**Título: Análisis visual y analítico de los genomas.**

**Introducción**

Actualmente, la biología molecular genera una alta cantidad de datos (ej. secuencias de ADN, ARN o aminoácidos) (Shaw, 2014) gracias al avances de nuevas tecnologías como las técnicas de secuenciación de nueva generación (NGS por sus siglas en inglés) por lo cual se presenta el gran desafío de analizar datos biológicos existentes en bases de datos y datos en el futuro. Sin embargo, estos métodos de análisis en muchas ocasiones no permiten a los usuarios, en este caso científicos, tener la información de forma amigable para su estudio (Marx, 2013). En otras palabras, la representación de los resultados no es simple de analizar porque frecuentemente son archivos de texto que no indican una tendencia clara (ej. tabla tabuladas, matrices de puntaje, entre otras). No obstante en el presente, se han generado avances en la representación de resultados permitiendo a los usuarios identificar patrones en sus datos y este proceso se beneficia de las aptitudes visuales de los seres humanos (Donald & Lindsay, 2013)**.** Por esta razón, es importante generar herramientas que exploten las aptitudes visuales humanas en la interpretación de la información biológica.

En vista que la información biológica es redundante y compleja, es necesario diseñar métodos más efectivos para el análisis biológico teniendo en cuenta los cambios de paradigma actuales sobre el papel del “ADN basura”(Pennisi, 2012)y las capacidades visuales humanas de acuerdo al concepto de alfabetización visual (*visual literacy* en inglés). En el 2012, científicos del proyecto ENCODE proponía que el ADN basura si cumple una funcionalidad pues entendieron que el 80% del genoma humano tiene un propósito bioquímico (Pennisi, 2012). Por otro lado, la alfabetización visual es “la habilidad de decodificar e interpretar mensajes visuales y codificar y componer las comunicaciones visuales significativas” (Metros & Dehoney, 2006). Entonces la dificultad de aclarar el misterio del genoma, especialmente del genoma humano, está en el hecho de que no hay formas efectivas que aprovechen las habilidades humanas para ver, analizar y comprender la información. A causa de esto, la transformación de variables biológicas intrínsecas del genoma como sesgos de k-meros a una representación visual como una imagen ultra HD (imagen con 4 a 16 veces el número pixeles del formato Full HD actual), es una opción con un alto impacto científico. En otras palabras, si se transforma cada 400 nucleótidos del genoma humano en 1 pixel, en una sola imagen podríamos ver, dibujar y colorear todo lo que somos, nuestras secuencias codificantes y no codificantes, repeticiones, centrómeros, telómeros, entre otras, para finalmente tener una fotografía del ADN humano en HD.

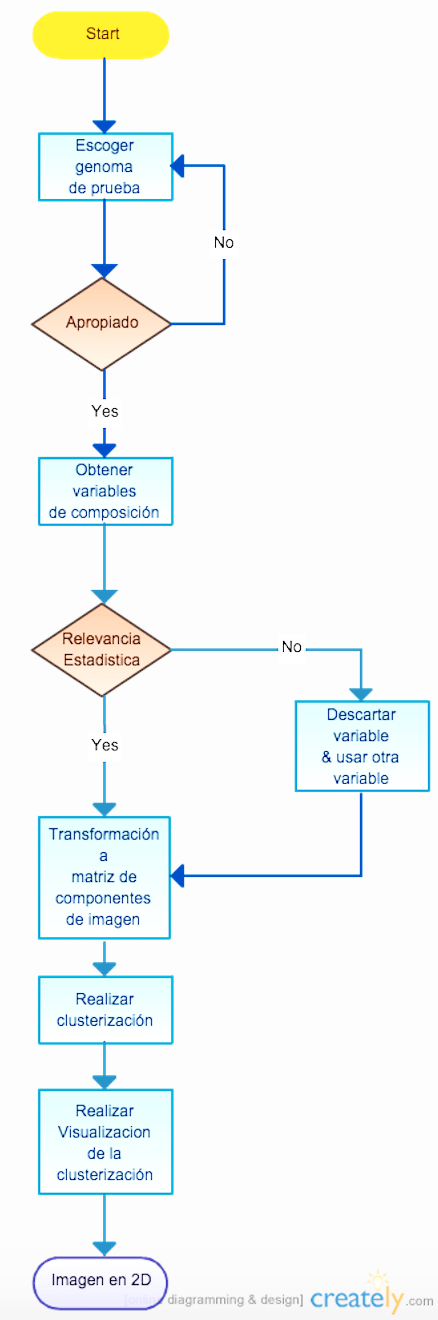
**Objetivo General**

* Generar una función que tome como datos de entrada información basada en variables de composición biológica a lo largo del genoma retornando una imagen que sea específica para el genoma pero permita visualmente identificar características propias del mismo y su similitud con otros genomas relacionados.

