Programa de la asignatura:

Bioinformática

U1 Introducción a la bioinformática

DCSBA

BIOTECNOLOGÍA

Índice

Presentación de la Unidad 3
Propósitos de la unidad 4
Competencia específica 4
1.1. Genómica funcional 5
1.1.1. Dogma central de la Biología Molecular 4
1.1.2. Secuenciación genómica masiva 6
1.1.3. La era de las “Omicas” 9
1.1.4. Concepto de Bioinformática 10
1.2. Recursos disponibles en internet 11
1.2.1. Bases de datos: utilidad, concepto y clasificación 11
1.2.2. Secuencias de ADN o de proteínas como datos biológicos 13
1.2.3. La base de datos del NCBI 13
Actividades 28
Autorreflexiones 28
Cierre de la Unidad 29
Para saber más 30
Fuentes de consulta 30

Presentación de la Unidad

Las células son las unidades de construcción de todos los seres vivos. Dentro de cada célula existen dos macromoléculas fundamentales que contienen la información necesaria para definir las características biológicas de los organismos, llamadas por lo tanto, macromoléculas informacionales. Estas macromoléculas son los ácidos nucleicos y las proteínas, los cuales contienen información genética e información funcional, respectivamente. Dicha información está reunida en forma de secuencias de nucleótidos o de aminoácidos en un código determinado.

Para que esta información pueda reflejarse en un fenotipo, tiene que ser procesada en una serie de eventos específicos que tienen cierta dirección. El dogma central de la biología molecular establece la dirección de este flujo de información. El avance científico permitió descifrar la naturaleza de estos códigos. Estos códigos se reflejaron en secuencias de aminoácidos o de nucleótidos que tenían un cierto orden. Paralelamente, también se hizo posible dilucidar la estructura de las macromoléculas, con lo cual, su caracterización y manipulación fue posible, surgiendo la ingeniería genética o metodología del ADN recombinante. Entre las metodologías que sustentan a la ingeniería genética se destaca la técnica de secuenciación de ácidos nucleicos, que ha permitido conocer la secuencia de genomas completos en un periodo de tiempo relativamente corto. Debido al gran número de datos biológicos que la secuenciación genómica provee, se ha vuelto necesario el perfeccionamiento y uso de operaciones y sistemas basados en tecnologías de la información para almacenar, acceder y analizar este gran cúmulo de información.

La creación de bases de datos ha permitido el almacenamiento y acceso a los datos biológicos.

Así, de la necesidad de operar la gran cantidad de datos biológicos que estaban siendo generados por al avance científico y tecnológico de la biología molecular y la genética en las últimas décadas, surge la Bioinformática. En términos generales, la Bioinformática es un área interdisciplinar de reciente creación que se encarga del análisis computacional de datos biológicos.

Propósitos de la unidad

- Reconocer el marco histórico que llevó al surgimiento y desarrollo de la Bioinformática.
- Describir la naturaleza de los datos generados y comprender la necesidad de crear bases datos.
- Abordar el uso de una las de bases de datos de secuencias genéticas más comunes en la actualidad.

Competencia específica

Analizar la utilidad del uso de recursos informáticos actuales para almacenar, acceder y obtener información biológica del gran número de datos generados en los últimos años, a partir del estudio del contexto histórico.

1.1. Genómica funcional

Con el objeto de adentrarnos en el campo de la Bioinformática, será preciso, primeramente, puntualizar algunos conceptos básicos de biología molecular.

1.1.1. Dogma central de la Biología Molecular

El dogma central de la biología molecular constituye el paradigma que establece la dirección del flujo de la información genética para que ésta pueda ser expresada, permitiendo obtener un fenotipo a partir de un genotipo. Determina que en los sistemas celulares, el ADN tiene la capacidad de replicarse y de ser transcrito a ARN, el cual a su vez es traducido a proteínas. Fue establecido por Francis Crick (Ondarza, 1994; Nelson & Cox, 2005).

En asignaturas anteriores (Biología Molecular 1 y Genética Molecular Bacteriana) se describió en detalle la naturaleza de las macromoléculas y procesos involucrados en este flujo de información, por lo que no es el objetivo de esta Unidad describirlos. Sin embargo, es importante destacar, que a este esquema del dogma clásico, se han adicionado algunos procesos particulares que han demostrado que la unidireccionalidad del flujo de la información genética no es absoluta, como anteriormente se creía. Así, hoy en día se sabe que tanto el ADN como el ARN pueden funcionar como material genético, teniendo ambos la capacidad de hacer copias de sí mismos y que se puede obtener ADN a partir de ARN. Dichos descubrimientos se realizaron en virus, donde se demostró que su ARN puede replicarse mediante la acción de la enzima ARN polimerasa dependiente de ARN. Por otro lado, también en virus (específicamente en retrovirus), se determinó que por la acción de la enzima transcriptasa reversa, el ARN puede ser transcrito a ADN, proceso conocido como transcripción reversa o retro-transcripción. Para lograr el proceso de transcripción reversa, las enzimas reverso transcriptasas catalizan tres distintas reacciones: la síntesis de ADN dependiente de ARN; la degradación de ARN; y la síntesis de ADN dependiente de ADN. Estas enzimas, a diferencia de las ADN polimerasas, presentan una alta tasa de error (1 en cada 20,000 nucleótidos adicionados), lo cual genera más altas tasas de mutación y evolución, contribuyendo a la frecuente aparición de cepas nuevas en retrovirus (Nelson & Cox, 2005).

En la Figura 1 se muestra el esquema del dogma central de la biología molecular ampliada donde se compendian tanto los procesos generales que se llevan a cabo en todas las células, como aquéllos característicos de los virus. A pesar de que los fenómenos de replicación y transcripción reversa de ARN están restringidos a virus, su importancia está dada en términos evolutivos, ya que han contribuido a la generación de cambios en el genoma de las células, pues una vez que un virus ha infectado a un

hospedero, éste tiene la capacidad de integrar ADN retro-transcrito a partir de ARN en el genoma de la célula (Nelson & Cox, 2005; Krebs, et al., 2010).

El planteamiento del dogma central de la biología molecular no fue una tarea fácil, sino que sólo tras años de intensa investigación fue que pudo establecerse de manera correcta la dirección del flujo de la información biológica. Importantemente, este hecho sentó una de las piedras angulares que llevó al surgimiento de una nueva rama de la biología, que en un principio fue llamada biología computacional, para después constituir lo que actualmente conocemos como Bioinformática.

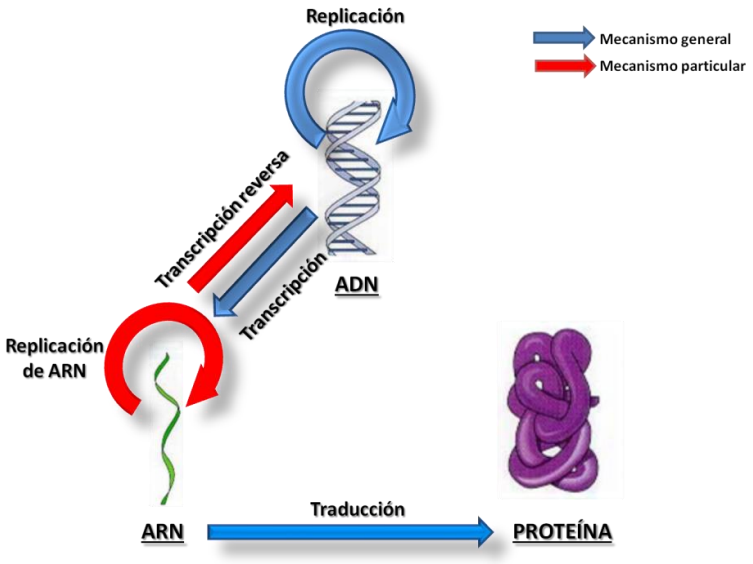


Figura 1. Esquema donde se representa el dogma central de la biología molecular mostrando el flujo de la información genética, señalando los procesos generales exhibidos por todas las células, como aquéllos procesos específicos llevados a cabo por virus.

1.1.2. Secuenciación genómica masiva

Una vez que se estableció que dentro de la célula existen macromoléculas informacionales que determinan su función y evolución, el siguiente paso fue descifrar el orden en que esta información se encontraba contenida. Así, se empezaron a desarrollar metodologías que permitieran conocer tanto la secuencia del ADN o ARN, conformada por nucleótidos, como la de las proteínas, conformada por aminoácidos.

Las técnicas de secuenciación de ADN fueron las que se desarrollaron más rápidamente, siendo la descrita por Sanger (1977) la más usada desde su creación, hace ya más de 30 años. Está basada en la síntesis química y se le conoce también como método de terminación de cadena. Debido a que en la asignatura de Biología Molecular II se

describió con detalle dicha técnica, no se ahondará más sobre este tema aquí. Sin embargo, en la sección “Para saber más” podrás encontrar un recurso que te ayudará a recordar el procedimiento de manera general.

A lo largo del tiempo, dicha técnica experimentó una importante transformación que impactó grandemente en el desarrollo de la biología actual. En sus inicios, la secuenciación Sanger permitió descifrar únicamente genes completos, pero más tarde, tras su optimización y automatización, llevó a la lectura de genomas enteros. Recordemos que el genoma es la colección completa de las secuencias de ADN que conforman a un organismo (Krebs, et al., 2010). Desde el punto de vista biológico, ya no sólo se quería conocer la secuencia de algunos cuantos genes aislados, si no que se pretendía conocer la secuencia de todos los genes que estaban presentes en los organismos. Así, el primer organismo vivo del cual se conoció por primera vez su secuencia completa en el año de 1995 fue el de la bacteria patógena causante de la gripa, *Haemophilus influezae* (Fleischmann, et al., 1995). Este acontecimiento marcó el nacimiento de la era genómica, pues a partir de ese momento se empezaron a secuenciar rápidamente otros muchos, miles de genomas de distintos organismos y cuyo objetivo principal era conocer la secuencia de nucleótidos completa que los conformaban.

