información acerca de la identificación de la evolución de diversos organismos biológicos a través de la comprobación de las reglas de Susumu Ohno mediante el uso de técnicas de Minería de Patrones Frecuentes, demostrando que de manera general dichas reglas se cumplen. Sin embargo, mediante las reglas de Ohno se identificó que para el organismo Cactaceae no cumple con todas las reglas, en especial la regla que involucra a los dímeros CG y CT. Este caso en particular, pudo ser comparado a través de un estudio que muestra la concentración de aminoácidos, los cuales son traducidos a partir de los dos dímeros en cuestión. Por medio de este estudio se muestra como la minería de datos puede ser aplicada en la bioinformática como apoyo al estudio del análisis de organismos biológicos, en este caso, para la biología evolutiva. Con el uso de las técnicas de minería de patrones frecuentes, se pudo tener un indicio de que el organismo Cactaceae tiene aspectos muy especiales en términos de su evolución. El presente trabajo puede servir como apoyo para estudios más profundos que comprueben, y si es posible expliquen, la razón de porqué este organismo tiene particularidades en su evolución.

English Summary

Frequent pattern mining to evaluate Susumu Ohno's rules and its use on the identification of evolution for biological organisms

Abstract

Within development of Bioinformatics, pattern mining has become a critical point of attention. Patterns on Biological sequences, specially repeated patterns, usually shows relevant structural o functional features. As a result of several studies, Susumu Ohno proposed a set of rules that most species should fulfill in order to accomplisg their evolution aspects. The present work validates such rules using frequent pattern mining and Markov Chain Techniques assessing 32,074 DNA sequences from diverse biological organisms from GenBank Biological database as a part of National Center of Biotechnology Information from National Health Institute of the United States. This could identify organisms that possess particularities that differs on their evolution

Keywords:

Data mining, Frequent Pattern, Markov Chain, DNA, Bioinformatics

Tabla 2: Matriz de probabilidades de dímeros y comprobación de reglas de Susumu Ohno

Base	AH1N1	Flu	Zika	HIV	Cactaceae	Chikungunya	Diabetes	Petroleum
Secuencias	83	163	725	1654	1677	4459	8313	150000
pTA	0.2046	0.1866	0.1709	0.1647	0.2673	0.2557	0.1578	0.2246
pCG	0.1091	0.1454	0.1452	0.1074	0.3190	0.2043	0.1121	0.0843
pTG	0.3197	0.3077	0.3528	0.3187	0.3618	0.3165	0.3291	0.2617
pCT	0.2650	0.2768	0.2831	0.3040	0.2363	0.2385	0.3004	0.3282
pCA	0.4181	0.3568	0.3292	0.3083	0.2456	0.3292	0.3030	0.3628
pGA	0.3757	0.3576	0.3077	0.2804	0.2424	0.2566	0.2738	0.2844
pAA	0.3318	0.3238	0.2854	0.2981	0.3039	0.2646	0.2837	0.3246
fTA<fTG	Si	Si	Si	Si	Si	Si	Si	Si
fCG<fTG	Si	Si	Si	Si	Si	Si	Si	Si
fCG<fCT	Si	Si	Si	Si	No	Si	Si	Si
fTA < (fCA,fGA,fAA)	Si	Si	Si	Si	Si	Si	Si	Si
CumpleTodas	Si	Si	Si	Si	No	Si	Si	Si