**Objetivos específicos**

1. Facilitar la comparación de genomas de acuerdo a las representaciones gráficas basadas en sesgos de k-meros
2. Estudiar la posibilidad de la correlación entre distancias de clústeres con distancias filogenéticas de acuerdo a la organización final de los datos

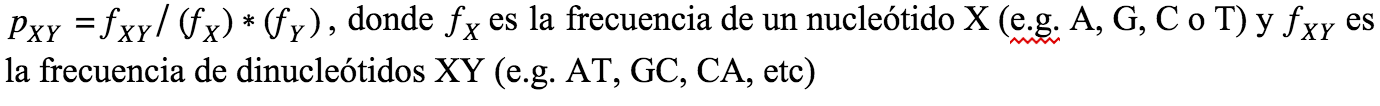
**Metodología**



*Figura 1. Diagrama de Flujo de la Metodología*

*Sesgo*

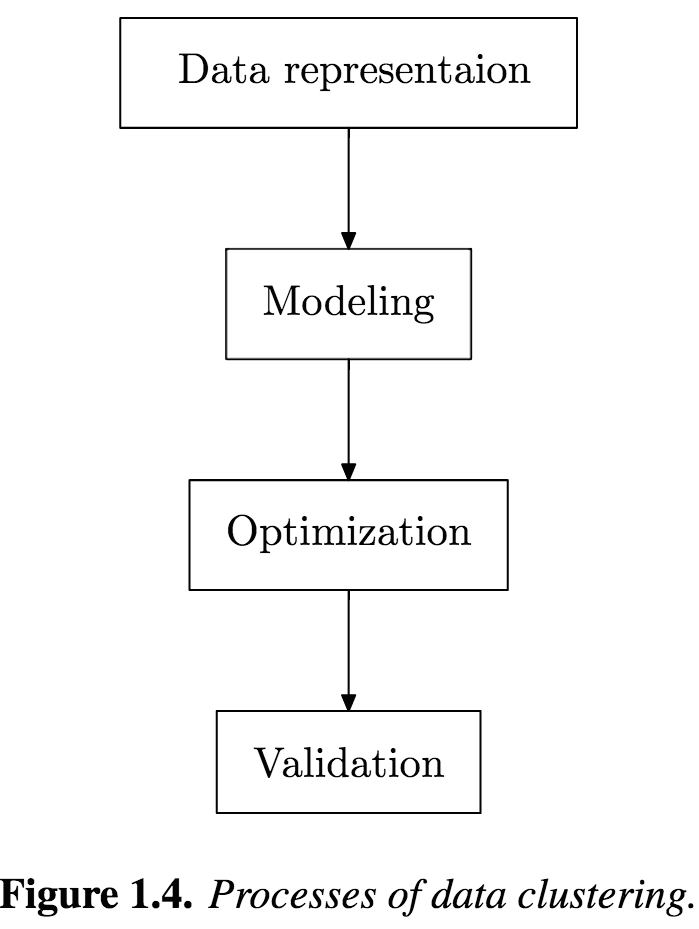
El estudio de la composición de secuencias biológicas (ej. frecuencias de di-nucleótidos, tri-nucleótidos, treta-nucleótidos, etc.) ha sido un tema que se ha estudiado en los últimos veinte años. Exactamente, las frecuencias o conteo de k-meros, (un k-mero se define como el tamaño de una cadena de caracteres), a nivel de genoma o metagenoma son una forma alternativa de clasificar un organismo (Karlin, Mrázek, & Campbell, 1997). Igualmente, el conteo de k-meros no depende de métodos basados en homología para su procesamiento, es decir que no hay limitaciones relacionadas con la base de datos como su tamaño o baja calidad (Mathe, Sagot, Schiex, & Rouze, 2002). Sin embargo, no sólo las frecuencias de k-meros han resultado un predictor alternativo para nuevas secuencias también los sesgos de las frecuencias de estos conteos han permitido un mayor nivel de sensibilidad para encontrar relaciones evolutivas entre organismos (Karlin, Mrázek, & Campbell, 1997). Por ejemplo, un perfil de sesgo de abundancias de di-nucleótidos (k-mero = 2) puede ser determinado por la siguiente ecuación según Karlin y Mrazek (1997):



