

OPTIMIZACIÓN DE COSTOS EN LA PRODUCCIÓN DE PROTEÍNAS EN EL GENOMA HUMANO

Araxi Urrutia Odabachian
bspauo@bath.ac.uk

*Estudiante de doctorado en el Departamento de Biología y Bioquímica
de la Universidad de Bath, Reino Unido*

Realizó sus estudios de licenciatura en la Facultad de Psicología, UNAM

OPTIMIZACIÓN DE COSTOS EN LA PRODUCCIÓN DE PROTEÍNAS EN EL GENOMA HUMANO

RESUMEN

Casi la totalidad de las células en nuestro cuerpo contienen una copia de nuestro genoma. Los genes contienen la información necesaria para producir todas las proteínas que realizan todas las actividades en la célula. Enfermedades como el cáncer o defectos hereditarios pueden resultar de errores en la secuencia de los genes. Hoy a más de 50 años del descubrimiento de la estructura del ADN, los investigadores cuentan con la secuencia completa del genoma humano. En contraste, el conocimiento acerca de la organización del genoma y la función de los genes es muy pobre. Dado que el tamaño de la población de las especies de mamíferos es reducida, generalmente se asume que la fuerza de la selección natural es relativamente débil. A continuación se presentan los resultados más recientes que evalúan esta hipótesis. La comparación de los niveles de actividad con las características físicas (como su longitud o composición) de los genes, ha revelado que aquellos genes con mayor actividad tienen características peculiares. Estas observaciones se cuentan entre las primeras en mostrar que las secuencias génicas muestran patrones consistentes con la acción de la selección natural para aumentar la eficiencia de la síntesis de proteínas.

Palabras clave. Genoma humano, gene, cromosoma, codon, intron, nucleótido, aminoácido, actividad génica, selección natural, teoría neutral.

COST OPTIMIZATION OF PROTEIN SYNTHESIS IN THE HUMAN GENOME

ABSTRACT

Virtually all cells in our body hold a complete copy of the genome. The 35000 human genes contained in the genome are required to produce all the proteins in the cell allowing us to carry out the activities of our daily lives. If our genes go wrong then diseases such as cancer might develop or inherited defects past on to future generations. On the 50th anniversary of the discovery of the DNA structure, the human genome sequence has already been cracked. Still very little is known about its organization and the function of most genes it contains. It is usually assumed that selection mediated by expression is a weak force in determining gene characteristics in humans. Here we present recent evidence of a relationship between physical characteristics of genes with levels of activity in human genes. We found that highly active genes tend to be smaller and have a particular nucleotide composition compared to other genes of lower activity. Together these results constitute the first systematic studies to show that gene sequences in mammalian genomes show signatures consistent with the action of natural selection to increase efficiency of protein synthesis.

Keywords. Human genome, gene, chromosome, codon, intron, nucleotide, aminoacid, gene expresion, natural selection, neutral theory.

LA ERA GENÓMICA

Una de las áreas de la Biología con mayor expansión desde mediados del siglo XX es la Genética. La importancia del estudio de los genes y su función son fundamentales para el entendimiento de los procesos celulares y el funcionamiento de los órganos y sistemas del cuerpo. Hace cinco décadas fue resuelta la estructura de la molécula que guarda la información hereditaria (ADN). Años después se descifró el código genético que permitió conocer la correspondencia entre la secuencia de un gen y la proteína que se produce a partir de este. A casi medio siglo de estos importantes avances, la comunidad científica cuenta con la secuencia completa del genoma humano (Lander et al. 2001; Venter et al. 2001), así como el de otras especies de eucariontes (Levadura, Mosca de la fruta, nematodo, el pez fugu, el ratón y la planta *arabidopsis*). La conclusión del proyecto del genoma humano que se avocó a descifrar la secuencia genómica ha permitido conocer características generales del reservorio de información de nuestra especie, así como la inferencia de funciones para los productos génicos. Sin embargo, también se ha hecho evidente que la secuencia por si sola provee de información muy limitada acerca del funcionamiento de los productos génicos en los procesos fisiológicos que permitan controlar la función de las proteínas para el diagnóstico y tratamiento de padecimientos hereditarios y adquiridos, así como, el mejoramiento de especies animales y vegetales relevantes para el consumo humano. La explotación de la información de nuestro acervo génico, requiere de un análisis sistemático de las características del genoma.