Así, las técnicas de secuenciación fueron optimizadas básica y principalmente para producir datos a gran escala, es decir, para poder secuenciar fragmentos de ADN muy grandes, pues debido a la complejidad de los organismos que se pretendía estudiar, era muy probable que estos estuvieran conformados por miles, hasta millones de nucleótidos. De esta forma, con la creación y puesta en marcha del proyecto de secuenciación del genoma humano, la demanda en la producción de secuencias de ADN aumentó considerablemente, por lo que las estrategias de secuenciación fueron optimizadas considerablemente. Pero el genoma humano no fue el único en el que la comunidad científica estaba interesada en secuenciar, sino que muchísimos otros organismos pretendían ser secuenciados, por lo que se necesitaban tener métodos que permitieran una secuenciación más barata y de alto rendimiento o eficiente, creándose entonces la secuenciación de segunda generación, también llamada secuenciación paralela masiva. Constituye una “tecnología que permite acumular información genética tanto de manera cualitativa como cuantitativa de cualquier tipo de ácido nucleico en una muestra dada, con un rendimiento enormemente alto” (Reis-Filho, 2009).

Usando distintas aproximaciones, esta tecnología ha sobrepasado la limitada capacidad de la secuenciación Sanger, permitiendo que millones de reacciones de secuenciación puedan llevarse a cabo de una vez y de forma paralela y no sólo 96, como en la secuenciación por terminación de cadena (Mardis, 2007). En la actualidad, existen en el mercado varias de estas tecnologías, las cuales son enlistadas en la Tabla 1, donde se anotan los datos más sobresalientes de cada tecnología que permiten compararla con la secuenciación Sanger tradicional, quedando de manifiesto la gran diferencia en el número

de secuencias obtenidas con las tecnologías de segunda generación, así como el tiempo requerido, que es mucho menor, obteniéndose con ello, un mayor número de secuencias en un tiempo relativamente corto y adicionalmente, a menor costo .

En los siguientes enlaces encontrarás videos explicativos sobre las metodologías de secuenciación de segunda generación comercialmente disponibles:

Secuenciación 454:
http://www.my454.com/

Solexa:
http://www.youtube.com/watch?v=HMyCqWhwB8E#t=58

SOLiD:
http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-4-system.html?CID=FL-091411_solid4

Tabla 1. Tecnologías de segunda generación comercialmente disponibles.

Método	Longitud de la lectura (pares de bases)	Templados secuenciados	Producción de datos/día	Reacción de la secuencia	Referencia
ABI 3730xl	900 a 1,100	96	1 Mb/día	Método Sanger	www.appliedbiosystems.com
454 FLX Roche	400	1,000,000	400 Mb/corrida/7-8 h	Pirosecuenciación	www.rocheapplied-science.com
Illumina (Solexa)	36 a 175	40,000,000	>17 Gb/corrida/3-6 días	Terminador reverso	www.illumina.com
ABI SOLiD	50	85,000,000	10-15 Gb/corrida/6 días	Secuenciación de ligación	www.appliedbiosystems.com
Helicos Heliscope	30 a 35	800,000,000	21-28 Gb/corrida/8 días	Secuencia de molécula única por síntesis	www.helicosbio.com

Tabla modificada de Reis-Filho, 2009.

Actualmente, se cuenta con 7,581 proyectos de secuenciación terminados y 28,913 por terminar (http: //www.ncbi.nlm.nih.gov/, 2013). Con esta revolución en las técnicas de secuenciación, uno puede apenas imaginar el gran número de datos que se han obtenido desde la creación de estas plataformas tecnológicas, lo cual puede reflejarse en el número de secuencias que se han depositado en diferentes fuentes de información en los últimos años. Como puede observarse en la gráfica de la Figura 2, el número de pares de bases y de secuencias que se depositaron en 10 años en centros de información especializados, aumentaron en número de manera considerable, duplicando el número de entradas aproximadamente cada 18 meses (Posada, 2009).

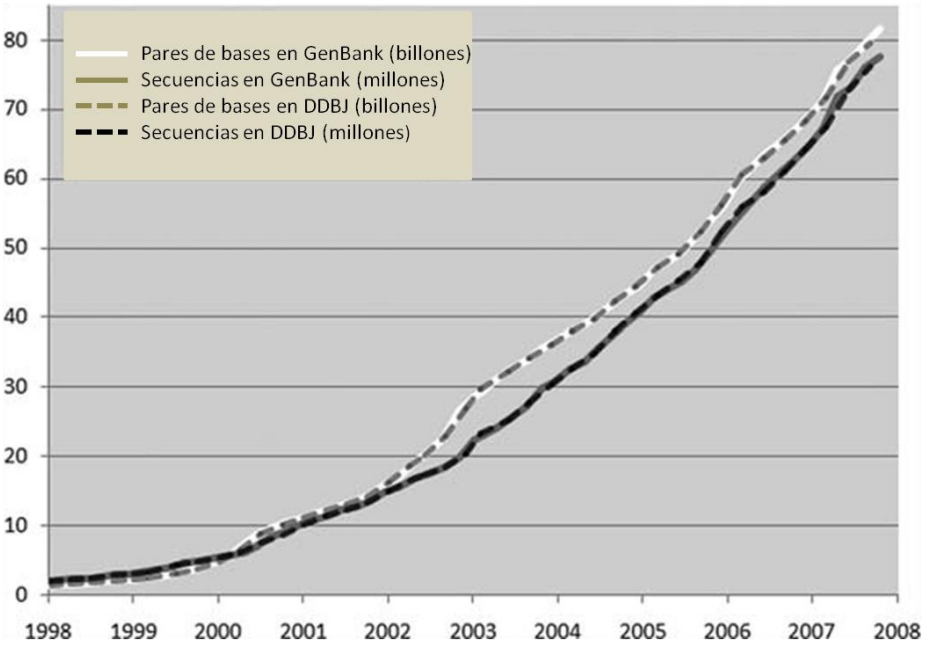


Figura 2. Crecimiento en el número de secuencias que se introdujeron en dos de las bases de datos más populares, el Gen Bank y DDBJ a lo largo de 10 años (modificada de Posada, 2009).

1.1.3. La era de las “Omicas”

El objetivo inicial de la era genómica fue determinar la secuencia genética completa de los organismos para hacer un análisis de su organización estructural, es decir, cómo estaban constituidos los genes, si poseían regiones regulatorias, si presentaban exones, dónde se encontraban localizados y por supuesto para qué proteínas codificaban, así como el estudio de secuencias no codificantes. Sin embargo, con el estudio individual de estas secuencias genéticas no es posible determinar si se expresarán o no y en qué condiciones, cuál sería su papel en la célula, cuál sería su mecanismo de acción o cómo ayudarían a conformar a la célula, etc. Por lo anterior, se volvió necesario el desarrollo de metodologías que permitieran analizar y explotar toda esta información generada, surgiendo entonces la **genómica funcional o era post-genómica,** que pretende

determinar la función de los genes y descubrir su comportamiento en los sistemas biológicos, procurando estudiarlos no como entes individuales, sino siendo parte de un todo en un momento determinado.

La genómica funcional está relacionada íntimamente con otras “-omicas”. Dicho sufijo, aplicado a las ciencias de la vida, se refiere a la medición de la totalidad de las moléculas biológicas e información en un nivel biológico particular (Schneider & Orchard, 2011). Su principal objetivo es tener una visión global de cuándo y cómo ocurren diversos procesos en los sistemas biológicos. Las omicas han incidido en una amplia variedad de campos y entre las principales encontramos a la transcriptómica, la proteómica, la metabolómica, la interactómica. Cada una de estas se vale de metodologías particulares. La transcriptómica se refiere al estudio del transcriptoma, es decir, a los ARN mensajeros o transcritos presentes en células, tejidos u organismos. Por lo tanto, estudia la expresión de los genes. Por otro lado, la proteómica, también analiza la expresión de genes, pero a nivel de proteína. La metabolómica se refiere al estudio del metaboloma, siendo este último el conjunto de metabolitos (moléculas de bajo peso molecular) presentes en un tiempo determinado. El interactoma estudia la interacción que puede ocurrir entre las proteínas en una condición específica (Kandpal, et al., 2009). Mientras que el genoma de un organismo no varía en distintas condiciones ambientales, el transcriptoma, proteoma, metaboloma o interactoma variarán dependiendo de las distintas condiciones de crecimiento, el estado fisiológico, el desarrollo, etc. Así, estas metodologías permiten el análisis de sistemas biológicos completos.

Es claro entonces, que tras el desarrollo en las últimas décadas de las metodologías mencionadas, se han generado un inmenso número de datos. Sin embargo, su generación sólo cobraría sentido si se cuenta con recursos que permitan analizarlos y entenderlos. Por lo que, paralelamente al desarrollo científico y tecnológico, se deben mejorar también las metodologías analíticas y desarrollar herramientas que permitan manejar estos datos.

1.1.4. Concepto de Bioinformática

La bioinformática es una disciplina que se originó con el propósito utilitario de introducir orden en el conjunto masivo de datos que se estaban generando gracias a las nuevas tecnologías de la biología molecular, entre ellas la secuenciación a gran escala (Polanski & Kimmel, 2007). Se basa en usar aproximaciones computacionales de la tecnología de la información como software, simulación gráfica, algoritmos (fórmulas matemáticas), para responder problemas biológicos. Lo anterior implica tomar ventaja de un gran y complejo número de datos de forma rigurosa, con el firme propósito de que las conclusiones biológicas concebidas sean válidas. Su potencial incluso está ligado a guiar de manera eficiente, el diseño experimental en el laboratorio (Baxevanis, Ouellette, 2001).

1.2. Recursos disponibles en internet

Con el auge de las técnicas de secuenciación, un gran número de secuencias biológicas fueron generadas en las últimas décadas, resultando imposible mantener toda esta gran cantidad de información disponible en forma impresa, por lo que se volvió necesario empezar a digitalizarla, con el propósito de poder acceder a ella de una manera mucho más rápida. Además, para cumplir el mismo fin de accesibilidad, también era necesario mantenerla ordenada. Una de las formas de conservar a la información en este formato es mediante la creación de **bases de datos**. Una base de datos es, en términos simples, una **colección de datos dispuestos de manera ordenada**.