Además, otras ventajas de usar frecuencias y sesgos de k-meros para la clasificación de organismos son usar todo el genoma en contraste a un solo gen (Karlin, Mrázek, & Campbell, 1997) y obtener información más precisa de zonas de baja complejidad. Cuando se quiere hacer comparaciones entre genomas ya sea entre el mismo dominio o no, estas se pueden beneficiar de un análisis composicional. Por ejemplo, al hacer comparaciones intra- e inter-especies entre procariotas, no es claro si sus diferencias se deban solamente al contenido de GC de estas zonas. Sin embargo, se puede realizar perfiles de uso de tetra-nucleótidos (k-mero = 4) desde expectativas (TUD por sus siglas en inglés) para describir las relaciones de estos perfiles con el contenido de GC (Pride, Meinersmann, Wassenaar, & Blaser, 2003). Un perfil TUD es esencialmente la distribución del uso de tetra-nucleótidos en el genoma y se puede obtener usando un método de Markov de orden cero o un método de cadenas de Markov (Pride, Meinersmann, Wassenaar, & Blaser, 2003). De acuerdo a Pride et al, las diferencias de los perfiles TUD entre organismos como *C. jejuni* y *H. pylori* no depende de su similitud en contenido de GC, la cual es relativamente alta, por lo cual puede haber una alta correlación entre sus perfiles TUD. Así, la conservación del uso de nucleótidos es relativamente independiente del cambio en el contenido de GC (Pride, Meinersmann, Wassenaar, & Blaser, 2003). De esta manera, el uso a la información que un genoma puede proveer es más eficiente debido que hay un mayor nivel de sensibilidad usando sesgos de k-meros o en este caso perfiles TUD.

*Clusterización*

Por otro lado, dado que los sesgos de k-meros se han definido como firmas, es relevante realizar metodologías para representar gráficamente estas firmas pues esto permitiría estimar patrones de una forma más sencilla. Uno de los métodos más usados para la representación visual de sesgos de k-meros es la clusterización, definida como “el proceso de generar grupos de objetos en una forma tal que los objetos en un clúster son similares y los objetos en diferentes clústeres son significativamente diferentes” (Gan, Ma, & Wu, 2007, pp. 8-12). La aplicación de análisis de clústeres en Biología ha sido un proceso en constante cambio debido al avance de las tecnologías como expresión génica (Gan, Ma, & Wu, 2007, pp. 323-324). Sin embargo, es posible adaptar las firmas de sesgos de k-meros, específicamente sus distribuciones, a las principales etapas de una clusterización que son representación de datos, modelamiento, optimización y validación (Figura 2).

