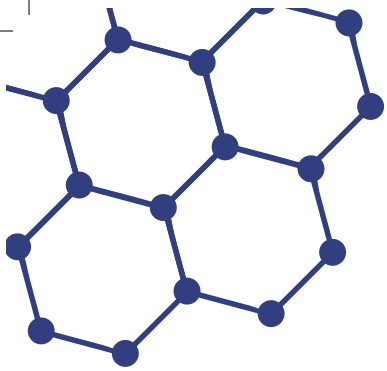
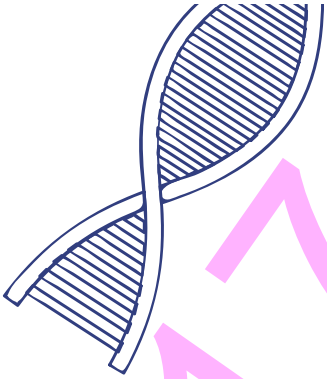
Elementos de **Bioinformática y Genómica Computacional**

para Ingenieros de Sistemas

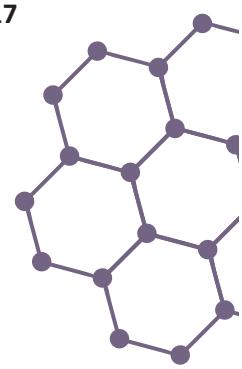
Lic. Carlos Noceda-Alonso, PhD.
Lic. Jesennia Cárdenas-Cobo, MBA.
Ing. Mirella Correa-Peralta, MBA.
Ing. Oscar León-Granizo.
Ing. Rafael Lazo-Sulca, MGTI.

UNIVERSIDAD ESTATAL DE MILAGRO
UNEMI
Evolución Académica





Título de la Obra: Elementos de Bioinformática y Genómica Computacional para Ingenieros De Sistemas
Derecho de Propiedad Intelectual: GYE-008404
Depósito Legal: GYE-000256
ISBN: 978-9942-969-73-6



Autores:
Lic. Carlos Noceda-Alonso, PhD.
Lic. Jesennia Cárdenas-Cobo, MBA.
Ing. Mirella Correa-Peralta, MBA.
Ing. Oscar León-Granizo.
Ing. Rafael Lazo-Sulca, MGTI.

Rector: Ing. Fabricio Guevara-Viejó, PhD.
Director del Proyecto: Ing. Richard Ramirez-Anormaliza, PhD.
Coordinadora del Proyecto: Ing. Mayra D'Armas-Regnault, PhD.
Revisores Pares: MSc. Jorge Vera MSc. Ailet Ávila

Ediciones Holguín S.A., equipo editorial:
Directora Editorial: Lic. Lucrecia Resabala Manosalvas, MSc.
Editor de Área: MSc. Bolívar Duchi.
Coordinador Editorial: Ing. Danilo Holguín Cabezas, MBA.
Asistente Editorial: Ing. Johanna Coronel Vélez
Revisora de Ortografía y Estilo: Lic. Pilar Huayamave Navarrete, MSc.
Diseño Gráfico y Diagramación: Lic. Vanessa Landin

Universidad Estatal de Milagro – UNEMI
Cdla. Universitaria Km. 1.5 vía Milagro Km. 26
Teléfonos: (593) 04 2715081- 04 2715079
http://www.unemi.edu.ec/
Milagro – Ecuador

DEDICATORIA

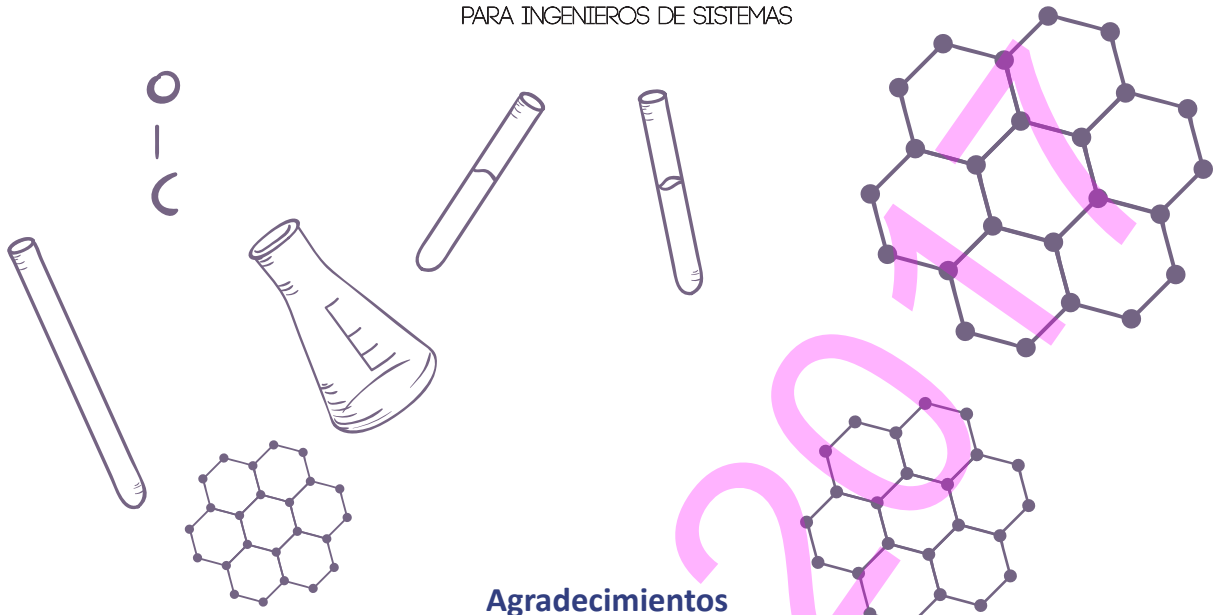
Carlos Noceda:
A las tres personas que más quiero en este mundo: mi madre, mi esposa y mi hijo.

Jesennia Cárdenas:
A mis padres, mi abuelita, mi esposo y a mis tesoros: mis hijos.

Mirella Correa Peralta:
Al regalo maravilloso que me da Dios, mi familia.

Óscar Darío León Granizo:
A mi familia, fortaleza de mi vida.

Rafael Lazo Sulca:
A mi madre, que me inculcó el amor y temor a Dios.



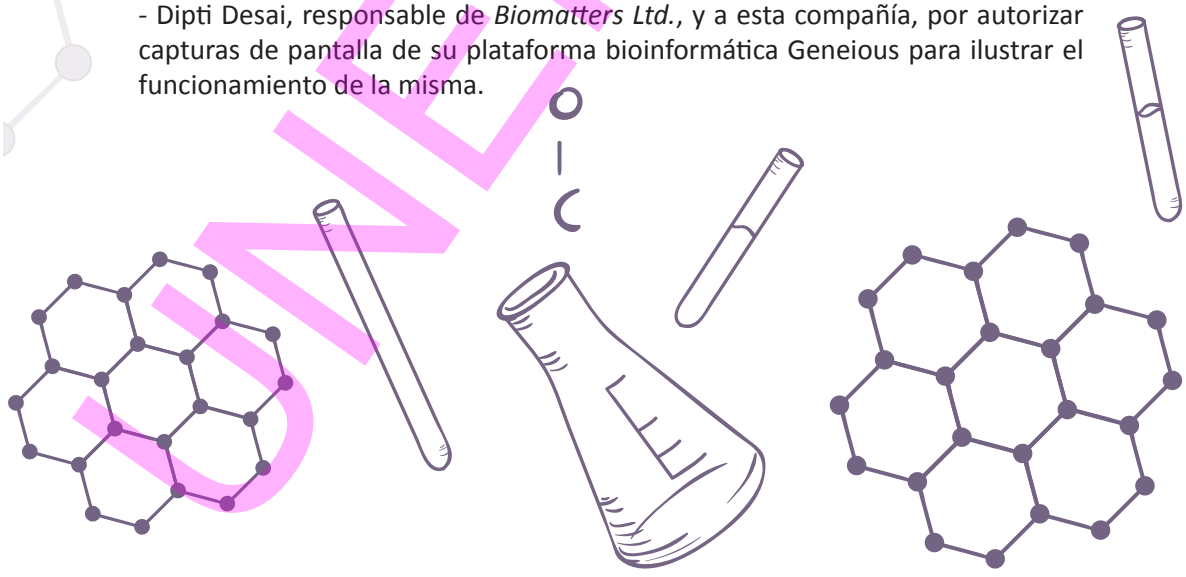
Agradecimientos

Queremos mostrar nuestro agradecimiento a:

- La *Nationmal Library of Medicine* (Estados Unidos), ya que de los recursos de su rama National Center for Biotechnology Information (NCBI) se han efectuado varias capturas de pantalla útiles en la exposición.

- Galaxy, TAIR, VectorBase, UniProt, EcoCyc y sitios relacionados, por las capturas de pantalla necesarias para las explicaciones sobre sus sitios.

- Dipti Desai, responsable de *Biomatters Ltd.*, y a esta compañía, por autorizar capturas de pantalla de su plataforma bioinformática Geneious para ilustrar el funcionamiento de la misma.



AUTORES

Carlos Noceda es profesor titular de Biología Vegetal en la Universidad de las Fuerzas Armadas-ESPE, investigador en los grupos de Biotecnología Celular y Molecular de Plantas y de Biotecnología Industrial y Bioproductos de dicha Universidad, y docente e investigador en la Facultad de Ciencias de la Ingeniería y el Departamento de Investigación de la Universidad Estatal de Milagro (Ecuador).

Jesennia Cárdenas Cobo es docente titular de Ingeniería de Software en la Facultad Ciencias de la Ingeniería de la Universidad Estatal de Milagro (Ecuador).

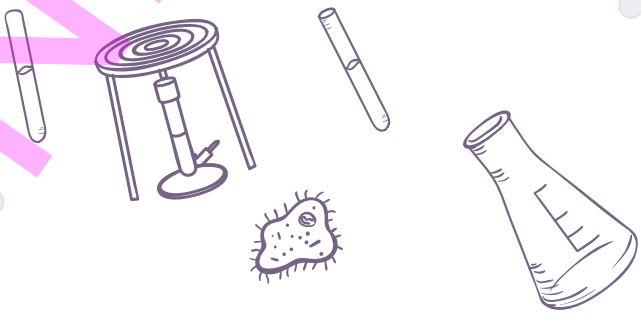
Mirella Correa Peralta es docente titular de Gestión de Proyectos y de Tecnología de la Facultad de Ciencias de la Ingeniería de la Universidad Estatal de Milagro (Ecuador).

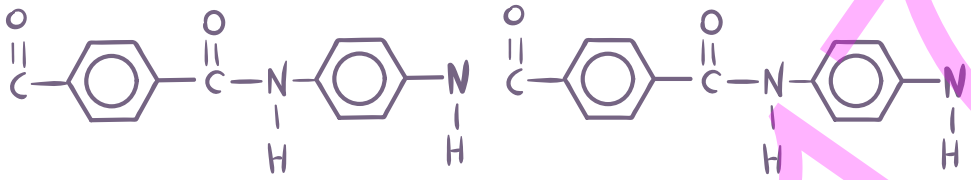
Óscar Darío León Granizo es docente contratado de Introducción al Algoritmo y Programación de la sección de Admisión y Nivelación de la Universidad Estatal de Milagro (Ecuador).

Rafael Lazo es docente titular de Bioinformática y de Ecología y Medio Ambiente de la Facultad de Ciencias de la Ingeniería de la Universidad Estatal de Milagro (Ecuador).

ÍNDICE

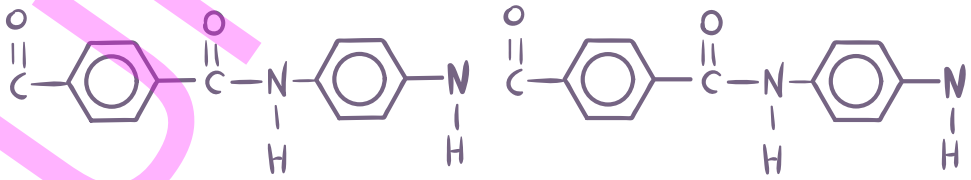
ÍNDICE DE FIGURAS





PRÓLOGO

Tanto la Biología de Sistemas como la Informática se encuentran en una fase de desarrollo sin precedentes cuyas expectativas a futuro son tan prometedoras como impredecibles. En la intersección de ambos campos de conocimiento se encuentra la Biología Computacional, área multidisciplinar que se apoya en la Informática para dar respuesta a complejos problemas biológicos. Desde esta perspectiva, la Biología Computacional se convierte en una especialidad idónea para la aplicación de la Ingeniería de Sistemas. En este libro se abordan los fundamentos biológicos necesarios para que el ingeniero de sistemas pueda tanto comprender apasionantes preguntas biológicas como enfrentar problemas bioinformáticos básicos, entendiendo la Bioinformática como el conjunto de tecnologías que se ocupan del procesamiento y análisis de datos biológicos. Por simplicidad, la estructura de este texto de iniciación se ha ceñido a la Genómica Computacional, aunque las estrategias expuestas son ampliamente extrapolables al uso de otras herramientas ´ómicas`, en especial a la Transcriptómica y la Proteómica. Así, el planteamiento se basa en una breve y necesaria introducción a la Biología Molecular, que se desarrolla a medida que se van describiendo los distintos aspectos bioinformáticos y problemas biológicos relacionados. Tanto el lector autodidacta como el estudiante que desee utilizar este libro como texto de partida, apoyados en una plataforma bioinformática mínima, habrán adquirido finalmente una esencial capacitación para, a un nivel inicial, responder a preguntas bioinformáticas y crear algoritmos afines, así como para lanzarse por sí mismos a un aprendizaje autónomo sobre aspectos biocomputacionales más complejos.



CAPÍTULO 1:

INTRODUCCIÓN A LA BIOLOGÍA MOLECULAR

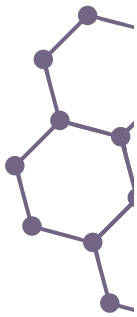
Lic. Carlos Noceda-Alonso, PhD.
Lic. Jesennia Cárdenas-Cobo, MBA.
Ing. Mirella Correa-Peralta, MBA.

1. INTRODUCCIÓN A LA BIOLOGÍA MOLECULAR

1.1. LOS FLUJOS DE LA INFORMACIÓN CONTENIDA EN EL GENOMA

Los seres vivos (**organismos**) funcionan fundamentalmente siguiendo instrucciones contenidas en su **genoma**. Este consiste en un conjunto de moléculas con unas características químicas definidas. La más importante de las cuales es la capacidad de almacenamiento de información. El genoma, en la mayoría de los organismos, está contenido en las unidades básicas estructurales y funcionales de estos, es decir, en las denominadas **células**. El genoma es prácticamente idéntico en casi todas las células de un organismo individual. Y, cuanto más se parezcan los organismos entre sí, más parecido será su genoma y, por tanto, la información contenida en él.

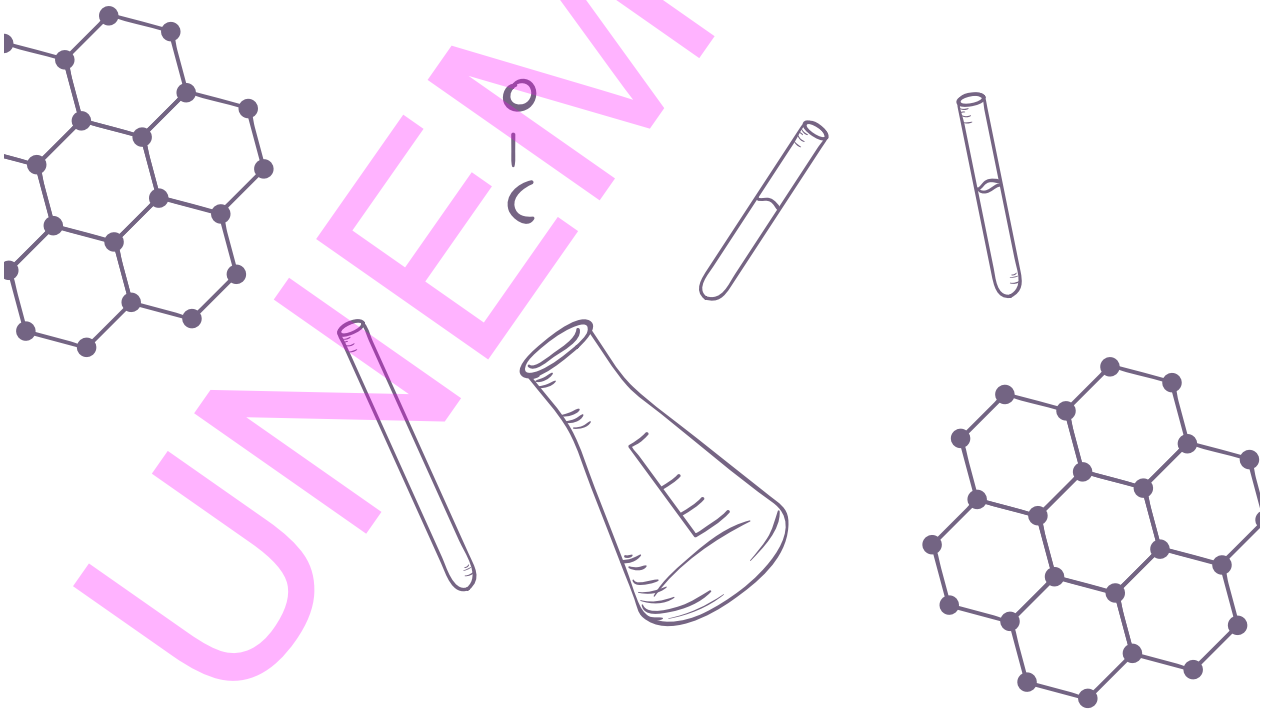
Haciendo un pequeño esfuerzo de abstracción, podríamos considerar que las células son unas pequeñas *computadoras* que reciben señales de su entorno (*inputs*), y que regulan sus funciones (*respuestas*) a partir de la información contenida en su genoma. Este vendría a ser el *disco duro* en el que se encuentran todos los programas de la microcomputadora denominada célula. Al igual que en las computadoras, en distintas células, en distintos tiempos, se está procesando un conjunto de programas diferentes. Las instrucciones de los programas contenidos en el genoma se ejecutan mediante *copiado* de la información de esas instrucciones genómicas a otras moléculas que se llaman **transcritos.** El conjunto de transcritos en un momento dado en una célula equivaldría por tanto a las instrucciones cargadas en la denominada "memoria *RAM*" de la célula. Además, la intensidad de lectura de cada instrucción del genoma influirá en el número de transcritos (copias) de esa instrucción. Por último, la información finalmente contenida en los transcritos puede *traducirse* para generar lo que denominaremos el *output* de todo este proceso, es decir, **polipéptidos**. Cada transcrito finalmente obtenido porta la información para construir un polipéptido. Los polipéptidos son un tipo simple de proteínas. Las proteínas son un grupo de moléculas con amplia variedad de funciones en la célula, y determinan el comportamiento biológico de la misma. Dentro de la célula, las proteínas más importantes tienen función estructural (construyendo la célula) o enzimática (permitiendo que las reacciones bioquímicas ocurran a la velocidad adecuada para la vida).

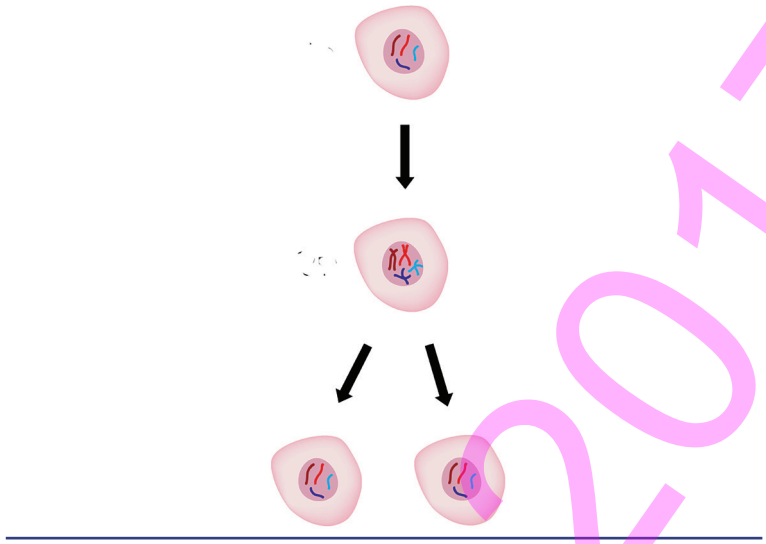


Este panorama simplificado es más complejo en la realidad. Así, existen mecanismos moleculares para que se lean, en un momento dado del desarrollo del organismo, en una zona del mismo, unas instrucciones específicas y no

otras, y con la intensidad adecuada para que las proteínas resultantes finales se encuentren en la cantidad precisa. Por otra parte, hay instrucciones o partes del genoma cuyo fin no es la generación de una proteína, sino la regulación de la lectura de otras instrucciones, o la interferencia con transcritos que contienen copia de otras instrucciones. A su vez, los polipéptidos formados pueden sufrir modificaciones posteriores (**post-traduccionales**) en función de otros programas del genoma que se estén ejecutando, generando así proteínas de características y funciones muy concretas.

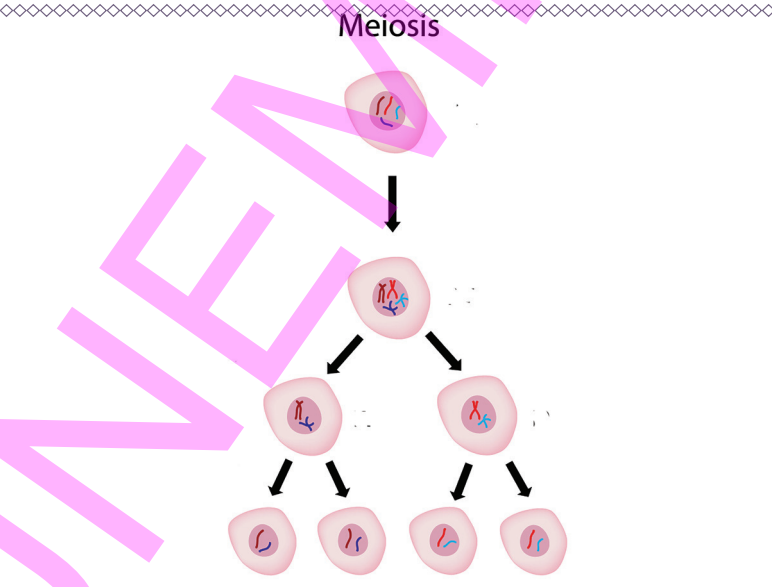
Otra peculiaridad de la *microcomputadora* célula es que en su disco duro (genoma) lleva las instrucciones para *autorreplicarse*, mediante la gestión de la materia y la energía del entorno. Gracias a esta autorreplicación, que incluye al genoma (que heredaría por tanto cada célula hija), un organismo se desarrolla y se mantiene. Además, el organismo entero se puede reproducir, o bien a partir de un conjunto de células que generan un organismo nuevo y genómicamente idéntico al original (**reproducción asexual**), o bien a partir de una célula (**gameto**) a la que se transmite sólo la mitad de la información contenida en el genoma, para que combine dicha información con el gameto de otro individuo (**reproducción sexual**). De esta manera, la información genómica se **hereda** entre generaciones de individuos.





Adaptado de shutterstock.com[1]

Figura 1. Reproducción asexual de una célula

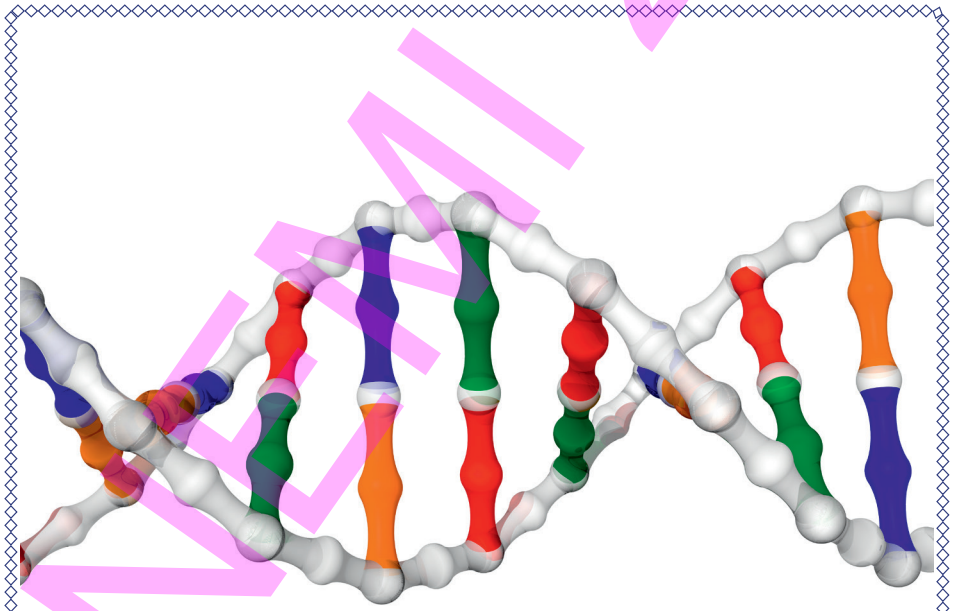


Adaptado de shutterstock.com[2]

Figura 2. Formación de cuatro gametos a partir de una célula germinal

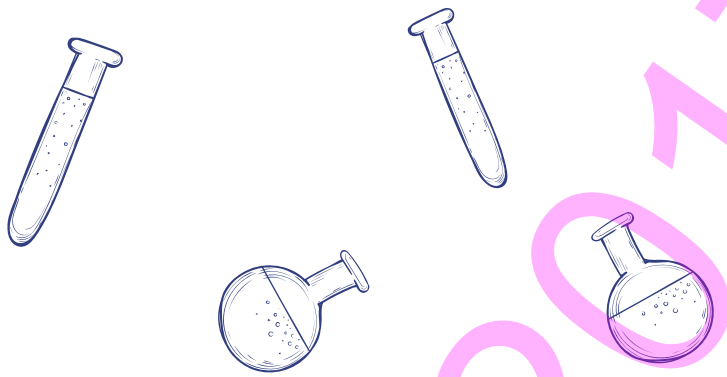
1.2. ÁCIDOS NUCLEICOS Y PROTEÍNAS

Químicamente, el genoma consiste en varias moléculas de ácido desoxirribonucleico (**ADN**). Cada molécula de ADN está formada por dos hebras moleculares que se enrollan helicoidalmente entre sí. Las **hebras**, y por lo tanto la molécula de ADN, pueden alcanzar longitudes enormes a escala molecular. Cada hebra posee un esqueleto molecular de fosfatos y de azúcares (desoxirribosas), dispuestos alternadamente, con dos extremos, denominados 5´ (donde se sitúa "el primer" fosfato) y 3´ (donde se sitúa "el último" azúcar). Las hebras de una misma molécula de ADN se disponen antiparalelamente, es decir, en el extremo de la molécula en que está el extremo 5´de una hebra, está el 3` de la otra, y viceversa.

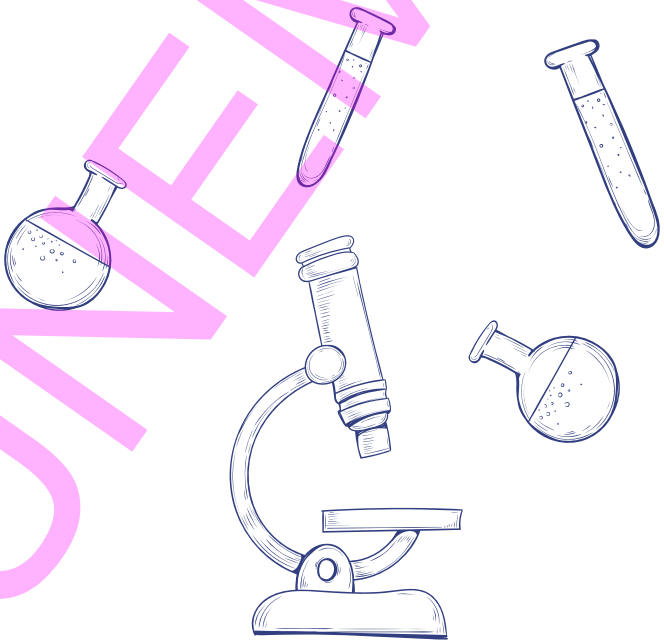


Fuente: shutterstock.com[3]

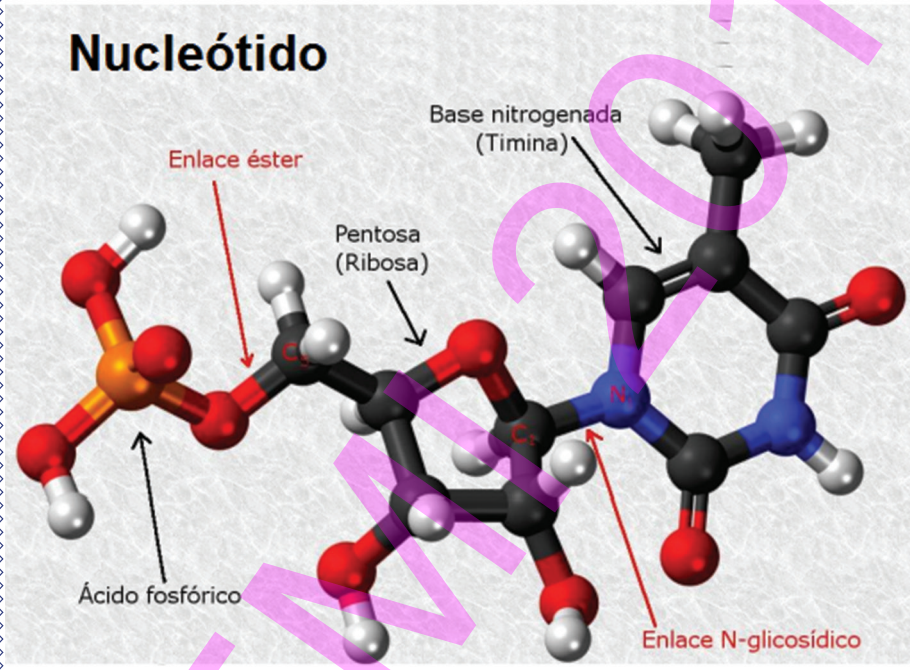
Figura 3. Modelo de doble hélice de la molécula de ADN



A lo largo del esqueleto molecular de cada hebra del ADN, que constituiría el soporte físico de la información que lleva el genoma, están enlazadas las moléculas "informativas", es decir, las **bases nitrogenadas**, que son cuatro moléculas distintas y pueden disponerse linealmente en cualquier orden a lo largo de la longitud de la secuencia azúcar-fosfato. Las bases nitrogenadas se unen al esqueleto molecular de cada hebra a través de un enlace con el azúcar (formando un **nucleósido**).



El conjunto base nitrogenada-azúcar-fosfato, es decir, base nitrogenada-nucleósido, se denomina **nucleótido**, y por ende cada hebra es una yuxtaposición de nucleótidos. Por tanto, si hay cuatro bases nitrogenadas en el ADN, hay cuatro tipos de nucleósidos y cuatro tipos de nucleótidos.

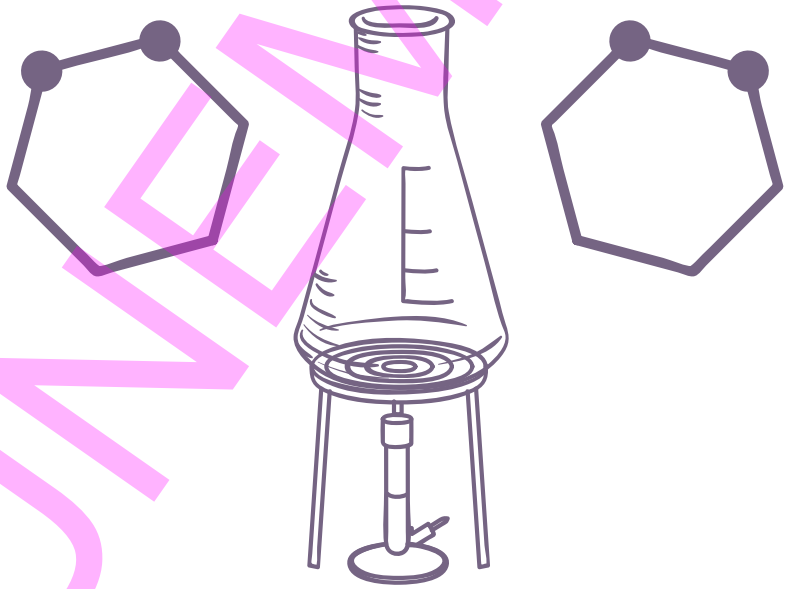


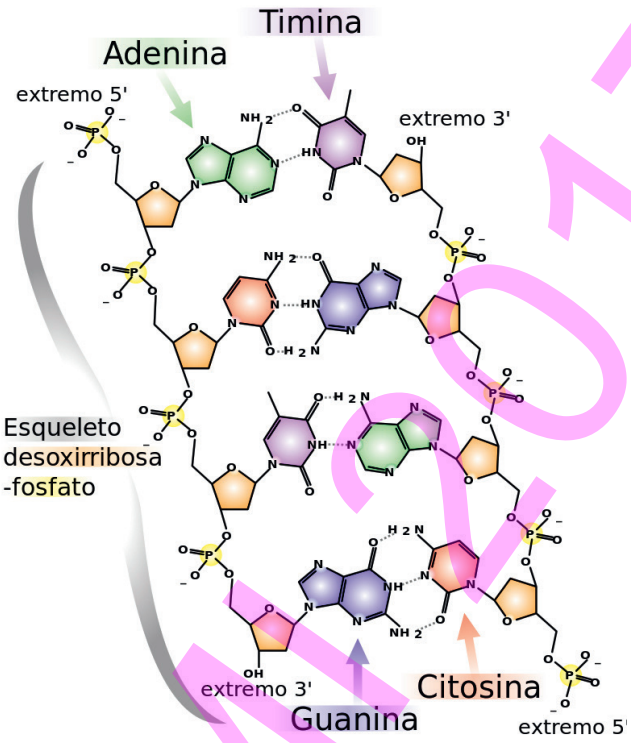
Fuente: bionova.org.es[4]

Figura 4. Estructura química de un nucleótido

La disposición longitudinal de las cuatro bases en las hebras de ADN confiere al genoma un código de cuatro elementos que, para su **expresión** final en proteínas, se leen de tres en tres. Es decir, cada **triplete** de nucleótidos es una unidad de información. Puesto que hay cuatro bases distintas (y por tanto cuatro nucleótidos distintos), el número de tripletes diferentes es 64, de acuerdo a la fórmula de variaciones con repetición de cuatro elementos tomados de tres en tres. A su vez, distintas posibles variaciones con repetición de estos 64 tripletes constituyen instrucciones para **biosintetizar** diferentes polipéptidos. Cuando una secuencia de estos tripletes de una hebra de ADN determina un polipéptido por intermedio de su copiado a transcrito, esos tripletes se llaman **codones**, tanto en el ADN como en el trascrito resultante.

Las cuatro bases nitrogenadas que intervienen en la codificación y estructura del ADN se denominan adenina, timina, guanina y citosina, denotadas típicamente como A, T, G y C, respectivamente. Las dos hebras de la molécula de ADN se aparean a través de estas bases, de manera que siempre la A de una hebra se aparea con una T de la otra, y de igual manera G con C, de modo que la secuencia de bases (o de nucleótidos) de una hebra determina la secuencia de la otra de manera complementaria. Es decir, las dos hebras antiparalelas son complementarias entre sí, aunque esta complementariedad, cuando se ordenan artificialmente (por ejemplo, *in silico*) en ambas hebras los nucleótidos en sentido 5´a 3´, es de naturaleza reversa. Así, dos hebras antiparalelas son complementarias, pero al ordenar ambas en un sentido químico 5´-3´, son **complementarias reversas** (o reversas complementarias) entre sí. (Ver Sección 3.3.1. Secuencias reversa, complementaria y reversa complementaria).

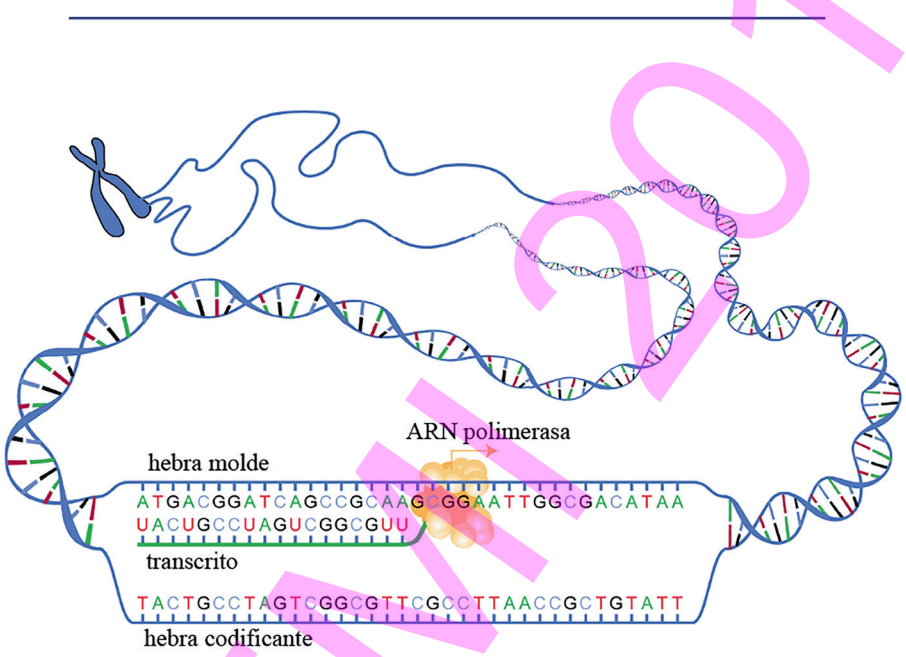




Fuente: wikimedia.org[5]

***Figura** 5. Representación bidimensional de la disposición de las hebras de ADN, en la que se observan las estructuras moleculares explicadas en el texto, así como la complementariedad de los nucleótidos en función de sus bases nitrogenadas*

De esta manera, en un fragmento de la molécula de ADN, una hebra será la denominada **codificante.** La hebra antiparalela a esta, que simplemente es su "espejo", su complementaria, se denomina para esta región **hebra molde** (que porta **anticodones**, tripletes complementarios a los codones), pues es sobre la que se construye el transcrito mediante la yuxtaposición consecutiva de nucleótidos con bases nitrogenadas complementarias a esa hebra molde. El transcrito se va así desarrollando en sentido 3´-5´ para la hebra molde, pero antiparalelamente, es decir, en sentido 5´-3´tanto para el transcrito como para la hebra codificante. Con todo ello, el transcrito queda en forma de *copia* del fragmento de la hebra codificante. Pueden copiarse fragmentos de ambas hebras siempre que se guarde la direccionalidad indicada. Así, en la naturaleza, se encuentran moléculas de ADN con hebras que poseen tanto regiones codificantes como regiones molde.



Adaptado de genome.gov[6]

Figura 6. Hebras codificante, molde y transcrito (ARN). Este último posee una secuencia que es prácticamente una copia de la información de la hebra molde, salvo que en el transcrito siempre hay una base U (uracilo, ver más bajo) donde en la hebra molde había una T

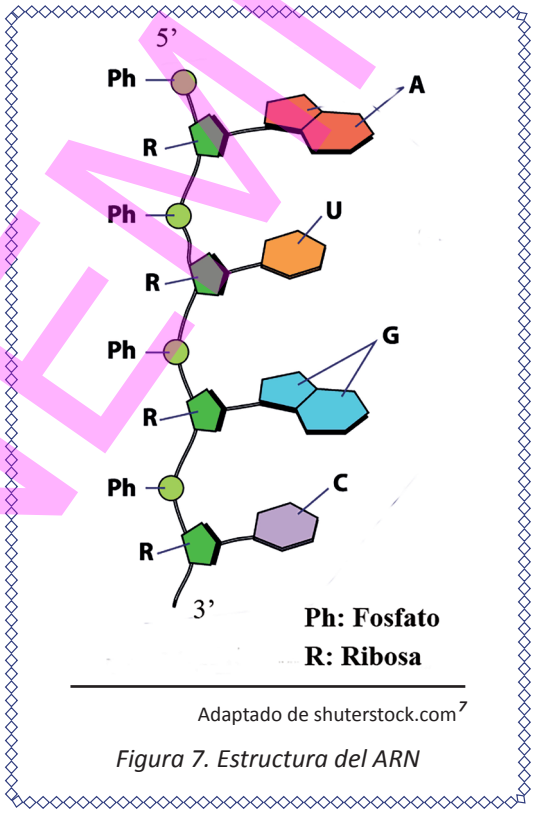
Aunque a efectos de información nucleotídica puede considerarse al transcrito como una copia exacta de un fragmento de la hebra codificante, a nivel molecular existen diferencias:

1. El azúcar consta de un oxígeno más (denominándose por tanto ribosa en lugar de desoxirribosa)

2. La base nitrogenada que se coloca complementariamente sobre la C de la hebra molde no es T, sino uracilo (U).

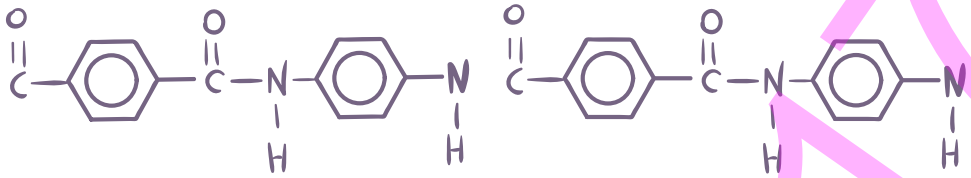
3.Estas características hacen que en general no se establezca posteriormente una hebra antiparalela, por lo que el transcrito suele tener sólo una hebra. (No obstante, muchas veces ocurre que parte de una hebra de ARN es reversa complementaria de otra parte, y ambas partes se unen complementariamente, configurando estructuras moleculares con distintas funciones biológicas).

Así, los transcritos no son moléculas de ADN, sino de **ácido ribonucleico** (ARN).



Adaptado de shuterstock.com[7]

Figura 7. Estructura del ARN

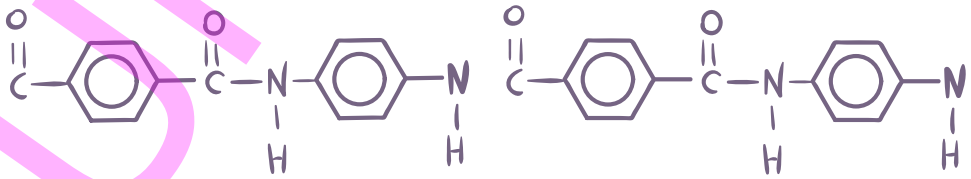


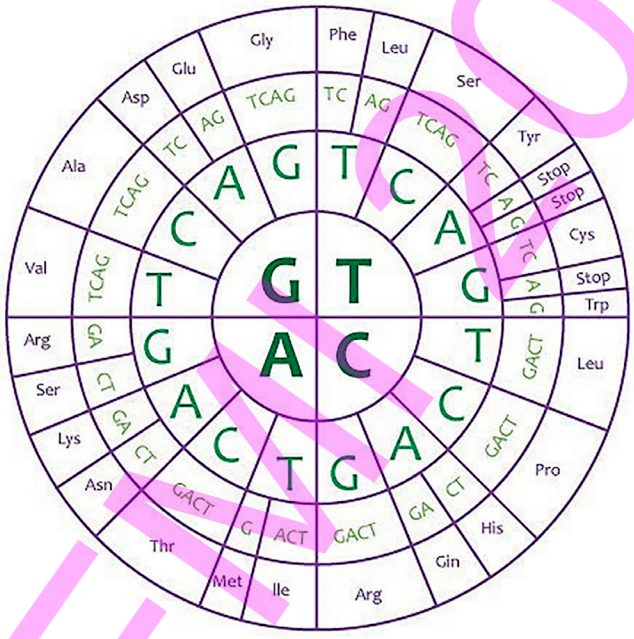
Los principales tipos de transcritos son:

1. ARNs reguladores, que controlan el proceso de expresión génica (transcripción de ADN a ARN y traducción de ARN a polipéptidos)

2. ARN ribosomal (ARNr), que llega a formar parte de los **ribosomas,** unidades funcionales de la maquinaria molecular que biosintetiza polipéptidos.

3. ARN mensajero (ARNm), el que con más puridad podría asimilarse a las instrucciones que se están ejecutando en la RAM de la microcomputadora célula, que lleva la información copiada del ADN (disco duro) a la maquinaria que comienza la construcción de proteínas, para especificar cómo es el polipéptido inicial que se forma (*output*). En el ARNm, los tripletes de nucleótidos o unidades básicas de información se denominan, **codones de ARN**, que, a efectos de transmisión de información, cada codón del ARN es copia de cada uno de los 64 tripletes del ADN, los codones del ARNm son también 64.