1.2.1. Bases de datos: utilidad, concepto y clasificación

El primer esfuerzo que involucró la recopilación de secuencias biológicas fue realizado a mediados de los 60´s en EUA por Margaret Dayhohh y sus colaboradores, quienes reunieron todas las secuencias y estructuras tridimensionales de proteínas conocidas en un atlas de secuencias y estructuras de proteínas que fueron clasificadas de acuerdo con su similitud en secuencia (Posada, 2009). Conforme la información biológica correspondiente a secuencias se siguió acumulando con el tiempo, finalmente en 1982, el Laboratorio Europeo de Biología Molecular o EMBL (por sus siglas en inglés, ***E****uropean* ***M****olecular* ***B****iology* *L**aboratory*) generó una base de datos unificada de secuencias biológicas. Posteriormente, otros dos centros se unieron a este esfuerzo, el Centro Nacional de Información en Biotecnología o NCBI (por sus siglas en inglés, ***N****ational* ***C****enter for* ***B****iotechnology* ***I****nformation*) y el Banco de Datos de ADN de Japón o DDBJ (por sus siglas en inglés, ***D****NA* ***D****ata* ***B****ank of* ***J****apan*). El Instituto de Bioinformática Europeo tiene su sede en Hinxton, Inglaterra, mientras que el Centro Nacional para la Información en Biotecnología se encuentra en Bethesda, Maryland, EUA. El Banco de Datos de ADN de Japón es mantenido por el Instituto Nacional de Genética y el Centro para la Información Biológica o NIG/CIB (por sus siglas en inglés, ***N****ational* ***I****nstitute of* ***G****enetics*, ***C****enter for* ***I****nformation* ***B****iology*), los cuales se localizan en Mishima, Japón (Lemey et al., 2009). Es preciso decir que en la actualidad, cada una de estas bases de datos se nutre principalmente de los datos introducidos por científicos de manera individual. Cabe mencionar que desde hace aproximadamente 18 años, estos 3 grupos han formado un consorcio colaborativo denominado Colaboración Internacional de Bases de Datos de Secuencias de Nucleótidos o INSDC (por sus siglas en inglés, ***I****nternational* ***N****ucleotide* ***S****equence* ***D****atabase* ***C****ollaboration*), que tiene el propósito de compartir toda la información depositada en sus bases de datos respectiva y establecer estándares que permitan garantizar una uniformidad en los datos, de hecho, cada servidor actualiza recíprocamente su información diariamente.

Entonces, las bases de datos permiten manejar información de diversos tipos. La mayoría de ellas está conformada por un archivo de información, una organización lógica de esa información y herramientas para acceder a ella.

Dependiendo del tipo de datos que opera, el tipo de información implicada puede pertenecer a uno o varios de tres distintos niveles:

- En el primer nivel, la información únicamente está almacenada.
- En el segundo nivel, hay un acceso a esa información, ya que las bases de datos tienen cierta estructuración que lo permite.
- En el tercer nivel se consigue el análisis de los datos, en este nivel, que depende de los dos anteriores, se obtiene nueva información mediante la comparación de los datos disponibles.

Es importante mencionar que aquéllas bases de datos que permiten un rápido acceso a la información poseen una estructura, donde la unidad estructural más pequeña es un registro, el cual contiene información relacionada, en un formato determinado. A dicho registro generalmente se le asigna un código que lo caracteriza, con el fin de poder localizarlo fácilmente en el futuro (Baxevanis, Ouellette, 2001).

Las bases de datos en biología molecular incluyen aquéllas que contienen secuencias de ácidos nucléicos, de aminoácidos, modelos de estructuras, datos de expresión, entre otros, constituyendo una herramienta Bioinformática indispensable para mantener y recuperar información biológica. Pueden ser clasificadas de la siguiente forma (Lesk, 2002):

1. **Bases de datos de registros de información biológica**: Contienen una gran cantidad de información adicional, como por ejemplo de qué organismo proviene la secuencia, el método de secuenciación utilizado, etc. Ejemplos de estas incluyen bases de datos con los siguientes contenidos:
    - Secuencias de ácidos nucléicos y proteínas
    - Estructuras de ácidos nucléicos y proteínas
    - Patrones de expresión de proteínas

2. **Bases de datos derivadas**: Contienen información colectada a partir de las bases de datos de registros de información biológica y de los análisis de su contenido: Ejemplos de estas incluyen bases de datos con los siguientes contenidos:
    - Motivos de secuencia (motivos señal característicos de una familia de proteínas)
    - Mutaciones y variaciones en secuencias de ADN y proteínas.

1.2.2. Secuencias de ADN o de proteínas como datos biológicos

El desarrollo de técnicas para secuenciar ADN, como la terminación en cadena de Sanger mencionada en el tema anterior, introdujo un nuevo tipo de dato: las secuencias de ADN. Una vez definido este dato biológico, la técnica se utilizó para adquirir miles de secuencias de organismos muy diversos. La información que se origina directamente del laboratorio se conoce como información primaria. Así, los datos primarios son tomados en un laboratorio sobre una molécula allí existente y no son una recopilación de datos de otros autores. Una vez que una secuencia es obtenida en el laboratorio y compartida en una base de datos, esta debe de ser curada.

Las secuencias de ADN se refinan para asegurar que son de buena calidad y además se debe adicionar información relacionada con ella, como por ejemplo la posición de genes y proteínas que codifica. A este proceso se le conoce como **curación**. El proceso de curación está pensado para dar valor agregado a una secuencia. Si se elimina cualquier otro dato y sólo se deja la secuencia o por el contrario, se incluye en un registro toda la información posible, pero sin que esta haya sido verificada, producirá que la secuencia se quede sin interpretación o que los datos no tengan credibilidad (Baxevanis, Ouellette, 2001).

Una vez que se crearon y adquirieron los datos deseados, estos pueden pasar por distintos procesos, siendo uno de los más importantes, su uso. La adquisición de datos tiene un costo que se justifica por el uso que se le dará. Cuando se mencionan secuencias en los artículos científicos, lo más común es que se pretenda probar alguna hipótesis mediante dicha secuencia. Un ejemplo sería establecer la hipótesis de que nuestra secuencia de estudio actúa como una región promotora de la transcripción, por lo que este será el uso que le daremos en nuestro contexto particular. Sin embargo, el uso inicial de un dato puede resultar distinto al que se le dé subsecuentemente. Usemos el ejemplo de los marcadores genéticos de la mosca de la fruta (*Drosophila*) que Morgan encontró y usó para explicar la herencia genética. Esos mismos marcadores se utilizaron mucho tiempo después para secuenciar el genoma de esta mosca. Así, se puede hacer uso de información ya depositada, en contextos distintos al original.

1.2.3. La base de datos del NCBI

Con el fin de que te familiarices con el uso de las bases de datos más importantes en biología, en este tema nos enfocaremos en la aplicación práctica de la búsqueda de secuencias en una de las principales bases de datos, el NCBI.

Como ya se mencionó anteriormente, en los servidores del NCBI se encuentra almacenada la mayor cantidad de información primaria en biología molecular. Fue

fundada hace aproximadamente 30 años con la intención de almacenar información biológica, hacer investigación en biología computacional, desarrollar herramientas de análisis de información biológica.

Las bases de datos principales que son encontradas en el NCBI son las siguientes:

- Bases de datos de literatura científica
- Bases de datos moleculares
- Bases de datos de genomas

Y los principales tipos de datos biológicos que se pueden analizar son: secuencias de nucleótidos, secuencias de proteínas, estructuras en tres dimensiones, genes, expresión de genes y taxonomía.

Cabe mencionar que esta base de datos se encuentra soportada por los Institutos Nacionales de Salud de EUA, de los que depende a su vez, la Librería Nacional de Medicina. Así, al ser una extensión de una librería médica, su enfoque principal está dirigido a la salud.

Búsqueda de una secuencia de nucleótidos

Ahora es el momento de visitar por primera vez la página principal del NCBI.

Primeramente, ingresa a la página principal del sitio: http://www.ncbi.nlm.nih.gov/. Debido a que esta URL será consultada en la Unidad siguiente, te recomendamos añadirla a tus favoritos del navegador de Internet.

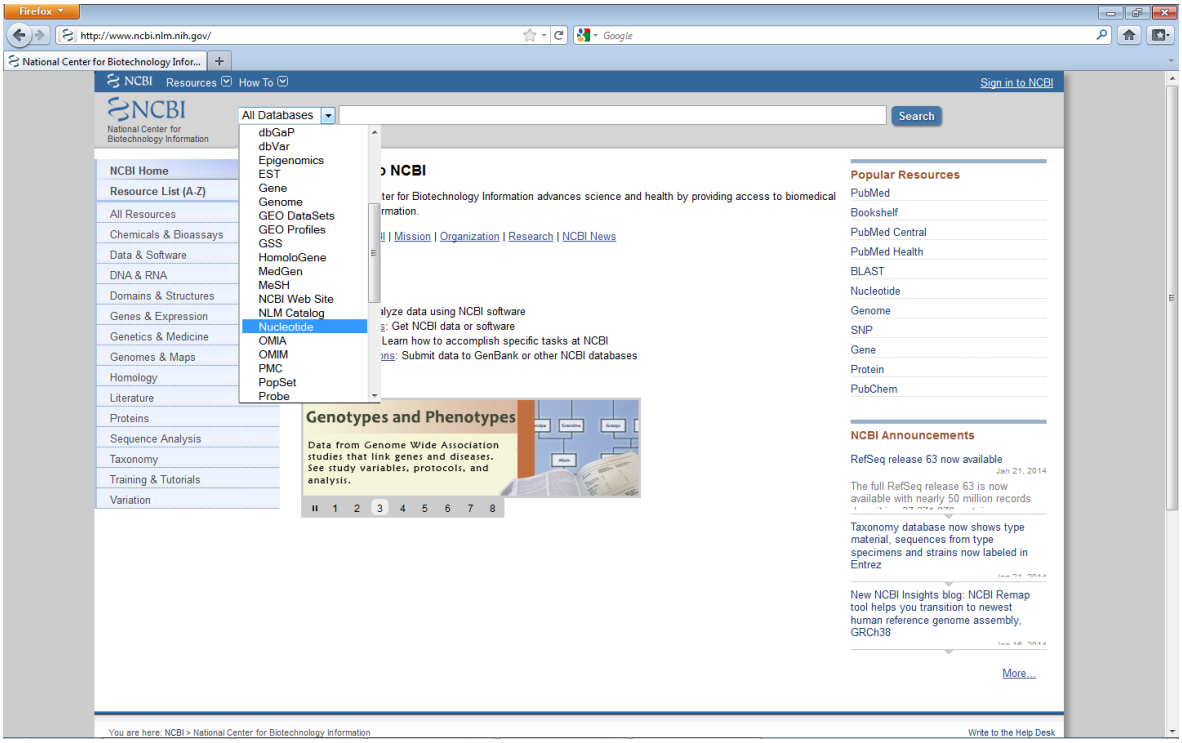
Desmenuzando la URL tenemos que la extensión nlm viene de ***N****ational* ***L****ibrary of* ***M****edicine*. La extensión nih viene de ***N****ational* ***I****nstitutes of* ***H****ealth*. La terminación gov es asignada sólo a entidades gubernamentales, lo cual significa que esta base de datos se sustenta con recursos públicos.