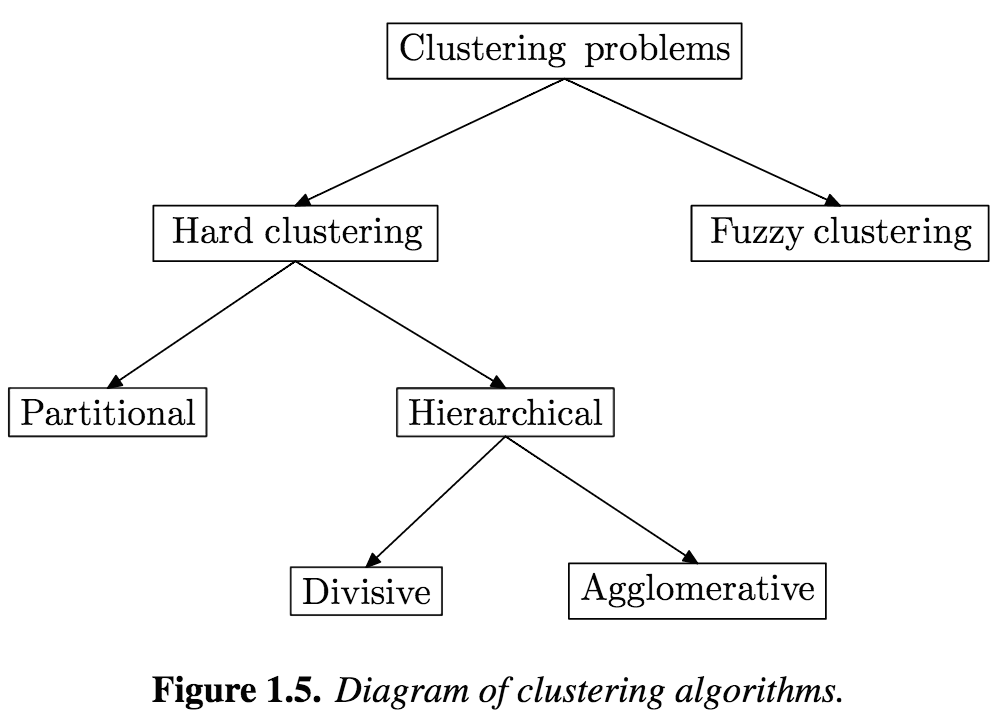


*Figura 2. Obetnido de* (Gan, Ma, & Wu, 2007)

*Pre-procesamiento*

De igual manera, es valioso mencionar que los datos biológicos brutos (ej. microarreglos de cDNA, microarreglos de oligonucleótidos, sesgos de k-meros, etc.) tienden a ser muy ruidosos por lo cual es necesario ejecutar un pre-procesamiento a los datos. Este pre-procesamiento se puede entender como una transformación necesaria del set de datos para disminuir su complejidad y así facilitar la clusterización (Gan, Ma, & Wu, 2007, pp. 43-47). Los métodos de transformación más comunes son análisis de componentes principales y descomposición en valores singulares (PCA y SVD por sus siglas en inglés respectivamente) pero no son los únicos. En pocas palabras, un PCA consiste en entender las observaciones de un set de datos por variables dependientes que están correlacionadas (Abdi & Williams, 2010) mientrasun SVD es una técnica de factorización de una matriz para acoplarse a un espacio deseado y es muy usada en el perfil completo de expresión de un genoma (Alter, Brown, & Botstein, 2000).

Luego de hacer el pre-procesamiento con algunas de las técnicas anteriores u otras, se puede hacer la clusterización entablando los parámetros más adecuados. Sin embargo, en este escrito no se va a exponer todos los posibles parámetros dado que estos dependen del algoritmo de clusterización a implementar. Los algoritmos de clusterización se pueden clasificar de acuerdo al problema a resolver, es decir en *hard clustering* o *fuzzy clustering*. *Hard clustering* es “cuando un punto del set de datos solo pertenece a un solo clúster o agrupación” (Gan, Ma, & Wu, 2007, pp. 8-9) mientras *fuzzy clustering* es lo contrario y “un punto del set de datos puede pertenecer a más de un clúster” (Gan, Ma, & Wu, 2007, pp. 8-9). Por consiguiente, es probable que se ejecute un algoritmo tipo *fuzzy* en este proyecto de grado porque los resultados tendrían mayor sentido biológico (Laczny, Pinel, Vlassis, & Wilmes, 2014). Asimismo, estos algoritmos pueden tener subdivisiones como se indica en la Figura 3.



*Figura 3. Obtenido de* (Gan, Ma, & Wu, 2007)