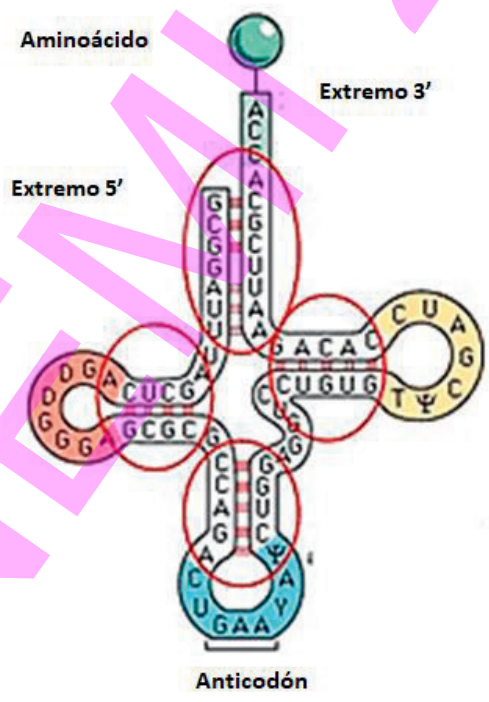




Fuente: flickr.com[8]

Figura 8. Aminoácidos (indicados con su abreviatura más común, de tres letras) especificados por cada codón

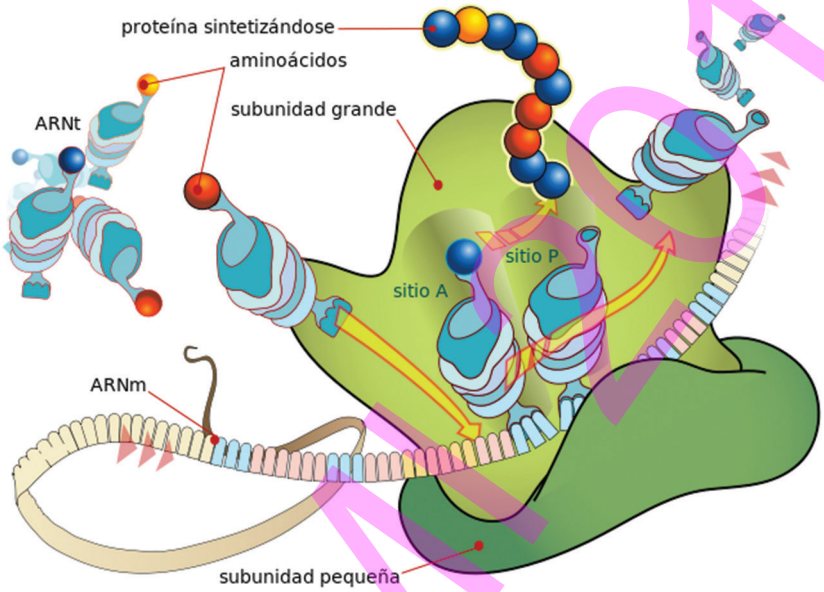
4. ARN de transferencia (ARNt), que transfiere las pequeñas moléculas (aminoácidos) que van a formar parte de los polipéptidos. A efectos de trasferencia de la información molecular, existen casi tantos tipos de ARNt como codones, pues cada ARNt posee en su secuencia nucleotídica un triplete de nucleótidos (**anticodón**) complementario de un codón concreto. Los ARNt que llevan un anticodón concreto portan para su transferencia un aminoácido concreto. Es decir, un codón (del ARNm) determinará un anticodón que estará en un ARNt concreto con un aminoácido específico, determinando así un anticodón un aminoácido. Es decir, un codón determina un anticodón, y éste un aminoácido. O sea, un codón también determina un aminoácido. La única excepción a esta generalidad son los denominados codones de terminación, que sólo indican el fin de la cadena de polipéptidos, y no están asociados a ningún ARNt y, por tanto, a ningún aminoácido.



Adaptado de: bdigital.unal.edu.co[9]

Figura 9. Esquema de ARN mensajero portando un aminoácido

El polipéptido saliente de los ribosomas consiste en la yuxtaposición de los aminoácidos mediante un tipo de uniones químicas denominadas enlaces peptídicos. En la mayoría de cada uno de los organismos vivos existen sólo 20 aminoácidos, que son transferidos por los casi 64 tipos (según el anticodón que lleven) de ARNt para formar el polipéptido. Puesto que hay casi 64 tipos de ARNt (anticodones) y 20 aminoácidos, algunos anticodones –y, por ende, sus correspondientes codones de ARNm y del ADN– pueden considerarse *sinónimos.* Es decir, varios tipos de ARNt pueden transferir el mismo aminoácido para el polipéptido naciente. De hecho, un aminoácido puede estar codificado por entre 1 y 4 codones. En general, existen tres codones "de terminación", es decir, de señal de finalización del polipéptido, y que no poseen anticodón complementario, por lo que en realidad sólo existen 61 tipos de ARNt en la mayoría de los organismos.



Fuente: wikimedia.org[10]

Figura 10. Síntesis de proteínas

Al igual que la secuencia nucleotídica codificante y la secuencia nucleotídica de ARNm. La secuencia de aminoácidos del polipéptido naciente posee una direccionalidad: se origina en el extremo del polipéptido denominado N (amino) y finaliza en el extremo C (carboxilo). Es decir, el extremo 5´ de la secuencia de ADN codificante, y por lo tanto el extremo 5´ del ARNm se corresponden con el extremo N de la proteína, y consecuentemente los extremos 3´ de aquellas secuencias nucleotídicas (de ácidos nucleicos) se corresponden con el extremo C de las secuencias aminoacídicas (peptídicas, proteicas).

Los **virus** son considerados organismos no celulares que precisan infectar una célula para que el genoma de los mismos (de ADN o ARN) entre en funcionamiento. El genoma de una célula (siempre ADN) se encuentra:

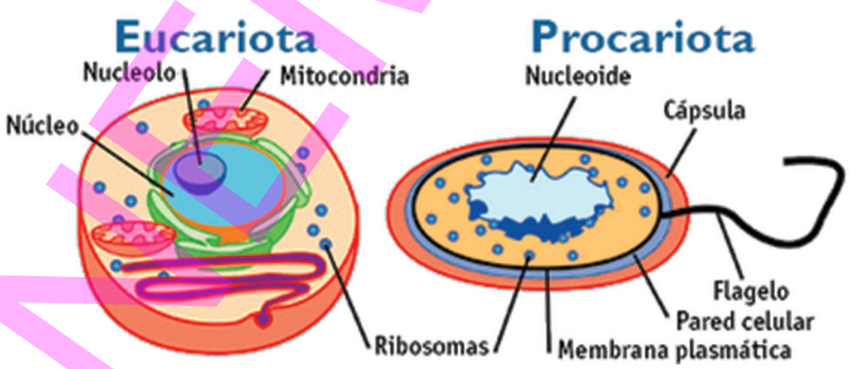
1. En el caso de **procariotas** (bacterias), en el cromosoma bacteriano (nucleoide), que consiste fundamentalmente en una cadena circular de ADN que se encuentra suspendida en el interior (citoplasma) de la célula bacteriana. Las bacterias pueden tener pequeños genomas extracromosómicos denominados **plásmidos**.

2. En el caso de **eucariotas** (resto de organismos celulares que no son bacterias), en determinados compartimentos celulares, que no poseen las bacterias, rodeados con doble membrana:

2.1. **Núcleo** celular, donde se encuentra casi todo el genoma de la célula, generalmente compartimentalizado dentro de unas estructuras subnucleares denominadas **cromosomas.**

2.2. **Mitocondrias.**

2.3. **Cloroplastos**, que se encuentran sólo en vegetales.

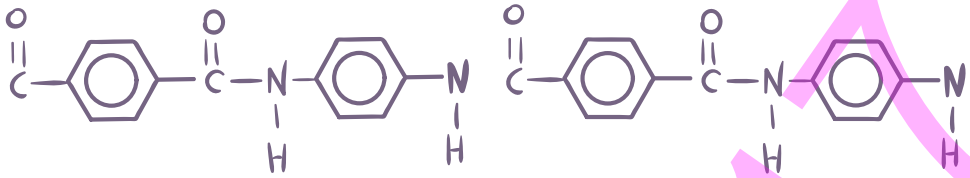


Fuente: wikimedia.org[11]

Figura 11. Estructura de células eucariota y procariota

1.3. GENES

El **código genético** es casi universal, con ligeras variaciones en algunos organismos y entre mitocondrias pertenecientes a distintos grupos taxonómicos de organismos. Asimismo, como se ha mencionado anteriormente, los aminoácidos son los mismos para casi todos los organismos, con una diferencia fundamental entre procariotas y eucariotas, poseyendo los primeros el aminoácido N-formilmetionina en lugar de metionina.



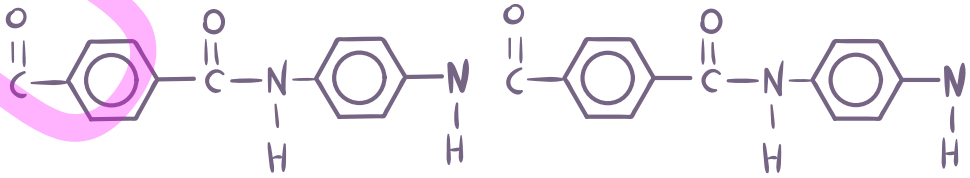
Pude definirse un **gen** como una unidad del genoma funcional a escala celular a través de una secuencia de transcrito. Es decir, un gen equivaldría a la unión de secuencias genómicas continuas que contienen informaciones necesarias para biosintetizar al menos una macromolécula (RNA, proteínas) con función celular específica.

La mayor parte del genoma no está formada por genes, lo que no quiere decir que no posea funciones estructurales o reguladoras. El resto del genoma son secuencias nucleotídicas requeridas para que se forme un transcrito funcional a nivel celular, es decir, consiste en grupos de genes contiguos (operones). En general, este número de genes se reduce a uno en eucariotas. En el caso de transcritos como los ARN reguladores, el gen no lleva secuencias codificantes de proteínas, ya que el propio transcrito tiene la función celular en sí mismo sin necesidad de traducirse a polipéptido.

Un gen codificante de polipéptidos consiste generalmente en una secuencia de nucleótidos que básicamente se divide en:

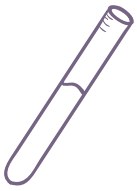
1. **Secuencias reguladoras que no se transcriben**, a ambos lados de la secuencia que se transcribe, controlando el proceso de la transcripción en dependencia del entorno molecular.

2. **Cuerpo del gen**, que es la secuencia que se transcribe dando lugar: a) en procariotas, generalmente al ARNm policistrónico (que porta de manera yuxtapuesta la información necesaria para la síntesis de varias proteínas en sendos **cistrones**); o b) en eucariotas, al **ARN heterogéneo** nuclear (**ARNhn** o **pre-ARNm**, generalmente monocistrónico, precursor del ARNm maduro). El cuerpo del gen se divide en los siguientes tipos de secuencias:



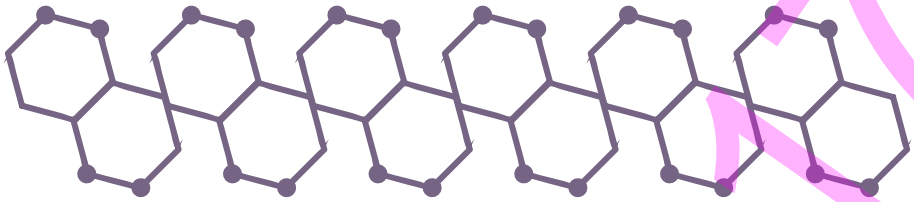
2.1. **Intrones:** regiones del genoma nuclear de los eucariotas que se transcriben pero que inmediatamente se eliminan del ARNhn, un fenómeno denominado ***splicing*** (corte y empalme) génico, dando lugar al ARNm maduro y a las regiones cortadas, que constituirán a su vez un tipo de ARN reguladores. El sitio de corte de intrones puede variar en algunos genes, dando lugar así al fenómeno de ***splicing alternativo***, originando de esta forma distintos transcritos (y distintos polipéptidos) a partir de un gen.

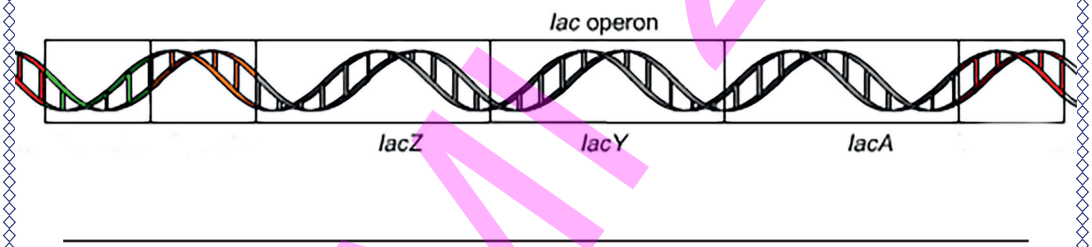
2.2. **Exones:** regiones del genoma cuya transcripción da lugar a la secuencia de ARNm y que por tanto portan la información que finalmente irá a los ribosomas para la biosíntesis de polipéptidos. Un gen procariota sólo posee un exón, y el de un eucariota puede tener un número indefinido de exones. El conjunto de exones posee a su vez dos tipos de regiones:



2.2.1. Regiones que se transcriben pero que no se traducen, es decir, regiones trasncribibles no codificantes de proteína. El primer y último exón de una secuencia transcribible ininterrumpida siempre portan en sus extremos inicial y final, respectivamente, regiones que no se traducirán (*untraslated regions*, **UTRs**). Así, la secuencia policistrónica o el gen tienen siempre una 5'-UTR y una 3´-UTR. El origen y el final de las UTRs también puede variar en algunos genes, dando esto lugar a distintos ARNm, y hasta a distintos polipéptidos. En procariotas, entre dos genes del mismo operón, se encuentran también UTRs que portan regiones de unión al ribosoma.

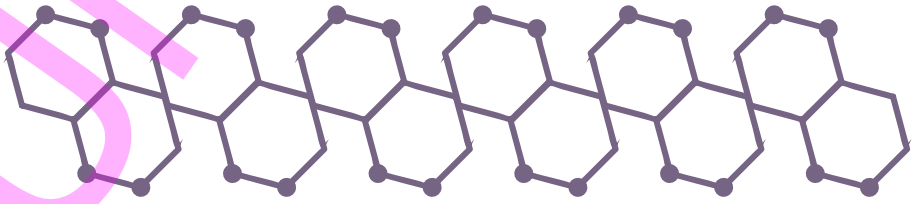
2.2.2. Las regiones del gen que finalmente codifican el polipéptido, incluyendo la señal de terminación, y que son las que portan por tanto los codones, se denominan en conjunto secuencia codificante (*codifying sequence*, **CDS**) del gen.

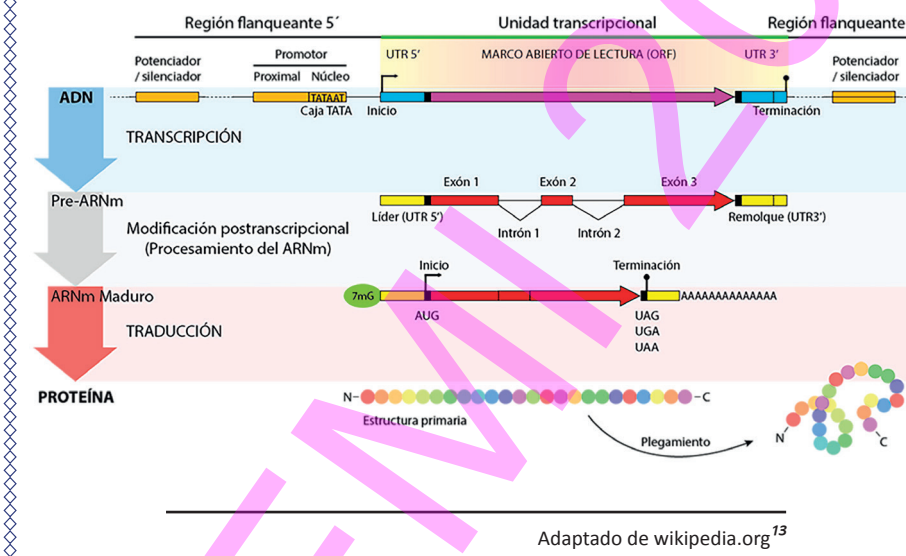




Adaptado de: wikimedia.org [12]

Figura 12. Esquema de un operón en procariotas12. Se indican regiones reguladoras (a la izquierda) y el cuerpo de los genes policistrónicos (genes estructurales) a la derecha.





Adaptado de wikipedia.org[13]

Figura 13. Estructura y transcripción de gen eucariota

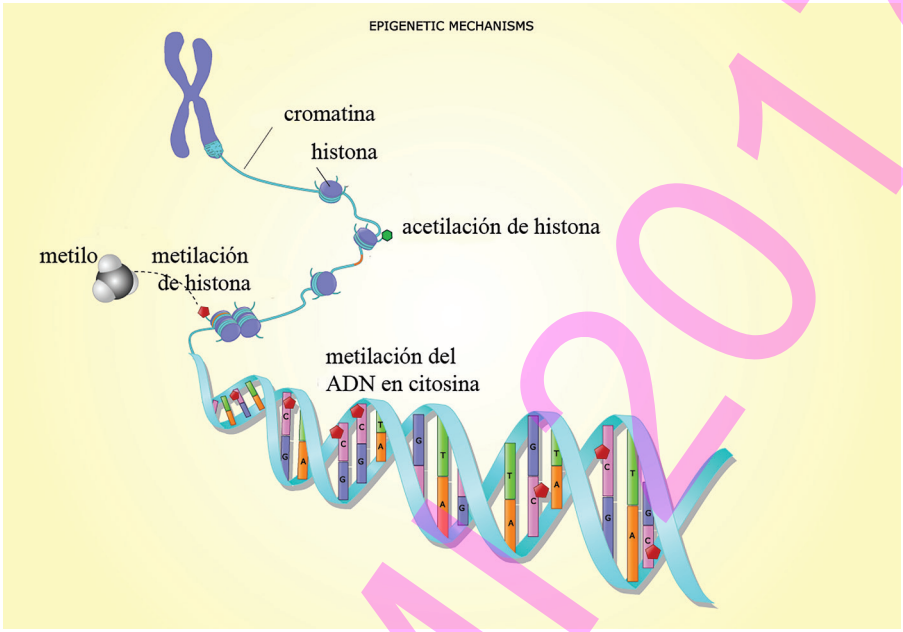
1.4. EPIGENÉTICA: UN RECIENTE PARADIGMA EN BIOLOGÍA

Es bien conocido el papel que desempeña el genoma en la herencia biológica. Aunque dicho rol no es tan importante como cabría suponer, y de hecho se suponía, en un principio. Esto es así porque las secuencias genómicas, que pasan de una generación de células u organismos a otra, no siempre indican una predeterminación del organismo para adquirir un cierto rasgo, ya que en la mayoría de los casos indican simplemente una predisposición.

Así pues, ¿qué es lo que hace que una determinada secuencia de material hereditario confiera un rasgo en un individuo y en otro no?

Un ejemplo típico para ilustrar la respuesta es un caso extremo: el de los gemelos univitelinos. Estos son idénticos al nacer, pero a lo largo del tiempo se van tornando cada vez más distintos. Esto no sucede porque su genoma vaya acumulando mutaciones; la causa fundamental son las marcas epigenéticas que cada gemelo va adquiriendo diferencialmente a medida que se van exponiendo cada uno a situaciones distintas en su vida. Esas marcas no son más que diferentes tipos de alteraciones químicas, que se disponen alrededor de las secuencias genómicas inalteradas de manera distinta para cada gemelo, modificando de forma diferente la lectura que efectúa la célula de las instrucciones que dichas secuencias encriptan, y rindiendo finalmente diferentes rasgos para cada uno de los hermanos.

Tanto la copia del DNA a una molécula igual para la división celular (replicación) como la copia de su información a transcritos (transcripción) requiere que las hebras de DNA se separen. La separación de hebras de DNA para que ocurra la transcripción (en cuanto a iniciación e intensidad) está básicamente regulada por las marcas epigenéticas, es decir, alteraciones químicas del DNA (fundamentalmente metilación de citosinas) o de las proteínas que lo empaquetan (histonas, en eucariotas) en forma de cromosomas, y que pueden heredar las células descendientes.



Adaptado de shutterstock.com[14]

Figura 14. Modificaciones epigenéticas. Se representa la metilación de la citosina en las hebras del ADN, y algunas modificaciones químicas de las histonas

Estrictamente hablando, las marcas epigenéticas se mantienen entre generaciones celulares. Sin embargo, a lo largo del ciclo de vida de un individuo van cambiando de naturaleza y posición, a veces de manera regulada independientemente por el propio organismo, y otras veces por efecto del ambiente, como en el caso expuesto de los gemelos. Ha de decirse que la mayoría de las marcas epigenéticas adquiridas por los tejidos que originan las células reproductoras se borran, hasta adquirir un estatus epigenético primigenio idóneo para el desarrollo de un futuro descendiente. Quizá ésta es una de las razones por las cuales la clonación de mamíferos ha sido hasta hoy sólo parcialmente exitosa. Así, se ha llegado a decir que la oveja Dolly murió joven porque no hubo una reprogramación hasta ese estado epigenético de partida por parte de las células adultas empleadas en el proceso de clonación.

Por otra parte, si bien es cierto que a lo largo de la vida de un ser vivo pueden manifestarse algunas enfermedades que están determinadas por secuencias genéticas, muchas otras dolencias son adquiridas, por ejemplo, mediante cambios epigenéticos, ya sea por un proceso de desarrollo, envejecimiento, o exposición a ambientes (naturales e incluso sociales) adversos.

Por último, las implicaciones de los descubrimientos sobre epigenética en bilogía poblacional y evolutiva son incluso sorprendentes. Las marcas epigenéticas pueden conferir ventajas de adaptación al ambiente, e incluso pueden direccionar ciertos tipos de alteraciones, esta vez sí, genéticas. Esto ilustra otro dogma que se ha derrumbado: hoy en día se sabe que las mutaciones en el genoma podrían ser hasta cierto punto dirigidas por el ambiente, resucitándose así desechadas ideas antiguas sobre la evolución biológica.

Los ejemplos expuestos muestran cómo la epigenética ofrece una nueva perspectiva en el modo de entender el funcionamiento de la vida y, por tanto, de cómo controlarla sustentablemente para el beneficio de la sociedad. Las implicaciones del desarrollo de esta nueva disciplina en la ciencia y en la biotecnología van desde una revolución en la biología de sistemas hasta el desarrollo de nuevas aplicaciones en, por ejemplo, biomedicina. En cuanto a la agrobiotecnología, se espera que el amplio conocimiento que se está obteniendo sobre la epigenética vegetal se aplique a medio plazo al control de enfermedades, pestes y productividad en plantas de interés.

Así pues, cualquier proyecto de investigación en biología molecular debería tener en cuenta los factores de tipo epigenético. Por esta razón son esenciales los estudios bioinformáticos epigenómicos que rentabilicen resultados y datos dispersos en bases accesibles. Es cierto que, por las características y estado actual de conocimiento del tema aquí brevemente esbozado, en general las investigaciones relacionadas serán más bien fundamentales y sus conclusiones no tendrán una aplicación inmediata, pero cabe reflexionar que la investigación básica es el mejor trampolín para realizar saltos cualitativos en el conocimiento científico y por ende en el desarrollo tecnológico.

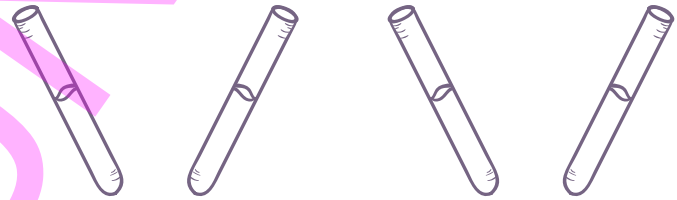
El estudio bioinformático de la metilación genómica (metiloma) se reduce, como en el caso de los fragmentos de ácidos nucleicos y polipéptidos, al estudio de secuencias, pero con un mayor número de elementos que las secuencias genómicas, es decir, añadiendo a los nucleótidos normales nucleótidos metilados.

El cambio metilómico más generalizado es la interconversión de la base nitrogenada citosina en 5-metilcitosina. Si bien existen técnicas modernas para detectar citosinas metiladas directamente en una secuencia de ADN, los métodos de análisis de metilación de citosinas generalmente implican la transformación de las citosinas no metiladas en timina, lo cual se consigue tras un tratamiento del ADN con bisulfito. Después, por comparación con la secuencia genómica, se pude deducir cuáles son las citosinas metiladas.

El análisis cuantitativo de sitios metilados se complejiza al considerar el grado de metilación de un nucleótido en un tejido, ya que los sitios de metilación pueden cambiar entre células contiguas, variación que no suele ocurrir cuando se analizan secuencias genómicas. Así, se pueden detectar distintos sitios de metilación en el mismo fragmento del genoma para una muestra que contenga el ADN nuclear de un grupo de células del mismo tejido.

[Para una profundización actual sobre epigenómica se puede consultar la revisión de Stricker et al (2016).]

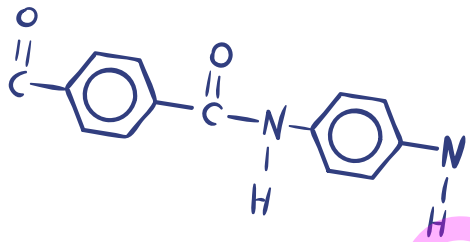
[Para adquirir más conocimientos de Bioquímica y Biología Molecular puede consultare el libro de texto de Papachristodoulou et al (2014).]



CAPÍTULO 2:

BASES DE DATOS BIOINFORMÁTICAS PÚBLICAS

Lic. Jesennia Cárdenas-Cobo, MBA.
Ing. Mirella Correa-Peralta, MBA.
Lic. Carlos Noceda-Alonso, PhD.



2. BASES DE DATOS PÚBLICAS.

Las bases de datos importantes en biología incluyen genes, nucleótidos, proteínas, estructura de proteínas, genomas, bibliografía, taxonomía, metabolismo, herramientas de análisis de datos y herramientas para manejar y recuperar información.

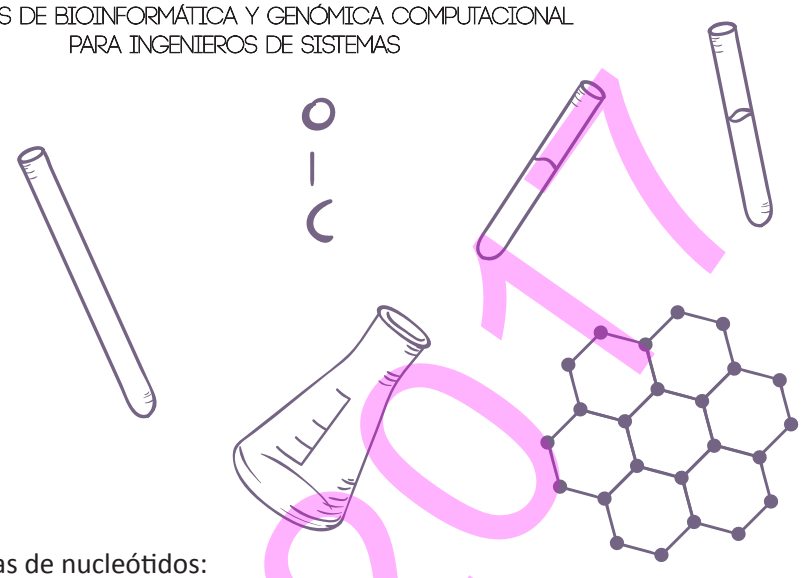
Las principales bases de datos bioinformáticas se encuentran en los siguientes centros:

- NCBI (*National Center for Biotechnology Information,* Estados Unidos).
- EMBL (*European Laboratory of Molecular Biology,* Europa).

Existen bases de datos especializadas, por ejemplo, para:

1. Bibliografía:
   - PubMed (Acceso a bases de datos bibliográficas del NLM (*National Library of Medicine*), como MEDLINE y PreMEDLINE.

2. Especies concretas:
   - TAIR (planta *Arabidopsis*)
   - SGD (levadura *Saccharomyces*)
   - Flybase (mosca *Drosophila*)



3. Secuencias de nucleótidos:
   - GenBank (Base de datos de secuencias genéticas de los NIH (*National Institutes of Health,* Estados Unidos)
   - DDBJ (*DNA Data Bank of Japan*)

4. Proteínas:
   - Uniprot (*Universal Protein Resource*)
   - PDB (*Pakistan Data Base*)

La información se encuentra almacenada en formatos estructurados, pudiendo realizarse la consulta a través de internet, pues los sitios web cuentan con herramientas propias para obtener, mostrar e incluso manipular la información.

2.1. NCBI

El **Centro Nacional para la Información Biotecnológica (*National Center for Biotechnology Information,* NCBI**) forma parte de la Biblioteca Nacional de Medicina de los Estados Unidos (*National Library of Medicine*), perteneciente a los Institutos Nacionales de Salud (*National Institutes of Health,* NIH)

Además de proveer el acceso a datos biológicos de relevancia en diversas bases de datos que están disponibles de manera gratuita, el NCBI ofrece algunas herramientas bioinformáticas, siendo BLAST (Ver Sección 3.4. Búsqueda de secuencias: BLAST) la más usada. Así, el NCBI almacena información de las secuencias genómicas en Genbak (base de datos de secuencias genéticas del NIH), y publicaciones referentes a biomedicina, biotecnología, bioquímica, genética y genómica en PubMed.

La *International Nucleotide Sequence Database Collaboration* intercambia datos a partir de la integración de la base de datos de ADN de Japón (*DNA DataBank of Japan, DDBJ*), la del Laboratorio Europeo de Biología Molecular (*European Molecular Biology Laboratory, EMBL*) y del GenBank del *National Center for Biotechnology Information*. Así, el GenBank del NCBI recepta y recibe secuencias genéticas de los laboratorios de todo el mundo.

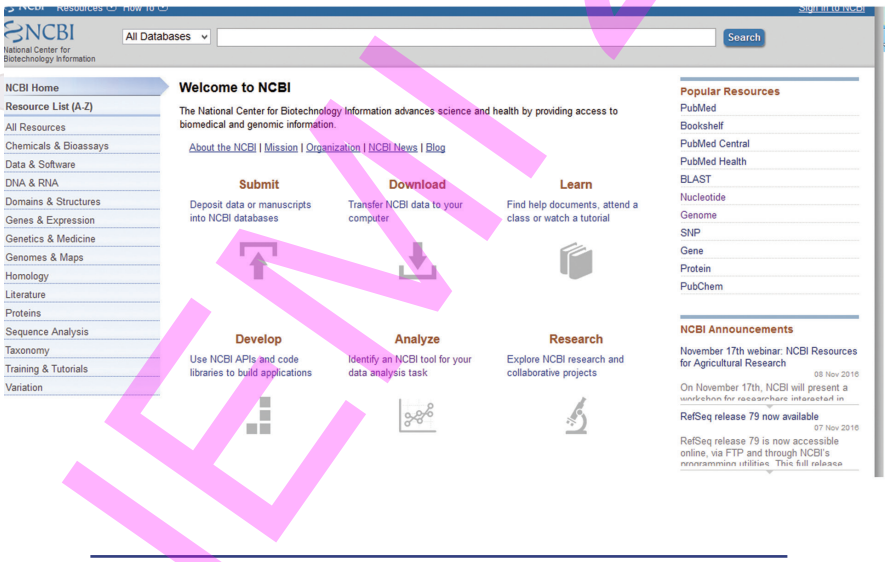


Figura 15. Portal del sitio web de NCBI.

En el NCBI, al ingresar a todos los recursos (*All Resources*) y escoger *databases* se presentan varias bases de datos. Por ejemplo, pueden explorarse:

Gene: Es la base de datos para genes del NCBI. Incluye los genes que han sido completamente secuenciados.

Genome: Información de genomas sobre la que se pueden hacer búsquedas a partir de secuencias de nucleótidos

Nucleotide: Reúne información de otras tres bases de datos: GenBank, EMBI y DDBJ.

Popset: Contiene conjuntos de secuencias alineadas que son el resultado de estudios de mutaciones, poblacionales o filogenéticos (evolutivos).

Proteins: Contiene secuencias de aminoácidos y traducciones de secuencias nucleotídicas de EMBL y DDBJ.

PubMed: Servicio de la NLM, de Estados Unidos, que incluye más de 16 millones de citas de revistas científicas sobre ciencias de la vida, como las incluidas en MEDLINE.

SNP: Contiene datos sobre variaciones de nucleótidos individuales en el genoma humano. En colaboración con el Instituto Nacional de Investigación del Genoma Humano, el Instituto Nacional de Información sobre Biotecnología (*National Human Genome Research Institute*) ha establecido la base de datos dbSNP como repositorio central.

Structure: La Base de Datos de Modelado Molecular (MMDB) contiene datos y herramientas para determinación de estructuras y dimensiones de biomoléculas a partir de información del *Protein Data Bank*

Taxonomy: Contiene los nombres de todos los organismos que están representados, al menos por una secuencia de ácido nucleico o proteína, en las bases de datos de secuencias del NCBI.

2.1.1. Cómo registrar una cuenta de usuario en NCBI

1. Ingresar a *Sign into NCBI*.

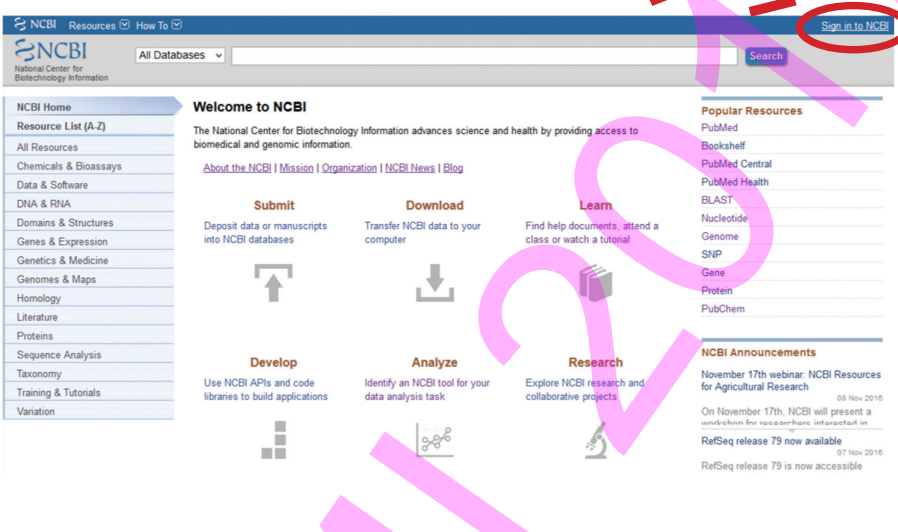


Figura 16. Crear una cuenta de usuario en NCBI.

2. Registrar una cuenta de usuario o acceder a su cuenta de usuario.

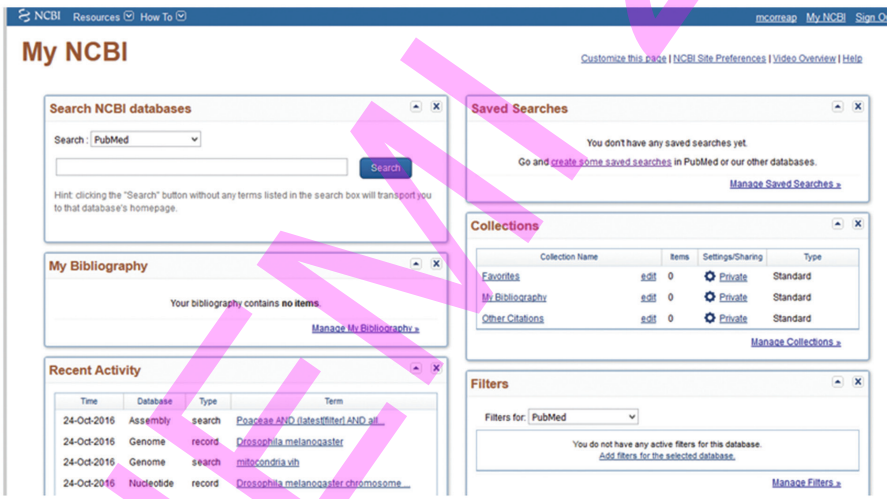
Figura 17. Acceder a la cuenta de usuario de NCBI.

3. Al introducir los datos personales se obtendrá su registro en su correo electrónico.

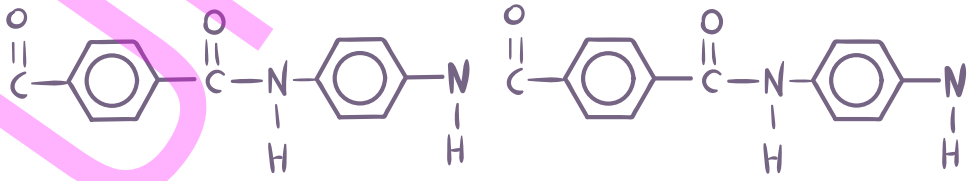
4. Ingresar a su cuenta de correo electrónico y activar su cuenta.

2.1.2. Mi cuenta en NCBI

Una vez creada su cuenta en NCBI, habrá de ingresarse en ella:



18. Mi cuenta en NCBI.



2.1.3. Agregando citas de PubMed a la cuenta de Mi Bibliografía *(My Bibliography)*

1. Ingresar a PubMed desde la página principal de NCBI.

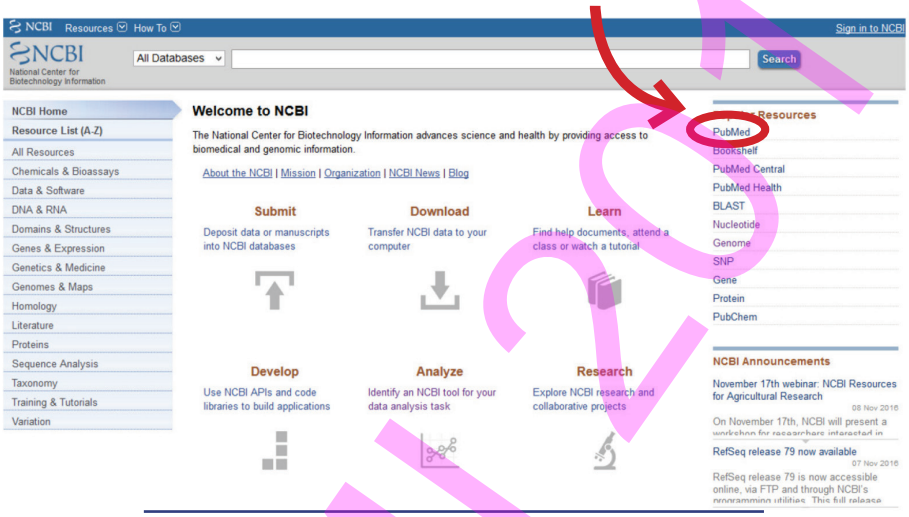


Figura 19. Acceder a PubMed a través de NCBI.

2. En PubMed, realizar una búsqueda, por ejemplo "**DNA fragmentation**"

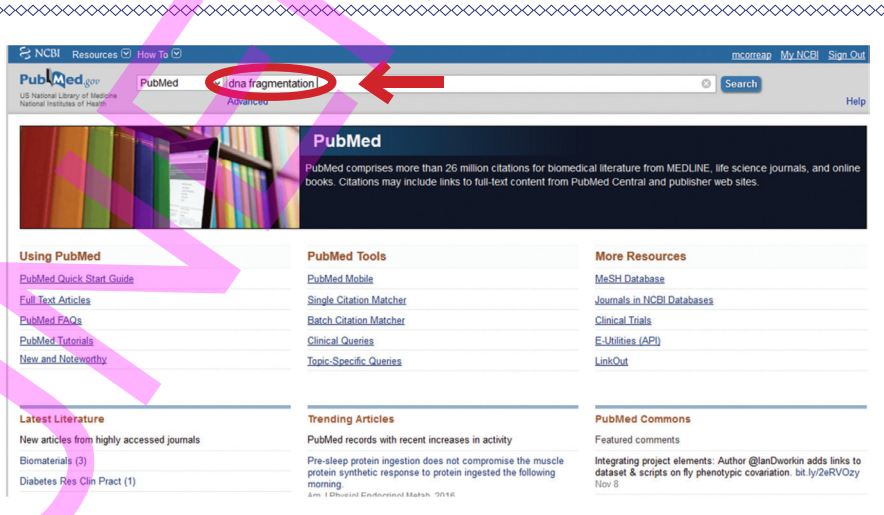


Figura 20. Búsqueda de Información en PubMed.

3. A partir de los resultados de búsqueda obtenidos, marcar y enviar a "Su bibliografía" (*My bibliography*)

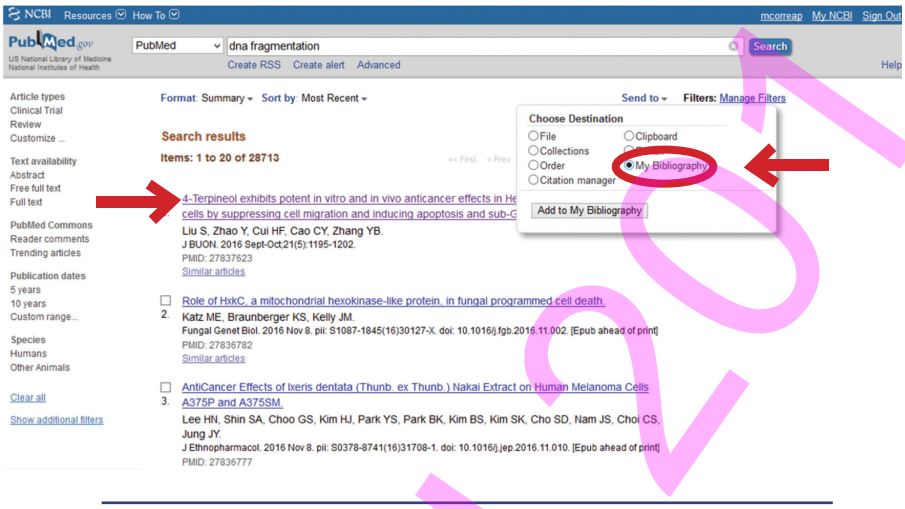


Figura 21. Agregar citas de PubMed a la cuenta de Mi Bibliografía.

4. Confirmar que se guardará en *su bibliografía*

5. Ingresar a *My NCBI* y la bibliografía estará actualizada. Verificar su historial

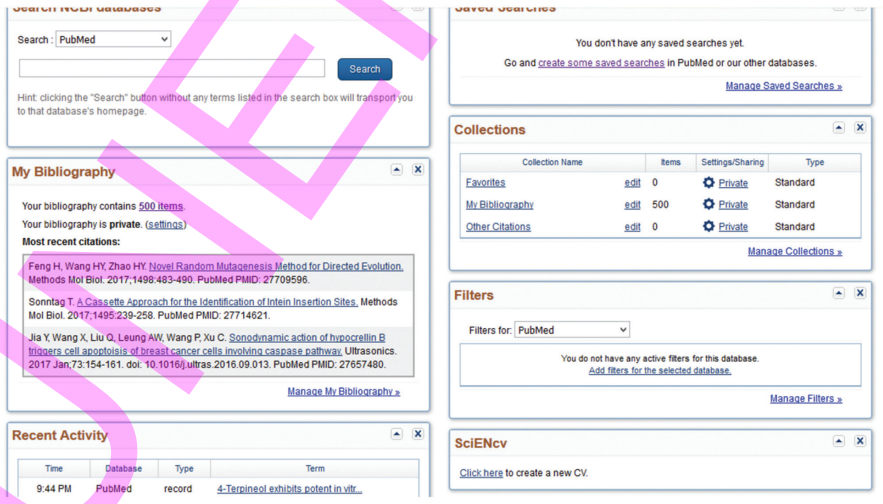


Figura 22. Verificación de citas de PubMed en Mi Bibliografía (My Bibliography).

2.1.4. Revisión de estructuras moleculares

1. Escribir el objeto de búsqueda (por ejemplo, "*Homo sapiens*") en la base *Structure*, y revisar las imágenes.

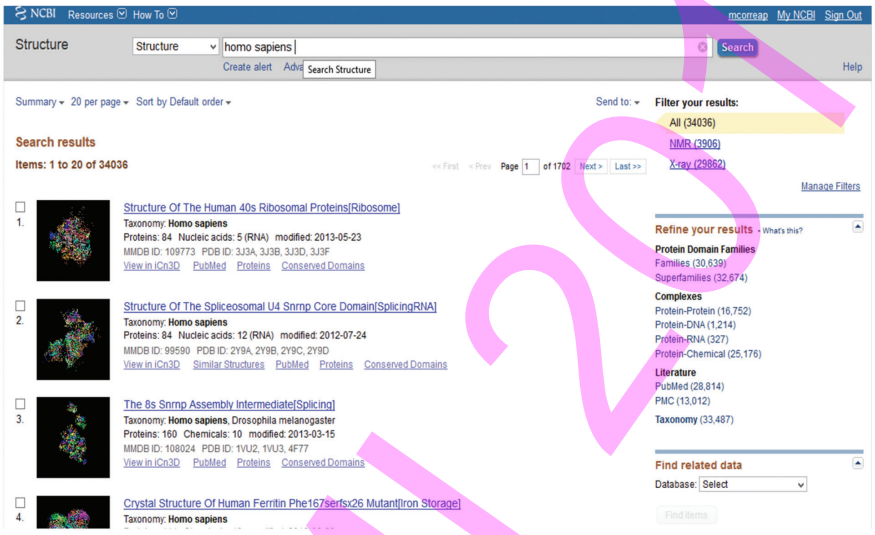


Figura 23. Estructura de biomoléculas del Homo Sapiens en el NCBI.