Así, la interface general de la página principal del NCBI es la siguiente:

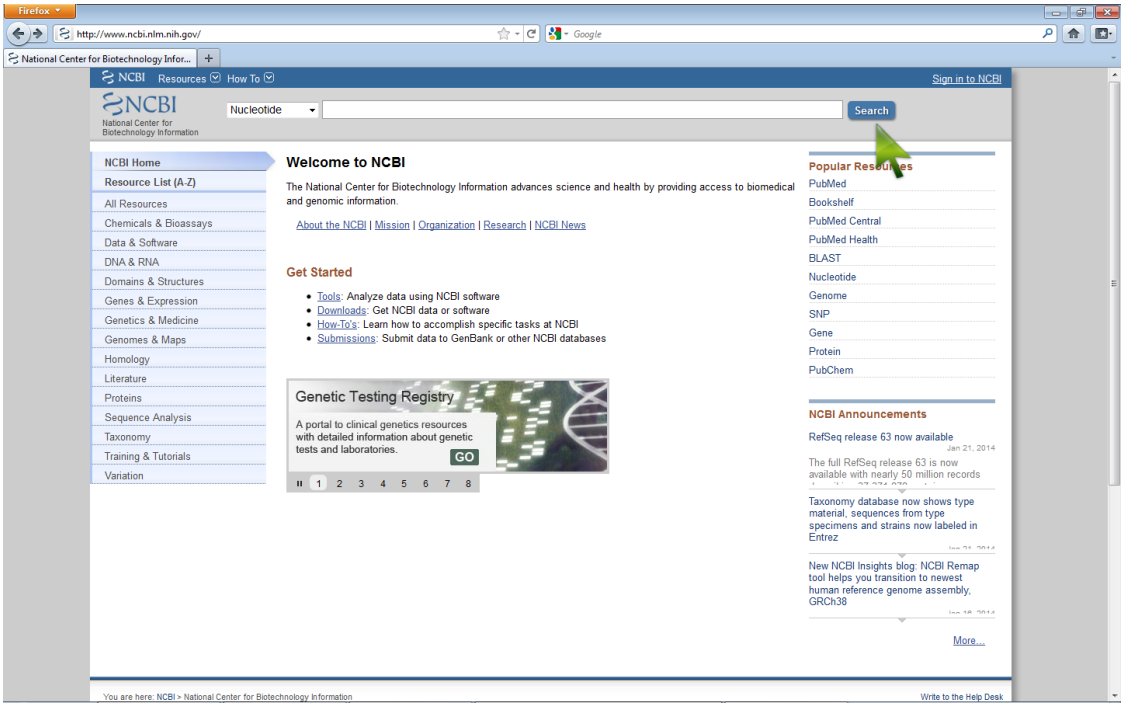
Nuestro primer ejercicio consiste en acceder a la secuencia de nucleótidos de una proteasa con potencial biotecnológico, la papaína. Dicha enzima ha sido utilizada para aclarar la cerveza.

Esta opción de búsqueda se nutre de la base de datos del GenBank, la base de datos genómica más conocida, que es mantenida por el NCBI y contiene todas las secuencias de ácidos nucléicos y aminoácidos. Constituye una base de datos con una gran cantidad de información, que sería imposible abordar en su totalidad en este curso. Por lo tanto, en esta Unidad sólo te guiaremos para que te introduzcas en una de las aplicaciones que te permitirán obtener secuencias con las que trabajarás en la siguiente Unidad.

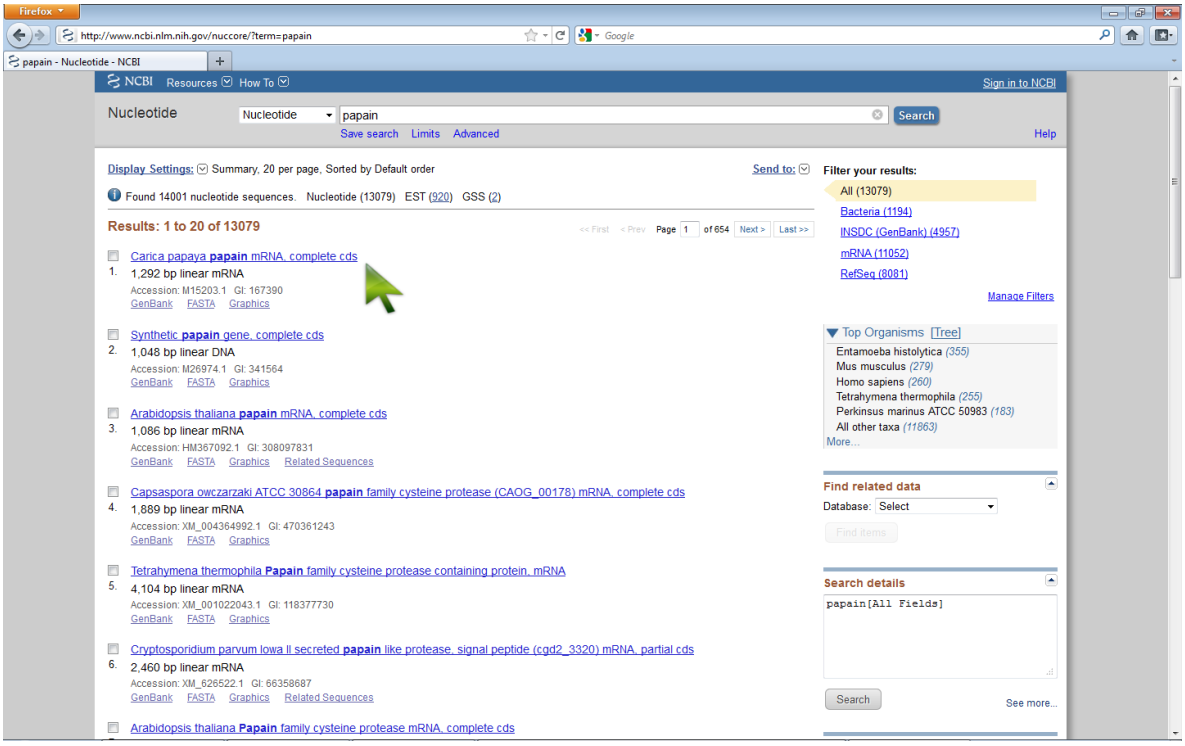
Usando la pestaña desplegable del lado superior izquierdo donde se leen las palabras “All Databases”, despliega la pestaña y selecciona la opción de “Nucleotide”:



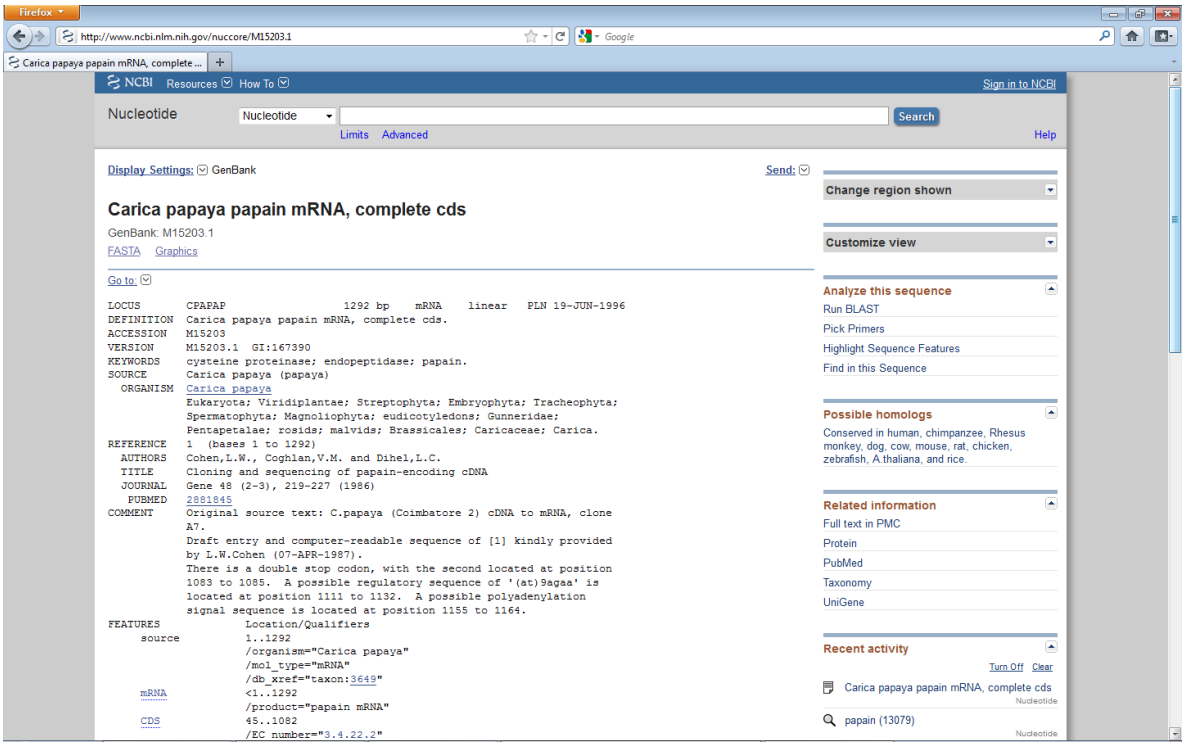
Una vez realizado lo anterior, anota el nombre del gen que quieres buscar, en este caso, papain (por papaína) y presiona Search “Search” o Enter. Recuerda que el idioma de las palabras de búsqueda debe ser el inglés.