*Visualización*

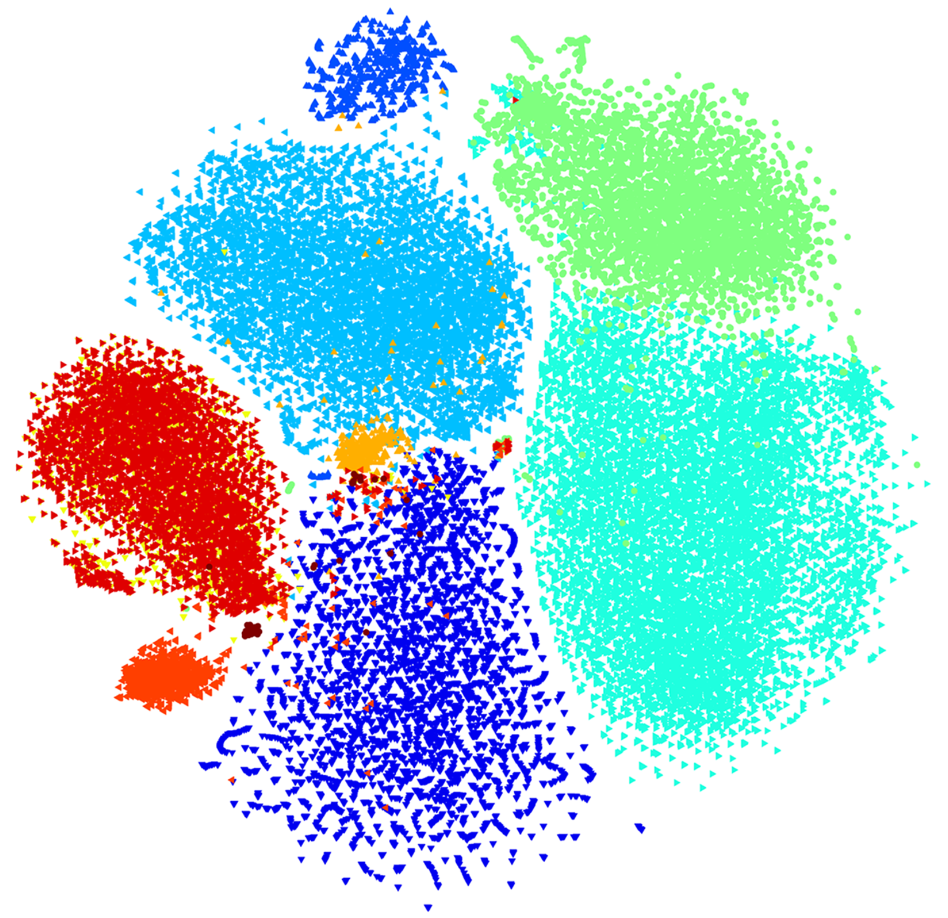
Además cuando se ha cumplido con los pasos de representación de datos, modelamiento, optimización y validación, se debe hacer la visualización del *clustering* para permitir el análisis cualitativo. Como es de esperarse, hay varios métodos para la visualización de clústeres; no obstante, recientemente se ha propuesto usar un método basado en reducción de dimensión no lineal (*nonlinear dimension reduction* en inglés). Según Laczny et al, es posible visualizar datos metagenómicos usando *nonlinear dimension reduction* y un algoritmo auxiliar denominado en *inglés Barnes-Hut Stochastic Neighbor Embedding* de firmas de k-meros(Laczny, Pinel, Vlassis, & Wilmes, 2014)**.** Específicamente, Laczny et al usaron la relación logarítmica del sesgo de oligonucleótidos, donde los k-mero = 4 y k-mero = 5 ofrecieron los mejores resultados (Laczny, Pinel, Vlassis, & Wilmes, 2014). En consecuencia, la visualización de los clústeres por *nonlinear dimension reduction* tampoco se apoya en la implementación de procedimientos basados en homología por lo cual esta metodología representa un gran potencial.