DIVERSIDAD EN LAS CARACTERÍSTICAS DE LOS GENES

Las instrucciones para la producción de proteínas necesarias en el desarrollo y funciones metabólicas del organismo están contenidas en los genes. En el núcleo de la célula los genes se encuentran distribuidos entre los 22 pares, más dos cromosomas sexuales (XX en las mujeres y XY en los hombres). Los genes humanos presentan una gran variabilidad tanto en sus características tales como longitud y composición así como en su distribución a lo largo del genoma. Así por ejemplo, los genes que codifican proteínas ribosomales importantes suelen no ser mayores de 1500 bases mientras que el gen más largo sobrepasa las 250000 bases. La proporción del uso de cada una de las cuatro bases nitrogenadas que constituyen el alfabeto de los genes para codificar proteínas también varía de un gen a otro, mientras que algunos genes utilizan todas las bases en las mismas cantidades, otros genes, dos de las cuatro bases constituyen más del 85% de la secuencia.

¿A qué se deben estas variaciones? ¿Qué factores determinan las características de los genes? A continuación describiré los estudios que se han avocado a dar respuesta a estas preguntas pero antes es justo preguntar ¿Por qué es importante el estudio de las características de los genes y el genoma humano? El conocimiento de los procesos que contribuyen a la creación de los patrones generales observados en genoma es potencialmente relevante para proyectos de terapia génica y modificación genética pero también en la evaluación y previsión de riesgos de estos procedimientos.

SELECCIÓN VS. AZAR

Tras este breve paréntesis volvamos a las preguntas que nos ocupan ¿Qué determina la longitud de un gen y su composición de bases? Es posible delinear dos hipótesis generales para explicar las diferencias de composición y longitud entre los genes. Estas dos hipótesis corresponden a las dos posiciones que mantienen los investigadores en el área de evolución molecular. La primera se deriva del modelo de evolución neutral propuesto por Kimura en 1969 (para una revisión reciente ver [1991]). Kimura propuso que para la mayoría de los cambios que se observan a nivel molecular durante la evolución, la selección natural es muy débil como para determinar su futuro dentro de la población. Bajo este modelo, además de las restricciones en la secuencia derivadas de las funciones que cumplen las proteínas, las características que observamos en los genes son resultado del azar.

La segunda hipótesis, seleccionista, se deriva de la posición complementaria a la teoría de evolución neutral. En este caso, se asume que la mayoría de los cambios que ocurren, son favorecidos o rechazados por presiones selectivas (Li 1997). Bajo este argumento, podríamos esperar que las características de los genes muestren los efectos de la selección natural para maximizar su eficiencia en la síntesis de proteínas. Gracias a que las dos hipótesis predicen diferentes patrones en las características de los genes es posible distinguir entre estas dos posibilidades. La teoría neutral predice distribuciones aleatorias mientras que los seleccionistas predicen orden y patrones definidos.

El análisis génico de especies de bacterias y organismos multicelulares invertebrados, han revelado que las características generales de los genes están en parte determinadas por presiones selectivas para la optimización en la síntesis de proteínas. Hace apenas unos años, la idea aceptada era que para el caso de humanos y mamíferos en general, las características génicas eran determinadas por procesos estocásticos y de deriva génica. Esto debido a lo reducido de las poblaciones de especies de mamíferos en comparación con otros grupos de organismos. Sin embargo, no se había realizado estudio alguno a escala genómica para aceptar o rechazar esta teoría. Los datos disponibles hasta hace muy poco no permitían el estudio sistemático de factores generales que influyen características génicas, en particular estimaciones de niveles de actividad. En los últimos años, los avances tecnológicos en las áreas de biología molecular e informática aunados a los esfuerzos de numerosos grupos de científicos alrededor del mundo, han permitido la acumulación de una gran cantidad de datos sobre características diversas de un gran número de genes.