2. Observar y consultar la estructura molecular que interese

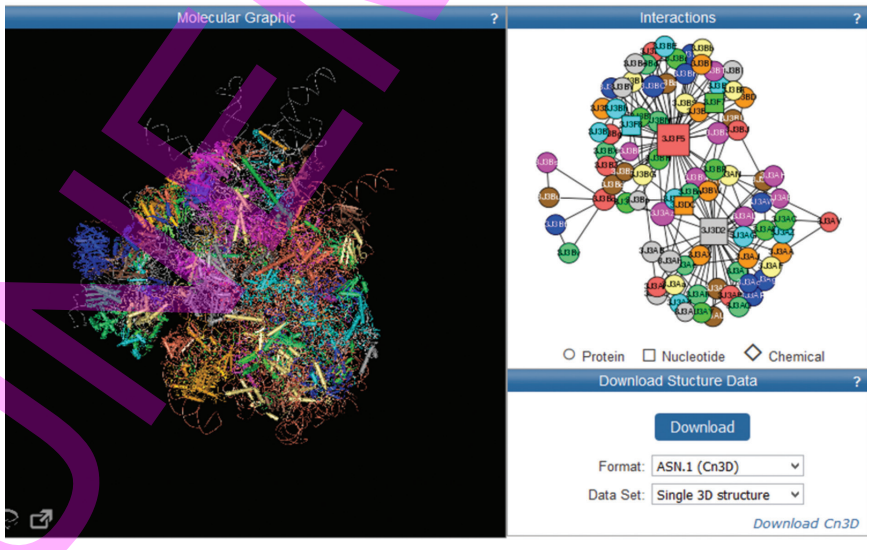


Figura 24. Una de las estructuras moleculares, con información complementaria, obtenida a partir de la base Structure del NCBI.

2.1.5. Secuencias de ADN en distintos formatos y manejo de las mismas

1. Ingresar a https://www.ncbi.nlm.nih.gov/.

2. Escoger nucleótidos (*nucleotide*).

3. Efectuar una búsqueda en *Homo sapiens.*

4. Seleccionar una secuencia.

5. Aparece por defecto la información en formato de la base de datos GenBank (Ver Sección 3.2.2. Formato GenBank).

6. A partir de aquí se pueden seleccionar varias opciones, como escoger el formato FASTA (Ver Sección 3.2.1. Formato FASTA).

7. Copiar la secuencia para utilizarla en otros programas.

2.1.6. Obtener secuencias de cromosomas completos

1. Visitar en el sitio web https://www.ncbi.nlm.nih.gov.

2. Escoger *Genome* y consultar el organismo *Homo sapiens.* Se presentará una tabla con enlaces a secuencias de referencia de los 23 cromosomas del ser humano, entre otros datos.

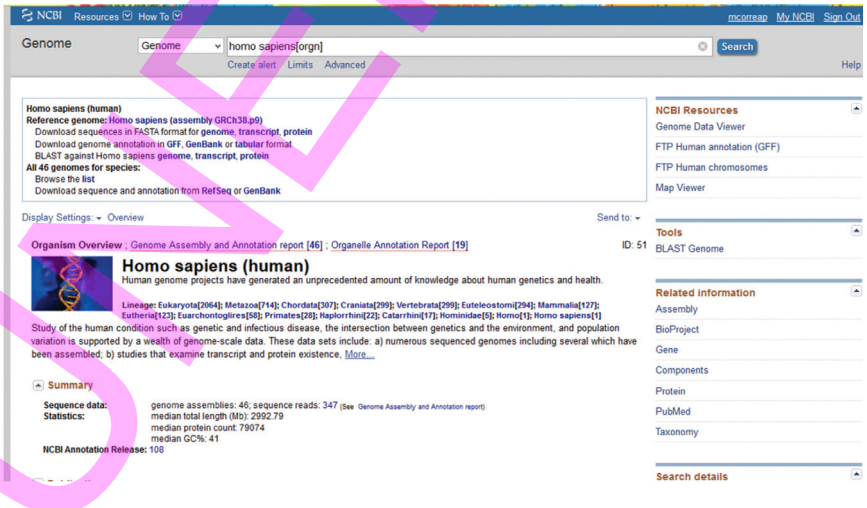
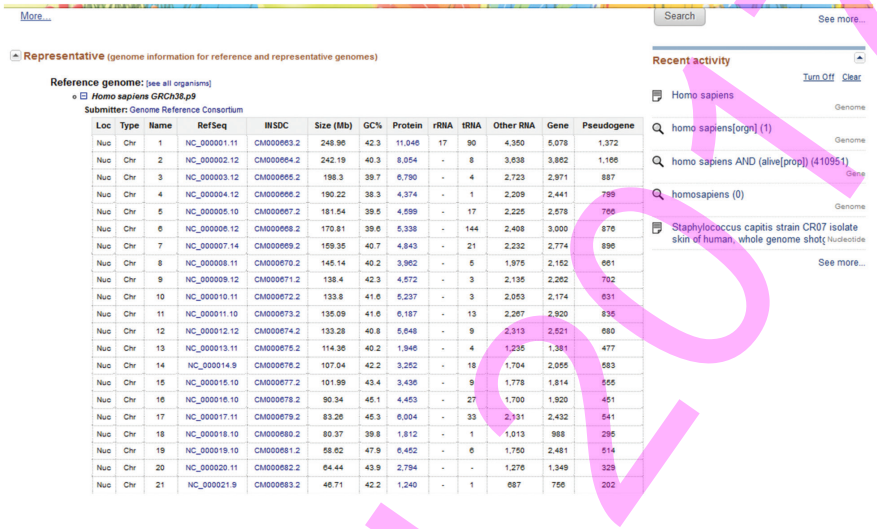


Figura 25. Genoma del Homo Sapiens en NCBI.

3. Ingresar al cromosoma con la referencia NC_000001.11



Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Nuc	Chr	1	NC_000001.11	CM000663.2	248.96	42.3	11,046	17	90	4,350	5,078	1,372
Nuc	Chr	2	NC_000002.12	CM000664.2	242.19	40.3	8,054	-	8	3,638	3,862	1,166
Nuc	Chr	3	NC_000003.12	CM000665.2	198.3	39.7	6,790	-	4	2,723	2,971	887
Nuc	Chr	4	NC_000004.12	CM000666.2	190.22	38.3	4,374	-	1	2,209	2,441	799
Nuc	Chr	5	NC_000005.10	CM000667.2	181.54	39.5	4,599	-	17	2,225	2,578	766
Nuc	Chr	6	NC_000006.12	CM000668.2	170.81	39.6	5,338	-	144	2,408	3,000	876
Nuc	Chr	7	NC_000007.14	CM000669.2	159.35	40.7	4,843	-	21	2,232	2,774	896
Nuc	Chr	8	NC_000008.11	CM000670.2	145.14	40.2	3,962	-	5	1,975	2,152	661
Nuc	Chr	9	NC_000009.12	CM000671.2	138.4	42.3	4,572	-	3	2,135	2,262	702
Nuc	Chr	10	NC_000010.11	CM000672.2	133.8	41.6	5,237	-	3	2,053	2,174	631
Nuc	Chr	11	NC_000011.10	CM000673.2	135.09	41.6	6,187	-	13	2,267	2,920	836
Nuc	Chr	12	NC_000012.12	CM000674.2	133.28	40.8	5,648	-	9	2,313	2,521	680
Nuc	Chr	13	NC_000013.11	CM000675.2	114.36	40.2	1,946	-	4	1,235	1,381	477
Nuc	Chr	14	NC_000014.9	CM000676.2	107.04	42.2	3,252	-	18	1,704	2,055	583
Nuc	Chr	15	NC_000015.10	CM000677.2	101.99	43.4	3,436	-	9	1,778	1,814	555
Nuc	Chr	16	NC_000016.10	CM000678.2	90.34	45.1	4,453	-	27	1,700	1,920	451
Nuc	Chr	17	NC_000017.11	CM000679.2	83.26	45.3	6,004	-	33	2,131	2,432	541
Nuc	Chr	18	NC_000018.10	CM000680.2	80.37	39.8	1,812	-	1	1,013	988	295
Nuc	Chr	19	NC_000019.10	CM000681.2	58.62	47.9	6,452	-	6	1,750	2,481	514
Nuc	Chr	20	NC_000020.11	CM000682.2	64.44	43.9	2,794	-	-	1,276	1,349	329
Nuc	Chr	21	NC_000021.9	CM000683.2	46.71	42.2	1,240	-	1	687	756	202

Figura 26. Consulta del cromosoma de Homo Sapiens (RefSeq NC_000001.11) del NCBI.

4. Obtener información en formato FASTA (Ver Sección 3.2.1. Formato FASTA).

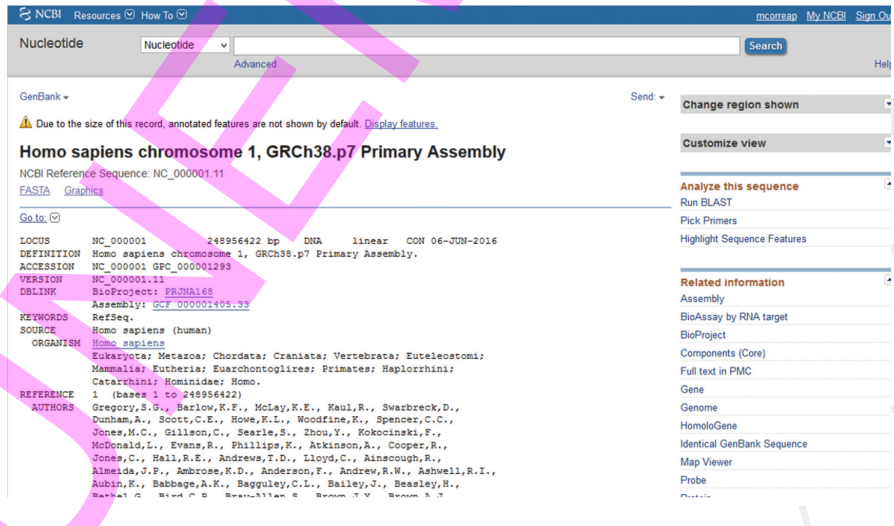


Figura 27. Cromosoma 1 del Homo Sapiens contenido en archivo FASTA en el NCBI.

5. Hacer clic en *graphics*. Aparecerán muchos datos genómicos del cromosoma seleccionado, y diversos recursos.

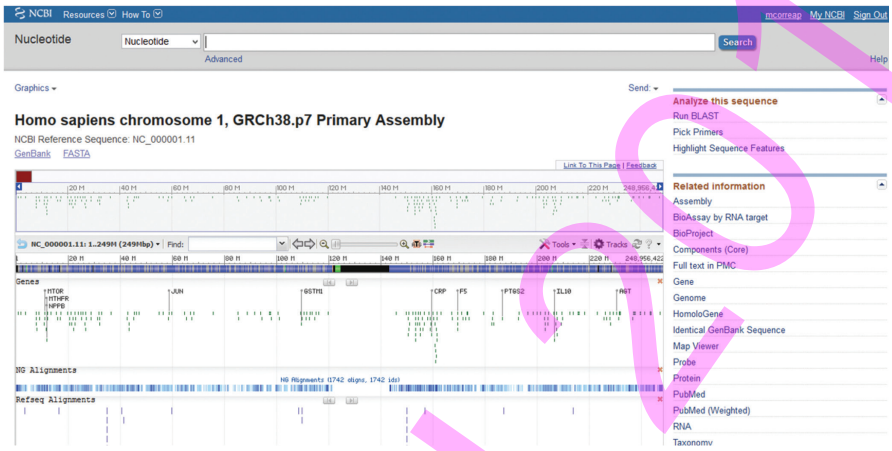
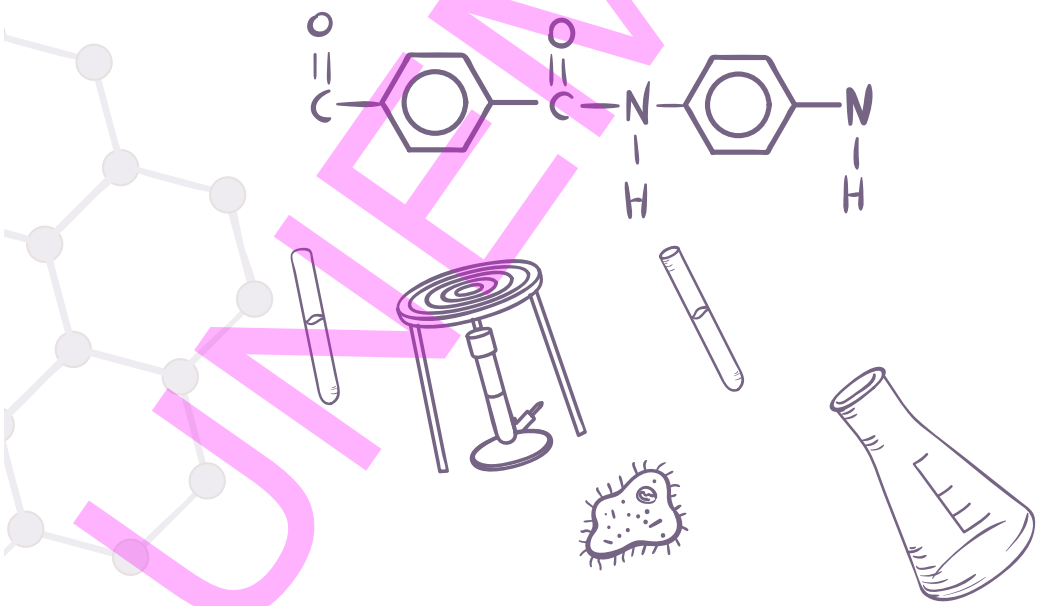


Figura 28. Gráfico del Cromosoma 1 de Homo Sapiens en el NCBI.



2.1.7 Búsqueda avanzada en NCBI

1. Ingresar a la base PMC (*PubMed Central*) y consultar información avanzada acerca del hombre publicada en el 2016.

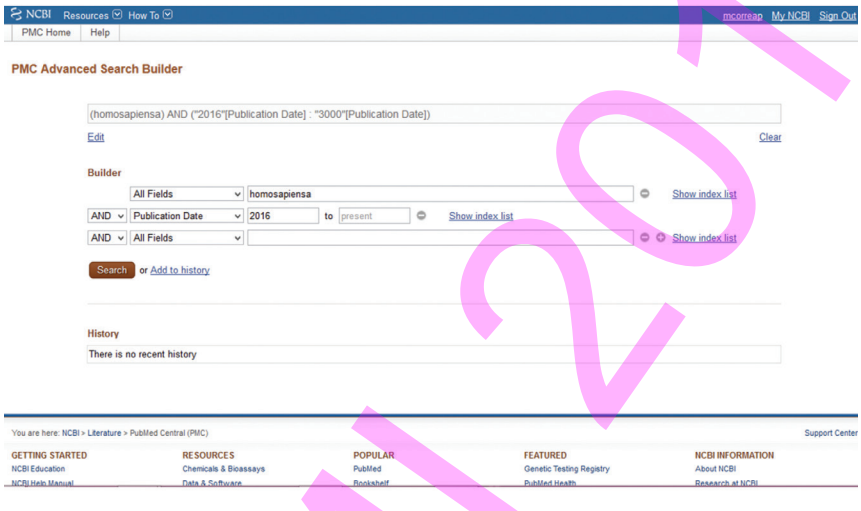


Figura 29. Consulta avanzada en la base de datos PMC del NCBI.

2. Al obtener los resultados relacionados con la consulta, hacer clic en una de las informaciones.

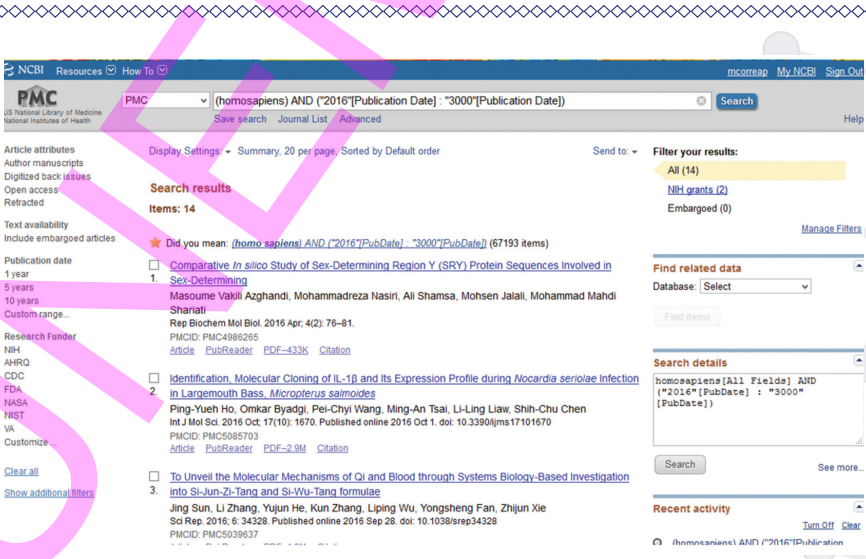


Figura 30. Resultado sobre publicaciones sobre "Homo sapiens" en 2016.

3. También se pueden efectuar consultas avanzadas con la ayuda de operadores Booleanos en mayúscula (AND, OR, NOT). Además se puede reducir la búsqueda con *Limits->Limits to* (indicando rangos de búsqueda).

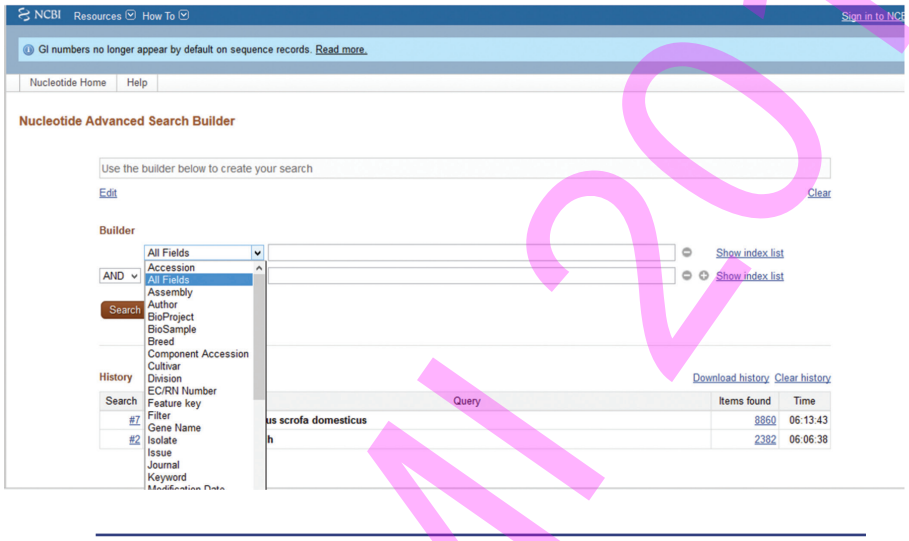


Figura 31. Consulta avanzada en NCBI.

2.2. TAIR

El Recurso de Información sobre *Arabidopsis* (TAIR) presenta una base de datos de la genética e información de biología molecular sobre *Arabidopsis thaliana*. Los datos disponibles de TAIR incluyen la secuencia completa del genoma de esta especie con su estructura génica, expresión génica, ADN de diversos ecotipos, reservas de semillas, mapas del genoma, marcadores genéticos, publicaciones e información sobre la comunidad de investigación de *Arabidopsis.* Integrada en TAIR se encuentra la información del Centro de Recursos sobre la Biología de *Arabidopsis* en la Universidad Estatal de Ohio, que recoge, reproduce, conserva y distribuye recursos de ADN de *Arabidopsis thaliana* y especies relacionadas. El acceso completo a TAIR requiere una suscripción.

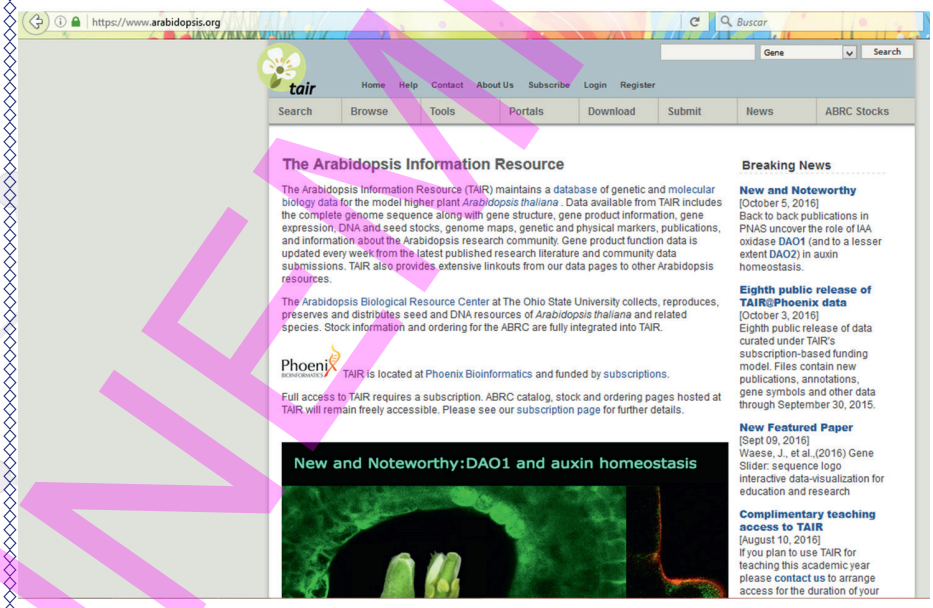


Figura 32. Portal del sitio web de TAIR.

2.2.1. Registro de cuenta en TAIR

1. Acceder a https://www.arabidopsis.org/ y escoger registro

2. Ingresar sus datos para crear su cuenta de registro

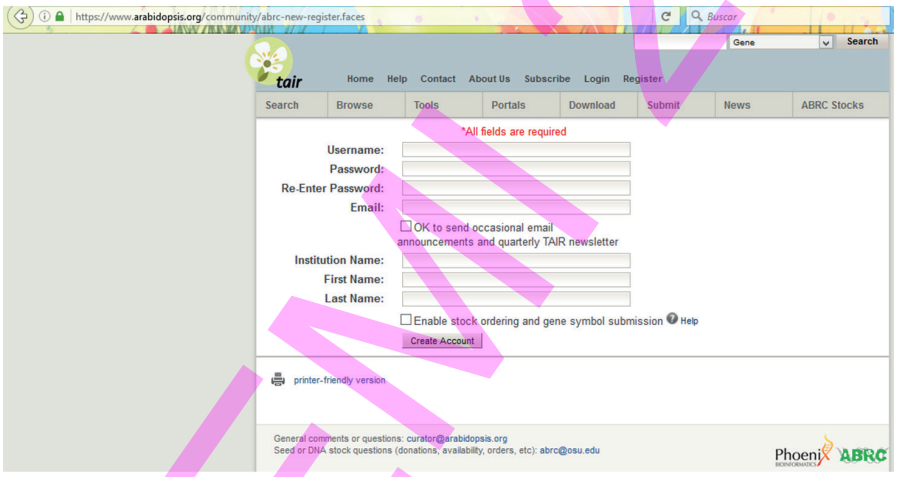


Figura 33. Registro de usuario en TAIR.

3. Ingresar a su cuenta de usuario y activar con el pago de la suscripción

2.3. VECTORBASE

Esta base de datos pertenece al Centro de Recursos de Bioinformática (BRC) del Instituto Nacional de Alergias y Enfermedades Infecciosas (NIAID), que proporciona a la comunidad científica datos genómicos, fenotípicos y centrados en la población de vectores invertebrados de patógenos humanos. (https://www.vectorbase.org)

Figura 34. Portal del sitio web de VectorBase.

VectorBase es uno de los cuatro Centros de Recursos de Bioinformática que proporcionan recursos basados en la web a la comunidad científica, y apoya la investigación sobre organismos causantes de enfermedades emergentes o re-emergentes. Todos los recursos VectorBase están a libre disposición de la comunidad de investigación bajo la Licencia Pública General del sistema operativo libre GNU.

Este proyecto recibe fondos federales del propio Instituto Nacional de Alergia y Enfermedades Infecciosas (NIAID), así como de otros Institutos Nacionales de Salud (NIH) y del Departamento de Salud y Servicios Humanos (HHS), todas instituciones estadounidenses.

La base VectorBase incorpora datos de bases de datos públicas de secuencias de nucleótidos (ENA -Archivo de Nucleótidos Europeo, que proporciona información de secuencias de nucleótidos del mundo, GenBank y DDBJ) o de otro tipo de informaciones (dbSNP, KEGG vía, UniProt, InterPro, RefSeq o ArrayExpress). Otros datos se calculan derivados de datos primarios. Específicamente, VectorBase contiene datos como secuencias de genomas de referencia, anotaciones estructurales (por ejemplo, modelos de genes codificantes de proteínas o de dominios funcionales) o datos fenotípicos. VectorBase proporciona información sobre especies de vectores, acceso a datos de recursos para esos organismos, y además posee herramientas para analizar estos conjuntos de datos.

Para acceder a los genomas alojados en VectorBase, ingresar a https://www.vectorbase.org/genomes

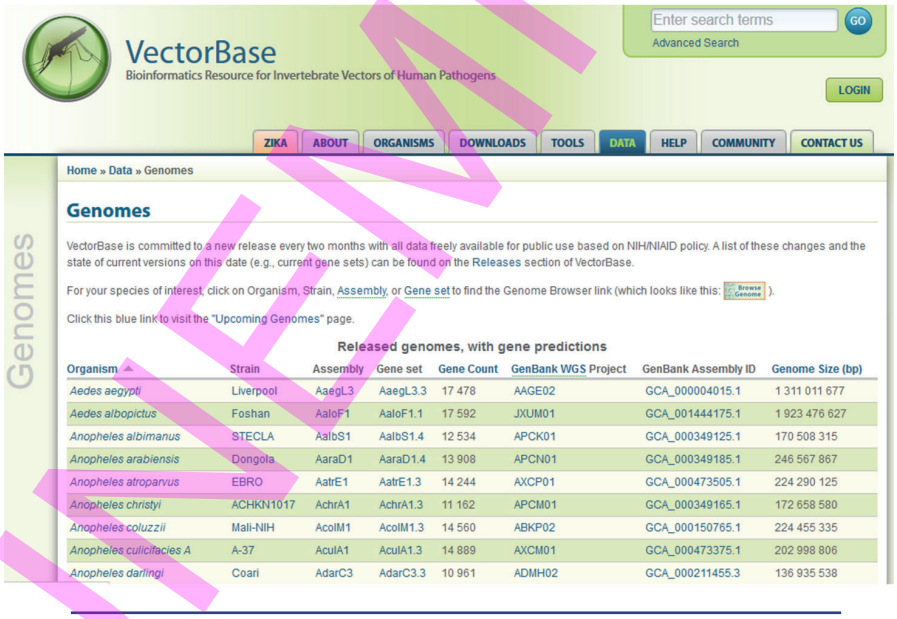


Figura 35. Lista de genomas alojados en Vectorbase.

2.3.1. Registro de cuenta de usuario

1. Ingresar a Login.

2. Crear nueva cuenta.

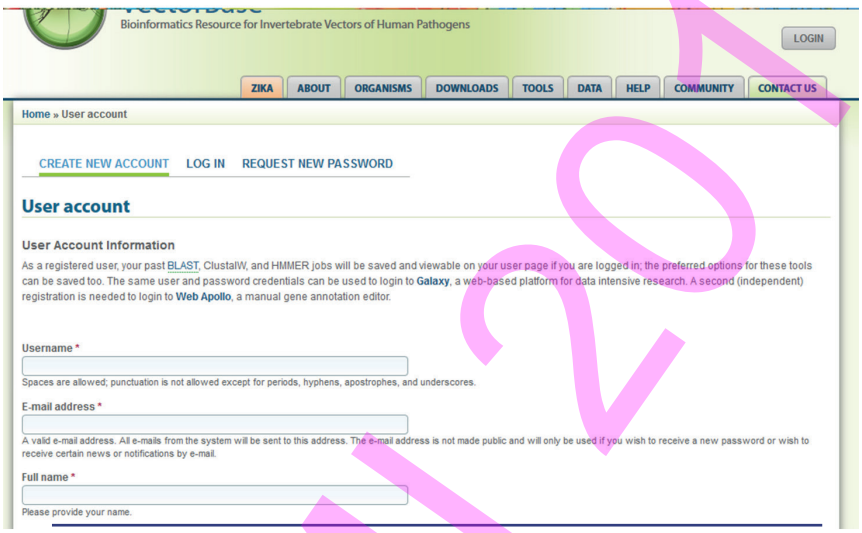


Figura 36. Crear usuario en VectorBase.

3. Activar su cuenta confirmando desde su correo electrónico registrado

2.3.2. Ejemplos de consultas y análisis en VectorBase

- Búsqueda de genes.
- Búsqueda, análisis y visualización de genomas y de otros tipos de datos con la herramienta *Genome Browser.*
- Descarga de conjuntos de datos complejos, incluyendo secuencias de nucleótidos y proteínas con la herramienta *BioMart.*
- Consulta con *BioMart* para cargar secuencias de proteínas de una familia de genes bien anotada en los genomas correspondientes.
- Utilización de datos de ARN secuenciado y/o publicados previamente para estudiar los patrones de transcripción diferencial en un organismo en diferentes condiciones experimentales.

- Utilizar BLAST (Ver sección 3.4. Búsqueda de secuencias: BLAST) para localizar una secuencia.

- Utilizar la plataforma bioinformática Galaxy con datos de secuenciación para estudiar polimorfismos (diferencias en nucleótidos en secuencias similares).

2.3.3. Ejemplo: Consultar en VectorBase sobre el virus del Zika

1. Ingresar a https://www.vectorbase.org.

2. Acceder a *Zika* y consultar la historia del virus.

Figura 37. Página web de información sobre zika.

2.4. ECOCYC

Es una base de datos (Karp et al 2014) de la bacteria *Escherichia coli* que recolecta genomas de la misma, así como datos de control transcripcional y metabólico. Describe el catálogo molecular completo de la célula de *E. coli*, así como las funciones de cada una de sus partes. Muchas referencias bibliográficas dentro de EcoCyc se obtuvieron de PubMed, mientras que la importación de proteínas en EcoCyc se realiza a través de las funciones de UniProt.

EcoCyc es una fuente de referencia electrónica para biólogos especialistas en *E. coli* o que trabajan con microorganismos relacionados.

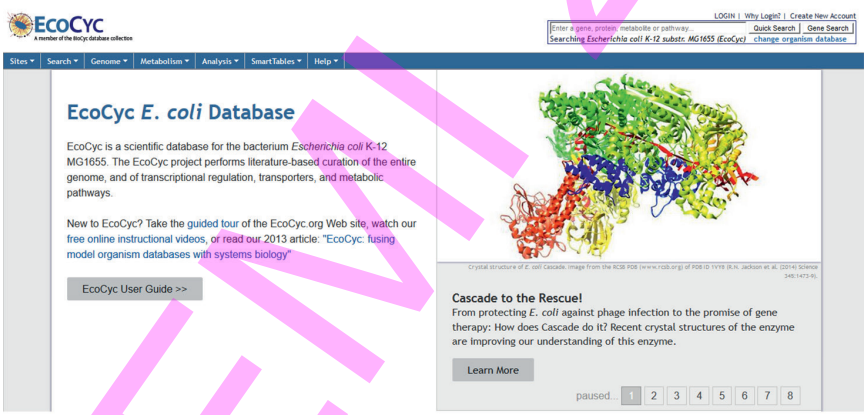


Figura 38. Portal del sitio web de EcoCyc.

2.4.1. Ejemplos de entrada en sitios relacionados con EcoCyc

1. Ingresar a https://ecocyc.org/

2. Ingresar a otros sitios relacionados:

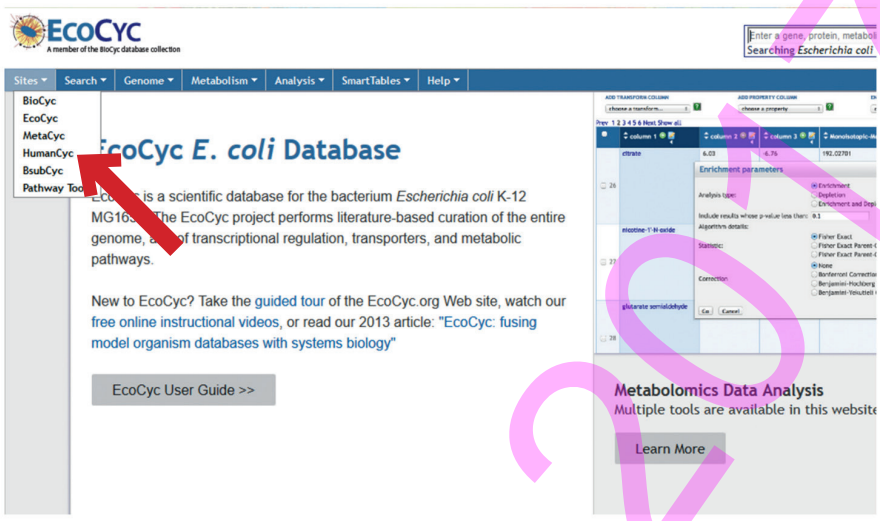


Figura 39. Sitios web desde Ecocyc.

- **BioCyc:** Proporciona genomas de referencia y rutas metabólicas (relacionadas con síntesis y degradación de moléculas) de organismos .

- **Metacyc:** Contiene información sobre estructura de enzimas (moléculas proteicas que intervienen en el metabolismo) y sus datos pueden usarse como referencia para la predicción computacional de rutas metabólicas de organismos a partir de sus genomas.

- **Humancyc:** Proporciona referencias sobre las vías metabólicas humanas y facilita diagramas del mapa metabólico humano para generar modelos cuantitativos en estado estacionario del metabolismo humano.

- **BsubCyc:** Es una base de datos de la bacteria modelo (desde el punto de vista del estudio científico) *Bacillus subtilis.*

2.4.2. Ejemplo de aplicación: Acceso a datos del NCBI a través de BioCyc

1. Ingresar a https://biocyc.org/

2. Escoger *Tier 2 and Tier 3 databases*

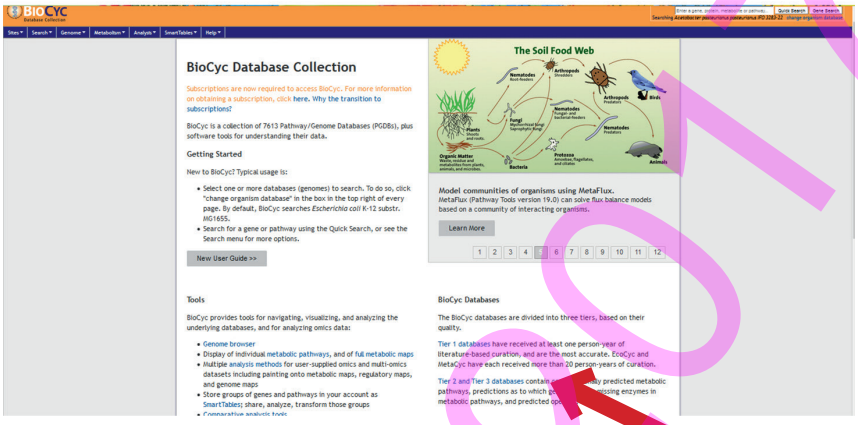


Figura 40. Base de datos de BioCyc.

3. Ingresar a la base de datos de *Escherichia coli* B str. REL606

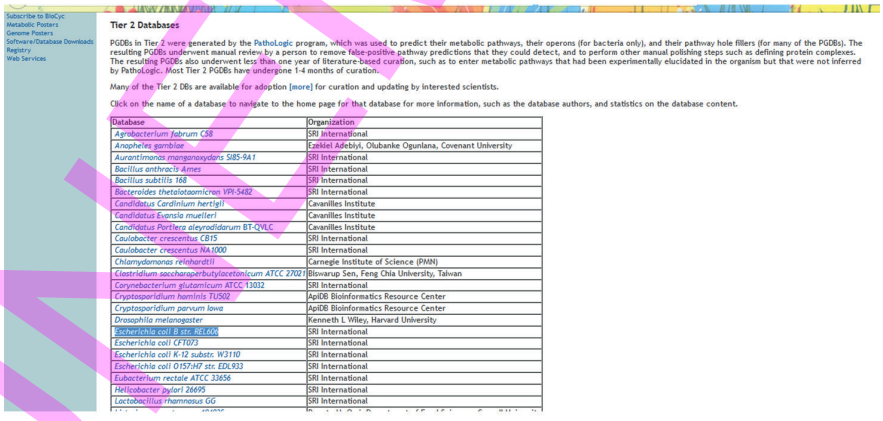


Figura 41. Base de datos de BioCyc.

4. Observar la información de *Escherichia coli* y escoger la referencia del NCBI.

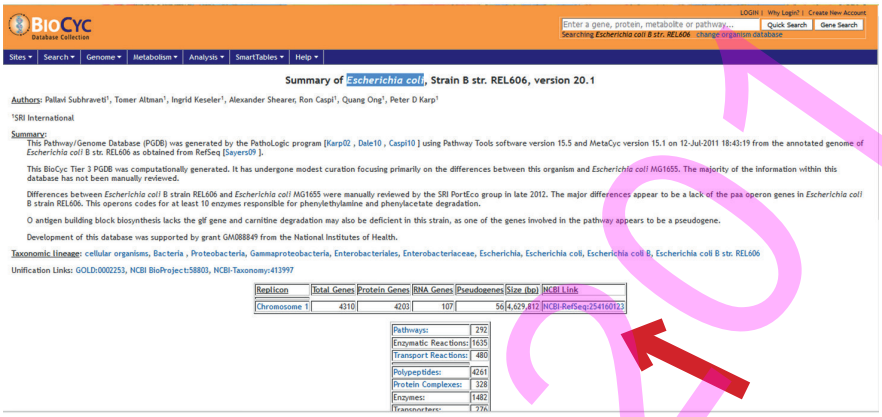


Figura 42. Información de Escherichia coli Strain B str. REL606, versión 20.1 del sitio web BioCyc.

5. Visualizar el genoma completo de *Escherichia coli* B str. REL606 en NCBI en la secuencia NC_012967.1.

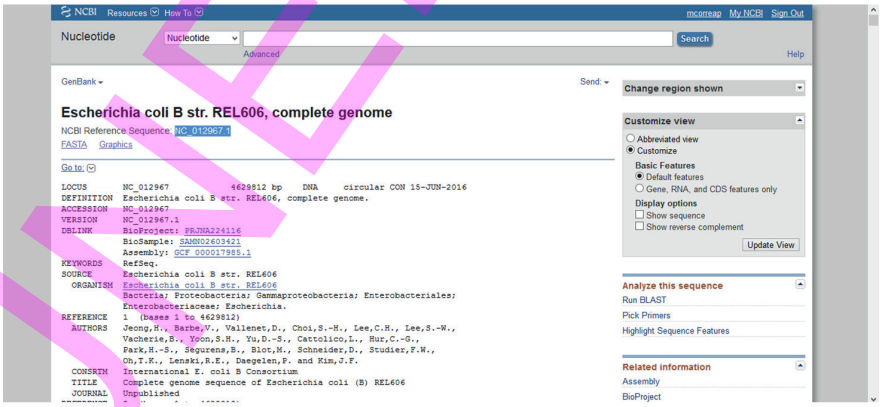


Figura 43. Genoma de Escherichia coli B str. REL606 del NCBI.

6. Retornar a la información del *Escherichia coli,* Strain B str. REL606, versión 20.1 del sitio web BioCyc y escoger Chromosome 1.

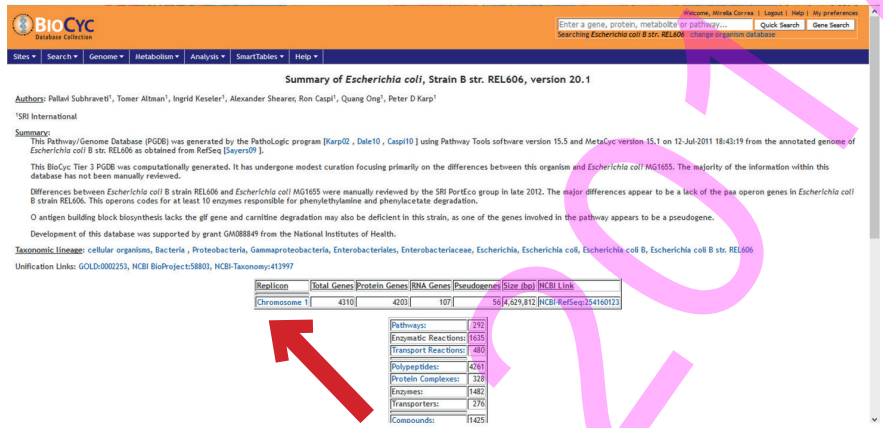


Figura 44. Información de Escherichia coli Strain B str. REL606, version 20.1 del sitio web BioCyc.

7. Observar la sede manera gráfica la secuencia del cromosoma seleccionado.

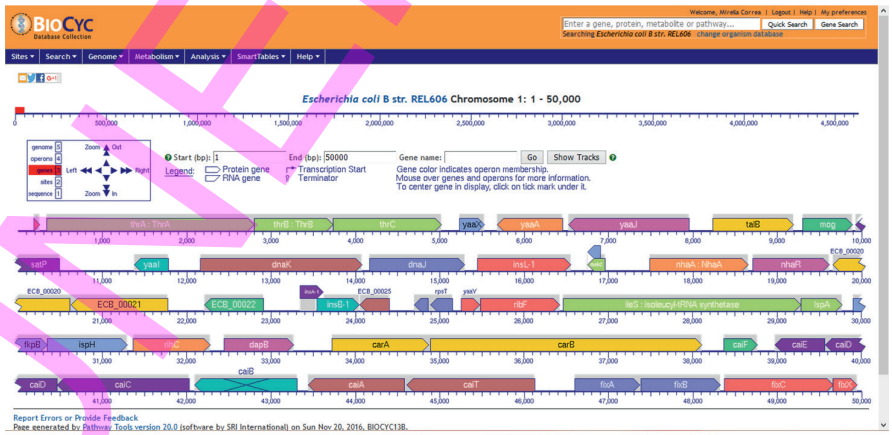


Figura 45. Secuencia en forma gráfica de un cromosoma de una cepa de Escherichia coli de manera gráfica.

8. Obtener la ubicación del gen *ThrA* desde la secuencia 336-2798 pares de bases. Al presionar obtendrá información del diagrama del gen.

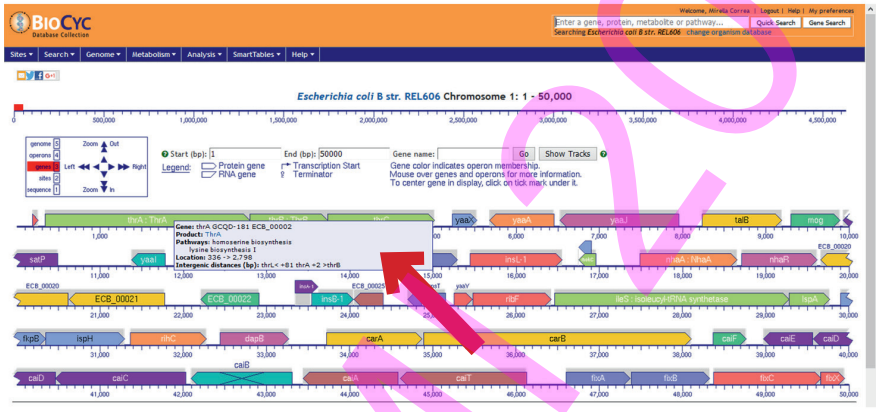


Figura 46. Ubicación del gen ThrA en el cromosoma bacteriano de Escherichia coli B str. REL606, localizado en la secuencia 336-2798 pares de bases.

2.4.3. Ejemplo de aplicación: Análisis comparativo de genomas en EcoCyc

1. Ingresar a https://ecocyc.org/, que permite calcular las estadísticas de una sola base de datos o comparaciones entre múltiples bases de datos.

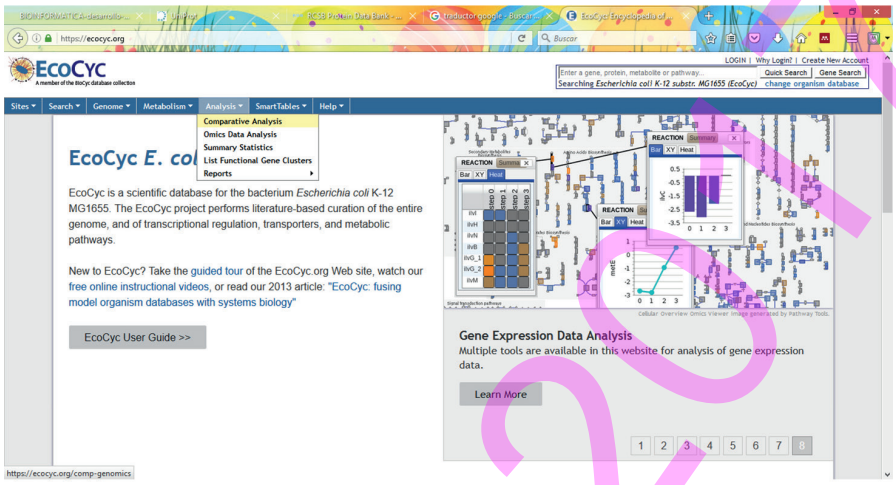


Figura 47. Análisis comparativo en EcoCyc.