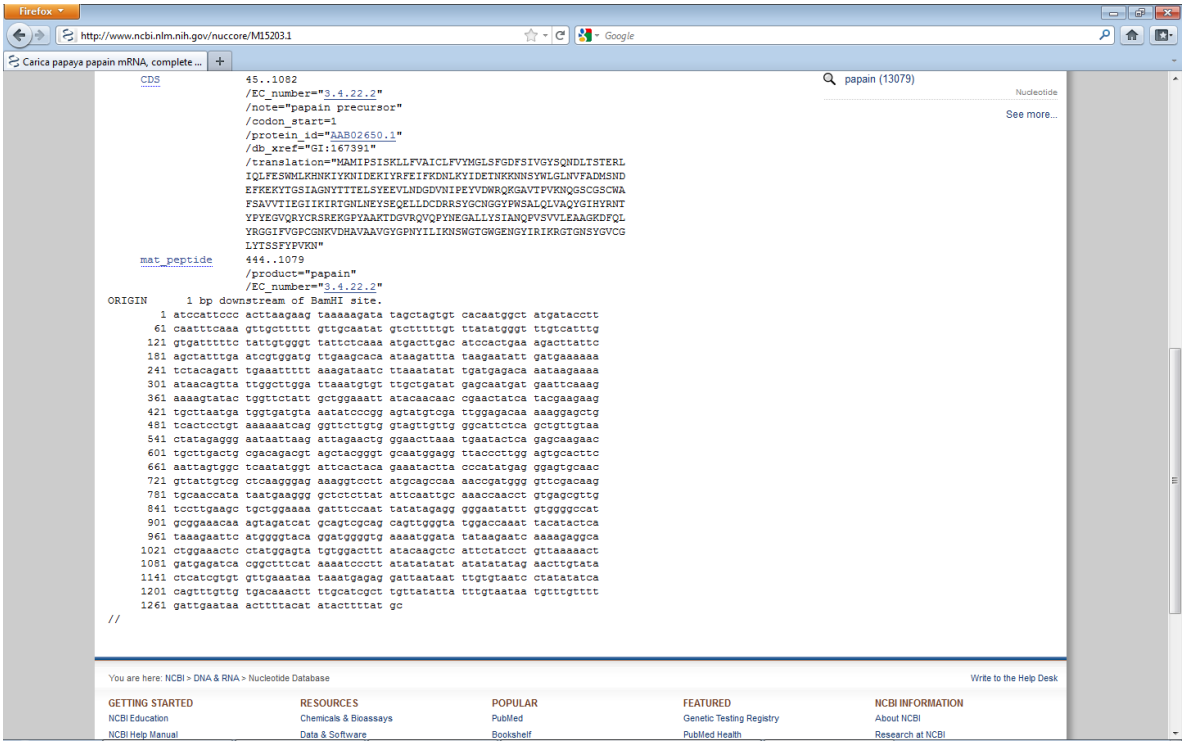


Acto seguido, en la ventana se desplegarán todas aquéllas secuencias donde se encontró la palabra *papain*. En este caso, el número de secuencias encontrado fue de 12,857. Este dato puede ser checado al inicio de los resultados, donde dice “Results”: Selecciona con el cursor la primera de las secuencias mostradas.



Una vez que seleccionaste la secuencia, este es el tipo de registro de resultados que se muestra y que es común para todas las secuencias:





Cada registro de secuencias de ácidos nucleicos es desplegado en GenBank con una serie de datos acompañantes que a continuación se describen (Lemey, et al., 2009):

En la primera línea se incluye el LOCUS, el número de bases que contiene la secuencia, la molécula que se secuenció (ADN, ARN, ARNm, etc.), la división a la que pertenece, que en este caso se refiere al grupo de organismos específico, PLN: secuencias de algas, fúngicas o vegetales y finalmente, la fecha de la última modificación.

LOCUS: Anteriormente era una clave que involucraba en las primeras tres letras el organismo del que provenía la secuencia y otras tres o cuatro letras para designar el nombre del gen o la probable función. En nuestro ejemplo, el organismo es ***C****arica* ***pa****paya* (CPA), y el gen es la **pap**ína (PAP), obteniendo el código CPAPAP. Sin embargo, fue complicado mantener un mismo LOCUS porque había casos en los que se descubría que en realidad ese gen codificaba o funcionaba para algo distinto a lo que originalmente se había pensado, con lo cual habría que cambiarle el nombre. Además, esta clave podía variar entre las tres bases de datos, es decir, EMBL, DDBJ y Gen Bank. Desde el 2006 entró en desuso en EMBL.

DEFINITION: Indica el nombre científico del organismo al que corresponde la secuencia así como una breve descripción del gen o fragmento de la secuencia.

ACCESSION: Muestra el número que se le asignó en la base de datos al momento de ingresarse por el usuario. Permanece constante a lo largo de nuevas liberaciones de secuencia y es el mismo en las tres bases de datos.

VERSION: Se construye a partir del número asignado a ACCESSION más un número que se adiciona cada vez que la secuencia se modifica. Con este número, uno puede acceder a versiones anteriores de la secuencia.

GI: Abreviación de GenInfo. Número asignado únicamente en la base de datos del GenBank y que permite identificar solamente una secuencia particular. Es un identificador único.

Además de estos tipos de identificadores, cada secuencia codificante (CDS por CoDing Sequence) tiene un identificador de proteína (ptotein_id), el cual es compartido por las tres bases de datos.

KEYWORDS: Línea opcional que puede utilizarse para listar palabras adicionales que relacionen la secuencia con posibles búsquedas sistemáticas.

SOURCE: Indica el origen biológico de la secuencia.

ORGANISM: Muestra la clasificación biológica del organismo del cual proviene la secuencia lo más detallada posible, desde lo más general (dominio) hasta lo más particular (especie, si es el caso).

REFERENCE: Comentarios literarios que incluyen:

- AUTHORS: La (s) persona (s) que obtuvo y está publicando las secuencias.
- TITLE: El nombre del trabajo de investigación publicado en donde se hace referencia a esta secuencia.
- JOURNAL: La fuente (Revista, Libro, etc.) en donde aparece publicada la secuencia.
- PUBMED: Numero de referencia asignado en esta base de datos. Esta es una interface al servicio de citas bibliográficas del Medline.
- REMARK: Aquí puede indicarse el estatus de la publicación, esto es, enviado, aceptado, en prensa, etc.

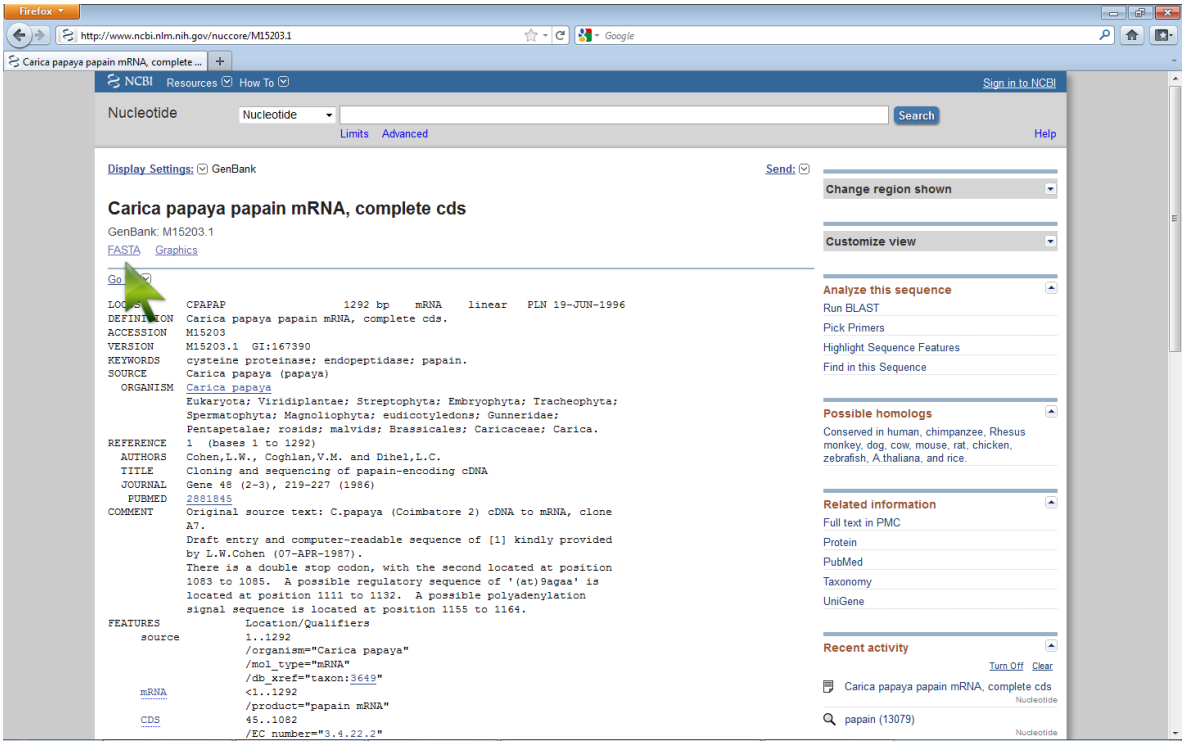
REFERENCE: Comentarios adicionales que involucran a la secuencia:

- AUTHORS: Nombre de quien obtuvo la secuencia (no necesariamente es el mismo al autor que publica).
- TITLE: Indica si el autor capturo directamente dicha secuencia.
- JOURNAL: Indica la fecha en que se sometió el artículo a publicación, así como la dirección postal del autor principal.

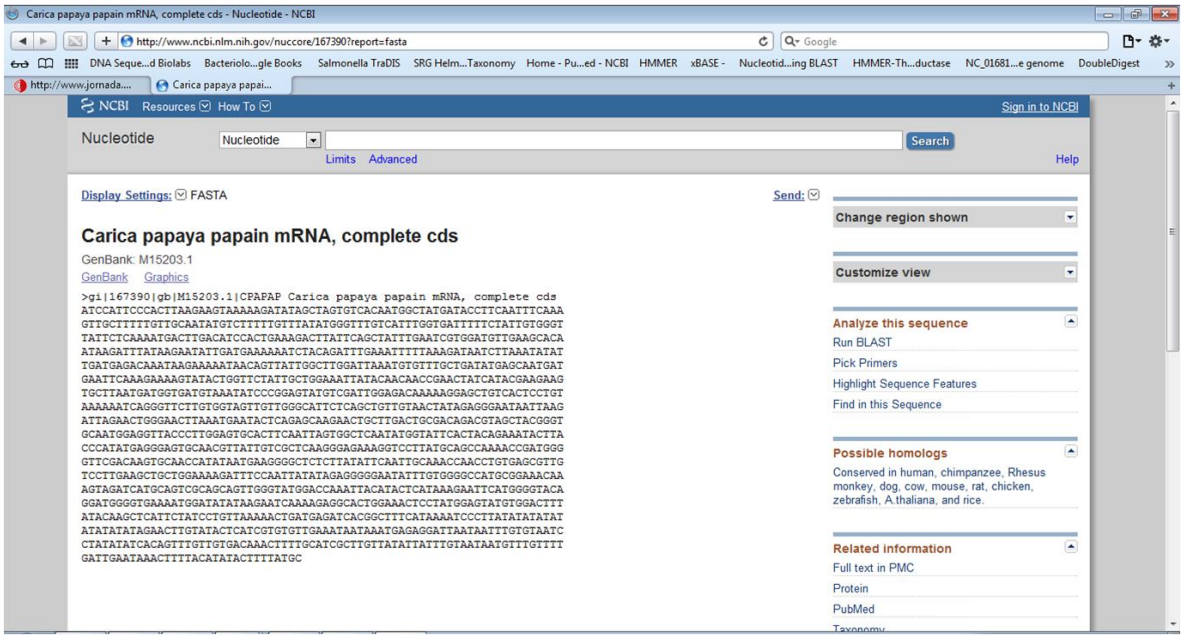
FEATURES: Indica el número de bases que comprenden esta secuencia; un resumen de la clasificación del organismo fuente y de qué tipo de gen (o fragmento de este) se trata.