A continuación se presentan los resultados recientes derivados de análisis sistemáticos de bases de datos de millares de genes humanos. En estos, se comparan los datos de niveles de actividad génica provenientes de dos bases de datos (Su et al. 2002; Velculescu et al. 1995) y se comparan características tales como longitud, contenido de intrones, composición de bases, frecuencias de amino ácidos en la proteína producida, etc. Los resultados derivados de estos estudios, lejos de reflejar la ocurrencia de eventos estocásticos en las características génicas, muestran efectos compatibles con la acción de procesos de selección natural y apoyan la visión de que muchas de las características génicas se encuentran estrechamente relacionadas con los niveles de actividad de los genes.

OPTIMIZACIÓN DE LA LONGITUD DE GENES HUMANOS

Para el correcto funcionamiento de las células, distintas proteínas se requieren en distintas cantidades. Dado que los recursos para las células son limitados podemos asumir que la síntesis de proteínas implica un costo para la célula. De ser el caso, podría esperarse que aquellas proteínas que se producen en grandes cantidades fueran más cortas. Así, desde una perspectiva seleccionista, podríamos esperar una relación inversa entre los niveles de actividad de un gene y la longitud de la proteína que codifica. La hipótesis nula predice que la reducción de costos asociadas con el acortamiento de las proteínas, no son suficientes como para ser favorecidos por la selección, por lo tanto no se espera relación entre las dos variables. Nuestros resultados (Urrutia y Hurst, 2003), muestran que aquellos genes de alto nivel de expresión producen proteínas relativamente cortas comparadas con los tamaños alcanzados por genes de baja expresión.

En organismos eucariontes como los mamíferos, muchos genes son interrumpidos por secuencias que no contribuyen a la codificación de las proteínas, estas secuencias se conocen como intrones. Muchos de los intrones se han conservado durante evolución de los genes, por lo que se ha especulado que estas secuencias proveen alguna ventaja. Sin embargo, ninguna ha probado ser una explicación general para el mantenimiento de los intrones y menos aún de su gran tamaño en muchos de los genes. Con un argumento similar al presentado al respecto de la longitud de las proteínas, la existencia de estas secuencias podría tener un efecto negativo para la actividad de los genes, dado que la transcripción de estas secuencias consume energía. El costo asociado a intrones de gran longitud, debe ser mayor para aquellos genes con alto nivel de actividad. Dos estudios (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Versteeg et al. 2003) muestran que los genes de alto nivel de actividad tienen intrones más cortos que los genes de bajo nivel de actividad.

al. 2003) utilizando datos de expresión provenientes de fuentes y metodologías independientes (SAGE y microarrays) mostraron que el tamaño de los intrones en la muestra se encuentra inversamente relacionado con el nivel de actividad de los genes y ningún gene de alta actividad tiene intrones de longitud elevada (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Versteeg et al. 2003).

Los estudios de la longitud génica apoyan la noción de que parece haber una relación entre las características de los genes como el nivel de expresión. Los resultados son consistentes con la hipótesis de que los genes largos imponen un costo en la síntesis de proteínas.

USO DE CODONES SINÓNIMOS

Los genes contienen la información necesaria para establecer la cantidad y el orden de los amino ácidos en las proteínas. Hasta hace unos años era comúnmente aceptado que los sitios de los genes que no contribuyen a la codificación de la cadena de aminoácidos en las proteínas, eran neutros para la selección natural y podían variar libremente de acuerdo con los patrones de mutación. Tal es el caso de los sitios sinónimos. Las proteínas están formadas por 20 aminoácidos diferentes. La traducción del gene a proteína se realiza con base en el código genético, cada conjunto de tres nucleótidos codifica para un amino ácido. Debido a que solo hay 4 nucleótidos que pueden formar 64 tripletes (tres de ellos no codifican ningún amino ácido) y solo hay 20 amino ácidos, varios amino ácidos son codificados por más de un triplete (codones sinónimos). Los sitios sinónimos son aquellos nucleótidos dentro de las regiones que codifican proteínas que pueden variar sin modificar la secuencia de aminoácidos que se produce.