2. Seleccionar los organismos a comparar. En este ejemplo se utilizan *Escherichia coli* 042 *y Escherichia coli* 07798. Presionar *OK* y *Submit.*

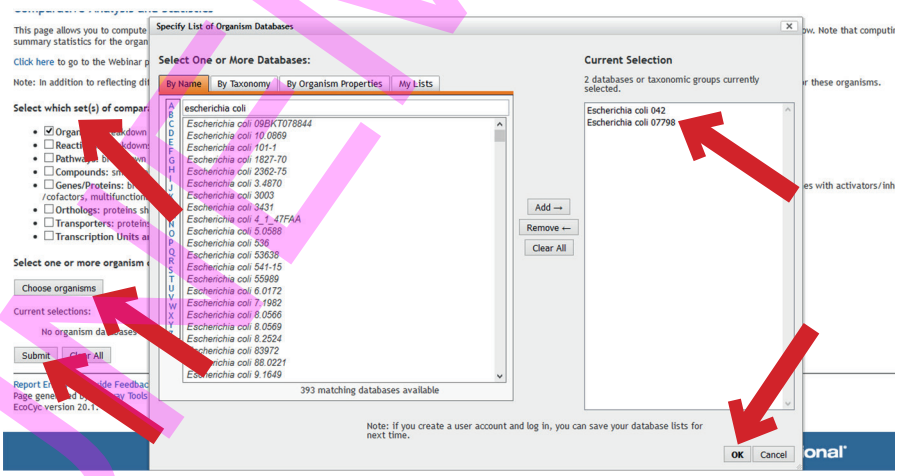


Figura 48. Comparación de organismos en EcoCyc.

3. Obtener información de la comparación.

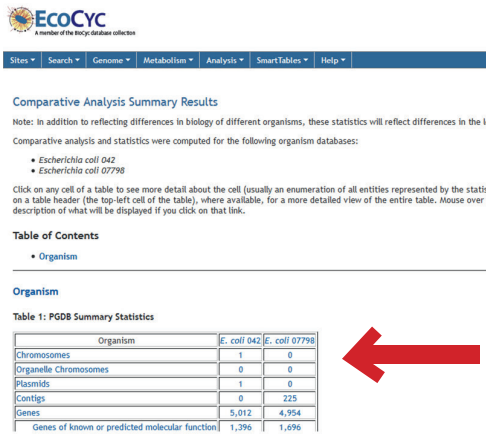


Figura 49. Cuadro de comparación entre Escherichia coli 042 y Escherichia coli 07798.

2.5. UNIPROT

UniProt (Universal Protein; http://www.uniprot.org) es el repositorio central de datos sobre proteínas de la combinación de bases Swiss-Prot, TrEMBL y PIRt.

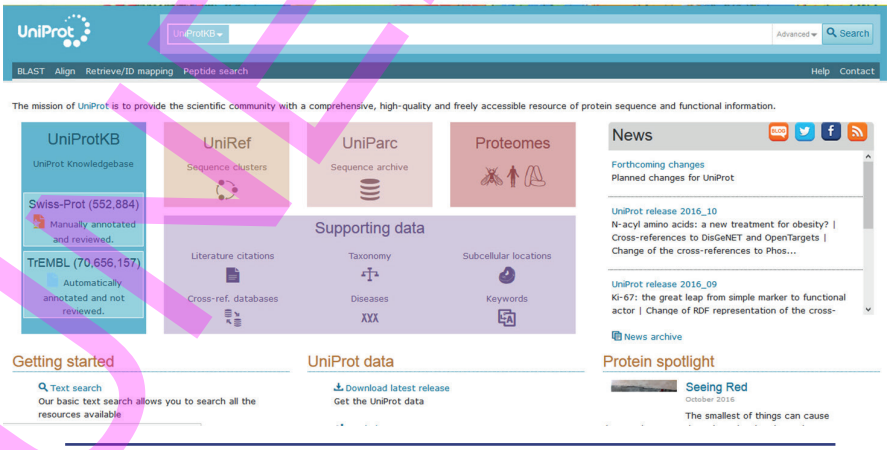


Figura 50. Sitio web de UniProt.

2.5.1. Búsqueda de proteínas en organismos en UniProt

1. Consultar proteínas de *Homo Sapiens.*

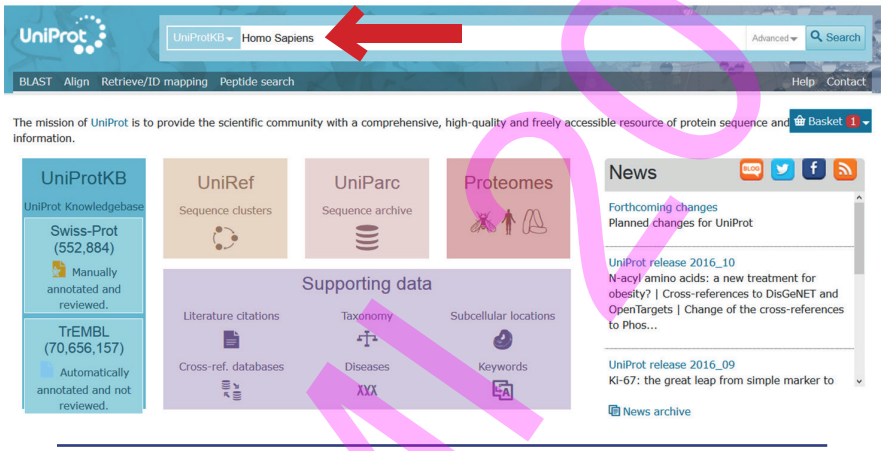
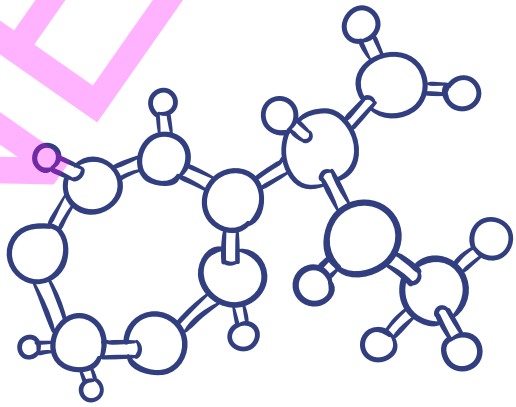


Figura 51. Consultar sobre el Homo Sapiens en UniProt.

2. Seleccionar dos proteínas que se quieran estudiar o comparar. En este ejemplo se selecciona A0A1B1R4L5 y A0A1B1R4L8 añadiendo a la cesta.



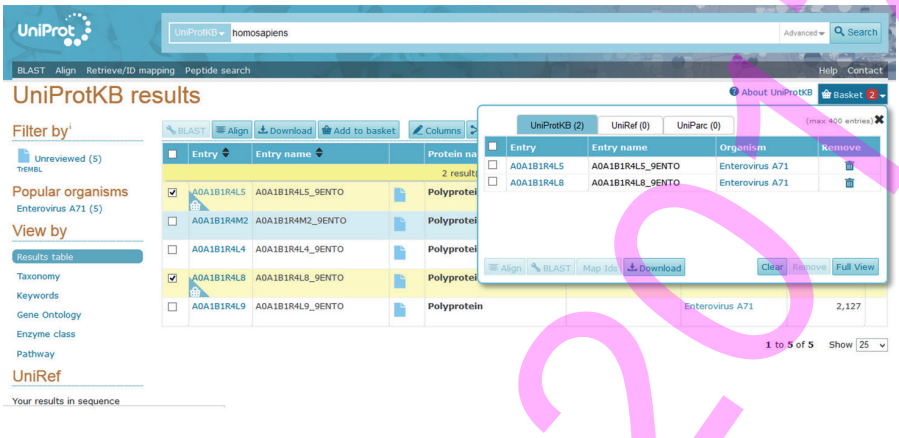


Figura 52. Selección de dos proteínas en UniProt.

3. Las proteínas seleccionadas pueden alinearse para ver similitudes o diferencias (Ver Sección 3.3. Alineamiento de secuencias), o ser sometidas a BLAST para buscar proteínas relacionadas (Ver Sección 3.4. Búsqueda de secuencias: BLAST).

4. Descargar los resultados.

2.5.2. Comparar archivos obtenidos desde Uniprot y NCBI

1. Ingresar a *Proteínas del NCBI* y consultar *Homo Sapiens* RAD23A (que es una proteína), y obtener el archivo FASTA (Ver Sección 3.2.1. Formato FASTA).

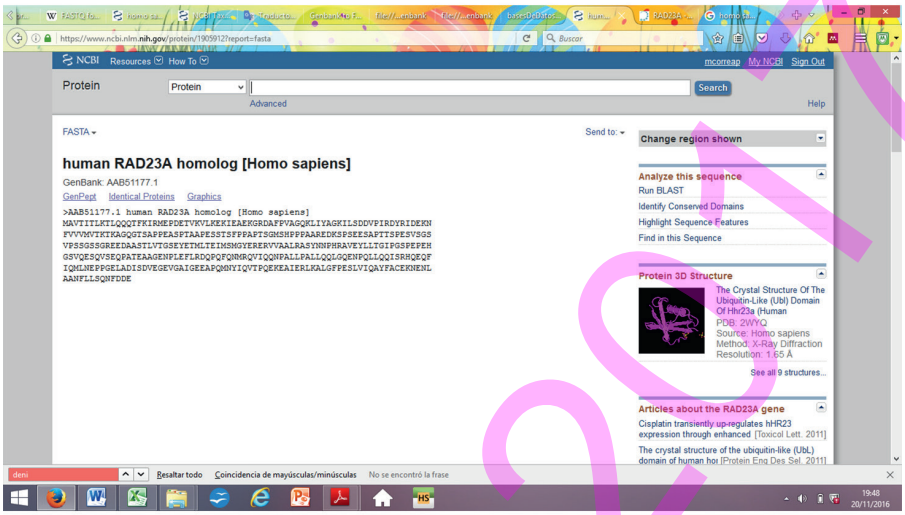


Figura 53. Proteína del Homo Sapiens RAD23A en NCBI.

2. Ingresar a UniProt y consultar *Homo Sapiens.* Descargar el archivo en formato FASTA.

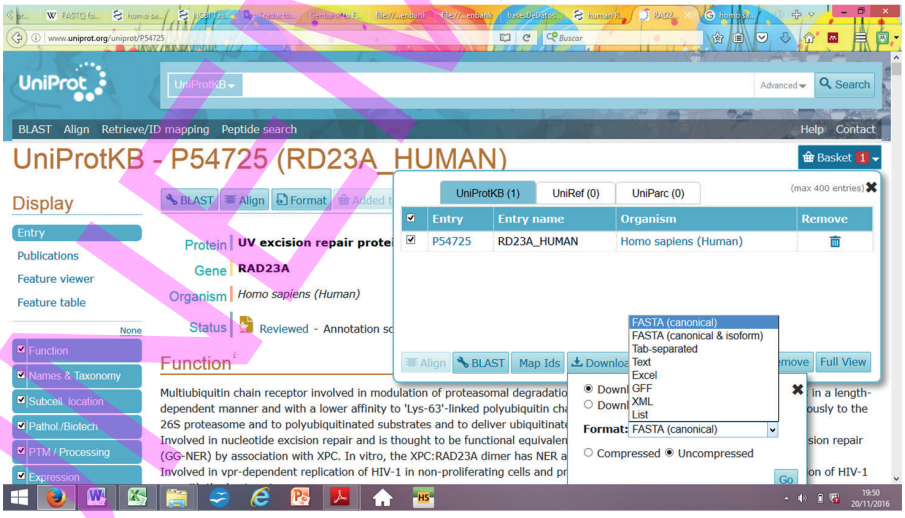
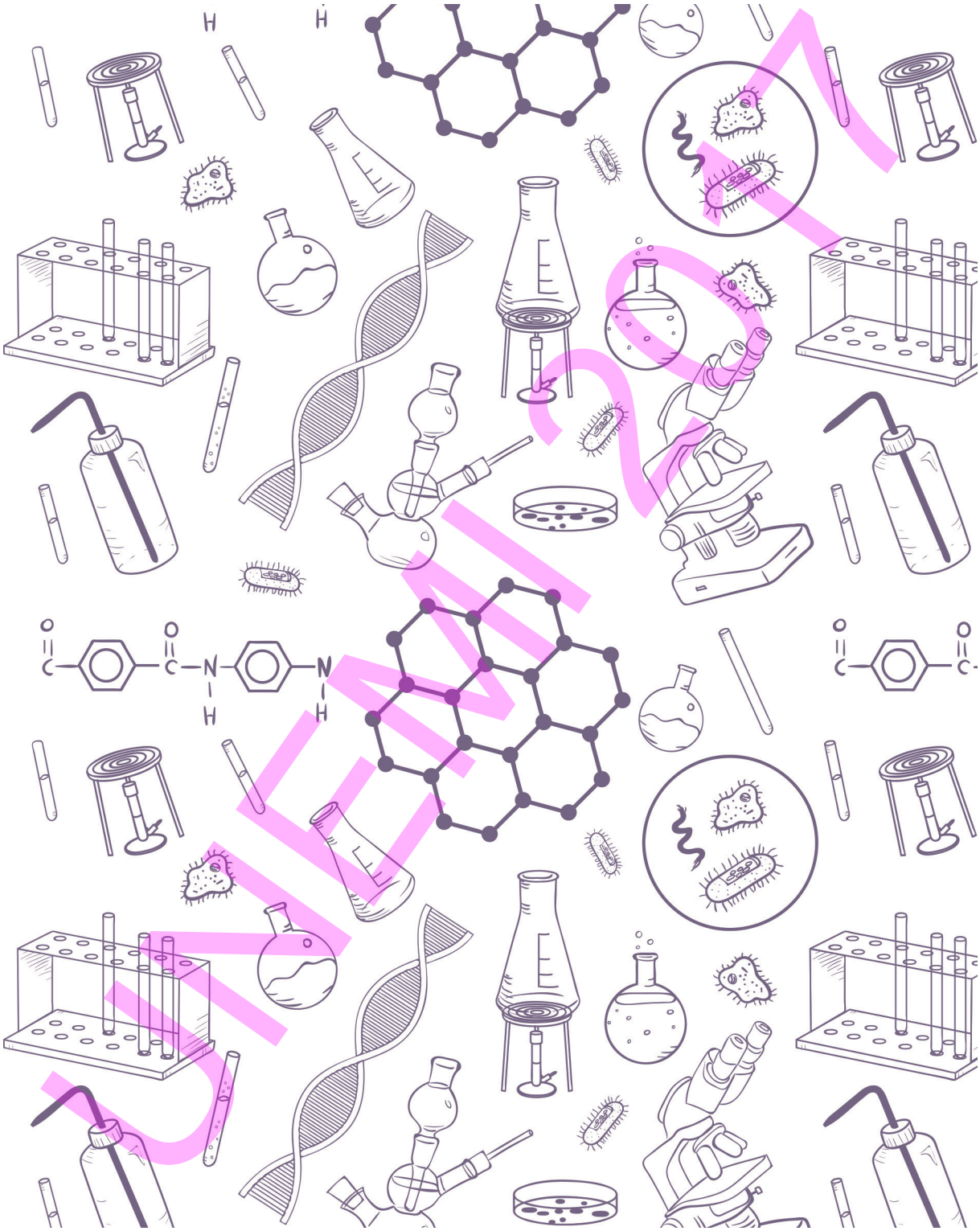
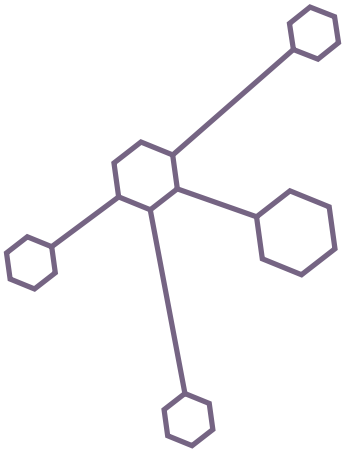
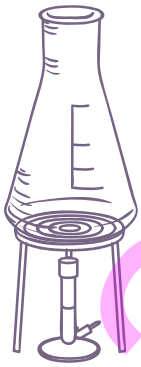


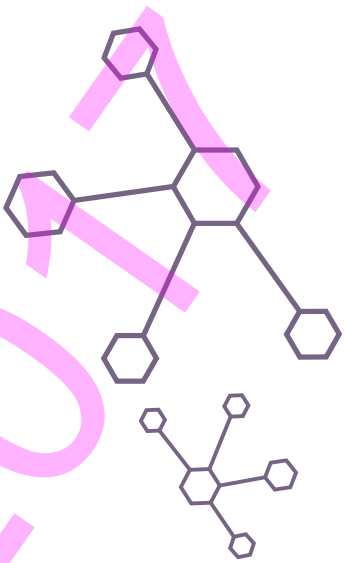
Figura 54. Proteína del Homo Sapiens en Uniprot.

3. Comparar los dos archivos: ambos tienen el mismo contenido.





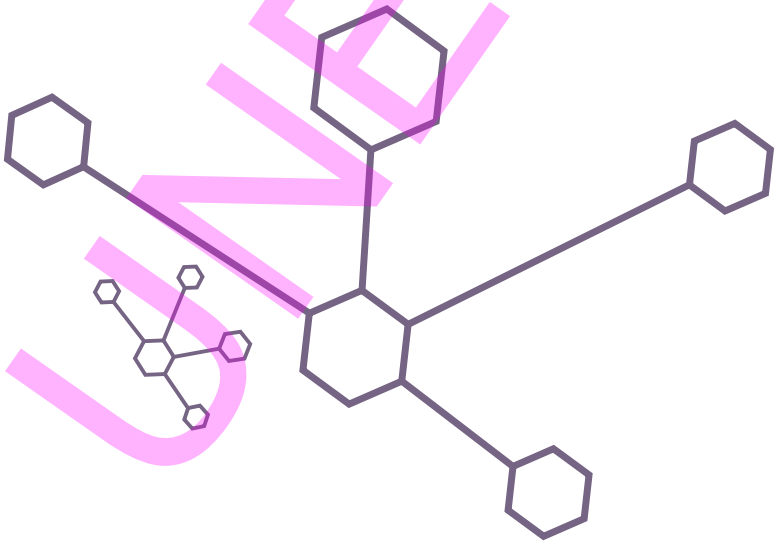


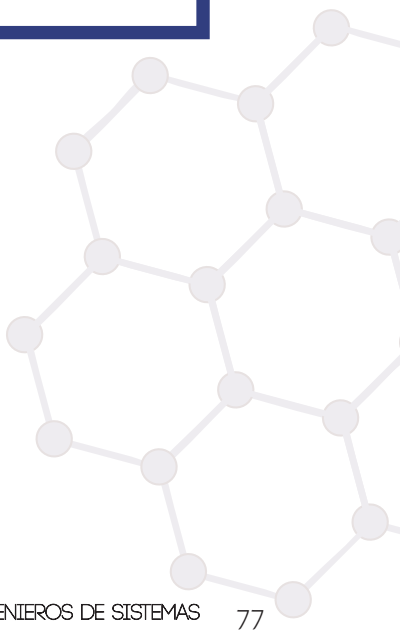


CAPÍTULO 3:

OBTENCIÓN Y TRATAMIENTO BIOINFORMÁTICO DE SECUENCIAS

Ing. Mirella Correa-Peralta, MBA.
Lic. Carlos Noceda-Alonso, PhD.
Ing. Oscar León-Granizo.
Lic. Jesennia Cárdenas-Cobo, MBA.





3. OBTENCIÓN Y TRATAMIENTO BIOINFORMÁTICO DE SECUENCIAS

3.1. OBTENCIÓN DE SECUENCIAS: FUNDAMENTOS BÁSICOS DE SECUENCIACIÓN DE PRÓXIMA GENERACIÓN

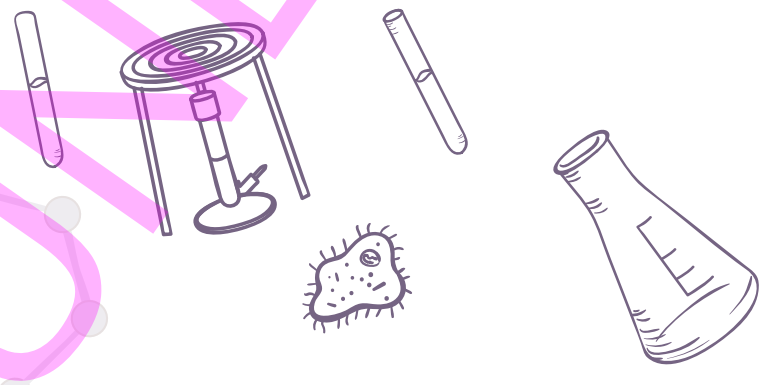
La secuenciación de ácidos nucleicos determina su secuencia de nucleótidos. Las primeras técnicas de secuenciación se realizaron sobre ADN, a partir de dos métodos:

- Método de Maxam-Gilbert (1976).
- Método de terminación de cadena de Sanger (1977), más sencillo.

Estos métodos implican acortar las secuencias problema a pequeños fragmentos que luego se separan en función de su tamaño, para finalmente determinar por algún procedimiento bioquímico cuál es el nucleótido terminal de cada uno de ellos.

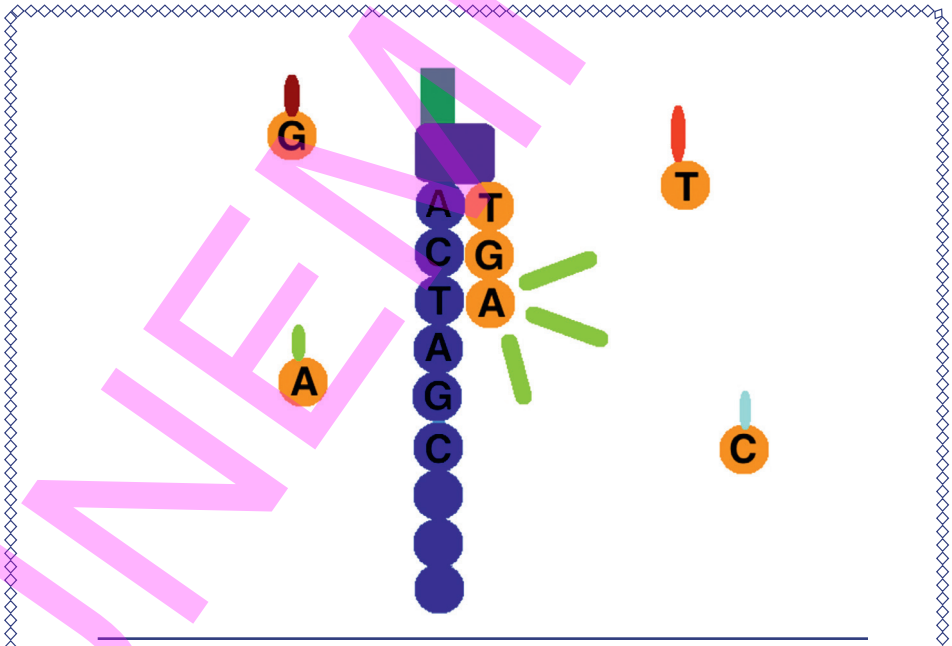
Las tecnologías denominadas de Secuenciación de Próxima Generación (*Next Generation Sequencing,* NGS) más en boga hasta la actualidad han utilizado estrategias basadas en:

1. **Secuenciación por síntesis**, es decir, en determinación en tiempo real de los nucleótidos que se van polimerizando (añadiendo) en una reacción de replicación de ADN. Las tecnologías NGS más utilizadas hasta la actualidad basadas en esta estrategia son:



1.1. Roche 454, que se basa en la detección de luz liberada en una reacción química que ocurre tras la incorporación de un nucleótido concreto a la hebra que se está sintetizando, entre los cuatro que se ensayan consecutivamente. Esta técnica se denomina **pirosecuenciación**, y está últimamente en desuso.

1.2. Illumina, que se basa en la detección de fluorescencia de un determinado color, asociado al nucleótido que se incorpora a la hebra en síntesis.



Fuente: wikimedia.org[15]

Figura 55. Estrategia de detección de nucleótidos de Illumina de BioSystem.

2. **Secuenciación por ligación** (SOLid *sequencing*), que se basa en la determinación en tiempo real de dobletes de nucleótidos que se complementan con dobletes de nucleótidos de una hebra molde de DNA, y que a continuación se ligan a la hebra *naciente*. La detección también implica el uso de florescencia de distintas longitudes de onda.

[Para una actualización de los distintos métodos de secuenciación se puede consultar la revisión de Goodwin et al (2016).]

Con las tecnologías NGS pueden analizarse desde fragmentos simples de ácidos nucleicos a genomas o transcriptomas (conjunto de transcritos de un sistema) completos. Para secuenciar el RNA, comúnmente éste se **retrotranscribe** a ADN. Las técnicas más utilizadas en la actualidad, como las mencioandas, implican trocear el ADN en fragmentos pequeños, de decenas a cientos de pares de nucleótidos. Así, la técnica registra a nucleótido la composición de cada fragmento, rindiendo una secuencia final para cada fragmento. Cuantas más veces se lee un fragmento, menor posibilidad de error. La información de secuencia y calidad se almacena en archivos FastQ (ver Sección 3.2.3. Formato FastQ).

Cuando se analizan varias muestras al mismo tiempo, estas se preparan previamente por separado añadiendo a los extremos de los fragmentos unos marcadores, que son secuencias de nucleótidos específicas para cada muestra, y tras la secuenciación los fragmentos son clasificados bioinformáticamente.

Un gran reto bioinformático consiste en **ensamblar** los fragmentos leídos (armar el rompecabezas del genoma completo a partir de los pequeños fragmentos secuenciados) de un genoma que se ha de construir *de novo*, es decir, que nunca se ha secuenciado. Esto se hace más complejo al considerar que existen grandes regiones del genoma que se repiten, incluso en distintos cromosomas. Además, muchas veces faltan piezas, es decir, fragmentos que no se han secuenciado o de calidad baja. Un nivel de dificultad aún mayor se añade en proyectos de **metagenómica**, que implica la secuenciación de varios genomas mezclados, generalmente de grupos de microorganismos colectados en un mismo lugar (suelo, intestino, etc.).

Tarea más sencilla es la **resecuenciación,** es decir, ubicar los fragmentos secuenciados de una muestra genómica sobre un genoma similar (de la misma especie, por ejemplo) ya secuenciado.

Cuando hay que analizar a nivel genómico o epigenómico un gran número de individuos, los recursos pueden no alcanzar para analizar todos los genomas enteros. En estos casos, puede recurrirse a técnicas que implican reducción de la complejidad genómica, o que analizan regiones concretas a lo largo de todo el genoma.

Una de estas técnicas es la secuenciación de ADN asociado a sitios e restricción (*Restriction Site Associated DNA sequencing,* **RAD-seq** [Miller et al 2007]), que se basa en partir el ADN (**restricción**), efectuando el corte en el interior de determinadas secuencias de nucleótidos cortas, y después secuenciar aproximadamente un centenar de nucleótidos a partir de los extremos de los fragmentos resultantes. Así, las secuencias obtenidas, que pueden pertenecer a multitud de individuos, se pueden alinear entre sí en base a su similitud sin necesidad de mapearlas sobre un genoma de referencia (que con frecuencia no existe). Estos alineamientos se pueden usar para estudios poblacionales o filogenéticos (evolutivos). El investigador debe establecer determinados parámetros de este tipo de técnica, llegando a un compromiso entre los siguientes factores: longitud del total de regiones del genoma que han de ser secuenciadas (como es obvio, dependiente la cantidad de fragmentos que se obtengan con la restricción), número de veces que se lee (que se secuencia) un mismo fragmento (más lecturas reducen las posibilidades de error) y recursos económicos.

La complejidad epigenómica también puede reducirse con técnicas como la secuenciación que, tras la aplicación de bisulfito (Ver Sección 1.4. Epigenética: un reciente paradigma en Biología), se realiza sobre una representación reducida del genoma (*Reduced Representation Bisulfite Sequencing*, **RRBS**, [Meissner et al 2005]), y que se centra en regiones genómicas que se sabe están muy metiladas.

3.2. FORMATOS DE ARCHIVOS DE SECUENCIAS NUCLEOTÍDICAS Y AMINOACÍDICAS

Las secuencias nucleótidicas y aminoacídicas se presentan en diversos tipos de formatos que se utilizan en bases de datos y programas bioinformáticos, tanto para entradas (*input*) como para salidas (*output*), empleándose comúnmente ficheros de texto. A continuación se clasifican distintos tipos de formato para secuencias nucleotídicas y/o aminoacídicas en función del uso que se le quiera dar a la secuencia:

- Multipropósito: FASTA, GenBank, XML, Nexus
- Alineamiento y filogenia: Phylip, phyloXML
- Next Generation Sequencing: FastQ, SRA
- Estructuras: PDB

3.2.1. Formato FASTA

Es el formato común en la secuencia del ADN, ARN y proteínas, y utilizado en los ejemplos del Capítulo 2. Es un formato de texto (los archivos tienen extensión .fasta, .fa, .fna, .fnn, .faa):

- **Las secuencias de ácidos nucleicos** se presentan, como es habitual, en 4 caracteres (uno para cada nucleótido): A, C, G, y T (en el caso del ADN o ARN) o U (en muchos casos para el ARN). En casos de desconocimiento preciso del nucleótido se aplican otras letras.

CÓDIGO	BASE NITROGENADA DEL NUCLEÓTIDO
A	Adenina
C	Citosina
G	Guanina
T	Timina
U	Uracilo
R	G A (pu**R**ínica)
Y	T C (pirimidínica/*p**Y**rimidinic*)
K	G T (cetona/***Ketone***)
M	A C (grupo a**M**ino)
S	G C (interacción fuerte/***S**trong interaction*)
W	A T (interacción débil/***W**eak interaction*)
B	G T C (no A) (**B** viene tras la A)
D	G A T (no C) (**D** viene tras la C)
H	A C T (no G) (**H** viene tras la G)
V	G C A (no T, no U) (**V** viene tras la U)
N	A G C T (cualquiera/*a**N**y*)
X	máscara
-	hueco (*gap*) de longitud indeterminada

Figura 56. Símbolos convencionales para nucleótidos.

- **Una secuencia de polipéptido** se presenta básicamente a partir de 20 caracteres (letras mayúsculas, cada una correspondiente a un aminoácido distinto).

  Un archivo FASTA comienza con el signo mayor que (>), seguido de una descripción, un cambio de línea y finalmente la secuencia con 80 caracteres por línea.

```
>L11285.1 Homosapiens ERK activator kinase (MEK2) mRNA
GAATTCGAGCCGACCGACCGCTCCCGGCCCGCCCCCTATGGGCCCCGGCTAGAGGCGCCGCCGCCGCCGG
CCCGCGGAGCCCCGATGCTGGCCCGGAGGAAGCCGGTGCTGCCGGCGCTCACCATCAACCCTACCATCGC
CGAGGGCCCATCCCCTACCAGCGAGGGCGCCTCCGAGGCAAACCTGGTGGACCTGCAGAAGAAGCTGGAG
GAGCTGGAACTTGACGAGCAGCAGAAGAAGCGGCTGGAAGCCTTTCTCACCCAGAAAGCCAAGGTTGGCG
AACTCAAAGACGATGACTTCGAAAGGATCTCAGAGCTGGGCGCGGGCAACGGCGGGGTGGTCACCAAAGT
CCAGCACAGACCCTCGGGCCTCATCATGGCCAGGAAGCTGATCCACCTTGAGATCAAGCCGGCCATCCGG
AACCAGATCATCCGCGAGCTGCAGGTCCTGCACGAATGCAACTCGCCGTACATCGTGGGCTTCTACGGGG
CCTTCTACAGTGACGGGGAGATCAGCATTTGCATGGAACACATGGACGGCGGCTCCCTGGACCAGGTGCT
GAAAGAGGCCAAGAGGATTCCCGAGGAGATCCTGGGGAAAGTCAGCATCGCGGTTCTCCGGGGCTTGGCG
TACCTCCGAGAGAAGCACCAGATCATGCACCGAGATGTGAAGCCCTCCAACATCCTCGTGAACTCTAGAG
GGGAGATCAAGCTGTGTGACTTCGGGGTGAGCGGCCAGCTCATAGACTCCATGGCCAACTCCTTCGTGGG
CACGCGCTCCTACATGGCTCCGGAGCGGTTGCAGGGCACACATTACTCGGTGCAGTCGGACATCTGGAGC
ATGGGCCTGTCCCTGGTGGAGCTGGCCGTCGGAAGGTACCCCATCCCCCCGCCCGACGCCAAAGAGCTGG
AGGCCATCTTTGGCCGGCCCCGTGGTCGACGGGGAAGAAGGAGAGCCTCACAGCATCTCGCCTCGGCCGAG
GCCCCCCGGGCGCCCCGTCAGCGGTCACGGGATGGATAGCCGGCCTGCCATGGCCATCTTTGAACTCCTG
GACTATATTGTGAACGAGCCACCTCCTAAGCTGCCCAACGGTGTGTTCACCCCCGACTTCCAGGAGTTTG
TCAATAAATGCCTCATCAAGAACCCAGCGGAGCGGGCGGACCTGAAGATGCTCACAAACCACACCTTCAT
CAAGCGGTCCGAGGTGGAAGAAGTGGATTTTGCCGGCTGGTTGTGTAAAACCCTGCGGCTGAACCAGCCC
GGCACACCCACGCGCACCGCCGTGTGACAGTGGCCGGGCTCCCTGCGTCCCGCTGGTGACCTGCCCACCG
TCCCTGTCCATGCCCCGCCCTTCCAGCTGAGGACACGTGGCGCCTCCACCCACCCTCCTGCCTCACCCTG
CGGAGAGCACCGTGGCGGGGCGACAGCGCATGCAGGAACGGGGGTCTCCTCTCCTGCCAGTCCTGGCCGG
GGTGCCTCTGGGGACGGGCGACGCTGCTGTGTGGTCTCAGAGGCTCTGCTTCCTTAGGTTACAAAACA
AAACAGGGAGAGAAAAGCAAAAAAAAAAAAAAAAAA
```

Figura 57. Ejemplo de archivo en formato FASTA.

3.2.1.1. Obtener un fichero en formato FASTA

1. Ingresar a NCBI en la base de datos de nucleótidos y escribir *Homo Sapiens chromosome 7*.

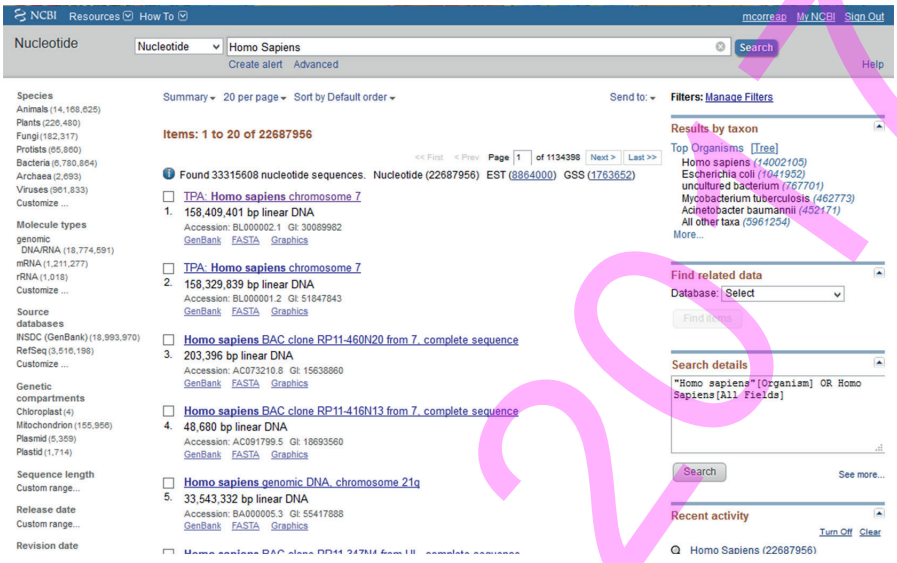


Figura 58. Consulta en NCBI en nucleótidos el organismo Homo Sapiens, y obtener una secuencia nucleotídica en formato FASTA.

2. Hacer clic en el primer nombre de la consulta de *Homo Sapiens, cromosoma 7* y escoger *FASTA*.

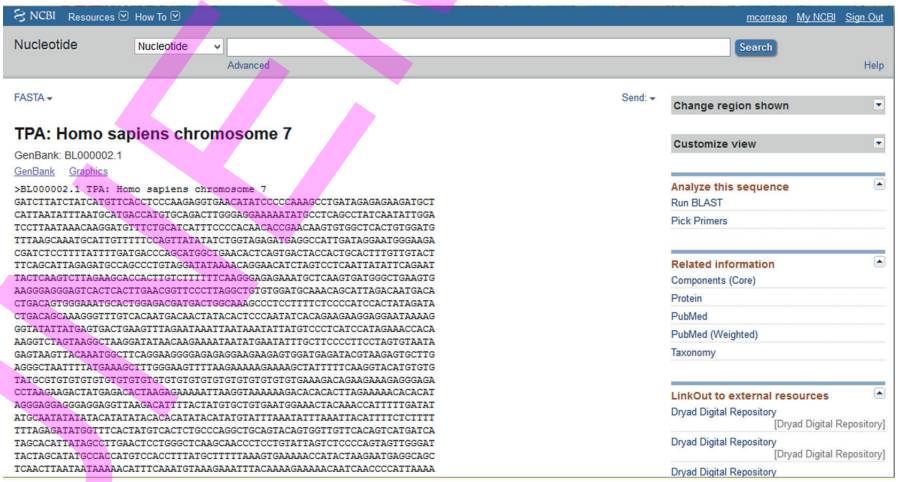


Figura 59. Secuencia nucleotídica del cromosoma 7 humano en formato FASTA en NCBI.

3.2.2. Formato GenBank

El formato GenBank tiene más información que el FASTA. Se presenta en forma de secuencias con 60 caracteres por línea, divididos en grupos de 10. Las letras representando los nucleótidos equivalen a los del formato FASTA, pero en minúsculas. El archivo tiene extensión .gb.

```
LOCUS       HUMMEK2NF               1576 bp    mRNA    linear   PRI 26-JUL-1993
DEFINITION  Homosapiens ERK activator kinase (MEK2) mRNA.
ACCESSION   L11285
VERSION     L11285.1
KEYWORDS    ERK activator kinase; MEK kinase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1576)
  AUTHORS   Zheng,C.F. and Guan,K.L.
  TITLE     Cloning and characterization of two distinct human extracellular
            signal-regulated kinase activator kinases, MEK1 and MEK2
  JOURNAL   J. Biol. Chem. 268 (15), 11435-11439 (1993)
   PUBMED   8388392
COMMENT     Original source text: Homo sapiens cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..1576
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
ORIGIN
        1 gaattcgagc cgaccgaccg ctccccgccc gccccctatg ggccccggct agaggcgccg
       61 ccgccgccgg cccgcggagc cccgatgctg gcccggagga agccggtgct gccggcgctc
      121 accatcaacc ctaccatcgc cgagggccca tcccctacca gcgagggcgc ctccgaggca
      181 aacctggtgg acctgcagaa gaagctggag gagctggaac ttgacgagca gcagaagaag
      241 cggctggaag cctttctcac ccagaaagcc aaggttggcg aactcaaaga cgatgacttc
      301 gaaaggatct cagagctggg cgcgggcaac ggcgggggtgg tcaccaaagt ccagcacaga
      361 ccctcgggcc tcatcatggc caggaagctg atccaccttg agatcaagcc ggccatccgg
      421 aaccagatca tccgcgagct gcaggtcctg cacgaatgca actcgccgta catcgtgggc
      481 ttctacgggg ccttctacag tgacggggag atcagcattt gcatggaaca catggacggc
      541 ggctccctgg accaggtgct gaaagaggcc aagaggattc ccgaggagat cctggggaaa
      601 gtcagcatcg cggttctccg gggcttggcg tacctccgag agaagcacca gatcatgcac
      661 cgagatgtga agccctccaa catcctcgtg aactctagag gggagatcaa gctgtgtgac
      721 ttcggggtga gcggccagct catagactcc atggccaact ccttcgtggg cacgcgctcc
      781 tacatggctc cggagcggtt gcagggcaca cattactcgg tgcagtcgga catctggagc
      841 atgggcctgt ccctggtgga gctggccgtc ggaaggtacc ccatcccccc gcccgacgcc
      901 aaagagctgg aggccatctt tggccggccc gtggtcgacg gggaagaagg agagcctcac
      961 agcatctcgc ctcggccgag gccccccggg cgccccgtca gcggtcacgg gatggatagc
     1021 cggcctgcca tggccatctt tgaactcctg gactatattg tgaacgagcc acctcctaag
     1081 ctgcccaacg gtgtgttcac ccccgacttc caggagtttg tcaataaatg cctcatcaag
     1141 aacccagcgg agcgggcgga cctgaagatg ctcacaaacc acaccttcat caagcggtcc
```

Figura 60. Ejemplo de archivo en formato FASTA.

GenBank, accesible a través del NCBI, facilita las secuencias de nucleótidos intercambiando datos con el EMBL y DDBJ, obteniendo una cobertura mundial.

3.2.3. Formato FastQ

Es un formato diseñado para almacenar secuencias incluyendo un parámetro de la calidad de secuenciación de cada nucleótido (es decir, indicando un valor de probabilidad de que sea correcta o no una base indicada). Presenta extensión .fastq. Estos archivos contienen:

- Cabecera, que empieza con @
- Secuencia de nucleótidos.
- "+" , opcionalmente seguido de alguna descripción.
- Valores de calidad de secuencia (para cada nucleótido) en código ASCII .

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Figura 61. Ejemplo de estructura de archivo FastQ.

3.2.3.1. Evaluación de calidad

El control de calidad de una secuenciación ayuda a decidir qué hacer con la secuencia obtenida. Para ello, han de filtrarse los datos según los siguientes criterios:

- Tamaño pequeño de las secuencias.
- Límite de calidad para cada nucleótido.
- Descarte de adaptadores.

Existen diversas herramientas para ello, como:

- Fastx-toolkit
- GalaxySeqTK
- Cutadapt

El **SRA** *(Sequence Read Archive)* pone a disposición los datos de secuencias biológicas para permitir nuevos descubrimientos a partir de la comparación de conjuntos de datos. El SRA almacena datos de secuenciación sin procesar e información de alineamientos con datos obtenidos en plataformas de secuenciación de alto rendimiento, incluyendo Roche 454 o Illumina.

3.2.3.2. Ejemplo de análisis de calidad de secuencia en la plataforma bioinformática Galaxy

1. Ingresar al SRA (https://www.ncbi.nlm.nih.gov/sra)

2. Descargar / copiar datos de la secuencia nucleotídica SRR030252

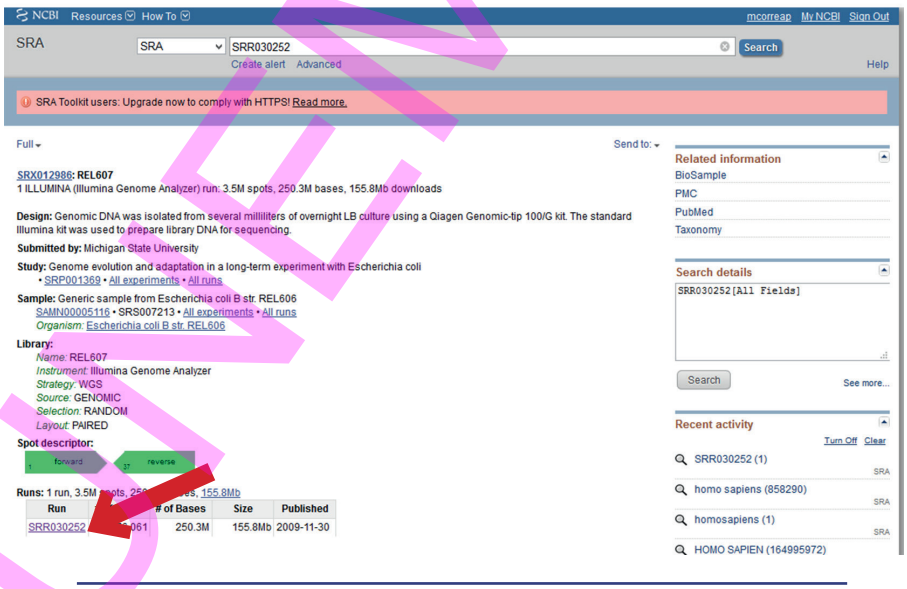


Figura 62. Base de SRA en NCBI.

3. Ingresar a Galaxy.org (https://usegalaxy.org/) y cargar el archivo descargado de SRA .

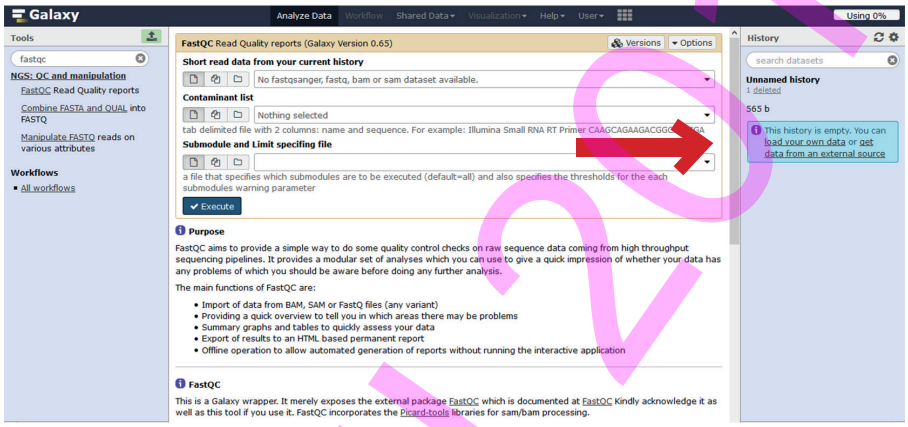


Figura 63. Sitio web de Galaxy.org.