ORIGIN: Muestra la secuencia completa en el lenguaje de bases nitrogenadas, esto es, Timina (T), Citosina (C), Guanina (G), Adenina (A) y Uracilo (U).

Finalmente, es el momento de obtener tu secuencia y guardarla en un formato que te permitirá manipularla posteriormente. En la ventana del registro de los resultados, selecciona la opción FASTA.



El formato FASTA es un formato estándar con el que trabajan muchos programas que analizan secuencias de ADN. Inicia con el signo de mayor que (>), seguido de una breve descripción de la secuencia de longitud variable. El siguiente renglón corresponde a la secuencia de nucleótidos.

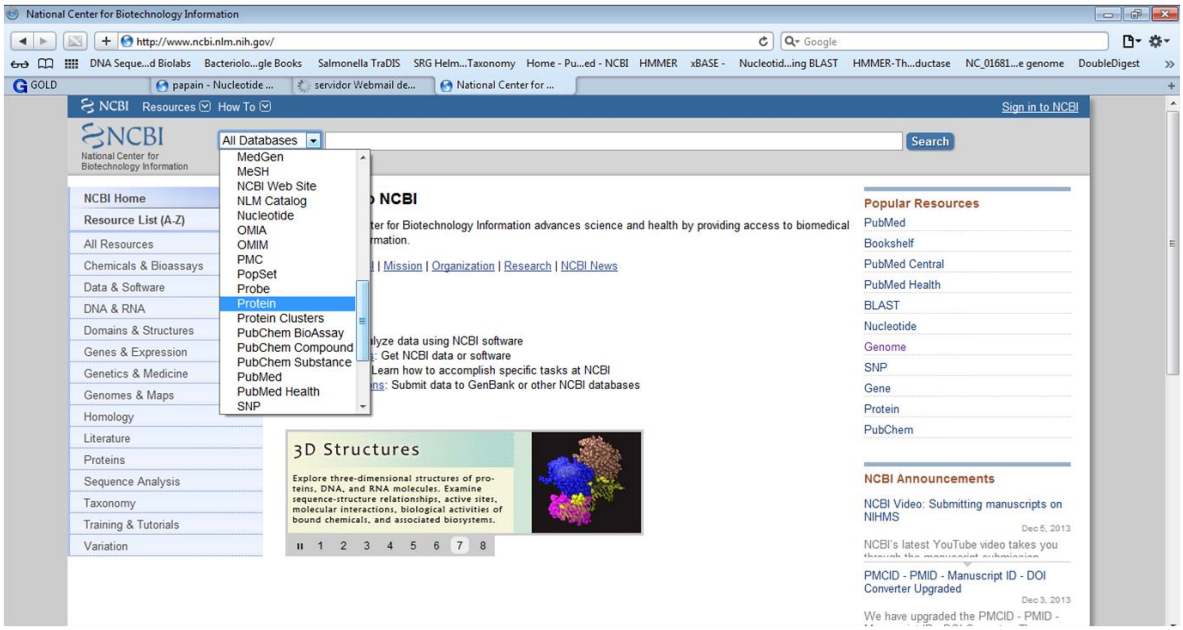


Selecciona tu secuencia, cópiala y pégala en un procesador de textos para que la puedas salvar. ¡Felicidades, has obtenido tu primer secuencia de nucleótidos!

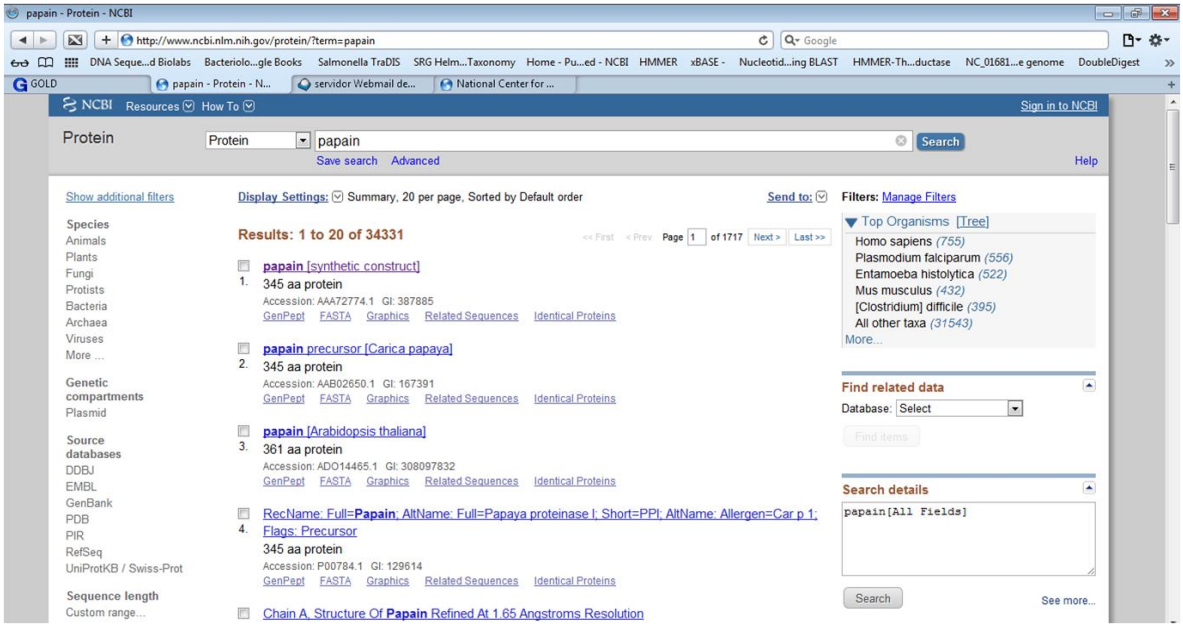
Otra forma de buscar secuencias específicas es introducir directamente su GI o Accession en la opción de “Nucleotide”. Esto se realiza cuando ya conocemos previamente esos datos de nuestra secuencia.

Búsqueda de una secuencia de aminoácidos

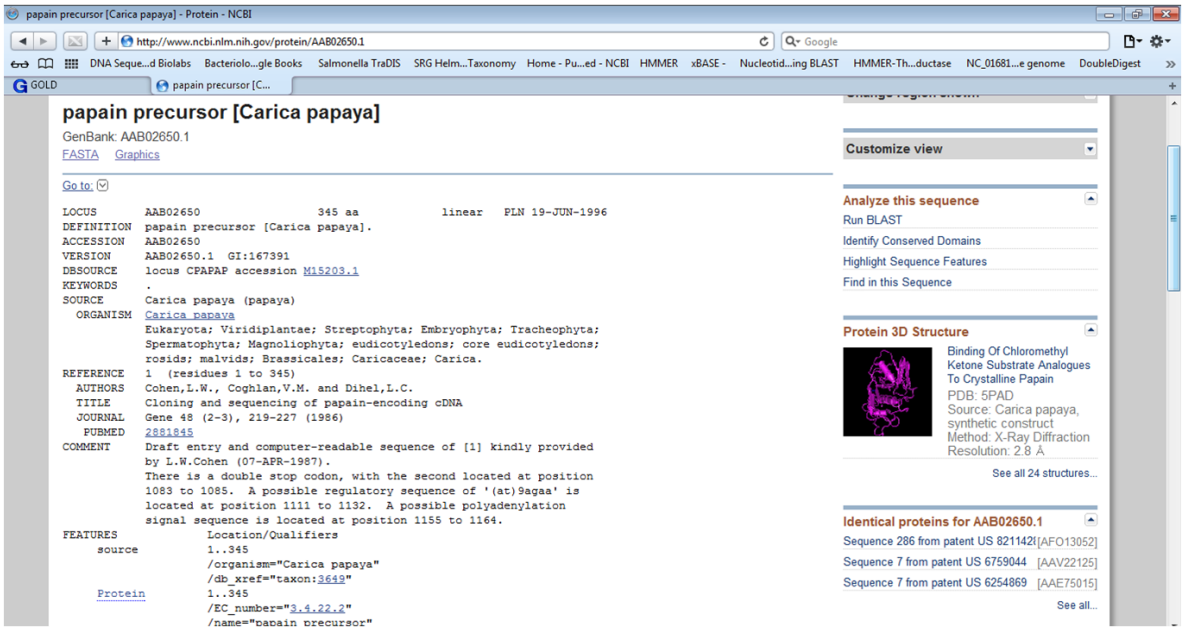
Como siguiente paso, buscaremos el mismo gen pero en su forma de proteína. Para esto, en la página inicial del NCBI, en la pestaña desplegable ahora selecciona “Protein”

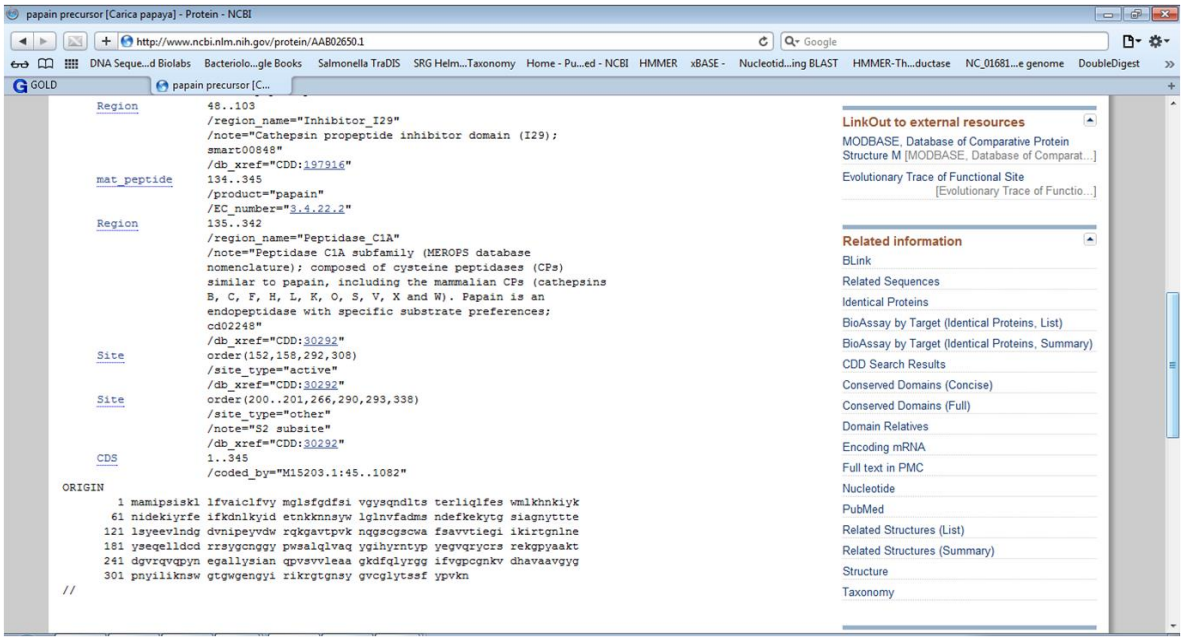


Introduce el término papain y presiona “Search” o Enter.