USO DE CODONES EN ESPECIES DE INVERTEBRADOS

Los sitios sinónimos se consideraban candidatos ideales para ejemplo de sitios selectivamente neutros. En caso de que los sitios sinónimos fueran neutros para la evolución natural entonces, podría esperarse que los distintos codones que codifican para un mismo aminoácido estuvieran presentes en las mismas cantidades. Sin embargo, hay una creciente acumulación de evidencia que muestra que, al menos para especies unicelulares e invertebrados como la mosca de la fruta, hay presiones selectivas actuando sobre los sitios sinónimos. En estas especies ciertos codones sinónimos son selectivamente favorecidos por la selección natural. El conjunto particular de codones favorecidos son específicos a cada especie. Los sesgos en la utilización de bases en los sitios sinónimos correlaciona con los niveles de actividad genética, probablemente en respuesta a sesgos en la concentración de RNAs de transferencia que median la traducción a proteína (Kanaya et al. 1999; Moriyama and Powell 1997; Sharp et al. 1995).

OPTIMIZACIÓN DEL USO DE CODONES EN MAMÍFEROS

En mamíferos, debido al menor tamaño de las poblaciones efectivas, se considera que los sitios sinónimos no se encuentran bajo presiones selectivas significativas. Las especies de mamíferos también presentan sesgo en el uso de codones sinónimos (Eyrewalker 1991; Ikemura and Wada 1991). Sin embargo, los sesgos se han atribuido a las variaciones en concentraciones de nucleótidos en el genoma (i.e. isocoros: Bernardi 1995; Bernardi et al. 1997). Estas variaciones son, aparentemente, particulares a los organismos que autorregulan su temperatura. (ver (Hughes et al. 1999). Debido a que ciertas regiones en el genoma son más ricas en guanina y citosina (GC), los genes en estas regiones muestran una mayor preferencia hacia los codones que terminan en G o C. En regiones ricas en adenina y timina (AT), ocurre el caso contrario. De este modo, la composición local de bases induce un sesgo en el uso de codones. Pero, ¿todo el sesgo en la distribución de codones sinónimos es producto de las variaciones en las concentraciones de nucleótidos?

Con el objetivo investigar si el sesgo en el uso de codones (SUC) observado en mamíferos puede explicarse

completamente por el sesgo en el uso de nucleótidos y en caso contrario, investigar qué variables podrían estar implicadas en el SUC residual observado, hemos analizado (Urrutia y Hurst, 2001) el uso de codones alternativos en una muestra de 2396 secuencias de genes humanos. Inicialmente analizamos si el uso de codones observado en los genes humanos podía explicarse por las distribuciones de nucleótidos. Para ello realizamos tres análisis contrastando las preferencias observadas, con las expectativas derivadas de las concentraciones de nucleótidos. Encontramos que 81% de los genes presentaban un sesgo superior al esperado de acuerdo a la distribución de las secuencias aleatorias indicando que los sesgos observados en la mayoría de los genes no pueden explicarse por los niveles basales de nucleótidos. Adicionalmente, encontramos preferencias sistemáticas en los genes por ciertos codones (Urrutia y Hurst, 2001).

Si el uso de codones particulares incrementa la eficiencia en la síntesis de proteínas, entonces se puede esperar que los genes de mayor actividad y por lo tanto requieran mayor eficiencia, tendrán un sesgo mayor en el uso de codones. Se utilizó un método novedoso (Urrutia and Hurst 2001) para medir el sesgo en el uso de codones corrigiendo las variaciones entre proporciones de nucleótidos, en más de 8000 genes humanos. Dicho método implica el cálculo de frecuencias esperadas de codones para cada aminoácido con base en la frecuencia de uso de nucleótidos en los sitios sinónimos de los otros aminoácidos. De este modo, se consigue corregir el impacto de las variaciones locales de la composición de nucleótidos. Los resultados mostraron una relación entre los niveles de uso preferencial de codones con los niveles de expresión, los genes con mayor actividad presentaban mayores sesgos en el uso de codones (Urrutia and Hurst 2003).