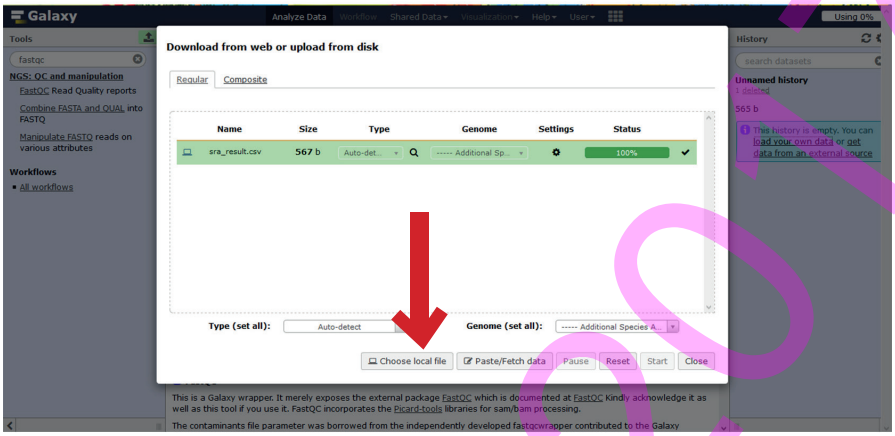


Figura 64. Selección de archivo en Galaxy.org.

4. Verificar que el formato esté en FastQ y seleccionar el archivo resultante de la carga.

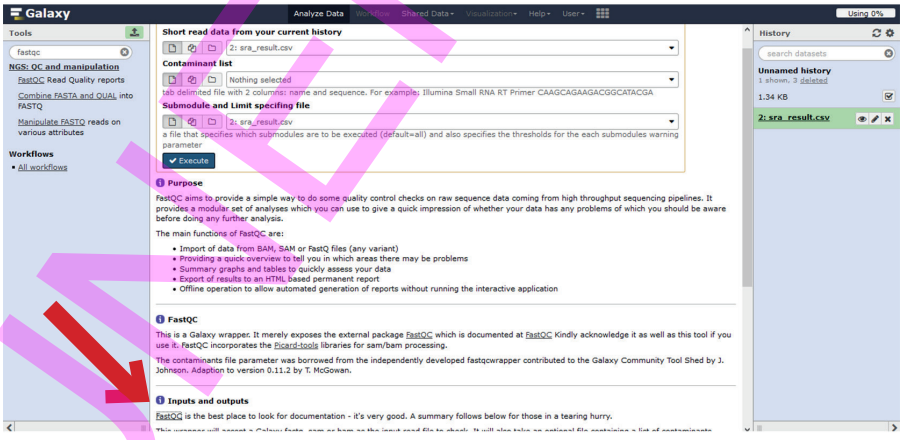


Figura 65. Carga del archivo en galaxy.org.

5. Observar el nivel de calidad.

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

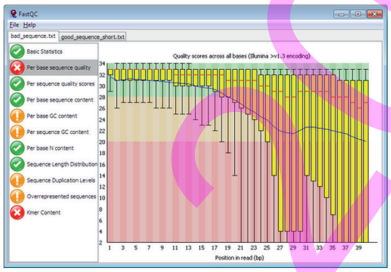


Figura 66. Calidad en FastQ en Galaxy.org.

3.2.3.3. Ejemplo de análisis de calidad de secuencia con la herramienta FastQC

1. Obtener un archivo descargado con extensión FastQC.

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

2. Ingresar a la dirección
www.bioinformatics.Babraham.ac.uk/projects

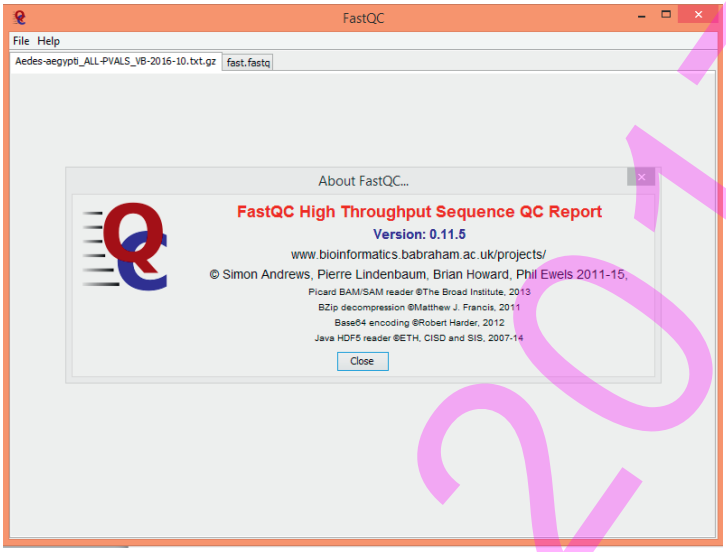


Figura 67. FastQC Software para análisis de calidad de secuencia.

3. Abrir el archivo FastQC.

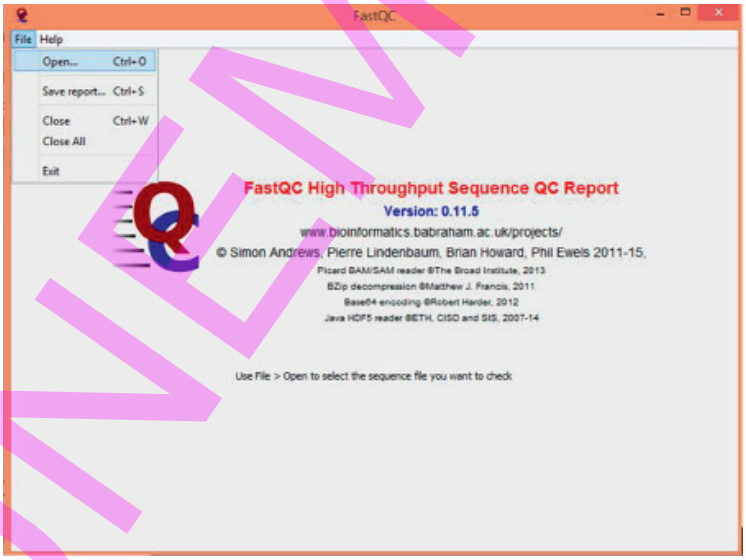


Figura 68. Apertura de archivo de secuencia FastQC.

4. Observar el resultado obtenido en cuanto a porcentaje de cada base leída por el secuenciador para cada posición en la secuencia.

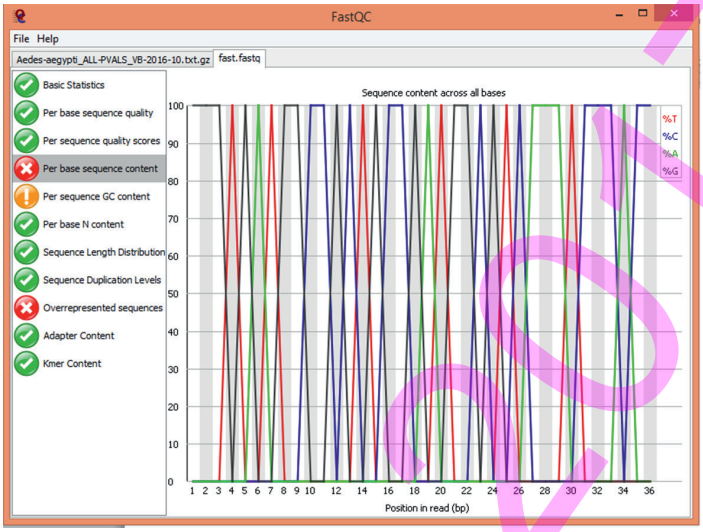


Figura 69. Resultado del FastQC para la evaluación de contenido.

5. Observar resultados en cuanto a la calidad (que depende de un porcentaje alto de lecturas de la misma base en cada posición). La zona verde indica buena claridad.

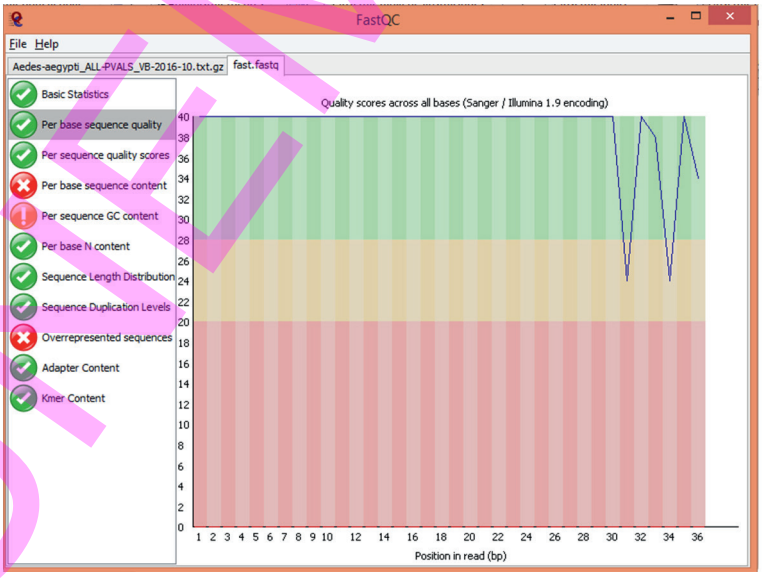


Figura 70. Resultados de calidad en FastQC.

3.2.4 .Conversión de formatos de secuencia de archivos de nucleótidos

1. Ingresar a http://sequenceconversion.bugaco.com/converter/biology/sequences/

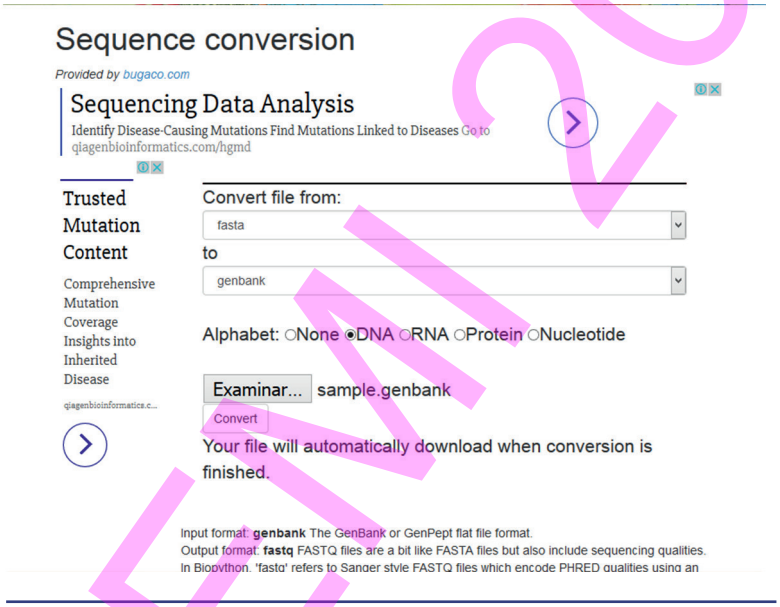


Figura 71. Conversión de archivos.

2. Escoger el archivo y los formatos para realizar la conversiónr, por ejemplo FASTA a GenBank y presionar *Convert*.

3. Obtener el archivo en GenBank.

```
LOCUS       L11285.1                1576 bp    DNA             UNK 01-JAN-1980
DEFINITION  L11285.1 Homosapiens ERK activator kinase (MEK2) mRNA
ACCESSION   L11285
VERSION     L11285.1
KEYWORDS    .
SOURCE      .
  ORGANISM  .
            .
FEATURES             Location/Qualifiers
ORIGIN
        1 gaattcgagc cgaccgaccg ctcccggccc gccccctatg ggccccggct agaggcgccg
       61 ccgccgccgg cccgcggagc cccgatgctg gccccggagga agccggtgct gccggcgctc
      121 accatcaacc ctaccatcgc cgagggccca tcccctacca gcgagggcgc ctccgaggca
      181 aacctggtgg acctgcagaa gaagctggag gagctggaac ttgacgagca gcagaagaag
      241 cggctggaag cctttctcac ccagaaagcc aaggttggcg aactcaaaga cgatgacttc
      301 gaaaggatct cagagctggg cgcgggcaac ggcggggtgg tcaccaaagt ccagcacaga
      361 ccctcgggcc tcatcatggc caggaagctg atccaccttg agatcaagcc ggccatccgg
      421 aaccagatca tccgcgagct gcaggtcctg cacgaatgca actcgccgta catcgtgggc
      481 ttctacgggg ccttctacag tgacggggag atcagcattt gcatggaaca catggacggc
      541 ggctccctgg accaggtgct gaaagaggcc aagaggattc ccgaggagat cctggggaaa
      601 gtcagcatcg cggttctccg gggcttggcg tacctccgag agaagcacca gatcatgcac
      661 cgagatgtga agccctccaa catcctcgtg aactctagag gggagatcaa gctgtgtgac
      721 ttcggggtga gcggccagct catagactcc atggccaact ccttcgtggg cacgcgctcc
      781 tacatggctc cggagcggtt gcagggcaca cattactcgg tgcagtcgga catctggagc
      841 atgggcctgt ccctggtgga gctggccgtc ggaaggtacc ccatcccccc gcccgacgcc
      901 aaagagctgg aggccatctt tggccggccc gtggtcgacg gggaagaagg agagcctcac
      961 agcatctcgc ctcggccgag gccccccggg cgccccgtca gcggtcacgg gatggatagc
     1021 cggcctgcca tggccatctt tgaactcctg gactatattg tgaacgagcc acctcctaag
     1081 ctgcccaacg gtgtgttcac ccccgacttc caggagtttg tcaataaatg cctcatcaag
     1141 aacccagcgg agcgggcgga cctgaagatg ctcacaaacc acaccttcat caagcggtcc
     1201 gaggtggaag aagtggattt tgccggctgg ttgtgtaaaa ccctgcggct gaaccagccc
     1261 ggcacaccca cgcgcaccgc cgtgtgacag tggccgggct ccctgcgtcc cgctggtgac
     1321 ctgcccaccg tccctgtcca tgccccgccc ttccagctga ggacacgtgg cgcctccacc
     1381 caccctcctg cctcaccctg cggagagcac cgtggcgggg cgacagcgca tgcaggaacg
     1441 ggggtctcct ctcctgccag tcctggccgg ggtgcctctg gggacggggcg acgctgctgt
     1501 gtgtggtctc agaggctctg cttccttagg ttacaaaaca aaacagggag agaaaagcaa
     1561 aaaaaaaaaa aaaaaa
//
```

Figura 72. Formato del archivo convertido de FASTA a GenBank.

3.3. PROCESAMIENTO BIOINFORMÁTICO DE SECUENCIAS

El manejo de secuencias de nucleótidos o aminoácidos es necesario para encontrar nueva información emergente o resolver problemas de biología molecular. Cuando se estudian secuencias muy largas y/o muchas secuencias, es preciso recurrir para su procesamiento a herramientas bioinformáticas.

3.3.1. Secuencias reversa, complemenaria y reversa complementaria

Todas las plataformas bioinformáticas comprenden multitud de herramientas de procesamiento de secuencias de nucleótidos y aminoácidos. Entre las aplicaciones más sencillas de los programas bioinformáticos se encuentra la generación de los siguientes tipos de secuencias:

- Secuencia reversa a una secuencia dada: es decir, la misma secuencia en orden inverso. Útil, por ejemplo, si la secuencia de una hebra codificante viene dada en orden inverso.

- Secuencia complementaria de una secuencia de ADN. Determina la secuencia de una hebra antiparalela a la de una secuencia dada de ADN, pero en la direccionalidad inversa a la de la secuencia dada. Es decir, si la secuencia dada está en el orden convencional 5´-3´, la secuencia complementaria se generará con el orden 3´-5´, y viceversa.

- Secuencia reversa complementaria de una secuencia de ADN. Determina la secuencia de una hebra antiparalela a la de una secuencia dada de ADN en la misma direccionalidad que la de la secuencia dada. Es decir, si la secuencia dada está en el orden convencional 5´-3´, la secuencia complementaria se generará también con esa direccionalidad 5´-3´. Así, se puede visualizar el orden convencional en ambas secuencias y estudiarlas y manipularlas bioinformáticamente para diversas aplicaciones, como búsqueda de genes o diseño de **cebadores** (secuencias cortas de nucleótidos que complementan reversamente con una secuencia de ADN para que se pueda iniciar una replicación en laboratorio).

3.3.2. Búsqueda de secuencias codificantes de proteínas

3.3.2.1. Marcos de lectura

Un marco de lectura (*reading frame*) es una división de una secuencia de nucleótidos, dispuestos en el orden convencional, 5´-3´, en tripletes consecutivos. Así, una secuencia rinde tres posibles marcos de lectura, como se ilustra en el este ejemplo:

Fuente: wikimedia.org[16]

Figura 73. Los tres posibles marcos de lectura en una secuencia

Los posibles marcos de lectura en este ejemplo son los siguientes:

AGG·TGA·CAC·CGC·AAG·CCT·TAT·ATT·AGC
A·GGT·GAC·ACC·GCA·AGC·CTT·ATA·TTA·GC
AG·GTG·ACA·CCG·CAA·GCC·TTA·TAT·TAG·C

Tras la obtención de los marcos de lectura, las conclusiones son diferentes en dependencia de a qué corresponda la secuencia ordenada de nucleótidos:

- Si corresponde a una región de un ARNm, uno de los tres marcos de lectura corresponderá a una secuencia de codones en el ARNm.

- Si se sabe que corresponde a una secuencia codificante para proteínas (CDS), uno de los tres marcos de lectura corresponderá a una secuencia de codones en el ADN.

- Si corresponde a una región de la hebra codificante de proteínas, podrá haber intrones. Si los intrones tienen un número de nucleótidos múltiplo de tres, los marcos de lectura se mantienen entre exones. Si no, los marcos de lectura cambiarán entre exones.

- Si corresponde a una región CDS, pero se ignora si corresponde a la hebra codificante o a la hebra molde, entonces habrá que obtener otros tres marcos de lectura para la hebra antiparalela (reversa complementaria), y alguno del total de los seis marcos de lectura se corresponderá a una secuencia de codones del ADN.

- Si pertenece a una región de cuerpo del gen, pero se ignora si corresponde a la hebra codificante o a la hebra molde, también han de obtenerse los otros tres marcos de lectura para la hebra antiparalela (reversa comlementaria). Si hay alguna porción exónica, alguno del total de los seis marcos de lectura se corresponderá al menos parcialmente, con una secuencia de codones del ADN.. Si existen intrones y éstos poseen un número de nucleótidos múltiplo de tres, los marcos de lectura se mantienen entre exones. Si no, los marcos de lectura cambiarán entre exones.

3.3.2.2. Marco de lectura abierto

El marco de lectura abierto (*Open Reading Frame, ORF*) es la parte de un marco de lectura, en una secuencia de ADN o ARN, que ***a priori*** tiene el potencial de ser traducido. Es decir, que comienza con un codón de iniciación y finaliza con un codón de terminación, respectivamente. Sin embargo, no tiene que corresponder necesariamente con la auténtica CDS, ya que:

- Puede que el marco de lectura no sea el correcto, el real en la célula.
- Puede que el codón de iniciación no sea el auténtico en la célula.
- Puede contener intrones, que habría que eliminar.

Existen varias herramientas online para encontrar ORFs en una secuencia dada y procesarlos. Una de ellas es ORFfinder, en la web del NCBI, que identifica todas las ORFs de una secuencia dada.

3.3.2.3. Traducción y retrotraducción in sílico (en computadora)

ORFfinder también genera las traducciones a polipéptidos de cada ORF. Asimismo, existen aplicaciones que generan las traducciones de cada marco de lectura.

También hay herramientas para retrotraducir una secuencia de nucleótidos. Naturalmente, al existir aminoácidos que pueden ser codificados por un grupo de codones distintos, las soluciones de retrotraducción son en general variadas. Sin embargo, son útiles para diversos propósitos en biología molecular, como por ejemplo para el diseño de cebadores degenerados (ver Sección 3.3.1. Secuencias reversa, complementaria y reversa complementaria), que en general consisten en mezclas de cebadores similares en secuencia, pero que en algunas posiciones poseen nucleótidos distintos.

3.3.1. Alineamiento de secuencias

Alinear secuencias consiste en disponer al menos dos secuencias, de manera que se puedan detectar fácilmente las similitudes y/u homología. Los algoritmos a propósito se diseñan de manera que las similitudes y/o las diferencias queden resaltadas. A partir de estos alineamientos, se pueden obtener:

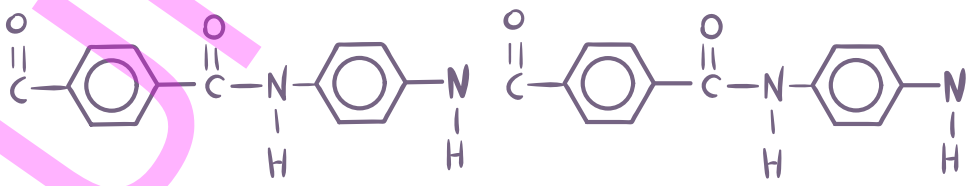
- **Valores de similitud:** Mediante análisis cuantitativo de la estructura primaria de dos o más secuencias que pueden ser ácidos nucleicos (ADN o ARN) o polipéptidos.

- **Gráficos de homología:** Mediante análisis cualitativo de las secuencias, de nucleótidos o aminoácidos, que por razones generalmente evolutivas (de evolución de los organismos vivos) se han conservado en mayor o menor medida.

Las comparaciones se efectúan a partir de miles de combinaciones, lo que sería demasiado complejo sin ayuda de un equipo especializado, y se invertiría un tiempo muy elevado tan solo para lograr ordenar las secuencias, sin contar el tiempo invertido en el análisis e interpretación que podrían extraerse en la comparación.

Una **aplicación del alineamiento** y comparación de secuencias de cadenas de ADN o ARN es su utilización en análisis biomédico para emitir conclusiones acerca de posibles cambios de secuencia (mutaciones) debidas a efectos de exposiciones a contaminación directa o indirecta del sujeto analizado. Algunas mutaciones de CDSs son *silenciosas*, es decir, no alteran el aminoácido que se expresa a partir del codón de ADN en que ocurrió la mutación. Sin embargo, muchas de las mutaciones pueden modificar el aminoácido, y por tanto la proteína y su función, en la mayoría de los casos causando alteranciones negativas en el organismo.

La cuantificación del parecido entre dos secuencias se hace a través del parámetro *score* (puntaje del alineamiento). Si el valor del *score* es alto significa que hay alta similitud en un tramo largo de las secuencias comparadas.



Existen dos tipos de alineamiento:

- Pareado: (2 secuencias).
- Múltiple: (más de 2 secuencias).

Ejemplo 1:

```
Secuencia 1:        A A A A C T T T

Secuencia 2:        A A A C T T
```

Secuencias alineadas (6 correspondencias de similitud, 2 diferencias y 2 huecos):

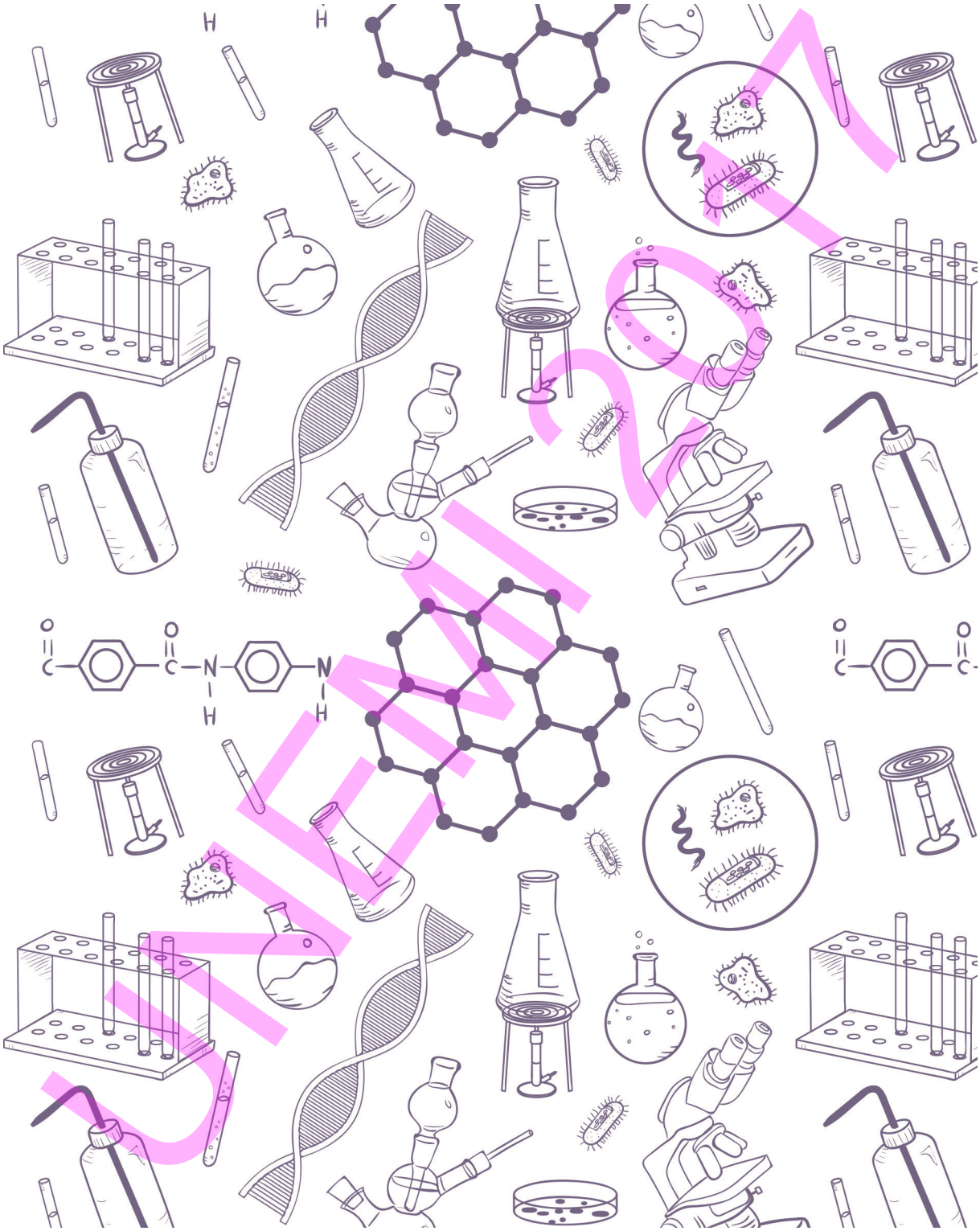
```
Alineamiento:
Secuencia 1:        A A A A C T T T
                    | | |   | | |
Secuencia 2:        A A A   C T T

                    A A A – C T T -
```

Ejemplo 2:

```
Secuencia 1:        A A A A C T T T

Secuencia 2:        A A A C C T T
```

Secuencias alineadas (7 correspondencias por similitud)

```
Secuencia 1:        A A A A C T T T
                    | | |   | | | |
Secuencia 2:        A A A G C T T T
                    A A A A C T T T
```



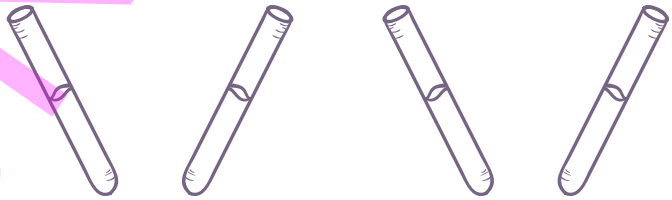
3.4. BÚSQUEDA DE SECUENCIAS: BLAST

BLAST (*Basic Local Alignment Search Tool*) representa un grupo de algoritmos desarrollados por Stephen Altschul, Warren Gish, David Lipman (del NCBI), Webb Millar (Universidad Estatal de Pennsylvania), y Gene Myers (Universidad de Arizona). Los programas **BLAST** comparan una secuencia dada (*query*) con secuencias de bases de datos diana (*target data bases, target DB*). Las secuencias, tanto las dadas como las de las bases de datos, pueden ser de nucleótidos (ya sean de ADN o de ARN) o de aminoácidos (proteínas). La secuencia de entrada ha de estar en formato FASTA (nucleótidos o aminoácidos) o similar. En el caso de secuencias nucleotídicas, BLAST busca también homologías con las secuencias reversas complementarias de las que están en la base de datos.

Las comparaciones detectan polimorfismos (elementos no coincidentes) y grados de parecido, y uno de los resultados es una lista de secuencias ordenadas de mayor a menor parecido. Existen opciones para delimitar las búsquedas en las *target DB*. La velocidad y precisión del BLAST son clave en la búsqueda.

Entre las diversas aplicaciones del BLAST, se encuentran, por ejemplo:

- Buscar y encontrar regiones conservadas (que por tanto tienen alta similitud) de un gen en especies evolutivamente relacionadas para poder diseñar cebadores (ver Sección 3.3.1. Secuencias reversa, complementaria y reversa complementaria) de un DNA aún no secuenciado de una de esas especies, obtener una amplificación de un fragmento de ese gen en esa especie, para poder entonces secuenciarlo.



- Buscar y encontrar, a partir de una secuencia de un gen con función celular conocida en una especie, secuencias similares aún no caracterizadas de otras especies, y que por tanto podrían pertenecer a un gen con la misma función de aquél o parecida.

- Encontrar las secuencias más parecidas, aplicando filtros o no, a una dada, para hacer, por ejemplo, estudios de similitud o evolutivos (obtener a partir de ahí árboles filogenéticos, por ejemplo; ver Sección 3.6. Matriz de distancias genéticas. Árboles filogenéticos).

3.4.1. Ingreso a la herramienta BLAST a través de NCBI

1. Ingresar a https://www.ncbi.nlm.nih.gov/

2. Escoger BLAST en Recursos populares (*Popular resources*). Por ejemplo, con esta herramienta puede calcularse que más del 95% de la composición genética de los chimpacés es similar a la humana, por lo que son considerados nuestros parientes evolutivos más cercanos.

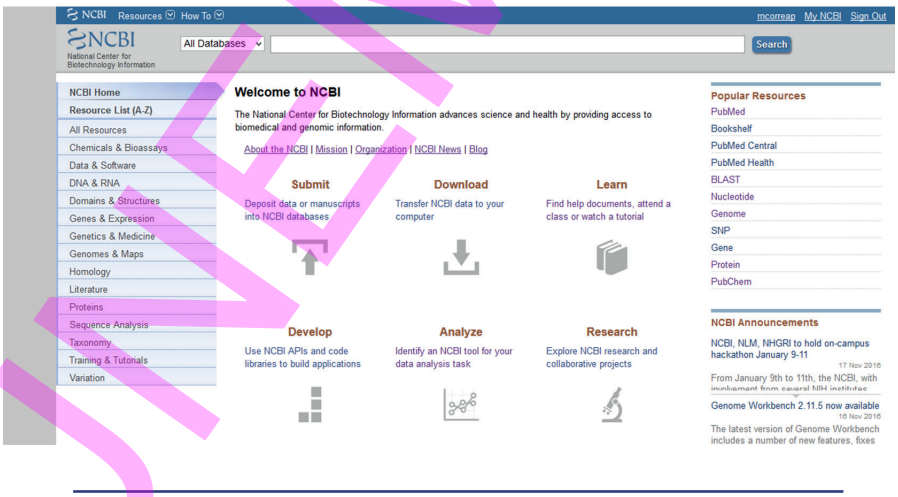


Figura 74. Ingresar a BLAST desde NCBI.

3.4.2. Tipos de búsquedas BLAST en NCBI

Una vez en el sitio de BLAST se observa la siguiente consola:

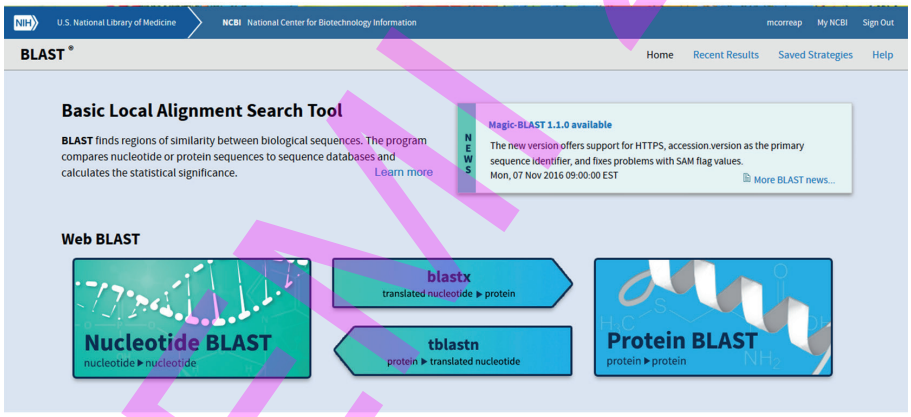


Figura 75. Consola de BLAST.

Programa	*Query*	*Target DB*	Descripción
BLASTn	Nucleótidos	Nucleótidos	Una secuencia de nucleótidos se compara con una base de datos de secuencias de nucleótidos.
BLASTp	Proteína	Proteína	Una secuencia de aminoácidos se compara con una base de datos de secuencias de aminoácidos. Por ejemplo, para identificar regiones comunes entre proteínas.
BLASTx	Nucleótidos traducidos a proteína	Proteína	Una secuencia de nucleótidos se compara con una base de datos de secuencias de aminoácidos, previa traducción de la secuencia de nucleótidos y su reversa complementaria a las 6 posibles proteínas*.
tBLASTn	Proteína	Nucleótidos traducidos a proteína	Una secuencia de aminoácidos se compara con una base de datos de secuencias de nucleótidos, previa traducción de las secuencias de nucleótidos y sus reversas complementarias a las 6 posibles proteínas*.
tBLASTx	Nucleótidos traducidos a proteína	Nucleótidos traducidos a proteína	Una secuencia de nucleótidos se compara con una base de datos de secuencias de nucleótidos, previa traducción en ambos casos de las secuencias de nucleótidos y sus reversas complementarias a las 6 posibles proteínas*.

*Basadas en los 6 posibles marcos de lectura.

Adaptado de: (“IBM Knowledge Center - ¿Qué es BLAST?”).

3.4.3. Ejemplo de BLAST

3.4.3.1. BLAST en NCBI

1. Ingresar a Nucleotide BLAST (nBLAST) y copiar una secuencia de archivo FASTA.

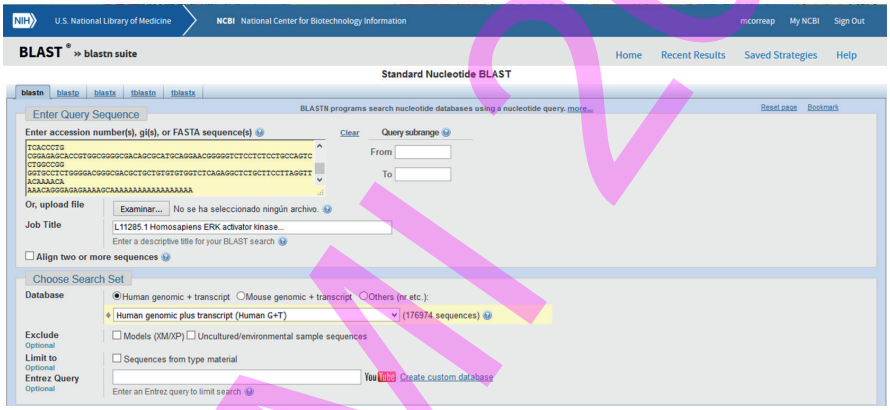


Figura 76. BLAST con secuencia de archivo FASTA de Homo Sapiens.

2. Presionar el botón BLAST para obtener las secuencias más parecidas. BLAST busca las coincidencias a lo largo de una secuencia problema, en ambas direcciones (hebras paralela y antiparalela), y finalmente crea un alineamiento en los tramos de secuencia parecidos.

3. En el panel que se obtiene se pueden seleccionar opciones distintas a las que aparecen por defecto. Al cabo de un tiempo se obtiene el gráfico de coincidencia.

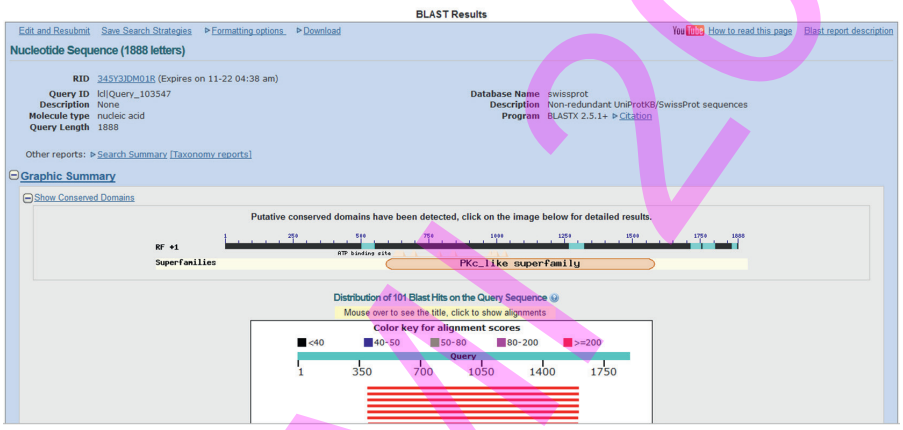
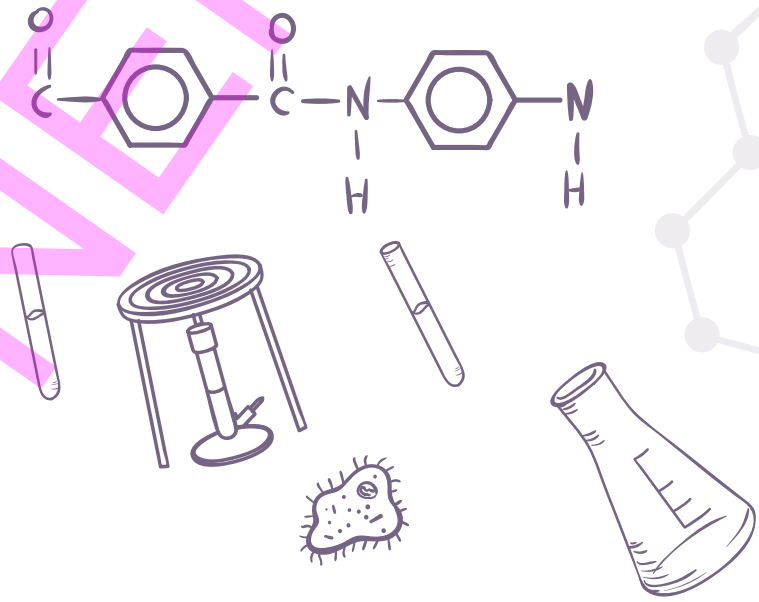


Figura 77. Resultados del BLAST anterior para una secuencia de Homo Sapiens.

4. Descargar los alineamientos entre la secuencia problema y cada una de las secuencias encontradas.

Figura 78. Alineamiento en BLAST.



3.4.3.2. BLAST en VectorBase con una secuencia obtenida en NCBI

1. Ingresar al sitio web de NCBI y escoger la base de datos *Nucleotide* introduciendo, por ejemplo, *Homo Sapiens.*

2. Descargar en formato FASTA una secuencia del organismo *Homo Sapiens.*

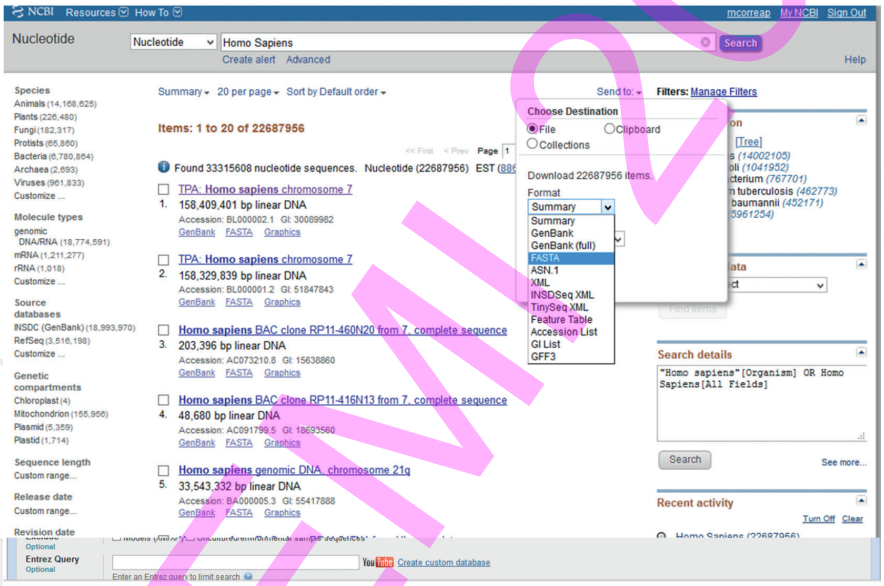


Figura 79. Descarga de un archivo de Homo Sapiens desde el sitio NCBI.

3. Ingresar al sitio web de VectorBase

4. Ingresar a la herramienta BLAST del sitio web de VectorBase

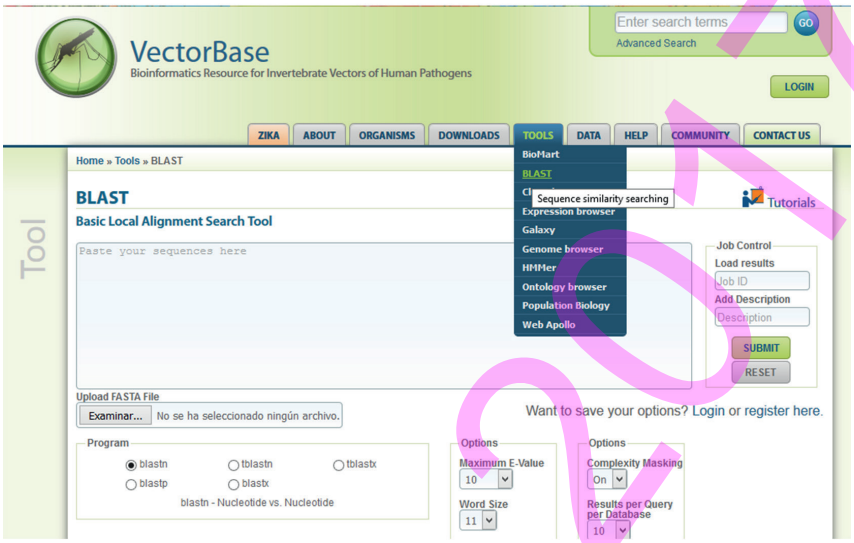


Figura 80. Herramienta BLAST de VectorBase.

5. Cargar el archivo FASTA de NCBI en la herramienta BLAST de VectorBase

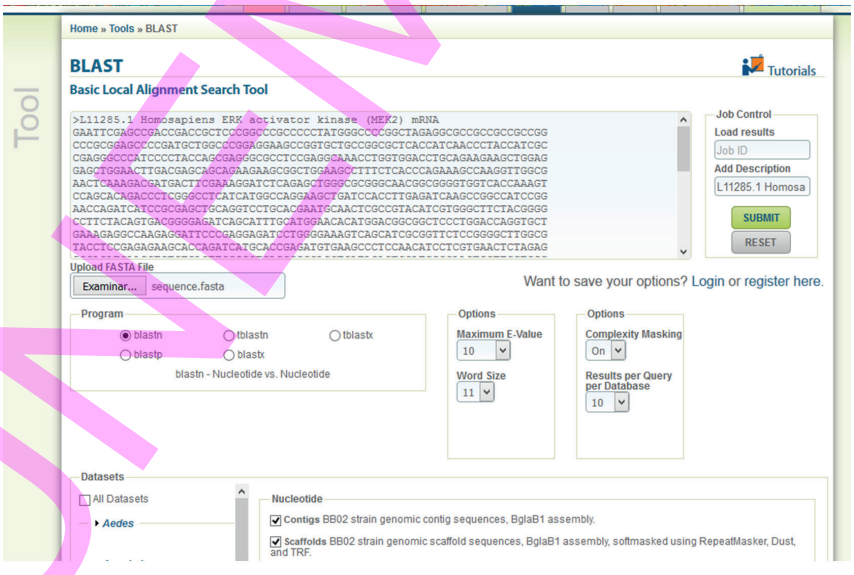


Figura 81. Carga del archivo FASTA de NCBI en la herramienta BLAST de VectorBase.

6. Seleccionar el programa *BLASTn* y la base de datos de nucleótidos.

7. Escoger *SUBMIT* para la presentación de la secuencia.

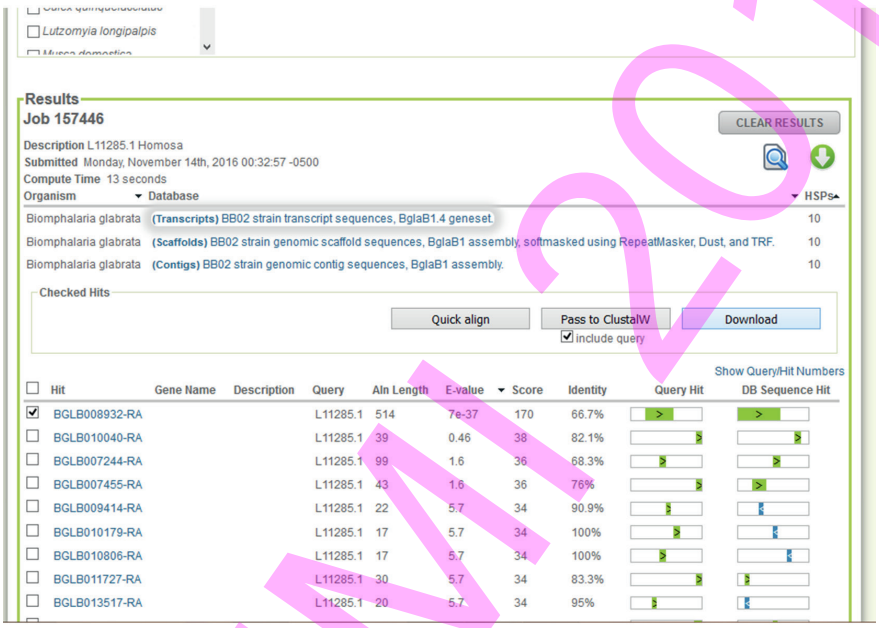


Figura 82. Secuencias obtenidas con BLAST de VectorBase.