Selecciona el primer resultado. ¿Cuántos resultados obtuviste? El registro de resultados es muy parecido al obtenido anteriormente cuando buscabas tu secuencia de nucleótidos. En este caso, se mostrará la secuencia en aminoácidos.

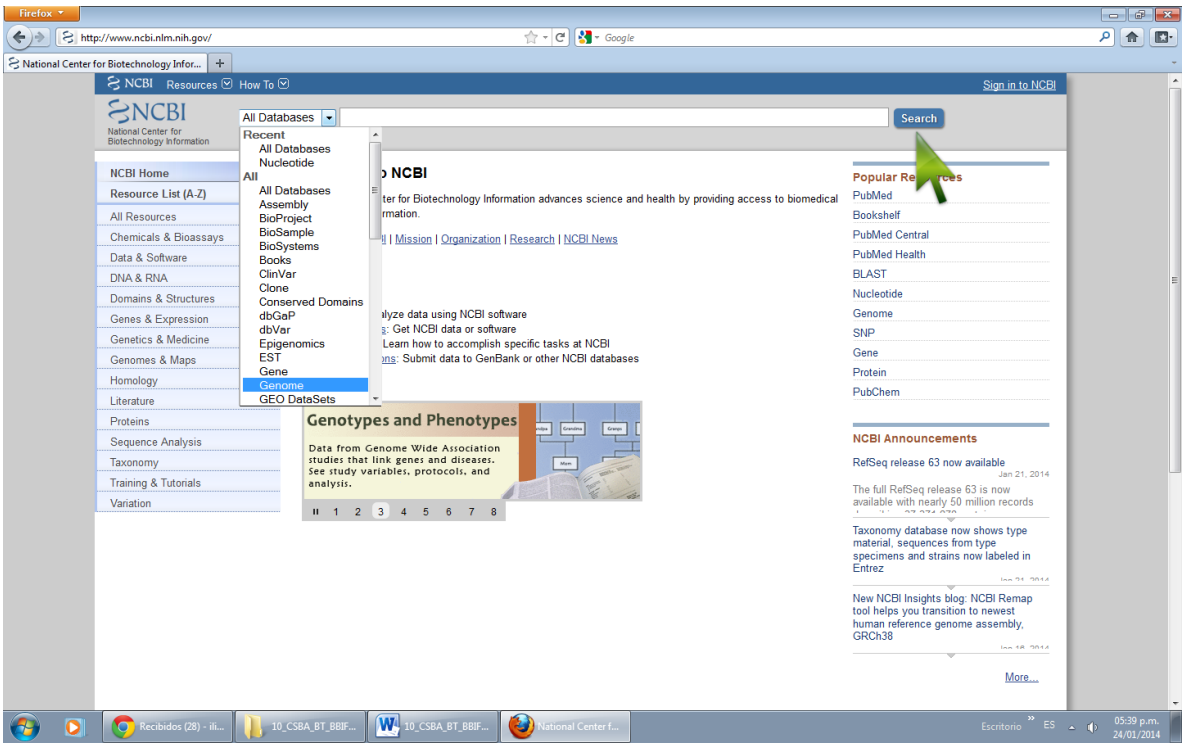




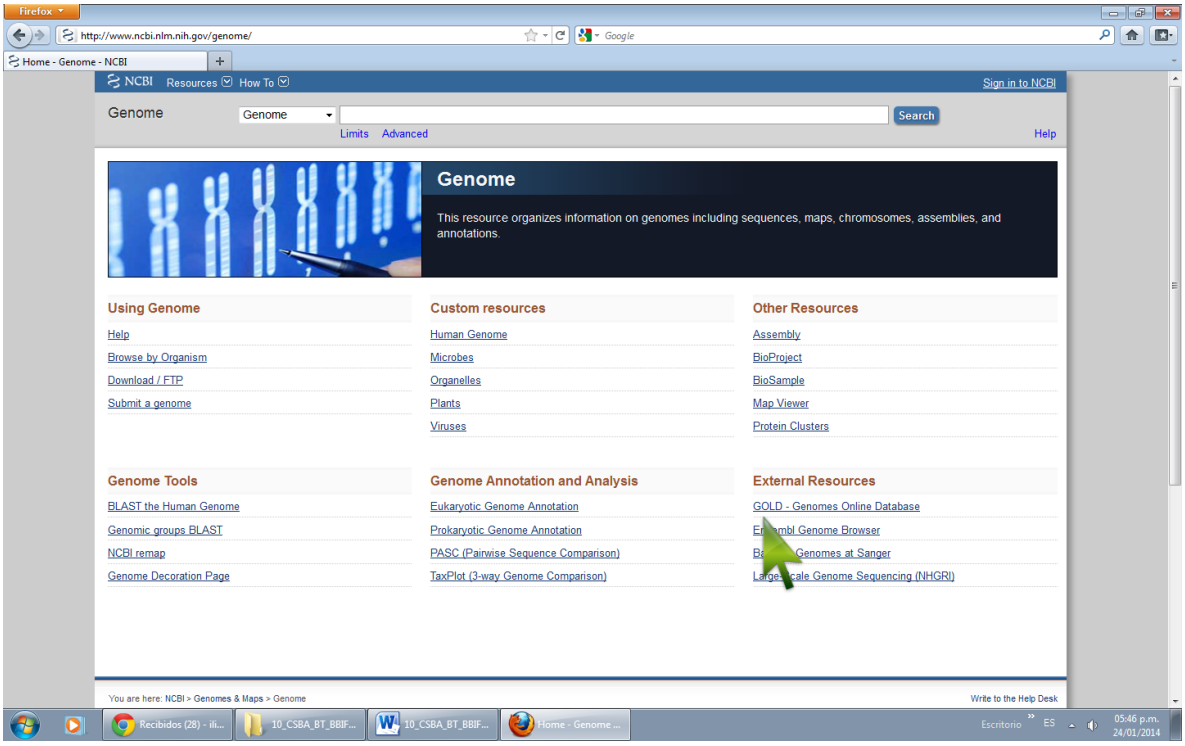
Guarda tu secuencia de aminoácidos de la misma forma que lo hiciste para la de nucleótidos, ¿verdad que es una tarea fácil? ¡Enhorabuena por la secuencia de aminoácidos que acabas de obtener!

Búsqueda de genomas secuenciados

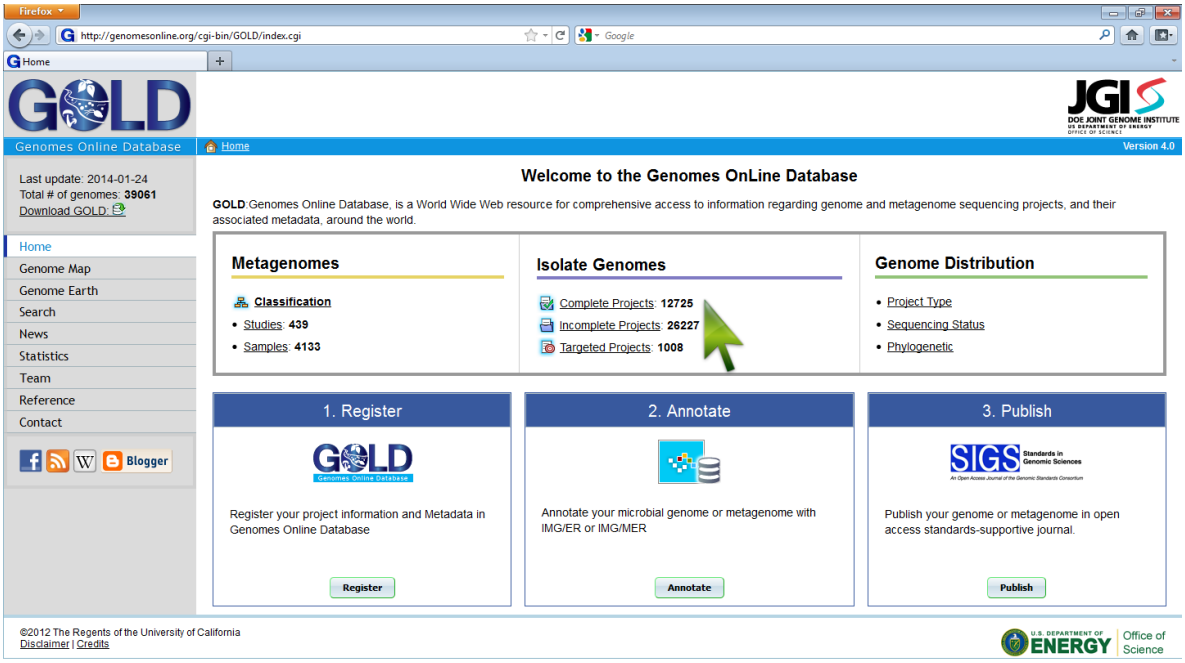
Finalmente, con el objeto de conocer aquéllos organismos para los cuales su genoma ha sido completamente secuenciado, utilizaremos otra de las herramientas del NCBI. Nuevamente, en la página principal, en la pestaña “All Databases”, selecciona la opción Genome.