Estos resultados sugieren que, a pesar de que las proporciones en las que se usan los codones alternativos están determinadas en gran medida por variaciones en las concentraciones de nucleótidos a lo largo del genoma, las concentraciones de nucleótidos observadas no son suficientes para explicar los sesgos en el uso de codones en los genes humanos. La correlación entre los niveles de actividad génica y el nivel de sesgo en el uso de codones sugieren la existencia de presiones selectivas significativas actuando sobre la elección de codones presumiblemente relacionadas con la eficiencia de la síntesis de proteínas.

CONSIDERACIONES FINALES

En conclusión se ha presentado evidencia de que los genes que codifican proteínas requeridas en grandes cantidades, tienden a un sesgo mayor en el uso de codones, codifican proteínas más cortas y tienen menor contenido de secuencias de intrones. Estas observaciones, en lugar de reflejar la ocurrencia de procesos estocásticos, de acuerdo con las posiciones anteriormente aceptadas, son compatibles con un escenario donde la selección natural actúa para optimizar la síntesis de proteínas, particularmente en los genes altamente expresados. Estos patrones son consistentes con un ahorro de recursos por parte del organismo en la producción de proteínas requeridas en cantidades mayores.

Los estudios anteriormente referidos han permitido el avance en el entendimiento de los factores que determinan las características de los genes y, podrían ser de particular relevancia tanto en el futuro desarrollo de métodos de diagnóstico como tratamientos médicos. En conjunto, estos resultados son de potencial relevancia en la ingeniería genética. Por ejemplo, al introducir genes ajenos para su expresión en células de mamíferos, tomar en cuenta como las características del gene afectarán sus niveles de expresión en la célula huésped. Es conocido que genes acompañados insertados tienen distintos niveles de actividad incluso si van acompañados de secuencias de regulación similares. En el caso del uso de codones, se podría esperar que aquellos genes insertados que contengan codones que son traducidos a proteína más eficientemente tengan una mayor actividad. En efecto, varios grupos de investigación han encontrado que la substitución de los codones menos comunes en el genoma por los codones más utilizados en el genoma se puede incrementar los niveles de expresión de genes introducidos artificialmente en células de mamíferos (Levy et al. 1996; Wells et al. 1999; Zhou et al. 1999; Zolotukhin et al. 1996). Efectos similares son de esperarse en el caso de la longitud de proteínas.