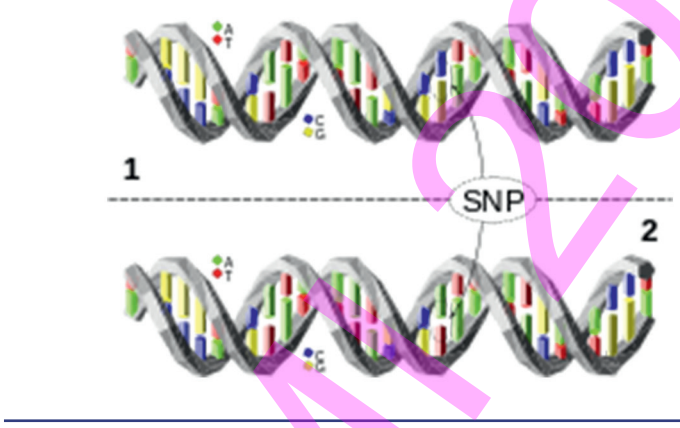
8. Descargar las secuencias obtenidas con BLAST de VectorBase.

3.5. DETECCIÓN DE POLIMORFISMOS

En genómica y proteómica, los polimorfismos son diferencias entre secuencias similares que se comparan. Pueden ser secuencias del genoma o del proteoma de un individuo, o comúnmente secuencias de individuos de la misma especie o de especies distintas.

3.5.1. SNPs

El polimorfismo de un nucleótido o SNP *(Single Nucleotide Polymorphism,* pronunciado *snip*) es una variación en la secuencia de ADN que afecta a un solo nucleótido.



Fuente: wikimedia.org[17]

Figura 83. Cadenas de ADN con el cambio de una base nitrogenada (y por tanto, de nucleótido).

Se pueden buscar polimorfismos de un nucleótido para el estudio de enfermedades complejas, lo que implicaría, si se encuentra por ejemplo en individuos de una misma familia, que la enfermedad es heredable. Si una enfermedad multifactorial (cuyo desarrollo parece depender de varios factores) es heredable, muchas de sus variantes estarán asociadas a SNPs.

3.5.1.1. Ejemplo de Búsqueda de SNPs en NCBI

1. Ingresar a https://www.ncbi.nlm.nih.gov en *SNP.*

2. Consultar sobre *Homo Sapiens.*

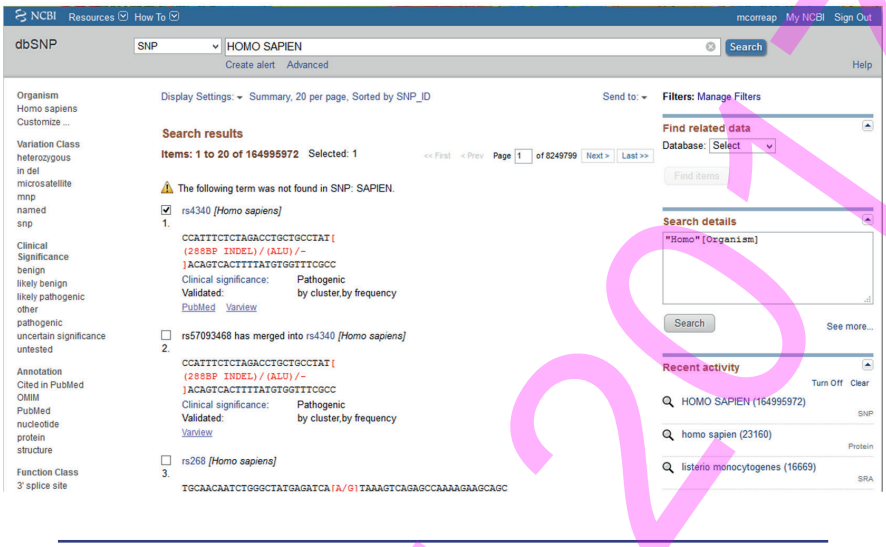


Figura 84. SNPs en NCBI.

3.5.2. InDels.

En el genoma se producen fenómenos de inserción o deleción (eliminación de trozos de secuencias), que en conjunto se denominan InDels, y constituyen mutaciones genéticas que pueden abarcar más de un nucleótido. Los InDels son considerablemente menos frecuentes en las regiones codificantes que en las no codificantes, ya que en estas su repercusión en la biología del organismo es menor. Los InDels se estudian desde la genética forense porque pueden utilizarse como marcadores polimórficos (marcan diferencias entre varios tipos de individuos de distinta procedencia geográfica en determinados grupos poblacionales.

La mutación más común de la fibrosis quística es una eliminación, en el gen de la proteína receptora transmembrana asociada a la fibrosis quística (CFTR), de un triplete de nucleótidos que codifican el aminoácido fenilalanina en la posición 508 de la proteína (Δ F508 o p.F508del).

Fuente: github.com[18]

Figura 85. Representación de un InDel (en este caso, una inserción en una secuencia CDS).

3.6. MATRIZ DE DISTANCIAS GENÉTICAS. ÁRBOLES FILOGENÉTICOS.

La **filogenética** estudia la relación evolutiva de los seres vivos, y la **filogenia** es la representación del árbol de divergencia evolutiva.

Estos árboles filogenéticos pueden ser dicotómicos o politómicos. La construcción de los árboles filogenéticos estaba originalmente basada en distancias fenotípicas, que no reconstruyen necesariamente la evolución sino más bien la similitud entre los seres. En la actualidad, estos árboles son más rigurosos al estar representados sobre la base de similitudes o diferencias de secuencias genómicas, como genes, o proteínas.

Existen dos métodos fundamentales de construcción de árboles filogenéticos:

- UPGMA (método de agrupamiento de pares no ponderados), basado en la similitud en pares.
- Neighbor-joining, que aplica algoritmos y produce ramas proporcionales en el árbol.

A partir de la obtención de matrices de distancias genéticas se pueden aplicar distintos métodos de reconstrucción de las relaciones entre secuencias: determinístico, probabilístico, inferencia bayesiana. (Ver Sección 4.3. Algoritmos aplicados a alineamientos de secuencias).

3.6.1. Un software de alineamiento y filogenia: Clustal Omega

1. Ingresar a Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)

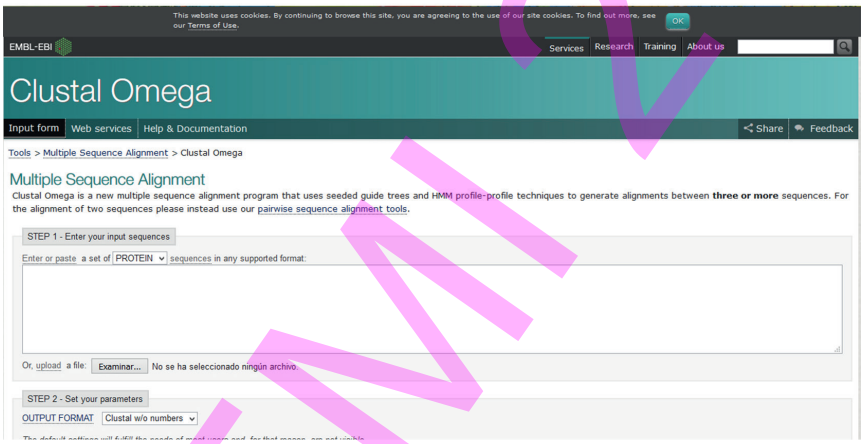
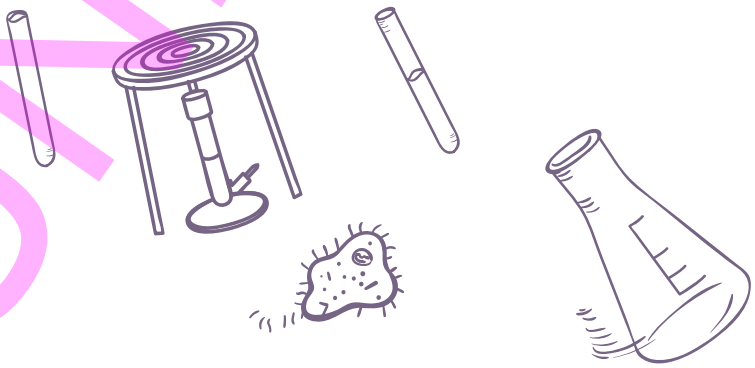


Figura 86. Sitio web de Clustal Omega.



2. Seleccionar los archivos FASTA a comparar y obtener el resultado del alineamiento.

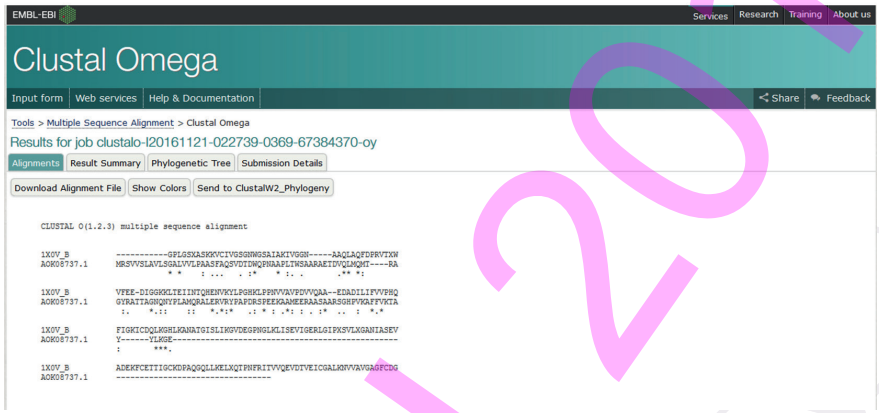


Figura 87. Alineamiento de dos secuencias en Clustal Omega.

3. Visualizar e árbol creado por Clustal Omega que no debe ser considerado como un ábol filogenético científicamente riguroso, sino como una guía de asociación genética entre especies o individuos.

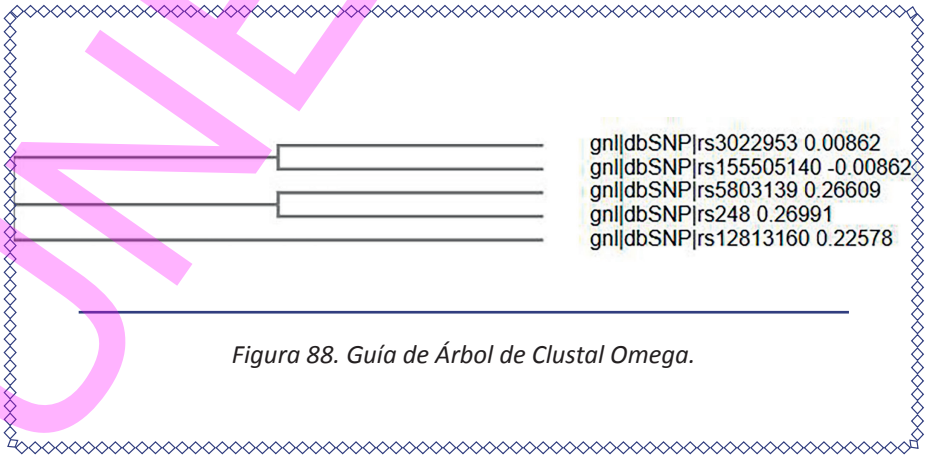
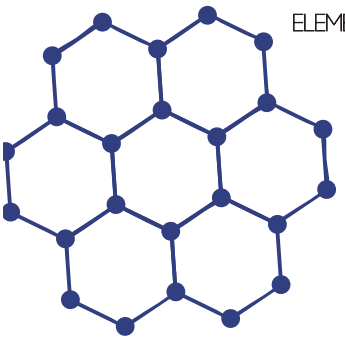
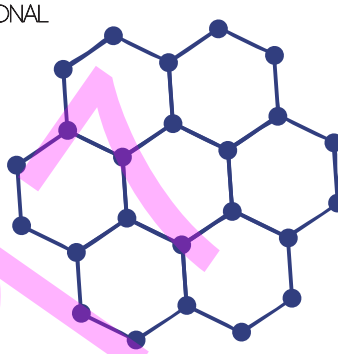


Figura 88. Guía de Árbol de Clustal Omega.



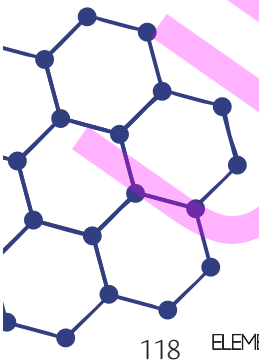


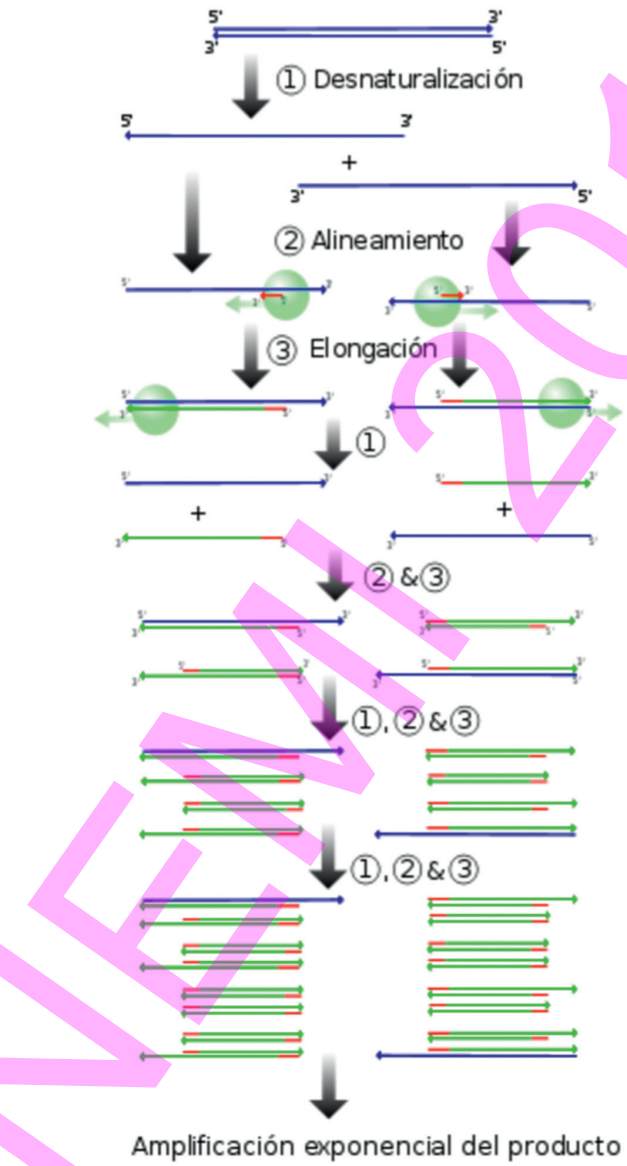
3.7. DISEÑO DE CEBADORES

K. B. Mullis (1983)[47] revolucionó el modo de detectar y amplificar fragmentos de ADN mediante la técnica de Reacción en Cadena de la Polimerasa (*Polymerase Chain Recation,* PCR), sin la cual no hubiera tenido lugar el enorme desarrollo de la Bioilogía Molecular. La PCR consiste en la replicación sucesiva de un fragmento de ADN (**amplicón**), según el fundamento que a continuación se explica.

Una molécula de ADN se **desnaturaliza** (sufre la separación de sus dos hembras) a unos 95°C. El extremo 5' de cada fragmento de hebra dentro del amplicón es una secuencias nucleotídica corta que es complementaria con un oligonucleótido (hebra corta) denominado **cebador** (***primer***). Por lo tanto, ese extremo de hebra del amplicón, cuando el ADN está desnaturalizado y se ha bajado la temperatura lo suficiente (40-60°C), **anilla** (se *alinea* manera complementaria reversa) con el cebador. De esta manera, el enzima **ADN-polimerasa** se apoya sobre esta estructura de anillamiento para, a una temperatura 72°C, ir sintetizando el resto de la hebra complementaria en sentido 5'-3' (**elongación**).

Naturalmente, los procesos consecutivos de anillamiento y elongación ocurren en ambas hebras, a partir de sendas moléculas de cebador. El ciclo desnaturalización-anillamiento-elongación se puede repetir indefinidamente, de manera que se obtiene una amplificación exponencial del amplicón.





Fuente: wikimedia.org [19]

Figura 89. Amplificación de un fragmento de ADN utilizando PCR.

3.7.1. Características ideales de los cebadores

Los cebadores se utilizan, por tanto, con el objetivo de obtener muchas copias de un fragmento de ADN, y también en las tecnologías de secuenciación por síntesis. Con estos fines, han de diseñarse pares de cebadores para cada amplicón. Las características ideales de los cebadores son:

- 20-25 nucleótidos de longitud, aunque podrían ser de 18 a 30 nucleótidos.
- G o C en el extremo 3'.
- Temperaturas de fusión (Tm, *melting tempertaure*), es decir, los valores de temperatura en que, según la misma desciende, los cebadores comienzan a anillar a sus regiones complementarias en el ADN, que se suelen encontrar en un rango de 50-65 °C. La diferencia de Tm entre los dos cebadores para un amplicón no debe ser mayor de 2 °C.
- Contenido GC: 40-60%.
- Debe tener un 100% de apareamiento con el molde complementario del amplicón.

3.7.2. Diseño de cebadores

Los programas bioinformáticos de diseño de cebadores rastrean a lo largo de las secuencias de dos hebras complementarias de ADN y encuentran sub-secuencias cortas que se ajusten a las características deseables de las regiones diana de los cebadores. Esos programas pueden encontrar varios pares de cebadores y ayudar a seleccionar el par más apropiado para una amplificación concreta. A continuación se comentan algunas herramientas bioinformáticas para el diseño de cebadores:

3.7.2.1. Oligo

Oligo (National Biosciences, Plymouth, MN, USA; distribuido por sus desarrolladores: Molecular Biology Insights, Plymouth, MN, USA <http://www.mbinsights.com>) es una herramienta valiosa valioso para el diseño de sondas de hibridación (secuencias de nucleótidos que complementan a una secuencia diana) y cebadores para PCR y secuenciación. Aplica algoritmos que tienen en cuenta aspectos termodinámicos, con lo cual se logra una mayor probabilidad de éxito en la amplificación.

3.7.2.2. Ejemplo con Primer3Plus

1. Consultar en internet el enlace de Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi)

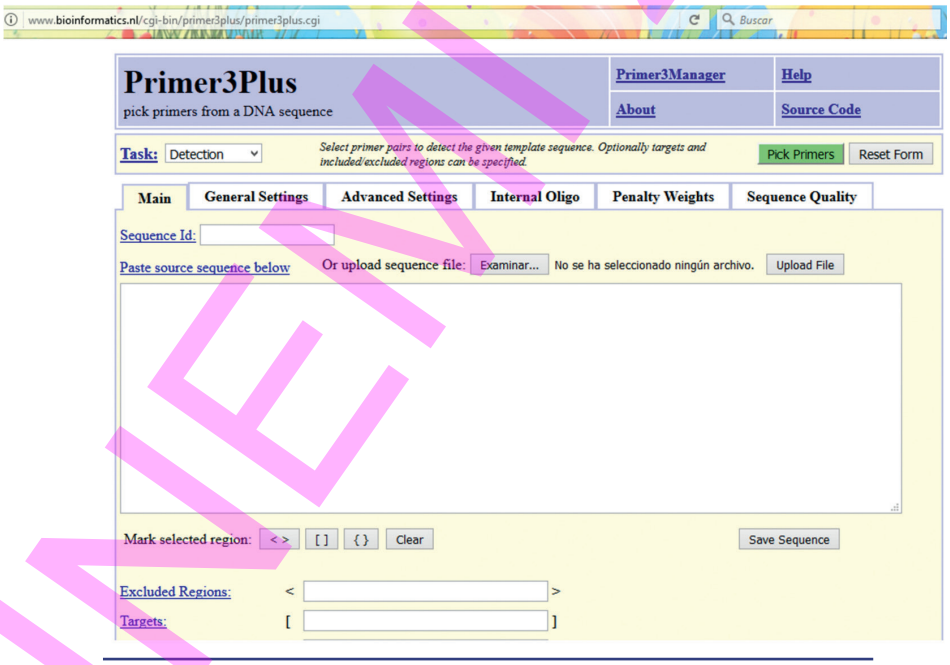


Figura 90. Pantalla de Primer3Plus.

2. Copiar y pegar la secuencia objetivo, en este ejemplo con *Homo Sapiens.*

3. Presionar el botón *Pick Primers.*

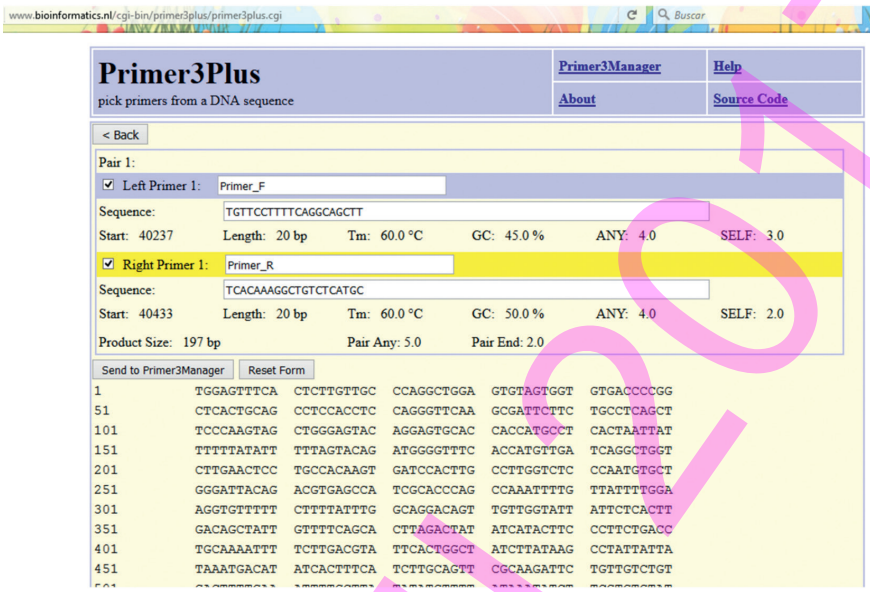


Figura 91. Diseño de primers en Primer3Plus.

3.7.2.3. NetPrimer:

Esta herramienta permite visualizar en los cebadores posibles defectos que dificulten el anillamiento con la hebra molde.

3.7.2.4. Ejemplo de diseño de cebadores desde el NCBI

1. Ingresar al sitio web del NCBI (https://www.ncbi.nlm.nih.gov/)

2. Ingresar en la secuencia de nucleótidos (*Nucleotide)* de "*Homo Sapiens* RP11-416N13 from 7" (puede teclearse en la sección Search)

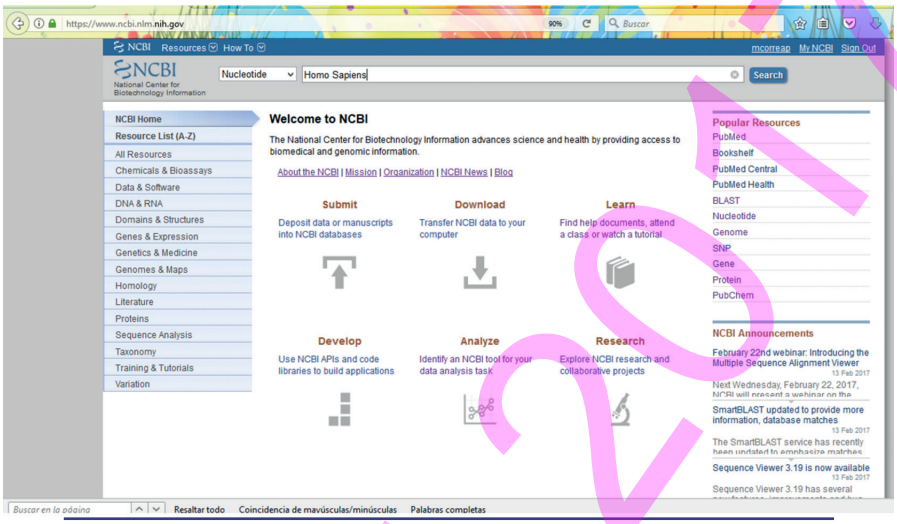


Figura 92. Ingreso a NCBI.

3. Seleccionar *Pic Primers.*

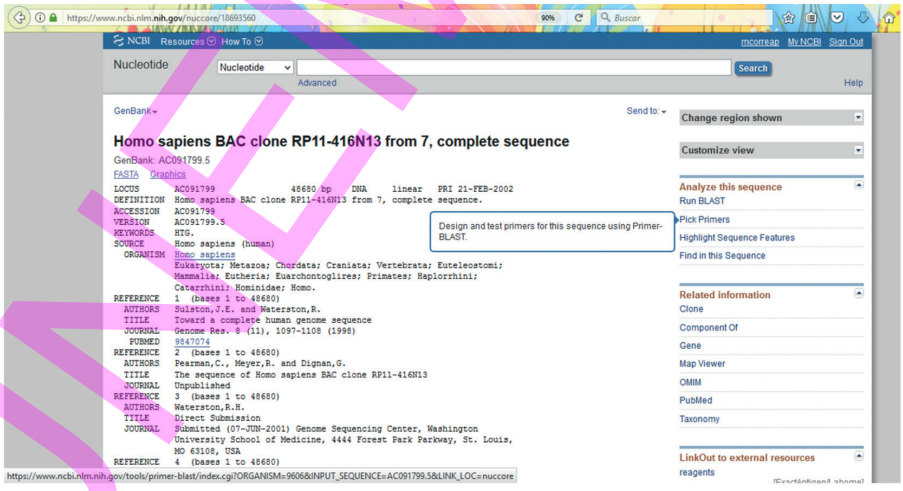


Figura 93. Consulta de Primer desde NCBI.

Se obtendrá la pantalla de la búsqueda específica de Primer.

4. Presionar el botón *Get Primers*

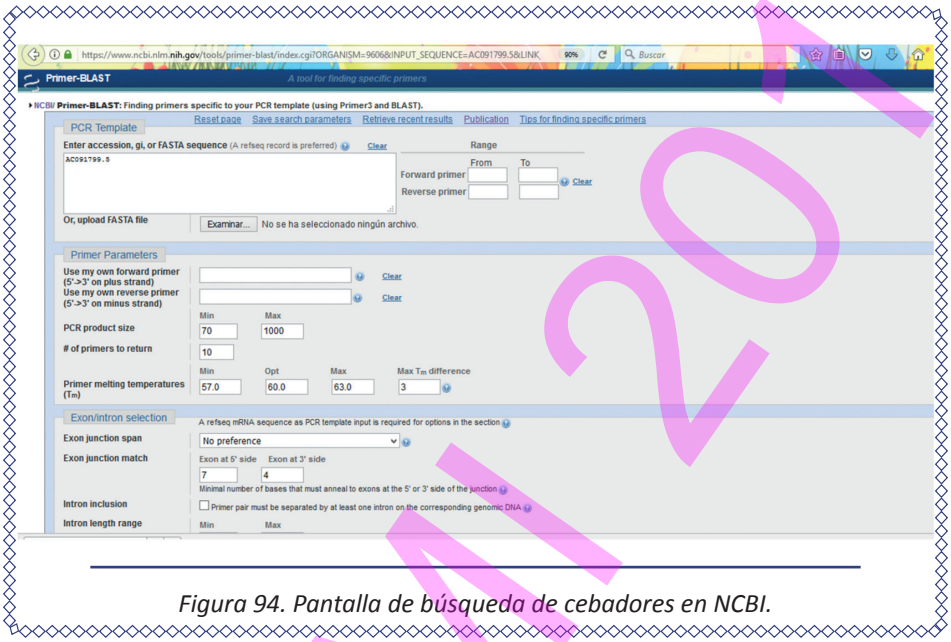


Figura 94. Pantalla de búsqueda de cebadores en NCBI.

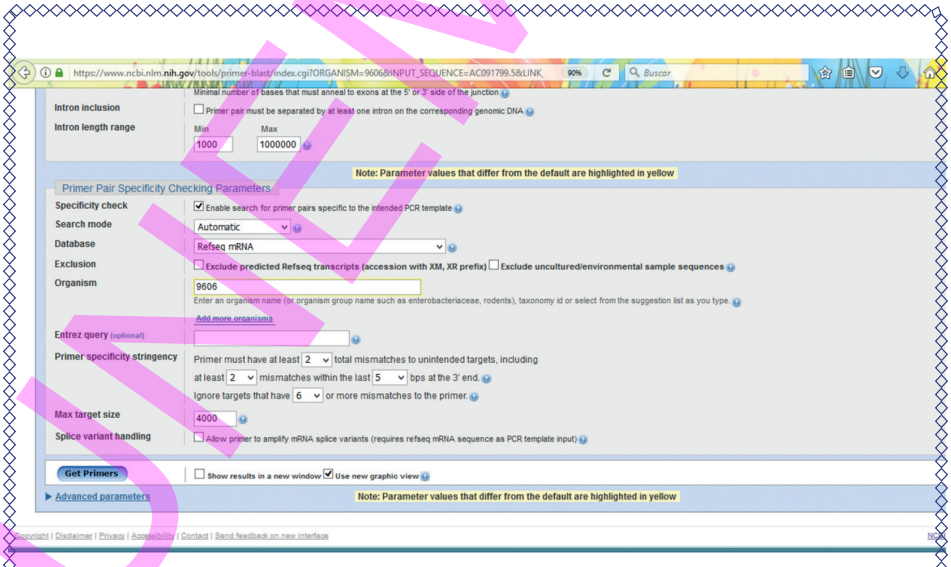


Figura 95. Generación de cebadores en NCBI.

5. Presionar *check* para obtención de los cebadores

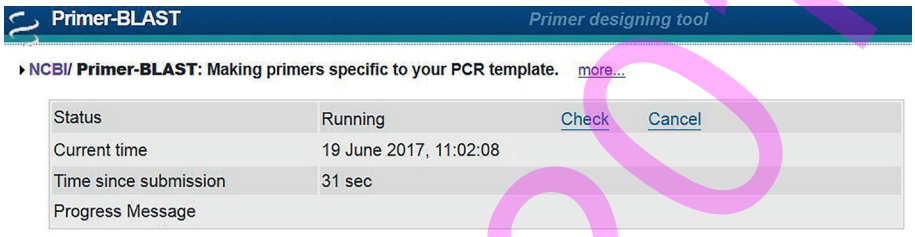


Figura 96. Obtención de cebadores.

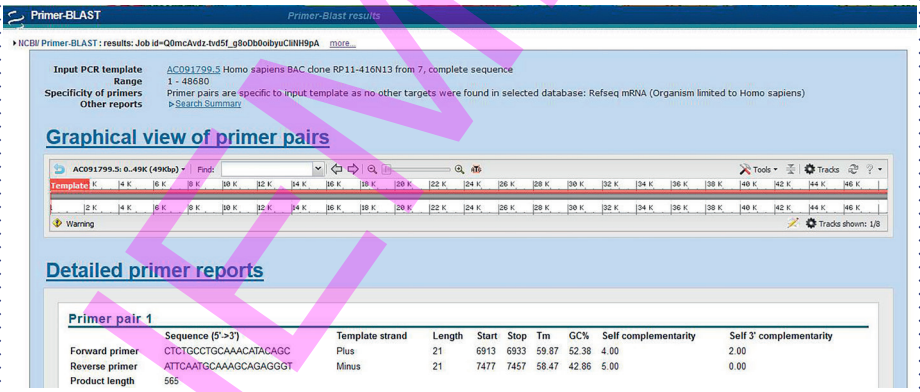
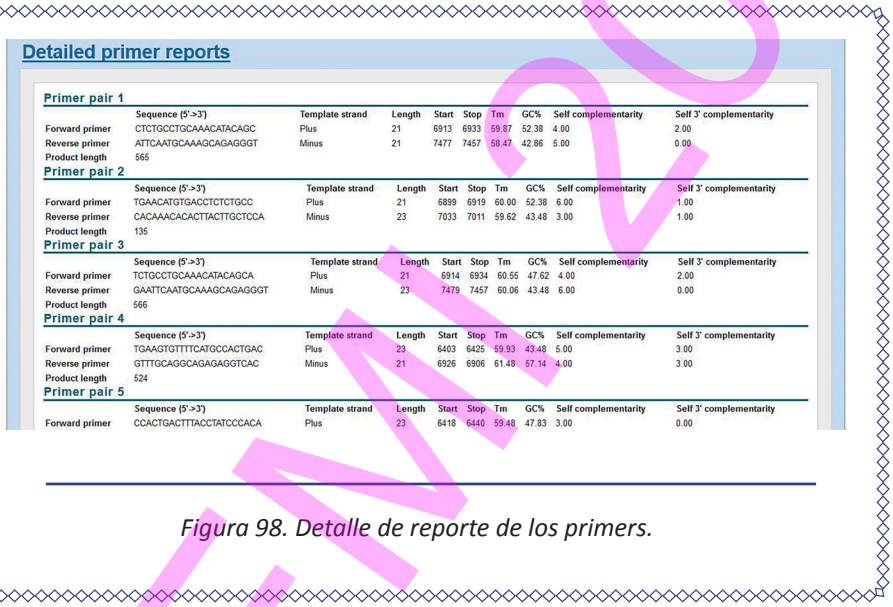


Figura 97. Primer par de cebadores propuesto para la secuencia RP11-416N13 de Homo sapiens, en NCBI.

Se presentarán 10 opciones de pares de cebadores con características similares para las cuales se especifica su porcentaje en guanina y citosina (GC%), la región que se está amplificando y la longitud del amplicón.



Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTCTGCCTGCAAACATACAGC	Plus	21	6913	6933	59.87	52.38	4.00	2.00
Reverse primer	ATTCAATGCAAAGCAGAGGGT	Minus	21	7477	7457	58.47	42.86	5.00	0.00
Product length	565								

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TGAACATGTGACCTCTCTGCC	Plus	21	6899	6919	60.00	52.38	6.00	1.00
Reverse primer	CACAAACACACTTACTTGCTCCA	Minus	23	7033	7011	59.62	43.48	3.00	1.00
Product length	135								

Primer pair 3

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TCTGCCTGCAAACATACAGCA	Plus	21	6914	6934	60.55	47.62	4.00	2.00
Reverse primer	GAATTCAATGCAAAGCAGAGGGT	Minus	23	7479	7457	60.06	43.48	6.00	0.00
Product length	566								

Primer pair 4

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TGAAGTGTTTTCATGCCACTGAC	Plus	23	6403	6425	59.93	43.48	5.00	3.00
Reverse primer	GTTTGCAGGCAGAGAGGTCAC	Minus	21	6926	6906	61.48	57.14	4.00	3.00
Product length	524								

Primer pair 5

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CCACTGACTTTACCTATCCCACA	Plus	23	6418	6440	59.48	47.83	3.00	0.00

Figura 98. Detalle de reporte de los primers.

3.7.2.5. Geneious

(Ver Sección 3.8.12. Diseño de Primers en Geneious)

3.8. UNA PLATAFORMA PARA EL TRATAMIENTO BIOINFORMÁTICO INTEGRAL DE SECUENCIAS: GENEIOUS

La plataforma Geneious (Biomatters Ltd.) es un software que comprende una amplia gama de herramientas bioinformáticas integradas. Para descargarlo debe ingresar a http://geneious.com/download.

Figura 99. Página de descarga de Geneious.

[Para profundizar en el manejo de la plataforma Geneious, existen diversos manuales, algunos de los cuales se mencionan en el subapartado "Tutoriales de Geneiuos", en el apartado "Referncias Bibliográficas", al final del libro]

3.8.1. Requisitos mínimos para el funcionamiento de Geneious

Geneious es compatible con los sistemas operativos más comunes: Windows, Mac y Linux. Requisitos mínimos:

- Procesador: Intel x86/x86-64 MHz.
- Memoria: 2048 Mb o superior.
- Disco duro: mínimo 2 Gb.
- Resolución en pixels: 1024 x 768 o superior.
- Lenguaje: Java 1.8 o superior.

3.8.2. Licencias en Geneious

Se pueden activar una licencia personal o elegir conectarse a un servidor de licencia a través del menú Ayuda en Geneious, mediante las siguientes opciones:

- En caso de haber adquirido una licencia personal, utilizar una clave de licencia. Para esto ha de introducirse el nombre del licenciatario tal y como aparece en el correo electrónico donde recibió su registro de activación. Se requiere conexión a Internet para activar licencias personales, y debe configurarse el firewall / proxy para permitir el acceso a http://licensing.biomatters.com en el puerto 80.

- Usar el servidor de licencias, en caso de que se hubiera adquirido una licencia adminsitrada a través de un servidor de licencias FLEXnet.

- Usar el gestor de licencia y activos de software Sassafras KeyServer, administrado a través de Sassafras (https://www.sassafras.com/).

Una licencia para descargar la versión libre es remitida a un correo electrónico, y facilita 14 días de prueba. En el presente tutorial se ha utilizado la versión R10.0.2

Figura 100. Pantalla inicial de Geneious R10.

3.8.3. Instalación de Geneious

3.8.3.1. En Linux

1. Acceder a la pestaña de descarga en el sitio de Geneious: http://www.geneious.com/download

2. Por defecto se mostrará el enlace para descargar el instalador para Windows. Para cambiar al el instalador de Linux, hacer clic en Linux y descargar

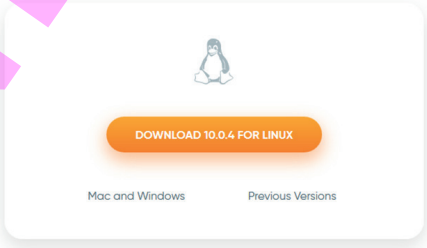


Figura 101. Descarga de Geneious en Linux.

Se descargará el archivo **Geneious_linux64_10_0_4_with_jre.sh** (última versión disponible en el momento de elaborar este documento)

3. Abrir una ventana del terminal de comandos en Linux.

4. Si el archivo se guardó en Descargas (o *Downloads*), ejecutar lo siguiente:

usuario@computador~$ sudo su
root@computador:/home/usuario/# cd Descargas
root@computador:/home/usuario/Descargas# sh Geneious_linux64_10_0_4_with_jre.sh

Nota: después de invocar al superusuario << **sudo su**>> el terminal pedirá la contraseña del sistema.

3.8.3.2. Instalación en Windows

Para ello, descargar el instalador para Windows, y realizar doble clic para ejecutarlo, y seguir las instrucciones. Si se instala por defecto, Geneious se alojará en *Archivos de programa.*

Figura 102. Descarga de Geneious para Windows.

Se pueden instalar plugins y personalizar funciones adicionales de Geneiuos. Para ello, ingresar a Herramientas *(Tools)* → *Plugins*.

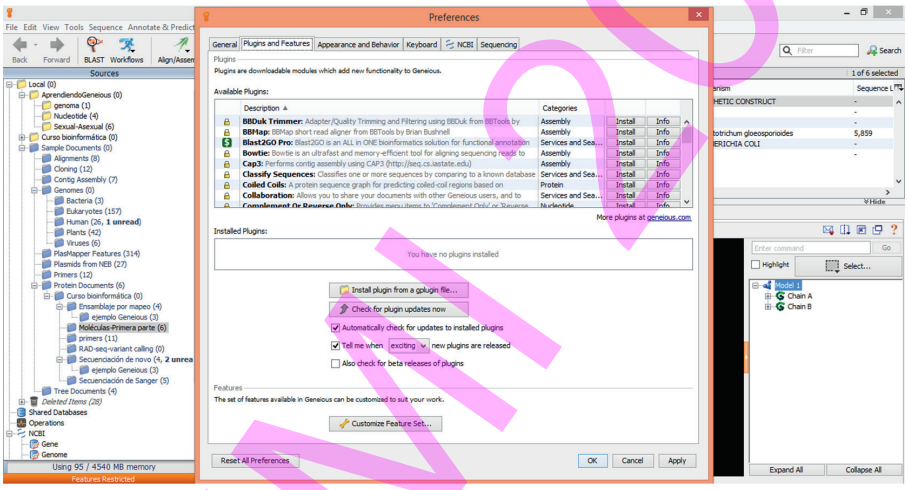


Figura 103. Los plugins en Geneious.

3.8.4. Elección de la ubicación para almacenar datos

Geneious almacena datos en una carpeta llamada Geneious. La ubicación de esta carpeta es solicitada en el momento de la instalación; es posible almacenar datos en una red o en una unidad USB para que pueda acceder a ella desde otros ordenadores, pero no es recomendable pues esto puede afectar al rendimiento. Para cambiar la ubicación de la base de datos Geneious ir a *Tools* → *Preferences* → *General.*

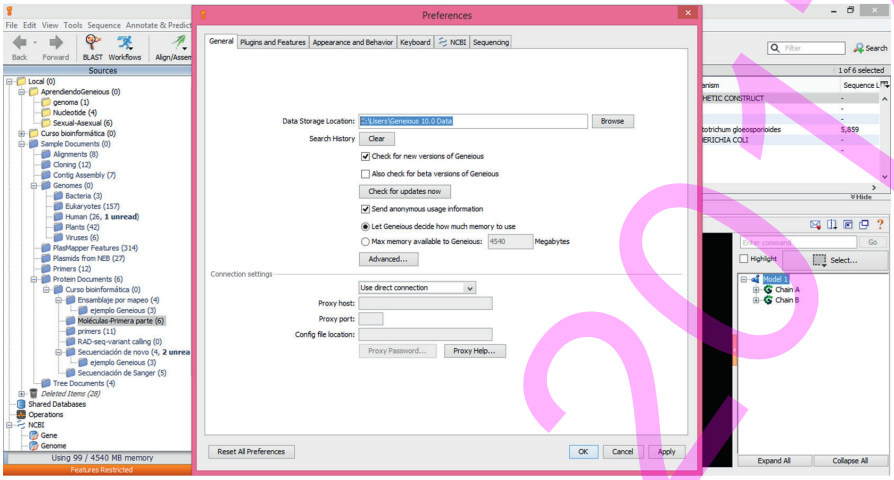
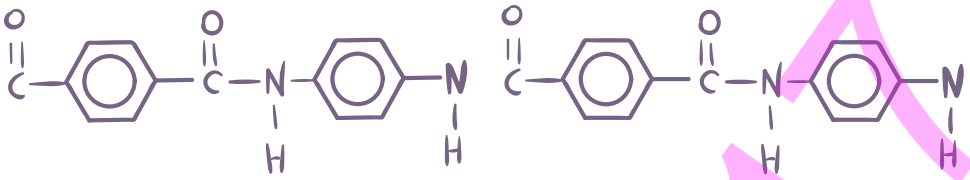


Figura 104. Localización de almacenamiento de datos en Geneious.

3.8.5. Panel de Fuentes

Figura 105. La ventana principal de Geneious.



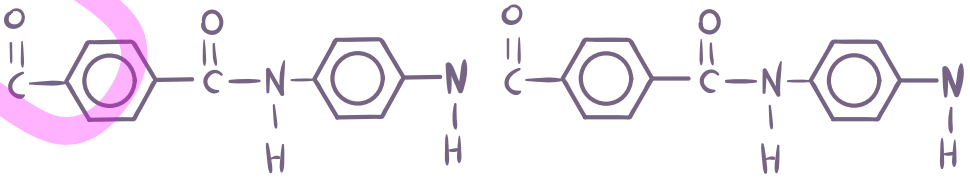
El panel Fuentes (*Sources*) muestra los documentos almacenados y los servicios de Geneious para almacenar y recuperar datos. El símbolo más (+) indica que una carpeta contiene subcarpetas, con acceso a:

- Documentos locales.
- Bases de datos NCBI - *Gene, Genome, Nucleotide, PopSet, Protein, PubMed, SNP, Structure* y *Taxonomy*. Geneious, a través de conexión a Internet, se direcciona a la URL de bases de datos de NCBI, para que el usuario pueda especificar, por ejemplo, qué campo de un documento GenBank se copia en Geneious.
- Bases de datos compartidas, si está configurado.
- Bases de datos Geneious de contactos, si se ha instalado la colaboración.

El panel de visor de documentos muestra el contenido de cualquier documento haciendo clic en el documento y así permitiendo ver secuencias u obtener alineamientos, árboles, estructuras en 3D, resúmenes de artículos de revistas y otros tipos de documentos en una vista de texto o gráfica sencilla. Las opciones para controlar el aspecto y el diseño de un documento dado se muestran en el menú de la derecha.

3.8.6. Mover archivos

Los archivos se pueden mover de varias maneras entre carpetas: arrastrar y soltar, copiar y pegar, o eliminar archivos o carpetas.



3.8.7. Controles generales de visualización

En la barra de herramientas, en la parte superior derecha del visor, existen los controles:

- ***Share*** *(compartir):* Permite compartir la visualización actual en Twitter, Facebook o correo electrónico.
- ***Split View*** *(dividir vista):* Proporciona varias opciones para dividir la vista, por lo que se pueden mostrar varias vistas al mismo tiempo para un documento.
- ***Expand*** *(ampliar vista):* Expande el panel de vista del documento para llenar la ventana principal ocultando el panel de fuentes. Al hacer clic nuevamente, regresa al diseño original.
- ***New window*** *(nueva ventana)*: Abre otra vista del documento actual en una ventana separada.
- ***Help*** *(ayuda):* Abre el Panel de Ayuda.