**Cronograma de actividades**

|  |  |
| --- | --- |
| **Semanas** | **Actividad a desarrollar** |
| 1 (Julio 27 – Julio 1) | -Escoger el genoma prueba y obtener diferentes variables de la composición (e.g. conteo de diferentes k-mers, sesgos de diferentes k-mers, perfiles TUD, etc)  -Determinar la arquitectura del script en python para las responsabilidades de cada función dentro de la función general. |
| 2 (Agosto 3 – Agosto 8) | -Escoger el genoma prueba y obtener diferentes variables de la composición (e.g. conteo de diferentes k-mers, sesgos de diferentes k-mers, perfiles TUD, etc) |
| 3 (Agosto 10 - Agosto 15) | -Realizar pruebas estadísticas (e.g. PCA) para determinar la relevancia de cada variable de la composición |
| 4 (Agosto 17 – Agosto 22) | -Realizar pruebas estadísticas (e.g. PCA) para determinar la relevancia de cada variable de la composición |
| 5 (Agosto 24 – Agosto 29) | -Programar (puede ser en Python o Perl), la función que permita la entrada de las variables de composición biológica |
| 6 (Agosto 31 – Septiembre 5) | -Programar (puede ser en Python o Perl), la función que permita la entrada de las variables de composición biológica |
| 7 (Septiembre 7 – Septiembre 12) | -Programar (puede ser en Python o Perl) o implementar herramientas para la transformación de las variables de composición del genoma en la matriz de componentes de imagenes. |
| 8 (Septiembre 14 – Septiembre 19) | -Programar (puede ser en Python o Perl) o implementar herramientas para la transformación de las variables de composición del genoma en la matriz de componentes de imagenes. |
| 9 (Septiembre 21 – Septiembre 26) | -Programar (puede ser en Python o Perl) o implementar herramientas para la transformación de las variables de composición del genoma en la matriz de componentes de imagenes. |
| 10 (Septiembre 28 – Octubre 3) | -Analizar las condiciones de los datos de entrada del *cluster analysis* (i.e. implementar *data transformation* si es necesario) |
| 11 (Octubre 5 - Octubre 10) | -Implementar (puede ser funciones establecidas en Python, C++ o MATLAB) para la etapas del *cluster analysis* en la función final. |
| 12 (Octubre 12 – Octubre 17) | -Implementar (puede ser funciones establecidas en Python, C++ o MATLAB) para la etapas del *cluster analysis* en la función final. |
| 13 (Octubre 19 - Octubre 24) | -Implementar (puede ser funciones establecidas en Python, C++ o MATLAB) para la etapas del *cluster analysis* en la función final.  -Empezar el análisis general de los resultados para el escrito final del proyecto de grado |
| 14 (Octubre 26 - Octubre 31) | -Implementar (puede ser funciones establecidas en Python, C++ o MATLAB) para la etapas del *cluster analysis* en la función final.  -Empezar el análisis general de los resultados para el escrito final del proyecto de grado |
| 15 (Noviembre 2 - Noviembre 7) | Análisis de resultados |
| 16 (Noviembre 9- Noviembre 14) | Ajustar Presentación y Escrito Final |

**Resultados esperados**

Los resultados esperados de este proyecto son imágenes en 2 dimensiones de al menos un genoma. Las imagen se obtendrán luego de aplicar algoritmos para la representación gráfica de clústeres como en el articulo *Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction* de Laczny et al (Figura 4). Los colores y las posiciones de los pixeles se determinarán según características genómicas específicas a lo largo del genoma en estudio. Esta función se podría aplicar a cualquier genoma completo y la comparación de las imágenes resultantes representarán las variaciones genómicas más características entre los genomas usados.



*Figura 4. Obtenido de* (Laczny, Pinel, Vlassis, & Wilmes, 2014)

**Presupuesto**

Debido que este proyecto no necesita materiales adicionales, el único costo que tiene es el del uso de la infraestructura computacional de la Universidad de los Andes. Esta infraestructura en la actualidad es de uso gratuito para los estudiantes de pregrado y posgrado de la Universidad. Específicamente, la infraestructura consiste en clusters, los cuales se denominan “cluster gate”, “hellsgate” y “tado”.

# **Referencias**

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wires: Computational Statistics* , 433.

Alter, O., Brown, P. O., & Botstein, D. (29 de Agosto de 2000). Singular value decomposition for genome-wideexpression data processing and modeling. *PNAS* , 10101–10106.

Donald, A. N., & Lindsay, P. H. (2013). *Human Information Processing: An Introduction to Psychology.* Academic Press.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications.* American Statistical Association and the Society for Industrial and Applied Mathematics.

Karlin, S., Mrázek, J., & Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology* *, 179* (12), 3899-3913.

Laczny, C. C., Pinel, N., Vlassis, N., & Wilmes, P. (31 de Marzo de 2014). Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Nature: Scientific Reports* , 1-12.

Marx, V. (2013). Biology: The big challenges of big data. *Nature: Technology Feature* , 255-260.

Mathe, C., Sagot, M.-F., Schiex, T., & Rouze, P. (2002). Current methods of gene prediction, their strengthsand weaknesses. *Nucleic Acids Research* *, 30* (19), 4103-4117.

Metros, S., & Dehoney, J. (2006). New Media Design Rubric. *The Educator’s Role* , 108.

Pennisi, E. (7 de Septiembre de 2012). ENCODE Project Writes Eulogy For Junk DNA. *Science: News & Analysis* , 1159-1161.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M., & Blaser, M. J. (2003). Evolutionary Implications of Microbial GenomeTetranucleotide Frequency Biases. *GenomeResearch* , 145-158.

Shaw, J. (Abril de 2014). Why “Big Data” Is a Big Deal. *Harvard Magazine* , pp. 30-35.