BIBLIOGRAFÍA

- Bernardi, G. 1995. The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* 29: 445-476.
- Bernardi, G., D. Mouchiroud, and C. Gautier. 1997. Isochores and synonymous substitutions in mammalian genes. In *DNA and Protein Sequence Analysis* (eds. M.J. Bishop and C.J. Rawlings). IRL Press, Oxford.
- Castillo-Davis, C.I., S.L. Mekhedov, D.L. Hartl, E.V. Koonin, and F.A. Kondrashov. 2002. Selection for short introns in highly expressed genes. *Nat Genet* 31: 415-418.
- Eyrewalker, A.C. 1991. An Analysis of Codon Usage in Mammals - Selection or Mutation Bias. *J. Mol. Evol.* 33: 442-449.
- Hughes, S., D. Zelus, and D. Mouchiroud. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Mol. Biol. Evol.* 16: 1521-1527.
- Ikemura, T. and K. Wada. 1991. Evident Diversity of Codon Usage Patterns of Human Genes with Respect to Chromosome-Banding Patterns and Chromosome-Numbers - Relation between Nucleotide-Sequence Data and Cytogenetic Data. *Nucleic Acids Res* 19: 4333-4339.
- Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238: 143-155.
- Kimura, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet* 66: 367-386.
- Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
- Levy, J.P., R.R. Muldoon, S. Zolotukhin, and C.J. Link. 1996. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat. Biotechnol.* 14: 610-614.
- Li, W.-H. 1997. *Molecular Evolution*. Sinauer, Sunderland, Mass.
- Moriyama, E.N. and J.R. Powell. 1997. Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* 45: 514-523.
- Sharp, P.M., M. Averof, A.T. Lloyd, G. Matassi, and J.F. Peden. 1995. DNA-Sequence Evolution - the Sounds of Silence. *Philos. Trans. R. Soc. Lond. B* 349: 241-247.
- Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. U.S.A.* 99: 4465-4470.
- Urrutia, A.O. and L.D. Hurst. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159: 1191-1199.
- . 2003. The signature of selection mediated by expression on human genes. *Genome Res* 13: 2260-2264.
- Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler. 1995. Serial Analysis of Gene-Expression. *Science* 270: 484-487.
- Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* 291: 1304-1351.
- Versteeg, R., B.D. Van Schaik, M.F. Van Batenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, and A.H. Van Kampen. 2003. The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Res* 12: 12.

- Wells, K.D., J.A. Foster, K. Moore, V.G. Pursel, and R.J. Wall. 1999. Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch. *Transgenic Res.* 8: 371-381.
- Zhou, J., W.J. Liu, S.W. Peng, X.Y. Sun, and I. Frazer. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J. Virol.* 73: 4972-4982.
- Zolotukhin, S., M. Potter, W.W. Hauswirth, J. Guy, and N. Muzyczka. 1996. A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* 70: 4646-4654.

GLOSARIO

- Genoma humano.** Totalidad del material genético contenido en el núcleo de cada célula de un individuo y que está dividido en 23 pares de cromosomas. La secuencia genómica, así como el número de cromosomas es particular a cada especie. Los miembros de una misma especie también tienen pequeñas diferencias en su material genético, pero son mínimas y contribuyen con las diferencias entre los propios individuos.
- ADN.** Ácido Desoxirribonucleico. Molécula utilizada en las células para el almacenamiento de información genética. El genoma está constituido por ADN.
- ARN.** Ácido Ribonucleico. Molécula utilizada en las células como intermediario en el paso de información de los genes hacia la producción de proteínas. Las moléculas de ARN se producen a partir de la lectura de la información de ADN y posteriormente pueden ser traducidas a proteínas.
- Cromosoma.** Unidades en las que se condensa la información genética de los seres vivos y entre las que el genoma se haya físicamente subdividido. Están conformados por moléculas de ADN y proteínas. Estas últimas son importantes para el mantenimiento de la estructura y la regulación de la actividad de los genes contenidos en cada cromosoma.
- Gene.** Unidad de ADN que contiene la información necesaria para sintetizar una molécula de ARN que cumpla una función en la célula o se traduzca en una proteína.
- Proteína.** Son cadenas de amino ácidos con estructuras tridimensionales particulares que cumplen funciones particulares en la célula. Los genes contienen la información necesaria para determinar el orden de amino ácidos de cada proteína.
- Intron.** Son secciones de la secuencia de un gene que no forman parte del mensaje que finalmente se traduce en la secuencia de amino ácidos en una proteína.
- Codon.** Son tripletes de nucleótidos. Estos tripletes o codones son la unidad de ADN que corresponde a un amino ácido en particular. Dado que el ADN está constituido por 4 diferentes nucleótidos, entonces hay 64 posibles combinaciones de tres nucleótidos o codones.
- Código genético.** Es la correspondencia de cada uno de los 64 codones o tripletes del ADN con los 20 aminoácidos que pueden ser usados en las proteínas.
- Actividad génica.** Se refiere a la lectura de los genes para la generación de moléculas de ARN conteniendo el mensaje para la producción de proteínas. Aquellos genes que son leídos con mayor frecuencia se consideran como más activos y los que son leídos con menor frecuencia son genes de baja actividad. La medición de los niveles de actividad génica se obtiene a través de la cuantificación de moléculas de ARN producto de la lectura de los genes. Existen diversos métodos para este fin, dos de los más utilizados son SAGE y micro arreglos. Estos dos métodos permiten el análisis de expresión de miles de genes a la vez.
- Síntesis de proteínas.** Es el proceso de producción de proteínas que implica la lectura de un gene para producir la molécula intermediaria ARN que a su vez será traducida a una molécula proteica hecha de aminoácidos.
- Selección natural.** Es el proceso mediante el cual caracteres hereditarios (es decir, variantes en la secuencia genómica) que confieren alguna ventaja a los organismos de una especie aumentan en proporción dentro de la población como resultado de esa ventaja. Esto ocurre porque aquellos individuos poseedores de estas características tenderán a dejar mayor descendencia.
- Teoría neutral.** Es la teoría que sostiene que dado que la fuerza de la selección natural es débil comparada con los procesos de azar que afectan las proporciones de los distintos caracteres hereditarios, la mayoría de los cambios que se registran en el ADN a lo largo del tiempo se acumulan por azar. La importancia principal de esta teoría es que permite formular predicciones acerca de las características de los genes si se asume que están determinadas por azar y luego se pueden comparar estas con las observaciones en el