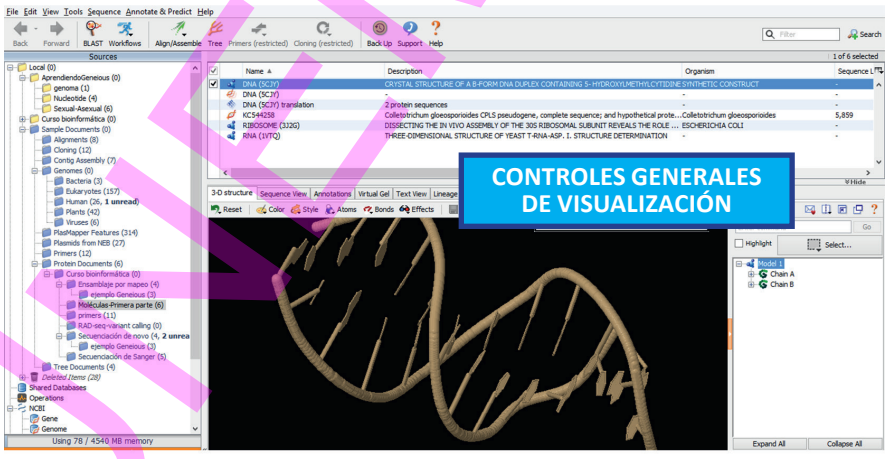


Figura 106. Vista general de los controles en Geneious.

3.8.8. La barra de menú de Geneious

Se puede acceder a las funciones en Geneious desde la barra de menú, dividido en siete submenús:

- ***File*** *(archivo):* Se encuentran opciones para la gestión de archivos, administración de carpetas, copias de seguridad de archivos, etc.
- ***Edit*** *(editar):* Contiene opciones para encontrar y cambiar el nombre de los documentos y su contenido.
- ***View*** *(ver):* Presenta varias opciones y comandos para cambiar la forma de ver los datos en Geneious, como regresar a los documentos que se hayan seleccionado anteriormente, o controles generales de visualización.
- ***Tools*** *(herramientas):* Presenta opciones para Alinear / Ensamblar, Árbol, Clonación, BLAST, Concatenar secuencias o Alineaciones, Plugins y Preferencias.
- ***Sequence*** *(secuencia):* Contiene varias operaciones para la manipulación de secuencias de nucleótidos y proteínas, incluyendo extraer la parte seleccionada de una secuencia o de una alineación en un nuevo documento, invertir dirección de una secuencia, crear un nuevo documento de proteínas a partir del ADN traducido, establecer si las secuencias seleccionadas son circulares, convertir entre ADN y ARN (es decir cambia todas las T de una secuencia a U, o viceversa), etc.
- ***Annotate & Predict*** *(anotar y predecir):* Contiene muchas herramientas para encontrar la predicción y la anotación de regiones de secuencias y alineaciones.
- ***Help*** *(ayuda):* Se compone de las opciones de Ayuda estándar que ofrece Geneious, como Recursos en línea, Comprobación de nuevas versiones de Geneious, Soporte o Activar licencias.

3.8.9. Importar a Geneious archivos de bases de datos públicas

Geneious es capaz de comunicarse con un número de bases de datos públicas alojadas por el NCBI y la base de datos UniProt. Desde 1988, NCBI es un recurso público para obtener información sobre biología molecular y Geneious puede descargar información de nueve de sus bases de datos (Gene, Genome, Nucleotide, Popset, Protein, PubMed, SNP, Structure, Taxonomy) para llevar a cabo búsquedas NCBI BLAST.

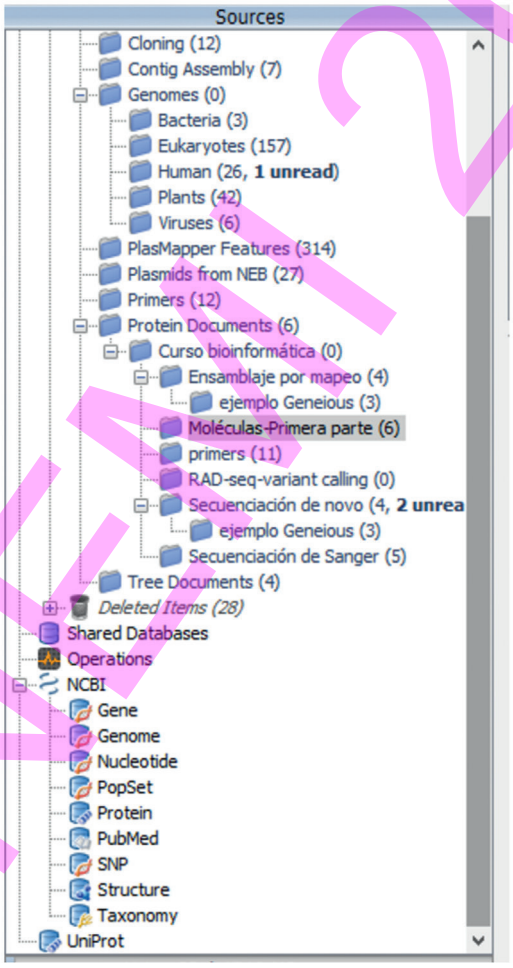


Figura 107. Bases de datos de NCBI a las que puede acceder Geneious.

Estos depósitos de NCBI son utilizados en los datos de biología molecular y están conectados a Geneiuos. Con el fin de activar una licencia, descargar plugins y buscar bases de datos externas como NCBI, Geneious debe tener conexión a Internet y verificar el firewall o el servidor proxy , pudiendo llegar a tener que configurarse manualmente la conexión.

3.8.9.1. Obtención desde Geneious de archivos de secuencias nucleotídicas GenBank en FASTA

1. Ingresar a https://www.ncbi.nlm.nih.gov/

2. Ingresar a *Nucleotide* y consultar, por ejemplo, en *Homo sapiens.*

3. Obtener formatos de archivos de secuencias nucleotídicas GenBank y FASTA.

3.8.9.2. Obtención de secuencias de nucleótidos desde NCBI con Geneious y extracción de un sector

1. Crear la carpeta *AprendiendoGeneious/Nucleotide.*

2. Ingresar a *Nucleotide* de la base de datos de NCBI.

3. Seleccionar una especie biológica (un ejemplo sencillo es el hombre –*Homo sapiens*).

4. Presionar *Search* para conectarse a la base de datos correspondiente.

5. Una vez que se complete la búsqueda, arrastrar la secuencia a la carpeta creada *AprendiendoGeneious/Nucleotide.*

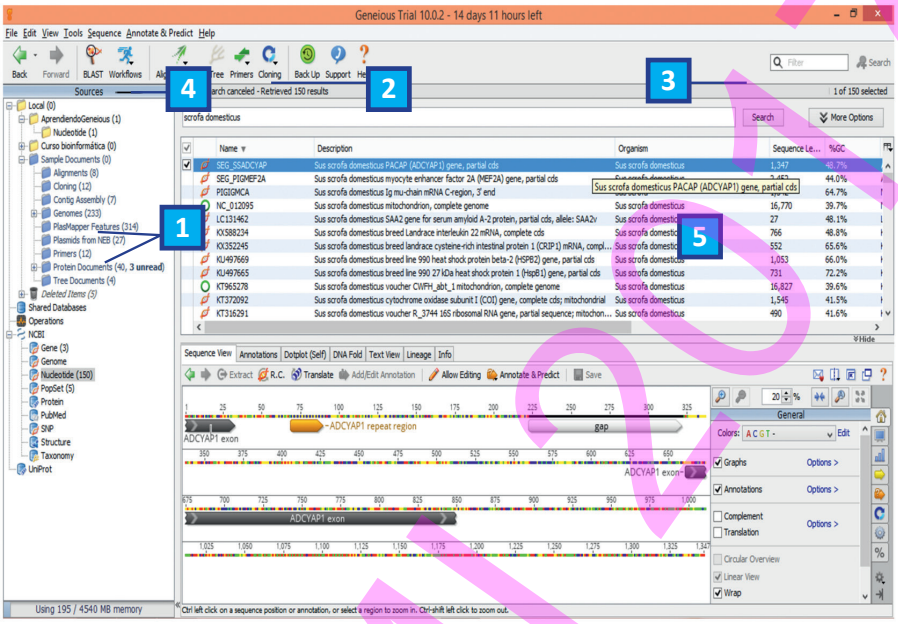


Figura 108. Consulta de secuencias nucleótídicas en NCBI desde Geneious.

6. Para ampliar las secuencias debe utilizarse el zoom (en el caso de la captura de pantalla mostrada se ha ampliado al 90%).

Figura 109. Secuencia obtenida desde Geneious.

7. Seleccionar desde el nucleótido 1 a 10: AATACTGCA

8. Presionar *Extract* para obtener esta selección e incluir un nombre, en este caso *SEG_SSADCYAP Sector1a10.*

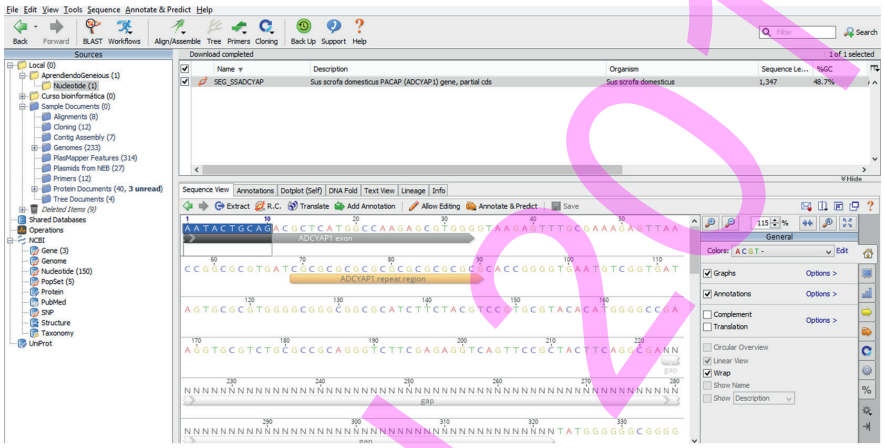


Figura 110. Selección de los nucleótidos 1 al 10 en Geneious.

8. Presionar *Extract* para obtener esta selección e incluir un nombre -en este caso *SEG_SSADCYAP Sector1a1.*

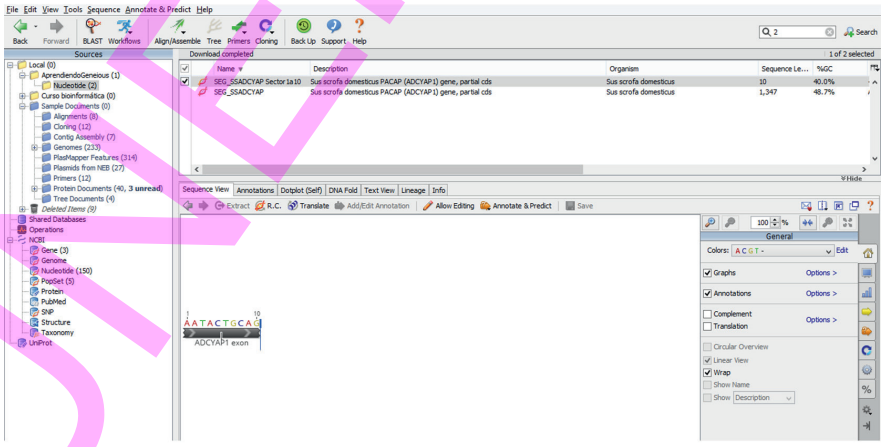


Figura 111. Extracción del de los nucleótidos 1 al 10 del gen PACAP con el descriptor SEG_SSADCYAP Sector1a10.

3.8.10. Alineamiento de secuencias en Geneious

1. Seleccionar los dos sectores *SEG_SSADCYAP Sector1a10* y *SEG_SSADCYAP Sector1a5.*

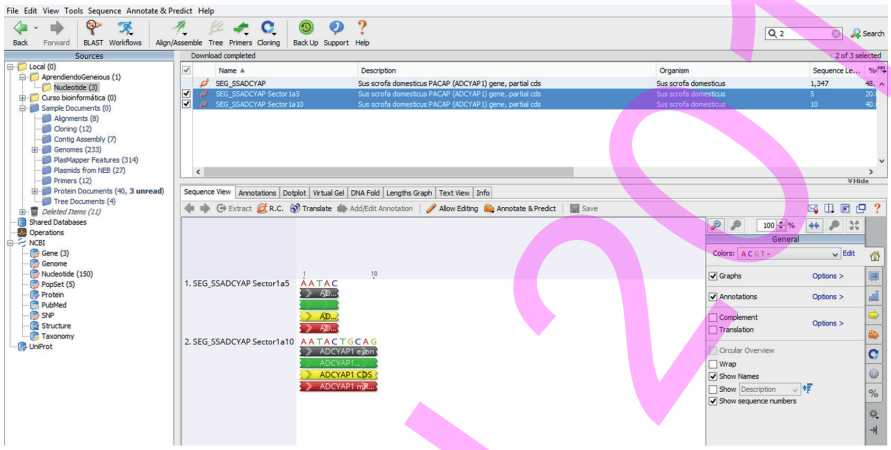


Figura 112. Selección de los dos sectores en una secuencia: SEG_SSADCYAP Sector1a10 y SEG_SSADCYAP Sector1a5.

2. Para el alineamiento, seleccionar *Align/Ensemble* y escoger *Pairwise Align.*

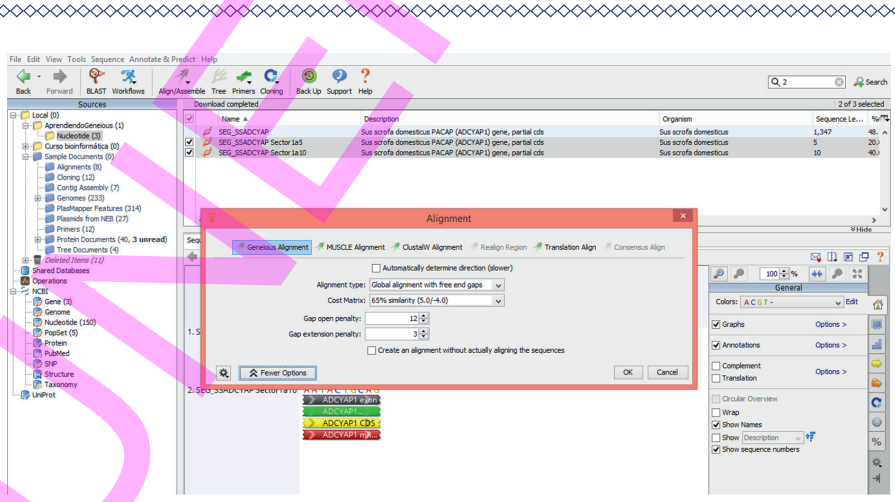


Figura 113. Selección para alineamiento.

3. Se obtiene la siguiente pantalla:

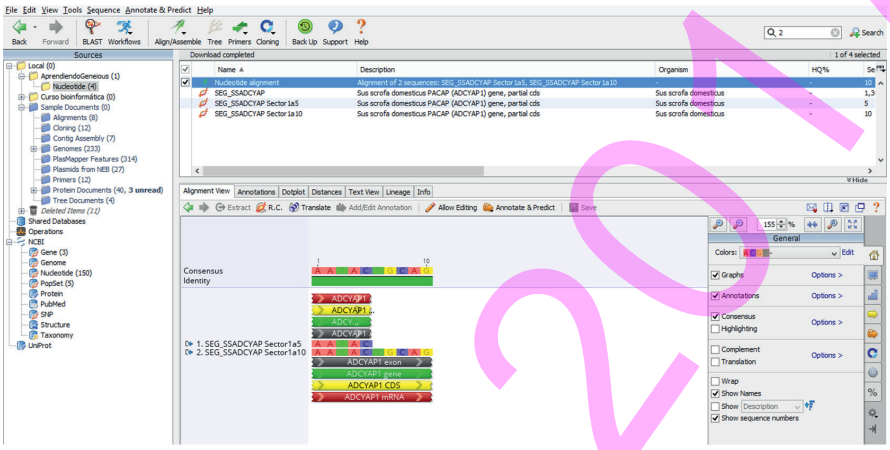


Figura 114. Resultado del alineamiento SEG_SSADCYAP Sector1a10 y SEG_ SSADCYAP Sector1a5.

4. Escoger el tipo de alineamiento.

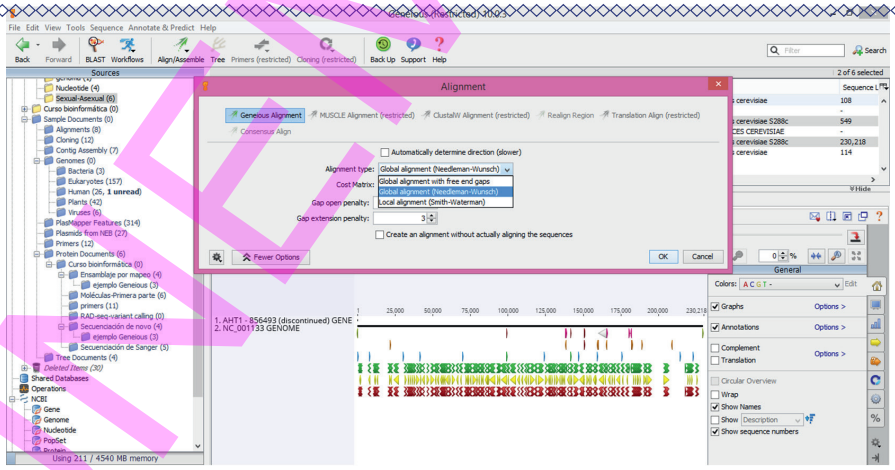


Figura 115. Alineamiento global de Needleman-Wunsches (Ver Sección 4.3.1. Algoritmos globales).

3.8.11. Visualización de genomas completos

1. Descargar en Genious, desde NCBI, el genoma de la mitocondria de *Homo sapiens*, y copiar el archivo en la carpeta *AprendiendoGeneious/genoma.*

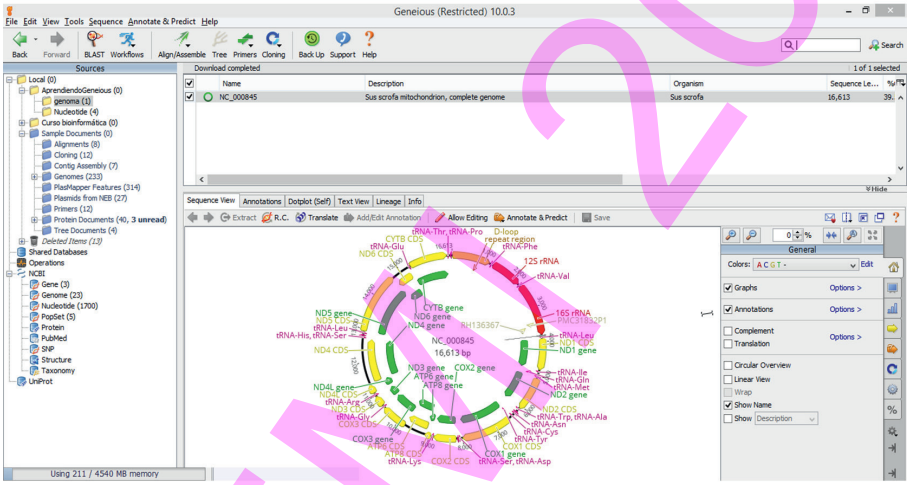


Figura 116. Obtención del genoma mitocondrial de Homo Sapiens desde Geneious.

3.8.12. Diseño de Primers en Geneious

1. Hacer click en la secuencia RP11-416N13 del genoma de *Homo Sapiens*. Hacer clic en el botón *Primers* de las opciones del menú y seleccionar *Design New Primers.*

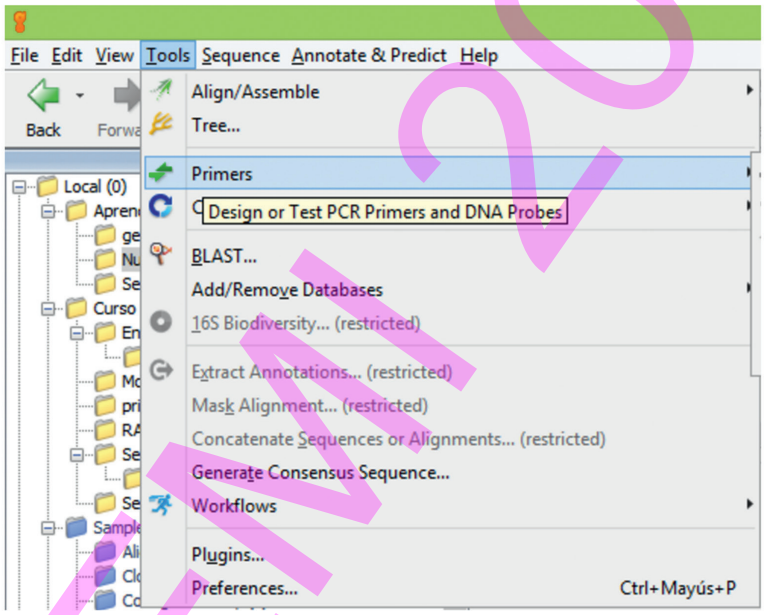


Figura 117. Diseño de los primers en Geneious.

Se puede especificar dónde colocar los primers en la secuencia objetivo, el tamaño de amplicón o la Tm. Para amplificar una parte específica de la secuencia objetivo, establecer en *Included region.*

2. Desmarcar la casilla *Included region* y establecer el tamaño del producto (amplicón) (por ejemplo 200-300 pares de bases –pb– con un tamaño óptimo de 250 pb).

Amplificar una región de menos de 300 pb haría que los cebadores puedan funcionar con ADN degradado (es decir, troceado).

3. Defininir el número de pares de cebadores que se generarán.

4. Se pueden establecer rangos para Tm (*melting temperature*).
Tm calculation proporciona detalles del cálculo del punto de fusión de oligos.

5. En *Characteristics*, también se pueden establecer penalizaciones para los cebadores en función del no cumplimiento de requisitos deseados.

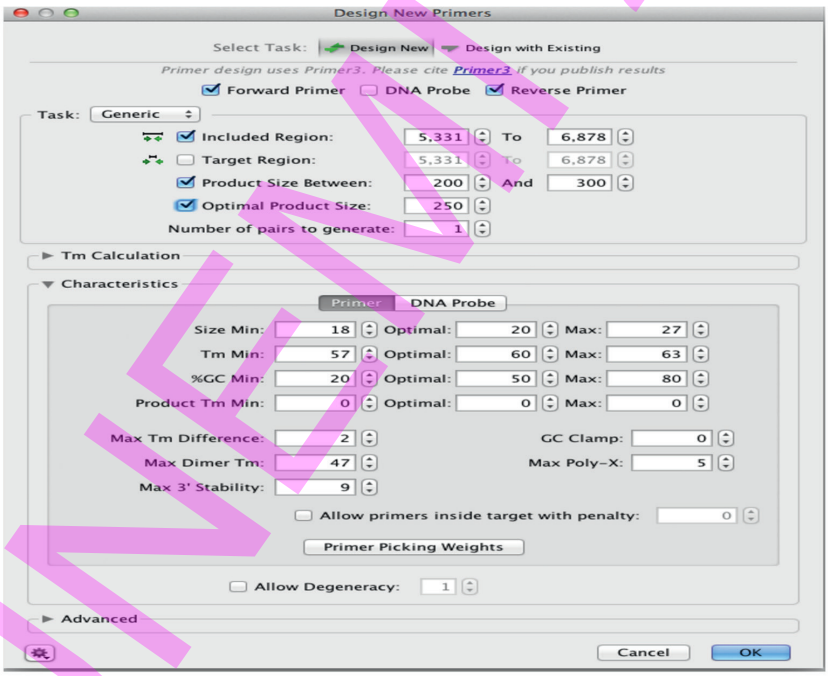


Figura 118. Pantalla para cebadores en Geneious.

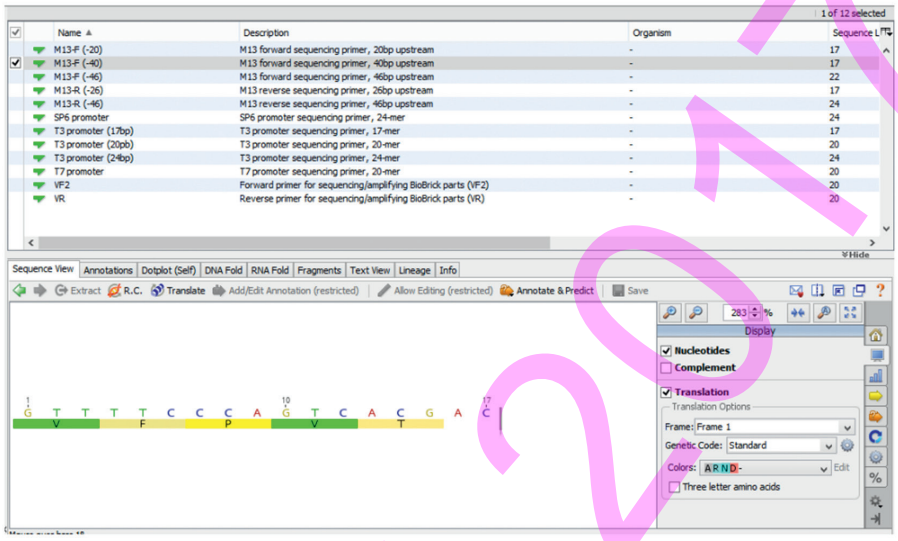


Figura 119. Ejemplo de cebadores obtenidos con Geneious.

3.8.12.1. Otras opciones de Geneious PARA DISEÑO DE PRIMERS

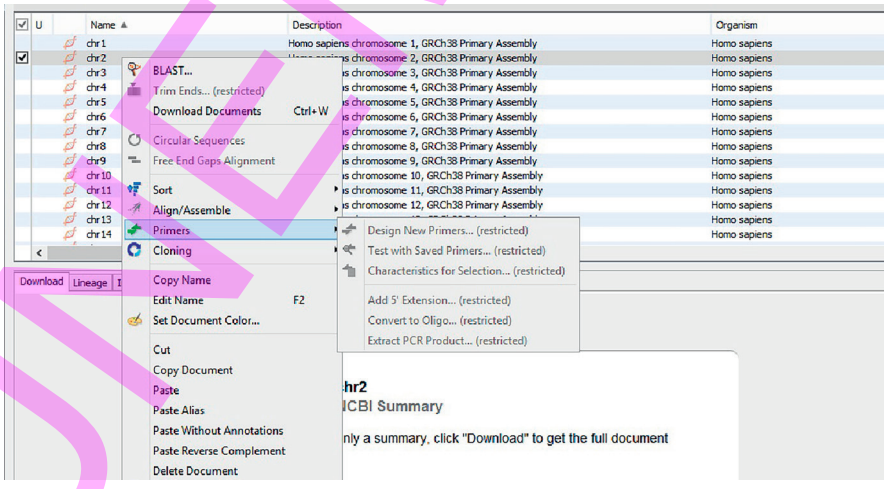


Figura 120. Diseño de primers en Geneoius.

- **Add 5' Extension** *(extender en 5').* Se puede adicionar un corto tramo de secuencia en el extremo 5' del cebador.

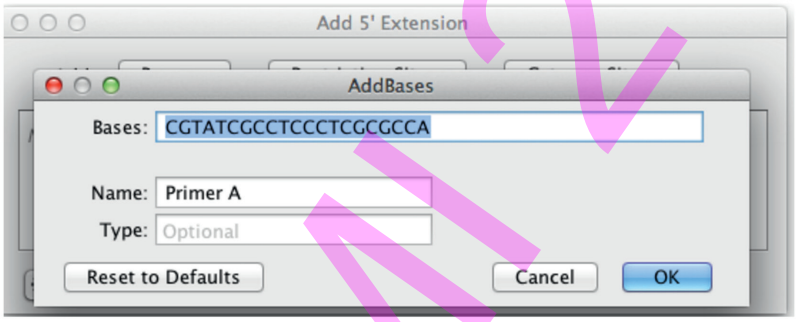


Figura 121. Pantalla para añadir 5' en Geneious.

- **Test with Saved Primers** permite probar los cebadores en la base de datos en la secuencia objetivo y probar la amplificación. Para ello, primero seleccionar *Primers* → *Test with Saved Primers*, y marcar el cuadro *Forward*. Después, hacer clic en *Choose*. La lista que aparece contiene todos los cebadores que hay en la base de datos.

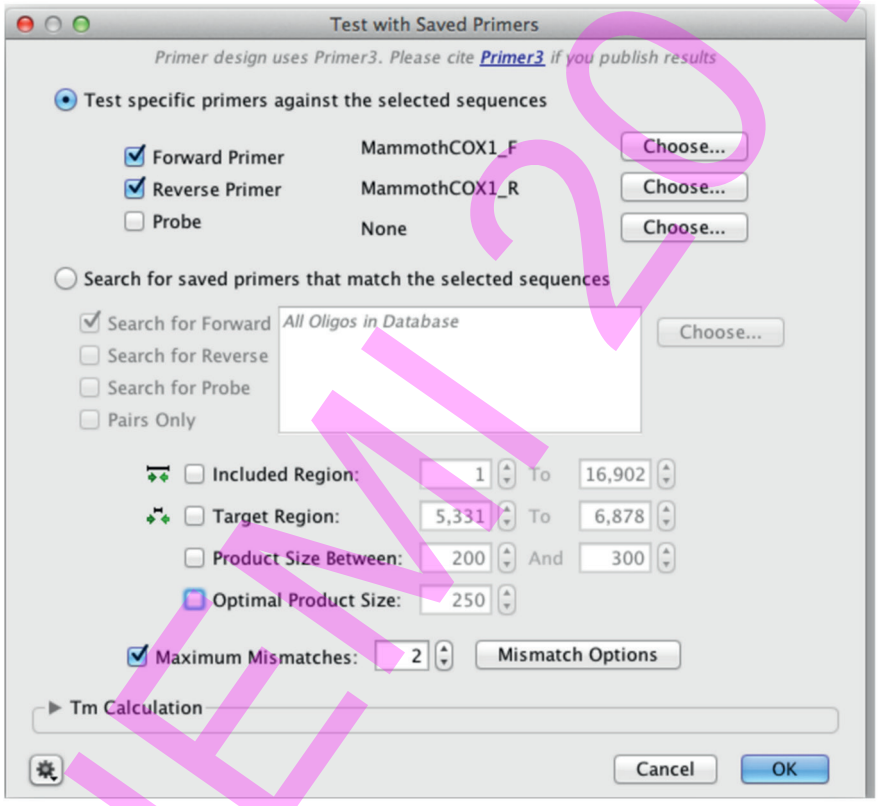
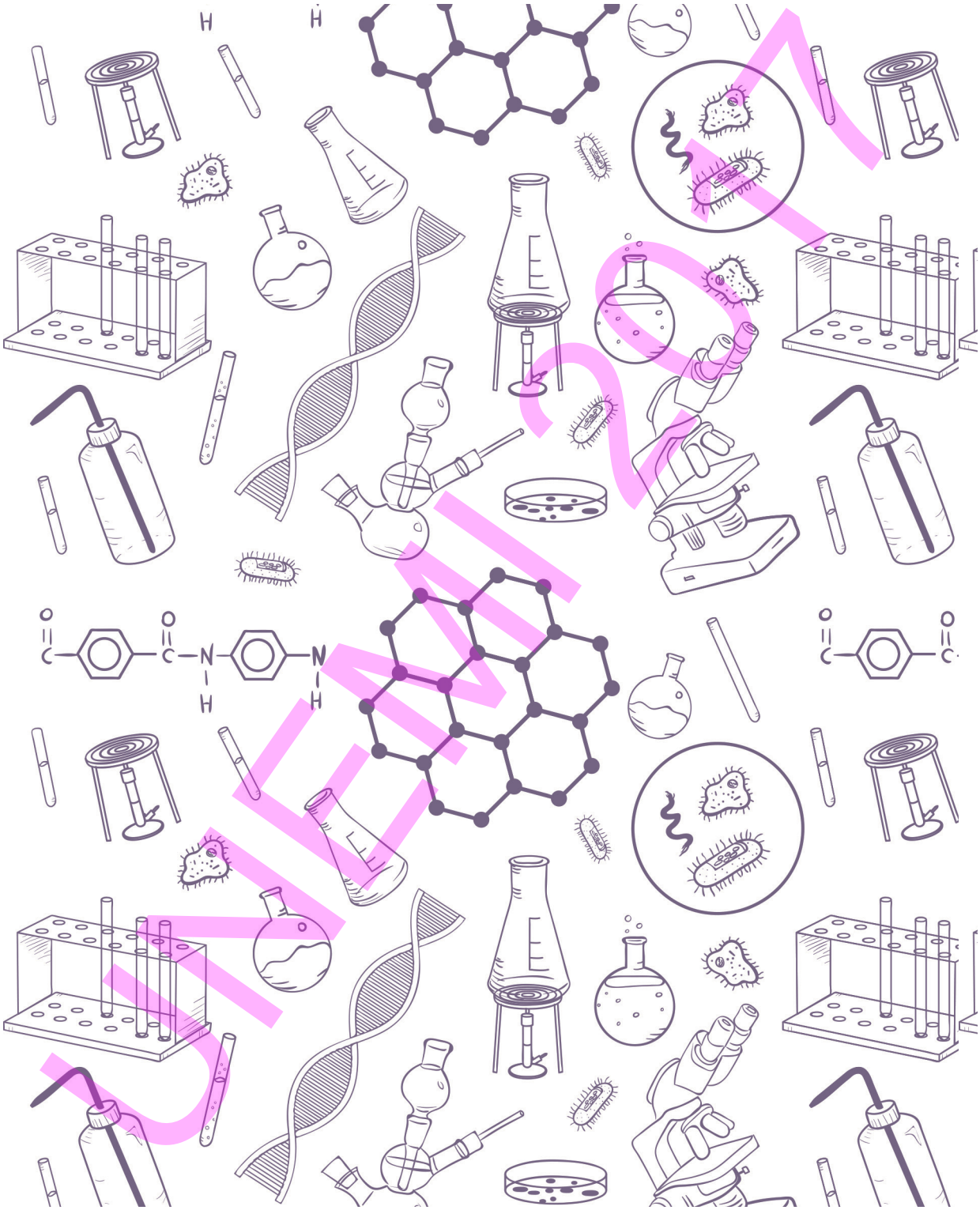


Figura 122. Ejemplo de pantalla Test with Saved Primers.



CAPÍTULO 4:

FUNDAMENTOS DE PROGRAMACIÓN Y HERRAMIENTAS ALGORÍTMICAS EN PROCESAMIENTO DE SECUENCIAS.

Ing. Rafael-Lazo, MGTI.
Ing. Oscar León-Granizo.
Lic. Carlos Noceda-Alonso, PhD.
Lic. Jesennia Cárdenas-Cobo, MBA.
Ing. Mirella Correa-Peralta, MBA.

4. FUNDAMENTOS DE PROGRAMACIÓN Y HERRAMIENTAS ALGORÍTMICAS EN PROCESAMIENTO DE SECUENCIAS.

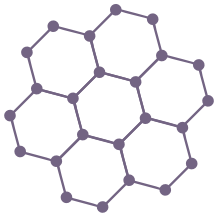
4.1. PROGRAMACIÓN DINÁMICA

En los procesos de análisis e interpretación de resultados de las comparaciones entre secuencias aplican criterios específicos que dependen del tipo de evaluación que se requiera realizar. Para llevar a cabo dichos procesos se utilizan programas informáticos que en su dinámica muestran resultados concretos según las aplicaciones para las cuales fueron creados, o según la naturaleza de la solución a un problema específico.

Los biotécnicos, al analizar datos, llegan a un punto de estancamiento al no poder generar más información acerca de un modelo planteado, al ver que la aplicación, aunque muy útil en muchos aspectos, está orientada en algunos casos a modelos genéricos o a modelos específicos que no son compatibles con el planteado por el investigador.

En este punto se requiere contar con herramientas informáticas que permitan que los modelos creados por los investigadores, que denominaremos modelos bioinformáticos, puedan generar diversas perspectivas de los datos analizados. De esta manera el investigador podrá visualizar las diferentes características de los datos, combinando variables que logren hallar la información relevante que busca en su proceso de investigación.

Para dar solución a problemas específicos se desarrollan herramientas tecnológicas basadas en programas y modelado de datos, que son diseñados a medida, según requiera el investigador. Para el diseño de estas herramientas, el técnico informático o de computación requiere aplicar conceptos asociados a la resolución de este tipo de problemas, originados en las ciencias básicas, que son diferentes a los planteados más comúnmente en automatización de procesos comerciales o industriales, y por tanto las técnicas habituales usadas para resolución de problemas, programación, estructuramodular y/u orientada objeto, no prestan frecuentemente un aporte al reto bioinformático.



La programación dinámica es un método de recursividad utilizado para la resolución de problemas en donde el procesamiento de información y los cálculos asociados consumen tiempo y recursos físicos de los equipos que los procesan. En este tipo de programación, un problema puede generar varias soluciones alternativas, y es seleccionada aquélla cuya sinergia entre tiempo y consumo de procesamiento es óptima. Se consideraría a ésta la solución más adecuada. La solución de problemas mediante esta técnica se basa en el llamado principio de optimización, enunciado por Bellman en 1957 y que dice: "En una secuencia de decisiones óptima toda sub-secuencia ha de ser también óptima".

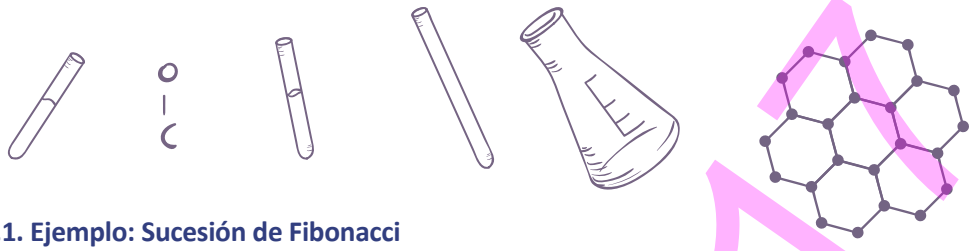
4.2. MODELO DE APLICACIÓN DE UN ALGORITMO EN PROGRAMACIÓN DINÁMICA

1. Planteamiento de la solución como una sucesión de decisiones y verificación de que esta cumple el principio de optimización.

2. Definición recursiva de la solución.

3. Cálculo del valor de la solución óptima mediante una tabla en donde se almacenan soluciones a problemas parciales para reutilizar los cálculos.

4. Construcción de la solución óptima haciendo uso de la información contenida en la tabla mencionada.



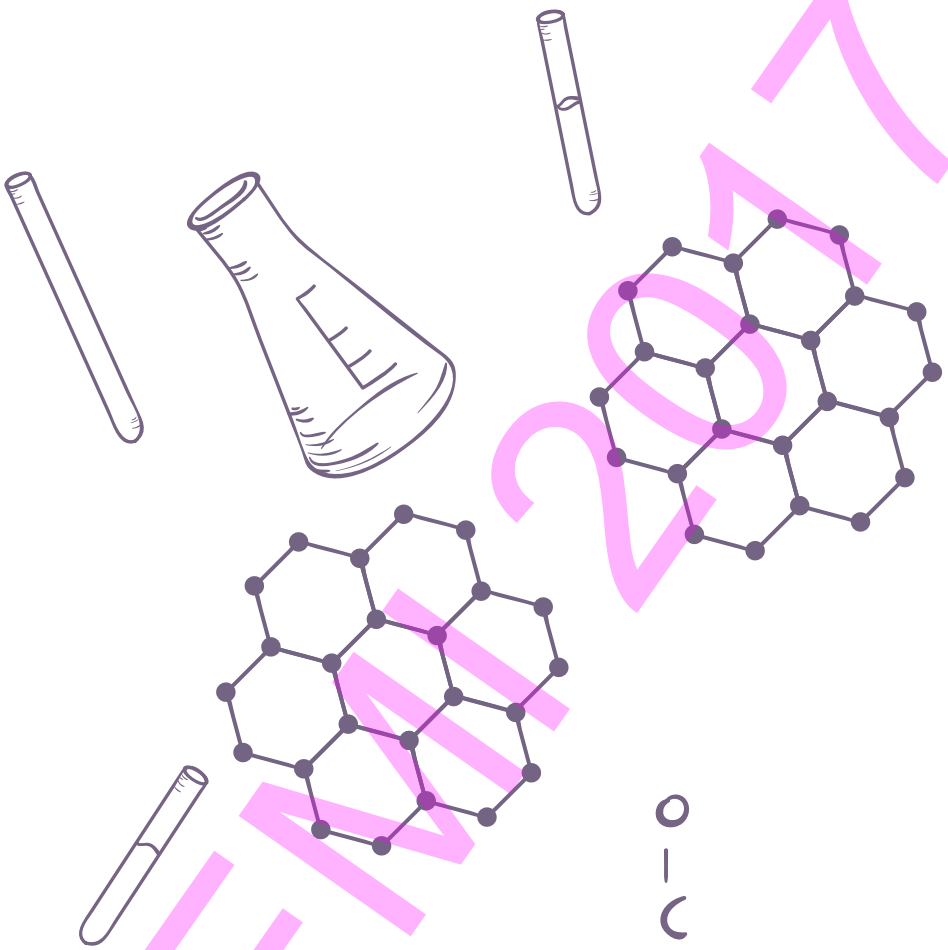
4.2.1. Ejemplo: Sucesión de Fibonacci

Apliquemos un ejemplo en donde la programación dinámica busca la solución más óptima en el problema planteado. El siguiente problema es paradigmático en la aplicación de este tipo de programación:

La sucesión de Fibonacci puede expresarse en términos matemáticos como una secuencia recursiva; desarrollar un algoritmo que genere esta sucesión.

Aplicando las técnicas de programación estructurada, modular, orientada a objetos, utilizando incluso lenguaje unificado de modelado, que comúnmente se usan para resolver problemas, consideraríamos como solución el siguiente código con base en lenguaje C/C++/Java:

```
int T[n];
int FibIter(int *T, int n)
{
   int i;
   if (n<=1)
      return(1);
   else
   {
     T[0]=1;
     T[1]=1:
     for (i:=2;i<n;i++)
     {
        T[i]=T[i-1]+T[i-2];
     }
     RETURN T[n];
    }
}
```



Aunque el código planteado genera la secuencia de la sucesión, el algoritmo no es óptimo, ya que el consumo de tiempo que genera y los recursos de procesamiento utilizados crecen en forma exponencial. Esto se debe a que se producen invocaciones a la función de forma recursiva repetida para calcular valores de la secuencia que, habiéndose calculado previamente, no son guardados, y por ello es necesario volver a calcular de nuevo.

Una segunda solución podría ser el uso de arreglos que permitan no volver a calcular desde el primer paso, cada vez que se quiera iniciar una secuencia.

Así, se puede seguir mejorando el código, ya que únicamente son necesarios los dos últimos valores calculados para determinar cada término, lo que permite

eliminar la tabla entera y quedarnos solamente con dos variables para almacenar los dos últimos términos:

```
    int FibIter2(int n)
    {
    int i,suma,x,y;
    if (n<=1)
    return (1);
    else
    {


     x=1;
     y=1;
     for (i=2; i<n;i++)
     {
          suma=x+y;
          y=x;
          x=suma;
      }
      return (suma);
  }
}
```

En resumen, el mismo problema puede tener infinitas soluciones. La programación dinámica trata de seleccionar cuál de ella es la más óptima para la sinergia tiempo y procesamiento.

4.3. ALGORITMOS APLICADOS A ALINEAMIENTOS DE SECUENCIAS

En cuanto a la solución aportada, el algoritmo aplicado para efectuar alineamientos puede ser:

- **Determinístico:** Optimiza tanto el tiempo como los recursos, y la solución ofrecida se considera la más adecuada, siendo determinada por las entradas.

- **Heurístico:** No necesariamente rinde la velocidad ni la solución óptimas.

En cuanto a la aproximación computacional al problema del alineamiento, el algoritmo aplicado puede ser:

- **Global:** Las secuencias a alinear se relacionan en una matriz de comparación en la que el resultado es determinístico. Estos algoritmos son más rigurosos pero ocupan mayor memoria y tiempo de procesamiento. Los algoritmos globales son utilizados más comúnmente en el ordenamiento de dos secuencias para observar similitudes.

- **Local:** Las secuencias se dividen en partes para comparar, siendo heurísticos al ser de resultados múltiples, pero no hay certeza de que el alineamiento resultante sea el mejor. Los algoritmos locales son más útiles cuando las secuencias son muy diferentes.

4.3.1. Algoritmos globales

Como ejemplo de algoritmo global puede tomarse el de Needleman-Wunsch, propuesto por primera vez en 1970, y muy utilizado para alinear dos secuencias de proteínas o de ácidos nucleicos. El algoritmo parte de aquéllas secciones que son iguales entre secuencias, y se desarrolla como se indica a continuación.

4.3.1.1. Desarrollo

a) Inicialización

1. Crear la matriz base S o de similitud entre los símbolos que aparecen en las secuencias. Es decir, cada valor S(u,v) en su respectiva celda (u,v) indicará la similitud entre los símbolos u y v del alfabeto usado.