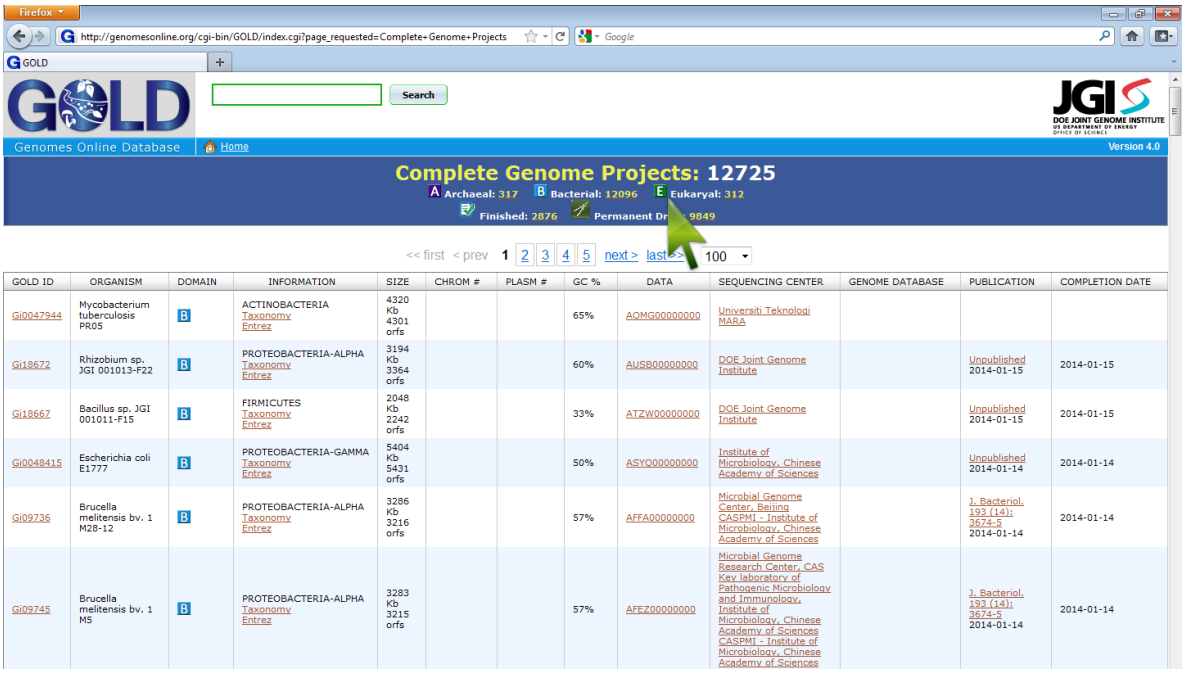
Una vez en esta ventana, en el apartado de “External Resources”, selecciona la opción de GOLD-Genomes Online Database.



Una vez desplegada la información de esta página, en el apartado “Isolate Genomes” presiona “Complete Projects” para poder acceder a todos los proyectos de secuenciación genómica completamente terminados o “Incomplete Projects” para conocer cuáles están en curso.



Si seleccionamos los que ya están completos, obtendremos la siguiente información:



GOLD ID	ORGANISM	DOMAIN	INFORMATION	SIZE	CHROM #	PLASM #	GC %	DATA	SEQUENCING CENTER	GENOME DATABASE	PUBLICATION	COMPLETION DATE
Gi0047944	Mycobacterium tuberculosis PR0S	B	ACTINOBACTERIA Taxonomy Entrez	4320 Kb 4301 orfs			65%	AOMG00000000	Universiti Teknologi MARA			
Gi18672	Rhizobium sp. JGI 001013-F22	B	PROTEOBACTERIA-ALPHA Taxonomy Entrez	3194 Kb 3364 orfs			60%	AUSB00000000	DOE Joint Genome Institute		Unpublished 2014-01-15	2014-01-15
Gi18667	Bacillus sp. JGI 001011-F15	B	FIRMICUTES Taxonomy Entrez	2048 Kb 2242 orfs			33%	ATZW00000000	DOE Joint Genome Institute		Unpublished 2014-01-15	2014-01-15
Gi0048415	Escherichia coli E1777	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez	5404 Kb 5431 orfs			50%	ASYQ00000000	Institute of Microbiology, Chinese Academy of Sciences		Unpublished 2014-01-14	2014-01-14
Gi09736	Brucella melitensis bv. 1 M28-12	B	PROTEOBACTERIA-ALPHA Taxonomy Entrez	3286 Kb 3216 orfs			57%	AFFA00000000	Microbial Genome Center, Beijing CASPMI - Institute of Microbiology, Chinese Academy of Sciences		J. Bacteriol. 193 (14): 3674-5 2014-01-14	2014-01-14
Gi09745	Brucella melitensis bv. 1 M5	B	PROTEOBACTERIA-ALPHA Taxonomy Entrez	3283 Kb 3215 orfs			57%	AFEZ00000000	Microbial Genome Research Center, CAS Key laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences CASPMI - Institute of Microbiology, Chinese Academy of Sciences		J. Bacteriol. 193 (14): 3674-5 2014-01-14	2014-01-14

La catalogación de los organismos con genomas secuenciados se encuentra de acuerdo con su taxonomía, es decir, Archea, Eubacteria o Eukarya. Se muestra la lista de los organismos cuyo genoma se ha completado, incluyendo algunas características

relevantes como el tamaño del genoma, el número de cromosomas, el lugar de secuenciación, entre otros.

Video explicativo para usar otra de las bases de datos del NCBI, el PubMed, que contiene información bibliográfica.

http://www.youtube.com/watch?v=2FDjQ6vuARg

Actividades

La elaboración de las actividades estará guiada por tu figura académica, mismo que te indicará, a través de la Planificación de actividades, la dinámica que tú y tus compañeros (as) llevarán a cabo, así como los envíos que tendrán que realizar.

Para el envío de tus trabajos usarás la siguiente nomenclatura: BIIN_U1_A1_XXYZ, donde BIIN corresponde a las siglas de la asignatura, U1 es la etapa de conocimiento, A1 es el número de actividad, el cual debes sustituir considerando la actividad que se realices, XX son las primeras letras de tu nombre, Y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno.

Autorreflexiones

Para la parte de **autorreflexiones** debes responder las *Preguntas de Autorreflexión* indicadas por tu figura académica y enviar tu archivo. Cabe recordar que esta actividad tiene una ponderación del 10% de tu evaluación.

Para el envío de tu autorreflexión utiliza la siguiente nomenclatura:

BIIN_U1_ATR _XXYZ, donde BIIN corresponde a las siglas de la asignatura, U1 es la unidad de conocimiento, XX son las primeras letras de tu nombre, y la primera letra de tu apellido paterno y Z la primera letra de tu apellido materno

Cierre de la Unidad

Durante la presente Unidad te adentraste en el contexto histórico mediante el cual la Bioinformática fue desarrollándose.

Primeramente, se incluyeron procesos adicionales al ya conocido dogma central de la biología molecular, con el fin de establecer el flujo de la información genética en su sentido más amplio.

Posteriormente se hizo una breve descripción de las tecnologías de secuenciación, desde la descrita por Sanger hace ya más de 30 años, hasta las más novedosas técnicas de la actualidad. La secuenciación Sanger constituyó el caballito de batalla para los esfuerzos genómicos iniciales y hoy en día, gracias al avance de la tecnología, es posible acceder a técnicas de secuenciación de segunda generación, las cuales permiten secuenciar una enorme cantidad de fragmentos de ADN de manera eficiente en un tiempo relativamente corto y a menores costos.

Finalmente, utilizaste la base de datos del NCBI, con el fin de aprender a utilizarla y entender el tipo de información que puedes obtener. Así, buscaste y accediste a una secuencia de nucleótidos y a otra de aminoácidos, además accediste a la lista de aquéllos organismos cuyos genomas han sido completamente secuenciados, así como de aquéllos que están en proceso de secuenciación.

Toda esta información te será de gran utilidad para las siguientes Unidades, en las cuales seguirás buscando secuencias para ahora sí poder analizarlas, ¡felicidades por terminar esta primera Unidad!

Para saber más

Video explicativo sobre la técnica de secuenciación de Sanger:
http://www.dnalc.org/files/swfs/animationlib/sangerseq.exe

Fuentes de consulta

- Baxevanis, A. E. & Ouellette, B. F. F. (Eds.). (2001). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Second Edition*. New York: John Wiley & Sons, Inc.

- Fleischmann, R. D., Adams, M. D., White, O., et al. (1995). *Whole-genome random sequencing and assembly of* Haemophilus influenzae *Rd*. *Science* 269(5223):496-512.

- Kandpal, R. P., Saviola, B., Felton, J. (2009). *The era of 'omics unlimited. Biotechniques* 46(5):351-355.

- Krebs, J. E., Goldstein, E. S., Kilpatrick, S. T. (2010). *Lewin's essential genes.* Second Edition. Massachusetts: Jones and Bartlett Publishiers.

- Lemey, P., Salemi, M., Vandamme A. M. (Eds.). (2009). *The Phylogenetic Handbook. A practical Approach to Phylogenetic Analysis and Hypothesis testing*. Second Edition. Cambridge.

- Lesk, A. M. (2002). *Introduction to Bioinformatics.* New York: Oxford University Press.

- Mardis, E. R. (2007). *The impact of next-generation sequencing technology on genetics*. *Trends in Genetics* 24(3):133-141.

- Nelson, D. L. & Cox, M. M. (2005). *Lehninger. Principles of Biochemistry*. Fourth Edition. New York: W.H. Freeman and Company.

- Ondarza, R. N. (1994). *Biología molecular. Antes y después de la doble hélice*. México: Siglo veintiuno editores.

- Polanski, A. & Kimmel, M. (2007). *Bioinformatics*. New York: Springer.

- Posada, D. (Ed.). (2009). *Bioinformatics for DNA sequence analysis.* New York: Humana Press.

- Reis-Filho, J. S. (2009). *Next-generation sequencing. Breast Cancer Research* 11(Suppl 3):S12 (doi: 10.1186/bcr2431).

- Sanger, F.,Nicklen, S.,Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors. Proceeding of the Natural Academy of Science U S A* 74(12):5463-5467.

- Schneider, M. V. & Orchard, S. Chapter 1. *Omics Technologies, Data and Bioinformatics Principles,* en Mayer, B. (Ed.). (2011). Bioinformatics for Omics Data: Methods and Protocols. Springer Science.