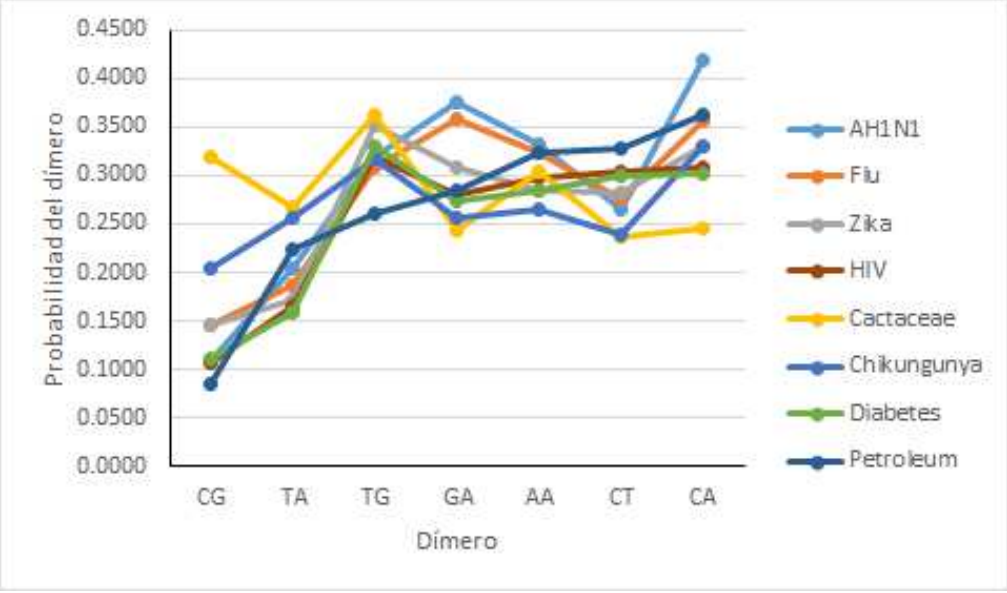

Figura 6: Probabilidades de los dímeros

Tabla 3: Contenido relativo de aminoácidos

Aminoácido	Contenido relativo	Aminoácido	Contenido relativo
Alanina	1.4	**Leucina**	1.7
Arginina	1.8	Lisina	1
Asparagina	0.7	Metionina	0.4
Aspartato	2.4	Fenilalanina	0.9
Cisteína	0.5	Prolina	1.1
Glutamina	0.6	Serina	1
Ácido glutámico	2.8	Treonina	1
Histidina	0.5	Triptófano	0.3
Isoleucina	1.7	Tirosina	0.7
Valina	1.2		

Agradecimientos

Agradecemos el apoyo del Área Académica de Biología del Instituto de Ciencias Básicas e Ingeniería de la Universidad Autónoma del Estado de Hidalgo, en especial al Dr. Francisco Nuñez de Cázares y al Dr. Julian Bueno Villegas por su orientación en los temas de biología.

Referencias

Aggarwal, C. C., Han, J., 2014. Frequent pattern mining. Springer.

Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Acm sigmod record. Vol. 22. ACM, pp. 207–216.

Calvo, A., 2015. Biología celular biomédica + StudentConsult en español. Elsevier, Barcelona.

Campbell, N., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B., Reece, J. B., 2014. Campbell biology. Pearson, Boston.

Crick, F., 1958. On protein synthesis. In: Symposium of the Society for Experimental Biology XII. New York: Academic Press.

Hall, B. K., 2003. Unlocking the black box between genotype and phenotype: Cell condensations as morphogenetic (modular) units. Biology and Philosophy 18 (2), 219–247.

Hossain, K. S. M. T., Patnaik, D., Laxman, S., Jain, P., Bailey-Kellogg, C., Ramakrishnan, N., Sept 2013. Improved multiple sequence alignments using coupled pattern mining. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10 (5), 1098–1112.
DOI: 10.1109/TCBB.2013.36

Junyan, Z., Chenhui, Y., 2015. Sequence pattern mining based on markov chain. In: Information Technology in Medicine and Education (ITME), 2015 7th International Conference on. IEEE, pp. 234–238.

Kumar, V., Sharma, A., Kaur, R., Thukral, A. K., Bhardwaj, R., Ahmad, P., 2017. Differential distribution of amino acids in plants. Amino Acids 49, 821–869.

Liu, X., Wu, J., Gong, H., Deng, S., He, Z., Sept 2014. Mining conditional phosphorylation motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 11 (5), 915–927.
DOI: 10.1109/TCBB.2014.2321400

Medeiros, I., Neves, N., Correia, M., March 2016. Detecting and removing web application vulnerabilities with static analysis and data mining. IEEE Transactions on Reliability 65 (1), 54–69.
DOI: 10.1109/TR.2015.2457411

Mutakabbir, K. M., Mahin, S. S., Hasan, M. A., 2014. Mining frequent pattern within a genetic sequence using unique pattern indexing and mapping techniques. In: Informatics, Electronics & Vision (ICIEV), 2014 International Conference on. IEEE, pp. 1–5.

National Center of Biotechnology Information, 11 2017. Genbank and wgs statistics. consultado el 15-04-2019 desde https://www.ncbi.nlm.nih.gov/genbank/.

Ohno, S., 1988. Universal rule for coding sequence construction: Ta/cg deficiency-tg/ct excess. Proceedings of the National Academy of Sciences 85 (24), 9630–9634.

Osawa, S., Honjo, T., 8 2015. Evolution of Life. Fossils, Molecules, and Culture. Springer?Verlag.
DOI: 10.1046/j.1420-9101.1992.5040725.x

Tahir, M., Hayat, M., Kabir, M., 2017. Sequence based predictor for discrimination of enhancer and their types by applying general form of chou's trinucleotide composition. Computer Methods and Programs in Biomedicine 146, 69 – 75.
DOI: https://doi.org/10.1016/j.cmpb.2017.05.008