2. Establecer el valor de penalización (correspondencia de elementos de una secuencia cuando aparea con huecos de la otra)

3. Colocar las secuencias a ordenar en una tabla de doble entrada, dejando en ambas secuencias un hueco como primer elemento.

- Secuencia 1: fila A[1;:::;n+1] (proporcionará las filas de una matriz).
- Secuencia 2: columna B[1;:::;m+1] (proporcionará las columnas de una matriz).

4. Generar en la tabla de doble entrada una matriz M, de m+1 filas y n+1 columnas.

b) Llenado de la matriz

1. El método más sencillo parte de asignar el valor de 0 a las celdas de la primera fila y de la primera columna.

2. Para el resto de las celdas (i,j), seleccionar como valor M(i,j) el mayor entre los siguientes valores:

a. M(i-1,j-1) + S(u,v), siendo u y v los símbolos correspondientes, respectivamente, a la columna B y fila A de la tabla de doble entrada.

b. M(i-1,j) + penalización

c. M(i,j-1) + penalización

c) Determinación del camino mediante un proceso de TRACE-BACK

1. Se establecerá un conjunto de celdas que denominaremos "***camino***", que se iniciará a partir de la celda inferior derecha (m,n)

2. Determinar de cuál de las tres celdas contiguas utilizadas en el cálculo del valor de la celda (i,j) –es decir, de las celdas M(i-1,j-1), M(i-1,j) y M(i,j-1)- proviene dicho valor. En este cálculo podrían llegar a obtenerse más de un ***camino.***

d) Alineamiento

1. Marcar el recorrido del camino generado para encontrar el alineamiento óptimo, de izquierda a derecha y en función del sentido del recorrido.

2. Hacer una matriz con dos filas.

3. Escribir en la matriz los símbolos correspondientes a las secuencias filas y columnas de acuerdo al siguiente criterio:

a. Si se produce un avance hacia la diagonal izquierda, se colocan:

- En la parte superior, el símbolo de la secuencia de la FILA
- En la parte inferior: el símbolo de la secuencia de la COLUMNA

b. Si se produce un avance hacia la izquierda, se colocan:

- En la parte superior: hueco en la secuencia FILA
- En la parte inferior: el símbolo de la secuencia de la COLUMNA

c. Si se produce un avance hacia arriba, se colocan:

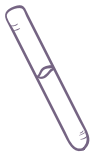
- En la parte superior: el símbolo de la secuencia de la FILA
- En la aparte inferior: hueco en la secuencia de la COLUMNA

4.3.1.2. Ejemplo práctico de un algoritmo de Needleman-Wunsch

Partiendo de las siguientes dos secuencias:

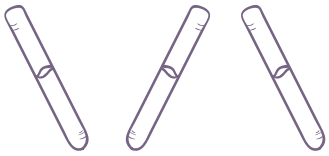
AGGGATTCCGG

GACCATGAG



a) **Inicialización**

Se obtiene como matriz base:



	A	C	G	T
A	3	-1	1	-1
C	-1	3	-1	1
G	1	-1	3	-1
T	-1	1	-1	3

Otorgamos un valor de penalización de -5

b) Llenado

Se obtiene:

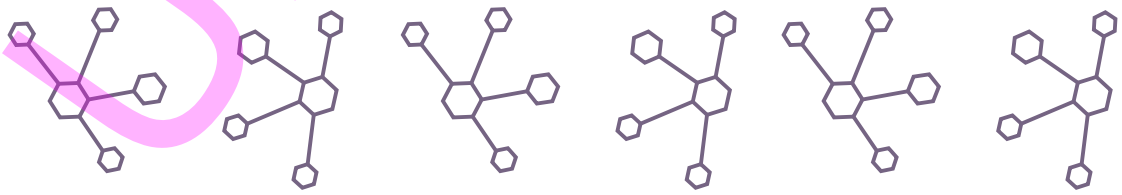
		A	G	G	G	A	T	T	C	C	G	G
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
A	0											
C	0											
C	0											
A	0											
T	0											
G	0											
A	0											
G	0											

Para determinar los valores de la matriz, se procede a calcular tal como fue definido en el algoritmo, siendo los valores para la primera fila los siguientes:

i	J	Cálculo	ASIGNAMOS VALOR DE
1	1	VALOR DIAGONAL ZIQUIERDA ARRIBA: 0 +(1)=1 VALOR IZQUIERDO: 0 + (-5)= -5 VALOR ARRIBA: 0 + (-5)= -5	1
1	2	↖:0 +(3)=3 ←:1+(-5)=-4 ↑:0+(-5)= -5	3
1	3	↖:0 +(3)=3 ←:3+(-5)=-2 ↑:0+(-5)= -5	3
1	4	↖:0 +(3)=3 ←:3+(-5)=-2 ↑:0+(-5)= -5	3
1	5	↖:0 +(1)=1 ←:3+(-5)=-2 ↑:0+(-5)= -5	1
1	6	↖:0 +(-1)=-1 ←:1+(-5)=-4 ↑:0+(-5)= -5	-1
1	7	↖:0 +(-1)=-1 ←:-1+(-5)=-6 ↑:0+(-5)= -5	-1
1	8	↖:0 +(-1)=-1 ←:-1+(-5)=-6 ↑:0+(-5)= -5	-1
1	9	↖:0 +(-1)=-1 ←:-1+(-5)=-6 ↑:0+(-5)= -5	-1
1	10	↖:0 +(3)=3 ←:-1+(-5)=-6 ↑:0+(-5)= -5	3
1	11	↖:0 +(3)=3 ←:3+(-5)=-2 ↑:0+(-5)= -5	3

Aplicando los cálculos para la primera fila se obtiene:

		A	G	G	G	A	T	T	C	C	G	G
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	3	3	3	1	-1	-1	-1	-1	3	3
A	0											
C	0											
C	0											
A	0											
T	0											
G	0											
A	0											
G	0											



Después de realizar el barrido completo de la matriz, fila por fila, se habrán asignado todos los valores correspondientes:

		A	G	G	G	A	T	T	C	C	G	G
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	3	3	3	1	-1	-1	-1	-1	3	3
A	0	3	2	4	4	6	1	-2	-2	-2	-2	4
C	0	-1	2	1	3	3	7	2	1	1	-3	-1
C	0	-1	-2	1	3	2	4	8	5	4	0	-4
A	0	3	0	-1	2	3	1	3	7	4	5	1
T	0	-1	2	-1	-2	1	6	4	4	8	3	4
G	0	1	2	5	2	-1	1	5	3	3	11	6
A	0	3	2	3	6	5	0	0	4	2	6	12
G	0	1	6	5	6	7	4	-1	-1	3	5	9

Finalmente se establece el camino óptimo, que se inicia en la celda que se encuentra en la parte inferior derecha. El recorrido se realiza seleccionando la celda de la que proviene (arriba, derecha, diagonal) el valor actual de la celda. Este proceso se sigue realizando sucesivamente, hasta alcanzar una celda de la primera fila.

		A	G	G	G	A	T	T	C	C	G	G
	0	0	0	0←	0	0	0	0	0	0	0	0
G	0	1	3	3	↖3	1	-1	-1	-1	-1	3	3
A	0	3	2	4	4	↖6	1	-2	-2	-2	-2	4
C	0	-1	2	1	3	3	↖7	2	1	1	-3	-1
C	0	-1	-2	1	3	2	4	↖8	5	4	0	-4
A	0	3	0	-1	2	3	1	3	↖7	4	5	1
T	0	-1	2	-1	-2	1	6	4	4	↖8	3	4
G	0	1	2	5	2	-1	1	5	3	3	↖11	6
A	0	3	2	3	6	5	0	0	4	2	↑6	12
G	0	1	6	5	6	7	4	-1	-1	3	5	↖9

d) Alineamiento

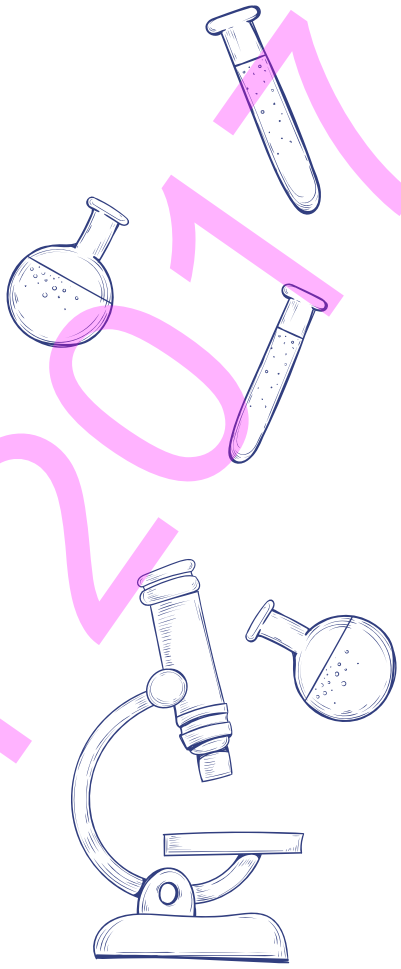
Por lo tanto, las secuencias quedan ordenadas de la siguiente manera:

←	↖	↖	↖	↖	↖	↖	↖	↑	↖
0	3	6	7	8	7	8	11	6	9
-	G	A	T	T	C	C	G	G	G
G	A	A	C	C	A	T	G	-	G

De lo expuesto se sugiere la siguiente **pseudo-codificación** para la aplicación del algoritmo Needleman-Wunsch:

```
for i=0 to length(A)-1
    F(i,0) <- d*i
  for j=0 to length(B)-1
    F(0,j) <- d*j
  for i=1 to length(A)
    for j = 1 to length(B)
    {
      Choice1 <- F(i-1,j-1) + S(A(i), B(j))
      Choice2 <- F(i-1, j) + d
      Choice3 <- F(i, j-1) + d
      F(i,j) <- max(Choice1, Choice2, Choice3)
    }
AlignmentA <- “”
  AlignmentB <- “”
  i <- length(A)
  j <- length(B)
  while (i > 0 AND j > 0)
  {
    Score <- F(i,j)
    ScoreDiag <- F(i - 1, j - 1)
    ScoreUp <- F(i, j - 1)
    ScoreLeft <- F(i - 1, j)
    if (Score == ScoreDiag + S(A(i), B(j)))
    {
      AlignmentA <- A(i-1) + AlignmentA
      AlignmentB <- B(j-1) + AlignmentB
      i <- i - 1
      j <- j - 1
    }
    else if (Score == ScoreLeft + d)
    {
      AlignmentA <- A(i-1) + AlignmentA
```

```
  AlignmentB <- "-" + AlignmentB
  i <- i - 1
 }
 otherwise (Score == ScoreUp + d)
 {
  AlignmentA <- "-" + AlignmentA
  AlignmentB <- B(j-1) + AlignmentB
  j <- j - 1
 }
}
while (i > 0)
{
 AlignmentA <- A(i-1) + AlignmentA
 AlignmentB <- "-" + AlignmentB
 i <- i - 1
}
while (j > 0)
{
 AlignmentA <- "-" + AlignmentA
 AlignmentB <- B(j-1) + AlignmentB
 j <- j - 1
}
```



4.3.2. Algoritmos locales

El algoritmo de Smith-Waterman, de tipo local, fue propuesto por Temple Smith y Michael Waterman en 1981, y se desarrolla de la misma manera que el algoritmo global de Needleman-Wunsch, con algunas modificaciones en su desarrollo:

4.3.2.1. Modificaciones en el desarrollo del algoritmo de Smith-Waterman respecto al de Needleman-Wunsch

1. En el llenado de la matriz M que se genera a partir de las dos secuencias ***siempre*** se asigna el valor de 0 a las celdas de la primera fila y de la primera columna.

2. Si el valor obtenido para una celda resulta negativo, entonces dicho valor se sustituye por 0.

3. El ***camino*** se construye a partir del valor más alto de la tabla, y se detiene en la celda antes de llegar a un 0 (como en el caso de Needleman-Wunsch, podría llegar a obtenerse más de un ***camino***).

4.3.2.2. Ejemplo práctico del algoritmo de Smith-Waterman

Basado en Pérez M. (ver Referencias Bibliográficas).

Sean las siguientes las dos secuencias a alinear:
ABCDEFG
EMBCFEFGHI

a) **Inicialización**
La matriz base será:

	A	B	C	D	E	F	G	H	I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

Se otorga un valor de penalización de -2

b) Llenado

Aplicando el proceso de llenado se obtiene:

		E	M	B	C	F	E	F	G	H	I
	0	0	0	0	0	0	0	0	0	0	0
A	0										
B	0										
C	0										
D	0										
E	0										
F	0										
G	0										

i	J	Cálculo	ASIGNAMOS VALOR DE
1	1	VALOR DIAGONAL ZIQUIERDA ARRIBA: 0 +(0)=0 VALOR IZQUIERDO: 0 + (-2)= -2 VALOR ARRIBA: 0 + (-2)= -2	0
1	2	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	3	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	4	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	5	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	6	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	7	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0
1	8	↖:0 +(0)=-0 ←:-0+(-2)=-2 ↑:0+(-2)= -2	0
1	9	↖:0 +(0)=0 ←:-0+(-2)=-2 ↑:0+(-2)= -2	0
1	10	↖:0 +(0)=0 ←:0+(-2)=-2 ↑:0+(-2)= -2	0

Aplicando los cálculos para la primera fila se tiene:

		E	M	B	C	F	E	F	G	H	I
	0	0	0	0	0	0	0	0	0	0	0
A	0										
B	0										
C	0										
D	0										
E	0										
F	0										
G	0										

De lo expuesto se sugiere la siguiente **pseudo-codificación** para la aplicación del algoritmo Needleman-Wunsch:

Llenado de la matriz base:

Tras realizar el barrido completo de la matriz, fila por fila, quedan asignados todos los valores:

		E	M	B	C	F	E	F	G	H	I
	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	1	0	0	0	0	0	0
C	0	0	0	0	0	2	0	0	0	0	0
D	0	0	0	0	0	0	2	0	0	0	0
E	0	1	1	0	0	0	0	3	1	0	0
G	0	0	0	1	0	0	1	1	4	2	0
F	0	0	0	0	1	0	0	1	2	5	3

c) Establecimiento del camino óptimo

Finalmente se establece el camino óptimo, que se inicia en la celda que se encuentra en la parte inferior derecha, y continúa como se describe para el algoritmo anterior.

		E	M	B	C	F	E	F	G	H	I
	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	↖ 1	0	0	0	0	0	0	0
C	0	0	0	0	↖ 2	0	0	0	0	0	0
D	0	0	0	0	0	↖ 2	0	0	0	0	0
E	0	1	0	0	0	0	↖ 3	1	0	0	0
F	0	0	1	0	0	1	1	↖ 4	2	0	0
G	0	0	0	1	0	0	1	2	↖ 5	3	1

d) Alinemaiento

Las secuencias quedan ordenadas de la siguiente manera:

↖	↖	↖	↖	↖	↖
1	2	2	3	4	5
B	C	D	E	F	G
B	C	F	E	F	G

4.4. LENGUAJES DE PROGRAMACIÓN EN BIOINFORMÁTICA. ALGORITMOS APLICADOS AL TRATAMIENTO DE SECUENCIAS.

Existen complementos para ciertos lenguajes de programación, que ofrecen funcionalidades orientadas a la bioinformática. A continuación se expone un resumen.

4.4.1. BioPerl

PERL (*Practical Extraction and Report Language* – Lenguaje de Extracción y Reporte Prácticos) es un lenguaje de programación con múltiples aplicaciones.

BioPerl es una aplicación bioinformática escrita en PERL, cuya última versión puede ser descargada desde su sitio web oficial: http://bioperl.org/ (Contreras-Moreira, n.d.)

4.4.1.1. instalación

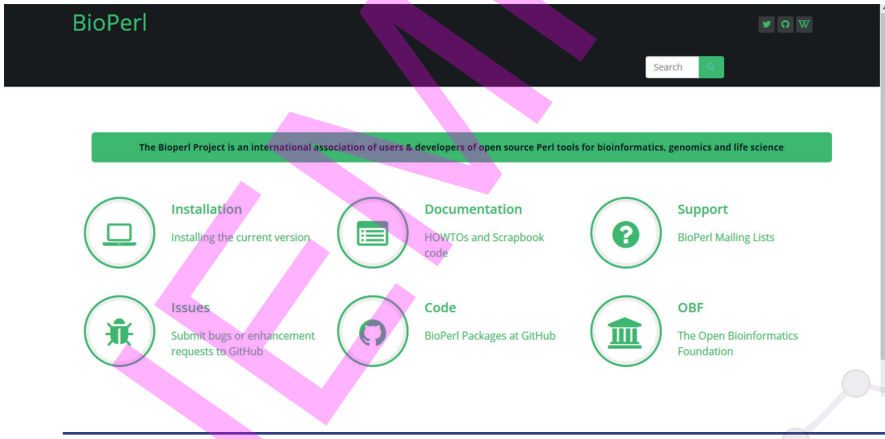


Figura 123. Sitio Web de BioPerl.

Para la instalación típica en Linux, utilizaremos las siguientes instrucciones:

1. Descargar la última versión de BioPerl de su sitio web oficial.

2. Ejecutar la siguiente línea de comandos en una ventana de terminal de comandos:

```
>gunzip bioperl-1.5.2_102.tar.gz
>tar xvf bioperl-1.5.2_102.tar
>cd bioperl-1.5.2_102
```

3. Ocupar los comandos de compilación:
```
>perl Build.PL
>./Build test
```

Si estamos usando Ubuntu como distribución de Linux, la instalación es más sencilla.

1. Abrir una ventana de terminal de comandos y escribimos lo siguiente:
```
sudo apt-get update
sudo apt-get install bioperl
```

4.4.1.2. Ejecución del tutorial

Una vez instalado el módulo de BioPerl, podemos ejecutar el programa bptutorial.pl, que nos muestra las funcionalidades de esta aplicación.

A continuación podrán introducirse los argumentos para trabajar en BioPerl.

1. => sequence_manipulations
2. => seqstats_and_seqwords
3. => restriction_and_sigcleave
4. => other_seq_utilities
5. => run_perl
6. => searchio_parsing
7. => bplite_parsing
8. => hmmer_parsing
9. => simplealign
10. => gene_prediction_parsing

11. => access_remote_db
12. => index_local_db
13. => fetch_local_db
14. => sequence_annotation
15. => largeseqs
16. => liveseqs
17. => run_struct
18. => demo_variations
19. => demo_xml
20. => run_tree
21. => run_map
22. => run_remoteBLAST
23. => run_standaloneBLAST
24. => run_clustalw_tcoffee
25. => run_psw_bl2seq

4.4.1.3. Acceso remoto a bases de secuencias y a listas BLAST

BioPerl permite acceder a diferentes bases de datos, ya sean locales o bien en la web, pero no a ambas al mismo tiempo. Así, es posible la conexión a diferentes bases de datos web de manera secuencial, es decir, si falla la primera conexión la aplicación se conectará a otra base de datos automáticamente, siempre y cuando se haya configurado de esa manera previamente en la consulta (*Query)*, permitiendo así análisis bioinformáticos complejos.

Obtener una secuencia de GenBank
(www.ncbi.nlm.nih.gov/GenBank)

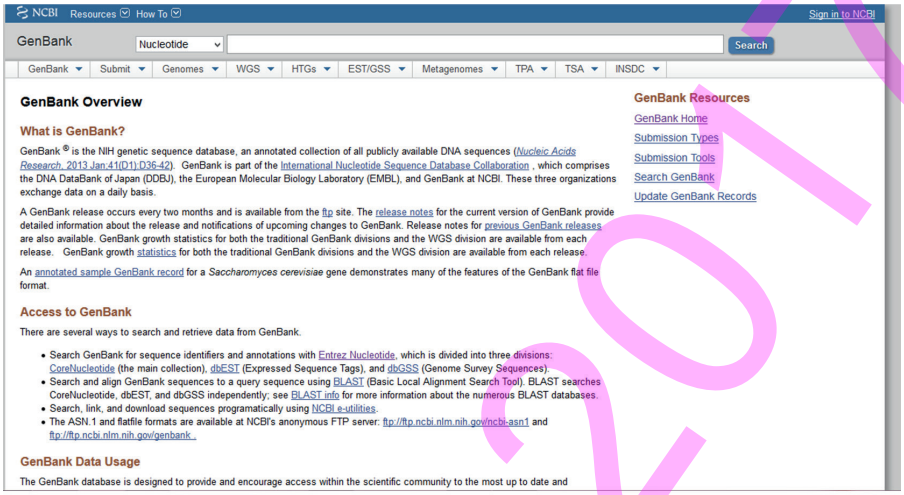


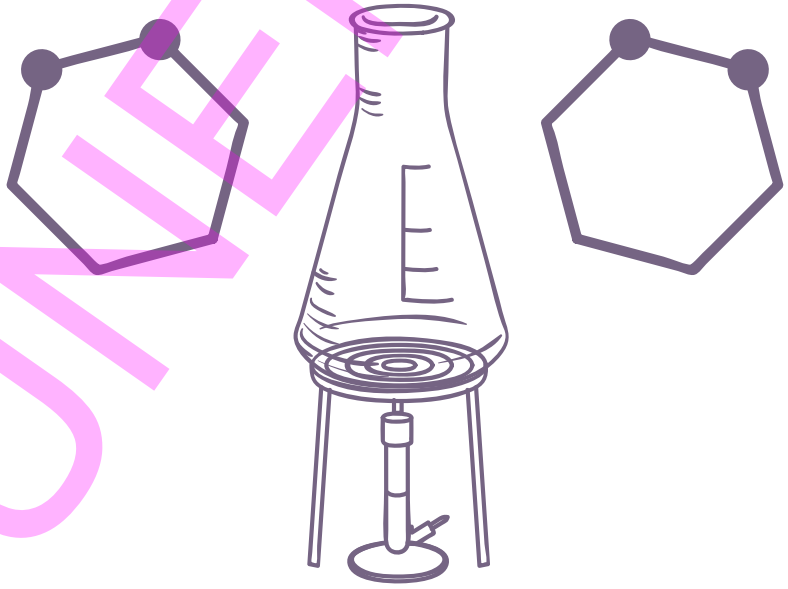
Figura 124. Sitio web de GenBank.

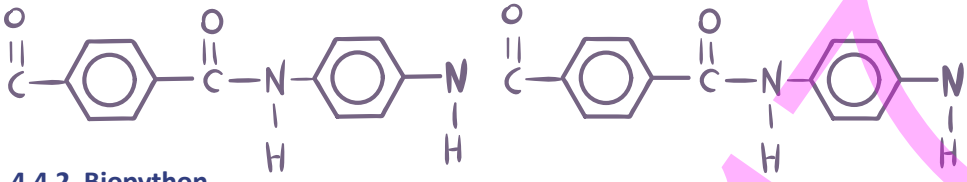
```
#!/usr/bin/perl -w

# Ejemplo para obtener a través de la red y guardar en formato FASTA una secuencia GenBank
use strict;
use Bio::Perl;
my $gi = “NP_417816”;
# ahora una de GenBank
my $secuenciaGenBank = get_sequence( ‘GenBank’, ‘$gi’ );
# escribe la secuencia en formato FASTA a un archivo
write_sequence(“>$gi\.fas”, ‘fasta’, $secuenciaGenBank);
```

Obtener una selección de datos de BLAST (usando SQL con la instrucción *Select*)

```
use strict;
use DBI;
```

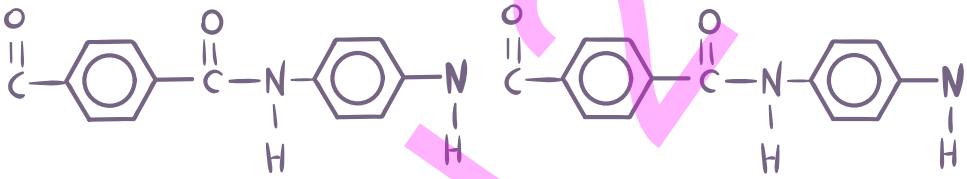
```
my $conexion = DBI->connect(“dbi:mysql:test:servidor.
tuservidor”,”usuario”,”clave”,{PrintError,0,RaiseError,1});
## manipulando la base de datos mediante consultas SQL
my $tabla = “BLAST”;
my $consulta=”SELECT hit FROM $tabla WHERE evalue < 1e-03”;
my $manipulador = $conexion->prepare($consulta);
$manipulador->execute();
# imprime todos los resultados obtenidos de la consulta
$manipulador->dump_results();
# guarda los resultados en un archivo en formato de 80 columnas
$manipulador->execute();
open(CONSULTA,”>resultados.txt”) || die “no puedo crear resultados.txt\n”;
$manipulador->dump_results(80, “\n”, ‘:’, \*CONSULTA);
$conexion->disconnect();
```





4.4.2. Biopython

Python es un lenguaje de programación de alto nivel que se centra principalmente en la facilidad de lectura y escritura de código, contrastando con otros lenguajes de programación que contienen una sintaxis más compleja, como es el caso de PERL. **Biopython** es una librería especializada en datos biológicos. Sus diferentes funciones permiten trabajar con estructuras de macromoléculas o secuencias, entre otros datos. Biopython puede descargarse directamente desde su sitio web: http://biopython.org/wiki/Download



4.4.2.1. Instalación

Podemos instalar Biopython por medio de la ventana de consola de Python (bajo Windows o Linux) o descargar el instalador para Windows (COMAV institute, n.d.).

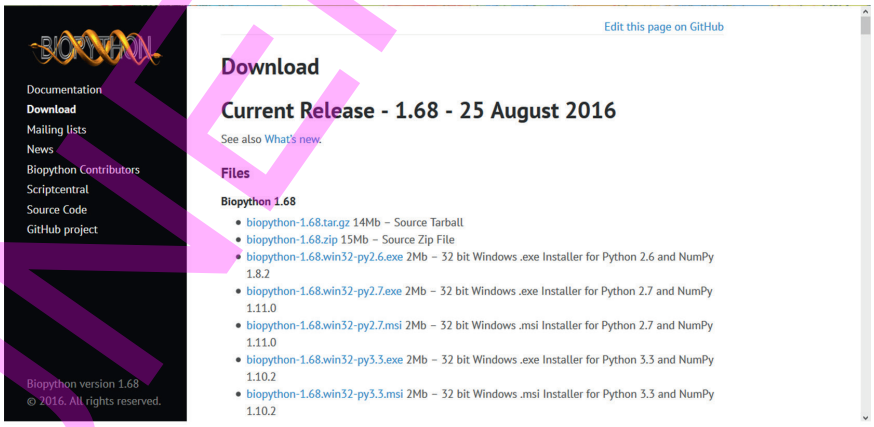
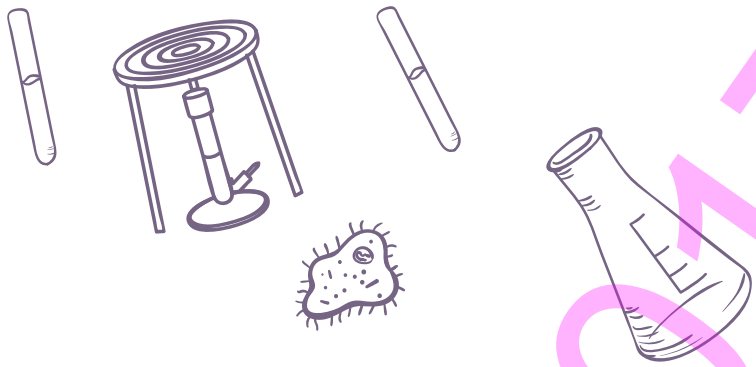


Figura 125. Sitio web de descarga de Biopython.



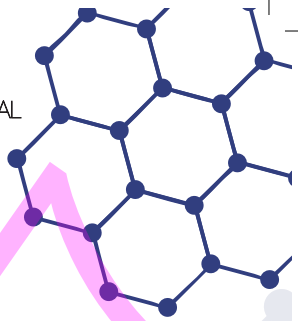
A través de la consola de Python se usa el gestor de paquetes easy_install, ejecutando lo siguiente:

- $ easy_install -f http://biopython.org/DIST/ biopython

En caso de necesitar privilegios *sudo*, ejecutar la siguiente sentencia:
$ sudo easy_install -f http://biopython.org/DIST/ biopython

Para instalar Biopython por medio de instalador de Windows es necesario descargar el archivo Biopython-version.exe que se encuentra en sitio web oficial.

A continuación se exponen algunos ejemplos basados en algoritmos que se encuentran en http://bioinf.comav.upv.es



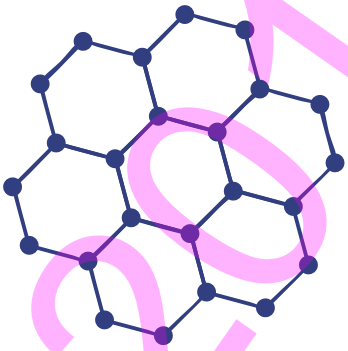
4.4.2.2. Crear un objeto Seq

Para trabajar con una secuencia, como 'ATCG', se crea un objeto Seq, que contiene la cadena de secuencia y un alfabeto:

```
>>> from Bio.Seq import Seq
>>> secuencia = Seq('ATCG')
>>> secuencia
Seq('ATCG', Alphabet())
>>> print secuencia
ATCG //muestra la secuencia asignada
>>> secuencia = Seq('ATCG') 'creamos la secuencia reversa y complementaria de 'ATCG''
>>> secuencia.reverse_complement()
Seq('CGAT', Alphabet())
>>> secuencia = Seq('ATGGCCATTGT') 'traduce la secuencia 'ATGGCCATTGT' a proteina'
>>> secuencia.translate()
Seq('MAI', ExtendedIUPACProtein())
```

4.4.2.3. Cambios de formato de secuencias. Pasar un fichero GenBank a formato FASTA

```
#!/usr/bin/env python
from Bio import SeqIO
def convertir_secuencia(fichero_entrada, formato_entrada,
        i.fichero_salida,  formato_salida):
'Convierte un fichero de secuencia de un formato a otro'
secuencias = SeqIO.parse(fichero_entrada, formato_entrada)
SeqIO.write(secuencias, fichero_salida, formato_salida)
#Alternativamente se podria utilizar la funcion convert
#esto es mas eficiente
#SeqIO.convert(fichero_entrada, formato_entrada,
#        fichero_salida, formato_salida)
```



```
if __name__ == '__main__':
formato_entrada = 'GenBank'
fichero_entrada = 'sequence.gb'
formato_salida  = 'fasta'
fichero_salida  = 'sequence.fasta'
convertir_secuencia(fichero_entrada, formato_entrada,
        ii. fichero_salida,  formato_salida
```

4.4.2.4. Recortar un fichero FASTA

```
from Bio import SeqIO

def recortar_secuencias(in_fname, out_fname, left_clip, right_clip, min_len):

seqs_cortadas = []

for seq in SeqIO.parse(in_fname, 'fasta'):

        left = left_clip
        right = len(seq) - right_clip
        if right <= left:
        #no queda secuencia
        continue
        elif (right - left) < min_len:
        #la secuencia es demasiado corta
        continue
        else:
        seq = seq[left:right]
        seqs_cortadas.append(seq)
SeqIO.write(seqs_cortadas, out_fname, 'fasta')

if __name__ == '__main__':
in_fname = 'seqs.fasta'
out_fname = 'seqs.trimmed.fasta'
left_clip = 50
right_clip = 10
min_len = 40
recortar_secuencias(in_fname, out_fname, left_clip, right_clip, min_len)
```

4.4.2.5. Extraer una lista de secuencias de un grupo de ficheros

```
from Bio import SeqIO
def extraer_secuencias(seq_names, in_fname, out_fname):
seqs_en_input = SeqIO.index(in_fname, 'fasta')
seqs_extraidas = []
for seq_name in seq_names:
una_seq = seqs_en_input[seq_name]
seqs_extraidas.append(una_seq)
SeqIO.write(seqs_extraidas, out_fname, 'fasta')

if __name__ == '__main__':
seq_names = ['seq1', 'seq3', 'seq7']
in_fname = 'seqs.fasta'
out_fname = 'seqs.out.fasta'
extraer_secuencias(seq_names, in_fname, out_fname)
```

REFERENCIAS BIBLIOGRÁFICAS

¿Qué es un árbol filogenético? (2016). Recuperado el 20 de octubre de 2016, de Universidad Nacional de Córdova: http://www.efn.uncor.edu/

Alejandro, P. (s.f.). Nucleótido. Obtenido de Curso de Biología: http://www.bionova.org.es/

Babraham Institute. (s.f.). FastQC. Recuperado el 2 de noviembre de 2016, de Babraham Bioinformatics: https://www.bioinformatics.babraham.ac.uk/projects/

Benson, D., Karsch, I., Lipman , D., Ostell, J., & Wheeler, D. (2007). GenBank. Nucleic Acids Research. Recuperado el 20 de Octubre de 2016, de Vía Clínica: http://viaclinica.com/

Bioinformatics. (s.f.). Bases de datos biológicas. Recuperado el 20 de octubre de 2016, de Bioinformatics at COMAV: https://bioinf.comav.upv.es/courses/intro_bioinf/

Biomatters Limited. (2015). A powerful and comprehensive suite of molecular biology and NGS analysis tools. Obtenido de Geneious: http://www.geneious.com/

Biopython. (2017). Biopython 1.70. Recuperado el 12 de enero de 2017, de Biopython Python Tools for Computational Molecular Biology: http://biopython.org/

Blais, S., Kornblatt, J., Barbeau, X., Bonnaure, G., Lagüe, P., & Lapointe, J. (10 de abril de 2015). tRNAGlu increases the affinity of glutamyl-tRNA synthetase for its inhibitor glutamyl-sulfamoyl-adenosine, an analogue of the aminoacylation reaction intermediate glutamyl-AMP: mechanistic and evolutionary implications. (PubMed, Productor) doi:10.1371

BsubCyc. (2016). BsubCyc Overview. Obtenido de https://bsubcyc.org/

Centro de Bioinformática - Instituto de Biotecnología. (2006). MultAlignment CBIB. Obtenido de EMBnet Colombia: http://bioinf.ibun.unal.edu.co/

Clancy, S. (2008). Chemical Structure of RNA. Obtenido de Scitable by Nature Education: https://www.nature.com/

Clancy, S., & Brown, W. (2008). Translation: DNA to mRNA to Protein. Obtenido de Scitable by Nature Education: https://www.nature.com/

Cohen, J. (2004). Bioinformatics—An Introduction for Computer Scientists. En J. Cohen, ACM Computing Surveys (Vol. 35, págs. 122-158). Obtenido de https://www.cs.indiana.edu/

Cohen, W. (2007). How Cells Work . En W. Cohen, A Computer Scientist's Guide to Cell Biology.

COMAV Institute. (s.f.). Paquetes útiles en biología. Recuperado el 9 de noviembre de 2016, de Bioinformatics at COMAV: https://bioinf.comav.upv.es/

Contreras Moreira, B. (2016). Perl en bioinformática. Recuperado el 9 de noviembre de 2016, de Laboratory of Computational & Structural Biology: http://www.eead.csic.es/compbio

EcoCyc. (s.f.). EcoCyc E. coli Database. Recuperado el 2 de octubre de 2016, de EcoCyc, a member of the BioCyc database collection: https://ecocyc.org/

Embley, M., & Martin, W. (2006). Eukaryotic evolution, changes and challenges. En Nature (págs. 623-630). doi:10.1038/nature04546

ESUKADI. (2016). Los Órganos Celulares. Recuperado el 20 de octubre de 2016, de Hiru: http://www.hiru.eus/biologia/los-organulos-celulares

Fraga, M., Ballestar, E., Paz, M., Ropero, S., Setien, F., Ballestar, M., . . . Spector, T. (26 de julio de 2005). Epigenetic differences arise during the lifetime of monozygotic twins. Obtenido de PubMed: https://www.ncbi.nlm.nih.gov/pubmed/16009939

Guerequeta, R., & Vallecillo, A. (2000). Técnicas de Diseño de Algoritmos (segunda ed.). Malagá. Obtenido de http://www.lcc.uma.es/

HumanCyc. (2016). BioCyc Database Collection. Recuperado el 18 de octubre de 2016, de BioCyc: https://biocyc.org/

HumanCyc. (s.f.). HumanCyc: Encyclopedia of Human Genes and Metabolism. Recuperado el 18 de octubre de 2016, de HumanCyc, a member of the BioCyc database collection: https://humancyc.org/

Hunter, L. (2011). Molecular Biology for Computer Scientists. En Artificial Intelligence and Molecular Biology. doi:10.1609/aimag.v11i4.867

IBM. (s.f.). ¿Qué es BLAST? Recuperado el 2 de noviembre de 2016, de IBM: https://www.ibm.com/support/knowledgecenter/es

Iglesias, G. (2008). Desde Mendel hasta las Moléculas. Recuperado el 20 de septiembre de 2016, de Gen Molecular: https://genmolecular.com/principios-debiologiamolecular/

Illumina. (2016). Next-Generation Sequencing (NGS). Recuperado el 14 de noviembre de 2016, de Illumina: https://www.illumina.com/science/technology/next-generation-sequencing.html

Instituto de Ciencias Forenses. (s.f.). Tipos de marcadores genéticos. Recuperado el 2 de noviembre de 2016, de Universidad de Santiago de Compostela: http://www.usc.es/

International Therapeutics. (s.f.). International Therapeutics. Recuperado el 30 de septiembre de 2016, de http://www.internationaltherapeutics.com/index.html

Kari, L., Kitto, R., & Gloor, G. (2001). A computer scientist´s guide to molecular biology. Soft Computer 5, 95-101.

Karp, P., Weaver, D., Paley, S., Fulcher, C., Kubo, A., Kothari, A., . . . Paulsen, I. (14 de mayo de 2016). The EcoCyc Database. doi:10.1128/ecosalplus.ESP-0009-2013

Mandal, A. (2016). Estructura de ARN. Recuperado el 10 de noviembre de 2016, de News Medical Life Sciences: http://www.news-medical.net/life-sciences/RNA-Structure-(Spanish).aspx

McClements, D., Mohan, A., & Udenigwe, C. (diciembre de 2016). Encapsulation of bioactive whey peptides in soy lecithin-derived nanoliposomes: Influence of peptide molecular weight. Obtenido de Science Direct: http://www.elsevier.com/locate/foodchem

McPherson, J., McCombie, R., & Goodwin, S. (17 de mayo de 2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 333-351. doi:10.1038/nrg.2016.49

Meissner, A., Gnirke, A., Bell, G., Ramsahoye, B., Lander, E., & Jaenisch, R. (13 de octubre de 2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. doi:10.1093/nar/gki901

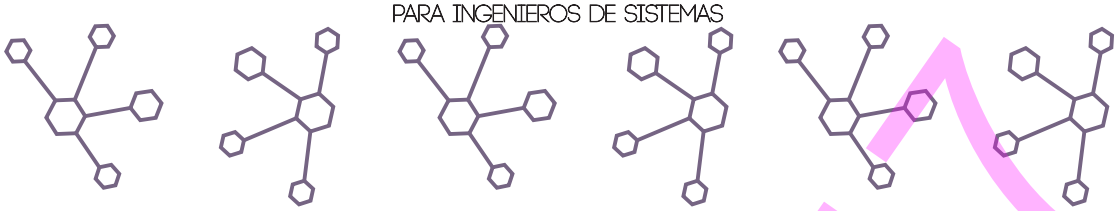
Miller, M., Dunham, J., Amores, A., Cresko, W., & Johnson, E. (22 de diciembre de 2006). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. doi:10.1101/gr.5681207

Nature Education. (2014). Cells Are the Basic Units of Living Organisms. Recuperado el 20 de octubre de 2016, de Scitable by Nature Education: https://www.nature.com/

Nature Education. (2014). Glossary. Recuperado el 5 de noviembre de 2016, de Scitable by Nature Education: https://www.nature.com/

NCBI. (enero de 2013). GenBank Overview. Recuperado el 28 de septiembre de 2016, de https://www.ncbi.nlm.nih.gov/genbank/

Needleman, S., & Wunsch, C. (julio de 1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Recuperado el octubre de 2016, de PubMed: https://www.ncbi.nlm.nih.gov/



Pérez Jiménez, M. (2013). Alineamiento de secuencias de genes/proteínas. Recuperado el 15 de noviembre de 2016, de Técnicas inteligentes en Bioinformática: https://www.cs.us.es/

Planelló Carro, R. (2008). Metabolismo celular: Reacciones anabólicas y catabólicas. Obtenido de Universidad Nacional de Educación a Distancia: http://ocw.innova.uned.es/biologia/

Plaza Hidalgo, I. (1998). Teoría de colas y programación dinámica. (U. N. Distancia, Ed.)

Pray, L. (2008). Discovery of DNA Structure and Function: Watson and Crick. Obtenido de Genetics and Bioengineering: https://gbe.ius.edu.ba/

Pray, L. (2008). Discovery of DNA Structure and Function: Watson and Crick. Obtenido de Nature Education: https://www.nature.com/

Preparata, F. (2000). A Biology Primer for Computer Scientists. Obtenido de Computer Science: https://cs.brown.edu

Regalia, M. (febrero de 2015). Bioinformática - Clase 1 - Introducción al Curso. Bogotá, Colombia. Obtenido de https://www.youtube.com/

Sánchez, A., & Vilardell, M. (s.f.). Modelos de Markov ocultos predicción de genes. Recuperado el 15 de octubre de 2016, de Universidad de Barcelona: http://www.ub.edu/

Santamaría, R. (2016). Secuenciación. Recuperado el 17 de noviembre de 2016, de VisUsal: http://vis.usal.es/

Snape, A., Papachristodoulou, D., Elliott, W., & Elliott, D. (2014). Biochemistry and Molecular Biology (quinta ed.). Oxford.

Stricker, S., Köferle, A., & Beck, S. (21 de noviembre de 2016). From profiles to function in epigenomics. Nature Reviews Genetics, 51-66. doi:10.1038/nrg.2016.138

Subhraveti, P., Altman, T., Keelser, I., Shearer, A., Caspi, R., Ong, Q., & D Karp, P. (12 de julio de 2011). Summary of Escherichia coli, Strain B str. REL606, version 21.0. Obtenido de BioCyc Database Collection: https://biocyc.org/

Tiscornia, M. (2013). Bioinformática. Recuperado el 19 de febrero de 2017, de Universidad Nacional de Misiones: http://www.aulavirtual-exactas.dyndns.org/

UniProt. (s.f.). UniProt Features. Recuperado el 10 de octubre de 2016, de UniProtKB: http://www.uniprot.org/

Vector Base. (s.f.). BLAST. Recuperado el 20 de octubre de 2016, de Vector Base Bioinformatics Resource for Invertebrate Vectors of Human Pathogens: https://www.vectorbase.org/blast

Watson, J., & Crick, F. (30 de mayo de 1953). Genetical Implications of The Structure of Deoxyribonucleic Acid. En Nature (Vol. 171, págs. 964-967). Obtenido de https://www.nature.com/

Young, F., Gerard, H., Jevous, W., Pauling, J., Furberg, S., Corey, R., & Wyatt, G. (1953). Nature. En Nature (págs. 735-738). Obtenido de https://www.nature.com/

Tutoriales de Geneious que se pueden encontrar en:
http://www.geneious.com/tutorials

01-BLAST_Searching.tutorial

02-Accessing_GenBank.tutorial

03-GenBank_submission.tutorial

04-Microsatellites.tutorial

05-Assembling_Chromatograms.tutorial

06-De_novo_assembly.tutorial

07-Expression.analysis.tutorial

08-Map_to_reference.tutorial

09-3D_structure.tutorial

10-Pairwise_alignments.tutorial

11-Primer_Design_R7.tutorial

12-Sequence_classifier.tutorial

13-CRISPR.tutorial

14-Cloning.tutorial

15-Creating_tutorials.tutorial

16-HIV_case_study.tutorial

17-Plant_Molecular_Evolution.tutorial

Referencia de figuras

1. Adaptado de: Shutterstock, 2017.

2. Adaptado de: Shutterstock, 2017.

3. Fuente: Shutterstock, 2017.

4. Fuente: Bionova.org.es, 2017.
   Licencia CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/legalcode

5. Fuente: Wikimedia, 2017.
   Licencia CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/legalcode

6. Adaptado de: Genome, 2017.
   Licencia: https://creativecommons.org/publicdomain/mark/1.0/deed.es

7. Adaptado de: Shutterstock, 2017.

8. Fuente: Flick, 2017.
   Licencia: https://creativecommons.org/licenses/by/2.0/legalcode

9. Fuente: Bdigital, 2017.
   Licencia: https://creativecommons.org/licenses/by-sa/3.0/legalcode

10. Fuente: Wikimedia, 2017.
    Licencia: https://creativecommons.org/publicdomain/zero/1.0/deed.es

11. Fuente: Wikimedia, 2017.
    Soporte: Science Primer (National Center for Biotechnology Information) Licencia: https://creativecommons.org/licenses/by-sa/3.0/legalcode

12. Fuente: Wikimedia, 2017.

13. Fuente: Wikimedia, 2017.
    Licencia CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/legalcode

14. Adaptado de: Shutterstock, 2017.

15. Fuente: Wikimedia, 2017.
    Licencia CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/Legalcode

16. Fuente: Wikimedia, 2017.
    Licencia CC BY-SA 3.0: https://creativecommons.org/licenses/by-sa/3.0/Legalcode

17. Fuente: Wikimedia, 2017.
    Licencia CC BY 4.0: https://creativecommons.org/licenses/by/4.0/legalcode

18. Fuente: Github, 2017.

19. Fuente: Wikimedia, 2017.
    Soporte: Madprime Licencia CC BY-SA 4.0: https://creativecommons.org/licenses/by-sa/4.0/legalcode