gene o genoma estudiado. Si las observaciones se desvían de las expectativas de azar, entonces se puede suponer que otra fuerza está determinando las características génicas, posiblemente un proceso de selección natural.

Eucarionte. Organismos con características celulares particulares, incluyendo entre otras, la concentración del material genético en cromosomas lineares (no circulares como las bacterias) dentro de un compartimiento nuclear.

Población efectiva. Es el número de individuos cuya descendencia contribuirá a la siguiente generación. Esta fracción de la población es particularmente importante en los estudios de evolución ya que la fuerza de la selección natural varía como función de este valor. Mientras mayor sea la población efectiva, mayor será la presión selectiva que actúe sobre los genes. Esto se debe a que en poblaciones pequeñas es posible que caracteres que no confieren ninguna ventaja y puede llegar a propagarse en toda la población por simple azar.

Amino ácido. Componente básico de las proteínas (sin embargo varios de estos son utilizados en otras funciones celulares como intermediarios metabólicos y la señalización). En los seres vivos, combinaciones de 20 amino ácidos se utilizan para sintetizar toda la variedad de proteínas requeridas en el organismo. Muchos organismos, incluyendo a los seres humanos, son incapaces de sintetizar varios de los 20 amino ácidos y deben adquirirlos mediante el consumo de proteínas sintetizadas por otros organismos. Las proteínas adquiridas de este modo, son degradadas en amino ácidos individuales para servir después en la síntesis de proteínas útiles al organismo.

Nucleótidos. Componentes o eslabones básicos de la cadena de ADN y ARN. Ambas moléculas utilizan cuatro nucleótidos distintos. En el caso del ADN, los nucleótidos utilizados son Adenina, Timina, Guanina y Citosina. En el caso del ARN se utilizan los mismos nucleótidos excepto Timina que se reemplaza por Uracilo.

Sitio sinónimo. Son los sitios dentro de la secuencia de un gene en los que el nucleótido especificado puede sustituirse por otro sin que varíe la cadena de aminoácidos especificada. Esto ocurre porque hay 64 posibles codones y solo 20 aminoácidos por codificar. Esto significa algunos amino ácidos son codificados por más de un codon. Cuando un cambio en el ADN conlleva al cambio de un codon por otro donde ambos codifican el mismo amino ácido, se dice que el sitio es sinónimo.

Traducción. Proceso en el que se produce una proteína a partir de la secuencia de bases en una molécula de ARN. En este proceso los componentes de la proteína (los aminoácidos) se añaden de acuerdo a la secuencia de bases o nucleótidos del RNA. Cada triplete de nucleótidos o codon corresponde a un amino ácido en particular.